AN INVESTIGATION OF NOISE SHAPING WITH PERCEPTUAL WEIGHTING FOR WAVENET-BASED SPEECH GENERATION

Kentaro Tachibana[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

tomoki@icts.nagoya-u.ac.jp, {yoshi.shiga, hisashi.kawai}@nict.go.jp

ABSTRACT

We propose a noise shaping method to improve the sound quality of speech signals generated by WaveNet, which is a convolutional neural network (CNN) that predicts a waveform sample sequence as a discrete symbol sequence. Speech signals generated by WaveNet often suffer from noise signals caused by the quantization error generated by representing waveform samples as discrete symbols and the prediction error of the CNN. We analyze these noise signals and show that 1) since the prediction error is much larger than the quantization error, the effect of the quantization error on the noise signals is practically negligible, and 2) noise signals tend to cause large spectral distortion in a high-frequency band. To alleviate the adverse effect of these noise signals on the generated speech signals, the proposed noise shaping method applies a perceptual weighting filter to WaveNet, making it possible to use the frequency masking properties of the human auditory system. We conducted objective and subjective evaluations to investigate the effectiveness of the proposed method and demonstrated that it significantly improved the sound quality of the generated speech signals.

Index Terms— speech synthesis, WaveNet, noise analysis, noise shaping, perceptual weighting

1. INTRODUCTION

Statistical parametric speech synthesis has been actively researched as a framework to generate a synthetic speech signal based on a statistical model [1–3]. In this framework, a natural speech signal is usually parameterized into acoustic features, which are then statistically modeled using a probabilistic generative model. In the synthesis, the acoustic features are generated from the model, and the synthetic speech signal is reconstructed from the generated acoustic features using a speech production approach, such as a source-filter model. One of the biggest issues in this framework is that the synthetic speech signal always suffers from errors in the parameterization and reconstruction processes.

As a new framework free of this issue, WaveNet was recently proposed [4]. In WaveNet, the speech signal is represented as a discrete symbol sequence using a quantization process with a μ-law companding transformation [5]. Then its probability mass function is modeled with a Markov model (*i.e.*, n-gram) using a convolutional neural network (CNN). In the synthesis, the discrete symbol sequence, which corresponds to the speech waveform samples, is directly generated from the CNN in the same manner as the random sampling from the n-gram model. WaveNet was originally proposed as a text-to-speech (TTS) method and its outstanding performance surpassed the existing state-of-the-art TTS methods [6–8]. It has also been successfully applied to the reconstruction process of speech signals from acoustic features, making it possible to develop a new vocoder, the WaveNet vocoder [9, 10]. Since the WaveNet vocoder no longer needs to use the source-filter model, it has great potential to outperform such conventional high-quality vocoders as STRAIGHT [11] and WORLD [12]. The direct waveform modeling approach, which includes not only WaveNet but also SampleRNN [13], has attracted attention and its effectiveness has been confirmed in several speech synthesis systems [14–16].

Although the direct waveform modeling approach tends to generate natural sounding speech signals, the generated speech signals usually suffer from the noise signals caused by two types of errors: the quantization error caused by representing waveform samples as discrete symbols and the prediction error of the generative model. To further improve the sound quality of the generated speech signals, it will be useful to analyze the statistical characteristics of these noise signals and develop a technique to suppress them.

In this paper, we first analyze the noise signals generated in the WaveNet vocoder and show the following two results: 1) since the prediction error is much larger than the quantization error, the effect of the latter on the noise signals is practically negligible, and 2) the noise signals tend to cause large spectral distortion in high-frequency bands. These results inspire us to use the frequency masking properties of the human auditory system to improve the sound quality of the generated speech signals. That is, the human auditory system has a limited capability to detect noise in frequency bands in which the speech signal has high energy, *e.g.*, near the formant peaks [17, 18], which is expected to be helpful for perceptually reducing the adverse effects of the noise signals on the generated speech signals. As a trial with this auditory masking effect, we propose a noise shaping method based on a perceptual weighting filter for reallocating the frequency components of the noise signals in relation to the global spectral shape of the speech signals. We will conduct objective and subjective evaluations and investigate the effectiveness of the proposed method, thereby demonstrating that it significantly improves the sound quality of the generated speech signals.

2. ANALYSIS OF NOISE GENERATED IN WAVENET

2.1. WaveNet [4]

A speech signal is typically represented as 16-bit integer values. In WaveNet, they are quantized into 8 bits by μ-law companding transformation, which is a non-linear function that roughly captures the characteristics of a sample distribution of the speech signals. Then a sample-by-sample speech waveform generation process is handled as a sample-by-sample multi-class classification task for 256 ($= 2^8$) classes. Thus, the speech signals are represented as discrete symbol sequences consisting of 256 discrete values.

K. Tachibana is currently with the DeNA Co. Ltd., Japan.

WaveNet is a Markov model that predicts the current discrete symbol from a fixed number of past discrete symbols. Therefore, the probability mass function of the discrete symbol sequence (*i.e.*, a quantized speech signal), $x_1, \cdots, x_T$, is given by

$$p(x_1, \cdots, x_{\mathrm{T}} | \boldsymbol{h}_1, \cdots, \boldsymbol{h}_T) = \prod_{t=1}^{T} p(x_t | x_{t-R}, \cdots, x_{t-1}, \boldsymbol{h}_t), \quad (1)$$

where $\boldsymbol{h}_t$ denotes an auxiliary feature vector at time t, *e.g.*, the acoustic feature vector in the WaveNet vocoder. Conditional probability mass function $p(x_t | x_{t-R}, \cdots, x_{t-1}, \boldsymbol{h}_t)$ is modeled by the CNN using many stacked convolutional layers, a dilated causal convolution, gated activation units [19], and residual and skip connections [20]. The CNN parameters are optimized by minimizing the cross entropy between the observed data and the model distributions.

In the generation, the discrete symbol sequence is randomly generated sample by sample from the conditional probability mass function $\hat{x}_t \sim p(x_t | \hat{x}_{t-R}, \cdots, \hat{x}_{t-1}, \boldsymbol{h}_t)$, which is predicted from previously generated samples $\hat{x}_{t-R}, \cdots, \hat{x}_{t-1}$ as well as the auxiliary feature. In other words, CNN is used as an auto-regressive model. Finally, the μ-law expansion transformation is applied to the generated discrete symbol sequence to reconstruct the speech signal.

2.2. Analysis of noise signals

The speech signals generated by WaveNet suffer from the noise signals caused by the quantization error $e_t^{(q)} = s_t - x_t$, where s_t denotes a 16-bit integer value at time *t* and the prediction error $e_t^{(p)} = x_t - \hat{x}_t$. Here, we analyze the statistical characteristics of these two errors using the WaveNet vocoder [9].

We trained the WaveNet vocoder and calculated the following distortion metrics: 1) the signal-to-noise-ratio (SNR) to evaluate the signal distortion including both the amplitude and the phase distortion, 2) the mel-cepstral distortion (MCD) to evaluate the power spectral envelope distortion that corresponds relatively well to human perceptual distortion, and 3) the frequency-dependent log-spectral distance (fLSD)[1] to evaluate the frequency-dependent power spectral distortion. These metrics were calculated between the target speech signal and two types of speech signals: 1) the quantized speech signal that only suffers from the quantization error $e_t^{(q)}$ and 2) the generated speech signal that suffers from both the quantization error $e_t^{(q)}$ and the prediction error $e_t^{(p)}$, *i.e.* $e_t^{(q)} + e_t^{(p)}$. For the SNR calculation, a linear phase compensation was performed frame by frame by adjusting a sample shift so that a correlation coefficient between the target and the generated speech signals was maximized in the same manner as in the previous work [9]. For the MCD and fLSD calculations, we performed a standard frame-by-frame calculation. The experimental conditions except for the speech dataset were the same as in Section 4.

The SNR and MCD results and the fLSD results are shown in Table 1 and Fig. 1, respectively. Table 1 shows that the noise signals caused by the quantization error are much smaller than those by the prediction error. Therefore, the quantization error is practically negligible compared to the prediction error. In Fig. 1, the quantization error increases fLSD in the high-frequency band (around 5.0 kHz) and decreases it in the low-frequency band (around 200 Hz). This is because the quantization error usually has a flat spectral structure and makes a fLSD curve shape over the frequency axis that resembles the inverse of an averaged spectral structure of the speech signals; *i.e.*, the power of the frequency components of the quantization error often exceeds that of the speech signals in the high-frequency band, which causes a large fLSD, and vice versa in the low-frequency band. Such a tendency is also observed in the noise caused by both the prediction and quantization errors, and fLSD is consistently larger than only in the quantization error. On the other hand, the shape of the fLSD curve becomes flatter than only in the quantization error. This means that the spectral structure of the prediction error of WaveNet is not as flat as the quantization error.

These results imply that audible noise signals are mainly caused by the prediction error and are more easily audible in the high-frequency band than in the low-frequency band.

[1] An averaged value of fLSD over the frequency is equivalent to the log-spectral distance (LSD).

Table 1. SNR and MCD between target and generated speech signals

Errors in noise signal	SNR (dB)	MCD (dB)
$e_t^{(q)}$	33.87 ± 0.38	1.63 ± 0.012
$e_t^{(q)} + e_t^{(p)}$	2.90 ± 0.24	4.12 ± 0.022

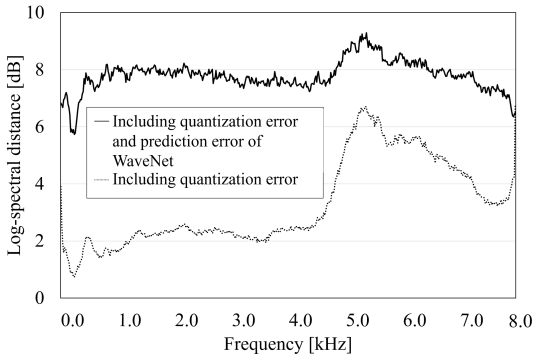

Fig. 1. fLSD between target and generated speech signals

3. PROPOSED NOISE SHAPING METHOD BASED ON PERCEPTUAL WEIGHTING FOR WAVENET

To reduce the adverse effects of the noise signals generated by WaveNet on the sound quality of the generated speech signals, we introduce a perceptual weighting technique to WaveNet based on predictive pulse code modulation (PPCM)[2] [21], which is a well-known technique in analysis-by-synthesis speech coders. An overview of the proposed method is shown in Fig. 2.

3.1. Noise shaping by PPCM

A block diagram of the PPCM is shown in Fig. 3. To quantize current input sample x_t in the encoder part, its value is predicted as $\tilde{x}_t$ from previously reconstructed output samples $\hat{x}_{t-p}, \cdots, \hat{x}_{t-1}$ using a simple linear prediction as the predictor, and then residual signal e_t between x_t and $\tilde{x}_t$ is quantized as $\hat{e}_t$ to be sent to the decoder part. In the decoder part, reconstructed output sample $\hat{x}_t$ can be calculated from quantized residual signal $\hat{e}_t$ and signal $\tilde{x}_t$, predicted from previously reconstructed ones $\hat{x}_{t-p}, \cdots, \hat{x}_{t-1}$, using the same predictor as in the encoder part.

In the PPCM, the predictor uses a time-invariant linear predictive filter whose transfer function usually models the global spectral

[2] It is also known as the differential PCM (DPCM) with a predictor.

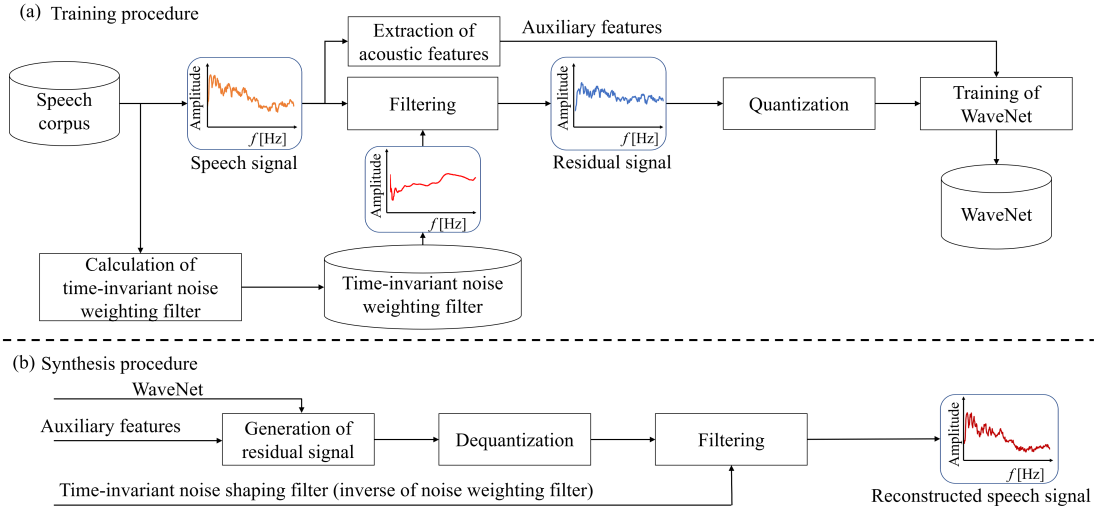

Fig. 2. Overview of proposed method

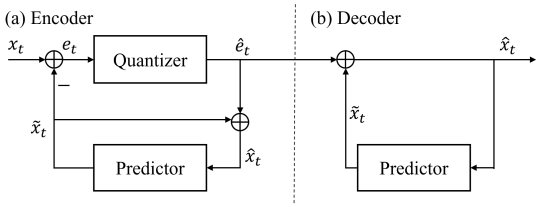

Fig. 3. Block diagram of PPCM

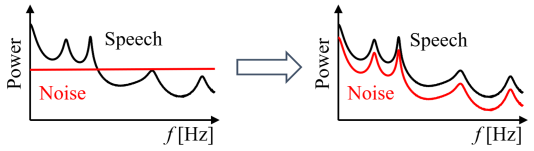

Fig. 4. Effectiveness of noise shaping

structure of the speech signals. On the other hand, the quantizer usually generates error signal $\hat{e}_t - e_t$ that has a flat spectral structure. In such a case, the predictor also works as a noise shaping filter based on perceptual weighting [21] because it makes a spectral structure of the error signal that resembles that of the speech signals, as shown in Fig. 4. Consequently, the filtered noise components are easily masked by the original speech components.

Inspired by this traditional technique, we propose a noise shaping method for WaveNet. In our proposed method, WaveNet is used as the quantizer in the PPCM. In training, residual signal e_t is generated from speech signal s_t using a time-invariant linear filter called a noise weighting filter or a perceptual weighting filter, which is given by an inverse filter of the noise shaping filter. Then the residual signal is quantized into a discrete symbol sequence to be modeled by WaveNet. In a synthesis, WaveNet generates the discrete symbol sequence that corresponds to synthetic residual signal $\hat{e}_t$, which is filtered using a noise shaping filter to reconstruct a synthetic speech signal.

3.2. Design of noise shaping/weighting filters

A mel-generalized cepstrum [22] is used to design both the noise shaping and weighting filters. The transfer function of the noise shaping filter is given by

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

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

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

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad (3)$$

where α denotes the frequency warping parameter.

We calculate the averaged mel-generalized cepstral coefficients in advance over all the frames extracted from the training data and use them as the coefficients of the time-invariant noise shaping filter, which is designed with the mel-log spectrum approximation filter [23]. The time-invariant noise weighting filter is also designed in advance as the inverse filter of the time-invariant noise shaping filter, which is easily derived by multiplying the mel-generalized cepstral coefficients by -1.

As expected from Fig. 1, since the spectral structure of the error signals depends on the prediction accuracy of WaveNet, it is not always flat, as shown in Fig. 4. This situation causes a mismatch between the spectral shape of the weighted noise signals and the desired spectral shape and reduces the auditory masking effect. Adjusting parameter β might alleviate this mismatch. In this paper, we investigate the effect of its setting on the sound quality of the generated speech signals.

3.3. Training and generation

In training the WaveNet vocoder, the acoustic features are extracted from the speech signals in the training data. Then a time-invariant noise weighting filter is designed and applied to the speech signals for generating residual signals, which are quantized using the μ-law companding transformation. Note that the dynamic range of

the residual signals needs to be adjusted so that they are effectively quantized by the μ-law companding. Then the quantized residual signals are modeled by the WaveNet vocoder using the acoustic features as auxiliary features.

In the generation, the quantized residual signal is made by the WaveNet vocoder using the given acoustic features as auxiliary features. Then the time-invariant noise shaping filter is applied to it after the μ-law expansion and the dynamic range adjustment.

4. EXPERIMENTAL EVALUATION

4.1. Experimental conditions

We used a Japanese speech dataset from a male speaker consisting of travel conversations and ATR phonetically balanced sentences [24]. From this dataset, 6,031 sentences (approximately 3.9 hours) were used for training and 25 were used for the test. The sampling frequency was 16 kHz. A high-pass filter was applied to the speech signals to remove frequency components under 50 Hz.

We used the 0th-through-39th mel-cepstral coefficients, log-scaled f_{o}, and a voiced/unvoiced binary symbol as the auxiliary features of the WaveNet vocoder. The f_{o} values were extracted by integrating the results of multiple f_{o} extractors [12, 25, 26] and linearly interpolated during the unvoiced regions. The mel-cepstral coefficients were analyzed by WORLD (D4C edition [27]). The frame shift was set to 5 ms. These features were duplicated from a frame to the samples to match the length of the speech signals for training the WaveNet vocoder.

For the parameters of the noise weighting filter, M_{c}, α, and γ were set to 39, 0.42, and 0.0, respectively, and β was set to 0.1, 0.5, and 1.0. The residual signals generated by applying the noise weighting filter to the speech signals were normalized in a range from -1.0 to 1.0 and quantized by 8-bit μ-law companding. We used a one-hot vector representation to represent the resulting discrete symbol.

We set the network configuration of WaveNet, as previously described [9]. The filter length of the causal convolution was set to 2. Three dilated convolution blocks were used, each of which had ten dilated convolution layers with dilations of 1, 2, 4, and 8 up to 512. Thus, our network formed a total of 30 dilated causal convolution layers, and the receptive field was 192 ms (3,070 samples). Both the numbers of channels of the dilated causal convolution and the 1×1 convolution in the residual block were set to 256. The number of 1×1 convolution channels between the skip-connection and the softmax layer was set to 2,048. Adam [28] was used as the optimizer, the learning coefficient was $1.0e^{-3}$, the batch size was 20 k samples, and the number of iterations was 200 k. We used an Inter Xeon(R) CPU E5-2670 and a single GPU (GeForce GTX 1080) to train the WaveNet vocoder. A fast WaveNet algorithm [29] was used in the generation.

4.2. Experimental result

4.2.1. *Objective evaluation*

To investigate the effectiveness of our proposed noise shaping, the fLSD between the target and reconstructed signals was calculated over all the frames in the test set, as in Section 2.2. Its averaged values are shown in Fig. 5. The increase of β reduces fLSD in the high-frequency band over around 5.0 kHz and increases it in the low-frequency band around 200 Hz, *i.e.*, making the fLSD curve close to flat. This result shows that the noise signals are shaped so that their spectral shape resembles the averaged one of the speech signals. fLSD in the low-frequency band also tends to be larger than that in the high-frequency band if β is set to 1.0 due to the mismatch between the spectral shape of the shaped noise and the desired spectral shape, as described in Section 3.2.

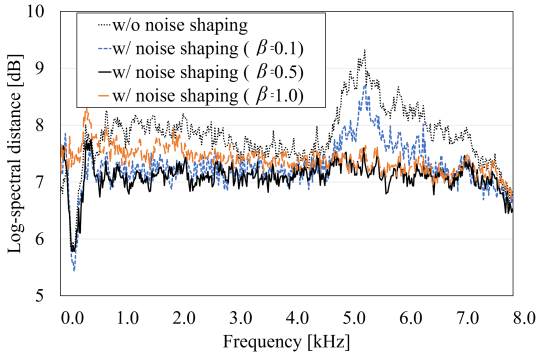

Fig. 5. fLSD between target and reconstructed speech signals

Table 2. Objective evaluation results

Methods	SNR (dB)	LSD (dB)	MCD (dB)
w/o NS	1.79 ± 0.19	10.45 ± 0.47	3.80 ± 0.017
w/ NS (β=0.1)	$\mathbf{2.27 \pm 0.21}$	9.64 ± 0.36	3.52 ± 0.016
w/ NS (β=0.5)	2.02 ± 0.20	8.70 ± 0.28	$\mathbf{3.21 \pm 0.015}$
w/ NS (β=1.0)	1.59 ± 0.18	$\mathbf{8.66 \pm 0.29}$	3.34 ± 0.017

Table 3. Preference score (%). "np" denotes no preference. Significant difference at $p < 0.01$ level is shown in bold

w/ NS	np	w/o NS	p-value	z-score
45.3	30.8	23.7	$< 10^{-6}$	0.29

We also calculated SNR, LSD, and MCD in the test set, as in Section 2.2. The results are shown in Table 2. The 95% confidence interval is also shown in the table (mean and lower/upper bound).[3] The proposed noise shaping (w/ NS) effectively improves these distance metrics.

4.2.2. *Subjective evaluation*

We conducted a preference test on the naturalness of synthesized speech to compare the conventional method without noise shaping and the proposed method with noise shaping where β was set to 0.5 based on an informal listening test. After our 15 subjects listened to each pair of samples, they made a decision, but they were allowed to choose "neutral" if they had difficulty.

Table 3 shows the result. The proposed method (w/ NS) significantly outperforms the conventional method (w/o NS).

5. CONCLUSION

We analyzed the statistical characteristics of the noise signals generated in WaveNet-based speech generation and proposed a noise shaping method based on perceptual weighting. Our experimental results demonstrated that the proposed method effectively improved the naturalness of the generated speech signals by successfully using the noise masking properties of the human auditory system. Since the experimental setting remains limited, we plan to conduct more investigations in various conditions, such as multiple speakers [10], various sizes of training data, and various applications [30–33].

[3]The SNR and MCD values are different from those in Table 1 since the datasets used in the experiments were different.

6. REFERENCES

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

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

[3] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," *in Proc. ICASSP*, pp. 7962–7966, 2013.

[4] 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, 2016.

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

[6] H. Zen, Y. Agiomyrgiannakis, N. Egberts, F. Henderson, and P. Szczepaniak, "Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices," *in Proc. Interspeech*, pp. 2273–2277, 2016.

[7] X. Gonzalvo, S. Tazari, C. Chan, M. Becker, A. Gutkin, and H. Silen, "Recent advances in Google real-time HMM-driven unit selection synthesizer," *in Proc. Interspeech*, pp. 2238–2242, 2016.

[8] V. Wan, Y. Agiomyrgiannakis, H. Silen, and J. Vit, "Google's next-generation real-time unit-selection synthesizer using sequence-to-sequence LSTM-based autoencoders," *in Proc. Interspeech*, pp. 1143–1147, 2017.

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

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

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

[12] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

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

[14] J. Sotelo, S. Mehri, K. Kumar, J. Felipe Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2wav: end-to-end speech synthesis," *in Proc. ICLR*, https://openreview.net/forum?id=B1VWyySKx, 2017.

[15] S. 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," *arXiv preprint*, arXiv:1702.07825, 2017.

[16] S. 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," *arXiv preprint*, arXiv:1705.08947, 2017.

[17] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *The Journal of the Acoustical Society of America*, vol. 66, no. 6, pp. 1647–1652, 1979.

[18] P. Kroon and B. S. Atal, "Predictive coding of speech using analysis by-synthesis techniques," *Advances in Speech Signal Processing*, pp. 141–164, 1992.

[19] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with PixelCNN decoders," *arXiv preprint*, arXiv:1606.05328, 2016.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *in Proc. CVPR*, pp. 770–778, 2016.

[21] B. S. Atal and M. R. Scröder, "Predictive coding of speech signals and subjective error criteria," *in Proc. ICASSP*, pp. 247–254, 1978.

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

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

[24] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[25] A. Camacho, "SWIPE: A sawtooth waveform inspired pitch estimator for speech and music," *Ph.D. Thesis*, University of Florida, 2007.

[26] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *in Speech Coding & Synthesis*, W.B. Kleijn and K.K. Pailwal (Eds.), Elsevier, pp. 495–518, 1995.

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

[28] D. Kingma and J. Ba, "Adam: a method for stochastic optimization," *arXiv preprint*, arXiv:1412.6980, 2014.

[29] T. Le Paine, P. Khorrami, S. Chang, Y. Zhang, P. Ramachandran, M.A. Hasegawa-Johnson, and T.S. Huang, "Fast Wavenet generation algorithm," *arXiv preprint*, arXiv:1611.09482, 2016.

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

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

[32] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florencio, and M. Hasegawa-Johnson, "Speech Enhancement Using Bayesian Wavenet," *in Proc. Interspeech*, pp. 2013–2017, 2017.

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