Deep Speech Analysis-Modification-Synthesis based on Quasi-Harmonic Modeling

Shaowen Chen

Graduate School of Informatics

Nagoya University

Abstract

Speech signal modeling, a fundamental component of modern speech processing, has attracted considerable attention due to its critical role in representing speech signals as intermediate acoustic features and reconstructing speech signals from these features, which enables a wide range of applications, including speech synthesis, voice conversion, and speech compression. Vocoders are considered as fundamental tools for achieving such encoding and decoding processes, which have been broadly classified into two categories: conventional vocoders grounded in conventional signal processing (CSP) techniques and neural vocoders leveraging the modeling capabilities of deep learning.

As a typical representative of conventional vocoders, Quasi-Harmonic Modeling (QHM) methods, including QHM, adaptive QHM (aQHM), and extended aQHM (eaQHM), are powerful methods to model the speech as sparse sinusoidal components by extracting the three fundamental elements of each component, i.e., amplitude, phase, and frequency, in a frame-by-frame scheme. Even with errors in initial frequency estimates from fundamental frequency extractors, the frequency correction mechanism in QHM is capable of mitigating such frequency errors. The extracted framewise parameters (amplitude, phase, and frequency) of the sparse components are interpretable and controllable, making QHM methods widely used for speech resynthesis and modification in terms of pitch and duration. Unfortunately, QHM methods struggle to extract the parameters accurately; thus, the quality of generated speech is limited. Additionally, the iterations in aQHM and eaQHM destroy the efficiency.

On the other hand, the recent advances in deep learning have ushered in the era of neural vocoders. With their powerful fitting capabilities, they can accurately generate speech waveforms. Furthermore, with proper design, neural networks often require only simple matrix operations, making neural vocoders highly efficient. However, they are inherently black-box models, lacking interpretability. As a result, they do not provide insights into speech structure or vocalization principles, and are unable to facilitate speech modification.

Both conventional vocoders and neural vocoders exhibit distinct advantages and limitations. These characteristics stand in contrast to, yet also complement, the QHM methods, which are inherently interpretable and capable of providing valuable insights into the underlying mechanisms of speech. This observation motivates the integration of the aforementioned approaches (conventional vocoders and neural vocoders), with the objective of retaining their respective strengths while mitigating their inherent weaknesses. Accordingly, this thesis proposes a novel hybrid framework designed to achieve efficient speech modeling, characterized by high quality, computational efficiency, and flexibility in speech analysis, modification, and synthesis.

First, we review the theory of QHM and some representatives of neural vocoders. To overcome that the frequency correction of QHM methods is limited, we propose a spectrogram-based frequency correction method, which relies on the spectrogram of the speech, instead of estimated parameters like

QHM methods. Inspired by the powerful gradient descent method for deep learning, we propose a backpropagation-based QHM (BP-QHM), which uses the gradient descent to obtain the required parameters by directly minimizing the reconstruction waveform error of the entire speech. Additionally, we propose a novel spectrogram loss to increase the convexity and accelerate the convergence during the optimization. The experimental evaluations being conducted to investigate the effectiveness show that our method achieves an improvement in speech resynthesis quality and frequency correction. The successful use of backpropagation in the QHM framework also implies the potential of combining the QHM framework and deep learning.

Second, although BP-QHM enhances the resynthesis quality, its iterative nature remains time-consuming. Inspired by the success of BP in propagating the loss back through the QHM synthesis process, we are motivated to integrate the neural network and QHM framework to propose a novel framework for the neural vocoder, leading to the development of QHM-GAN. QHM-GAN integrates the interpretability of CSP and the high quality and robustness of neural networks. However, the resonance characteristic modeling of QHM-GAN is limited. Thus, based on this framework, speech signals can be further encoded into autoregressive moving average (ARMA) functions to model the resonance characteristics, which gives birth to QHARMA-GAN, enabling accurate amplitude and phase estimation at arbitrary frequencies. This allows for high-quality synthesis and flexible speech modifications in terms of pitch shifting and time stretching, while also reducing time consumption and network size. Experimental evaluations indicate that the proposed method leverages the strengths of QHM, the ARMA modeling, and neural networks, outperforming existing methods in generation speed, synthesis quality, and modification flexibility.

Eventually, the specific speech modification algorithms for both the conventional QHM-based methods and the proposed neural-based methods are described in detail. The fundamental concepts of time-scale and pitch-scale modification are introduced, and the core idea of shape-invariant modification is consistently extended across all methods. Each method adopts a different strategy to preserve the relative phase, thereby maintaining a consistent waveform shape during modification. As a result, the speech can be modified while preserving a spectral envelope sequence, which is essential for ensuring the perceptual quality of the modified output. Experimental evaluations indicate that the proposed method QHARMA-GAN is powerful in resonance characteristics modeling, leading to a higher performance in pitch-scale speech modification than QHM methods, particularly from the perspectives of generation efficiency and the quality of the modified speech.

Overall, in this thesis, the QHM framework and neural network were successfully integrated to propose a hybrid speech analysis-synthesis-modification system, which can extract the acoustic parameters from the speech signals with a high accuracy, synthesize the speech from the extracted parameters with a high quality, and modify the speech with a high controllability.

Contents

List of Figures

Chapter 1

Introduction

Speech is the most natural and essential way of communication, including human-human communication and machine-human communication. Consequently, speech signal processing has emerged as one of the most dynamic and influential fields within signal processing. Over the past few decades, research in speech processing has driven the development of numerous technologies that benefit human society. For instance, the advances in speech recognition has enabled machines to comprehend not only human speech but also human language, greatly enhancing machine-human interaction, while speech compression has significantly improved the efficiency and effectiveness of telecommunication and storage of the speech. Speech modeling is also one of the aspects of speech processing, helping human to understand the structure of speech and human vocal mechanisms. These advances in speech modeling have further promoted the development of various fields, such as doctors diagnosing by analyzing voices and providing voice conversion for vocal patients. The applications of speech processing are extensive and continuously expanding. With the increasing development of computing, communication, and the Internet, the role of speech processing is expected to grow even more prominent in the future.

1.1 Formulation of Speech Signals

The speech signals are usually considered as the nonstationary signals, consisting of voiced parts and unvoiced parts. The voiced part of speech is often referred to as the deterministic part, as it typically exhibits a periodic structure that can be effectively characterized through spectral analysis. In contrast, the unvoiced part is considered stochastic in nature, as it resembles random noise both perceptually and acoustically. The combination of these two parts forms the complete human speech signal, which can be represented as follows:

$$x(t) = \sum_{k=0}^{K} x_k(t) + \varepsilon(t) = \sum_{k=0}^{K} A_k(t) e^{i\varphi_k(t)} + \varepsilon(t), \tag{1.1}$$

where K is the number of components, $A_k(t)$ and $\varphi_k(t)$ denote the instantaneous amplitude and instantaneous phase of the k-th component at instantaneous time t (t is continuous), and $\varepsilon(t)$ is stochastic noise. This equation implies that the deterministic voiced speech can be further decomposed into several sinusoidal components. To gain a clearer understanding of the composition and intrinsic nature of speech signals, researchers have continuously explored various methods and models for parameter extraction, such as the parameters of sinusoidal components or the statistical properties of noise distributions, to represent and characterize the speech signals.

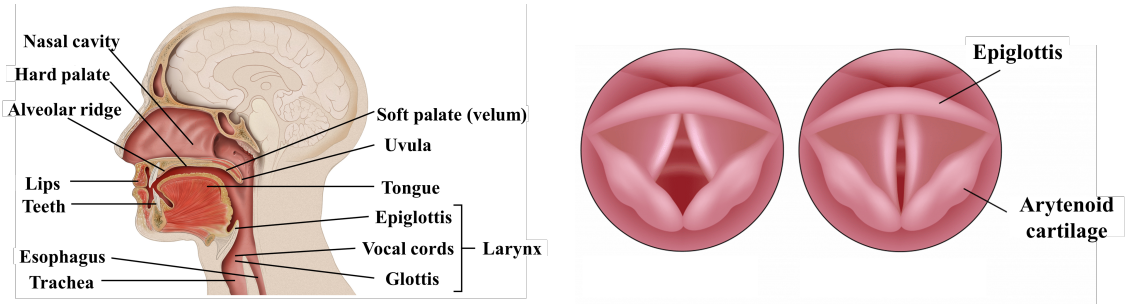
1.2 Speech Modeling

As the foundation of modern speech processing, speech modeling has driven the development of related fields such as speech enhancement and speech compression. It enables a better understanding of speech by visualizing its structure with acoustic features, and allows effective manipulation by reconstructing signals from these features. Since the structure of raw speech waveforms is often difficult to interpret directly, speech modeling extracts or transforms them into acoustic features that are more interpretable and physically meaningful. These features provide insights into the nature of speech signals. By leveraging such interpretable acoustic features together with the objectives of other speech processing tasks, related problems can be addressed from a physical perspective. For example, the essence of voice conversion lies in modifying the fundamental frequency (also referred to as f_0) or timbre of speech. Speech modeling allows for the extraction of acoustic features such as loudness and frequency, enabling voice conversion to be achieved through the manipulation of these physically meaningful parameters.

The tool for completing speech modeling is called a vocoder, which is responsible for the tasks of acoustic feature extraction and speech reconstruction, i.e., encoding and decoding. With the development of the computer, vocoders have consequently seen substantial advancements over the years. Throughout their development, various types of vocoders have emerged, all evolving toward several overarching objectives:

- Accurately extracting acoustic features, such as the three essential elements of the signal, i.e., amplitude, frequency, and phase;
- Reconstructing speech with the highest possible quality;
- Performing analysis and synthesis with increasing speed to enable real-time processing;
- Enhancing the controllability of acoustic features to better support downstream tasks.

These vocoders can be categorized into different types based on their characteristics. For example, according to whether the model adheres to the speech production mechanism, vocoders can be classified into models, such as the source-filter model and the sinusoidal model. Alternatively, based on the algorithmic architecture employed, vocoders can be divided into conventional



(a) Human vocal tract

(b) Human vocal cords

Figure 1.1: Human speech production mechanism [1].

vocoders and neural vocoders. In the following sections, we provide a detailed introduction to these vocoder types.

1.2.1 Source-Filter Model versus Sinusoidal Model

To model speech signals, the most intuitive approach is to adopt a mathematical perspective, i.e., representing speech using formal mathematical models. One of the most representative examples is the sinusoidal model, which characterizes speech as a sum of sinusoidal components. This modeling framework facilitates a deeper understanding of the mathematical essence of speech by enabling the extraction of physically interpretable parameters, such as amplitude, frequency, and initial phase. These parameters not only offer insights into the nature of speech signals but also allow for flexible artificial control, namely synthesis and modification.

On the other hand, unlike many other types of signals, speech signal exhibits inherent structural characteristics. For instance, the energy distribution and relative phases of each component somewhat follow some specific regularities, which are primarily determined by the configuration of the speaker's vocal tract, which can be considered as a filter. To uncover these underlying regularities, researchers have investigated the human speech production mechanism from medical and physical perspectives and have attempted to formulate mathematical models that reflect the entire speech production mechanism. Source-filter models are built based on such an uncovered speech production mechanism. Before introducing these two modeling methods in detail, we first briefly review the human speech production mechanism.

Figs 1.1(a) and 1.1(b) illustrate the principal anatomical parts involved in speech generation, which are conventionally categorized into three major subsystems: the lungs, the larynx, and the vocal tract.

The speech production process initiates in the lungs, which serve as the primary energy source by generating an airflow that travels upward through the trachea. Upon reaching the glottis, this airflow is modulated, as depicted in Fig. 1.1(b). Depending on the state of the vocal cord, this airflow can be modulated as two types of source signals, namely a quasi-periodic signal, which is produced by the vibration of the vocal cord, as shown in the right part of Fig. 1.1(b), or a tur-

bulence excitation signal, which is generated from the abducted vocal cord. These source signals then propagate into the vocal tract, which plays a vital role in shaping the speech characteristics and determining the type of phonation: the quasi-periodic signals will evolve into voiced speech, while turbulence excitation signals will evolve into unvoiced speech. The vocal tract comprises three interconnected resonating cavities: the pharyngeal, oral, and nasal cavities. Within this tract, the initial source is spectrally filtered and transformed into the distinct timbral and articulatory human speech. Finally, the acoustically shaped signal is radiated from the lips into the external environment.

For voiced speech production, the larynx parts can be considered as the excitation source, with the vocal folds (or vocal cords) playing a central role. The process begins when the vocal folds periodically adduct and abduct to yield a train of glottal airflow pulses, sustaining a quasi-periodic excitation. The temporal frequency of these pulses defines the f_0 of the voiced speech signal, which directly influences the perceived pitch. Dynamic variations in f_0 over time convey critical prosodic information, distinguishing between different linguistic forms (e.g., questions or statements) as well as reflecting the emotional status of the speaker. Acoustically, the vocal tract acts as a time-varying filter that selectively amplifies or attenuates each component of the source signals. Frequencies exhibiting significant energy concentrations are referred to as formants, whereas those with substantial energy suppression are termed anti-formants. The interaction between formants and anti-formants plays a crucial role in defining the spectral envelope of the signal, and thereby determines the perceived timbre of the sound.

In the case of unvoiced speech, no periodic excitation is generated at the glottis. Instead, the turbulent airflow is produced at constrictions within the vocal tract, such as the teeth, lips, or tongue, and subsequently shaped by the surrounding cavities to form voiceless consonants, such as /s/, /f/, /h/, /p/, /t/, and /k/.

Source-Filter Model

As the name suggests, the source-filter theory of speech production, initially proposed by Dudley [8] in the context of vocoder design and later formalized and extended by Fant [9], posits that speech signals can be decomposed and modeled mathematically as a combination of a source excitation signal and a filter. The excitation signal mainly represents the glottal source generated by the vocal folds, while the filter mainly represents the vocal tract. The speech waveform can be conceived as the result of the source excitation signal being shaped by the time-varying filter whose characteristics are determined by the configuration of the vocal tract, which can be formulated as

$$x(t) = u(t) * h(t), \tag{1.2}$$

where $*$ denotes the convolution operator. $x(t)$ denotes the observed speech waveform while $u(t)$ and $h(t)$ represents the source excitation signal and filter response at t instant in time domain, respectively. This interaction produces a spectral envelope exhibiting broadband energy peaks, forming what is widely known as the source-filter model of speech. Assuming that the filter is time-invariant, i.e., considering the interaction within a short term, such an interaction can be spectrally derived from the Fourier transform of Eq. (1.2):

$$X(\omega) = U(\omega)H(\omega), \tag{1.3}$$

where $X(\omega)$, $U(\omega)$, and $H(\omega)$ denote the Fourier transform of $x(t)$, $u(t)$, and $h(t)$, respectively. This equation indicates that the filter produces a spectral envelope exhibiting broadband energy peaks to shape the source excitation signals envelope, forming what is widely known as the source-filter model of speech.

For voiced speech, the excitation source is typically modeled as a sequence of glottal pulses representing the glottal volume velocity, where the temporal spacing of the pulses defines the output f_0. Thus, the Liljencrants-Fant (LF) model was proposed to model the generation of the glottal pulses [10]. Although the LF model was capable of modeling the source excitation signal from a fluid dynamics perspective of glottal airflow, it needs too many parameter to represent the signal in a simple way. Thus, many studies were gradually conducted to more specifically explore the source excitation signals and simplify them, such as the improved LF model [11] and the reshaped LF model [12]. In contrast, unvoiced speech lacks periodic glottal pulses, and is instead represented using zero-mean white noise to reflect its stochastic nature. Therefore, the source excitation signal can be similarly formulated as

$$u(t) = \sum_{k=0}^{K} u_k(t) + \varepsilon_u(t) = \sum_{k=0}^{K} A_k^u(t) e^{i\varphi_k^u(t)} + \varepsilon_u(t), \tag{1.4}$$

where $A_k^u(t)$ and $\varphi_k^u(t)$ denote the amplitude and phase of the k-th excitation component and $\varepsilon_u(t)$ is the noise in the source excitation signal. Typical examples of source-filter models include STRAIGHT and WORLD, both of which model the source and filter components using different approaches. In the following sections, we provide a detailed introduction to each of these models.

STRAIGHT [13, 14] is a prominent example of vocoders grounded in the source–filter model paradigm, in which speech is modeled as a spectral envelope (serving as the filter) derived from short-time Fourier transform (STFT), and excited by a spectrally flat source signal constructed in the frequency domain.

The main contribution of STRAIGHT lies in its two-stage spectral envelope analysis technique, designed to suppress periodicity-induced artifacts both in the time and frequency domains. In the first stage, STRAIGHT applies two pitch-synchronous complementary analysis windows, $w_p(t)$

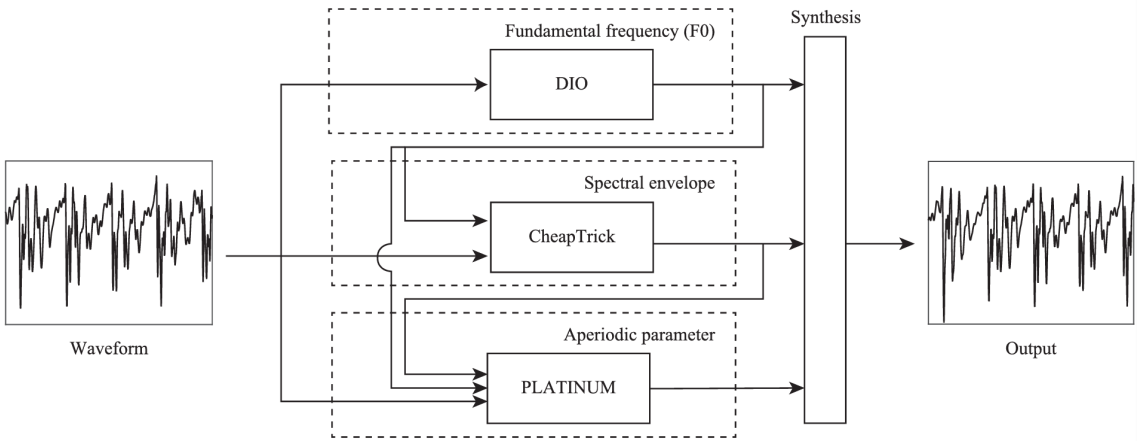


Figure 1.2: Structure of WORLD vocoder in [2].

and $w_c(t)$, to extract two magnitude spectra, $S_p(t)$ and $S_c(t)$, respectively:

$$w_p(t) = e^{(t-t_0)^2} h(t-t_0), \quad w_c(t) = w_p(t) \cdot \sin\left(\frac{\pi t}{t_0}\right) \tag{1.5}$$

where t denotes the time index, $t_0 = 1/f_0$ is the pitch period, and $h(t)$ is the second-order cardinal B-spline function:

$$h(t) = \begin{cases} 1-|t|, & \text{if } |t| < 1 \\ 0, & \text{otherwise} \end{cases} \tag{1.6}$$

These complementary spectra are then combined to form a smoothed spectrogram:

$$S_U(t) = \sqrt{S_p^2(t) + \xi S_c^2(t)} \tag{1.7}$$

where $\xi = 0.13655$ is an empirically determined blending factor to minimize temporal fluctuation. This stage effectively smooths the spectrogram along the time axis, reducing pitch-synchronous artifacts. In the second stage, STRAIGHT applies a frequency-domain smoothing process to S_U, using a smoothing window whose shape is determined by the second-order B-spline and whose bandwidth is proportional to f_0. This operation suppresses artifacts caused by the pitch-dependent frequency resolution, resulting in a cleaner and more stable spectral envelope. With the estimated f_0, smoothed spectral envelope, and aperiodicity spectrogram, STRAIGHT can resynthesize high-quality speech and flexibly modify pitch. However, due to its multi-stage processing and high computational cost, STRAIGHT is not well-suited for real-time applications.

To improve the STRAIGHT vocoder, the WORLD vocoder [2] was proposed with improvements in terms of effectiveness, implementability, and robustness. The structure of the WORLD vocoder is illustrated in Fig. 1.2, showing that the WORLD consists of a pitch detector (DIO) [15], a spectral envelope extractor (CheapTrick) [16], and an aperiodic parameter estimator (PLATINUM) [17]. Sometimes, the pitch detector was further replaced by Harvest [18] to obtain a

better result of pitch and then further processing. The workflow of the WORLD vocoder is given below. First, the pitch will be extracted by DIO or Harvest, as

$$f_0 = P_{\text{DIO}}(x) \text{ or } f_0 = P_{\text{Harvest}}(x). \tag{1.8}$$

where P_{DIO} and P_{Harvest} are the functions of DIO and Harvest, respectively. The extracted f_0 is subsequently utilized, together with the original speech waveform x, as input to the *CheapTrick* algorithm in order to estimate the time-varying spectral envelope on a frame-by-frame basis. This process is formally described as:

$$S_{\text{spec}} = P_{\text{CheapTrick}}(x, f_0), \tag{1.9}$$

where S_{spec} denotes the resulting spectral envelope and $P_{\text{CheapTrick}}$ is the function of CheapTrick. This envelope is then employed as an additional input to the *PLATINUM* module, which estimates the aperiodicity parameter. The aperiodicity parameter characterizes the relative bandwidths of the periodic and aperiodic components across the full frequency range. The computation is expressed as:

$$P_{\text{aperiodic}} = P_{\text{PLATINUM}}(x, f_0, S_{\text{spec}}). \tag{1.10}$$

where P_{PLATINUM} is the function of PLATINUM. After obtaining the f_0, spectral envelope S_{spec}, and aperiodicity parameter $P_{\text{aperiodic}}$, the speech can be reconstructed with an overlap-add (OLA) method [19]. Additionally, this OLA method is based on f_0, also called pitch; thus, this method is referred to as pitch-synchronous OLA (PSOLA) [20]. The linear interpolation will be used in the spectral envelope, while the f_0 will determine the positions of source impulses in the time domain. Then, the spectral envelope at the source impulse positions will be determined to obtain the corresponding impulse response, which is similar to the speech waveform. One of the key advantages of this approach is that it allows for the manual modification of f_0; the corresponding source impulse locations and their associated spectral envelopes can then be determined accordingly, enabling the synthesis of pitch-modified speech, namely, prosodically altered speech. Additionally, the low computation of the WORLD vocoder enables it to be applied in real-time processing, thus, it became a well-known and widely used vocoder.

In addition to STRAIGHT and WORLD, numerous other source-filter vocoders have been developed. While we do not provide an exhaustive overview here, these vocoders share a common objective: to accurately model both the vocal tract transfer function and the glottal excitation signal. This precise modeling is crucial for achieving high-quality speech synthesis as well as greater flexibility in speech modification tasks.

Sinusoidal Model Family

In contrast to the source-filter model, the sinusoidal models do not aim to explicitly model the vocal tract; instead, they fundamentally process speech from the perspective of signal representation. As Eq. (1.1) shows, the speech signal can be represented as the combination of several sinewaves and noise. Consequently, the essence of both speech generation and modification lies in the extraction, manipulation, and reconstruction of sinusoidal and noise parameters. As such, parameter extraction becomes a critical issue in the overall process.

Before introducing the sinusoidal model, we first review the parameter extraction. Parameter extraction techniques are broadly classified into two categories: nonparametric and parametric methods. Nonparametric approaches, such as the short-time Fourier transform (STFT) [21] and the Wigner–Ville distribution [22], are based on time–frequency (TF) analysis, where the signal is transformed from the time domain into the TF domain to reveal its internal structure. These methods provide an intuitive visualization of how signal energy evolves over time and frequency, and parameters such as instantaneous frequency can typically be inferred by detecting ridges or local maxima within the TF representation. Nevertheless, these methods are fundamentally constrained by the time–frequency uncertainty principle, which states that it is impossible to achieve arbitrarily high resolution simultaneously in both time and frequency domains. This trade-off leads to blurred TF representations, thereby degrading the accuracy of ridge detection and, consequently, the precision of parameter estimation.

To mitigate the resolution limitations inherent in conventional TF methods, various postprocessing techniques have been proposed, including the reassignment method [23],[24], and synchrosqueezing transforms [25], [26]. These techniques aim to enhance the concentration of energy in the TF representation by reallocating energy to more appropriate locations, thereby yielding sharper and more interpretable TF structures. While such enhancements can significantly improve visual clarity and reduce spectral smearing, they remain limited in that they do not involve an explicit demodulation process. As a result, their ability to isolate and extract modulated signal components remains suboptimal, particularly in the presence of noise or overlapping components.

In contrast, parametric methods adopt a model-based framework, where the signal is assumed to adhere to a specific analytical form characterized by a set of parameters. Representative examples include the chirplet transform [27], the matching demodulation transform [28], and sinusoidal modeling (SM) [29]. These methods generally require an initial estimation of certain key parameters, such as instantaneous frequency, chirp rate, or amplitude envelope, which serve as priors to guide the subsequent analysis. By incorporating such prior knowledge, parametric approaches can more accurately adapt to the signal's structure, thereby enabling more precise and robust parameter extraction. Although this increased accuracy often comes at the expense of higher computational complexity and a dependency on reliable prior estimates, parametric methods offer considerable advantages in scenarios requiring high-resolution analysis or in conditions

where nonparametric methods are rendered ineffective.

Among the various parametric approaches, sinusoidal modeling (SM) has garnered considerable attention and has been widely adopted in speech analysis and synthesis tasks. The core idea of SM is to represent the speech signal as a sum of sinusoidal components, each characterized by its framewise amplitude, frequency, and phase, as formulated as

$$x(t) = \left(\sum_{k=-K}^{K} a_k e^{i2\pi f_k t} \right) w(t),\ t \in [-T_l, T_l], \tag{1.11}$$

where a_k and f_k denote the complex amplitude and frequency of the k-th component, respectively. $w(\cdot)$ denotes the moving window whose length is $2T_l$. Note that k and $-k$ correspond to symmetric frequency components, which are complex conjugates of each other, ensuring that $x(t)$ is a real-valued signal. These parameters are typically estimated during the analysis stage on a frame-by-frame basis. Subsequently, in the synthesis stage, their instantaneous counterparts are derived through interpolation across frames, which are then used to reconstruct the time-domain speech signal. This modeling framework offers a compact and interpretable representation of voiced speech components, making it particularly suitable for applications such as speech coding, transformation, and synthesis.

However, a critical limitation of SM lies in its dependency on the short-time Fourier transform (STFT) for parameter extraction. Due to the limited time–frequency resolution of STFT and the inherent trade-offs imposed by the uncertainty principle, the resulting time–frequency representation is often blurred. Consequently, the detected amplitude and frequency trajectories may deviate from their true values, leading to inaccuracies in the subsequent modeling process. These inaccuracies inevitably propagate to the synthesis stage, thereby degrading the perceptual quality of the reconstructed speech. Moreover, the phase component is also affected by the imprecise extraction, further contributing to synthesis errors.

Another notable disadvantage of conventional SM arises when modeling unvoiced speech segments. Unlike voiced sounds, unvoiced speech is inherently stochastic, characterized by energy distributed irregularly across the spectrum. Since SM is fundamentally designed to model periodic structures via deterministic sinusoids, it fails to capture the aperiodic and broadband nature of unvoiced sounds. As a result, the synthesized unvoiced segments tend to sound unnatural or overly smoothed, lacking the richness and variation of natural speech.

To overcome these deficiencies, the harmonic plus noise model (HNM) was proposed [30],[31], offering a more comprehensive modeling framework. HNM decomposes the speech signal into two distinct components: deterministic harmonic components for the voiced segments, and stochastic noise components for the unvoiced or aperiodic segments. The noise component is typically modeled using an all-pole filter driven by white noise, enabling it to better capture the random spectral characteristics of unvoiced sounds. By combining the harmonic component with a noise model, HNM achieves greater flexibility and can produce more natural-sounding synthetic speech

across both voiced and unvoiced segments [32]. This dual-component structure also allows for more effective speech modification and transformation, such as pitch shifting and time stretching, while maintaining high quality.

Nevertheless, HNM introduces its own set of challenges. A major issue is its reliance on the accurate estimation of f_0. Since the harmonic component (denoted as $x^v_{\text{HNM}}(t)$) is constructed based on f_0 and its multiples, as formulated as

$$x^v_{\text{HNM}}(t) = \left(\sum_{k=-K}^{K} a_k e^{i2\pi k f_0 t} \right) w(t),\ t \in [-T_l, T_l]\,, \tag{1.12}$$

any errors in f_0 estimation can lead to significant mismatches in the harmonic structure, thereby compromising the quality of the synthesized speech. To mitigate this problem, various f_0 refinement strategies have been proposed and are often employed prior to synthesis [33], aiming to correct initial estimation errors and stabilize the harmonic modeling process.

Despite these improvements, a fundamental limitation remains: both SM and HNM assume a purely harmonic structure for voiced speech. In practice, however, natural speech is only approximately periodic. Variations in vocal fold vibration, articulation dynamics, and coarticulation effects introduce deviations from ideal harmonicity. As a result, modeling voiced speech using a fixed set of harmonics still leads to notable resynthesis errors, particularly in expressive or emotionally colored speech. These challenges highlight the need for more flexible and adaptive modeling techniques capable of capturing the full complexity of natural speech production.

To address the aforementioned limitations inherent in both SM and HNM, the quasi-harmonic model (QHM) was proposed by Pantazis et al. [34]. QHM introduces a more flexible modeling framework by leveraging quasi-harmonic components whose frequencies are not strictly constrained to be integer multiples of a single f_0. This is achieved through the incorporation of a frequency correction mechanism, allowing the model to treat each component's frequency as an independent parameter rather than being solely tied to f_0. As a result, QHM is capable of jointly modeling both the harmonic (voiced) and inharmonic (unvoiced or transitional) segments of speech, thus providing a unified representation for a broader class of signals.

Despite this flexibility, a key assumption in QHM limits its effectiveness: the model assumes that the signal is locally stationary within each analysis frame. In practice, however, speech signals, particularly in expressive speech or singing voice, often exhibit rapid variations in frequency and amplitude, even within short time intervals. This stationarity assumption can lead to inaccurate parameter estimation, especially when modeling fine-grained dynamics of speech or rapidly modulated components.

To overcome this limitation, the adaptive quasi-harmonic model (aQHM) was introduced [35], extending the QHM framework by replacing its fixed-phase exponential formulation with a non-stationary phase function. This modification allows aQHM to more accurately track time-varying instantaneous frequencies, thereby capturing the dynamic nature of speech signals across consec-

utive frames. By modeling the phase evolution in a flexible manner, aQHM significantly improves the fidelity of frequency estimation and enhances the overall modeling capability.

Nevertheless, while aQHM addresses the limitations related to nonstationary frequency, it still employs a relatively simplistic amplitude model, typically assuming linear amplitude modulation within each frame. This assumption becomes inadequate in scenarios such as singing voice modeling, where the amplitude envelope can exhibit nonlinear behavior, often governed by cubic or higher-order modulations. To address this, the extended adaptive quasi-harmonic model (eaQHM) was proposed [36], which further augments aQHM by introducing an adaptive amplitude modulator. This additional component enables the model to capture complex, nonlinear amplitude variations, thereby facilitating more accurate extraction of both frequency and complex amplitude trajectories.

The enhanced modeling capabilities of eaQHM make it a powerful tool not only for speech analysis but also for a wide range of speech processing applications. Notably, it enables high-quality speech synthesis [37],[38] by accurately reconstructing both voiced and unvoiced components with accurate time-varying parameters. Furthermore, it supports advanced speech modification techniques, such as time-scale modification [39] and pitch-shifting [40], by providing a detailed and manipulable representation of the speech signal. Taken together, QHM and its adaptive extensions offer a principled and flexible framework that bridges the gap between conventional modeling methods and the rich variability of natural speech. Finally, we summarize the evolution of the sinusoidal model family in Fig. 1.3, where the regular (upright) text denotes the analysis methods, and the italicized text indicates the specific limitations addressed by each newly proposed method.

1.2.2 Conventional Vocoder versus Neural Vocoder

Vocoders can be categorized based on their modeling objectives into source-filter models and sinusoidal models. Similarly, from the perspective of methodological approaches employed during the modeling process, vocoders can be broadly classified into two main categories: conventional vocoders, which are grounded in conventional signal processing (CSP) techniques, and neural vocoders, which leverage the representational power of deep learning frameworks. In the following parts, we provide a detailed introduction to each of these two categories of vocoders.

Conventional Vocoder

Conventional vocoders typically employ CSP algorithms for feature or parameter extraction, which often results in faster computational performance. Representative examples include spectrogram-based approaches, such as those relying on STFT and SST, source-filter vocoders, such as the STRAIGHT and the WORLD, as well as sinusoidal vocoders such as HNM and QHM-based methods.

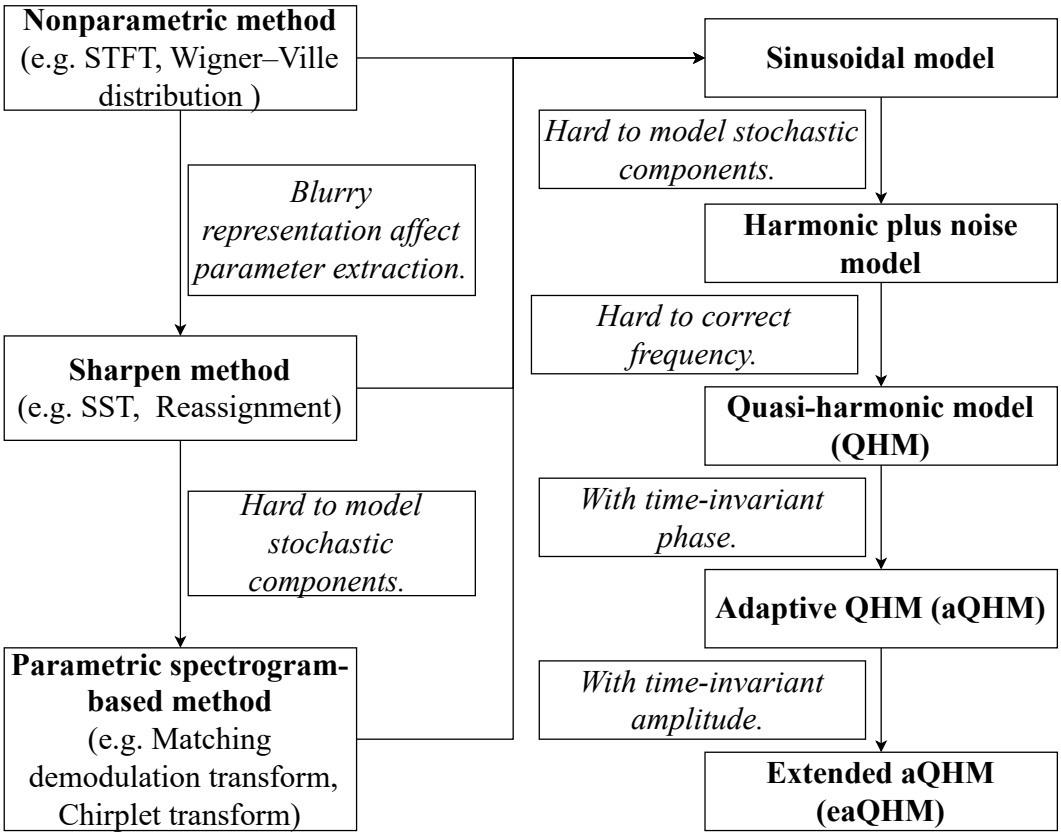


Figure 1.3: Development of the sinusoidal model family

Spectrogram-based methods offer the most straightforward and efficient means of exploring the structure of speech signals. However, as previously discussed, the uncertainty principle results in blurred TF representations when using STFT or wavelet transform. Even post-processing techniques such as the SST, which aim to sharpen spectrograms, are limited in their ability to produce more readable TF representations, making accurate parameter extraction challenging.

To address this, numerous studies have focused on enhancing the TF concentration of spectrograms. For instance, Oberlin et al. [41] proposed the second-order synchrosqueezing transform (2nd-SST), based on the STFT and extendable to the wavelet domain, which is designed for multicomponent signals. This method includes both the vertical synchrosqueezing transform (VSST) and the oblique synchrosqueezing transform (OSST), significantly improving time–frequency concentration while maintaining the reconstruction capability of conventional time–frequency reassignment methods. Based on Oberlin's work, Pham et al. [42] introduced the high-order synchrosqueezing transform (High-order SST), which elevates instantaneous frequency estimation from second-order to higher-order, enabling the characterization of multicomponent signals with higher-order frequency modulations. This not only enhances time–frequency concentration but also improves reconstruction accuracy of individual components. Furthermore, a synchroextracting transform [43] was proposed, which uses the instantaneous frequency estimator to locate the ridges of signal components and only extracts the parameters at the ridges, significantly sharp-

ening the spectrogram for better ridge detection and parameter extraction.

Although the constantly emerging studies have improved the TF concentration a lot, such concentrated TF representation can not be straightway modified to obtain the modified speech from the perspectives of pitch-shifting and time-stretching. Thus, the parameter can be preliminarily extracted and then further improved by parametric methods. Among them, the sinusoidal model family was more widely used because of directly extracts the amplitudes, frequencies, and phases of each signal component. Subsequently, in the sinusoidal model family, QHM and HNM have become two major branches, and many researchers have successively studied how to extract parameters and reconstruct speech signals more accurately [30], [44], [45], [34], [35], [36].

As the parameter estimation accuracy improves, a corresponding increase in computational complexity becomes inevitable. For instance, models such as the HNM rely on iterative procedures for refining f_0 to ensure convergence toward perceptually natural results. More advanced QHM variants, including aQHM and eaQHM, further intensify this computational burden by employing multi-stage iterative optimization algorithms, which aim to progressively capture the intricate amplitude and frequency modulations present in natural speech signals. These iterative processes are essential for achieving high-fidelity modeling and resynthesis, particularly in complex acoustic situations, including rapid pitch transitions, vibrato, or breathy voice. However, the computational cost associated with such high-resolution modeling significantly limits the practical applicability of QHM-based methods, especially in resource-constrained environments such as embedded systems, mobile devices, or real-time communication platforms. This gives rise to a trade-off between computational efficiency and synthesis quality: while more accurate modeling leads to perceptually improved output, it also demands more processing time and hardware resources.

Neural Vocoder

With the rapid advancement of neural networks (DNNs), these technologies have also been widely applied to speech modeling, which has significantly accelerated the development of neural vocoders, leveraging their strong nonlinear modeling capabilities and hierarchical architectures to learn complex mappings between input acoustic features and output waveforms. DNNs utilize complex fitting functions to mimic the learning mechanisms of the human brain, and thus typically require large amounts of training data to achieve optimal performance. Selecting an appropriate neural network architecture is akin to choosing an effective learning strategy for a specific task: a well-designed model is capable of acquiring the desired knowledge, while an unsuitable architecture may fail to learn altogether. Consequently, the design of neural network architectures has become a major focus of research, further propelling the development of artificial intelligence. Much like the human brain, once a neural network has successfully acquired knowledge, it can quickly and effectively apply it to new data. Moreover, because DNNs are trained on a large dataset, they

exhibit a degree of robustness against noise and perturbations, enabling them to make accurate inferences even under challenging conditions. In the context of text-to-speech (TTS) tasks, for example, DNNs function similarly to the human brain: after extensive training on the pronunciation of text, they can rapidly respond to new textual input and synthesize corresponding speech. Furthermore, even if the input text contains minor distortions or noise, the model is often capable of producing accurate and intelligible output.

In the early stages, WaveNet [46] pioneered neural waveform modeling by introducing an autoregressive (AR) architecture that generates speech sample-by-sample with high fidelity. Despite its outstanding synthesis quality, WaveNet suffers from prohibitive computational complexity and latency due to its sequential generation nature. To address this limitation, Parallel WaveNet [47] was proposed, employing a knowledge distillation mechanism to achieve parallel waveform generation while preserving synthesis quality. Subsequently, WaveGlow [48] integrated Glow-based normalizing flows with WaveNet-style components to achieve a more favorable balance between speed and fidelity. Meanwhile, WaveRNN [49] adopted a recurrent neural network (RNN) framework to reduce computational cost significantly while maintaining high-quality waveform generation. In parallel, the emergence of generative adversarial networks (GANs) [50] further revolutionized neural vocoder design by introducing an adversarial learning paradigm. Models such as Parallel WaveGAN [51] and MelGAN [52], [53] applied GAN architectures to non-autoregressive vocoder frameworks, where a generator is trained to produce waveforms that fool a discriminator trained to distinguish between real and synthetic speech. Among these, HiFi-GAN [4] stands out for its ability to generate high-quality speech at real-time speeds, employing multi-scale discriminators and multi-resolution STFT loss to ensure both waveform fidelity and perceptual quality. HiFi-GAN transforms mel-spectrograms into waveforms through multi-stage upsampling in the time domain and spectral compression, enabling it to efficiently model fine-grained waveform details.

1.2.3 Application of Speech Modeling

Speech modeling plays a pivotal role across various speech processing tasks, where the objective is to construct representations that accurately reflect the acoustic, prosodic, or linguistic information of speech. Below, we describe several key application areas and how speech modeling is integrated into each task.

In speech coding, the goal is to compress the speech signal for efficient storage or transmission while preserving perceptual quality. For instance, the telecommunication technique has significantly shortened the distance between people, enabling speech information to be transmitted via wireless communication systems. However, due to technical limitations, bandwidth constraints have become a bottleneck that restricts transmission efficiency. Therefore, it is crucial to use speech modeling to encode the speech signals. Conventional approaches, such as linear predic-

tive coding (LPC) [54], model the spectral envelope of speech using a linear filter and separately encode the excitation signal. This modeling approach enables low-bitrate coding and forms the basis of widely used codecs such as code excited linear prediction (CELP) [55]. By leveraging a parametric model of speech production, these methods reduce redundancy while maintaining intelligibility.

In automatic speech recognition (ASR), speech modeling is used to match acoustic observations to phonetic or linguistic units. Hidden Markov models (HMMs) [56] combined with Gaussian mixture models (GMMs) [57] were long considered the standard framework, where HMMs model the temporal dynamics of phonemes and GMMs characterize the emission distributions. The speech signal is first converted into acoustic features, such as MFCCs (mel-frequency cepstral coefficients) [58, 57], which are then aligned with phonetic state sequences using the model. This process enables robust speech-to-text conversion and has been applied in a wide range of real-world ASR systems.

In speaker recognition, the task is to determine or verify a speaker's identity from speech. For example, with the rapid development of smartphones, these devices have become increasingly similar to computers in their ability to store vast amounts of information and media, including sensitive personal data. Consequently, speaker recognition systems can serve as a form of biometric authentication, enhancing the protection of user privacy and data security. Conventional systems often employ Gaussian mixture models [59] to represent the distribution of acoustic features for each speaker. During enrollment, a GMM is trained per speaker using their speech data; during inference, the likelihood of the test utterance under each speaker model is evaluated. This statistical modeling approach effectively captures speaker-specific characteristics and supports both identification and verification cases.

Speech synthesis, particularly statistical parametric speech synthesis, also heavily relies on speech modeling. HMM-based synthesis methods [60] model the distributions of spectral, pitch, and duration parameters conditioned on linguistic input. In synthesis, these parameters are generated from the model and passed to a vocoder to reconstruct the waveform. The parametric nature of the model allows explicit control over speaking rate, pitch, and other prosodic attributes, making it suitable for applications such as audiobook generation or assistive speech devices.

Speech modification is another meaningful task that aims to modify the intonation or speed of the speech while keeping the retained speech information unchanged. The essence of intonation modification is actually to change the frequency of the voiced speech, whereas the key to changing speech speed is to extend or shorten the duration of the local phonemes. As mentioned before, the methods in the sinusoidal model family, such as sinusoidal model [61], are the typical approaches that represent speech as a sum of sinusoids with time-varying amplitudes, frequency, and phases. Thanks to their capacity for extracting framewise amplitudes, frequency, and phases, these models are particularly effective in vocoding and speech modification tasks. The model's

interpretability makes it useful for analyzing and modifying specific frequency bands or formants, which is important in speech modification applications.

Additionally, in voice conversion, the task is to modify the speech of a source speaker to sound as if it were spoken by a target speaker, without changing the linguistic content. Conventional approaches often decompose speech into spectral envelope, pitch, and duration, and then map the source speaker's features to those of the target using statistical models such as Gaussian mixture models [62]. The converted features are then resynthesized into a waveform using vocoders. This technique finds application in personalized text-to-speech, dubbing, and assistive communication systems.

With the advent of deep learning, many of these conventional methods have been augmented or replaced by neural networks, since DNN can bring superior quality and robustness that conventional algorithms cannot achieve. Many techniques that were previously difficult or infeasible to implement, such as voice conversion, text-to-speech, and speech enhancement, can now be effectively realized through the application of DNNs. Additionally, many well-established techniques, such as speech coding and speaker recognition, have also experienced substantial performance improvements through the integration of DNN.

In text-to-speech (TTS), systems such as Tacotron [63] generate mel-spectrograms from text, which are then converted to waveforms using neural vocoders like WaveNet [46] or HiFi-GAN [4]. These models learn the entire mapping from text to audio, enabling natural and high-quality speech generation that outperforms earlier parametric approaches. Besides, in speech enhancement, neural speech models are trained to suppress noise while preserving clean speech characteristics. Given noisy speech as input, the model predicts a clean signal [64], either in the time domain or spectral domain. Such models are applied in hearing aids, telecommunication systems, and robust speech interfaces. Additionally, DNN helped voice conversion methods directly learn mappings between source and target speaker characteristics without requiring parallel data. Models such as CycleGAN-VC [65] and AutoVC [66] disentangle speaker identity from linguistic content, enabling flexible many-to-many voice conversion. These models extract a content representation from the input speech and then generate speech in the target speaker's voice using a neural decoder or vocoder. Neural approaches greatly improve naturalness and speaker similarity, making them suitable for applications like personalized assistants, robot voice generation, and expressive speech synthesis.

DNN is also key in neural speech coding, where end-to-end models [67] are trained to compress and reconstruct speech. Unlike the conventional versions, these systems jointly optimize feature extraction and quantization, enabling higher quality at lower bit rates. In speaker recognition, DNN-based embeddings like *x*-vectors [68] are learned to represent speaker identity in a compact form. These embeddings are then used in back-end classification systems, supporting robust verification across different languages and acoustic conditions.

As speech modeling continues to evolve, it remains at the core of enabling high-quality, controllable, and interpretable speech processing across both conventional and modern neural frameworks.

1.3 Potential Limitations and Problem Formulation

1.3.1 Limitations of Existing Speech Modeling Methods

Although speech modeling technologies have made significant strides, each approach still has its limitations. It is precisely these limitations that have drawn the attention of many researchers, greatly advancing the field. In this subsection, we analyze the aforementioned methods and models in detail, examining their respective advantages and disadvantages.

First, we focus on the source-filter model and the sinusoidal model family. Source-filter vocoders, such as STRAIGHT [69] and WORLD [2], decompose speech into excitation and spectral envelope components, mimicking the human vocal production mechanism. These models provide interpretable control over pitch, timbre, and duration with efficient computing, making them well-suited for applications requiring flexible prosody manipulation and voice conversion in a real-time scheme. However, their performance degrades when the assumptions of stationarity and minimum-phase filtering are violated. In other words, if the pitch varies rapidly over time, it often results in artifacts such as buzzy or over-smoothed speech. More importantly, most existing source-filter models exhibit significant limitations in accurately modeling the phase information, which inevitably leads to a reconstructed speech waveform that deviates considerably from the ground truth, further degrading the quality of synthesis and modification.

Additionally, sinusoidal models, including HNM and quasi-harmonic models such as QHM, aQHM, and eaQHM, represent speech as a sum of time-varying sinusoids with optional noise components. Likewise, these models also offer an interpretable framework, which is particularly effective for high-quality synthesis and modification of voiced segments, especially when phase information is accurately modeled. Advanced variants, such as aQHM, eaQHM, further improve modeling accuracy and stability to both voiced and unvoiced segments by introducing adaptive refinement and energy-aware iteration, leading to the perfect modeling and reconstruction quality but a substantial computation cost. Nonetheless, sinusoidal vocoders often struggle with unvoiced or transient segments, although eaQHM can provide a considerable capacity to fit unvoiced speech. Moreover, the reliance on accurate f_0 estimation of them can limit performance in real-world cases, for instance, if the f_0 initially detected by the pitch detector significantly deviates from the ground truth, the performance of HNM degrades sharply, and even though eaQHM attempts to iteratively refine the frequency estimates, it often fails to converge to the accurate values. In some cases, it may even amplify the estimation errors instead of correcting them, implying that the robustness is limited. What's more, the iterative estimation brings a considerable

computation cost, making it unable to be applied in real-time processing, which has distanced aQHM and eaQHM from the advantage of the high computation speed of conventional methods. However, since they do not require data-driven training, their implementation and deployment are relatively straightforward.

On the other hand, neural vocoders such as WaveNet [46], HiFi-GAN [4], and WaveGlow [48] directly learn the mapping from acoustic features (e.g., mel-spectrograms) to waveforms using DNNs. These models achieve near-natural speech quality and are highly data-driven, enabling them to generalize across diverse speaking styles and languages. Due to the large amount of training data, the results are usually considerably better than the results of conventional methods, especially in terms of multi-style and multi-speaker tasks. However, neural vocoders require large amounts of data and computational resources for training. Since collecting training data requires considerable effort, this becomes a disadvantage of neural methods compared to conventional approaches. Moreover, neural vocoders often lack interpretability and fine control over speech parameters because of their black-box nature, which limits their applicability in prosody-sensitive tasks such as expressive synthesis or speech modification. In other words, most neural vocoders fail in speech modeling, including analysis and modification. Moreover, their inference effectiveness is usually influenced by their structure: a simple architecture of DNN allows for high-speed or even real-time processing, while a complicated structure will bring a heavy computation cost, leading to unnecessary waste in device configuration. Moreover, their performance can degrade under mismatched or noisy conditions if the training data is insufficient, where conventional parametric models may instead offer more stability and robustness.

Overall, source-filter models offer interpretable structure, with strong controllability and efficiency, but they suffer from limited spectral detail modeling and assumptions that modulation should be weak. Sinusoidal models focus on the speech signal itself to precisely represent speech signals as interpretable parameters and are particularly well-suited for periodic signals, yet they are sensitive to pitch estimation and less robust for aperiodic or unvoiced segments. On the other hand, neural vocoders achieve high-quality, robust, and natural synthesis through end-to-end learning, but they lack interpretability, require large datasets, and high computational resources. In contrast, conventional approaches are usually fast, easy to implement, and not data-driven. Their performance in analysis is limited, which then degrades the synthesis quality. Therefore, hybrid frameworks that combine the interpretability of conventional models with the learning capacity of neural networks are an active area of research.

A comprehensive summary of the advantages and limitations of these methods is presented in Fig. 1.4, illustrating that the ultimate goal of vocoders is to integrate the strengths of different approaches while addressing their weaknesses.

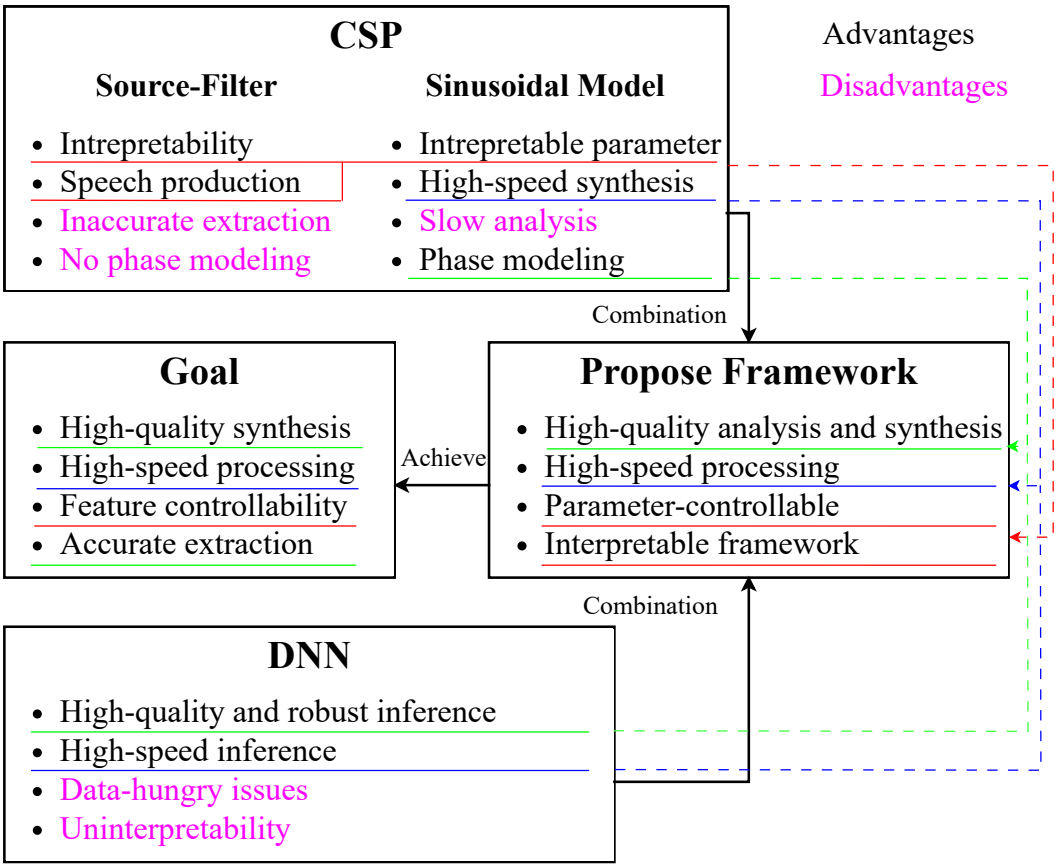


Figure 1.4: Overview of existing speech modeling methods, summarizing their advantages and disadvantages. The figure also illustrates the motivation of hybrid approaches that aim to combine the strengths of different paradigms while mitigating their weaknesses.

1.3.2 Problem Formulation

Taking into account the advantages and disadvantages of each method shown in Fig. 1.4, it can be concluded that the ultimate goal of speech modeling is to construct systems that enable

1) accurately extracting acoustic features,

2) high-quality synthesis,

3) high-speed processing,

4) acoustic feature controllability.

Conventional vocoders, including source-filter models and sinusoidal-based approaches (e.g., WORLD and QHM), offer interpretability and explicit control of speech parameters, which is useful for tasks requiring precise modification. Neural vocoders, on the other hand, achieve superior synthesis quality and robustness through data-driven learning, and can even operate in real time with efficient designs.

These characteristics are inherently complementary: conventional methods provide transparency and controllability, while neural vocoders excel in naturalness and generalization. This comple-

mentarity motivates us to explore hybrid approaches that integrate the strengths of both paradigms while compensating for their individual weaknesses.

Based on this perspective, several key research questions arise:

- Is it possible to effectively combine conventional modeling with neural vocoders to simultaneously leverage their respective strengths to achieve the main goals of speech modeling?
- How to design neural architectures and training schemes that preserve interpretability and controllability while ensuring synthesis quality and speed?
- How to build a unified system and algorithms capable of achieving parameter controllability and speech modification?

Addressing these questions is critical for advancing the field of speech modeling, particularly for applications requiring high-quality, editable, and real-time synthesis.

1.4 Purpose of Research

Given that no single existing method satisfies all four requirements, the purpose of this thesis is to develop a unified speech modeling framework, namely a novel vocoder framework, that simultaneously satisfies four essential requirements: accurate extraction of acoustic features (amplitude, frequency, phase), high-fidelity reconstruction and modification of speech, real-time processing capability, and strong controllability over prosodic and temporal parameters.

To achieve this, this thesis adopts an integrative perspective to explore the integration of conventional signal processing-based models with neural vocoder architectures. The objective is not merely to combine the two, but to design a framework where interpretable representations, typically found in source-filter or sinusoidal models, can be effectively incorporated into modern neural systems without sacrificing synthesis quality or efficiency.

Specifically, the QHM framework models speech as a sum of sparse sinusoids and allows for efficient waveform generation based on interpretable parameters, namely, amplitude, frequency, and phase. This makes it a powerful tool for speech analysis, synthesis, and modification. Given our goal of leveraging deep learning for more accurate and efficient signal processing (speech modeling), we first investigate the feasibility of the integration of neural networks and QHM architecture by applying backpropagation to QHM structures for extracting more accurate interpretable parameters, paving the way for the subsequent combination of the conventional and neural systems. Then, we integrate QHM structure with neural networks to propose a novel hybrid neural vocoder framework to achieve the four essential requirements of speech modeling.

By doing so, the proposed framework enables both high-quality synthesis and flexible, semantically meaningful speech modification, such as pitch and time scaling. Furthermore, the system is designed to support real-time applications like voice conversion and prosody editing, making it

suitable for deployment in practical, latency-sensitive scenarios. Additionally, the ways to alleviate the data-hungry issue are also explored.

Ultimately, this research seeks to advance the field of speech modeling by demonstrating that interpretability and learning-based expressiveness can coexist within a cohesive, efficient, and extensible system.

To clearly delineate the shortcomings of current methods and motivate the proposed solution, an overview is presented in Fig. 1.4. The advantages and disadvantages of CSP and DNN are highlighted in black and pink, respectively, where the dotted lines linking the CSP block and DNN block mean the combination of the advantages. This figure shows the inspiration of the combination of CSP and DNN, which is the main idea of this thesis.

1.5 Thesis Organization

Fig. 1.5 gives the structure of this thesis, showing that the rest of this thesis is organized in five parts, as follows:

1. In Chapter II, the theories of related methods are reviewed.
   (a) The development history and derivative process of the QHM methods will be elaborated in detail.
   (b) The details of some typical neural vodoers are introduced.
   (c) The limitations of QHM methods and neural vocoders are listed.
2. Chapter III demonstrates an attempt to combine backpropagation and QHM framework for improving the performance of parameter extraction, which further hastens the development of BP-QHM and the improvement of the speech synthesis quality.
   (a) The inspiration and derivative process of the QHM methods are demonstrated in detail.
   (b) Some experiments were conducted, and their results indicate that our proposed method outperforms conventional QHM methods in terms of parameter extraction and speech resynthesis.
3. In Chapter III, the success of combining backpropagation and the QHM framework shows the potential of the integration of neural networks and the QHM framework. Chapter IV introduces novel neural vocoders (QHM-GAN and QHARMA-GAN) that integrates the advantages of neural networks and the QHM framework, significantly accelerating the processing and improving the parameter extraction accuracy and the quality of synthesis.
   (a) The inspiration and integration process are demonstrated in detail.

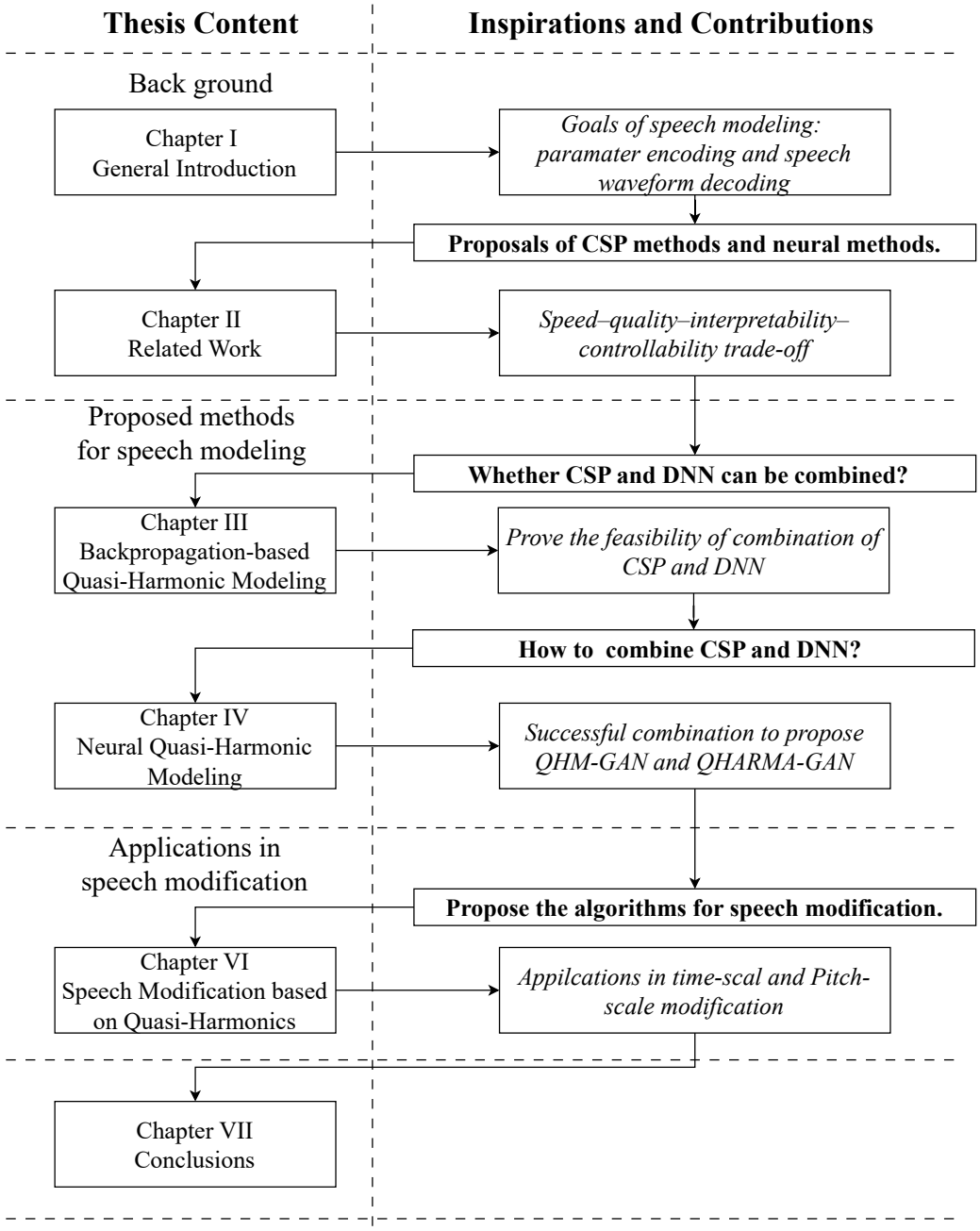


Figure 1.5: The structure of the thesis.

(b) Some experiments were conducted, and their results indicate that our proposed method integrates the advantages of conventional vocoders and neural vocoders, from the outperformance of our proposed vocoder in interpretable parameter extraction, speech resynthesis, and the processing effectiveness.

4. In Chapter V, the speech modification algorithms for various methods are illustrated, including the QHM methods and our proposed methods (QHM-based neural methods).

(a) The algorithms and pseudocode of speech modification for various methods are elaborated in detail.

(b) Some experimental results will indicate the advantages and disadvantages of all methods and their applicable occasions.

5. Eventually, in Chapter VI, the conclusions are summarized. The future directions will also be discussed.

Chapter 2

Related Work

In this chapter, we review representative vocoders for speech synthesis, encompassing both conventional vocoders and neural vocoders, along with a discussion of their respective limitations.

As introduced in Chapter I, speech signals can be viewed as a combination of harmonic components and stochastic noise. This observation inspired the development of sinusoidal modeling (SM), which represents speech waveforms as a sum of sinewaves. While this approach marked a significant advancement in vocoder design, it suffers from two key limitations that impair both modeling accuracy and resynthesis quality. First, unvoiced speech segments do not exhibit harmonic structure but instead resemble stochastic noise, which SM cannot model effectively. Second, SM lacks mechanisms for refining frequency estimates; therefore, if the initial frequency estimation is inaccurate, the quality of the synthesized speech will be significantly compromised. To address these issues, a series of QHM-based methods were proposed, which model both voiced and unvoiced components simultaneously. These methods progressively improve the precision of parameter extraction and achieve higher-quality speech synthesis.

Despite the advancements brought by conventional methods such as the QHM family, their robustness remains limited. In particular, their performance degrades significantly when speech signals are contaminated by noise. Moreover, the computational complexity introduced by their iterative algorithms results in considerable time consumption, making them less suitable for real-time applications. With the emergence of deep learning, neural network-based vocoders have gained increasing attention. Among them, HiFi-GAN has emerged as a prominent and widely adopted neural vocoder, offering both high-quality synthesis and efficient inference through a GAN-based architecture.

Overall, QHM methods and neural vocoders such as HiFi-GAN exemplify two different paradigms in vocoder development: conventional signal processing and data-driven modeling, respectively. Each paradigm offers unique advantages and faces distinct limitations in terms of synthesis quality, computational efficiency, and robustness. In the subsequent sections, we present a comprehensive overview of these two representative approaches, detailing their underlying principles, technical implementations, and practical limitations.

2.1 Quasi-Harmonic Model Family

As discussed in Chapter I, speech signals can be considered as a combination of harmonic components and stochastic noise, as expressed in Eq. (1.1). Based on this assumption, the SM represents the speech signal using a set of pure sinewaves. Although SM is conceptually intuitive and offers interpretability, it suffers from two major limitations, significantly reducing both the modeling accuracy and the quality of the resynthesized speech:

1) **Inaccurate frequency estimation.**

   The SM heavily depends on prior frequency estimates f_k, typically extracted from the ridges in the spectrogram obtained by STFT. These ridges correspond to peaks in the magnitude spectrum and are used to approximate the signal's instantaneous frequencies. However, due to the inherent time–frequency resolution trade-off, especially in cases of rapid pitch fluctuation or closely spaced harmonics, the detected ridges often deviate from the true frequencies. Accurate estimation further requires a large number of samples per frame, which is impractical given the high sampling rates (e.g., 16000 Hz, 24000 Hz). Even with longer frames, true frequencies are often non-integer values, making perfect resolution unattainable. Additionally, the speech is usually strongly modulated in terms of frequency, since the vocal fold usually vibrates with different frequencies to make the speech exhibit time-varying pitch. However, SM uses fixed frequencies to model the speech within a frame, leading to inevitable mismatches in frequency estimation. Consequently, the resulting estimation errors degrade the overall synthesis quality.

2) **Inadequate modeling of unvoiced segments.**

   Unvoiced speech lacks harmonic structure, as it is produced without vocal fold vibration. These segments exhibit broadband, noise-like characteristics and are inherently stochastic. The sparse and frequency-fixed nature of sinusoidal components fails to capture such spectral properties, making the SM ineffective for modeling or reconstructing unvoiced sounds. This leads to noticeable artifacts and degraded perceptual quality in the synthesized speech.

To address these challenges, the QHM and its extensions have been proposed. With a frequency refinement mechanism in the modeling process, QHM methods significantly improve the accuracy of parameter estimation. As a result, it enhances both the representational fidelity and the resynthesis quality, making it more powerful to adapt various speech types, including both voiced and unvoiced segments.

2.1.1 Quasi-Harmonic Model

To improve the performance of SM in modeling, the two main focuses should be addressed:

1) How to improve the frequency estimations?

2) How to model the unvoiced part?

To address the aforementioned two limitations, a number of studies have undertaken extensive investigations to improve the modeling accuracy. One notable advancement is proposed in [70] to address the limitation that individual frequencies are treated as fixed within a frame and often obtained from inaccurate pitch. Thus, the QHM [34] introduces a complex slope term to augment the constant amplitude. This enhancement allows the model to more flexibly represent the target speech signal and to adaptively refine the individual frequencies, thereby bringing them closer to their true values. QHM is formulated as:

$$x(t) = \left[\sum_{k=-K}^{K} (a_k + t\, b_k)\, e^{i2\pi \hat{f}_k t} \right] w(t), \quad t \in [-T_l, T_l], \tag{2.1}$$

where $\hat{f}_k$ represents the initial estimate of the k-th component's frequency, and b_k denotes its complex slope. This formulation is referred to as QHM. In this formulation, the term tb_k introduces a time-dependent component that enables the model to capture subtle frequency deviations and amplitude modulations. Consequently, the model is capable of representing speech signals in a non-harmonic manner, thereby increasing its flexibility and improving its ability to fit the waveform more precisely.

The Fourier transform of Eq. (2.1) can be readily derived as:

$$X(f) = \sum_{k=-K}^{K} \left[a_k W(f - \hat{f}_k) + i \frac{b_k}{2\pi} \frac{\partial W}{\partial f}(f - \hat{f}_k) \right], \tag{2.2}$$

where W is the Fourier transform of the window function $w(t)$, and $\frac{\partial W}{\partial f}$ denotes its derivative with respect to frequency f. Assuming that all frequency components are sufficiently isolated such that their frequency trajectories do not overlap, we can analyze each component independently. For the k-th component, Eq. (2.2) simplifies to:

$$X_k(f) = a_k W(f - \hat{f}_k) + i \frac{b_k}{2\pi} \frac{\partial W}{\partial f}(f - \hat{f}_k). \tag{2.3}$$

To interpret b_k in terms of the geometric relationship with a_k, we consider both as vectors in the complex plane. The complex slope b_k can be decomposed into components parallel and perpendicular to a_k:

$$b_k = \rho_{p,k} a_k + \rho_{v,k} i a_k, \tag{2.4}$$

where $ia_k = (-a_k^I, a_k^R)$ represents a 90° rotation of a_k in the complex plane. Here, a_k^R and a_k^I denote the real and imaginary parts of a_k, respectively. The scalar coefficients $\rho_{p,k}$ and $\rho_{v,k}$ correspond to the projections of b_k onto a_k and ia_k, respectively. These projections are computed using the

inner product as:

$$\rho_{p,k} = \frac{\langle a_k, b_k \rangle}{|a_k|^2} = \frac{a_k^R b_k^R + a_k^I b_k^I}{|a_k|^2}, \tag{2.5a}$$

$$\rho_{v,k} = \frac{\langle ia_k, b_k \rangle}{|a_k|^2} = \frac{a_k^R b_k^I - a_k^I b_k^R}{|a_k|^2}, \tag{2.5b}$$

where $\langle \cdot, \cdot \rangle$ denotes the complex inner product. Substituting Eq. (2.5) into Eq. (2.3) allows us to express the Fourier transform of the k-th component as:

$$X_k(f) = a_k \left[W(f - \hat{f}_k) - \frac{\rho_{v,k}}{2\pi} \frac{\partial W}{\partial f}(f - \hat{f}_k) + i \frac{\rho_{p,k}}{2\pi} \frac{\partial W}{\partial f}(f - \hat{f}_k) \right]. \tag{2.6}$$

Using a first-order Taylor expansion of $W(f - \hat{f}_k - \frac{\rho_{v,k}}{2\pi})$, we obtain:

$$W(f - \hat{f}_k - \frac{\rho_{v,k}}{2\pi}) = W(f - \hat{f}_k) - \frac{\rho_{v,k}}{2\pi} \frac{\partial W}{\partial f}(f - \hat{f}_k) + O(\rho_{v,k}^2 \frac{\partial^2 W}{\partial f^2}(f - \hat{f}_k)). \tag{2.7}$$

where $O(\rho_{v,k}^2 \frac{\partial^2 W}{\partial f^2})$ accounts for the higher-order frequency-shift effects. Since $\rho_{v,k}$ is assumed small, these terms are negligible in the current approximation. Neglecting the imaginary term $i\frac{\rho_{p,k}}{2\pi}\frac{\partial W}{\partial f}$ due to its relatively small magnitude, Eq. (2.3) can be approximated as:

$$\begin{aligned} X_k(f) &= a_k W(f - \hat{f}_k) + i \frac{b_k}{2\pi} \frac{\partial W}{\partial f}(f - \hat{f}_k) \\ &\approx a_k W(f - \hat{f}_k - \frac{\rho_{v,k}}{2\pi}), \end{aligned} \tag{2.8}$$

which implies that the main effect of the tb_k term is to introduce a frequency shift. The inverse Fourier transform of Eq. (2.8) gives:

$$x_k(t) \approx a_k e^{i(2\pi \hat{f}_k + \rho_{v,k})t}. \tag{2.9}$$

This result reveals that the frequency of the k-th component has been adjusted by an amount $\rho_{v,k}/(2\pi)$ relative to the initial estimate $\hat{f}_k$. Denoting this frequency deviation as η_k, we have:

$$\eta_k = f_k - \hat{f}_k = \frac{\rho_{v,k}}{2\pi} = \frac{a_k^R b_k^I - a_k^I b_k^R}{2\pi |a_k|^2}. \tag{2.10}$$

Therefore, once a_k and b_k are obtained from the time-domain model, we can directly compute the frequency estimation error η_k for each component. It is precisely the introduction of the tb_k term that empowers the model to account for frequency modulation through a first-order Taylor expansion in the frequency domain. This enriched structure effectively addresses the issue of mismatch between the analysis model and the ground truth, thereby resolving the first of the two problems identified earlier [34].

Apart from refining the frequency components, QHM also estimates the complex amplitudes and slopes, namely a_k and b_k, to further determine the amplitude and phase parameters. This

is achieved by solving Eq. (2.1) via the least squares (LS) method, whose details are presented below.

Let a_k and b_k be represented as K-dimensional vectors $\mathbf{a}$ and $\mathbf{b}$, respectively. By concatenating them, the parameter to be estimated is defined as:

$$\boldsymbol{\chi} = \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}. \tag{2.11}$$

The cost function is defined as the error between the ground-truth signal and the QHM-generated signal:

$$\begin{aligned} J(\mathbf{a},\mathbf{b}) &= \sum_{n=-N}^{N} |x[n] - x_{\mathrm{QHM}}[n]|^2 \\ &= \sum_{n=-N}^{N} (x[n] - x_{\mathrm{QHM}}[n])^* (x[n] - x_{\mathrm{QHM}}[n]). \end{aligned} \tag{2.12}$$

Let the original signal without windowing be denoted as $s[n]$, i.e., $x[n] = s[n]w[n]$. Then the cost function can be rewritten as:

$$\begin{aligned} J(\mathbf{a},\mathbf{b}) &= \sum_{n=-N}^{N} (s[n] - s_{\mathrm{QHM}}[n])^* w^*[n] w[n] (s[n] - s_{\mathrm{QHM}}[n]) \\ &= (\mathbf{s} - \mathbf{s}_{\mathrm{QHM}})^H \mathbf{W}^H \mathbf{W} (\mathbf{s} - \mathbf{s}_{\mathrm{QHM}}), \end{aligned} \tag{2.13}$$

where $\mathbf{W}$ is a diagonal matrix formed from the window function, and $\mathbf{s}_{\mathrm{QHM}}$ is the synthesized signal vector constructed based on Eq. (2.1):

$$\begin{aligned} \mathbf{s}_{\mathrm{QHM}} = s_{\mathrm{QHM}}[n] &= \sum_{k=-K}^{K} (a_k + n b_k) e^{i2\pi f_k n / f_s} \\ &= \sum_{k=-K}^{K} a_k e^{i2\pi f_k n / f_s} + n b_k e^{i2\pi f_k n / f_s} \\ &= \mathbf{E}_0 \mathbf{a} + \mathbf{E}_1 \mathbf{b} = \begin{bmatrix} \mathbf{E}_0 | \mathbf{E}_1 \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \mathbf{E}\boldsymbol{\chi}, \end{aligned} \tag{2.14}$$

where $\mathbf{E}_0$ and $\mathbf{E}_1$ are $(T \times K)$ matrices consisting of sinusoidal basis functions corresponding to a_k and b_k, respectively.

To minimize the cost function $J(\mathbf{a},\mathbf{b})$, we take its derivative with respect to $\boldsymbol{\chi}$ and set it to zero:

$$\frac{\partial J(\mathbf{a},\mathbf{b})}{\partial \boldsymbol{\chi}} = \frac{\partial}{\partial \boldsymbol{\chi}} (\mathbf{s} - \mathbf{E}\boldsymbol{\chi})^H \mathbf{W}^H \mathbf{W} (\mathbf{s} - \mathbf{E}\boldsymbol{\chi}) = 0. \tag{2.15}$$

Solving the above equation yields the optimal solution:

$$\hat{\boldsymbol{\chi}} = \begin{bmatrix} \hat{\mathbf{a}} \\ \hat{\mathbf{b}} \end{bmatrix} = (\mathbf{E}^H \mathbf{W}^H \mathbf{W} \mathbf{E})^{-1} \mathbf{E}^H \mathbf{W}^H \mathbf{W} \mathbf{s}. \tag{2.16}$$

After estimating a_k and b_k, these parameters can be used to compute the frequency deviation, which in turn allows for the adaptive correction of each component's frequency. The updated frequency f_k is computed as:

$$f_k = \hat{f}_k + \hat{\eta}_k = \hat{f}_k + \frac{1}{2\pi} \cdot \frac{\hat{a}_k^R \hat{b}_k^I - \hat{a}_k^I \hat{b}_k^R}{|\hat{a}_k|^2}, \tag{2.17}$$

where η_k denotes the estimated frequency mismatch between the true and estimated values for the k-th component. The terms $\hat{a}_k^R$ and $\hat{a}_k^I$ represent the real and imaginary parts of $\hat{a}_k$, respectively, and $\hat{b}_k^R$ and $\hat{b}_k^I$ represent those of $\hat{b}_k$. This correction helps reduce the error between the original and synthesized speech in the time domain.

For the second challenge that unvoiced speech is difficult to represent using purely harmonic components, prior studies have investigated alternative approaches. Specifically, [38] and [3] explored the feasibility of approximating unvoiced speech using combinations of sinewaves. The underlying assumption is that unvoiced speech, often characterized as stochastic noise, can be viewed as a rapidly time-varying signal with strongly modulated frequency and amplitude. To better capture these variations, a chirp-based spectrogram can be employed using the Short-Time Fan-Chirp Transform (STFChT) [71, 72], which is defined as:

$$\mathrm{STFChT}(t, \omega, \alpha) = \int_{\mathbb{R}} x(u)\, g(u-t)\, \xi(t, \omega, \alpha)\, \mathrm{d}u, \tag{2.18}$$

where the kernel function $\xi(t, \omega, \alpha)$ is given by:

$$\xi(t, \omega, \alpha) = e^{-i\omega\left[(u-t) + \frac{\alpha}{\omega}(u-t)^2\right]}, \tag{2.19}$$

and α denotes the chirp rate that controls the instantaneous frequency variation. The STFChT enables enhanced visualization of time-varying frequency trajectories by incorporating the chirp rate α, which is pre-estimated to match the dynamics of the target speech signal. As a result, frequency components in the spectrogram are sharpened, improving interpretability. Figure 2.1 presents a comparison between conventional STFT and STFChT on a sample speech signal. In the voiced segments, the harmonic trajectories become notably sharper and more distinct, facilitating harmonic structure detection. Interestingly, in the unvoiced segments, the STFChT also induces structured, harmonic-like patterns in what is conventionally considered stochastic noise. This observation suggests that such noise components may also be decomposed into sums of modulated sinewaves.

Motivated by this insight, [38, 3] proposed using QHM to perform unified full-band modeling

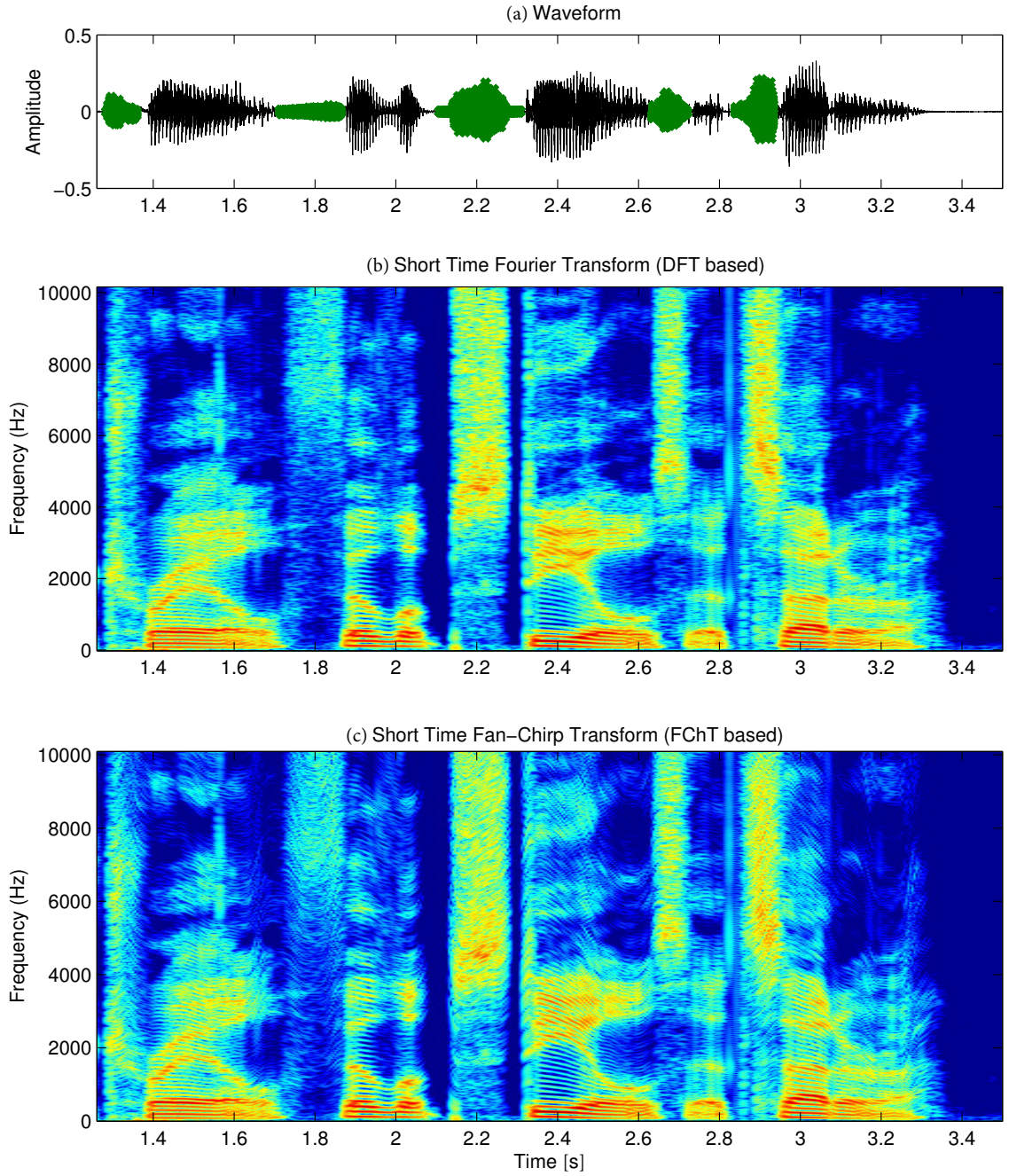


Figure 2.1: (a) The waveform of a speech sample in [3]. (b) and (c) are the STFT and STFChT of the speech signal, respectively.

of both voiced and unvoiced speech. Since QHM is capable of refining frequency estimates with high precision, it can be extended to unvoiced segments by initializing the model with harmonic frequencies. These initial frequencies are then iteratively adjusted to better conform to the complex, noise-like structure of the unvoiced signal. In doing so, both voiced and unvoiced speech can be jointly represented using sinewave components. This unified modeling approach simplifies the analysis process and ensures continuity across different speech segments.

Although the frequencies of the components exhibit a ”stochastic” pattern in the unvoiced seg-

ments after the frequency correction step, the estimated frequencies still remain insufficiently close to the ground truth, even within the voiced segments. This is primarily because the initially provided frequency values are significantly misaligned from their true counterparts, leading QHM to struggle with accurately estimating the amplitude a_k and slope b_k. These inaccuracies, in turn, compromise the frequency correction process.

To address this issue, an iterative refinement scheme based on QHM was proposed in [73]. This approach involves multiple rounds of estimation and correction to progressively improve alignment with the true frequency components. The overall procedure of this iterative framework can be summarized as follows: in the first iteration, the initial frequency estimates are provided to the QHM model (Eq. (2.1)), from which the amplitude $\hat{a}_k$ and slope $\hat{b}_k$ are estimated using Eq. (2.16). These estimates are then used to update the frequencies via Eq. (2.17). This completes the first iteration. In subsequent iterations, the newly corrected frequencies are re-fed into the QHM model to repeat the process. Through this iterative update mechanism, the estimated frequencies gradually converge toward the ground truth. Notably, in unvoiced segments, the refined frequencies begin to display a stochastic distribution, which better captures the intrinsic randomness and variability of such segments. This enhances the model's ability to represent natural unvoiced speech. It is important to emphasize that all of these iterative refinements are based on the standard QHM framework, which inherently assumes frequency stationarity within each frame, that is, it treats frequencies as constant over the analysis window. Consequently, some residual mismatch may still exist between the corrected frequencies and their true values, particularly in highly non-stationary segments of speech. In practice, stochastic noise can be modeled as a superposition of multiple sinewaves whose frequencies and amplitudes vary rapidly over time. A model based on fixed-frequency assumptions is inherently limited in its capacity to represent such signals. This limitation is not confined to unvoiced speech alone; even voiced segments such as singing voices often exhibit rapidly time-varying pitch trajectories. Unlike normal speech, where pitch typically changes gradually, singing voices often require rapid transitions from low to high notes, resulting in much faster pitch modulation. These dynamic pitch variations are, in fact, essential to the musicality and expressiveness of singing.

To better handle these non-stationary characteristics, an adaptive extension of QHM was proposed. This adaptive version is designed to track time-varying frequencies more accurately, thereby enhancing the model's ability to represent both voiced and unvoiced segments with greater fidelity.

2.1.2 Adaptive Quasi-Harmonic Model

To enhance the ability of models to align with the time-varying, modulated frequencies present in natural speech, as mentioned above, the adaptive quasi-harmonic model (aQHM) [35] modifies the exponential term of the QHM model by incorporating non-stationary phases, as proposed in

[35]. The aQHM model is defined as:

$$x(t) = \left\{ \sum_{k=-K}^{K} (a_k + t b_k)\, e^{i[\hat{\varphi}_k(t+t_l) - \hat{\varphi}_k(t_l)]} \right\} w(t), \quad t \in [-T_l, T_l] \tag{2.20}$$

where $\hat{\varphi}_k(t)$ denotes the phase function of the k-th harmonic component, and t_l is the center of the l-th analysis frame ($l = 1, \cdots, L$). By replacing the stationary phase term in QHM with the dynamically evolving phase $\hat{\varphi}_k(t)$, aQHM aims to more accurately capture the intra-frame frequency modulations present in natural speech. The main point of aQHM able to capture the frequency modulations within a frame, is that aQHM uses the phase functions in the exponential part, which is calculated from time-varying frequencies instead of the fixed frequencies. The frequencies are obtained by interpolation, causing the time variations of the frequencies. Therefore, during the LS, aQHM can more accurately match the signals with strong frequency modulations.

aQHM operates under an iterative framework, similar to the iterative variant of QHM, where frequency estimates are progressively refined to approximate the true speech characteristics. Initially, QHM is used to generate a coarse estimation of the frequency trajectories. The resulting estimates are then interpolated over successive time frames using cubic interpolation, yielding smooth instantaneous frequencies. The corresponding instantaneous phase is subsequently computed by integration:

$$\varphi_k(t) = \varphi_k(t_l) + \int_{t_l}^{t_l+t} 2\pi f_k(u)\, \mathrm{d}u, \quad t \in [-T_l, T_l]. \tag{2.21}$$

In the next iteration, this time-varying phase function replaces the constant-phase exponential term in the synthesis model, allowing for improved alignment with the actual speech waveform. Then, similar to Eq. (2.14), the modified sinusoidal basis functions can be expressed as:

$$\mathbf{E}_{\mathbf{a},\mathbf{0}} = e^{i(\varphi_k(n+t_l) - \varphi_k(t_l))}, \quad \mathbf{E}_{\mathbf{a},\mathbf{1}} = n e^{i(\varphi_k(n+t_l) - \varphi_k(t_l))} \tag{2.22}$$

where $\varphi_k(t)$ is computed from Eq. (2.21). These updated basis functions, $\mathbf{E}_{\mathbf{a},\mathbf{0}}$ and $\mathbf{E}_{\mathbf{a},\mathbf{1}}$, are substituted into Eq. (2.16) to estimate the complex amplitude and slope. The frequencies are then updated and serve as the input for the subsequent iteration, where a new phase trajectory is again computed using Eq. (2.21). In each iteration, the refined model in Eq. (2.20) is applied to obtain more accurate estimates of a_k and b_k. Through this recursive procedure, aQHM progressively improves both the frequency and amplitude parameters. The estimated frequencies converge to their true trajectories, and the complex amplitudes (a_k, b_k) better reflect the underlying modulated nature of the speech signal. As a result, aQHM can synthesize speech with higher perceptual fidelity and more accurate spectral content than conventional QHM.

Unfortunately, in many practical scenarios, the amplitude envelope of speech exhibits nonlinear temporal variation, which cannot be effectively captured by linear models. Both QHM and aQHM operate under the assumption that amplitude varies linearly within a frame, a simplification

that limits their ability to track more intricate amplitude modulations frequently encountered in natural speech. This limitation becomes particularly prominent in expressive or dynamic speech segments, such as screaming or vibrato in singing, where amplitude fluctuations are rapid and highly nonlinear. To overcome this challenge, it is necessary to extend the amplitude modeling framework beyond the linear assumption, potentially by incorporating higher-order polynomial terms or leveraging data-driven adaptive approaches that are capable of tracking the true nonlinear amplitude dynamics of speech signals.

2.1.3 Extended Adaptive Quasi-Harmonic Model

To address the limitations of QHM and aQHM in modeling nonlinear amplitude variations, an extended adaptive quasi-harmonic model (eaQHM) was proposed, which incorporates an explicit amplitude modulation mechanism into the model formulation [36]:

$$x(t) = \left\{ \sum_{k=-K}^{K} (a_k + t b_k) \frac{A_k(t+t_l)}{A_k(t_l)} e^{i[\hat{\varphi}_k(t+t_l) - \hat{\varphi}_k(t_l)]} \right\} w(t),$$
$$t \in [-T_l, T_l] \tag{2.23}$$

where $A_k(t)$ denotes the time-varying amplitude of the k-th harmonic component. Unlike aQHM, which assumes a linear amplitude variation within an analysis frame, eaQHM introduces an amplitude amplifier derived as the ratio between the instantaneous amplitude $A_k(t+t_l)$ and the reference amplitude at the frame center $A_k(t_l)$. This amplitude modulation term serves to nonlinearly scale the basis functions, thereby enabling the model to more accurately capture fast or nonlinear amplitude fluctuations that are prevalent in expressive or emotional speech, such as singing, shouting, or stressed phonation. The key point of eaQHM able to extract the signals with nonlinearly modulated frequencies, is that, similar to the frequency process in aQHM, the amplitude is interpolated into an instantaneous version, which exhibits a time-varying pattern. Then, the time-varying amplitude will be considered during the LS optimization.

The inclusion of the amplitude amplifier enhances the flexibility of the model without changing the analysis workflow. In practice, the implementation of eaQHM proceeds similarly to that of aQHM. First, the QHM framework is used to obtain initial estimates of the model parameters, including the frequencies and complex amplitudes of all components. Then, the eaQHM uses such pre-extracted parameters to start the iterations for eaQHM. In each iteration, the model updates the complex amplitude parameters a_k and b_k using the modified sinusoidal basis that now incorporates both time-varying phase and amplitude modulation. Similar to Eq. (2.14), the sinusoidal basis metrics for eaQHM can be obtained by

$$\mathbf{E}_{\mathbf{e},\mathbf{0}} = \frac{A_k(n+t_l)}{A_k(t_l)} e^{i(\varphi_k(n+t_l) - \varphi_k(t_l))}, \quad \mathbf{E}_{\mathbf{e},\mathbf{1}} = n \frac{A_k(n+t_l)}{A_k(t_l)} e^{i(\varphi_k(n+t_l) - \varphi_k(t_l))}. \tag{2.24}$$

The corresponding harmonic frequencies f_k are also re-estimated and serve as inputs to the next it-

eration, ensuring that both amplitude and frequency characteristics are progressively aligned with the true underlying speech signal. Then, substituting these bases into Eq. (2.16), the estimated complex amplitudes and slopes for eaQHM will be estimated by LS. Then, the frequencies will be refined through the a_k and b_k estimated in the current iteration to get more accurate instantaneous frequency and instantaneous phase. The same process will be conducted for amplitude. Then, the amplitude and phase containing time variations will be the input of the analysis in the next iteration. This iterative refinement continues until convergence, typically determined by the stability of the frequency and amplitude estimates or the minimization of a modeling error criterion.

By accounting for intra-frame amplitude modulations explicitly, eaQHM significantly improves the model's ability to represent the dynamic nature of real speech signals. In comparison to QHM and aQHM, which rely on more restrictive assumptions, eaQHM achieves superior performance in both analysis and resynthesis, particularly in segments with strong prosodic variation or rapid energy fluctuations. The enhanced model fidelity leads to synthesized speech with improved perceptual quality and spectral consistency.

It is worth noting that in all QHM-related frameworks, including eaQHM, the modification of frequency components across frames leads to a relaxation of the strict harmonicity condition. That is, the resulting sinusoidal components are not constrained to be exact integer multiples of a single fundamental frequency. This quasi-harmonic structure is especially beneficial in modeling unvoiced or noise-like segments of speech, where purely harmonic models fail to capture the inherent stochasticity. Consequently, these models are termed quasi-harmonic models, emphasizing their ability to represent both voiced and unvoiced speech components within a sinusoidal framework.

2.1.4 Synthesis Process of QHM Methods

The synthesis part is important for a vocoder, since it is crucial for generating high-quality speech. As for QHM methods, the speech can be reconstructed after obtaining the complex amplitudes and updated frequencies from the eaQHM analysis. The reconstructed speech signal is synthesized from estimated harmonic parameters, as shown in

$$\hat{x}(t) = \sum_{k=-K}^{K} \hat{A}_k(t) e^{i\hat{\varphi}_k(t)}, \tag{2.25}$$

where $\hat{A}_k(t)$ and $\hat{\varphi}_k(t)$ denote the instantaneous amplitude and phase of the k-th component, respectively, and the notation $\hat{(\cdot)}$ indicates that the quantities are estimated values. This synthesis formula mirrors the quasi-harmonic assumption by reconstructing the signal as a sum of modulated sinusoids, each representing a spectral component of speech. The details of the synthesis part are demonstrated below.

First, we need to determine the framewise parameters. Assuming that the estimated parameters

correspond to those used in the signal representation in Eq. (2.1), and considering framewise analysis with local coordinates centered at $t = 0$, the instantaneous amplitude and phase within the frame can be expressed in terms of the complex amplitude trajectory as

$$\hat{A}_k(t) = |\hat{a}_k + t\hat{b}_k|, \quad \hat{\varphi}_k(t) = 2\pi f_k t + \angle\left(\hat{a}_k + t\hat{b}_k\right), \tag{2.26}$$

where $\hat{a}_k$ and $\hat{b}_k$ represent the complex amplitude and slope, estimated within the frame. Such framewise amplitude and phase jointly determine the instantaneous amplitude envelope and the instantaneous phase of each component. In practical implementations, since QHM-based models operate in a framewise manner, the amplitude and phase are typically estimated and emphasized at the center of each analysis window (i.e., $t = 0$). Therefore, the amplitude and phase for the l-th frame at time t_l can be written as

$$\hat{A}_k(t_l) = |\hat{a}_k^l|, \quad \hat{\varphi}_k(t_l) = \angle\hat{a}_k^l, \tag{2.27}$$

where $\hat{a}_k^l$ is the estimated complex amplitude of the k-th component at the l-th frame. Given these values at the centers of all frames, the full instantaneous amplitude $\hat{A}_k(t)$ can be approximated across the frames by linear interpolation, while the instantaneous phase trajectory $\hat{\varphi}_k(t)$ can be obtained by integrating the estimated instantaneous frequency $\hat{f}_k(u)$ over time:

$$\tilde{\varphi}_k(t) = \hat{\varphi}_k(t_l) + \int_{t_l}^{t_l+t} 2\pi\hat{f}_k(u)\mathrm{d}u. \tag{2.28}$$

To ensure a smooth and natural-sounding instantaneous phase, the frequency $\hat{f}_k(u)$ is interpolated using cubic splines. This smooth interpolation preserves the continuity and natural variation of the frequency curve, which is crucial for high-fidelity synthesis, particularly in expressive speech. However, it is well known that instantaneous frequency estimation is prone to small errors, and even with smooth interpolation, these discrepancies can accumulate during the interpolation, potentially leading to discontinuities in the phase across adjacent frames. For instance, focusing on two adjacent framewise phase $\varphi(t_l)$ and $\varphi(t_{l+1})$, we can find that these two phases are computed from the framewise complex amplitude by Eq. (2.27). Thus, if these two phases and the frequency between the l-th and the $l+1$-th frames are correct, they satisfy that

$$\varphi_k(t_{l+1}) = \varphi_k(t_l) + \int_{t_l}^{t_{l+1}} 2\pi f_k(u)\mathrm{d}u. \tag{2.29}$$

Nevertheless, it is hard to ensure the above equation. There must be mismatches arising during the analysis. Such phase mismatches, especially at frame boundaries, often manifest as audible frequency jitters in the reconstructed signal, namely the perceivable frequency distortions.

To overcome this issue, an adaptive phase compensation strategy is introduced in [35] to spread the phase error over time in a perceptually smooth manner. Specifically, the instantaneous phase is

adjusted using a sinusoidal correction term, leading to the modified phase reconstruction equation:

$$\hat{\varphi}_k(t) = \hat{\varphi}_k(t_l) + \int_{t_l}^{t_l+t} \left[2\pi \hat{f}_k(u) + z \sin \frac{\pi(u - t_{l-1})}{t_l - t_{l-1}} \right] \mathrm{d}u, \tag{2.30}$$

where the added sinusoidal term dynamically adjusts the frequency trajectory such that the resulting phase aligns better with the target value at the next frame. The magnitude of the correction is controlled by the coefficient z, which is calculated to minimize the phase discontinuity at the frame boundary. The correction term z is computed by

$$z = \frac{\pi \left[\hat{\varphi}_k(t_{l+1}) + 2\pi Q - \tilde{\varphi}_k(t_{l+1}) \right]}{2(t_{l+1} - t_l)}, \tag{2.31}$$

where Q is the integer closest to $|\hat{\varphi}_k(t_{l+1}) - \tilde{\varphi}_k(t_{l+1})|/2\pi$, which is calculated by

$$Q = \textbf{round} \left[\frac{|\hat{\varphi}_k(t_{l+1}) - \tilde{\varphi}_k(t_{l+1})|}{2\pi} \right]. \tag{2.32}$$

This adjustment ensures that the integrated phase at t_{l+1} smoothly converges to the desired phase estimate, avoiding sudden jumps.

This phase correction mechanism, initially proposed in [35], provides robustness against minor errors in frequency estimation, thereby enhancing the temporal coherence of the synthesized waveform. As a result, the speech reconstructed with this method exhibits smoother transitions and improved perceptual quality, particularly across frame boundaries where phase mismatches are most problematic. Additionally, aQHM and eaQHM need to be iteratively extracted to obtain the complex amplitude and complex slope, in which the accurate instantaneous frequencies and phases are used as the input to the analysis. Thus, the accuracy of the instantaneous frequencies and phases is crucial. Such a synthesis process with a phase compensation mechanism helps to provide robust and accurate instantaneous phases, further ensuring the stability of the analysis.

2.1.5 Limitations of QHM Methods

QHM methods are capable of accurately extracting complex amplitudes through frequency adaptation, enabling high-quality speech modeling and resynthesis. Nevertheless, several limitations still persist.

Inadequate Frequency Optimization

QHM methods rely on the f_0 (also referred to as pitch) to generate individual harmonic frequencies a priori, i.e., $f_k = kf_0$, where f_0 is typically estimated by a pitch detection algorithm. However, most pitch detectors assume that the analyzed signal is locally stationary within a given window, leading to pitch estimation errors. As a result, the detected f_0 is often mismatched from the true value, introducing frequency mismatches, particularly for $k = 0$, as denoted by η_k in Eq. (2.10). This error, even if it is small, becomes increasingly pronounced for higher harmonics,

as it scales linearly with the harmonic index k:

$$\eta_k = f_k - \hat{f}_k = k(f_0 - \hat{f}_0) = k\eta_0, \tag{2.33}$$

which inevitably deteriorates the accuracy of harmonic component extraction. This phenomenon is especially critical in speech signals, where even subtle pitch estimation errors can lead to significant spectral distortions in higher-order harmonics.

Fig. 2.2(a) shows the STFT of a speech signal along with the corresponding detected harmonic frequencies. To improve readability, Fig. 2.2(b) presents a sectional view of the STFT magnitude (in decibels) at $t = 0.34$ sec. We also employed the YAAPT [74] to detect the f_0 trajectory, which is also depicted in Fig. 2.2(a) and pointed out in Fig. 2.2(b). From Fig. 2.2(a), it is obvious that a f_0 mismatch exists over the frames, which increases the frequency mismatch with the increase of the order of the harmonic components. To more specifically demonstrate this phenomenon, an enlarged view was made in Fig. 2.2(b). As seen in the zoomed region, the detected f_0 deviates from the true value by approximately 6 Hz, which results in a significantly larger error in the higher harmonics; for instance, the 19th harmonic exhibits a deviation of about 114 Hz.

Although QHM methods allow for iterative refinement of frequencies to reduce such mismatches, the convergence of this process is sensitive to the initial frequency error and the rate of frequency variation. When the initial deviation is large or the frequency changes rapidly, the iterative process may struggle to correct the estimates and may even exacerbate the mismatch. This difficulty arises because harmonics interact with their adjacent components, effectively narrowing the main lobe width and limiting the range within which frequency corrections are feasible. A large mismatch may cause the estimated frequency $\hat{f}_k$ to fall within the dominant region of an adjacent harmonic component (e.g., f_{k-1}), rendering it uncorrectable.

Furthermore, QHM methods update the frequency by estimating the complex amplitude a_k and slope b_k via LS optimization. However, when the initial frequency estimate is far from the correct value, the LS solution tends to be biased toward neighboring spectral components, further amplifying the frequency deviation. In other words, the result of LS is unstable if the frequency mismatch is large, where the LS results are not accurate, further leading to the biased estimation of frequency mismatch, which increases the frequency mismatch. Therefore, large initial deviations may enter a feedback loop of error propagation, in which inaccurate frequency estimates corrupt the LS solution, which in turn reinforces the incorrect frequency, leading to a deterioration in both frequency tracking and harmonic component extraction.

To show the phenomenon mentioned above, a simulated signal with time-varying frequencies

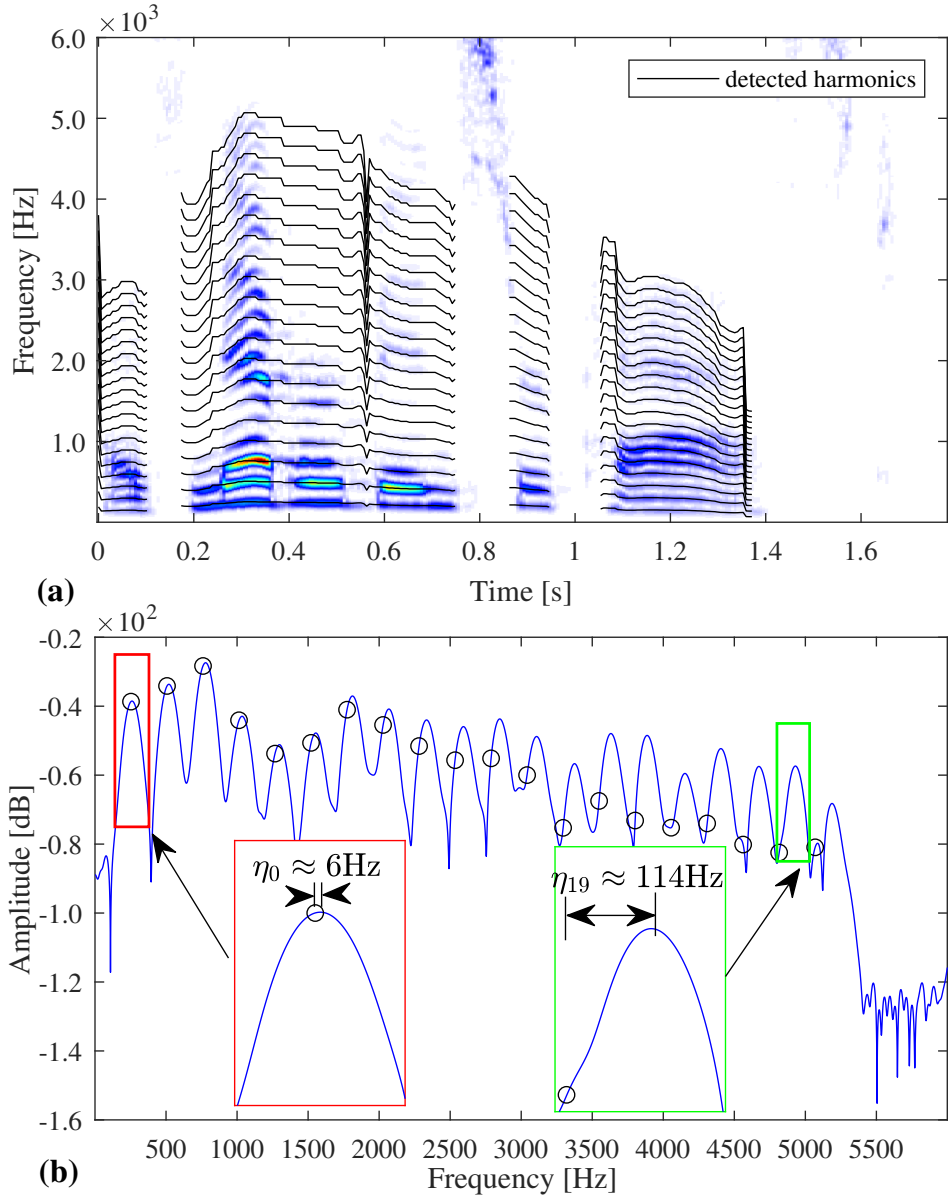


Figure 2.2: (a) STFT of a speech signal and the detected harmonic frequencies; (b) sectional view at $t = 0.34$ sec. The f_0 mismatch (6 Hz) results in a 114-Hz deviation at the 19th harmonic.

was analyzed, as

$$x(t) = \sum_{k=1}^{5} A_k(t) e^{i\varphi_k(t)}, \tag{2.34}$$

$$\text{with } \varphi_k(t) = \phi_k + k \int (300t + 150) dt,$$

$$A_1 = 0.4, A_2 = 0.5, A_3 = 0.8, A_4 = 0.5, A_5 = 0.7,$$

where ϕ_k is randomly initialized. The f_0 can be easily obtained as $f_0(t) = 300t + 150$. We added different frequency mismatches to f_0, i.e., $[-10, 10]$ and $[-45, 45]$. We used eaQHM, which

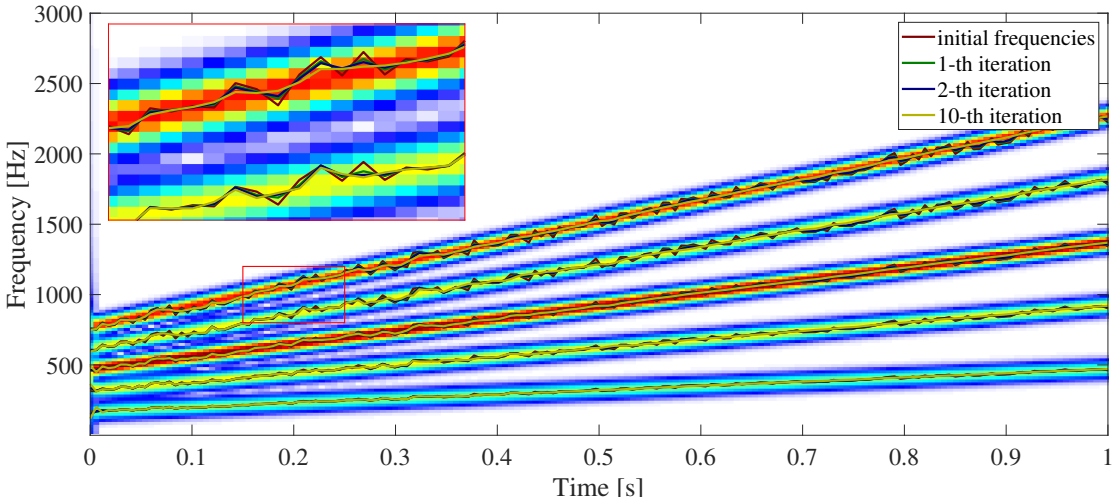


Figure 2.3: Frequency estimates by eaQHM using initial f_0 whose mismatch $\in [-10, 10]$. Only the first two iterations and the final result, namely the 10th iteration, are shown for clarity.

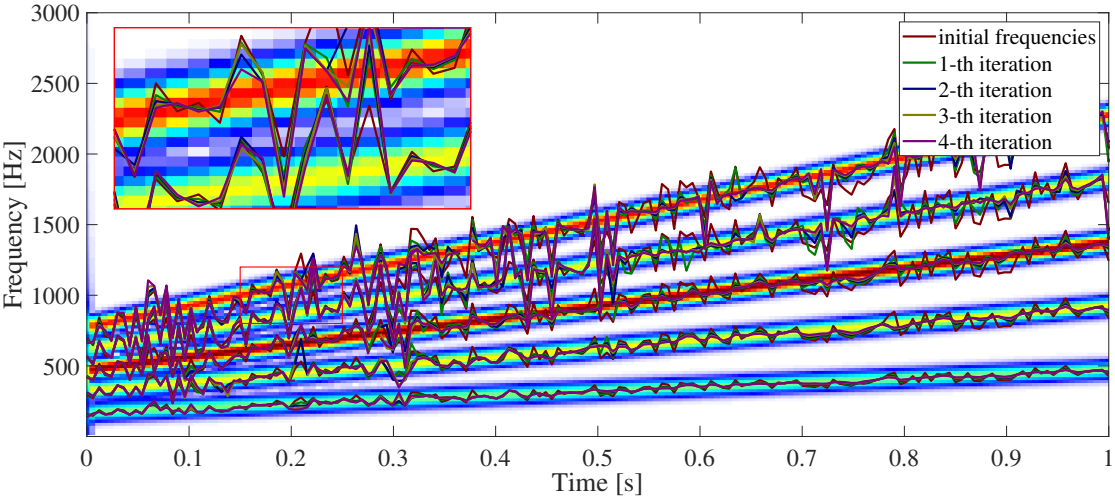


Figure 2.4: Frequency estimates by eaQHM using initial f_0 whose mismatch $\in [-45, 45]$.

requires an iterative process. We depicted the frequencies of all components extracted by eaQHM in Figs. 2.3 and 2.4.

Fig. 2.3 shows that when the mismatch is small, eaQHM can iteratively reduce the mismatch. On the contrary, Fig. 2.4 shows that eaQHM found it hard to reduce the mismatch when the mismatch is large. Besides, it can also be found in Fig. 2.4 that, in the front part, the frequency correction was not optimal, as the small frequency spacing limited eaQHM's ability to reduce the mismatch. In the back part, however, the larger frequency spacing led to improved frequency correction results.

Additionally, from these two figures, it is apparent that the frequencies were improved, but the frequency mismatches still exist, even through the iterative improvement. That is caused by the biased LS results, which unsatisfactorily correct the frequencies.

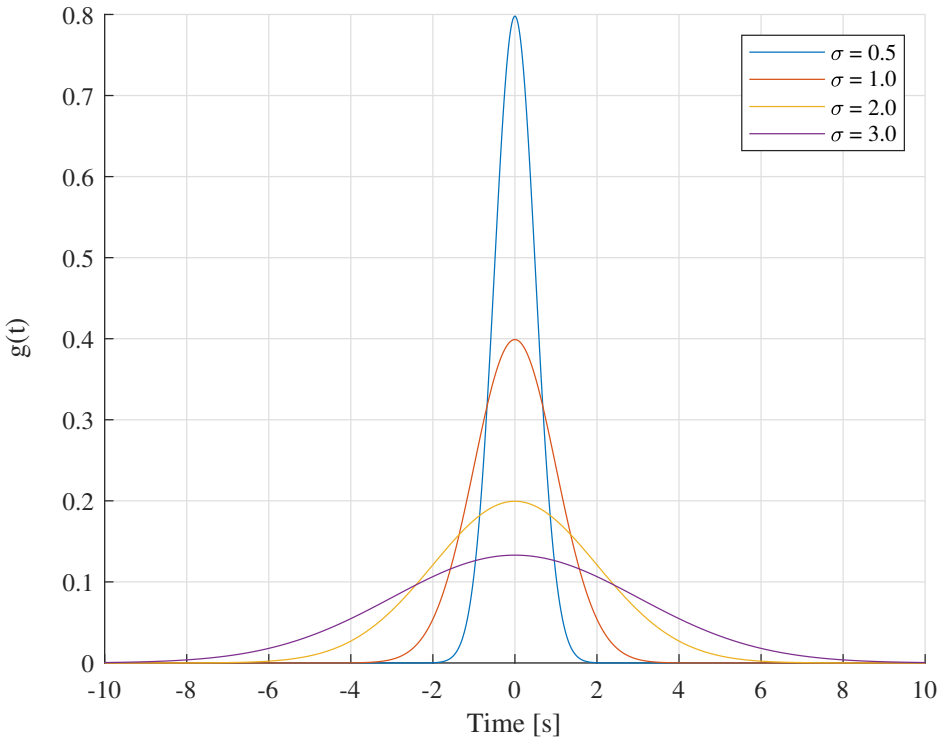


Figure 2.5: Gaussian windows with different σ values.

Framewise Modeling

Most existing speech modeling approaches segment the signal into frames and analyze each frame independently using a set of basis functions for parameter estimation. The performance of such forward, framewise methods heavily depends on the design of the window function and the chosen frame-shift. An inappropriate window configuration may impair the accuracy of parameter estimation. For instance, the shape of the Gaussian window, which is formulated as

$$g(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2\sigma}}, \tag{2.35}$$

will be affected by the coefficient σ. Fig. 2.5 depicts the Gaussian windows with different σs. This will affect the time resolution and frequency resolution of STFT results. It affects the connection between frames, further influencing the extraction in each frame. For instance, when the σ is small, such as the blue line in Fig. 2.5, signals at the center of the window are amplified, while those near the edges are attenuated. In this way, the time resolution is high, which is beneficial for the exact analysis of the current frame. However, when conducting LS, it tends to focus more on the signals at the center of the window, while the signals at the edges are largely disregarded. Thus, this window is inappropriate.

Additionally, an excessively large frame-shift risks overlooking the information between adjacent frames, as only truncated segments of the waveform are processed individually. In QHM,

parameters such as amplitude, frequency, and phase are extracted at the center of each frame for resynthesis. While frequency refinement can be performed within each frame, estimation errors in any single frame can cause deviations in interpolated parameters between frames. This issue becomes more severe with larger frame-shifts, especially when estimation errors (e.g., η_k) are large. For instance, if the error is significant enough to shift $\hat{f}_k$ into the main lobe of f_{k-1}, the LS estimation may incorrectly adjust $\hat{f}_k$ toward f_{k-1}. During resynthesis, this misalignment leads to discontinuities in the interpolated instantaneous frequency, thereby degrading the quality of the synthesized speech.

In summary, both the reliance on initially accurate pitch detection and the inherent limitations of framewise modeling pose challenges for QHM methods, particularly in terms of frequency precision and inter-frame continuity. Overcoming these challenges requires either improved initialization strategies for f_0, robust to local non-stationarities, or model structures that better exploit temporal dependencies across frames to enhance estimation stability.

2.2 Neural Vocoder

As mentioned in the discussion of the limitations of QHM methods, the model structure determines the performance of the speech modeling. For instance, from QHM to eaQHM, the complexity of the model increases while the performance of speech modeling is gradually improved. However, they can not get rid of the unrobustness against to the disturbances from noise in the waveform or the detected pitch. Namely, once the speech signals contain heavy noise or the frequency estimations exhibit substantial mismatches, QHM methods struggle to model the speech accurately. Such a limitation is widespread in conventional methods.

With the rapid development of deep learning, neural vocoders have emerged as a powerful alternative to conventional signal processing-based speech synthesis frameworks. Unlike sinusoidal models that rely on a predefined parametric form to approximate the speech signal, neural vocoders adopt a data-driven approach that learns the complex mapping from acoustic features, typically mel-spectrograms, to the corresponding time-domain waveform. By leveraging large-scale training data and powerful neural architectures, these models are capable of capturing both the harmonic structure and the stochastic characteristics of speech with significantly higher fidelity, even in a noisy case. Therefore, the main advantage of data-driven methods, i.e., they are usually robust, is fully demonstrated in the neural vocoders.

Early neural vocoders such as WaveNet [46] and WaveGlow [48] demonstrated the possibility of synthesizing highly natural speech by modeling the waveform sample by sample using autoregressive or flow-based frameworks. While autoregressive models like WaveNet suffer from high computational complexity and slow inference, non-autoregressive models such as WaveGlow can achieve real-time synthesis through parallel generation, although their model sizes remain large due to the normalizing flow architecture. To overcome remaining limitations, a class of non-

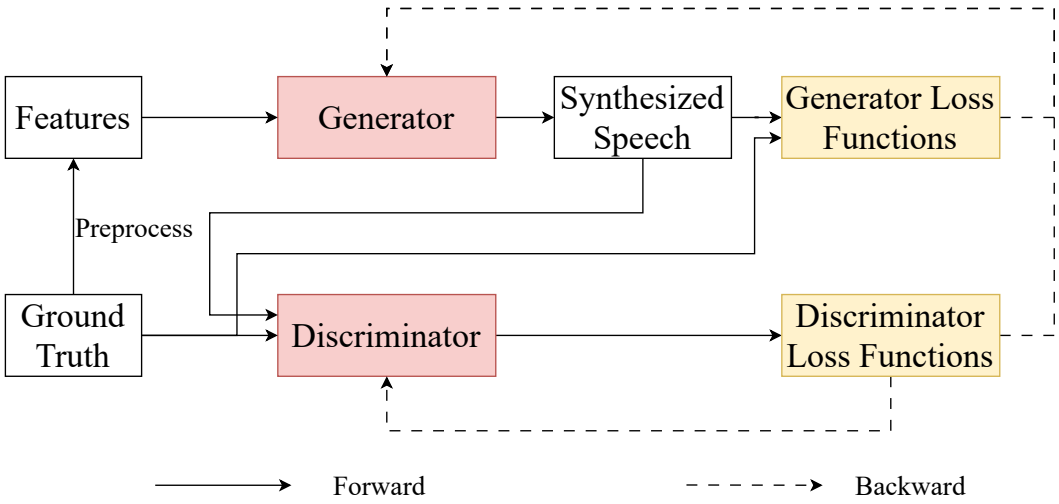


Figure 2.6: The overall structure of GAN-based methods.

autoregressive and GAN-based vocoders has been developed. Among them, MelGAN [52] and HiFi-GAN [4] are two representative architectures that achieve both high-quality synthesis and real-time inference. These models are built upon a generator–discriminator paradigm, in which the generator transforms mel-spectrograms into waveforms, while the discriminator guides the training through adversarial learning, encouraging the generator to produce speech that is perceptually indistinguishable from natural speech. Fig. 2.6 is the diagram of such GAN-based methods. Both the generator and the discriminator were trained simultaneously during the training. Additionally, in order to further accelerate the inference speed of vocoders, the conventional signal processing algorithms, such as iSTFT, are combined with neural networks to narrow the gap between framewise parameters and sequence-wise speech signals and alleviate the pressure on the network in learning. Typical representations, such as iSTFTnet [75] and Vocos [5], adopt neural networks to estimate complex spectrograms, which will be inverted into speech waveforms by iSTFT, based on the GAN structure.

In the following, we introduce the architecture and training strategies of HiFi-GAN and Vocos, which are recent successful GAN-based vocoders, and analyze how its key components contribute to the high-quality synthesis of speech signals across both voiced and unvoiced segments.

2.2.1 HiFi-GAN: High-Fidelity Generative Adversarial Network

HiFi-GAN [4] is a high-fidelity neural vocoder that directly converts mel-spectrograms to time-domain waveforms, achieving high-quality and efficient waveform generation. The overall architecture of HiFi-GAN consists of a generator and a set of discriminators trained adversarially. The generator is responsible for synthesizing waveforms from acoustic features, while the discriminators guide the generator by distinguishing real speech from synthetic ones from multiple perspectives.

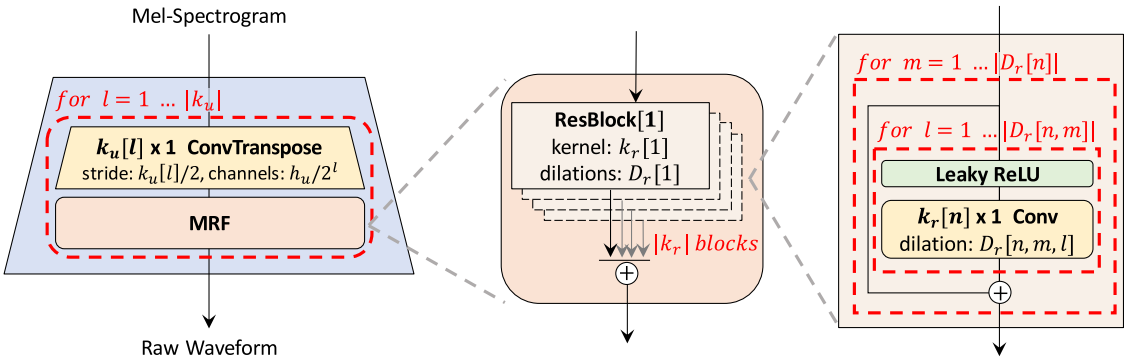


Figure 2.7: The structure of the HiFi-GAN generator in [4]. The generator upsamples mel-spectrograms up to $|k_u|$ times to match the temporal resolution of raw waveforms. A MRF module adds features from $|k_r|$ residual blocks of different kernel sizes and dilation rates. Lastly, the n-th residual block with kernel size $k_r[n]$ and dilation rates $D_r[n]$ in an MRF module is depicted.

Generator Architecture: To introduce the structure of the HiFi-GAN generator in detail, Fig. 2.7 is shown to specify the details. The generator in HiFi-GAN adopts a hierarchical structure that progressively upsamples the input mel-spectrogram to waveform resolution. The upsampling is performed by a cascade of transposed convolution layers, each increasing the temporal resolution by a fixed factor k_u. After each transposed convolution, the intermediate feature is passed through a Multi-Receptive Field Fusion (MRF) module, which is designed to enrich the receptive field and capture temporal dependencies at multiple resolutions. Each MRF module comprises several parallel residual blocks, and each residual block uses a different kernel size k_r and dilation D_r configuration. This design allows the model to capture fine-grained and long-term dependencies simultaneously, as smaller kernels focus on local detail while larger dilations cover broader contexts. Specifically, each residual block within MRF follows a structure of two 1D convolution layers with LeakyReLU activations and skip connections. The use of grouped convolutions in these blocks reduces the computational cost while maintaining representational capacity. This multi-resolution design is particularly beneficial in modeling the quasi-periodic and multi-scale nature of speech signals, allowing the generator to flexibly model both the harmonic structure and stochastic components of speech. Moreover, the progressive upsampling architecture ensures that each stage works at an appropriate temporal scale, facilitating stable training and fast inference.

Discriminator Design: To effectively supervise the generator and ensure high perceptual quality, HiFi-GAN employs a set of discriminators: the Multi-Scale Discriminator (MSD) and the Multi-Period Discriminator (MPD), whose structures were illustrated in Fig. 2.8. The MSD consists of several sub-discriminators operating on waveforms downsampled to different temporal resolutions. This enables the model to evaluate both the local and global structure of the waveform, ensuring coherence at different timescales.

The MPD, on the other hand, slices the waveform into periodic segments with different predefined periods (e.g., 2, 3, 5, 7 samples per period) and applies discriminators to each segment.

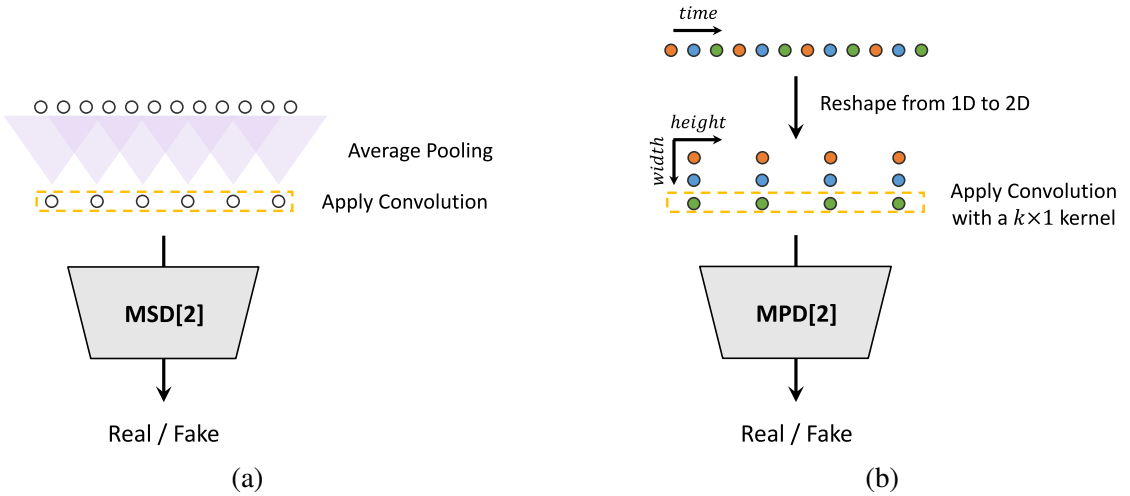


Figure 2.8: (a) The second sub-discriminator of MSD and (b) the second sub-discriminator of MPD with period 3 in [4].

This periodic slicing is motivated by the observation that voiced speech contains strong periodicity, particularly in voiced segments where the pitch structure dominates. By evaluating periodic patterns explicitly, MPD can guide the generator to produce harmonically consistent waveforms, especially in voiced segments.

Training Objectives: The generator and discriminator are jointly optimized in an adversarial framework. The generator is trained using a weighted combination of adversarial loss $L_{\text{g,adv}}$, feature matching loss L_{fm}, and mel-spectrogram reconstruction loss L_{mel}:

$$L_{\text{G}} = L_{\text{g,adv}} + \lambda_{\text{fm}} L_{\text{fm}} + \lambda_{\text{mel}} L_{\text{mel}}, \tag{2.36}$$

where λ_{fm} and λ_{mel} are hyperparameters that balance the contribution of each term. The feature matching loss L_{fm} stabilizes training by minimizing the L^1 distance between the discriminator feature maps of real and generated samples, while the mel-spectrogram loss L_{mel} ensures that the generated waveform maintains acoustic fidelity in the perceptual domain.

The discriminators are trained with a standard adversarial loss $L_{\text{d,adv}}$, summed over all sub-discriminators in both MPD and MSD. This multi-view adversarial supervision encourages the generator to produce speech that is both globally natural and locally periodic.

Advantages and Limitations: Compared with earlier GAN-based vocoders such as MelGAN, HiFi-GAN significantly improves speech fidelity while maintaining fast inference speed. The key improvement lies in the introduction of the MRF modules and the dual discriminator architecture. In MelGAN, each residual block has a fixed receptive field, which limits its ability to model both local detail and global structure. HiFi-GAN overcomes this by combining multiple receptive fields in parallel, thereby capturing speech characteristics at multiple temporal scales.

Additionally, MelGAN relies solely on MSD, whereas HiFi-GAN introduces the MPD to exploit pitch periodicity, which proves highly effective in producing periodically rich speech. Moreover, although models like WaveNet or WaveGlow achieve high audio quality, they are computationally expensive and unsuitable for real-time applications. HiFi-GAN, through careful architectural simplification and the use of grouped convolutions and transposed convolutions, offers a balance between audio quality and computational efficiency. It supports real-time inference on modern GPUs.

Despite its advantages, HiFi-GAN also has some limitations. Its generator is an end-to-end structure, lacking an interpretable structure that reflects the intrinsic structure of speech signals, such as the source-filter structure. Without external f_0 information, HiFi-GAN cannot extrapolate pitch contours beyond the training distribution. Although some variants [76, 77] incorporate f_0 as an additional condition, such control remains constrained by the coverage of the training data. Furthermore, although HiFi-GAN almost supports real-time processing, the generation speed is still limited, hindering the deployment on a lightweight device. This is caused by the transposed convolutions between MRF modules. The convolutions in MRF modules increasingly become heavy as the number of upsampling increases. For example, in the first MRF, the convolutions are conducted based on the framewise data, while the convolutions in the last MRF are based on sequence-wise data.

In summary, the MRF modules and MPD enable a high-fidelity and fast waveform generation, setting a strong baseline for neural vocoding.

2.2.2 Vocos

HiFi-GAN has already achieved a relative high-quality and high-speed speech synthesis, however, the upsampling process in HiFi-GAN somewhat hinders the generation speed while its end-to-end architecture increases the pressure of study. HiFi-GAN needs to learn the intrinsic structures of speech, such as harmonic patterns, from scratch, since the input (mel-spectrogram) and the output (speech waveform) are in different domains, meaning that the network should face a challenge in converting data across domains, which inevitably increases the burden of learning. To overcome such issues, many studies were conducted to employ the conventional algorithm in the neural vocoder, such as Vocos [5] and iSTFTnet [75].

Among recent advances in such vocoders, Vocos proposes a novel Fourier-based architecture that effectively closes the gap between conventional spectral vocoders and modern time-domain GAN-based methods. Unlike HiFi-GAN and its variants, which directly generate time-domain waveforms using a series of transposed convolutions and adversarial training, Vocos formulates speech synthesis in the frequency domain. Specifically, Vocos learns to predict both the magnitude and phase of the complex spectrogram and reconstructs the waveform using an inverse STFT module that is fully differentiable. Similarly, iSTFTnet also adopts such a framework; however,

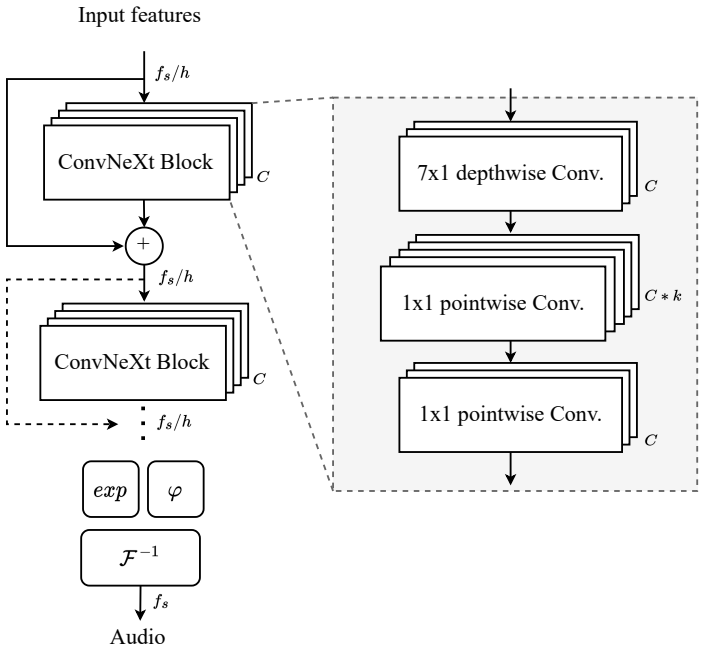


Figure 2.9: The structure of Vocos generator in [5].

it still employs the transposed convolution layers to upsample the data. Therefore, the generated speed is still limited. In contrast, the non-upsampling architecture of Vocos benefits from the fast inversion of STFT, allowing for the high-efficiency generation. Besides, such a framework also provides the interpretability from the spectrogram, which contains the physical meaning, allowing the model to focus on spectral coherence rather than waveform precision.

Generator Architecture: The generator architecture of Vocos consists of two main components: a neural spectrogram estimator and an iSTFT-based reconstruction module, as shown in Fig. 2.9. The spectrogram estimator is typically implemented using a stack of ConvNeXt [78], which maps a sequence of mel-spectrogram frames to a sequence of complex-valued STFT coefficients. To model the phase component more effectively, Vocos adopts a phase prediction strategy where the real and imaginary parts can be obtained from the estimated phases and amplitudes, allowing the network to learn harmonic phase relationships in a data-driven manner. Once the complex spectrogram is predicted, waveform reconstruction is performed by applying the inverse STFT using overlap-add synthesis.

Discriminator Design: Regarding the discriminator, compared to HiFi-GAN, a novel sub-discriminator is employed, i.e., the multi-resolution discriminator [6]. The discriminator consists of MPD and MSD, where the configuration of MPD is the same as that of HiFi-GAN, as shown in Fig. 2.10. MSD uses different sets of parameters, such as frame-shift and number of Fast Fourier Transform (nfft), to generate the spectrograms of generated speech in different time-frequency

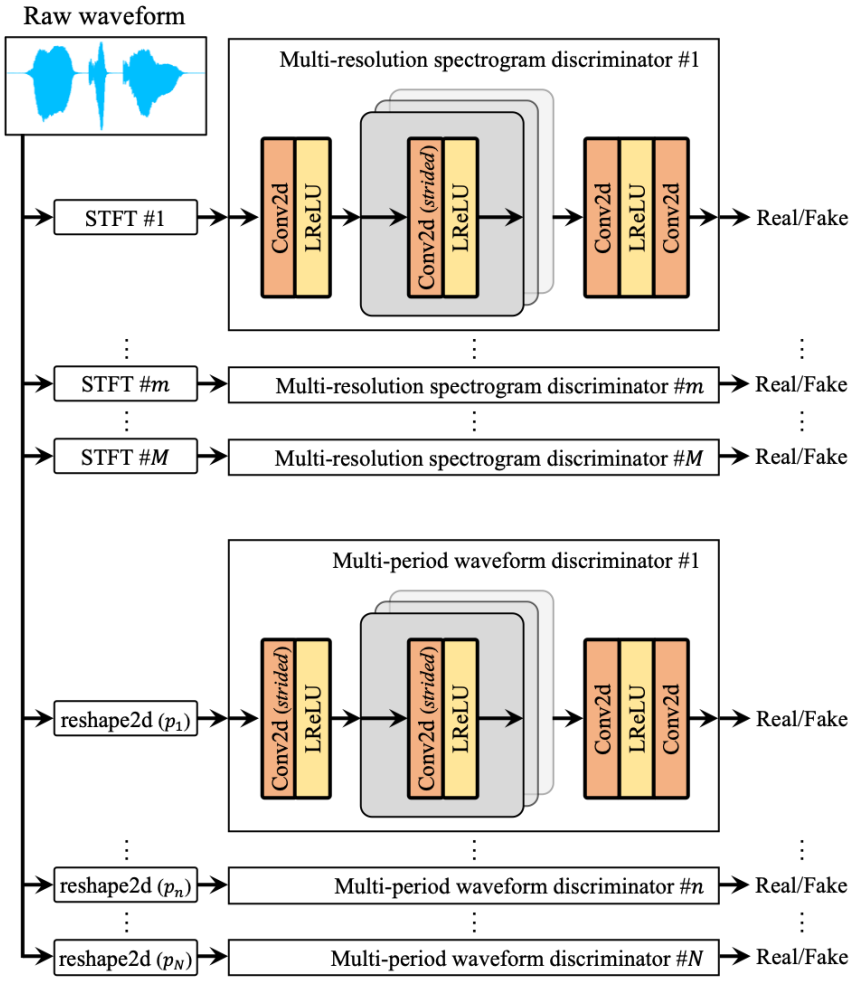


Figure 2.10: The structure of Vocos discriminators in [6].

resolutions and distinguish them from those of ground truth. In such a way, the spectral structure is more focused, and the generated speech will be perceptually better. Since the humans is much more sensitive to frequencies instead of the waveform in the time domain.

The configurations of training objectives are the same as those of HiFi-GAN.

Advantages and Limitations: One of the key contributions of Vocos is to reduce the burden on the neural network. Time-domain models like HiFi-GAN must implicitly learn waveform periodicity, harmonicity, and phase structure through deep upsampling layers, which are computationally expensive and prone to overfitting. In contrast, Vocos explicitly leverages the Fourier transform, allowing it to work in a domain where harmonic and noise components are more naturally separable. As a result, the generator only needs to model the spectral structure of speech, while the deterministic iSTFT ensures accurate waveform recovery. By getting rid of transposed convolutions, Vocos simplifies the inference flow, significantly accelerates inference speed, and achieves faster generation than HiFi-GAN.

However, Vocos still does not adopt a source-filter-based framework; thus, the controllable parameters are still not transparent, leading to the failure of the speech modification, such as pitch-scale modification and time-scale modification. Moreover, Vocos still suffers from the data-hungry issue, which means that Vocos needs a large amount of training data to avoid overfitting and ensure a good generalization ability. Although HiFi-GAN's generalization is highly data-dependent, apparently worse than Vocos, Vocos tends to degrade significantly in low-resource or out-of-distribution scenarios.

2.2.3 Hn-NSF: Harmonic plus noise Neural Source-Filter model

HiFi-GAN and Vocos have achieved impressive speech synthesis quality and speed. Since HiFi-GAN and Vocos have no input of the extra acoustic feature, such as f_0, to the neural network, it faces the challenging task of speech manipulation, especially f_0 extrapolation. To mitigate these challenges, neural vocoders that integrate conventional signal processing insights with the source-filter framework have been proposed, among which the Harmonic plus noise Neural Source-Filter model (hn-NSF) [7] stands out as a principled approach based on the classic source-filter theory of speech production.

Hn-NSF explicitly decomposes speech synthesis into two components: a harmonic source module that generates periodic excitation signals driven by the f_0, and a noise source module that captures stochastic components such as unvoiced segments and aperiodicities. These excitation signals are then shaped by a neural network-based filter module that models the vocal tract resonances. By explicitly incorporating this physical prior knowledge, hn-NSF reduces the difficulty of directly modeling complex waveforms and improves interpretability and controllability in synthesis.

Generator Architecture: The generator of hn-NSF consists of two parts. The source module generates an excitation signal conditioned on the input f_0 and the noise. Finally, a neural vocal tract filter network, often implemented as a stack of convolutional layers or residual blocks, filters the excitation signals and predicts the time-domain waveform. Fig. 2.11 shows the details of the hn-NSF generator. The acoustic feature is input into the condition module to generate the vocal tract filter, while the excitation signal is generated from the source module. Finally, the speech is generated with voicing flags.

Loss and Discriminator Design: Unlike some GAN-based vocoders that employ multi-scale or multi-period discriminators directly on waveforms, Hn-NSF often uses a multi-resolution spectral amplitude distance as the loss function to measure the distance between generated speech and the ground truth. However, with the development of the GAN-based methods, the discriminator and its auxiliary loss functions are also employed during the training. A straightforward way is to use the same configurations of discriminators and loss functions as HiFi-GAN (MSD + MPD).

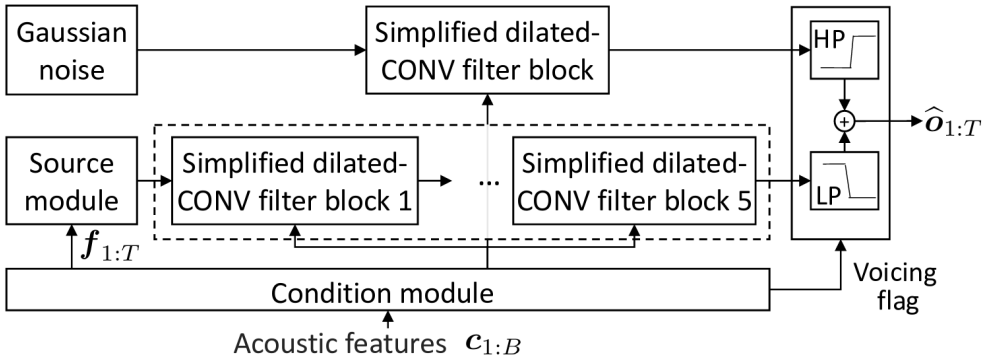


Figure 2.11: The structure of hn-NSF generator in [7].

Advantages and Limitations: A key advantage of Hn-NSF lies in its physically motivated source-filter decomposition, which reduces the complexity of learning by providing clear inductive biases. This results in improved interpretability and controllability: parameters such as f_0 and noise components can be explicitly manipulated for expressive synthesis or pitch-scale speech modification. Moreover, Hn-NSF often exhibits better robustness in low-resource or multi-speaker settings compared to fully end-to-end time-domain models.

However, hn-NSF remains inherently constrained by the black-box nature of neural networks, which inevitably limits its capacity for interpretable acoustic feature manipulation. For example, although hn-NSF supports pitch (f_0) modification, it does not explicitly extract or model amplitude information, which restricts its ability to control loudness. Moreover, the filter module in hn-NSF is implemented as a fully data-driven, non-transparent neural component, offering no guarantee of accurate or stable spectral envelope modeling. As a result, the overall synthesis quality may suffer, particularly under large pitch-scale transformations. These limitations are evident in the experimental analyses presented in Chapters IV and V.

Therefore, it becomes essential to explore a more interpretable neural vocoder framework that not only retains the flexibility and robustness of deep learning models but also enables fine-grained and transparent control over acoustic features. Such a framework would significantly enhance both the performance and versatility of speech modeling, paving the way for more effective synthesis, modification, and voice transformation applications.

2.2.4 Limitations of Neural Vocoders

While neural vocoders such as HiFi-GAN and Vocos have achieved impressive performance in terms of speech naturalness and synthesis speed, several inherent limitations persist. These limitations come from their data-driven nature, network structure, lack of interpretability, and difficulty in explicit control over acoustic features. In the following part, we discuss the key challenges faced by current neural vocoders.

Requirement of Extensive Training

Conventional vocoders, such as QHM methods, WORLD, and STRAIGHT, are built on signal processing, even the established principles of speech production in source-filter vocoders. These models typically rely on signal processing algorithms and do not require learning from data. In contrast, neural vocoders are entirely data-driven and must be trained on large-scale paired datasets of acoustic features and corresponding waveforms. The quality of a neural vocoder is thus highly dependent on the size, diversity, and quality of its training corpus. Without sufficient data, especially across speakers or prosodic conditions, the performance of neural vocoders degrades. In contrast, conventional vocoders offer greater modularity for a system and are more deployable in practical engineering applications.

Poor Generalization Ability

As discussed above, neural vocoders need sufficient speech data to learn. Moreover, neural vocoders often struggle to generalize satisfactory speech that is unseen during the training. In general, neural vocoders demand substantial amounts of training data; otherwise, they easily suffer from overfitting. For example, although HiFi-GAN is able to generate speech of great high quality, it has shown difficulty in the generation of speech unseen during the training. Even extrapolating f_0 values beyond the range of the training data is hard for HiFi-GAN, leading to poor synthesis in high-pitch singing voices. Similarly, Vocos, while leveraging the STFT framework to ease spectral modeling, relies on neural networks to predict the amplitude and phase of the complex spectrogram, and these predictions are also biased by the training distribution. The same limitations happen in hn-NSF. As a result, neural vocoders are prone to overfitting and may generate distortions or unnatural speech when presented with unseen inputs, even out-of-distribution inputs. A straightforward way to overcome this issue is to prepare sufficient data for training, i.e., including different languages, different genders, and different speakers. However, constructing such a large, clean, and effective speech dataset often requires considerable human and material resources, involving careful recording, manual annotation, and quality control, all of which are time-consuming and costly. As a result, obtaining large-scale, high-fidelity speech data suitable for robust modeling remains a major challenge. Thus, another way to alleviate this problem is to try to improve the generalization ability of neural vocoders.

Lack of Interpretability

One of the most significant limitations of neural vocoders is their black-box nature. Unlike conventional vocoders, where each parameter (e.g., pitch, formant frequency, amplitude, and phase) has a clear acoustic meaning, the internal activations of neural networks are often opaque. Although components, such as the MRF module in HiFi-GAN, attempt to represent the meaningful waveforms and decompose the modeling task into interpretable stages, the learned representations

themselves remain abstract. Fortunately, vocoders such as Vocos, which combine conventional transformations (STFT), can generate complex spectrograms relatively well to initially reveal the structure of the speech signal, while hn-NSF absorbs the f_0 to achieve pitch control of the synthesized speech. However, the effect is still not significant. This lack of interpretability makes it difficult to achieve the goal of speech modeling. The intrinsic models or components for speech cannot be extracted; therefore, the structure of the speech is still not transparent. This hinders the exploration of speech.

Absence of Explicit, Editable Parameters

Conventional vocoders allow for explicit control over acoustic features such as pitch, duration, or timbre, enabling flexible manipulation of speech. For instance, pitch can be shifted by scaling f_0, and time duration can be altered by modifying segment boundaries or interpolating the features. This enables conventional methods to be applied in various practical engineering tasks, such as pitch modification in singing or slowing down speech to improve intelligibility. In contrast, neural vocoders typically take mel-spectrograms as input and directly produce waveform samples. Even Vocos estimates the spectrogram using the opaque networks. These intermediate features do not retain parameters that correspond to intuitive aspects of speech. Although the source-filter-based neural vocoders, such as hn-NSF, also input the f_0 to control the intonation of the speech, the performance is still limited due to their black box nature. As a result, it is challenging to perform freewheeling edits, such as pitch modification or time-stretching, since there are no accessible control handles. Particularly, the f_0 extrapolation is in great need of human communication, for instance, the application in speaking aids for patients with laryngeal illness. The most neural vocoders fail to achieve this task. Although some attempts have been made to incorporate f_0 or prosody embeddings into the model, i.e., some neural vocoders are built on the basis of the source-filter structure, the common weak point of neural models in generalization ability hinder them from achieving a satisfactory performance. Since such controls are often effective only within the statistical bounds of the training data and do not guarantee interpretability or reliability.

2.3 Summary

This chapter shows the details and its limitations of some representations of conventional vocoders and neural vocoders.

For the conventional vocoders, the QHM and its extensions, i.e., aQHM and eaQHM, are fully introduced and discussed in terms of advantages and disadvantages. QHM methods has the capacity of modeling the structures of speech signals, enabling humans to manipulate the speech. Specifically, the speech signals can be modeled as the sum of several sinewaves with their amplitudes, frequencies, and phases accurately extracted. Moreover, although the analysis part (modeling the speech waveforms to the acoustic features) is time-consuming, the synthesis part is ef-

ficient, which generates the speech using only the framewise amplitudes, frequencies, and phase, and can be applied to achieve real-time speech generation. Unfortunately, they are fundamentally limited by their dependence on accurate pitch (f_0) estimation, which is often compromised by local non-stationarities in speech. Even small pitch errors can cause significant frequency mismatches in higher harmonics, degrading spectral precision. Although iterative refinement (e.g., eaQHM) can reduce these mismatches, convergence is highly sensitive to the initial frequency error and local frequency spacing. Large deviations can lead to biased LS estimates that reinforce incorrect frequencies, creating a feedback loop of error amplification. Furthermore, QHM operates in a framewise manner, where modeling performance is tightly coupled to window design and frame-shift configuration. Poor choices in these parameters can compromise both intra-frame estimation and inter-frame continuity, ultimately degrading synthesis quality.

Neural vocoders, on the other hand, have greatly advanced speech synthesis quality but face their own challenges. They utilize a large amount of data to achieve a great generalization ability, which is usually greater than conventional methods. Besides, a sophisticatedly designed structure only needs several simple computations, such as addition and convolutions, to efficiently estimate the output. This undoubtedly accelerates the speech synthesis if such a neural network can be employed. On the contrary, they face limitations that include heavy dependence on large, diverse training datasets, limited generalization to unseen conditions, lack of interpretability, and the difficulty in controlling acoustic features such as pitch and timbre. Such drawbacks hinder their adaptability and controllability in fine-grained speech synthesis tasks.

In essence, QHM offers interpretability and explicit parameter control but struggles with robustness and continuity, while neural vocoders offer high-quality synthesis but at the cost of transparency and controllability. These complementary limitations motivate the development of hybrid or alternative models that combine the interpretability and precision of signal models with the generative power of neural networks. As a preliminary step toward such integration, the next section investigates whether QHM can be combined with the backpropagation method, allowing us to examine the feasibility of bridging explicit quasi-harmonic modeling with data-driven optimization. This exploratory attempt lays the groundwork for the more comprehensive hybrid vocoder framework developed in Chapter IV.

Chapter 3

Backpropagation-based Quasi-Harmonic Modeling

3.1 Introduction

Building on the discussion in Chapter II, this chapter explores the integration of the backpropagation (BP) method into the QHM framework as a means to overcome the limitations of existing QHM methods. Specifically, conventional QHM methods suffer from large frequency mismatches that deteriorate performance, as well as framewise modeling strategies that compromise continuity and accuracy in speech representation. By reformulating QHM within a differentiable framework, we aim to investigate whether explicit sinusoidal modeling can benefit from data-driven optimization, thereby setting the stage for the more comprehensive hybrid model developed in the next work.

First, a spectrogram-based frequency refinement algorithm is introduced to enhance frequency correction performance. Rather than relying on least-squares optimization results, this method leverages a fixed spectrogram extracted from the speech waveform to refine frequency estimates. An iterative correction mechanism is employed to progressively align the frequency estimates with the ground truth. This approach is particularly effective in handling nonstationary signals characterized by strong modulations in both frequency and amplitude.

Second, backpropagation is applied to the QHM synthesis process to obtain more accurate estimates of complex amplitudes and frequencies. Unlike conventional frame-by-frame forward estimation, the proposed approach defines a sequence-level waveform loss function to quantify the discrepancy between the synthesized and reference waveforms. Gradients are propagated through the entire quasi-harmonic synthesis pipeline, enabling the frame-wise parameters to be updated in a global and coherent manner. To promote effective convergence and maintain a balance between frequency and amplitude estimation, two separate optimizers are employed, each tailored to its respective parameter type. This coordinated optimization strategy allows both parameter sets to converge stably and complementarily toward their true values.

Third, to improve the convergence speed and optimization efficiency, a novel loss function is designed. This function directly generates a time–frequency representation (TFR) from the parameters under optimization, namely the complex amplitudes, and compares it with the TFR derived from the ground-truth waveform. By circumventing both waveform reconstruction and subsequent TFR computation, the proposed loss function reduces computational complexity and encourages the estimated amplitudes and frequencies to more accurately capture the underlying vocal tract characteristics.

To assess the effectiveness of the proposed method, a series of experiments are conducted on real speech utterances. The results show that the BP-based QHM approach significantly improves speech resynthesis quality, yielding the highest signal-to-reconstruction accuracy and the lowest mel-cepstral distortion. These findings demonstrate the high potential of QHM-based synthesis in speech modeling and highlight its differentiable nature, making it a promising candidate for integration with neural network-based architectures.

3.2 Frequency Correction based on Spectrogram

First, we begin by addressing the primary limitation of QHM-based approaches, namely their relatively low accuracy in frequency estimation. This limitation stems from the intrinsic structure of the QHM algorithm, wherein instantaneous frequency is not computed directly, but rather inferred from the estimated complex amplitudes of harmonic components. Specifically, QHM expresses a signal as a superposition of AM–FM components, and computes the instantaneous frequency via the phase derivatives of the complex envelope. As a consequence, any noise, error, or instability present in the estimation of these complex envelopes will propagate into the frequency domain, thereby degrading the reliability and precision of the estimated instantaneous frequencies. This sensitivity becomes particularly problematic in cases involving nonstationary signals or strong frequency modulation, which are common in natural speech.

To mitigate this issue, we propose a novel frequency refinement strategy that bypasses the complex amplitude estimation step entirely. Instead of estimating frequency indirectly, we directly correct the initial frequency estimates by leveraging the local structure of the time–frequency representation (TFR), specifically the Short-Time Fourier Transform (STFT). The core idea is to exploit the fact that, in the STFT domain, each harmonic component manifests as a localized peak, and the frequency of each component can be refined by analytically locating the center of its corresponding spectral blob. This approach is not only theoretically justified under reasonable signal assumptions, but also practically effective in improving the accuracy of frequency estimation. In the following part, we give the specific derivation process of this algorithm and prove its effectiveness even for nonstationary signals with strongly modulated frequencies.

We use a simulated signal to demonstrate the details of the proposed refinement method. Con-

sidering the STFT of a signal, $x \in L^2(\mathbb{R})$, it can be expressed with a phase shift $e^{i\omega t}$ as

$$S_x(t,\omega) = \int_{\mathbb{R}} x(u)g(u-t)e^{-i\omega(u-t)}\mathrm{d}u, \tag{3.1}$$

where $g(\cdot)$ is the moving window and $g \in L^2(\mathbb{R})$. For the deterministic part of the speech, Eq. (3.1) can be rewritten as

$$S_x(t,\omega) = \sum_{k=-K}^{K} \int_{\mathbb{R}} x_k(u)g(u-t)e^{-i\omega(u-t)}\mathrm{d}u, \tag{3.2}$$

which shows that each component can be considered a well-separated intrinsic mode-type component if an appropriate window is chosen to prevent the main lobes of adjacent harmonics from overlapping. Therefore, assuming that the amplitude and frequency of $x_k(u)$ are weakly modulated, Eq. (3.2) can be approximated according to Parseval's Theorem as

$$\begin{aligned} S_x(t,\omega) &= \frac{1}{2\pi} \sum_{k=-K}^{K} \int_{\mathbb{R}} \hat{X}_k(\xi)\hat{G}(\xi-\omega)e^{i\xi t}\mathrm{d}\xi \\ &\approx \sum_{k=-K}^{K} A_k(t) \int_{\mathbb{R}} \delta(\xi-\omega_k)\hat{G}(\xi-\omega)e^{i\xi t}\mathrm{d}\xi \\ &= \sum_{k=-K}^{K} A_k(t)\hat{G}(\omega_k-\omega)e^{i\omega_k t}, \end{aligned} \tag{3.3}$$

where ω_k is the angular frequency, measured in radians per second and $\omega_k = 2\pi f_k(t)$. $\hat{G}(\cdot)$ and $\hat{X}_k(\xi)$ denote the Fourier transform of the window and the k-th component, respectively, where $\hat{X}_k(\xi) \approx 2\pi A_k(t)\delta(\xi-\omega_k)$ is considered in the equation. This illustrates that the TFR of a harmonic signal is composed of multiple Fourier transforms of window functions concentrated in the trajectories $\omega_k = 2\pi f_k(t)$. Since all components are well separated, here we only use the k-th component

$$S_{x_k}(t,\omega) = A_k(t)\hat{G}(\omega_k-\omega)e^{i\omega_k t}, \tag{3.4}$$

as an example to derive the method. In this paper, the Gaussian function is chosen to represent the window function $g(t) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{t^2}{2\sigma^2}}$, whose Fourier transform is $\hat{G}(\omega) = e^{-\frac{\sigma^2\omega^2}{2}}$. Therefore, the STFT of the k-th component gives

$$\begin{aligned} S_{x_k}(t,\omega) &= A_k(t)\hat{G}(\omega_k-\omega)e^{i\omega_k t} \\ &= A_k(t)e^{-\frac{\sigma^2(\omega_k-\omega)^2}{2}}e^{i\omega_k t} \\ &= A_k(t)e^{-\frac{\sigma^2(\omega-\omega_k)^2}{2}+i\omega_k t}. \end{aligned} \tag{3.5}$$

Let us examine the frequency axis. The k-th component can be interpreted as a complex Gaussian distribution. Although the shape of the Gaussian function (i.e., its width and height) varies with

the parameter σ, its overall contour remains consistent. This observation inspires the idea that the distance from any point on the ω-axis to the center of the Gaussian distribution can be determined, according to the specific "bell shape" of the window. Accordingly, the partial derivative of $S_{x_k}(t,\omega)$ with respect to ω can be computed as

$$\begin{aligned}\frac{\partial S_{x_k}(t,\omega)}{\partial \omega} &= A_k(t)\frac{\partial\left[e^{-\frac{\sigma^2(\omega_k-\omega)^2}{2}}\right]}{\partial \omega}e^{i\omega_k t}\\ &= A_k(t)\sigma^2(\omega_k-\omega)e^{-\frac{\sigma^2(\omega-\omega_k)^2}{2}+i\omega_k t}.\\ &= \sigma^2(\omega_k-\omega)S_{x_k}(t,\omega)\end{aligned} \tag{3.6}$$

Defining the distance from the point to the center of the Gaussian function as Δ_{ω_k}, we have

$$\Delta_{\omega_k} = \omega-\omega_k = -\frac{\partial S_{x_k}(t,\omega_k^{\text{init}})/\partial\omega}{\sigma^2 S_{x_k}(t,\omega)}. \tag{3.7}$$

This indicates that the distance from any arbitrary bin to the center of the Gaussian function is known. Thus, assuming that the frequency detected by the pitch detector ω_k^{init} is located within the k-th Gaussian window lobe, accordingly, the distance from ω_k^{init} to the center can be determined and adding this distance to ω_k^{init} yields the accurate frequency ω_k^{refine} as

$$\omega_k^{\text{refine}} = \omega_k^{\text{init}} - \Delta_{\omega_k} = \omega_k^{\text{init}} + \frac{\partial S_{x_k}(t,\omega_k^{\text{init}})/\partial\omega}{\sigma^2 S_{x_k}(t,\omega_k^{\text{init}})}. \tag{3.8}$$

In this way, the frequencies can be corrected.

As we all know, speech signals are usually strongly frequency-modulated, meaning that the frequencies of speech signal components are time-varying, even within a frame, and Eq. (3.3) is unsatisfied. To prove the effectiveness of frequency update in Eq. (3.8) for strongly frequency-modulated signals, we analyze a chirp signal with a fixed amplitude A:

$$x_0(t) = Ae^{i\varphi_0(t)}, \varphi_0^{(n)} = 0 \text{ for } n \geq 3, \tag{3.9}$$

where n is the derivative order. The Taylor expansion of the exponential part at time t can be expressed as

$$x_0(u) = Ae^{i\left[\varphi_0(t)+\varphi_0'(t)(u-t)+\frac{1}{2}\varphi_0''(t)(u-t)^2\right]}. \tag{3.10}$$

Substituting Eq. (3.10) into Eq. (3.1), we have

$$
\begin{aligned}
S_{x_0}(t,\omega) &= \int_{-\infty}^{+\infty} x_0(u)g(u-t)e^{-i\omega(u-t)}du \\
&= \int_{-\infty}^{+\infty} x_0(t+u)g(u)e^{-i\omega u}du \\
&= \int_{-\infty}^{+\infty} Ae^{i\left[\varphi_0(t)+\varphi_0'(t)u+\frac{1}{2}\varphi_0''(t)u^2\right]}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{u^2}{2\sigma^2}}e^{-i\omega u}du \\
&= \frac{A}{\sqrt{2\pi}\sigma}e^{i\varphi_0(t)}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left[\frac{1}{\sigma^2}-i\varphi_0''(t)\right]u^2-i[\omega-\varphi_0'(t)]u}du \\
&= \frac{A}{\sqrt{2\pi}\sigma}e^{i\varphi_0(t)}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}D(t)u^2+E(t,\omega)u}du \\
&= \frac{A}{\sqrt{2\pi}\sigma}e^{i\varphi_0(t)+\frac{E(t,\omega)^2}{2D(t)}}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}D(t)\left[u-\frac{E(t,\omega)}{D(t)}\right]^2}du \\
&= \frac{A}{\sqrt{2\pi}\sigma}\sqrt{\frac{2\pi}{D(t)}}e^{i\varphi_0(t)+\frac{E(t,\omega)^2}{2D(t)}} \\
&= \frac{A}{\sqrt{1-i\sigma^2\varphi_0''(t)}}e^{i\varphi_0(t)-\frac{[\omega-\varphi_0'(t)]^2}{2\left[\frac{1}{\sigma^2}-i\varphi_0''(t)\right]}}
\end{aligned}
\tag{3.11}
$$

where $D(t)=\frac{1}{\sigma^2}-i\varphi_0''(t)$, $E(t,\omega)=-i[\omega-\varphi_0'(t)]$, $\int_{-\infty}^{+\infty}e^{-\frac{1}{2}D(t)\left[u-\frac{E(t,\omega)}{D(t)}\right]^2}du=\sqrt{\frac{2\pi}{D(t)}}$ is consider [79]. Considering $\tilde{A}=\frac{Ae^{i\varphi_0(t)}}{\sqrt{1-i\sigma^2\varphi_0''(t)}}$ and $\tilde{\sigma}^2=1/\left[\frac{1}{\sigma^2}-i\varphi_0''(t)\right]$, Eq. (3.11) can be considered a Gaussian function as

$$
\begin{aligned}
S_{x_0}(t,\omega) &= \frac{A}{\sqrt{1-i\sigma^2\varphi_0''(t)}}e^{i\varphi_0(t)-\frac{[\omega-\varphi_0'(t)]^2}{2\left[\frac{1}{\sigma^2}-i\varphi_0''(t)\right]}} \\
&= \tilde{A}e^{-\frac{\tilde{\sigma}^2[\omega-\varphi_0'(t)]^2}{2}}
\end{aligned}
\tag{3.12}
$$

which is also a complex Gaussian function. Hence, the partial derivative of $S_{x_0}(t,\omega)$ with respect to ω can be computed as

$$
\begin{aligned}
\frac{\partial S_{x_0}(t,\omega)}{\partial\omega} &= \tilde{A}\frac{\partial\left[e^{-\frac{\tilde{\sigma}^2[\omega-\varphi_0'(t)]^2}{2}}\right]}{\partial\omega} \\
&= \tilde{A}e^{-\frac{\tilde{\sigma}^2[\omega-\varphi_0'(t)]^2}{2}}\tilde{\sigma}^2[\varphi_0'(t)-\omega] \\
&= S_{x_0}(t,\omega)\tilde{\sigma}^2[\varphi_0'(t)-\omega].
\end{aligned}
\tag{3.13}
$$

Subsequently, the distance from the point to the center of the Gaussian function (the ground truth of frequency) can be computed by

$$
\varphi_0'(t)-\omega=\frac{\partial_\omega S_{x_0}(t,\omega)}{\tilde{\sigma}^2 S_{x_0}(t,\omega)}. \tag{3.14}
$$

Finally, the frequency of the signal can be estimated on the basis of any point by adding the distance from that point to the center, as

$$\varphi_0'(t) = \omega - \Delta_{f_0} = \omega + \frac{\partial_\omega S_{x_0}(t,\omega)}{\tilde{\sigma}^2 S_{x_0}(t,\omega)}, \tag{3.15}$$

where Δ_{f_0} is the distance from the bin to the frequency of $x_0(u)$. As discussed previously, voiced speech can be characterized as a quasi-harmonic signal, where the instantaneous frequencies of the harmonics evolve continuously over time. This time-varying harmonic nature allows the speech signal to be effectively approximated by a superposition of multiple chirp-like components, each occupying distinct and well-separated frequency bands. Owing to this structure, it becomes feasible to perform a refinement process on the estimated frequencies of individual harmonic components to improve their alignment with the underlying signal characteristics.

To demonstrate the effectiveness of the proposed frequency refinement strategy, we present a comparison between the initial frequency estimates and their refined counterparts in Fig. 3.1. Specifically, the black frequency trajectories shown in Fig. 2.2 are refined through the refinement algorithm, resulting in the more accurate frequency curves depicted in Fig. 3.1. As illustrated, each harmonic frequency is realigned such that it coincides with the peak of its corresponding Gaussian distribution, which models the uncertainty or dispersion in the frequency estimation.

By adopting the refined frequencies as initialization for subsequent optimization, such as gradient descent, it is possible to significantly enhance both the convergence speed and the final estimation accuracy. This is because starting from values that are closer to the true solution reduces the number of iterations required and mitigates the risk of convergence to suboptimal local minima. Therefore, the refinement process serves as an essential step toward more precise and efficient harmonic parameter extraction.

To further illustrate the performance of the proposed frequency refinement method in correcting the frequency of modulated signals (frequency-modulated), we use such a spectrogram-based refinement method to correct the frequency in different noisy cases. The same simulated signal in Eq. (2.34) is employed to verify the performance, which is corrupted by different levels of noise. The results are plotted in Figs. 3.2 and 3.3. It can be easily observed that the proposed method can effectively correct the frequencies in Fig. 3.2. Even though the frequencies of the signal vary rapidly over time, the frequency of each component can be located at the peak of the corresponding Gaussian distribution. Fig. 3.3 shows the performance in the extremely noisy condition, demonstrating that the frequency of each component can be accurately relocated at the peak of the corresponding Gaussian distribution, even when the frequency mismatch is large, making the initial frequency located in the main lobe of the adjacent component.

This is due to a two-stage refinement strategy: first, the pitch initially detected by the pitch detector will be corrected independently to reduce the pitch mismatch. Second, the frequencies of individual harmonic components will be calculated by multiplying the order of the harmonic

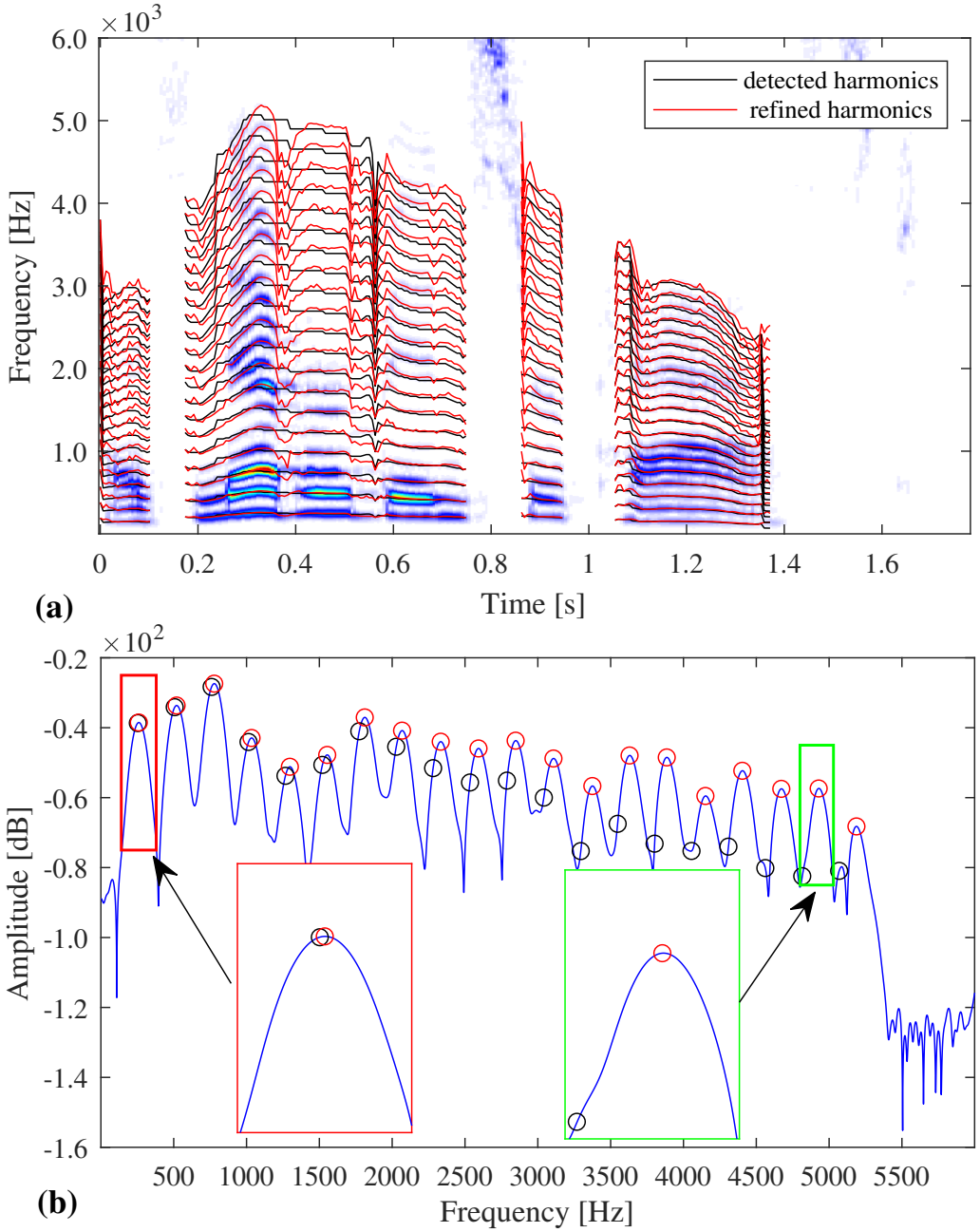


Figure 3.1: (a) STFT of a speech signal and the refined frequencies, and (b) sectional view at $t = 0.34$ sec. In (b), the black dots represent the detected frequencies by the pitch detector, while the red dots indicate the refined frequencies. The mismatches for each harmonic are reduced, locating the frequencies at their corresponding peaks.

components. Then, the frequency of the individual harmonic component will be refined through the spectrogram again. In such a way, the pitch mismatch will be preliminarily reduced to avoid the increment of frequency mismatch of the individual frequency, as shown in Eq. (2.33). Then, the frequency mismatch of the individual harmonic component can be easily estimated for the correction. Besides, if the frequency is not linearly modulated but in a higher order, the proposed spectrogram can still correct the frequency accurately with an iterative strategy, i.e., the corrected

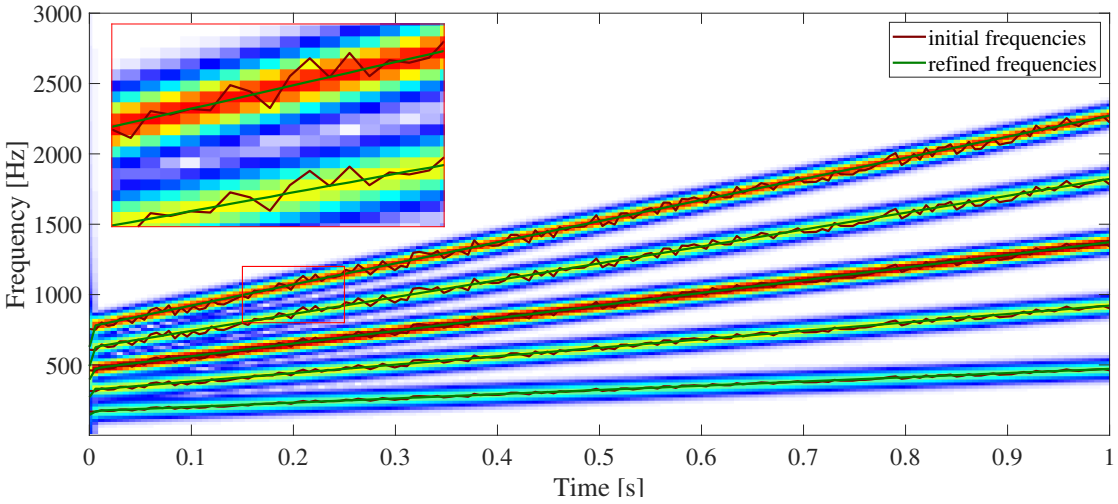


Figure 3.2: Frequency estimates by proposed refinement method using initial f_0 whose mismatch $\in [-10, 10]$.

frequency can be considered as the initial frequency of the next iteration. Thus, after the iterations, the frequency will be gradually approximated towards the peak of the corresponding Gaussian distribution, namely the ground truth of the frequency.

It is worth noting that even in such iterative or two-stage cases, the spectrogram $(S_{x_k})(t,\omega)$ and its partial derivative with respect to ω ($\partial_\omega S_{x_k}(t,\omega)$) only needs to be computed only once. Therefore, the efficiency of the proposed method is higher than that of QHM methods, since QHM methods need to correct the frequency according to the result of LS. LS is time-consuming, implying that the iterative frequency correction of QHM methods is extremely time-consuming, which is unsuitable to be applied in practical applications. In contrast, once the two complex spectrogram are computed, the frequency can be iteratively corrected by Eq. (3.8) without any extra heavy computation.

3.3 Parameter Refinement based on Backpropagation

In this section, we focus on addressing the second limitation discussed earlier, namely, the degradation of resynthesis quality caused by framewise modeling. With the rapid advancement of deep learning techniques, backpropagation (BP) has become a fundamental tool for optimizing model parameters. BP works by computing the gradients of a loss function that quantifies the discrepancy between the generated output and the ground truth, and then updating the parameters accordingly. Inspired by this paradigm, we extend the notion of trainable parameters to include the complex amplitudes and instantaneous frequencies used in signal modeling. Rather than estimating these parameters independently for each frame, we jointly optimize the complex amplitude and frequency of each harmonic component across all frames using gradient-based optimization.

To achieve this, we compute the gradients of a loss function defined on the entire waveform

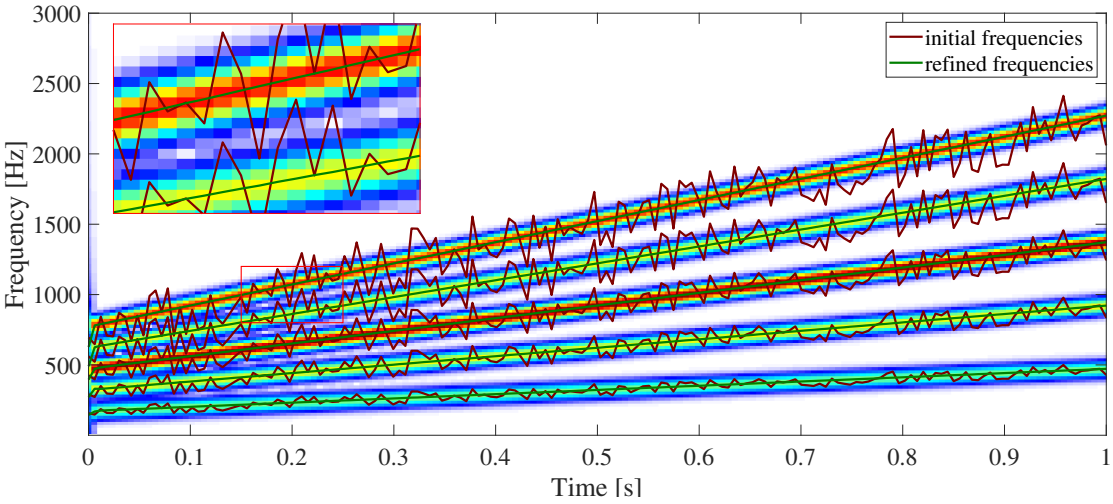


Figure 3.3: Frequency estimates by proposed refinement method using initial f_0 whose mismatch $\in [-45, 45]$.

sequence, rather than on individual frames, and backpropagate the errors through the synthesis process. This sequence-wise modeling framework allows for holistic parameter adjustment, thereby overcoming the inherent discontinuities and inconsistencies introduced by framewise estimation. However, implementing this approach necessitates addressing several key considerations, as outlined below.

3.3.1 Differentiability of the Synthesis Process

For the gradients computed by the loss function to be effectively backpropagated, it is essential that the synthesis process be differentiable with respect to the parameters to be optimized. Specifically, the synthesized speech signal must be differentiable with respect to both the complex amplitude and the instantaneous frequency. In this work, we adopt the synthesis process previously employed in QHM, which includes both deterministic and stochastic components. Notably, this synthesis process incorporates a phase compensation algorithm that enhances robustness against errors in parameter estimation.

Here, we focus on the differentiability of the entire synthesis process. Noting that the input of the synthesis process is the framewise complex amplitudes and frequencies, the output is the sequence-wise speech waveform. To explore the differentiability, we start from the speech waveform. Eq. (2.25) shows that the speech waveform is the sum of harmonics, showing that the speech waveform is differential with respect to individual instantaneous amplitudes and phases. Eq. (2.30) shows that the individual instantaneous phases is differential with respect to individual frequencies. Subsequently, thanks to the use of cubic interpolation for frequency trajectories and linear interpolation for amplitude evolution, the individual instantaneous amplitudes and frequencies are differential with respect to their own framewise versions. Then, Eq. (2.27) indicates that

the framewise amplitudes and phases are obtained from the complex amplitudes, proving that the framewise amplitudes and phases are differential with respect to framewise complex amplitude. Consequently, the entire synthesis pipeline is differentiable, which implies that the gradients of the loss function with respect to framewise parameters, i.e., the complex amplitude a_k^l and frequency f_k^l, can be obtained via standard backpropagation.

3.3.2 Design of an Appropriate Loss Function

Once the differentiability of the synthesis process has been established, the next step is to define an appropriate loss function for guiding the optimization. Departing from conventional framewise analysis methods such as QHM, we adopt a sequence-wise loss that evaluates the discrepancy between the generated and reference signals over the entire waveform. Specifically, we define a waveform-domain loss function as

$$L_{\text{wave}} = f_{\text{wave}}[\hat{x}(t); x(t)], \tag{3.16}$$

where $f_{\text{wave}}[\cdot]$ denotes a suitable distance metric, and $x(t)$ and $\hat{x}(t)$ represent the ground truth and synthesized speech signals, respectively. This formulation enables the error information between adjacent frames to be jointly utilized, thereby facilitating consistent and coherent parameter optimization across frames.

It is well known that in optimization problems, the presence of a non-convex loss function often leads to the solution becoming trapped in local minima, thereby hindering convergence to the global optimum. This phenomenon poses a significant challenge, as it may result in suboptimal parameter estimation and degraded model performance. Consequently, a considerable body of research has been dedicated to the design and formulation of loss functions exhibiting convexity or quasi-convexity properties. Such convex loss functions are advantageous because they increase the likelihood that iterative optimization algorithms will converge to the global minimum, thereby enhancing the stability and reliability of the training process. In the following part, we also check the convexity of our loss function.

Beforehand, a well-known convexity theorem should be concerned.

Theorem 1. *A fundamental result in convex analysis states the following: if each function $f_i(x)$ is convex and the corresponding weights α_i are non-negative real numbers, i.e., $\alpha_i \geq 0$, then their weighted sum*

$$f(x) = \sum_i \alpha_i f_i(x) \tag{3.17}$$

is also convex. This result critically relies on both the convexity of the functions f_i and the positivity of weights α_i within the real domain.

In the waveform synthesis process considered here, the reconstructed waveform at time t is

expressed as (same as Eq. (2.25))

$$\hat{x}(t) = \sum_k A_k(t) e^{i\varphi_k(t)}, \tag{3.18}$$

where $A_k(t) > 0$ are instantaneous amplitudes (positive real numbers), but $e^{i\varphi_k(t)}$ are complex numbers on the unit circle in the complex plane, thus not positive real values.

Because the weights $A_k(t)$ multiply complex-valued terms $e^{i\varphi_k(t)}$, the overall summation

$$\sum_k A_k(t) e^{i\varphi_k(t)} \tag{3.19}$$

is a sum of complex numbers rather than a sum of non-negative real values. Therefore, the positivity condition on weights required by the convexity theorem does not hold.

As a consequence, even though each amplitude $A_k(t)$ is positive, the function $x \mapsto \hat{x}(t)$ is a nonlinear mapping involving complex exponentials, which are not convex functions over the real-valued phase parameters $\varphi_k(t)$. Although $f_{\text{wave}}[\cdot]$ can be selected as a convex function, it is still hard to ensure the convexity of the loss function with respect to framewise complex amplitudes and frequencies.

This argument demonstrates rigorously that, due to the inherent nonlinearity of the synthesis process, the waveform loss function does not satisfy the conditions for convexity, and hence poses challenges for optimization methods that rely on convexity assumptions, causing optimization algorithms to converge to poor local minima, particularly during the early stages of training. To mitigate this risk, it is desirable to introduce an auxiliary convex loss function that can stabilize the optimization trajectory. Convex loss functions can guide the parameters towards globally favorable regions of the solution space, thereby accelerating convergence and enhancing robustness.

Moreover, it is crucial that the loss function ensure consistency between the spectral characteristics of the synthesized and target signals. In this context, spectrogram-based loss functions have gained widespread popularity. However, recent studies [80, 81] have demonstrated that most conventional spectrogram-based losses used in neural speech processing are also non-convex with respect to the model parameters. This observation has spurred a wave of research aimed at addressing the limitations of spectrogram losses. For example, [82] examines how different loss configurations influence harmonic parameter estimation, while [83] explores the application of optimal transport to enable more meaningful comparisons between discrete spectral distributions.

In addition to the issue of non-convexity, standard spectrogram-based losses are computationally expensive, which can lead to slow convergence. To address both challenges, non-convexity and computational inefficiency, we are motivated to propose a novel loss function that operates directly on the model parameters, bypassing the need to generate the full waveform.

Recall from Eq. (2.1) that the absolute value of the complex amplitude at the frame center ($t = 0$) corresponds to the instantaneous amplitude. Meanwhile, Eq. (3.3) reveals that the peak of each

Gaussian window in the time-frequency domain aligns with the amplitude of the corresponding signal component at the frame center. This observation motivates us to construct a spectrogram-like representation directly from the parameters, leading to the definition of a parameter-based spectrogram loss:

$$L_{\text{spec}} = f_{\text{spec}}[\hat{M}_x(t_l, \omega); M_x(t_l, \omega)], \tag{3.20}$$

where $f_{\text{spec}}[\cdot]$ can be either the l^1 or l^2 norm, $M_x(t_l, \omega)$ denotes the ground-truth spectrogram magnitude at the center of frame l. $\hat{M}_x(t_l, \omega)$ is the corresponding estimated magnitude, which can be considered as the sum of several Gaussian windows distributed at all harmonic frequencies. It can be computed as

$$\begin{aligned} \hat{M}_x(t_l, \omega) = |\hat{S}_x(t_l, \omega)| &= \sum_{k=-K}^{K} \hat{A}_k(t_l) \hat{G}(\omega - \hat{\omega}_k) \\ &= \sum_{k=-K}^{K} \hat{A}_k(t_l) e^{-\frac{\sigma^2[\omega - \hat{\omega}_k]^2}{2}}, \end{aligned} \tag{3.21}$$

where $\hat{\omega}_k = 2\pi \hat{f}_k(t_l)$ representing the angular frequency of the k-th component at frame l.

Focusing on that $\hat{A}_k(t_l)$ is always larger than 0 ($\hat{A}_k(t_l) > 0$ holds forever) and the Gaussian window is also always larger than 0 ($e^{-\frac{\sigma^2[\omega - \hat{\omega}_k]^2}{2}} > 0$ holds forever), thus, L_{spec} is always convex with respect to $\hat{A}_k(t_l)$. Thus, this loss can be employed to increase the convexity of the entire loss function.

Importantly, this formulation avoids the synthesis process entirely, which significantly speeds up the computation of gradients during backpropagation. Moreover, the expression for $\hat{M}_x(t_l, \omega)$ is convex with respect to $\hat{A}_k(t_l)$, and when f_{spec} is chosen to be a convex function, the overall loss L_{spec} is also convex. This convexity plays a vital role in ensuring the stability and reliability of the optimization process. Thus, during the optimization, we use the proposed spectrogram loss and waveform loss as the entire loss function.

3.3.3 Alternating Backpropagation

During the training of neural models, optimization algorithms such as Adagrad [84] and Adam [85] are commonly employed to adaptively adjust the learning rate based on the magnitude and history of the gradients. These optimizers facilitate the simultaneous and coordinated updating of model parameters, ideally steering them toward their respective optimal values. A fundamental prerequisite for the effectiveness of such adaptive optimization techniques is that the parameters involved should possess comparable distributions and similar magnitudes. When this assumption is violated, the efficacy of the optimization process can be severely compromised.

In the context of speech parameter extraction, specifically the joint estimation of complex amplitudes and instantaneous frequencies, this issue becomes particularly pronounced. These two

types of parameters differ significantly in scale and statistical distribution: the complex amplitude typically exhibits relatively small and bounded values, whereas the frequency may vary over a much broader range. This disparity introduces substantial difficulties in optimization.

For instance, employing a unified learning rate across both parameter types can lead to conflicting behaviors. A relatively low learning rate may suffice for updating the complex amplitude toward its optimal value but results in the frequency component remaining stagnant. Conversely, a higher learning rate may enable effective frequency adjustment but causes the complex amplitude to overshoot, thereby failing to converge. Moreover, due to their inherently different distributions, these parameters are more susceptible to being trapped in saddle points or poor local optima during optimization.

Our empirical observations corroborate this analysis. Preliminary experiments reveal that the optimization trajectories of complex amplitude and frequency often diverge, with each becoming confined to distinct local optima. This divergence ultimately prevents the model from achieving a globally optimal solution and underscores the necessity for tailored optimization strategies that account for the unique characteristics of each parameter type.

3.4 BP-QHM Implementation

After introducing the ideas of the frequency refinement and the parameter refinement, in this section, we combine the ideas mentioned above to propose a new speech modeling method named backpropagation-based quasi-harmonic model (BP-QHM) to surmount the limitations of QHM methods, i.e., QHM methods correct the frequency inadequately, and their framewise process causes the degeneration of speech resynthesis. BP-QHM is designed to jointly address both frequency inaccuracy and temporal discontinuity by embedding gradient-based optimization within a differentiable synthesis pipeline. The framework leverages both deterministic signal modeling and modern deep learning tools to achieve high-fidelity speech reconstruction.

Fig. 3.4 shows the details of forward methods (QHM methods) and our BP-QHM. The workflow of BP-QHM has two steps:

- First, the pitch is estimated by the pitch detector and improved by the TFR-based refiner. This refinement ensures that the harmonic structure aligns better with the actual spectral peaks in the signal, effectively reducing frequency deviations that would otherwise distort the synthesized output.
- Second, the random values of complex amplitude and frequency are initialized and subsequently optimized by gradient descent. The use of random initialization, followed by principled optimization, allows BP-QHM to explore a wide parameter space and converge towards a global optimum that preserves both the harmonic content and the phase continuity of the speech waveform.

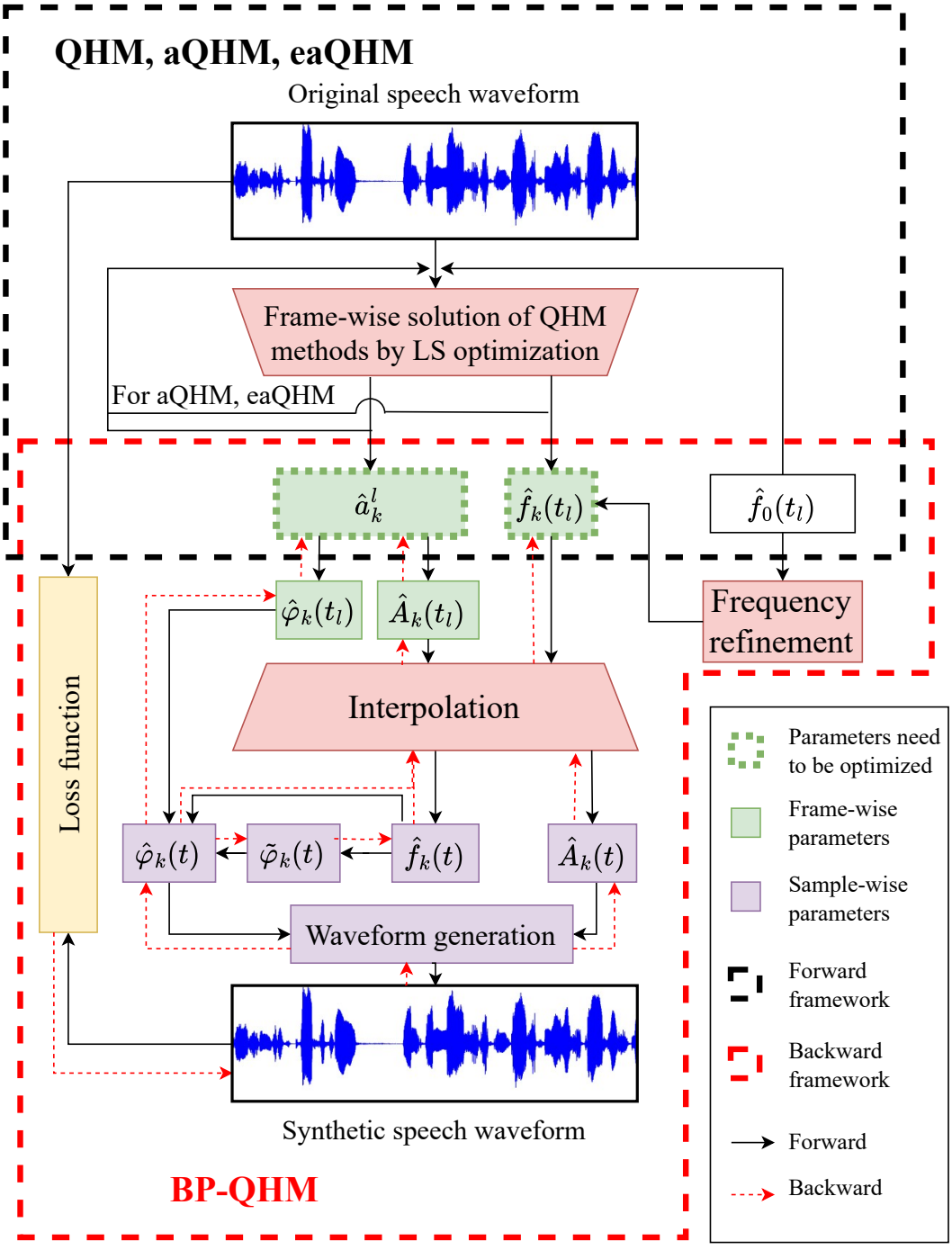


Figure 3.4: Workflows of QHM, aQHM, eaQHM, and BP-QHM. The black solid lines indicate the flow of QHM, aQHM, and eaQHM, including analysis and synthesis, whereas the red dotted lines indicate that the gradient of the loss function is propagated backward along the synthesis flow of QHM to the parameters being optimized, allowing for their adjustment and optimization.

In the first step, the initial framewise pitch $\hat{f}_0^{\text{init}}(t_l)$ ($l = 1, ..., L$) is detected by a pitch detector, such as YAAPT, YIN, and Harvest. This initial estimation serves as a coarse representation of the f_0 contour, and while it may capture general pitch trends, it is often insufficient for precise harmonic reconstruction. Then, the frequency should be improved by the refiner on the basis of Eq. (3.8). It is worth noting that $\partial_\omega S_x(t,\omega)$ is the partial derivative of $S_x(t,\omega)$. Therefore, a straightforward way to calculate the $\partial_\omega S_x(t,\omega)$ in a discrete way is to use the *diff*. However, in this way, the result will not align with $S_x(t,\omega)$ along the time axis. Thus, some works will use the original $S_x(t,0)$ to compensate for the result, inevitably increasing the computation error. Therefore, to reduce the computation error, we establish $\partial_\omega S_x(t,\omega)$ in a novel way, which directly uses the modified window function $g^t(t) = tg(t)$ to conduct the STFT, i.e.,

$$\begin{aligned}\frac{\partial S_x(t,\omega)}{\partial \omega} &= \int_{\mathbb{R}} x(t)g(u-t)\frac{\partial[e^{i\omega(u-t)}]}{\partial \omega}\mathrm{d}u \\ &= i\int_{\mathbb{R}} x(t)(u-t)g(u-t)e^{i\omega(u-t)}\mathrm{d}u \\ &= i\int_{\mathbb{R}} x(t)g^t(u-t)e^{i\omega(u-t)}\mathrm{d}u \\ &= iS_x^{g^t}(t,\omega). \end{aligned} \tag{3.22}$$

This approach improves numerical stability by incorporating time-weighted analysis, which is less sensitive to minor perturbations or window truncation effects.

Immediately, $\hat{f}_0^{\text{init}}(t_l)$ is regarded as $\hat{f}_0^{\text{in}}(t_l)$ and refined by calculating

$$\hat{f}_0^{\text{out}}(t_l) = \hat{f}_0^{\text{in}}(t_l) + \frac{iS_x^{g^t}[t_l, 2\pi\hat{f}_0^{\text{in}}(t_l)]}{2\pi\sigma^2 S_x[t_l, 2\pi\hat{f}_0^{\text{in}}(t_l)]}, \tag{3.23}$$

which effectively shifts the pitch value towards the center of energy in the frequency domain. This process is analogous to performing Newton-like updates in the spectral domain to achieve local alignment of signal energy.

As discussed before that an iterative strategy can be considered to obtain a more accurate result. Therefore, $\hat{f}_0^{\text{out}}(t_l)$ can be re-inputted into Eq. (3.23) for an iterative refinement, which acts similarly to the multi-instantaneous frequency estimator in [86]. Empirically, two iterations are sufficient. Using the refined pitch, we can obtain the individual frequencies as $\hat{f}_k^{\text{init}}(t_l) = k\hat{f}_0^{\text{out}}(t_l)$, and perform a masking on $\hat{f}_k^{\text{init}}(t_l)$ to remove the aliased frequencies[1]. This masking ensures that the synthesized signal remains physically meaningful and compliant with the Nyquist theorem. Then, the obtained individual frequencies will be corrected with the same process as that for pitch to approximate their corresponding ridges, namely, an iterative correction for each frequency. So far, such a two-stage and iterative frequency refinement has been completed to obtain the accurate frequencies for each quasi-harmonic component.

[1] In [87], the frequencies were not refined and aliased frequency components (i.e., frequencies exceeding the Nyquist limit) were inadvertently included. In this work, a mask is applied to ensure that only frequencies below the Nyquist limit are utilized during correction and backpropagation optimization.

After improving the frequency as the initial value of BP, the complex amplitude should be initialized. The amplitude is initialized with random values sampled from a uniform distribution or a Gaussian distribution, depending on the implementation. Subsequently, an optimizer for the frequency Opt_F with the learning rate α_F and another optimizer for the complex amplitude Opt_A with the learning rate α_A are employed to update their values iteratively. These optimizers are selected to support adaptive learning rates, allowing for flexible adjustments as the loss landscape evolves.

In this paper, f_wave and f_spec are defined as the l^2 norm to allow the gradients to adaptively self-adjust, i.e., reducing the gradients when approaching the optimum, which especially helps the frequencies to be updated in the unvoiced part while keeping those in voiced part fixed. The adoption of the l^2 norm ensures smooth gradients, which are critical for stable convergence in high-dimensional parameter spaces.

We suggest the use of a two-stage optimization to optimize the framewise complex amplitude $\hat{\mathbf{a}}$ and the individual frequency of each harmonic component $\hat{\mathbf{f}}$:

(1) Preliminary optimization (Consider $\hat{\mathbf{a}}$ as a variable and fix $\hat{\mathbf{f}}$ for a preliminary result.)

In the early stage, we use $\hat{f}_k^{\mathrm{out}}(t_l)$ as $\hat{\mathbf{f}}$ and fix it for the preliminary optimization. Then, the modulus of $\hat{\mathbf{a}}$, i.e., the real amplitude, will rapidly converge from a random initial value to the global optimum since we adopt the proposed convex spectrogram loss function, which also prevents the optimization from being trapped in the local optimum. This stage effectively calibrates the spectral envelope and sets a solid foundation for subsequent phase and frequency optimization.

(2) Alternate optimization.

(2.1) Consider $\hat{\mathbf{f}}$ as a variable and fix $\hat{\mathbf{a}}$.

Since the current frequencies and amplitudes are close to the ground truth, there is less concern about the result remaining in other unreasonable local optima. Therefore, the update of frequencies can be started with a low learning rate to further fit the unvoiced speech. This gradual refinement allows the model to delicately capture high-frequency stochastic patterns without introducing artifacts.

(2.2) Consider $\hat{\mathbf{a}}$ as a variable and fix $\hat{\mathbf{f}}$.

To promptly adjust the complex amplitude to synchronize with the frequency updates, an amplitude update is performed after each frequency update. This step allows the phase of each harmonic to be fine-tuned to match the updated instantaneous frequency, maintaining both temporal and spectral coherence.

Here, we specifically introduce the steps for optimization. We first use Opt_A to conduct the preliminary optimization for J_p iterations to obtain a preliminary result of $\hat{\mathbf{a}}$. This stage focuses

solely on refining the amplitude magnitudes without involving the phase or frequency variables, which helps to stabilize the training and provide a good initialization for the subsequent alternating optimization. Currently, we aim to optimize the modulus of complex amplitude to match the magnitude spectrogram of the ground truth. Thus, only the proposed spectrogram loss function $L_{\text{spec}}(\hat{\mathbf{a}})$ is employed in the optimization, avoiding getting trapped in the local optimum. The convexity of this loss function ensures that the optimization process converges to a global solution, which is particularly important in the early stage where the parameters are far from their true values.

After obtaining the gradient from the loss function, accordingly, the optimizer updates the complex amplitude as

$$\hat{\mathbf{a}}_{j+1} = \hat{\mathbf{a}}_j - \text{Opt}_{\text{A}}(\nabla L_{\text{spec}}(\hat{\mathbf{a}}_j); \alpha_{\text{A}}), \tag{3.24}$$

where j means the number of iterations and ∇ denotes the gradient. The learning rate α_{A} controls the step size of the update, and its value must be carefully chosen to ensure convergence without oscillation.

After the preliminary optimization, the energy of each harmonic can be determined. This energy initialization plays a crucial role in determining the relative importance of each frequency component and provides a stable reference for later phase and frequency optimization. Next, the phases of each component should be optimized to approximate the real waveform. Phase alignment is essential for capturing the temporal fine structure of the signal, which greatly affects perceptual quality. Afterward, to adjust the frequency and fine-tune the phase for matching the waveform, especially the unvoiced speech, which is not deterministic, we start applying the waveform loss function. The waveform loss enables direct comparison between the synthesized and original speech signals in the time domain, thus incorporating both phase and energy errors.

Therefore, on the basis of the improved $\hat{\mathbf{a}}$, the alternate optimization is conducted iteratively for several iterations to update $\hat{\mathbf{f}}$ and $\hat{\mathbf{a}}$ alternately. Alternating updates allow the model to iteratively adjust one set of parameters while holding the other fixed, which helps avoid interference between amplitude and frequency during gradient descent. In each iteration, we first optimize the frequency in step (2.1). As the proposed spectrogram loss is based on harmonic characteristics, which do not align with the unvoiced part, here we use the waveform loss function $L_{\text{wave}}(\hat{\mathbf{f}})$ to compute gradients of frequency and update the frequency as

$$\hat{\mathbf{f}}_{j+1} = \hat{\mathbf{f}}_j - \text{Opt}_{\text{F}}(\nabla L_{\text{wave}}(\hat{\mathbf{f}}_j); \alpha_{\text{F}}). \tag{3.25}$$

This stage is crucial for fitting the instantaneous frequency trajectory, particularly in unvoiced or noisy segments, where phase discontinuities or stochastic variations make spectral loss unreliable. By leveraging waveform-level loss, BP-QHM ensures time-domain accuracy without relying on unstable or undefined harmonics.

Immediately, in step (2.2), the gradient is computed on the basis of the updated frequency with the combination of waveform and spectrogram loss functions, i.e., $L_{\text{w+s}}(\hat{\mathbf{a}}) = L_{\text{spec}}(\hat{\mathbf{a}}) + L_{\text{wave}}(\hat{\mathbf{a}})$. This combination provides a comprehensive optimization objective, balancing spectral envelope accuracy with waveform fidelity. Then, the complex amplitude is optimized similarly to Eq. (3.24) to be updated with the latest frequency. In this way, the phase can be adaptively updated while maintaining the spectral characteristics. This back-and-forth updating ensures that both magnitude and phase are coherently adjusted, resulting in a speech reconstruction that is perceptually natural and spectrally consistent.

During the iterations, we pick the result corresponding to the minimum of $L_{\text{w+s}}$ as the output. This strategy acts as a safeguard against divergence and captures the best performance achieved throughout the entire optimization process. After iterative optimization, the complex amplitudes and frequencies converge to the optimum, with which the speech can be perfectly resynthesized in both voiced and unvoiced parts. The optimization process thereby achieves the joint refinement of all parameters in a holistic manner, maximizing reconstruction accuracy and preserving signal integrity.

Algorithm 1 provides the pseudocode for BP-QHM, encapsulating the entire process that has been outlined thus far. It serves not only as a practical guide for implementation but also as a summary of the theoretical framework described. Each step in the algorithm corresponds to a well-motivated mathematical operation, ensuring reproducibility and clarity for future researchers.

3.5 Experimental Evaluations

3.5.1 Experimental Design and Evaluation Aspects

To explore the performance and practical effectiveness of the proposed BP-QHM framework, we carry out a comprehensive evaluation across multiple dimensions. Specifically, the following key aspects are investigated:

(1) The availability and computational feasibility of the proposed spectrogram loss function.

Regarding the loss function, we conduct a direct comparison between the proposed spectrogram loss and the conventional magnitude-based spectrogram loss. This evaluation aims to validate whether the newly designed loss not only accelerates convergence but also enables the model to escape from suboptimal local minima by leveraging its convex nature. The computational cost of calculating each loss is also recorded to demonstrate the practicality of the proposed function when integrated into iterative gradient-based training.

(2) The time complexity and execution efficiency of BP-QHM compared with baseline QHM variants.

In terms of time complexity, we systematically measure the actual runtime of each method,

Algorithm 1 BP-QHM.

Step 1: Initialization and Preprocess

Input the target signal $x(t)$, the sampling rate f_{s}, the refinement number I_{f_0} and I_f, and the iteration numbers J and J_{p}, and compute $S_x(t,\omega)$ and $\partial_\omega S_x(t,\omega)$;

Choose the harmonic number K, the frame-shift h, the σ of Gaussian function g, and the learning rates α_{A} and α_{F};

Step 2: Detection and Refinement of Frequency

Detect the pitch $\hat{f}_0^{\mathrm{init}}(t)$ and set it as $\hat{f}_0^{\mathrm{in}}$;

for $i = 1 : I_{f_0}$ **do**

$\hat{f}_0^{\mathrm{out}}(t_l) \leftarrow \hat{f}_0^{\mathrm{in}}(t_l) + \frac{\partial_\omega S_x[t_l, \hat{f}_0^{\mathrm{in}}(t_l)]}{2\pi\sigma^2 S_x[t_l, \hat{f}_0^{\mathrm{in}}(t_l)]}$;

end for

Get the frequencies $\hat{f}_k^{in}(t_l) = k\hat{f}_0^{\mathrm{out}}(t_l)$;

for $i = 1 : I_f$ **do**

$\hat{f}_k^{\mathrm{out}}(t_l) \leftarrow \hat{f}_k^{\mathrm{in}}(t_l) + \frac{\partial_\omega S_x[t_l, \hat{f}_k^{\mathrm{in}}(t_l)]}{2\pi\sigma^2 S_x[t_l, \hat{f}_k^{\mathrm{in}}(t_l)]}$;

end for

Get the refined frequencies $\hat{\mathbf{f}}_0 \leftarrow \hat{f}_k^{\mathrm{out}}(t_l)$

Step 3: BP Optimization

Initialize complex amplitudes $\hat{\mathbf{a}}_0$;

for $j = 1 : J$ **do**

if $j < J_{\mathrm{p}}$ **then**

(1) Preliminary Optimization

Get $\nabla L_{\mathrm{spec}}(\hat{\mathbf{a}}_j)$ and update $\hat{\mathbf{a}}$ by (3.24);

else

(2) Alternate Optimization

Get $\nabla L_{\mathrm{wave}}(\hat{\mathbf{f}}_j)$ and update $\hat{\mathbf{f}}$ by (3.25);

Get $\nabla L_{\mathrm{w+s}}(\hat{\mathbf{a}}_j)$ and use it as the gradient in (3.24), then update $\hat{\mathbf{a}}$;

end if

if $L_{\mathrm{w+s}}(\hat{\mathbf{a}}_j) < L_{\mathrm{w+s}}(\hat{\mathbf{a}}_{j-1})$ **then**

$\hat{\mathbf{a}}_{\mathrm{fin}} \leftarrow \hat{\mathbf{a}}_j$ and $\hat{\mathbf{f}}_{\mathrm{fin}} \leftarrow \hat{\mathbf{f}}_j$;

end if

end for

Output: $\hat{\mathbf{a}}_{\mathrm{fin}}$ and $\hat{\mathbf{f}}_{\mathrm{fin}}$

including QHM, aQHM, eaQHM, and the proposed BP-QHM, under consistent experimental conditions. This analysis reveals the trade-offs between modeling accuracy and computational demand, which is crucial for deploying these methods in real-world applications where latency and efficiency are of paramount concern.

(3) The performance of frequency estimation in terms of harmonic structure alignment.

To evaluate the accuracy of frequency estimation, we examine the corrected individual frequencies produced by QHM-related methods and BP-QHM. These frequencies are visualized as trajectories superimposed on time–frequency representations (e.g., spectrograms), allowing for qualitative assessment. Additionally, we introduce a novel quantitative metric named "harmonic deviation," which measures the deviation of estimated harmonic frequencies from their theoretical positions derived from the estimated pitch. This metric provides a fine-grained and interpretable evaluation of harmonic alignment, particularly important in high-fidelity speech synthesis and analysis tasks.

(4) The overall quality of reconstructed speech in both objective and subjective terms.

Finally, for reconstruction quality, we assess how well the synthesized speech reproduces the original signal across various experimental setups. Specifically, we compare BP-QHM with QHM and eaQHM by analyzing the perceptual and signal-level quality of reconstructed speech under different frame-shift settings, varying numbers of harmonics (K), and different sampling rates. These experiments are designed to test the robustness and generalization ability of each method under both standard and challenging scenarios. The resynthesis results are evaluated using established metrics such as signal-to-reconstruction error, mel-cepstral distortion (MCD), and, where applicable, human listening tests, thereby providing a holistic view of the proposed model's capabilities.

Through these evaluations, we aim to demonstrate that BP-QHM not only addresses the shortcomings of conventional QHM methods, such as limited frequency correction and framewise modeling errors, but also achieves superior performance in terms of both modeling fidelity and synthesis quality.

3.5.2 Experiment Conditions

Here, we introduce the experiment conditions in detail.

Optimizer: To empirically validate the effectiveness of BP-QHM and ensure the robustness of our evaluation across various conditions, we conduct experiments using diverse datasets and rigorously defined optimization settings. The Adam optimizer [85] is employed throughout the optimization process, owing to its adaptive learning rate adjustment and widespread applicability in training deep neural models. As for the optimization setup, a step-decay learning rate schedule is

adopted. Specifically, the initial learning rate for complex amplitude updates is set to $\alpha_{\mathrm{A}} = 0.1$, while that for frequency updates is $\alpha_{\mathrm{F}} = 4$. To improve convergence and avoid overfitting, both learning rates are reduced periodically: α_{A} is decreased by a factor of 0.1 and α_{F} by a factor of 0.5 every 100 epochs. The total number of optimization iterations is fixed at $J = 500$, with the first $J_p = 200$ iterations dedicated to preliminary optimization of the complex amplitude using the proposed convex spectrogram loss function.

Dataset: The speech utterances analyzed in our experiments are randomly selected from three representative open-source corpora with varying sampling rates: the LJSpeech dataset [88] sampled at 22.05 kHz, the LibriTTS corpus [89] sampled at 24 kHz, and the AISHELL corpus [90] sampled at 44.1 kHz. From each dataset, we extract 32 utterances to ensure a fair and statistically reliable evaluation. The results are averaged across utterances to obtain representative performance indicators, and all average scores are accompanied by 95% confidence intervals to reflect the variability and statistical significance of the outcomes.

Initialization and Parameter Setting: For spectrogram-based loss computation, we set the fast Fourier transform length to $N_{\mathrm{fft}} = 1024$, which determines the temporal resolution and window length for time–frequency analysis. In this study, pitch values are provided by the YAAPT algorithm [91], which is known for its high accuracy in voiced–unvoiced detection and pitch tracking. The complex amplitudes, on the other hand, are randomly initialized unless otherwise specified. However, it is widely acknowledged that initial parameter values can significantly influence the convergence behavior and final outcome of optimization algorithms [92]. To this end, we recommend computing the short-time Fourier transform (STFT) coefficients using a truncated signal with length $N_{\mathrm{fft}} = 2K$ as a more stable and informed initialization strategy for the complex amplitude, where K denotes the number of harmonic components modeled in BP-QHM.

Measurements: To quantitatively evaluate the performance of the proposed BP-QHM framework as well as the baseline QHM methods, we adopt a series of objective and subjective metrics that reflect different aspects of system behavior, including reconstruction quality, frequency estimation accuracy, intelligibility, and computational efficiency. These indicators offer a comprehensive understanding of the strengths and limitations of each candidate method. It is worth emphasizing that for all metrics, we follow a consistent notation where the upward arrow (↑) indicates that higher values are preferable (i.e., better performance), whereas the downward arrow (↓) denotes that lower values are desirable.

1) **SRER** [dB] ↑: The Signal-to-Reconstruction Error Ratio (SRER) serves as an objective measure to assess the fidelity of waveform reconstruction. It quantifies how much of the original signal energy is preserved relative to the reconstruction error. Formally, it is defined

as

$$SRER(\hat{x},x) = 20\log_{10}\frac{\text{std}(x)}{\text{std}(x-\hat{x})}, \tag{3.26}$$

where std$(\cdot)$ denotes the standard deviation, x represents the ground-truth waveform, and $\hat{x}$ denotes the reconstructed waveform. A higher SRER indicates that the reconstructed waveform more closely matches the original signal, reflecting better reconstruction quality.

2) **RMSE** [Hz] $\downarrow$: The Root Mean Squared Error (RMSE) provides a classical and widely-used metric to quantify the average magnitude of reconstruction error in the time domain. It is computed as

$$\text{RMSE}(\hat{x},x) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n-\hat{x}_n)^2}, \tag{3.27}$$

where x_n and $\hat{x}_n$ denote the n-th samples of the reference and reconstructed waveforms, respectively, and N is the total number of samples. A lower RMSE implies smaller reconstruction error and hence better performance.

3) **STOI** $\uparrow$: The Short-Time Objective Intelligibility (STOI) [93] is a perceptually-motivated metric designed to estimate the intelligibility of speech signals. It compares the short-time temporal envelope features of clean and degraded signals across time–frequency units using a correlation-based method. STOI scores range between 0 and 1, with higher values indicating higher intelligibility, and are particularly useful for assessing the impact of processing on speech understanding.

4) **HD**: To evaluate the accuracy of harmonic structure modeling, we introduce a novel metric called *Harmonic Deviation* (HD). This metric assesses the deviation of each individual harmonic frequency f_k^l from its ideal harmonic position kf_0^l in each frame, normalized by the f_0. It is defined as

$$HD(f_0^1,\ldots,f_K^L) = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{2K}\sum_{k=-K}^{K}\left(\frac{f_k^l - kf_0^l}{f_0^l}\right)^2}, \tag{3.28}$$

where L denotes the number of frames and K is the number of harmonics. A smaller HD value reflects a more precise alignment of harmonics with the ideal harmonic grid, which is essential for generating perceptually natural and high-quality speech.

5) **MCD** $\downarrow$: The Mel-Cepstral Distortion (MCD) is a widely-used spectral distance measure for evaluating the difference between the synthesized and reference mel-cepstral features. It

is computed as

$$\mathrm{MCD}(v_{\mathrm{gen}}, v_{\mathrm{ref}}) = \frac{10\sqrt{2}}{\ln 10} \sqrt{\sum_{d=1}^{24} (v_{\mathrm{gen}}^{d} - v_{\mathrm{ref}}^{d})^2}, \tag{3.29}$$

where v_{gen} and v_{ref} are the mel cepstrum coefficient vectors of the generated and reference speech signals, respectively. Lower MCD values indicate smaller spectral envelope distortions and are associated with higher perceptual quality.

6) **MOS** $\uparrow$: The Mean Opinion Score (MOS) is a subjective metric derived from human listening tests. Participants are asked to rate the naturalness or overall quality of synthesized speech samples on a 5-point scale, where 1 corresponds to "bad" and 5 corresponds to "excellent". The MOS provides an intuitive and human-centric evaluation of speech quality, complementing the objective metrics.

7) **RTF** $\downarrow$: The Real-Time Factor (RTF) measures the computational efficiency of the system by comparing the processing time to the duration of the input signal. It is defined as

$$\mathrm{RTF} = \frac{T_{\mathrm{processing}}}{T_{\mathrm{input}}}, \tag{3.30}$$

where $T_{\mathrm{processing}}$ denotes the total processing time and T_{input} denotes the duration of the input speech signal. An RTF value below 1.0 indicates that the system can operate in real time or faster, which is important for practical deployment in speech synthesis and processing applications.

This experimental design ensures that the performance of BP-QHM is thoroughly evaluated under realistic and diverse acoustic conditions, while the structured optimization strategy and principled initialization contribute to both the reproducibility and reliability of the results.

3.5.3 Rapid Convergence with Proposed Spectrogram Loss

In this section, to assess the effectiveness and efficiency of the backpropagation process (BP) in the proposed BP-QHM framework, we first focus on evaluating the impact of different spectrogram loss functions during the preliminary optimization phase to show the superiority of the proposed spectrogram loss in backpropagation speed and the guidance for results not stuck in the local optimum. In particular, we compare the conventional spectrogram loss (denoted as $L_{\mathrm{spec}}^{\mathrm{conv}}$) with our proposed spectrogram loss (denoted as $L_{\mathrm{spec}}^{\mathrm{prop}}$), isolating their effects by running the optimization process exclusively based on these losses. For this comparative analysis, we conduct experiments with a fixed number of iterations, setting the preliminary optimization length to $J_p = 200$ epochs. During this process, we measure two key indicators:

- The average computation time per epoch.

- The quality of the optimized amplitude, as measured by the mel-cepstral distortion (MCD).

Table 3.1 presents the average results obtained from these experiments. As shown, the proposed spectrogram loss significantly reduces the computation time, achieving a processing speed approximately 3.8 times faster than that of the conventional spectrogram loss. This remarkable improvement is primarily attributed to the structural simplicity of the proposed loss, which eliminates the need to explicitly reconstruct the time-domain waveform during optimization. This not only accelerates each iteration but also reduces the computational overhead associated with signal synthesis.

Furthermore, the MCD results demonstrate the superior optimization capability of the proposed spectrogram loss. A lower MCD score suggests that the optimized complex amplitude more accurately captures the spectral envelope of the target signal, thereby indicating more effective convergence towards the global optimum. The reduction in MCD further corroborates the suitability of the proposed spectrogram loss in guiding the optimization to achieve high-quality signal reconstruction.

To provide a more intuitive illustration of the convergence behavior, we visualize the optimization trajectories of both spectrogram losses in Fig. 3.5. Specifically, we perform waveform loss analysis by generating synthesized speech using the complex amplitudes optimized with $L_{\text{spec}}^{\text{conv}}$ and $L_{\text{spec}}^{\text{prop}}$, respectively, and then compute the waveform loss L_{wave} for each epoch. These waveform losses are plotted as solid lines in the figure, while the dashed lines indicate the corresponding spectrogram loss values during the optimization process.

As can be observed in Fig. 3.5, although both loss functions demonstrate a similar decreasing trend in their respective spectrogram loss values, the waveform loss associated with $L_{\text{spec}}^{\text{prop}}$ decreases more rapidly and converges to a lower value. This suggests that the amplitude estimates optimized with the proposed loss are closer to the true amplitude values that best match the reference waveform. This improved convergence behavior can be attributed to the convexity of the proposed loss function, which ensures a well-behaved gradient landscape and avoids local minima during optimization. In contrast, the conventional spectrogram loss is inherently non-convex, often leading to unstable optimization behavior and suboptimal convergence, where the amplitude may either oscillate between different local optima or become trapped in a poor local minimum.

Overall, these results strongly support the use of the proposed spectrogram loss in the preliminary optimization stage of BP-QHM, offering both faster convergence and better reconstruction performance.

3.5.4 Study of Efficiency

In this subsection, we aim to thoroughly investigate the efficiency of the proposed BP-QHM framework by conducting a comparative analysis of the time consumption involved in both the analysis and synthesis stages, as compared to conventional QHM methods, including the original

Table 3.1: Time required for BP and MCD scores. The proposed spectrogram loss consumes less time and achieves a better MCD score.

Method	Conventional L_{spec}	Proposed L_{spec}
Time [ms/epoch] ↓	2.622±0.53	0.689±0.15
MCD [dB] ↓	2.42±0.11	1.86±0.08

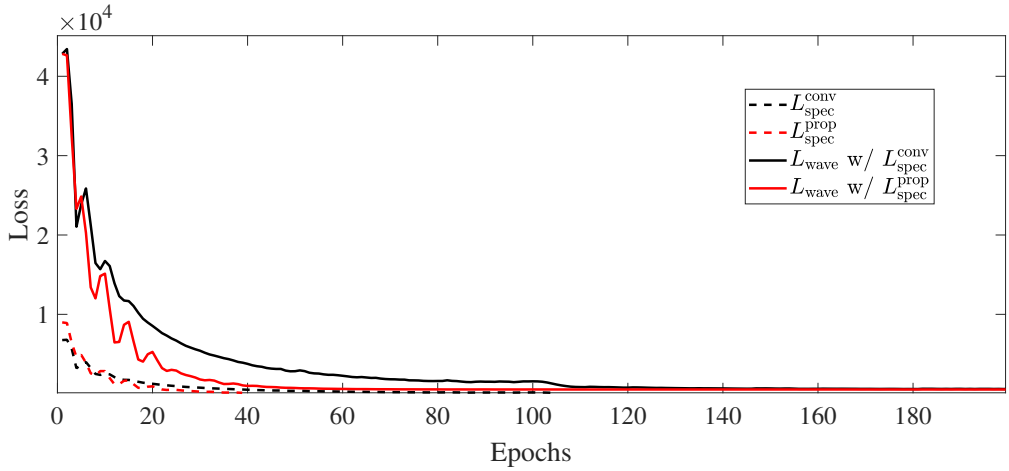


Figure 3.5: Curves of losses. Black dotted line, conventional spectrogram loss; black solid line, waveform loss optimized with conventional spectrogram loss; red dotted line, proposed spectrogram loss; red solid line, waveform loss optimized with proposed spectrogram loss.

QHM and the extended eaQHM. The primary objective is to quantify the computational demands of each method under consistent experimental conditions and to provide insight into their real-time applicability for speech modeling and synthesis.

It is important to recognize that the total time consumption is inherently influenced by various external factors, such as the total duration of the input speech signal, the selected frame-shift interval, and the specific hardware configurations employed. To ensure fairness and reproducibility, we design a controlled experiment to serve as a reference benchmark. In this experiment, we measure the time required by each method to analyze and synthesize a set of standardized speech inputs under identical settings.

Specifically, the evaluation is conducted on a platform equipped with a single AMD EPYC 7542 CPU and an NVIDIA GeForce RTX 3090 GPU. A total of 32 utterances are randomly selected from a speech dataset sampled at 22.05 kHz, and the frame-shift is fixed at 6 ms for all methods. Since the original implementations of QHM and eaQHM methods[2] do not incorporate parallel processing capabilities and are purely CPU-based, they are executed entirely on the CPU. In contrast, BP-QHM is evaluated in two configurations: one using the CPU only, and the other

[2]https://github.com/Antibas/eaQHM-analysis-and-synthesis-in-Python/tree/main

leveraging GPU acceleration.

To enable a standardized comparison across methods and platforms, we employ the real-time factor (RTF) as the performance metric. The RTF is defined as the ratio between the time required to process the signal and the actual duration of the signal, thereby representing how many seconds of computation are needed to process one second of input audio. A lower RTF value corresponds to a more efficient implementation and indicates a stronger potential for real-time or near-real-time deployment.

The RTF results for both the analysis and synthesis stages are summarized in Table 3.2. As shown, for the analysis stage, BP-QHM executed on the GPU achieves the best performance, with an average RTF of 74.79, significantly outperforming all other configurations. In contrast, eaQHM exhibits the highest RTF of 683.03, suggesting that it is the most computationally demanding method under the given experimental conditions. This is primarily due to the iterative nature of eaQHM, which involves repeated estimation and refinement of both the complex amplitude and the harmonic frequency trajectories, thereby incurring considerable computational cost.

The original QHM method, although more efficient than eaQHM, still shows a relatively high RTF of 101.03. This is mainly because it also operates frame-by-frame without parallelism, which limits its scalability. BP-QHM on the CPU exhibits a similar level of inefficiency to eaQHM, highlighting the crucial role of GPU acceleration in achieving practical performance.

It is worth noting that the original QHM methods could benefit significantly from GPU-based parallelization. Given their relatively simple and well-structured algorithmic design, it is plausible that an optimized GPU implementation would drastically reduce their RTF, potentially enabling them to surpass BP-QHM in terms of speed. However, such improvements are outside the scope of this current study, which adheres to the original open-source CPU-only implementations for a fair baseline comparison.

Turning to the synthesis stage, we observe that all methods achieve relatively low and comparable RTFs, indicating that the synthesis process is generally lightweight across different methods. Nevertheless, BP-QHM with GPU again demonstrates the best performance, achieving an RTF of 0.07, which is marginally faster than the others. The slight advantage can be attributed to the unified GPU memory access and the reuse of optimized amplitude parameters from the analysis stage.

In conclusion, these findings indicate that the proposed BP-QHM framework, when accelerated by a GPU, offers significant computational advantages in the analysis stage while maintaining competitive performance in synthesis. This makes it a promising candidate for scalable, high-quality speech modeling applications.

Table 3.2: Average RTFs of QHM, eaQHM, and BP-QHM.

RTF ↓	QHM	eaQHM	BP-QHM-GPU	BP-QHM-CPU
Analysis	101.03±5.65	683.03±24.40	**74.79±1.36**	679.46±2.49
Synthesis	0.22±0.53	0.27±0.49	**0.07±0.12**	0.25±0.22

3.5.5 Performance of Individual Frequency Estimation

In this part, we further explore the characteristics of individual frequency estimation results obtained by different methods. As formulated in Eq. (1.1), the speech signal is composed of two parts: the voiced part and the unvoiced part. Thus, the ideal behavior of speech and expected HD patterns can be concluded:

- In voiced regions, the speech spectrum exhibits a harmonic structure: spectral peaks occur near integer multiples of f_0, and their trajectories vary smoothly over time. Accordingly, an accurate estimator should place component frequencies close to kf_0, with minimal deviation and temporal continuity, yielding low HD in voiced frames.
- In unvoiced regions, the signal is noise-like without a stable harmonic grid. A proper estimator should therefore avoid aligning frequencies to kf_0; instead, any extracted components should not systematically coincide with the harmonic lattice, resulting in high HD in unvoiced frames.

First, for visual comparison, we compute the STFT of a speech signal and provide an enlarged time-frequency representation, overlaid with the individual frequencies estimated by QHM, eaQHM, and BP-QHM, as illustrated in Fig. 3.6. For consistency and fairness, all frequency estimation results are obtained using the proposed spectrogram loss as the optimization criterion.

From Fig. 3.6, we can observe that while QHM performs certain frequency adjustments during the optimization, the magnitude of these modifications remains relatively limited and often imperceptible. In particular, even in non-harmonic or unvoiced sections, the individual frequencies estimated by QHM still tend to align with the expected harmonic structure. This behavior suggests a tendency of QHM to less correct the frequencies, thereby limiting its adaptability to non-harmonic components in speech, such as those present in unvoiced segments or transient sounds.

In contrast, the individual frequency estimates produced by eaQHM appear to exhibit a higher degree of stochasticity, especially in unvoiced segments. While this randomness allows the method to better capture the inherent variability and noise-like nature of unvoiced speech, it comes at the cost of reduced stability in the voiced segments. In these segments, particularly at higher frequencies, the estimated individual frequencies deviate significantly from their corresponding

harmonic values. This instability arises due to an accumulation of frequency mismatches in the estimation of f_0, which is further amplified across higher-order harmonics. Consequently, the initial frequency of a given harmonic may inadvertently approach the true frequency of an adjacent harmonic. When the least-squares estimation step is subsequently applied, it tends to refine the estimated frequency toward this adjacent harmonic, thus introducing a systematic error. This results in excessive frequency updates and mismatch amplification, particularly in repeated estimation scenarios.

In comparison, BP-QHM demonstrates a more robust and adaptive behavior across different regions of the speech signal. Specifically, it accurately preserves the harmonic structure in voiced segments while exhibiting stochasticity in unvoiced segments, enabling it to adaptively match the spectral characteristics of both types of speech components. This is largely attributable to the frequency refinement strategy employed by BP-QHM, which enables individual frequencies to approximate their true values more accurately in voiced segments. During the backpropagation process, the gradient magnitudes of the frequency components differ between voiced and unvoiced segments, being relatively smaller in voiced segments and larger in unvoiced ones. This selective gradient behavior allows BP-QHM to focus frequency updates on non-harmonic segments while stabilizing the harmonic structure in voiced parts, thereby achieving a balanced and adaptive estimation across the signal.

Second, to provide a quantitative measure of the harmonic integrity maintained by each method, we define a new metric, termed "harmonic deviation" (HD), to characterize the deviation of individual frequencies from ideal harmonic structures, which is formulated in Eq. (3.28). The HD values computed for voiced and unvoiced segments across QHM, eaQHM, and BP-QHM are presented in Table 3.3. As shown in the table, QHM achieves the smallest HD values in both voiced and unvoiced segments, indicating a strict preservation of harmonic structure. However, this strictness may be counterproductive in non-harmonic segments, where a more flexible representation could be desirable. On the other hand, eaQHM exhibits the largest HD values, especially in voiced segments, suggesting that its frequency estimation process disrupts the harmonic structure due to instability and mismatch propagation.

In contrast, BP-QHM exhibits an intermediate result. In voiced parts, BP-QHM yields a much lower HD value than that of eaQHM, only slightly higher than QHM. In unvoiced parts, BP-QHM obtains a much higher HD value than that of QHM. This indicates that BP-QHM is capable of maintaining sufficient harmonic consistency in voiced segments while also permitting flexibility in unvoiced segments. Such a balance aligns well with the characteristics of natural speech and demonstrates the effectiveness of BP-QHM in approximating the underlying generative structure of the signal. This also indicates that BP-QHM is more accurate in frequency estimation compared to QHM and eaQHM. In essence, BP-QHM achieves a more realistic and adaptive modeling of speech frequencies by enabling localized and selective refinement, which is critical for high-

Table 3.3: HDs of speech signals resynthesized by QHM, eaQHM, and BP-QHM.

Method	QHM	eaQHM	BP-QHM
Voiced	3.42×10^{-5}	15.29	7.37×10^{-2}
Unvoiced	3.59×10^{-4}	9.79	1.54×10^{-1}

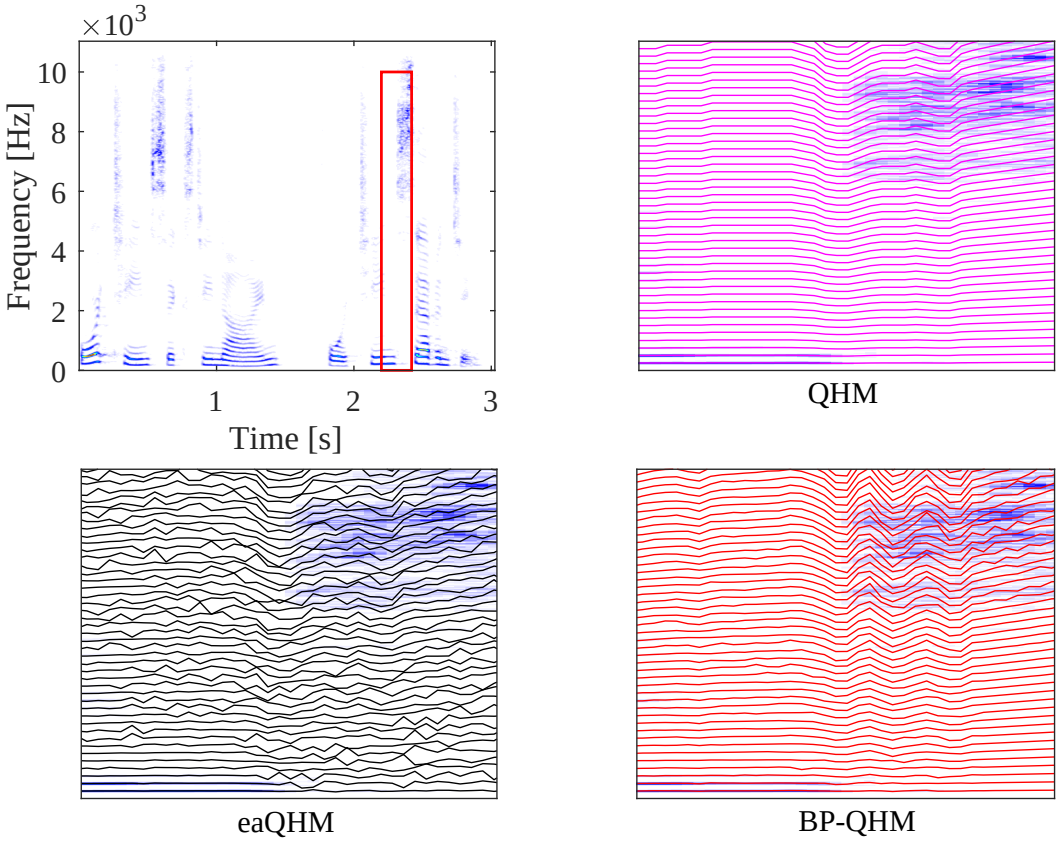


Figure 3.6: The STFT of a speech signal and individual frequencies estimated by various methods. Pink lines: estimated frequencies by QHM, black lines: estimated frequencies by eaQHM, red lines: estimated frequencies by BP-QHM.

quality resynthesis and downstream processing tasks.

3.5.6 Evaluation of Speech Resynthesis

In this subsection, we evaluate the resynthesis capability of different methods by systematically comparing the reconstructed speech with the ground truth reference in terms of signal fidelity, intelligibility, and perceived quality. Specifically, we adopt several objective and subjective metrics to comprehensively assess the reconstruction performance, signal similarity, and intelligibility, such as RMSE (Eq. (3.27)), SRER (Eq. (3.26)), STOI, MCD (Eq. (3.29)), and MOS.

Before comparing BP-QHM with other QHM-based approaches, we first perform an internal ablation study to investigate how two key design choices, namely the type of spectrogram loss and

Table 3.4: Average RMSE, SRER, STOI, and MCD scores of BP-QHM in different conditions.

BP-QHM with	CSL-ABP	PSL-nonABP	PSL-ABP
RMSE ↓	0.025±0.002	0.027±0.002	**0.017±0.001**
SRER [dB] ↑	12.3±0.91	11.3±0.70	**14.9±0.71**
STOI ↑	0.87±0.02	0.85±0.01	**0.89±0.01**
MCD [dB] ↓	2.63±0.10	2.55±0.11	**2.04±0.09**

the use of alternating backpropagation (ABP), affect the performance of BP-QHM. Specifically, we examine three configurations:

1) **CSL-ABP**: Using the conventional spectrogram loss with alternating backpropagation.

2) **PSL-nonABP**: Using the proposed spectrogram loss without alternating backpropagation.

3) **PSL-ABP**: Using the proposed spectrogram loss with alternating backpropagation.

In this experiment, 32 utterances randomly selected from the LJSpeech corpus (22.05 kHz) are analyzed under identical conditions: 8 ms frame-shift and 128 harmonics. Table 3.4 presents the averaged scores of RMSE, SRER, STOI, and mel-cepstral distortion (MCD) for the three settings.

The results clearly demonstrate that the PSL-ABP configuration significantly outperforms the other settings across all evaluation metrics. Specifically, the conventional spectrogram loss (CSL) leads to suboptimal performance due to its inherent non-convexity, which can cause the optimization process to get stuck in local minima. In contrast, the proposed spectrogram loss (PSL), by bypassing the waveform synthesis during loss computation, improves the convexity of the objective landscape and thereby facilitates faster and more reliable convergence to a global optimum. Furthermore, the inferior performance of PSL-nonABP confirms the necessity of alternating backpropagation, which enables the method to adaptively transition between harmonic and stochastic components in unvoiced segments. Based on these findings, the PSL-ABP configuration is selected to represent BP-QHM in subsequent comparisons.

We now compare the proposed BP-QHM method with QHM and eaQHM under identical experimental conditions. Table 3.5 presents the average performance metrics obtained from 32 LJSpeech utterances. The results indicate that although QHM performs well in SRER and MCD, eaQHM achieves better scores due to its iterative and nonlinear estimation of both frequency and amplitude components. Nevertheless, BP-QHM surpasses both baselines across all metrics. Specifically, the global optimization framework of BP-QHM enables it to consider the entire speech sequence holistically rather than modeling it frame by frame. This comprehensive optimization allows BP-QHM to iteratively refine frequency estimates toward their true values and simultaneously adjust amplitude and phase, resulting in improved parameter accuracy and signal

Table 3.5: Average RMSE, SRER, STOI, and MCD scores. The MOS of the ground truth samples was 4.25 ± 0.02.

Method	QHM	eaQHM	BP-QHM
RMSE ↓	0.029±0.002	0.022±0.002	**0.017±0.001**
SRER [dB] ↑	10.6±0.54	12.9±0.71	**14.9±0.71**
STOI ↑	0.84±0.01	0.88±0.01	**0.89±0.01**
MCD [dB] ↓	2.30±0.07	2.18±0.08	**2.04±0.09**
MOS ↑	3.78±0.03	3.83±0.03	**4.05±0.02**

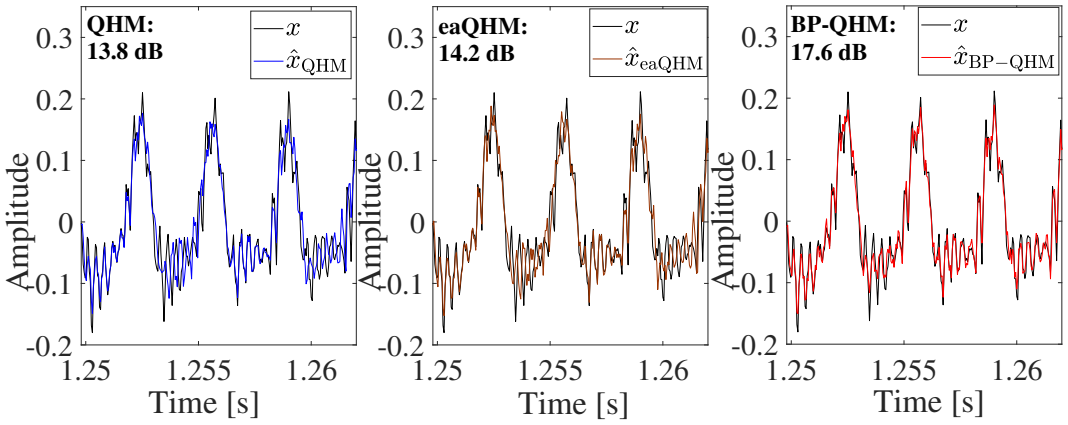


Figure 3.7: The waveforms of resynthesized speech signals. Black: reference, blue: QHM (left), brown: eaQHM (middle), red: BP-QHM (right).

reconstruction. For instance, BP-QHM achieves a notable SRER of 14.9 dB and an MCD of 2.04 dB, indicating superior spectral and waveform fidelity.

In addition to numerical metrics, the qualitative performance of each method is illustrated in Fig. 3.7, where waveforms reconstructed by QHM, eaQHM, and BP-QHM are displayed alongside the original reference waveform. It is evident from the figure that the waveform generated by BP-QHM is most similar to the reference, both in terms of shape and dynamics. Furthermore, the SRER values for the displayed example are also provided in the figure, confirming that BP-QHM achieves the highest reconstruction fidelity.

To comprehensively investigate the performance of the proposed BP-QHM method under different analysis conditions, we conduct a series of experiments under three distinct scenarios: varying frame-shift lengths, different numbers of harmonic components, and multiple sampling rates. These experiments aim to evaluate the robustness, generalization, and adaptability of the methods under various practical configurations.

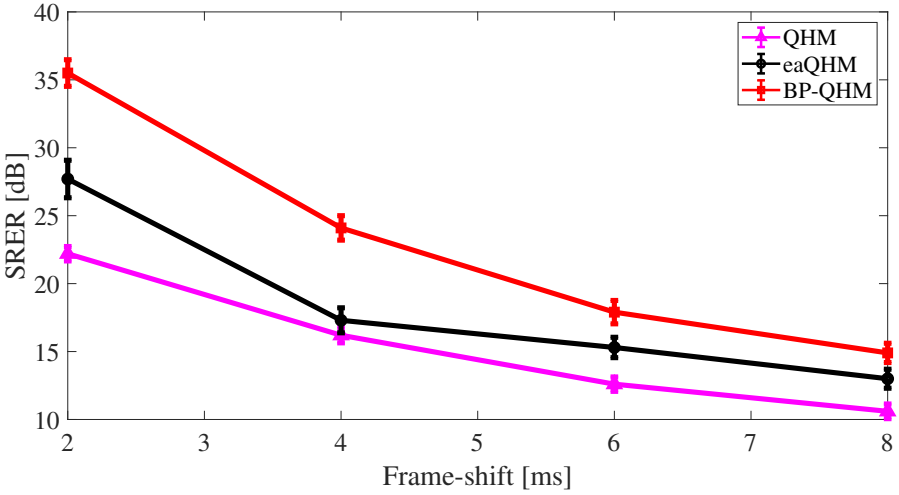


Figure 3.8: Curves of SRER scores obtained by various methods as a function of time shift.

Effect of Frame-Shift Lengths

We first examine how the analysis frame-shift affects the performance of each method. Specifically, we evaluate the models under frame-shift durations of 2 ms, 4 ms, 6 ms, and 8 ms. The average SRER, MCD, and MOS scores are computed over 32 test utterances and shown in Figs. 3.8, 3.9, and 3.10, respectively.

As illustrated in the figures, both QHM and eaQHM suffer significant performance degradation as the frame-shift increases. This is primarily attributed to their inherent framewise estimation process, which fails to capture temporal dependencies across frames. In contrast, BP-QHM maintains superior performance, with more gradual deterioration as the frame-shift increases. Even at the largest frame-shift (8 ms), BP-QHM achieves an SRER of 14.9 dB and an MCD of 1.64, which remain superior to those of the other methods. Moreover, BP-QHM consistently achieves the highest MOS values, further confirming its ability to reconstruct perceptually natural speech.

These findings suggest that the sequence-level modeling adopted by BP-QHM enables it to retain high-quality speech resynthesis while reducing the number of frames required for analysis. This not only enhances computational efficiency but also improves modeling robustness.

Effect of the Number of Harmonics

Next, we investigate how the number of harmonic components (K) used during analysis affects performance. The experiments are conducted on 22.05 kHz speech data using $K = 32$, 64, and 128 harmonics. The average SRER, STOI, MCD, and MOS scores are summarized in Table 3.6.

Overall, BP-QHM consistently outperforms the other methods when sufficient harmonic components are available (i.e., $K = 128$ and $K = 64$). However, as K decreases, particularly to $K = 32$, BP-QHM's performance notably declines in both SRER and MOS. In contrast, QHM and eaQHM exhibit only minor performance drops. This is likely because the LS-based estimation in QHM

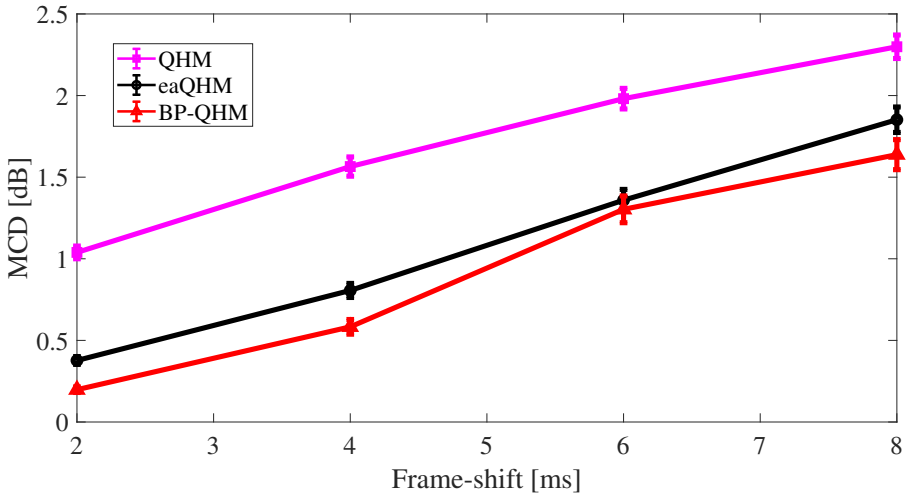


Figure 3.9: Curves of MCD scores obtained by various methods as a function of time shift.

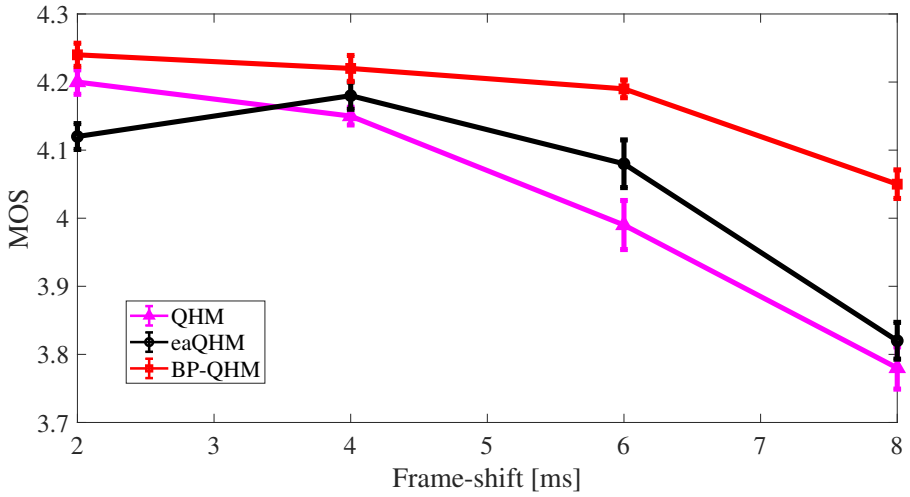


Figure 3.10: Curves of MOS values obtained by various methods as a function of time shift. The MOS value of ground truth is 4.25 ± 0.02.

and eaQHM can match the waveform reasonably well even with fewer harmonic components, due to their flexible local fitting.

BP-QHM, however, relies on modeling the full signal spectrum through harmonic decomposition and requires a sufficient number of harmonics to accurately reconstruct both spectral and temporal characteristics. The results highlight the importance of selecting an appropriate K, particularly for higher-pitched signals such as those from female speakers, to ensure high synthesis quality.

Effect of Sampling Rate

Finally, we assess how the sampling rate (f_s) of the input speech affects the performance of the different methods. Using a fixed frame-shift of 6 ms and $K = 128$ harmonics, we evaluate speech

Table 3.6: Average SRER, STOI, and MCD scores at various K values. The MOS of the ground truth samples was 4.33 ± 0.02.

Harmonic number	Method	SRER [dB] ↑	STOI ↑	MCD [dB] ↓	MOS ↑
$K = 128$	QHM	12.6±0.53	0.88±0.01	1.98±0.06	4.12±0.03
	eaQHM	15.3±0.75	0.92±0.01	1.41±0.07	4.05±0.02
	BP-QHM	**17.9±0.86**	**0.93±0.01**	**1.40±0.08**	**4.18±0.02**
$K = 64$	QHM	12.4±0.55	0.88±0.01	2.03±0.06	4.11±0.02
	eaQHM	**14.8±1.03**	0.92±0.01	1.52±0.07	4.18±0.02
	BP-QHM	14.8±0.96	**0.93±0.01**	**1.49±0.08**	**4.23±0.02**
$K = 32$	QHM	11.3±0.69	0.88±0.01	2.37±0.10	4.06±0.02
	eaQHM	**12.2±0.86**	**0.92±0.01**	2.00±0.10	**4.11±0.03**
	BP-QHM	11.0±1.06	**0.92±0.01**	**1.94±0.10**	4.09±0.02

signals sampled at 22.05 kHz, 24 kHz, and 44.1 kHz. The results are summarized in Table 3.7.

As shown in the table, BP-QHM consistently yields higher SRER and lower MCD across all sampling rates, confirming its robustness. Notably, BP-QHM significantly benefits from increased sampling rate in terms of frequency resolution, allowing for better modeling of high-frequency components. This is evident from the performance at 24 kHz, where BP-QHM achieves its best overall results.

However, at 44.1 kHz, the advantage of BP-QHM becomes less pronounced. This is likely due to the predominance of male speech in the test samples, where f_0 values are relatively low, and most energy is concentrated in the lower-frequency band. Under such conditions, the higher sampling rate provides little additional benefit, since fewer high harmonics are active. Nonetheless, BP-QHM still remains competitive or superior to QHM and eaQHM.

In summary, across all experimental conditions, including varying frame-shifts, harmonic numbers, and sampling rates, the proposed BP-QHM consistently demonstrates robust and high-quality performance. These results underscore its flexibility and adaptability to a wide range of speech signal conditions.

3.6 Summary

In this chapter, we demonstrated that conventional QHM methods struggle to effectively extract the amplitudes and frequencies of individual signal components, often resulting in unsatisfactory reconstruction of the speech waveform. To address these limitations, we proposed a frequency

Table 3.7: Average SRER, STOI, and MCD scores at various f_s values. The MOS values of ground truth with 22.05 kHz, 24 kHz, and 44.1 kHz were 4.25 ± 0.01, 3.90 ± 0.03, and 3.67 ± 0.03, respectively.

Sampling rate	Method	SRER [dB] ↑	STOI ↑	MCD [dB] ↓	MOS ↑
$f_s = 22.05$ kHz	QHM	12.6±0.53	0.88±0.01	1.98±0.06	4.08±0.02
	eaQHM	15.3±0.75	0.92±0.01	1.41±0.07	4.14±0.03
	BP-QHM	**17.9±0.86**	**0.93±0.01**	**1.40±0.08**	**4.21±0.02**
$f_s = 24$ kHz	QHM	17.0±0.66	0.93±0.01	1.23±0.08	3.78±0.03
	eaQHM	21.6±1.03	0.96±0.01	0.54±0.08	3.80±0.03
	BP-QHM	**26.9±1.12**	**0.98±0.01**	**0.38±0.09**	**3.88±0.04**
$f_s = 44.1$ kHz	QHM	16.9±0.27	0.92±0.01	1.01±0.03	3.56±0.02
	eaQHM	20.5±0.31	**0.95±0.01**	0.73±0.04	**3.65±0.02**
	BP-QHM	**21.2±0.26**	**0.95±0.01**	**0.71±0.03**	3.60±0.02

refinement strategy based on the time–frequency representation (TFR). This approach aims to refine the initially estimated f_0 obtained via pitch detection and subsequently enhance the accuracy of the associated harmonic frequencies. When the initial f_0 estimation deviates moderately from the ground truth, the proposed method is capable of achieving precise frequency estimation, even for signals with rapidly varying frequency trajectories.

Furthermore, we reformulated the QHM framework by discarding the conventional framewise analysis and instead performing parameter estimation over the entire speech utterance. Specifically, we adopted a gradient descent approach to jointly solve for the complex amplitudes and frequencies of all components, utilizing gradients propagated from the original waveform. To promote faster and more accurate convergence to the optimal parameter values, we introduced a novel spectrogram-based loss function. This loss function guides the optimization by aligning the estimated parameters with the spectral characteristics of the target speech signal, thereby enabling nearly perfect waveform reconstruction.

Despite its effectiveness, the proposed BP-QHM framework typically requires a large number of iterations to converge, making it computationally intensive and less suitable for real-time applications. Nonetheless, through extensive experiments on various speech datasets, the proposed method has been validated to achieve superior accuracy in both frequency and complex amplitude estimation, resulting in significantly improved waveform reconstruction. These promising results not only indicate the potential of BP-QHM for downstream applications in speech analysis, transformation, and synthesis, but also demonstrate that backpropagation can be effectively integrated into the QHM framework. This finding further suggests the feasibility of combining

neural network models with QHM structure, which lays the foundation for achieving the ideal speech modeling in the next chapter, i.e., the vocoder that satisfies high-speed, high-quality, and high-controllability extraction and synthesis.

Chapter 4

Neural Quasi-Harmonic Modeling

4.1 Introduction

The BP-QHM framework effectively optimizes the quasi-harmonic parameters via backpropagation, enabling high-quality speech resynthesis. This success suggests that the synthesis mechanism underlying QHM methods holds great potential for integration with neural networks. With careful architectural design, it becomes possible to harness the strengths of both QHM methods and neural networks, thereby facilitating the development of more advanced neural vocoders and contributing to practical applications in human-centered speech processing.

Building on this insight, this chapter introduces a novel neural vocoder framework that incorporates the QHM structure into neural networks. Two neural vocoders are proposed under this framework to further improve speech synthesis performance. This integration addresses several limitations associated with conventional and neural vocoders. Specifically, conventional vocoders often suffer from insufficient parameter estimation accuracy and limited robustness, while neural vocoders tend to operate as black-box models, lacking the ability to extract interpretable parameters or to provide controllable features, such as explicit pitch manipulation. The proposed approach aims to overcome these challenges by combining interpretability and structure-awareness with the modeling power of neural networks.

To integrate the interpretability and controllability of CSP with the robustness and high-fidelity generation capabilities of deep neural networks, we propose a novel vocoder framework that combines the QHM with DNN-based modeling. Specifically, the proposed framework simplifies the original QHM synthesis process by introducing a phase compensation mechanism to replace the conventional instantaneous phase computation. This simplification significantly accelerates the synthesis process while maintaining the frequency correction mechanism inherent in QHM. The DNN is employed to estimate the necessary quasi-harmonic parameters, including complex amplitudes and individual frequencies, enabling fast and high-quality speech synthesis that retains interpretable and controllable acoustic structures.

Second, to further enhance the modeling capability of the proposed framework, particularly

in representing the spectral resonance characteristics of speech, we incorporate an autoregressive moving average (ARMA) model into the architecture, resulting in the QHM-GAN vocoder. In this design, the ARMA parameters are directly predicted by the DNN, allowing each quasi-harmonic component to acquire amplitude and phase delay characteristics from the frequency response of the estimated ARMA filter. This facilitates a more accurate and compact representation of the vocal tract response, improving both synthesis quality and the flexibility of speech modification.

In this chapter, we evaluate the proposed methods by comparing them with QHM, the canonical baseline of conventional vocoders, as well as several neural vocoders, rather than BP-QHM. This choice is motivated by several considerations:

- First, QHM is the origin of the entire QHM family and a widely recognized conventional vocoder in the literature. Since our work aims to integrate QHM with neural networks, it is natural and most reasonable to compare with QHM rather than BP-QHM.
- Second, while BP-QHM achieves accurate parameter estimation through waveform-level optimization, it relies on iterative gradient descent, resulting in substantial computational cost and making it less suitable for large-scale or real-time evaluation. As shown in Table 3.2, the RTFs of BP-QHM and eaQHM are much higher than that of QHM, which makes large-scale experimental evaluation impractical.
- Third, the success of BP-QHM has already demonstrated that integrating QHM with neural networks is feasible and effective, which serves as the cornerstone for subsequent integration. Therefore, there is no need for further comparison with BP-QHM in this Chapter.

To comprehensively evaluate the effectiveness of the proposed method, a series of experiments are conducted using real speech utterances from publicly available corpora. The experimental results demonstrate that our framework successfully leverages the complementary strengths of QHM, ARMA filtering, and deep neural networks. Specifically, the integration of these components enables the proposed method to achieve significant improvements over conventional vocoders and state-of-the-art neural vocoders in multiple aspects, including waveform generation speed, synthesis quality, and modification flexibility. These advantages affirm the practicality and generalizability of the proposed method in diverse speech processing scenarios such as high-fidelity synthesis, pitch modification, and prosody control.

4.2 CSP-DNN Hybrid Vocoder

To simultaneously overcome the limitations of conventional vocoders and neural vocoders, we target the design of a vocoder capable of efficiently and accurately compressing speech signals into sparse representations during encoding. This is particularly important for high-sampling-rate signals, which pose challenges due to their large data volumes. During decoding, the vocoder is expected to rapidly and reliably reconstruct high-quality speech waveforms.

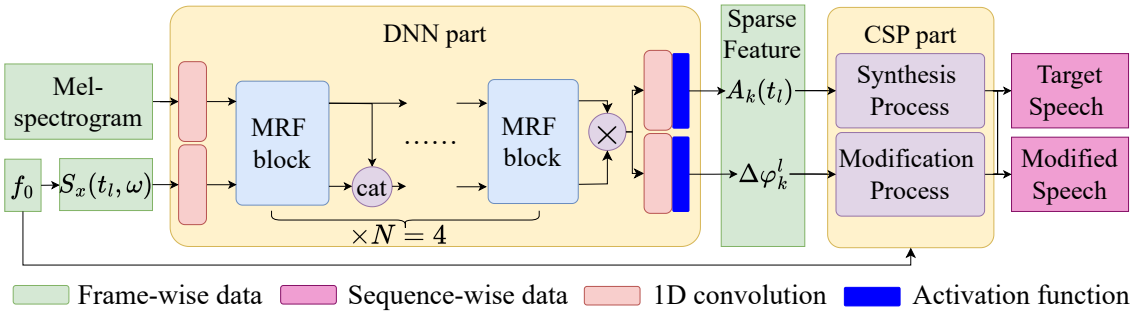


Figure 4.1: Structure of QHM-GAN generator, which takes mel-spectrogram as inputs and outputs sparse framewise amplitude and phase compensation. The framewise amplitude and phase compensation will be used to generate the speech waveform along with the frequency.

Motivated by the capability of QHM methods to sparsely represent both voiced and unvoiced speech, support efficient synthesis, and offer precise control over the f_0, we incorporate these advantages into a robust deep learning framework. Specifically, we propose a novel vocoder architecture named **QHM-GAN**, which integrates CSP with a generative adversarial network (GAN) structure. By leveraging the inference power of deep neural networks and the interpretability and controllability of QHM-based representations, QHM-GAN enables accurate and flexible speech generation.

In contrast to conventional mel-spectrogram-to-waveform paradigms, QHM-GAN adopts a mel-spectrogram-to-parameter approach, where the generator predicts quasi-harmonic parameters from input mel-spectrograms. These parameters are then used for efficient waveform reconstruction and enable explicit speech modification capabilities, such as f_0 extrapolation. The architecture of the generator is illustrated in Fig. 4.1. From Fig. 4.1, it is obvious that the proposed QHM-GAN framework mainly consists of two components: a DNN module and a CSP module.

The DNN module takes the mel-spectrogram as input and predicts sparse quasi-harmonic parameters, specifically the framewise amplitude and phase compensation for each harmonic component. These parameters are compact and interpretable, allowing for efficient representation and manipulation of speech signals.

The CSP module is responsible for decoding, where the predicted quasi-harmonic parameters are utilized to synthesize the time-domain speech waveform. This module inherits the benefits of QHM-based synthesis, enabling high-quality reconstruction and precise control over fundamental frequency and harmonic structure.

The following subsections describe the DNN and CSP modules in detail.

4.2.1 Synthesis Process Simplification

In the synthesis process, we are inspired to integrate the QHM structure, namely, that the speech waveform is the sum of several quasi-harmonic sinewaves, with neural networks. In this manner,

the speech waveform is generated from framewise parameters of quasi-harmonics, including pitch (f_0), amplitude, and phase, which analytically describe the structure of the speech waveform. A straightforward approach is to directly employ the synthesis process of conventional QHM methods. However, such a process is computationally complex and may degrade both the generation efficiency and backpropagation performance.

To address this, we propose a novel synthesis approach by introducing new definitions that simplify the original QHM synthesis while retaining its advantageous properties.

First, we briefly review the synthesis process of QHM methods. Once the instantaneous amplitude and phase are obtained, the speech waveform can be reconstructed by

$$\hat{x}(t) = \sum_{k=-K}^{K} \hat{A}_k(t) e^{i\hat{\varphi}_k(t)}, \tag{4.1}$$

where $\hat{(\cdot)}$ denotes the estimated value of each variable. Here, $\hat{x}(t)$ is the generated speech waveform, and $\hat{A}_k(t)$ and $\hat{\varphi}_k(t)$ are the instantaneous amplitude and phase of the k-th component, respectively. The instantaneous phase is typically computed according to Eq. (2.30). However, this relies on the assumption that the error between $\tilde{\varphi}_k(t_{l+1})$ and the unwrapped $\hat{\varphi}_k(t_{l+1})$ is sufficiently small. In practice, the frequencies estimated by pitch detectors often deviate from the ground truth, and such deviations can result in significant phase errors. Furthermore, unvoiced speech segments are inherently non-harmonic, and forcing them into a harmonic structure results in inevitable mismatches.

Given these issues, it becomes essential to introduce a phase compensation mechanism, akin to that in QHM methods, while simplifying the synthesis process and maintaining the frequency correction capability. To this end, we are inspired by the fact that the phase is the integral of the instantaneous frequency, and frequency mismatches can thus be interpreted as phase errors. We therefore define a novel concept of phase compensation to represent this mismatch.

Definition 1. ***Phase compensation.*** *Within the l-th frame, for the k-th component, let the estimated instantaneous frequency be $\hat{f}_k(t)$ and the phase compensation be defined as $\Delta\varphi_k^l$. Then, the phase at the frame center t_l is calculated as*

$$\hat{\varphi}_k(t_l) = \hat{\varphi}_k(t_{l-1}) + \int_{t_{l-1}}^{t_l} 2\pi \hat{f}_k(u)\,\mathrm{d}u + \Delta\varphi_k^l, \tag{4.2}$$

where $\Delta\varphi_k^l$ accounts for the discrepancy between the estimated phase (derived from the estimated frequency) and the true phase.

To provide a visual explanation of phase compensation, Fig. 4.2 illustrates its definition. Assuming that the estimated frequency at $t = t_l$ is accurate, i.e., $\hat{f}_k(t_l) = f_k(t_l)$, an inaccurate frequency estimate at $t = t_{l+1}$, denoted by $\hat{f}_k(t_{l+1})$, leads to an erroneous phase $\tilde{\varphi}_k(t_{l+1})$. This deviation introduces a phase error relative to the true phase, which is represented by $\Delta\varphi_k^l$ in Fig. 4.2.

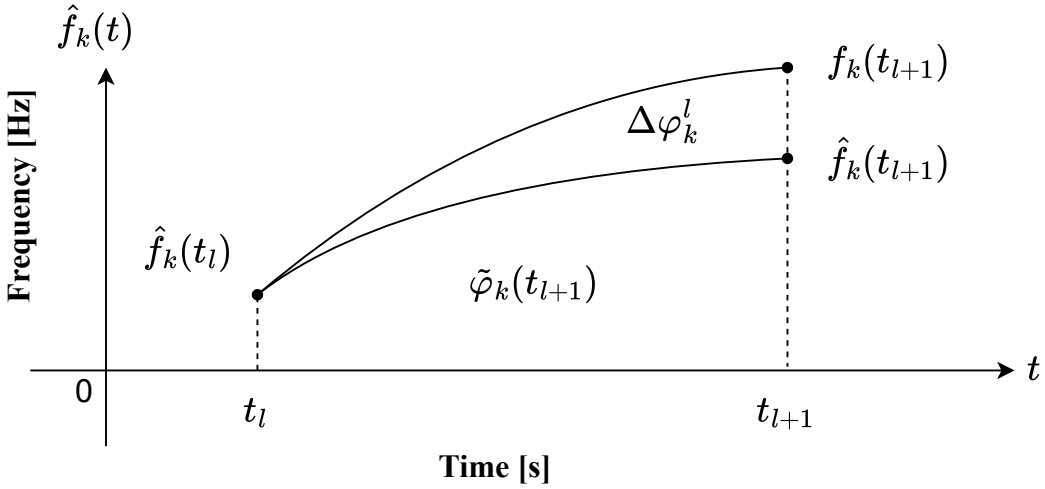


Figure 4.2: Illustration of phase compensation.

With this definition, we enable the DNN to estimate $\Delta\varphi_k^l$ for each component at each frame center. Once the instantaneous frequency is computed via $\hat{f}_k(t) = k\hat{f}_0(t)$, the phase at t_l can be corrected by

$$\hat{\varphi}_k(t_l) = \int_0^{t_l} 2\pi \hat{f}_k(u)\,\mathrm{d}u + \sum_{i=0}^{l} \Delta\varphi_k^i. \tag{4.3}$$

To facilitate implementation, the integral in Eq. (4.3) is approximated using a trapezoidal rule, averaging the adjacent framewise frequencies:

$$\int_0^{t_l} 2\pi \hat{f}_k(u)\,\mathrm{d}u \approx \pi \sum_{i=1}^{l} \left[\hat{f}_k^{i-1} + \hat{f}_k^i\right](t_i - t_{i-1}), \tag{4.4}$$

where $\hat{f}_k^l$ is the estimated frequency of the k-th component at frame l.

Subsequently, the instantaneous phase is obtained via cubic interpolation across frame centers, and the amplitude is interpolated linearly. This combination allows accurate waveform synthesis. By introducing phase compensation, the generator acquires the capability to adaptively adjust frequencies, akin to QHM methods. As a result, the synthesized speech waveform aligns more closely with the target speech, improving the quality of both voiced and unvoiced segments.

4.2.2 Neural Structure for Acoustic Feature Estimation

The DNN component aims to estimate framewise amplitude and phase compensation from the mel-spectrogram for subsequent resynthesis. Given that speech is a temporal sequence, temporal DNNs, such as RNNs, LSTMs, or CNNs, are preferred for their ability to model both past and future contexts. As QHM-GAN requires harmonic parameters conditioned on the input f_0, the generator receives both the mel-spectrogram and f_0. To capture rich linguistic and speaker information, multi-receptive field (MRF) modules are employed as the main network structure. Each

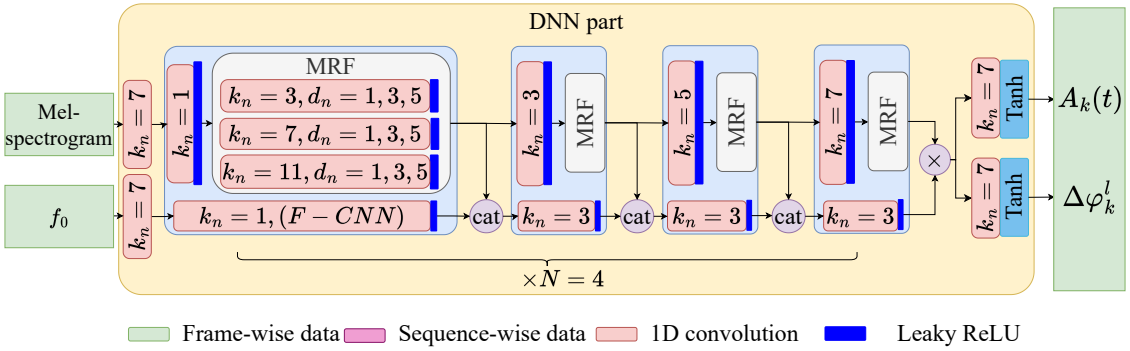


Figure 4.3: Structure of the DNN part of QHM-GAN generator, which employs several MRF blocks to estimate the output. k_n and d_n are the kernel size and dilation of the corresponding convolution layer, respectively, while Tanh denotes the hyperbolic tangent function. If unlabeled, $d_n = 1$ by default. Note that the configurations of all MRFs in the generator are the same.

MRF module comprises three 1D convolution layers with kernel sizes of 3, 7, and 11, and dilation rates of 1, 3, and 5. Additionally, CNNs process f_0 to extract harmonic features that facilitate pitch control. Since both the generator input and output are framewise, standard CNN layers connect adjacent MRF blocks. Leaky ReLU is used as the activation function throughout. Details of the DNN component are illustrated in Fig. 4.3.

We now describe how the mel-spectrogram and f_0 are input to and processed by the network. The mel-spectrogram first passes through a standard CNN layer, followed by N sequential MRF modules (typically $N = 4$) to extract hidden features.

To incorporate f_0 into the network in a spectrogram-like form, a pseudo-STFT is generated by setting all amplitudes to 1 and placing Gaussian peaks at the harmonic frequencies:

$$S_x(t_l, \omega) = \sum_{k=-K}^{K} e^{-\frac{\sigma^2[\omega - 2\pi\hat{f}_k(t_l)]^2}{2}}, \tag{4.5}$$

where σ is the standard deviation of the Gaussian window. This approach is motivated by the observation that when the frequency and amplitude vary slowly within a frame, the STFT can be approximated as a sum of window functions centered at the harmonic frequencies.

The pseudo-STFT is processed through a series of CNN layers (F-CNNs). After each F-CNN, its output is concatenated with the corresponding MRF module output and passed to the next F-CNN, enabling effective integration of frequency information with linguistic and contextual features. Following several F-CNN layers, the final output is element-wise multiplied with the hidden features from the last MRF module to emphasize harmonically relevant information.

The fused representation is subsequently fed into two separate standard CNN branches to produce phase compensation and amplitude values. As phase compensation inherently falls within $[-\pi, \pi]$, a hyperbolic tangent function is applied to constrain the output:

$$\Delta\varphi_k^l = \pi \cdot \tanh(\cdot). \tag{4.6}$$

The generated framewise amplitude is linearly interpolated to obtain the instantaneous amplitude, and the generated phase compensation is used to compute the framewise phase via Eq. (4.3), which is then cubically interpolated to obtain the instantaneous phase.

4.2.3 QHM-GAN

By integrating the two designs mentioned above, the generator of QHM-GAN is established by combining the advantages of conventional signal processing and neural networks. In such a case, QHM-GAN can robustly and efficiently estimate the framewise acoustic features of quasi-harmonics, with which the speech waveform can be quickly and flexibly generated. In the following parts, we introduce the QHM-GAN in detail.

Generator Architecture: The generator architecture of QHM consists of a neural estimator and a QHM-based synthesis module, as shown in Fig. 4.1. Neural estimator absorbs the mel-spectrogram and f_0 and estimates the framewise amplitudes and phase compensations of individual quasi-harmonics. Then, the QHM-based synthesis module utilizes the framewise phase compensations and frequencies to generate the instantaneous phases of individual quasi-harmonics. With the instantaneous amplitudes linearly interpolated from the framewise amplitudes, the speech can be synthesized by summing the sinewaves of individual quasi-harmonics using Eq. (4.1). The complete pseudocode of the speech synthesis process in QHM-GAN is detailed in Algorithm 2.

The differentiability of the entire framework allows the gradients from the loss function to be backpropagated through the synthesis module. Importantly, the introduction of phase compensation enables the model to correct the phase at each frame center, effectively adjusting the frequency of each sparse component.

Algorithm 2 Speech synthesis based on QHM-GAN.

Step 1: Preprocessing and Setting
Extract the $\hat{f}_k$ and mel-spectrogram from speech $x(t)$ and compute $S_x(t,\omega)$; set sampling rate f_s, harmonic number K, and frame-shift; obtain $\hat{A}_k(t_l)$ and $\Delta\varphi_k^l$ from DNN;
Step 2: Instantaneous amplitude and phase
Compute raw rotation angle by Eq. (4.4);
Compute unwrapped framewise phase $\hat{\varphi}_k(t_l)$ by Eq. (4.3);
Cubically interpolate $\hat{\varphi}_k(t_l)$ into $\hat{\varphi}_k(t)$;
Linearly interpolate $\hat{A}_k(t_l)$ into $\hat{A}_k(t)$;
Step 3: Generation
$\hat{x}(t) \leftarrow \sum_{k=-K}^{K} \hat{A}_k(t) e^{i\hat{\varphi}_k(t)}$;
Output: $\hat{x}(t)$

Discriminator Design: The discriminator of QHM-GAN follows the general paradigm of Vocos. Since the human auditory system is more sensitive to frequency components than to time-domain waveform shapes, we adopt a spectrogram-based discriminator, as proposed in [6], to guide the

learning process. This discriminator operates on multiple spectrogram resolutions and is designed to distinguish between the generated and target speech in the frequency domain.

The spectrogram-based discriminator effectively serves as a surrogate for the human auditory perception system, focusing on discrepancies in the spectral structure rather than the raw waveform. By evaluating the spectrogram at different resolutions, the discriminator captures both coarse and fine-grained spectral patterns. This encourages the generator to produce speech with smoother and more natural frequency transitions, thereby improving the perceptual quality of the synthesized waveform.

High-speed Generation of QHM-GAN: Because QHM-GAN does not require upsampling, it avoids the computational cost associated with transposed convolutions. Moreover, the generator produces outputs at the same temporal resolution as the input mel-spectrogram, so convolutional kernels are not applied to enlarged feature maps, preventing additional computation. These design choices allow QHM-GAN to achieve fast inference.

In contrast to methods like HiFi-GAN, which map mel-spectrograms directly to time-domain waveforms, the DNN component of QHM-GAN is designed to produce sparse spectral parameters, namely, amplitude and phase, greatly reducing the network's learning burden. As a result, the generator can be simplified by decreasing either the number of multi-receptive field (MRF) modules or the dilation rates in each MRF block, while still retaining adequate capacity to accurately estimate the parameters.

Preliminary experimental results indicate that decreasing the number of MRF modules and dilation rates has minimal impact on the quality of the synthesized speech. This observation demonstrates the feasibility of employing QHM-GAN in real-time speech synthesis applications.

4.2.4 Potential Limitation of QHM-GAN

To preliminarily validate the effectiveness of integrating QHM with neural networks, we conducted exploratory experiments to evaluate the performance of QHM-GAN, showing that QHM-GAN successfully integrates the strengths of neural networks and QHM. On the one hand, QHM-GAN achieves the fast inference capability of neural networks; on the other hand, it inherits the interpretability of CSP-based methods while maintaining high-quality speech reconstruction. These findings suggest that combining QHM with neural networks is both feasible and promising. Nevertheless, since the speech generation process is still grounded in the QHM framework, certain inherent limitations of QHM remain, leaving QHM-GAN with some shortcomings. The main potential issues can be summarized as follows:

- QHM models unvoiced segments using quasi-harmonics, despite the fact that unvoiced speech are inherently random noise. This mismatch can lead to inaccurate modeling and degraded synthesis quality, particularly when the frame shift is large. QHM-GAN suffers

from the same issue.

- QHM does not explicitly model formant characteristics, and thus requires additional methods, such as linear predictive coding (LPC) [54] or discrete all-pole (DAP) [94], for formant modeling. This limitation also applies to QHM-GAN. Although QHM-GAN provides interpretable parameters, the estimation process is still carried out by a black-box model, making the modeling of the spectral envelope implicit. As a result, the modeling of the spectral envelope is affected by the data-hungry nature of the system. For example, although QHM-GAN has a good generalization ability, when generating speech with a f_0 that lies far outside the range observed in the training data, it becomes difficult to accurately model the spectral envelope. This leads to abnormal amplitude and phase compensation, ultimately degrading the quality of the synthesized speech.

- Owing to its quasi-harmonic modeling, QHM-GAN is well suited for such tasks. However, the implicit spectral envelope modeling limits the performance of QHM-GAN in speech modification, which involves altering f_0 or duration while preserving the short-time spectral envelope. For time-scale modification, framewise parameters can be temporally stretched or compressed via interpolation with a modified frame-shift, easily preserving the spectral envelope and yielding high-quality waveform scaling. In contrast, pitch-scale modification is more challenging, as QHM-GAN does not explicitly model the spectral envelope, limiting its ability to accurately estimate amplitudes for modified f_0 values, since the modified f_0 is usually out of the distribution of the training dataset. Informal experiments show that directly inputting modified f_0 leads to inaccurate amplitude predictions, especially for extreme pitch values, due to the lack of training references outside the original pitch distribution. To address this limitation, the spectral envelope must be estimated to constrain amplitude during pitch modification. One approach is to fix the envelope from the original speech and modify pitch accordingly, using conventional methods such as LPC or DAP. However, these methods often lack robustness and accuracy, reducing synthesis quality and limiting pitch-scale modification effectiveness.

To address the limitations discussed above, we are motivated to employ neural networks to model formant information. This approach allows for accurate estimation of amplitudes for arbitrary f_0 values, including those that fall outside the distribution of the training data. A detailed discussion of this method is provided in the next section.

4.3 ARMA Embedding

As discussed above, QHM-GAN lacks the ability to model the spectral envelope and therefore exhibits limited performance in pitch-scale speech modification. Therefore, additional LPC or DAP should be employed to take over this task. Whatever LPC or DAP, they are autoregressive

(AR). However, as introduced in Chapter I, the speech is generated by vocal tract filtering the excitation signals from the glottis. Therefore, speech signals are influenced not only by preceding samples but also by the excitation. To address this, we propose an autoregressive moving average-based (ARMA) approach to model formant characteristics. Additionally, we also embed this ARMA structure into the neural networks to improve the accuracy and efficiency of the estimation.

To overcome the limitation of QHM-GAN in resonance characteristic modeling, we enhance the generator by integrating an ARMA model. The ARMA model is capable of capturing the spectral envelope characteristics of speech, which are closely associated with the resonances of the vocal tract and are essential for maintaining the naturalness and intelligibility of modified speech.

In the enhanced framework, the generator estimates the coefficients of the ARMA model from the mel-spectrogram and fundamental frequency. The quasi-harmonic parameters are obtained from ARMA and control the fine structure of the speech signal. Specifically, the estimated amplitude is modulated by the learned spectral envelope in the frequency domain to ensure consistency with the target resonance characteristics. This allows the model to preserve the spectral shape even under pitch-scale modifications, where the harmonic structure is changed but the envelope should remain fixed.

In this section, we present a detailed formulation of the enhanced QHM-GAN architecture, describe the embedding strategy of the ARMA model within the generator, and demonstrate its advantages in various speech modification tasks.

4.3.1 ARMA Modeling

In this section, the modeling process of the autoregressive moving average model will be introduced in detail.

Linear predictive coding has long been regarded as a fundamental technique in speech processing due to its proficiency in analyzing discrete time-series signals. It models the current sample of a signal as a linear combination of its previous samples, effectively capturing the short-term spectral envelope of speech. However, the LPC model, which corresponds to a pure autoregressive process, often falls short in modeling the stochastic components of real-world signals, particularly in representing resonances with sharp anti-resonant behavior. Additionally, the speech signals are generated from filtering the excitation signals from the glottis. This indicates that considering the previous samples from itself is not enough for a speech signal. The previous samples of excitation signals are also significant for the modeling.

To address this limitation, the autoregressive moving average (ARMA) model will be introduced. Considering a system, the output signal is the result by filtering the input signal, and the system itself is the filter. Unlike LPC, which only considers the previous samples of output signals, ARMA incorporates both past output signals and past input signals, thereby offering a more

flexible framework for signal representation. The ARMA model predicts the current output $x(t)$ by combining a weighted sum of previous outputs and a weighted sum of past inputs as

$$x(t) = -\sum_{p=1}^{P} a_p x(t-p) + G\sum_{q=0}^{Q} b_q u(t-q), \tag{4.7}$$

where a_p and b_q are the coefficients of the autoregressive and moving average parts, respectively. Here, $b_0 = 1$ for normalization, G denotes the global gain of the system, and $u(t)$ is the input signal, where $x(t)$ is the output signal. The ARMA model is parameterized by the orders P and Q, which determine the memory length of the autoregressive and moving average processes. The z transform of Eq. (4.7) can be easily obtained by

$$\begin{aligned}\left(1+\sum_{p=1}^{P} a_p z^{-p}\right) X(z) &= G\left(1+\sum_{q=1}^{Q} b_q z^{-q}\right) U(z) \\ X(z) &= G\frac{1+\sum_{q=1}^{Q} b_q z^{-q}}{1+\sum_{p=1}^{P} a_p z^{-p}} U(z),\end{aligned} \tag{4.8}$$

where $X(z)$ and $U(z)$ are the z transforms of $x(t)$ and $u(t)$. Then, the transfer function of such a system, namely, the ARMA, can be modeled in the z domain as

$$H(z) = \frac{X(z)}{U(z)} = G\frac{1+\sum_{q=1}^{Q} b_q z^{-q}}{1+\sum_{p=1}^{P} a_p z^{-p}}, \tag{4.9}$$

where $H(z)$ is the transfer function. Its frequency response can be modeled as

$$H(\omega) = H(z)|_{z=e^{iw}} = \frac{X(e^{i\omega})}{U(e^{i\omega})} = G\cdot\frac{1+\sum_{q=1}^{Q} b_q e^{-i\omega q}}{1+\sum_{p=1}^{P} a_p e^{-i\omega p}}, \tag{4.10}$$

where $z = e^{i\omega}$ is considered.

In the context of speech, which is inherently a time-varying signal, the ARMA model serves as a powerful tool to model the resonant behavior of the vocal tract. By treating the excitation $u(t)$ as the quasi-periodic source from the vocal folds, and $x(t)$ as the corresponding output waveform, the ARMA model essentially acts as a framewise resonance filter. Within each analysis frame, we assume the speech is locally stationary within a frame and denote the frame-dependent ARMA parameters as a_p^l, b_q^l, and G_l for the l-th frame. Thus, the sequence-wise speech signals can be modeled by a time-varying ARMA function, which models the signal in each frame with time-varying coefficients. Next, we focus on the l-th frame.

To formulate the excitation signal at the l-th frame, we follow the harmonic assumption and express the input as a sum of sinewaves (i.e., harmonic components) with unit amplitudes:

$$u^l(t) = \sum_{k=-K}^{K} e^{i2\pi\hat{f}_k^l t}, \quad t \in [-T_l, T_l], \tag{4.11}$$

where $\hat{f}_k^l = k\hat{f}_0^l$ represents the harmonic frequency components derived from the estimated pitch $\hat{f}_0^l$ in the l-th frame.

Given this excitation, we can consider the time-varying ARMA transfer function as a time-frequency representation in the time-frequency domain, whose value at the l-th frame can be computed as:

$$H(t_l, \omega) = \frac{X(t_l, \omega)}{U(t_l, \omega)} = G_l \cdot \frac{1 + \sum_{q=1}^{Q} b_q^l e^{-i\omega q}}{1 + \sum_{p=1}^{P} a_p^l e^{-i\omega p}}. \tag{4.12}$$

This transfer function characterizes the magnitude and phase response of the vocal tract at each frame.

The amplitude and phase of each harmonic component in the output signal $x^l(t)$ can then be computed directly from $H(t_l, \omega_k)$, where $\omega_k = 2\pi\hat{f}_k^l$:

$$\hat{A}_k(t_l) = |H(t_l, \omega_k)| = |G_l| \cdot \frac{\left|1 + \sum_{q=1}^{Q} b_q^l e^{-i\omega_k q}\right|}{\left|1 + \sum_{p=1}^{P} a_p^l e^{-i\omega_k p}\right|}, \tag{4.13}$$

$$\hat{\varphi}_k(t_l) = \varphi_k^u(t_l) + \angle H(t_l, \omega_k). \tag{4.14}$$

Here, $\varphi_k^u(t_l)$ denotes the phase of the excitation signal at frame center t_l, obtained by integrating the instantaneous frequency:

$$\varphi_k^u(t_l) = \int_0^{t_l} 2\pi\hat{f}_k(u)\,\mathrm{d}u, \tag{4.15}$$

where $\hat{f}_k(u)$ is the interpolated instantaneous frequency corresponding to the k-th harmonic.

Clearly, once the ARMA coefficients are known, the amplitude and phase of each frequency component can be derived analytically using Eqs. (4.13) and (4.14). This formulation provides a principled and efficient method for speech resynthesis. Furthermore, it offers a direct pathway for speech modification: when modifying the f_0, the new harmonic components can be mapped through the same ARMA filter to compute their corresponding amplitude and phase responses. This ensures consistency of the spectral envelope during pitch modification, a feature that conventional vocoders often struggle to preserve.

To estimate the ARMA parameters, we can extend the generator structure of QHM-GAN by appending three separate convolutional heads to predict the gain G_l, the AR coefficients a_p^l, and the MA coefficients b_q^l, respectively. The output layer thus consists of 1, P, and Q channels. These parameters, once obtained, are used to compute the harmonic amplitudes and phases for speech generation. The process remains fully differentiable and can be optimized end-to-end using standard gradient descent.

In summary, the integration of ARMA into the QHM-GAN framework allows for effective modeling of the spectral envelope, thereby addressing one of the key limitations of the original QHM-GAN. The resulting model can achieve accurate and flexible speech synthesis and modi-

fication, particularly in pitch-scaling scenarios where envelope preservation is critical. Thus, we are motivated to integrate the ARMA into the neural networks.

4.3.2 Wide-band Frequency Correction via Cascaded ARMA Modeling

In the original QHM-GAN framework, phase correction is implemented via a learnable phase compensation term, which is designed to offset the error between the predicted and the actual rotation angle within each frame. This phase compensation operates on a per-frame basis and can be accumulated across frames during the frequency refinement stage. As this term is generated independently for each frame, it is advisable to constrain it within a principal value range, i.e., $[-\pi, \pi]$, to ensure stability and prevent phase wrapping issues, especially when performing wide-band frequency correction.

In contrast, phase delay arises from the frequency response of the resonance filter (e.g., ARMA), and it affects the unwrapped phase directly. Importantly, this phase delay is inherently continuous across frames and can influence both the current and adjacent frames during frequency estimation. Thus, this will somewhat hinder the performance of frequency correction.

Here, we compare phase compensation and phase delay in frequency correction with their formulations. Assuming that the frequency remains constant within each frame, the correction of frequency error Δf_k^l for the k-th harmonic at frame l can be derived from the difference in unwrapped phase between two consecutive frames:

$$\Delta f_k^l = \frac{\hat{\varphi}_k(t_l) - \hat{\varphi}_k(t_{l-1})}{2\pi(t_l - t_{l-1})} - \hat{f}_k^l, \tag{4.16}$$

where $\hat{\varphi}_k(t_l)$ is the synthesized phase at time t_l and $\hat{f}_k^l$ is the estimated instantaneous frequency at the l-th frame. Let $\Delta t = t_l - t_{l-1}$ be the constant frame shift.

First, we focus on the phase delay. If we denote the frequency correction derived from phase delay as $\Delta \check{f}_k^l$, then the cumulative correction over L frames can be expressed as:

$$\begin{aligned} \sum_{l=1}^{L} \Delta \check{f}_k^l &= \sum_{l=1}^{L} \frac{\angle H(t_l, \omega_k) - \angle H(t_{l-1}, \omega_k)}{2\pi\Delta t} \\ &= \frac{\angle H(t_L, \omega_k) - \angle H(t_0, \omega_k)}{2\pi\Delta t} \in \left[\frac{-1}{\Delta t}, \frac{1}{\Delta t}\right], \end{aligned} \tag{4.17}$$

assuming that the phase delay from the ARMA filter, i.e., $\angle H(t_l, \omega_k)$, lies within $[-\pi, \pi]$.

On the other hand, for the phase compensation term used in QHM-GAN (denoted here as $\Delta\varphi_k^l$), the corresponding cumulative frequency correction is given by:

$$\sum_{l=1}^{L} \Delta \acute{f}_k^l = \frac{\sum_{l=1}^{L} \Delta\varphi_k^l}{2\pi\Delta t} \in \left[\frac{-L}{2\Delta t}, \frac{L}{2\Delta t}\right]. \tag{4.18}$$

A direct comparison between Eq. (4.17) and Eq. (4.18) reveals that the correction range induced

by phase delay is inherently limited when $L > 2$. This limitation hinders the ability to perform wide-band frequency correction using only the ARMA phase response. To enhance the correction range, we propose extending the permissible phase delay range from $[-\pi, \pi]$ to $[-r\pi, r\pi]$, where r is a scaling factor that determines the degree of expansion.

To this end, we propose a novel mechanism by factorizing the ARMA model into a cascade of r mini ARMA modules. This is achieved by evenly dividing the AR and MA coefficients of the entire ARMA functions into r groups, denoted as $a_{j,p}$ and $b_{j,q}$ for $j = 1, \ldots, r$. The original frequency response in Eq. (4.10) is then restructured as a product of r mini-ARMA systems:

$$H(t_l, \omega) = G_l \prod_{j=1}^{r} \tilde{H}_j(t_l, \omega), \tag{4.19}$$

where each mini-model $\tilde{H}_j(t_l, \omega)$ is defined as

$$\tilde{H}_j(t_l, \omega) = \frac{1 + \sum_{q=1}^{Q/r} b_{j,q}^{l} e^{-i\omega q}}{1 + \sum_{p=1}^{P/r} a_{j,p}^{l} e^{-i\omega p}}. \tag{4.20}$$

This cascaded formulation allows each mini-ARMA model to contribute its own phase delay, and their summation yields an aggregated phase delay:

$$\hat{A}_k(t_l) = G_l \prod_{j=1}^{r} |\tilde{H}_j(t_l, \omega_k)| = G_l \prod_{j=1}^{r} \left| \frac{1 + \sum_{q=1}^{Q/r} b_{j,q}^{l} e^{-i\omega q}}{1 + \sum_{p=1}^{P/r} a_{j,p}^{l} e^{-i\omega p}} \right|, \tag{4.21a}$$

$$\angle H(t_l, \omega_k) = \sum_{j=1}^{r} \angle \tilde{H}_j(t_l, \omega_k). \tag{4.21b}$$

As a result, the total phase delay becomes unbounded by $[-\pi, \pi]$, and is instead scaled up to $[-r\pi, r\pi]$, significantly enlarging the dynamic range of frequency correction. This enables the model to perform robust wide-band refinement, especially beneficial for tasks involving large pitch shifts or expressive speech synthesis. Then, we can integrate this mechanism into QHM-GAN to improve the ability to model resonance characteristics.

4.3.3 QHARMA-GAN

Building upon the cascaded formulation of the ARMA model, we proceed to integrate this enhanced resonance modeling mechanism into a neural vocoder architecture that inherits the QHM structure. A natural and effective approach is to embed the cascaded ARMA module directly into the generator of QHM-GAN, thereby extending its modeling capacity while retaining its core advantages.

In this section, we present the complete integration of the cascaded ARMA model into the QHM-GAN framework. The resulting vocoder, referred to as **QHARMA-GAN**, is a novel archi-

tecture that unifies harmonic component modeling, deep neural estimation, and cascaded spectral envelope modeling. By leveraging the resonance modeling ability of ARMA and the frequency-continuous representation of QHM, QHARMA-GAN offers a powerful solution for high-fidelity and controllable speech synthesis.

The detailed structure, training strategy, and synthesis procedure of QHARMA-GAN will be elaborated in the following subsections.

Generator Structure: Similar to QHM-GAN, the generator of QHARMA-GAN is composed of two main components: a DNN module and a CSP module. In the case of QHM-GAN, the DNN is responsible for estimating interpretable framewise parameters, namely, the amplitude and phase compensation for each harmonic component, while the CSP module reconstructs the final waveform from these parameters. In QHARMA-GAN, a similar division of labor is adopted. The DNN module estimates the ARMA coefficients, and the CSP module utilizes these coefficients, along with the harmonic structure inferred from the f_0, to synthesize the final speech waveform.

The overall generator architecture of QHARMA-GAN is illustrated in Fig. 4.4. As shown, its structure closely mirrors that of QHM-GAN, particularly in terms of the use of MRF modules. In fact, the configuration and design of MRF modules in QHARMA-GAN are kept identical to those in QHM-GAN to ensure consistent temporal modeling capacity.

A key distinction, however, lies in the input representation and parameter prediction strategy. Unlike QHM-GAN, QHARMA-GAN no longer requires the input of f_0 into the DNN module. This is because the ARMA coefficients inherently model the frequency-dependent spectral shaping characteristics, effectively absorbing both the resonance behavior and the frequency mismatches. As a result, f_0 is only utilized within the CSP module to construct harmonic excitation components, as shown in Fig. 4.5. Consequently, the frequency-related F-CNNs used in QHM-GAN are no longer necessary, simplifying the network architecture and reducing the computational load of the DNN component.

Nonetheless, while the DNN part of QHARMA-GAN is simplified compared to QHM-GAN, the computational complexity of the CSP module increases slightly due to the additional steps involved in deriving amplitude and phase from the ARMA frequency response. Specifically, the ARMA-based spectral shaping requires the calculation of the magnitude and phase delay for each harmonic component using Eqs. (4.13) and (4.14). These framewise amplitude and phase values are then interpolated to obtain the instantaneous amplitude (via linear interpolation) and instantaneous phase (via cubic interpolation). Finally, the speech waveform is synthesized by summing all quasi-harmonic components generated from these interpolated parameters.

Despite the increased complexity in the CSP module, the overall computational efficiency of QHARMA-GAN remains comparable to that of QHM-GAN. The reduction in DNN complexity compensates for the additional operations in the CSP stage. Furthermore, since both architectures avoid upsampling and transposed convolutions, they are well-suited for real-time speech

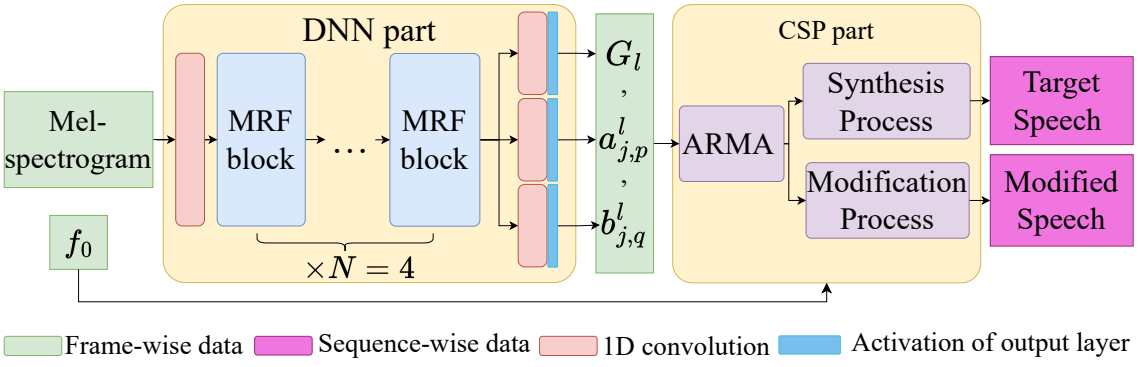


Figure 4.4: Structure of QHARMA-GAN generator, which takes mel-spectrogram as inputs and outputs framewise ARMA coefficients. The framewise amplitude and phase will be calculated by the corresponding ARMA function to generate the speech waveform along with the frequency.

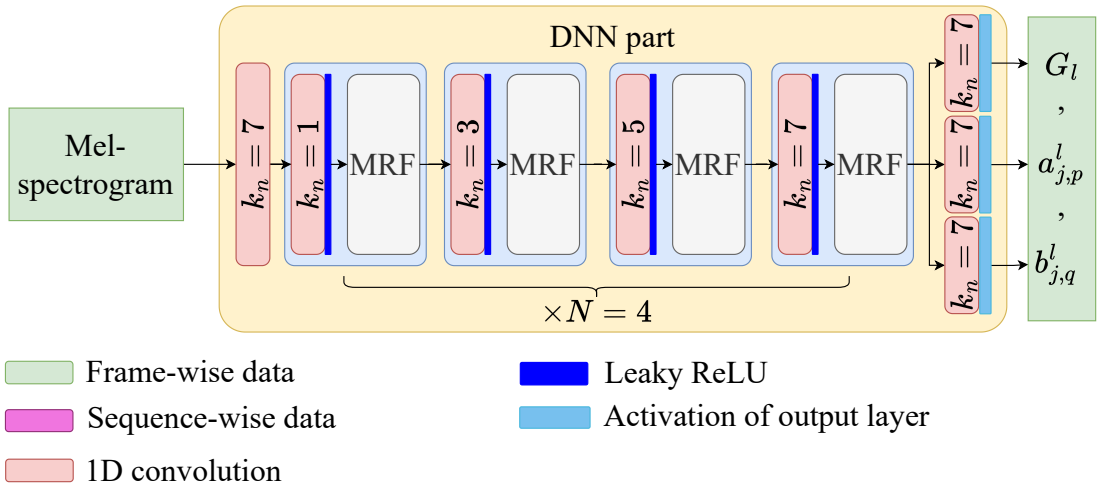


Figure 4.5: Structure of DNN part of QHARMA-GAN generator. The configurations of all MRFs in the generator are the same as those of QHM-GAN in Fig. 4.3.

Algorithm 3 Speech Synthesis based on QHARMA-GAN.

Step 1: Preprocessing and Setting
Extract the $\hat{f}_k$ and mel-spectrogram from speech $x(t)$; set sampling rate f_s, harmonic number K and frame-shift Δt; obtain $a^l_{j,p}$ and $b^l_{j,q}$ from DNN;
Step 2: Instantaneous amplitude and phase
Get framewise amplitude and phase delay by Eq. (4.19)-(4.21);
Compute unwrapped framewise phase $\hat{\varphi}_k(t_l)$ by Eq. (4.14);
Cubically interpolate $\hat{\varphi}_k(t_l)$ into $\hat{\varphi}_k(t)$;
Linearly interpolate $\hat{A}_k(t_l)$ into $\hat{A}_k(t)$;
Step 3: Generation
$\hat{x}(t) \leftarrow \sum_{k=-K}^{K} \hat{A}_k(t) e^{i\hat{\varphi}_k(t)}$;
Output: $\hat{x}(t)$

synthesis.

The complete synthesis procedure of QHARMA-GAN is summarized in Algorithm 3.

Discriminator Design: The training of QHARMA-GAN adopts a carefully designed adversarial framework to effectively supervise both voiced and unvoiced speech generation. Based on our preliminary experimental analysis, we observe that different discriminator architectures offer complementary advantages. Specifically, the multi-resolution discriminator (MRD) and multi-period discriminator (MPD) exhibit strong capability in capturing the periodic structure of speech, thereby providing effective guidance for the generator to synthesize high-fidelity voiced speech characterized by rich harmonic structures. In contrast, the multi-scale discriminator (MSD), which analyzes speech signals at different temporal resolutions, proves more effective in guiding the generator to produce natural-sounding unvoiced speech, which is inherently stochastic and lacks clear periodicity.

Motivated by these observations, QHARMA-GAN integrates all three discriminators, i.e., MRD, MPD, and MSD, into its discriminator module. This ensemble allows for comprehensive adversarial feedback, ensuring that both harmonic and stochastic components of the synthesized speech are perceptually and structurally plausible. The combination of these discriminators has been empirically shown to strike a good balance between harmonic continuity and spectral richness across different phonetic contexts.

4.3.4 Source-Filter Modeling based on QHARMA-GAN

As the introduction above states, QHARMA-GAN uses neural networks to build the ARMA model to explicitly shape the resonance characteristics, which can be considered as the filter. The framewise frequencies are used to generate the framewise phase of the excitation signal, which can be considered as the source. Thus, QHARMA-GAN also has a source-filter structure and the ability to separate the source part and filter part.

To qualitatively evaluate the capability of QHARMA-GAN in modeling the resonance characteristics of speech, we present a detailed spectral analysis based on a representative utterance. As shown in Fig. 4.6(a), the ground truth waveform of the selected utterance is depicted as the reference for subsequent analysis.

To further investigate the spectral properties, we compute and visualize both the magnitude and phase spectra of the speech signal, as well as those of the estimated ARMA filter. Specifically, Figs. 4.6 (b) and (d) illustrate the decibel-scale magnitude spectrum and wrapped phase spectrum of the ground truth speech, respectively. In contrast, Figs. 4.6 (c) and (e) depict the corresponding spectral response generated by the ARMA model inferred by QHARMA-GAN, where the autoregressive and moving average orders are set to $P = 128$ and $Q = 128$, respectively, and the number of cascaded stages is set to $r = 8$.

A visual comparison reveals that the magnitude and phase responses reconstructed by the QHARMA-GAN exhibit a smooth spectral envelope that aligns closely with the harmonic structure of the original signal. The reconstructed spectral envelope effectively captures the resonance

characteristics, which are independent of the excitation signal. In particular, even when the f_0 is arbitrarily specified, the smoothness and coherence of the generated spectral envelope remain consistent, highlighting the model's generalization capability. Thus, QHARMA-GAN can synthesize high-quality speech with ease and proficiency, even when the f_0 is out of the distribution of the training data.

These observations demonstrate that the ARMA-based spectral envelope modeled by QHARMA-GAN is capable of approximating the true resonance characteristics of speech. This further implies that QHARMA-GAN successfully separates the excitation and system components in the source-filter framework, and enables flexible, high-fidelity speech synthesis and modification.

4.4 Experimental Evaluations

4.4.1 Experimental Design and Evaluation Aspects

To comprehensively assess the performance of the proposed methods, including QHM-GAN and QHARMA-GAN, we conduct a series of experiments and compare them against several state-of-the-art neural vocoders. The evaluation focuses on multiple aspects such as speech synthesis quality, generation speed, and model efficiency. Both objective and subjective metrics are employed to quantify the synthesis performance.

The detailed experimental conditions and results are presented and discussed in the following subsections. The experiments consist of four parts.

(1) The confirmatory experiment for QHM-GAN.

Before conducting comprehensive evaluations across all methods, we first design a dedicated experiment specifically for QHM-GAN. The objective of this preliminary experiment is to verify the feasibility of integrating the QHM structure with neural network architectures, and to investigate whether the proposed QHM-GAN can simultaneously inherit the advantages of both conventional signal-processing-based approaches and modern neural vocoders. In this experiment, we employ both objective and subjective evaluation metrics to assess the synthesis quality, providing an initial validation of the effectiveness and potential of the QHM-GAN framework.

(2) The preliminary experiments to show the best vocoders in candidate methods.

First, we uses objective measurement indicators to assess the performances of all neural vocoders in terms of synthesis quality. Subsequently, the screened-out methods, including the best one of other neural vocoders and the best one between QHM-GAN and QHARMA-GAN, will be compared with the conventional vocoders, such as WORLD and QHM.

(3) The synthesis experiment of selected neural and conventional vocoders.

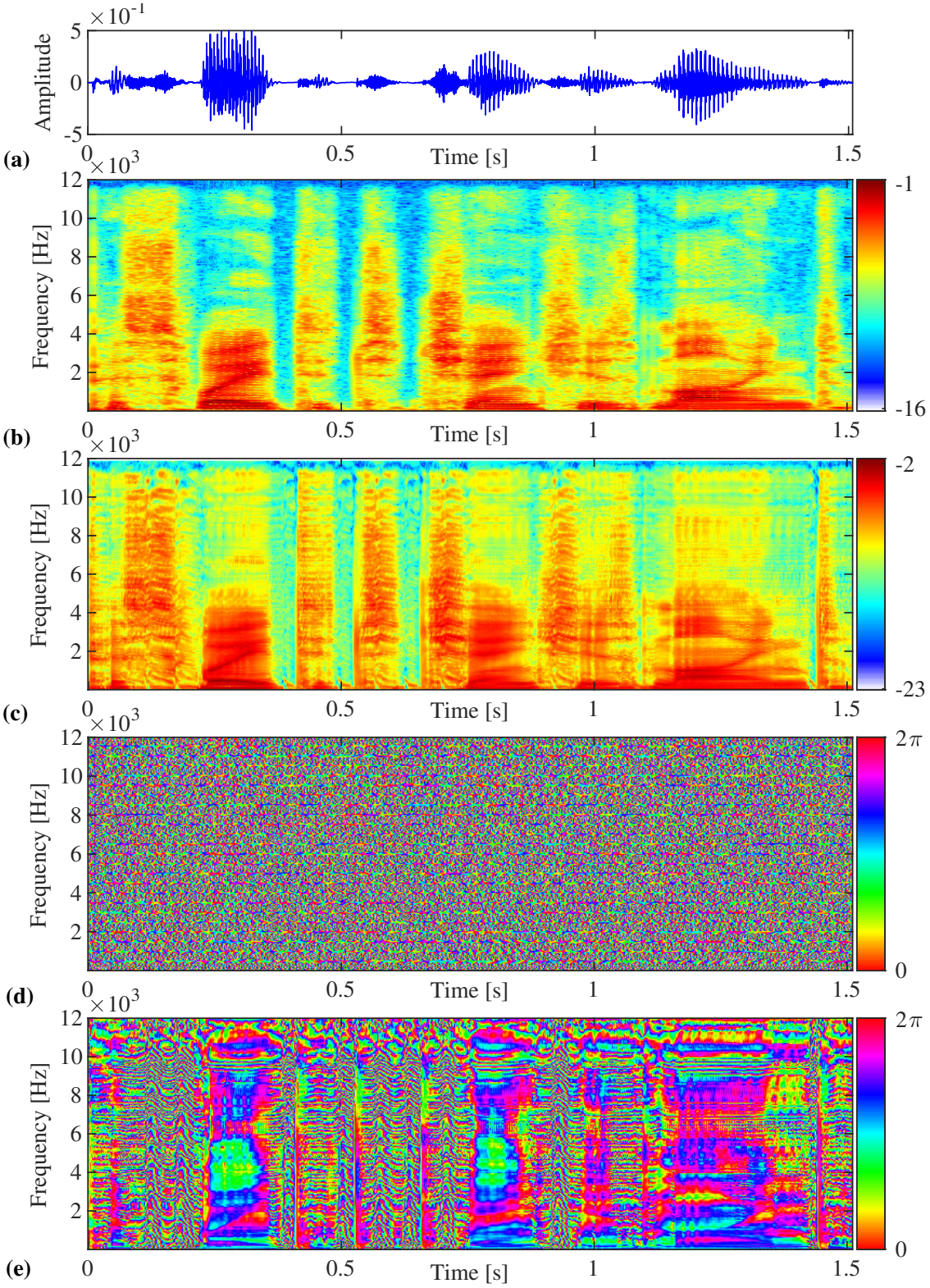


Figure 4.6: (a) Ground truth of utterance sample. Magnitude spectra of (b) ground truth and (c) corresponding ARMA response. Phase spectra of (d) ground truth and (e) corresponding ARMA response.

Second, both objective and subjective evaluation metrics are jointly employed to comprehensively assess the quality of the speech synthesized by all candidate methods. The objec-

tive metrics provide quantifiable measurements of intelligibility and spectral fidelity, while the subjective evaluations reflect human perceptual preferences and naturalness. This combined evaluation strategy ensures a thorough and balanced assessment of the synthesis performance.

(4) The time complexity and execution efficiency of QHM-GAN and QHARMA-GAN compared with baseline neural vocoders and conventional vocoders.

Third, in terms of time complexity, the actual runtime of each method, including all conventional, neural, and proposed approaches, is systematically measured under consistent experimental conditions. This comprehensive analysis highlights the trade-offs between modeling accuracy and computational efficiency, which are critical considerations for real-world applications where latency and resource constraints are of paramount importance.

(5) The generalization ability of neural vocoders.

Finally, to evaluate the generalization capability, which is a critical property for neural vocoders, we assess the performance of all candidate neural vocoders under two challenging scenarios: out-of-distribution (OOD) generalization and few-shot learning. These evaluations aim to verify whether the models can maintain synthesis quality when confronted with unseen conditions or when trained with limited data. Since the data-hungry issue is essential for real-world applications.

Through these evaluations, we aim to demonstrate that QHM-GAN and QHARMA-GAN not only possess the advantages of conventional QHM methods, such as pitch controllability and source-filter modeling, but also achieves robust and high-quality generation of the synthesized speech.

4.4.2 Experiment Conditions

Here, we introduce the experiment conditions in detail.

Optimizer: To empirically validate the effectiveness of all neural methods and to ensure the robustness of the evaluation across diverse conditions, experiments are conducted using multiple datasets under rigorously defined optimization settings. The Adam optimizer [85] is utilized throughout the training process due to its adaptive learning rate mechanism and proven effectiveness in deep neural network training.

A step-decay learning rate schedule is adopted in the optimization setup. Specifically, the initial learning rates for both the generator and discriminator are set to 0.0002. To facilitate convergence and mitigate overfitting, the learning rates are halved every 200,000 epochs. The total number of optimization iterations is fixed at 1,000,000.

Dataset: The speech utterances used in the experiments are randomly selected from three representative open-source corpora, covering both single-speaker and multi-speaker datasets: the LJSpeech dataset [88], sampled at 22.05 kHz; the VCTK corpus [95], comprising 110 speakers sampled at 24 kHz; and the Japanese Versatile Speech (JVS) corpus [96], consisting of 100 speakers sampled at 24 kHz.

For the JVS dataset, the utterances are divided into training, validation, and test sets using a ratio of 80% / 10% / 10%. In the case of the VCTK corpus, all utterances from two unseen speakers (p271 and p300) are reserved exclusively for testing and excluded from the training process. The remaining speakers in the VCTK corpus follow the same 80% / 10% / 10% split strategy as applied to the JVS and LJSpeech datasets. Finally, for the LJSpeech dataset, a ratio of 919/250/250 for training, validation, and test sets is used.

Initialization and Parameter Setting: For all neural methods, the same loss function configuration is adopted to ensure consistency in evaluation. Specifically, both the discriminator loss and generator loss are employed. The generator loss comprises the mel-spectrogram loss, feature matching loss, and adversarial loss, each weighted by the same predefined coefficients across methods. The Fast Fourier Transform (FFT) length for computing the mel-spectrogram loss is set to $N_{\text{fft}} = 2048$.

Regarding preprocessing, the frame shift for both the mel-spectrogram input and the f_0 is fixed at 8 ms. The f_0 values are extracted using the Harvest algorithm [18]. For the initialization of neural network parameters, standard Gaussian noise is applied to generate initial values.

For QHM-GAN and QHARMA-GAN, the number of harmonic components K is determined by the sampling rate of each dataset. Specifically, $K = 128$ is used for LJSpeech (22.05 kHz), while $K = 256$ is applied to both the VCTK and JVS corpora (24 kHz). Additionally, for QHARMA-GAN, the orders of the autoregressive and moving average filters are both set to 256, and the downsampling parameter r is set to 8.

Measurements: To quantitatively evaluate the performance of the proposed QHM-GAN and QHARMA-GAN, along with baseline neural and conventional methods, a series of objective and subjective metrics are employed. These metrics are designed to capture multiple aspects of system behavior, including reconstruction quality, frequency accuracy, voiced/unvoiced (V/UV) classification accuracy, intelligibility, and computational efficiency. This comprehensive evaluation framework facilitates a thorough comparison of the strengths and limitations of each candidate method.

It is important to note that a consistent notation is adopted across all metrics: an upward arrow (↑) indicates that higher values correspond to better performance, while a downward arrow (↓) signifies that lower values are preferable.

1) **V/UV Error Rate** [%] ↓: This metric quantifies the percentage of incorrect classifications

between voiced and unvoiced segments. Voiced speech refers to segments with vocal fold vibration, whereas unvoiced speech corresponds to segments generated by turbulent airflow without vocal fold activity. Voicing decisions are based on whether the f_0 value detected by Harvest [18] is zero (unvoiced) or non-zero (voiced). The V/UV rate is computed by comparing the voicing decisions between the generated and reference speech signals.

2) f_0 **RMSE** [Hz] $\downarrow$: The root mean squared error of the logarithmic f_0 values between the generated and reference speech, both detected using Harvest. This metric evaluates pitch reconstruction accuracy and pitch controllability.

3) **PESQ** $\uparrow$: The Perceptual Evaluation of Speech Quality (PESQ) assesses the perceptual quality of synthesized speech by computing the perceptual spectral distance between the generated and reference signals. Higher PESQ scores are associated with improved speech quality.

4) **UTMOS**[1] $\uparrow$: UTMOS is a neural network-based model [97] designed to predict the Mean Opinion Score (MOS) from synthesized speech signals. It serves as a non-intrusive, automated estimate of subjective naturalness and is used to screen for perceptual quality across systems.

It is worth noting that **STOI**, **MCD**, **MOS**, **RTF** were also employed to measure the performance of synthesis quality and efficiency, which were introduced in Section IV of Chapter III.

For the model candidates, HiFi-GAN, Vocos [5], DDSP [98], and hn-NSF [7] are chosen as representative neural vocoders, whereas QHM and WORLD serve as conventional CSP vocoders. These are compared with the proposed QHM-GAN and QHARMA-GAN. Additionally, a lightweight version of QHARMA-GAN, referred to as QHARMA-GAN-small, is included to explore the trade-off between model complexity and performance. A summary of all evaluated models is presented below:

1) **WORLD** [2]: A traditional CSP-based source-filter vocoder that allows flexible manipulation of acoustic features, including f_0. It decomposes speech into spectral envelope, aperiodicity, and f_0, and resynthesizes speech by combining these components through signal processing. WORLD is widely used as a baseline in vocoding and voice conversion tasks due to its interpretability and reasonable synthesis quality.

2) **QHM** [34]: A quasi-harmonic modeling method that includes a frequency correction mechanism to enhance the harmonic structure of speech. It estimates frame-wise complex amplitudes for each quasi-harmonic component, enabling adaptive frequency correction. The corrected frequencies and amplitudes are then used to synthesize high-fidelity speech, which can be further modified in time and frequency domains using Discrete Amplitude Phase (DAP) representations.

[1] `https://github.com/sarulab-speech/UTMOS22`

3) **HiFi-GAN** [4]: A widely adopted high-quality neural vocoder that synthesizes speech from acoustic features through a generator network based on progressive transposed convolutions. It employs MRF modules to capture temporal context at different scales. To improve perceptual quality, it incorporates adversarial training with MPD and MSD, guiding the generator to produce realistic and natural-sounding speech. We included HiFi-GAN as a baseline because it is one of the most widely used neural vocoders and is known for its strong performance across various tasks. Although HiFi-GAN does not condition on f_0, it serves as a representative baseline to compare with our proposed method.

4) **Vocos** [5]: An efficient neural vocoder that replaces conventional upsampling modules with ConvNeXt blocks to directly predict complex-valued spectrograms. The waveform is reconstructed using inverse Short-Time Fourier Transform (iSTFT). MPD and MRD are used as discriminators during training. In addition, to test if Vocos is able to modify f_0, we introduced f_0 conditioning to enable pitch modification, which is proposed in [99], and ensure a fair comparison with our f_0-conditioned models.

5) **DDSP** [98]: A neural vocoder based on a sinusoidal signal model with explicit f_0 input. It decomposes audio into harmonic and noise components, which are parameterized by interpretable features such as amplitude, phase, and spectral shape. The differentiable DSP modules synthesize the waveform by summing the sinusoidal and filtered noise signals. This model enables fine-grained pitch control and interpretable synthesis through structured priors.

6) **hn-NSF** [7]: A neural source-filter vocoder that synthesizes speech using a harmonic-plus-noise excitation signal derived from f_0 priors. The excitation is passed through a learnable filter, which adapts its spectral response to produce the final waveform. During training, adversarial supervision is applied using HiFi-GAN discriminators, helping the model generate high-quality, natural speech. This architecture supports direct pitch modification and has been successfully applied in pitch-editable synthesis tasks.

7) **QHM-GAN**: Our proposed model that integrates CSP-based QHM with neural networks for flexible and high-quality speech synthesis. The architecture consists of four MRF modules, each containing three convolutional kernels with dilation rates of 1, 3, and 5. A neural network maps mel-spectrograms to frame-wise complex amplitudes and phase compensation terms, which are then passed to the QHM synthesizer for waveform generation. Adversarial training is applied using MRD and MPD to enhance the fidelity and realism of the output.

8) **QHM-GAN-S**: A variant of QHM-GAN designed to investigate the effect of different input representations. Instead of mel-spectrograms, it uses spectral envelopes extracted by WORLD as input to the neural network. This enables direct comparison between conventional and learned spectral representations in QHM-based synthesis.

9) **QHM-GAN-small**: A lightweight version of QHM-GAN targeting reduced model size and improved computational efficiency. It contains three MRF modules, each composed of three convolutional layers with a fixed dilation rate of 1. Despite its compactness, adversarial training with MRD, MSD, and MPD is retained to preserve synthesis quality.

10) **QHARMA-GAN**: Our proposed extension of QHM-GAN that incorporates autoregressive moving average (ARMA) modeling for better spectral envelope representation. The architecture remains similar to QHM-GAN with four MRF modules, but instead of directly predicting complex amplitudes, the network maps mel-spectrograms to ARMA coefficients. These coefficients control the synthesis process, providing explicit control over spectral shaping. MRD, MSD, and MPD are employed to enhance both stability and perceptual quality.

11) **QHARMA-GAN-small**: A compact version of QHARMA-GAN optimized for efficiency. It comprises three MRF modules with fixed dilation rates, similar to QHM-GAN-small, and predicts ARMA parameters from mel-spectrograms. MRD, MSD, and MPD are adopted during training to ensure quality is maintained despite reduced model capacity.

This experimental design facilitates a comprehensive evaluation of QHM-GAN and QHARMA-GAN under realistic and diverse acoustic scenarios. Furthermore, the structured optimization protocol and principled parameter initialization enhance both the reproducibility and the reliability of the experimental results.

4.4.3 Confirmatory Experiment for QHM-GAN

As mentioned earlier in Section 4.2.4 in this Chapter, we carried out a preliminary experiment on QHM-GAN to examine its potential benefits. In this section, we present the detailed results and analysis of this experiment, which confirm and extend the initial observations. The experiment was based on the comparison between QHM-GAN and some widely used methods (including WORLD, QHM, HiFi-GAN, and DDSP), where all methods were trained using LJSpeech [88]. We also use different types of QHM-GAN as the candidates, i.e., QHM-GAN with spectral envelope as the input (QHM-GAN-S) and QHM-GAN with a simplified DNN structure (QHM-GAN-small).

To evaluate the generation efficiency, we employed the real-time factor (RTF) on both a single CPU and a GPU as the primary metric. For assessing intelligibility and perceptual quality, we adopted the short-time objective intelligibility (STOI) [93], mel-cepstral distortion (MCD), and mean opinion scores (MOS). The MOS evaluation involved 15 human subjects, each of whom was asked to rate 12 utterances generated by each method.

Table 4.1 summarizes the quantitative results. QHM achieves the best objective quality, with the highest STOI and lowest MCD, owing to its direct waveform modeling and adaptive frequency ad-

Table 4.1: Results of objective and subjective evaluations. The MOS of the ground truth samples was 3.96 ± 0.09.

	WORLD	QHM	HiFi-GAN	DDSP	QHM-GAN	QHM-GAN-S	QHM-GAN -small
STOI ↑	0.959	**0.984**	0.968	0.947	0.972	0.973	0.970
MCD [dB] ↓	2.815	**1.670**	2.975	3.486	2.747	2.440	2.842
MOS ↑	3.74 ± 0.06	$\mathbf{3.86 \pm 0.11}$	3.78 ± 0.09	3.27 ± 0.14	3.81 ± 0.09	3.73 ± 0.12	3.80 ± 0.14
RTF(CPU) ↓	2.265	50.245	0.909	0.789	0.888	0.891	**0.512**
RTF(GPU) e-3 ↓			14.5	6.6	5.8	5.8	**1.9**

justment. QHM-GAN follows closely, with QHM-GAN-S showing favorable performance among its variants. Compared to HiFi-GAN, DDSP, and WORLD, all QHM-GAN variants achieve superior MCD and STOI, confirming the effectiveness of the proposed method. In perceptual evaluation, QHM attains the highest MOS, followed by QHM-GAN; notably, QHM-GAN-3-1 provides competitive quality with a substantially smaller model size. HiFi-GAN achieves high MOS but introduces frequency distortions, whereas QHM-GAN produces smoother and more natural speech by leveraging continuous frequency trajectories.

In terms of efficiency, QHM-GANs outperform other neural vocoders on a single GPU. QHM-GAN-3-1 further reduces computational costs and achieves near real-time CPU performance (RTF = 0.512). In contrast, QHM shows the highest RTF due to frame-by-frame analysis, making it unsuitable for real-time applications.

In summary, the experimental results demonstrate that QHM-GAN can simultaneously inherit the strengths of neural networks and QHM. Nevertheless, the limitations of QHM still remain, degrading the performance of QHM-GAN in terms of synthesis. Besides, the lack of resonance characteristic modeling inevitably limits the performance of speech modification, particularly from the perspective of pitch-scale modification. Therefore, the ARMA model was employed to be embedded into the QHM-GAN for achieving the resonance characteristic modeling and improving the quality of pitch-scale speech modification.

4.4.4 Preliminary Experiments for Baseline Screening

Based on the findings of the preliminary experiment of QHM-GAN in Section 4.1.4, it is evident that the QHM structure can be effectively integrated with neural networks to form a novel framework for neural vocoders. Building upon this validation, we proceed to conduct comparative evaluations between our proposed methods and other existing neural vocoders. Given the rapid emergence of numerous neural vocoders in recent years, it becomes impractical to perform exhaustive comparisons with all existing methods, particularly when subjective evaluations such

Table 4.2: Results of objective and subjective evaluations for speech synthesis. The UTMOS values of the ground truth samples for VCTK and JVS were 4.04 and 3.63, respectively.

Metric	Dataset	HiFi-GAN	Vocos	hn-NSF	QHM-GAN	QHARMA-GAN
V/UV rate [%] ↓	VCTK	**11**	**11**	**11**	13	**11**
	JVS	11	**9**	**9**	14	**9**
f_0 RMSE [Hz] ↓	VCTK	**0.06**	**0.06**	**0.06**	**0.06**	**0.06**
	JVS	0.11	0.11	0.10	**0.09**	**0.09**
PESQ ↑	VCTK	3.14	**3.15**	2.91	3.00	3.14
	JVS	3.29	**3.42**	3.23	3.29	3.26
MCD ↓	VCTK	**3.61**	3.62	4.21	4.14	4.09
	JVS	3.6	**3.45**	3.91	4.08	4.00
UTMOS ↑	VCTK	**3.91**	3.89	3.76	3.50	3.76
	JVS	3.16	3.25	3.21	3.15	**3.30**

as MOS tests are involved, which impose significant cognitive load on human raters.

To mitigate the burden on listeners while maintaining fairness and comprehensiveness in evaluation, a two-stage evaluation strategy is adopted. In the first stage, we rely solely on objective metrics to preliminarily screen out the top-performing models from among all neural vocoders. This allows us to identify the best-performing baseline from existing neural methods as well as our most effective proposed model. In the second stage, we focus on a more detailed comparison involving subjective assessments, where the selected neural vocoders are compared against conventional signal-processing-based methods, such as WORLD and QHM.

This subsection presents the results of the first-stage screening, where the objective performance of all neural vocoders is evaluated on the VCTK and JVS datasets. The results are summarized in Table 4.2. HiFi-GAN and Vocos show comparable performance, with minor differences across metrics and datasets, whereas hn-NSF consistently underperforms in both mel-cepstral distortion (MCD) and utterance-level mean opinion score (UTMOS), reflecting lower synthesis quality from spectral and perceptual perspectives.

Among the proposed models, QHARMA-GAN achieves the best performance, attaining the highest UTMOS on the JVS dataset while maintaining strong results across both corpora. Considering HiFi-GAN's prevalence as a standard benchmark in recent literature and its performance being close to Vocos, it is chosen as the representative baseline for subsequent comparisons with QHARMA-GAN in the main synthesis evaluation.

4.4.5 Evaluations of Synthesis Quality

According to the results discussed in the previous subsection, it is evident that HiFi-GAN serves as an appropriate baseline among neural vocoders, while QHARMA-GAN consistently outperforms QHM-GAN in various aspects. Building upon this foundation, we conduct a more comprehensive evaluation in this subsection to investigate the synthesis performances of all methods, including both conventional signal-processing-based methods and neural vocoders. In addition to widely used objective metrics, we employ the Mean Opinion Score (MOS) to assess the perceptual quality of the synthesized speech, offering a more holistic understanding of each method's performance.

The results of the main experiment are summarized in Table 4.3, which includes both objective and subjective evaluation metrics. From the objective standpoint, QHM achieves the best overall performance, as evidenced by the highest PESQ score and the lowest f_0 root mean square error (RMSE). This superior performance is attributed to QHM's ability to directly model the speech waveform while adaptively correcting the frequency of each harmonic component. Consequently, QHM excels at capturing complex amplitude structures and reconstructing speech waveforms with high fidelity. However, QHM exhibits a relatively high voiced/unvoiced (V/UV) error rate, which stems from its limited frequency correction range. When significant frequency mismatches are present, QHM often requires multiple iterative refinements for accurate correction [73]. This issue is particularly pronounced when analyzing unvoiced segments, where the corrected frequencies may still exhibit harmonic-like patterns. These patterns can mislead pitch detectors into falsely classifying unvoiced segments as voiced, thus increasing the V/UV error.

In contrast, HiFi-GAN demonstrates slightly lower PESQ scores compared to QHM but avoids the iterative correction process. QHARMA-GAN, on the other hand, achieves the lowest V/UV error among all evaluated methods, although it yields intermediate values in terms of PESQ and f_0 RMSE. This suggests that QHARMA-GAN strikes a balance between frequency modeling and speech quality.

Turning to the subjective MOS results, QHARMA-GAN attains the highest MOS score, surpassing even the well-established QHM and HiFi-GAN. This outcome highlights the perceptual superiority of QHARMA-GAN. One major reason for QHM's lower MOS is its difficulty in adjusting frequency trajectories for unvoiced segments. The harmonic patterns in these segments degrade the perceived naturalness of unvoiced speech. It is worth noting that the PESQ scores are higher than those of other methods, which is inconsistent with the MOS values. This is caused by that QHM methods not only model the magnitudes but also the phase, leading to an extremely close waveform to the ground truth without phase delays, while the other methods only model the magnitude, even though they can achieve higher-quality synthesis, reducing the PESQ scores. While HiFi-GAN produces waveforms directly in the time domain, this method does not ensure the frequency stability and continuity required for high-quality synthesis. As a result, some generated samples suffer from audible frequency distortions.

Table 4.3: Results of objective and subjective evaluations. The MOS values of the ground truth samples for VCTK and JVS datasets were 4.27 ± 0.022 and 3.88 ± 0.038, respectively.

Metric	Dataset	WORLD	QHM	HiFi-GAN	QHARMA-GAN	QHARMA-GAN -small
V/UV rate [%] ↓	VCTK	**11**	13	**11**	**11**	12
	JVS	12	15	11	**9**	10
f_0 RMSE [Hz] ↓	VCTK	0.06	**0.03**	0.06	0.06	0.06
	JVS	0.11	**0.05**	0.11	0.09	0.09
PESQ ↑	VCTK	2.54	**3.45**	3.14	2.72	2.49
	JVS	2.97	**3.61**	3.29	3.26	3.18
MOS ↑	VCTK	4.12 ± 0.025	4.07 ± 0.035	4.08 ± 0.028	$\mathbf{4.21 \pm 0.025}$	4.10 ± 0.054
	JVS	3.67 ± 0.042	3.40 ± 0.047	3.64 ± 0.030	$\mathbf{3.80 \pm 0.029}$	3.68 ± 0.030

In contrast, QHARMA-GAN inherently ensures frequency smoothness and stability during synthesis by leveraging the ARMA-based formulation. This leads to a significant reduction in frequency-related distortions. Additionally, QHARMA-GAN is capable of adaptively modulating the phase delay in unvoiced segments based on the mel-spectrogram input. This process introduces a quasi-harmonic structure in these segments, which improves the perceived naturalness of unvoiced speech. Notably, even the simplified variant, QHARMA-GAN-small, which employs a more compact neural architecture, achieves a MOS score only marginally lower than that of the full version. This finding further validates the effectiveness of the proposed hybrid framework that combines interpretable signal modeling with neural generative capabilities.

4.4.6 Evaluations of Model Efficiency

According to the results presented in the previous subsection, both QHARMA-GAN and its simplified variant (QHARMA-GAN-small) demonstrate synthesis quality that is comparable to, and in certain cases superior to, state-of-the-art neural vocoders such as HiFi-GAN and Vocos. These findings suggest that the proposed methods are competitive not only in perceptual and objective metrics but also in robustness across datasets.

To further validate the practicality of the proposed models, this subsection evaluates the efficiency of all candidate vocoders in terms of real-time factor (RTF). We separately compare the QHM-GAN and QHARA-GAN with conventional vocoders (WORLD and QHM) and neural vocoders (HiFi-GAN, Vocos, and hn-NSF). First, in Table 4.4, the RTF values for both analysis and synthesis stages are reported, enabling a detailed comparison of computational performance across different processes. Second, in Table 4.5, the RTFs of all neural vocoders are given.

For the first stage, to ensure a fair comparison between conventional methods, the time required for f_0 detection is excluded from the RTF computation for WORLD and QHM. This is because

modern neural vocoders typically do not perform explicit f_0 estimation during inference, and including this step would create a bias against conventional methods due to its high computational cost. Additionally, WORLD and QHM are fed by the speech waveform, whereas the neural methods are fed by the mel-spectrogram. Thus, to further improve the fair comparison, the process of calculating the mel-spectrogram from the speech waveform is included for all neural methods.

As shown in Table 4.4, QHM exhibits the highest analysis RTF among all methods, primarily due to its frame-by-frame least-squares fitting, which incurs significant computational overhead. Similarly, WORLD also exhibits relatively high analysis latency as a result of its frame-wise spectral envelope estimation. On the contrary, QHM-GAN achieves a faster analysis RTF of 0.133, while QHARMA-GAN yields a slightly higher value of 0.139. Despite this, the simplified structure of QHARMA-GAN-small leads to a noticeable reduction in analysis time, confirming that architectural simplifications can effectively mitigate computational demands.

For the synthesis stage, QHARMA-GAN achieves a lower RTF (0.048) than conventional vocoders such as WORLD and QHM. This improvement is largely due to the optimized synthesis pipeline, in which frequency interpolation is omitted and replaced with direct phase interpolation, thus reducing the computational complexity. However, QHM methods need to interpolate amplitude and frequency at the first stage, and then conduct a special integration of the frequency, in a frame-by-frame scheme, to obtain the instantaneous phase. That's the main reason that the RTF of synthesis for QHM methods is much higher than QHM-GAN and QHARMA-GAN. Therefore, for the speech modeling task, such hybrid methods can efficiently and accurately extract the parameters and synthesize the speech.

For the second stage, we focus on Table 4.5, which shows that Vocos achieves the lowest analysis RTF of 0.040, significantly outperforming both HiFi-GAN and hn-NSF. This efficiency can be attributed to its streamlined architecture, which eliminates the use of transposed convolutions and directly synthesizes complex spectrograms, which are subsequently converted to waveforms via an efficient iSTFT module. Although QHM-GAN and QHARMA-GAN are slightly slower than HiFi-GAN in synthesis speed, they remain significantly faster than hn-NSF. The higher synthesis RTF of hn-NSF stems from its reliance on upsampled input sequences and the additional computations introduced by its source excitation module.

It is also worth noting that while QHARMA-GAN benefits from improved analysis efficiency by avoiding upsampling during analysis, it introduces additional synthesis overhead due to the requirement of interpolating both instantaneous amplitude and phase. This trade-off is effectively addressed by QHARMA-GAN-small, which simplifies the DNN architecture and accelerates the overall inference pipeline. Further reductions in computational cost could potentially be achieved by continuing to streamline the DNN architecture, especially in the analysis component.

In summary, QHARMA-GAN strikes a balance between synthesis quality and computational efficiency, demonstrating that the proposed quasi-harmonic ARMA framework is not only ef-

Table 4.4: RTF comparison with conventional methods computed on a single Intel Xeon Gold 6230.

RTF ↓	Analysis	Synthesis	Overall
WORLD	0.727	0.406	1.133
QHM	23.246	1.118	24.364
QHM-GAN	0.133	**0.045**	0.180
QHARMA-GAN	0.139	0.048	0.187
QHARMA-GAN-small	**0.034**	0.049	**0.084**

Table 4.5: RTF comparison with neural methods computed on a single Intel Xeon Gold 6230.

Methods	HiFi-GAN	Vocos	hn-NSF	QHM-GAN	QHARMA-GAN	QHARMA-GAN-small
RTF ↓	0.153	**0.040**	0.192	0.179	0.187	0.084

fective but also practical for real-world applications where latency and throughput are critical considerations.

4.4.7 Evaluations of Generalization Ability

The generalization capability of neural vocoders is a critical aspect for their practical deployment. Most neural vocoders, including HiFi-GAN, tend to be data-hungry, requiring large amounts of training data to prevent overfitting. To evaluate the generalization ability of various vocoders, we conduct two sets of experiments: out-of-distribution (OOD) evaluation and few-shot learning.

Out-of-Distribution Evaluation

In the OOD setting, the generalization of vocoders trained on the JVS corpus (Japanese speech) is evaluated using a distinct task: singing voice synthesis. Singing samples are taken from the OpenSinger dataset [100], with songs in Mandarin and Cantonese sung by male and female vocalists. To ensure gender and linguistic diversity, 20 songs comprising 727 utterances were randomly selected for evaluation.

Table 4.6 summarizes the objective evaluation results. Among all models, QHARMA-GAN achieves the best performance in most metrics, particularly in f_0 RMSE and UTMOS, indicating superior pitch stability and perceptual quality. Although Vocos achieves slightly better PESQ and MCD scores, its notably higher f_0 RMSE suggests instability in pitch trajectory. UTMOS values further confirm that Vocos and HiFi-GAN often suffer from frequency jitter and phase discontinuities, resulting in unnatural synthesized singing voices. This issue is especially pronounced in high-frequency regions.

Table 4.6: Results of objective evaluations for OOD evaluation. The UTMOS value of the ground truth was 2.42.

Metric	HiFi-GAN	Vocos	hn-NSF	QHM-GAN	QHARMA-GAN
V/UV rate [%] ↓	8	**7**	8	8	**7**
f_0 RMSE [Hz] ↓	0.16	0.28	0.16	**0.11**	**0.11**
PESQ ↑	2.54	**3.02**	2.68	2.87	2.94
MCD ↓	6.42	**5.78**	6.44	6.80	5.96
UTMOS ↑	1.99	2.04	2.05	2.00	**2.20**

hn-NSF benefits from the explicit use of f_0 as a prior and exhibits more stable pitch patterns than HiFi-GAN and Vocos. However, it still struggles to model accurate f_0s and spectral envelopes, particularly in singing voices with extreme pitch values. In contrast, QHM-GAN and QHARMA-GAN adopt an interpolation-based instantaneous phase reconstruction strategy, enabling them to produce smooth and continuous frequency trajectories. The use of cubic interpolation helps mitigate phase discontinuities caused by phase compensations or ARMA phase delay adjustments, thereby preserving the naturalness of synthesized audio.

Notably, HiFi-GAN, Vocos, and hn-NSF all show poor performance on soprano voices with extremely high f_0 values due to the absence of such pitch ranges in the training data. While hn-NSF includes pitch priors, its inability to capture fine-grained resonance characteristics limits its performance in such cases. Same limitation happens in QHM-GAN, since QHM-GAN cannot fully get rid of the black-box nature, particularly in terms of the spectral envelope modeling. Thus, once the f_0 is outside the training distribution, the synthesis quality degrades. In contrast, QHARMA-GAN, leveraging the QHM-based architecture, effectively generalizes to these challenging scenarios, preserving harmonic structures and producing stable output even under pitch conditions far from the training distribution.

To visually support this observation, Fig. 4.7 displays the spectrograms of a soprano utterance synthesized by different methods. HiFi-GAN, Vocos, and hn-NSF only reproduce low-frequency harmonics and fail to generate high-frequency components, resulting in audible distortions. On the other hand, the inaccurate estimation of amplitudes and phase compensations in QHM-GAN gives rise to artifacts similar to aliasing, ultimately distorting the perceived quality of the synthesized singing voice, especially in the latter part of the spectrogram. In contrast, QHARMA-GAN generates harmonics across the full frequency range, with smooth frequency curves that reflect higher naturalness and pitch fidelity.

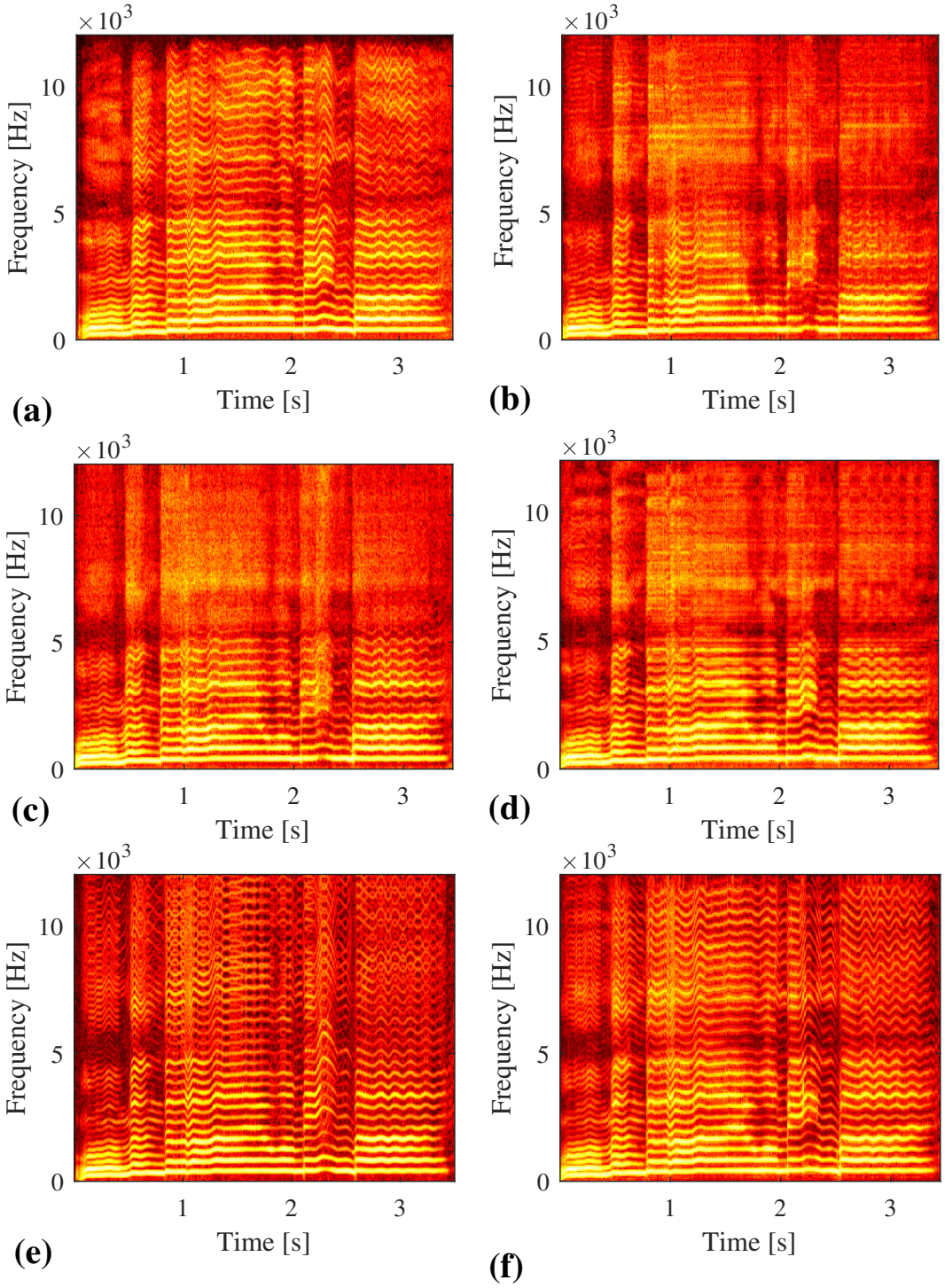


Figure 4.7: The spectrograms of the soprano voice generated by (a) ground truth (b) HiFi-GAN, (c) Vocos, (d) hn-NSF, (e) QHM-GAN, and (f) QHARMA-GAN.

Few-Shot Learning

We further assess the few-shot learning ability by training HiFi-GAN and QHARMA-GAN on a limited subset of the LJSpeech dataset. Specifically, 1419 utterances are split into training /

Table 4.7: Results of objective and subjective evaluations of small LJSpeech. The MOS of the ground truth samples was 3.96 ± 0.018.

Metric	WORLD	QHM	HiFi-GAN	QHARMA-GAN	QHARMA-GAN-small
V/UV rate [%] ↓	11	10	**9**	**9**	**9**
f_0 RMSE [Hz] ↓	0.14	**0.08**	0.12	0.09	0.09
PESQ ↑	2.64	**3.63**	3.13	3.18	3.18
MOS ↑	3.63±0.020	3.86±0.029	3.53±0.016	3.85±0.024	**3.90±0.022**

validation / test sets with a 919 / 250 / 250 ratio. Table 4.7 presents the evaluation results under this low-resource setting.

Among all methods, QHM yields the best performance in most objective metrics, benefiting from its direct waveform modeling and adaptive frequency correction. However, its V/UV rate remains higher, as it lacks explicit mechanisms to model unvoiced segments. In contrast, QHARMA-GAN and HiFi-GAN achieve the lowest V/UV rates due to their superior ability to capture unvoiced speech characteristics.

Despite using a small training set, QHARMA-GAN achieves a MOS score comparable to QHM, while QHARMA-GAN-small even slightly outperforms all methods in subjective quality. HiFi-GAN, on the other hand, exhibits the worst performance in both objective and subjective measures, largely due to overfitting caused by insufficient training data, which leads to frequency distortions and unnatural prosody.

These results indicate that QHARMA-GAN possesses strong generalization capabilities under data-scarce conditions. The hybrid design combining DNNs with the interpretable quasi-harmonic model effectively reduces the dependency on large training corpora. Human auditory perception is highly sensitive to frequency discontinuities, and any instability can significantly degrade speech quality. QHARMA-GAN's use of sine-based synthesis ensures inherently smooth frequency trajectories. Although the phase delays introduced by the ARMA structure may disrupt phase continuity, the subsequent cubic interpolation step ensures continuity in both phase and frequency, yielding natural and stable synthesized speech.

4.5 Summary

In this chapter, we proposed a novel hybrid neural vocoder framework, QHARMA-GAN, to address the limitations of conventional CSP-based vocoders, such as poor robustness and low synthesis quality, as well as the instability of f_0 control in existing neural vocoders. The proposed QHARMA-GAN bridges deep neural networks (DNNs) and quasi-harmonic modeling (CSP) via ARMA-based resonance characteristic modeling.

QHARMA-GAN comprises two components: a DNN module that estimates ARMA coefficients to capture resonance structures, and a CSP module that synthesizes or modifies speech based on these coefficients. First, the robustness of DNNs is utilized to alleviate the quality degradation resulting from the limited resonance modeling in conventional CSP algorithms. Second, we introduce a novel CSP-based synthesis algorithm that efficiently reconstructs speech waveforms by interpolating only amplitude and phase, enabling fast and high-fidelity synthesis.

By integrating the strengths of both DNNs and CSP, QHARMA-GAN effectively exploits the quasi-harmonic structure of speech to support high-quality synthesis and flexible parameter manipulation without sacrificing stability. Experimental results demonstrate that QHARMA-GAN produces speech with smoother frequency trajectories than HiFi-GAN. Furthermore, the model achieves competitive computational efficiency due to its non-upsampling architecture and simplified DNN structure, supporting real-time processing and making it suitable for deployment in tasks such as text-to-speech (TTS) and voice conversion.

Notably, unlike mel-spectrogram-based vocoders, QHARMA-GAN requires an additional f_0 prediction module when applied in TTS systems. Nonetheless, the proposed framework offers improved interpretability and controllability, making it a promising candidate for applications involving emotional TTS and prosody editing in future research.

Chapter 5

Speech Modification in Quasi-Harmonic Framework

5.1 Introduction

After extracting the quasi-harmonic parameters, the speech signal cannot only be accurately resynthesized, but also flexibly modified in terms of both pitch and duration. Speech modification plays a critical role in various speech processing applications, such as voice conversion, prosody editing, emotional speech synthesis, and singing voice manipulation. This chapter delves into the fundamental concepts and practical implementations of pitch-scaling and time-scaling, emphasizing their role within quasi-harmonic modeling frameworks.

Conventional vocoder-based methods often suffer from artifacts and loss of naturalness due to inaccurate modeling of phase or spectral fine structure during modification. In contrast, quasi-harmonic modeling provides a more interpretable and structured representation of speech signals by explicitly modeling amplitude and phase trajectories of harmonic components. This makes it particularly well-suited for high-quality and precise speech modification.

In the following sections, we first introduce the definition of speech modification and the corresponding applications in the real world, including time-scale modification and pitch-scale modification. Secondly, we provide detailed implementation strategies for the methods within the QHM frameworks, including conventional QHM methods, QHM-GAN, and QHARMA-GAN. We also discuss the advantages and limitations of each method under different application scenarios, such as extreme pitch shifts, tempo changes, and cross-lingual voice adaptation. Finally, we present a comprehensive evaluation of these methods through both objective metrics (e.g., pitch accuracy, spectral distortion) and subjective listening tests (e.g., naturalness, intelligibility). The results confirm that quasi-harmonic modeling not only supports high-quality speech synthesis but also enables flexible and interpretable modification capabilities, making it a promising framework for controllable speech generation in modern speech processing systems.

5.2 Speech Modification

Speech modification refers to the process of altering the f_0 or the duration of a speech signal while preserving its short-time spectral envelope characteristics, commonly known as resonance characteristics or timbre. These modifications are essential for various speech applications, such as prosody adjustment in TTS, expressive speech synthesis, singing voice transformation, and speaker identity preservation during voice conversion. Within the quasi-harmonic modeling framework, speech is represented in terms of amplitude, frequency, and phase trajectories of harmonic components, which provides a powerful and interpretable basis for performing high-quality pitch and time modifications.

Two primary types of speech modification are considered in this context: time-scale modification and pitch-scale modification, both of which are intrinsically linked to the quasi-harmonic parameters.

5.2.1 Time-scale Speech Modification

Time-scale modification refers to the adjustment of speech duration without altering its perceived pitch or timbral characteristics. Within the quasi-harmonic modeling framework, this is accomplished by manipulating the temporal alignment of framewise quasi-harmonic parameters, namely, the amplitude and phase trajectories, via interpolation according to a modified frame-shift schedule. For instance, decreasing the frame-shift yields a compressed (faster) signal, whereas increasing the frame-shift results in an expanded (slower) signal. Because the instantaneous frequency trajectories remain unchanged, the harmonic structure and spectral envelope are preserved, ensuring prosody adjustment while maintaining formant structure and speaker identity. Thus, QHM allows time-scaling to be executed without introducing typical artifacts such as spectral smearing, pitch fluctuation, or prosodic discontinuities.

5.2.2 Pitch-scale Speech Modification

Pitch-scale modification, on the other hand, involves altering the f_0 contour of speech, thereby changing the spacing and absolute positions of harmonic components. Unlike time-scaling, pitch-scaling directly impacts the alignment between harmonic frequencies and the spectral envelope, which, if not handled properly, can lead to unnatural timbre, aliasing, or even a loss of speaker identity, since the spectral envelope contains both speech and speaker information. To address this, the QHM framework incorporates explicit modeling of resonance characteristics, typically represented by a smooth spectral envelope or an all-pole filter. This allows the system to re-estimate the amplitude and phase of each harmonic component based on its new frequency location under the modified f_0 trajectory. Furthermore, smoothness in both amplitude and phase trajectories is maintained through interpolation, which is critical for avoiding phase discontinuities and preserving

the naturalness of the synthesized speech.

In summary, the QHM framework offers a principled and interpretable approach for high-fidelity speech modification. Time-scaling is realized through temporal interpolation of quasi-harmonic parameters, while pitch-scaling resynthesizes the harmonic structure guided by the preserved spectral envelope, enabling flexible and high-quality control over speech attributes.

5.3 Core Principles and Implementation Strategies

After establishing the fundamental concepts and objectives of speech modification, we now proceed to describe the core principles and implementation strategies in detail.

Based on how the modification is carried out, speech modification methods based on QHM can be generally divided into two categories:

1. Post-modification using extracted quasi-harmonic parameters:

   In this approach, the quasi-harmonic parameters (i.e., amplitudes, frequencies, and phases) are first extracted from the original speech signal. These parameters are then manually or algorithmically modified to achieve the desired pitch or duration changes. For example, time-scaling can be realized by stretching or compressing the temporal axis of the amplitude and phase trajectories, while pitch-scaling can be achieved by proportionally shifting the harmonic frequencies and adjusting phases and amplitudes to maintain continuity and the spectral envelope. The modified parameters are then used to resynthesize the speech signal. This method provides fine-grained control and interpretability but may require additional signal processing (e.g., LPC) to avoid distortion.

2. Direct generation of modified speech via neural networks:

   This approach utilizes neural networks to directly estimate the quasi-harmonic parameters that correspond to the target pitch or time scale. In other words, instead of first extracting and then modifying, the neural model is conditioned on the desired modification parameters (e.g., target pitch contour or duration factor) and learns to generate the corresponding amplitude and phase patterns. This enables an end-to-end pipeline for speech modification and is particularly suitable for real-time or highly variable applications. However, this method relies on the generalization capability of the neural model and may require more data for training or fine-tuning under different conditions.

Obviously, “Post-modification based on extracted quasi-harmonic parameters” corresponds to the conventional QHM methods, since they are only able to extract the parameters for the original speech signal and need the manual manipulation of the QHM parameters, where “Neural

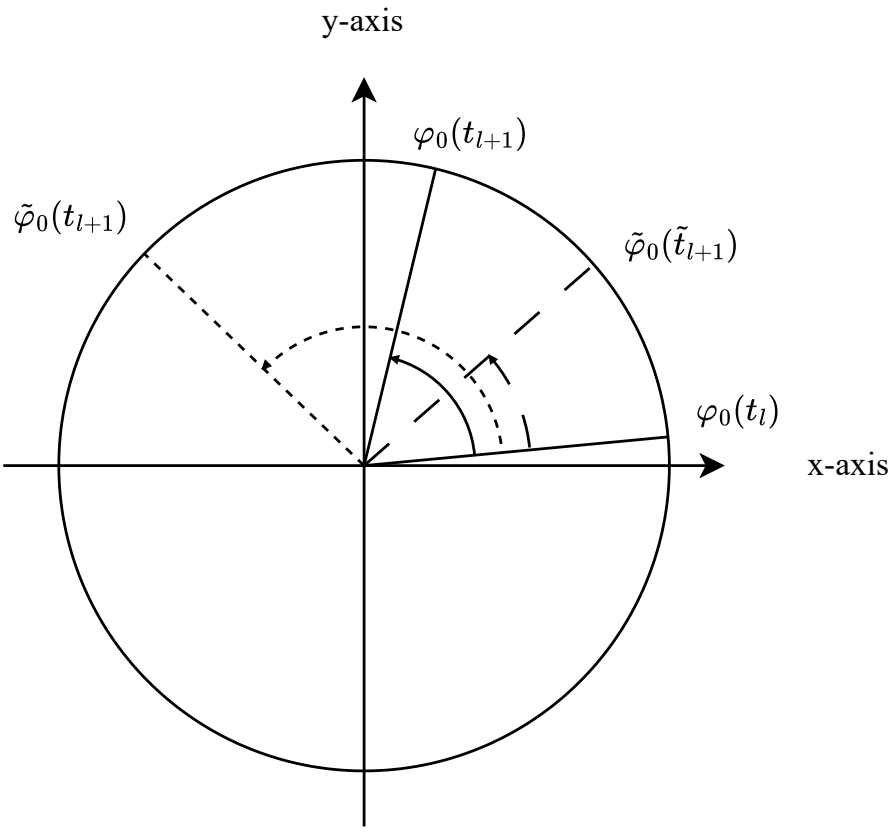


Figure 5.1: The rotation in speech modification.

network-based direct modification" corresponds to the neural methods, namely QHM-GAN and QHARMA-GAN, since they are trained by a large amount of data, encompassing various types of speech signals, and they can directly estimate the QHM parameters for modified speech.

In the subsequent sections, we first give the principle of the speech modification in the QHM framework. According to the ways of modification, we second elaborate on the implementation details of each approach, highlighting their respective advantages, technical challenges, and applicable scenarios.

5.3.1 Core Principles of Speech Modification

The core principle of time-scale and pitch-scale modification in the QHM framework lies in adjusting the rotation angle of all components' phase across frames. Specifically, both types of modification can be interpreted as controlled changes to the phase increment, i.e., the angular displacement, of all components between adjacent frames. At the same time, the spectral shapes must remain fixed and smooth across frames.

Taking the pitch component as an example, Fig. 5.1 schematically illustrates its rotation behavior during speech modification. Here, $\varphi_0(t_l)$ denotes the phase at the center of the l-th frame. The phase at the next frame, $\varphi_0(t_{l+1})$, reflects the original phase increment and is visualized as a solid

arc.

For time-scale modification, changing the frame-shift alters the phase rotation. For example, halving the frame-shift halves the phase increment between frames. This scenario is illustrated by $\tilde{\varphi}_0(\tilde{t}_{l+1})$ in the figure, where the phase increment is reduced accordingly. Importantly, the instantaneous frequency remains the same, but the duration shortens. In such a shorter duration, the original sentence is compressed, thus it sounds like the speech is accelerated.

Conversely, in pitch-scale modification, the frame-shift remains fixed, but the instantaneous frequency is scaled (e.g., doubled), leading to a proportional increase in the phase increment per frame. This is illustrated by $\tilde{\varphi}_0(t_{l+1})$ in the figure, where the phase rotation angle is doubled due to the doubled frequency.

Thus, both modifications can be uniformly described as adjustments to the inter-frame phase increment of each component. Maintaining consistent relative positions across harmonics preserves the spectral envelope during modification. This phase-centric view provides a compact and effective basis for high-quality speech modification in the QHM framework. In the following sections, we detail the ways of QHM methods and our proposed extensions (QHM-GAN and QHARMA-GAN) to maintain the spectral shapes, respectively.

5.3.2 Post-Modification of Parameters for QHM Methods

In this approach, speech parameters (i.e., framewise amplitudes, frequencies, and phases) are first extracted by QHM methods from the original signal. Based on them, new parameters are estimated for the modified speech, and finally, resynthesis is performed to achieve time- or pitch-scaling. Since preserving the spectral shape requires phase consistency, QHM first modifies the pitch component's phase and then derives the phases of higher harmonics to maintain harmonic coherence and the spectral envelope. This provides fine-grained control with high fidelity and minimal distortion, as illustrated in Fig. 5.2. Next, techniques for phase re-adjustment are introduced.

Time-scale Modification

For time-scale modification, to preserve the spectral envelope, the waveform shape must remain unchanged. This requires interpolation of instantaneous parameters when rescaling harmonically related sinusoids. Instantaneous amplitudes and frequencies can be interpolated directly, typically using linear or spline methods. In contrast, phase interpolation is more challenging because phase naturally rotates with time across analysis frames and is sensitive to frequency estimation errors.

To address this issue, the concept of relative phase [101, 39] is introduced. By subtracting the integral of kf_0 from the phase (where k is the harmonic index and f_0 the fundamental frequency), the rotational component is removed, and the effect of frequency estimation errors is mitigated. As a result, the relative phase varies smoothly across consecutive frames, assuming the signal

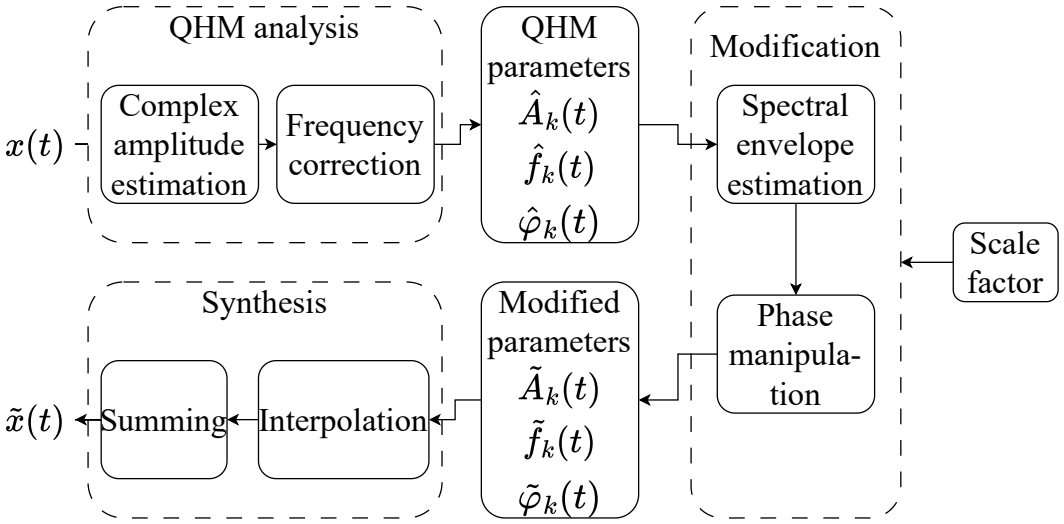


Figure 5.2: The workflow of the speech modification with a post-modification strategy.

evolves smoothly over time.

Once the relative phase is obtained, it can be interpolated to form a continuous trajectory. To ensure that the derivative (instantaneous frequency) is also smooth, higher-order interpolation schemes such as spline or cubic interpolation are required. This preserves the waveform shape under time-scaling, enabling shape-invariant synthesis.

Relative Phase Here, we present the theoretical foundation of the relative phase, which is essential for ensuring waveform consistency under time or pitch modifications. Consider a harmonic signal within the l-th frame. The fundamental (pitch) component can be represented as

$$x_0(t) = e^{i\left[2\pi \int_{t_l}^{t_l+t} f_0(u)du + \theta_0\right]}, \tag{5.1}$$

where θ_0 denotes the initial phase of the pitch component at the beginning of the l-th frame. This value can also be interpreted as the phase at the end of the preceding frame, i.e., the $(l-1)$-th frame.

Similarly, the k-th harmonic component is expressed as

$$x_k(t) = e^{i\left[2\pi k \int_{t_l}^{t_l+t} f_0(u)du + \theta_k\right]}, \quad k \in \mathbb{Z}^+, \tag{5.2}$$

where $f_k = kf_0$ is assumed and θ_k denotes the initial phase of the k-th harmonic component. Accordingly, the instantaneous phases of the pitch and the k-th harmonic components are, respectively, given by

$$\varphi_0(t) = 2\pi \int_{t_l}^{t_l+t} f_0(u)du + \theta_0, \tag{5.3}$$

$$\varphi_k(t) = 2\pi k \int_{t_l}^{t_l+t} f_0(u)du + \theta_k. \tag{5.4}$$

Assuming $\theta_0 = 0$ without loss of generality, we can subtract k times the pitch phase from the k-th component's phase to obtain

$$\theta_k = \varphi_k(t) - k\varphi_0(t), \tag{5.5}$$

which reveals a fundamental property: the initial phase of the k-th harmonic can be determined by its instantaneous phase and that of the pitch component. This formulation ensures that, regardless of how the pitch phase θ_0 evolves across frames, the waveform shape within each frame can be preserved by adjusting θ_k accordingly. Therefore, θ_k (when $\theta_0 = 0$) or, more generally, $\theta_k - k\theta_0$ defines the *relative phase* of the k-th harmonic component, which is critical for maintaining harmonic phase coherence.

When performing time- or pitch-scale modifications, it is often necessary to redefine the time axis. Let $\tilde{t}_l$ denote the scaled center time of the l-th frame, and let $\tilde{t}$ represent a local time variable relative to $\tilde{t}_l$. The instantaneous phase of the pitch component in the scaled time domain is given by

$$\tilde{\varphi}_0(\tilde{t}) = 2\pi \int_{\tilde{t}_l}^{\tilde{t}_l+\tilde{t}} \tilde{f}_0(u)du, \tag{5.6}$$

and accordingly, the phase of the k-th harmonic becomes

$$\tilde{\varphi}_k(\tilde{t}) = \theta_k + 2\pi k \int_{\tilde{t}_l}^{\tilde{t}_l+\tilde{t}} \tilde{f}_0(u)du. \tag{5.7}$$

It is important to note that this formulation remains valid even when k is not an integer, which is essential for representing non-integer harmonics or frequency components during fine-scale pitch manipulation.

In summary, the use of relative phase allows for consistent reconstruction of harmonic components regardless of time or pitch modifications. By anchoring the harmonic phases relative to the pitch component, waveform coherence is preserved across frames, enabling natural and artifact-free speech synthesis and transformation.

Pitch-scale Modification

For the time-scale modification, it is easy to use the relative phase to keep the waveform shape invariant. However, when the modification involves rapid changes in frequency components (e.g., in pitch-scale modification), the approach can suffer from phase dispersion, where the relative phase relationships among harmonic components are altered. This occurs because simply preserving frequency magnitudes is insufficient to maintain the waveform shape; fine phase alignment between harmonics must also be preserved for perceptually natural reconstruction. To mitigate this issue, the phase delay method has been proposed [102].

Relative Phase Delay The phase delay of the k-th component in the l-th frame can be defined as

$$\tau_k^l = \frac{\varphi_k(t_l)}{\omega_k(t_l)} = \frac{\varphi_k(t_l)}{2\pi f_k(t_l)}, \tag{5.8}$$

where $\varphi_k(t_l)$ denotes the instantaneous phase of the k-th component at the frame center t_l. This quantity represents the temporal shift required to reach the current phase, essentially indicating the alignment of the sinusoid within the frame.

To preserve the waveform shape, it is important to maintain the relative alignment among components. Thus, the *relative phase delay* is introduced as the difference between the phase delays of the k-th component and the fundamental component (pitch), expressed as

$$\Delta\tau_k^l = \tau_k^l - \tau_0^l. \tag{5.9}$$

This value reflects the time lag of the k-th harmonic component relative to the pitch component. Once the phase delay of the pitch component is known, the phase delay of any other harmonic component can be computed using its relative delay.

In the case of time- or pitch-scaled speech, we denote the scaled phase delay as $\tilde{\tau}_k^l$, and assume that the relative delay remains invariant:

$$\begin{aligned} \Delta\tau_k^l &= \Delta\tilde{\tau}_k^l \\ \tau_k^l - \tau_0^l &= \tilde{\tau}_k^l - \tilde{\tau}_0^l \\ \tilde{\varphi}_k(t_l) &= \left(\tilde{\tau}_0^l + \left(\tau_k^l - \tau_0^l\right)\right) \cdot 2\pi f_k(t_l). \end{aligned} \tag{5.10}$$

Therefore, once the phase delay of the pitch component is determined in the scaled signal, the absolute phase of all harmonic components can be calculated by preserving their relative phase delays. This strategy ensures coherent phase alignment across harmonics, effectively preventing phase dispersion and preserving waveform fidelity under time and pitch modifications.

Apart from the estimation of phase, the amplitude of each harmonic component also plays a crucial role in pitch-scale modification. Unlike time-scale modification, where the framewise amplitudes can be retained to preserve the spectral envelope, pitch-scaling alters the frequencies of all components. As a result, using the original amplitudes at the shifted frequencies may distort the spectral envelope, especially when the pitch shift is large. This leads to timbral inconsistencies and a loss of speaker identity in the synthesized speech.

To address this issue, the spectral envelope must be re-estimated to adaptively determine the amplitudes at the new harmonic positions. However, sinusoidal models alone are insufficient for modeling the smooth spectral envelope due to their discrete and component-wise nature. Therefore, additional estimation methods, such as linear predictive coding (LPC) [54] or discrete all-pole (DAP) [94] modeling, are employed to represent the spectral envelope as a continuous function over frequency.

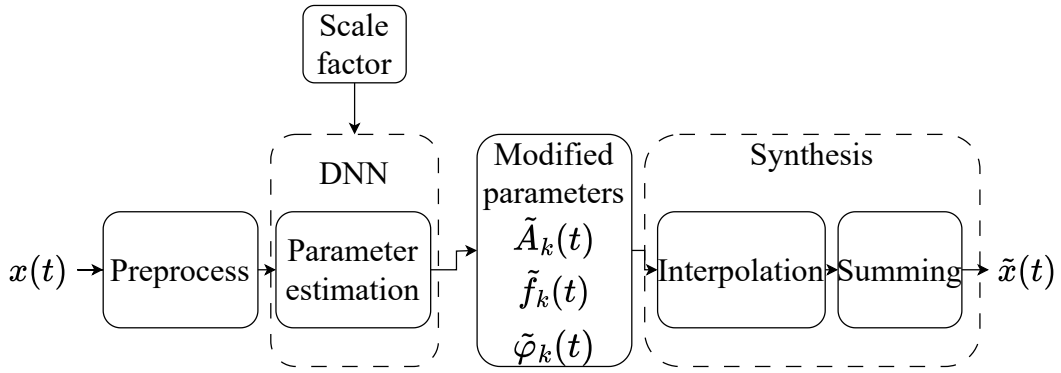


Figure 5.3: The workflow of the speech modification with neural-based methods.

Using these techniques, the amplitude $\tilde{A}_k(t_l)$ of the k-th component in the l-th frame, after pitch-scaling, is computed by evaluating the spectral envelope at the scaled frequency:

$$\tilde{A}_k(t_l) = f_{\mathrm{DAP}}(t_l, \rho_l f_k), \tag{5.11}$$

where ρ_l is the pitch-scale factor for the l-th frame, and f_k is the original frequency of the k-th component. Here, $f_{\mathrm{DAP}}(\cdot,\cdot)$ denotes the spectral envelope estimation process by discrete all-pole at a specific time and frequency.

This strategy ensures that the amplitude distribution across frequencies remains consistent with the original spectral shape, thereby maintaining the timbral characteristics of the speech signal after pitch transformation. Combined with relative phase delay alignment, this amplitude correction forms a complete and coherent approach to high-quality pitch-scale modification within the quasi-harmonic modeling framework.

5.3.3 Neural Network-based Modification for QHM-GAN and QHARMA-GAN

Neural vocoders such as QHM-GAN and QHARMA-GAN can be trained to predict quasi-harmonic parameters for a target modification (e.g., pitch or speaking rate). Instead of manually modifying extracted parameters, the network learns a mapping from input features (e.g., mel-spectrograms and f_0) to modified quasi-harmonic parameters. This enables end-to-end generation of modified speech without additional techniques like relative phase or relative phase delay. The workflow is illustrated in Fig. 5.3.

Time-scale Modification

A straightforward strategy is to interpolate mel-spectrograms and feed them to the network. Although effective, this increases computation for long utterances and introduces interpolation errors that cannot be corrected by the black-box model.

Fortunately, vocoders with quasi-harmonic modeling architectures, such as QHM-GAN and

QHARMA-GAN, generate waveforms from interpretable parameters, namely the framewise amplitudes and phases of quasi-harmonics. This structure naturally enables a more efficient and controllable time-scale modification. Instead of interpolating neural network inputs, we adopt a QHM-style parametric approach: the network first estimates frame-level amplitudes and phases from the mel-spectrogram under the original frame alignment, after which these parameters are temporally interpolated according to a modified frame-shift schedule, yielding amplitude and phase trajectories consistent with the new temporal structure.

This strategy reduces computational complexity during inference by decoupling the time-scaling operation from the network and applying it directly in the parameter space. Since interpolation operates on smooth, low-dimensional quasi-harmonic parameters rather than high-dimensional acoustic features, the generated speech maintains fidelity and avoids artifacts common in time-domain resampling. In contrast, interpolating mel-spectrograms depends on the black-box inference of the neural network, leading to inevitable errors that cannot be corrected. Parameter-space interpolation, however, is more direct and accurate, producing higher-quality scaled speech. Therefore, in this thesis, time-scale modification for QHM-GAN and QHARMA-GAN is implemented based on modifying the output parameters estimated by the DNN.

Pitch-scale Modification

For pitch-scale modification, the neural method strategy is conceptually similar to time-scale modification but differs due to the nature of pitch alteration. Unlike post-modification methods, which require spectral envelope re-estimation and phase adjustment, neural vocoders can directly estimate the spectral envelope from the input, simplifying the pipeline.

In vocoders with quasi-harmonic modeling, such as QHM-GAN, the input includes a mel-spectrogram and an f_0 contour. Pitch-scaling is achieved by scaling the f_0 trajectory to a target contour, which is fed into the network along with the original mel-spectrogram. The network then directly estimates the quasi-harmonic parameters, i.e., framewise amplitudes and phase correction terms, corresponding to the desired pitch-modified speech.

The scaled speech waveform is reconstructed using the sinusoidal synthesis engine of QHM-GAN, eliminating explicit post-processing steps such as harmonic re-alignment or amplitude compensation, as these are implicitly learned. Embedding pitch modification within the inference pipeline ensures efficiency, controllability, and high-quality output without noticeable artifacts.

This approach is especially advantageous for large-scale speech synthesis or voice conversion systems, where interpretability and efficiency are important. Moreover, the quasi-harmonic structure preserves the resonance characteristics, maintaining the perceptual identity and timbral quality of the speaker even under substantial pitch scaling.

5.4 Speech Modification Algorithms

In this section, we elaborate on the specific algorithms used to achieve time-scale and pitch-scale modifications within the quasi-harmonic modeling framework: conventional post-modification and neural vocoder-based approaches.

First, the post-modification strategy for conventional QHM methods directly operates on the extracted amplitudes and phases of harmonic components, enabling interpretable and controllable modifications but often requiring additional modeling of spectral characteristics for accurate pitch manipulation.

Second, in contrast, neural vocoder-based approaches (e.g., QHM-GAN and QHARMA-GAN) estimate the quasi-harmonic parameters of the modified speech directly from scaled inputs such as mel-spectrograms and modified f_0. These methods provide a unified, data-driven framework that is efficient and effective for large-scale or real-time applications, though with reduced transparency compared to parametric methods.

The following subsections discuss the implementation details, interpolation mechanisms, and phase-handling techniques of the two strategies, highlighting their relative advantages and suitable use cases.

5.4.1 Speech Modification based on QHM Methods

In this subsection, we focus on the modification for on the QHM methods, primarily following [3]. These methods rely on extracting frame-wise quasi-harmonic parameters, i.e., amplitude, frequency, and phase, from the original speech. Speech modification is then achieved by adjusting these parameters and resynthesizing the signal. We particularly emphasize the estimation process of the modified parameters, which is crucial for achieving high-quality synthesis results under time-scale or pitch-scale transformations for QHM methods.

Time-scale Modification

As described in Section 5.3.2, for time-scale modification, the pitch contour is expected to be stretched or compressed temporally, while the formant structure is modified at a proportionally adjusted rate. Let t_l represent the time instant at the l-th frame of the original signal, and let β_l denote the time-scaling factor at that frame. The corresponding scaled time $\tilde{t}_l$ is calculated by

$$\tilde{t}_l = \sum_{i=1}^{l} \beta_i (t_i - t_{i-1}), \tag{5.12}$$

where $\tilde{t}_0 = 0$ is defined as the starting time of the scaled signal. A time-scaling factor $\beta_l > 1$ increases the duration of the signal, while $\beta_l < 1$ shortens it.

Let $\hat{A}_k(t)$, $\hat{f}_k(t)$, and $\hat{\varphi}_k(t)$ represent the instantaneous amplitude, frequency, and phase of the k-th component in the original quasi-harmonic signal. The corresponding parameters for the time-

scaled signal are denoted by $\tilde{A}_k(\tilde{t})$, $\tilde{f}_k(\tilde{t})$, and $\tilde{\varphi}_k(\tilde{t})$, respectively. The time-scaled speech signal can thus be expressed as

$$\tilde{x}(\tilde{t}) = \sum_{k=-K}^{K} \tilde{A}_k(\tilde{t}) e^{i\tilde{\varphi}_k(\tilde{t})}. \tag{5.13}$$

To obtain the time-scaled waveform, the following steps are performed:

1. **Amplitude adjustment:** The framewise amplitudes are directly mapped from the original to the scaled time instants:

$$\tilde{A}_k(\tilde{t}_l) = \hat{A}_k(t_l). \tag{5.14}$$

2. **Frequency adjustment:** The instantaneous frequencies are similarly mapped:

$$\tilde{f}_k(\tilde{t}_l) = \hat{f}_k(t_l). \tag{5.15}$$

3. **Pitch phase rescaling:** The phase of the fundamental component ($k = 0$) is adjusted using the scaling factor:

$$\tilde{\varphi}_0(\tilde{t}_l) = \tilde{\varphi}_0(\tilde{t}_{l-1}) + \beta_l \left[\hat{\varphi}_0(t_l) - \hat{\varphi}_0(t_{l-1})\right]. \tag{5.16}$$

4. **Relative phase reconstruction:** The relative phase θ_k of each component is first obtained from Eq. (5.5), and the full phase at each scaled frame is reconstructed as

$$\tilde{\varphi}_k(\tilde{t}_l) = \theta_k + k\tilde{\varphi}_0(\tilde{t}_l). \tag{5.17}$$

5. **Instantaneous phase computation:** The instantaneous phase at an arbitrary time $\tilde{t}$ is obtained by integrating the instantaneous frequency:

$$\tilde{\varphi}_k(\tilde{t}) = \tilde{\varphi}_k(\tilde{t}_l) + \int_{\tilde{t}_l}^{\tilde{t}_l + t} \left[2\pi\tilde{f}_k(u) + c(u)\right] du, \tag{5.18}$$

   where $\tilde{f}_k(\tilde{t})$ is the interpolated frequency using a cubic spline, and $c(\tilde{t})$ is a compensation term ensuring smooth phase transitions. It is defined by

$$c(\tilde{t}) = z \sin\left(\frac{\pi(\tilde{t} - \tilde{t}_{l-1})}{\tilde{t}_l - \tilde{t}_{l-1}}\right), \tag{5.19}$$

   which follows the same form as that used in Eq. (2.30).

6. **Final synthesis:** The complete time-scaled speech signal $\tilde{x}(\tilde{t})$ is synthesized by substituting the interpolated amplitudes and computed phases into Eq. (5.13).

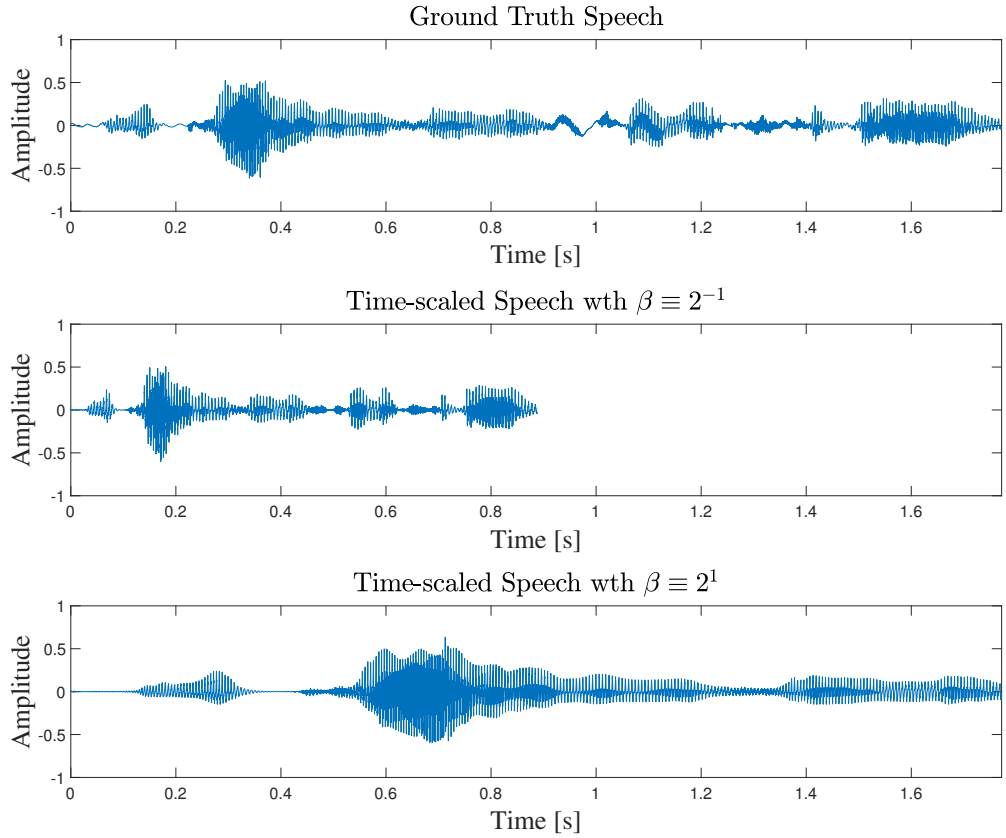


Figure 5.4: The waveforms of ground truth and time-scaled speech based on QHM methods.

To illustrate the workflow more specifically, a pseudocode is given in Algorithm 4. Additionally, an example of time-scale speech modification based on QHM methods is illustrated in Fig. 5.4, where the corresponding spectrograms are demonstrated in Fig. 5.5.

Algorithm 4 Time-scale speech modification based on QHM methods.

Step 1: Preprocessing and Setting
Extract the framewise parameters, including amplitudes $\hat{A}_k(t_l)$, frequencies $\hat{f}_k(t_l)$, and phase $\hat{\varphi}_k(t_l)$, from speech $x(t)$ by QHM methods;
Step 2: Parameters modification
Scale the time instant by Eq. (5.12);
Adjust the amplitudes of all components by Eq. (5.14);
Adjust the frequencies of all components by Eq. (5.15);
Scale the phase difference of pitch components by Eq. (5.16);
Calculate the relative phase of all components by Eq. (5.5);
Calculate the phase of all components for scaled speech by Eq. (5.17);
Linearly interpolate $\tilde{A}_k(\tilde{t}_l)$ into $\tilde{A}_k(\tilde{t})$;
Cubically interpolate $\tilde{f}_k(\tilde{t}_l)$ into $\tilde{f}_k(\tilde{t})$;
Calculate the instantaneous phase $\tilde{\varphi}_k(\tilde{t}_l)$ by Eq. (5.18);
Step 3: Generation
$\tilde{x}(\tilde{t}) \leftarrow \sum_{k=-K}^{K} \tilde{A}_k(\tilde{t}) e^{i\tilde{\varphi}_k(\tilde{t})}$;
Output: $\hat{x}(t)$

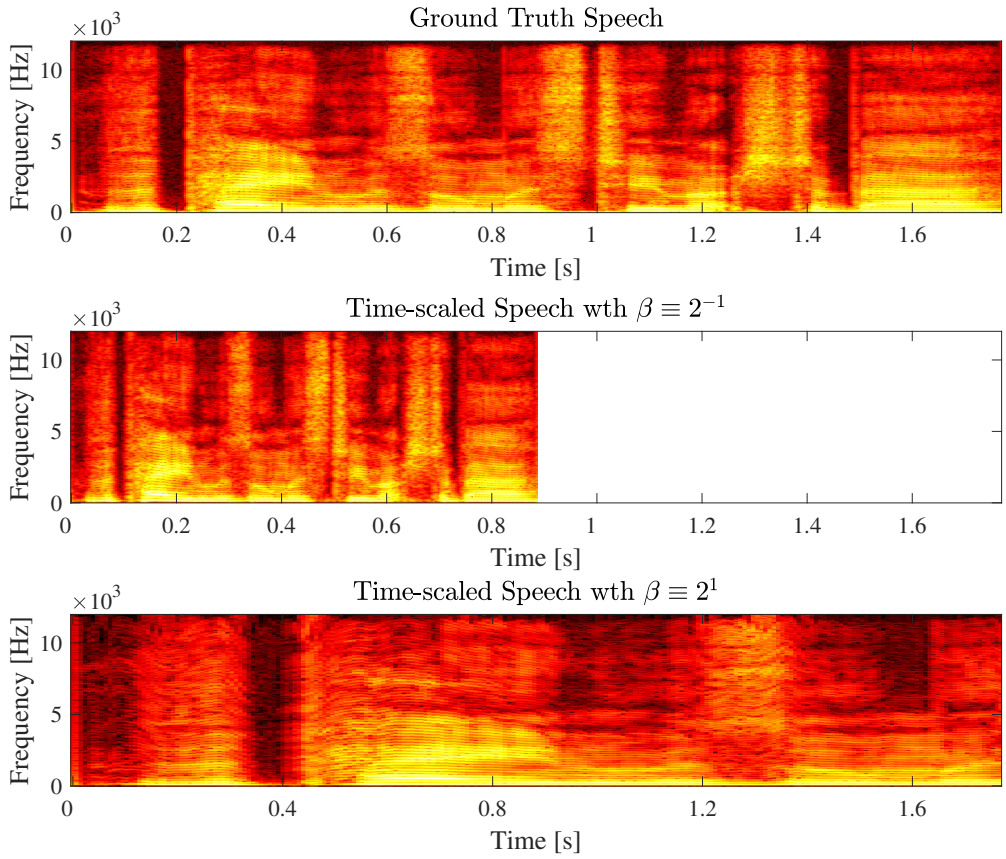


Figure 5.5: The spectrograms of ground truth and time-scaled speech based on QHM methods.

Pitch-scale Modification

As described in Section 5.3.2, in pitch-scale modification, the f_0 and its harmonics are scaled by a pitch factor, requiring accurate amplitude and phase estimation to avoid distortions. To this end, amplitude estimation is performed using the discrete all-pole (DAP) method, and the relative phase delay is applied to preserve coherent harmonic alignment.

Likewise, the parameters of quasi-harmonics for original speech are denoted as $\hat{A}_k(t)$, $\hat{f}_k(t)$, and $\hat{\varphi}_k(t)$, respectively, whereas those for pitch-scaled speech are denoted as $\tilde{A}_k(t)$, $\tilde{f}_k(t)$, and $\tilde{\varphi}_k(t)$, respectively. Denoting the pitch-scale factor of the l-th frame as ρ_l, it can be known that the pitch increases when $\rho_l > 1$, where the opposite happens when $\rho_l < 1$. The workflow of pitch-scale modification will be demonstrated in the following part.

1. **Frequency modification:** The frequency will be modified by

$$\tilde{f}_k(t_l) = \rho_l \hat{f}_k(t_l) \tag{5.20}$$

2. **Amplitude estimation:** The amplitudes of all components will be estimated by DAP with

their shifted frequencies:

$$\tilde{A}_k(t_l) = f_{\mathrm{DAP}}(t_l, \tilde{f}_k(t)), \tag{5.21}$$

3. **Pitch phase rescaling:** The phase of the fundamental component ($k = 0$) will be modified by

$$\tilde{\varphi}_0(t_l) = \tilde{\varphi}_0(t_{l-1}) + \rho_l \left[\hat{\varphi}_0(t_l) - \hat{\varphi}_0(t_{l-1})\right]. \tag{5.22}$$

4. **Phase reconstruction with relative phase delay:** The relative phase delay of each component $\hat{\tau}_k^l$ is first obtained from Eq. (5.8), where the relative phase delay of the fundamental component for scaled speech $\tilde{\tau}_0^l$ is also computed. The phase at each frame for pitch-scaled speech is reconstructed as

$$\tilde{\varphi}_k(t_l) = \left[\tilde{\tau}_0^l + \left(\hat{\tau}_k^l - \hat{\tau}_0^l\right)\right] 2\pi \tilde{f}_k(t_l). \tag{5.23}$$

5. **Instantaneous phase computation:** The instantaneous phases of all components are obtained by integrating the instantaneous frequencies:

$$\tilde{\varphi}_k(t) = \tilde{\varphi}_k(t_l) + \int_{t_l}^{t_l+t} \left[2\pi \tilde{f}_k(u) + c(u)\right] du, \tag{5.24}$$

   where $\tilde{f}_k(t)$ is the interpolated frequency using a cubic spline, and $c(t)$ is a compensation term ensuring smooth phase transitions, as used in Eq. (2.30).

6. **Final synthesis:** The complete time-scaled speech signal $\tilde{x}(t)$ is synthesized by

$$\tilde{x}(t) = \sum_{k=-K}^{K} \tilde{A}_k(t) e^{i\tilde{\varphi}_k(t)} \tag{5.25}$$

To illustrate the process of pitch-scale more specifically, a pseudocode is given in Algorithm 5. Additionally, an example of pitch-scale speech modification based on QHM methods is illustrated in Fig. 5.6, where the corresponding spectrograms are demonstrated in Fig. 5.7.

5.4.2 Speech Modification based on QHM-GAN and QHARMA-GAN

Second, we consider modification based on neural vocoders, specifically QHM-GAN and QHARMA-GAN. These models directly estimate quasi-harmonic parameters (amplitude, frequency, and phase) of modified speech from scaled inputs such as f_0 and mel-spectrogram, leveraging training data to achieve robust and efficient parameter estimation compared to conventional methods. Given their distinct synthesis processes, the modification algorithms for QHM-GAN and QHARMA-GAN are introduced separately.

Algorithm 5 Pitch-scale speech modification based on QHM methods.

Step 1: Preprocessing and Setting
Extract the framewise parameters, including amplitudes $\hat{A}_k(t_l)$, frequencies $\hat{f}_k(t_l)$, and phase $\hat{\varphi}_k(t_l)$, from speech $x(t)$ by QHM methods;
Step 2: Parameters modification
Scale the frequencies of all components by Eq. (5.20);
Adjust the amplitudes of all components by Eq. (5.21);
Scale the phase difference of pitch components by Eq. (5.22);
Calculate the relative phase delay of all components by Eq. (5.8);
Calculate the phase of all components for scaled speech by Eq. (5.23);
Linearly interpolate $\tilde{A}_k(t_l)$ into $\tilde{A}_k(t)$;
Cubically interpolate $\tilde{f}_k(t_l)$ into $\tilde{f}_k(t)$;
Calculate the instantaneous phase $\tilde{\varphi}_k(t_l)$ by Eq. (5.24);
Step 3: Generation
$\tilde{x}(t) \leftarrow \sum_{k=-K}^{K} \tilde{A}_k(t) e^{i\tilde{\varphi}_k(t)}$;
Output: $\hat{x}(t)$

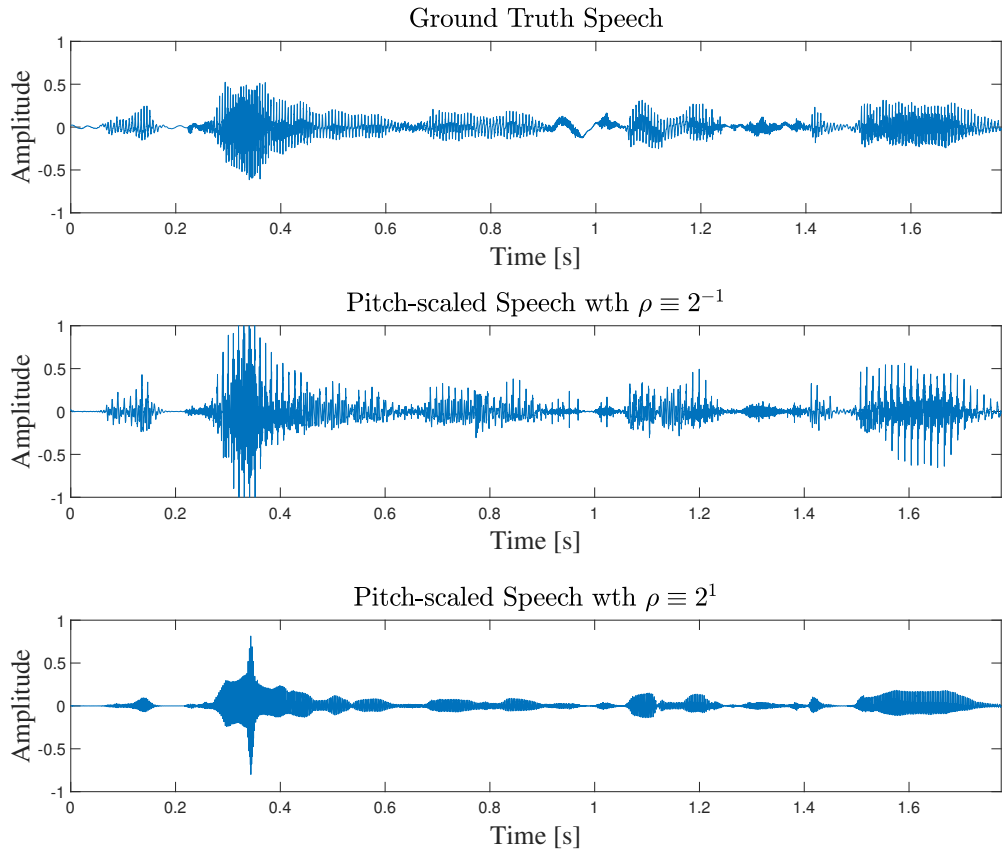


Figure 5.6: The waveforms of ground truth and pitch-scaled speech based on QHM methods.

Speech Modification based on QHM-GAN

As mentioned before, the central principle underlying QHM-GAN-based speech modification lies in scaling the rotation angles of all quasi-harmonic components simultaneously. Given that each quasi-harmonic component is represented as a sinusoidal waveform, it can be individually modified by applying an appropriate scale factor. As with conventional QHM-based modification

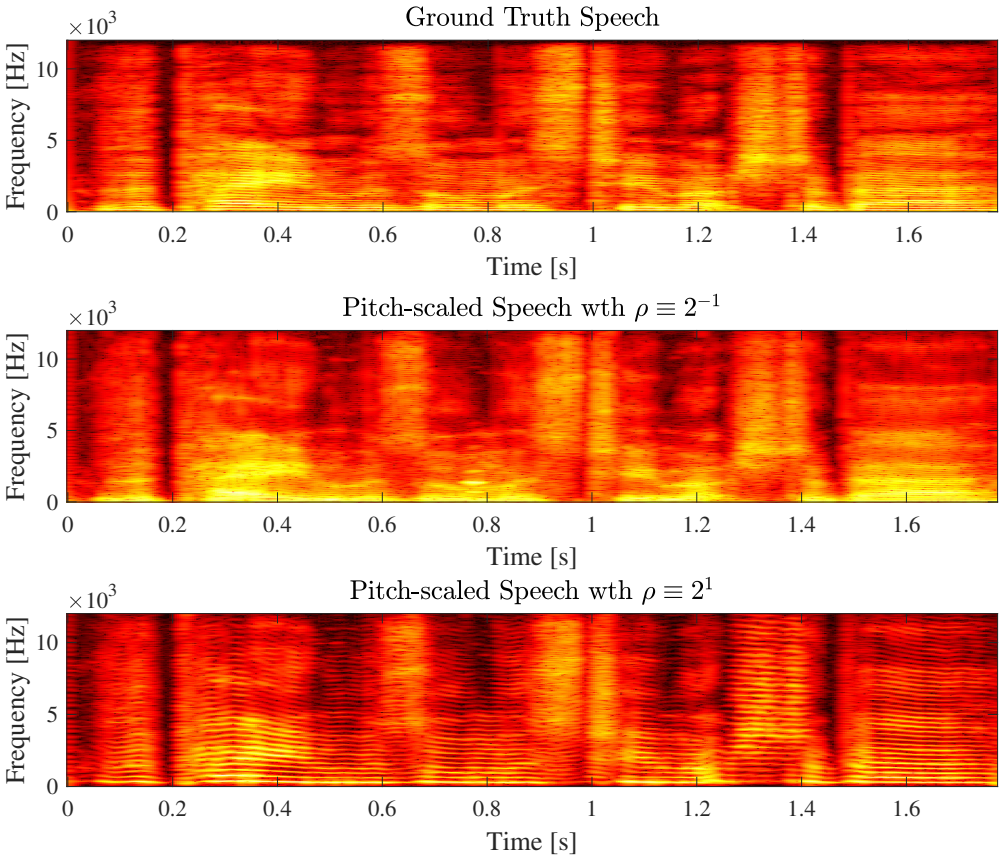


Figure 5.7: The spectrograms of ground truth and pitch-scaled speech based on QHM methods.

methods, it is first necessary to calculate the phase difference between adjacent frames. These framewise phase differences are then scaled accordingly. Notably, the phase estimated by QHM-GAN is unwrapped across frames, allowing direct computation of the phase difference from the estimated instantaneous frequencies and phase compensation terms.

Based on Eq. (4.2), the phase difference of the k-th harmonic component at frame l can be formulated as

$$\begin{aligned}\Delta\phi_k^l &= \hat{\varphi}_k(t_l) - \hat{\varphi}_k(t_{l-1}) \\ &= \int_{t_{l-1}}^{t_l} 2\pi \hat{f}_k(u)\,\mathrm{d}u + \Delta\varphi_k^l,\end{aligned} \tag{5.26}$$

where $\Delta\varphi_k^l$ represents the phase compensation term estimated by the neural network, and $\hat{f}_k(u)$ denotes the instantaneous frequency of the k-th component.

We first consider time-scale modification. Let $\tilde{t}_l$ denote the modified time instant at frame l, and let β_l be the time-scale factor at that frame. Under time scaling, the frequencies of each component remain unchanged. Therefore, using the scaled phase differences, the phase at each

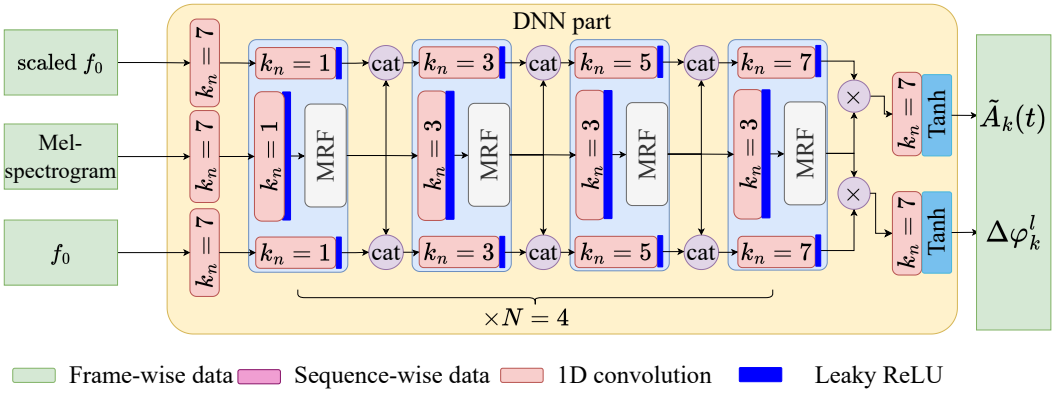


Figure 5.8: Generator architecture of QHM-GAN for pitch-scale modification.

frame for the modified speech is given by

$$\begin{aligned}\tilde{\varphi}_k(\tilde{t}_l) &= \sum_{i=1}^{l} \beta_i \Delta\phi_k^i \\ &= \int_0^{\tilde{t}_l} 2\pi \hat{f}_k(u)\,\mathrm{d}u + \sum_{i=1}^{l} \beta_i \Delta\varphi_k^i,\end{aligned} \tag{5.27}$$

where the summation accumulates the scaled phase compensation over time. After the modified framewise phase and frame-shift are obtained, the time-scaled waveform can be generated by substituting these parameters into the QHM-GAN synthesis process.

Likewise, pitch-scale modification can be formulated as a scaling of the rotation angles, since shifting the frequencies directly changes the rate of phase accumulation. Let ρ_l be the pitch-scale factor at frame l, and let $\tilde{f}_k(u) = \rho_l \hat{f}_k(u)$ denote the pitch-shifted frequency. The phase at each frame center under pitch scaling can then be written as

$$\begin{aligned}\tilde{\varphi}_k(t_l) &= \sum_{i=1}^{l} \rho_i \Delta\phi_k^i \\ &= \int_0^{t_l} 2\pi \tilde{f}_k(u)\,\mathrm{d}u + \sum_{i=1}^{l} \rho_i \Delta\varphi_k^i,\end{aligned} \tag{5.28}$$

where the integral reflects the accumulation of the scaled instantaneous frequency, and the summation accounts for the scaled phase compensation.

It is important to note that the phase compensation term $\Delta\varphi_k^i$ in Eq. (5.28) is predicted using the original f_0 as input, rather than the pitch-scaled f_0. In contrast, the harmonic amplitudes should be estimated using the scaled input, i.e., $\rho_l f_0$. To address this inconsistency, a dual-input structure is adopted, as illustrated in Fig. 5.8. In this design, both the original f_0 and the scaled f_0 are simultaneously input to the neural network. The original f_0 is used to predict the phase compensation, while the scaled f_0 is used to estimate the harmonic amplitudes. This strategy enables high-fidelity pitch- and time-scale speech modification using the QHM-GAN architecture.

Likewise, the parameters of quasi-harmonics for original speech are denoted as $\hat{A}_k(t)$, $\hat{f}_k(t)$, and $\hat{\varphi}_k(t)$, respectively, whereas those for pitch-scaled speech are denoted as $\tilde{A}_k(t)$, $\tilde{f}_k(t)$, and $\tilde{\varphi}_k(t)$, respectively. Then, we demonstrate the specific modification algorithm, which can modify the speech signals in terms of both time-scale and pitch-scale simultaneously. The details are as follows:

1. **Time instant scaling:** The time instants of each frame center should be scaled; in other words, the frame-shift is scaled, as

$$\tilde{t}_l = \tilde{t}_0 + \sum_{i=1}^{l} \beta_i (t_i - t_{i-1}), \;\; \tilde{t}_0 = 0. \tag{5.29}$$

2. **Amplitude adjustment:** The amplitudes of each component for modified speech can be obtained by

$$\tilde{A}(\tilde{t}_l) = G_{QHM-GAN}(l, c_l, \tilde{f}(\tilde{t}_l)), \tag{5.30}$$

   where $G_{QHM-GAN}$ is the generator of QHM-GAN and c is the mel-spectrogram.

3. **Phase compensation estimation:** The phase compensation for the original synthesis should be estimated by

$$\Delta\varphi_k^l = G_{QHM-GAN}(l, c_l, \tilde{f}(\hat{t}_l)). \tag{5.31}$$

4. **Phase adjustment:** The phase of all components for scaled speech should be estimated by

$$\begin{aligned} \tilde{\varphi}_k(t_l) &= \sum_{i=1}^{l} \beta_i \rho_i \Delta\phi_k^i \\ &= \int_0^{\tilde{t}_l} 2\pi \tilde{f}_k(u)\, \mathrm{d}u + \sum_{i=1}^{l} \beta_i \rho_i \Delta\varphi_k^i, \end{aligned} \tag{5.32}$$

5. **Instantaneous parameters:** After obtaining the framewise amplitudes and phases, their instantaneous version can be obtained by linear interpolation and cubic interpolation, respectively, i.e., $\tilde{A}_k(\tilde{t})$ and $\tilde{\varphi}_k(\tilde{t})$.

6. **Final synthesis:** Finally, the modified speech can be generated by

$$\tilde{x}(t) = \sum_{k=-K}^{K} \tilde{A}_k(t) e^{i\tilde{\varphi}_k(t)}. \tag{5.33}$$

A pseudocode is also elaborated in Algorithm 6.

Algorithm 6 Time- and pitch-scale speech modification based on QHM-GAN.

Step 1: Preprocessing and Setting
Train the QHM-GAN with the dataset; set the time-scale factor β_l and the pitch-scale factor ρ_l;
Step 2: Parameters modification
Scale the time instants by Eq. (5.29);
Adjust the framewise amplitudes of all components by Eq. (5.30);
Linearly interpolate the framewise amplitudes to obtain the instantaneous amplitudes;
Obtain the phase compensations of all components at all frame centers by Eq. (5.31);
Calculate the phase of all components at all frame centers by Eq. (5.32);
Cubically interpolate the framewise phases to obtain the instantaneous phases;
Step 3: Generation
$\tilde{x}(t) \leftarrow \sum_{k=-K}^{K} \tilde{A}_k(t) e^{i\tilde{\varphi}_k(t)}$;
Output: $\hat{x}(t)$

Speech Modification based on QHARMA-GAN

Similar to QHM-GAN, QHARMA-GAN also modifies the speech by directly estimating the quasi-harmonic parameters for modified speech. However, there are some differences. First, due to the estimation of ARMA models, the spectral envelope can always be smooth over frequency. Therefore, the amplitudes of arbitrary frequencies can be easily estimated once the coefficients are obtained. Additionally, f_0 is not used during the process of the DNN; therefore, the modification of QHARMA-GAN is more flexible because of the lack of repeated process, e.g., the phase compensation and amplitude are separately estimated with different inputs of the DNN, as shown as Fig. 5.8.

For the time-scale modification, the time instants are also scaled, while the framewise phase can be obtained by the phase delay calculated from the ARMA model. Let us consider the phase increment at the l-th frame, we can easily get the phase increment as

$$\begin{aligned}\hat{\varphi}_k(t_l) - \hat{\varphi}_k(t_{l-1}) &= \varphi_k^u(t_l) + \angle H(t_l, \omega_k) - [\varphi_k^u(t_{l-1}) + \angle H(t_{l-1}, \omega_k)] \\ &= \int_{t_{l-1}}^{t_l} 2\pi \hat{f}_k(u)\mathrm{d}u + [\angle H(t_l, \omega_k) - \angle H(t_{l-1}, \omega_k)], \end{aligned} \tag{5.34}$$

where Eqs. (4.14) and (4.15) are considered. Observing the definition of relative phase from Eq. (5.3) to Eq. (5.5), we can obtain a similar relationship between fundamental component and harmonic components, as

$$\hat{\varphi}_0(t_l) - \hat{\varphi}_0(t_{l-1}) = \int_{t_{l-1}}^{t_l} 2\pi \hat{f}_0(u)\mathrm{d}u + \theta_0' \tag{5.35}$$

$$\hat{\varphi}_k(t_l) - \hat{\varphi}_k(t_{l-1}) = \int_{t_{l-1}}^{t_l} 2\pi k \hat{f}_0(u)\mathrm{d}u + \theta_k' \tag{5.36}$$

where $\theta_0' = \angle H(t_l, \omega_0) - \angle H(t_{l-1}, \omega_0)$ and $\theta_k' = \angle H(t_l, \omega_k) - \angle H(t_{l-1}, \omega_k)$. θ_0' and θ_k' can be

considered as the initial phase of the current frame. Therefore, it is easy to obtain

$$\theta'_k - k\theta'_0 = [\hat{\varphi}_k(t_l) - \hat{\varphi}_k(t_{l-1})] - k\,[\hat{\varphi}_0(t_l) - \hat{\varphi}_0(t_{l-1})] \tag{5.37}$$

Similar to QHM methods, once the relative phase is obtained, the phases of all harmonic components can be determined to ensure waveform shape invariance, regardless of how the fundamental phase is modified. In QHARMA-GAN, this property is further guaranteed since the ARMA coefficients uniquely determine the phase delay, i.e., $\theta'_k - k\theta'_0$. Thus, even when the frame-shift is scaled for time modification, the synthesis process preserves the relative phase, maintaining both waveform shape and spectral envelope without additional adjustments.

For pitch-scaled speech modification, unlike the conventional QHM-based approaches, where the spectral envelope is typically estimated using the DAP method, which approximates the amplitude of each harmonic component based on a discretely estimated autoregressive model and is inherently limited by its dependency on framewise estimation and its inability to capture fine-grained temporal dynamics, QHARMA-GAN adopts a more advanced strategy by directly learning the parameters of an autoregressive moving average (ARMA) model from data. Since the ARMA coefficients perform full-band modeling over 0–$\frac{f_s}{2}$, this end-to-end neural estimation enables simultaneous prediction of both amplitude and phase delay for all components at any position, thereby achieving higher fidelity in capturing resonance structures. Benefiting from the availability of sufficient training data, QHARMA-GAN is capable of estimating more accurate phase delays for shifted frequencies. As a result, QHARMA-GAN achieves better spectral-envelope-invariant pitch modification even without explicitly modeling the relative phase delay, offering a significant advantage over conventional QHM methods in pitch-scaling tasks.

However, it is important to note a common limitation inherent to all methods based on the QHM structure, including QHM, BP-QHM, QHM-GAN, and QHARMA-GAN, when applied to pitch-scale modification. These methods synthesize both voiced and unvoiced components using sums of sinewaves. While this is effective for modeling periodic (voiced) signals, it presents a challenge for unvoiced segments, which are inherently aperiodic and noise-like. When the pitch is increased, the number of quasi-harmonics within the analysis bandwidth decreases, leading to a sparse harmonic representation. Consequently, the synthesized unvoiced speech may undesirably exhibit harmonic artifacts, giving it an unnatural tonal or "voiced-like" quality.

To mitigate this issue, a dedicated strategy is employed that involves the separate processing of voiced and unvoiced segments. A voiced/unvoiced (V/UV) detection algorithm is first applied to segment the speech signal accordingly. For voiced frames, pitch-scaling is performed by adjusting both frequency and amplitude parameters based on the modified f_0. For unvoiced frames, the original parameters are retained without modification, preserving the stochastic nature of noise components. After both voiced and unvoiced segments are synthesized independently using their respective parameter sets, they are seamlessly concatenated to produce the final modified speech

waveform. This hybrid strategy effectively balances flexibility in pitch manipulation with the preservation of naturalness in unvoiced segments, further enhancing the quality of pitch-scaled speech synthesized within the QHM framework.

In the following part, we present a detailed speech modification algorithm capable of achieving high-quality results for both time-scale and pitch-scale transformations simultaneously. The quasi-harmonic parameters of the original speech are denoted as $\hat{A}_k(t)$, $\hat{f}_k(t)$, and $\hat{\varphi}_k(t)$, representing the amplitude, frequency, and phase, respectively. Similarly, the parameters for the modified (pitch-scaled) speech are denoted as $\tilde{A}_k(t)$, $\tilde{f}_k(t)$, and $\tilde{\varphi}_k(t)$. After obtaining the ARMA coefficients from a DNN, we define the time-scale and pitch-scale factors at the l-th frame as β_l and ρ_l, respectively. The detailed process is described as follows:

1. **Time instant scaling:** The temporal locations of frame centers are rescaled according to the time-scale factor, i.e.,

$$\tilde{t}_l = \tilde{t}_0 + \sum_{i=1}^{l} \beta_i (t_i - t_{i-1}), \;\; \tilde{t}_0 = 0. \tag{5.38}$$

2. **Frequency modification:** The frequencies are adjusted based on whether the frame is voiced or unvoiced:

$$\text{Voiced: } \tilde{f}^l_{k,v} = \rho_l \hat{f}^l_k, \;\; \text{Unvoiced: } \tilde{f}^l_{k,uv} = \hat{f}^l_k, \tag{5.39}$$

where $\tilde{f}^l_{k,v}$ and $\tilde{f}^l_{k,uv}$ denote the modified frequencies for voiced and unvoiced segments, respectively.

3. **Amplitude estimation:** The amplitudes for the voiced and unvoiced segments are estimated as:

$$\tilde{A}^v_k(\tilde{t}_l) = VUV_l \times G_l \prod_{j=1}^{r} |\tilde{H}_j(\tilde{t}_l, 2\pi \tilde{f}^l_{k,v})| = VUV_l \times |G_l| \prod_{j=1}^{r} \frac{\left|1 + \sum_{q=1}^{Q/r} b^l_{j,q} e^{-i2\pi \tilde{f}^l_{k,v} q}\right|}{\left|1 + \sum_{p=1}^{P/r} a^l_{j,p} e^{-i2\pi \tilde{f}^l_{k,v} p}\right|}, \tag{5.40}$$

$$\tilde{A}^{uv}_k(\tilde{t}_l) = (1 - VUV_l) \times G \prod_{j=1}^{r} |\tilde{H}_j(\tilde{t}_l, 2\pi \tilde{f}^l_{k,uv})| = VUV_l \times |G_l| \prod_{j=1}^{r} \frac{\left|1 + \sum_{q=1}^{Q/r} b^l_{j,q} e^{-i2\pi \tilde{f}^l_{k,uv} q}\right|}{\left|1 + \sum_{p=1}^{P/r} a^l_{j,p} e^{-i2\pi \tilde{f}^l_{k,uv} p}\right|}, \tag{5.41}$$

where VUV_l is a binary flag indicating whether the l-th frame is voiced ($VUV_l = 1$) or unvoiced ($VUV_l = 0$).

4. **Phase delay estimation:** To ensure smooth phase continuity, the phase delays are calculated separately for voiced and unvoiced components. This is done by applying the frequencies from Eq. (5.39) into the phase response function in Eq. (4.14), i.e., $\angle H(\tilde{t}_l, 2\pi \tilde{f}^l_{k,v})$ and

$\angle H(\tilde{t}_l, 2\pi\tilde{f}^l_{k,uv})$.

5. **Phase estimation:** The resulting phase delays are employed to calculate the framewise phases for both voiced and unvoiced speech using Eq. (4.15), i.e., $\tilde{\varphi}^u_k(\tilde{t}_l)$ and $\tilde{\varphi}^{uv}_k(\tilde{t}_l)$. The phase of the excitation signal in Eq. (4.14) can then be computed efficiently as follows:

$$\begin{aligned}\varphi^u_k(\tilde{t}_l) &= \int_0^{\tilde{t}_l} 2\pi\tilde{f}_k(u)\mathrm{d}u \\ &\approx \pi\sum_{i=1}^{l}[\tilde{f}^{i-1}_k + \tilde{f}^i_k](t_i - t_{i-1})\beta_l,\end{aligned} \tag{5.42}$$

where $\tilde{f}^i_k$ should be substituted with either $\tilde{f}^l_{k,v}$ or $\tilde{f}^l_{k,uv}$ depending on the voicing. The framewise phases of all components can be computed as

$$\tilde{\varphi}_{k,v}(\tilde{t}_l) = \pi\sum_{i=1}^{l}[\tilde{f}^{i-1}_{k,v} + \tilde{f}^i_{k,v}](t_i - t_{i-1})\beta_l + \angle H(t_l, 2\pi\tilde{f}^l_{k,v}) \tag{5.43}$$

$$\tilde{\varphi}_{k,uv}(\tilde{t}_l) = \pi\sum_{i=1}^{l}[\tilde{f}^{i-1}_{k,uv} + \tilde{f}^i_{k,uv}](t_i - t_{i-1})\beta_l + \angle H(\tilde{t}_l, 2\pi\tilde{f}^l_{k,uv}) \tag{5.44}$$

6. **Instantaneous parameter interpolation:** Linear and cubic interpolation methods are applied to obtain the instantaneous amplitude and phase for each harmonic component:

$$\tilde{A}^v_k(\tilde{t}), \quad \tilde{A}^{uv}_k(\tilde{t}), \quad \tilde{\varphi}_{k,v}(\tilde{t}), \quad \tilde{\varphi}_{k,uv}(\tilde{t}).$$

7. **Final synthesis:** The modified speech waveform is synthesized by summing the quasi-harmonic components for both voiced and unvoiced parts:

$$\begin{aligned}\tilde{x}(\tilde{t}) =& \tilde{x}_{uv}(\tilde{t}) + \tilde{x}_v(\tilde{t}) \\ =& \sum_{k=-K}^{K}\tilde{A}^v_k(\tilde{t})e^{i\tilde{\varphi}_{k,v}(\tilde{t})} + \sum_{k=-K}^{K}\tilde{A}^{uv}_k(\tilde{t})e^{i\tilde{\varphi}_{k,uv}(\tilde{t})}.\end{aligned} \tag{5.45}$$

A pseudocode is also elaborated in Algorithm 7. Additionally, an example of time-scale speech modification and an example of pitch-scaled speech modification based on QHARMA-GAN methods are illustrated in Figs. 5.9 and 5.11, respectively, where the corresponding spectrograms are demonstrated in Figs. 5.10 and 5.12, respectively.

5.5 Experimental Evaluations

5.5.1 Experimental Design and Evaluation Aspects

To comprehensively assess the performance of the proposed methods, including QHM-GAN and QHARMA-GAN, we conduct a series of experiments and compare them against several state-of-

Algorithm 7 Time- and pitch-scale speech modification based on QHARMA-GAN.

Step 1: Preprocessing and Setting
Train the QHARMA-GAN with the dataset and infer the ARMA coefficients from DNN part; set the time-scale factor β_l and the pitch-scale factor ρ_l;
Step 2: Parameters modification
Scale the time instants by Eq. (5.38);
Shift the frequencies of unvoiced and voiced segments by Eq. (5.39);
Obtain the framewise amplitudes of all components at unvoiced and voiced segments, respectively, by Eqs. (5.40) and (5.41);
Linearly interpolate the framewise amplitudes to obtain the instantaneous amplitudes;
Obtain the phase delay of all components at all frame centers, i.e., $\angle H(\tilde{t}_l, 2\pi\tilde{f}^l_{k,v})$ and $\angle H(\tilde{t}_l, 2\pi\tilde{f}^l_{k,uv})$.;
Calculate the phase of all components at all frame centers by Eqs. (5.43) and (5.44);
Cubically interpolate the framewise phases to obtain the instantaneous phases;
Step 3: Generation
$\tilde{x}(t) \leftarrow \sum_{k=-K}^{K} \tilde{A}^v_k(\tilde{t}) e^{i\tilde{\varphi}_{k,v}(\tilde{t})} + \sum_{k=-K}^{K} \tilde{A}^{uv}_k(\tilde{t}) e^{i\tilde{\varphi}_{k,uv}(\tilde{t})}$;
Output: $\hat{x}(t)$

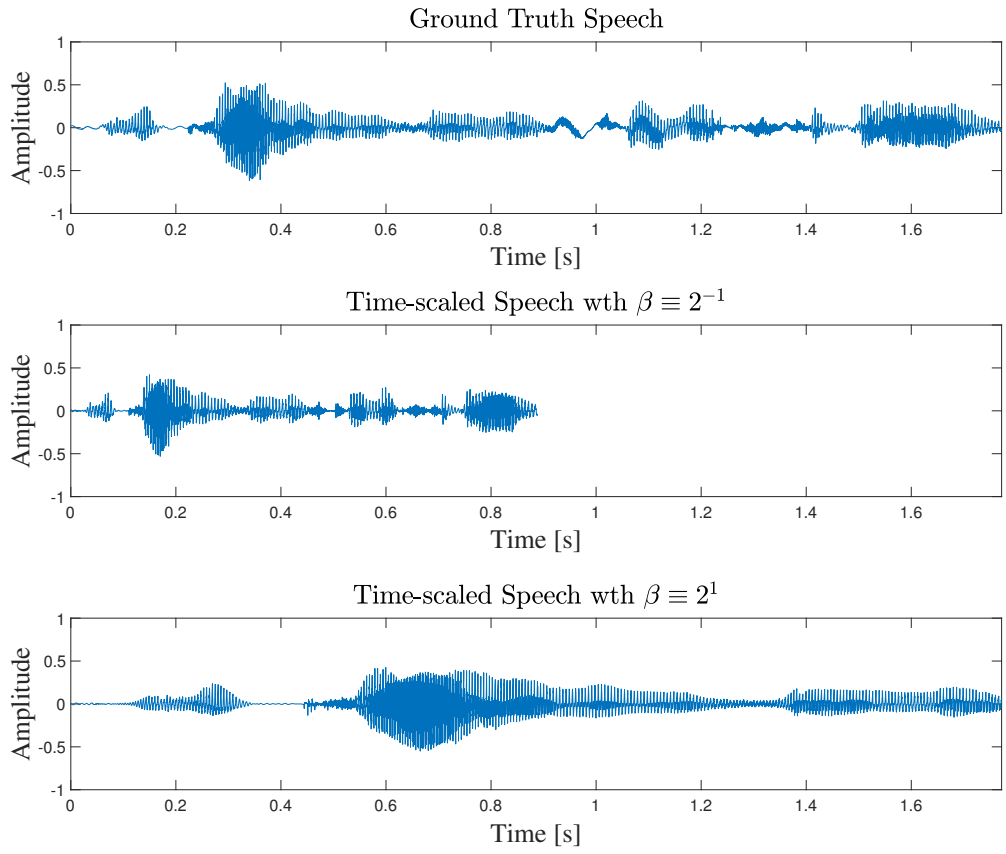


Figure 5.9: The waveforms of ground truth and time-scaled speech based on QHARMA-GAN.

the-art neural vocoders. The evaluation focuses on time-scale and pitch-scale speech modification quality. Both objective and subjective metrics are employed to quantify the modification performance.

The detailed experimental conditions and results are presented and discussed in the following subsections. The experiments consist of four parts.

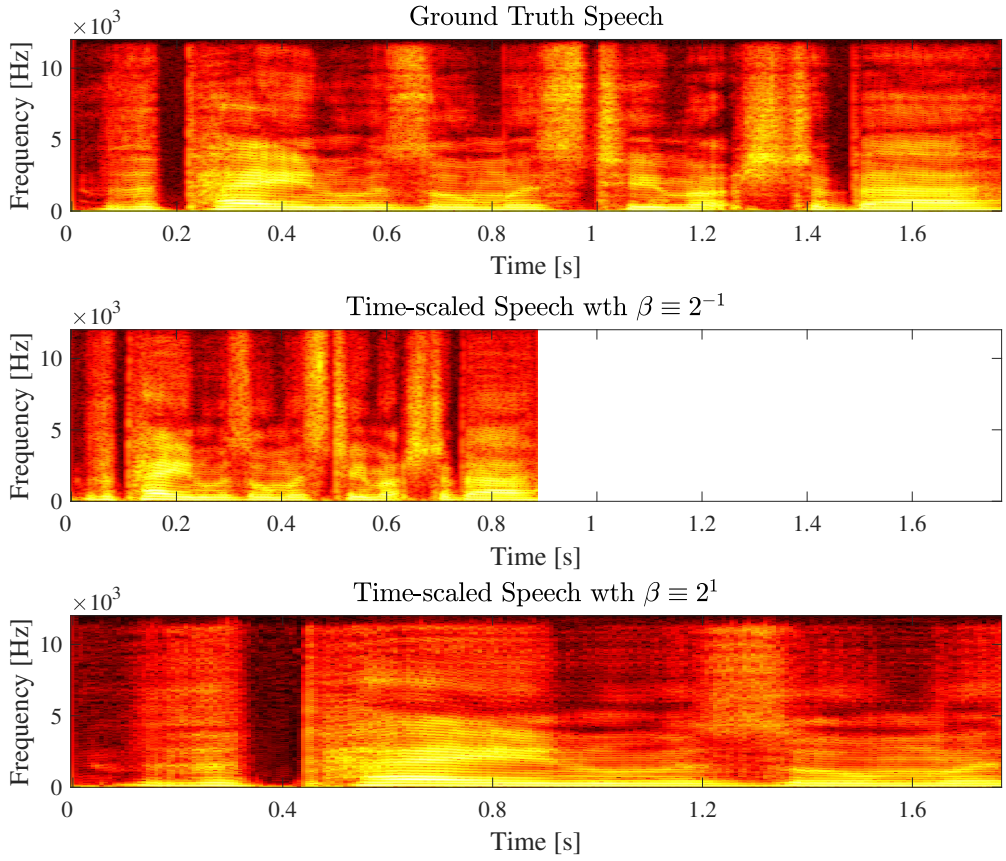


Figure 5.10: The spectrograms of ground truth and time-scaled speech based on QHARMA-GAN.

(1) **The experiments for time-scale modification.**

For the time-scale modification, we use objective measurement indicators to assess the performance of all methods (including WORLD, QHM, QHM-GAN, and QHARMA-GAN) in terms of time-scale modification quality. The experimental results show the comparable performances of QHM and QHARMA-GAN, both outperforming WORLD and QHM-GAN.

(2) **The preliminary experiments for pitch-scale modification to show the best vocoders in the candidate methods.**

For the pitch-scale modification, first, we use objective measurement indicators to assess the performances of all neural vocoders in terms of modification quality. Subsequently, the screened-out methods, including the best one of other neural vocoders and the best one between QHM-GAN and QHARMA-GAN, will be compared with the conventional vocoders, such as WORLD and QHM.

(3) **The main experiment for pitch-scale modification of selected neural and conventional vocoders.**

Second, both objective and subjective evaluation metrics are jointly employed to comprehensively assess the quality of the speech modified by all candidate methods. The objective

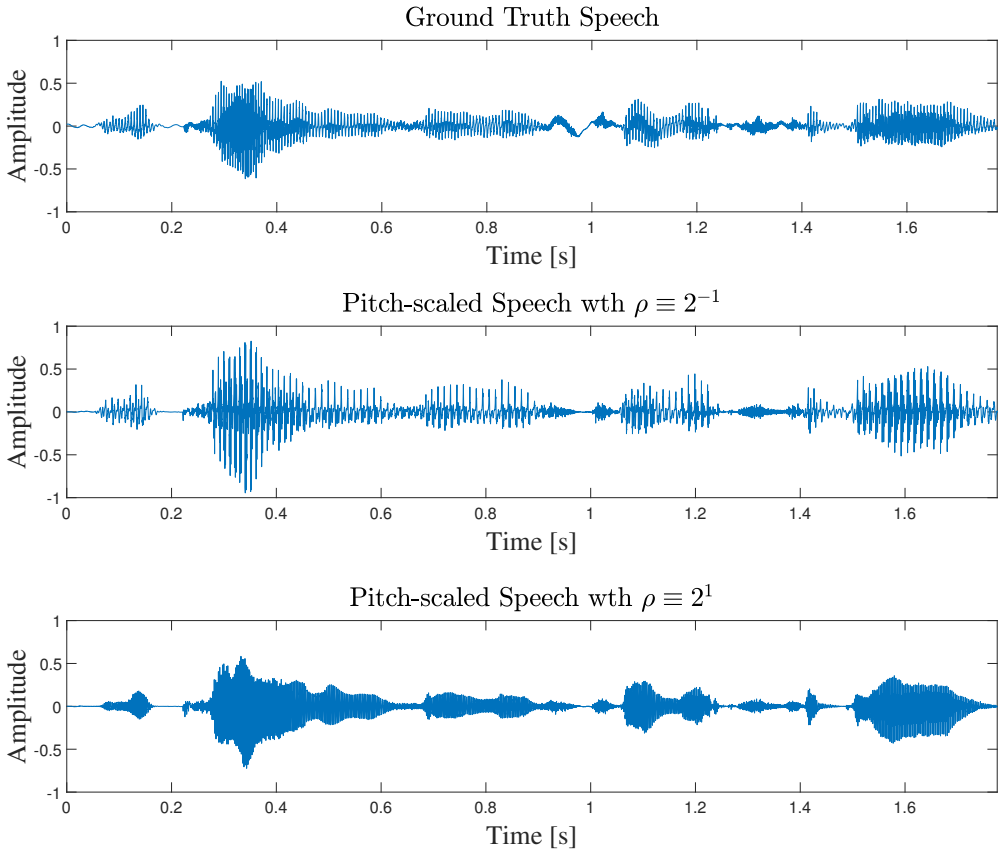


Figure 5.11: The waveforms of ground truth and pitch-scaled speech based on QHARMA-GAN.

metrics provide quantifiable measurements of intelligibility and spectral fidelity, while the subjective evaluations reflect human perceptual preferences and naturalness. This combined evaluation strategy ensures a thorough and balanced assessment of the synthesis performance.

(4) **The generalization ability of neural vocoders in terms of pitch modification.**

Finally, to evaluate the generalization capability, which is a critical property for neural vocoders, we assess the performance of all candidate neural vocoders with an out-of-distribution (OOD) generalization. These evaluations aim to verify whether the models can maintain modification quality when confronted with unseen conditions.

Through these evaluations, we aim to demonstrate the superior performance of QHM-GAN and QHARMA-GAN in pitch-scaled speech modification.

5.5.2 Experimental Conditions

This section provides a detailed description of the experimental setup.

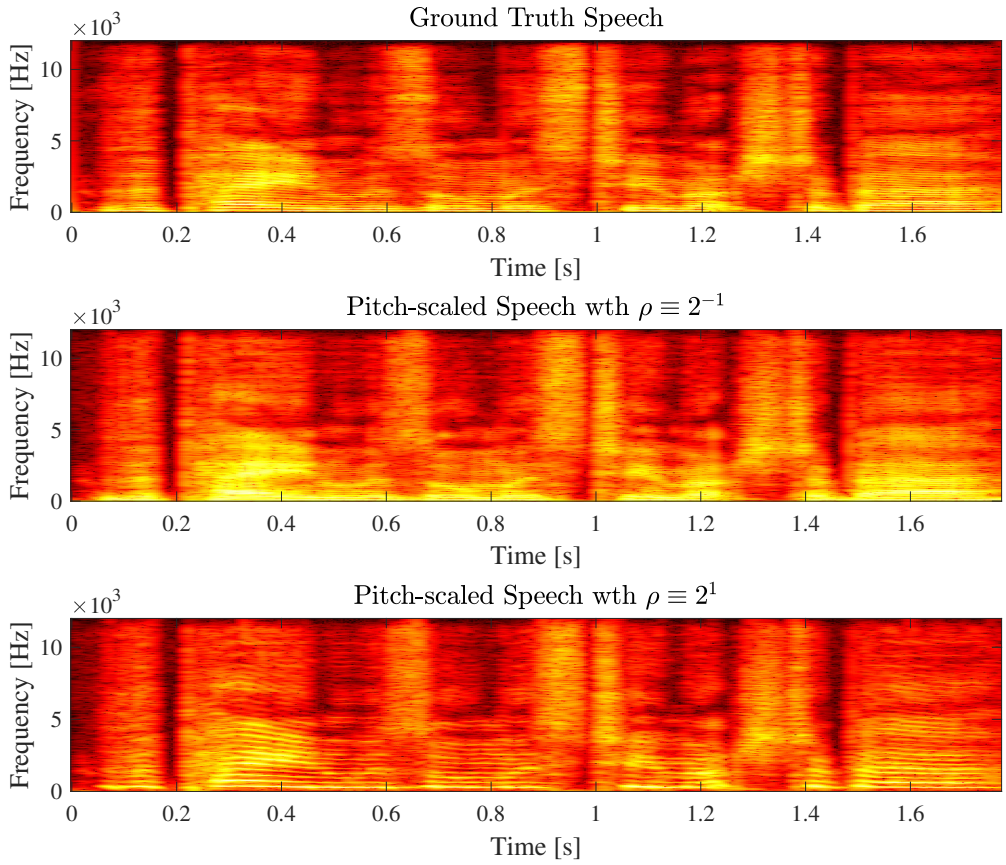


Figure 5.12: The spectrograms of ground truth and pitch-scaled speech based on QHARMA-GAN.

Model Settings: All models used in the experiments are those trained in Section III of Chapter IV without any additional fine-tuning.

Dataset: The experimental utterances are randomly selected from three widely used open-source corpora, covering both single-speaker and multi-speaker scenarios. Specifically, we use the VCTK corpus [95], which contains recordings from 110 English speakers sampled at 24 kHz, and the Japanese Versatile Speech (JVS) corpus [96], which includes speech from 100 Japanese speakers, also sampled at 24 kHz. The data splitting strategy follows the same configuration as in Section III of Chapter IV.

Evaluation Metrics: To objectively and subjectively assess the performance of QHM-GAN and QHARMA-GAN against baseline neural and conventional vocoders, we adopt a comprehensive set of evaluation metrics, including **V/UV Error Rate**, f_0 **RMSE**, **UTMOS**, **MCD**, and **MOS**. These indicators are used to quantify the pitch accuracy, spectral fidelity, V/UV classification accuracy, intelligibility, and perceptual naturalness.

Regarding the model candidates, Vocos [5] and hn-NSF [7] are selected as representative neural vocoders, while QHM and WORLD are adopted as conventional CSP-based vocoders.

These models are compared with the proposed QHM-GAN and QHARMA-GAN. Additionally, a lightweight variant of QHARMA-GAN, referred to as QHARMA-GAN-small, is included to investigate the trade-off between model complexity and performance. An overview of all evaluated models is provided below:

1) **WORLD**: A conventional source-filter vocoder based on the CSP framework that enables flexible manipulation of acoustic features such as f_0. It decomposes speech into spectral envelope, aperiodicity, and f_0, and reconstructs the waveform through signal processing. Due to its interpretability and decent synthesis quality, WORLD is widely adopted as a baseline in vocoding and voice conversion tasks.

2) **QHM**: A quasi-harmonic modeling method that incorporates frequency correction to refine the harmonic structure of speech. It estimates frame-wise complex amplitudes for each quasi-harmonic component, allowing for adaptive frequency adjustment. The corrected amplitudes and frequencies are then used to generate high-fidelity speech, which can be further modified in both time and frequency domains via the DAP representation.

3) **Vocos**: The pitch-controllable variant of Vocos proposed in [99] integrates external harmonic components derived from f_0 priors, enabling flexible pitch editing under a source-filter framework. In this study, we adopt a similar approach for f_0 extrapolation. Multi-Period Discriminator (MPD) and Multi-Resolution Discriminator (MRD) are utilized during training.

4) **hn-NSF**: A neural source-filter vocoder that employs a harmonic-plus-noise excitation signal derived from f_0 priors. The excitation is passed through a learnable filter that adjusts its spectral response to produce the final waveform. Adversarial training with HiFi-GAN discriminators is employed to improve naturalness and fidelity. This architecture allows for direct pitch modification and has been widely applied in pitch-controllable synthesis tasks.

5) **QHM-GAN**: Our proposed model that integrates CSP-based QHM with neural networks to achieve flexible and high-quality speech synthesis. A neural network maps mel-spectrograms to frame-wise complex amplitudes and phase compensation terms, which are then fed into the QHM synthesizer to generate the waveform. Adversarial training using MPD and MRD improves output fidelity and perceptual quality.

6) **QHARMA-GAN**: An enhanced version of QHM-GAN that incorporates autoregressive moving average (ARMA) modeling to improve spectral envelope representation. The architecture remains similar to QHM-GAN, consisting of four MRF modules, but instead of predicting complex amplitudes directly, the network estimates ARMA coefficients from the input mel-spectrogram. These coefficients guide the synthesis process, allowing explicit

control over spectral shaping. To improve both stability and perceptual quality, MRD, MSD (Multi-Scale Discriminator), and MPD are used during training.

7) **QHARMA-GAN-small**: A compact and efficient variant of QHARMA-GAN. It comprises three MRF modules with fixed dilation rates, following a similar design to QHM-GAN-small. The model predicts ARMA coefficients from mel-spectrograms, and MRD, MSD, and MPD are adopted to maintain synthesis quality despite the reduced model capacity.

5.5.3 The Experiments for Time-scale Modification

This section presents a comprehensive comparison of time-scale speech modification performance among several widely used vocoders, including conventional vocoders (WORLD and QHM) and our proposed models (QHM-GAN and QHARMA-GAN).

Table 5.1 reports the performance under four time-scale factors: $\beta_l \equiv 2^{-1}$, $\beta_l \equiv 2^{-0.5}$, $\beta_l \equiv 2^{0.5}$, and $\beta_l \equiv 2^{1}$. Among all methods, WORLD achieves the best V/UV error rate, indicating its strength in preserving voicing structure. However, it exhibits relatively high f_0 RMSE, suggesting inaccurate pitch preservation during time-scaling. This is mainly because WORLD is unable to correct the frequency estimates, which are usually insufficiently accurate.

In contrast, QHM shows the lowest f_0 RMSE among all models, demonstrating its excellent ability to maintain pitch accuracy during time-scale modification. This stems from its waveform-shape-invariant modeling, which directly controls frequency trajectories. In terms of spectral fidelity, QHM achieves the lowest MCD due to its inherent waveform preservation, which ensures a consistent spectral envelope after time-scale modification. However, QHM yields poor V/UV accuracy, as it uses sinewaves to model both voiced and unvoiced segments. This modeling leads to confusion during voiced/unvoiced classification, where unvoiced segments are often misidentified as voiced.

A similar tendency is observed in QHM-GAN and QHARMA-GAN. Their QHM-based structures inherit the same limitations regarding unvoiced segment modeling. As a result, their V/UV error rates are only slightly better than that of QHM, though still far worse than WORLD.

The proposed QHARMA-GAN also demonstrates competitive performance: it achieves relatively low f_0 RMSE, similar to QHM, and its UTMOS score is close to QHM, indicating comparable perceptual quality. Although its MCD is slightly higher, suggesting mild spectral distortion, QHARMA-GAN still maintains better overall balance among all evaluation metrics.

In summary, QHM and QHARMA-GAN demonstrate the best time-scale modification performance. QHM excels in pitch accuracy and spectral fidelity, while QHARMA-GAN offers comparable quality with added flexibility and robustness from neural modeling. Both significantly outperform conventional neural vocoders in this task.

Table 5.1: Results of objective evaluations for time-scale modification. The UTMOS values of the ground truth samples for VCTK and JVS without modification were 4.04 and 3.63, respectively.

Metric	Dataset	Scale	WORLD	QHM	QHM-GAN	QHARMA-GAN
V/UV rate [%] ↓	VCTK	$\beta_l \equiv 2^{-1}$	**13**	17	18	15
		$\beta_l \equiv 2^{-0.5}$	**12**	16	16	14
		$\beta_l \equiv 2^{0.5}$	**11**	17	15	13
		$\beta_l \equiv 2^{1}$	**11**	20	15	15
	JVS	$\beta_l \equiv 2^{-1}$	**12**	16	18	**12**
		$\beta_l \equiv 2^{-0.5}$	**11**	16	16	**11**
		$\beta_l \equiv 2^{0.5}$	**11**	18	16	**11**
		$\beta_l \equiv 2^{1}$	**11**	22	17	**11**
f_0 RMSE [Hz] ↓	VCTK	$\beta_l \equiv 2^{-1}$	0.07	**0.05**	**0.05**	**0.05**
		$\beta_l \equiv 2^{-0.5}$	0.06	**0.04**	0.05	0.05
		$\beta_l \equiv 2^{0.5}$	0.06	**0.04**	0.05	0.05
		$\beta_l \equiv 2^{1}$	0.06	**0.04**	0.06	0.06
	JVS	$\beta_l \equiv 2^{-1}$	0.12	**0.08**	0.09	0.11
		$\beta_l \equiv 2^{-0.5}$	0.11	**0.07**	0.09	0.09
		$\beta_l \equiv 2^{0.5}$	0.11	**0.05**	0.10	0.09
		$\beta_l \equiv 2^{1}$	0.10	**0.05**	0.11	0.09
MCD [dB] ↓	VCTK	$\beta_l \equiv 2^{-1}$	3.73	**3.48**	4.27	5.28
		$\beta_l \equiv 2^{-0.5}$	3.25	**3.03**	4.17	4.31
		$\beta_l \equiv 2^{0.5}$	2.68	**2.66**	4.33	4.67
		$\beta_l \equiv 2^{1}$	**2.54**	2.67	4.46	5.21
	JVS	$\beta_l \equiv 2^{-1}$	3.55	**2.99**	4.01	4.91
		$\beta_l \equiv 2^{-0.5}$	3.19	**2.63**	3.99	4.29
		$\beta_l \equiv 2^{0.5}$	2.73	**2.34**	4.25	4.72
		$\beta_l \equiv 2^{1}$	2.62	**2.34**	4.41	5.54
UTMOS ↑	VCTK	$\beta_l \equiv 2^{-1}$	2.41	**2.54**	2.15	2.50
		$\beta_l \equiv 2^{-0.5}$	3.34	**3.56**	3.09	3.49
		$\beta_l \equiv 2^{0.5}$	3.25	**3.57**	2.61	**3.57**
		$\beta_l \equiv 2^{1}$	2.76	3.13	1.89	**3.19**
	JVS	$\beta_l \equiv 2^{-1}$	1.68	**1.87**	1.46	1.85
		$\beta_l \equiv 2^{-0.5}$	2.52	**2.91**	2.09	2.86
		$\beta_l \equiv 2^{0.5}$	2.61	3.23	2.26	**3.24**
		$\beta_l \equiv 2^{1}$	2.29	2.89	1.90	**2.90**

5.5.4 Preliminary Pitch-scale Modification Experiments for Baseline Screening

This section presents the preliminary comparison between the performances of several widely used neural vocoders (including Vocos and hn-NSF) and our proposed models in terms of speech

modification quality.

Table 5.2 presents the performance of f_0 extrapolation under different pitch-scale factors: $\rho_l \equiv 2^{-1}$, $\rho_l \equiv 2^{-0.5}$, $\rho_l \equiv 2^{0.5}$, and $\rho_l \equiv 2^{1}$. Among all methods, Vocos attains the best MCD score, with UTMOS comparable to hn-NSF. However, its notably high f_0 RMSE indicates a failure in accurate pitch modification.

By contrast, hn-NSF shows relatively stable f_0 RMSE values, suggesting effective pitch control, but it struggles to maintain synthesis quality under extreme pitch-scale conditions (i.e., $\rho \equiv 2^{-1}$ or 2^{1}).

For the proposed methods, QHM-GAN achieves a satisfactory f_0 RMSE, reflecting reliable pitch control. Nevertheless, it has the worst MCD scores due to insufficient resonance modeling and inaccurate amplitude estimation, resulting in considerable spectral distortion and lower perceptual quality, as evidenced by its lowest UTMOS.

QHARMA-GAN, on the other hand, exhibits consistently strong performance across all pitch-scale factors. Its low V/UV error rates and minimal f_0 RMSE demonstrate accurate pitch extrapolation with little spectral distortion. The higher UTMOS further confirms that QHARMA-GAN provides the most effective and perceptually natural f_0 extrapolation among the evaluated methods.

To further demonstrate f_0 extrapolation performance, Fig. 5.13 shows spectrograms of a pitch-scaled speech sample ($\rho \equiv 2^{1}$) generated by different methods. It is clear that Vocos fails to produce the correct pitch, and both Vocos and hn-NSF have difficulty reconstructing high-frequency harmonics. In contrast, QHM-GAN and QHARMA-GAN successfully synthesize these harmonics, producing clear and sharp high-frequency trajectories in the spectrogram. Nonetheless, QHM-GAN exhibits some aliasing artifacts. Overall, QHARMA-GAN achieves the best performance in f_0 extrapolation.

Based on the above analysis, the main experiment focuses on comparing QHARMA-GAN with conventional vocoders (QHM and WORLD) regarding their f_0 extrapolation capabilities.

5.5.5 Evaluation of Pitch-scale Modification Quality

This section presents the performance of selected methods, including WORLD, QHM, QHARMA-GAN, and QHARMA-GAN-small, in terms of f_0 manipulation, specifically focusing on f_0 extrapolation. A total of 1885 utterances from VCTK and 980 utterances from JVS were pitch-scaled using factors 2^{-1}, $2^{-0.5}$, $2^{0.5}$, and 2^{1}, i.e., $\rho_l \equiv 2^{-1}$, $\rho_l \equiv 2^{-0.5}$, $\rho_l \equiv 2^{0.5}$, and $\rho_l \equiv 2^{1}$. The averaged quantitative results are summarized in Table 5.3.

Consistent with previous findings, QHM achieved the best performance in terms of f_0 RMSE due to its effective frequency adaptation, while QHARMA-GAN obtained the best V/UV detection rate. From the perspective of the pitch-scaled speech quality, QHM exhibited the poorest performance among the compared methods. This is attributed to its lack of explicit spectral enve-

Table 5.2: Results of objective evaluations for f_0 modification. The UTMOS values of the ground truth samples for VCTK and JVS without modification were 4.04 and 3.63, respectively.

Metric	Dataset	Scale	Vocos	hn-NSF	QHM-GAN	QHARMA-GAN
V/UV rate [%] ↓	VCTK	$\rho_l \equiv 2^{-1}$	15	15	18	**14**
		$\rho_l \equiv 2^{-0.5}$	**12**	**12**	16	13
		$\rho_l \equiv 2^{0.5}$	14	14	17	**11**
		$\rho_l \equiv 2^{1}$	**13**	27	18	**13**
	JVS	$\rho_l \equiv 2^{-1}$	15	19	17	**14**
		$\rho_l \equiv 2^{-0.5}$	14	14	13	**12**
		$\rho_l \equiv 2^{0.5}$	13	13	13	**12**
		$\rho_l \equiv 2^{1}$	**14**	15	15	**14**
f_0 RMSE [Hz] ↓	VCTK	$\rho_l \equiv 2^{-1}$	0.65	0.29	0.32	**0.11**
		$\rho_l \equiv 2^{-0.5}$	0.33	0.09	**0.08**	**0.08**
		$\rho_l \equiv 2^{0.5}$	0.34	0.10	**0.08**	0.09
		$\rho_l \equiv 2^{1}$	0.67	0.32	**0.07**	0.10
	JVS	$\rho_l \equiv 2^{-1}$	0.64	0.28	0.20	**0.15**
		$\rho_l \equiv 2^{-0.5}$	0.34	0.15	0.14	**0.12**
		$\rho_l \equiv 2^{0.5}$	0.35	0.15	**0.14**	**0.14**
		$\rho_l \equiv 2^{1}$	0.69	0.19	**0.11**	0.12
MCD [dB] ↓	VCTK	$\rho_l \equiv 2^{-1}$	**3.59**	5.52	8.81	5.28
		$\rho_l \equiv 2^{-0.5}$	**3.76**	4.60	5.51	4.31
		$\rho_l \equiv 2^{0.5}$	**3.78**	4.98	4.96	4.67
		$\rho_l \equiv 2^{1}$	**3.95**	5.83	7.27	5.21
	JVS	$\rho_l \equiv 2^{-1}$	**3.62**	5.19	6.97	4.91
		$\rho_l \equiv 2^{-0.5}$	**3.63**	4.49	5.51	4.29
		$\rho_l \equiv 2^{0.5}$	**3.82**	4.78	5.34	4.72
		$\rho_l \equiv 2^{1}$	**3.88**	5.73	6.71	5.54
UTMOS ↑	VCTK	$\rho_l \equiv 2^{-1}$	**3.49**	3.16	1.53	3.14
		$\rho_l \equiv 2^{-0.5}$	2.78	3.60	2.22	**3.61**
		$\rho_l \equiv 2^{0.5}$	2.81	2.93	2.47	**3.25**
		$\rho_l \equiv 2^{1}$	**3.03**	2.00	1.62	2.68
	JVS	$\rho_l \equiv 2^{-1}$	**2.80**	2.34	1.32	2.66
		$\rho_l \equiv 2^{-0.5}$	1.83	2.79	1.57	**3.05**
		$\rho_l \equiv 2^{0.5}$	1.73	2.22	1.87	**2.23**
		$\rho_l \equiv 2^{1}$	1.83	1.37	1.46	**1.84**

lope modeling and its reliance solely on a shape-invariant modification algorithm, which hinders accurate modification of the speech signal. Furthermore, the insufficient number of components representing unvoiced segments becomes more problematic when increasing f_0, further degrading

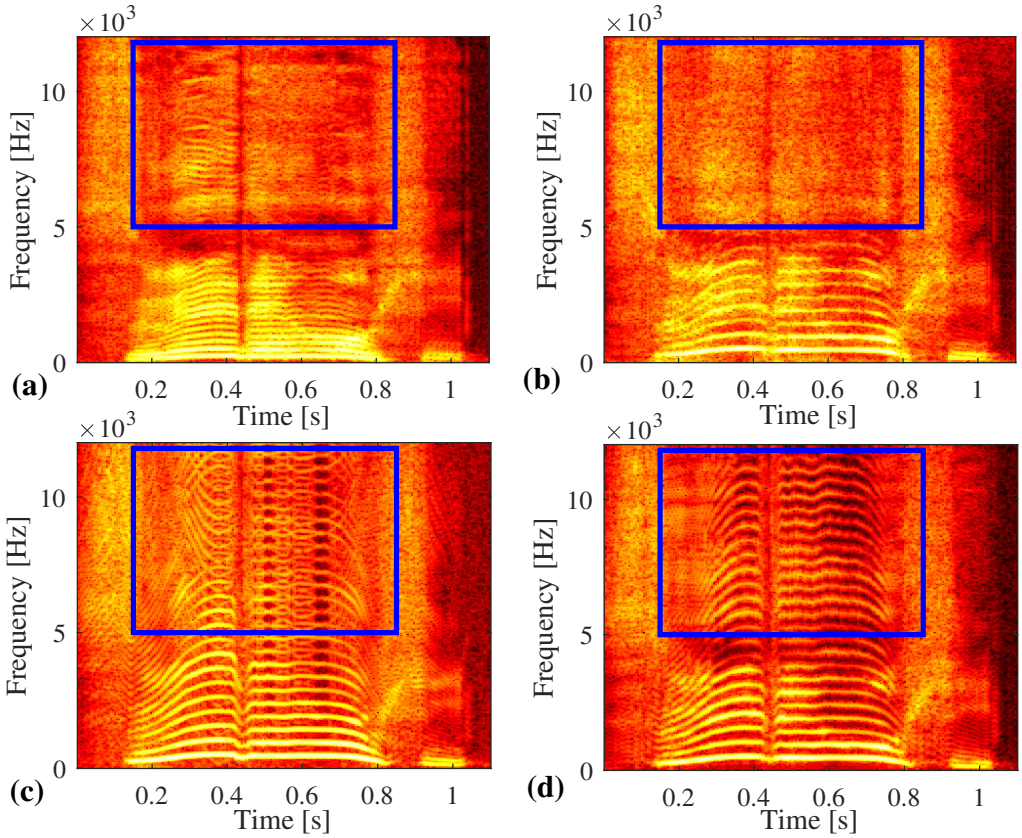


Figure 5.13: The spectrograms of the speech generated by (a) Vocos, (b) hn-NSF, (c) QHM-GAN, and (d) QHARMA-GAN with $\rho \equiv 2^1$.

the quality of pitch-scaled speech.

By contrast, QHARMA-GAN and WORLD achieve better results due to their explicit modeling of the spectral envelope. As shown by the MOS scores in Table 5.3, the overall performances are comparable: QHARMA-GAN, including its smaller variant QHARMA-GAN-small, notably outperforms WORLD in pitch-lowering scenarios, whereas WORLD performs better in pitch-raising cases.

A preliminary analysis indicates that this pattern arises since QHARMA-GAN depends on V/UV detection to selectively modify speech segments. Errors in V/UV detection can lead to unvoiced segments being misclassified as voiced, reducing the number of components used in synthesis and consequently degrading speech quality. Therefore, accurate V/UV estimation remains a key challenge and an important direction for future research.

5.5.6 Evaluation of Generalization Ability in Pitch-scale Modification

In this part, we evaluate the generalization ability of neural vocoders in pitch-scale modification when applied to speech signals that are unseen during training. Specifically, all models, including Vocos, hn-NSF, and the proposed QHARMA-GAN, are trained using the JVS corpus, which consists of Japanese speech. Cross-domain robustness is assessed using the OpenSinger dataset

Table 5.3: Results of objective and subjective evaluations. The MOS of the ground truth samples for VCTK and JVS datasets without modification were 4.27 ± 0.022 and 3.88 ± 0.038, respectively.

Metric	Dataset	Scale	WORLD	QHM	QHARMA-GAN	QHARMA-GAN-small
V/UV rate [%] ↓	VCTK	$\rho_l \equiv 2^{-1}$	15	18	**14**	16
		$\rho_l \equiv 2^{-0.5}$	14	18	**13**	14
		$\rho_l \equiv 2^{0.5}$	13	19	**11**	13
		$\rho_l \equiv 2^{1}$	**13**	21	**13**	14
	JVS	$\rho_l \equiv 2^{-1}$	16	16	**14**	**14**
		$\rho_l \equiv 2^{-0.5}$	14	14	**12**	13
		$\rho_l \equiv 2^{0.5}$	14	18	12	**11**
		$\rho_l \equiv 2^{1}$	15	22	**13**	**13**
f_0 RMSE [Hz] ↓	VCTK	$\rho_l \equiv 2^{-1}$	0.13	**0.11**	**0.11**	**0.11**
		$\rho_l \equiv 2^{-0.5}$	0.10	**0.06**	0.08	0.09
		$\rho_l \equiv 2^{0.5}$	0.11	**0.04**	0.09	0.10
		$\rho_l \equiv 2^{1}$	0.16	**0.04**	0.10	0.11
	JVS	$\rho_l \equiv 2^{-1}$	0.17	**0.13**	0.15	0.17
		$\rho_l \equiv 2^{-0.5}$	0.14	**0.08**	0.12	0.13
		$\rho_l \equiv 2^{0.5}$	0.16	**0.05**	0.12	0.12
		$\rho_l \equiv 2^{1}$	0.21	**0.05**	0.12	0.14
MOS ↑	VCTK	$\rho_l \equiv 2^{-1}$	2.98±0.056	2.43±0.043	**3.01±0.050**	2.99±0.044
		$\rho_l \equiv 2^{-0.5}$	3.84±0.052	3.11±0.036	**3.86±0.033**	3.62±0.057
		$\rho_l \equiv 2^{0.5}$	**3.78±0.084**	3.29±0.072	3.66±0.066	3.29±0.072
		$\rho_l \equiv 2^{1}$	**2.82±0.071**	2.33±0.064	2.72±0.057	2.69±0.053
	JVS	$\rho_l \equiv 2^{-1}$	3.03±0.052	2.07±0.025	**3.11±0.044**	3.08±0.045
		$\rho_l \equiv 2^{-0.5}$	3.64±0.042	2.85±0.052	**3.75±0.049**	3.72±0.039
		$\rho_l \equiv 2^{0.5}$	**3.82±0.033**	3.19±0.058	3.60±0.053	3.63±0.054
		$\rho_l \equiv 2^{1}$	**3.08±0.040**	2.50±0.040	2.91±0.051	2.89±0.056

[100], which includes Mandarin and Cantonese songs by both male and female singers. Twenty songs were randomly chosen to ensure gender balance and stylistic diversity, resulting in 727 utterances for evaluation.

Table 5.4 summarizes the f_0 extrapolation performance under four pitch-scale factors: $\rho_l \equiv 2^{-1}$, $\rho_l \equiv 2^{-0.5}$, $\rho_l \equiv 2^{0.5}$, and $\rho_l \equiv 2^{1}$. Among all models, Vocos achieves the lowest MCD scores, suggesting a strong ability to reconstruct the general spectral shape. Its UTMOS is also comparable to that of hn-NSF. However, Vocos exhibits significantly higher f_0 RMSE values across all

scaling factors, which indicates a failure to track and modify the pitch trajectory accurately. This problem is particularly pronounced under large pitch-shift conditions and was also observed in the preliminary experiments reported in Table 5.2. The discrepancy between its spectral fidelity (low MCD) and pitch accuracy (high f_0 RMSE) highlights a fundamental limitation of Vocos in prosodic control.

In contrast, hn-NSF demonstrates relatively stable f_0 RMSE values, implying its effectiveness in pitch control. However, its synthesis quality degrades under extreme pitch manipulations, as evidenced by increased MCD and reduced UTMOS when $\rho \equiv 2^{-1}$ or $\rho \equiv 2^{1}$. For instance, hn-NSF yields f_0 RMSE values of 0.37 and 0.22 under these conditions, respectively, which are considerably higher than those of QHARMA-GAN. These results suggest that although hn-NSF retains basic pitch-tracking capability, it fails to preserve the naturalness and harmonic structure of the singing voice under substantial pitch changes.

On the other hand, QHARMA-GAN achieves consistently strong performance across all pitch-scale factors. Its low V/UV error rates and stable f_0 RMSE values demonstrate accurate control over pitch modification, even for out-of-distribution singing voices. Furthermore, its ability to model both amplitude and phase delay through ARMA mechanisms allows it to synthesize harmonics more naturally and preserve the spectral envelope effectively. As a result, QHARMA-GAN not only minimizes spectral distortion but also yields the highest UTMOS scores across all conditions, confirming its superiority in both objective and perceptual metrics.

In summary, QHARMA-GAN exhibits better generalization in cross-domain pitch modification tasks, achieving both accurate f_0 extrapolation and high-quality synthesis. These results highlight the robustness and adaptability of the proposed framework when applied to real-world applications involving diverse languages, musical content, and vocal styles.

5.6 Summary

In this chapter, we introduce speech modification methods based on QHM methods. These methods estimate the parameters of quasi-harmonics and employ additional processing to modify them while preserving the spectral envelope, with the aim of generating modified speech. However, the additional processing methods are typically based on conventional signal processing techniques, which often lead to inaccurate parameter modification.

Given the well-known robustness of neural-based approaches, we are motivated to utilize neural networks to estimate parameters for speech modification. Accordingly, we propose QHM-GAN and QHARMA-GAN, two novel speech modification methods based on a hybrid neural vocoder framework. These models enable real-time speech modification in terms of both time-scale and pitch-scale transformation. Benefiting from the robustness of deep neural networks trained on large-scale datasets, our methods are capable of estimating parameters with high accuracy, which are then used to synthesize modified speech.

Table 5.4: Results of objective evaluations for OOD in terms of f_0 modification. The UTMOS of the ground truth samples without modification was 2.42.

Metric	Scale	Vocos	hn-NSF	QHARMA-GAN
V/UV rate [%] ↓	$\rho_l \equiv 2^{-1}$	**9**	13	10
	$\rho_l \equiv 2^{-0.5}$	**9**	10	**9**
	$\rho_l \equiv 2^{0.5}$	**9**	**9**	**9**
	$\rho_l \equiv 2^{1}$	9	15	**8**
f_0 RMSE [Hz] ↓	$\rho_l \equiv 2^{-1}$	0.71	0.37	**0.13**
	$\rho_l \equiv 2^{-0.5}$	0.45	0.29	**0.11**
	$\rho_l \equiv 2^{0.5}$	0.49	0.17	**0.13**
	$\rho_l \equiv 2^{1}$	0.79	0.22	**0.13**
MCD [dB] ↓	$\rho_l \equiv 2^{-1}$	**5.81**	7.33	6.37
	$\rho_l \equiv 2^{-0.5}$	**5.99**	6.81	6.43
	$\rho_l \equiv 2^{0.5}$	**6.29**	7.93	7.61
	$\rho_l \equiv 2^{1}$	**6.26**	10.42	7.67
UTMOS ↑	$\rho_l \equiv 2^{-1}$	1.98	1.58	**2.04**
	$\rho_l \equiv 2^{-0.5}$	1.44	1.69	**2.20**
	$\rho_l \equiv 2^{0.5}$	1.43	1.54	**1.58**
	$\rho_l \equiv 2^{1}$	**1.48**	1.36	1.43

Both proposed methods directly estimate the quasi-harmonic parameters required for speech generation. QHM-GAN, however, requires additional modification of the phase compensation terms for each component at every frame to avoid phase distortion, resulting in extra computational cost. Moreover, due to the lack of explicit spectral envelope modeling, QHM-GAN struggles with pitch modification, particularly under large pitch-scale factors. Since it cannot guarantee the preservation of the spectral envelope, the modified speech is often perceptually distorted.

To address these limitations, we introduce an autoregressive moving average (ARMA) model into the QHM-GAN architecture, resulting in QHARMA-GAN. The ARMA model enables accurate modeling of the time-varying spectral envelope of speech, allowing for simultaneous estimation of amplitude and phase delay while preserving the spectral shape. As a result, QHARMA-GAN can generate pitch-modified speech without introducing spectral distortion, thus achieving high-quality modification.

Thanks to the avoidance of conventional parameter modification and the real-time parameter estimation enabled by the DNN, QHARMA-GAN supports real-time pitch-scale and time-scale

modification. This makes it suitable for various practical applications, such as singing voice transformation.

Experimental results show that QHARMA-GAN produces smoother and more accurate frequency trajectories than baseline methods, and enables stable f_0 extrapolation. Moreover, the model exhibits strong generalization capabilities, suggesting that it can be trained with limited data and still generate high-quality modified speech for unseen inputs.

Chapter 6

Conclusion

6.1 Summary of Thesis

This thesis presents a comprehensive exploration into the integration of conventional signal processing algorithms with modern deep learning techniques, highlighting a synergistic framework that advances the development of neural network-based speech modeling. By leveraging the strengths of both paradigms, we proposed a hybrid system in which conventional signal processing was employed to model and synthesize the intrinsic structures of speech, including its underlying acoustic mechanisms, while neural networks were used to robustly and accurately infer the required parameters. Through this combination, the proposed neural vocoders were capable of achieving real-time speech analysis, synthesis, and modification with high fidelity. These vocoders held potential for deployment in various applications, such as text-to-speech (TTS) systems, speech restoration for patients with vocal impairments, and general-purpose speech transformation tools.

The main contributions of this thesis can be summarized in three aspects:

1. We show that the backpropagation algorithm can be effectively used in the QHM framework to optimize quasi-harmonic parameters, demonstrating the feasibility of integrating QHM into neural systems;

2. We successfully incorporated the QHM structure into neural vocoders, thereby combining the interpretability and controllability of conventional signal processing with the efficiency and accuracy of deep learning;

3. We propose real-time speech modification algorithms capable of simultaneously handling both pitch-scale and time-scale transformations while preserving the naturalness and quality of the output.

Chapter II provided a review of conventional QHM-based methods (QHM, aQHM, and eaQHM) and modern neural vocoders, analyzing their respective advantages and limitations. Conventional

QHM-based methods are shown to be effective in producing high-quality synthesis due to their strong interpretability and fine-grained control over harmonic components. However, their reliance on frame-by-frame analysis leads to computational inefficiency and limited robustness in parameter estimation. In contrast, neural vocoders demonstrate robustness and high inference efficiency, particularly when trained on large-scale datasets. Nonetheless, their black-box nature often results in a lack of interpretability, hindering acoustic feature manipulation. This contrast inspired the proposed integration strategy, which aims to exploit the complementary properties of the two approaches.

Chapter III presented the first core contribution of this thesis, i.e., demonstrating that QHM structures can be optimized through gradient-based learning. Although conventional QHM methods can refine frequencies using complex amplitudes, they are constrained by the inaccuracy of amplitude estimation, which in turn limits frequency correction and speech quality. Moreover, their frame-by-frame structure disrupts temporal coherence. To overcome these issues, we introduced a novel frequency refinement approach that iteratively corrects frequency estimates using spectrogram-based supervision, independent of amplitude estimation. In addition, we applied backpropagation to jointly optimize complex amplitudes and frequencies over entire utterances, rather than frame-wise, and propose a spectrogram-based loss function that bypasses the synthesis process to increase convexity and accelerate convergence. These innovations enhanced parameter accuracy and synthesized quality, while also confirming the feasibility of integrating QHM into differentiable deep learning architectures.

Chapter IV elaborated on the second contribution: the development of a novel vocoder framework that integrates QHM with neural networks. While QHM methods offer strong interpretability and support for acoustic manipulation, their parameter estimation is often inaccurate and time-intensive. In contrast, neural networks can learn complex parameter mappings from large datasets but typically lack interpretability. Our proposed framework combined these strengths by using neural networks to infer quasi-harmonic parameters, followed by QHM-based synthesis for waveform generation. Additionally, we incorporated autoregressive and moving average (ARMA) models to estimate the spectral envelope, which further enhanced the preservation of vocal timbre and speaker identity. The resulting system achieved high-quality, real-time synthesis and modification, making it well-suited for assistive technologies such as voice prostheses for laryngectomy patients.

Chapter V discussed in detail the design of speech modification algorithms for both conventional QHM-based methods and neural vocoder-based methods such as QHM-GAN and QHARMA-GAN. Conventional methods relied on a two-step process involving parameter estimation and post-modification. However, they required auxiliary models (e.g., DAP) to estimate spectral envelopes, and their limited accuracy can impair modification quality. In contrast, the neural methods were designed to directly infer the parameters of modified speech from modified inputs, en-

abling real-time processing. Owing to their data-driven nature, these models exhibited superior generalization and synthesis quality in both pitch and time-scale transformations. Consequently, the proposed neural vocoders provided a practical framework for high-quality, real-time voice transformation applications.

In conclusion, this thesis demonstrated that the combination of conventional signal processing methods and neural networks not only bridges the interpretability gap in current neural vocoders but also unlocked new possibilities for real-time, high-quality speech modeling and transformation. The proposed hybrid architecture paved the way for further research in interpretable and controllable neural speech synthesis, with promising applications in assistive technology, expressive speech synthesis, and voice transformation.

6.2 Future Perspective

Although this thesis has achieved substantial progress, there remain numerous opportunities to further advance the neural vocoder framework. Beyond methodological improvements, the approaches developed here hold significant potential for real-world applications that can benefit human life, such as expressive speech synthesis, assistive technologies, and personalized voice restoration. The details are as follows:

More Accurate Voiced/Unvoiced Detection: As previously mentioned, the modification mechanism of QHARMA-GAN is fundamentally based on the detection of Voiced/Unvoiced (V/UV) segments in speech. This is because unvoiced segments, which primarily consist of stochastic, noise-like components, lack the structured harmonicity required by quasi-harmonic models. Due to the sparse nature of the harmonic representation, these unvoiced segments are particularly difficult to model accurately, and are therefore excluded from the modification process. Only the voiced segments, which exhibit quasi-periodic structures, are considered for modification.

However, the performance of QHARMA-GAN is highly dependent on the accuracy of the V/UV detection. Existing methods often produce classification errors that degrade the quality of the speech modification. In particular, when originally voiced segments are misclassified as unvoiced, the system fails to apply the intended modification, resulting in unnatural discontinuities or insufficient transformation. Conversely, when unvoiced segments are incorrectly classified as voiced, they are over-modified by the harmonic model, introducing artificial periodicity into inherently aperiodic segments, which negatively affects the naturalness of the unvoiced sounds.

To address these limitations, it is essential to develop a more accurate V/UV detection method. An improved classification system would enhance the reliability of the modification process by ensuring that only truly voiced segments are modified, while unvoiced segments are appropriately excluded. This, in turn, would significantly improve the perceptual quality and consistency of the speech modified by QHARMA-GAN.

Text-to-Speech Application: Although the proposed vocoder is not specifically designed for end-to-end text-to-speech (TTS) synthesis, its design philosophy aligns with that of conventional vocoders, such as WORLD, which emphasize the interpretability and controllability of speech modeling. The focus of this work has been on analyzing and generating speech signals through the modeling of f_0 and spectral resonance structure. Nevertheless, the method also holds great potential for downstream applications, particularly in TTS systems.

Incorporating this vocoder into a complete TTS pipeline would require an additional f_0 prediction module. While this introduces some implementation complexity compared to typical mel-spectrogram-based approaches, previous research [103] has demonstrated that such integration is feasible. Notably, the advantage of using an explicit f_0 and spectral parameter representation lies in its ability to support controllable and expressive speech synthesis. Compared with non-trainable vocoders like WORLD, the proposed method also benefits from trainable components, allowing fine-tuning for different tasks. In particular, the proposed vocoder is expected to be effective for emotional TTS and prosody manipulation, where control over pitch and timbre is crucial.

Future work will explore the integration of this vocoder into full TTS systems, particularly in the context of emotional speech synthesis and prosodic editing, where its controllability can be fully exploited. This line of research will further validate the practical applicability of the proposed vocoder beyond analysis-by-synthesis evaluation.

Voice Conversion Application: The proposed method utilizes autoregressive moving average (ARMA) modeling to extract the spectral envelope of speech signals, which inherently captures speaker-specific acoustic characteristics. This capability provides a robust and promising foundation for voice conversion tasks, where the goal is to transform one speaker's voice to sound like another while preserving linguistic content. By manipulating the ARMA coefficients, the method enables controlled modification of the spectral envelope, facilitating speaker conversion with fine-grained control over prosodic attributes such as pitch and intonation.

This approach offers significant advantages over conventional voice conversion techniques based on mel-spectrogram representations. In particular, the explicit parametric form of the ARMA spectral envelope affords greater interpretability and precision in modifying speaker identity and timbre. Moreover, it allows for continuous and direct control of pitch through parameter transformation, thereby enhancing the flexibility and naturalness of the converted speech.

Future research directions will focus on exploring the full potential of the ARMA-based spectral envelope representation to achieve high-quality and pitch-controllable voice conversion. This includes developing end-to-end frameworks that integrate ARMA parameter manipulation with neural vocoder synthesis, aiming to extend the applicability of the proposed vocoder architecture beyond conventional speech synthesis toward sophisticated speaker transformation and voice conversion applications.

Resonance Characteristics Mapping: The proposed method utilizes an ARMA model composed of multiple cascaded smaller ARMA components. Each of these smaller ARMA units can be interpreted as corresponding to distinct physical parts of the vocal tract; for example, one ARMA component may represent the tongue, another the pharynx, and others the teeth or oral cavity structures. This decomposition enables the ARMA coefficients to reflect physical characteristics of the speaker's vocal tract in a more interpretable and localized manner.

By selectively modifying the coefficients of specific ARMA components, it becomes possible to realize fine-grained speech modifications that correspond to changes in particular vocal tract parts. For instance, variations such as a larger or smaller tongue can be modeled by adjusting the ARMA parameters linked to the tongue component, allowing controllable modification of speech timbre or articulation.

Moreover, this interpretable mapping between ARMA coefficients and physical vocal tract structures also holds potential for diagnostic applications, such as identifying anatomical differences like macroglossia (enlarged tongue) directly from speech signals.

Establishing and refining the mapping between ARMA components and vocal tract anatomy will likely require data-driven approaches, such as neural networks or other machine learning methods, to learn the complex relationships involved. Future work will focus on developing these mappings and applying them to achieve physically interpretable and controllable speech modification.

Hybrid System: The method proposed in this thesis constitutes a hybrid framework that integrates conventional algorithms with neural network-based approaches. This framework leverages the advantages of both paradigms while mitigating their respective limitations. Many contemporary studies prioritize the use of neural networks at the expense of system interpretability, often overlooking underlying physical principles. In contrast, conventional algorithms, although sometimes inferior in performance to neural networks, typically possess clear physical meaning, which facilitates a deeper understanding of the system and enables more controlled manipulation and modification.

Consequently, the proposed hybrid framework is not only applicable to the speech analysis, modification, and synthesis tasks addressed in this work but also has potential extensions to other fields, such as image detection and generation. By combining the interpretability of traditional methods with the expressive power of neural networks, this framework supports a more comprehensive understanding of algorithmic mechanisms, human physiology, and real-world phenomena. In this sense, the hybrid approach offers new insights and opens avenues for further research across diverse scientific and engineering fields.

Other Applications: Beyond technical improvements, quasi-harmonic vocoders hold substantial potential for real-world applications that directly impact human life. For example, individuals

undergoing laryngectomy due to throat cancer lose their natural voice. By recording their speech prior to surgery, quasi-harmonic vocoders combined with TTS systems can restore their voice postoperatively, enabling patients to communicate with their original vocal characteristics. The controllable and interpretable nature of these vocoders ensures that the reconstructed voice preserves speaker identity, intonation, and expressiveness, which are critical for social interactions and psychological well-being.

Moreover, the proposed framework employs a decomposition of the ARMA-based spectral envelope into multiple smaller ARMA components, each of which can be associated with distinct vocal or phonetic factors. In principle, these components could correspond to individual phonemes or sub-phonemic features, making it possible to generate speech at a finer granularity. Such a design enables future systems to synthesize speech from text or phonetic inputs more flexibly, where each letter or phoneme is mapped to specific ARMA parameters. This could significantly improve both the precision and naturalness of synthetic speech, particularly in applications requiring personalized or adaptive voice synthesis.

Taken together, these capabilities suggest that quasi-harmonic vocoders not only advance speech modeling technology but also offer practical benefits to everyday life. They can enhance accessibility for individuals with speech impairments, support personalized virtual assistants, improve expressive TTS for education and entertainment, and provide tools for creative voice transformation in music and media. Future research should continue to develop both algorithmic improvements and human-centered applications, emphasizing controllable, interpretable, and high-fidelity speech synthesis for societal benefit.

Acknowledgement

Completing this thesis marks not only the end of a long and challenging academic journey, but also a moment of reflection and gratitude for all those who have supported, guided, and accompanied me throughout these years. I am deeply indebted to many people who made this achievement possible.

I would first like to express my most sincere gratitude to my supervisor, Professor Tomoki Toda. From the very beginning of my doctoral studies, he welcomed me into the lab with kindness and humility, and gave me the precious opportunity to explore the rich and complex field of speech signal processing, especially speech modeling, which I am fully interested in. His academic rigor, careful thinking, and deeply human approach to research have been a lasting influence on me. More than once, when I faced conceptual confusion or setbacks in implementation, it was his patient encouragement and methodical guidance that helped me see a path forward. His words, encouraging bold hypotheses and cautious validation, have become my personal research philosophy. I am especially grateful for the freedom he gave me to explore ideas independently, while always being there to offer clarity and direction when I needed it most. Working under his mentorship not only helped me grow as a researcher but also as a person.

I am also sincerely thankful to the members of my thesis committee: Professor Zhenhua Ling and Professor Kensaku Mori. Their time, attention, and constructive feedback during my defense were invaluable. Professor Ling's deep expertise and sharp insights provided critical perspectives that pushed me to refine both the technical and theoretical aspects of this work. Professor Mori offered unique interdisciplinary viewpoints that encouraged me to consider broader implications.

I would also like to express my sincere gratitude to Professor Keisuke Fujii, whose thoughtful questions reminded me of the importance of clarity and interpretability in scientific work. I am equally grateful to Assistant Professor Huang Wen-Chin, who offered practical advice and insightful guidance on my presentations throughout my time in the lab.

I am also deeply grateful to Ms. Nami Noro and Ms. Mayuko Hayashi, whose dedicated administrative support made my life in Japan smooth and worry-free. Whether it was visa paperwork, conference reimbursements, or simply navigating daily life in a new country, they were always ready to help with warmth and efficiency. Their kindness, especially during my early months here, gave me peace of mind and allowed me to focus on my research without distraction.

My deepest personal gratitude goes to my wife and my parents, who have been my strongest and most enduring source of support. To my wife: your patience, your strength, and your unwavering belief in me, even when I doubted myself, carried me through the hardest moments of this journey. Being apart for long periods was not easy, and I know it required great emotional strength on your part to endure the distance and sacrifice your own comfort to support my goals. This thesis is just as much yours as it is mine. To my parents: your love has always been quiet, steadfast, and unconditional. From my earliest days, you nurtured my curiosity, encouraged me to dream, and supported me at every turning point. Though thousands of kilometers separated us, your support was always felt. I hope this work makes you proud.

I also wish to express my heartfelt appreciation to my master's advisor, Professor Shibin Wang, who played a formative role in my academic life. It was under his guidance that I first encountered the field of signal processing and was encouraged to pursue research. His mentorship gave me the foundation upon which this thesis is built.

To all my colleagues in the Sound Information Processing Laboratory, thank you for your collaboration, insights, and friendship. The atmosphere of the lab, serious when needed, but full of mutual respect and humor, was an ideal environment in which to grow. In particular, I would like to thank my fellow Chinese colleagues and dear friends. Whether we were troubleshooting late at night, debating ideas on papers, or just sharing meals and stories after long days, your companionship made this journey meaningful. I look forward to many more shared memories on the road, on two wheels, or over the dinner table.

To all the people who offered their help, whether through academic collaboration, practical support, or emotional strength, I extend my sincerest thanks. This dissertation is not just the result of solitary effort, but a reflection of all the kindness, wisdom, and generosity I have received along the way.

As this chapter of my life closes, I move forward with deep appreciation and renewed commitment, inspired by those who have helped me reach this milestone.

References

[1] Patrick J. Lynch. Mouth anatomy [medical illustration]. Licensed under CC BY 2.5, 2005.

[2] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, 99(7):1877–1884, 2016.

[3] George Kafentzis. *Adaptive sinusoidal models for speech with applications in speech modifications and audio analysis*. PhD thesis, Université de Rennes; Panepistīímio Krīítīs, 2014.

[4] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Proc. NeurIPS*, 33:17022–17033, 2020.

[5] Siuzdak Hubert. Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis. In *Proc. ICLR*, 2024.

[6] Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. UnivNet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation. In *Proc. Interspeech*, pages 2207–2211, 2021.

[7] Xin Wang, Shinji Takaki, and Junichi Yamagishi. Neural source-filter waveform models for statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:402–415, 2019.

[8] Homer Dudley. Remaking speech. *The Journal of the Acoustical Society of America*, 11(2):169–177, 1939.

[9] Gunnar Fant. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*. Walter de Gruyter, 2 edition, 1971.

[10] Gunnar Fant, Johan Liljencrants, Qi-guang Lin, et al. A four-parameter model of glottal flow. *STL-QPSR*, 4(1985):1–13, 1985.

[11] Gunnar Fant. The LF-model revisited. transformations and frequency domain analysis. *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, 2(3):40, 1995.

[12] Christer Gobl. Reshaping the transformed LF Model: Generating the glottal source from the waveshape parameter rd. In *Proc. Interspeech*, 2017.

[13] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain 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*, 27(3–4):187–207, 1999.

[14] Hideki Kawahara, Jo Estill, and Osamu Fujimura. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. In *Proc. MAVEBA*, pages 59–64. Firenze, 2001.

[15] Masanori Morise, Hideki Kawahara, and Haruhiro Katayose. Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech. In *Proc. ICAG*, 2009.

[16] Masanori Morise. Cheaptrick, a spectral envelope estimator for high-quality speech synthesis. *Speech Communication*, 67:1–7, 2015.

[17] Masanori Morise. Platinum: A method to extract excitation signals for voice synthesis system. *Acoustical Science and Technology*, 33(2):123–125, 2012.

[18] Masanori Morise et al. Harvest: A high-performance fundamental frequency estimator from speech signals. In *Proc. Interspeech*, pages 2321–2325, 2017.

[19] Malah David. Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):121–133, 1979.

[20] Eric Moulines and Francis Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, 9(5-6):453–467, 1990.

[21] Thomas Oberlin, Sylvain Meignen, and Valérie Perrier. The Fourier-based synchrosqueezing transform. In *Proc. IEEE ICASSP*, pages 315–319, 2014.

[22] Boualem Boashash and Peter Black. An efficient real-time implementation of the Wigner-Ville distribution. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(11):1611–1618, 1987.

[23] François Auger and Patrick Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Transactions on Signal Processing*, 43(5):1068–1089, 1995.

[24] François Auger, Patrick Flandrin, Yu-Ting Lin, Stephen McLaughlin, Sylvain Meignen, Thomas Oberlin, and Hau-Tieng Wu. Time-frequency reassignment and synchrosqueezing: An overview. *IEEE Signal Processing Magazine*, 30(6):32–41, 2013.

[25] Ingrid Daubechies, Jianfeng Lu, and Hau-Tieng Wu. Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool. *Applied and computational harmonic analysis*, 30(2):243–261, 2011.

[26] Gaurav Thakur, Eugene Brevdo, Neven S Fučkar, and Hau-Tieng Wu. The synchrosqueezing algorithm for time-varying spectral analysis: Robustness properties and new paleoclimate applications. *Signal Processing*, 93(5):1079–1094, 2013.

[27] Steve Mann and Simon Haykin. The chirplet transform: A generalization of Gabor's logon transform. In *Proc. Vision Interface*, volume 91, pages 205–212. Citeseer Princeton, NJ, USA, 1991.

[28] Shibin Wang, Xuefeng Chen, Gaigai Cai, Binqiang Chen, Xiang Li, and Zhengjia He. Matching demodulation transform and synchrosqueezing in time-frequency analysis. *IEEE Transactions on Signal Processing*, 62(1):69–84, 2014.

[29] T Quatieri and Rl McAulay. Speech transformations based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(6):1449–1464, 1986.

[30] Yannis Stylianou. Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification. *Ph.D. dissertation, E.N.S.T- Paris*, 1996.

[31] Yannis Pantazis and Yannis Stylianou. Improving the modeling of the noise part in the harmonic plus noise model of speech. In *Proc. IEEE ICASSP*, pages 4609–4612, 2008.

[32] Jean Laroche, Yannis Stylianou, and Eric Moulines. HNS: Speech modification based on a harmonic + noise model. In *Proc. IEEE ICASSP*, volume 2, pages 550–553, 1993.

[33] Gilles Degottex and Yannis Stylianou. Analysis and synthesis of speech using an adaptive full-band harmonic model. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2085–2095, 2013.

[34] Yannis Pantazis, Olivier Rosec, and Yannis Stylianou. On the properties of a time-varying quasi-harmonic model of speech. In *Proc. Interspeech*, 2008.

[35] Yannis Pantazis, Olivier Rosec, and Yannis Stylianou. Adaptive AM-FM signal decomposition with application to speech analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(2):290–300, 2010.

[36] George P Kafentzis, Yannis Pantazis, Olivier Rosec, and Yannis Stylianou. An extension of the adaptive quasi-harmonic model. In *Proc. IEEE ICASSP*, pages 4605–4608, 2012.

[37] George P Kafentzis, Olivier Rosec, and Yannis Stylianou. Robust full-band adaptive sinusoidal analysis and synthesis of speech. In *Proc. IEEE ICASSP*, pages 6260–6264, 2014.

[38] George P Kafentzis and Yannis Stylianou. High-resolution sinusoidal modeling of unvoiced speech. In *Proc. IEEE ICASSP*, pages 4985–4989, 2016.

[39] George P Kafentzis, Gilles Degottex, Olivier Rosec, and Yannis Stylianou. Time-scale modifications based on a full-band adaptive harmonic model. In *Proc. IEEE ICASSP*, pages 8193–8197, 2013.

[40] George P Kafentzis, Gilles Degottex, Olivier Rosec, and Yannis Stylianou. Pitch modifications of speech based on an adaptive harmonic model. In *Proc. IEEE ICASSP*, pages 7924–7928, 2014.

[41] Thomas Oberlin, Sylvain Meignen, and Valérie Perrier. Second-order synchrosqueezing transform or invertible reassignment? towards ideal time-frequency representations. *IEEE Transactions on Signal Processing*, 63(5):1335–1344, 2015.

[42] Duong-Hung Pham and Sylvain Meignen. High-order synchrosqueezing transform for multicomponent signals analysis-with an application to gravitational-wave signal. *IEEE Transactions on Signal Processing*, 65(12):3168–3178, 2017.

[43] Gang Yu, Mingjin Yu, and Chuanyan Xu. Synchroextracting transform. *IEEE Transactions on Industrial Electronics*, 64(10):8042–8054, 2017.

[44] Yannis Stylianou, Jean Laroche, and Eric Moulines. High-quality speech modification based on a harmonic+ noise model. In *Proc. Eurospeech 1995*, pages 451–454, 1995.

[45] Yannis Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on speech and audio processing*, 9(1):21–29, 2002.

[46] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *Proc. SSW 2016*, pages 125–125, 2016.

[47] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *Proc. ICML*, pages 3918–3926, 2018.

[48] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *Proc. IEEE ICASSP*, pages 3617–3621, 2019.

[49] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *Proc. ICML*, pages 2410–2419, 2018.

[50] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Proc. NeurIPS*, 27, 2014.

[51] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *Proc. IEEE ICASSP*, pages 6199–6203, 2020.

[52] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. MelGAN: Generative adversarial networks for conditional waveform synthesis. *Proc. NeurIPS*, 32, 2019.

[53] Ahmed Mustafa, Nicola Pia, and Guillaume Fuchs. StylemelGAN: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization. In *Proc. IEEE ICASSP*, pages 6034–6038, 2021.

[54] Markel John D. and Gray Jr. Augustine H. Linear prediction of speech. *Springer*, 1976.

[55] Schroeder Manfred R. and Atal Bishnu S. Code-excited linear prediction (celp): High-quality speech at very low bit rates. In *Proc. ICASSP*, volume 10, pages 937–940. IEEE, 1985.

[56] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[57] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[58] Paul Mermelstein. Distance measures for speech recognition, psychological and instrumental. *Pattern recognition and artificial intelligence*, 116:374–388, 1976.

[59] Douglas A Reynolds. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, 1995.

[60] Keiichi Tokuda, Heiga Zen, and Alan W Black. Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, 101(5):1234–1252, 2013.

[61] Gilles Quiniou and Jean-Noël Fouquet. Sinusoidal modeling of voiced speech using time-varying harmonic amplitudes and phases. In *Proc. Interspeech*, 2002.

[62] Tomoki Toda, Alan W Black, and Keiichi Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2222–2235, 2007.

[63] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. In *Proc. Interspeech*, pages 4006–4010, 2017.

[64] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018.

[65] Takuhiro Kaneko and Hirokazu Kameoka. CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks. In *Proc. EUSIPCO*, pages 2100–2104. IEEE, 2018.

[66] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *Proc. ICML*, pages 5210–5219, 2019.

[67] Srihari Kankanahalli. End-to-end optimized speech coding with deep neural networks. In *ICASSP*, pages 6799–6803, 2020.

[68] David Snyder et al. X-vectors: Robust dnn embeddings for speaker recognition. In *ICASSP*, pages 5329–5333, 2018.

[69] Hideki Kawahara. STRAIGHT, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology*, 27(6):349–353, 2006.

[70] Robert McAulay and Thomas Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):744–754, 1986.

[71] Marian Kepesi and Luis Weruaga. Adaptive chirp-based time–frequency analysis of speech signals. *Speech communication*, 48(5):474–492, 2006.

[72] Luis Weruaga and Márian Képesi. The Fan-chirp transform for non-stationary harmonic signals. *Signal Processing*, 87(6):1504–1522, 2007.

[73] Yannis Pantazis, Olivier Rosec, and Yannis Stylianou. Iterative estimation of sinusoidal signal parameters. *IEEE Signal Processing Letters*, 17(5):461–464, 2010.

[74] Stephen A Zahorian and Hongbing Hu. A spectral/temporal method for robust fundamental frequency tracking. *The Journal of the Acoustical Society of America*, 123(6):4559–4571, 2008.

[75] Takuhiro Kaneko, Kou Tanaka, Hirokazu Kameoka, and Shogo Seki. iSTFTNet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time Fourier transform. In *ICASSP*, 2022.

[76] Keisuke Matsubara, Takuma Okamoto, Ryoichi Takashima, Tetsuya Takiguchi, Tomoki Toda, and Hisashi Kawai. Harmonic-net: Fundamental frequency and speech rate controllable fast neural vocoder. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1902–1915, 2023.

[77] Shimizu Souta, Okamoto Takuma, Takashima Ryouichi, Takiguchi Tetsuya, Toda Tomoki, and Kawai Tsune. Initial investigation of fundamental frequency controllable HiFi-GAN conditioned on mel-spectrogram. In *Proc. Acoustical Society of Japan*, pages 1137–1140, 2022.

[78] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proc. IEEE/CVF CVPR*, pages 11966–11976, 2022.

[79] Shaowen Chen, Shibin Wang, Botao An, Ruqiang Yan, and Xuefeng Chen. Instantaneous frequency band and synchrosqueezing in time-frequency analysis. *IEEE Transactions on Signal Processing*, 71:539–554, 2023.

[80] Joseph Turian and Max Henry. I'm sorry for your loss: Spectrally-based audio distances are bad at pitch. In *Proc. NeurIPS*, 2020.

[81] Ben Hayes, Charalampos Saitis, and György Fazekas. Sinusoidal frequency estimation by gradient descent. In *Proc. IEEE ICASSP*, pages 1–5, 2023.

[82] Simon Schwär and Meinard Müller. Multi-scale spectral loss revisited. *IEEE Signal Processing Letters*, 30:1712–1716, 2023.

[83] Bernardo Torres, Geoffroy Peeters, and Gaël Richard. Unsupervised harmonic parameter estimation using differentiable dsp and spectral optimal transport. In *Proc. IEEE ICASSP*, pages 1176–1180, 2024.

[84] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[85] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. ICLR*, 2015.

[86] Gang Yu, Zhonghua Wang, and Ping Zhao. Multisynchrosqueezing transform. *IEEE Transactions on Industrial Electronics*, 66(7):5441–5455, 2018.

[87] Shaowen Chen and Tomoki Toda. Sequence-wise optimization for quasi-harmonic speech waveform modeling. In *Proc. APSIPA ASC*, pages 1661–1666, 2022.

[88] Keith Ito and Linda Johnson. The LJ Speech dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[89] Heiga Zen, Viet Dang, Robert A. J. Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Z. Chen, and Yonghui Wu. LibriTTS: A corpus derived from librispeech for text-to-speech. In *Proc. Interspeech*, 2019.

[90] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *Proc. O-COCOSDA*, pages 1–5, 2017.

[91] Kavita Kasi and Stephen A. Zahorian. Yet another algorithm for pitch tracking. In *Proc. ICASSP*, volume 1, pages 361–364, 2002.

[92] John Kolen and Jordan Pollack. Back propagation is sensitive to initial conditions. *Proc. NeurIPS*, 3, 1990.

[93] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *Proc. IEEE ICASSP*, pages 4214–4217, 2010.

[94] Amro El-Jaroudi and John Makhoul. Discrete all-pole modeling. *IEEE Transactions on signal processing*, 39(2):411–423, 1991.

[95] Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. CSTR VCTK Corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, pages 271–350, 2019.

[96] Shinnosuke Takamichi, Kentaro Mitsui, Yuki Saito, Tomoki Koriyama, Naoko Tanji, and Hiroshi Saruwatari. JVS corpus: free japanese multi-speaker voice corpus. *arXiv preprint arXiv:1908.06248*, 2019.

[97] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: Utokyo-sarulab system for voicemos challenge 2022. In *Proc. Interspeech*, pages 4521–4525, 2022.

[98] Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, and Adam Roberts. DDSP: Differentiable digital signal processing. In *Proc. ICLR*, 2020.

[99] Reo Yoneyama, Atsushi Miyashita, Ryuichi Yamamoto, and Tomoki Toda. Wavehax: Aliasing-free neural waveform synthesis based on 2d convolution and harmonic prior for reliable complex spectrogram estimation. *arXiv preprint arXiv:2411.06807*, 2024.

[100] Huang Rongjie, Chen Feiyang, Ren Yi, Liu Jinglin, Cui Chenye, and Zhao Zhou. Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus. In *Proc. ACM MM*, pages 3945–3954, 2021.

[101] Gilles Degottex and Yannis Stylianou. A full-band adaptive harmonic representation of speech. In *Interspeech*, pages 382–385, 2012.

[102] Riccardo Di Federico. Waveform preserving time stretching and pitch shifting for sinusoidal models of sound. In *Proc. DAFx*, pages 44–48, 1998.

[103] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: 2000-speaker neural text-to-speech. In *Proc. ICLR*, volume 79, pages 1094–1099, 2018.

List of Publications

Journal Papers

[1] **S. Chen**, T. Toda, "QHARMA-GAN: quasi-harmonic neural vocoder based on autoregressive moving average model," IEEE Transactions on Audio, Speech and Language Processing, vol. 33, pp. 3703–3719, 2025.

[2] **S. Chen**, T. Toda, "Sequence-wise speech waveform modeling via gradient descent optimization of quasi-harmonic parameters," IEEE Transactions on Audio, Speech and Language Processing, vol. 33, pp. 319–332, 2025.

[3] **S. Chen**, S. Wang, B. An, R. Yan, and X. Chen, "Instantaneous Frequency Band and Synchrosqueezing in Time-Frequency Analysis," IEEE Transactions on Signal Processing, vol. 71, pp. 539–554, 2023.

[4] B. An, Z. Zhao, S. Wang, **S. Chen**, X. Chen. "Sparsity-assisted bearing fault diagnosis using multiscale period group lasso," ISA transactions, vol. 98, pp. 338–348, 2020.

International Conference Papers

[1] **S. Chen** and T. Toda, "Sequence-wise optimization for quasi-harmonic speech waveform modeling," in Proc. APSIPA ASC, pp. 1661–1666, 2022.

[2] **S. Chen** and T. Toda, "QHM-GAN: Neural vocoder based on quasiharmonic modeling," in Proc. Interspeech, pp. 3889–3893, 2024.

Domestic Conference Papers

[1] **S. Chen**, T. Toda, "Sequence-wise parameter extraction of quasi-harmonic model for speech waveform generation,"日本音響学会春季研究発表会講演論文集 , 1-8-7, pp. 1129-1130, 2022.

Technical Report

[1] **S. Chen**, "QHM-GAN: Neural vocoder based on quasi-harmonic modeling," 2024 APSIPA China-Japan Joint Symposium on Speech and Language Processing, 2024.