Incorporating Prior Knowledge on Speech Production Mechanism into Neural Speech Waveform Generation

Department of Intelligent Systems
Graduate School of Informatics
Nagoya University

Yi-Chiao Wu

Contents

Abstract

Speech generation techniques including text-to-speech (TTS), speech enhancement, and voice conversion ones are widely applied to current daily applications such as a personal mobile assistant and car navigation. The naturalness of synthesized speech and the flexibility of acoustic controllability are the main challenges of speech generation. That is, high-fidelity synthesized speech sounding like natural speech and speech components being flexibly manipulated are important to a speech generation system. A speaker voice conversion (VC) task is adopted in this thesis as a paradigm of speech generation systems, and the VC task involves converting the speaker identity of input speech to a specific target speaker while keeping the same speech content. A general VC system is composed of analysis, manipulation, and synthesis modules, and the thesis focuses on improving the synthesis module using prior knowledge of speech.

A baseline VC system for the non-parallel VC (SPOKE) task of voice conversion challenge 2018 (VCC2018) has been established in this study. The analysis module of the VC system parameterizes speech into spectral and prosodic features using the WORLD vocoder, which is a conventional source–filter-based vocoder. Since the training corpus is non-parallel, the speech contents of the source and target utterances of the SPOKE task are different. A two-stage spectral conversion model with TTS-generated reference speech has been adopted to map the non-parallel source and target utterances. In contrast to conventional VC systems, the baseline system replaces the

synthesizer of the conventional vocoder with a neural-based speech generation model, WaveNet (WN). The WN as a vocoder directly transfers the converted acoustic features to speech waveforms without many ad hoc designs of speech production imposed on the conventional vocoder. Both the internal and external evaluation results in this thesis show the better performance of the WN vocoder than the conventional vocoder.

However, because of the data-driven nature, generic network architecture, and lack of speech-related prior knowledge, the WN vocoder sometimes generates unexpected outputs such as non-speechlike noise while the input acoustic features are unseen or distorted. To avoid the collapsed speech problem of the WN vocoder, a collapsed speech detection and suppression approach has been studied in this thesis. The method is based on the prior knowledge of speech continuity and the stability of conventional vocoders. Specifically, although the naturalness of the WORLD-generated speech is worse, the speech is more stable than the WN-generated speech. The proposed detection method segmentally compares the waveform envelope difference between the WORLD- and WN-generated utterances to detect the collapsed speech segments. The WN vocoder regenerates the detected segments with a waveform-based constraint derived from the continuity extracted from the WORLD-generated speech.

On the other hand, because of the implicit pitch modeling of the WN vocoder, the lack of pitch controllability is a problem. That is, regardless of whether the input fundamental frequency feature F_0 is scaled or not in the F_0 range of the training data, the WN vocoder usually cannot generate speech with accurate pitches. To improve the pitch controllability of the WN vocoder, which is an essential vocoder feature, a pitch-dependent dilated convolutional neural network (PDCNN) and a quasi-periodic (QP) structure have been studied in this thesis. The PDCNN introduces the prior periodicity knowledge of speech to the WN vocoder for dynamically adapting the network

architecture according to the input F_0. The QP structure based on the prior knowledge of speech production applies a source–filter-like structure to the WN vocoder for modeling pitch- and spectral-related components. With the PDCNN and QP structure, QPNet has been proposed, which markedly improves the pitch controllability and speech modeling efficiency of the WN vocoder.

Furthermore, to achieve real-time generation, a non-autoregressive neural-based speech generation model, parallel WaveGAN (PWG), with a compact network has also been studied in this thesis. Since the training process and network designs of the PWG are similar to those of the WN, the PWG also suffers from the difficulty of pitch controllability. With the proposed PDCNN and QP structure, the proposed QPPWG also markedly improves the pitch controllability and speech modeling efficiency of the PWG. Because of the direct waveform output, the internal generative mechanisms are easily revealed by the intermediate outputs of the PWG and QPPWG. The visualized results confirm the effectiveness of the QP structure to make the QPPWG like a source–filter model with a unified neural network. That is, the QPPWG simultaneously attains high-fidelity speech generation as the PWG with better tractability and interpretability.

To summarize, the studies show that applying the speech-related prior knowledge to the neural-based speech generation models significantly improves the robustness against the distorted and unseen acoustic features, pitch controllability, and speech modeling efficiency of these speech generation models.

1 Introduction

1.1 Background

Speech generation is a technique used to generate specific speech samples corresponding to given inputs such as text (text-to-speech, TTS) [1–3], noisy speech (speech enhancement, SE) [4–6], and source speaker speech (speaker voice conversion, VC) [7–9]. A general speech synthesis system includes three modules, analysis, manipulation, and synthesis. Specifically, the analysis module parameterizes the given input into a specific representation, and then the manipulation module converts the representation according to the target requirement. The final synthesis module generates the target speech on the basis of the converted representation.

Taking VC as an example, the target is to change the speaker identity of the given source speech to the identity of a specific target speaker while keeping the linguistic content the same. As shown in Fig. 1.1, a conventional VC system first parameterizes the source speech into acoustic features such as spectral and prosodic features. That is, the analysis module disentangles the input speech into several speech components such as pitch and timbre. Second, these source acoustic features are converted to target acoustic features by changing the corresponding speech components. Last, the synthesis module generates the converted speech on the basis of the converted acoustic features.

Different speech generation tasks usually have specific analysis and manipulation

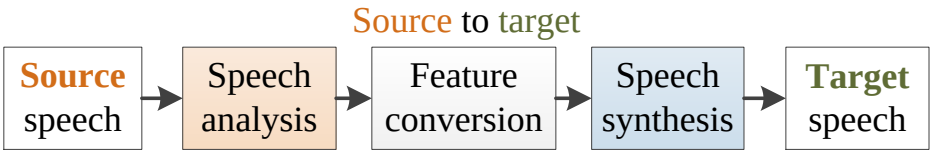


Figure 1.1: *Voice conversion flowchart.*

modules. For example, the analysis and manipulation modules of a TTS system usually transfer the input text into various acoustic features such as a mel-spectrogram [10, 11] corresponding to its synthesis module. However, the synthesis modules of different tasks, which map specific acoustic features to speech waveforms, are usually similar. In my research, I focus on improving the synthesis module. There are several common challenges for arbitrary synthesis modules, and four main challenges are explored in this thesis. The details are described in the next section.

1.2 Thesis Scope

A useful synthesis module is usually general for arbitrary front-end modules, which may involve various speech component transformations such as spectral and pitch conversions. In this thesis, I explore techniques of tackling speech generation with distorted acoustic features and transformed pitch, and the aim of this thesis is to develop robust synthesis modules with high-fidelity generated speech for different speech manipulation tasks. Specifically, to explore speech generation techniques, a non-parallel VC application is first adopted. In this thesis, I start with building a baseline VC system for the non-parallel VC task in voice conversion challenge 2018 (VCC2018) and continually improve the synthesis module of the baseline VC system. Furthermore, a pitch transformation scenario is also explored. Although the research is applied to VC and pitch-transform systems, the proposed methods are generally applicable to other speech

generation tasks. That is, the thesis explores four fundamental challenges of speech generation. First, the quality of the generated speech is usually important. Many of speech generation research studies focus on improving the naturalness of the generated speech and making the generated speech indistinguishable from natural speech as much as possible. Second, the robustness of distorted input is also important for the synthesis module. Since the input representations of the synthesis module are predicted by the manipulation module, the converted representations usually include some distortion. Third, the controllability of speech components is an essential feature of the synthesis module. Last, the efficiency of generation such as real-time generation is necessary for many practical speech synthesis applications.

To tackle these challenges, the main concept of this thesis is applying speech-related prior knowledge in speech generation systems. Most of the state-of-the-art speech generation models [12–26] focus on pure data-driven and end-to-end networks without many ad hoc assumptions of the speech generation process. The advantage of adopting these models is that they prevent networks from suffering quality degradation of some oversimplified designs imposed on them. However, the disadvantage of these models is that the generated speech sometimes includes some significant errors, which are obviously non-speechlike, but it is difficult to directly adjust the models to fix the errors because of their data-driven nature without explicitly controlling the speech components. Moreover, the lack of speech controllability also degrades the flexibility of these speech generation models for speech synthesis. However, since speech is a sequential signal with long-term correlations, there are many specific characteristics of speech signals such as continuity and periodicity. These specific characteristics are speech-related prior knowledge, which can advance the speech modeling ability of the speech generation models and prevent the generated signals from violating the inherent

patterns of speech [27, 28].

Furthermore, although signal-processing-based conventional speech generation models and codecs [29–37] usually suffer from naturalness degradation because of many oversimplified speech modeling mechanisms, the underlying knowledge and empirical designs are still valuable for learning more about human speech production mechanisms. In this thesis, speech modeling mechanisms based on conventional speech modeling techniques are adopted to advance neural-based speech generation models. The proposed architectures [38–42] based on the knowledge of speech production mechanisms make the neural-based speech generation models more tractable and interpretable. Because of the improved tractability and interpretability, the proposed models also realize speech components with higher controllability, which markedly improves the flexibility of the speech generation models.

To summarize, a basic VC system is adopted in this thesis to explore the quality, robustness, controllability, and generation efficiency challenges of speech generation. Specifically, a basic VC system [43] for the VCC2018 non-parallel VC task is first introduced. Both a popular conventional speech generative technique [37] and a state-of-the-art neural-based speech generation model [12] have been combined with the VC system to show the significant improvement of the neural-based speech generation model. However, although the neural-based speech generation model usually achieves high-fidelity speech generation, the robustness of these models against unseen acoustic features such as the distorted acoustic features from the conversion model is insufficient. A postprocessing method [27, 28] based on the prior knowledge of speech continuity is introduced to prevent the neural-based speech generation model from generating unexpected noise while being conditioned on the VC acoustic features. Furthermore, the insufficient pitch controllability of the neural-based speech generation model is

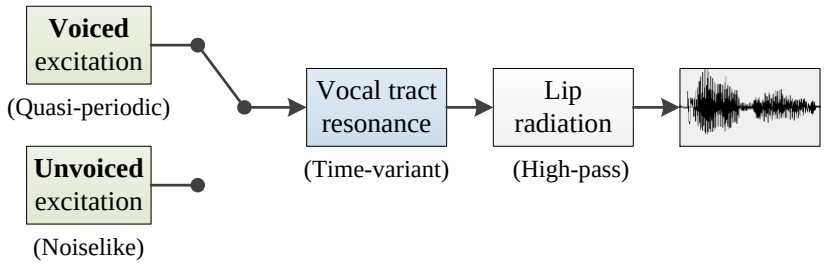


Figure 1.2: *source–filter model.*

also tackled using a novel network [38, 39], which introduces the prior knowledge of speech periodicity into the neural-based speech generation model. Last, the real-time generation [41, 42] has also been explored in this thesis. More details of the speech generation challenges and the corresponding work in this thesis are as follows.

1.2.1 Quality

The basic technique applied in speech analysis and synthesis is the use of a voice coder, vocoder [29–31]. A vocoder includes an analyzer to parameterize speech into specific representations (acoustic features) for different speech components, such as spectral and prosodic components, and a synthesizer to generate speech on the basis of these representations. One of the most general speech modeling techniques for vocoders is the source–filter model [34]. As shown in Fig. 1.2, speech production is formulated as a convolution of an excitation (source) signal and a spectral filter with a high-pass filter in the final stage. The excitation signal models vocal fold vibration, the spectral filter models vocal tract resonance, and the high-pass filter models lip radiation. Specifically, speech includes voiced and unvoiced sounds, the excitation signal of a voiced sound is quasi-periodic and the excitation signal of an unvoiced sound is similar to that of white noise. The fundamental frequency (F_0) of the voiced sound is pitch, and the F_0 of the

unvoiced sound is set to zero. The spectral filter is time-variant, which is assumed to be pitch-independent and timbre-dependent; however, it is difficult to achieve complete pitch independence in a practical vocoder. Because of the time-invariant characteristics of the lip radiation, the lip radiation modeling is usually integrated into a time-variant spectral envelope parameter.

Conventional vocoders such as STRAIGHT [35, 36] and WORLD [37], which are adopted in many VC works, usually model vocal fold vibrations on the basis of F_0 and the aperiodicity feature (ap) and model vocal tract resonance on the basis of the spectral envelope feature (sp). However, because many oversimplified assumptions of speech production such as a fixed length of the analysis window, a time-invariant linear filter, and a stationary Gaussian process are imposed on these conventional vocoders, phase information and temporal details are lost during the analysis and synthesis processes. This loss of phase information and temporal details causes significant naturalness degradation of the generated speech such as buzzy noise. Therefore, many neural-based speech generation models [12–26] have been proposed to directly model speech waveforms without many ad hoc assumptions imposed on their speech production mechanisms. Moreover, integrating these advanced neural-based speech generation models into a VC system [43–46] also results in significant improvements. That is, the conventional vocoders of the speech synthesis module are replaced with these neural-based speech generation models to greatly recover the lost phase information and temporal details. Since these neural-based speech generation models generate speech condition on the acoustic features from the analysis and manipulation modules, these generative models are called neural vocoders [47–50].

In this thesis, the quality of the WaveNet vocoder [47, 48] is first compared with that of the conventional vocoder WORLD for our baseline VC system to show the

effectiveness of neural vocoders. The quality of the pitch-transformed speech of the WaveNet and WORLD vocoders is also presented. Furthermore, another state-of-the-art neural vocoder parallel WaveGAN [24] is also compared with the WORLD vocoder for pitch transformation.

1.2.2 Robustness

Since the analysis and manipulation modules of a speech synthesis system are imperfect, the converted or predicted acoustic features usually suffer from distortions and prediction errors. For instance, the converted acoustic features of conventional statistical VC models usually suffer from the oversmoothing problem [9] because of their statistical nature. Moreover, because of insufficient training data, these conversion models sometimes generate distorted acoustic features. Since the prediction and conversion errors from the analysis and manipulation modules will propagate to the synthesis module, these distorted acoustic features usually cause significant naturalness degradation of the synthesis module. Therefore, the robustness of the neural vocoders in a VC system against the distorted acoustic features is important.

Although training the neural vocoders with these distorted acoustic features will ease the oversmoothing problem [44, 51–53], the prediction and conversion errors still make the neural vocoders generate unexpected speech samples such as those with marked discontinuity [27, 28]. Since speech is a sequential signal with strong continuity and these neural vocoders directly model speech waveforms, even a few prediction errors of the neural vocoders will cause significant perceptually quality degradation. As a result, a waveform-based constraint is introduced in the models. Specifically, the proposed technique takes advantage of the prior knowledge of speech continuity to derive a constraint making the output waveform samples of the neural vocoders follow the prior

speech continuity. With the waveform-based constraint, the WaveNet vocoder of our VC system greatly eases the discontinuity problem and generates a more stable speech.

1.2.3 Controllability

To prevent the degradation of these neural vocoders caused by many ad hoc assumptions of speech production, these neural vocoders are usually a unified neural network trained in a data-driven manner to directly model the transformation from acoustic features to speech waveforms. Although these neural vocoders usually achieve high-fidelity speech generation, the insufficient speech component controllability of these neural vocoders is a problem. That is, since these neural vocoders model the relationships between acoustic features and speech waveforms with a multilayer complicated nonlinear mapping function, it is difficult to directly control specific speech components by adjusting the input acoustic features, especially when the changed acoustic features are unseen data. For example, when the input F_0 is outside the observed F_0 training range of the training data or the input F_0 and other features are an unseen combination, it is difficult for the WaveNet vocoder to generate speech with accurate pitch and good quality [38, 39]. That is, the WaveNet vocoder lacks pitch controllability, which is an essential feature of a vocoder.

Not only the WaveNet vocoder but also the parallel WaveGAN model has an insufficient pitch controllability problem, so the proposed pitch-dependent convolutional neural network (PDCNN) and quasi-periodic (QP) structure have been applied to them. Specifically, since speech is a quasi-periodic signal, it is reasonable to adjust the neural networks based on the input F_0. The proposed PDCNN is a pitch-adaptive network, whose network architecture is dynamically changed according to the input F_0. The proposed QP structure is a cascaded network architecture that captures the hier-

archical information of speech signals. The adaptive subnetwork of the QP structure adopts PDCNNs to model pitch-related information with long-term dependence, and the fixed subnetwork of the QP structure adopts dilated convolutional neural networks (DCNNs), which are also adopted in the WaveNet and parallel WaveGAN models, to model spectral-related information with short-term dependence. With the PDCNN and QP structure, the quasi-periodic WaveNet and parallel WaveGAN attain higher pitch controllability and more efficient speech modeling because of the introduced prior knowledge of speech periodicity by the PDCNN and QP structure. To summarize, this thesis focuses on improving the pitch controllability of data-driven and unified neural vocoders by introducing the prior speech periodicity knowledge to the neural networks.

1.2.4 Generation Efficiency

For practical speech synthesis systems, real-time and streaming generations are important. Specifically, the generation time of a real-time system should be equal to or less than the length of the generated speech, and streaming generation is a more difficult technique to continuously generate speech according to the streaming inputs in a very lowlatency manner. Since speech is a real-time communication medium, and the conversation is continuous, real-time generation is the minimum requirement in most speech generation scenarios. Moreover, since a conversation continues in a back and forth manner, streaming speech generation with very low latency is preferred.

I start with a high-quality but extremely slow generative model, WaveNet [12], and then adopt a batch-type generative model, parallel WaveGAN [24], simultaneously generating all samples in one utterance to improve generation efficiency. Specifically, autoregression and the huge network architecture make the generation speed of the WaveNet vocoder extremely low. Even with the proposed PDCNN and QP structure

to improve speech modeling efficiency resulting in a 50 % model size, the generation speed of the quasi-periodic WaveNet [38,39] is still far away from real-time generation. Therefore, the non-autoregressive parallel WaveGAN model with a compact network size is also adopted in this thesis with the proposed PDCNN and QP structure [41,42]. The advantage of non-autoregression is that the network can simultaneously generate all samples, which significantly improves the generation speed. Only the batch-type real-time generation is explored in this thesis, and the streaming generation remains for future work.

1.3 Thesis Overview

High-quality speech generation models for VC and pitch transformation applications are the main topics of this thesis, and the thesis overview is shown in Fig. 1.3. Specifically, a speech generation system usually includes three modules, analysis, manipulation, and synthesis, and this thesis focuses on improving the synthesis module for VC and pitch transformation scenarios. For the VC scenario, a basic VC system submitted to the non-parallel VC task of VCC2018 is taken as the baseline system of this thesis. For the pitch transformation scenario, controlling the pitch of the generated speech on the basis of the input F_0 is explored.

Four fundamental challenges of the synthesis module, quality, robustness, controllability, and generation efficiency, are explored. For quality, two neural-based speech generation models, WaveNet and parallel WaveGAN, are taken as vocoders to generate a high-fidelity speech on the basis of acoustic features, and the fundamentals for these neural vocoders are introduced in Chapter 2. In Chapter 3, the early success of the baseline VC system combining a basic VC module with the WaveNet vocoder in VCC2018 is shown. In Chapters 4, 5, and 6, WaveNet, parallel WaveGAN, and

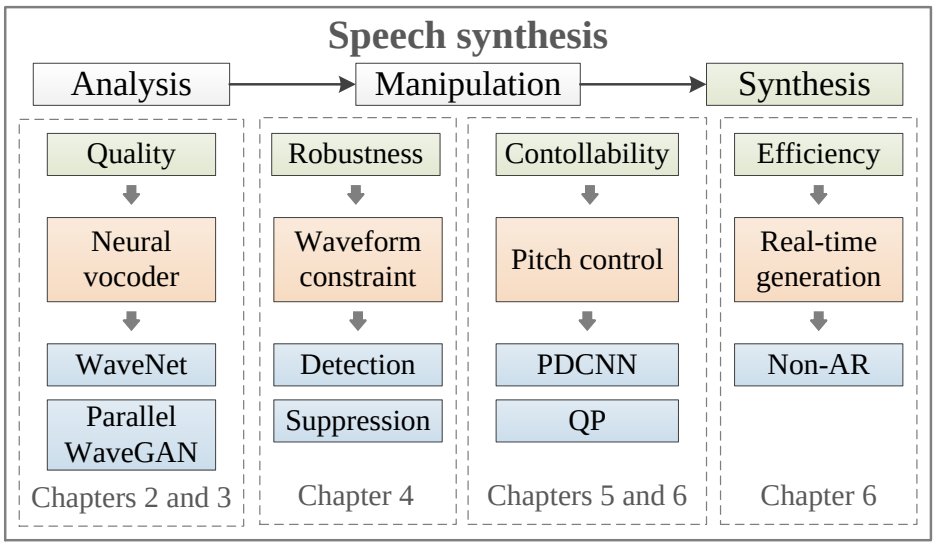


Figure 1.3: *Thesis overview.*

the conventional vocoder WORLD are compared to show the effectiveness of the neural vocoders. For robustness, a waveform-based constraint, which includes abnormal speech detection and suppression modules, for the WaveNet vocoder is presented in this thesis. In Chapter 4, the proposed waveform-based constraint is applied to the WaveNet vocoder of the baseline VC system, and the obtained subjective results show speech quality improvements achieved with the proposed method dealing with the distorted VC acoustic features.

For the controllability of speech components, this thesis focuses on pitch controllability, and a pitch-adaptive network (PDCNN) and a cascaded network structure (QP structure) are described. The PDCNN and QP structure are applied to the WaveNet and parallel WaveGAN vocoders to improve their pitch controllability and speech modeling efficiency. In Chapters 5 and 6, both objective and subjective results obtained show the higher pitch accuracy of the utterances generated by the QP WaveNet (QPNet) and QP parallel WaveGAN (QPPWG) vocoders in the pitch trans-

formation scenario while keeping the speech quality similar with smaller model sizes. For the generation efficiency, the parallel WaveGAN, which is a non-autoregressive model with a compact network size, is explored in this thesis for real-time generation. The real-time factor (RTF) results in Chapter 6 show that the proposed QPPWG achieves real-time generation on both GPU and CPU. To summarize, the high-fidelity and real-time speech generation techniques combining VC and pitch transformation applications have been explored in this thesis. With the proposed waveform-based constraints, the speech quality of the baseline VC system has been improved by the more robust WaveNet vocoder. With the proposed QP structure, the pitch controllability of the WaveNet and parallel WaveGAN vocoders has been enhanced.

This thesis is organized as follows. The related work is reviewed in Chapter 2, and the baseline VC system is introduced in Chapter 3. The proposed waveform-based constraint for the WaveNet vocoder is described in Chapter 4. The proposed PDCNN and QP structure are applied to the WaveNet in Chapter 5 and the parallel WaveGAN in Chapter 6. Lastly, the summary and future work are presented in Chapter 7.

2 Related Work

The main technique involved in this thesis is a vocoder with speaker voice conversion (VC). The proposed methods in the following chapters focus on improving the quality, robustness, and controllability of neural-based vocoders for VC and pitch transformation scenarios. In this chapter, the review from conventional parametric-based vocoders to recent neural-based vocoders is first presented. There are three main topics of the vocoder review. First, since this thesis focuses on incorporating prior knowledge of speech production into speech generation models, a review of speech production mechanisms and speech modeling is introduced. Second, the baseline neural vocoders, WaveNet and parallel WaveGAN, are introduced. Third, the controllability of each speech component is an essential feature of a vocoder, and I will review the speech manipulation of conventional vocoders. I will also discuss the fundamental techniques of VC and introduce two neural-based VC models adopted in this thesis.

2.1 Vocoder

A vocoder [29–31] models the human vocal system, which is composed of the vocal fold, vocal tract (the space between the vocal fold and the lips) couping with the nasal tract, and the lips. As shown in Fig. 2.1, a general speech production flow includes three main stages. First, an excitation signal is generated. Second, the generated excitation signal is modulated by the resonance of the vocal and nasal tracts. Third,

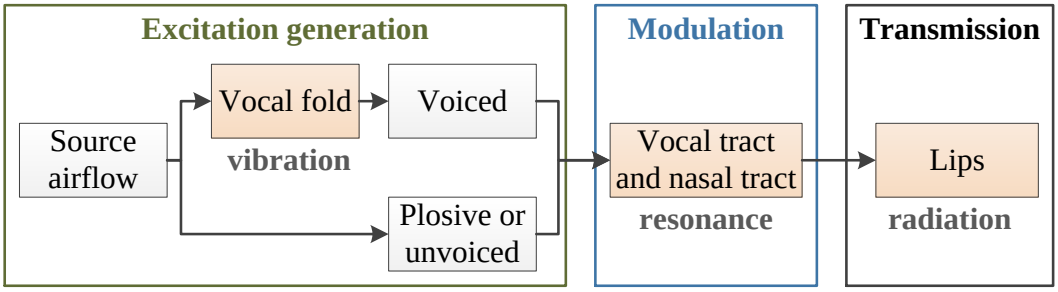


Figure 2.1: *Human vocal system.*

the modulated signal is transferred by the lips. Specifically, the speech includes three categories, voiced, unvoiced, and plosive, corresponding to different types of excitation signal [54]. A source airflow is first generated by the lungs, bronchi, and trachea and then transformed into different excitation signals through different mechanisms. The excitation signal of a voiced sound has a quasi-periodic waveform, which is due to the air vibration produced by the vocal fold vibrations and source airflow. The excitation signals of the unvoiced and plosive sounds are noiselike waveforms, which are the turbulence and burst of the airflow respectively produced by the constriction and closure of specific points along the vocal tract without vocal fold movements. The vocoder techniques introduced in this chapter can be divided into two categories, source–filter and unified vocoders, and the details are as follows.

2.1.1 Source–filter Vocoder

For a discrete-time digital system, one of the most general speech modeling techniques is the source–filter model [34]. The excitation signal is represented as a digital signal, and the spectral properties of vocal and nasal tracts resonances and lip radiation are represented as a digital filter. The digital source signal excites the digital filter to generate speech signals. The spectral envelope of the speech signal can be estimated by linear

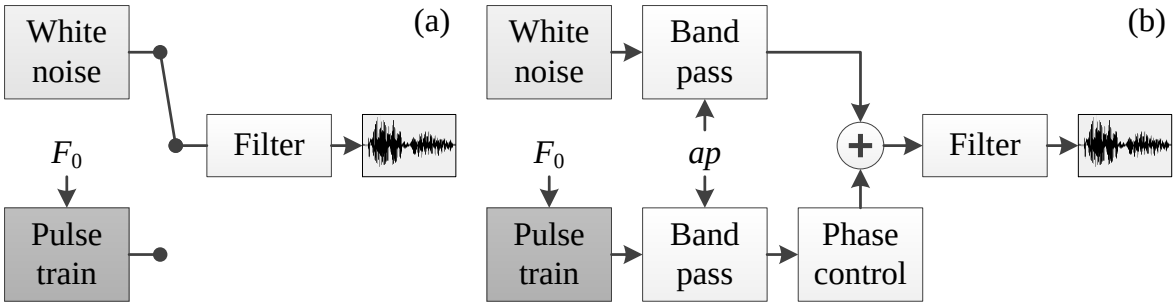


Figure 2.2: *(a) Simple vocoder; (b) STRAIGHT vocoder.*

prediction coding (LPC) or using a mel-generalized cepstrum (MGC) as in the case of LPC [55, 56] and MGC [57, 58] vocoders. As shown in Fig. 2.2 (a), the simplest way to model the excitation signal is by using a pulse train and white noise, and the pulse train is generated according to the fundamental frequency (F_0) of the speech signal. However, the simplest vocoders usually suffer from "buzzy" noise caused by unnatural harmonic components and "hissy" noise caused by missed harmonic components.

To improve the excitation signal modeling, many advanced methods such as mixed excitation [8, 35–37, 59–61] and glottal source modeling [62, 63] have been proposed. For speech generation applications such as text-to-speech (TTS) and voice conversion (VC), the STRAIGHT [35, 36] (Fig. 2.2 (b)) and WORLD [37] vocoders are two of the most popular conventional parametric vocoders. Both STRAIGHT and WORLD vocoders are multiband mixed excitation vocoders, which model excitation signals using the Gaussian noise, F_0, and subband aperiodicity (ap) of the speech signal. Although the high flexibility and tractability of STRAIGHT and WORLD make these vocoders applicable to many TTS and VC systems, the imperfect excitation signal modeling still causes significant naturalness degradation.

Owing to the thriving development of neural networks (NNs), many source–filter-based neural vocoders shown in Fig. 2.3 have been proposed to improve the naturalness

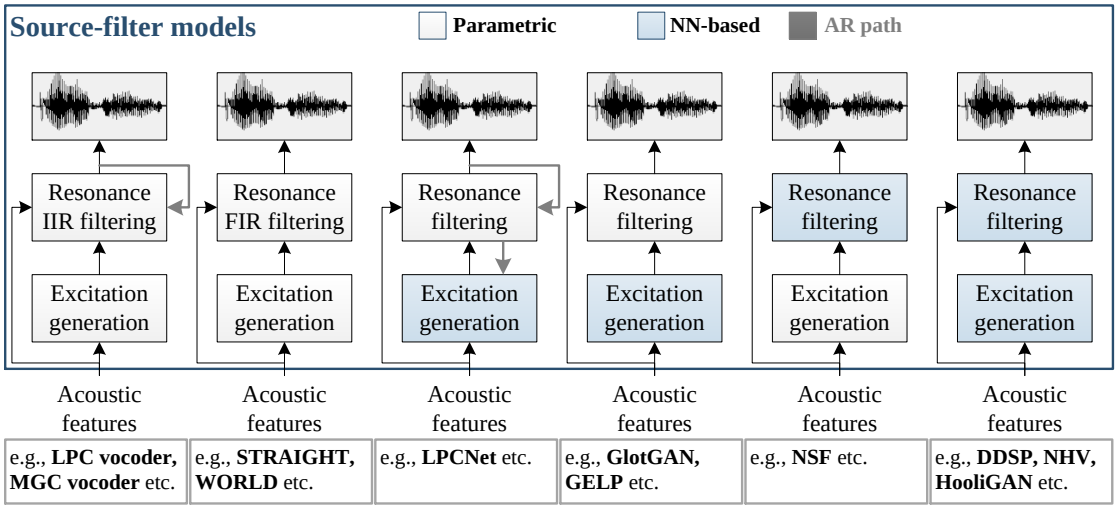


Figure 2.3: *Source–filter vocoder.*

of vocoder-generated speech. For instance, to generate a better excitation signal, LPCNet [16] adopts a recurrent neural network (RNN) to model the LPC residual signal in an autoregressive (AR) manner. GlotGAN [64, 65] and GELP [66] adopt non-AR convolutional neural networks (CNNs) with a generative adversarial network (GAN) [67] structure to generate glottal source signals. Furthermore, to advance the spectral filtering, the authors of [68] and [69] also proposed a neural source–filter (NSF) network to adopt an advanced CNN-based neural filter. The authors of [70] also proposed another GAN-based vocoder with tailored periodic and aperiodic inputs, and the model was trained with the GAN loss of the generated waveform and the Gaussian loss of its aperiodic components. Recently, on the basis of the neural excitation generation of differentiable digital signal processing (DDSP) [71] and the neural spectral filtering of NSF, completely differentiable source–filter vocoders with a GAN structure such as neural homomorphic vocoder (NHV) [72] and HooliGAN [73] have also been proposed.

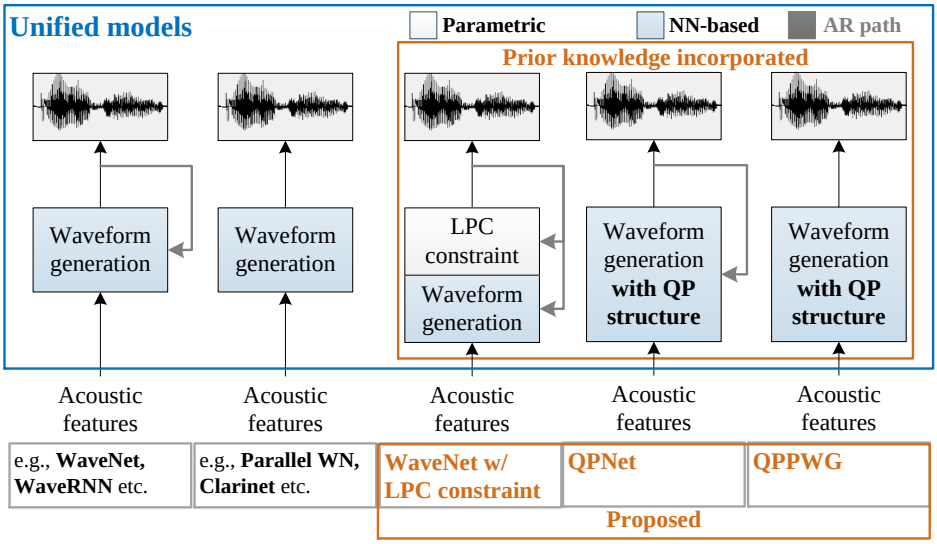


Figure 2.4: *Unified vocoder.*

2.1.2 Unified Vocoder

On the other hand, to avoid many ad hoc assumptions of speech production, many NN-based unimodal vocoders have been proposed. As shown in Fig. 2.4, in contrast to the source–filter-based vocoders, the unified vocoders directly model the relationships among speech waveform samples. Specifically, AR models such as CNN-based WaveNet (WN) [12] and RNN-based SampleRNN [13] achieve high-fidelity speech generation by modeling the probability distribution of each speech sample with the given auxiliary features and previous samples. Taking conventional-vocoder-extracted acoustic features as the auxiliary features for the unified vocoders [47–50, 74], which replace the synthesizer of the conventional vocoders, also achieved early success. However, the AR mechanism and huge network architectures of WN and SampleRNN result in a slow generation speed. To overcome these problems, many compact AR models with specific knowledge [14, 15] and non-AR models such as flow-based [17–22] and generative adversarial network (GAN)-based [23–26] models have been proposed.

Although these unified vocoders achieve high-fidelity speech generation without many ad hoc assumptions of speech production, the data-driven nature, the generic network architecture, and the lack of prior speech-related knowledge make most of these models lose their acoustic controllability and robustness against unseen auxiliary features. For instance, the WN vocoder sometimes generates non-speechlike noisy segments when the input is VC acoustic features [43, 44, 46], and it is difficult for the WN vocoder to generate speech with accurate pitches when the input F_0 is outside the F_0 range of training data [38, 39].

As shown in Fig. 2.4, for the robustness of the WN vocoder with distorted acoustic features such as VC features, a waveform-domain LPC constraint based on the prior knowledge of speech continuity [27, 28] is introduced in this thesis. For pitch controllability, although the authors of [68–70] proposed different NN-based models to explicitly control the excitation signal with the input F_0, carefully designed mixed periodic and aperiodic inputs are required. A pitch-adaptive unified vocoder, QPNet [38, 39], which improves the pitch controllability of the WN vocoder without the requirements of specific inputs, is first introduced in this thesis. The QPNet vocoder introduces the prior knowledge of speech periodicity and source–filter modeling into the network by dynamically adapting the network architecture according to the input F_0 and the proposed QP structure. Moreover, to achieve real-time generation, a non-AR unified vocoder, parallel WaveGAN (PWG) [24], is adopted, and the proposed QP structure is also applied to the PWG vocoder (QPPWG) to improve its pitch controllability and speech modeling efficiency [41, 42].

2.1.3 Baseline WaveNet Vocoder

The AR unified WN vocoder is taken as a baseline AR neural vocoder in this thesis. That is, the WN vocoder is adopted as the baseline vocoder of the baseline VC system submitted to VCC2018 in Chapter 3, the collapsed speech detection and suppression technique in Chapter 4, and the QPNet vocoder in Chapter 5. The details of the WN model and WN vocoder are as follows.

WaveNet

Autoregression is adopted in WN to sequentially predict the probability distribution of each waveform sample conditioned on previous samples for modeling the relationships among these audio samples. The conditional probability function is formulated as

$$P(\boldsymbol{x}) = \prod_{t=1}^{T} P(x_t \mid x_{t-1}, \ldots, x_{t-r}), \tag{2.1}$$

where t is the sample index, x_t is the current audio sample, and r is a specific length of previous samples called the receptive field. Instead of the general recurrent structure for AR modeling, WN applies stacked CNNs with a dilated mechanism [75] and a causal structure to model the very long term dependence and causality of audio signals. Since the modeling capability of WN is highly related to the amounts of previous samples considered to predict the current sample, the dilated mechanism is crucial to the efficient extension of the receptive field. Moreover, a categorical distribution is applied to model the conditional probability, whereas audio signals are encoded into 8 bits by using the μ-law algorithm. The categorical distribution is flexible to model an arbitrary distribution of the target speech.

As shown in Fig. 2.5, the data flow of WN is as follows: previous audio samples first pass through a causal layer and several residual blocks with skip connection out-

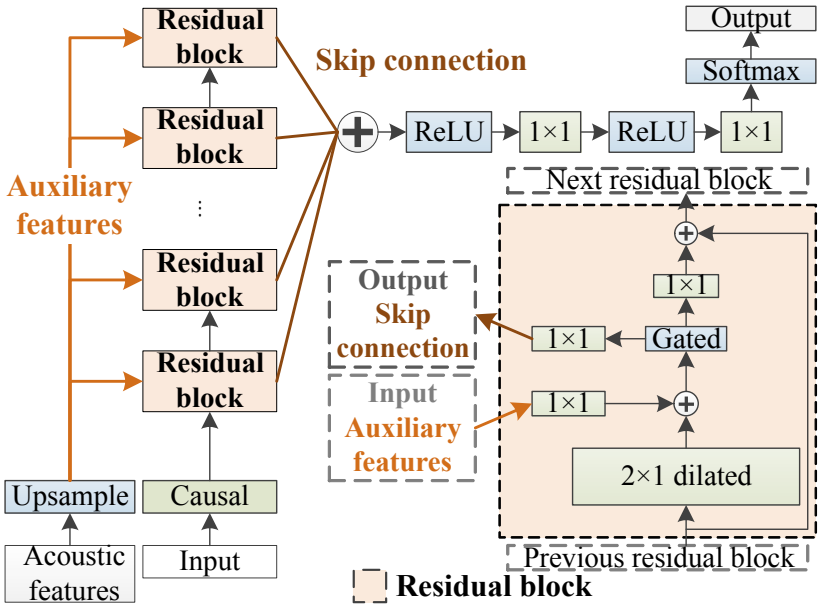


Figure 2.5: *WaveNet vocoder.*

puts, and then the summation of all skip connections is processed by two ReLU [76] activations with 1×1 convolutions and one softmax layer to output the predicted distribution of the current audio sample. Each residual block includes a DCNN layer, a gated structure, a residual connection, and a skip connection output. The gated structure for enhancing the modeling capability of the network is formulated as

$$\boldsymbol{y}^{(\mathrm{o})} = \tanh\left(\boldsymbol{V}_{f,k} * \boldsymbol{y}^{(\mathrm{i})}\right), \odot \sigma\left(\boldsymbol{V}_{g,k} * \boldsymbol{y}^{(\mathrm{i})}\right), \tag{2.2}$$

where $\boldsymbol{y}^{(\mathrm{i})}$ and $\boldsymbol{y}^{(\mathrm{o})}$ are the input and output feature maps of the gated structure, respectively. $\boldsymbol{V}$ is a trainable convolution filter, $*$ is the convolution operator, $\odot$ is an elementwise multiplication operator, σ is a sigmoid function, k is the layer index, and f and g are the filter and gate, respectively.

Furthermore, to guide the WN model to generate desired contents, the vanilla WN

is also conditioned on linguistic and F_0 features. Equation 2.1 is modified as

$$P\left(\boldsymbol{x} \mid \boldsymbol{h}\right) = \prod_{t=1}^{T} P\left(x_t \mid x_{t-1}, \ldots, x_{t-r}, \boldsymbol{h}\right), \tag{2.3}$$

where $\boldsymbol{h}$ is the vector of the auxiliary features (linguistic and F_0 features), and Eq. 2.2 with auxiliary features becomes

$$\boldsymbol{y}^{(\mathrm{o})} = \tanh\left(\boldsymbol{V}_{f,k}^{(1)} * \boldsymbol{y}^{(\mathrm{i})} + \boldsymbol{V}_{f,k}^{(2)} * \boldsymbol{h}'\right) \odot \sigma\left(\boldsymbol{V}_{g,k}^{(1)} * \boldsymbol{y}^{(\mathrm{i})} + \boldsymbol{V}_{g,k}^{(2)} * \boldsymbol{h}'\right), \tag{2.4}$$

where $\boldsymbol{V}^{(1)}$ and $\boldsymbol{V}^{(2)}$ are trainable convolution filters and $\boldsymbol{h}'$ is the temporal extended auxiliary features, whose temporal resolution matches the speech samples.

WaveNet Vocoder

Many speech synthesis systems adopt source–filter-based conventional vocoders such as STRAIGHT [35] and WORLD [37] because of their flexibility and controllability. However, the oversimplified assumptions, such as analysis windows with a fixed length, time-invariant linear filters, and stationary Gaussian processing, imposed on the conventional vocoders make these vocoders lose some essential information of speech such as phase and temporal details causing marked quality degradation. To address this problem, the authors of [47, 48] proposed the WN vocoder to replace the synthesis part of conventional vocoders to synthesize high-fidelity speech on the basis of the prosodic and spectral acoustic features extracted by conventional vocoders. With the WN conditioned on the acoustic feature, most of the lost phase information and temporal details are recovered by the WN vocoder. Furthermore, conditioning WN on the acoustic features greatly reduces the requirements of the amount of training data, and it makes WN more tractable.

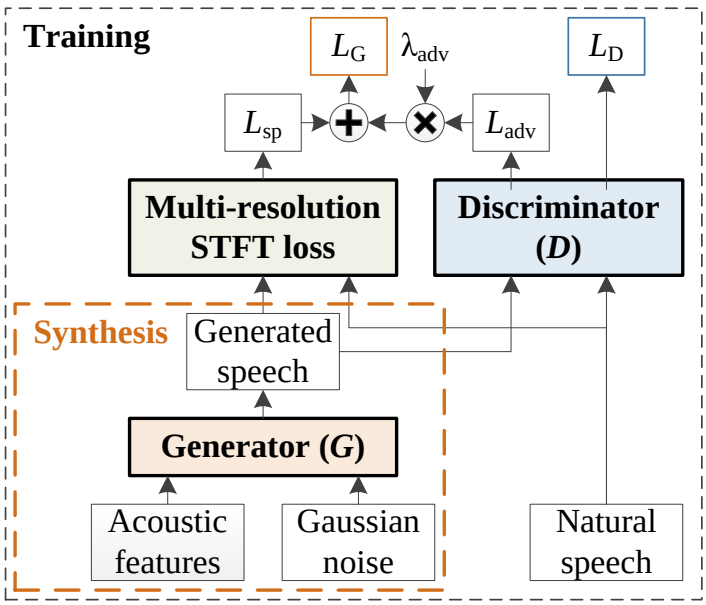


Figure 2.6: *Parallel WaveGAN.*

2.1.4 Baseline Parallel WaveGAN Vocoder

In addition to the WN vocoder, PWG is been taken as a baseline non-AR neural vocoder in this thesis. As shown in Fig. 2.6, PWG includes a classical GAN structure, which consists of CNN-based discriminator (D) and generator (G) modules, and an additional multi-resolution STFT loss module. The details are as follows.

GAN-based Waveform Generation

A WN-like architecture is adopted for the generator of PWG. The main differences between the PWG generator and the WN are a Gaussian noise input instead of previous samples, a raw waveform output instead of a probability distribution, and a non-AR manner. Specifically, the inputs of the generator are a Gaussian noise sequence $\boldsymbol{z}$ and auxiliary acoustic features, and $\boldsymbol{z}$ is taken from a Gaussian distribution with zero mean and standard deviation, denoted as $N(0, I)$. The output of the generator is waveform

samples. The generator, which tries to generate realistic speech samples, is trained in a manner adversarial to the discriminator, which attempts to distinguish natural (*real*) and generated (*fake*) speech waveforms. The adversarial loss of the generator (L_{adv}) is formulated as

$$L_{\text{adv}}(G, D) = \mathbb{E}_{\boldsymbol{z} \in N(0,I)} \left[(1 - D(G(\boldsymbol{z})))^2 \right]. \tag{2.5}$$

Note that all auxiliary features of the generator are omitted in this section for simplicity. Unlike some flow-based models [19, 20], which adopt an invertible network to map real data into the Gaussian noise sequence, the generator of PWG learns to transfer the input noise sequence to the output waveforms via the feedback from the discriminator.

Furthermore, a simple architecture consisting of stacked DCNN layers with the activation function LeakyReLU [77] is adopted for the discriminator of PWG, and the dilation size of each DCNN layer increases exponentially with a base of 2 and the exponent of its layer index. The discriminator is trained to minimize the adversarial loss (L_{D}) formulated as

$$L_{\text{D}}(G, D) = \mathbb{E}_{\boldsymbol{x} \in p_{\text{data}}} \left[(1 - D(\boldsymbol{x}))^2 \right] + \mathbb{E}_{\boldsymbol{z} \in N(0,I)} \left[D(G(\boldsymbol{z}))^2 \right], \tag{2.6}$$

where $\boldsymbol{x}$ denotes the natural samples and p_{data} denotes the data distribution of the natural samples.

Multi-resolution STFT Loss

Since it is difficult to stably train PWG with only adversarial losses, an additional STFT-based loss (L_{sp}) is adopted to improve the stability and efficiency of GAN training. Specifically, a spectral convergence loss (L_{sc}) is formulated as

$$L_{\text{sc}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{\| |\text{STFT}(\boldsymbol{x})| - |\text{STFT}(\hat{\boldsymbol{x}})| \|_F}{\| |\text{STFT}(\boldsymbol{x})| \|_F}, \tag{2.7}$$

and a log STFT magnitude loss (L_{mag}) is formulated as

$$L_{\text{mag}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{N} \left\| \log |\text{STFT}(\boldsymbol{x})| - \log |\text{STFT}(\hat{\boldsymbol{x}})| \right\|_{L1}, \tag{2.8}$$

where $\hat{\boldsymbol{x}}$ denotes the PWG-generated samples, $\|\cdot\|_F$ is the Frobenius norm, $\|\cdot\|_{L1}$ is the $L1$ norm, $|\text{STFT}(\cdot)|$ denotes the STFT magnitudes, and N is the number of magnitude elements. The multi-resolution STFT-based loss L_{sp} is formulated as

$$L_{\text{sp}}(G) = \frac{1}{M} \sum_{m=1}^{M} (L_{\text{sc}}^{(m)}(G) + L_{\text{mag}}^{(m)}(G)), \tag{2.9}$$

where M denotes the number of STFT setting groups, and each group includes different FFT sizes, frame lengths, and frame shifts. The losses $L_{\text{sc}}^{(m)}$ and $L_{\text{mag}}^{(m)}$ are calculated on the basis of the STFT features extracted using the settings of the m group. The multiple STFT losses prevent a suboptimal problem for the generator and enhance the modeling capability of the generator by making it capture hierarchical speech structures. Taken together, the overall training loss of the PWG generator (L_{G}) is formulated as

$$L_{\text{G}}(G, D) = L_{\text{sp}}(G) + \lambda_{\text{adv}} L_{\text{adv}}(G, D), \tag{2.10}$$

which is a weighted sum of L_{adv} and L_{sp} with weight λ_{adv}, and the hyperparameter λ_{adv} is empirically set to 4.0 in this thesis.

WaveNet-like Generator

As shown in Fig. 2.7, a WN-like architecture including DCNN, auxiliary acoustic features, skip connections, and gated structures is adopted for the PWG generator. The main differences between the PWG generator and the WN are a Gaussian noise input instead of previous samples, a raw waveform output instead of a probability

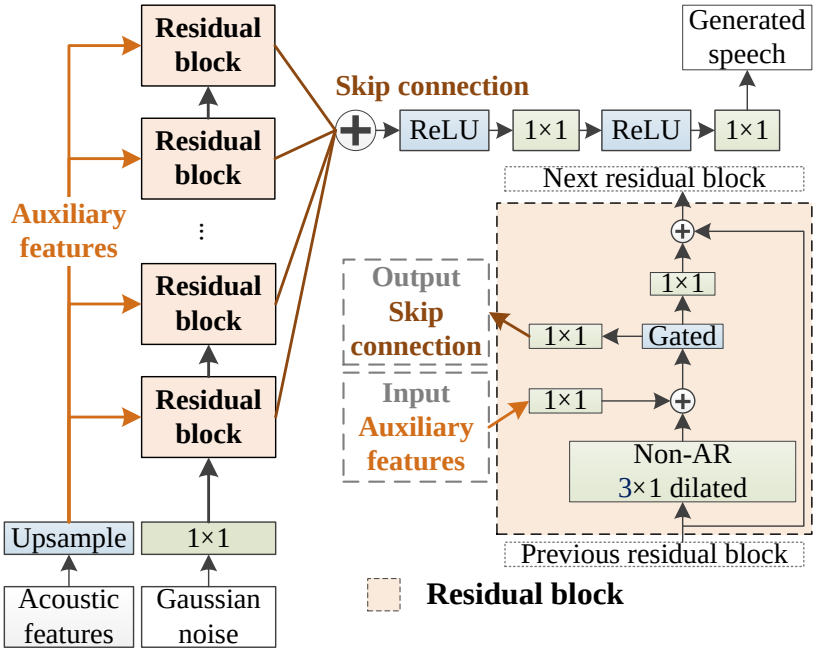


Figure 2.7: *Generator of parallel WaveGAN.*

distribution, and non-AR and noncausal manners. Moreover, a more compact model including fewer CNN channels is adopted in the PWG generator. Therefore, the generation speed of the PWG is much higher than the WN because of the parallel generation by the non-AR structure and the much smaller model.

2.1.5 Speech Manipulation of STRAIGHT and WORLD

To flexibly manipulate speech components such as pitch and timbre, many source–filter vocoder techniques have been proposed. However, the spectral estimation of early approaches such as the linear predictive coding (LPC) vocoder technique [55, 56] are susceptible to signal periodicity [78]. Specifically, obtaining a stable spectral envelope regardless of windowing temporal positions is difficult for voiced speech analysis. The time-variant pitch and natural fluctuations result in the periodicity interferences in

spectral analysis because of the fixed window length.

To address this problem, STRAIGHT [35] and WORLD [37] have been proposed. The STRAIGHT vocoder adopts a pitch-synchronized mechanism [79] with phasic interference reduction and oversmoothing compensation to extract stable spectra, which are highly uncorrelated to the instantaneous F_0. Specifically, when extracting features, the window of each frame has a different length according to the F_0 of this frame to prevent the periodicity interferences from the voiced speech. Furthermore, as an improved and real-time version, the WORLD vocoder also adopts the pitch-synchronized concept for spectral analysis [80].

Although the STRAIGHT and WORLD vocoders achieve speech manipulation with high flexibility, the lost details and phase information problems cause speech quality degradation. The recent neural vocoders greatly improve speech quality but suffer from the limited flexibility of speech manipulation. As a result, we propose a pitch-adaptive component, PDCNN, and a cascaded structure to improve the pitch controllability of the WN and PWG vocoders while trying to maintain a similar speech quality. The proposed QPNet and QPPWG vocoders are also conditioned on the WORLD-extracted features, and we expect that the QPNet and QPPWG vocoders are capable of manipulating the pitch similarly to the WORLD vocoder.

2.2 Voice Conversion

Voice conversion is a technique of converting speech characteristics such as the speaker identity and emotion of an input speech while maintaining the same linguistic content. Speaker voice conversion is the most general voice conversion scenario converting a source speaker identity to a specific target speaker. For simplicity, VC is used in this thesis to refer to speaker voice conversion. Parallel and non-parallel VC tasks

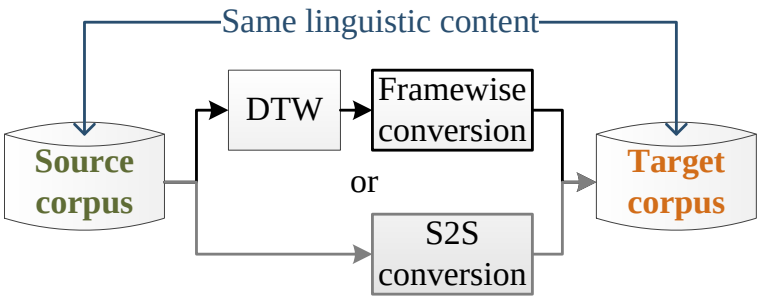


Figure 2.8: *Parallel voice conversion.*

are two main VC categories according to the content of each speaker in the training corpus. The details are as follows.

2.2.1 Parallel Voice Conversion

All speakers in a parallel corpus have identical utterances with the same linguistic contents. Although the data lengths may differ among the parallel utterances because of speaker-dependent (SD) speaking rates, it is easy to train a mapping function between source and target speakers with implicit one-to-one correlations. As shown in Fig. 2.8, the simplest method of tackling the mismatched data lengths is the use of dynamic time warping (DTW) [81] to align framewise the source and target acoustic features. With the aligned source-target acoustic features, a framewise conversion function such as the Gaussian mixture model (GMM) [7–9], deep neural network (DNN) [82–84], or exemplar-based model [85–87] can be easily built. However, because of the extra aligned errors introduced by the DTW and the limited prosody conversion capability of the frame-based conversion techniques, many advanced sequence-to-sequence (S2S) models [88–90] have been proposed. Since prosodic characteristics such as speaking rate are highly related to the speaker identity, the S2S conversion models usually attain better speaker similarity of the converted speech.

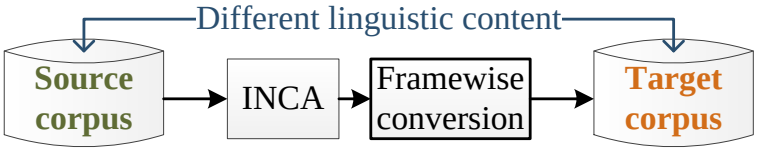


Figure 2.9: *Voice conversion with INCA.*

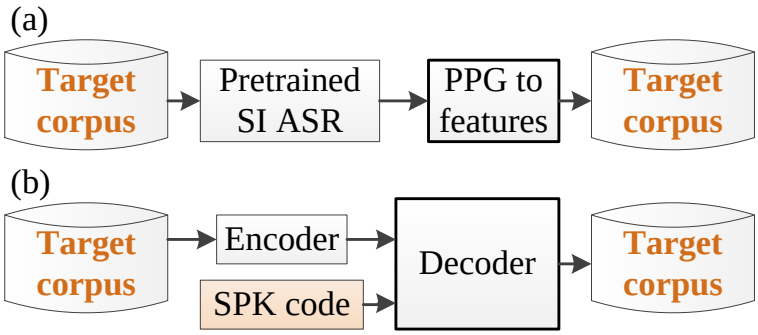


Figure 2.10: *(a) PPG-based voice conversion; (b) VAE-based voice conversion.*

2.2.2 Non-parallel Voice conversion

The collection of a parallel corpus for VC training is however time-consuming and expansive; thus, many non-parallel VC methods have been proposed. Erro et al. [91] proposed the INCA algorithm to iteratively align the non-parallel corpus for conventional parallel GMM-based VC, as shown in Fig. 2.9. Sun et al. [92] and Xie et al. [93] proposed similar frameworks using a well-trained automatic speech recognition (ASR) system to extract speaker-independent (SI) phonetic posteriorgrams (PPGs) and adopt an SD PPG-to-spectrum model to generate converted spectra, as shown in Fig. 2.10 (a). Restricted Boltzmann machine (RBM)- [94] and variational autoencoder (VAE)-based [95–97] (Fig. 2.10 (b)) models have also been proposed to disentangle the acoustic features into SD and SI components for VC. Moreover, inspired by the success of cycle consistent adversarial networks (CycleGAN) for image translation [98], cycleconsistency has been widely applied to non-parallel VC [53, 99, 100] as shown in Fig. 2.11.

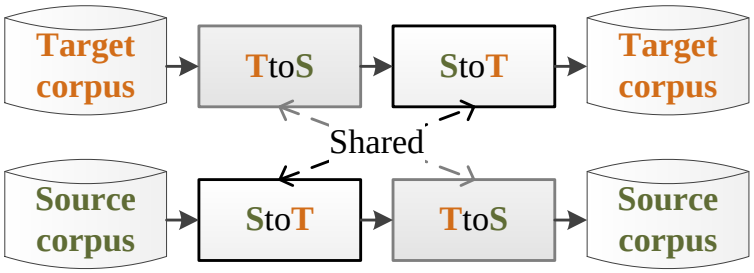


Figure 2.11: *Cycle voice conversion.*

In addition, non-parallel VC with external reference speakers has also been widely surveyed. For instance, building a VC model of reference speakers with a parallel corpus and adapting it for source and target speakers with a non-parallel corpus [101,102] achieved early success. On the basis of the GMM-based speaker verification technique [103], speaker adaptation from a universal background model (UBM) [104] for non-parallel VC also shows the effectiveness of assistance from reference speakers. Representing the source and target utterances on the basis of weighted reference dictionaries also attains good quality for exemplar-based VC [105]. Eingenvoice conversion (EVC) technique [106–108] also adopts limited parallel data including several pre-stored speakers and one reference speaker to build an initial model, and then the model is adapted for arbitrary source speakers to the reference speaker and the reference speaker to arbitrary target speakers. The final many-to-many process is performed by the cascaded source-to-reference and reference-to-target models. The reference speaker is treated as a hidden variable in the EVC system, which is similar to the latent variable in the VAE-based VC. In this thesis, a two-stage non-parallel VC system with a reference speaker is taken as the baseline VC system. The cascaded VC system is also built on the basis of a combination of many-to-reference and reference-to-many models [108–110].

2.2.3 DNN-based Voice Conversion

A general DNN-based framewise spectral conversion model [111,112] including training and conversion stages has been adopted in the baseline VC system submitted to VCC2018 [43]. In the training stage, given the paired source feature vector $\boldsymbol{S}_n = \left[\boldsymbol{s}_n^\top, \Delta\boldsymbol{s}_n^\top\right]^\top$ and target feature vector $\boldsymbol{T}_n = \left[\boldsymbol{t}_n^\top, \Delta\boldsymbol{t}_n^\top\right]^\top$, which include static and delta spectral features with the frame index n, the DNN-based conditional probability is formulated as

$$P\left(\boldsymbol{T}_n \mid \boldsymbol{S}_n, \boldsymbol{\lambda}, \boldsymbol{\Sigma}\right) = \mathcal{N}\left(\boldsymbol{T}_n; f_{\boldsymbol{\lambda}}(\boldsymbol{S}_n), \boldsymbol{\Sigma}\right), \tag{2.11}$$

where $\boldsymbol{\lambda}$ and $f_{\boldsymbol{\lambda}}$ respectively denote the parameters and nonlinear transformation function of the DNN model, $\mathcal{N}$ is the Gaussian distribution, and $\boldsymbol{\Sigma}$ is the diagonal covariance matrix of the training data. In the training stage, the DNN parameter $\hat{\boldsymbol{\lambda}}$ is updated as

$$\begin{aligned}\hat{\boldsymbol{\lambda}} &= \arg\max_{\boldsymbol{\lambda}} \sum_{n=1}^{N} \log P\left(\boldsymbol{T}_n \mid \boldsymbol{S}_n, \boldsymbol{\lambda}, \boldsymbol{\Sigma}\right) \\ &= \arg\min_{\boldsymbol{\lambda}} \sum_{n=1}^{N} (\boldsymbol{T}_n - f_{\boldsymbol{\lambda}}(\boldsymbol{S}_n))\boldsymbol{\Sigma}^{-1}(\boldsymbol{T}_n - f_{\boldsymbol{\lambda}}(\boldsymbol{S}_n)).\end{aligned} \tag{2.12}$$

In the conversion stage, the maximum likelihood parameter generation (MLPG) [113] is adopted to alleviate the discontinuity caused by the framewise approach, and the global variance (GV) postfilter [9] is applied to minimize the oversmoothing effect caused by the statistical nature.

2.2.4 DMDN-based Voice Conversion

Because the weakness of the DNN-based VC model is its unimodal nature without variances, a multimodal approach, deep mixture density network (DMDN) [114], has

also been applied to the baseline VC system [28]. The DMDN model attains the variance predicting capability and enhances the model capacity by modeling the conditional probability with mixtures of Gaussian distributions instead of a single Gaussian distribution. Given the same condition as those in Eq. 2.11, the DMDN-based conditional probability is formulated as

$$P\left(\boldsymbol{T}_n \mid \boldsymbol{S}_n, \boldsymbol{\theta}\right) = \sum_{m=1}^{M} \alpha_m(\boldsymbol{S}_n)\, \mathcal{N}\left(\boldsymbol{T}_n \mid \mu_m(\boldsymbol{S}_n), \sigma_m^2(\boldsymbol{S}_n)\right), \tag{2.13}$$

where $\boldsymbol{\theta}$ denotes the parameters of the DMDN model, $\mathcal{N}\left(\cdot \mid \mu, \sigma^2\right)$ denotes a single Gaussian mixture with the mean μ and covariance matrix σ^2, M is the number of mixture components, m is the mixture index, and α_m denotes the mixture weight of the m_{th} component given $\boldsymbol{S}_n$. The DMDN outputs are as follows:

$$\alpha_m(\boldsymbol{S}_n) = \frac{\exp\left(\psi_m^{(\alpha)}\left(\boldsymbol{S}_n, \boldsymbol{\theta}\right)\right)}{\sum_{j=1}^{M} \exp\left(\psi_j^{(\alpha)}\left(\boldsymbol{S}_n, \boldsymbol{\theta}\right)\right)}, \tag{2.14}$$

$$\mu_m(\boldsymbol{S}_n) = \psi_m^{(\mu)}\left(\boldsymbol{S}_n, \boldsymbol{\theta}\right), \tag{2.15}$$

$$\sigma_m(\boldsymbol{S}_n) = \exp\left(\psi_m^{(\sigma)}\left(\boldsymbol{S}_n, \boldsymbol{\theta}\right)\right), \tag{2.16}$$

where $\psi^{(\alpha)}$, $\psi^{(\mu)}$, and $\psi^{(\sigma)}$ respectively denote the weight, mean, and variance activations of the DMDN output layer. The updated form of the DMDN parameters is defined as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \log P\left(\boldsymbol{T}_n \mid \boldsymbol{S}_n, \boldsymbol{\theta}\right). \tag{2.17}$$

The MLPG and GV techniques are also adopted in the DMDN conversion stage.

2.3 Summary

In this chapter, the development of vocoders and VC techniques is reviewed. The baseline WN and PWG vocoders and DNN and DMDN VC models are also introduced.

Although this thesis focuses on improving unified vocoders such as WN and PWG, prior knowledge related to the source–filter model and speech production mechanisms is applied to the baseline unified vocoders to improve their pitch controllability in Chaptesr 5 and 6. The basic DNN and DMDN VC models is adopted to develop the baseline non-parallel VC model in Chapter 3. Since in this thesis I attempt to improve the neural vocoders for VC and pitch transformation scenarios, a simple cascaded non-parallel VC model is adopted in the next chapter, and the following chapters focus on techniques of improving the quality, robustness, and controllability of the neural vocoders.

3 Non-parallel Voice Conversion with Reference Speaker

The main motivation of this thesis is developing high-quality neural-based speech synthesis modules for different speech generation applications. In addition to the basic analysis-synthesis scenario of vocoders, voice conversion (VC) and pitch transformation applications have been constructed to evaluate the performance of the adopted neural vocoders. In this chapter, a basic non-parallel VC system using a WaveNet (WN) vocoder is introduced, and this system is the VC baseline of Chapter 4 and 5. Since the baseline system has been submitted to Voice Conversion Challenge 2018 (VCC2018), and the organizer provides many subjective results from crowdsourced evaluations, the system is a convincing reference for the following evaluations in this thesis.

3.1 Introduction

Voice conversion is a technique to generate specific target speech based on given source speech while maintaining the same linguistic content. One of the most general VC scenarios is speaker conversion, which converts speaker identity from a source speaker to a target speaker. For simplicity, we use the term VC in this thesis to denote speaker conversion. For conventional VC [7–9,82–87], a parallel corpus including paired source and target utterances is required, and the paired utterances have the

same linguistic contents. However, collecting a parallel corpus is expensive, time-consuming, and impractical for real-world applications. Therefore, many non-parallel VC techniques [53, 91–97, 99–102, 104, 105, 108–110] have been proposed.

In this chapter, the NU (Nagoya University) non-parallel VC system [43] for the SPOKE task of VCC 2018 is presented. A non-parallel corpus and corresponding transcripts of source and target speakers are provided in the SPOKE task. The key idea of our proposed system is taking advantage of the provided transcripts to generate a parallel corpus using a text-to-speech (TTS) system. That is, the TTS speaker is taken as a reference speaker, so source-to-reference (StoR) and reference-to-target (RtoT) conversion models can be respectively built. With the cascaded StoR and RtoT models, a non-parallel VC system is available. In VCC2018 [115], our system adopted a deep neural network (DNN)-based [111, 112] spectral conversion model and a WN [12] vocoder [47,48], and achieved the second-best score for similarity and an above-average score for naturalness among all submitted systems.

Since the cascaded VC with a TTS reference still caused performance degradation, an AutoEncoder (AE) framework [116] has been adopted to compensate for the acoustic mismatch between training and testing stages. Moreover, because of the lack of variances and the unimodal nature of a DNN-based VC model, the effectiveness of a deep mixture density network (DMDN)-based [114] VC model has also been explored. The main contributions of this work are three-fold:

- The effectiveness of a neural network (NN)-based non-parallel VC system using TTS-generated reference utterances has been shown.
- The effectiveness of DNN- and DMDN-based VC models have been explored.
- An AE to compensate for the acoustic mismatch between the training and testing stages has been proposed.

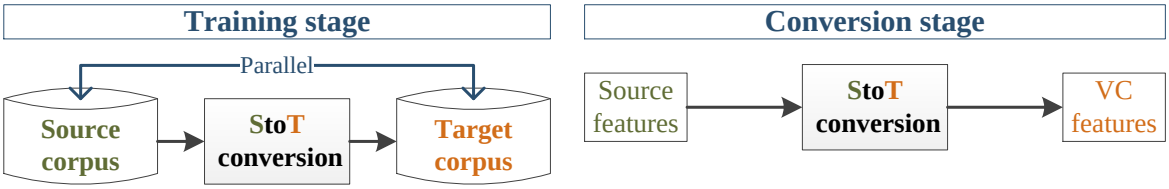


Figure 3.1: *Parallel voice conversion system.*

The rest of this chapter is organized as follows. The proposed cascaded VC model is presented in Section 3.2. Both objective and subjective evaluation results are presented in Section 3.3. Last, we give a summary in Section 3.4.

3.2 Cascaded Voice Conversion

As shown in Fig. 3.1, a typically parallel VC model usually includes training and conversion stages. In the training stage, the source-to-target (StoT) mapping function of the StoT conversion model is constructed with a parallel corpus, which has an inherent one-to-one relationship between source and target data. However, an inherent one-to-one relationship does not exist in a non-parallel corpus such as the SPOKE set of VCC2018 [115]. To address this issue, since the transcripts of the SPOKE set are available, it is feasible to use a TTS system to respectively generate corresponding parallel utterances for the source and target speakers of the SPOKE set. With the TTS-generated parallel corpus, a cascaded non-parallel VC system is available [43]. Moreover, to ease the training and testing mismatch, we also proposed an AE for mismatch compensation [116]. The details are as follows.

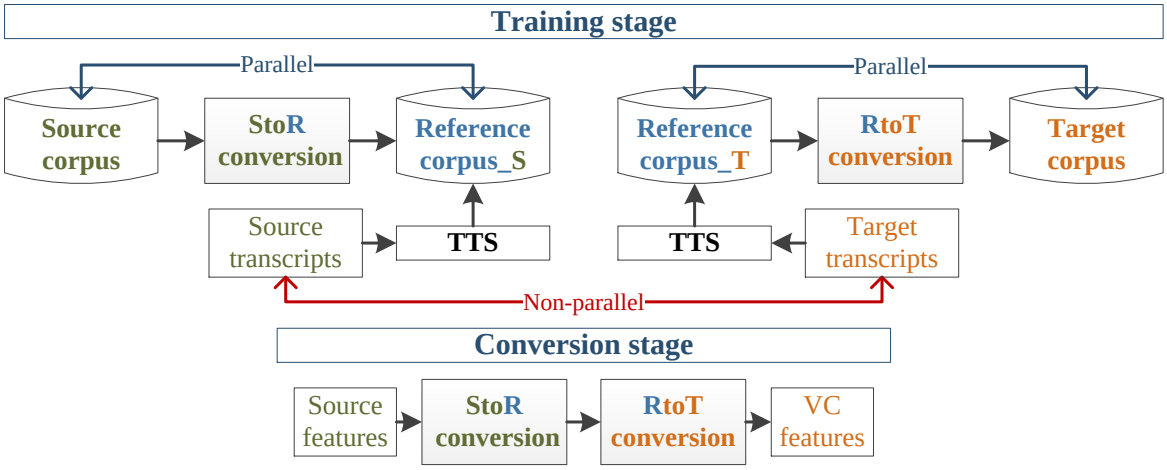


Figure 3.2: *Cascaded voice conversion system with reference speaker.*

3.2.1 VC with Reference Speaker

As shown in Fig. 3.2, source-to-reference (StoR) and reference-to-target (RtoT) models are respectively developed using the TTS speaker as a reference speaker in the training stage, and source features are converted to target features via the cascaded StoR and RtoT models in the conversion stage. Specifically, the parallel corpus for training the StoR model includes the source corpus and reference corpus_S, which is established by a unit-selection-based single-speaker TTS system with the source transcripts. The RtoT model is trained with the parallel corpus including the target corpus and reference corpus_T, which is established by the same TTS system but with the target transcripts. In the conversion stage, the input source features are converted to reference features by the StoR model, and then the reference features are further converted to the target features by the RtoT model.

To alleviate the alignment mismatch between the source/target and reference utterances, human-labeled short pauses and silences of the training utterances are adopted to handle the short pauses and silences of the TTS-generated speech to match that

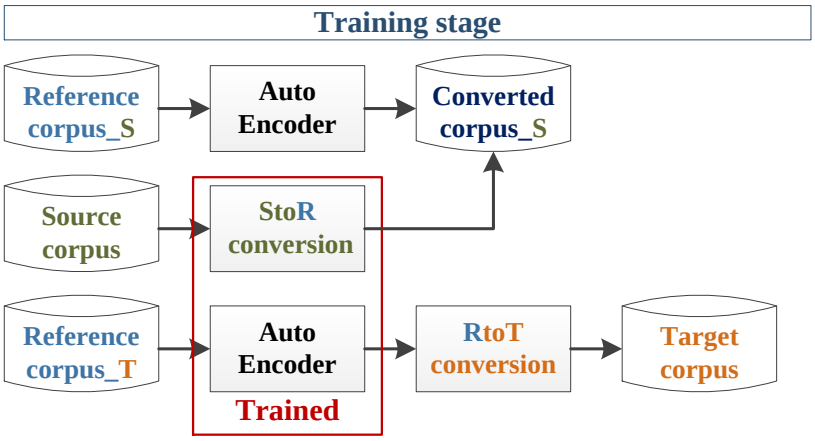


Figure 3.3: *Cascaded voice conversion system with mismatch compensation.*

of the corresponding source/target speech. After that, because the TTS system uses hidden Markov model (HMM)-state alignments, the framewise DTW technique is still applied to alleviate the spectral mismatch between the natural acoustic features and the acoustic features extracted from the TTS-generated speech. In conclusion, arbitrary parallel VC models can be adopted for non-parallel VC using the TTS-generated parallel corpus and cascaded two-stage approach. In this chapter, we respectively apply DNN- and DMDN-based VC models to the cascaded VC system.

3.2.2 Cascaded VC with Mismatch Compensation

The proposed cascaded VC system converts features in a two-stages manner, and these cascaded conversion models are independently trained. Therefore, the proposed system suffers a mismatch problem of the two-stage conversion. Specifically, the mismatch between the outputs of the StoR model in the conversion stage and the inputs of the RtoT model in the training stage will cause quality degradation. Therefore, we proposed an AE module for mismatch compensation [116]. As shown in Fig. 3.3,

after the training of the StoR model is finished, an AE is trained for mapping original reference features to converted reference features. When we train the RtoT model, the input reference features are processed by the trained AE to simulate the distorted input features from the StoR model outputs in the conversion stage. With the aid from the AE, we can ease the mismatch between the inputs of the RtoT model in the training and conversion stages.

3.3 Experimental Evaluation

In this section, we present the internal objective evaluations and the external subjective evaluations carried out in VCC2018. The details are as follows.

3.3.1 Experimental Setting

VCC2018 Corpus

The VCC18 corpus [115] is an English dataset provided by the VCC2018 organizer. The corpus includes two subsets, HUB and SPOKE. The HUB subset consists of four male speakers and four female speakers. Two males and two females are the source speakers, and the remaining four speakers are the target speakers for the parallel VC task (HUB task). Each speaker in the HUB set has 81 utterances for training and 35 utterances for testing. The linguistic contents of the utterances in the HUB set are parallel, which means that each HUB speaker recorded their utterances based on the same transcription. On the other hand, the SPOKE subset includes another two male and two female speakers as the source speakers for the non-parallel VC task (SPOKE task). Each speaker in the SPOKE set also has 81 utterances for training and 35 parallel

utterances for testing. Although each SPOKE speaker also recorded their utterances based on the same transcription, the SPOKE transcription is different from the HUB transcription. The total number of source-target pairs in the SPOKE task is 16 (four SPOKE source speakers and four HUB target speakers), which includes four female-to-female (F–F) pairs, four female-to-male (F–M) pairs, four male-to-female (M–F) pairs, and four male-to-male (M–M) pairs. The sampling rate of speech signals is 22,050 Hz and the resolution per sample is 16 bits. Both HUB and SPOKE transcripts are also provided by the VCC2018 organizer.

Internal TTS Corpus and TTS-generated Reference Corpus

The reference utterances used to construct the cascaded VC system was generated by a concatenative unit-selection TTS system, which was trained by around 3000 utterances from a single male speaker. Notably, although the linguistic contexts are the same, each speaker still has different prosody such as short-pause positions. The different prosodic patterns result in significantly different spectral characteristics. To alleviate these acoustic mismatches, we controlled the short-pause positions of the TTS-generated utterances using human-labeled pauses. Therefore, each speaker in the SPOKE task has its specific TTS-generated reference utterances.

WORLD Acoustic Feature

The WORLD [37] vocoder was adopted to extract a 513-dimensional aperiodicity (*ap*), 513-dimensional spectral envelope (*sp*), and one-dimensional fundamental frequency (F_0) with 5 ms frameshift. The *sp* feature was further parameterized into a 34-dimensional mel-spectrum (*mcep*), and the *ap* feature was coded into a two-dimensional aperiodic component (*codeap*). Joint spectral features were aligned via dynamic time

warping (DTW). For non-parallel VC, each source *mcep* was converted to a target *mcep* by the cascaded VC model. The source F_0 sequence was linearly transformed into a target one in the logarithm domain. The source *codeap* was kept the same.

Network Architecture

Both DNN- and DMDN-based VC models contained four hidden layers with 1024 hidden units, and the mixture number of the DMDN model was 16. The nonlinear activation functions were rectified linear units (ReLUs) [76] and the optimization algorithm was Adam [117]. The weights were randomly initialized by Xavier [118] and the biases were initially set to zero. The learning rate was 6×10^{-4} without decay, the training epochs was 15, and the utterance-based mini-batch was adopted. The settings of the compensation AEs followed the VC models, but the numbers of training epoch were increased to 85.

3.3.2 Objective Evaluation

To easily compare the proposed non-parallel VC models with parallel VC models, a parallel corpus consisting of 12 speaker pairs was adopted. The 12 speaker pairs were constructed on the basis of the four speakers in the SPOKE subset. Although the parallel corpus was adopted, the proposed non-parallel VC models did not use any parallel information. Moreover, the results of [91] also confirm that taking a parallel corpus for non-parallel training achieves almost the same experimental tendency as taking a non-parallel corpus for that. Therefore, four SPOKE-to-reference and four reference-to-SPOKE conversion models were trained with the SPOKE subset and the corresponding TTS-generated utterances. These eight VC models formed the 12 simu-

lated non-parallel VC (two-stage) paired models, which did not adopt any source-target parallel information.

In this section, a comparison among the parallel-VC and the proposed two-stage VC systems is first presented. Secondly, the proposed method is also compared with a basic any-to-one (AtoO) VC system adopted in the VCC2018 baseline system [43]. The objective measurement was the mel-cepstrum distortion (MCD), which is defined as

$$\text{MCD[dB]} = \frac{1}{N}\sum_{n=1}^{N}\frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}\left(\hat{y}_{n,d}-y_{n,d}\right)^2}\,, \tag{3.1}$$

where $\hat{y}$ is the converted *mcep*, y is the target *mcep*, n is the frame index, and d is the dimension index.

Comparison of Proposed Methods

In this comparison, four main arguments were explored. First, the degradation of the proposed non-parallel (two-stage) VC compared with the parallel (one-stage) VC was investigated. Secondly, the effectiveness of TTS-generated speech as a reference was evaluated. Thirdly, the effectiveness of the proposed AE was also evaluated. Last, the evaluation of a non-parallel corpus was also conducted. The six systems built for the evaluation are as follows.

- One-stage: the basic one-to-one parallel VC system.
- Two-stage (TTS): the proposed cascaded VC system with TTS-generated speech as the reference.
- Two-stage (Natural): the proposed cascaded VC system with parallel natural speech as the reference speech

Table 3.1: *MCD of DNN- and DMDN-based VC models*

	DNN-based		DMDN-based	
	w/o GV	w/ GV	w/o GV	w/ GV
One-stage	5.48	6.12	5.39	6.00
Two-stage (TTS)	5.64	6.22	5.59	6.13
Two-stage (Natural)	5.67	6.19	5.55	6.10
Two-stage (AE)	5.70	6.35	5.54	6.15
Two-stage (VC)	5.61	6.25	5.57	6.20
Non-parallel (TTS)	5.54	6.01	5.46	5.90

- Two-stage (AE): the proposed cascaded VC system with the compensation AE.
- Two-stage (VC): the RtoT model of the proposed cascaded VC system was directly trained with the parallel converted features from the StoT model.
- Non-parallel: the proposed cascaded VC system with the non-parallel corpus (SPOKE source and HUB target speakers).

Since the parallel natural speech and the parallel converted features from the StoT model do not exist in a real non-parallel scenario, the two-stage (Natural) and (VC) are the systems in a control group.

As shown in Table 3.1, the one-stage VC system still outperforms all two-stage VC systems for both DNN- and DMDN-based VC models. The results show that the parallel information is still a strong prior knowledge for framewise spectral conversion. However, the results also show the effectiveness of the two-stage VC system to achieve an acceptable spectral prediction accuracy. Furthermore, the two-stage systems with TTS or natural speech as the reference attain similar MCDs, which indicate that TTS-generated speech already contains sufficient acoustic components for being the reference speech.

However, the MCD differences between the one-stage and two-stage (TTS) systems still imply that the mismatch between the training and conversion stages causes performance degradation. Therefore, we applied the compensation AE in the training stage for easing the input mismatch of the RtoT model between the training and conversion stages. As shown in Table 3.1, we can find that both the two-stage (AE) and (VC) systems outperform the two-stage (TTS) system in the DMDN-based VC without GV. The results confirm the degradation caused by the two-stage mismatch and show the effectiveness of the compensation AE. Although the two-stage (AE) system achieves higher MCDs than the two-stage (TTS) system in the DNN-based VC, the two-stage (VC) system still outperforms the two-stage (VC). The results also imply the speech degradation caused by the mismatch problem. However, the settings or hyper-parameters of DNN-based AE might not be optimized, so the compensation AE only shows the effectiveness in the DMDN-based VC.

Furthermore, the DMDN-based VC models achieve slightly higher spectral prediction accuracies than the DNN-based VC models. Although applying the GV postfilter leads to higher MCD values, the tendencies are still the same. The results confirm a reasonable spectral prediction accuracy of the DMDN-based system, and the DMDN-based system is comparable to our DNN-based system submitted to VCC2018. In conclusion, although the parallel VC (one-stage) models exhibit a higher conversion performance, the two-stage models still achieve an acceptable conversion accuracy, and the well-trained compensation AE can ease the mismatch problem.

Comparison Between Cascaded VC and Any-to-one VC

In this comparison, we compared the proposed two-stage VC framework with a basic AtoO VC system, which was adopted in the VCC2018 baseline system [43]. The basic

Table 3.2: *MCD of DNN- and GMM-based VC models*

	DNN-based		GMM-based	
	w/o GV	w/ GV	w/o GV	w/ GV
One-stage	5.54	6.16	5.46	6.13
Two-stage (TTS)	5.68	6.23	5.54	6.14
AtoO	6.09	6.90	5.95	6.79
Source		8.46		

AtoO VC system used the parallel data of two source speakers and a target speaker to train a VC model for each target speaker. Since our TTS system was built with a male speaker, we constructed an M–F AtoO VC model for each SPOKE male speaker using the utterances from the other SPOKE male speaker and the TTS system. The evaluation involved cross-validation of the two male speakers in the SPOKE subset. Because the VCC2018 baseline system was GMM-based, we evaluated the performance of both GMM-based and DNN-based models. The GMMs were 32 mixtures with a full covariance matrix. As shown in Table 3.2, the proposed system outperforms the AtoO system for both the DNN-based and GMM-based models w/o and w/ GV scenarios. That is, for the speaker spectral conversion, the proposed method achieves a comparable performance to the one-stage system and outperforms the basic AtoO system in the objective evaluation. Note that since we only conducted this evaluation with the M–F pairs, the DNN-based MOS results were a little different from the results in Table 3.1. To summarize, the evaluation results show the effectiveness of using TTS-generated reference speech for non-parallel VC while the training data is very limited.

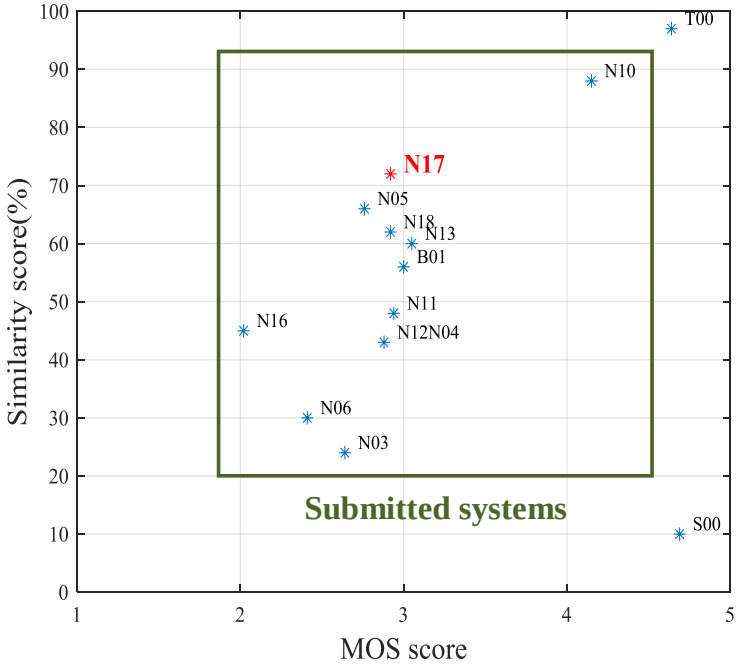


Figure 3.4: *Overall evaluation results of VCC2018 SPOKE task.*

3.3.3 External Subjective Evaluation

The VCC2018 organizer conducted a crowdsourced perceptual evaluation on all submitted systems for both the HUB task and the SPOKE task [115]. The number of unique listeners was 267 (146 male and 121 female). The evaluation included naturalness and speaker similarity tests. In the naturalness test, the measurement was the five-point mean opinion score (MOS), where "5" stood for "completely natural" and "1" stood for "completely unnatural". In the similarity test, listeners were asked to decide whether or not the converted utterances and target utterances were spoken by the same person. A four-point scale was given to listeners: "definitely the same", "probably the same", "probably different" and "definitely different". The final similarity scores were the percentage of the summation of "definitely the same" and "probably the same".

Table 3.3: *p-values of the naturalness evaluation results of VCC2018 SPOKE task.*

	Null hypothesis: quality is better or worse than N17
p-value >0.6	B01 (Baseline), N11, N18, N04, N12
p-value = 0.073	N13
p-value = 0.016	N05
p-value <2-16	S00 (Source), T00 (Target), N10, N06, N16

Furthermore, the VCC2018 organizer also provided a GMM-based baseline system (B01) [119]. For the spectral conversion, the baseline system constructed four AtoO GMM-based VC models for the HUB target speakers. The AtoO VC models were trained using the parallel utterances in the HUB subset. For example, when training an AtoO VC model for a female target speaker in the HUB subset, all female source speaker in the HUB subset were adopted to train this model. For the converted speech waveform generation, a vocoder-free method [120,121] was adopted for the intra-gender conversions, and an excitation generation and a mel-log-spectral-approximation (MLSA) filter [122, 123] were adopted for the inter-gender conversions.

On the other hand, the submitted NU non-parallel VC system (N17) [43] adopted the proposed DNN-based two-stage (TTS) VC model, and the final converted speech waveforms are respectively generated by four speaker-dependent (SD) WN vocoders [47,48]. The details of the SD WN vocoders can be found in Section 4.6. Figure 3.4 shows the overall results and Table 3.3 presents the significant relationships of our N17 system with others in terms of the p-values in the naturalness evaluations. Our system is about third place in the naturalness evaluations and second place in the speaker similarity measurements. The average MOS of the proposed system is about 3, and the average similarity accuracy is about 70 %. The details are described as follows.

Table 3.4: *Crowdsourced Perceptual Evaluation Results of VCC2018 SPOKE Task.*

	MOS of naturalness						Speaker similarity				
	F–F	F–M	M–F	M–M	Avg.		F–F	F–M	M–F	M–M	Avg.
S00	4.69	4.69	4.69	4.69	4.69	S00	10 %	10 %	10 %	10 %	10 %
T00	4.64	4.64	4.64	4.64	4.64	T00	97 %	97 %	97 %	97 %	97 %
N10	4.24	4.19	4.02	4.15	4.15	N10	83 %	94 %	74 %	86 %	88 %
N13	3.12	3.05	2.90	3.11	3.05	**N17**	**79 %**	**71 %**	**70 %**	**59 %**	**72 %**
B01	3.60	2.66	2.46	3.29	3.00	N05	65 %	68 %	56 %	78 %	66 %
N11	3.28	2.83	2.93	2.73	2.94	N18	66 %	72 %	37 %	55 %	62 %
N18	3.25	2.77	2.94	2.72	2.92	N13	55 %	66 %	53 %	70 %	60 %
N17	**3.20**	**2.86**	**2.75**	**2.85**	**2.92**	B01	66 %	59 %	49 %	35 %	56 %
N04	2.89	2.93	2.69	3.03	2.88	N11	46 %	60 %	18 %	50 %	48 %
N12	3.38	3.05	2.08	3.00	2.88	N16	38 %	63 %	14 %	56 %	45 %
N05	3.20	2.49	2.56	2.82	2.76	N12	25 %	60 %	9 %	68 %	43 %
N03	2.70	2.81	2.13	2.92	2.64	N04	24 %	69 %	17 %	57 %	43 %
N06	2.93	2.21	2.05	2.46	2.41	N06	28 %	33 %	0 %	50 %	30 %
N16	2.20	1.93	1.82	2.13	2.02	N03	17 %	37 %	3 %	34 %	24 %

Naturalness

As shown in Table 3.4, the similar MOS scores for all pairs indicate the generality of our N17 system under different conversion conditions. Compared with the B01 baseline system, the higher MOS scores of the cross-gender pairs imply the higher spectral prediction accuracy of the proposed two-stage VC model than that of the AtoO GMM-based VC model and the higher naturalness of the WN-generated speech than the MLSA-generated speech. However, our N17 system achieved worse performance in the intra-gender pairs, and the possible reason was the vocoder-free framework [119–121] of the B01 system. Specifically, since the vocoder-free framework directly modified the input source waveform to the target waveform bypassing the conventional vocoder

process, the naturalness of the converted speech was well maintained. However, since the WN vocoder was conditioned on the WORLD-extracted acoustic features, the extraction errors propagated to the WN-generated speech. For example, the pitch of a SPOKE male speaker is very low, and it easily causes voiced/unvoiced decision errors when extracting the F_0 by WORLD. According to the MOS results, there was a marked naturalness gap between the F–F pair and the other pairs of our N17 system. When listening to the converted speech, we found that the flawed F_0 values caused many unexpected scratchy voices, which significantly degraded speech naturalness. Furthermore, because the WN vocoder was trained with natural acoustic features but tested with converted acoustic features, the WN-generated speech sometimes became unstable. The instability made the generated speech include some unexpected noise, which is called a collapsed speech problem. This collapsed speech problem sometimes markedly degrade the speech quality, so the instability of the WN-generated speech might be another possible reason for the worse naturalness. The details of the collapsed speech problem will be described in Chapter 4.

Speaker Similarity

As shown in Table 3.4, our N17 system significantly outperforms the B01 baseline system for both inter-gender and intra-gender tasks. The results show that the WN vocoder retained the target's timbre better than the conventional vocoding process and the vocoder free frameworks [119–121]. Nonetheless, although our system achieves an above-average accuracy for speaker similarity in cross-gender and F–F pairs, the performances of our M–M pairs are seriously degraded. Broken excitation signals caused by the flawed F_0 might be a possible reason because the WORLD vocoder often has the difficulty in extracting the correct F_0 for male speakers. We also found that

the WN vocoder was more sensitive to flawed F_0 than the WORLD vocoder, and it resulted in many scratchy voices of the WN-generated speech. The scratchy voices usually cause significant blurring of speaker identity, particularly in same-gender pairs. Therefore, the speaker similarity of the M–M pairs was significantly degraded than that of the other pairs of our N17 system. To summarize, the proposed two-stage VC model with the WN vocoder shows the effectiveness of speaker identity conversion.

3.4 Summary

In this chapter, a non-parallel VC framework with a two-stage conversion and TTS-generated reference speech is described. The main concept is that using a TTS system to respectively generate parallel utterances for source and target speakers. The TTS-generated speech is a bridge to connect the non-parallel source and target utterances. However, the mismatch between the two stages of the cascaded conversion causes performance degradation, so we adopt a compensation AE in the training stage to ease the mismatch. On the other hand, because of the unimodal nature and the lack of variance of a DNN-based model, the effectiveness of a DMDN-based model for the proposed two-stage VC is also explored. Moreover, to easily compare with the VCC2018 baseline system, we also explore the proposed VC framework with a GMM-based model. Internal experimental results show that the proposed VC framework achieves acceptable spectral prediction accuracy, which is slightly worse than a parallel VC system, as well as the effectiveness of the compensation AE and the DMDN-based model. External subjective evaluation results provided by the VCC2018 organizer are also presented. The crowdsourced perceptual evaluation results demonstrate that our submitted non-parallel VC system, which is developed based on the DNN-based two-stage VC models and the WN vocoder, achieves a beyond average performance in both quality and

similarity measurements in VCC2018.

Although the evaluation results of the proposed VC system show the early effectiveness of the WN vocoder for generating high-quality VC speech, conditioned on the defective VC acoustic features usually causes instability of the WN vocoder. That is, the WN vocoder sometimes generates non-speech like and very noisy speech segments when the input acoustic features are distorted by a manipulation module such as VC models. These discontinue and noisy segments usually significantly degrade the generated speech quality. Therefore, to improve the quality and robustness of the proposed VC system, a waveform-based constraint for the WN vocoder will be presented to ease the negative effects of the unstable problem in the next chapter.

4 Collapsed Speech Detection and Suppression

Because of the rapid developments of neural-based speech generations, the state-of-the-art speech generative model such as WaveNet (WN) achieves very high-fidelity speech generation with natural acoustic features for the basic analysis-synthesis application. However, for most speech generation applications, the auxiliary acoustic features are distorted because of the manipluations. These distorted acoustic features usually cause unstable problem of the WN-generated speech. Therefore, to improve the robustness of our baseline voice conversion (VC) system introduced in Chapter 3, the defect of the WN vocoder combined with the VC model and a waveform-based constraint will be explored in this chapter.

4.1 Introduction

Conventional VC systems usually adopt parametric-based (conventional) vocoders such as STRAIGHT [35] and WORLD [37], which encode (analyze) speech into acoustic features such as spectral and prosodic features and decode (synthesize) speech based on these acoustic features. However, the oversimplified assumptions of the speech generation mechanism, such as the fixed length of analysis windows, a time-invariant linear filter, and a stationary Gaussian process, imposed on conventional vocoders lead

to loss of phase and temporal details of the original speech, which cause significant speech quality degradation of the synthesized speech signals.

Recently, many neural-based autoregressive (AR) models directly modeling raw speech waveforms such as WN [12] and SampleRNN [13] have been proposed. These neural-based models achieve high-fidelity speech generation by modeling the conditional probability distribution of each speech sample conditioned on past speech samples. The authors of [47, 48] applied WN to replace the synthesis part of conventional vocoders to markedly improve the quality of the synthesized speech. Specifically, the WN vocoder generates speech conditioned on not only previous speech samples but also conventional-vocoder-extracted acoustic features without various ad hoc assumptions, so the lost phase and temporal details can be greatly recovered by the WN vocoder.

However, directly integrating the WN vocoder into a VC system causes serious mismatch problems. Because of the length difference between the source and target data of a VC speaker pair, the WN vocoder is usually trained with natural target acoustic features and waveform pairs. In the testing stage, the trained WN vocoder is conditioned on the converted acoustic features, which have the same data length as the source acoustic features, to generate the converted waveforms, so the acoustic mismatch between the natural and converted acoustic features leads to significant quality degradation such as a waveform-based discontinuity [27,28,43]. Moreover, the inherent exposure bias problem [124,125] caused by the AR nature of the WN vocoder sometimes leads to unexpected noisy segments, especially when the WN vocoder is conditioned on artificial acoustic features such as those used in VC. The discontinuous waveforms and unexpected noisy segments caused by the acoustic and temporal mismatches and the exposure bias are called the collapsed speech problem [27, 28].

To address this problem, a simple and low cost collapsed speech suppression frame-

work [27, 28] using the prior knowledge of speech continuity is presented in this chapter. The framework includes a collapsed speech detection technique and a collapsed speech suppression constraint in waveform-domain. Since speech is a sequential signal with strong continuity, the WN-generated speech might follow the speech continuity. Furthermore, the conventional-vocoder-generated speech is a good reference, which is usually stable and collapsed-speech-free, so the predicted distributions of the WN vocoder might be constrained by the sequential correlations of the reference speech. Specifically, WN-generated speech is segmentally inspected using WORLD-generated speech as a reference. If collapsed speech is detected, the WN will regenerate this segment with a distribution constraint derived from the WORLD-generated speech and previous samples.

In this chapter, the proposed collapsed speech detection and suppression approach is evaluated with a baseline non-parallel VC system [43] submitted to the non-parallel VC (SPOKE) task of Voice Conversion Challenge 2018 (VCC2018) [115]. This chapter is organized as follows. In Section 4.2, the details of the collapsed speech problem are described. In Section 4.3, the collapsed speech detection technique is presented. In Section 4.4, the collapsed speech suppression constraint is described. In Section 4.5, the WN vocoder with the proposed framework is introduced. In Section 4.6, we report objective and subjective tests carried out to evaluate the effectiveness of the proposed framework. Finally, the conclusion is given in Section 4.7.

4.2 Collapsed Speech Problem

To directly model a speech waveform, WN [12] adopts an AR approach to generate the speech waveform sample by sample. Specifically, WN models the probability distribution of each speech sample conditioned on a segment of previous samples called a

receptive field. To guide WN to generate the desired speech content, WN is conditioned on the previous samples in the receptive field and auxiliary features such as linguistic features. Furthermore, taking WN as a vocoder [47–49], which adopts the conventional-vocoder-extracted acoustic features as the auxiliary features, greatly reduces the huge training data requirement and makes it easy to combine WN with conventional VC systems [43, 44, 46]. The conditional probability of the WN vocoder is formulated as

$$P\left(\boldsymbol{x} \mid \boldsymbol{h}\right) = \prod_{t=1}^{T} P\left(x_t \mid x_{t-r}, \ldots, x_{t-1}, \boldsymbol{h}\right), \tag{4.1}$$

where t is the sample index, r is the length of the receptive field, x_t is the current audio sample, and $\boldsymbol{h}$ is the vector of the acoustic features. In this chapter, the μ-law is applied to encode speech waveforms into 8 bits, so the output of the WN vocoder is a logistic distribution with 256 levels. The details about WN and the WN vocoder are presented in Section 5.2.

Although the effectiveness of the WN vocoder for generating high-fidelity speech on the basis of acoustic features has been proved, the AR nature and waveform-domain modeling make the WN vocoder vulnerable to prediction errors. Specifically, because the WN vocoder is conditioned on previous samples to predict the current sample, a prediction error will propagate through the sequential speech samples. The negative ripple effect easily leads to the WN vocoder generating very noisy speech, which is similar to white noise, especially when conditioned on acoustic features with high amplitudes. This white-noise-like speech is defined as Type I collapsed speech as shown in Fig. 4.1 (a). Furthermore, even if the prediction errors only occur in a few samples, it still leads to the significant discontinuity of the WN-generated waveform because of the direct waveform modeling. This waveform-domain discontinuity usually causes short impulse noise with significant perceptual quality loss. We define the short impulse noise as Type II collapsed speech as shown in Fig. 4.1 (a).

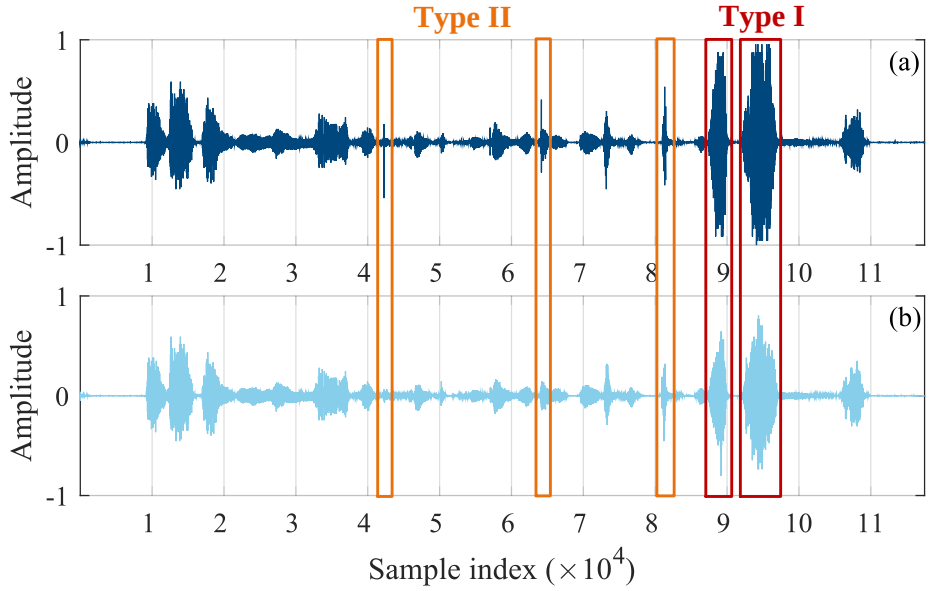


Figure 4.1: *(a) WN-generated waveform w/ collapsed speech. (b) WN-generated waveform w/ proposed collapsed speech suppression.*

The possible reasons for collapsed speech are a lack of training data, conditioned on artificial acoustic features, and exposure bias [124, 125]. Specifically, because of the different lengths of the source and target utterances for VC, the WN vocoder is usually trained with natural target acoustic features and waveforms but tested with converted acoustic features. The acoustic mismatch between the training and testing stages usually leads to the WN vocoder generating unexpected speech waveforms. Even if a data alignment technique such as dynamic time warping (DTW) is adopted to allow the WN vocoder to be trained with the VC acoustic features and natural waveforms pair, extra errors caused by the imperfect alignment are introduced. Furthermore, in our previous work [126], we found that even if the source and target utterances are length-matched, the temporal structures of the source and target utterances are still different. Training the WN vocoder using this temporal mismatched data still causes performance degradation. Moreover, because the AR WN model is usually trained

with ground-truth natural waveforms but tested with self-generated waveforms, the different decoding behavior, which is called the exposure bias problem, sometimes leads to unexpected generation results.

4.3 Collapsed Speech Detection

To tackle the collapsed speech problem, the first step is collapsed speech detection. According to the observation in our previous work [27, 28, 43], although the quality of WN-generated speech is usually higher than that of WORLD-generated speech, the WN vocoder is more sensitive to imperfect converted acoustic features while the WORLD-generated speech is usually stable and collapsed-speech-free. Moreover, although the perceptual qualities are different, the waveform envelopes and powers of the utterances generated by these vocoders are similar because of the same input acoustic features. Therefore, it is reasonable to take the utterance generated from the WORLD vocoder as a reference to evaluate whether or not the WN-generated utterance contains collapsed speech.

In our VC system submitted to VCC2018 [43, 46], a simple power-based collapsed speech detection technique is adopted. Since the powers of WN-generated and WORLD-generated utterances are similar, a large difference of the maximum powers usually indicates the WN-generated utterance suffers from collapsed speech, particularly in the high-frequency band. Following this observation, an utterance-based detection criterion has been proposed. Specifically, given the frame-based power sequences $\boldsymbol{p}^{(\mathrm{wn})} = \left[p_1^{(\mathrm{wn})}, \ldots, p_N^{(\mathrm{wn})}\right]$ and $\boldsymbol{p}^{(\mathrm{wd})} = \left[p_1^{(\mathrm{wd})}, \ldots, p_N^{(\mathrm{wd})}\right]$, and the power sequences of Nyquist frequency $\boldsymbol{p}_{\mathrm{L}}^{(\mathrm{wn})} = \left[p_{\mathrm{L},1}^{(\mathrm{wn})}, \ldots, p_{\mathrm{L},N}^{(\mathrm{wn})}\right]$ and $\boldsymbol{p}_{\mathrm{L}}^{(\mathrm{wd})} = \left[p_{\mathrm{L},1}^{(\mathrm{wd})}, \ldots, p_{\mathrm{L},N}^{(\mathrm{wd})}\right]$, the detec-

tion measurements are defined as

$$\Delta \boldsymbol{p} = \max\left(\boldsymbol{p}^{(\mathrm{wn})}\right) - \max\left(\boldsymbol{p}^{(\mathrm{wd})}\right) \tag{4.2}$$

and

$$\Delta \boldsymbol{p}_{\mathrm{L}} = \max\left(\boldsymbol{p}_{\mathrm{L}}^{(\mathrm{wn})}\right) - \max\left(\boldsymbol{p}_{\mathrm{L}}^{(\mathrm{wd})}\right), \tag{4.3}$$

where (wn) denotes powers of a WN-generated utterance, (wd) denotes powers of a utterance generated by the conventional WORLD vocoder, and N is the frame number of these utterances. If both $\Delta \boldsymbol{p}$ and $\Delta \boldsymbol{p}_{\mathrm{L}}$ are higher than an empirical threshold, the collapsed speech is detected. According to our internal experiments, the differences between maximum powers are more stable than the frame-based power differences.

However, there are several problems with the power-based method. First, impulse noise (Type II collapsed speech) is easily ignored by the utterance-based detection. Secondly, the utterance-based detection is inefficient for WN generation. Because collapsed speech usually occurs in a few samples, a segmental detection manner is more efficient for WN tackling the collapsed speech. Moreover, a segmental collapsed detection and suppression also restricts the side effects from the collapsed speech suppression in a short period. Therefore, we proposed a collapsed speech segment detection (CSSD) [27, 28] technique to segmentally detect the collapsed speech and only apply our collapsed speech suppression technique to the problematic segments.

The motivation of the proposed CSSD is that even without listening to audio samples, people still easily detect collapsed speech segments from the waveform shape. Therefore, the core of the CSSD is to segmentally compare the waveform envelopes of WN- and WORLD-generated utterances. A segment is detected as a collapsed segment when the difference between the two envelopes is larger than a threshold, which is determined by a detection error tradeoff (DET) curve. Note that because the conditional acoustic features of the WN vocoder already contain a power component, which

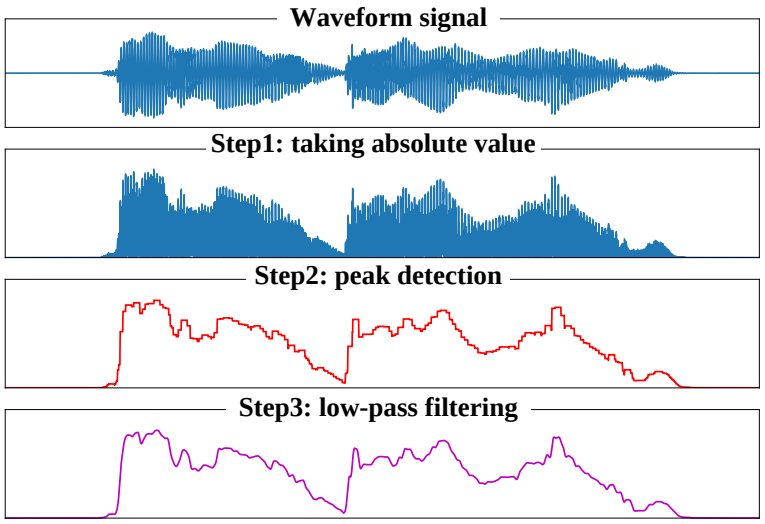


Figure 4.2: *Three steps of waveform shape detection.*

is consistent with that of the acoustic features for WORLD synthesis, the amplitudes of the WN- and WORLD-generated waveform envelopes should be similar. Therefore, the CSSD is employed without any waveform normalization.

Because of the frequent detection requirements, a low-computational-cost approach [127] is adopted to obtain the waveform envelopes for the CSSD. As shown in Fig. 4.2, we first take the absolute value of waveform signals. Secondly, a peak detection is performed by dividing the whole absolute sequence into non-overlapping slots and replacing all signals in each slot with the one with the maximum value in that slot. Finally, the final waveform envelope is obtained by processing the detected peak sequence with a low-pass filter. To achieve a lower collapsed speech detection error rate, the Hilbert transform (HT) instead of taking the absolute value is adopted in the first step. The length of the speech segment of the CSSD was 4000 samples, which means that the system checked for collapsed speech every time the WN vocoder generated 4000 new samples. The length of the peak detection window was 200 samples and the cutoff frequency of the low-pass filter in the CSSD was 300 Hz.

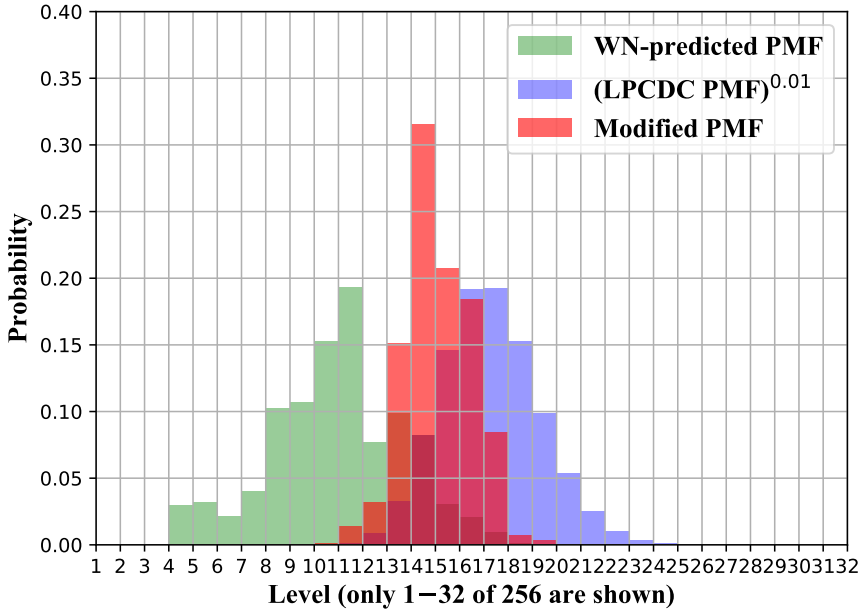


Figure 4.3: *Probability distributions of WN-predicted PMF, LPCDC PMF with regularizer* $\rho = 0.01$*, and LPCDC-modified PMF.*

4.4 Collapsed Speech Suppression

Since speech waveform is a sequential signal with strong continuity, any waveform-domain discontinuity usually causes marked noisy voices, collapsed speech. However, because of the prediction errors caused by the limited training data, distorted auxiliary features, and exposure bias problem, WN-generated speech sometimes suffers from waveform-domain discontinuities. Since the WN vocoder implicit models speech continuity because of the data-driven nature, we proposed a postprocessing module [27, 28] to explicitly constrain the WN output following the speech continuity extracted from WORLD-generated speech. Specifically, each WN output probability distribution is multiplied by another constraint probability distribution derived from the WORLD-generated speech to explicitly make WN-generated speech follow the continuity.

In our previous work [27, 28], we adopted a simple codec, linear prediction coding (LPC), to model speech continuity and derive the constraint probability distribution, LPC distribution constraint (LPCDC). The main concept of LPC is that the current speech sample can be represented as a linear combination of previous speech samples, so the relationship between the current and previous samples is described by the LPC coefficients. The proposed LPCDC extracts the relationships from WORLD-generated speech and constrains the corresponding outputs of the WN vocoder with these relationships. As shown in Fig. 4.3, the LPCDC-constrained (modified) form of equation 4.1 is derived as

$$P\left(\boldsymbol{x} \mid \boldsymbol{h}, \boldsymbol{\phi}\right) = \prod_{t=1}^{T} P\left(x_t \mid x_{t-r}, \ldots, x_{t-1}, \boldsymbol{h}, \boldsymbol{\phi}\right)$$

$$= \prod_{t=1}^{T} P\left(x_t \mid x_{t-r}, \ldots, x_{t-1}, \boldsymbol{h}, \right) \left(P\left(x_t \mid x_{t-d}, \ldots, x_{t-1}, \boldsymbol{\phi}\right)\right)^{\rho}, \tag{4.4}$$

where d is the number of LPC dimensions, $\boldsymbol{\phi}$ denotes the LPC coefficients extracted from the corresponding WORLD-generated speech, and ρ is a regularization hyper-parameter. That is, the probability distribution of each speech sample is constrained by the LPCDC mask $\left(P\left(x_t \mid x_{t-d}, \ldots, x_{t-1}, \boldsymbol{\phi}\right)\right)^{\rho}$, which is a probability mass function (PMF) approximating a Gaussian distribution with the mean μ_{lpc} and variance σ^2_{lpc}. The mean μ_{lpc} is the LPC-predicted value of the current sample, which is given by the weighted sum of the past samples multiplied by the d-dimensional LPC coefficients. The variance σ^2_{lpc} is the variance of the prediction errors derived from the corresponding frames of the WORLD-generated speech utterance. ρ is the weight used to control the balance between the LPCDC mask and the WN-predicted probability distribution.

For efficient WN generation, the waveform envelope and LPC coefficients of the WORLD-generated speech are extracted in advance. Only the WN vocoder segmentally generates speech samples, which are checked by the CSSD, in the testing stage. To

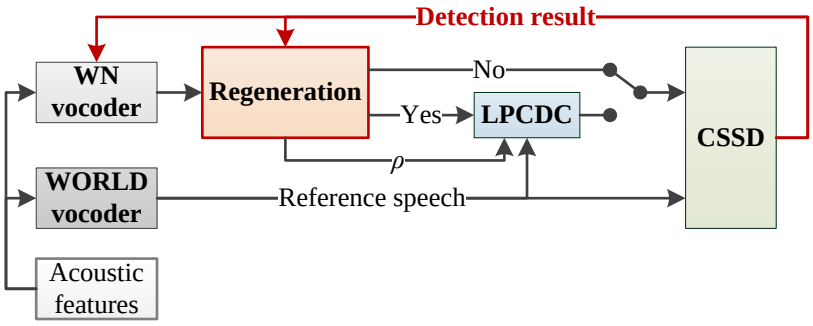


Figure 4.4: *WN vocoder with LPCDC and CSSD.*

simulate the effect of the μ–law codec, the WORLD-generated speech is also encoded and decoded by the μ–law. The 8-bit μ–law encoding is as follows:

$$f(x) = \mathrm{sgn}(x)\frac{\ln(1+255*|x|)}{\ln(1+255)}, \tag{4.5}$$

where x is the input speech sample. The output of the WN vocoder is the μ–law-encoded 256-level logistic distribution. To derive an LPCDC mask, we first obtain the real waveform amplitude of each level y_{level} as

$$y_{\mathrm{level}} = f^{-1}(q), \quad q \in [0, 1, \ldots, 255]. \tag{4.6}$$

Then, the value of each level of the LPCDC mask is approximated as

$$\mathrm{lpcdc}(y_{\mathrm{level}}) = \frac{1}{\sigma_{\mathrm{lpc}}\sqrt{2\pi}} \exp\left(\frac{-(y_{\mathrm{level}} - \mu_{\mathrm{lpc}})^2}{2\sigma_{\mathrm{lpc}}^2}\right). \tag{4.7}$$

Last, the LPCDC mask is normalized to make the summation equal to 1.

4.5 WaveNet Vocoder with Collapsed Speech Detection and Suppression

As shown in Fig. 4.4, the proposed system includes the WN and WORLD vocoders and the LPCDC and CSSD modules. Given acoustic features, reference speech is gen-

erated by the WORLD vocoder, and then LPC coefficients and reference waveform envelopes are extracted from the WORLD-generated reference speech. The WN vocoder sequentially generates nonoverlapping speech segments, and each WN-generated segment is examined by the CSSD. If collapsed speech is detected, the WN vocoder will regenerate this segment with the LPCDC. The maximum number of the regeneration times is three, and the regularizer ρ is respectively set as 0.01, 0.1, and 1 for the first, second, and third regenerations. The system preserves the latest results. The additional computational costs of the proposed system are mainly from WN regenerations compared with other fast modules (LPC extraction, LPCDC distribution derivation, waveform envelope detections and comparisons for CSSD, and WORLD synthesis). Therefore, if the WN generation time can be markedly reduced, a robust low-latency segmental generation system might be implemented based on the proposed system.

We take the WN-generated utterance of Fig. 4.5 (a) as an example, which is the same utterance as that in Fig. 4.1 (a). The corresponding WORLD-generated waveform is shown in Fig. 4.5 (b), the extracted waveform envelopes are shown in Fig. 4.5 (c), and the difference in waveform envelope is shown in Fig. 4.5 (d). The WORLD-generated waveform is stable and without collapsed speech segments, but the WN-generated one contains several Type I and II collapsed speech segments. The results in Figs. 4.5 (c) and (d) confirm the effectiveness of the CSSD module, which detects the collapsed speech segment based on the envelope difference.

Moreover, as shown in Fig. 4.6 (a), if we zoom in on a partial PMF sequence from the first Type II collapsed speech segment in Fig. 4.1 (a), we can find that a few samples with unexpected prediction errors lead to serious discontinuity and unexpected impulse noise. However, after applying the LPCDC PMF sequence shown in Fig. 4.6 (b) to constrain the WN vocoder outputs, the modified PMF sequence in Fig. 4.6 (c) is free

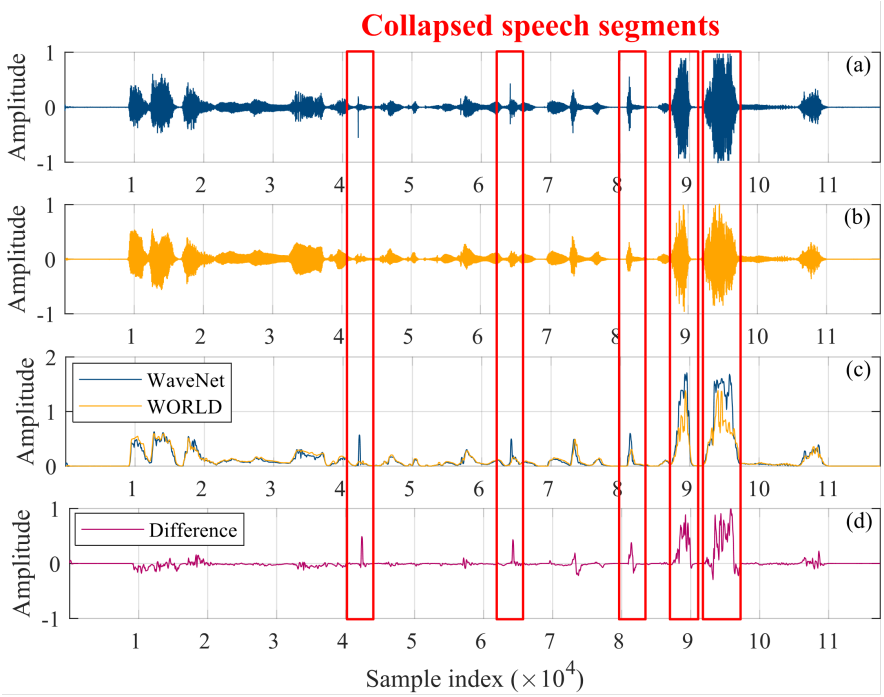


Figure 4.5: *(a) WN-generated waveform w/ collapsed speech. (b) WORLD-generated waveform. (c) Extracted waveform envelopes. (d) Difference in waveform envelope.*

from the unexpected prediction error. Furthermore, as shown in Fig. 4.7, the PMF sequences corresponding to the last Type I collapsed speech segment in Fig. 4.1 (a) also show the effectiveness of the proposed system. Specifically, most PMF values of the collapsed segment in Fig. 4.7 (a) are close to extremums resulting in continuous maximum amplitudes, but the modified PMF sequence in Fig. 4.7 (c) is normal and speech-like. Note that the modified predicted PMF sequence of Fig. 4.6 (c) / 4.7 (c) is not the result of directly multiplying the predicted PMF sequence of Fig. 4.6 (a) / 4.7 (a) by the LPCDC PMF sequence of Fig. 4.6 (b) / 4.7 (b). Because of the AR manner of WN, when the first sample in this segment is changed by the LPCDC, the distributions of the following samples are also affected. Finally, the refined speech waveform is shown in Fig. 4.1 (b), which is free from collapsed speech.

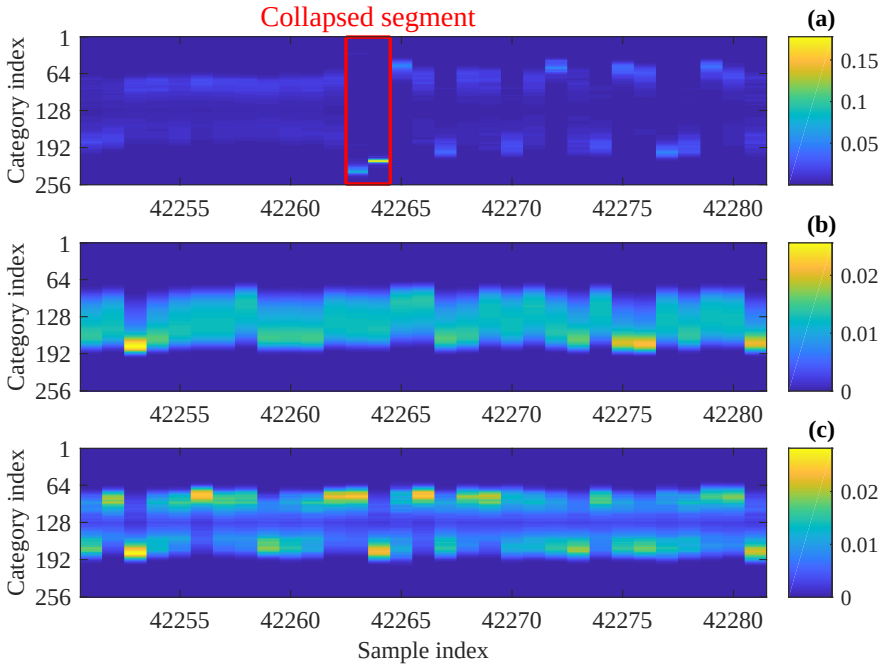


Figure 4.6: *(a) Predicted PMF sequence of WN w/ Type II collapsed speech. (b) PMF sequence from LPCDC. (c) Modified PMF sequence of WN w/ LPCDC and CSSD.*

4.6 Experimental Evaluation

In this section, our non-parallel VC system submitted to VCC2018 is taken as a baseline system. Both collapsed speech detection and perceptual quality evaluations are presented to respectively show the effectiveness of the proposed CSSD module and the proposed WN vocoder with the LPCDC and CSSD for improving the baseline system.

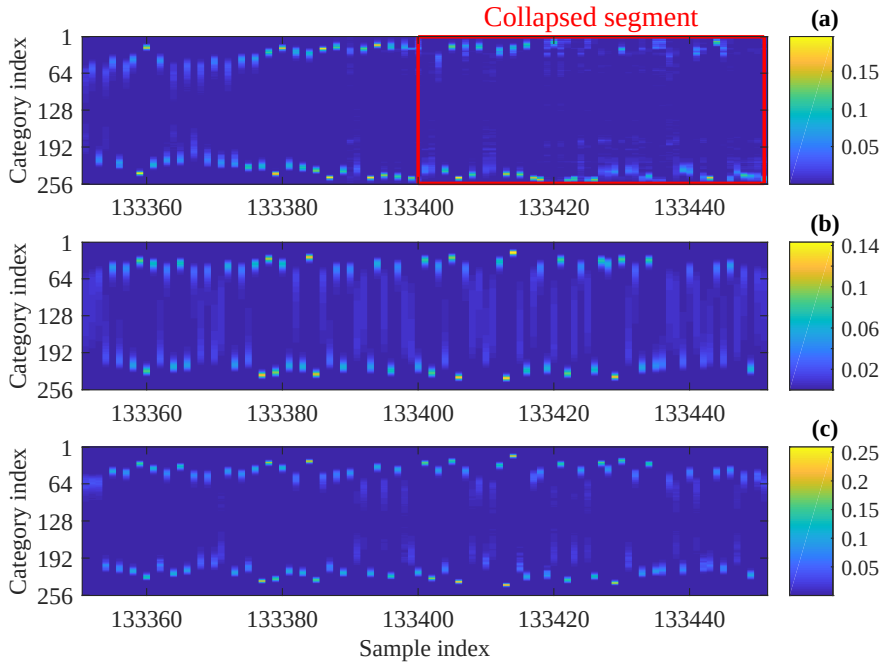


Figure 4.7: *(a) Predicted PMF sequence of WN w/ Type I collapsed speech. (b) PMF sequence from LPCDC. (c) Modified PMF sequence of WN w/ LPCDC and CSSD.*

4.6.1 Experimental Setting

Corpus and Acoustic Feature

For the evaluations, three English corpora, VCC2018 [115], CMU-ARCTIC [128], and an internal TTS corpus, were adopted. The details of the VCC2018 and internal TTS corpora can be found in Section 3.3.1, and both of them were set to a 22,050 Hz sampling rate and 16-bit quantization. To advance the WN voder capacity, the partial CMU-ARCTIC corpus was involved in the WN vocoder training. Since the sampling rate of most data of the CMU-ARCTIC corpus was 16 kHz, and only speakers "bdl" and "slt" had 32 kHz data, only these two speakers' data were adopted for the requirement of the target 22,050 Hz sampling rate. Speaker "bdl" had 1131 utterances and speaker "slt" had 1132 utterances. All data of speakers "bdl" and "slt" were down-

sampled from 32 kHz to 22,050 Hz, and the quantization number was also 16 bits. The WORLD acoustic features mentioned in Section 3.3.1 were adopted for the non-parallel VC models and the WN vocoders. The auxiliary features of the WN vocoders included mcep, coded ap, interpolated continuous F0, and a voice/unvoice binary code. Moreover, the LPC coefficients for the LPCDC were 30-dimensional with 20 ms frame length and 5 ms frameshift.

Architecture and Hyperparameter

A standard WN architecture [12] was adopted. Each WN vocoder included 30 residual blocks, and the dilated and 1×1 convolutions in each residual block had 512 channels. The dilation sizes of these dilated convolutions were set to 2^0–2^9 with three cycles (one cycle included 10 residual blocks). The 1×1 convolutions between the skip connections and softmax of a WN vocoder had 256 channels. The number of trainable parameters of a WN vocoder was 44 million. A multi-speaker WN vocoder was first trained based on the training data of all VCC2018 speakers and speakers "bdl" and "slt" of the CMU-ARCTIC corpus. Four speaker-dependent (SD) WN vocoders were fine-tuned by updating the output layers of the multi-speaker WN vocoder with the training data of the corresponding target speakers. The number of training iterations was 200,000 and the training learning rate was initially 0.001 with 50 % decay per 50,000 iterations. The number of updating iterations was 50,000 and the updating learning rate was 0.001 without decay. The mini-batch size was one and the batch length is 20,000 samples. Adam [117] was adopted for optimization. Furthermore, the noise shaping (NS) technique [74] was applied to the WN vocoders.

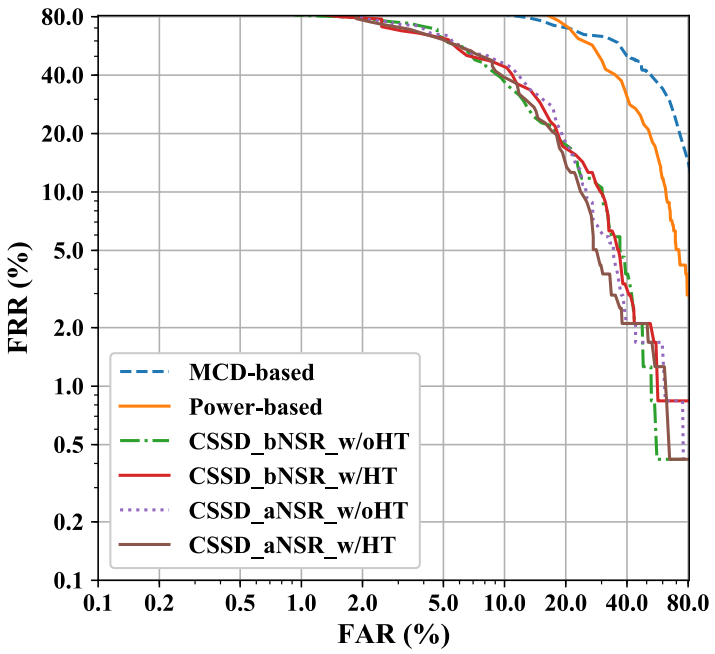


Figure 4.8: *DET curve for overall collapsed speech detection.*

4.6.2 Collapsed Speech Detection Evaluation

To evaluate the performance of the proposed CSSD, a human-labeled test set of the VCC18 SPOKE task was adopted. The test set was established using DNN-based non-parallel VC models with SD WN vocoders. The number of speaker pairs of the SPOKE task was 16, so the total number of utterances in this test set was 560. According to the labeled results, 46 utterances suffered from the Type I collapsed speech problem and 276 utterances had the Type II short impulse noise. Although more than 50 % of the utterances suffered from the collapsed speech problem, some utterances with the Type II short impulse noise did not cause perceptual degradation. This is because the label criterion was only based on the waveform shape and the perceptual loss was not considered.

Collapsed speech detection was formulated as a verification problem in the evaluation. The detection performance was measured via the false acceptance rate (FAR) and false rejection rate (FRR). Specifically, the rate of the collapsed utterances that were not detected was FAR, and the rate of the normal utterances that were detected was FRR. As shown in Fig. 4.8, the proposed CSSD method was compared with a mel-cepstrum distortion (MCD)-based detection method and the power-based detection method described in Section 4.3. Both MCD- and power-based methods conducted utterance-based detections based on the MCD/power differences between the generated and reference utterances. Because we adopted the NS technique for the WN vocoder and the HT for waveform envelope extraction, four CSSD variants of the waveform envelope extraction before (bNSR) and after NS restoration (aNSR) and with (w/ HT) and without the HT (w/o HT) were considered in the evaluation.

The detection performance of all utterances including the Type I and II collapsed speech segments is shown in Fig. 4.8, and the results show that the proposed CSSD significantly outperforms the MCD- and power-based methods. Because of the known weakness of utterance-based detection methods for Type II collapsed speech detection, we also present the results of the utterances only including Type I collapsed speech segments in Fig. 4.9. However, the CSSD-series methods still achieve a lower equal error rate (EER), especially the methods with the HT. In conclusion, the experimental results confirm the effectiveness of the proposed CSSD with the HT, which detects Type I collapsed speech segments with an EER lower than 5 % and both Type I and Type II collapsed speech segments with an EER of 20 %. Because of the convenience of implementation and the similar detection performance, the following evaluations were conducted on the system with the CSSD applied with the HT before the NSR.

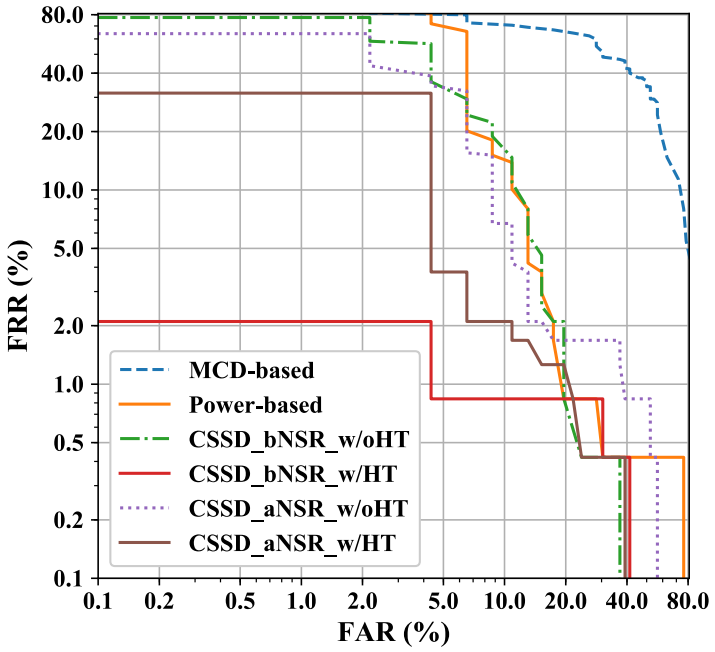


Figure 4.9: *DET curve for Type I collapsed speech detection.*

4.6.3 Subjective Evaluation

To evaluate the perceptual performance of the proposed system, we conducted a speech quality evaluation measured by a mean opinion score (MOS) and a speaker similarity evaluation measured by a similarity score [115]. In this subsection, both DNN- and DMDN-based non-parallel VC models mentioned in Section 3 were trained with a non-parallel corpus, which took the speakers of the SPOKE set as the sources and the four target speakers of the HUB set as the targets to form 16 speaker pairs. The demo utterances can be found on our demo page [1].

[1] https://bigpon.github.io/LpcConstrainedWaveNet_demo/

Speech Quality

Three sets, upper bound, collapse-free, and collapsed, including 10 systems were evaluated in the MOS test for speech quality. The upper bound set included natural speech and the WN vocoder conditioned on the target natural acoustic features. For the collapse-free and collapsed sets, utterances of DNN- and DMDN-based non-parallel VC were first generated by the WN vocoders and then partitioned into the corresponding sets by the CSSD results. Collapsed utterances were detected in 377 of 560 utterances generated by the DNN-based system, and the number of detected collapsed utterances of the DMDN-based system was 335. Possible reasons for the high ratio of collapsed utterances were the 20 % EER of the CSSD and the unoptimized threshold. Furthermore, the utterances of the DMDN-based VC generated by the WN vocoder with only LPCDC and the WORLD vocoder were also included in this evaluation as a control group. Since the proposed system only applied the LPCDC to the CSSD detected segments, only the collapsed set included the results of the DMDN-based VC with the WN vocoder, LPCDC, and CSSD (the proposed system).

An evaluation set including 800 ($5 \times 16 \times 10$) utterances was collected by randomly selecting five utterances of each system and speaker pair. The evaluation set was evenly portioned into five subsets, and each subset was evaluated by three listeners with the same device in a quiet environment. Although the 15 listeners were not native English speakers, most of them had worked on speech or audio generation research. The speech quality of each utterance was evaluated by the listeners, who assigned a MOS of 1–5, where the higher the MOS, the higher the speech quality of the utterance.

As shown in Fig. 4.10, although the synthesized speech of the WN vocoder suffers from a slight speech quality degradation, it still achieves a MOS of 4.5, which confirms the effectiveness of the WN vocoder for generating high-fidelity speech. For

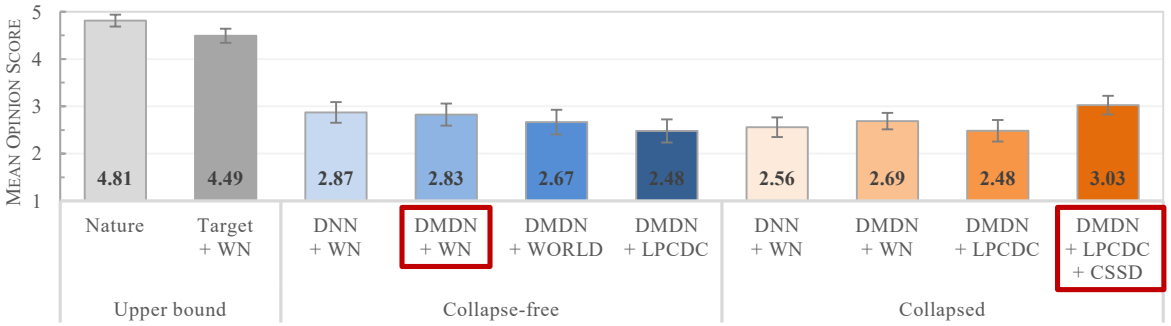


Figure 4.10: *MOS evaluation of speech quality with 95 % CI. (The performance of the proposed system is the combination of DMDN + WN in the collapse-free set and DMDN + LPCDC + CSSD in the collapsed set.)*

the collapse-free set, the results also show the effectiveness of the vanilla WN vocoder (DMDN + WN) for achieving higher speech quality than the WORLD vocoder (DMDN + WORLD). However, the results of the collapsed set indicate that the collapsed speech significantly degrades the speech quality for both the DNN- and DMDN-based systems. We also find that the WN vocoder (DMDN + LPCDC), which always applies the LPCDC, also suffers from a severe speech quality degradation. The same MOSs of the collapse-free and collapsed sets generated by the DMDN-based system with the WN vocoder applying the LPCDC imply that although the LPCDC alleviates the collapsed speech problem, it causes extra speech quality degradation.

However, when the LPCDC was applied to only the CSSD detected segments, the negative effect of the LPCDC was well limited into very short periods. The system with the LPCDC and CSSD in the collapsed set attains a similar MOS to the systems with the vanilla WN vocoder for the collapse-free set. The results (DMDN + LPCDC + CSSD) show that the proposed system not only markedly alleviates the collapsed speech problem but also prevents the WN vocoder from speech degradation caused by applying the LPCDC. In conclusion, the proposed LPCDC and CSSD modules significantly

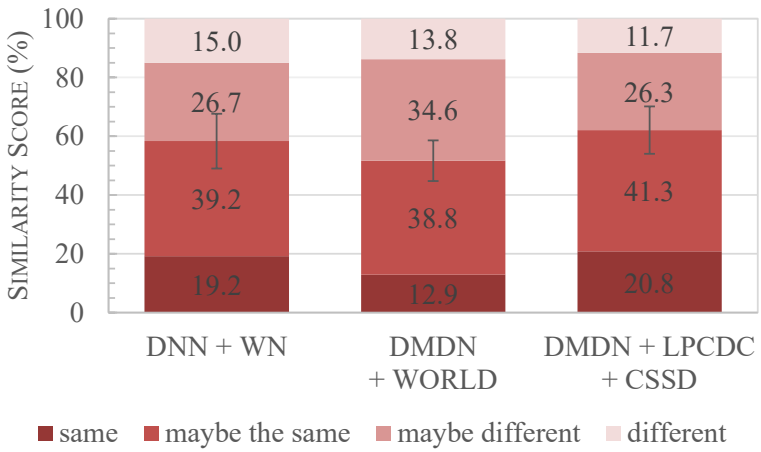


Figure 4.11: *Speaker similarity evaluations with 95 % CI of the same speaker (same and maybe the same) and different speakers (different and maybe different).*

alleviate the collapsed speech problem of the WN vocoder while maintaining a similar speech quality.

Speaker Similarity

A speaker similarity test on the proposed system, the DNN-based system with the vanilla WN vocoder, and the DMDN-based system with the WORLD vocoder was conducted. The listeners, devices, and environment were the same as in the MOS test. The same evaluation set including five subsets was adopted, and each subset was also evaluated by three listeners. The similarity measurement was the same as that in Section 3.3.3. Each listener was asked to determine whether the speakers of the natural and converted utterances are the same or not, and the final similarity score is the sum of the definitely the same and maybe the same scores. As shown in Fig. 4.11, the proposed method achieves a higher speaker similarity than the DMDN-based system with the WORLD vocoder, which is consistent with previous comparisons with the WN

and WORLD vocoders [47, 48]. Moreover, the proposed system also attains a similar speaker similarity to the DNN-based system with the WN vocoder, which confirms that the proposed LPCDC with the CSSD can simultaneously ease the collapsed speech problem, greatly alleviate the negative effect of the LPCDC, and maintain the same speaker similarity as the vanilla WN vocoder without the collapsed speech problem

Comparison with NU VCC2018 System

Our non-parallel VC system submitted to VCC2018 (NU non-parallel VC system) [43] was a DNN-based non-parallel VC system with the vanilla WN vocoder. The collapsed speech utterances were detected using power differences between the generated and reference utterances, which achieves a lower accuracy than the CSSD method with waveform envelope detection. Moreover, the LPCDC was applied to the whole collapsed speech utterance, which also caused speech quality degradation. However, the system still attained second place in the speaker similarity test and above-average speech quality as shown in Section 3.3.3. In this chapter, a DMDN-based system with the proposed CSSD and LPCDC is introduced. Since the system applies the LPCDC to only the CSSD detected collapsed speech segments, the speech degradation caused by the LPCDC was greatly alleviated. The experimental results also confirm the effectiveness of the proposed CSSD and LPCDC modules. To sum up, the proposed system in this chapter obviously outperforms the previously submitted system.

4.7 Summary

In this chapter, we first introduce the collapsed speech problem of the WN vocoder, and then the proposed collapsed speech detection and suppression framework are pre-

sented. The collapsed speech is caused by the prediction errors of the WN vocoder, and the possible reasons are limited training data, conditioned on distorted features, and exposure bias. The WN prediction errors result in waveform-domain discontinuity causing noisy voices, and the phenomena can be observed in the WN-predicted PMF sequences. Moreover, significant quality degradation caused by the collapsed speech is also shown in the subjective results.

An utterance-based detection method using a maximum power difference criterion and a segmental detection method using a waveform envelope difference criterion, CSSD, are introduced. Because of the segmental detection and the straightforward criterion, the proposed CSSD outperforms the power-based method. Moreover, the segmental detection also makes the system apply the proposed collapsed speech suppression method to only the detected segments, which greatly restricts the negative effects of the suppression method in short periods.

An LPC coefficients-derived distribution constraint, LPCDC, is introduced to suppress the collapsed speech. The prior speech continuity knowledge is extracted from the WORLD-generated reference speech and applied to the WN-predicted distribution. Although the LPCDC also introduces extra degradation, the LPCDC combined with the CSSD still significantly improves the speech quality of the utterances suffering the collapsed speech problem.

The proposed LPCDC combined with the CSSD was applied to our non-parallel VC system submitted to VCC2018, and the subjective results show that the advanced system in this chapter outperforms the submitted system. Since the proposed framework can be applied to any predicted probability distribution, the framework might be easily extended to other AR neural vocoders such as WaveRNN [15], which sequentially predicts the probability distribution of each speech sample. Furthermore, the LPC codec

of the proposed distribution constraint also can be easily replaced by other advanced speech coding techniques.

To summarize, in this chapter, the evaluation results show that directly combining the WN vocoder with a VC model sometimes causes the collapsed speech problem. To improve the robustness of the baseline VC system, a waveform-based constraint has been introduced to the WN vocoder to tackle the collapsed speech problem. The proposed method adopts the prior knowledge of the speech continuity and the stability of the WORLD-generated speech to design the detection and suppression techniques. The evaluation results also show the improved speech quality of the WN-generated VC utterances adopting the proposed LPCDC and CSSD, which indicates the robustness of the WN vocoder against the distorted acoustic features has been enhanced by the proposed techniques. Furthermore, besides the robustness of the WN vocoder, another essential feature, controllability of the speech components, of a vocoder will be explored in the next chapter.

5 Quasi-Periodic WaveNet for Audio Waveform Generation

In Chapter 3 and 4, the proposed baseline speaker voice conversion (VC) system, the instability of the WaveNet (WN) vocoder combined with the VC model, and the techniques for easing the unstable problem were introduced. In addition to the VC scenario and the robustness enhancement of the WN vocoder, the pitch controllability, which is an essential feature of a vocoder, and the pitch transformation scenario will be presented in this chapter. Specifically, the proposed collapsed speech detection and suppression techniques in Chapter 4 focused on only the speech continuity and avoiding the collapsed speech problem, but the match of the generated speech and the input VC acoustic features was not considered. However, a vocoder should precisely generate speech according to the changes of the input acoustic features. In this chapter, the insufficient pitch controllability of the WN vocoder and the improved WN vocoder with pitch-dependent architectures will be introduced. Moreover, the performance of the improved WN vocoder combined with the baseline VC model will also be evaluated.

5.1 Introduction

Since audio waveform is a sequential signal with extremely high temporal resolution (sampling rates are usually higher than 16 kHz), directly modeling the long term dependence of audio waveform is challenging. Conventional audio analysis and synthesis

techniques, which are called the vocoder [29–31], usually encode audio into low temporal resolution acoustic features and decode audio waveforms on the basis of these acoustic features. However, because of the lost temporal details and phase information during the analysis and synthesis, conventional vocoders [35, 37] usually suffer from buzz noise and naturalness degradation.

Owing to the recent development of deep learning, many neural-based audio generation models [12–20, 68–70] have been proposed to directly model raw audio waveforms without many over-simplified assumptions of speech generation imposed on conventional vocoders. In this chapter, we focus on the state-of-the-art audio generation model, WN [12]. The core of WN is a convolutional neural network (CNN)-based autoregressive (AR) network modeling the probability distribution of each audio sample conditioned on auxiliary features and a fixed number of previous samples called a receptive field. A variety of applications such as music generation [129], text-to-speech (TTS) [11, 130], speech coding [131], speech enhancement [132, 133], and voice conversion [43, 44, 46] have adopted WN. Furthermore, taking WN as a vocoder [47–49, 74] to generate speech waveforms conditioned on conventional-vocoder-extracted acoustic features greatly ease the lost information problem of conventional vocoders.

Although WN achieves high-fidelity speech generation, the fixed architecture without prior knowledge of audio periodicity is inefficient and limits the pitch controllability of the WN vocoder. For instance, because of the quasi-periodicity of speech, each sample may have a specific dependent field related to its periodicity instead of a fixed receptive field that presumably includes many redundant previous samples. Furthermore, the data-driven architecture without prior periodic knowledge only implicitly models the relationship between the periodicity of waveform signals and the auxiliary fundamental frequency (F_0) features, which may not explicitly generate speech with the precise

pitch corresponding to the auxiliary F_0 values, especially in an unseen F_0 case [38, 39]. However, the pitch controllability is an essential feature for the definition of a vocoder.

To address these problems, we proposed Quasi-Periodic WaveNet (QPNet) [38, 39] with a pitch-dependent dilated convolution neural network (PDCNN) inspired by the source–filter model [34] and code-excited linear prediction (CELP) codec [32, 33]. Specifically, the generation process of periodic signals can be modeled as the generation of a single periodic cycle signal (short-term correlation) and then extending this single cycle signal to form the whole periodic sequences based on pitches (long-term correlation). Therefore, QPNet including two cascaded networks has been proposed. The first network is vanilla WN with a fixed network architecture modeling the short-term correlations of the nearest samples within one periodic cycle. The second network is a pitch-dependent WN with an adaptive network architecture utilizing the PDCNNs to link the correlations of the relevant segments in the current and previous periodic cycles. With the cascaded network structure, the proposed QPNet achieves higher pitch controllability and maintains similar speech quality with a more compact model.

This chapter is organized as follows. In Section 5.2, a brief introduction of WN and the limitations of the WN vocoder is presented. In Section 5.3, the concepts and details of the QPNet are described. In Sections 5.4, the effectiveness of the QPNet for generating high-temporal-resolution periodic sinusoid signals was evaluated. In Section 5.5, objective and subjective tests were conducted to evaluate the speech generation with pitch manipulations performance. In Section 5.6, the results of the QPNet combined with VC are reported. Finally, the conclusion is given in Section 5.7.

5.2 WaveNet and Limitations of WaveNet Vocoder

WaveNet [12] is a CNN-based AR model sequentially predicting the probability distribution of each waveform sample conditioned on a fixed number of previous samples and auxiliary features. The probability distribution is formulated as

$$P\left(\boldsymbol{x} \mid \boldsymbol{h}\right) = \prod_{t=1}^{T} P\left(x_t \mid x_{t-1}, \ldots, x_{t-r}, \boldsymbol{h}\right) \tag{5.1}$$

where t is the sample index, x_t is the current audio sample, $\boldsymbol{h}$ denotes the auxiliary features, and the receptive field length is r. Since the model capacity is highly related to the length of the receptive field, WN applies stacked dilated CNNs (DCNN) [75] to efficiently attain a large receptive field. Furthermore, taking WN as a vocoder [47, 48] to generate speech conditioned on the auxiliary acoustic features extracted by conventional parametric-based vocoders also achieves marked speech quality improvements.

However, the core of a vocoder is the controllability of each speech component such as pitch. Although the WN vocoder achieves high-fidelity speech generation, the WN vocoder lacks pitch controllability. Specifically, when conditioned on the F_0 values that are not observed in the F_0 range of training data, the WN vocoder usually has difficulties in generating speech with precise pitch [38,39]. Moreover, even if the F_0 and spectral features are within the observed range, an unseen combination of the auxiliary features still markedly degrades the generation performance of the WN vocoder [27, 28, 43, 44, 46]. The possible reasons for this problem are that WN lacks prior speech knowledge and does not explicitly model the relationship between the auxiliary F_0 feature and pitch. Moreover, because of the fixed architecture of the WN, each sample has the same receptive field length. However, making each sample have a specific receptive field length corresponding to its pitch is more reasonable. The inefficient receptive field extending of the WN may lead to the costly requirements of a huge

network and lots of computation power.

5.3 Quasi-Periodic WaveNet

To improve speech modeling efficiency and pitch controllability, the proposed QPNet introduces the prior knowledge of speech periodicity into WN by dynamically changing the network architecture according to the auxiliary F_0 features. As shown in Fig. 5.1, the main differences between WN and QPNet are the pitch-dependent dilated convolution mechanism handling the periodicity of audio signals and the cascaded fixed and adaptive residual blocks simultaneously modeling the long- and short-term correlations of speech samples. The pitch filtering in CELP, which is the basis of the PDCNN, and the details of QPNet are described as follows.

5.3.1 Pitch Filtering in CELP

As shown in Fig. 5.2, the CELP system [32, 33] includes an innovation signal codebook and two cascaded time-varying linear recursive filters, and the speech modeling is formulated into three steps. First, each innovation signal in the codebook is scaled and passed to the pitch filter (long delay) to generate the pitch periodicity of the speech, and then the linear-prediction filter (short delay) restores the spectral envelope to obtain the synthesized speech. Secondly, the mean-square errors between the original and synthesized speech signals are weighted by a linear filter to attenuate/amplify frequency components that are less/more perceptually important. Finally, the optimum innovation signal and the scaled factor are determined by minimizing the weighted

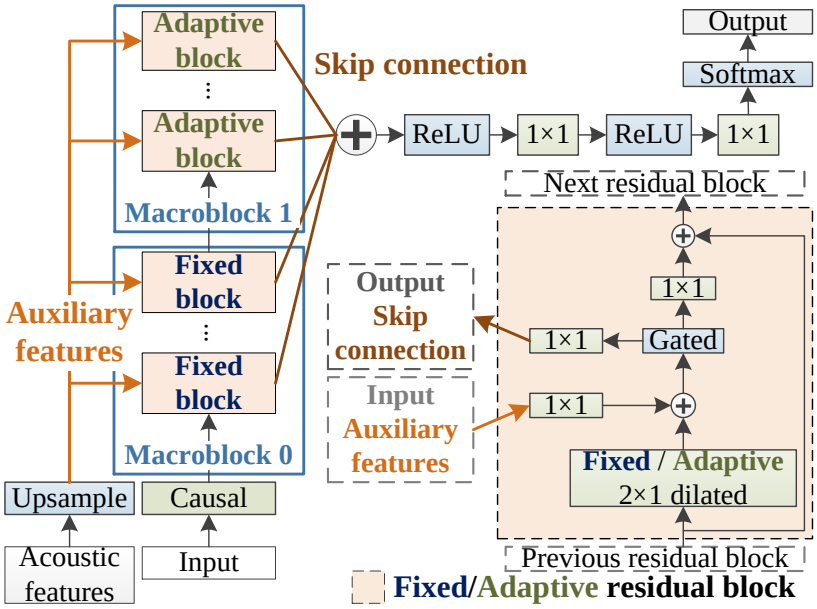


Figure 5.1: *QPNet vocoder architecture.*

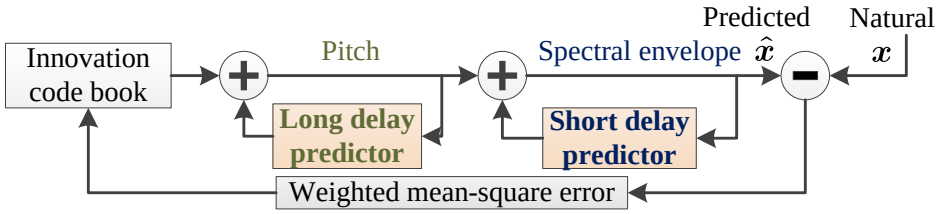


Figure 5.2: *Code-excited linear prediction system.*

mean-square error. Specifically, the pitch-filtering process can be formulated as

$$c_t^{(\mathrm{o})} = g \times c_t^{(\mathrm{i})} + b \times c_{t-t_d}^{(\mathrm{o})} \tag{5.2}$$

where $c^{(\mathrm{i})}$ is the input, $c^{(\mathrm{o})}$ is the output, t_d is the pitch delay, g is the gain, and b is the pitch filter coefficient. This periodic feedback structure handling speech periodicity is the basis of the proposed PDCNN, and the cascaded recursive structure modeling hierarchical speech information is also applied to QPNet.

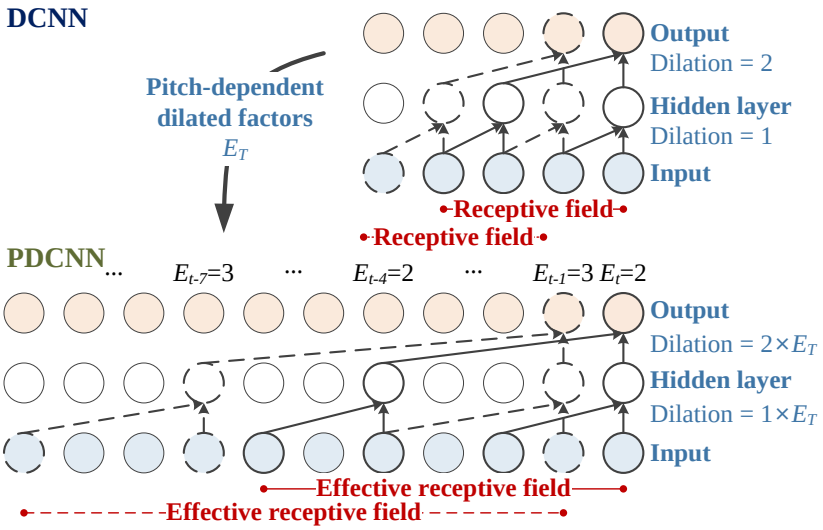


Figure 5.3: *Fixed and pitch-dependent dilated convolution.*

5.3.2 Causal Pitch-dependent Dilated Convolution

The main idea of the PDCNN is that since audio signals have the quasi-periodic property, the network architecture can be dynamically adapted using the prior periodic information. Specifically, a causal convolution with a size two kernel can be formulated as

$$\boldsymbol{y}_t^{(\mathrm{o})} = \boldsymbol{W}^{(\mathrm{c})} \times \boldsymbol{y}_t^{(\mathrm{i})} + \boldsymbol{W}^{(\mathrm{p})} \times \boldsymbol{y}_{t-d}^{(\mathrm{i})}, \tag{5.3}$$

where $\boldsymbol{y}_t^{(\mathrm{o})}$ is the output of the DCNN layer at sample t, $\boldsymbol{y}_t^{(\mathrm{i})}$ is the input of the DCNN layer at sample t, and d is the dilation size. The trainable 1×1 convolution filters $\boldsymbol{W}^{(\mathrm{c})}$ and $\boldsymbol{W}^{(\mathrm{p})}$ are respectively for the current and previous samples. The dilation size of the vanilla CNN is set to one, and the dilation size of the DCNN is predefined and constant. However, the dilation size of the PDCNN is pitch-dependent and time-variant.

To efficiently enlarge the receptive field length, stacked chunks including DCNN layers with different dilation sizes are adopted in the vanilla WN. Each chunk contains a specific number of DCNN layers, and the dilation sizes of the DCNN layers in each

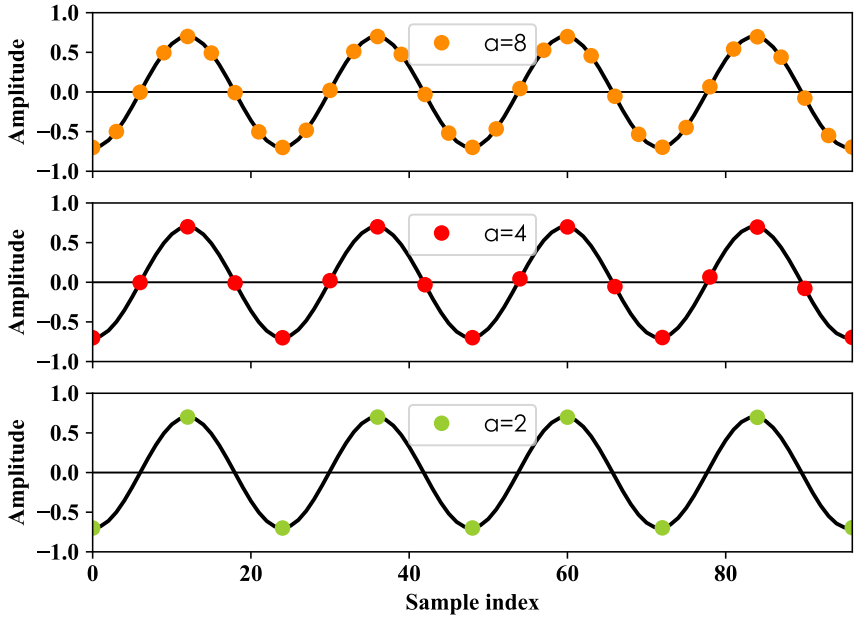


Figure 5.4: *Sampling sparsity of different dense factor a.*

chunk are exponentially increased with base two. As shown in Fig. 5.3, the dilation sizes of PDCNN layers in the stacked adaptive chunks of QPNet follow the same extension rule but multiplied by an extra dilated factor to match the instantaneous pitch of the current sample. The pitch-dependent dilated factor E_t is derived from

$$E_t = F_s/(F_{0,t} \times a), \tag{5.4}$$

where F_s is the utterance-wise constant sampling rate, $F_{0,t}$ is the fundamental frequency with speech sample index t, and a is a hyperparameter called the dense factor, which indicates the number of samples in one periodic cycle taken into consideration as shown in Fig. 5.4 when predicting the current sample.

Specifically, the grid sampling locations of each DCNN is controlled by the dilation size d, and the dilation size d' of each PDCNN is controlled by the dilated factor E_t as

$$d' = E_t \times d. \tag{5.5}$$

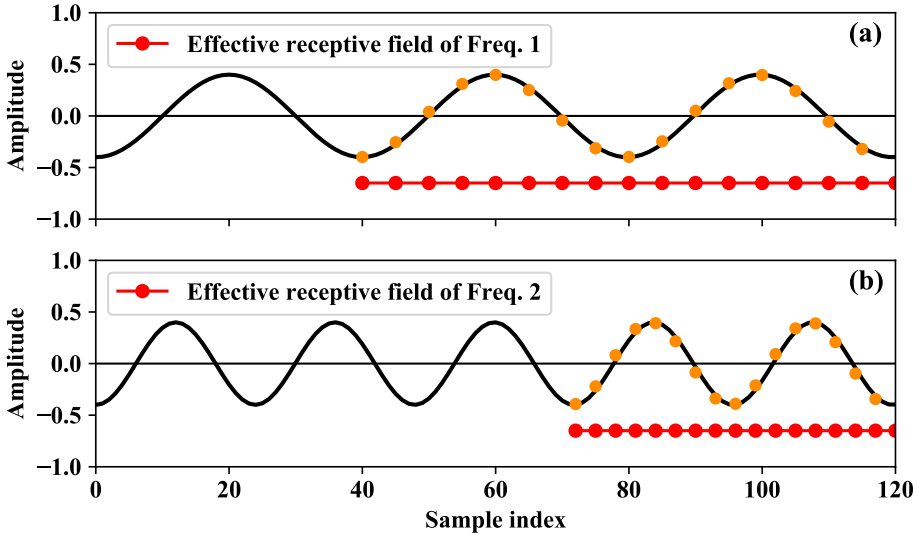


Figure 5.5: *Effective receptive fields with different F_0 values.*

The sparsity of the CNN sampling grids is controlled by set specific F_0 values and dense factor a to attain the desired effective receptive field length. As shown in Fig. 5.5, because of the same dense factors and sampling rates of the sinusoids in Fig. 5.5 (a) and (b), even though the frequencies of these sinusoids are different, their effective receptive fields still include the same numbers of periodic cycles. The difference is the temporal sparsity of the effective receptive field. The lower the frequency, the higher the sparsity. That is, fixing the number of sampling grids in each periodic cycle by the dense factor and changing the gaps between the grid sampling locations by the instantaneous F_0 values lead to pitch-dependent and time-variant effective receptive field lengths.

In summary, the dilated factor E_t is the ratio of the effective receptive field length to the receptive field length, and the ratio of the receptive field length to the dense factor a is the number of past period cycles in the effective receptive field. With the pitch-dependent structure, each sample has an exclusive effective receptive field

length, which is efficiently enlarged according to the auxiliary F_0 values. In addition, since speech has voiced and unvoiced segments, we have tried to set E_t to one or the value calculated by interpolating the F_0 values of the adjacent voiced segments for the unvoiced segments, and the results in Section 5.5 show that QPNet with the continuous E_t from interpolated F_0 values achieves higher speech quality.

5.3.3 Cascaded Autoregressive Network

Audio generative models are usually capable to simultaneously model the long-term (periodicity) and short-term (aperiodicity) correlations of audio samples because most audio signals are sequential and quasi-periodic. As shown in Fig. 5.1, a fixed macroblock and an adaptive (pitch-dependent) macroblock are adopted in the proposed QPNet. The fixed macroblock models the sequential relationship between the current sample and a segment of the most recent samples. The adaptive macroblock models the periodic correlations of the current and related past segments in the successive periodic cycles. Specifically, the fixed macroblock (macroblock 0 in Fig. 5.1) of the QPNet is composed of several fixed chunks. Each fixed chunk consists of several stacked residual blocks with DCNNs (fixed blocks), conditional auxiliary features, gated activations, and residual and skip connections, similarly to the vanilla WN. The adaptive macroblock (macroblock 1 in Fig. 5.1) also contains several adaptive chunks, which also have similar stacked residual blocks but with PDCNNs (adaptive blocks). In summary, the cascaded structure of QPNet presumably mimics a similar generative procedure of CELP for quasi-periodic audio signals generation.

Table 5.1: *Architecture of sinusoidal generative model*

	WNf	WNc	(r)QPNet	pQPNet
Fixed chunk	3	4	3	-
Fixed block	10	4	4	-
Adaptive chunk	-	-	1	4
Adaptive block	-	-	4	4
CNN channel (Causal and dilated CNN)		128		
CNN channel (CNN in residual block)		128		
CNN channel (CNN in output layer)		64		
Number of trainable parameter ($\times 10^6$)	2.4	1.5	1.5	1.5

5.4 Periodic Signal Generation Evaluation

To evaluate the pitch controllability of the proposed QPNet with the PDCNNs, generative evaluations of simple periodic but high-temporal-resolution signals were conducted. The training data of QPNet were sine waves within a specific frequency range and the corresponding F_0 values. In the test stage, QPNet was conditioned on an F_0 value and a small piece of the related sine wave for initializing the receptive field to generate sinusoid waveforms.

5.4.1 Experimental Setting

Model Architecture

Three types of QPNet with two types of WN were involved in this evaluation. Specifically, in addition to the basic QPNet, because a sinusoid is a simple periodic signal that can be modeled well by a pure pitch-dependent structure, the QPNet model with only adaptive residual blocks (pQPNet) was taken into account. The QPNet model with the reverse order of the fixed and adaptive macroblocks (rQPNet) was also con-

sidered. Moreover, a compact-size WN (WNc) and a full-size WN (WNf) models were evaluated as a control group.

The details of the network architectures are shown in Table 5.1. Since the numbers of CNN channels were the same for all models, the model sizes were proportional to the numbers of the chunks and residual blocks. For instance, the WNf contained 3 chunks and each chunk included 10 residual blocks, so the model size of the WNf was larger than that of the WNc, which only had 4 chunks with 4 residual blocks in each chunk. The learning rate was 1×10^{-4} without decay, the minibatch size was one, the batch length was 22,050 samples, the training epochs were two, and the optimizer was Adam [117] for all models.

Evaluation Setting

The pitch range of the training sine waves was set to be in the same range as most speech, which was 80–400 Hz with a step size of 20 Hz (ex: 80, 100, 120 ... Hz). A related one-dimensional F_0 value was adopted as the auxiliary feature for each model. Since the single-tone generation was evaluated, the auxiliary features of all samples in one utterance were the same. To prevent the networks from suboptimal training and lacking the generality for sinusoid generations with unseen F_0 values, both sinusoid and auxiliary signals were mixed with white noise.

The signal-to-noise ratio (SNR) of the sine waves was around 20 dB, and the noise of the auxiliary feature was a random sequence between -1 and 1. Random initial phases were also applied to the sinusoid signals. The number of training utterances was 4000, and each utterance was one second. The ground truths were clean sinusoid signals, so each model was trained as a denoising network. The test data included 20 different F_0 values, which were 10–80 Hz with a step size of 10 Hz, 100–400 Hz with a step size of

100 Hz, and 450–800 Hz with a step size of 50 Hz, and each F_0 value contained 10 test utterances with different phase shifts. Both training and test data were encoded using the μ-law into 8 bits, and the sampling rate was 22,050 Hz.

In the test stage, the initial receptive field of each network was initilized with the noisy test sine wave, and the length of the generated sinusoid was set to 1s. The quality of each generated waveform was evaluated based on the SNR and the root-mean-square error (RMSE) of the log F_0 value measured from the peak of the power spectral density (PSD). Moreover, the test data were divided into 10–40 Hz (under $1/2L$), 50–80 Hz (above $1/2L$), 100–400 Hz (inside), 450–600 Hz (under $3/2U$), and 650–800 (above $3/2U$) subsets. L is the lower bound and U is the upper bound of the inside F_0 range, which was the F_0 range of the training data. As a result, the under $1/2L$ and above $1/2L$ F_0 ranges are the lower outside F_0 range, and the under $3/2U$ and above $3/2U$ F_0 ranges are the higher outside F_0 range.

5.4.2 Performance Measurement

The quality of each generated waveform was evaluated on the basis of the SNR and the root-mean-square error (RMSE) of the log F_0 value measured from the peak of the power spectral density (PSD). Specifically, because the SNRs are related to the noisy degrees of the generated signals, the SNR values will indicate the generated signals are clear sinusoids or not. Since it was a single-tone sinusoid generation test, the high log F_0 RMSEs might imply that the generated signals include much harmonic noise or the frequencies of these signals are incorrect. In other words, the generated signal with a high SNR and a high RMSE might be a clear sinusoid with an inaccurate frequency, the generated signal with a low SNR and a high RMSE might be a noisy sinusoid with much harmonic noise, and the generated signal with a very low SNR might be a

noise-like signal.

5.4.3 Experimental Result

The periodic signal generation evaluation includes two parts. First, to explore the efficient dense factor value of the PDCNNs, several pQPNet models with different dense factors were evaluated. Secondly, three types of QPNet with the most efficient dense factor according to the first evaluation and two types of WN were evaluated. The details are as follows.

Dense Factor

Since the chunk and block numbers of the pQPNets were set to four, the length of the receptive field was 61 samples. That is, the receptive fields included from 61 past periodic cycles to less than one periodic cycle according to the dense factors from 2^0 to 2^6. Moreover, in contrast to containing a fixed number of past cycles for sinusoids with arbitrary pitch, the receptive fields of the WNf contained 11 past cycles for 80 Hz sinusoids and 56 past cycles for 400 Hz sinusoids when the sampling rate was 22,050 Hz. As a result, the effective receptive fields of the pQPNet with a dense factor 2 already contained a comparative number of the past periodic cycles as the WNf. Since the pQPNets introduced prior periodicity knowledge into the network, the required number of the past cycles for modeling the sinusoids might be less than that of the WNf.

The number of training epochs of the pQPNet models with dense factors from 2^2 to 2^6 was two. For dense factors of 2^0 and 2^1, the pQPNets required at least 10 training epochs to attain stable results. As shown in Tables 5.2 and 5.3, even though

Table 5.2: *SNR (*dB*) of sinusoid generation with different dense factors*

Dense	2^0	2^1	2^2	2^3	2^4	2^5	2^6
Under $1/2L$	6.7	14.4	20.8	21.9	25.8	**28.0**	27.9
Above $1/2L$	19.8	11.9	21.5	26.6	24.5	**28.9**	26.4
Inside	17.1	19.1	19.4	26.0	**29.9**	23.2	17.5
Under $3/2U$	1.1	6.7	3.0	19.9	**23.2**	17.1	-17.7
Above $3/2U$	-8.1	-0.8	-0.3	2.7	**8.3**	3.0	-23.5
Average	7.3	10.3	12.9	19.4	**22.3**	20.0	6.1

the pQPNet with the dense factor of 2^0 was trained with 10 epochs, the network was still very unstable. The results indicate that although the small dense factor made the network have long effective receptive fields, the overbrief information of each past periodic cycle might make it difficult to capture audio information well. For the inside and lower outside F_0 ranges, the networks with dense factors greater than 2^1 achieved high SNR values. However, the performance of the network with a dense factor of 2^6 markedly degraded when the auxiliary F_0 values were in the higher outside F_0 range. The possible reason is that the PDCNNs of the network degenerated to DCNNs because the E_t became one when the dense factor was 2^6 and the F_0 values were higher than 350 Hz. Moreover, the log F_0 RMSE results show a similar tendency to the SNR results. The networks with dense factors of 2^0 and 2^6 achieved the lowest pitch accuracies while the networks with dense factors of 2^2 and 2^3 achieved the highest pitch accuracies.

Furthermore, according to the Nyquist–Shannon sampling theorem [134], a signal can be perfect reconstructed if the bandwidth of the signal is less than the halved sampling rate. Therefore, the dense factor 2^1 is theoretically enough to model the periodic signals. The instability and markedly high RMSE results of the pQPNet with dense factor 2^0 also confirm this theory. However, in signal processing, oversampling

Table 5.3: *Log F_0 RMSE of sinusoid generation with different dense factors*

Dense	2^0	2^1	2^2	2^3	2^4	2^5	2^6
Under $1/2L$	0.26	**0.00**	**0.00**	**0.00**	0.03	0.05	0.14
Above $1/2L$	0.00	0.01	**0.00**	**0.00**	0.01	0.01	0.10
Inside	0.42	**0.00**	**0.00**	0.01	0.01	0.02	0.03
Under $3/2U$	1.95	0.08	**0.03**	0.04	0.08	0.09	0.89
Above $3/2U$	0.61	**0.04**	0.05	0.06	0.09	0.15	1.97
Average	0.65	0.03	**0.02**	**0.02**	0.04	0.06	0.63

usually improves resolution and SNR, and relaxes filter performance requirements to avoid aliasing. The higher SNR and lower RMSE of the pQPNets with dense factor 2^2 and 2^3 have shown this tendency, and the performance degradation of the pQPNet with dense factor 2^6 is caused by the PDCNN degeneration issue, which is irrelevant to the sampling theorem.

In conclusion, the PDCNN with an appropriate dense factor was found to be robust against the conditions in the outside F_0 range, especially in the lower outside F_0 range conditions. For the higher outside F_0 range conditions, the networks still had acceptable quality until the F_0 value exceeded 600 Hz. Therefore, we set the dense factors to 2^3 for the models in the following evaluations because of the balance between the generative performance and the number of past periodic cycles covered in its receptive fields.

Network Comparison

As shown in Tables 5.4, the PDCNNs significantly improved pitch controllability. The PDCNNs made the QP-series networks achieve much higher SNR and lower log F_0 RMSE values than the same-size WNc network in both higher and lower outside F_0 ranges, and it shows the effectiveness of the PDCNNs to enlarge the effective receptive

Table 5.4: *SNR (*dB*) and Log F_0 RMSE of sinusoid generation with different models*

	WNc		WNf		pQPNet		QPNet		rQPNet	
	SNR	RMSE	SNR	RMSE	SNR	RMSE	SNR	RMSE	SNR	RMSE
Under 1/2 L	-18.1	2.93	**24.3**	1.75	21.9	**0.00**	-8.1	2.00	18.4	0.18
Above 1/2 L	8.1	0.55	23.0	0.58	26.6	**0.00**	28.2	0.02	**28.7**	**0.00**
Inside	28.8	0.01	**34.5**	**0.00**	26.0	0.01	25.9	0.01	27.0	**0.00**
Under 3/2 U	13.7	**0.04**	17.6	0.50	**19.9**	**0.04**	8.7	0.11	19.3	0.11
Above 3/2 U	-14.1	0.12	-0.4	0.48	**2.7**	**0.06**	-18.6	0.48	-8.2	**0.06**
Avg.	3.7	0.73	**19.8**	0.66	19.4	**0.02**	7.2	0.53	17.0	0.07

field length. Although the full-size WNf attained similar SNRs to the pQPNet, the log F_0 RMSE of WNf was much higher in the outside F_0 ranges. The results indicate that the WNf tended to generate the signals in the inside F_0 range instead of being consistent with the auxiliary F_0 feature. That is, the generated waveform of the WNf might still be a perfect sinusoid signal but with an incorrect pitch. The results also imply that the PDCNNs improved the periodical modeling capability using prior periodicity knowledge. Furthermore, because of the simple periodic signal generation scenario, the pQPNet with the longest effective receptive fields and the pure PDCNN structure attained the best generative performance among all QP-series networks. The QPNet and the rQPNet showed some quality degradations when the auxiliary F_0 values were far away from the inside F_0 range, but they still outperformed the WNc in both measurements and the WNf in terms of log F_0 RMSE.

5.4.4 Discussion

To further explore the physical phenomena behind the objective results, several sinusoid generation examples are presented. As shown in Figs. 5.6 (a) and (b), the

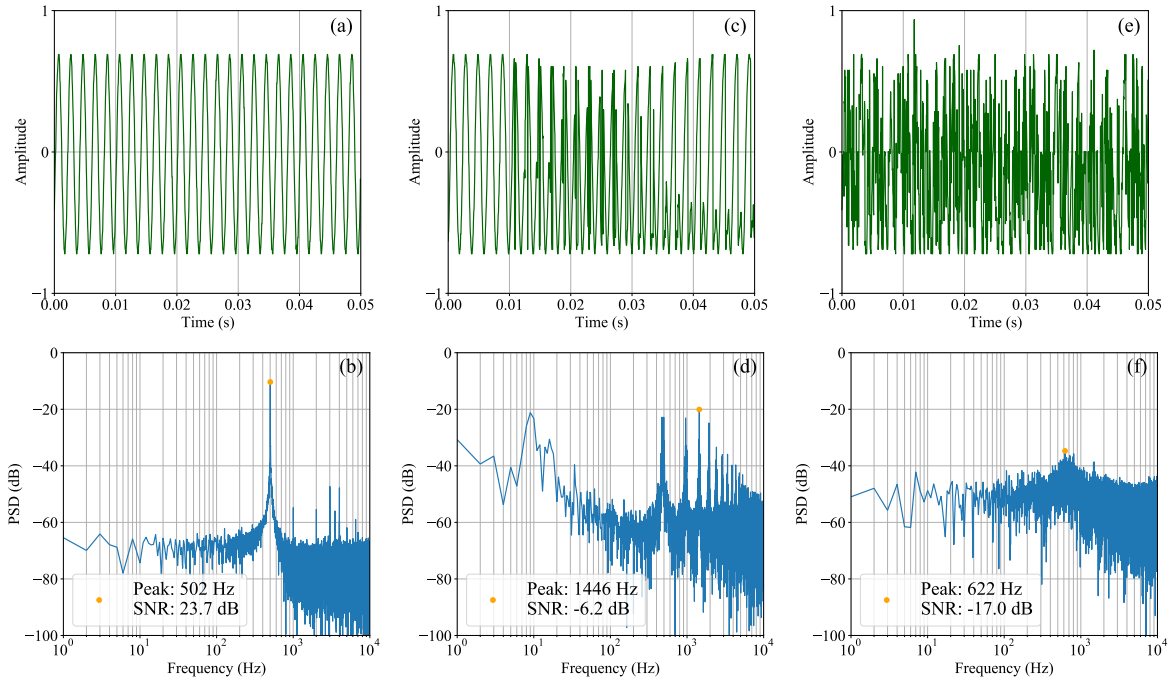


Figure 5.6: *Waveform and PSD of 500 Hz sinusoid generated by pQPNets with dense factors (a, b)* 2^3*, (c, d)* 2^0*, and (e, f)* 2^6*.*

pQPNet with a dense factor 2^3 generated clear sine waves with an SNR 23.7 dB when conditioned on an outside auxiliary value of 500 Hz (under $3/2U$). The peak value of the PSD of the pQPNet-generated signal is 502 Hz, which is very close to the ground truth, and the log F_0 RMSE is less than 0.01. However, the results in Figs. 5.6 (c) and (d) show that the sine wave generated by the pQPNet with a dense factor 2^0 includes much harmonic noise, which results in a low SNR. Even if the generated sine wave is still like a periodic signal, the wrong peak value from the second harmonic component of the PSD also causes a high log F_0 RMSE. Moreover, the results in Figs. 5.6 (e) and (f) show that the pQPNet with a dense factor 2^6 generated a very noisy signal, which results in a low SNR and an incorrect peak value of its PSD.

In addition, as shown in Figs. 5.7 (a) and (b), the pQPNet with a dense factor 2^3

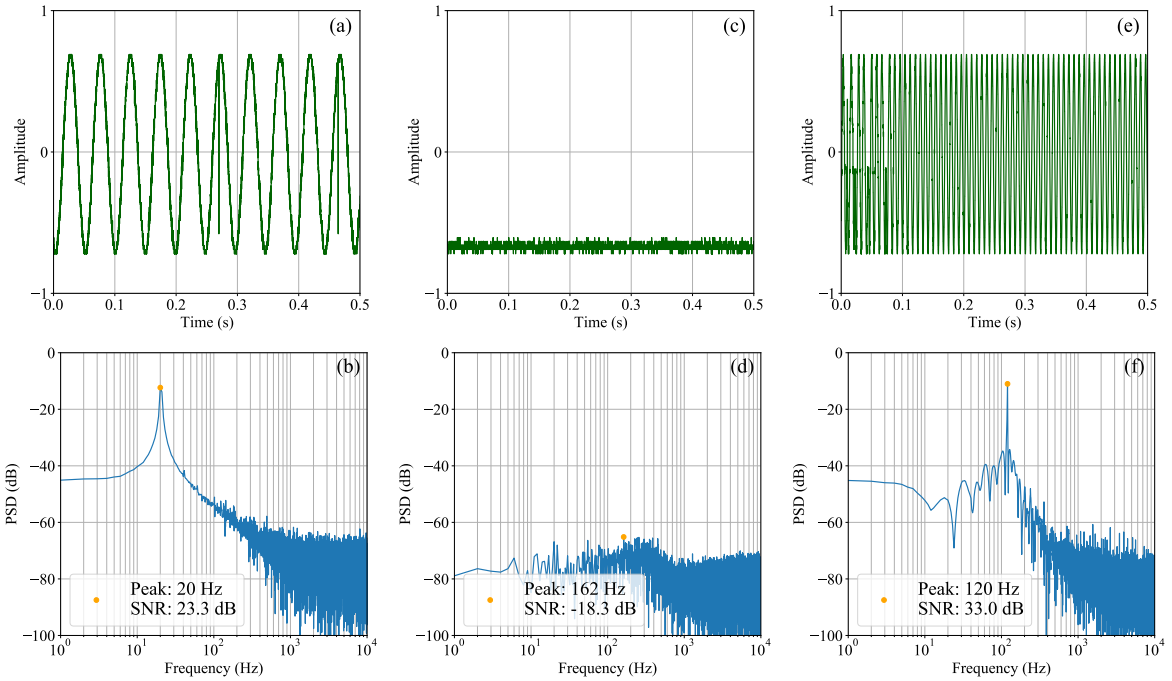


Figure 5.7: *Waveform and PSD of 20 Hz sinusoid generated by (a, b) pQPNet with a dense factor* 2^3 *(c, d) WNc, and (e, f) WNf.*

still generated a clear sine wave with an SNR 23.3 dB and a correct peak value of its PSD when conditioned on an outside 20 Hz (under $1/2L$) auxiliary value. However, the same-size WNc could not generate any meaningful signal, and the SNR of the WNc-generated signal is very low as shown in Figs. 5.7 (c) and (d). on the other hand, the WNf still generated a clear sine wave with an SNR 33 dB but its frequency is incorrect as shown in Figs. 5.7 (e) and (f). Specifically, the PSD peak value is 120 Hz, and it implies that the WNf tends to generate seen signals even if conditioned on an unseen auxiliary feature.

The results confirm our assumptions that the high SNR and RMSE signal like Fig. 5.7 (e) is a clear sinusoid with an inaccurate frequency, the low SNR and high RMSE signal like Fig. 5.6 (c) is a noisy sinusoid with much harmonic noise, and the very low

SNR signal like Figs. 5.6 (e) or 5.7 (c) is a noise-like signal. More results of different frequencies can be found on our demo page[1].

5.5 Speech Generation Evaluation

In this section, the effectiveness of the PDCNNs for speech generation are evaluated. The appropriate proportions of adaptive and fixed residual blocks, the continuous pitch-dependent dilated factor, and the order of the macroblocks are explored.

5.5.1 Experimental Setting

Model Architecture

Three types of vocoder, WN, QPNet, and WORLD (WD), were involved in the speech generation evaluation, and the total number of all vocoder variants was 11. First, to explore the efficient receptive field extension by the PDCNNs, the compact-size QPNet vocoders were compared with the same-size WNc and double-size WNf vocoders. Secondly, eight variants of QPNet were adopted. Four compact-size QPNet models and four full-size QPNet (QPNetf) models, which were full-size WN vocoders cascaded with four extra adaptive residual blocks, were also taken into consideration to explore the effect of the ratio of adaptive to fixed residual blocks. Both fixed-to-adaptive (QPNet) and reversed adaptive-to-fixed (rQPNet) macroblock orders were evaluated. Moreover, continuous and discrete E_t sequences were also explored. For the unvoiced frames, the discrete E_t sequence was set to ones, and the continuous E_t sequence was calculated using interpolated F_0 values as mentioned in Section 5.3.2. Last, the conventional WD vocoder was adopted as a reference.

[1] https://bigpon.github.io/QuasiPeriodicWaveNet_demo/

Table 5.5: *Architecture of speech generative model*

	WNf	WNc	(r)QPNet	(r)QPNetf
Fixed chunk	3	4	3	3
Fixed block	10	4	4	10
Adaptive chunk	-	-	1	1
Adaptive block	-	-	4	4
CNN channel (Causal and dilated CNN)		512		
CNN channel (CNN in residual block)		512		
CNN channel (CNN in output layer)		256		
Number of trainable parameter ($\times 10^6$)	44	24	24	50

The network architectures and model sizes are shown in Table 5.5. The learning rate was 1×10^{-4} without decay, the minibatch size was one, the batch length was 20,000 samples, and the optimizer was Adam [117] for all models. Since even the compact-size WNc had tens of millions parameters, which was the same order of magnitude as that of WNf, the training iterations were empirically set to 200,000 for all models. Note that we did not evaluate speech generation using the pQPNet model because it failed to model the short-term correlation of speech according to our internal experiments.

Evaluation Setting

All models were trained in a multispeaker manner. The training corpus of these multispeaker NN-based vocoders consisted of the training sets of the "bdl" and "slt" speakers of CMU-ARCTIC [128] and all speakers of voice conversion challenge 2018 (VCC2018) [115]. The total number of training utterances was around 3000, and the total training data length was around three hours. The evaluation corpus was composed of the SPOKE set of VCC2018, which included two female and two male speakers, and each speaker had 35 test utterances. All speech data were set to a sampling rate of

22,050 Hz and a 16-bit resolution. The waveform signals for the categorical output of the NN-based vocoders were further encoded into 8 bits using the μ-law. The 513-dimensional spectral (*sp*) and aperiodicity (*ap*) and one-dimensional F_0 features were extracted using WD. The *sp* feature was further parameterized into 34-dimensional *mcep*, *ap* was coded into two-dimensional components, and F_0 was converted into continuous F_0 and the voice/unvoice (U/V) binary code for the auxiliary features [47]. The F_0 range of the SPOKE set was around 40–330 Hz, and the F_0 mean was around 150 Hz. The unseen outside auxiliary features were simulated by replacing the original F_0 values of the acoustic features with the scaled F_0 values, and the scaling ratios were 1/2, 3/4, 5/4, 3/2, and 2. A demo and open-source QPNet implementation can be found on our demo page[2].

5.5.2 Objective Evaluation

For the objective evaluations, the ground truth acoustic features were extracted from natural speech utterances using WD, and the extraction error from WD was neglected. A speaker-dependent F_0 rage was applied to the feature extraction of each speaker to improve the extraction accuracy, and the F_0 range was set following the process in a open-source VC system (sprocket[3]). Since WD was developed to extract F_0 independent spectral features [37], the WD-extracted *sp* feature was assumed to be highly uncorrelated to the F_0 feature in this chapter. Therefore, the ground truth acoustic features for the scaled F_0 scenarios were the same natural spectral features with the F_0 feature scaled by an assigned ratio. The auxiliary features of the evaluated vocoders were the ground truth acoustic features. Mel-cepstral distortion (MCD) was

[2]https://bigpon.github.io/QuasiPeriodicWaveNet_demo/
[3]https://github.com/k2kobayashi/sprocket

Table 5.6: *QPNet with different dense factors*

Dense	2^0	2^1	2^2	2^3	2^4	2^5	2^6
MCD (dB)	4.05	**4.02**	4.03	4.08	4.17	4.63	4.26
F_0RMSE	0.23	0.17	0.15	**0.13**	0.14	0.21	0.24
U/V (%)	21.8	16.0	14.2	**13.2**	13.5	20.9	19.3

applied to measure the spectral reconstruction capability of the vocoders, and the MCD was calculated between the auxiliary *mcep* and the WD-extracted *mcep* from the generated speech. The pitch accuracy of the generated speech was evaluated using the RMSE of the auxiliary F_0 and the WD-extracted F_0 value from the generated speech in the logarithmic domain. The unvoiced/voiced (U/V) decision error was also taken into account in the evaluation of the prosodic prediction capability, which was the percentage of the unvoiced/voiced decision difference of each utterance.

An objective evaluation of the QPNet models with different dense factors for speech generation was first conducted to check the consistency of the efficient dense factor value. As shown in Table 5.6, the tendency of the objective evaluation is similar to the results of the sinusoid generation evaluation. That is, the QPNets with dense factors from 2^1–2^4 achieved similar generative performance while the speech quality and pitch accuracy of the QPNets with dense factors 2^5 and 2^6 markedly degraded because of the much shorter effective receptive field lengths. Specifically, as shown in Table 5.7, the average effective receptive field lengths of the QPNets with the dense factors 2^5 and 2^6 are much shorter than others, and the lengths were too short to cover at least one cycle of the signal with 150 Hz, which was the F_0 mean of the SPOKE set.

Furthermore, although the QPNet with a 2^0 dense factor had the longest average effective receptive field length and achieved an acceptable MCD, the higher RMSE of log F_0 and U/V error indicate its instability, which was also observed in the sinusoid

Table 5.7: *Average effective receptive field length (samples).*

Dense	2^0	2^1	2^2	2^3	2^4	2^5	2^6
Length	**2753**	1399	723	384	215	130	88

Table 5.8: *MCD (*dB*) of different speech generative models*

	WD	WNc	WNf	QPNet		rQPNet		QPNetf		rQPNetf	
E_t	-	-	-	cont.	disc.	cont.	disc.	cont.	disc.	cont.	disc.
$1 \times F_0$	2.51	4.34	3.58	4.08	4.16	3.59	3.60	3.91	3.97	**3.54**	3.58
$1/2 \times F_0$	3.88	5.02	4.56	4.79	4.90	4.49	4.46	4.66	4.79	4.43	**4.40**
$3/4 \times F_0$	2.91	4.58	3.95	4.34	4.43	3.95	3.91	4.19	4.26	**3.87**	3.88
$5/4 \times F_0$	2.76	4.39	3.62	4.16	4.25	**3.54**	3.60	3.98	4.03	3.60	3.63
$3/2 \times F_0$	3.04	4.50	3.68	4.27	4.35	**3.56**	3.64	4.06	4.12	3.65	3.67
$2 \times F_0$	3.75	4.75	3.86	4.59	4.64	**3.82**	3.88	4.33	4.37	3.92	3.90
Average	3.14	4.60	3.87	4.37	4.45	**3.83**	3.85	4.19	4.26	3.84	3.84

generation evaluation. The results also confirm our assumption that the QPNet with a 2^0 *dense factor* cannot model the periodic components well because the Nyquist frequency of the QPNet adaptive macroblock is lower than the bandwidth of the periodic components. Moreover, because of the natural fluctuations of speech, F_0 extraction errors, etc., the oversampling models with an appropriate *dense factors* such as 2^2–2^4, which keep long enough *effective receptive fields*, also achieve better performance. As a result, the dense factors of the following QPNet-series models were set to 2^3 because of the lowest RMSE of log F_0 and U/V error with an acceptable MCD. Our internal subjective evaluation results also show the preference of the utterances generated by the QPNet with the dense factor 2^3.

As shown in Table 5.8, in terms of spectral prediction capability, the compact-size

Table 5.9: *Log F_0 RMSE of different speech generative models*

	WD	WNc	WNf	QPNet		rQPNet		QPNetf		rQPNetf	
E_t	-	-	-	cont.	disc.	cont.	disc.	cont.	disc.	cont.	disc.
$1 \times F_0$	0.09	0.26	0.14	**0.13**	0.14	0.15	0.15	0.16	0.16	0.15	0.15
$1/2 \times F_0$	0.13	0.38	0.30	**0.23**	0.24	0.33	0.34	0.26	0.26	0.33	0.33
$3/4 \times F_0$	0.10	0.32	0.20	**0.17**	0.18	0.22	0.22	0.21	0.22	0.21	0.21
$5/4 \times F_0$	0.09	0.25	0.17	0.14	**0.13**	0.15	0.15	0.16	0.16	0.15	0.16
$3/2 \times F_0$	0.09	0.27	0.21	0.16	**0.15**	0.19	0.19	0.18	0.19	0.20	0.20
$2 \times F_0$	0.09	0.28	0.26	0.18	**0.17**	0.26	0.26	0.18	0.20	0.29	0.33
Average	0.10	0.29	0.21	**0.17**	**0.17**	0.22	0.22	0.19	0.20	0.22	0.23

(r)QPNet vocoders with the proposed PDCNNs significantly outperformed the same-size WNc vocoder. The results confirm the effectiveness of the QP structure to skip some redundant samples using the prior periodicity knowledge for a more efficient receptive field extension. However, the MCDs of the double-size WNf vocoder are lower than that of the compact-size (r)QPNet vocoders, and the full-size (r)QPNetf vocoders with the largest network size also outperformed the WNf vocoder in terms of MCD. The results indicate that the MCD values are highly related to the network sizes, so a deeper network attains a more powerful spectral modeling capability. Furthermore, the systems with continuous pitch-dependent dilated factors achieved better MCDs than those with discrete ones, and the result is consistent with our internal subjective evaluation for speech quality. However, the MCD differences of the rQPNet and QPNet vocoders were not reflected in the perceptual quality, and they had similar speech qualities according to the internal evaluation.

The log F_0 RMSE results in Table 5.9 also show that both the compact-size QPNet and rQPNet vocoders attained markedly higher pitch accuracy than the same-size WNc vocoder, particularly when conditioned on the unseen F_0 with a large shift.

Since the WNf vocoder usually generates seen signals even conditioned on unseen auxiliary features, the compact-size QPNet vocoder achieved higher pitch accuracies than the WNf vocoder as expected. The results indicate that the PDCNNs with the prior periodicity knowledge improved the pitch controllability of these vocoders against the unseen F_0. However, the pitch accuracies of the full-size QPNetf and rQPNetf vocoders are lower than that of the (r)QPNet vocoders. The possible reason is that the unbalanced proportion of the adaptive and fixed residual blocks impaired the pitch controllability. That is, for the full-size (r)QPNetf vocoders, the number of the fixed blocks is markedly larger than the number of the adaptive blocks. Therefore, the network might be dominated by the fixed blocks, which degraded the influence from the adaptive blocks. For the (r)QPNet vocoders with a dense factor 2^3, the receptive field length of the fixed blocks is 46 samples (The details of the receptive field length can be found in Discussion), and the average effective receptive field length of the adaptive blocks is 384 samples as shown in Table 5.7. However, for the full-size (r)QPNetf vocoders, the receptive field length of the fixed blocks is 3070 samples, which was much longer than the 384 samples of the extra four adaptive blocks. Therefore, the influence of the adaptive blocks might be very limited.

As shown in Table 5.10, the compact-size QPNet vocoder attained the lowest U/V decision error among all NN-based vocoders, and it indicates a higher capability to capture U/V information. In conclusion, the compact-size QPNet vocoder with the proposed PDCNNs and continuous pitch-dependent dilated factors attained the highest accuracy of pitch and U/V information among the evaluated NN-based vocoders. Although the compact-size QPNet vocoder did not achieve the same spectral prediction capability as the WNf vocoder according to the MCD results, it is difficult to measure a perceptual quality difference only based on MCD. As a result, subjective evaluations

Table 5.10: *U/V Decision error rate (%) of different speech generative models*

	WD	WNc	WNf	QPNet		rQPNet		QPNetf		rQPNetf	
E_t	-	-	-	cont.	disc.	cont.	disc.	cont.	disc.	cont.	disc.
$1 \times F_0$	9.9	23.6	14.5	**13.2**	13.9	14.9	14.3	15.7	15.2	14.0	14.7
$1/2 \times F_0$	16.0	35.0	26.6	**22.3**	22.8	29.9	30.1	27.6	26.3	29.5	30.4
$3/4 \times F_0$	12.2	29.1	18.2	**16.4**	17.5	20.2	20.2	19.8	20.2	18.5	19.5
$5/4 \times F_0$	9.6	24.9	13.3	**13.1**	13.9	13.9	13.5	14.5	14.2	14.1	13.9
$3/2 \times F_0$	9.9	27.9	13.8	14.7	15.5	13.6	14.8	16.3	15.7	**13.3**	14.8
$2 \times F_0$	10.5	36.7	**20.3**	21.9	20.6	26.2	24.3	25.3	26.3	29.6	33.4
Average	11.3	29.5	17.8	**16.9**	17.4	19.8	19.5	19.8	19.7	19.8	21.1

of the compact-size QPNet (with continuous pitch-dependent dilated factors), WNc, and WNf vocoders are presented in the next section. Moreover, although the WD vocoder had the best objective evaluation results, the WD-generated speech usually lacks naturalness and contains buzz noise, which may not be reflected in the objective measurements. Therefore, the WD vocoder was also involved in the subjective evaluations.

5.5.3 Subjective Evaluation

The subjective evaluations included the mean opinion score (MOS) test for speech quality and the ABX preference test for perceptual pitch accuracy. Specifically, the naturalness of each utterance in the evaluation set for the MOS test was evaluated by several listeners by assigning scores of 1–5 to each utterance; the higher the score, the greater naturalness of the utterance. The MOS evaluation set was composed of randomly selected utterances generated by the WD, WNf, WNc, and QPNet vocoders, and the auxiliary features with 1/2 F_0, 3/2 F_0, and unchanged F_0. The compact-size

QPNet vocoder with the continuous dilated factors was adopted and abbreviated as QPNet in the subjective evaluations. We randomly selected 20 utterances from the 35 test utterances of each condition and each speaker to form the MOS evaluation set, so the number of utterances in the set was 960. The mean, standard deviation, longest, and shortest lengths of the selected utterances were 4 s, 1.6 s, 8 s, and 1 s, respectively. The MOS evaluation set was divided into five subsets, and each subset was evaluated by two listeners, so the total number of listeners was 10. All listeners took the test using the same devices in the same quiet room. Although the listeners were not native speakers, they had worked on speech or audio generation research.

In the ABX preference test, the listeners compared two test utterances (A and B) with one reference utterance (X) to evaluate which testing utterance had a pitch contour more consistent with that of the reference utterance. Because the natural speech with the desired scaled F_0 does not exist, and the conventional vocoders usually have high pitch accuracy, we took the WD-generated speech as the reference. The ABX evaluation set consisted of the same generated utterances of the WNf, QPNet, and WD vocoders as the MOS evaluation set. The number of ABX utterance pairs was 240, and each pair was evaluated by two of the same 10 listeners as in the MOS test. Since the ABX test focus on pitch accuracy, all listeners were asked to focus on the pitch differences and ignore the quality differences.

MOS of Speech Quality

As shown in Fig. 5.8, for the female speaker set, the QPNet vocoder significantly outperforms the same-size WNc vocoder in all cases. Although the QPNet vocoder achieves slightly lower naturalness than the WNf vocoder in the unchanged F_0 (inside) case, the QPNet vocoder still attains markedly better naturalness than the WNf

vocoder in the 1/2 F_0 (outside) case. The results indicate that halving the network size markedly degrades the speech modeling capability of the WN vocoder. However, the proposed PDCNNs significantly improves the modeling capacity with the halved network size, especially in the 1/2 F_0 case which makes QPNet obtain a long effective receptive field length. On the other hand, owing to the small dilated factors caused by the high F_0 values, many of the PDCNNs may degenerate to DCNNs in the 3/2 F_0 case. Specifically, when the dilated factors are less than or equal to one because of the high F_0 values, the dilation sizes of PDCNN are also less than or equal to DCNN. As a result, while the F_0 values of the auxiliary features are scaled by 3/2, although the QPNet vocoder still outperforms the WNc vocoder, the naturalness of the WNf- and WORLD-generated utterances is higher than that of the QPNet-generated utterances because of the much shorter effective receptive field length of the QPNet vocoder.

Furthermore, as the results of the male speaker set shown in Fig. 5.9, the naturalness of the QPNet-generated utterances is comparable to that of the WNf-generated utterances and significantly better than that of the WNc-generated utterances in all F_0 cases. Specifically, even if the F_0 values are scaled, most of the 3/2 F_0 values of the male utterances are still within the range of the normal female F_0. Therefore, the effective receptive field lengths of the QPNet vocoder are still much longer than the receptive field lengths of the WNc vocoder for most male utterances with scaled F_0. On the other hand, the WORLD vocoder shows a similar tendency in the evaluations of both female and male speaker sets. In the unchanged F_0 case, the naturalness of the WORLD-generated utterances is slightly lower than the WNf- and QPNet-generated utterances. In the scaled F_0 cases, the WORLD vocoder achieves even much lower naturalness in the 1/2 F_0 case, but comparable naturalness in the 3/2 F_0 case.

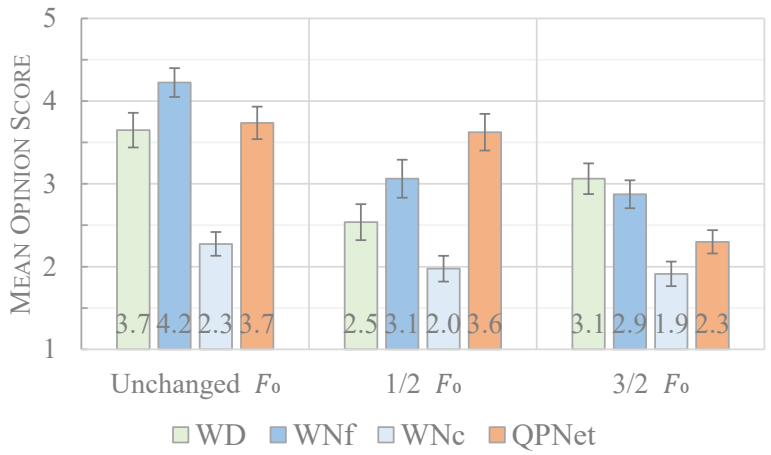


Figure 5.8: *Sound quality MOS evaluation of female speakers with 95 % CI.*

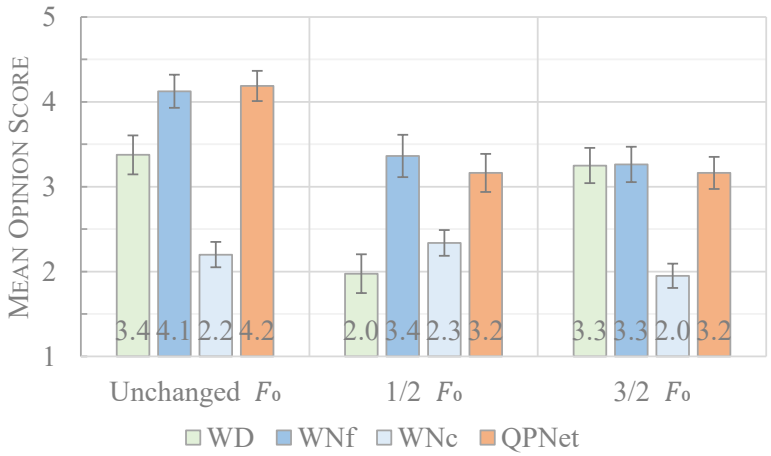


Figure 5.9: *Sound quality MOS evaluation of male speakers with 95 % CI.*

ABX of Pitch Accuracy

As shown in Figs. 5.10 and 5.11, the QPNet vocoder significantly outperforms the WNf vocoder in terms of pitch accuracy in most F_0 cases and both the female and male sets except in the unchanged F_0 cases of the female set, which may be caused by the naturalness degradation. The results confirm the pitch controllability improvement of the QPNet vocoder with the PDCNNs. In summary, the QPNet vocoder with the more compact network size achieves comparable speech quality to the WNf vocoder under

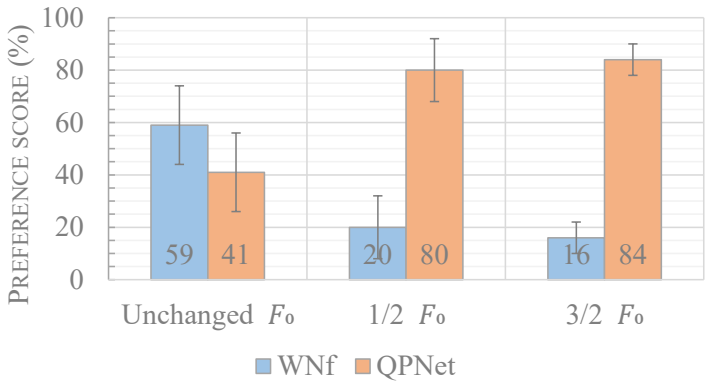


Figure 5.10: *Pitch accuracy ABX evaluation of female speakers with 95 % CI.*

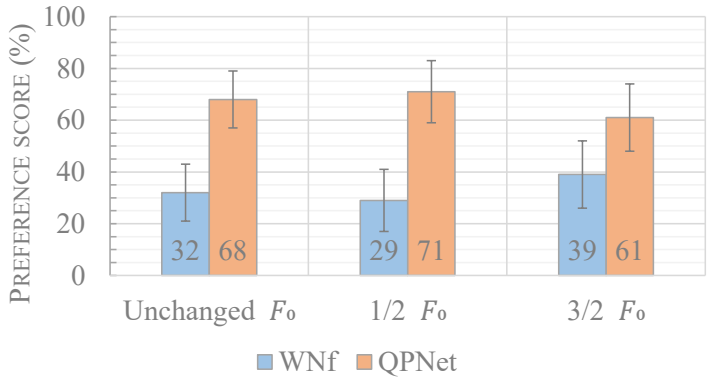


Figure 5.11: *Pitch accuracy ABX evaluation of male speakers with 95 % CI.*

most conditions except for the female set with 3/2 F_0 because the higher F_0 values may make the PDCNNs degenerate to the DCNNs. The QPNet vocoder conditioned on the unseen F_0 also gets the markedly higher pitch accuracy than the WNf vocoder. Moreover, the QPNet vocoder achieved higher or comparable speech quality than the WORLD vocoder under most conditions except conditioned on the acoustic features with the unseen 3/2 female F_0.

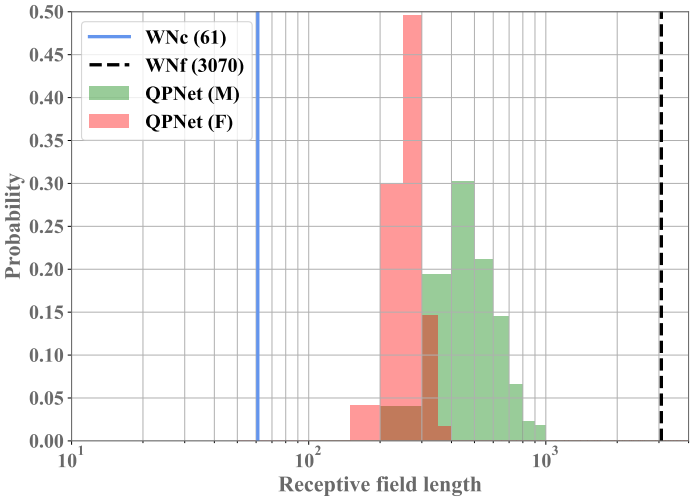


Figure 5.12: *Distributions of receptive field lengths of different vocoders.*

5.5.4 Discussion

As shown in Fig. 5.12, the length of the receptive fields of WNf is 3070 samples (The receptive field length of 10 blocks in each chunk is $2^0 + 2^1 + \cdots + 2^9 = 1023$ samples, so the total length is 1023×3 samples with an extra one sample from the causal layer), that of WNc is 61 samples (Each chunk contains $2^0 + 2^1 + 2^2 + 2^3 = 15$ samples, so the total receptive field length is $15 \times 4 + 1 = 61$ samples), and that of QPNet is 100–1000 samples (The receptive field length of the fixed blocks and the causal layer is $15 \times 3 + 1 = 46$ samples, and that of the adaptive blocks is $15 \times E_t$ samples. The pitch-dependent dilated factor E_t with a dense factor 8 was around 60 for 50 Hz and 6 for 500 Hz). Specifically, the receptive field lengths of WNf and WNc are constant because of the fixed network structure, and the receptive field length of QPNet is time-variant and pitch-dependent because of the QP structure.

Furthermore, the results in Fig. 5.12 also show that the effective receptive field lengths of both SPOKE male and female speakers are longer than the receptive field length of WNc, which are consistent with the evaluation results showing that QPNet

significantly outperforms WNc. Furthermore, most of the effective receptive field lengths of the female set are shorter than that of the male set, and it is caused by the higher F_0 values of the female speakers. The distribution results also imply that the effective receptive field lengths of QPNet are close to the receptive field length of WNc when conditioned on the female 3/2 F_0 because most PDCNNs degenerate to DCNNs. In conclusion, the performance of AR models is highly related to the length of the receptive fields.

However, the length of the receptive fields may be more strongly correlated to the quality of the generated speech, whereas a balanced proportion of the adaptive and fixed modules may be an essential factor for the pitch accuracy. Specifically, although the full-size QPNet has the longest effective receptive field lengths and achieves the lowest MCD, the pitch accuracy of full-size QPNet is still lower than that of compact-size QPNet. The possible reason is that the full-size QPNet is dominated by the fixed blocks because the number of the fixed blocks is much larger than the number of the adaptive blocks while the compact-size QPNet has more balanced numbers of the fixed and adaptive blocks.

In addition, as shown in Tables 5.1 and 5.5, the number of the trainable parameters of the compact-size QPNet model is around half of that of the WNf model, so only about 75 % of the training time and 40 % of the generation time were required. However, because of the very long effective receptive fields, the memory usage of QPNet in the training stage was almost the same as that of WNf. The huge memory requirement in the training process limits the possible ratio of the fixed to adaptive modules, which leads to an unbalanced proportion problem. Therefore, improving the efficiency of memory usage will be one of the main tasks of future QPNet research.

5.6 Voice Conversion Evaluation

In this section, the performance of the QPNet vocoder combined with a basic VC model is presented [40]. According to the previous work [43, 46, 48], we know that target speaker adaptation is necessary for the WN vocoder combined with VC models. Therefore, the investigation of the speaker adaptation strategy of the WNf, WNc, and QPNet vocoders are first described. Then, the objective tests to evaluate the waveform generation capability of these vocoders and subjective tests to evaluate the performance of the whole VC system are presented.

5.6.1 Experimental Setting

The training corpus and hyperparameters of these multispeaker vocoders are the same as that in Section 5.5. The adopted VC models are the DNN-based non-parallel VC models in Section 3, which are the same as the VC models of the NU non-parallel VC system [43] for the VCC2018 SPOKE task. For the VC flow, a source *mcep* is converted to a specific target *mcep* by the trained DNN-VC model, and then the speaker-dependent (SD) QPNet vocoder generates the converted speech waveforms conditioned on the converted *mcep*, linearly transformed F_0, and source *ap*.

5.6.2 Speaker Adaptation

Two strategies to adapt the speaker-independent (SI) WN-based vocoders to SD ones are presented. The first one is updating all network parameters (SDa) and the second one is only updating the final output layers of the network (SDo) with the training data of the target speakers. Figure 5.13 shows the training loss (cross-entropy) of the SD vocoders with speaker TM1 while the other target speakers have the same tendency.

Figure 5.13: *Adaption training losses of different vocoders (SD: speaker-dependent; o: only update output layers; a: update whole network).*

The utterances used for adaptation were only the 81 utterances of TM1, the updating batch size was one, the batch length was 20,000 samples, and the number of iterations was from 100 to 50,000. As shown in Fig. 5.13, the training losses of the SDa vocoders start to markedly decrease when beyond 1000 iterations, whereas the training losses of the SDo vocoders are stable regardless of the number of iterations. The results indicate that updating the whole network with very limited data will cause serious overfitting. Furthermore, the adaptation performance using the training loss of the validation data while fixing the network parameters denoting the validation loss was evaluated. Figure 5.14 shows that the validation losses of the SDa vocoders start to increase from around 500 iterations ($\sim$ 2 epochs), whereas the SDo vocoders exhibited stable validation losses. Therefore, we set the number of updating iteration as 500 for the SDa vocoders and 50,000 for the SDo vocoders (our NU system submitted to VCC2018 was SDo-WNf with 50,000 iterations) in the following evaluations.

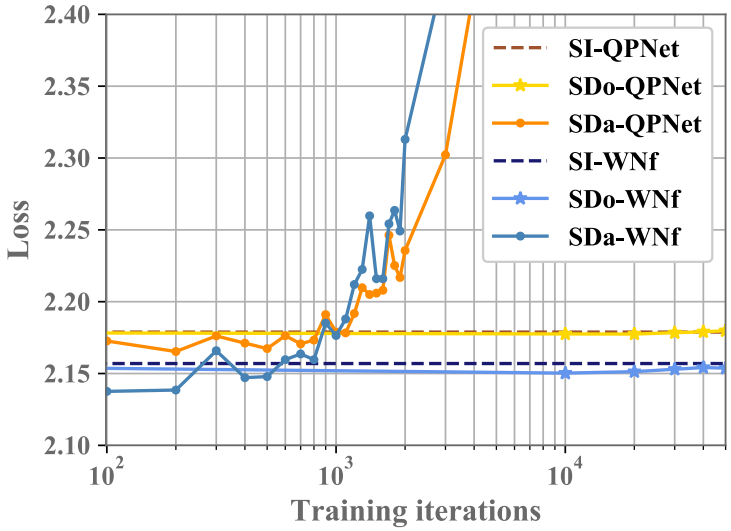


Figure 5.14: *Validation losses of different vocoders (SI: speaker-independent; SD: speaker-dependent; o: only update output layers; a: update whole network).*

5.6.3 Objective Evaluation

To evaluate the performance of these vocoders with statistically converted acoustic features, we measured the mel-cepstral distortion (MCD) and the root-mean-square error (RMSE) of logarithmic F_0 between the acoustic features extracted from the converted speech and the auxiliary features of these vocoders. Specifically, we computed MCD between the conditional and extracted *mcep* to evaluate the robustness of spectrum reconstruction with the vocoders conditioned on the VC acoustic features. Moreover, to evaluate the generation pitch accuracy of each vocoder corresponding to the conditional linearly transformed F_0, we calculated the RMSE between the conditional F_0 and the F_0 extracted from the converted speech in the logarithmic domain.

As shown in Table 5.11, the QPNet vocoder significantly outperforms the same-size WNc vocoder in both MCD and RMSE measurements. Even compared with the WNf vocoder with double the network size, the QPNet vocoder still achieves slightly higher

Table 5.11: *MCD and RMSE of* $\log F_0$ *with different vocoders.*

	WNf			WNc			QPNet		
	SI	SDo	SDa	SI	SDo	SDa	SI	SDo	SDa
MCD	3.25	3.11	**3.02**	3.83	3.73	3.68	3.57	3.51	3.46
RMSE	0.15	0.15	0.15	0.21	0.20	0.19	0.15	**0.13**	0.14

pitch accuracy. Although the WNf vocoder had the highest spectrum prediction capability because of its longest receptive filed length, the QPNet vocoder still outperforms the same-size WNc vocoder. That is, the much shorter receptive field length caused by the halved network size degrades the spectral prediction capability, but the PDCNNs of the QPNet markedly improve the spectral modeling capability. In summary, the objective evaluations show that the QPNet vocoder achieves higher pitch accuracy and more efficient spectral modeling than the WN vocoders. Furthermore, all SDo and SDa vocoders achieve better MCD than the relevant SI vocoders, and the results confirm the effectiveness of the target speaker adaptation. Because the SDa vocoders attain the highest spectrum prediction capabilities, we applied the adaptation strategy of SDa to the VC systems in the following evaluations.

5.6.4 Subjective Evaluation

To evaluate the speech quality and speaker similarity of the converted waveforms generated by the different vocoders conditioned on the converted acoustic features, MOS and speaker similarity tests were conducted. Specifically, we randomly selected 20 utterances from 35 testing utterances of each speaker pair and vocoder to establish an evaluation set. Then, we divided this set into 10 non-overlapping subsets for 10 listeners, and each subset was evaluated by one listener. As a result, each listener

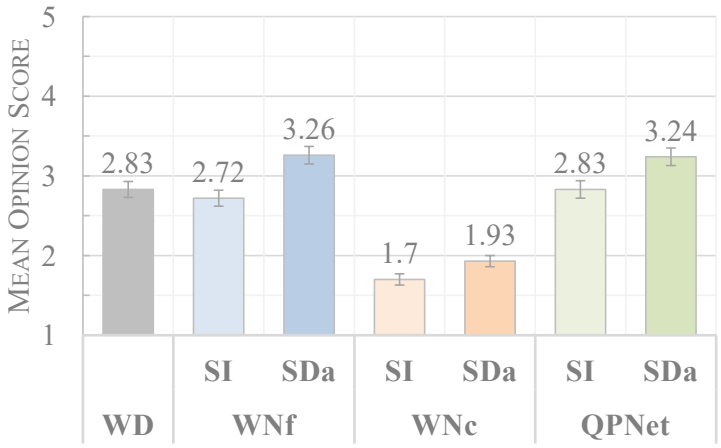


Figure 5.15: *MOS evaluation of sound quality with 95% CI. (SI: speaker-independent vocoder; SDa: speaker-dependent vocoder with fine-tuning of whole network).*

evaluated 224 different utterances generated by seven vocoders including the SI and SDa WN-based and WD vocoders in the MOS test. The speech quality was assigned a value of 1–5; the higher the score, the better the naturalness. Moreover, the speaker similarity evaluation followed the test flow of VCC2018 [115]. That is, a subject was first asked to listen to a natural speech and a converted speech and then asked to evaluate the speaker similarity of the two speech files using four labels: definitely the same, maybe the same, maybe different, and definitely different. The final speaker similarity scores were the sum of the percentages of definitely the same and maybe the same and the sum of definitely different and maybe different.

As shown in Fig. 5.15, the MOS evaluation results of WNc and QPNet indicate that the PDCNNs significantly improve the speech quality of converted speech even though the network sizes of these two vocoders were the same. Furthermore, the overall results confirm the effectiveness of the SD adaptation of all WN-based vocoders to achieve significantly better speech naturalness. Compared with the full-size WN vocder, SI-QPNet attained slightly better performance than SI-WNf, and the perceptual qualities

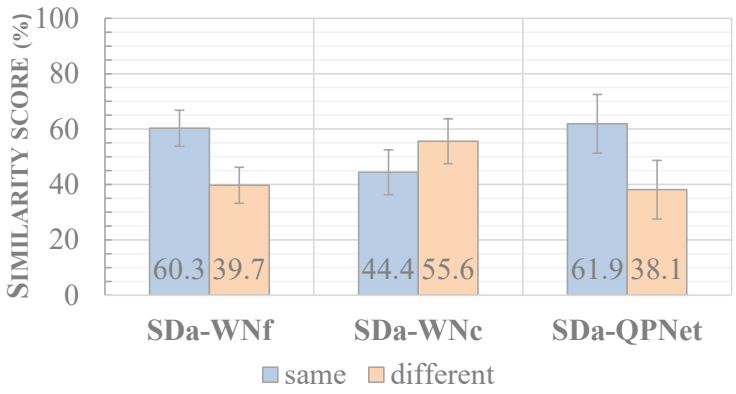


Figure 5.16: *Speaker similarity evaluation with 95% CI. (SI: speaker-independent vocoder; SDa: speaker-dependent vocoder with fine-tuning of whole network).*

of SDa-QPNet and SDa-WNf were comparable despite the network size of QPNet being only half of that of WNf. Moreover, SDa-QPNet also achieved markedly better conversion speech generation capability than the traditional WORLD vocoder. To further evaluate the conversion accuracy of speaker identity among these WN-based vocoders, we conducted the speaker similarity tests on the SDa-WNf, SDa-WNc, and SDa-QPNet vocoders. The results in Fig. 5.16 demonstrate the same tendency as the speech naturalness results. SDa-QPNet markedly outperforms SDa-WNc for speaker similarity and achieved similar performance to SDa-WNf. The demo utterances can be found on our demo page[4].

5.7 Summary

In this chapter, we introduce a WaveNet-like audio waveform generation model named QPNet, which models quasi-periodic and high-temporal-resolution audio signals based on an NN-based AR model with the PDCNN component and cascaded

[4]https://bigpon.github.io/QuasiPeriodicWaveNet_demo/

structure. Specifically, the PDCNN component is a variant of a DCNN that dynamically changes the dilation size corresponding to the instantaneous F_0 for modeling the long-term correlations of audio samples. The cascaded adaptive blocks with the PDCNNs and fixed blocks with the DCNNs respectively modeling the periodic and aperiodic audio components are adopted in the QPNet.

The proposed QPNet was respectively evaluated with the generations of periodic sinusoids, F_0-transformed speech, and VC speech. According to the sinusoid generation evaluation results, the PDCNNs significantly improves the periodicity-modeling capability of the generation network using the introduced prior periodicity information. Furthermore, the QPNet vocoder models the short- and long-term correlations of speech samples based on the cascaded fixed and adaptive macroblocks, respectively. The speech generation evaluation results indicate that the proposed QPNet vocoder attains a much higher pitch accuracy and comparable speech quality to the WN vocoder especially when conditioning on the unseen auxiliary F_0 values. Moreover, the network size and generation time requirements of the QPNet vocoder are only half of those of the WN vocoder.

For VC, the effectiveness of two speaker adaption methods for SD WN-based vocoders was first evaluated. Both objective and subjective evaluations confirm the effectiveness of the speaker adaption technique and the QPNet vocoder, which takes advantage of the PDCNNs to attain better pitch controllability and achieve comparable quality to the WN vocoder with only half the network size. That is, the proposed QPNet model with the PDCNN component and compact cascaded network architecture significantly improves the pitch controllability of the vanilla WN model, and it makes the QPNet vocoder more in line with the definition of a vocoder, which precisely manipulates the pitch of the generated speech according to the F_0.

To summarize, this chapter focuses on the essential feature of a vocoder, pitch controllability. The evaluation results show the insufficient pitch controllability of the vanilla WN vocoder and the markedly improved pitch controllability of the proposed QPNet with the PDCNN and QP structure. Since the network architecture of the QPNet is dynamically adapted to the instantaneous input F_0, the pitch accuracy of the QPNet-generated speech is improved. Furthermore, because of the more efficient speech modeling of the pitch-dependent structure, the QPNet-generated speech achieves similar speech quality as the WN-generated speech while the model size of the QPNet vocoder is only half of that of the WN vocoder. However, although the generation speed is also increased because of the smaller model, the AR nature still makes the generation of the QPNet vocoder far away from real-time generation, which degrades the practicality of the QPNet vocoder. Therefore, a non-AR model for real-time generations will be explored in the next chapter.

6 Quasi-Periodic Parallel WaveGAN for Speech Waveform Generation

In Chapter 5, the effectiveness of the proposed pitch-dependent architecture for the WaveNet (WN) vocoder has been shown. However, the autoregressive (AR) generation manner makes the generation of the proposed quasi-periodic WN (QPNet) very slow. To improve the generation efficiency for practical applications, this chapter adopts a non-AR model, parallel WaveGAN (PWG). Since the PWG model also suffers from the insufficient pitch controllability problem, the proposed pitch-dependent architecture will be applied to the PWG model in this chapter.

6.1 Introduction

Speech generation is a technique to generate specific speech according to given inputs such as texts (text-to-speech, TTS). The core of speech generation is the controllability of speech components, and the fundamental technique is called a vocoder [29–31]. Conventional vocoders such as STRAIGHT [35] and WORLD [37] are based on a source–filter model [34], which models speech with vocal fold movements (excitation) and vocal tract resonances (spectral envelope). However, these conventional vocoders usually suffer from the losses of phase information and temporal details caused by

ad hoc designs result in speech quality degradation.

In contrast to the conventional source–filter-based vocoders, neural network (NN)-based models have been proposed to directly model speech waveforms. Specifically, AR models such as WN [12] and SampleRNN [13] achieve high-fidelity speech generation by modeling the probability distribution of each speech sample with the given auxiliary features and previous samples. Taking conventional-vocoder-extracted acoustic features as the auxiliary features for NN-based speech generation models [47–50, 74], which replace the synthesizer of the conventional vocoders, also achieved early success. However, the AR mechanism and huge network architectures of WN and SampleRNN result in a very slow generation, making these models impractical for realistic scenarios. To tackle these problems, many compact AR models with specific knowledge [14–16] and non-AR NN-based speech generation models such as flow-based [17–22] and generative adversarial network (GAN)-based [23–26, 41, 42, 64–66, 72, 73, 135] models have been proposed.

Although these NN-based speech generation models achieve high-fidelity speech generation without many ad hoc designs of speech generation, the data-driven nature, the generic network architecture, and the lack of prior acoustic knowledge of these models make most of them lose acoustic controllability and robustness to unseen auxiliary features [27, 28, 43, 44, 46]. For instance, without explicitly modeling the excitation signals as conventional source–filter models, it is difficult for WN to generate speech with accurate pitches outside the fundamental frequency (F_0) range of training data when conditioned on the scaled F_0 feature [38, 39]. However, using carefully designed mixed periodic and aperiodic inputs and source–filterlike architectures, the authors of [68–70] proposed different NN-based models attaining pitch controllability.

In Chapter 5, we also introduce QPNet [38,39], which improves the pitch controllabil-

ity of WN by dynamically changing the network architecture according to the auxiliary F_0 feature without the requirement of specific mixed inputs. However, the AR mechanism and the huge network requirement of the QPNet result in a slow generation. To address this problem, we applied the quasi-periodic (QP) structure to PWG [24], which is a compact non-AR model with a WN-like network architecture consisting of stacked dilated convolutional neural networks (DCNNs) [75]. The proposed QPPWG speech generation model [41, 42] attains pitch controllability using a simple pitch-dependent architecture without the requirement of specific mixed periodic and aperiodic inputs as in [68–70].

This chapter is organized as follows. In Section 6.2, a brief introduction of PWG and the limitations of the PWG vocoder is presented. In Section 6.3, the concepts and details of the proposed QPPWG are described. In Section 6.4, objective and subjective tests are reported to show the effectiveness of QPPWG for generating speech with scaled F_0. Further discussion of QPPWG is presented in Section 6.5. Finally, the conclusion is given in Section 6.6.

6.2 Parallel WaveGAN and Limitations of Parallel WaveGAN Vocoder

As shown in Fig. 6.1, the PWG model [24] consists of a generator, a discriminator, and a multi-resolution short-time Fourier transform (STFT) module. Specifically, the generator adopts a WN-like architecture to transfer the input noise sequence into speech samples conditioned on auxiliary features. Since the whole noise sequence and acoustic features are given, the generator achieves real-time generation with a non-AR manner. A DCNN-based discriminator is applied to guide the generator to generate high-fidelity

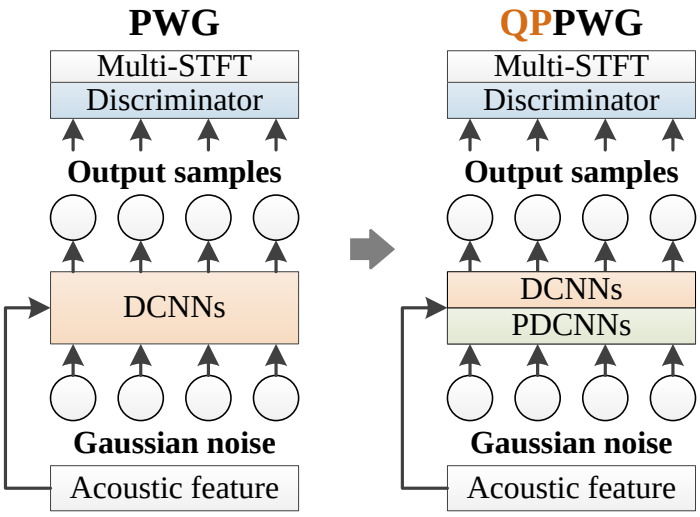


Figure 6.1: *Architectures of PWG and QPPWG.*

speech using adversarial losses. However, training a high-quality generator with only the adversarial losses is difficult. Therefore, a multi-STFT loss has been introduced for improving the stability and efficiency of the GAN training. The multi-STFT loss is calculated based on several different frame lengths and shifts to make the network capture the hierarchical information of speech signals. Furthermore, taking PWG as a vocoder to generate speech samples conditioned on conventional-vocoder-extracted acoustic features also achieves marked naturalness improvements than the conventional parametric-based vocoders [42].

However, although the PWG vocoder achieves high-fidelity speech generation, it is still vulnerable to unseen acoustic features such as scaled F_0. That is, the speech quality and pitch accuracy of the PWG-generated speech will markedly degrade when the F_0 of the auxiliary acoustic features is scaled or is outside the training data range [41, 42]. The possible reasons for the degradation are the generic architecture, data-driven nature, and lack of prior periodicity knowledge. Moreover, since speech is a quasi-periodic signal, which includes both periodic components with long-term correlations and ape-

riodic components with short-term correlations, modeling both components with the fixed network architecture of PWG is inefficient. Therefore, the QP structure [38, 39] has also been applied to PWG as show in Fig. 6.1 for improving pitch controllability and modeling efficiency. The details of the proposed QPPWG are as follows.

6.3 Quasi-Periodic Parallel WaveGAN

Since pitch controllability is an essential feature of a vocoder, QPPWG [41, 42] has been proposed to improve the pitch controllability and speech modeling efficiency of PWG. Specifically, since the effectiveness of the GAN structure and the multi-resolution STFT losses have been shown for PWG, the proposed QPPWG inherits the discriminator and L_{sp} of PWG and focuses on improving the generator. The QP structure of the proposed generator introduces periodicity information to the network via a PDCNN module and a cascaded structure. The details are as follows.

6.3.1 Noncausal Pitch-dependent Dilated Convolution

In Chapter 5, the causal PDCNN inspired by pitch filtering in code-excited linear prediction (CELP) [32,33] has been described. In this chapter, the noncausal extension of the PDCNN is presented. As shown in Fig. 6.2, a DCNN is a convolution layer with gaps between input samples, and the length of each gap is a predefined hyperparameter called the dilation size (rate). The noncausal dilated convolution can be formulated as

$$\boldsymbol{y}_t^{(\text{o})} = \boldsymbol{W}^{(\text{c})} \times \boldsymbol{y}_t^{(\text{i})} + \boldsymbol{W}^{(\text{p})} \times \boldsymbol{y}_{t-d}^{(\text{i})} + \boldsymbol{W}^{(\text{f})} \times \boldsymbol{y}_{t+d}^{(\text{i})}, \tag{6.1}$$

where $\boldsymbol{y}_t^{(\text{o})}$ is the DCNN output at sample t, $\boldsymbol{y}_t^{(\text{i})}$ is the DCNN input at sample t, and d is the dilation size. $\boldsymbol{W}^{(\text{c})}$, $\boldsymbol{W}^{(\text{p})}$, and $\boldsymbol{W}^{(\text{f})}$ are the trainable 1×1 convolution filters

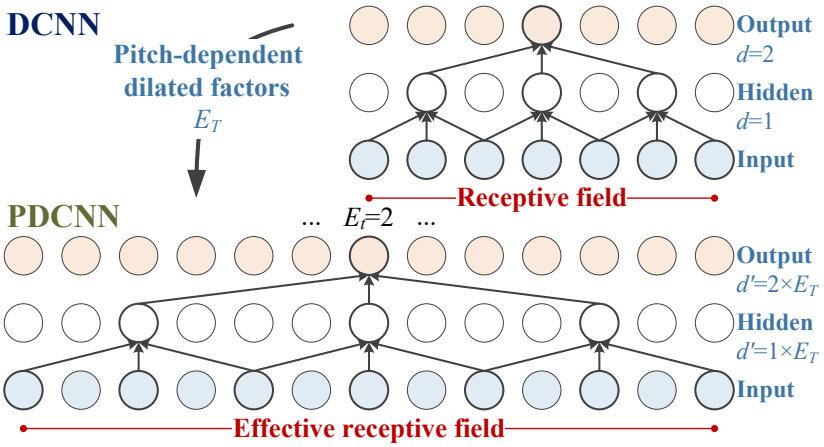


Figure 6.2: *Fixed and pitch-dependent noncausal dilated convolution.*

of the current, previous, and following samples, respectively. For the DCNN, d is a predefined time-invariant constant. As an extension of a DCNN, the dilation size d' of a PDCNN is pitch-dependent and time-variant. Specifically, the pitch-dependent dilated factor E_t is multiplied by the dilation size d in each time step t to dynamically set the dilation size d' as

$$d' = E_t \times d. \tag{6.2}$$

The dilated factor E_t is derived from

$$E_t = F_s/(F_{0,t} \times a), \tag{6.3}$$

where F_s is the sampling rate, $F_{0,t}$ is the fundamental frequency of the input sample at time step t, and a is the dense factor. The dense factor a is a hyperparameter that indicates the number of samples in one periodic cycle taken as the inputs of a PDCNN. The higher the dense factor, the lower the sparsity of the PDCNN. Using the pitch-dependent dilation size, the architecture of QPPWG with PDCNNs is dynamically changed according to the input F_0 feature.

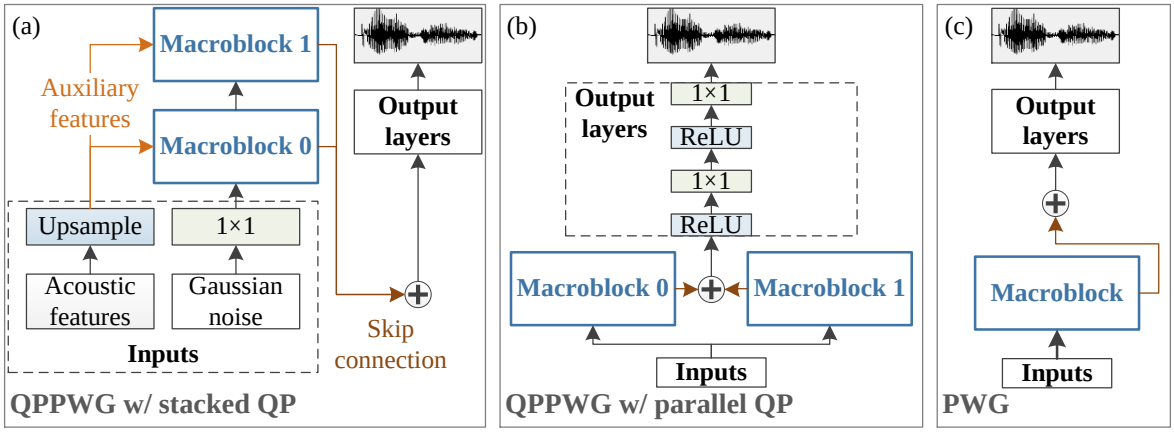


Figure 6.3: *Architecture of QPPWG generator.*

Furthermore, according to our previous work [38, 39], calculating E_t using the interpolated F_0 values of the adjacent voiced segments achieves higher speech quality than directly setting E_t to one for the unvoiced segments. Because our internal evaluation results of QPPWG also show the same tendency, all QPPWG models in this thesis adopt the interpolated F_0 values for calculating the E_t values of the unvoice segments. In conclusion, the adaptive architecture of QPPWG introduces prior periodicity knowledge to the network to improve the pitch controllability, allow each sample to have a specific receptive field size, and efficiently extend the receptive field.

6.3.2 QPPWG Generator with PDCNN

As shown in Fig. 6.3, a generator of QPPWG/PWG consists of input, macroblock, and output modules. The input module includes a Gaussian noise input with 1×1 CNN and upsampled acoustic features with the matched temporal resolution to the output waveform samples. As shown in Fig. 6.4, several stacked residual blocks compose a macroblock. The inputs of each residual block are the residual connection output of

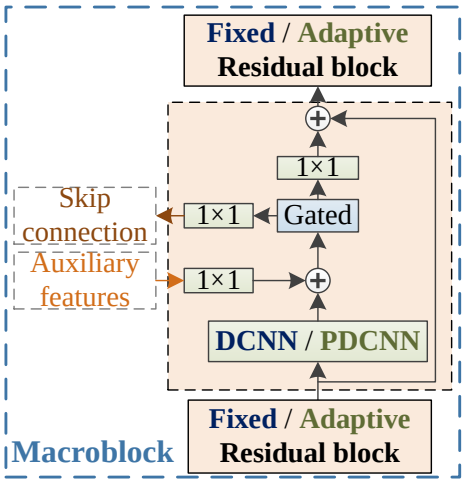


Figure 6.4: *Macroblock of QPPWG generator.*

the previous block and auxiliary features. The outputs of each residual block are the residual connection output for the next block input and the skip connection to the output module. The architecture of each residual block contains a DCNN/PDCNN layer, a gate structure, and a residual connection. Last, the summation of the skip connections from all residual blocks is processed by two ReLU [76] activations with 1×1 CNNs to directly output speech waveform samples.

The main difference between the QPPWG and PWG generators is the QP structure, and a QPPWG generator includes a fixed macroblock and an adaptive macroblock while a PWG generator includes only one fixed macroblock. The fixed macroblock consists of only fixed (residual) blocks with DCNN layers, and the adaptive macroblock consists of only adaptive (residual) blocks with PDCNN layers. Each fixed block adopts a DCNN with a fixed network architecture to model the aperiodic speech components such as spectral envelopes with short-term correlations. Each adaptive block adopts a PDCNN layer to model the periodic speech components such as excitation signals with long-term correlations, and the PDCNN layer makes the architecture of the block

adaptive to auxiliary F_0 values.

As shown in Fig. 6.3 (a), QPPWG adopts a cascaded architecture composed of two different macroblocks, which is unlike PWG consisting of residual blocks with only DCNNs. The cascaded architecture simultaneously models both periodic and aperiodic speech components efficiently by using prior pitch knowledge, which also improves its pitch controllability. The cascaded architecture with prior pitch knowledge is assumed to have better tractability and interpretability than the original PWG architecture since it models different speech components with related specific network structures. Furthermore, since we assume that the fixed and adaptive macroblocks respectively focus on aperiodic and periodic components, we also explore a parallel QP structure as shown in Fig. 6.3 (b) to better understand the internal speech production mechanisms.

6.4 Experimental Evaluation

6.4.1 Model Architecture

Several variants of PWG and QPPWG models and a baseline QPNet model were involved in the evaluations. To describe the different architecture of each model, several basic modules were introduced. Specifically, a macroblock module consisting of stacked chunks was adopted. Each chunk included several residual blocks, and the dilation sizes in each chunk were exponentially increased with base two. All chunks in a macroblock were only composed of one type of residual block namely, adaptive blocks ($\mathrm{B_A}$) or fixed blocks ($\mathrm{B_F}$). The PWG models only consisted of one macroblock (Macro 0) with several chunks including only fixed residual blocks. The models with a QP structure, namely, the QPPWG and QPNet models, were composed of two cascaded macroblocks (Macro 0 and 1) with chunks including different residual block types. Taking vanilla PWG as

an example, the architecture composed of three chunks, and each chunk includes 10 fixed blocks was denoted as $C3B_F10$

Moreover, all PWG and QPPWG models had the same discriminator architecture, which consisted of 10 noncausal DCNN layers with 64 convolution channels, three kernels, and LeakyReLU ($\alpha = 0.2$) activation functions. For each adaptive/fixed block of the QPPWG/PWG generator, a gated activation with tanh and sigmoid functions was adopted, and the number of CNN channels of residual and skip connections and auxiliary features was also 64. The QPNet structure followed that in Section 5.5, and the number of CNN channels of residual connections and auxiliary features was 512 and that of skip connections was 256.

6.4.2 Experimental Setting

All speech generation models in the evaluations were trained in a multi-speaker manner. The training corpus consisted of 2200 utterances of the “slt” and “bdl” speakers of the CMU-ARCTIC corpus [128] and 852 utterances of all speakers of the Voice Conversion Challenge 2018 (VCC2018) corpus [115]. The total size of the training corpus was around 3000 utterances and the data length was around 2.5 hours. The testing corpus was the SPOKE set of the VCC2018 corpus. The SPOKE set consists of two male and two female speakers, and each speaker has 35 testing utterances. The sampling rate of all speech data was set to 22,050 Hz, and the resolution of the speech data was 16-bit.

The WORLD (WD) vocoder was adopted to extract one-dimensional F_0 and 513-dimensional spectral (*sp*) and aperiodicity (*ap*) features with a frameshift of 5 ms. F_0 was interpolated to the continuous F_0 and converted to one-dimensional unvoiced/voiced binary code (U/V), *ap* was coded into two-dimensional coded aperiodicity (*codeap*),

and *sp* was parameterized into 35-dimensional mel-cepstrum (*mcep*). The auxiliary features of these speech generation models were composed of U/V, F_0, *mcep*, and *codeap*. To simulate unseen data, the continuous F_0 was scaled by ratios of 0.5, 1.5, and 2 while keeping the other features the same. Moreover, the dilated factor E_t of QPPWG was empirically calculated based on the continuous F_0 because of the better speech quality.

All PWG-like models were trained with the RAdam optimizer [136] ($\epsilon = 10^{-6}$) with 400 k iterations. Specifically, the generators were trained with only multi-resolution STFT losses for the first 100 k iterations and then jointly trained with the discriminators for the following 300 k iterations. The multi-resolution STFT losses were calculated on the basis of three different FFT sizes (1024/2048/512), frame shifts (120/240/50), and frame lengths (600/1200/240). The balanced weight λ_{adv} of L_{adv} was set to 4.0. The generators ' learning rate was 10^{-4} and the discriminators ' learning rate was 5×10^{-5}. Both learning rates decayed by 50 % every 200 k iterations. The mini-batch size was six and the batch length was 25,520 samples. Furthermore, the baseline QPNet [38,39] model was trained with the Adam optimizer [117] with 200K iterations. The learning rate of QPNet was 10^{-4} without decay, and the minibatch size was one with a batch length of 20,000 samples.

6.4.3 Objective Evaluation

The quality of these vocoders was evaluated by the mel-cepstral distortion (MCD), root mean square error (RMSE) of log F_0, and U/V decision error. These measurements were calculated using the auxiliary features and the acoustic features extracted from the generated speech. The following objective evaluations were conducted to explore the different hyperparameter settings, and the WD vocoder was used as a reference.

Table 6.1: *CNN channels of PWG generator*

	WD	PWG			
Channels	-	64	32	16	8
MCD (dB)	2.58	**3.69**	4.15	4.23	4.89
F_0RMSE	0.10	**0.12**	0.14	0.15	0.20
U/V (%)	10	**14**	16	16	15
Size ($\times 10^6$)	-	1.16	0.34	0.11	**0.04**

Number of CNN Channels

To explore the relationship between model capacities and the number of CNN channels, vanilla PWG generators with 8–64 CNN channels were evaluated. Note that because we focused on improving the generator, all PWG/QPPWG models in this section adopted the same discriminator, whose number of CNN channels was 64 and whose model size was 0.1 M. The results in Table 6.1 show that the original setting (64 CNN channels) predictably achieves the best performance characteristics of all objective measurements. However, even if the number of CNN channels is reduced to 16, which greatly reduces the training time because of the compact model size, the speech quality and pitch accuracy are still acceptable. To efficiently explore the efficiency of the network architectures, the objective evaluations in the following sections were conducted based on models with 16 CNN channels.

Numbers of Chunks and Blocks

Since one of the motivations for adopting the QP structure is taking advantage of the higher speech modeling capability to reduce the model size, the importance of the numbers of chunks and residual blocks was first evaluated. The results in Table 6.2

Table 6.2: *Blocks and chunks of PWG generator*

Chunk (C)	3	2	1	4	1
Block (B_F)	10	10	10	5	20
MCD (dB)	**4.23**	4.61	5.95	4.59	5.98
F_0RMSE	**0.15**	0.17	0.31	0.35	0.30
U/V (%)	**16**	17	33	44	27
Size ($\times 10^6$)	0.11	0.08	**0.04**	0.08	0.08

show that the model capacity is highly dependent on the total number of residual blocks, which is directly related to the *receptive field* length. However, the results of $C2B_F10$, $C1B_F20$, and $C4B_F5$ also imply that not only the number of residual blocks but also the number of chunks is important. Although the same number of residual blocks with fewer chunks result in a longer *receptive field*, the network may not capture the speech information well. By contrast, the larger the number of chunks, the shorter the *receptive field*. Since a longer *effective receptive field* can be achieved by replacing fixed blocks with adaptive blocks, we focus on improving the $C2B_F10$ and $C4B_F5$ PWG generators using the QP structure in this thesis.

Ratio of Fixed and Adaptive Blocks

Since speech is a quasi-periodic signal, speech modeling is theoretically required both fixed and adaptive blocks to respectively model aperiodic and periodic components. To explore the efficient ratio of fixed and adaptive blocks, four QPPWG models with dense factor 4 were evaluated (more discussions of dense factor are presented in the following subsection). Each QPPWG model was composed of four chunks, and each chunk included five residual blocks. As shown in Table 6.3, although the $C3B_F5$+$C1B_A5$ model achieves the lowest MCD, its F_0 and U/V accuracies are also lowest. On the

Table 6.3: *Ratios of fixed and adaptive blocks of QPPWG generator*

	Macro 0 Macro 1	$C3B_F5$ $C1B_A5$	$C2B_F5$ $C2B_A5$	$C1B_F5$ $C3B_A5$	- $C4B_A5$
MCD (dB)	$1 \times F_0$	**4.79**	**4.79**	5.58	7.48
	$1/2 \times F_0$	**5.22**	5.29	6.03	8.16
	$2 \times F_0$	**5.66**	6.03	7.13	8.47
	Average	**5.22**	5.37	6.24	8.04
RMSE of $\log F_0$	$1 \times F_0$	0.13	**0.12**	0.13	0.14
	$1/2 \times F_0$	0.22	**0.17**	**0.17**	0.19
	$2 \times F_0$	0.10	**0.12**	0.12	0.14
	Average	0.15	**0.14**	0.14	0.15
U/V error (%)	$1 \times F_0$	23	**16**	**16**	20
	$1/2 \times F_0$	26	21	**20**	22
	$2 \times F_0$	18	**15**	16	18
	Average	23	**17**	**17**	20

other hand, when the ratio of adaptive blocks increases, the F_0 and U/V accuracies become higher, but the MCD also becomes higher. The same tendency can also be observed in the spectral domain. The more adaptive blocks the model has, the more harmonic components the generated speech has. However, overenhanced harmonic structures generate significantly robotic and unnatural sounds.

Since the balanced $C2B_F5$+$C2B_A5$ model achieves the highest pitch accuracy and lowest U/V error while keeping acceptable spectral accuracy and attaining longer *receptive fields* than the $C3B_F5$+$C1B_A5$ model, the 20 residual blocks with balanced numbers of adaptive and fixed blocks was selected as the QPPWG paradigm. To summarize, the ratio of adaptive and fixed blocks is crucial to the network for avoiding over/undermodeling the harmonic structures, and this observation is consistent with the experimental results in Section 5.5. Moreover, since one chunk including 10 fixed

Table 6.4: *QP structure of QPPWG_20 generator*

QP structure	stacked			parallel		
	MCD	RMSE	U/V	MCD	RMSE	U/V
$1 \times F_0$	**5.10**	**0.14**	**18**	5.80	0.32	28
$1/2 \times F_0$	**5.49**	**0.18**	**21**	6.05	0.48	43
$2 \times F_0$	6.32	**0.14**	**26**	**6.00**	0.45	52
Average	**5.63**	**0.15**	**22**	5.95	0.42	41

blocks showed effectiveness in the PWG and WN models, and the *receptive fields* of 10 fixed blocks are longer than that of two chunks with five fixed blocks, the architecture of the following QPPWG models was set to $\mathrm{C1B_F10{+}C2B_A5}$. The QPPWG architecture is denoted as QPPWG_20.

QP Structure

Since the fixed and adaptive blocks are assumed to respectively model aperiodic and periodic components of speech signals, a parallel QP structure (Fig. 6.3 (b)) was evaluated in this section compared to the original stacked QP structure (Fig. 6.3 (a)). The CNN channel size was 16, and the dense factor was 4, too. The results in Table 6.4 show that the QPPWG_20 model with a parallel QP structure achieves very low pitch accuracy and high U/V errors, which indicate the very limited periodic component modeling capability of the parallel model. Observing the output waveforms of the skip connection summation from the adaptive/fixed blocks, we also find that the output waveforms are dominated by the fixed blocks in the parallel QP model while the outputs of the adaptive blocks are very small. In other words, these results show that only the fixed blocks are well activated for speech modeling when the parallel QP structure is adopted.

The possible reason is that the difficulty of modeling speech using a fixed network architecture is lower than that of the network adopting a more complicated pitch-adaptive architecture in the very initial stage. Since the gradient paths of the fixed and adaptive macroblocks are separated, this difference of modeling difficulty may make the whole adaptive macroblock inactive. On the other hand, because the adaptive and fixed macroblocks are cascaded in the stacked QP structure, these macroblocks are in the same gradient flow, which makes the entire network participates in the speech modeling. Furthermore, since the aperiodic and periodic components are not completely independent, the stacked QP structure takes advantage of the aperiodic and periodic information propagations between the fixed and adaptive macroblocks to get better speech modeling capability. As a result, the stacked QP structure was selected as the QPPWG paradigm. Further discussion and more details about the outputs of the adaptive and fixed macroblocks will be presented in Section 6.5. Moreover, the cascaded adaptive to fixed macroblock order is denoted as af, and the reversed macroblock order is denoted as fa. The effectiveness of the macroblock order will be presented in the overall objective evaluation.

Dense Factor

The dense factor is inversely proportional to the receptive field length, and the QPPWGaf_20 models with 16 CNN channels and 2^0–2^4 dense factors were evaluated. The results in Table 6.5 show that while the models with dense factors of 2^2–2^4 achieve similar generative performance, the models with dense factors of 2^0 and 2^1 achieve slightly worse performance. A similar tendency was also observed by listening to the generated speech. The generated utterances from the models with dense factors of 2^0 and 2^1 were more unstable. Furthermore, PDCNN degenerates to DCNN when E_t is

Table 6.5: *Dense factor of QPPWGaf_20 generator*

	Dense	2^0	2^1	2^2	2^3	2^4
MCD (dB)	$1 \times F_0$	5.36	5.35	**5.10**	5.26	5.26
	$1/2 \times F_0$	5.61	5.61	**5.49**	5.57	5.64
	$2 \times F_0$	6.03	5.99	6.32	6.06	**5.92**
	Average	5.67	5.65	5.63	5.63	**5.60**
RMSE of $\log F_0$	$1 \times F_0$	0.17	0.14	0.14	**0.13**	**0.13**
	$1/2 \times F_0$	0.28	0.23	0.18	**0.17**	0.21
	$2 \times F_0$	0.15	0.14	0.14	0.14	0.14
	Average	0.20	0.17	0.15	**0.14**	0.16
U/V error (%)	$1 \times F_0$	17	17	18	17	17
	$1/2 \times F_0$	27	24	**21**	**21**	25
	$2 \times F_0$	20	20	26	**19**	24
	Average	21	20	22	**19**	22

one, and a larger dense factor makes E_t closer to one for more F_0 values. Therefore, since a lower dense factor attains a longer receptive field expansion and a higher lower bound of F_0, which makes PDCNN degenerate to DCNN, the dense factors of the following QPPWG models were set to 2^2.

Overall Objective Evaluation

An overall objective evaluation was conducted including the WD, QPNet, PWG, and QPPWG models. Specifically, since the AR QP structure has shown effectiveness for the WN [38, 39] vocoder, it is interesting to explore the generality of the QP structure for non-AR models and the performance difference between the QPNet and QPPWG models. Because the QPNet adopts an architecture including four chunks with four residual blocks, the PWG and QPPWG models with the same architecture were also

Table 6.6: *Number of trainable parameters (G: Generator; D: Discriminator)*

	QPNet	PWG		
	-	30	20	16
Macro 0	$C3B_F4$	$C3B_F10$	$C2B_F10$	$C4B_F4$
Macro 1	$C1B_A4$	-	-	-
G ($\times 10^6$)	24	1.16	0.78	0.63
D ($\times 10^6$)	-	0.10	0.10	0.10
		QPPWG		
	af_20	af_16	fa_20	fa_16
Macro 0	$C2B_A5$	$C2B_A4$	$C1B_F10$	$C2B_F4$
Macro 1	$C1B_F10$	$C2B_F4$	$C2B_A5$	$C2B_A4$
G ($\times 10^6$)	0.79	0.63	0.79	0.63
D ($\times 10^6$)	0.10	0.10	0.10	0.10

evaluated. Moreover, the effectiveness of the different QPPWG macroblock orders was also explored. The number of CNN channels of the PWG and QPPWG models was set to 64 following the original setting. The model sizes (numbers of trainable parameters) are shown in Table 6.6. Since the model size is proportional to the square of the number of CNN channels, the model size of vanilla PWG is only 5 % of that of QPNet because of the greatly reduced number of CNN channels. The sizes of the QPPWG models were reduced further by 30–50 % because of the reduced number of residual blocks compared with that of vanilla PWG.

According to the MCD results shown in Table 6.7 and 6.8, the QPPWG models with the af order still achieve higher spectral accuracy than the QPPWG models with the fa order. The QPPWG models with 20 residual blocks also predictably outperform the QPPWG models with only 16 residual blocks. Moreover, the QPPWGaf_20 model achieves a comparable spectral accuracy with the PWG_30 and PWG_20 models. On

Table 6.7: *Overall comparison*

		WD	QPNet	PWG_30	PWG_20	PWG_16
MCD (dB)	$1 \times F_0$	2.58	4.20	**3.69**	3.74	4.25
	$1/2 \times F_0$	3.89	4.92	4.47	**4.39**	4.65
	$2 \times F_0$	3.79	4.61	5.24	5.06	4.56
	Average	3.42	4.58	4.46	**4.40**	4.49
RMSE of $\log F_0$	$1 \times F_0$	0.10	0.14	0.12	0.15	0.41
	$1/2 \times F_0$	0.14	0.23	0.27	0.32	0.42
	$2 \times F_0$	0.10	0.18	0.15	0.15	0.73
	Average	0.11	0.19	0.18	0.21	0.51
U/V error (%)	$1 \times F_0$	10	**14**	**14**	15	55
	$1/2 \times F_0$	15	26	**21**	22	45
	$2 \times F_0$	11	22	**12**	17	66
	Average	12	21	16	18	55

the other hand, although the average MCD of PWG_16-generated utterances is not very high, the very high RMSE of $\log F_0$ and the very high U/V error indicate that the speech quality is low. Specifically, the similar MCDs of PWG_16-generated utterances with different scaled F_0 values imply that the PWG_16 model tends to ignore the F_0 scaled ratio to generate similar speech waveforms. The very high RMSE of $\log F_0$ and the very high U/V error also indicate that the PWG_16-generated speech waveforms lack fine harmonic structures. On the other hand, compared to QPNet, although the model size of QPPWGaf_16 is much smaller than that of QPNet, the non-AR mechanism and GAN structure still make QPPWG achieve comparable spectral prediction accuracy.

The results of the F_0 RMSE and U/V error in Table 6.7 and 6.8 also show that the non-AR PWG models already achieve a comparable pitch accuracy with the AR QPNet model, and the possible reason is that the GAN structure greatly improves the speech modeling capability. However, the QP structure further improves the pitch

Table 6.8: *Overall comparison (continued)*

		QPPWGaf_20	QPPWGaf_16	QPPWGfa_20	QPPWGfa_16
MCD (dB)	$1 \times F_0$	3.80	4.18	4.54	4.99
	$1/2 \times F_0$	4.52	4.89	5.18	5.60
	$2 \times F_0$	4.92	5.42	5.61	5.97
	Average	4.41	4.83	5.11	5.52
RMSE of $\log F_0$	$1 \times F_0$	0.11	**0.10**	0.11	0.12
	$1/2 \times F_0$	0.19	**0.15**	0.20	0.19
	$2 \times F_0$	0.11	**0.10**	0.11	0.11
	Average	0.14	**0.12**	0.14	0.14
U/V error (%)	$1 \times F_0$	16	18	15	16
	$1/2 \times F_0$	23	22	23	22
	$2 \times F_0$	19	14	13	11
	Average	19	18	17	17

accuracy of the non-AR PWG models. The QPPWGaf_16 model even attains a similar pitch accuracy to the reference WD vocoder. Moreover, although the pitch and U/V accuracies of PWG_16 markedly degrade because of the short receptive field, the QPPWGaf_16 model significantly improves them to an acceptable level. In conclusion, the QP structure efficiently increases the effective receptive field size and introduces the prior periodicity information to the network, resulting in a comparable spectral accuracy, a much higher pitch accuracy, and a smaller model size. The objective results show the effectiveness of the proposed QP structure for the PWG models.

On the other hand, since the WD-extracted *mcep* and F_0 are not completely independent, taking *mcep* extracted from natural speech as the ground truth of the scaled F_0 scenarios might cause some mismatches. However, the objective evaluations still provide meaningful information about the performance of these vocoders, and we also conducted the subjective evaluation in the following subsection to provide convincing

results from different aspects.

6.4.4 Subjective Evaluation

The subjective evaluation set was composed of 1680 synthesized and 80 natural utterances. The synthesized utterances were generated by seven vocoders conditioned on three F_0 scaled ratios (unchanged, halved, and doubled) and four speakers (the VCC2018 SPOKE set). For each vocoder, speaker, and F_0 scaled ratio, we randomly selected 20 utterances from the 35 testing utterances for both mean opinion score (MOS) and ABX evaluations. Specifically, the speech quality of each utterance was evaluated by listeners assigning MOSs of 1–5. The higher the MOS, the better the speech quality. For each ABX, two testing utterances were compared with one reference, and the listeners chose the one whose pitch was more consistent with that of the reference. Eight listeners evaluated part of the subjective evaluation set in both MOS and ABX tests, and each utterance/pair was evaluated by at least two listeners. Although the listeners were not native English speakers, they worked on audio-related research. The demo utterances can be found on our demo page[1].

MOS of Speech Quality

The vocoders of WD, QPNet, PWG of three different sizes, and QPPWG of two different sizes were involved in the MOS evaluation. The results shown in Figs. 6.5 and 6.6 are presented for three different F_0 scaled ratios for male and female speakers, respectively. The overall results show that the proposed QP structure improves the speech modeling capacity of the PWG vocoders. In particular, while the PWG_16

[1]https://bigpon.github.io/QuasiPeriodicParallelWaveGAN_demo/

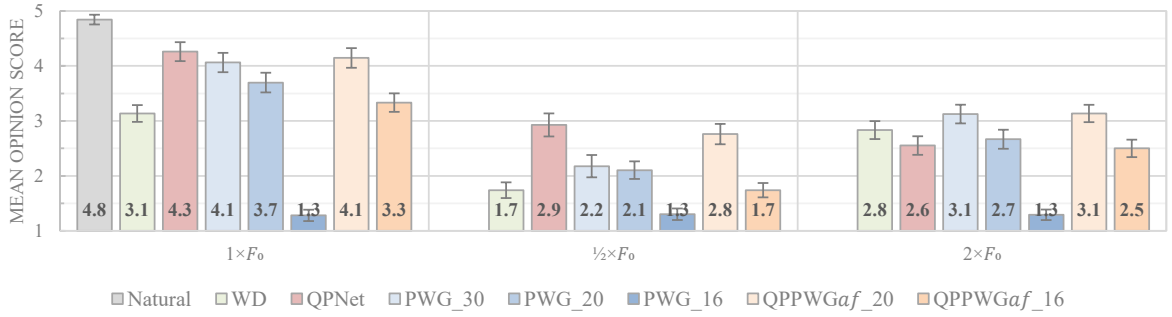


Figure 6.5: *Speech quality MOS evaluations of male speakers with 95 % CI.*

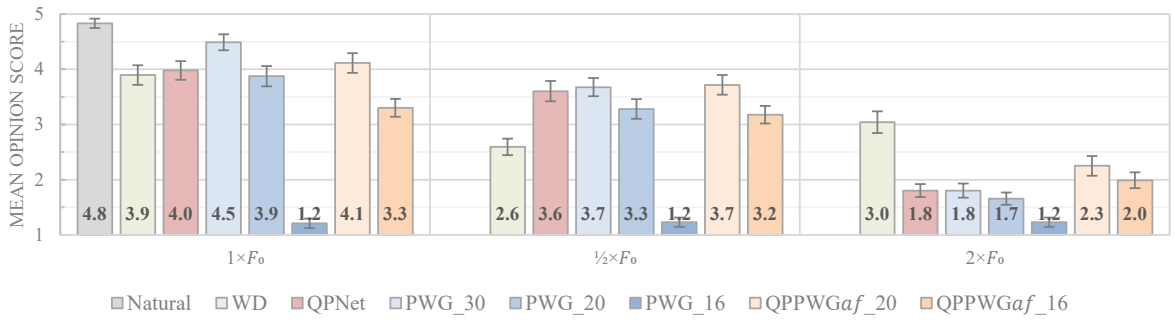


Figure 6.6: *Speech quality MOS evaluations of female speakers with 95 % CI.*

vocoder achieves very low quality because of the very small receptive field, the same-size QPPWG_16 markedly outperforms the PWG_16 for all scenarios in the MOS evaluation. However, since the performance of the QPPWG_16 still worse than the vanilla PWG (PWG_30), the following discussion focuses on comparisons among QPPWGaf_20, PWG_30, and QPNet.

For the halved F_0 scenario, the QPPWGaf_20 vocoder markedly outperforms the PWG_30 and WD vocoders and attains a similar speech quality to the QPNet vocoder for the male set. For the female set, the QPPWGaf_20 vocoder is comparable to the PWG_30 and QPNet vocoders while still outperforming the WD vocoder. The results indicate that the models with the QP structure are more robust for an unseen F_0 outside the F_0 range of the training data, such as most of the half F_0 values in

the male set. On the other hand, although the combination of the half F_0 and other acoustic features in the female set is still unseen, the scaled F_0 values are almost in the F_0 range of the training data. Therefore, the PWG_30 vocoder can still achieve a similar speech quality to the QPPWGaf_20 vocoder.

For the doubled F_0 scenario, because most of the scaled F_0 values of the male set are in the F_0 range of the training data, the performance of the QPPWGaf_20 vocoder is similar to that of the PWG_30 vocoder for the male set. The QPPWGaf_20 vocoder outperforms the WD and QPNet vocoders in the male set, while the QPNet vocoder achieves an inferior speech modeling capacity for the doubled F_0 scenario [38, 39]. On the other hand, although the QPPWGaf_20 vocoder predictably outperforms the PWG_30 and QPNet vocoders in the doubled female F_0 scenario, the WD vocoder achieves a higher speech quality than the QPPWGaf_20 vocoder. A possible reason for this is that many PDCNNs of the QPPWGaf_20 model might degenerate to DCNNs because of the values of E_t close to one due to the very high F_0 values.

In conclusion, the proposed QPPWG vocoder with 20 residual blocks attains comparable speech quality to the PWG vocoder with 30 residual blocks for natural auxiliary features even though the model size is only 70 % of that of the PWG model. When conditioned on the auxiliary features with the unseen F_0 values, which are outside the F_0 range of the training data, the proposed QPPWG vocoders achieve a higher speech quality than the PWG vocoders. The results confirm the effectiveness of the proposed QP structure for the PWG model in efficiently modeling speech signals and dealing with unseen F_0 features using the prior periodicity knowledge.

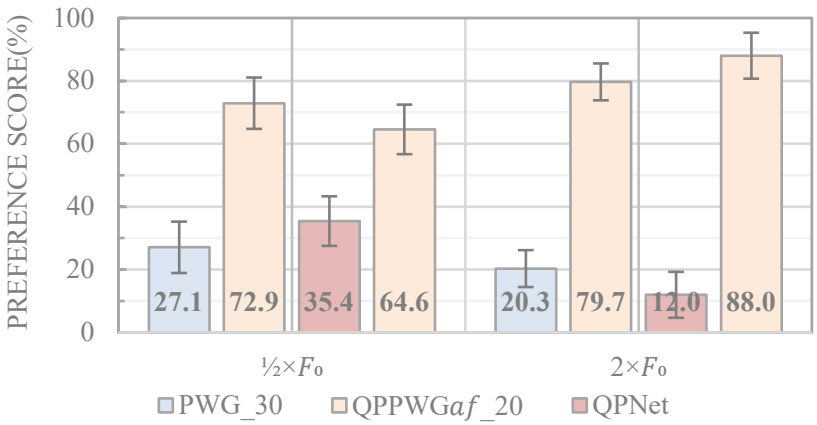


Figure 6.7: *Pitch accuracy ABX evaluations with 95 % CI.*

ABX of Pitch Accuracy

To evaluate the perceptual pitch accuracy, ABX tests of the QPPWGaf_20, PWG_30, and QPNet vocoders were conducted while the WD-generated utterances taken as references. Note that since there were no natural utterances with scaled F_0 and the conventional signal-processing-based vocoder usually attains accurate pitch controllability, the WD-generated utterances were an alternative ground truth. Since the speech quality of the WD-generated speech is usually worse than the neural-vocoder-generated-speech, we asked the listeners to focus on the pitch differences and ignore the speech quality differences. Because the results of the female and male sets have the same tendency, only the overall results are shown in Fig. 6.7. We find that the perceptual pitch accuracy of the proposed QPPWGaf_20 vocoder is much better than that of the PWG and QPNet vocoders for both halved and doubled F_0 scenarios. To summarize, the ABX results show perceptible pitch differences between QPPWG- and PWG-/QPNet-generated utterances, and the ABX experimental results are consistent with the objective results of the RMSE of $\log F_0$.

6.5 Discussion

In this section, we further explore the internal speech generative mechanism of the QPPWG. Specifically, the visualized cumulative intermediate outputs of residual blocks are presented to show the internal process of speech generation. Furthermore, the statistical results of the effective receptive fields are also presented for easily comparing the modeling capacity of the QPPWG with that of the PWG by the receptive field lengths. Last, a discussion about the PDCNN and classical deformable CNN are presented.

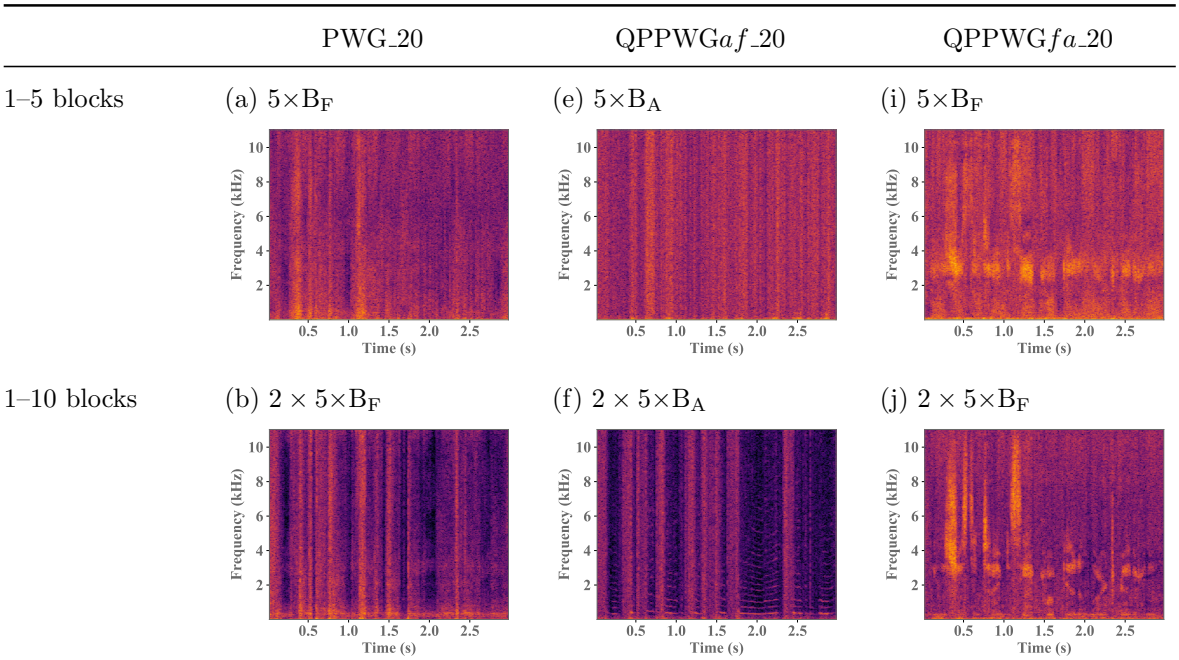


Figure 6.8: *Intermediate cumulative outputs.*

6.5.1 Understanding of QP Structure

Because of the direct waveform outputs of PWG/QPPWG, we can easily dissect the models to explore the internal speech modeling mechanisms. Specifically, the raw

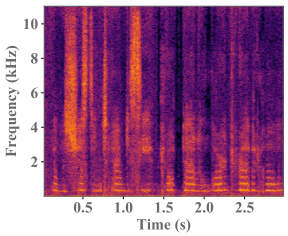
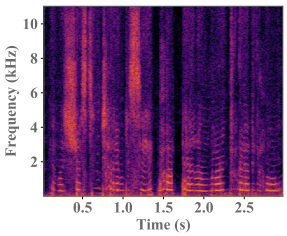
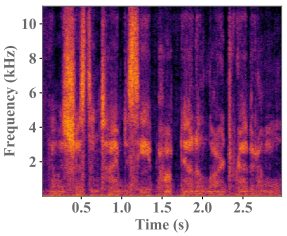
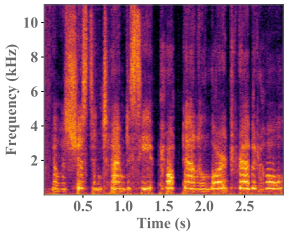
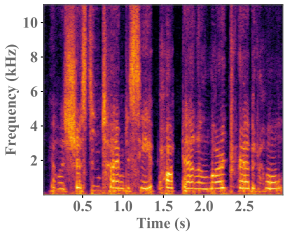
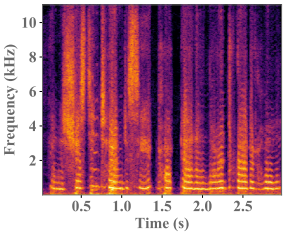
	PWG_20	QPPWGaf_20	QPPWGfa_20
1–15 blocks	(c) $3 \times 5{\times}\mathrm{B_F}$	(g) $2 \times 5{\times}\mathrm{B_A} + 5{\times}\mathrm{B_F}$	(k) $2 \times 5{\times}\mathrm{B_F} + 5{\times}\mathrm{B_A}$
			
1–20 blocks	(d) $4 \times 5{\times}\mathrm{B_F}$	(h) $2 \times 5{\times}\mathrm{B_A} + 2 \times 5{\times}\mathrm{B_F}$	(l) $2 \times 5{\times}\mathrm{B_F} + 2 \times 5{\times}\mathrm{B_A}$
			

Figure 6.8: *Intermediate cumulative outputs (continued).*

waveform outputs of the PWG/QPPWG models are the cumulative results of the skip connection outputs from the residual blocks. Therefore, the speech modeling behavior of the residual blocks can be explored via the visualized intermediate outputs of partial residual blocks. Spectrograms of the intermediate outputs of the cumulative residual blocks are presented in Fig. 6.8. For the PWG vocoder results (Figs. 6.8 (a)–(d)), the spectrogram contains more details and textures as the number of cumulative residual blocks increases. In contrast to the PWG vocoder, which gradually adds both harmonic and non-harmonic components to the spectrogram, the first 10 adaptive blocks of the QPPWGaf vocoder mostly focus on modeling the harmonic components as shown in Fig. 6.8 (f). By contrast, the first ten fixed blocks of the QPPWGfa vocoder mostly generate the non-harmonic part of the speech as shown in Fig. 6.8 (j). The results confirm our assumption that the adaptive blocks with the PDCNNs primarily model

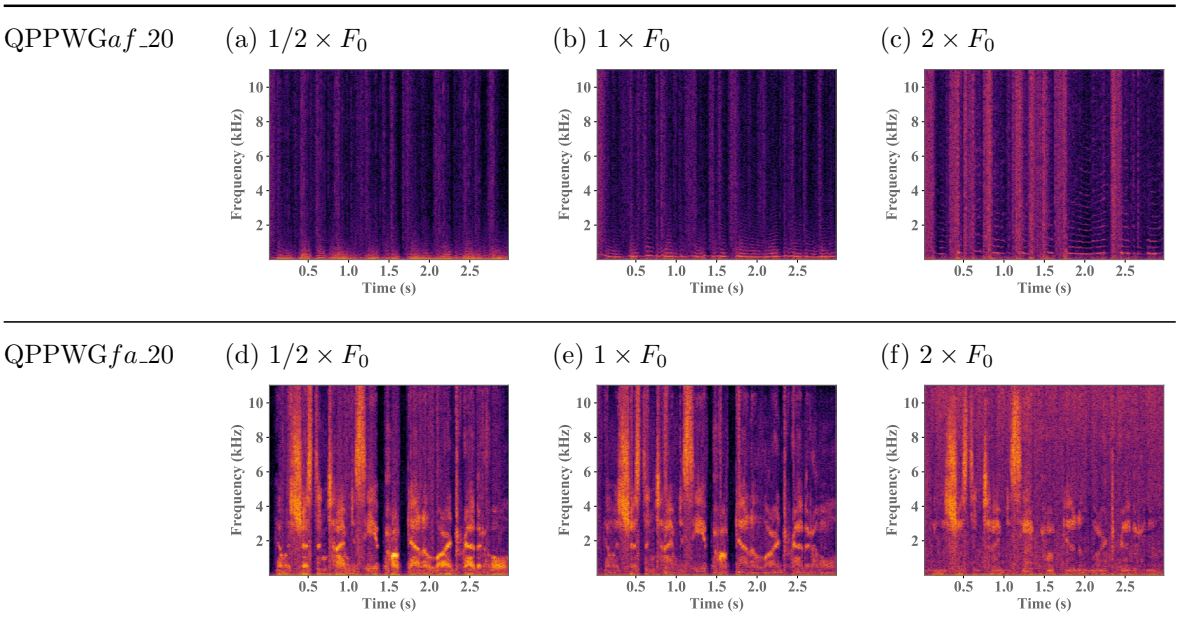


Figure 6.9: *Cumulative intermediate outputs of 1–10 blocks with scaled* F_0.

the pitch-related speech components with long-term correlations, while the fixed blocks with the DCNNs mainly focus on the spectrum-related components with short-term correlations.

In addition, to explore the behaviors of the adaptive and fixed blocks for different scaled F_0 features, comparisons among the visualized cumulative outputs of the first 10 residual blocks from the QPPWGaf and QPPWGfa vocoders are presented. The spectrograms of QPPWGaf shown in Figs. 6.9 (a)–(c) have similar structures along the time axis but increasingly stretched harmonic structures along the frequency axis as F_0 increases. By contrast, despite the different F_0 scaled ratios, both the frequency and temporal structures of the spectrograms of QPPWGfa shown in Figs. 6.9 (d)–(f) are similar. The results imply that the adaptive blocks primarily model the pitch-dependent harmonic components and the fixed blocks mainly focus on the pitch-independent non-harmonic components.

Although the QPPWG vocoder is a unified NN-based waveform generative model, the generative mechanism of its QP structure is similar to that of a source–filter model. The cascaded adaptive (pitch-dependent) and fixed macroblocks of the QP structure are analogous to the excitation generation and spectral filtering of the source–filter model. Furthermore, since scaling F_0 mostly affects the excitation generation parts, which is modeled by the adaptive blocks, it may be the possible reason for the QPPWGaf outperforming the QPPWGfa with the scaled F_0. Specifically, the adaptive macroblock of the QPPWGaf first models the corresponding excitation signal based on the scaled F_0, and then pass this information to the fixed macroblock. Therefore, when the fixed macroblock of the QPPWGaf models the spectral information, it already had plentiful information about the F_0 scaled excitation signal. However, when the fixed macroblock of the QPPWGfa models the spectral information, which is less related to the F_0, the generated spectral may not be well matched to the scaled F_0. In conclusion, because a vocoder is assumed to have the capability for independently controlling each speech component, the QPPWG vocoder is more consistent with the definition of a vocoder while having a more tractable and interpretable architecture. More details of the visualized intermediate outputs can be found on our demo page[2].

6.5.2 Effective Receptive Field

The experimental results in Chapter 5 show that the speech modeling capacity of an AR vocoder is strongly related to the length of its receptive field, and we argue that a non-AR vocoder has a similar tendency. Specifically, the receptive field length of PWG_30 is 6139 samples ($2^0 + \cdots + 2^9 = 1023$ samples for one side in each chunk. The total length includes three chunks with two sides plus one sample) and that of PWG_20

[2]https://bigpon.github.io/QuasiPeriodicParallelWaveGAN_demo/

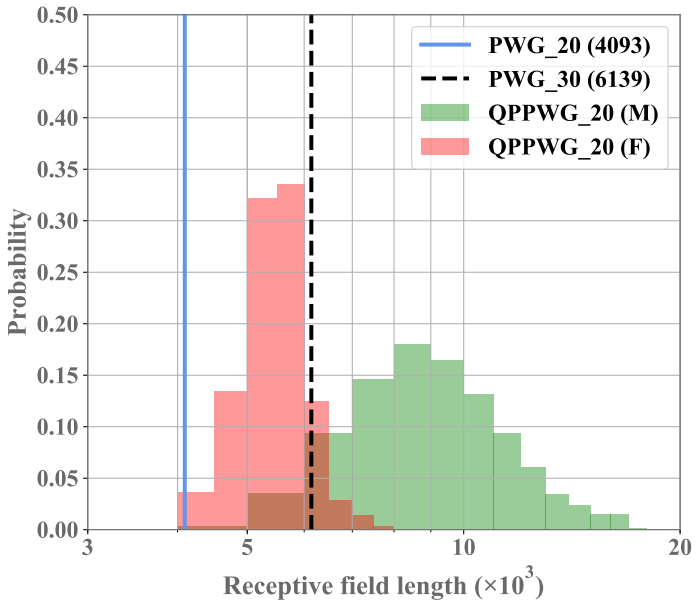


Figure 6.10: *Receptive field lengths of PWG_30, PWG_20, and QPPWG_20 for male (M) and female (F) sets.*

is 4093 samples. For the QPPWG, the effective receptive field length is the summation of 2047 samples for $\mathrm{C1B_F10}$ and $124 \times E_t$ samples (two chunks with two sides, and each side in a chunk has $2^0 + \cdots + 2^4 = 31$ samples) for $\mathrm{C2B_A5}$. The male F_0 range is around 40–240 Hz and the female F_0 range is around 100–400 Hz, so the E_t with the dense factor four of the male set is around 20–140 and that of the female set is around 10-–60. As shown in Fig. 6.10, most of the effective receptive filed lengths of QPPWG_20 for the male set are longer than the receptive filed length of PWG_30, which may result in the higher pitch accuracy and comparable speech quality of QPPWG. The slightly lower speech quality of QPPWG_20 than of PWG_30 for the female set may result from the shorter effective receptive fields of QPPWG_20. In conclusion, the speech modeling capacity of a non-AR vocoder is still strongly related to the receptive field length. The proposed QPPWG has longer effective receptive fields by skipping some

redundant samples of the periodic components. Although the network may also lose some details of the aperiodic components owing to the skipping mechanism, the overall experimental results still show the effectiveness of the QP structure.

6.5.3 Deformable Dilated Convolution

The idea of a dynamically updated attention mechanism, which makes a sequential network know "where to look" at each time step, is not new. Generative models [137–139] that utilize differentiable attention mechanisms to constrain the read and write operations of the network to specific parts of the scene have been proposed. To handle the limitation of the fixed geometric structure of the CNNs, the authors of [140] proposed a learnable spatial transformation of the input feature maps of the CNNs to regularize the input of each CNN layer. Moreover, the authors of [141] proposed a deformable convolution to enable the freeform deformation of the CNN sampling grid. The deformable convolution gives the network an adaptive receptive filed that focuses on different locations of the input feature map corresponding to the current conditions.

Since the offsets of the grid sampling locations in PDCNN are derived from the F_0 values, the proposed PDCNN is a special case of the deformable CNN. Specifically, both the deformable CNN and the PDCNN dynamically index the input feature map to implement the deformation of the CNN sampling grid, and the main difference between them is that the deformation index of the PDCNN is calculated from the sampling rate and instantaneous F_0 while that of the deformable CNN is predicted using a NN. As the deformable CNN with few additional parameters and computations, the PDCNN is implemented with a simple indexing technique without a large extra computational cost. As shown in Table 6.9, the average real-time factor (RTF) of the QPPWG_20 inferences is similar to that of PWG_20 and less than that of PWG_30 when running

Table 6.9: *RTF of Model Inference*

	PWG_20	PWG_30	QPPWG_20
Intel Xeon Gold 6142	**0.474**	0.579	0.512
Nvidia TITAN V	**0.011**	0.016	0.020

on an Intel Xeon Gold 6142 CPU (2.60 GHz and 32 threads). However, because of the different indexing processes of each CNN kernel, the parallelization of the CNN computation on a GPU is degraded. As shown in Table 6.9, although the model size of QPPWG_20 is only 70 % of that of PWG_30, the QPPWG_20 model has 170 % of the training time and 130 % of the inference time of the PWG_30 model when using an Nvidia TITAN V GPU. However, since the RTF of the PWG generation is much less than one, the additional inference time of QPPWG is insignificant.

6.6 Summary

In this chapter, the proposed QPPWG vocoder is introduced. Specifically, although the proposed QPNet in Chapter 5 has reduced 50 % model size of the WN, the generation speed of the QPNet is still far away from real-time generation. To achieve real-time generation, a compact NN-based vocoder, PWG, is adopted. However, because of the fixed geometric structure and data-driven nature without much prior speech knowledge, the PWG lacks pitch controllability. Therefore, the QPPWG vocoder has been proposed to introduce the prior periodicity information to the network using the QP structure. The QPPWG network architecture is dynamically adapted to the input F_0 feature of each input sample using the proposed non-AR PDCNN, and this pitch-dependent mechanism improves speech modeling efficiency and pitch controllability by

introducing the prior periodicity knowledge to the network.

Both objective and subjective experimental results show the effectiveness of the QP structure for the PWG vocoder. The QPPWG vocoder outperforms the PWG vocoder in pitch accuracy and speech quality for unseen scaled F_0 features while attaining a comparable speech quality to the PWG vocoder for natural F_0 features. Because of the more efficient receptive field expansion by PDCNNs, the model size of the QPPWG vocoder is only 70 % of that of the PWG vocoder.

Moreover, the visualized intermediate outputs of QPPWG vocoders confirm our assumption that adaptive blocks mainly model long-term correlations and fixed blocks focus on short-term correlations. The results also imply that although the QPPWG vocoder is a unified NN, the cascaded adaptive and fixed modules work like a source–filter model to respectively model excitation signal and spectral information. That is, the proposed QPPWG vocoder is a fast and simple waveform generative model with higher pitch controllability, smaller model size, and better interpretability and tractability than vanilla PWG. The effectiveness of the QPPWG vocoder also indicates the generality of the QP structure for different CNN-based speech generative models.

To summarize, in this chapter, the generation efficiency of our speech synthesis module has been significantly improved because of adopting a non-AR generative model, PWG. The proposed QPPWG vocoder also markedly improves the pitch controllability of the vanilla PWG vocoder because of the proposed PDCNN and QP structure. Both objective and subjective results show a superior pitch accuracy of the QPPWG-generated speech. Moreover, because of the more efficient speech modeling by the pitch-dependent architecture, the QPPWG vocoder reduces 30 % model size compared to the vanilla PWG vocoder while achieving similar speech quality with natural acoustic features and higher speech quality in the pitch transformation scenarios.

7 Conclusions

7.1 Summary of This Thesis

Speech generation is a technique of generating desired speech based on a specific input such as text. A speech generation system usually includes an analysis module to parametrize the input to a specific representation, a manipulation module to manipulate the representation according to the requirements or transfer the input representation to another proper representation(s) for speech synthesis, and a synthesis module to generate speech waveforms based on the manipulated representation. A speaker voice conversion (VC) task, which changes the speaker identity while maintaining the speech content, and a pitch transformation task are the two examples of speech generation systems discussed in this thesis, and the research mainly focuses on improving the synthesis module. Four fundamental challenges of the synthesis module, namely, quality, robustness, controllability, and generation efficiency, were studied, and the proposed methods introduced in this thesis are mainly related to improving the robustness against the distorted acoustic features and pitch controllability of neural-based speech generation models.

The main aim of this research is to advance neural-based speech generation models using prior knowledge of speech signals and speech production mechanisms. Specifically, since speech is a continuous sequential signal with long-term dependence, the neural model-generated speech should have the same characteristics such as continuity

and periodicity. As a result, a waveform constraint based on the speech continuity and a pitch-adaptive network based on the speech periodicity have been developed in my research. Moreover, since the human speech production mechanism is similar to a cascaded system with vocal fold vibrations and vocal tract resonance, conventional speech modeling techniques usually model speech production using a source–filter model. The source signal models the signal generated by vocal fold vibrations and the spectral filter models the vocal tract resonance. The cascaded structure with the prior knowledge of the speech production mechanism has also been applied to the proposed neural-based generative models, and it makes those models more tractable and interpretable.

In Chapter 2, the fundamental concepts of a vocoder and source–filter model were reviewed. The currently developed source–filter-based and unified-model-based neural vocoders with autoregressive and non-autoregression were also introduced. On the other hand, the background knowledge and techniques of VC with parallel and non-parallel training corpora were also reviewed.

In Chapter 3, the details of the baseline two-stage deep neural network (DNN)-based NU non-parallel VC system developed for VCC2018 were first described, and the main concept was to use text-to-speech (TTS) outputs as a bridge to connect non-parallel source and target speaker utterances. To improve the unimodal weakness and the lack of capability to predict the variance of the DNN-based model, the deep mixture density network (DMDN)-based VC model was also presented. Furthermore, since the mismatch between the two stages of the cascade conversion structure caused performance degradation, a compensation AutoEncoder to reduce the mismatch was described. Internal objective evaluation results showed that the baseline DNN-based VC system achieved a slightly worse spectral prediction accuracy than a parallel VC, and the DMDN-based model slightly improved the spectral prediction accuracy. The

evaluation results also showed the potential capability for the spectral prediction accuracy improvement of the compensation AutoEncoder. Furthermore, the subjective evaluations provided by the VCC2018 organizer showed that the submitted baseline VC system achieved an above-average performance in both quality and similarity measurements.

In Chapter 4, the phenomena, possible reasons, and negative effects of the collapsed speech problem of the WaveNet (WN) vocoder were first described. To prevent the WN vocoder from generating unexpected and non-speech-like outputs (collapsed speech), collapsed speech detection and suppression techniques were introduced. The collapsed speech segment detection (CSSD) technique segmentally detects the collapsed speech segments of the WN-generated speech, and the proposed suppression technique was applied to the detected segment to suppress the collapsed speech. The suppression technique is a waveform-based distribution constraint, which constraints the WN output distribution according to the relationships among reference speech samples. Since it is usually stable and collapse-free, the WORLD-generated speech was adopted as the reference speech. The sample-based relationships are described by linear prediction coding (LPC) coefficients, so the waveform-based constraint is called the LPC distribution constraint (LPCDC). The subjective evaluation results showed that the WN vocoder with the proposed CSSD and LPCDC significantly improved the speech quality and maintained the same speaker similarity of the baseline VC system.

In Chapter 5, a WN-like audio waveform generation model named QPNet was introduced. The QPNet models quasi-periodic and high-temporal-resolution audio signals with a pitch-dependent dilated convolutional neural network (PDCNN) and a cascaded network structure in an autoregressive (AR) manner. The PDCNN is a variant of a dilated convolutional neural network (DCNN), and the dilation size of the PDCNN is

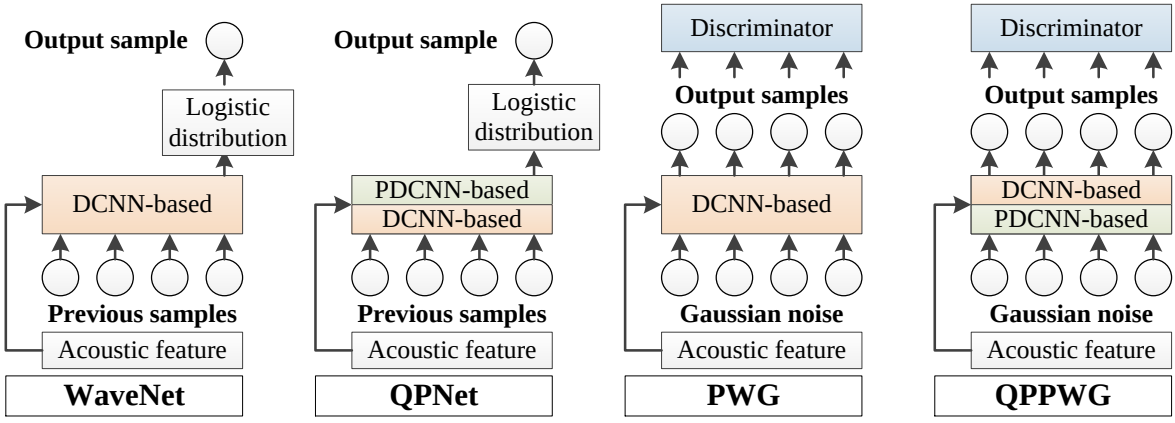


Figure 7.1: *Neural vocoders in this thesis.*

dynamically changed corresponding to the input fundamental frequency feature (F_0) for modeling the long-term correlations of audio samples. The cascaded structure was induced by the conventional source–filter model to respectively model the spectrum and pitch-related components of speech samples by the cascaded fixed and adaptive modules. Specifically, the adaptive module adopts PDCNNs to model the long-term correlations with a pitch-adaptive network architecture, and the fixed module adopts DCNNs to model the short-term correlations with a fixed network architecture. A sine wave generation task was presented to show the effectiveness of the proposed PDCNN for generating sine waves with unseen frequencies. In the pitch transformation task, the QPNet vocoder achieved higher pitch controllability and similar speech naturalness to the double-size WN vocoder. In the VC task, the QPNet vocoder also achieved similar naturalness and speaker similarity to the double-size WN vocoder. In conclusion, the QP structure with the PDCNN improves the pitch controllability and speech modeling efficiency of the WN vocoder.

In Chapter 6, quasi-periodic parallel WaveGAN (QPPWG) was presented. The QPPWG is a parallel WaveGAN (PWG)-based model with the proposed QP structure

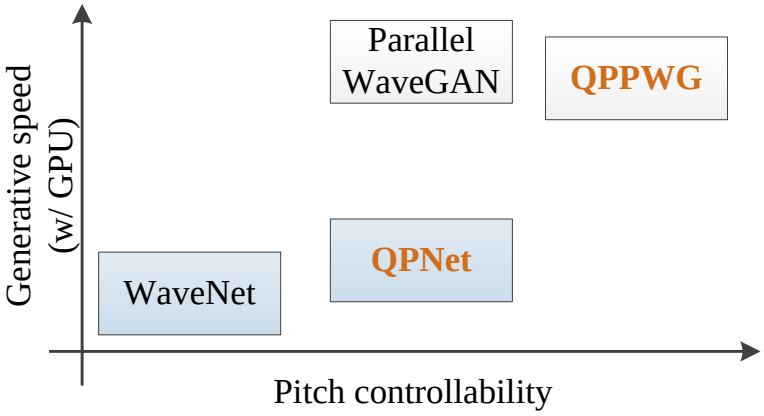


Figure 7.2: *Performance comparison of the neural vocoders in this thesis.*

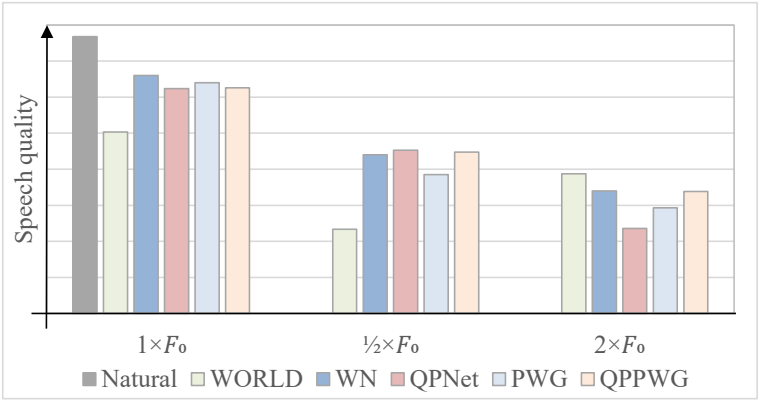


Figure 7.3: *Speech quality comparison of the vocoder-generated speech in this thesis.*

and non-AR PDCNNs to improve the pitch controllability of the PWG by introducing the prior periodicity information of speech signals. The architecture differences of the WN, QPNet, PWG, and QPPWG vocoders are shown in Fig. 7.1. Since the PWG is a non-AR speech generation model with a compact model size, which is less than 3 % of that of the WN, the generation speed of the PWG is much higher than the real time. Although the PDCNNs slightly degrade the parallelization of the non-AR CNN computation, the generation speed of the QPPWG is still much higher than the real time because of the non-AR mechanism and an even smaller model size.

In the pitch transformation task, although the PWG had already achieved similar pitch accuracy to the QPNet because of the assistance of the GAN structure, the QPPWG still markedly improved the pitch accuracy of the PWG. Moreover, the QPPWG vocoder even outperformed the PWG vocoder in speech quality for unseen scaled F_0 features while attaining a comparable speech quality to the PWG vocoder for natural F_0 features. Overall, for the natural acoustic features (unchanged F_0), the speech qualities of the WN, QPNet, PWG, and QPPWG-generated utterances are similar and markedly higher than those of the WORLD-generated utterances, which shows the effectiveness of the neural vocoders. For the scaled F_0 scenarios, the QPPWG-generated utterances achieve slightly higher speech quality and much higher pitch accuracy than other vocoder-generated utterances. The pitch controllability, generation speed, and speech quality performance characteristics of the neural vocoders in this thesis are shown in Figs. 7.2 and 7.3, and both objective and subjective results in Chapters 5 and 6 showed the effectiveness of the proposed QP structure and PDCNN for improving the speech modeling efficiency and pitch controllability of neural-based speech generation models. In addition, the visualized intermediate outputs of QPPWG vocoders also showed the higher tractability and interpretability of the neural vocoders with the QP structure.

In conclusion, this thesis demonstrates several ways to introduce the prior knowledge of speech signals and speech production mechanisms into data-driven unimodal neural-based vocoders to improve the robustness against the distorted acoustic features, pitch controllability, and speech modeling efficiency of the neural-based vocoders. Both objective and subjective results showed the effectiveness of the proposed methods. Moreover, visualized results are also provided to show the internal behaviors of the neural-based vocoders for understanding how the proposed methods work in the neu-

ral networks. The main contribution of this thesis is as follows. Although end-to-end neural networks are the mainstream of the current research for avoiding oversimplified assumptions causing mismatches, properly introducing the signal-related prior knowledge into the neural networks not only improves their capability but also makes these neural networks more than a black box. That is, the internal generative behaviors and mechanisms of the neural networks are more controllable and consistent with our understanding and observation of signals.

7.2 Future Work

Although the proposed methods greatly ease the collapsed speech issue and improve the pitch controllability and speech modeling efficiency of the WN and PWG vocoders, several challenges should be addressed in the future.

7.2.1 Collapsed Speech Detection and Suppression

For the collapsed speech detection, although the waveform-envelope-based detection is simple and efficient, the equal error rate (EER) is still around 20 %, which means that there is room for improvement. There are several possible directions such as a better reference speech than the WORLD-generated speech and a more detailed label of training data. For the collapsed speech suppression, since the proposed method is a Gaussian distribution with a specific mean and variance derived from linear prediction coding (LPC) results, the LPC can be easily replaced with other advanced speech coding techniques such as linear spectral pairs (LSP) to obtain a better mean and variance.

7.2.2 Pitch-dependent Dilated Convolution

The dilation size of the PDCNN is determined from the input F_0. In this thesis, the WORLD-extracted F_0 from the auxiliary acoustic features is directly adopted. However, since the performance of the PDCNN is assumed to be highly related to the accuracy of the input F_0, adopting advanced F_0 extraction methods may improve the performance of the PDCNN. Furthermore, only the PDCNN with the F_0 of natural speech and scaled F_0 has been evaluated, but the robustness of the PDCNN with the F_0 of noisy speech is an interesting question. On the other hand, the dilation size expansion of the stacked PDCNNs directly follows that of the stacked DCNNs in the WN. However, other possible expansion methods that are more consistent with the speech production mechanism may improve the tractability and interpretability of the networks.

7.2.3 Quasi-Periodic Structure

According to the evaluation results in Chapters 5 and 6, the ratio of the fixed to adaptive blocks of the QP structure is highly related to the performance. However, because of the GPU memory limitation, only a few possible combinations of the fixed and adaptive blocks were evaluated. A more efficient way to explore the optimized ratio of the fixed to adaptive blocks is necessary in future work. Furthermore, the adaptive-to-fixed macroblock order achieves a similar performance to the fixed-to-adaptive macroblock order in the QPNet while significantly outperforming the fixed-to-adaptive macroblock order in the QPPWG. The root cause will be another interesting research topic in future work. This thesis only introduces the most straightforward and simple way to apply the PDCNN and QP structure to the WN and PWG. However, there is still room

for improvement in the usage and hyperparameters of the PDCNN and QP structure.

7.2.4 Real-time Generation

Although the proposed QPPWG has already achieved real-time generation on both a GPU and a CPU with multiple cores and threads, the generation speed of the QPPWG will become much lower on devices with limited computation power such as a single CPU without multiple cores and threads. Furthermore, since only utterance-based generation was evaluated in this thesis, the capability of the QPPWG for streaming generation is still a question. Because many applications are running on mobile devices nowadays, the low resource requirement and streaming generation are very important features for a speech synthesis system.

7.2.5 Prior Knowledge

In this thesis, only the basic speech continuity and periodicity and the conventional source–filter model were applied to the generative models in the speech generation tasks. However, more advanced knowledge of audio signals such as music can be introduced to the generative networks for other audio generation tasks. This thesis is just the beginning of applying prior knowledge to dynamically adapt the outputs or architectures of networks. Further advanced exploration of prior knowledge with neural networks will be an interesting future work.

7.2.6 More than Audio Synthesis

Since not only audio signals have periodicity, the PDCNN concept can be applied to other sequential signals. For example, many web applications collect many user data to analyze users behaviors. Since user behaviors usually include specific patterns and cycles related to time (e.g., the specific day in each week), modeling user behaviors using a dilated CNN with time-dependent dilation size is reasonable. In conclusion, sequential signals usually have some specific patterns, and a sequential signal modeling network can benefit from the adaptation corresponding to these implicit patterns.

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 advice and support to broaden my research work.

I would like to acknowledge my earnest thanks to Dr. Takuma Okamoto and Dr. Hisashi Kawai of National Institute of Information and Communications Technology, Japan, for their supports to help me finish our journal paper.

I would like to express my sincere appreciation to Professor Junichi Yamagishi of National Institute of Informatics, Japan, and Professor Daisuke Deguchi of Nagoya University for their helpful comments on this thesis.

I would especially like to express my humble gratefulness to NEC C&C Foundation for the opportunity to receive 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 assistance.

I would also like to convey my thanks to laboratory colleagues for their support, especially Dr. Kazuhiro Kobayashi of TARVO, Inc., Dr. Tomoki Hayashi of Human Dataware Lab. Co., Ltd., and Dr. Patrick Lumban Tobing for their aids on the research works.

I would like to express my wholehearted recognition to my family and my friends, especially Pei-Ho Tang, for their encouragement and support throughout my study and research life. Without their help, I cannot earn my Ph.D. title alone.

Lastly, I would like to thank everything that happened and each people I meet in Japan during my Ph.D. life, those will be my lifetime treasure. I will miss those good time we had in Japan forever.

References

[1] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[2] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1. IEEE, 1996, pp. 373–376.

[3] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *speech communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[4] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[6] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2. IEEE, 1996, pp. 629–632.

[7] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, vol. 1. IEEE, 1998, pp. 285–288.

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

[9] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[10] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[11] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*. IEEE, 2018, pp. 4779–4783.

[12] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *Proc. SSW9*, Sept. 2016, p. 125.

[13] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," in *Proc. ICLR*, Apr. 2017.

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

[15] 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," in *Proc. ICML*, July 2018, pp. 2415–2424.

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

[17] 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," in *Proc. ICML*, July 2018, pp. 3915–3923.

[18] W. Ping, K. Peng, and J. Chen, "ClariNet: Parallel wave generation in end-to-end text-to-speech," in *Proc. ICLR*, May 2019.

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

[20] S. Kim, S.-G. Lee, J. Song, J. Kim, and S. Yoon, "FloWaveNet : A generative flow for raw audio," in *Proc. ICML*, June 2019, pp. 3370–3378.

[21] N.-Q. Wu and Z.-H. Ling, "WaveFFJORD: FFJORD-based vocoder for statistical parametric speech synthesis," in *Proc. ICASSP*, May 2020, pp. 7214–7218.

[22] H. Kim, H. Lee, W. H. Kang, S. J. Cheon, B. J. Choi, and N. S. Kim, "WaveN-ODE: A continuous normalizing flow for speech synthesis," in *Proc. ICML*, 2020.

[23] R. Yamamoto, E. Song, and J.-M. Kim, "Probability density distillation with generative adversarial networks for high-quality parallel waveform generation," in *Proc. INTERSPEECH*, Sept. 2019, pp. 699–703.

[24] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, May 2020, pp. 6199–6203.

[25] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Proc. NeurIPS*, Dec. 2019, pp. 14 910–14 921.

[26] M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, "High fidelity speech synthesis with adversarial networks," in *Proc. ICLR*, Apr. 2020.

[27] 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*, Sept. 2018, pp. 1988–1992.

[28] Y.-C. Wu, P. L. Tobing, K. Kobayashi, T. Hayashi, and T. Toda, "Non-parallel voice conversion system with WaveNet vocoder and collapsed speech suppression," *IEEE Access*, vol. 8, pp. 62 094–62 106, 2020.

[29] H. Dudley, "Remaking speech," *The Journal of the Acoustical Society of America*, vol. 11, no. 2, pp. 169–177, 1939.

[30] M. R. Schroeder, "Vocoders: Analysis and synthesis of speech," *Proc. IEEE*, vol. 54, no. 5, pp. 720–734, 1966.

[31] J. L. Flanagan and R. Golden, "Phase vocoder," *Bell System Technical Journal*, vol. 45, no. 9, pp. 1493–1509, 1966.

[32] S. Singhal and B. Atal, "Improving performance of multi-pulse LPC coders at low bit rates," in *Proc. ICASSP*, vol. 9. IEEE, 1984, pp. 9–12.

[33] M. Schroeder and B. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *Proc. ICASSP*, vol. 10, Apr. 1985, pp. 937–940.

[34] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

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

[36] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the nitech hmm-based speech synthesis system for the blizzard challenge 2005," *IEICE transactions on information and systems*, vol. 90, no. 1, pp. 325–333, 2007.

[37] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[38] Y.-C. Wu, T. Hayashi, P. L. Tobing, K. Kobayashi, and T. Toda, "Quasi-Periodic WaveNet vocoder: A pitch dependent dilated convolution model for parametric speech generation," in *Proc. INTERSPEECH*, Sept. 2019, pp. 196–200.

[39] Y.-C. Wu, T. Hayashi, P. L. Tobing, K. Kobayashi, and T. Toda, "Quasi-Periodic WaveNet: An autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, (submitted).

[40] Y.-C. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, "Statistical voice conversion with Quasi-Periodic WaveNet vocoder," in *Proc. SSW10*, Sept. 2019, pp. 63–68.

[41] Y.-C. Wu, T. Hayashi, T. Okamoto, H. Kawai, and T. Toda, "Quasi-Periodic Parallel WaveGAN vocoder: a non-autoregressive pitch dependent dilated convolution model for parametric speech generation," in *Proc. INTERSPEECH*, Oct 2020.

[42] Y.-C. Wu, T. Hayashi, T. Okamoto, H. Kawai, and T. Toda, "Quasi-Periodic Parallel WaveGAN: A non-autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[43] Y.-C. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, "The NU non-parallel voice conversion system for the Voice Conversion Challenge 2018," in *Proc. Odyssey*, June 2018, pp. 211–218.

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

[45] 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*, 2018, pp. 1983–1987.

[46] 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*, June 2018, pp. 219–226.

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

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

[49] N. Adiga, V. Tsiaras, and Y. Stylianou, "On the use of WaveNet as a statistical vocoder," in *Proc. ICASSP*, Apr. 2018, pp. 5674–5678.

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

[51] 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*. IEEE, 2018, pp. 297–303.

[52] W.-C. Huang, Y.-C. Wu, H.-T. Hwang, P. L. Tobing, T. Hayashi, K. Kobayashi, T. Toda, Y. Tsao, and H.-M. Wang, "Refined WaveNet vocoder for variational autoencoder based voice conversion," in *Proc. EUSIPCO*. IEEE, 2019, pp. 1–5.

[53] 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*. IEEE, 2019, pp. 6815–6819.

[54] M. Airaksinen, "Analysis/synthesis comparison of vocoders utilized in statistical parametric speech synthesis," *Master's thesis, Aalto University*, 2012.

[55] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The journal of the acoustical society of America*, vol. 50, no. 2B, pp. 637–655, 1971.

[56] D. Wong, B.-H. Juang, and A. Gray, "An 800 bit/s vector quantization LPC vocoder," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 5, pp. 770–780, 1982.

[57] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proc. ICASSP*, vol. 8. IEEE, 1983, pp. 93–96.

[58] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, vol. 1, 1992, pp. 137–140.

[59] J. Makhoul, R. Viswanathan, R. Schwartz, and A. Huggins, "A mixed-source model for speech compression and synthesis," *The Journal of the Acoustical Society of America*, vol. 64, no. 6, pp. 1577–1581, 1978.

[60] A. V. McCree and T. P. Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 242–250, 1995.

[61] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, 2001.

[62] J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Towards an improved modeling of the glottal source in statistical parametric speech synthesis," 2007.

[63] J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Glottal spectral separation for parametric speech synthesis," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[64] B. Bollepalli, L. Juvela, and P. Alku, "Generative adversarial network-based glottal waveform model for statistical parametric speech synthesis," in *Proc. INTERSPEECH*, 2017, pp. 3394–3398.

[65] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "Waveform generation for text-to-speech synthesis using pitch-synchronous multi-scale generative adversarial networks," in *Proc. ICASSP*, May 2019, pp. 6915–6919.

[66] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "GELP: GAN-excited linear prediction for speech synthesis from mel-spectrogram," in *Proc. INTERSPEECH*, Sept. 2019, pp. 694–698.

[67] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, Dec. 2014, pp. 2672–2680.

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

[69] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter waveform models for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 402–415, 2020.

[70] K. Oura, K. Nakamura, K. Hashimoto, Y. Nankaku, and K. Tokuda, "Deep neural network based real-time speech vocoder with periodic and aperiodic inputs," in *Proc. SSW10*, Sept. 2019, pp. 13–18.

[71] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable digital signal processing," in *Proc. ICLR*, Apr. 2020.

[72] Z. Liu, K. Chen, and K. Yu, "Neural homomorphic vocoder," in *Proc. INTERSPEECH*, Oct. 2020, pp. 240–244.

[73] O. McCarthy and Z. Ahmed, "HooliGAN: Robust, high quality neural vocoding," *arXiv preprint arXiv:2008.02493*, 2020.

[74] 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*, Apr. 2018, pp. 5664–5668.

[75] F. Yu and K. Vladlen, "Multi-scale context aggregation by dilated convolutions," in *Proc. ICLR*, May 2016.

[76] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc.ICML*, 2010, pp. 807–814.

[77] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, June 2013, pp. 3–11.

[78] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Transactions on signal processing*, vol. 39, no. 2, pp. 411–423, 1991.

[79] M. Mathews, J. E. Miller, and E. David Jr, "Pitch synchronous analysis of voiced sounds," *The Journal of the Acoustical Society of America*, vol. 33, no. 2, pp. 179–186, 1961.

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

[81] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[82] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 954–964, 2010.

[83] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order eigen space using deep belief nets," in *Proc. INTERSPEECH*, 2013, pp. 369–372.

[84] 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 Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.

[85] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *Proc. SLT*. IEEE, 2012, pp. 313–317.

[86] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.

[87] Y.-C. Wu, H.-T. Hwang, C.-C. Hsu, Y. Tsao, and H.-M. Wang, "Locally linear embedding for exemplar-based spectral conversion," in *Proc. INTERSPEECH*, 2016, pp. 1652–1656.

[88] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 631–644, 2019.

[89] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "Atts2s-vc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," in *Proc. ICASSP*. IEEE, 2019, pp. 6805–6809.

[90] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," in *Proc. INTERSPEECH*, 2020.

[91] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 944–953, 2009.

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

[93] F.-L. Xie, F. K. Soong, and H. Li, "A KL divergence and DNN-based approach to voice conversion without parallel training sentences," in *Proc. INTERSPEECH*, 2016, pp. 287–291.

[94] T. Nakashika, T. Takiguchi, and Y. Minami, "Non-parallel training in voice conversion using an adaptive restricted boltzmann machine," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2032–2045, 2016.

[95] 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*. IEEE, 2016, pp. 1–6.

[96] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *Proc. INTERSPEECH*, Aug. 2017, pp. 3364–3368.

[97] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," in *Proc. INTERSPEECH*, Aug. 2017, pp. 1273–1277.

[98] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

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

[100] 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*. IEEE, 2018, pp. 5279–5283.

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

[102] C.-H. Lee and C.-H. Wu, "MAP-based adaptation for speech conversion using adaptation data selection and non-parallel training," in *Ninth International Conference on Spoken Language Processing*, 2006.

[103] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

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

[105] R. Aihara, T. Takiguchi, and Y. Ariki, "Multiple non-negative matrix factorization for many-to-many voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1175–1184, 2016.

[106] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on gaussian mixture model," in *Proc. ICSLP*, Sep. 2006, p. 2446–2449.

[107] T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," in *Proc. ICASSP*, vol. IV, April 2007, p. 2446–24.

[108] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Many-to-many eigenvoice conversion with reference voice," in *Proc. INTERSPEECH*, 2009, pp. 1623–1626.

[109] T. Masuda and M. Shozakai, "Cost reduction of training mapping function based on multistep voice conversion," in *Proc. ICASSP*, vol. 4. IEEE, 2007, pp. IV–693.

[110] H. Doi, T. Toda, T. Nakano, M. Goto, and S. Nakamura, "Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system," in *Proc. APSIPA*. IEEE, 2012, pp. 1–6.

[111] H.-T. Hwang, Y. Tsao, H.-M. Wang, Y.-R. Wang, and S.-H. Chen, "A probabilistic interpretation for artificial neural network-based voice conversion," in *Proc. APSIPA*. IEEE, 2015, pp. 552–558.

[112] P. L. Tobing, H. Kameoka, and T. Toda, "Deep acoustic-to-articulatory inversion mapping with latent trajectory modeling," in *Proc. APSIPA*. IEEE, 2017, pp. 1274–1277.

[113] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from hmm using dynamic features," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1995, pp. 660–663.

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

[115] 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," in *Proc. Odyssey*, June 2018, pp. 195–202.

[116] Y.-C. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, "Development of NU non-parallel voice conversion system 2018 (応用音響)," **電子情報通信学会技術研究報告**, vol. 117, no. 515, pp. 385–390, 2018.

[117] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, May 2015.

[118] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[119] K. Kobayashi and T. Toda, "sprocket: Open-source voice conversion software," in *Odyssey*, 2018, pp. 203–210.

[120] 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 *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[121] K. Kobayashi, S. Takamichi, S. Nakamura, and T. Toda, "The NU-NAIST voice conversion system for the voice conversion challenge 2016," in *Proc. INTERSPEECH*, 2016, pp. 1667–1671.

[122] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (mlsa) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, pp. 10–18, 1983.

[123] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis-a unified approach to speech spectral estimation," in *Third International Conference on Spoken Language Processing*, 1994.

[124] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 1171–1179.

[125] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *Proc. ICLR*, 2016, pp. 1–16.

[126] Y.-C. Wu, P. L. Tobing, K. Yasuhara, N. Matsunaga, Y. Ohtani, and T. Toda, "A cyclical post-filtering approach to mismatch refinement of neural vocoder for text-to-speech systems," in *Proc. INTERSPEECH*, Oct 2020.

[127] C. Jarne, "A heuristic approach to obtain signal envelope with a simple software implementation," *arXiv preprint arXiv:1703.06812*, 2017.

[128] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases for speech synthesis research," in *Tech. Rep. CMU-LTI- 03-177*, 2003.

[129] S. Dieleman, A. van den Oord, and K. Simonyan, "The challenge of realistic music generation: modelling raw audio at scale," in *Advances in Neural Information Processing Systems*, 2018, pp. 7989–7999.

[130] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," *arXiv preprint arXiv:1710.07654*, 2017.

[131] W. B. Kleijn, F. S. Lim, A. Luebs, J. Skoglund, F. Stimberg, Q. Wang, and T. C. Walters, "WaveNet based low rate speech coding," in *Proc. ICASSP*. IEEE, 2018, pp. 676–680.

[132] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hasegawa-Johnson, "Speech enhancement using bayesian WaveNet," in *Proc. INTERSPEECH*, 2017, pp. 2013–2017.

[133] D. Rethage, J. Pons, and X. Serra, "A WaveNet for speech denoising," in *Proc. ICASSP*. IEEE, 2018, pp. 5069–5073.

[134] C. E. Shannon, "Communication in the presence of noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[135] Q. Tian, X. Wan, and S. Liu, "Generative adversarial network based speaker adaptation for high fidelity WaveNet vocoder," in *Proc. SSW10*, Sept. 2019.

[136] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *Proc. ICLR*, Apr. 2020.

[137] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[138] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," *arXiv preprint arXiv:1410.5401*, 2014.

[139] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "DRAW: A recurrent neural network for image generation," in *Proc. ICML*, July 2015, pp. 1462–1471.

[140] M. Jaderberg, K. Simonyan, A. Zisserman, and K. kavukcuoglu, "Spatial transformer networks," in *Proc. NIPS*, Dec 2015, pp. 2017–2025.

[141] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. ICCV*, Oct. 2017, pp. 764–773.

List of Publications

Journal Papers

1. Y.-C. Wu, P. L. Tobing, K. Kobayashi, T. Hayashi, and T. Toda, "Non-parallel voice conversion system with WaveNet vocoder and collapsed speech suppression," IEEE Access, vol. 8, pp. 62094–62106, Apr. 2020.

2. Y.-C. Wu, T. Hayashi, P. L. Tobing, K. Kobayashi, and T. Toda, "Quasi-Periodic WaveNet: an autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021. (Under review)

3. Y.-C. Wu, T. Hayashi, T. Okamoto, H. Kawai, and T. Toda, "Quasi-Periodic Parallel WaveGAN: a non-autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021. (Accepted)

4. H-T. Hwang, Y.-C. Wu, Y.-H. Peng, C.-C. Hsu, Y. Tsao, H.-M. Wang, Y.-R. Wang, and S.-H. Chen, "Voice conversion based on locally linear embedding," Journal of Information Science and Engineering, vol. 34, pp. 1469–1491, 2018.

5. H-T. Hwang, Y.-C. Wu, S.-S. Wang, C.-C. Hsu, Y. Tsao, H.-M. Wang, Y.-R. Wang, and S.-H. Chen, "Locally linear embedding based post-filtering for speech enhance-

ment," Journal of Information Science and Engineering, vol. 34, pp. 1493–1516, 2018.

6. P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Voice conversion eith cycleRNN-based spectral mapping and finely tuned WaveNet vocoder," IEEE Access, vol. 7, pp. 171114–171125, Apr. 2019.

7. X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K.A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. Le Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, Z.-H. Ling, "ASVspoof 2019: a large-scale public database of synthetic, converted and replayed speech," Computer Speech and Language, Vol. 64, Article 101114, 25 pages, Nov. 2020.

8. P. L. Tobing, Y.-C. Wu, K. Kobayashi, T. Hayashi, and T. Toda, "An evaluation of voice conversion with neural network spectral mapping models and WaveNet vocoder," APSIPA Transactions on Signal and Information Processing, vol. 9, e26, pp. 1-14, Nov. 2020.

9. W. -C. Huang and T. Hayashi and Y. -C. Wu and H. Kameoka and T. Toda, "Pretraining Techniques for Sequence-to-Sequence Voice Conversion," IEEE/ACM Transactions on Audio, Speech, and Language Processing, Jan. 2021.

International Conferences

1. Y.-C. Wu, H-T. Hwang, C.-C. Hsu, Y. Tsao, and H.-M. Wang, "Locally linear embedding for exemplar-based spectral conversion," Proc. INTERSPEECH, pp. 1652–165, Sept. 2016.

2. Y.-C. Wu, H.-T. Hwang, S.-S. Wang, C.-C. Hsu, Y.-H. Lai, Y. Tsao, and H.-M. Wang, "A locally linear embedding based postfiltering approach for speech enhancement," Proc. ICCASP, pp. 5555–5559, Mar. 2017.

3. Y.-C. Wu, H.-T. Hwang, S.-S. Wang, C.-C. Hsu, Y. Tsao, and H.-M. Wang, "A post-filtering approach based on locally linear embedding difference compensation for speech enhancement." Proc. INTERSPEECH, pp. 1953–1957, Aug. 2017.

4. Y.-C. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, "The NU non-parallel voice conversion system for the voice conversion challenge 2018," Proc. Speaker Odyssey, pp. 211–218, Les Sables d ' Olonne, France, Jun. 2018.

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

6. Y.-C. Wu, T. Hayashi, P. L. Tobing, K. Kobayashi, and T. Toda, "Quasi-Periodic WaveNet vocoder: a pitch dependent dilated convolution model for parametric speech generation," Proc. INTERSPEECH, pp. 196–200, Graz, Austria, Sep. 2019.

7. Y.-C. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, "Statistical voice conversion with quasi-periodic WaveNet vocoder," Proc. SSW10, pp. 63–68, Vienna, Austria, Sep. 2019.

8. Y.-C. Wu, T. Hayashi, T. Okamoto, H. Kawai, and T. Toda, "Quasi-Periodic Parallel WaveGAN vocoder: a non-autoregressive pitch-dependent dilated convolution model for parametric speech generation," Proc. INTERSPEECH, Full virtual, Oct. 2020.

9. Y.-C. Wu, P. L. Tobing, K. Yasuhara, N. Matsunaga, Y. Ohtani, and T. Toda, "A cyclical post-filtering approach to mismatch refinement of neural vocoder for text-to-speech systems," Proc. INTERSPEECH, Full virtual, Oct. 2020.

10. C.-C. Hsu, H-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Dictionary update for NMF-based voice conversion using an encoder-decoder network," Proc. ISCSLP, pp. 1–5, 2016.

11. 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," Proc. APSIPA, pp. 1–6, 2016.

12. C.-C. Hsu, H-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," Proc. INTERSPEECH, pp. 3364–3368, Aug. 2017.

13. Y.-H. Peng, C.-C. Hsu, Y.-C. Wu, H-T. Hwang, Y.-W. Liu, Y. Tsao, and H.-M. Wang, "Fast locally linear embedding algorithm for exemplar-based voice conversion," Proc. APSIPA, pp. 591–595, 2017.

14. 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, pp. 219–226, Les Sables d ' Olonne, France, Jun. 2018.

15. Y.-H. Peng, H-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Exemplar-based spectral detail compensation for voice conversion," Proc. INTERSPEECH, pp. 486–490, Hyderabad, India, Sep. 2018.

16. 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," Proc. SLT, pp. 297–303, Athens, Greece, Dec. 2018.

17. 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, pp. 6815–6819, Brighton, UK, May 2019.

18. W.-C. Huang, Y.-C. Wu, H.-T. Hwang, P. L. Tobing,T. Hayashi, K. Kobayashi, T. Toda, Y. Tsao, and H.-M. Wang, "Refined WaveNet vocoder for variational autoencoder based voice conversion," Proc. EUSIPCO, pp. 1–5, A Coruna, Spain, Sep. 2019.

19. W.-C. Huang, Y.-C. Wu, C.-C. Lo , P. L. Tobing,T. Hayashi, K. Kobayashi, T. Toda, Y. Tsao, and H.-M. Wang, "Investigation of F0 conditioning and fully convolutional networks in variational autoencoder based voice conversion," Proc. INTERSPEECH, pp. 709–713, Graz, Austria, Sep. 2019.

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

21. W.-C. Huang, Y.-C. Wu, K. Kobayashi, Y.-H. Peng, H.-T. Hwang, P.L. Tobing, Y. Tsao, H.-M. Wang, and T. Toda, "Generalization of spectrum differential based di-

rect waveform modification for voice conversion," Proc. SSW10, pp. 57–62, Vienna, Austria, Sep. 2019.

22. P.L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Efficient shallow WaveNet vocoder using multiple samples output based on Laplacian distribution and linear prediction," Proc. ICASSP, Full virtual, pp. 7204–7208, May 2020.

23. P.L. Tobing, T. Hayashi, Y.-C. Wu, K. Kobayashi, and T. Toda, "Cyclic spectral modeling for unsupervised unit discovery into voice conversion with excitation and waveform modeling," Proc. INTERSPEECH, Full virtual, Oct. 2020.

24. W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," Proc. INTERSPEECH, Full virtual, Oct. 2020.

25. P.L. Tobing, Y.-C. Wu, and T. Toda, "Baseline system of voice conversion challenge 2020 with cyclic variational autoencoder and parallel WaveGAN," Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020, Full virtual, Oct. 2020.

26. W.-C. Huang, P.L. Tobing, Y.-C. Wu, K. Kobayashi, and T. Toda, "The NU voice conversion system for the voice conversion challenge 2020: on the effectiveness of sequence-to-sequence models and autoregressive neural vocoders," Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020, Full virtual, Oct. 2020.

Domestic Conferences

1. Y.-C. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, “Development of NU non-parallel voice conversion system for voice conversion challenge 2018,” 日本音響学会講演論文集, 1-9-5, pp. 217-218, Mar. 2018.

2. Y.-C. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, “Development of NU non-parallel voice conversion system 2018,” 電子情報通信学会技術研究報告, vol. 117, no. 517, SP2017-155, pp. 385-390, Mar. 2018.

3. P.L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, ”Development of NU voice conversion system for Voice Conversion Challenge 2018,” 日本音響学会講演論文集, 1-9-4, pp. 215-216, Mar. 2018.

4. 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, SP2017-121, pp. 203-208, Mar. 2018.

5. W.-C. Huang, Y.-C. Wu, H.-T. Hwang, P.L. Tobing, T. Hayashi, K. Kobayashi, T. Toda, Y. Tsao, and H.-M. Wang, “Reducing mismatch of WaveNet vocoder for variational autoencoder based voice conversion,” 日本音響学会講演論文集, 3-5-14, pp. 1317-1318, Mar. 2019.

6. 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, 3-5-15, pp. 1319-1320, Mar. 2019.

7. K. Yasuhara, Y.-C. Wu, P. L. Tobing, N. Matsunaga, Y. Ohtani, and T. Toda, “テキスト音声合成におけるポストフィルタとしてのWaveNetボコーダ学習法,” 日本音響学会講演論文集, 1-2-5, pp. 1051-1052, Mar. 2020.

Awards

1. NEC C&C 2019 年度外国人研究員助成事業

2. INTERSPEECH 2019 Travel Grant