Speech Synthesis Using WaveNet Vocoder Based on Periodic/Aperiodic Decomposition

Takato Fujimoto*, Takenori Yoshimura*, Kei Hashimoto*,
Keiichiro Oura*, Yoshihiko Nankaku*, and Keiichi Tokuda*
* Department of Computer Science, Nagoya Institute of Technology, Nagoya, Japan
E-mail: {taka19, takenori, bonanza, uratec, nankaku, tokuda}@sp.nitech.ac.jp Tel: +81-52-735-5479

***Abstract*—This paper proposes speech synthesis using a WaveNet vocoder based on periodic/aperiodic decomposition. Normally, quasiperiodic and aperiodic components are contained in human speech waveforms. Therefore, it is important to accurately model periodic and aperiodic components. Periodic and aperiodic components are represented as the ratios of the energies in conventional statistical parametric speech synthesis. On the other hand, statistical parametric speech synthesis based on periodic/aperiodic decomposition has been proposed. Although the effectiveness of this approach has been shown, speech waveforms considering both periodic and aperiodic components cannot be generated directly. In this paper, we propose speech synthesis using a WaveNet vocoder based on periodic/aperiodic decomposition. In the proposed approach, separated periodic and aperiodic components are modeled by a single acoustic model based on deep neural networks, and then speech waveforms considering both periodic and aperiodic components are directly generated by a single WaveNet vocoder based on neural networks. Experimental results show that the proposed approach outperforms the conventional approach in the naturalness of the synthesized speech.**

I. Introduction

Statistical parametric speech synthesis [1] has grown in popularity over the last decade. In statistical parametric speech synthesis, the relationship between acoustic features and linguistic features is modeled by statistical models, which are generally called acoustic models. This approach has several advantages over concatenative speech synthesis [2], such as the flexibility to change voice characteristics [3]–[6], a reduced memory footprint [7]–[9], and robustness [10]. However, its major flaw is the quality of the synthesized speech. To improve synthesized speech quality, high-quality vocoders that can generate natural speech waveforms from acoustic features output by acoustic models and acoustic models that can predict acoustic features from linguistic features accurately are required.

Some vocoders (e.g., a simple mel-log spectrum approximate (MLSA) filter [11] based on mel-cepstrum [12] and high-quality ones such as STRAIGHT [13] and WORLD [14]) have been proposed. The vocoders are based on a source-filter model, which assumes that the characteristics of a vocal tract acoustic have no significant effect on the vibration of the vocal folds. Neural vocoders, which are neural networks modeling speech waveforms, have been recently proposed. A neural vocoder can be trained without assumptions based on prior knowledge of specific speech and can recover phase information. A WaveNet vocoder [15] is a neural vocoder that is a waveform generator that uses the acoustic features of existing vocoders as auxiliary features of WaveNet. WaveNet directly models audio waveforms and the predicted waveform samples are used to predict the next sample. Therefore, a WaveNet vocoder can recover phase information and detailed temporal structures. Consequently, the quality of speech synthesized by a WaveNet vocoder is significantly better than that of conventional vocoders.

Recently, statistical parametric speech synthesis based on deep neural networks (DNNs) [17] is one of the major approaches. In the training for DNN-based speech synthesis, a single DNN is trained to represent a mapping function from linguistic features to acoustic features. In the generation process of DNN-based speech synthesis, linguistic features extracted from a given text to be synthesized are mapped to acoustic features by the trained DNN. The DNN-based acoustic models can predict acoustic features accurately because DNNs can represent complex mapping functions from input features to output features.

In the human speech production process, vocal source signals containing quasiperiodic and aperiodic components are generated by vocal fold vibration and turbulent noise, respectively. This is particularly obvious in voiced fricatives, breathy voices, etc. Therefore, it is important to accurately model periodic and aperiodic components.

In conventional statistical parametric speech synthesis, periodic and aperiodic components are represented as the ratios of the energies; e.g., aperiodicity measures [18] and harmonics-to-noise ratio [19]. The aperiodicity parameters are used to generate a mixed excitation signal consisting of periodic and aperiodic components. The mixed excitation signal is formed by weighting the pulse sequence and white noise using an aperiodicity parameter. Statistical parametric speech synthesis based on periodic/aperiodic decomposition has been proposed in [20]. In this approach, periodic and aperiodic speech signals are independently generated by a vocoder, and then, the final speech signals are generated by adding the generated speech signals representing periodic and aperiodic signals. Therefore, this approach cannot directly generate speech waveforms considering periodic components and aperiodic components.

In this paper, we propose speech synthesis using a WaveNet vocoder based on periodic/aperiodic decomposition. The separated periodic and aperiodic components are modeled by the periodicity and aperiodicity based on spectral parameters. The periodicity based on spectral parameters that include the spectral envelope of the periodic voiced source, and the aperiodicity based on spectral parameters that include the spectral envelope of the aperiodic noise source, are predicted by the single DNN-based acoustic model. Then, speech waveforms are generated by the single WaveNet vocoder. The input features of the WaveNet vocoder include the predicted periodicity and aperiodicity based on spectral parameters.

The rest of this paper is organized as follows. Section 2 describes speech synthesis based on DNNs, Section 3 describes the proposed speech synthesis using a WaveNet vocoder based on periodic/aperiodic decomposition. Section 4 tells the experimental conditions and results. Section 5 presents the concluding remarks and future work.

II. Speech synthesis based on deep neural networks

In statistical parametric speech synthesis, the relationship between linguistic features and acoustic features is modeled by statistical models, which are generally called acoustic models. It has been shown that deep neural networks (DNNs) improve the performance of speech synthesis [17]. A single DNN is trained to represent a complex mapping function from linguistic features to acoustic features consisting of spectral and excitation parameters and their dynamic features. The weights of the DNN are optimized by minimizing the mean squared error between the output features of the training data $\boldsymbol{o}_t$ and predicted features $\hat{\boldsymbol{o}}_{\boldsymbol{\lambda},t}$ as follows:

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \frac{1}{2}\sum_{t=1}^{T} \|\boldsymbol{o}_t - \hat{\boldsymbol{o}}_{\boldsymbol{\lambda},t}\|^2, \quad (1)$$

where $\boldsymbol{\lambda}$ is a parameter set of the DNN. Assuming that outputs of a neural network are used as mean vectors and that an identity matrix is used as the covariance matrix independent of linguistic features, (1) is equivalent to maximize the likelihood defined by a Gaussian distribution with the mean vector $\tilde{\boldsymbol{\mu}}_t$ and the covariance matrix $\tilde{\boldsymbol{\Sigma}}$.

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{o}_t | \tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}), \quad (2)$$

where $\tilde{\boldsymbol{\mu}}_t = \hat{\boldsymbol{o}}_{\boldsymbol{\lambda},t}$. In the generation process, input features are mapped to output features by the trained DNN using forward propagation. Then, a smooth speech parameter sequence is generated from the output feature sequence by the maximum likelihood parameter generation (MLPG) algorithm [23]. Synthesized speech is output by putting the generated speech parameters into a vocoder.

III. Speech synthesis using WaveNet vocoder based on periodic/aperiodic decomposition

Speech signals contain periodic and aperiodic components. It is important to accurately model the periodic and aperiodic components for generating natural-sounding speech in speech synthesis. In this paper, we propose speech synthesis using a WaveNet vocoder based on periodic and aperiodic decomposition.

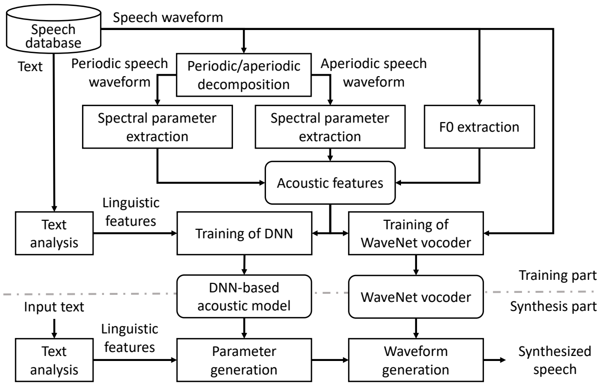

Fig. 1. An overview of the proposed system

A. Acoustic model based on periodic/aperiodic decomposition

An overview of the proposed system is shown in Fig. 1. The speech signals are decomposed into periodic speech signals and aperiodic speech signals. Various periodic/aperiodic decomposition methods have been proposed [25]–[28]. In this paper, harmonic plus residual model (HPR) [29] is applied for the periodic/aperiodic decomposition. The HPR is based on harmonic plus noise model [30] and defines the noise components as the residual components obtained by subtracting the harmonic components from the original speech signal. The periodicity and aperiodicity based on spectral parameters that respectively represent periodic and aperiodic source information with vocal tract characteristics are extracted from periodic and aperiodic speech signals. Then, the periodicity and aperiodicity based on spectral parameters, the logarithmic fundamental frequency, their dynamic features, and voiced/unvoiced binary value are modeled by a single DNN. In the synthesis part, these acoustic features are predicted from the trained DNN. The smooth speech parameters (the input features of the vocoder) are generated from the predicted acoustic features by the MLPG algorithm.

B. WaveNet vocoder

A WaveNet vocoder is a vocoder based on neural networks that generates audio waveforms from acoustic features. Inputs of WaveNet are a sequence of predicted waveform samples in the past and auxiliary features. The joint probability of a sequence of waveform samples $\boldsymbol{x} = (x_1, ..., x_T)$ can be written as

$$P(\boldsymbol{x}|\boldsymbol{h}) = \prod_{t=1}^{T} P(x_t|x_1, ..., x_{t-1}, \boldsymbol{h}), \quad (3)$$

where $\boldsymbol{h}$ represents the auxiliary features. The auxiliary features are used to predict the waveform sample at gated activation units. The gated activation function is defined as follows:

$$\boldsymbol{z} = \tanh(\boldsymbol{W}_f * \boldsymbol{x} + \boldsymbol{V}_f * \boldsymbol{y}) \odot \sigma(\boldsymbol{W}_g * \boldsymbol{x} + \boldsymbol{V}_g * \boldsymbol{y}), \quad (4)$$

where $\boldsymbol{x}$ and $\boldsymbol{z}$ are the input and output of the activation units, $*$ is a convolution operator, $\odot$ is an element-wise product operator, $\sigma(\cdot)$ represents a sigmoid function, f and g represent a filter and a gate, and $\boldsymbol{W}$ and $\boldsymbol{V}$ represent convolution weights for the input and auxiliary features, respectively. The variable $\boldsymbol{y}$ is a time series of the original auxiliary features $\boldsymbol{h}$ transformed into the same resolution as $\boldsymbol{x}$. In the WaveNet vocoder, acoustic features are used as auxiliary features, and waveform samples are generated from the input acoustic features. In the proposed method, acoustic features based on the periodic/aperiodic decomposition (i.e., the periodic and aperiodic based on spectral parameters, the logarithmic fundamental frequency, their dynamic features, and voiced/unvoiced binary value) are used as auxiliary features.

IV. Experiments

A. Experimental conditions

A Japanese speech database constructed by our research group was used in the experiments. The database includes a set of 503 phonetically balanced sentences uttered by a male speaker. The set is the same as the B-set of the ATR phonetically balanced Japanese speech database. The 450 utterances were used for training, and the remaining 53 utterances were used for tests. Speech signals were sampled at 16kHz.

Feed-forward neural networks that had three hidden layers with 1024 units per layer were used as acoustic models. The sigmoid activation function was used in the hidden layers, and the linear activation function was used in the output layer. The input features (i.e., the linguistic features) were normalized to be within 0.0–1.0 based on their minimum and maximum values in the training data. The output features (i.e., the acoustic features) were normalized to have zero-mean and unit-variance.

Acoustic feature vectors were extracted with a 5 ms shift, and mel-cepstral coefficients were extracted from the smoothed spectrum analyzed by STRAIGHT [13]. In these experiments, four speech synthesis systems shown in Table I were compared. MIX+RATIO in Table I represents a set of spectral parameters that consist of 39-dimensional mel-cepstral coefficients including the 0th coefficient, which were extracted from the smoothed spectrum analyzed by STRAIGHT, and 39-dimensional aperiodicity measures that were extracted from STRAIGHT aperiodicity measures by mel-cepstral analysis. Also, SEPARATED represents a set of spectral parameters consisting of 39-dimensional mel-cepstral coefficients extracted from periodic signals and 39-dimensional mel-cepstral coefficients extracted from aperiodic signals. The periodicity and aperiodicity based on the mel-cepstral coefficients of **STRAIGHT_SEPARATED** were extracted from the periodicity and aperiodicity spectrum that was calculated from spectrum and aperiodicity measures extracted by STRAIGHT, respectively. On the other hand, the periodicity and aperiodicity based on mel-cepstral coefficients of **HPR_SEPARATED** were extracted from the smoothed periodicity spectrum of the periodic speech and the smoothed aperiodicity spectrum of the aperiodic speech, respectively. Also, the mel-cepstral coefficients and aperiodicity measures of **HPR_MIX+RATIO** were extracted from the smoothed spectrum and aperiodicity measures that were calculated from the periodicity spectrum of the periodic speech and the aperiodicity spectrum of the aperiodic speech. In addition to the spectral parameters shown in Table I, logarithmic fundamental frequency, their dynamic and acceleration coefficients, and the voiced/unvoiced binary value were used as acoustic features for the four speech synthesis systems.

B. Objective Evaluation for DNN-based Acoustic Model

To objectively evaluate the distortion of the acoustic features between natural speech and synthesized speech, mel-cepstral distortion (MCD), root mean squared error (RMSE) for aperiodicity measures (RMSE$_{\text{ap}}$), RMSE for periodicity spectra (RMSE$_{\text{p-sp}}$), and RMSE for aperiodicity spectra (RMSE$_{\text{a-sp}}$) were used. MCD and RMSE were calculated by

$$MCD = \frac{10}{\ln 10}\sqrt{2\sum_{m=1}^{M}(c_m - \hat{c}_m)^2}, \quad (5)$$

$$RMSE = \sqrt{\frac{1}{F}\sum_{f=1}^{F}\left(20\log_{10}\frac{|Y(f)|}{|X(f)|}\right)^2}, \quad (6)$$

where c and $\hat{c}$ are mel-cepstrum from natural speech and synthesized speech, respectively, and M is the order of mel-cepstrum. Also, $X(f)$ and $Y(f)$ represent the aperiodicity measure, periodicity spectrum, or aperiodicity spectrum of natural speech and synthesized speech, respectively, and F is the number of frequency bins.

Table II lists the objective evaluation results of the STRAIGHT-based system and Table III lists the objective evaluation results of the HPR-based system. As spectra and aperiodicity measures extraction methods differ between STRAIGHT-based systems and HPR-based systems, the targets of the output features also differ. Therefore, a fair comparison between the results of Table II and Table III cannot made. The RMSE$_{\text{p-sp}}$ and RMSE$_{\text{a-sp}}$ results show that modeling accuracy of periodic/aperiodic components of SEPARATED-based systems achieved better than MIX+RATIO-based systems. However, the RMSE$_{\text{ap}}$ results show that SEPARATED-based systems deteriorate aperiodicity measures. This is because the MIX+RATIO-based systems model aperiodicity measures directly, whereas the SEPARATED-based systems do not consider the ratio of an aperiodic component to a speech signal. Also, SEPARATED-based systems model periodicity and aperiodicity spectra directly, whereas the MIX+RATIO-based systems do not model these. Nevertheless, the MCD results show that SEPARATED-based systems that do not directly model mel-cepstral coefficients are better than MIX+RATIO-based systems. These results suggest that the prediction of

TABLE I
COMPARATIVE SYSTEMS

System	Acoustic features	Periodic/aperiodic decomposition or extraction method
STRAIGHT_MIX+RATIO	mel-cepstral coefficients aperiodicity measures	STRAIGHT
STRAIGHT_SEPARATED	periodicity based on mel-cepstral coefficients aperiodicity based on mel-cepstral coefficients	
HPR_MIX+RATIO	mel-cepstral coefficients aperiodicity measures	HPR
HPR_SEPARATED	periodicity based on mel-cepstral coefficients aperiodicity based on mel-cepstral coefficients	

TABLE II
COMPARISON OF THE OBJECTIVE EVALUATION FOR STRAIGHT-BASED SYSTEMS

		STRAIGHT_ MIX+RATIO	**STRAIGHT_ SEPARATED**
MCD	[dB]	4.687 ± 0.012	$\mathbf{4.667 \pm 0.012}$
$RMSE_{ap}$	[dB]	$\mathbf{3.518 \pm 0.011}$	3.586 ± 0.011
$RMSE_{p\text{-}sp}$	[dB]	6.699 ± 0.033	$\mathbf{6.567 \pm 0.032}$
$RMSE_{a\text{-}sp}$	[dB]	6.742 ± 0.029	$\mathbf{6.631 \pm 0.028}$

TABLE III
COMPARISON OF THE OBJECTIVE EVALUATION FOR HPR-BASED SYSTEMS

		HPR_MIX+RATIO	**HPR_SEPARATED**
MCD	[dB]	4.434 ± 0.012	$\mathbf{4.430 \pm 0.011}$
$RMSE_{ap}$	[dB]	$\mathbf{3.335 \pm 0.017}$	3.437 ± 0.021
$RMSE_{p\text{-}sp}$	[dB]	9.626 ± 0.051	$\mathbf{8.764 \pm 0.045}$
$RMSE_{a\text{-}sp}$	[dB]	6.361 ± 0.032	$\mathbf{6.151 \pm 0.030}$

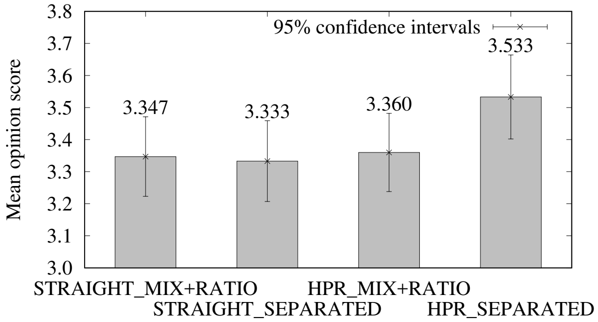

Fig. 2. Comparison of evaluated Mean Opinion Score (MOS) for the acoustic models

spectral parameters was affected by periodic/aperiodic decomposition.

C. Subjective Evaluation for DNN-based Acoustic Model

A subjective listening test to evaluate the naturalness of the synthesized speech was conducted. The naturalness of the synthesized speech was assessed by the mean opinion score (MOS) test method. The opinion score for the MOS tests was set on a five-point scale (5: natural – 1: poor). Fifteen sentences were chosen at random from the test sentences, and the ten subjects were Japanese.

We compared DNN-based acoustic models by employing a STRAIGHT vocoder. The input features of the STRAIGHT vocoder consisted of the smoothed spectrum, aperiodicity measures, and a fundamental frequency. Therefore, the smoothed spectrum and the aperiodicity measures of SEPARATED-based systems were recalculated. Fig. 2 shows the experimental results in MOS. **HPR_SEPARATED** significantly outperforms **STRAIGH_SEPARATED**, as shown in Fig. 2. These results indicate that periodic/aperiodic decomposition based on HPR is effective. This could be because the periodic and aperiodic components were extracted more accurately than STRAIGHT aperiodicity measures by decomposing the speech signals into periodic/aperiodic components. Also, as the difference between the periodic/aperiodic components of STRAIGHT and HPR was not modeled in MIX+RATIO-based systems, there was no significant difference between **STRAIGHT_MIX+RATIO** and **HPR_SEPARATED**.

D. Subjective Evaluation for WaveNet vocoder

The WaveNet vocoder was built from 30 residual blocks. Specifically, dilation in 10 layers was set to $2^0, 2^1, 2^2, ..., 2^9$ and repeated three times to form a total of 30 dilated causal convolution layers. The number of channels for dilated causal convolutions was set to 128 and the number of channels for residual and skip-connection were set to 256. The WaveNet vocoder was trained by using acoustic features extracted from training data as auxiliary features. In the generation process, the auxiliary features of the WaveNet vocoder were the acoustic features predicted by DNN-based acoustic model. Fig. 3 shows the MOS results for the WaveNet vocoder. It can be seen that **HPR_SEPARATED** significantly outperforms other systems, as shown in Fig. 3. These results indicate that speech synthesis using the WaveNet vocoder based on periodic/aperiodic decomposition improve the naturalness of synthesized speech. Therefore, modeling based on periodic/aperiodic decomposition is effective even in the WaveNet vocoder.

V. CONCLUSIONS

In this paper, speech synthesis using WaveNet vocoder based on periodic/aperiodic decomposition has been proposed

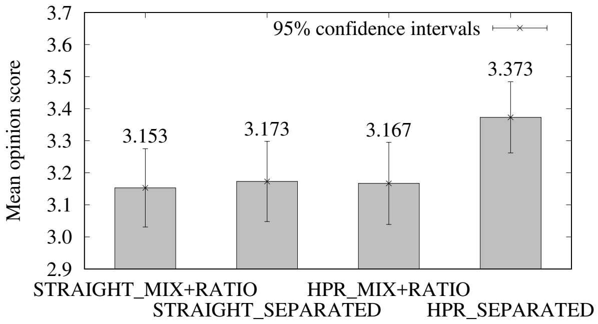

Fig. 3. Comparison of WaveNet vocoder

for statistical parametric speech synthesis. The proposed approach models separated periodic and aperiodic components by the DNN-based acoustic model. Additionally, the synthesized speech is directly generated considering both periodic and aperiodic components by the WaveNet vocoder. The results of experiments show the proposed approach can improve the naturalness of synthesized speech more than a conventional approach.

Future work will include extensive experiments to compare the proposed system to two WaveNet vocoders that generate periodic and aperiodic speech waveforms.

Acknowledgment

This work was supported by JSPS KAKENHI (A) Grant Number 18H04128, JSPS KAKENHI (C) Grant Number 18K11163, and CASIO SCIENCE PROMOTION FOUNDATION.

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] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” *Proceedings of ICASSP*, pp. 373–376, 1996.

[3] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR,” *Proceedings of ICASSP*, pp. 805–808, 2001.

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Speaker interpolation in HMM-based speech synthesis system,” *Proceedings of Eurospeech*, pp. 2523–2526, 1997.

[5] K. Shichiri et al., “Eigenvoices for HMM-based speech synthesis,” *Proceedings of ICSLP*, pp. 1269–1272, 2002.

[6] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, “A style controltechnique for HMM-based expressive speech synthesis,” *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 9, pp. 1406–1413, 2007.

[7] Y. Morioka et al., “Miniaturization of HMM-based speech synthesis,” *Autumn Meeting of the Acoustical Society of Japan*, pp. 325–326, 2004, (in Japanese).

[8] S. J. Kim, J. J. Kim, and M. S. Hahn, “HMM-based Korean speech-synthesis system for hand-held devices,” *IEEE Trans. Consum. Electron.*, vol. 52, no. 4, pp. 1384–1390, 2006.

[9] A. Gutkin, X. Gonzalvo, S. Breuer, and P. Taylor, “Quantized HMMsfor low footprint text-to-speech synthesis,” *Proceeding of Interspeech*, pp. 837–840, 2010.

[10] J. Yamagishi et al., “Robust speaker-adaptive HMM-based text-to-speechsynthesis,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 6, pp. 1208–1230, 2009.

[11] S. Imai, K. Sumita, and C. Furuichi, “Mel Log Spectrum Approximation (MLSA) filter for speech synthesis,” *Electronics and Communications in Japan*, vol. 66 no. 2, pp. 10–18, 1983.

[12] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech,” *Proceedings of ICASSP*, pp. 137–140, 1992.

[13] H. Kawahara, I. Masuda-Katsuse, A. de Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure insounds,” *Speech Communication*, pp. 187–207, 1999.

[14] 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.

[15] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, “Speaker-dependent WaveNet vocoder,” *Proceedings of Interspeech*, pp. 1118–1122, 2017.

[16] A. van den Oord et al., “WaveNet: A Generative Model for Raw Audio,” https://arxiv.org/abs/1609.03499, 2016.

[17] H. Zen, Andrew. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” *Proceedings of ICASSP*, pp. 7962–7966, 2013.

[18] H. Kawahara, J. Estill, and O. Fujimura, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT,” *Proceedings of MAVEBA*, pp. 13–15, 2001.

[19] Y. Long, Z. Yan, F. K. Soong, L. Dai, and W. Guo, “Speaker characterization using spectral subband energy ratio based on Harmonic plus Noise Model,” *Proceedings of ICASSP*, pp.4520–4523, 2011.

[20] K. Tachibana, Y. Shiga, T. Toda, and H. Kawai, “V/UV-decision-free statistical parametric speech synthesis based on periodic/aperiodic decomposition,” *Spring Meeting of the Acoustical Society of Japan*, pp. 209–212, 2017, (in Japanese).

[21] T. Drugman and T. Raitio, “Excitation modeling for HMM-based speech synthesis: breaking down the impact of periodic and aperiodic components,” *Proceeding of ICASSP*, pp. 260–264, 2014.

[22] R. Maia, M. Akamine, and M. J. F. Gales, “Complex cepstrum for statistical parametric speech synthesis,” *Speech Communication*, vol. 55, pp. 606–618, 2013.

[23] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” *IEEE International Conference on Acous-tics, Speech and Signal Processing*, pp. 936–939, 2000.

[24] K. Sawada, C. Asai, K. Hashimoto, K. Oura, and K. Tokuda, “The NITech text-to-speech system for the Blizzard Challenge 2016,” *Blizzard Challenge 2016 Workshop*, 2016.

[25] B. Yegnanarayana, C. d’Alessandro, and V. Darsinos, “An iterative algorithm for decomposition of speech signals into periodic and aperiodic components,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 1–11, 1998.

[26] B. Elie and G. Chardon, “Robust tonal and noise separation in presence of colored noise, and application to voiced fricatives,” *Proceedings of International Congress on Acoustics*, pp. 1–10, 2016.

[27] P. Zubrycki and A. Petrovsky, “Accurate speech decomposition into periodic and aperiodic components based on discrete harmonic transform,” *Proceedings of European Signal Processing Conference*, pp. 2336–2340, 2007.

[28] X. Serra and J. Smith, “Spectral Modeling Synthesis: A Sound Analysis/Synthesis Based on a Deterministic plus Stochastic Decomposition,” *Computer Music Journal*, vol. 14 pp. 12–24, 1990.

[29] sms-tools https://github.com/MTG/sms-tools

[30] A. Syrdal, Y. Stylianou, L. Garrison, A. Conkie, and J. Schroeter, “Td-Psola Versus Harmonic Plus Noise Model In Diphone Based Speech Synthesis,” *Proceedings of ICASSP*, 1998.