# Sequence-wise Optimization for Quasi-Harmonic Speech Waveform Modeling

Shaowen Chen* and Tomoki Toda†
* Nagoya University, Nagoya, Japan
E-mail: shaowen.chen@g.sp.m.is.nagoya-u.ac.jp
† Nagoya University, Nagoya, Japan
E-mail: tomoki@icts.nagoya-u.ac.jp

***Abstract*—Quasi-harmonic models (QHMs) are effective methods for representing a speech waveform with frame-wise parameters and flexibly resynthesizing a speech waveform from them. The original QHM methods analytically extract those parameters by directly minimizing an error between resynthesized and original speech waveform segments frame by frame. However, such a frame-wise parameter extraction process suffers from information loss between individual frames, causing the quality degradation of a resynthesized speech waveform. In this paper, we propose a sequence-wise parameter optimization method based on back propagation (BP) by directly minimizing the reconstruction error of a whole speech waveform. The proposed method is capable of specifically compensating for the missing information between frames by making the parameter extraction process and resynthesis process consistent. We investigate the effectiveness of the proposed method by conducting experimental evaluations using real speech utterances. The experimental results demonstrate that the proposed method achieves a great improvement of the speech resynthesis quality, i.e., from 11.7 dB to 36.2 dB of the signal-to-reconstruction error ratio and from 0.83 to 0.99 of short-time objective intelligibility.**

## I. Introduction

Sinusoidal modeling has been intensively studied, which represents a speech waveform as a sum of harmonic components [1], with corresponding amplitudes and frequencies parameters. Sinusoidal modeling compresses the speech waveform to the frame-wise parameters [2], possibly accelerating the generation speed of neural vocoder and easily modifying synthesized speech by changing the amplitudes and frequency parameters of harmonic components. On the other hand, it is well known that the speech waveform consists of voiced speech and unvoiced speech, where the voiced speech is a quasi-periodic signal while the unvoiced speech is a stochastic signal. Therefore, simple sinusoidal modeling is hard to accurately represent the speech waveform.

As one of the sophisticated speech waveform modeling methods based on the sinusoidal model, the harmonic plus noise model (HNM) [3] was proposed to represent the voiced speech using harmonic components corresponding to an initially given fundamental frequency ($f_0$) and the unvoiced speech using noise components. HNM is capable of representing a speech waveform with frame-wise parameters, such as $f_0$, complex amplitudes, and noise spectral envelope, making it possible to achieve high-quality and flexible speech modification [4]. Nevertheless, HNM is sensitive to $f_0$, i.e., accurate $f_0$ estimation is essential to resynthesize the speech waveform with high quality. Thus, HNM needs to refine the given $f_0$ to achieve reasonably good resynthesis performance [5]. Even so, using fully harmonic components to model a speech waveform, which is not a fully periodic signal, remains a considerable resynthesis error.

To address this issue, the quasi-harmonic model (QHM) [6] is proposed to use quasi-harmonic components to represent the voiced speech and unvoiced speech simultaneously with a frequency correction mechanism, allowing to use not only $f_0$ but also individual frequencies of components as free parameters. The individual frequency of each component can be updated after extracting the frame-wise complex amplitude parameter to get closer to the true value. Furthermore, the adaptive QHM (aQHM) [7] is proposed to model the amplitude- and frequency-modulated (AM-FM) signals more accurately by replacing the original exponential part with a non-stationary phase function, making the model adaptable to the real time-varying phase in each frame. In most cases, the amplitude modulation is usually not linear but cubical or even higher order. Therefore, the extension of aQHM (eaQHM) [8] is proposed and makes the model more matchable to the nonlinear amplitude modulation. In this way, the amplitude and frequency parameters can be accurately extracted in most cases, with which the speech can be well resynthesized [9], [10] and modified [11], [12].

Although these QHM-based methods are helpful for making parameter extraction and speech waveform resynthesis processes increasingly accurate, there remains a problem to be addressed. They estimate the parameters by minimizing the error between the resynthesized and original speech waveform segments frame by frame. Such a frame-wise parameter extraction process inevitably results in the loss of the information between individual frames, in turn causing the reconstructed speech waveform to deviate from the original speech waveform, particularly in the area between frames. As a consequence, the quality of the reconstructed speech is severely deteriorated, especially when the frame-shift is large.

In this paper, we propose an optimization method to improve the extracted parameters, i.e., the frequency and complex amplitude of each quasi-harmonic component, to improve the quality of synthetic speech. In this optimization, we apply the back propagation (BP) method [13], [14] to propagate the

978-616-590-477-3 ©2022 APSIPA

reconstruction error and update the parameters by the inverse of the synthesis process for minimizing the reconstruction error over a whole speech waveform instead of frame-by-frame speech, compensating for the lost information between frames. Results of experimental evaluations show that our proposed method significantly improves the quality of resynthesized speech.

The outline of this paper is organized as follows. In Section II, we provide a review of QHM methods and discuss their limitations. In Section III, the proposed method is presented. In Section IV, quantitative performance comparison between QHM methods and our proposed method is explored by analyzing the speech segments. The conclusion is given in Section V.

## II. Review of QHM Methods

The sinusoidal models work based on the assumption that each frame of signal can be represented as the combination of harmonic components:

$$x(t) = \left( \sum_{k=1}^{K} a_k e^{j2\pi f_k t} \right) w(t),\ t \in [-T_l, T_l] \tag{1}$$

where $a_k$ and $f_k$ are the complex amplitude and the frequency of the $k$-th component. $w(\cdot)$ denotes the moving window whose length is $2T_l$. These models extract the complex amplitude of the speech using the frequencies obtained by an initial fundamental frequency $f_0$ (or also called the pitch), i.e., $f_k = kf_0$. Unfortunately, most pitch detectors work under the assumption that the frequency of the windowed speech is constant, which is inconsistent with the fact that the speech is non-stationary, resulting in the deviations between the estimated pitch or the frequencies of each harmonic component and their true ones. This in turn causes the complex amplitude parameters not to be extracted accurately and the resynthesized speech to be substantially different from the original speech. To address this problem, QHM methods were proposed with a frequency correction mechanism to more precisely extract the complex amplitudes.

### A. QHM, aQHM and eaQHM

QHM matches the target signal by the following frame-wise model:

$$x(t) = \left[ \sum_{k=1}^{K} (a_k + tb_k)\, e^{j2\pi \hat{f}_k t} \right] w(t),\ t \in [-T_l, T_l], \tag{2}$$

where $\hat{f}_k$ and $b_k$ denote the initially estimated frequency and the complex slope of the $k$-th component. QHM utilizes $b_k$ to make the model capable of adaptively matching signal, compensating for errors between original speech and synthetic speech with inaccurately estimated frequency [6]. First, $a_k$ and $b_k$ are obtained by solving Eq. 2 via Least Squares (LS). Then, the more accurate frequency can be obtained as follows:

$$f_k = \hat{f}_k + \eta_k = \hat{f}_k + \frac{1}{2\pi} \frac{a_k^R b_k^I - a_k^I b_k^R}{|a_k|^2} \tag{3}$$

where $a_k^R, a_k^I$ and $b_k^R, b_k^I$ respectively denote the real and imaginary parts of $a_k, b_k$. It is well known that speech signals are non-stationary signals (AM-FM signals), whose amplitude and frequency vary over time. QHM considers the frequency of each harmonic component as constant, resulting in that the updated frequency still mismatches the real one in each frame.

To better match the modulated frequency of the speech, aQHM rewrites the exponential part of the model with non-stationary phases as follows:

$$x(t) = \left\{ \sum_{k=1}^{K} (a_k + tb_k)\, e^{j[\hat{\varphi}_k(t+t_l) - \hat{\varphi}_k(t_l)]} \right\} w(t),\quad t \in [-T_l, T_l] \tag{4}$$

where $\hat{\varphi}_k(t)$ denotes the phase function of the $k$-th harmonic component and $t_l$ is the center of the moving window. The aQHM first uses QHM to correct the frequency and integrate it to get the phase function by

$$\varphi_k(t) = \int_{t_l}^{t_l+t} 2\pi f_k(u) \mathrm{d}u,\ t \in [-T_l, T_l]. \tag{5}$$

Secondly, this phase function subtract the old phase value at $t_l$ (namely, $\hat{\varphi}_k(t_l)$) and the result will be the exponential part of aQHM. Note that the phase is no longer obtained by a constant frequency but computed by a time-varying frequency. That's why aQHM is capable of better adapting to the time-varying signals. Thirdly, the new $a_k$ and $b_k$ can be computed again by solving Eq. 4 and the frequency will be updated again to further approach the real one. The aQHM carries out these 3 steps iteratively until the generated speech is no longer getting closer to the original speech. Eventually, the more accurate complex amplitude parameter and frequency of each harmonic component can be acquired, with which the speech with good quality can be resynthesized.

In most cases, the amplitude of speech varies nonlinearly instead of linearly, QHM and aQHM cannot adapt the amplitude well. To solve this problem, an extension of aQHM (eaQHM) is proposed and suggest the model to be build adding an amplitude amplifier at the amplitude part of the model, as follows

$$x(t) = \left\{ \sum_{k=1}^{K} (a_k + tb_k)\, \frac{A_k(t+t_l)}{A_k(t_l)} e^{j[\hat{\varphi}_k(t+t_l) - \hat{\varphi}_k(t_l)]} \right\} w(t),\quad t \in [-T_l, T_l] \tag{6}$$

where $A_k(t)$ denotes the amplitude of the $k$-th component. Compared to aQHM, eaQHM multiplies the aQHM by a function which is the ratio of the instantaneous amplitude of the entire frame to the amplitude at the center of the frame, allowing the model to better match the rapidly modulated amplitude of the speech. Likewise, the eaQHM uses the QHM's result as the initial parameter to further adapt them in each frame. Then, the $a_k$, $b_k$ and frequency $f_k$ will be updated iteratively. After the update, the frequencies of individual components don't strictly obey the multiplicative relationship anymore, i.e., these frequencies are not the integer multiples of

$f_0$, especially in the unvoiced part of the speech. That's why they are called quasi-harmonic models. Finally, more accurate parameters can be obtained which can reconstruct the speech with a better quality than QHM and aQHM.

### B. Limitations of QHM methods

Although QHM methods can achieve a good result in terms of parameter extraction and speech resynthesis, the quality of the synthetic speech is not good enough. When the frame-shift (i.e., the shift of the analysis window in time domain) becomes larger, the quality of the synthetic speech drops dramatically. This is mainly owing to that QHM methods estimate the amplitude parameters frame by frame, ignoring the information between individual frames. More specifically, QHM methods use their own models to approximate the windowed speech in each frame and obtain the best-estimated amplitudes and update the frequencies of all components only at the center of each frame via LS. Although the amplitude, frequency, and phase can achieve their own instantaneous version by interpolation for the reconstruction of the whole speech waveform, the values between the centers of the individual frames are not accurate enough, leading to the underperformed speech resynthesis.

Besides, although the QHM methods compensate for the frequency mismatch adaptively, their performances deteriorate when the initial $f_0$ differs greatly from the true value. That's because the unvoiced speech is stochastic so it is hard to be matched by the harmonic components with the irrelevant frequencies and, in turn, the frequencies are hard to be corrected. What's worse, in this case, sometimes the individual frequencies of components in the unvoiced part will deviate more and more from the ideal value after the update, leading to the poorer extraction of unvoiced speech.

## III. BP Optimization for QHM

The back propagation (BP) algorithm can update the parameters to be optimized backwards through the generation process from the perspective of the final desired results, so that updated parameters can ultimately generate more accurate results. Therefore, we are motivated to employ BP to improve waveform modeling with QHM methods, especially in terms of speech resynthesis. Specifically, we propose a BP-based method to jointly optimize all parameters from QHM methods over multiple frames sequence by sequence. The BP method uses the loss function to calculate the error between synthetic speech and target speech, then propagates it to calculate its gradients with respective to the parameters by the inverse of the synthesis process to update the parameters by the optimizer. The workflow of the proposed method is illustrated in Fig. 1, and the synthesis process and optimization process are detailedly demonstrated in the following part.

### A. Synthesis process

After QHM methods' analysis, the final $a_k$, $f_k$ of each quasi-harmonic component for speech resynthesis can be acquired, which can resynthesize the speech waveform. However, they

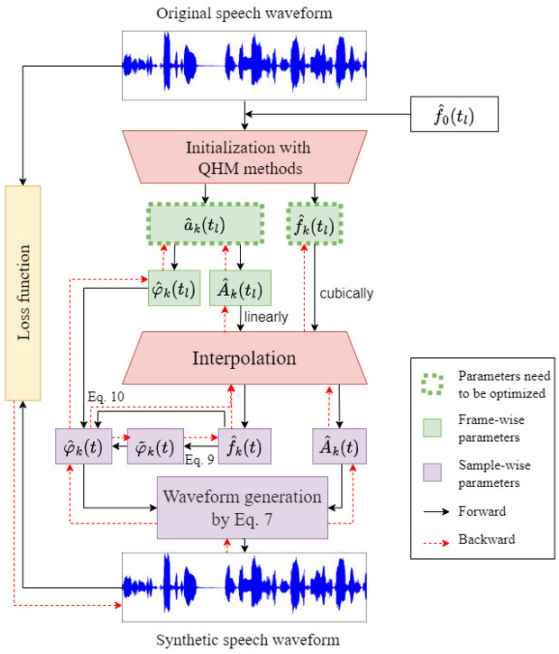


Fig. 1. The workflow of the proposed method.

are usually imperfect and need to be optimized. Inspired by [15], the speech can be resynthesized by summing all the quasi-harmonic components obtained by instantaneous amplitudes and phases, expressed as

$$\hat{x}(t) = \sum_{k=1}^{K} \hat{A}_k(t) e^{j\hat{\varphi}_k(t)}. \tag{7}$$

It is suggested that the instantaneous frequencies and amplitudes can be derived from the frame-wise values by interpolation. Therefore, we firstly need to get the frame-wise amplitude and phase by

$$\hat{A}_k(t_l) = |\hat{a}_k(t_l)|, \hat{\varphi}_k(t_l) = \angle \hat{a}_k(t_l), \tag{8}$$

where $\hat{a}_k(t_l)$ is initially given by QHM methods. For amplitude $\hat{A}_k(t)$, the values at the instants between two consecutive frames are recommended to be interpolated linearly. For instantaneous phases $\hat{\varphi}_k(t)$ between the consecutive frames, they can be computed by integrating frequency as

$$\tilde{\varphi}_k(t) = \hat{\varphi}_k(t_l) + \int_{t_l}^{t_l+t} 2\pi \hat{f}_k(u) \mathrm{d}u, \tag{9}$$

where the instantaneous frequency $\hat{f}_k(t)$ is suggested to be obtained from the frame-wise version by cubic spine interpolation and the integration result of the current frame is noted as $\tilde{\varphi}_k(t)$. Nevertheless, the frame-wise frequency and interpolated values are not strictly equal to the real ones, which is bound to cause the errors and discontinuity of the phase over the frame boundaries, i.e., it is hard to ensure $\tilde{\varphi}_k(t_{l+1}) = \hat{\varphi}_k(t_{l+1}) + 2\pi M$; where $M$ is the closest integer to $|\hat{\varphi}_k(t_{l+1}) - \tilde{\varphi}_k(t_{l+1})|/2\pi$ . To avoid this, we interpolate the

phase in the following special way:

$$\hat{\varphi}_k(t) = \hat{\varphi}_k(t_l) + \int_{t_l}^{t_l+t} 2\pi \hat{f}_k(u) + z \sin\left[\frac{\pi(u-t_{l-1})}{t_l - t_{l-1}}\right] \mathrm{d}u \tag{10}$$

where $z$ is computed by

$$z = \frac{\pi[\hat{\varphi}_k(t_{l+1}) + 2\pi M - \tilde{\varphi}_k(t_{l+1})]}{2(t_{l+1} - t_l)}.$$

Then we can get a smooth instantaneous phase. In this way, we can resynthesize the speech naturally by Eq. 7.

### B. Sequential optimization with BP

BP needs to measure the error between the generated speech and the target speech and propagate it backwards through the synthesis process to calculate the gradients of the frame-wise parameters ($a_k(t_l)$ and $f_k(t_l)$) to update them. Thus, it is essential to ensure that the error is differentiable with respect to those frame-wise parameters. From Eq. 7, we can know the synthetic speech $\hat{x}(t)$ is differentiable to the sample-wise parameters, $\hat{A}_k(t)$ and $\hat{\varphi}_k(t)$. And the interpolation is differentiable with respect to the frame-wise parameters, $\hat{A}_k(t_l)$ and $\hat{\varphi}_k(t_l)$, which are differentiable with respect to $\hat{a}_k(t)$ and $\hat{f}_k(t)$ according to Eq. 8. Thus, the synthesis process is differentiable and the gradients of the error with respect to the frame-wise parameters can be calculated by propagating the error backwards.

Aimed at improving the quality of synthetic speech, especially in terms of the waveform in the time domain, we use the loss function to measure the error between the whole speech sequences of synthesis and target (over all frames instead of the losses in individual frames), such as L1 loss or L2 loss. Then the parameters can be optimized, considering the information between frames.

BP needs initial input, i.e., the frame-wise amplitude parameter $a_k(t_l)$ and frequency $\hat{f}_k(t)$, and a proper initialization is essential. Although our experiments shows that random initial values of $\hat{a}_k(t)$ and $\hat{f}_k(t)$ can be also optimized to generate the speech with a good quality, the optimized $\hat{a}_k(t)$ and $\hat{f}_k(t)$ do not represent the structure of the speech. For instance, in the non-speech part, the $a_k(t)$ should be equal to 0, but the $a_k(t)$ optimized from random values are not equal to 0 and their summation is equal to 0. To avoid this unreasonable matter, regarding the frequency, we obtain the frequencies of individual components by $\hat{f}_k(t) = k\hat{f}_0(t)$. For the amplitude $\hat{a}_k(t)$, we can choose the results of QHM, aQHM, or eaQHM. In this paper, we choose QHM and eaQHM as the initialization way and note their corresponding optimizations as BP-QHM and BP-eaQHM.

## IV. Experiment Evaluation

In this section, we compare the performance of the proposed method with QHM and eaQHM to explore the superiority of the proposed method in different cases, including working with different frame-shifts, setting different harmonic numbers ($K$) and analyzing signals with different sampling rates ($f_s$).

TABLE I
AVERAGE SRER AND STOI SCORES.

| Method | QHM | eaQHM | BP-QHM | BP-eaQHM |
|---|---|---|---|---|
| SRER [dB] | 11.7 | 5.9 | 36.2 | 27.8 |
| STOI | 0.83 | 0.83 | 0.99 | 0.99 |

### A. Experimental conditions

To test the performance of the proposed method, we randomly drew 32 utterances from LJspeech [16] (resampled with 16 kHz), LibriTTS corpus [17] and AISHELL [18] as 16 kHz, 24 kHz, and 44.1 kHz samples respectively. And we analyzed these 32 utterances and averaged over all the results. We employed QHM and eaQHM to separately provide their frame-wise results as the initial input for BP. Regarding the settings of QHM methods, we provided the $f_0$ by a neural pitch detector named Crepe [19]. For BP's setting, we used a step decay learning rate scheme and set the initial value as 0.001 with the reduction by half after every 100 epochs (1000 epochs in total). And we chose the adaptive moment estimation as the optimizer. For a quantified comparison, we employed signal-to-reconstruction-error ratio (SRER) to measure the similarity between results and target, as

$$SRER = 20\log_{10}\frac{\sigma_{x(t)}}{\sigma_{x(t)-\hat{x}(t)}}$$

where $\sigma(\cdot)$ denotes the standard deviation and $\hat{x}(t)$ is the resynthesized speech. A larger SRER denotes a closer result to target. Besides, we used short-time objective intelligibility (STOI) [20] to measure the intelligibility of the synthetic speech, where a higher value means a better result.

### B. Results

Firstly, we compare the QHM methods with the proposed method by analyzing 16 kHz speech samples while setting the frame shit to 8 ms and setting the harmonic number to 128. All the speech samples will be analyzed and the quality of the synthetic results is measured by calculating the average of 32 SRER and STOI scores, which are listed in Table I. In Table I, we can find that the result of BP-QHM and BP-eaQHM are much better than QHM and eaQHM. And it can also be concluded that better initialization of BP yields better final results, as what [21] describes. Although eaQHM works better than QHM theoretically, QHM works better in this experiment. Note that the SRER of QHM is better than that of eaQHM while the STOIs of QHM and eaQHM are the same. It is possible that the initial $f_0$ in the unvoiced part differs greatly from the true one and eaQHM overcorrects the frequencies of the unvoiced speech, leading to the poor extraction in unvoiced parts and SRER degradation. It can be seen that the BP-eaQHM also improves the performance of the original eaQHM in this case.

Furthermore, we explore the performances of various methods when the frame-shift setting changes. We compute average SRERs of the results by all methods with different frame-shift (2 ms, 5 ms, 8 ms, 10 ms, 12 ms, and 15 ms), as shown in

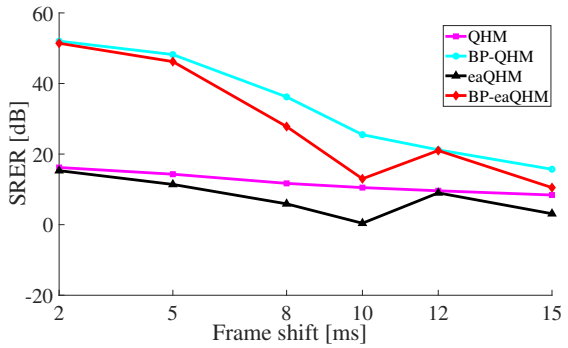


Fig. 2. Under different frame-shift, the SRERs of various methods.

TABLE II
AVERAGE SRER AND STOI SCORES WITH DIFFERENT $K$.

| Method | $K = 128$ | | $K = 64$ | | $K = 32$ | |
|---|---|---|---|---|---|---|
| | QHM | BP-QHM | QHM | BP-QHM | QHM | BP-QHM |
| SRER [dB] | 11.7 | 36.2 | 11.7 | 17.5 | 11.0 | 12.2 |
| STOI | 0.83 | 0.99 | 0.83 | 0.90 | 0.82 | 0.84 |

Fig. 2. Apparently, the performances of QHM and eaQHM degrade with the increase of the frame-shift. Although the performances of BP-QHM and BP-eaQHM also degrade, BP methods significantly outperform QHM methods, especially in terms of the reconstruction quality of speech. Even setting the frame shift to 12 ms, BP methods achieve more than 20 dB of SRER.

Subsequently, we investigate the effect of the harmonic number ($K$) on the various method. We set the frame-shift as 8 ms and use QHM and BP-QHM to analyze 16 kHz speech samples with different harmonic number settings (the harmonic number is 128, 64, and 32). Table II lists the average SRER and STOI scores of all methods, showing that BP-QHM works better than QHM even with fewer harmonics. However, the fewer harmonic number we set, the worse performances of the methods. Thus, it is necessary to select a proper harmonic number to guarantee the quality of the speech. This point is particularly important when analyzing the signals with a high sampling rate, as discussed below.

Eventually, we examine the properties of various methods when analyzing the speech signals with different sampling rates ($f_s$). Setting the frame-shift to 8 ms and harmonic number to 128, we apply QHM and BP-QHM for analyzing the signals with different sampling rates ($f_s$= 24 kHz and 44.1 kHz). The average SRER and STOI scores are given in Table III, indicating that BP-QHM works better than QHM in analyzing the speech sampled with 24 kHz. Note that BP-QHM performs worse than QHM in the analysis of 44.1 kHz speech. That is because the harmonic number is too limited for the BP method to adjust the parameters flexibly to approximate the target speech. To validate this argument, we conduct another analysis by setting the harmonic number to 256, whose result is listed in the right part of Table III. Apparently, BP-QHM

TABLE III
AVERAGE SRER AND STOI SCORES OF THE SPEECH WITH VARIOUS $f_s$.

| Method | $f_s = 24$ kHz | | $f_s = 44.1$ kHz | | $f_s = 44.1$ kHz and $K$=256 | |
|---|---|---|---|---|---|---|
| | QHM | BP-QHM | QHM | BP-QHM | QHM | BP-QHM |
| SRER [dB] | 14.9 | 21.8 | 15.1 | 14.2 | 15.1 | 29.7 |
| STOI | 0.83 | 0.93 | 0.87 | 0.87 | 0.87 | 0.99 |

outperforms QHM significantly, leading to a higher SRER and STOI. As a result, it can be concluded that it is necessary to increase the harmonic number when analyzing the speech with a high sampling rate.

Overall, the proposed BP methods can significantly optimize the extraction of amplitude and frequency and the quality of synthetic speech.

## V. CONCLUSIONS

In this paper, we propose a novel method to extract the frame-wise amplitude parameter of speech, which outperforms the conventional QHM methods. In the proposed method, the result of conventional QHM methods is considered as the initial input of the BP method. By measuring the error between synthetic speech and the original speech, and backwards propagating the gradients to update the parameters, the proposed method can get more accurate frame-wise amplitude parameters and frequency, with which the speech can be nearly flawlessly resynthesized. Experiments on real speech using various datasets verify the superiority of the proposed method.

## ACKNOWLEDGMENT

This work was partly supported by JST CREST Grant Number JPMJCR19A3 and JSPS KAKENHI Grant Number JP21H04892.

## REFERENCES

[1] T. Quatieri, R. McAulay. Speech transformations based on a sinusoidal representation[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1986, 34(6): 1449-1464.
[2] Sassan Ahmadi and Andreas S. Spanias. Low bit-rate speech coding based on an improved sinusoidal model. Speech Communication, 34(4):369 – 390, 2001.
[3] Yannis Stylianou. Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification. PhD thesis, Ecole Nationale Suprieure des T´el´ecommunications, 1996.
[4] J. Laroche Y. Stylianou and E. Moulines, "High-Quality Speech Modification based on a Harmonic + Noise Model.," Proceedings of EUROSPEECH, 1995.
[5] Yannis Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. IEEE Trans. on Speech and Audio Proc., 9:21–29, 2001.
[6] Y. Pantazis, O. Rosec and Y. Stylianou, "On the properties of a time-varying quasi-harmonic model of speech," in Proc. Interspeech, Brisbane, Australia, Sep. 2008, pp. 1044–1047.
[7] Y. Pantazis, O. Rosec, and Y. Stylianou. Adaptive AMFM signal decomposition with application to speech analysis. IEEE Trans. on Audio, Speech, and Language Processing, 19:290–300, 2011.
[8] G. P. Kafentzis, Y. Pantazis, O. Rosec, and Y. Stylianou. An Extension of the Adaptive Quasi-Harmonic Model. In Proc. IEEE ICASSP, Kyoto, March 2012.
[9] G. P. Kafentzis, O. Rosec, and Y. Stylianou, "Robust full-band adaptive sinusoidal analysis and synthesis of speech," in IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), 2014.

[10] G. P. Kafentzis and Y. Stylianou, "High-resolution sinusoidal modeling of unvoiced speech," in IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), 2016.
[11] G. P. Kafentzis, G. Degottex, O. Rosec, and Y. Stylianou, "Time-scale Modifications based on an Adaptive Harmonic Model," in IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), Vancouver, CA, May 2013.
[12] G. P. Kafentzis, G. Degottex, O. Rosec, and Y. Stylianou, "Pitch modifications of speech based on an adaptive harmonic model," in IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), 2014.
[13] Y. LeCun, D. Touresky, G. Hinton, et al. A theoretical framework for back-propagation[C]//Proceedings of the 1988 connectionist models summer school. 1988, 1: 21-28.
[14] R. Hecht-Nielsen. Theory of the backpropagation neural network[M]//Neural networks for perception. Academic Press, 1992: 65-93.
[15] R. J. McAulay and T. F. Quatieri. Speech Analysis/Synthesis based on a Sinusoidal Representation. IEEE Trans. on Acoustics, Speech and Signal Processing, 34:744–754, 1986.
[16] Keith Ito and Linda Johnson. The ljspeech dataset. https://keithito.com/LJ-Speech-Dataset/
[17] H. Zen, V. Dang, R. Clark, et al. LibriTTS: A corpus derived from LibriSpeech for text-to-speech[J]. arXiv preprint arXiv:1904.02882, 2019.
[18] H. Bu, J. Du, X. Na, et al. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline[C], 2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA). IEEE, 2017: 1-5.
[19] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CREPE: A convolutional representation for pitch estimation," in Proc. Int. Conf. Acoust., Speech, Signal Process., Feb. 2018, pp. 161–165. [Online]. Available: http://arxiv.org/abs/1802.06182
[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2010.
[21] J. Kolen, J. Pollack. Back propagation is sensitive to initial conditions[J]. Advances in neural information processing systems, 1990, 3.