INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Speaker-dependent WaveNet vocoder

Akira Tamamori[1], *Tomoki Hayashi*[2], *Kazuhiro Kobayashi*[3], *Kazuya Takeda*[2], *Tomoki Toda*[3]

[1]Institute of Innovation for Future Society, Nagoya University, Japan
[2]Graduate School of Information Science, Nagoya University, Japan
[3]Information Technology Center, Nagoya University, Japan

{tamamori, hayashi.tomoki, kazuhiro.kobayashi}@g.sp.m.is.nagoya-u.ac.jp,
takeda@is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp

Abstract

In this study, we propose a speaker-dependent WaveNet vocoder, a method of synthesizing speech waveforms with WaveNet, by utilizing acoustic features from existing vocoder as auxiliary features of WaveNet. It is expected that WaveNet can learn a sample-by-sample correspondence between speech waveform and acoustic features. The advantage of the proposed method is that it does not require (1) explicit modeling of excitation signals and (2) various assumptions, which are based on prior knowledge specific to speech. We conducted both subjective and objective evaluation experiments on CMU-ARCTIC database. From the results of the objective evaluation, it was demonstrated that the proposed method could generate high-quality speech with phase information recovered, which was lost by a mel-cepstrum vocoder. From the results of the subjective evaluation, it was demonstrated that the sound quality of the proposed method was significantly improved from mel-cepstrum vocoder, and the proposed method could capture source excitation information more accurately.

Index Terms: WaveNet, convolutional neural network, vocoder, deep neural network

1. Introduction

In recent years, the demand for speech waveform synthesis technique by computer is increasing. Speech waveform synthesis is an indispensable basic technique for applications with speech interface such as smartphone, car navigation system, speech translation and screen readers, etc. Various demands are also occurring, such as stable and easy to hear, not mechanical but highly naturally sounded human speech, control of speech speed and voice characteristics, multiple speakers, multiple languages, emotional speech, etc. Expectations for high-quality and diverse speech waveform synthesis technique will be increasing more and more in future.

Various speech waveform synthesis techniques have been proposed so far. One of the representative techniques is concatenative synthesis [1, 2]. In this method, many speech waveforms are divided into fine fragments such as phoneme and syllable units and then stored in a database beforehand. At synthesis phase, an optimum segment sequence to given text is extracted and combined. Since speech units are directly connected, it is advantageous that high-quality speech with high clarity can be obtained. However, it is often difficult to generate various speech with voice characteristics changed flexibly. Another technique is parametric waveform synthesis, focusing on speech generation process. The speech waveform can be synthesized from acoustic features; they represents sound source and vocal tract characteristics. The system which implements this process as a digital filter is often called as vocoder [3]. Various vocoders have been proposed so far [4–11], and they are imposed on many assumptions based on prior knowledge specific to speech [12]; e.g., fixed length of analysis window, time-invariant linear filter, stationary Gaussian process, etc. Moreover, since vocoders are accompanied by modeling source excitation signals and detailed temporal, phase information of the original speech will be lost. While the introduction of these assumptions simplifies mathematical formulation and has an advantage of implementation, the sound quality of the synthesized speech will be more or less deteriorated.

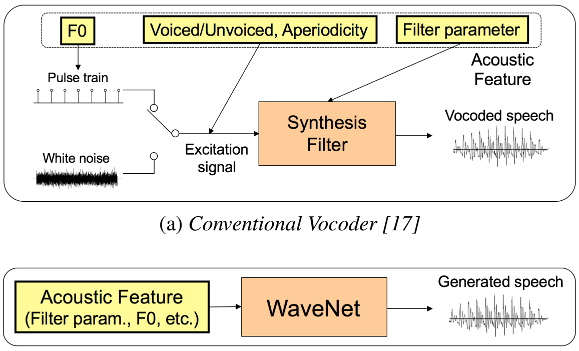
Following these studies, a neural network called WaveNet, which directly generates speech waveforms without vocoder [13], has been proposed. The WaveNet will be briefly reviewed in Section 2. One of its features is that it does not depend on the characteristics of the data to be applied, and can build a generative model in a data-driven manner. In the case of speech, various assumptions based on the prior knowledge specific to speech can be avoided. In the original literature [13], it was applied to text-to-speech (TTS), and the quality of synthesized speech exceeded that of state-of-the-art approaches [14, 15]. The input to the WaveNet was the linguistic feature, the fundamental frequency (f_o), and the phoneme duration, except for the waveform samples that it generated in the past. However, it was not specifically clarified what kind of features works effectively other than those.

Our aim is realization of a new vocoder which resolves various constraints imposed on existing vocoders and synthesizes high quality speech simultaneously. In this study, we propose a method that uses the acoustic features of existing vocoders as auxiliary features of WaveNet, which will be described in Section 3. Figure 1 shows the difference in speech waveform generation in conventional vocoder and proposed method. The proposed method does not require various assumptions imposed on existing vocoders. In particular, since the proposed method does not involve modeling a source excitation signal and driving an articulate filter required for the existing vocoders, it is expected to synthesize a high quality speech waveform with detailed temporal structure and with phase information recovered.

In this paper, we conducted both objective and subjective evaluations, which will be described in Section 4. From the results of objective evaluation, it was demonstrated that the proposed method could generate high quality speech with phase information recovered, which was lost by a mel-cepstrum vocoder. Moreover, from the results of subjective evaluation, it was demonstrated that the sound quality of the proposed method was significantly improved from mel-cepstrum vocoder, and the proposed method could capture source excitation information more accurately.

This paper is an extended version of our earlier work [16]. We conducted further additional objective evaluation experiments. We also newly conducted subjective evaluation experiments and added some discussions to these results.

http://dx.doi.org/10.21437/Interspeech.2017-314



(a) *Conventional Vocoder [17]*

(b) *Proposed*

Figure 1: *Difference in waveform generation between conventional vocoder and proposed method [16]*

2. WaveNet

WaveNet [13] is a neural network which directly generates audio signal. The input to the network is a sequence of waveform samples. WaveNet mainly consists of a stack of one dimensional convolution layers called dilated causal convolution layer. The input passes through these convolution layers and gated activation functions, and finally the softmax function outputs the posterior probability of waveform sample value encoded by μ-law algorithm [18]. The concrete form of the gated activation function is given by following equation:

$$\boldsymbol{z} = \tanh(W_f * \boldsymbol{x}) \odot \sigma(W_g * \boldsymbol{x}), \qquad (1)$$

where $\boldsymbol{x}$ and $\boldsymbol{z}$ is the input and output to the activation, respectively. The symbol $*$ is a convolution operator and the symbol $\odot$ is an element-wise product operator. $\sigma(\cdot)$ represents a sigmoid function, and W represents a convolution weight. The subscripts f and g represent a filter and a gate, respectively.

From a viewpoint of generative model, joint probability of waveform sample points $\boldsymbol{x} = \{x_1, x_2, \ldots, x_T\}$ can be written as

$$P(\boldsymbol{x} \mid \lambda) = \prod_{t=1}^{T} P(x_t \mid x_1, \ldots, x_{t-1}, \lambda), \qquad (2)$$

given model parameters λ. It can be regarded that WaveNet approximately calculates the above joint probability by repetition of linear operation by the causal convolution considering past waveform samples and nonlinear operation by the gated activation function. Here the model parameters corresponds to the network parameters of WaveNet. Synthesis of speech waveform can be performed by repetition of sampling from Eq. (2) a desired number of times. In this case the input to the network is the waveform samples that it generated in the past.

3. Speaker-dependent WaveNet vocoder

Various assumptions are imposed on the conventional vocoder as shown in Fig 1(a). First, each sample point in natural speech waveform follows a non-stationary process. However in fact, speech analysis is done based on the assumption of stationary process in the analysis window. The articulate filter is realized as a time-invariant linear filter assuming the stationary process. The filter is driven by excitation signals and the excitation itself is modeled under a assumption; for instance, it can be represented by pulse and white noise, or those with aperiodic components (mixed excitation). The acoustic features are often extracted based on the assumption that speech is distributed under a Gaussian process. However this assumption does not always hold in real speech. Since these assumptions are combined in existing vocoder, the detailed time structure such phase information in the original speech will be lost more or less, and the generated speech loses naturalness or clarity compared to the original.

In this study, a new speech waveform synthesis method with WaveNet is proposed. It uses the acoustic features of existing vocoder as auxiliary features of WaveNet. Since the acoustic features are extracted by considering the speech generation process, it is expected that correspondence between the speech waveform and the acoustic features is built in the network automatically, considering the physical restrictions on the speech generation process. Furthermore, the proposed method does not involve driving an articulatory filter by the excitation signals, and any mathematical assumptions to the data such as Gaussianity are not also required. Therefore it is expected that high quality speech can be synthesized which recovers detailed time information lost by various existing vocoders. In the next and subsequent sections, the concrete formulation of the proposed method will be described.

3.1. Formulation

Here we extend the Eq.(2) and consider the distribution conditioned by an additional variable $\boldsymbol{h}$ as a new target of modeling, where $\boldsymbol{h}$ represents the auxiliary features. Then all the gated activation functions shown in the network are modified as follows:

$$\boldsymbol{z} = \tanh(W_f * \boldsymbol{x} + V_f * \boldsymbol{y}) \odot \sigma(W_g * \boldsymbol{x} + V_g * \boldsymbol{y}), \qquad (3)$$

where V_f is the convolution weight for the auxiliary features, $V_f * \boldsymbol{y}$ and $V_g * \boldsymbol{y}$ represents 1×1 convolution calculation. The variable $\boldsymbol{y}$ is an extended time series of the original auxiliary features $\boldsymbol{h}$ where the time resolution of $\boldsymbol{h}$ is adjusted to $\boldsymbol{x}$. In the next section we will describe how to adjust the time resolution, which was adopted in this study.

3.2. Time resolution adjustment

Acoustic features of a vocoder are extracted from the windowed speech waveform. A series of the feature vectors can be obtained by shifting the analysis window at regular intervals along the time axis. The length of the series is generally shorter than the original speech. When they are used as auxiliary features for WaveNet, it is necessary to match the sequence length between the feature sequence and the speech waveform. In order to adjust time resolution, the authors of the original WaveNet paper adopted to use a transposed convolutional network and other authors of the Deep Voice 2 paper [19] applied a stack of bidirectional quasi-recurrent neural networks. In this study, we adopted a relatively simple method to match both sequence lengths of $\boldsymbol{x}$ and $\boldsymbol{h}$, which copies the feature vector of each frame by the shift amount of the analysis window. In other words, the original feature vectors $\boldsymbol{h}$ will be extended along the time axis in advance (see Fig. 2). This can be viewed as a special form of the transposed convolution.

4. Experimental evaluation

We conducted experiments for objective evaluation and subjective evaluation. In this study, mel-cepstrum and fundamental frequency were adopted as the auxiliary features of WaveNet, which are acoustic features from existing mel-cepstrum vocoder. As listed in the table 1, we varied the extraction method of the mel-cepstrum and synthesis method of speech waveform.

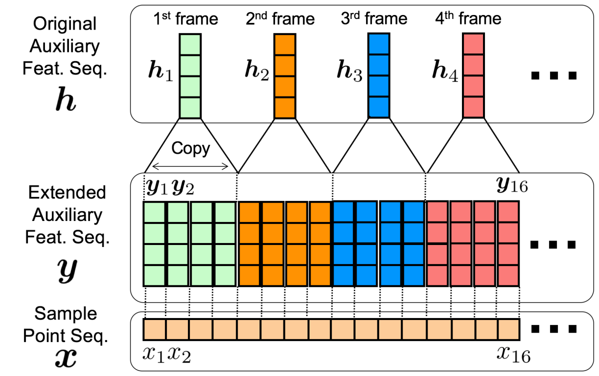


Figure 2: *Time resolution adjustment of auxiliary features; frame shift is 4 points in the figure [16].*

Table 1: *Comparative methods of waveform synthesis; spectrum envelop was extracted by STRAIGHT analysis.*

Comparative Method	**Source of mel-cepstrum**	**Waveform Synthesis**
Plain-MLSA	STFT	MLSA filter
STRAIGHT-MLSA	Spectrum envelop	MLSA filter
Plain-WaveNet	STFT	WaveNet
STRAIGHT-WaveNet	Spectrum envelop	WaveNet

4.1. Experimental conditions

We used speech data of four speakers in CMU-ARCTIC database [20]; slt, bdl, clb, and rms, where slt and clb are female and bdl and rms are male. The sampling frequency is set to 16kHz. The total number of utterances is 1,132 per speaker, and the total utterance duration is about 1 hour per speaker. For each speaker, the 1,082 utterances were used for training the speaker dependent network and the remaining 50 utterances were used for evaluation. In this study, mel-cepstrum and fundamental frequency (f_o) were used as vocoder parameters. The 0-24 th mel-cepstrum sequences were extracted from the sequence of short time Fourier Transform of speech, or from the smoothed spectrum obtained by STRAIGHT analysis. The f_o sequences were extracted by RAPT algorithm [21]. The length of analysis window was 25 ms and the amount of window shift was 5 ms.

Considering one layer of dilated causal convolution, gate activated function, and residual as one block, we connected the 30 residual blocks in total. Specifically, dilations in 10 layers were set to 2^0, 2^1, 2^2, ..., 2^9, and this was repeated three times to form a total of 30 dilated causal convolution layers. The number of channels of (dilated) causal convolution and 1×1 convolution in the residual block were set to 256. The number of 1×1 convolution channel between skip-connection and softmax layer was set to 2,048. Adam algorithm [22] was used for optimization, and its learning rate was manually adjusted to 0.001 as an initial value, and an attenuation schedule was adjusted. The mini-batch size was 20,000 samples and the number of parameter update was 300,000. We used Xeon(R) E5-2650 and a single GPU of TITAN X (Pascal). It takes about 2 days to train WaveNet for one speaker, and 6 minutes to synthesize a speech of 3 sec from it.

4.2. Objective evaluation

We evaluated distortion between the original speech and synthesized speech. The comparative methods listed in table 1 are distinguished by the source of mel-cepstrum and synthesis method. For the objective evaluation, we first applied the following SNR and RMSE for each frame:

$$SNR \quad = \quad 10 \ln_{10} \left(\frac{\sum_{n=1}^{N} y(n)^2}{\sum_{n=1}^{N} \left(x(n) - y(n)\right)^2} \right), \quad (4)$$

$$RMSE \quad = \quad \sqrt{\frac{1}{F} \sum_{f=1}^{F} \left(20 \log_{10} \frac{|Y(f)|}{|X(f)|} \right)^2}, \quad (5)$$

where $x(n)$ and $y(n)$ represents the windowed synthesized speech and natural speech at position n in a frame, respectively. Also, $X(f)$ and $Y(f)$ represents the short-time Fourier transform of synthesized speech and natural speech in a frame, frequency bin f, respectively. $| \cdot |$ represents an operator for absolute value. L is the total number of frames, N is the frame length, and F is the total number of frequency bins. For both SNR and RMSE, a linear phase compensation was performed in advance for each frame. That is, a time shift between ±200 points that maximizes the correlation between raw and synthesized speech was calculated for each frame, and applied it to the windowed synthesized speech. The above SNR and RMSE were calculated for each frame and averaged over total frames.

The SNR and RMSE for each speaker and comparative method are shown in table 2, where "P" indicates "Plain" and "ST" indicates "STRAIGHT" in accordance with table 1. Each numerical value in the table represents the 95% confidence interval (mean and lower/upper bound). It can be seen that WaveNet vocoder could improve SNR. On the other hand, RMSE could not be improved, which was especially the case in 'slt' or 'clb'. We also tested the statistical significance of improvement. From the test at significance level 5%, we confirmed that both SNRs of "Plain-WaveNet" and "STRAIGHT-WaveNet" were improved significantly for all speakers.

Next, in order to confirm whether WaveNet can reproduce the characteristics of original speech, we applied STRAIGHT analysis to the synthesized speech and extracted mel-cepstrum, and RAPT algorithm to extract f_o. Then we evaluated objectively the distortion of the acoustic features between raw and synthesized speech. For mel-cepstrum, following Mel-Cesptrum Distance (MCD) was applied:

$$MCD = \frac{10}{\log 10} \sqrt{2 \sum_{m=1}^{M} (c_r(m) - c_s(m))^2}, \quad (6)$$

where c_r and c_s is mel-cepstrum from raw and synthesized speech, respectively, and M is order of mel-cepstrum. For f_o, following RMSE were applied:

$$RMSE(f_o) \quad = \quad 1200 \sqrt{(\log_2(F_r) - \log_2(F_s))^2}, \quad (7)$$

where the subscript r and s represents raw and synthesized speech, respectively. The above MCD and RMSE were calculated for each frame and averaged over total frames.

The distortions between acoustic features from natural and synthesized speech are shown in table 3. First, from table 3(a), it can be seen that MCDs of the proposed method were deteriorated from mel-cepstrum vocoder, which means it could not reproduce the original spectrum. Next, from table 3(b), it can be seen that "Plain-WaveNet" could reproduce the original f_o with the relatively higher accuracy than mel-cepstrum vocoder. Finally, table 4 lists unvoiced/voiced (U/V) decision errors. This error is the ratio of the number of unmatched U/V frames between natural and synthesized speech to total frames. It was demonstrated that the proposed method could capture the U/V information with relatively higher accuracy except slt, compared to the mel-cepstrum vocoder.

Table 2: *Comparison of distortion between natural speech and synthesized speech*

(a) *SNR (dB); distortion in time domain*

Method	slt	bdl	clb	rms
MLSA (P)	−0.24 ± 0.31	−2.7 ± 0.19	−0.044 ± 0.35	−2.2 ± 0.52
MLSA (ST)	3.7 ± 0.32	−2.6 ± 0.16	−1.9 ± 0.31	−2.3 ± 0.45
WaveNet (P)	**4.1 ± 0.23**	**3.6 ± 0.21**	**3.8 ± 0.38**	**4.0 ± 1.0**
WaveNet (ST)	3.7 ± 0.32	2.2 ± 0.28	3.7 ± 0.32	2.6 ± 0.94

(b) *RMSE (dB); distortion in frequency domain*

Method	slt	bdl	clb	rms
MLSA (P)	**7.9 ± 0.13**	**7.9 ± 0.21**	**7.8 ± 0.23**	**8.1 ± 0.97**
MLSA (ST)	8.3 ± 0.31	8.6 ± 0.48	7.9 ± 0.43	8.4 ± 0.53
WaveNet (P)	8.8 ± 0.21	8.6 ± 0.21	9.2 ± 0.30	9.0 ± 1.3
WaveNet (ST)	9.0 ± 0.35	9.4 ± 0.30	9.1 ± 0.28	9.5 ± 1.3

Table 3: *Comparison of distortion between acoustic features of natural speech and synthesized speech*

(a) *Mel-cepstrum (MCD; dB)*

Method	slt	bdl	clb	rms
MLSA (P)	3.8 ± 0.027	3.8 ± 0.050	4.6 ± 0.050	3.6 ± 0.054
MLSA (ST)	**2.4 ± 0.047**	**2.3 ± 0.054**	**2.5 ± 0.049**	**2.5 ± 0.059**
WaveNet (P)	5.5 ± 0.052	5.5 ± 0.050	6.8 ± 0.11	4.9 ± 0.053
WaveNet (ST)	5.7 ± 0.045	5.7 ± 0.053	6.8 ± 0.045	5.1 ± 0.052

(b) *Fundamental frequency (RMSE; cent)*

Method	slt	bdl	clb	rms
MLSA (P)	2.9 ± 0.21	9.4 ± 1.6	2.4 ± 0.19	6.4 ± 0.63
MLSA (ST)	2.7 ± 0.18	8.7 ± 1.6	2.1 ± 0.13	6.2 ± 0.79
WaveNet (P)	**1.9 ± 0.22**	**7.5 ± 1.6**	**1.1 ± 0.087**	**3.7 ± 1.4**
WaveNet (ST)	2.3 ± 0.13	9.7 ± 2.0	1.1 ± 0.13	5.6 ± 1.5

Table 4: *Comparison of voiced/unvoiced decision error (%)*

Method	slt	bdl	clb	rms
MLSA (P)	**1.7 ± 0.19**	4.8 ± 0.40	2.5 ± 0.27	4.8 ± 0.44
MLSA (ST)	1.7 ± 0.19	5.0 ± 0.44	2.3 ± 0.23	4.7 ± 0.47
WaveNet (P)	2.1 ± 0.32	**2.9 ± 0.29**	**1.7 ± 0.28**	**3.1 ± 0.31**
WaveNet (ST)	3.1 ± 0.39	4.2 ± 0.41	1.8 ± 0.39	4.1 ± 0.48

4.3. Subjective evaluation

We evaluated the sound quality of the synthesized speech using a mean opinion score (MOS). The subjects rated the quality of the synthesized speech using a 5-point scale: “5” for excellent, “4” for good, “3” for fair, “2” for poor, and “1” for bad. The number of evaluation sentences in each subject was 100; 25 sentences × 4 methods. The number of subjects was 15 and they are all non-native speakers of English.

Figure 3 indicates the results of the MOS test for sound quality. The error bar represents 95% confidence interval. It can be seen that the proposed method could generate high quality speech compared to the mel-cepstrum vocoder. The significant difference between “Plain-WaveNet” and “STRAIGHT-WaveNet” could not be confirmed for the averaged scores. This result suggests that the proposed method can compensate the degradation of the acoustic feature caused by applying STFT spectrum as a source of mel-cepstrum. While the synthesized speech from male speakers (bdl and rms) achieved high quality, that from female speakers (slt and clb) achieved relatively low quality. We have already confirmed that, the noise contained in the synthesized speech, which is caused by the waveform prediction error of WaveNet, was relatively and strongly perceived for the two female speakers. It is considered that this noise affected the evaluation scores.

5. Conclusions

In this study, we proposed speaker-dependent WaveNet vocoder which utilizes acoustic features of an existing vocoder as auxiliary features of WaveNet. The advantage of the proposed method is that it does not require explicit modeling of excitation signals and various assumptions specific to speech generation process and speech analysis. The experimental results demonstrated that the proposed method could recover phase information which was lost by existing mel-cepstrum vocoder. It was also demonstrated that the sound quality of the proposed method was significantly improved compared to mel-cepstrum vocoder, and the proposed method could capture source excitation information more accurately.

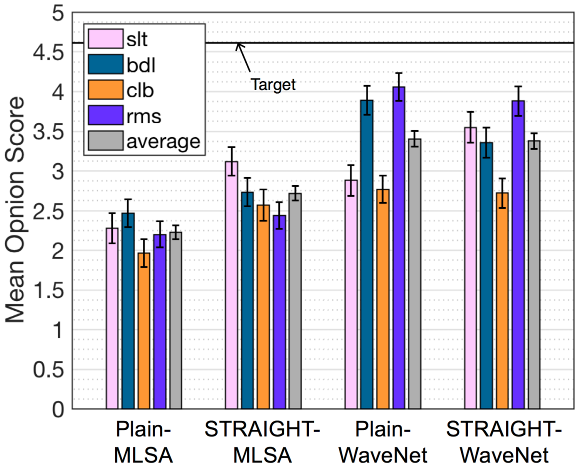


Figure 3: *Sound quality of synthesized speech*

The original WaveNet paper describes that the speaker dependent WaveNet for TTS was trained by using speech data over 20 hours and successfully synthesized high-quality speech close to natural one. Although detailed information about training time for the network and the author’s implementation are not described in the original paper, it can be supposed that abundant computer resources were required to carry out the training. In this paper, we have shown the reference of the achieved sound quality when using a small amount of speech data about 1 hour per speaker and limited computer resources. We expect that this reference will be a beneficial information for researchers and practitioners in related research fields who plan to adopt the WaveNet vocoder and incorporate it into their own applications.

Future works include additional investigations on the effectiveness of the proposed method when target vocoder and corresponding acoustic features are changed. Moreover, development of a technique to alleviate the noise caused by prediction error of WaveNet will be also a future work.

6. Acknowledgements

This research is partially supported by the Center of Innovation Program (Nagoya-COI) from Japan Science and Technology Agency, and by JSPS KAKENHI Grant Number 15H02726.

7. References

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

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

[3] H. Dudley, “The vocoder,” *Bell Labs Record*, vol. 18, no. 4, pp. 122–126, 1939.

[4] J. Flanagan, “Phase vocoder,” *Bell System Technical Journal*, vol. 45, no. 9, pp. 1493–1509, 1966.

[5] B. Gold and C. Rader, “The channel vocoder,” *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 4, pp. 148–161, Dec 1967.

[6] A. Oppenheim, “Speech analysis-synthesis based on homomorphic filtering,” *The Journal of the Acoustical Society of America*, vol. 45, no. 2, pp. 458–465, 1969.

[7] F. Itakura, “Line spectrum representation of linear predictor coefficients of speech signals,” *The Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S35–S35, 1975.

[8] S. Imai and C. Furuichi, “Unbiased estimation of log spectrum,” in *EURASIP*, 1988, pp. 203–206.

[9] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, “Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation,” in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2008, pp. 3933–3936.

[10] D. Erro, I. Sainz, E. Navas, and I. Hernaez, “Harmonics plus noise model based vocoder for statistical parametric speech synthesis,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 184–194, April 2014.

[11] 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.

[12] S. Furui, *Digital Speech Processing: Synthesis, and Recognition*, ser. Signal Processing and Communications. Taylor & Francis, 2000.

[13] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, “WaveNet: A Generative Model for Raw Audio,” http://arxiv.org/abs/1609.03499, 2016.

[14] X. Gonzalvo, S. Tazari, C. an Chan, M. Becker, A. Gutkin, and H. Silen, “Recent advances in google real-time hmm-driven unit selection synthesizer,” in *INTERSPEECH 2016*, Sep 8-12, San Francisco, USA, 2016, pp. 2238–2242.

[15] 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 *INTERSPEECH 2016*, San Francisco, CA, USA, 2016, pp. 2273–2277.

[16] A. Tamamori, T. Hayashi, T. Toda, and K. Takeda, “A method of speech waveform synthesis based on WaveNet considering speech generation process (Japanese edition),” *Technical Report of The Institute of Electronics, Information and Communication Engineers (IEICE)*, no. 475, pp. 1–6, 2017.

[17] HTS Working Group, “HTS Slides ver. 2.3,” http://hts.sp.nitech.ac.jp/.

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

[19] 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,” https://arxiv.org/abs/1705.08947, 2017.

[20] “CMU-ARCTIC Speech Synthesis Databases,” http://festvox.org/cmu_arctic/index.html.

[21] D. Talkin, “A robust algorithm for pitch tracking (RAPT),” *Speech coding and synthesis*, pp. 495–518, 1995.

[22] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2014.