INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Improved Time-Frequency Trajectory Excitation Vocoder for DNN-Based Speech Synthesis

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

[1]Microsoft Research Asia, Beijing, China
[2]Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
sewplay@dsp.yonsei.ac.kr

Abstract

We investigate an improved time-frequency trajectory excitation (ITFTE) vocoder for deep neural network (DNN)-based statistical parametric speech synthesis (SPSS) systems. The ITFTE is a linear predictive coding-based vocoder, where a pitch-dependent excitation signal is represented by a periodicity distribution in a time-frequency domain. The proposed method significantly improves the parameterization efficiency of ITFTE vocoder for the DNN-based SPSS system, even if its dimension changes due to the inherent nature of pitch variation. By utilizing an orthogonality property of discrete cosine transform, we not only accurately reconstruct the ITFTE parameters but also improve the perceptual quality of synthesized speech. Objective and subjective test results confirm that the proposed method provides superior synthesized speech compared to the previous system.

Index Terms: improved time-frequency trajectory excitation vocoder, speech synthesis, deep neural network

1. Introduction

Statistical parametric speech synthesis (SPSS) systems have been significantly advanced when combined with a deep neural network (DNN)-based training process. A centralized network enables compact modeling of complex dependencies between input contexts and output acoustic features, which not only improves the accuracy of acoustic models but also alleviates over-smoothing problems in generated parameters. Various analyses have also confirmed that DNN-based SPSS systems perform significantly better than conventional approaches based on hidden Markov models [1, 2, 3, 4].

However, the impact of vocoding techniques remains unclear even though it is undoubtedly a key component for synthesizing natural voices. In our previous work, we proposed a high-quality speech vocoder for SPSS systems by using a time-frequency trajectory excitation (TFTE) model [5]. By decomposing a pitch-dependent excitation signal into a slowly evolving waveform (SEW) and a rapidly evolving waveform (REW), the TFTE model extracts the periodicity distribution of the excitation signal in the time-frequency domain [6, 7, 8]. The SEW, which is the most important parameter in the TFTE model, represents the quasi-periodic/voiced portion of the excitation; in contrast, the REW represents the remaining noise-like components. Since the time-varying periodicity of various phonetic information is effectively controlled by SEW and REW, the perceptual quality of the TFTE model is much better than that of conventional band aperiodicity (BAP)-based methods [5].

An improved TFTE (ITFTE) vocoder, which provides a novel parameterization method for the TFTE model, also enhanced the perceptual quality of synthesized speech in SPSS systems [9, 10]. Note that the SEW and REW cannot be directly applied to the DNN training process because their parametric dimensions vary depending on the length of the pitch interval. In the ITFTE vocoder, SEW was first divided into fixed number of frequency sub-bands, and then transformed with a discrete cosine transform (DCT). The first DCT coefficient of each sub-band, which represents a mean (average) component, was used in the DNN training process; the remaining coefficients were stochastically generated by Gaussian random variables. In the case of REW, it was modeled via a power contour estimation method because perceptual significance did not differ greatly from that achieved using the REW parameter itself [11]. Consequently, this resolved the problem of dimensional variation in the training process.

To further improve the whole framework, we propose a full-band DCT-based parameterization method for the ITFTE vocoder. The previous sub-band DCT-based approach produced perceptually non-transparent sound when all the sub-bands were combined into a full-band SEW, because the stochastically generated random variables could not provide enough information for reconstruction. In the proposed method, rather than dividing the SEW into sub-bands, the DCT is applied to the full-band SEW directly. By using the energy compactness and invertibility properties of DCT [12], the entire frequency information of the SEW can be simply and accurately recovered from the fixed number of DCT coefficients. Furthermore, the reconstructed SEW presents smoother shape due to the orthogonality of the DCT basis functions. Various analyses verify that this approach significantly reduces reconstruction errors of the ITFTE parameters in the DNN training framework. Objective and subjective experimental results also confirm that the synthesized quality of the proposed system is superior to the previous approach.

2. Improved time-frequency trajectory excitation vocoder

2.1. TFTE model

In ITFTE vocoder, a speech signal is first inverse filtered by using linear prediction (LP) coefficients, and then an excitation signal is represented by a time-frequency distribution of periodicity. Let $u(n, \phi)$ denote a periodic function with ϕ extracted at the n-th frame, then the TFTE can be represented as follows:

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

where the phase function is defined as $\phi(m) = 2\pi m/P(n)$ with a pitch period $P(n)$; $A_k(n)$ and $B_k(n)$ denote the k-th discrete-time Fourier coefficients of the excitation signal [8].

The SEW, which represents the harmonic components of the excitation signal, is obtained by applying a low-pass filter

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

http://dx.doi.org/10.21437/Interspeech.2016-230

(LPF) to the TFTE along the time-domain axis as follows:

$$u_{SEW}(n,\phi) = \sum_{l=1}^{L} h(l)u(n-l,\phi), \tag{2}$$

where $h(l)$ denotes an L-th order LPF. Note that the SEW is the salient parameter in the ITFTE vocoder, since it contains most of the voicing information. The REW, which represents the noisy components beyond the cut-off frequency of the LPF, is obtained by removing the SEW from the TFTE as follows:

$$u_{REW}(n,\phi) = u(n,\phi) - u_{SEW}(n,\phi). \tag{3}$$

Therefore, the periodicity distribution is efficiently represented by the SEW and REW, whereby it can produce natural shape of the excitation signal.

2.2. Sub-band DCT-based parameterization method for the ITFTE vocoder

Although the TFTE model provides the high-quality analysis/synthesis performance, its parameters cannot be directly applied to the DNN training process because their parametric dimensions change over time due to the variation of pitch period. In our previous works [9, 10], the SEW was parameterized by a sub-band DCT (SB-DCT)-based modeling technique in order to impose a fixed dimension; in contrast, the REW was modeled as Legendre orthonormal polynomials.

In the SB-DCT method, the SEW magnitude spectrum is first divided into K number of frequency sub-bands,

$$\begin{bmatrix} c_{k,1} \\ \vdots \\ c_{k,J_k} \end{bmatrix}^T = \begin{bmatrix} u_{SEW}(n, J_{k-1}+1) \\ \vdots \\ u_{SEW}(n, J_{k-1}+J_k) \end{bmatrix}^T, \tag{4}$$

$$1 \le k \le K, \tag{5}$$

where $c_{k,j}$ denotes the j-th SEW magnitude of the k-th sub-band; J_k denotes a length of the k-th sub-band that satisfies the following condition:

$$\sum_{k=1}^{K} J_k = P(n)/_2, \tag{6}$$

where $P(n)/_2$ denotes the length of the SEW. Each sub-band is then transformed with the DCT as follows:

$$C_{k,m} = \frac{1}{J_k} \sum_{j=1}^{J_k} c_{j,k} \cos\left(\frac{\pi}{J_k}(j-0.5)(m-1)\right), \tag{7}$$

$$1 \le m \le J_k, \tag{8}$$

where $C_{k,m}$ represents the m-th DCT coefficient of the k-th sub-band. As the DCT has good decorrelation and energy compactness properties, most SEW magnitude-related information is concentrated in the first few coefficients. Therefore, the first coefficient ($C_{k,1},\ 1 \le k \le K$) of each sub-band, which is defined as an SB-DCT coefficient, is used for the DNN training process. On the other hand, the remaining coefficients in each sub-band are stochastically generated by Gaussian random variables in the synthesis step.

However, this approach cannot fully exploit the advantages of the DCT. Note that the first coefficient of DCT is the mean (average) of the input signal, whereas the remaining coefficients are the weights of the corresponding orthogonal basis functions at different frequencies, which can efficiently represent the shape of input signal. Although the first DCT coefficient of each sub-band (SB-DCT coefficient) can be well modeled by the DNN, stochastically generated random variables have limited

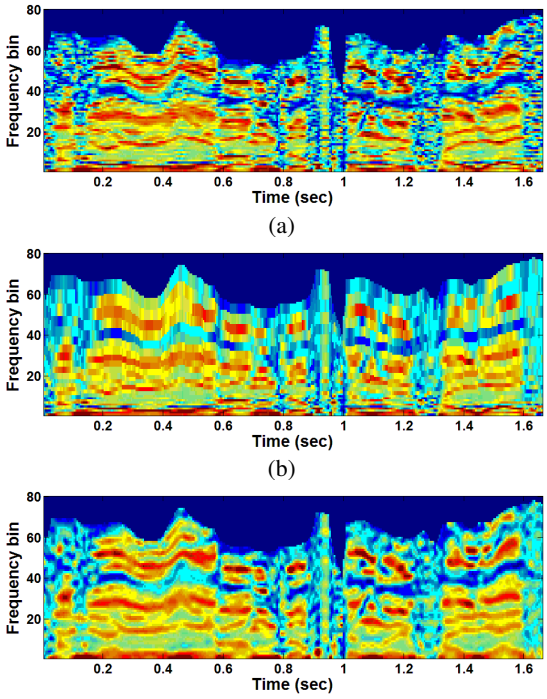

Figure 1: SEW magnitude spectrum (a) obtained from recorded speech, (b) reconstructed by SB-DCT coefficients, (c) reconstructed by FB-DCT coefficients (18 coefficients in both SB-DCT and FB-DCT).

information to represent the remaining coefficients. As a result, the reconstructed sub-band causes a blocky effect, which cannot produce a continuous SEW shape when all the sub-bands are combined into the full-band signal. Furthermore, because the DNN training process does not consider the discontinuity between the adjacent sub-bands, the quality of synthesized speech becomes perceptually non-transparent.

2.3. Full-band DCT-based parameterization method for the ITFTE vocoder

To alleviate the aforementioned problems of the previous approach, this paper proposes a full-band DCT (FB-DCT)-based parameterization method for the ITFTE vocoder. Rather than transforming the sub-band SEW magnitude into the DCT-domain, the DCT is applied to the full-band SEW magnitude as follows:

$$C_m = \frac{1}{J} \sum_{\phi=1}^{J} u_{SEW}(n,\phi) \cos\left(\frac{\pi}{J}(\phi-0.5)(m-1)\right), \tag{9}$$

$$1 \le m \le J, \tag{10}$$

where $J = P(n)/_2$ denotes the length of the SEW at n-th frame. By setting the higher-order DCT coefficients to zero, the full-band SEW magnitude is simply reconstructed by applying an inverse DCT as follows:

$$\widehat{u}_{SEW}(n,\phi) = \widetilde{C}_1 + 2\sum_{m=1}^{J} \widetilde{C}_m \cos\left(\pi(\phi-0.5)(m-1)\right), \tag{11}$$

$$1 \le \phi \le J, \tag{12}$$

$$\widetilde{C}_m = \begin{cases} C_m, & 1 \le m \le K \\ 0, & otherwise \end{cases}, \tag{13}$$

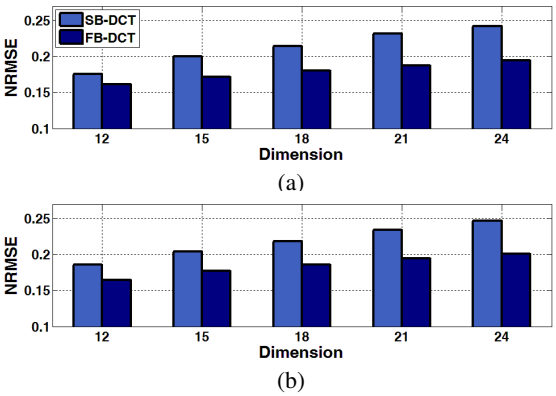

Figure 2: Average NRMSE results for trained SB-DCT and FB-DCT coefficients for (a) Korean male and (b) US female speaker.

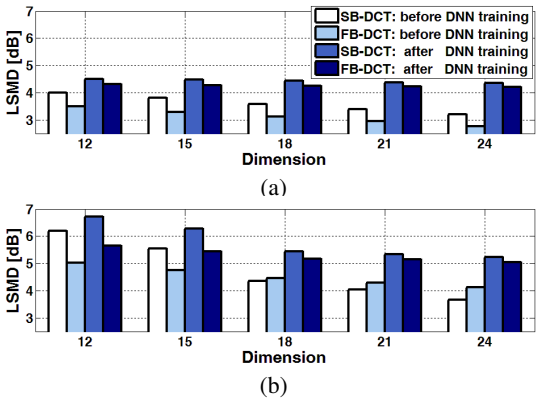

Figure 3: Average LSMD results for reconstructed SEW magnitude from SB-DCT and FB-DCT coefficients for (a) Korean male and (b) US female speaker.

where the lower K-th order non-zero DCT coefficients are defined as full-band DCT (FB-DCT) coefficients. Since the FB-DCT coefficients consist of comparatively large sets of smoothed orthogonal basis functions, the reconstructed SEW magnitudes have more accurate shape than the those derived from the SB-DCT coefficients.

Figure 1-(a) depicts an example of the SEW magnitude extracted from a recorded speech signal; Figure 1-(b) shows the reconstructed SEW magnitude from the SB-DCT coefficients (analysis/synthesis without DNN training, when there are 18 sub-bands; $K = 18$). As shown in these figures, the full-band information of SEW magnitude can be represented by using several SB-DCT coefficients, but there exist discontinuities at the sub-band boundaries, which degrades the perceptual quality of the synthesized speech. Figure 1-(c) depicts an example of the reconstructed SEW magnitude from the FB-DCT coefficients (lower 18 coefficients; $K = 18$). Compared with the previous approach shown in Figure 1-(b), it is clear that the proposed algorithm recovers the SEW magnitude very well.

3. Advantages of full-Band DCT-based parameterization method

This section describes the advantages of the proposed algorithm when it is combined with the DNN training process. The SB-DCT and FB-DCT coefficients extracted from Korean and English speech corpora are used to verify the effectiveness of the proposed method. The trainability of the DNN is measured in terms of normalized root mean square error (NRMSE) with respect to different dimensions of the SB-DCT and FB-DCT coefficients as follows:

$$NRMSE = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\frac{x(n,k)-\widehat{x}(n,k)}{x(n,k)}\right)^2}, \tag{14}$$

where N and K denote the number of frames and the dimension of parameters, respectively; $x(n,k)$ denotes either the SB-DCT or FB-DCT coefficients extracted from the recorded speech; $\widehat{x}(n,k)$ denotes the coefficients generated by the trained DNN. Note that the standard DNN-based SPSS system is used to generate the SB-DCT and FB-DCT coefficients [1], for which setup details are shown in section 4.1.

Figure 2 represents NRMSE results for each system with respect to different dimensions of SB-DCT and FB-DCT coefficients. For both the Korean male (upper) and US female (lower) speakers, the NRMSEs show consistent results, in which the FB-DCT method contains smaller training errors than the SB-DCT method. Therefore, it is clear that the proposed system is more robust than the previous approach if the parameters are trained with the DNN framework.

To further analyze the effect of DNN training, a log-SEW magnitude distance (LSMD) in dB between the original and reconstructed SEW magnitude is measured as follows:

$$LSMD\ [dB] = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{1}{J}\sum_{\phi=1}^{J}\left(20\,log\frac{u_{sew}(n,\phi)}{\widehat{u}_{sew}(n,\phi)}\right)^2}, \tag{15}$$

where $J = P(n)/_2$ denotes the length of the SEW at n-th frame; $u_{sew}(n,\phi)$ and $\widehat{u}_{sew}(n,\phi)$ denote the SEW magnitude extracted from the recorded speech and reconstructed by either the SB-DCT or FB-DCT coefficients, respectively.

Figure 3 shows LSMD results for the Korean male (upper) and US female (lower) speakers with respect to the dimension of parameters. Before performing the DNN training (i.e., only the process of analysis/synthesis), the reconstruction errors from the FB-DCT processing are consistently small in the case of the Korean male speaker, whereas for the US female speaker, the SB-DCT accurately reconstructs SEW magnitude when the dimension increases. However, following DNN training, all the LSMDs of the FB-DCT were smaller than those of the SB-DCT for both the Korean male and US female speakers. These results support the NRMSE analysis, which showed that the proposed system is more robust to the DNN training process than the previous method, resulting in more accurate reconstruction of SEW magnitude.

4. Experiments

4.1. Experimental setup

Two phonetically and prosodically rich speech corpora (Korean and English) were used in our experiment, where each corpus was recorded by the professional Korean male and US female speaker, respectively. The speech signals were sampled at 16 kHz and quantized by 16 bits. Each database was divided into

Table 1: Number of utterances for different sets.

Speaker	Training	Development	Test
Korean male	2500 (∼3.2 h)	200	200
US female	5114 (∼6.0 h)	200	200

Table 2: Objective test results for previous and proposed systems.

Speaker	System	LSD (dB)	F0 RMSE (Hz)	LSMD (dB)	LRMD (dB)	v/uv error rate (%)
Korean male	Previous SB-DCT	4.02	17.89	4.43	5.75	6.76
	Proposed FB-DCT	4.01	18.63	4.25	5.74	6.76
US female	Previous SB-DCT	3.21	17.66	5.44	5.20	4.88
	Proposed FB-DCT	3.21	17.71	5.17	5.11	4.90

three parts, namely training, development, and test sets. The size of each set is shown by the number of utterances in Table 1.

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 the line spectral frequencies (LSFs) for spectral parameters. 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) [13]. On the other hand, the 18-dimensional and 4-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 for the DNN training process.

In the DNN training step, all of these parameters, together with their time dynamics [14], composed the 193-dimensional output feature vectors. The corresponding input feature vectors of the Korean and English databases included 210 and 346-dimensional contextual information, consisting of: 203 and 311 binary features for categorical linguistic contexts; 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 comprised 4 layers of 1024 units and the sigmoid activation function was used for the hidden and output layers. In the training, the weights were first initialized by using a layer-wise back-propagation (BP) pre-training method [15], and then trained by using the BP procedure based on the mini-batch stochastic gradient descent algorithm [16]. The minibatch size was 128 and RMSProp was performed to determine the learning rate [17]. The training and test procedures were implemented by using the computational network toolkit (CNTK) [18]. Since the DNN could not predict the variance used for a speech parameter generation (SPG) algorithm [19], we used pre-computed global variances of output features from all the training data.

In the synthesis step, the mean vectors of all the output feature vectors were first predicted by DNN, and then the SPG algorithm was applied to generate smooth trajectories of acoustic parameters. To reconstruct the SEW, the generated SB-DCT or FB-DCT SEW coefficients were converted to SEW magnitude; a fixed phase spectrum drawn from speech was used for the SEW phase [7]. In contrast, the REW magnitude was converted from the generated polynomial or FB-DCT REW coefficients, whereas its 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 spectral clarity, LSF-sharpening [20, 21] and formant-enhancing [22] filters were also applied.

The previous and proposed systems shared same acoustic parameters such as LSFs, F0, energy, and gain. The difference between two systems was the method of parameterizing the SEW and REW magnitudes. In the previous system, the SB-DCT and polynomial coefficients were used to parameterize the SEW and REW magnitude, respectively; in the proposed system, the FB-DCT coefficients were used to parameterize the magnitudes of both SEW and REW.

4.2. Objective and subjective test results

In the objective test, we compared distortions in acoustic parameters obtained from the original speech with those estimated by DNNs. The metrics for measuring distortion were log-spectral distance (LSD) for LSFs (dB), root mean square error (RMSE) for F0 (Hz), and LSMD for SEW (dB), log-REW magnitude distance (LRMD) for REW (dB), and v/uv error rate (%).

The test results for the previous and proposed systems are shown in Table 2. The LSD, F0 RMSE, and v/uv error rate are similar in both systems, since same parameters were trained by the DNN. Note that the proposed system contains larger F0 RMSE than the previous system. This could be due to the initialization of the DNN weights. In the case of the LSMD, the proposed system has much smaller error, as described in the previous section. The LRMD is also smaller in the proposed system, which verifies that the FB-DCT method is also more effective than the polynomial curve-fitting method for parameterizing the REW magnitude.

To evaluate the perceptual quality of the proposed system, mean opinion score (MOS) listening tests were performed. In the tests, eight native Korean and eight native US listeners were asked to make quality judgments of synthesized Korean and English utterances, respectively (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent). Twenty utterances were randomly selected from the test set in both Korean and English database, which were then synthesized by the previous and the proposed systems. The results of the MOS test (Table 3) show that the proposed system provides superior perceptual quality than that of the previous system, where the proposed system achieved 3.06 and 3.63 MOS for the Korean and English database, respectively.

Table 3: MOS test results with 95% confidence interval for previous and proposed systems.

System	Korean male	US female
Recorded speech	4.96±0.04	4.33±0.18
Analysis/synthesis: SB-DCT	4.50±0.14	3.84±0.18
Analysis/synthesis: FB-DCT	4.48±0.13	3.92±0.18
DNN-SPSS: SB-DCT	2.62±0.19	3.54±0.15
DNN-SPSS: FB-DCT	3.06±0.21	3.63±0.15

5. Conclusion

A parameterization method for improved time-frequency trajectory excitation (ITFTE) vocoder has been proposed. To overcome the discontinuity problem of the previous approach, we utilized a full-band DCT (FB-DCT) method to model the ITFTE parameters. The proposed FB-DCT method was confirmed to be robust to the DNN training process; thus, the excitation signal was reconstructed accurately when the ITFTE parameters were combined with the DNN training process. Subjective listening tests also confirmed the superiority of the proposed system over the previous method.

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. Qian, Y. Fan, W. Hu, and F. K. Soong, "On the training aspects of deep neural network (DNN) for parametric TTS synthesis," in *Proc. ICASSP*, 2014, pp. 3829–3833.

[3] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "The effect of neural networks in statistical parametric speech synthesis," in *Proc. ICASSP*, 2015, pp. 4455–4459.

[4] O. Watts, G. E. Henter, T. Merritt, Z. Wu, and S. King, "From HMMs to DNNs: Where do the improvements come from?" in *Proc. ICASSP*, 2016, pp. 5505–5509.

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

[6] W. B. Kleijn, "Continuous representations in linear predictive coding," in *Proc. ICASSP*, 1991, pp. 201–204.

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

[8] E. L. Choy, "Waveform interpolation speech coder at 4 kb/s," Ph.D. dissertation, McGill University Montreal, Canada, 1998.

[9] 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.

[10] 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.

[11] G. Kubin, B. S. Atal, and W. B. Kleijn, "Performance of noise excitation for unvoiced speech," in *Proc. SCT Workshop*, 1993, pp. 35–36.

[12] J. C. Hardwick and J. S. Lim, "A 4.8 kbps multi-band excitation speech coder," in *Proc. ICASSP*, 1988, pp. 374–377.

[13] W. B. Kleijn and K. K. Paliwal, *Speech coding and synthesis*. Elsevier, 1995.

[14] 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.

[15] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. ASRU Workshop*, 2011, pp. 24–29.

[16] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," DTIC Document, Tech. Rep., 1985.

[17] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Networks for Machine Learning*, 2012.

[18] 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.

[19] 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.

[20] Y.-J. Wu, "Research on HMM-based speech synthesis," Ph.D. dissertation, University of Science and Technology of China, 2006.

[21] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang, "USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method," in *Proc. Blizzard Challenge Workshop*, 2006.

[22] 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.