PERCEPTUAL QUALITY AND MODELING ACCURACY OF EXCITATION PARAMETERS IN DLSTM-BASED SPEECH SYNTHESIS SYSTEMS

Eunwoo Song[1,2,3*], *Frank K. Soong*[2], *Hong-Goo Kang*[3]

[1]NAVER Corp., Seongnam, Korea
[2]Microsoft Research Asia, Beijing, China
[3]Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea

ABSTRACT

This paper investigates how the perceptual quality of the synthesized speech is affected by reconstruction errors in excitation signals generated by a deep learning-based statistical model. In this framework, the excitation signal obtained by an LPC inverse filter is first decomposed into harmonic and noise components using an improved time-frequency trajectory excitation (ITFTE) scheme, then they are trained and generated by a deep long short-term memory (DLSTM)-based speech synthesis system. By controlling the parametric dimension of the ITFTE vocoder, we analyze the impact of the harmonic and noise components to the perceptual quality of the synthesized speech. Both objective and subjective experimental results confirm that the maximum perceptually allowable spectral distortion for the harmonic spectrum of the generated excitation is $\sim$0.08 dB. On the other hand, the absolute spectral distortion in the noise components is meaningless, and only the spectral envelope is relevant to the perceptual quality.

Index Terms— Speech synthesis, long short-term memory, improved time-frequency trajectory excitation vocoder

1. INTRODUCTION

Statistical parametric speech synthesis (SPSS) systems have been greatly advanced when they are combined with a deep learning-based training process. As centralized networks such as a feed-forward neural network (FFNN) or a long short-term memory (LSTM) compactly model the complicated dependencies between input contexts and output acoustic features, they not only improve the accuracy of acoustic models but also alleviate over-smoothing problems in generated parameters [1, 2]. However, the impact of vocoding techniques, especially for modeling LPC excitation signals, remains unclear even though it is undoubtedly a key component for synthesizing natural voices. Because parameterizing the excitation signal has been overly simplified, i.e., a pulse or noise (PoN) or a mixed excitation [3, 4, 5], it has also been difficult to clearly analyze the influence of the excitation on the quality of the synthesized speech.

In our previous work, we proposed an improved time-frequency trajectory excitation (ITFTE) vocoder [6], and implemented deep learning-based SPSS systems with it [7, 8, 9]. In the ITFTE vocoder, a pitch-dependent excitation signal is first obtained by applying an LP inverse filter to an input speech signal, and then it is represented in the time-frequency domain. This excitation, i.e., the TFTE signal, is further decomposed into harmonic and noise components. The slowly evolving waveform (SEW); obtained by the frequency-dependent low-pass filtering of the TFTE signal along the temporal-axis, represents the quasi-periodic/harmonic components of the excitation; in contrast, the remaining noise-like components are represented by the rapidly evolving waveform (REW). Because the dimensions of these components vary depending on the pitch period, they are independently parameterized via a discrete cosine transform (DCT) before the training process; they are reconstructed by an inverse DCT after the generation process [8].

In this paper, we further investigate the relationships between the perceptual quality and the modeling accuracy of the excitation signal depending on the parametric dimension of the ITFTE vocoder. For the analysis, the excitation signals extracted from the Korean male and US English female speech database are decomposed into harmonic and noise components, and then they are trained by a deep LSTM (DLSTM)-based SPSS framework. By changing the parametric dimension of the DCT coefficients, the spectral distortions between the original and the reconstructed excitations are analyzed in detail. Finally, perceptual listening tests are performed to verify how the generation errors in the harmonic and noise components affect the quality of the synthesized speech. Experimental results confirm that even very small errors in the harmonic excitations are perceptually noticeable, whereas only the envelope information is important in the noise excitation. Based on the experimental results, we obtain the best parametric dimensions for the excitation parameters of the ITFTE vocoder, which is very important to design a high-quality speech synthesis system.

*Work partially performed as an intern in the Speech Group, Microsoft Research Asia (MSRA).

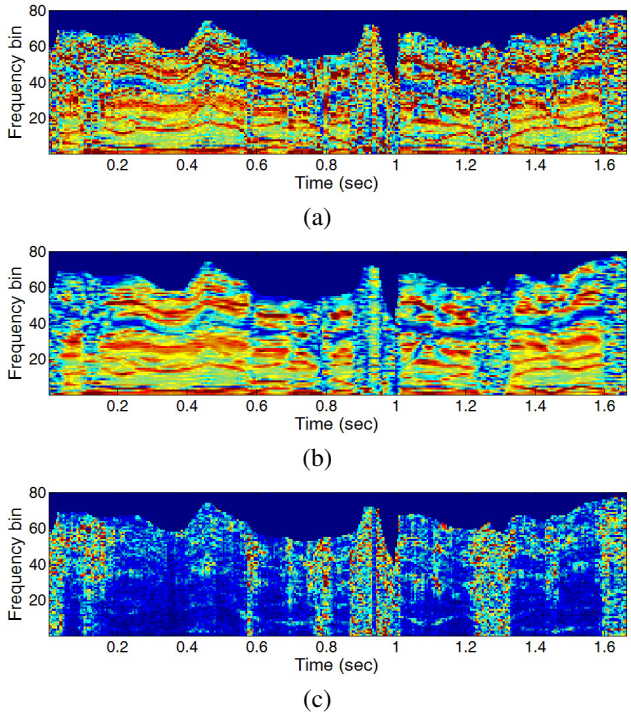

Fig. 1. Time-frequency representation of (a) the excitation signal (TFTE), (b) the harmonic components (SEW), and (c) the noise components (REW).

2. BACKGROUND

2.1. Decomposition method for the excitation signal

To effectively represent and model the excitation components, we adopt an ITFTE-based vocoding technique [6, 7, 8, 9]. In the ITFTE vocoder, a pitch dependent excitation signal is obtained by the inverse filtering of an input speech signal with the LP analysis filter, and it is transformed to the time-frequency domain. A two-dimensional TFTE signal is used to represent the spectral shape of excitation along the frequency axis and the evolution of this shape along the time axis as shown in Fig. 1-(a). To obtain the harmonic excitation spectrum, i.e., the SEW component shown in Fig. 1-(b), each frequency component of the TFTE is low-pass filtered along the time-domain axis. Beyond the cut-off frequency, the remaining noise spectrum, i.e., the REW component shown in Fig. 1-(c), is obtained by subtracting the SEW from the TFTE.

Before being applied to the LSTM training process, both components are parameterized by a full-band discrete cosine transform (FB-DCT)-based modeling method [8]. Note that the SEW and REW dimensions vary depending on the pitch period; therefore, a fixed dimension must be imposed. In the analysis step, the SEW and REW magnitude spectra are independently transformed into the DCT domain. As the DCT has good decorrelation and energy compactness properties, most information is concentrated in the first few coefficients. The lower K-th order coefficients, defined as FB-DCT-SEW and FB-DCT-REW coefficients, are used for the DLSTM training process. By setting the higher-order DCT coefficients to zero in the synthesis step, the full-band SEW and REW magnitude spectra are simply reconstructed by applying an inverse DCT.

2.2. DLSTM-based speech synthesis

In the proposed DLSTM-based SPSS system, rich contexts are first analyzed, and then converted to a sequence of input linguistic features that contain binary features for categorical contexts (e.g., phone labels) and numerical features for numerical contexts (e.g., the number of words in a phrase or the position of the current frame of the current phone). The corresponding output features include the ITFTE and spectral parameters with their time dynamics [10]. The pairs of input and output features are then used to train the weights of the DLSTM by using a *backpropagation through time* (BPTT) [11]. Note that the DLSTM architecture is built by stacking multiple LSTM layers. The LSTM already has a deep architecture in time, because its hidden state relates to a function of all previous hidden states. However, the depth in DLSTMs has an additional meaning in that the input to the network at a given time step goes through multiple LSTM layers in addition to propagation through time [12]. As the inputs go through more non-linear operations per time step, it is possible to find not only a sequential nature in time but also the mapping relationship between the input linguistic and the output acoustic features.

3. EXPERIMENTS

In this section, we analyze the reconstruction errors in the DLSTM-generated excitation signal and verify how the harmonic and noise excitations affect the perceptual quality of the synthesized speech. From the experimental results, we propose a high quality speech synthesis system based on the ITFTE vocoder and DLSTM-training methods.

3.1. Experimental setup

Two phonetically and prosodically rich speech corpora (Korean and English) were used in our experiment, where each corpus was recorded by a professional Korean male (KRM)

Table 1. Number of utterances in different sets.

Speaker	Training	Development	Test
KRM	2500 (∼3.2 h)	200	200
USF	5100 (∼6.0 h)	200	200

Table 2. Subjective preference test results (%) between the synthesized speech samples for the KRM and USF speakers with respect to different dimensions of FB-DCT-SEW coefficients. The systems that achieved significantly better preference at the $p < 0.01$ level are indicated in bold font. The last column represents the LSMD difference between two systems.

Speaker	SEW dimension					Neutral	p-value	LSMD difference (dB)
	8	16	24	32	40			
KRM	**14.6**	**39.6**				45.8	$< 10^{-7}$	**0.18**
	12.1		**57.5**			30.4	$< 10^{-19}$	**0.24**
		17.9	29.2			52.9	0.01	0.06
		9.6		**33.8**		56.6	$< 10^{-8}$	**0.08**
			17.5	22.9		59.6	0.19	0.02
			15.0		29.2	55.8	0.01	0.04
				18.3	22.1	59.6	0.36	0.01
USF	**18.3**	**46.3**				35.4	$< 10^{-7}$	**0.50**
	18.3		**52.9**			28.8	$< 10^{-10}$	**0.77**
		20.8	**42.5**			36.7	$< 10^{-4}$	**0.27**
		22.5		**51.2**		26.3	$< 10^{-6}$	**0.48**
			23.8	**41.7**		34.5	$< 10^{-3}$	**0.21**
			19.2		**46.3**	34.5	$< 10^{-6}$	**0.25**
				27.9	37.1	35.0	0.08	0.04

and a US female (USF) speaker, respectively. The speech signals were sampled at 16 kHz and quantized with 16 bits. Each database was divided into training, development, and test sets. Table 1 shows the number of utterances in each set.

In the analysis step, the frame length was set to 20 ms, and the spectral and excitation parameters were extracted every 5 ms. The 40-dimensional LP coefficients were extracted and converted to line spectral frequencies (LSFs). To prevent unnatural spectral peaks in the LP analysis filter, each coefficient $(a_i, i = 1, ..., 40)$ was multiplied by the bandwidth expansion factor (0.981^i) [9]. On the other hand, the $\{8, 16, 24, 32, 40\}$-dimensional and $\{4, 8, 16\}$-dimensional parameterized SEW and REW coefficients were extracted for the excitation parameters, respectively. The fundamental frequency (F0), energy, and v/uv information were also extracted.

In the DLSTM training step, the output feature vectors comprised all of these parameters together with their time dynamics [10]. The corresponding input feature vectors of the KRM and USF databases included 210- and 346-dimensional contextual information, consisting of: 203 and 311 binary features for categorical linguistic contexts and 7 and 35 numerical features for numerical linguistic contexts, respectively. Before training, both input and output features were normalized to have zero mean and unit variance. The hidden layers consisted of three unidirectional LSTM layers with 512 memory blocks. The weights were randomly initialized and were trained using the BPTT algorithm. The learning rate was set to 0.02 for the first 10 epochs, 0.01 for the next 20 epochs, and 0.005 for the remaining epochs. To parallelize and increase the speed, tens of utterances were randomly selected for each update and used to update the weights simultaneously. The training and test procedures were implemented by using the computational network toolkit (CNTK) [13].

In the synthesis step, the mean vectors of all output feature vectors were first predicted by the DLSTM, and then a speech parameter generation (SPG) algorithm was applied to generate smooth trajectories of acoustic parameters [14]. Because the DLSTM network structure could not predict the variance used for the SPG algorithm, we used pre-computed global variances of output features from all the training data. To reconstruct the SEW and REW, the generated FB-DCT-SEW and FB-DCT-REW coefficients were converted to SEW and REW magnitude spectra, respectively. A fixed phase spectrum drawn from speech was used for the SEW phase [15], whereas the REW phase was randomly selected. The TFTE was then obtained by combining the SEW and REW with its pitch period. Finally, a single pitch-based speech signal was synthesized by the generated LSFs and TFTE. To enhance the spectral clarity, LSF-sharpening [9] and formant-enhancing [16] filters were also applied.

3.2. Analyzing reconstruction errors of harmonic excitations

This section investigates the impact of the parameter dimension of the FB-DCT-SEW coefficients on the reconstruction errors of the harmonic excitation signal. The FB-DCT-SEW coefficients were extracted from the recorded speech, and then trained/generated by the DLSTM to reconstruct the harmonic excitation signals. To measure the reconstruction errors, we define the log-SEW magnitude distance (LSMD)

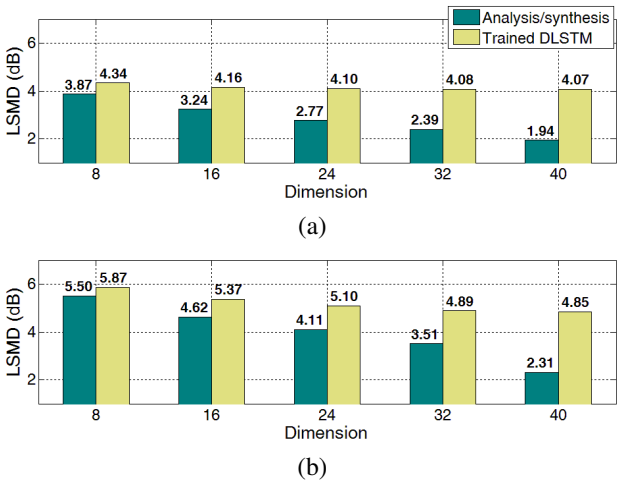

Fig. 2. Average LSMD results for reconstructed SEW magnitude from the FB-DCT-SEW coefficients for (a) the KRM and (b) the USF speakers.

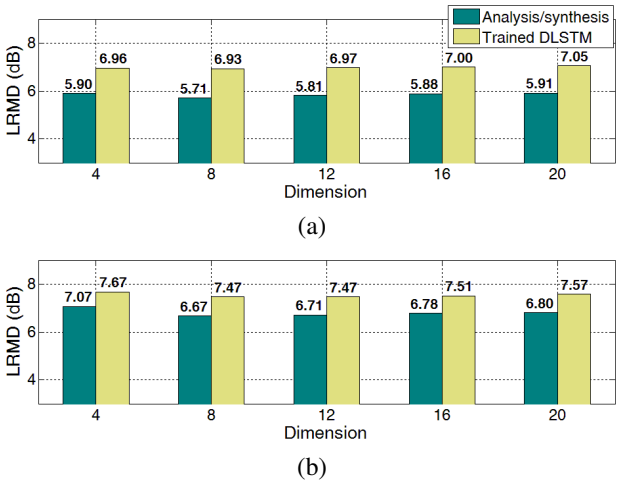

Fig. 3. Average LRMD results for reconstructed REW magnitude from the FB-DCT-REW coefficients for (a) the KRM and (b) the USF speakers.

between the original and the reconstructed spectra as follows:

$$LSMD\ [dB] = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{1}{J}\sum_{k=1}^{J}\left(20\ log\frac{u_{sew}(n,k)}{\widehat{u}_{sew}(n,k)}\right)^2}, \quad (1)$$

where $J = P(n)/_2$ denotes the length of the SEW defined as a half of the pitch period at the n-th frame; $u_{sew}(n,k)$ and $\widehat{u}_{sew}(n,k)$ denote the SEW magnitude spectrum at the k-th frequency-bin extracted from the recorded speech and reconstructed by FB-DCT-SEW coefficients, respectively.

Fig. 2 shows the LSMD results, for the KRM (upper) and USF (lower) speakers with respect to different dimensions of the FB-DCT-SEW coefficients. The findings can be analyzed as follows: First, the reconstruction errors in the harmonic spectra of the analysis/synthesis case consistently decreased as the parametric dimension increased, whereas those of the trained DLSTM case were saturated at a certain dimension in both speakers. This implies that we do not have to impose an infinitely large dimension to train the parameters owing to the limitation of the model accuracy. Second, the female speaker having dynamic high-pitched signals contained relatively larger errors than male speaker, mainly because of the difficulties in the pitch-synchronous analysis/synthesis framework. Applying a high-sampling-rate analysis or weighted LP approach can alleviate this problem [17, 18], which will be discussed in our future works.

To further analyze the influence of the harmonic excitation on the quality of the synthesized speech, A-B preference listening tests were performed. In these tests, 12 native Korean and 12 native US listeners were asked to make quality judgments of the synthesized Korean and English utterances, respectively. Twenty utterances were randomly selected from the test set from both the KRM and the USF databases. These were then synthesized by the generated parameters. In each experiment, the dimensions of the FB-DCT-SEW coefficients differ from each other; whereas all other parameters were kept the same in all systems. Table 2 shows the preference test results with respect to different dimensions of FB-DCT-SEW coefficients. The perceptual qualities in both KRM and USF speakers were also saturated even though the parametric dimension increased; the results exactly coincided with the LSMD results. In the tests, the listeners were able to distinguish the perceptual quality when that the difference was larger than 0.08 dB. It means the accurate reconstruction of the harmonic excitation is very important in the perceptual aspect.

3.3. Analyzing reconstruction errors of noise excitations

As in the harmonic excitation cases, the reconstruction errors of the noise excitations were measured using the log-REW magnitude distance (LRMD) between the original and the reconstructed spectra. Fig. 3 shows the LRMD results for the KRM (upper) and USF (lower) speakers with respect to different dimensions of the FB-DCT-REW coefficients. As shown in these figures, the LRMDs from the noise spectra are not affected by the parameter dimensions; in other words, a few FB-DCT-REW coefficients are sufficient to estimate the noise envelope. This result was also confirmed by the preference test results[1] shown in Table 3, where the listeners could not notice the perceptual difference among the different systems.

[1]The preference test setups were almost same as previous experiments given in section 3.2, but only the dimensions of the FB-DCT-REW coefficients differ from each other (all the other parameters were kept same).

Table 3. Subjective preference test results (%) between the synthesized speech samples for the KRM and USF speakers with respect to different dimensions of FB-DCT-REW coefficients.

Speaker	REW dimension			Neutral	p-value
	4	8	16		
KRM	16.3	13.8		69.9	0.48
	12.1		17.1	70.8	0.15
		15.4	14.2	70.4	0.72
USF	5.0	3.8		91.2	0.51
	5.0		4.2	90.8	0.67
		4.2	6.7	89.3	0.24

Table 4. MOS test results with 95% confidence interval for the final configurations of the ITFTE vocoder and DLSTM-based SPSS systems. For the comparison, the STRAIGHT-based system was included as a baseline.

Speaker	STRAIGHT	ITFTE
KRM	3.61±0.18	4.21±0.21
USF	3.54±0.19	3.72±0.19

3.4. ITFTE vocoder and DLSTM-based speech synthesis

In this section, we verify the performance of the proposed system based on the ITFTE vocoder using the DLSTM training approach. Based on the experimental results discussed in the previous section, we chose the 32-dimensional FB-DCT-SEW and 4-dimensional FB-DCT-REW coefficients to represent the excitation signal. To evaluate the quality of the proposed system, we performed subjective mean opinion score (MOS) tests. The setups were the same as for the preference tests, except that the listeners were asked to make quality judgments of the synthesized speech (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent).

In the test, the impact of excitation components was compared with an additional DLSTM-based system using a STRAIGHT vocoder [19]. Note that only the excitation parameters (e.g., SEW and REW) were replaced with the band aperiodicities (BAPs); whereas all the other parameters were kept the same as those of the ITFTE vocoder. The results of the MOS test (Table 4) show that the proposed system provided better perceptual quality than the BAP-based approach, which implies decomposing the SEW/REW is beneficial to improve the modeling accuracy of the excitation signals. However, as discussed in section 3.2, larger reconstruction errors in the female excitations resulted in relatively worse perceptual quality than that of the male speaker. This will be further analyzed in our future works to improve the vocoding technique for the high-pitched excitation signals.

4. CONCLUSIONS

In this paper, we investigated how the perceptual quality of synthesized speech was affected by reconstruction errors in excitation signals. The excitation signals extracted from the speech database were decomposed into the harmonic and noise components and then trained by the deep DLSTM-based SPSS framework. By changing the parametric dimension of each component, the reconstruction errors were analyzed in detail. Perceptual listening tests were also performed, and the results of the same confirmed that even very small errors in the harmonic excitations were perceptually noticeable, whereas only the envelope information was important in noise excitations. Finally, we successfully implemented a high-quality speech synthesis framework based on the ITFTE vocoder and the DLSTM-based statistical training process.

5. ACKNOWLEDGEMENTS

This research was supported by Microsoft Research and the MSIP (The Ministry of Science, ICT and Future Planning), Korea, under ICT/SW Creative research program supervised by the IITP (Institute for Information & Communications Technology Promotion). The authors would like to thank Feng-Long Xie, Microsoft Research Asia, Beijing, China, for conducting the listening tests.

6. REFERENCES

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

[2] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks." in *Proc. INTERSPEECH*, 2014, pp. 1964–1968.

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EUROSPEECH*, 1999, pp. 2347–2350.

[4] ——, "Mixed excitation for HMM-based speech synthesis," in *Proc. INTERSPEECH*, 2001, pp. 2263–2266.

[5] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE trans. Inf. Syst.*, vol. 90, no. 1, pp. 325–333, 2007.

[6] E. Song, Y. S. Joo, and H.-G. Kang, "Improved time-frequency trajectory excitation modeling for a statistical parametric speech synthesis system," in *Proc. ICASSP*, 2015, pp. 4949–4953.

[7] E. Song and H.-G. Kang, "Deep neural network-based statistical parametric speech synthesis system using improved time-frequency trajectory excitation model," in *Proc. INTERSPEECH*, 2015, pp. 874–878.

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

[9] ——, "Effective spectral and excitation modeling techniques for LSTM-RNN-based speech synthesis systems," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 11, pp. 2152–2161, 2017.

[10] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 34, no. 1, pp. 52–59, 1986.

[11] R. J. Williams and J. Peng, "An efficient gradient-based algorithm for on-line training of recurrent network trajectories," *Neural computat.*, vol. 2, no. 4, pp. 490–501, 1990.

[12] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." in *Proc. INTERSPEECH*, 2014, pp. 338–342.

[13] D. Yu, A. Eversole, M. Seltzer, K. Yao, Z. Huang, B. Guenter, O. Kuchaiev, Y. Zhang, F. Seide, H. Wang *et al.*, "An introduction to computational networks and the computational network toolkit," Microsoft Research, Tech. Rep., 2014.

[14] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000, pp. 1315–1318.

[15] W. B. Kleijn and J. Haagen, "A speech coder based on decomposition of characteristic waveforms," in *Proc. ICASSP*, 1995, pp. 508–511.

[16] J.-H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. Speech and Audio Process.*, vol. 3, no. 1, pp. 59–71, 1995.

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

[18] 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 *Proc. ICASSP*, 2016, pp. 5120–5124.

[19] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Proc. ICASSP*, 1997, pp. 1303–1306.