IMPROVING FFTNET VOCODER WITH NOISE SHAPING AND SUBBAND APPROACHES

Takuma Okamoto[1], *Tomoki Toda*[2,1], *Yoshinori Shiga*[1], *and Hisashi Kawai*[1]

[1]National Institute of Information and Communications Technology, Japan
[2]Information Technology Center, Nagoya University, Japan

ABSTRACT

Although FFTNet neural vocoders can synthesize speech waveforms in real time, the synthesized speech quality is worse than that of WaveNet vocoders. To improve the synthesized speech quality of FFTNet while ensuring real-time synthesis, residual connections are introduced to enhance the prediction accuracy. Additionally, time-invariant noise shaping and subband approaches, which significantly improve the synthesized speech quality of WaveNet vocoders, are applied. A subband FFTNet vocoder with multiband input is also proposed to directly compensate the phase shift between subbands. The proposed approaches are evaluated through experiments using a Japanese male corpus with a sampling frequency of 16 kHz. The results are compared with those synthesized by the STRAIGHT vocoder without mel-cepstral compression and those from conventional FFTNet and WaveNet vocoders. The proposed approaches are shown to successfully improve the synthesized speech quality of the FFTNet vocoder. In particular, the use of noise shaping enables FFTNet to significantly outperform the STRAIGHT vocoder.

Index Terms— speech synthesis, vocoder, WaveNet, FFTNet, noise shaping, subband processing

1. INTRODUCTION

In conventional statistical parametric speech synthesis (SPSS) [1] and voice conversion (VC) [2], source-filter vocoders are typically employed to synthesize speech waveforms from estimated and converted acoustic features. These features are mainly constructed from the fundamental frequency and vocal tract spectrums. To improve the synthesized speech quality of conventional SPSS and VC over that offered by a simple mel-log spectrum approximation (MLSA) filter [3], several sophisticated corpus-independent vocoders have been developed [4–8]. Compared with these corpus-independent vocoders, deep learning-based corpus-dependent approaches, such as acoustic feature extraction [9], glottal vocoder [10, 11], power spectrum reconstruction for vocoded speech [12], and speech waveform generation [13, 14] from power spectrums estimated using the Griffin–Lim algorithm [15], have been investigated as in deep learning-based acoustic models for SPSS and VC [16]. However, the quality of their synthesized speech is limited by the analysis errors, approximations, and assumptions inherent in conventional vocoders.

WaveNet [17, 18] is a neural network-based raw audio autoregressive generative approach. In text-to-speech synthesis (TTS), WaveNet directly synthesizes raw speech waveforms from linguistic features, allowing it to outperform state-of-the-art unit selection- and SPSS-based TTS systems. Other raw audio generative models such as SampleRNN [19] and WaveRNN [20] have also been proposed. Such raw audio generative models can realize end-to-end TTS, converting text to raw speech waveforms. Examples include Char2Wav [21], Deep Voice [22–24], and Tacotron 2 [25–27]. The speech quality of English synthesized by Tacotron 2 can match that of natural speech with a sampling frequency of 24 kHz [25].

Compared with TTS, a WaveNet-based neural vocoder that directly synthesizes raw speech waveforms from acoustic features [28] has been used to drive conventional source-filter vocoders within a raw audio generative model framework. Neural vocoders based on WaveNet [28] and SampleRNN [29] have been applied to SPSS [30–34] and VC [35–39], and also outperform conventional source-filter vocoders. In addition, several speaker-independent WaveNet vocoders have been investigated [23, 24, 34, 40, 41].

Although source-filter vocoders can synthesize speech waveforms in real time, the synthesis speed of WaveNet and SampleRNN neural vocoders remains problematic, because the sequential synthesis of each sample requires a huge number of network parameters [17, 19]. To overcome this problem, Deep Voice uses smaller networks that quickly synthesize speech waveforms in real time. However, there is a tradeoff between the synthesis speed and the synthesized speech quality [22]. Parallel WaveNet [18] and WaveRNN [20] use a probability density distillation and a single-layer recurrent neural network with sparse and subscale modifications, respectively. These methods enable real-time synthesis with 16-bit linear pulse code modulation (PCM) raw audio prediction without any degradation in synthesized speech quality. However, the detailed network structures of these methods, especially the linguistic feature input network, are not disclosed, and complicated training and synthesis strategies might be required.

An alternative raw audio autoregressive generative model, FFTNet [42], has a simpler structure based on a 1×1 convolutional network and rectified liner unit (ReLU) layers. As a result, FFTNet can realize real-time raw audio synthesis. To improve the synthesized speech quality, four modifications have been introduced, namely, zero padding and noise injection in the training stage, argmax sampling of voice segments, and spectral subtraction [43]. However, despite these modifications, the synthesized speech quality of FFTNet vocoders remains inferior to that of WaveNet vocoders. In particular, argmax sampling can only be used when the fundamental frequency is known, and spectral subtraction introduces additional musical noise. Therefore, alternative approaches are required.

To improve the synthesized speech quality of FFTNet vocoders while ensuring that the network model size remains small enough for real-time synthesis, this paper presents the following four approaches. 1) Residual connections are introduced into FFTNet to improve the prediction accuracy. 2) Time-invariant noise shaping [40, 44, 45] and 3) subband approaches [46, 47], which significantly improve the synthesized speech quality in WaveNet vocoders, are directly applied to FFTNet. 4) A subband FFTNet vocoder with multiband input is proposed for the direct compensation of the phase shift between subbands.

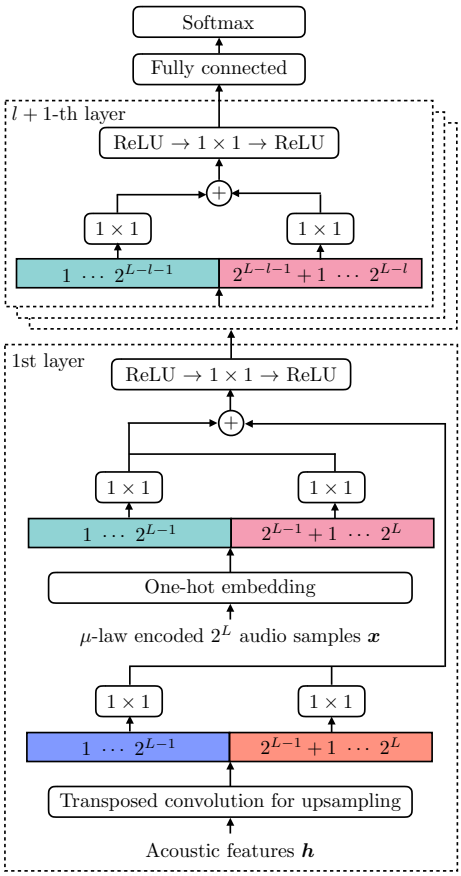


Fig. 1. Baseline FFTNet vocoder structure with L layers.

2. WAVENET AND FFTNET NEURAL VOCODERS

Given the acoustic features $\boldsymbol{h}$, the WaveNet and FFTNet neural vocoders [28, 42] model the conditional probability distribution $p(\boldsymbol{x}|\boldsymbol{h})$ of the raw audio waveform $\boldsymbol{x} = [x(1), \cdots, x(T)]$ as

$$p(\boldsymbol{x}|\boldsymbol{h}) = \prod_{t=1}^{T} p(x(t)|x(1), \cdots, x(t-1), \boldsymbol{h}). \quad (1)$$

In WaveNet, Eq. (1) is modeled by a stack of dilated causal convolution layers, allowing the efficient input of very long audio samples with relatively few layers. However, the network model size of WaveNet vocoders is still too large to synthesize speech waveforms in real time.

To significantly reduce the network model size, FFTNet uses simple 1×1 convolution layers instead of the dilated causal convolution layers, and can therefore synthesize speech waveforms in real time with a fast generation algorithm [48].

Rather than a continuous distribution, the WaveNet and FFTNet models output a categorical distribution of the next sample $x(t)$ through a final softmax layer. This approach is relatively flexible and can easily model arbitrary distributions, although raw waveform inputs are typically treated as continuous values. In vanilla WaveNet and FFTNet, a μ-law companding algorithm defined in G.711 [49] is introduced and raw audio waveforms are quantized into one of 256 possible values.

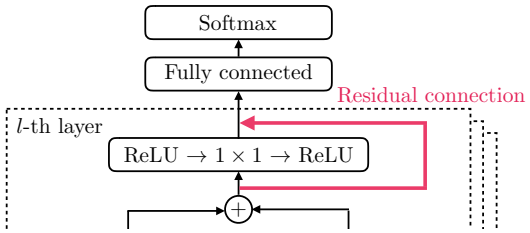


Fig. 2. Proposed FFTNet with residual connections.

In SPSS and VC, acoustic features for vocoders are typically analyzed every 5 ms. Some adjustment of the time resolution between speech waveform $\boldsymbol{x}$ and acoustic features $\boldsymbol{h}$ is then required. In WaveNet vocoders, a simple approach for matching the sequence lengths of $\boldsymbol{x}$ and $\boldsymbol{h}$ involves copying $\boldsymbol{h}$ in each frame according to the shift amount of the analysis window [28, 40, 47]. In FFTNet vocoders, a linear interpolation method is applied [42]. Based on empirical results, transposed convolution [17] is applied for upsampling the acoustic features in both the WaveNet and FFTNet vocoders, rather than the simple copy and linear interpolation approaches. The baseline L-layer FFTNet vocoder structure investigated in this paper is illustrated in Fig. 1, where the receptive field is 2^L samples.

3. IMPROVING FFTNET VOCODER BY INTRODUCING RESIDUAL CONNECTIONS AND SIGNAL PROCESSING APPROACHES

The simple network structure of FFTNet enables speech waveforms to be synthesized in real time, although the synthesized speech quality is not as good as that of WaveNet. To improve the synthesized speech quality of FFTNet while retaining a network model size that allows real-time synthesis, a network structure modification and two signal processing methods, time-invariant noise shaping and subband processing, are now described.

3.1. Introducing residual connections

To improve the FFTNet model while ensuring real-time synthesis, two network modifications inspired by the WaveNet model structure are investigated.

The first is the introduction of skip connections from all layers, as used in WaveNet. However, empirical results indicate that this cannot improve the FFTNet prediction accuracy, and so it is not used in the experiments.

The other modification introduces residual connections in all layers, an approach that is also employed in WaveNet, as shown in Fig. 2. Experimental results (see Sec. 4) demonstrate that this modification can significantly improve the prediction accuracy and synthesized speech quality of FFTNet.

3.2. FFTNet with time-invariant noise shaping method

The speech signals generated by WaveNet often suffer from noise caused by prediction errors, and these noise signals tend to cause large spectral distortions in high-frequency bands. Thus, the noise signals degrade the synthesized speech quality [44].

To reduce the adverse effects of the noise signals generated by neural vocoders, predictive pulse code modulation (PPCM) [50]-based time-invariant noise shaping, which is a perceptual weighting

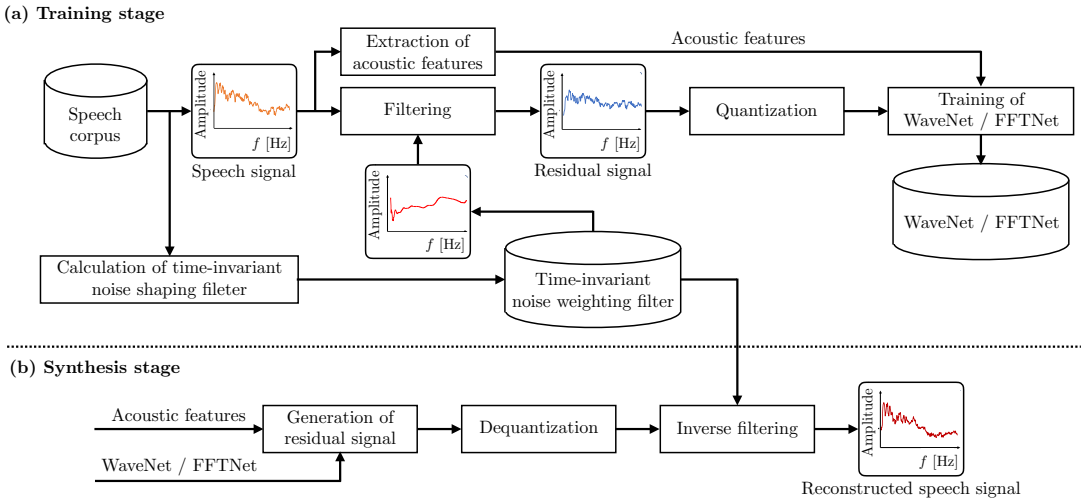


Fig. 3. Time-invariant noise shaping method for WaveNet and FFTNet neural vocoders.

technique, has been applied to WaveNet [40, 44]. This noise shaping method is expected to improve the synthesized speech quality of FFTNet vocoders, and is therefore directly applied to FFTNet. A block diagram of time-invariant noise shaping is depicted in Fig. 3.

Similar to WaveNet methods [40, 44], a mel-generalized cepstrum [51] is employed to calculate the noise shaping filter. The transfer function of the filter is given by

$$H(z) = s_\gamma^{-1}\left(c_\gamma(0) + \sum_{m=1}^{M_c} \beta c_\gamma(m)\tilde{z}^{-m}\right), \tag{2}$$

$$s_\gamma^{-1}(\omega) = \begin{cases} (1+\gamma\omega)^{\frac{1}{\gamma}}, & 0 < |\gamma| \leq 1 \\ e^{\omega}, & \gamma = 0 \end{cases}, \tag{3}$$

where $c_\gamma(m)$, γ, β, and M_c are the m-th mel-generalized cepstral coefficients, a power parameter of the mel-generalized cepstrum, a parameter to control noise energy in the formant regions, and the order of the mel-generalized cepstrum, respectively. $\tilde{z}^{-1}$ is the first-order all-pass function given by

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \tag{4}$$

where α is the frequency warping parameter.

The averaged mel-generalized cepstral coefficients are calculated in advance over all frames extracted from the training data, and are then used for the time-invariant noise shaping filtering implemented by the MLSA filter [3] in the training stage (Fig. 3(a)). In the synthesis stage, the inverse filter is easily derived by multiplying the mel-generalized cepstral coefficients by -1, and the reconstructed speech signal is finally obtained as shown in Fig. 3(b).

3.3. Subband FFTNet

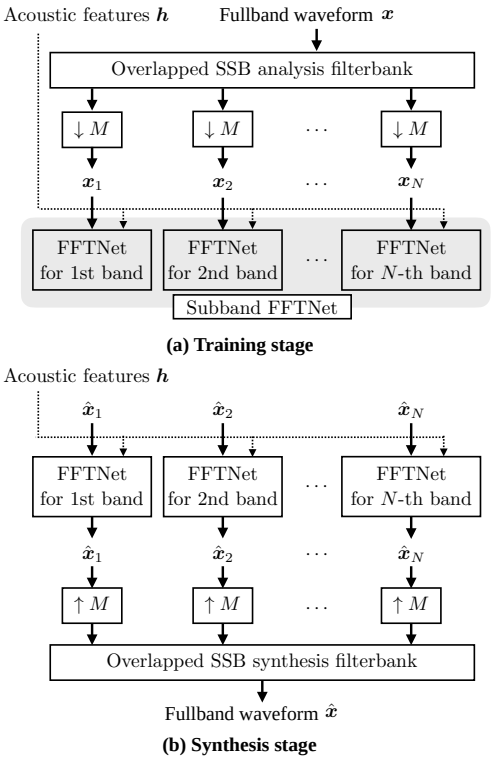
For rapid synthesis and improved synthesized speech quality, a subband WaveNet based on multirate signal processing [52,53] has been proposed [46, 47]. By introducing an overlapped single-sideband (SSB) filterbank based on a square-root Hann window [46, 47], subband WaveNet achieves better prediction accuracy compared with fullband WaveNet, resulting in accelerated synthesis speed and improved synthesized speech quality. Thus, the subband approach is also directly applied to FFTNet.

A block diagram of the proposed subband FFTNet vocoder is illustrated in Fig. 4. In the training stage, fullband speech waveforms $\boldsymbol{x} = [x(1), \cdots, x(T)]$ with a sampling frequency of fs are decimated by a factor of M and decomposed into N subband streams $\boldsymbol{x}_n = [x_n(1), \cdots, x_n(T/M)]$ of (short) length T/M and low sampling frequency fs/M by an overlapped SSB analysis filterbank. Each subband FFTNet network $p_n(\boldsymbol{x}_n|\boldsymbol{h})$ is then separately and efficiently trained by each subband waveform $\boldsymbol{x}_n$ with common acoustic features $\boldsymbol{h}$. In the synthesis stage, each subband stream $\hat{\boldsymbol{x}}_n = [\hat{x}_n(1), \cdots, \hat{x}_n(T/M)]$ is simultaneously generated by the trained network and upsampled by a factor of M, and the synthesized speech waveform with a sampling frequency of fs is obtained by an overlapped SSB synthesis filterbank. The proposed subband method can be combined with the noise shaping approach, and is evaluated in the experiments reported in Sec. 4.

In the subband WaveNet vocoder [47], there is a phase shift between subbands because each estimated sample $\hat{x}_n(t)$ is independently generated from already-estimated past samples $[\hat{x}_n(1), \cdots, \hat{x}_n(t-1)]$ and acoustic features $\boldsymbol{h}$ with random sampling based on $p_n(\boldsymbol{x}_n|\boldsymbol{h})$. Thus, a maximum correlation-based phase shift compensation between subbands is introduced to the subband WaveNet vocoder. However, the results of further investigations indicate that the maximum correlation-based approach cannot improve the synthesized speech quality when transposed convolution is used to upsample the acoustic features. Therefore, the maximum correlation-based approach is not included in the subband FFTNet vocoder, and another phase shift compensation approach is proposed in the next subsection.

3.4. Subband FFTNet with multiband input

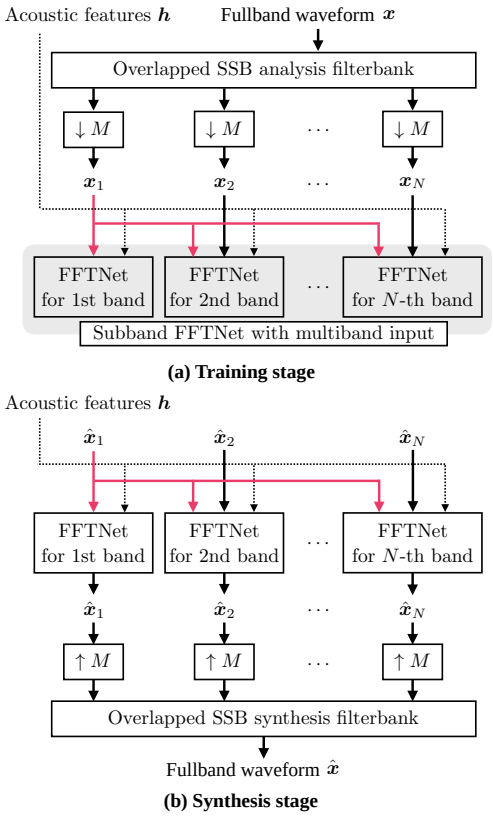
To directly compensate the phase shift between subbands within a neural network framework, a subband FFTNet vocoder with multiband input is proposed. In this approach, the n-th band training and synthesis uses other band waveforms as well as the n-th band, unlike the subband method with single-band input described in Sec. 3.3. The same strategy is used in several other methods: 1) pixelCNN [54, 55] for conditional color image generation, where a blue pixel is predicted from the previous red, green, and blue pixels; 2) the neural parametric singing synthesizer [56, 57], where



(a) Training stage

(b) Synthesis stage

Fig. 4. Block diagram of proposed subband FFTNet vocoder.



(a) Training stage

(b) Synthesis stage

Fig. 5. Block diagram of proposed subband FFTNet vocoder with multiband input.

a vocal tract spectrum is predicted using the previous spectrum and fundamental frequency trajectories; and 3) subscale WaveRNN [20].

However, when using many band waveforms, the network model size increases because the number of 1×1 convolution layers in the first layer corresponds to the number of multiband waveforms. To reduce the network model size, the use of only the first and n-th band waveforms for n-th band training and synthesis is investigated. A block diagram of the proposed subband FFTNet vocoder with multiband input and its first-layer network structure for n-th band training and synthesis are depicted in Figs. 5 and 6, respectively. Compared with the single-band input method, two additional 1×1 convolution layer components are required for the first-band waveform input, as shown in Fig. 6. For first-band training and synthesis, the standard subband method using only the first-band input is shown in Fig. 5. The proposed method is expected to directly compensate the phase shift between subbands in the synthesis with random sampling based on $p_n(\boldsymbol{x}_n|\boldsymbol{h})$. Compared with WaveNet, the subband method with multiband input can be easily implemented in FFTNet because of its simple network structure, as shown in Fig. 6.

The proposed subband method with multiband input can also be combined with the noise shaping approach. However, empirical results suggest that the synthesized speech waveforms sometimes include collapsed segments, which are also found in WaveNet vocoders [58]. Therefore, subband FFTNet with multiband input combined with noise shaping is not included in the experiments reported here, and further investigations will be conducted in future work.

4. EXPERIMENTS

4.1. Experimental conditions

To evaluate the effectiveness of the proposed methods for FFTNet vocoders, a series of objective and subjective experiments were conducted using a Japanese male speech corpus recorded with a sampling frequency of 48 kHz and downsampled to 16 kHz, as used in [46, 47]. In the experiments, 5697 utterances (about 3.7 h) were used as the training set and 20 utterances were used as the test set. The experiments compared (a) the baseline vanilla FFTNet vocoder [42] with (b) FFTNet with residual connections, (c) FFTNet with noise shaping, (d) subband FFTNet, (e) subband FFTNet with noise shaping, and (f) subband FFTNet with multiband input. In addition, results were computed for (g) the vanilla WaveNet vocoder [28], (h) the WaveNet vocoder with time-invariant noise shaping method [44], and (i) the conventional STRAIGHT source-filter vocoder without mel-cepstral compression [4]. Methods (c)–(f) also included residual connections.

In the experiments, acoustic features $\boldsymbol{h}$ were analyzed every 5 ms over a Hann window of length 25 ms. The fundamental frequency f_0, analyzed by an NDF algorithm implemented in STRAIGHT [59], was used in all the vocoders.

For the WaveNet and FFTNet vocoders, the 0–th to 24–th mel-cepstral coefficients (25 dimensions) were analyzed from a simple short-time Fourier transform of windowed speech waveforms with a sampling frequency of 16 kHz and warping coefficient $\alpha = 0.42$.

In the STRAIGHT vocoder, the original STRAIGHT smooth vo-

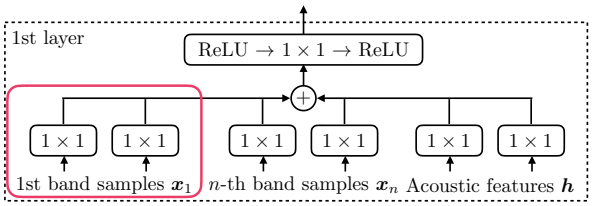


Fig. 6. First layer of subband FFTNet vocoder with multiband input for n-th band training and synthesis.

Table 1. Number of network model parameters.

Model	Num of params
WaveNet (g) and (h)	44,592,721
FFTNet (a) to (c)	2,251,857
Subband FFTNet (d) and (e)	1,857,105 (each subband)
Subband FFTNet with multiband input (f)	1,988,117 (each subband)

cal tract spectrum (1025 dimensions) and aperiodicity (1025 dimensions) coefficients without mel-cepstral compression were directly used. Although they are typically compressed as mel-cepstrum and band aperiodicity coefficients [60] for dimensional reduction, the original high-dimensional coefficients were used in the experiments to evaluate the potential of the proposed approaches.

In FFTNet vocoders (a)–(c), $L = 11$ layers were introduced and the receptive field was $2^{11} = 2048$ samples, as used in the original FFTNet [42]. A square-root Hann window-based overlapped SSB filterbank was also introduced for subband FFTNet vocoders (d)–(f). A decimation factor of $M = 4$ and division number $N = 2M + 1 = 9$ were also used. The length of the analysis and synthesis prototype FIR filters was 1024 samples. The sampling frequency of each subband waveform was $(16/4 =)$ 4 kHz. The frequency response of the filterbank is plotted in Fig. 7. In subband FFTNet vocoders (d)–(f), $L = 9$ layers were used and the receptive field was $2^9 = 512$ samples. The channel number of each FFTNet layer was 256 [42].

The dilation channel, residual channel, and skip channel of WaveNet vocoders (g) and (h) were set to 512, 512, and 256, respectively. Thirty layers (10 dilations $\times$ 3 cycles) with a kernel size of 2 were used for the dilated causal convolution layers, giving a receptive field of 3070 samples [17, 28].

In the WaveNet and FFTNet vocoders, $(1 + 1 + 25 =)$ 27-dimensional vectors constructed from the continuous logarithmic f_0, voice/unvoice one-hot vector, and mel-cepstrum coefficients (normalized to have a zero-mean and unit-variance) were used as the acoustic features $\boldsymbol{h}$. The WaveNet and FFTNet vocoders required 400,000 and 1,000,000 parameter updates, respectively, and an Adam optimization algorithm [61] updated the neural network parameters with a learning rate of 0.001. The minibatch sizes of WaveNet, FFTNet, and subband FFTNet were 1 $\times$ 20,000, 5 $\times$ 5,000, and 5 $\times$ 1,250 samples, respectively. They were trained using a single GPU of an NVIDIA Tesla P100. A value of $\beta = 0.5$ in Eq. (2) was used for the noise shaping methods (c), (e), and (h) according to the results of the WaveNet vocoder investigations [40, 44].

The number of network model parameters in the FFTNet and WaveNet vocoders is listed in Table 1. The FFTNet vocoders have about 1/20 the number of the WaveNet vocoders, and the FFTNet

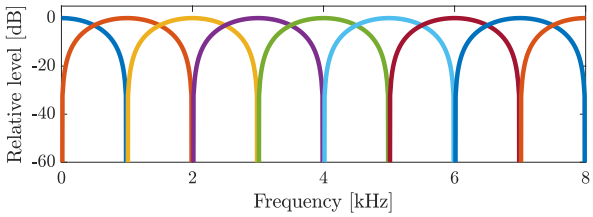


Fig. 7. Frequency response of a square-root Hann window-based overlapped SSB filterbank with decimation factor $M = 4$ and division number $N = 9$ for proposed subband FFTNet vocoder.

vocoders can synthesize speech waveforms in real time. In particular, the subband FFTNet vocoders can realize more than $M = 4$ times faster generation as a result of the decimation and smaller network, as in subband WaveNet [46, 47].

4.2. Objective evaluations

To objectively evaluate the synthesized test set speech waveforms, the signal-to-noise ratio (SNR) and the spectral distortion (SD) between the original waveform $x(t)$ and the synthesized $\hat{x}(t)$ were computed:

$$SNR = 10 \log_{10} \left(\frac{\sum_{t=1}^{T} \hat{x}(t)^2}{\sum_{t=1}^{T} (x(t) - \hat{x}(t))^2} \right), \tag{5}$$

$$SD = \frac{1}{A} \sum_{a=1}^{A} \sqrt{\frac{1}{F} \sum_{f=1}^{F} \left(20 \log_{10} \frac{|\hat{X}(f,a)|}{|X(f,a)|} \right)^2}, \tag{6}$$

where $X(f,a)$ and $\hat{X}(f,a)$ are the short-time Fourier spectrums of $x(t)$ and $\hat{x}(t)$ in frame a for frequency bin f, and A is the total number of frames. As in previous studies [28, 44, 47], a linear phase compensation for each frame was introduced to calculate the SNR. For acoustic feature analysis, the short-time Fourier transform analysis window function was also a Hann window with a frame length of 25 ms, a frameshift of 5 ms, and $F = 257$. To consider the human auditory perception criterion in the objective evaluation, the mel-cepstral distortion (MCD) was also computed. This is defined as:

$$MCD = \frac{10}{\log 10} \sqrt{2 \sum_{b=1}^{B} (c(b) - \hat{c}(b))^2}, \tag{7}$$

where $c(b)$ and $\hat{c}(b)$ are the b-th mel-cepstral coefficients obtained from $X(f,a)$ and $\hat{X}(f,a)$ with $\alpha = 0.42$ and $B = 24$. The results of the objective evaluations are presented in Table 2. In addition, the training softmax loss scores averaged over 10,000 iterations are given in Table 2 to evaluate the model accuracy. In subband methods (d)–(f), the loss scores were averaged over all nine bands.

4.3. Subjective evaluations

To subjectively evaluate the proposed approaches, mean opinion score (MOS) tests [62] were conducted. All 20 utterances of the test set were used as the evaluation set. These were presented through headphones to 10 Japanese adult native speakers without hearing loss (20 utterances $\times$10 conditions including the original test set waveforms $=$ 200 utterances). The MOS results are plotted in Fig. 8.

Table 2. Results of objective evaluations of 20 test set utterances. **Bold** and *italic* entries indicate best scores of all six FFTNet vocoders and of all nine methods, respectively.

	Training softmax loss score	SNR [dB]	SD [dB]	MCD [dB]
(a):vanilla FFTNet (baseline)	1.89	5.20 ± 0.26	10.29 ± 0.15	3.66 ± 0.11
(b):FFTNet with residual connections	1.81	5.50 ± 0.25	9.68 ± 0.12	3.33 ± 0.08
(c):FFTNet with noise shaping	2.19	4.00 ± 0.47	$\mathbf{8.19 \pm 0.05}$	$\mathbf{2.84 \pm 0.06}$
(d):subband FFTNet	1.39	4.00 ± 0.27	10.76 ± 0.30	2.96 ± 0.04
(e):subband FFTNet with noise shaping	1.55	2.90 ± 0.39	9.62 ± 0.22	$\mathbf{2.84 \pm 0.06}$
(f):subband FFTNet with multiband input	***1.35***	$\mathbf{5.80 \pm 0.36}$	10.84 ± 0.36	3.13 ± 0.39
(g):vanilla WaveNet	1.50	$\mathit{6.60 \pm 0.36}$	9.16 ± 0.12	2.50 ± 0.08
(h):WaveNet with noise shaping	1.80	5.50 ± 0.60	7.58 ± 0.06	$\mathit{2.00 \pm 0.07}$
(i):STRAIGHT	–	0.10 ± 0.47	$\mathit{7.09 \pm 0.07}$	2.78 ± 0.08

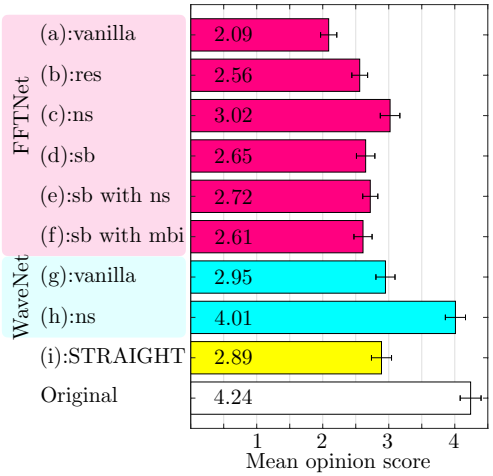


Fig. 8. Results of MOS test with 10 listening subjects. "res," "ns," "sb," and "mbi" denote residual connection, noise shaping, subband, and multiband input, respectively.

4.4. Discussion

Compared with the vanilla FFTNet (a), the MOS test results (Fig. 8) show that the synthesized speech quality can be significantly improved by approaches (b)–(f), although they cannot match the performance of the WaveNet with noise shaping (h). This verifies the effectiveness of the noise shaping and subband approaches for FFTNet vocoders. In particular, FFTNet with noise shaping (c) significantly outperforms the STRAIGHT vocoder without mel-cepstral compression (t-test result with $p = 0.049 < 0.05$). The SNR results for the noise shaping methods (c) and (h) were lower than those for the methods without noise shaping (b) and (g), as the noise-shaped speech signals are whiter than the original signals and the higher loss scores indicate that they are more difficult to predict. However, the synthesized speech quality of the methods with noise shaping was successfully improved. In contrast, the subband speech signals are covered by the square-root Hann window-based overlapped SSB filterbank, and the lower loss scores suggest that they are more easily predicted [46]. The residual connections can successfully improve the FFTNet model accuracy, and the noise shaping and subband approaches can improve the synthesized frequency response. In addition, the subband method with multiband input (f) can effectively compensate the phase shift between subbands because it produces higher SNR scores than the other FFTNet vocoders. However, method (f) cannot realize higher-quality synthesis because the synthesized speech waveforms include some additional noise signals in a higher frequency band compared with subband methods with a single-band input. Although the subband methods with a single-band input incorporate fewer noise components, the synthesized speech quality is not as high because there is a lack of phase shift compensation between subbands. Therefore, subband FFTNet with multiband input should be further investigated to reduce the noise components.

5. FUTURE WORK

To enhance the synthesized speech quality of FFTNet to that of WaveNet and the original speech signals, alternative network structure modifications such as the gated activation units used in WaveNet [17] and signal processing techniques such as subband methods with multiband input combined with noise shaping will be investigated at higher sampling frequencies, such as 24 [18, 25, 34] and 48 kHz [24, 47]. In addition, experiments using a female speech corpus will be conducted. Furthermore, bandwidth extension [41, 47, 63], 16 bit linear PCM raw audio prediction with discretized logistic mixture likelihood [18, 25, 34, 55] or a dual softmax layer [20], synthesis with mel-spectrogram input [25, 39], and synthesis speed evaluations for FFTNet vocoders offer significant scope for future studies.

6. CONCLUSIONS

To improve the synthesized speech quality of FFTNet vocoders while retaining the small network model size required for real-time synthesis, this paper has described the following four approaches. 1) Residual connections were introduced into FFTNet to improve the prediction accuracy. 2) Noise shaping and 3) subband approaches were directly applied to FFTNet vocoders. 4) A subband FFTNet vocoder with multiband input was used to directly compensate the phase shift between subbands. The proposed approaches were evaluated through a series of objective and subjective experiments using a Japanese male corpus with a sampling frequency of 16 kHz, and the results were compared with those from STRAIGHT without mel-cepstral compression as well as vanilla FFTNet and WaveNet vocoders. The results suggest that the proposed approaches can significantly improve the synthesized speech quality of FFTNet vocoders. In particular, the proposed FFTNet vocoder with noise shaping significantly outperforms the STRAIGHT vocoder.

7. REFERENCES

[1] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proc. IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

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

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

[4] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "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 Commun.*, vol. 27, no. 3–4, pp. 187–207, Apr. 1999.

[5] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE trans. Inf. Syst.*, vol. E99-D, no. 7, pp. 1877–1884, July 2016.

[6] C. Demiroglu, O. Buyuk, A. Khodabakhsh, and R. Maia, "Postprocessing synthetic speech with a complex cepstrum vocoder for spoofing phase-based synthetic speech detectors," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 4, pp. 671–683, June 2017.

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

[8] Y. Agiomyrgiannakis, "Vocaine the vocoder and applications in speech synthesis," in *Proc. ICASSP*, Apr. 2015, pp. 4230–4234.

[9] S. Takaki and J. Yamagishi, "A deep auto-encoder based low-dimensional feature extraction from FFT spectral envelopes for statistical parametric speech synthesis," in *Proc. ICASSP*, Mar. 2016, pp. 5535–5539.

[10] M. Airaksinen, B. Bollepalli, L. Juvela, Z. Wu, S. King, and P. Alku, "GlottDNN — A full-band glottal vocoder for statistical parametric speech synthesis," in *Proc. Interspeech*, Sept. 2016, pp. 2473–2477.

[11] M. Airaksinen, L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "Comparison between STRAIGHT, glottal, and sinusoidal vocoding in statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1658–1670, Sept. 2018.

[12] T. Okamoto, K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, "Deep neural network-based power spectrum reconstruction to improve quality of vocoded speech with limited acoustic parameters," *Acoust. Sci. Tech.*, vol. 39, no. 2, pp. 163–166, Mar. 2018.

[13] S. Takaki, H. Kameoka, and J. Yamagishi, "Direct modeling of frequency spectra and waveform generation based on phase recovery for DNN-based speech synthesis," in *Proc. Interspeech*, Aug. 2017, pp. 1128–1132.

[14] Y. Wang, RJ Skerry-Ryan, D. Stanton, Y. Wu, R. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, Aug. 2017, pp. 4006–4010.

[15] D. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[16] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 35–52, May 2015.

[17] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, Sept. 2016, (unreviewed manuscript).

[18] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. ICML*, July 2018, pp. 3915–3923.

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

[20] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *Proc. ICML*, July 2018, pp. 2415–2424.

[21] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2Wav: End-to-end speech synthesis," in *Proc. ICLR*, Apr. 2017.

[22] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep voice: Real-time neural text-to-speech," in *Proc. ICML*, Aug. 2017, pp. 195–204.

[23] S. O. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Proc. NIPS*, Dec. 2017, pp. 2966–2974.

[24] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," in *Proc. ICLR*, Apr. 2018.

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

[26] RJ Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proc. ICML*, July 2018, pp. 4700–4709.

[27] Y. Wang, D. Stanton, Y. Zhang, RJ Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. ICML*, July 2018, pp. 5167–5176.

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

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

[30] X. Wang, J. Lorenzo-Trueba, S. Takaki, L. Juvela, and J. Yamagishi, "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis," in *Proc. ICASSP*, Apr. 2018, pp. 4804–4808.

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

[32] J. Vít, Z. Hanzlíček, and J. Matoušek, "On the analysis of training data for WaveNet-based speech synthesis," in *Proc. ICASSP*, Apr. 2018, pp. 5684–5688.

[33] Y. Gu and Y. Kang, "Multi-task WaveNet: A multi-task generative model for statistical parametric speech synthesis without fundamental frequency conditions," in *Proc. Interspeech*, Sept. 2018, pp. 2007–2011.

[34] L. Juvela, V. Tsiaras, B. Bollepalli, M. Airaksinen, J. Yamagishi, and P. Alku, "Speaker-independent raw waveform model for glottal excitation," in *Proc. Interspeech*, Sept. 2018, pp. 2012–2016.

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

[36] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proc. NIPS*, Dec. 2017, pp. 6306–6315.

[37] J. Niwa, T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Statistical voice conversion based on WaveNet," in *Proc. ICASSP*, Apr. 2018, pp. 5289–5293.

[38] B. Sisman, M. Zhang, and H. Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted WaveNet vocoder," in *Proc. Interspeech*, Sept. 2018, pp. 1978–1982.

[39] K. Chen, B. Chen, J. Lai, and K. Yu, "High-quality voice conversion using spectrogram-based WaveNet vocoder," in *Proc. Interspeech*, Sept. 2018, pp. 1993–1997.

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

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

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

[43] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

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

[45] T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Mel-cepstrum-based quantization noise shaping applied to neural-network-based speech waveform synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 7, pp. 1173–1180, July 2018.

[46] T. Okamoto, K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, "Subband WaveNet with overlapped single-sideband filterbanks," in *Proc. ASRU*, Dec. 2017, pp. 698–704.

[47] T. Okamoto, K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, "An investigation of subband WaveNet vocoder covering entire audible frequency range with limited acoustic features," in *Proc. ICASSP*, Apr. 2018, pp. 5654–5658.

[48] P. Ramachandran, T. L. Paine, P. Khorrami, M. Babaeizadeh, S. Chang, Y. Zhang, M. Hasegawa-Johnson, R. Campbell, and T. Huang, "Fast generation for convolutional autoregressive models," in *Proc. ICLR*, Apr. 2017.

[49] ITU-T. Recommendation G. 711, *Pulse Code Modulation (PCM) of voice frequencies*, 1988.

[50] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 27, no. 3, pp. 247–254, June 1979.

[51] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis — A unified approach to speech spectral estimation," in *Proc. ICSLP*, Sept. 1994, pp. 1043–1046.

[52] R. E. Crociere and L. R. Rabiner, *Multirate Digital Signal Processing*, Prentice Hall, Englewood Cliffs, 1983.

[53] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall, Upper Saddle River, 1993.

[54] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with pixelCNN decoders," in *Proc. NIPS*, Dec. 2016, pp. 4790–4798.

[55] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "PixelCNN++: Improving the pixelCNN with discretized logistic mixture likelihood and other modifications," in *Proc. ICLR*, Apr. 2017.

[56] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer," in *Proc. Interspeech*, Aug. 2017, pp. 4001–4005.

[57] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer modeling timbre and expression from natural songs," *Appl. Sci.*, vol. 7, no. 12, 1313, Dec. 2017.

[58] Y.-C. Wu, K. Kobayashi, T. Hayashi, P. L. Tobing, and T. Toda, "Collapsed speech segment detection and suppression for WaveNet vocoder," in *Proc. Interspeech*, Sept. 2018, pp. 1988–1992.

[59] H. Kawahara, A. de Cheveigné, H. Banno, T. Takahashi, and T. Irino, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," in *Proc. Interspeech*, Sept. 2005, pp. 537–540.

[60] H. Zen, T. Toda, M. Nakamura, and T. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.

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

[62] ITU-T Recommendation P. 800, *Methods for subjective determination of transmission quality*, 1996.

[63] Y. Gu and Z.-H. Ling, "Waveform modeling using stacked dilated convolutional neural networks for speech bandwidth extension," in *Proc. Interspeech*, Aug. 2017, pp. 1123–1127.