Interspeech 2018
2-6 September 2018, Hyderabad

A New Glottal Neural Vocoder for Speech Synthesis

Yang Cui, Xi Wang, Lei He and Frank K. Soong

Microsoft AI & Research, Beijing, China
{yancu,xwang,helei,frankkps}@microsoft.com

Abstract

Direct modeling of waveform generation for speech synthesis, e.g. WaveNet, has made significant progress on improving the naturalness and clarity of TTS. Such deep neural network-based models can generate highly realistic speech but at high computational and memory costs. We propose here a novel neural glottal vocoder which tends to bridge the gap between the traditional parametric vocoder and end-to-end speech sample generation. In the analysis, speech signals are decomposed into corresponding glottal source signals and vocal tract filters by the glottal inverse filtering. Glottal pulses are parameterized into energy, DCT coefficients (shape) and phase. The phase trajectory of successive glottal pulses is rendered with a trainable weighting matrix to keep a smooth pitch synchronous phase trajectory. We design a hybrid, i.e., both feed-forward and recurrent, neural network to reconstruct the glottal waveform including the optimized weighting matrix. Speech is then synthesized by filtering the generated glottal waveform with the vocal tract filter. The new neural glottal vocoder can generate high-quality speech with efficient computations. Subjective tests show that it gets an MOS score of 4.12 and 75% preference over the conventional glottal vocoder with a perceived quality comparable to WaveNet and natural recording in analysis-by-synthesis.

Index Terms: statistical parametric speech synthesis, neural vocoder, glottal waveform generation, phase weighting matrix

1. Introduction

Statistical parametric speech synthesis (SPSS) is a widely used paradigm to produce successive and flexible high quality speech with low computational and memory cost in text-to-speech (TTS). The quality of SPSS system is mainly affected by three factors: vocoder, acoustic model accuracy and over-smoothing [1]. Recently, deep neural networks, especially the sequential neural network [2, 3], has improved the model accuracy and alleviate the over-smoothing issue. Despite those improvements, the synthetic speech quality is still limited by the vocoder, which causes the gap between SPSS and unit concatenation approaches.

Vocoders are used for speech parametrization and waveform generation in the SPSS system. The quality of analysis-by-synthesis reflects the final synthetic speech quality in naturalness and similarity. Source-filter based vocoder is one of the most popular and high quality ways to parameterize, modify, and reconstruct waveform, e.g. STRAIGHT [4, 5], glottDNN [6], IT-FTE [7], etc., which are proposed to improve the perceptual quality while alleviating the 'buzziness' and 'muffledness' problems [1]. The differences between them are mainly the extraction and parameterization methods of the excitation signal. STRAIGHT is a high quality channel vocoder, which analyzes the speech into smoothed time-frequency representation using pitch-adaptive way then model the excitation by the periodic and aperiodic component in several frequency bands. GlottDNN is one of series glottal vocoders [8, 9, 10]. The physiologically meaningful glottal waveform, which represents the time-derivative of the air flow generated from vocal folds [8], is extracted by glottal inverse filtering (GIF) [11] and parameterizes to glottal features. IT-FTE, a kind of waveform interpolation vocoder [12], decomposes the excitation into slowly evolved waveform (SEW) and rapidly evolved waveform (REW) and parameterizes their spectrum by DCT coefficients. Although the vocoders above have improved the perceptual quality of synthetic speech, an inevitable loss has been made during the parameterization and reconstruction stage, as there are some assumptions which are not accurate.

Recently, as the rapid development of deep learning and the increased computational power, some advanced and complicated autoregressive generative models have been successfully applied to complex distribution for wideband raw audio samples. Specifically, WaveNet [13], SampleRNN [14] and WaveRNN [15] can generate realistic and impressive voice quality with appropriate conditions [16]. These model architectures directly use audio samples with long-range temporal dependences by applying a very deep model with complicated non-linear activations. Different from traditional parametric vocoders, the autoregressive models address the statistical parametric speech synthesis problem from another angle, by skipping the difficulty of speech signal analysis and synthesis. However, such autoregressive models are a lot more computational and more memory expensive than traditional parametric vocoders.

In this paper, we propose a novel neural glottal vocoder which uses vocoder features with appropriate design of the neural network to achieve waveform-like voice quality as raw generative models in frame-level. By domain knowledge from speech signal processing, our approach largely improves the efficiency and flexibility of speech generation. More specifically, we design a glottal waveform generative neural network by referring to the generation process of waveform interpolation vocoders from given features. In analysis stage, speech signals are decomposed into corresponding glottal source signals and vocal tract filters by the glottal inverse filtering. The glottal closure instants (GCI) are detected while glottal pulse prototypes are extracted accordingly. Glottal pulses are parameterized into energy, DCT coefficients (shape) and phase. The phase trajectory of successive glottal pulses is rendered with a trainable weighting matrix to keep a smooth pitch synchronous phase trajectory. Speech is then synthesized by filtering the generated glottal source signal with the vocal tract filter. We expect the proposed glottal neural vocoder to generate high quality speech with low computational and memory cost.

This paper is organized as follows. Section 2 describes the details of speech signal analysis and the proposed neural network for glottal waveform generation, which is followed by evaluation of the proposed neural glottal vocoder in Section 3. Section 4 is the conclusion.

10.21437/Interspeech.2018-1757

2. The proposed glottal neural vocoder

The proposed glottal neural vocoder consists of a glottal feature extractor and a hybrid neural network for glottal waveform generation. The speech is first decomposed into the glottal source signal and the vocal tract filter by glottal inverse filtering. After glottal closure instant detection, glottal features such as phase, shape and energy of each pitch cycle are extracted in pitch-synchronized analysis. With these extracted glottal features, a hybrid neural network is designed according to the characteristic of each feature, which consists of a trainable phase weighting matrix, energy applied glottal pulse through a long-short-term memory (LSTM) and generation of glottal waveform. The speech is then synthesized by filtering the generated glottal waveform with the vocal tract filter.

2.1. Glottal signal analysis and features

2.1.1. glottal inverse filtering

Glottal inverse filtering is a procedure to estimate glottal source signal and vocal tract filters from the speech signal. Here we adopt the iterative adaptive inverse filtering (IAIF) [11] algorithm, which can automatically decompose the glottal source and the vocal tract in adaptive manner and converge with a few iterations. The vocal tract filters are then parameterized as line spectrum pair (LSP) coefficients.

2.1.2. glottal feature extraction

The block diagram of glottal signal analysis and extracted features is shown in Figure 1. We extract the glottal features by referring to the waveform interpolation vocoders. These features are the fundamental phase, shape and energy features, where the fundamental phase represents the time series and fundamental frequency information, the shape and energy feature represent the characteristic waveform (CW) information. In the perspective of waveform interpolation coding [17], the glottal pulse and the fundamental phase together form a characteristic waveform surface. Let $u(n, \phi)$ denote a periodic function with the fundamental phase ϕ extracted at the n-th frame. Then the period signal $u(n, \phi)$ can be represented as follows:

$$u(n,\phi) = \sum_{k=1}^{P(n)/2} [A_k cos(k\phi) + B_k sin(k\phi)], \qquad (1)$$

where the fundamental phase $\phi(n, m)$ which denotes the m-th component of the CW extracted at the n-th frame is defined as $\phi(n, m) = 2\pi m/P(n)$. $P(n)$ is the time-varying pitch period in the n-th frame. A_k and B_k denote the k-th discrete-time Fourier series coefficients of the characteristic waveform. Thus fundamental phase and CW features are necessary to reconstruct the glottal waveform.

A voice/unvoiced (V/UV) detection is applied to discriminate between the voiced frame and the unvoiced frame. We use GCI detection to mark anchor points which represent the beginning of each pitch cycle for voiced segments. For the unvoiced segments, pseudo anchor points are marked according to the interpolated F0 between the nearest voiced frame. Then the fundamental phase is linear interpolated between the neighboring anchor points from 0 to 2π on sample-level in both voiced and unvoiced frame according to the definition.

To extract shape and energy, the glottal pulse is extracted and interpolated to a fixed length. The energy of interpolated glottal pulse is calculated and transformed to Logarithm. The shape feature is extracted by normalizing the interpolated glottal pulse to unit energy and represented as DCT coefficients. Then the pitch-synchronized shape and energy features are rearranged into each frame by linear interpolation. The fundamental phase feature are also stacked together in frame-level. Now, all the features are represented in frame-level which can be directly used in glottal waveform generative model.

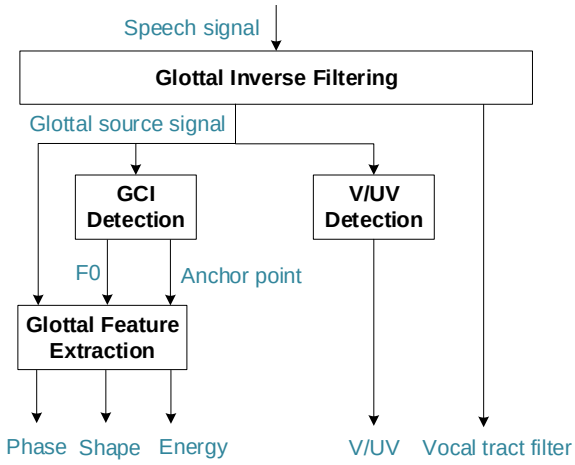

Figure 1: *Block diagram of glottal signal analysis.*

2.2. Hybrid neural network for glottal waveform generation

In this section, we introduce how the extracted glottal features work in the proposed neural network. The straight forward way is to put in all the useful features and calculate the loss regarding the final waveform. However, it doesn't work well in our initial experiment. Those features could not be effectively learned by the network, even worse than directly reconstruct the glottal waveform through definition. It reminds us to carefully design and use these different features. Thus, we design a trainable weighting matrix as part of the generative model to handle the phase feature and a LSTM neural network for shape and energy features. Then the product of weighting matrix and LSTM go through two fully connected (FC) layers to predict the target glottal waveform.

2.2.1. phase-based weighting matrix

As introduced above, phase information represents the timing for waveform interpolation. This feature should be separated and well handled before multiplying with energy and glottal pulse. Here we revisit a phase-based weighting matrix to reconstruct the glottal waveform. Let $\phi(n, k)$ denote the k-th component of phase in the n-th frame, and $c(n, l)$ denote the l-th component of the CW in the n-th frame. The glottal waveform $u(n, \phi(n, k))$ can be reconstructed as follows:

$$\begin{aligned} u(n,\phi(n,k)) &= \sum_{l=-\infty}^{\infty} c(n, lT_s) sinc(\phi(n,k) - lT_s) \\ &\approx \sum_{l=1}^{L} c(n, lT_s) f(\phi(n,k) - lT_s)), \end{aligned} \qquad (2)$$

where $sinc(t) = sin(t)/t$ represents the sinc function. L is the length of the CW. $T_s = 2\pi/L$ represents the sampling interval

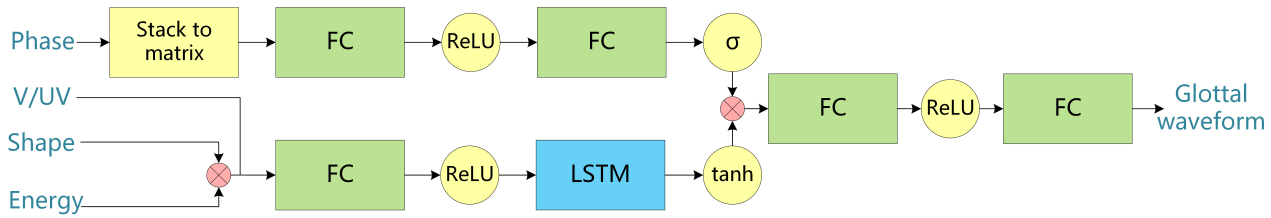

Figure 2: *Overview of the glottal waveform generative neural network.*

of the CW. The equation in (2) requests the CW satisfying the Nyquist sampling rate. As the length of CW is finite, the sinc function can be replaced by other interpolation functions in the local region, e.g. spline functions, represented as $f(t)$. Thus, we get the approximation equation in (2).

Let the following variables $\boldsymbol{\Phi}(n)$, $\boldsymbol{c}(n)$ and $\boldsymbol{u}(n, \boldsymbol{\Phi}(n))$ represent the vectorized phase $\phi(n,k)$, CW $c(n,l)$ and reconstructed waveform $u(n, \phi(n,k)$ in the n-th frame, respectively:

$$\begin{aligned} \boldsymbol{\Phi}(n) &= [\phi(n,1), \phi(n,2), \cdots, \phi(n,K)]^T_, \\ \boldsymbol{c}(n) &= [c(n,T_s), c(n,2T_s), \cdots, c(n,LT_s)]^T_, \\ \boldsymbol{u}(n, \boldsymbol{\Phi}(n)) &= [u(n, \phi(n,1)), \cdots, u(n, \phi(n,K))]^T_, \\ \boldsymbol{F}_{k,l}(\boldsymbol{\Phi}(n)) &= f(\phi(n,k) - lT_s), \end{aligned} \tag{3}$$

where K is the number of samples in one frame and L is the CW length. $\boldsymbol{F}(\boldsymbol{\Phi}(n))$ is defined as the phase-based weighting matrix, $k \in [1, K]$ and $l \in [1, L]$.

Then we have the vector version of equation (2):

$$\boldsymbol{u}(n, \boldsymbol{\Phi}(n)) = \boldsymbol{F}(\boldsymbol{\Phi}(n))\boldsymbol{c}(n). \tag{4}$$

The formula above shows that the reconstructed waveform $\boldsymbol{u}(n, \boldsymbol{\Phi}(n))$ can be decomposed to the product of the weighting matrix $\boldsymbol{F}(\Phi(n))$ and the CW vector $\boldsymbol{c}(n)$. It inspires us to leverage neural network to predict phase-based weighting matrix and the CW then multiply them to reconstruct the glottal waveform.

2.2.2. *hybrid neural network*

It is not efficient to simply stack the phase, shape and energy features as input feature vector and train the network regarding the glottal waveform as it is showed in our initial experiment. Hence, we design the network for each feature with consideration of each role who plays in the reconstruction process introduced above.

The designed hybrid glottal waveform generative neural network is shown in Figure 2. The phase-based weighting matrix is introduced to reconstruct the glottal waveform through weighting the CW component, as shown in equation (4). The equation (3) shows that the weighting matrix function $\boldsymbol{F}(\cdot)$ is a complicated non-linear function of the phase vector $\boldsymbol{\Phi}$(n). Thus, we use two fully connected layers followed by different nonlinear activations to simulate the phase-based weighting function $\boldsymbol{F}(\cdot)$. As the CW has been slowly changing in voiced segments and rapidly changing in unvoiced segments, we adopt LSTM to capture the history sequence information. We use the different activations as ReLU and sigmoid to increase the regularization and boundary smoothness for phase weighting matrix, and tanh for the LSTM.

To construct the weighting matrix, the phase vector is stacked to matrix in the same manner as the matrix $\boldsymbol{F}(\boldsymbol{\Phi}(n))$ defined in equation (3). The shape feature is multiplied by the energy after exponential operation to recover the original amplitude of CW. Then the energy modulated shape feature is feed to the LSTM with V/UV feature. After the sigmoid activation function, the weighting matrix is multiplied by the output of the LSTM, which represents the weighting multiplication in equation (4). The glottal waveform is generated after passing the product through two additional fully connected layers. Mean square error (MSE) is adopted as the loss function during training. We find the learning curve of proposed hybrid neural network performs much better than the simple network which stacks all the features together with the same model size.

Finally, the vocal tract feature represented as LSP coefficients are transformed to vocal tract filters. After glottal waveform generation, the final speech is synthesized by linear convolution using the vocal tract filters on glottal waveform.

3. Experiments

The data base consists of over 16,000 sentences approximately 20 hours recorded from a professional US female speaker. We choose 15,000, 500, and 500 utterances as training, validation and test set, respectively. The sampling rate of the corpus is 16 kHz.

In the analysis stage, the vocal tract features and glottal source features are extracted in pitch-synchronized analysis then interpolated to the frame as 2.5 ms (40 samples) in length. The phase feature is merged into one frame phase vector every 40 samples. As shown in Table 1, 30-dimension LSP coefficients are extracted as vocal tract features, 64-dimension DCT coefficients and 1-dimension energy in Logarithm and 1-dimension V/UV flag as glottal features.

In the experiment, the dimensionality of the hidden fully connected layers is set to 512, while the cell dimension of LSTM is set to 512 with project dimension 512. For the training, we use the RMSProp optimizer and set the truncation length to 10 and mini-batch size to 200. The hybrid neural network is trained on Tesla K80 using the Microsoft Cognitive Toolkit CNTK[1] [18].

Table 1: *Speech features and dimension per frame.*

Feature name	**Dimension per frame**
Fundamental phase	40
Glottal shape (DCT)	64
Glottal energy (Log)	1
V/UV	1
Vocal tract (LSP)	30

[1]https://github.com/Microsoft/CNTK

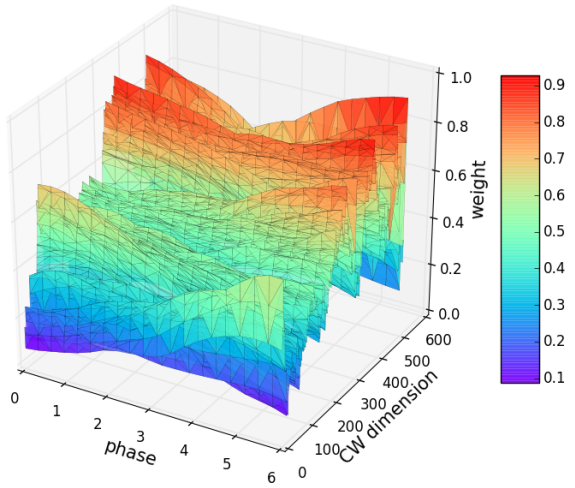

Figure 3: *Visualization of the phase-based weighting matrix.*

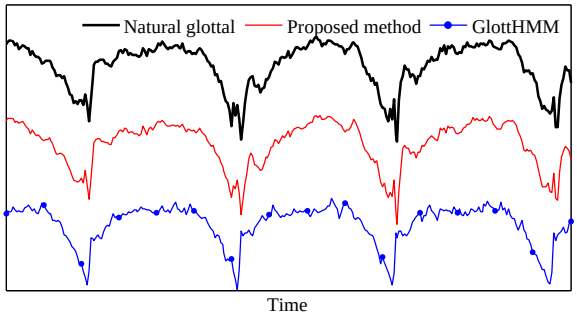

Figure 4: *Demonstration of the glottal waveform estimated by GIF(top), generated by the proposed glottal neural vocoder (middle) and generated by the glottal vocoder(bottom) in analysis-by-synthesis.*

The visualization of the phase-based weighting matrix in the fine tuned glottal waveform generative model is shown in Figure 3. The x,y and z axis represent the phase, the CW index and the weight value, respectively. As the phase changes from 0 to 2π, the weight vector slowly evolves while weighting different parts of CW components, which verifies our derivation and design of the hybrid generative model.

The perceptual quality of the proposed glottal neural vocoder is evaluated by performing an A/B preference test and a mean opinion score (MOS) test compared with two vocoders plus natural recordings. The first vocoder is the glottal vocoder of glottHMM [8], which has the released version[2]. The second one is the WaveNet with 20 layers depth conditioned on Mel-spectrum as a vocoder [19]. The analysis-by-synthesis result for glottal features using the proposed method is illustrated in Figure 4, which shows the original glottal waveform extracted by GIF from natural recording and the generated glottal waveform by the proposed glottal neural vocoder and the glottal vocoder.

In the A/B preference test, the proposed glottal neural vocoder and the glottal vocoder are tested through analysis-by-synthesis. 50 utterances are randomly selected from the evaluation set. 15 listeners are asked to provide quality judgements and their preferences in naturalness of two given synthesis speech. The results of the A/B preference test are presented in Figure 5. It can be clearly seen that our approach outperforms the glottal vocoder by 75% preference, which demonstrates the effectiveness of our neural glottal vocoder over the traditional parametric glottal vocoder.

[2]http://www.helsinki.fi/speechsciences/synthesis/glott.html

Table 2: *The MOS score and computational cost.*

Voice name	**FLOPS**	**MOS**
Recording	–	4.53 ± 0.08
WaveNet vocoder	209.7G	4.51 ± 0.05
Proposed neural vocoder	767.5M	4.12 ± 0.05
Glottal vocoder	101.3M	3.48 ± 0.05

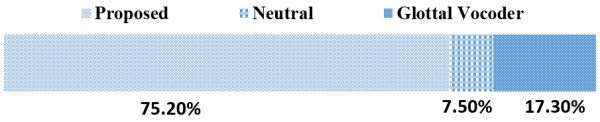

Figure 5: *Results of A/B preference test(%).*

In MOS test, the WaveNet vocoder and the glottal vocoder are chosen to compare with the proposed glottal neural vocoder through analysis-by-synthesis. The WaveNet vocoder is trained with the same corpus and condition as Mel-spectrum, 20 layers with discretized mixture Logistic loss [20]. 50 utterances are randomly selected from evaluation set. 20 listeners are asked to give scores from 1 (Bad) to 5 (Excellent) in naturalness of synthesis speech and recording. The MOS test in Table 2 shows that the WaveNet gets the highest score in these three vocoders and very close to realistic speech. The proposed glottal neural network gets much higher quality than the parametric glottal vocoder, which is consistent with the A/B preference test.

Further more, we also analysis the computational cost of these three vocoders in the MOS test. The results of the MOS score and the corresponding computational cost of each vocoder are presented in Table 2. We use the float-point operations per second (FLOPS) as measurement of computational cost, which shows the float-point operations required to synthesize one second speech waveform. The computational cost of WaveNet is the most expensive of all, nearly three hundred times of the proposed vocoder. The proposed glottal neural vocoder achieves significant higher quality than the glottal vocoder at a quite acceptable cost.

4. Conclusion

In this paper, we propose a novel neural glottal vocoder for speech synthesis, which explores the capability of glottal features for reconstructing glottal waveform with the utilization of a hybrid deep neural network. With the high quality of reconstructed glottal waveform, we could get highly natural synthesized speech quality comparing with traditional vocoder, getting rid of the 'vocoding' effect, closing to waveform-like voice. Although current voice quality is still not as good as raw generative mode like WaveNet, the computational cost is much lower. Besides glottal waveform generation, in the next work, we will continue to make the end-to-end neural vocoder combining with the vocal tract filter features to the final waveform, which will further improve the voice quality.

5. 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] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "Tts synthesis with bidirectional lstm based recurrent neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[3] X. Wang, S. Takaki, and J. Yamagishi, "An autoregressive recurrent mixture density network for parametric speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4895–4899.

[4] 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," in *Second International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2001.

[5] 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 *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 3933–3936.

[6] 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 *Interspeech*, 2016, pp. 2473–2477.

[7] E. Song, F. K. Soong, and H.-G. Kang, "Improved time-frequency trajectory excitation vocoder for dnn-based speech synthesis." in *INTERSPEECH*, 2016, pp. 2253–2257.

[8] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "Hmm-based speech synthesis utilizing glottal inverse filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 153–165, 2011.

[9] T. Raitio, A. Suni, L. Juvela, M. Vainio, and P. Alku, "Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[10] L. Juvela, B. Bollepalli, M. Airaksinen, and P. Alku, "High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5120–5124.

[11] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[12] C.-S. Jung, Y.-S. Joo, and H.-G. Kang, "Waveform interpolation-based speech analysis/synthesis for hmm-based tts systems," *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 809–812, 2012.

[13] 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.

[14] 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.

[15] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *arXiv preprint arXiv:1802.08435*, 2018.

[16] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Advances in Neural Information Processing Systems*, 2017, pp. 2966–2974.

[17] W. B. Kleijn and J. Haagen, "A speech coder based on decomposition of characteristic waveforms," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1. IEEE, 1995, pp. 508–511.

[18] F. Seide and A. Agarwal, "Cntk: Microsoft's open-source deep-learning toolkit," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 2135–2135.

[19] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent wavenet vocoder," in *Proceedings of Interspeech*, 2017, pp. 1118–1122.

[20] A. v. d. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. v. d. Driessche, E. Lockhart, L. C. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," *arXiv preprint arXiv:1711.10433*, 2017.