Effective Spectral and Excitation Modeling Techniques for LSTM-RNN-Based Speech Synthesis Systems

Eunwoo Song, Frank K. Soong, and Hong-Goo Kang

***Abstract*—In this paper, we report research results on modeling the parameters of an improved time-frequency trajectory excitation (ITFTE) and spectral envelopes of an LPC vocoder with a long short-term memory (LSTM)-based recurrent neural network (RNN) for high-quality text-to-speech (TTS) systems. The ITFTE vocoder has been shown to significantly improve the perceptual quality of statistical parameter-based TTS systems in our prior works. However, a simple feed-forward deep neural network (DNN) with a finite window length is inadequate to capture the time evolution of the ITFTE parameters. We propose to use the LSTM to exploit the time-varying nature of both trajectories of the excitation and filter parameters, where the LSTM is implemented to use the linguistic text input and to predict both ITFTE and LPC parameters holistically. In the case of LPC parameters, we further enhance the generated spectrum by applying LP bandwidth expansion and line spectral frequency-sharpening filters. These filters are not only beneficial for reducing unstable synthesis filter conditions but also advantageous toward minimizing the muffling problem in the generated spectrum. Experimental results have shown that the proposed LSTM-RNN system with the ITFTE vocoder significantly outperforms both similarly configured band aperiodicity-based systems and our best prior DNN-trainecounterpart, both objectively and subjectively.**

***Index Terms*—Speech synthesis, improved time-frequency trajectory excitation vocoder, long short-term memory, recurrent neural network.**

I. Introduction

As THE accuracy of the acoustic-modeling process has increased following the introduction of deep neural networks (DNNs), systems for statistical parametric speech synthesis (SPSS) have been popularly used for many applications [1]. Because the centralized DNN framework compactly models complicated dependencies between input contexts and output acoustic features, it not only improves the accuracy of acoustic models but also alleviates the over-smoothing problems of generated parameters [1]–[4]. Accordingly, many researchers have analyzed the effects of various vocoding techniques on the quality of synthesized speech [5]–[8], as the speech analysis/synthesis module itself determines the best achievable quality of text-to-speech (TTS) systems.

In many vocoding schemes, the parametric representation of speech is largely based on the source-filter theory of speech production [5]. This model assumes that the speech signal is composed of the vocal folds-related excitation signal and the vocal tract-related transfer function. The latter has been relatively well approximated by a digital filter, i.e., a linear prediction (LP) filter; however, the former has been overly simplified, i.e., a pulse or noise (PoN) [6], creating buzzy or whispery sounds. To reduce this kind of unnatural synthetic quality, several types of mixed excitation models have been adopted [7], [8]. By separating the whole frequency band into several fixed sub-bands, the excitation signal of each sub-band is represented by either PoN or band aperiodicities (BAP) [9]–[11]. However, it is still challenging to reliably model the excitation components as these methods cannot fully represent the time-varying periodicity of various types of phonetic information.

Our previous research proposed an improved time-frequency trajectory excitation (ITFTE) vocoder [12]; and implemented DNN-based SPSS systems with it [13], [14]. In the ITFTE vocoder, a pitch-dependent excitation signal is first obtained by applying an LP inverse filter to an input speech signal, which is then represented in the time-frequency domain. This excitation, i.e., the TFTE signal, is further decomposed into slowly evolving waveform (SEW) and rapidly evolving waveform (REW) components. The SEW, obtained from a frequency-dependent low-pass filtering of the TFTE signal through the temporal axis, represents the quasi-periodic/harmonic components of the excitation. In contrast, the REW represents the remaining noise-like components. Therefore, utilizing the SEW/REW decomposition enables effective estimation of the periodicity distribution of the excitation signal in the time-frequency domain.

In the DNN-based SPSS system, these excitation parameters are used to compose the output acoustic feature vectors together with the spectral envelope parameters. The whole neural net

Manuscript received March 29, 2017; revised July 12, 2017; accepted August 12, 2017. Date of publication August 29, 2017; date of current version September 20, 2017. This work 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 associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sin-Horng Chen. *(Corresponding author: Hong Goo Kang.)*

E. Song was with the Microsoft Research Asia, Beijing 100080, China. He is now with the Department of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749 South Korea, and also with NAVER Corp., Seongnam 13561, Korea (e-mail: sewplay@dsp.yonsei.ac.kr).

F. K. Soong is with the Microsoft Research Asia, Beijing 100080, China (e-mail: frankkps@microsoft.com).

H-G. Kang is with the Department of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, South Korea (e-mail: hgkang@yonsei.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2017.2746264

then learns a regression model to map the input linguistic feature vectors to the output vectors, which minimizes mean-square errors between the target and predicted output vectors. Thanks to its ability to effectively capture non-linear relationships, the DNN-based SPSS system has been verified to perform better than the conventional hidden Markov model (HMM)-based systems [13]. However, it is limited in representing sequential characteristics, because it adopts a *frame-by-frame* modeling method. Note that the SEW and REW represent the time-varying periodicity of phonetic information [15]–[17]; therefore, it is very important to accurately model the dynamic nature of evolving characteristics in the time domain. Although a speech parameter generation (SPG) algorithm alleviates the problem by producing a smooth trajectory of the parameters [18], this causes harmonic structures to be smeared at the phoneme boundaries [19].

To address these limitations, we propose acoustic modeling methods based on long short-term memory (LSTM) recurrent neural networks (RNNs). LSTM architectures are used to build a non-linear function that maps contextual information to the corresponding ITFTE excitation and spectral envelope parameters. It is well known that the LSTM structure has the capability to model temporal sequences and their long-term dependencies [20], [21]; therefore, it has demonstrated better quality than the DNN framework in speech-synthesis applications [22]–[24]. In this framework, the LSTM memory blocks inherently train the temporal variation of the strongly correlated features, such as SEW in consecutive frames. As a result, acoustic modeling methods based on LSTM reconstruct ITFTE and spectral parameters more accurately than the DNN-based method previously used. To verify the effectiveness of the proposed system, we analyze the trainability of the excitation parameters compared with that of the previously used approach. The findings confirm that the proposed system significantly reduces estimation errors in the generated parameters.

In addition to adopting LSTM-based acoustic models, we also introduce several enhancement techniques, especially for the spectral parameters: linear prediction bandwidth expansion (LP-BWE) and line spectral frequency-sharpening (LSF-S) filters. The first filter is used to prevent unnatural LP filter conditions. Increasing the order of LP coefficients improves the spectral intelligibility of the synthesized speech [25]; however, it sometimes makes the synthesis filter unstable. Since estimation errors in the excitation signal result in unstable outputs in the spectral filtering process, it is necessary to broaden the bandwidth at the overly emphasized spectral peaks. Similar to early speech coders [26], [27], the LP-BWE filter is applied to shift the poles of the synthesis filter radially toward the z-plain origin, thereby significantly reducing unstable frame rates in the cases of both analysis/synthesis and LSTM training. The second filter is used to alleviate the over-smoothing problem of generated spectral parameters. It operates based on the relationship between the LSFs and their corresponding spectral structures [28], [29]. Two consecutive LSFs near the spectral peaks tend to be close, whereas those near the spectral valleys are far from each other [30]. In the proposed LSF-S filter, the location of the target LSF is first compared with its adjacent left and right LSFs and then moved toward the closer one. As this process can reconstruct sharper spectral peaks and valleys, it not only alleviates the muffling problem caused by the statistic-oriented training process, but also enhances spectral clarity in a perceptual listening test.

The remainder of this article describes the key modules for designing the high-quality TTS system, including the ITFTE vocoder for extracting the excitation parameters (Section III), the LSTM structures for improving the acoustic models (Section IV), and the spectral filters for enhancing the perceptual quality of the synthesized speech (Section V). The article also includes the objective and subjective experimental results (Section VI), which confirm that the proposed system significantly outperforms both STRAIGHT and WORLD (D4C edition [31]) vocoders trained with the same LSTM configuration [10], [11], and our best prior DNN-trained counterpart [13].

II. Relationship to Prior Work

The idea of using an ITFTE vocoder with the SPSS systems is not very new. Statistical models such as the decision-tree-based HMM [12] and DNN [13] have been combined with the ITFTE vocoder, and we have verified its superior quality over other vocoders. However, limitations in modeling the vocoder parameters have been reported by several points: (1) HMMs cannot fully model cross-dimensional correlations between ITFTE parameters, because each state is modeled by a single Gaussian, diagonal covariance output distribution. (2) Having a prohibitively large tree and separating the training data set can lead to over-fitting problems when generating the ITFTE parameters. (3) Substituting the decision-tree-based HMMs into a centralized feed-forward DNN could address prior two problems; however, its frame-by-frame modeling paradigm is not suitable for representing the time-varying characteristics of the ITFTE parameters. (4) Although applying the SPG algorithm alleviates the discontinuity in generated parameters, it aggravates the over-smoothing problem.

To ameliorate the aforementioned issues, our aim here is to use the LSTM architectures to improve the model accuracy of the ITFTE and LPC parameters. There have been prior works in using the LSTM in the TTS applications [22]–[24]. However, our research differs from these studies in several ways: (1) we focus further on the effect of LSTMs in modeling excitation parameters, whereby the trainability represented by a reconstruction error in the excitation signal is analyzed in detail. (2) Our experiments verify the performance of various architectures of neural networks including a DNN, a hybrid system (DNN+LSTM), and a deep LSTM (DLSTM). The synthesis quality of each system is investigated by varying the amount of the training data set. Both the objective and the subjective test results could be usefully referred to when designing similarly configured TTS systems. (3) Regarding the vocoder itself, in a perceptual listening test, the proposed system shows superiority over BAP-based vocoders with the same LSTM model structure.

In addition to the above, we explore the use of spectral enhancements, including LP-BWE and LSF-S filters. Analysis shows that the LP-BWE filter reduces unstable synthesis filter

conditions in high-order LPC and that the LSF-S filter improves perceptual quality in terms of spectral clarity.

III. Improved Time-Frequency Trajectory Excitation Vocoder

In the ITFTE vocoder, an excitation signal is generated by an inverse filtering of an input speech signal with the LP analysis filter. A two-dimensional TFTE signal, $u(n,\phi)$, is used to represent the spectral shape of excitation along the phase axis and the evolution of this shape along the time axis 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 $A_k(n)$ and $B_k(n)$ denote the k-th discrete-time Fourier series coefficients of the excitation signal at the n-th frame; $\phi(m) = 2\pi m/P(n)$ denotes a phase function with a pitch period $P(n)$ [17].

The TFTE is further decomposed into two components by filtering each frequency component along the time-domain axis. The SEW that represents the quasi-periodic/harmonic component is obtained from a low-pass filter (LPF) as follows:

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

where $h(l)$ denotes an L-th order LPF. Beyond the cut-off frequency, the remaining noise-like components are represented by the REW as follows:

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

Therefore, the time-evolving periodicity is efficiently represented by the SEW and REW components, thereby producing the natural shape of the excitation signal.

However, these parameters cannot be directly applied to the DNN/LSTM training process, because their parametric dimensions change over time due to variation in the pitch period. In the ITFTE vocoder, the SEW and REW components are parameterized by the modeling technique based on full-band (FB) discrete cosine transform (DCT) to impose a fixed dimension [14]. In the analysis step, the SEW magnitude is first transformed into the DCT domain as follows:

$$C_m = \frac{1}{J}\sum_{\phi=1}^{J} u_{\text{SEW}}(n,\phi)\cos\left(\frac{\pi}{J}(\phi-0.5)(m-1)\right), \quad (4)$$

$$1 \le m \le J, \quad (5)$$

where $J = P(n)/2$ denotes the length of the SEW (one-half of one pitch period) at the n-th frame. As the DCT has good properties of decorrelation and energy compactness, most information related to SEW magnitude is concentrated in the first few coefficients. Therefore, the lower K-th order coefficients, defined as FB-DCT-SEW coefficients, are used for the DNN/LSTM training process. By setting the higher-order DCT coefficients to zero in the synthesis step, the full-band SEW magnitude is simply reconstructed by applying an inverse DCT, as follows:

$$\widehat{u}_{\text{SEW}}(n,\phi) = \widetilde{C}_1 + 2\sum_{m=2}^{J} \widetilde{C}_m \cos\left(\pi(\phi-0.5)(m-1)\right), \quad (6)$$

$$1 \le \phi \le J, \quad (7)$$

$$\widetilde{C}_m = \begin{cases} C_m, & 1 \le m \le K \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

In the case of REW, it is modeled via either a method of estimating power contour [17] or the FB-DCT approach, the latter of which has been verified by previous experiments as exhibiting better modeling performance [14].

IV. LSTM-Based Acoustic Modeling Methods for ITFTE Parameters

This section describes the previously used DNN-based and the proposed LSTM-based acoustic modeling methods that use ITFTE parameters as outputs. As the parameters represent the time-evolving characteristics of the excitation signal, their temporal characteristics should be accurately modeled in the training process. The LSTM architecture is beneficial for incorporating the sequential nature of the ITFTE parameters into the acoustic model.

A. DNN-Based Acoustic Modeling Method

Fig. 1(a) depicts the previous feed-forward DNN-based acoustic modeling framework for the ITFTE parameters [13]. In this system, rich contexts are first analyzed and then converted to a sequence of input linguistic features, $\mathbf{x} = (x_1, \ldots, x_N)$, which 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, $\mathbf{y} = (y_1, \ldots, y_N)$, are composed of the ITFTE and spectral parameters with their time dynamics. The pairs of input and output features are then used to train the weights of the DNN using a backpropagation (BP) procedure [32].

The DNN-based architecture is not suitable for representing temporal variation in the ITFTE parameters, because it uses only a *frame-by-frame* modeling method. As shown in the dependency graph of Fig. 1(a), there is no temporal interaction between adjacent frames, which results in discontinuities in the synthesized signal. Although applying the SPG algorithm relieves this problem [18], it cannot fully reflect the temporal variation in the strongly correlated features, including consecutive SEW frames [19]. As the time-frequency representation of SEW and REW directly affects the quality of the synthesized speech, it is desirable to design more elegant modeling methods that can effectively capture their sequential characteristics.

B. LSTM-Based Acoustic Modeling Methods

This study adopts an LSTM framework as the acoustic modeling method for the ITFTE and spectral parameters. The LSTM

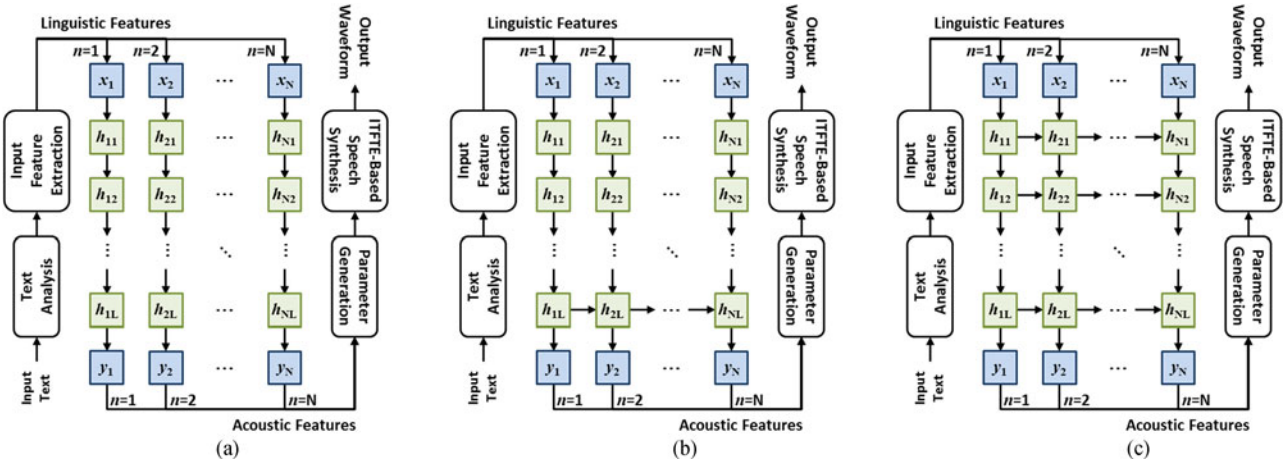

Fig. 1. Framework of the acoustic modeling methods for the ITFTE parameters: (a) previously used DNN, (b) proposed hybrid system of DNN and LSTM, and (c) proposed DLSTM.

contains special units called *memory blocks* in the recurrent hidden layer, and each memory block contains a self-connected *memory cell* having multiplicative units called *gates*. During the training process, the memory cell maintains its state over time, and the gates regulate the information flow to decide what to retain and what to erase from memory. By combining the previous state, current memory, and input sequence, the network inherently estimates the temporal information of the output sequence.

The training procedure is almost the same as that for the DNN framework, but the weights are updated based on the modified BP procedure, called *backpropagation through time* (BPTT) [33]. As the parameters are shared by all time steps in the network, the gradient at each output depends on both current and previous time steps. To train the weights successfully, the BPTT first unfolds the LSTM into the feed-forward network through time, and then trains the unfolded network using the BP procedure.

Fig. 1(b) depicts a hybrid neural network architecture [22], where the DNN and LSTM layers are positioned to the input linguistic and output acoustic layers, respectively. The figure shows the dependency between adjacent frames at the last hidden layer, which enables the compact modeling of the temporal correlations of the consecutive output features. As a result, ITFTE parameters are accurately predicted during the generation procedure.

To further improve accuracy in prediction, as shown in Fig. 1(c), a deep LSTM (DLSTM) architecture is built by stacking multiple LSTM layers. Note that 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 [34]. As the inputs go through more nonlinear operations per time step, it is possible to find not only a sequential nature in time but also the mapping relationship between the features of the linguistic input and the acoustic output.

C. Trainability Analysis

This section describes the advantages of employing acoustic modeling methods based on the LSTM for ITFTE parameters. To verify the effectiveness of the proposed systems, the trainability of the DNN and LSTMs (both hybrid and DLSTM) is measured in terms of the log-SEW magnitude distance (LSMD) as follows:

$$\text{LSMD}\,[dB] = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{1}{J}\sum\nolimits_{\phi=1}^{J}\left(20\log\frac{u_{\text{sew}}(n,\phi)}{\widehat{u}_{\text{sew}}(n,\phi)}\right)^2}, \tag{9}$$

where $J = P(n)/2$ denotes the length of the SEW at the n-th frame; $u_{\text{sew}}(n,\phi)$ and $\widehat{u}_{\text{sew}}(n,\phi)$ denote the SEW magnitude extracted from the recorded speech and reconstructed by the generated FB-DCT-SEW coefficients from either the trained DNN or LSTMs, respectively. Fig. 2 represents the LSMD results for each system with respect to the various dimensions of FB-DCT-SEW coefficients. The LSMDs show consistent results, in which the LSTM-based systems have much smaller training errors than does the system based on DNN. Among the LSTM systems, the DLSTM performs best in terms of accurately reconstructing SEW magnitudes. Therefore, the proposed LSTM structures are advantageous for modeling the ITFTE parameters compared to the previously used DNN-based system.

D. Determination of Parameter Dimensions

Since the loss in the higher-order DCT coefficients can result in smoothed SEW spectra, increasing the dimension of FB-DCT-SEW coefficients is beneficial to reconstructing more accurate SEW spectra. However, as Fig. 2 shows, there is no dramatic improvement at a certain dimension. The reason is a limited information of the higher-order DCT coefficients, which implies that their detailed characteristics could not be represented by the DNN/LSTM models. Therefore, we did not impose the entire dimension to represent SEW components, and a certain number

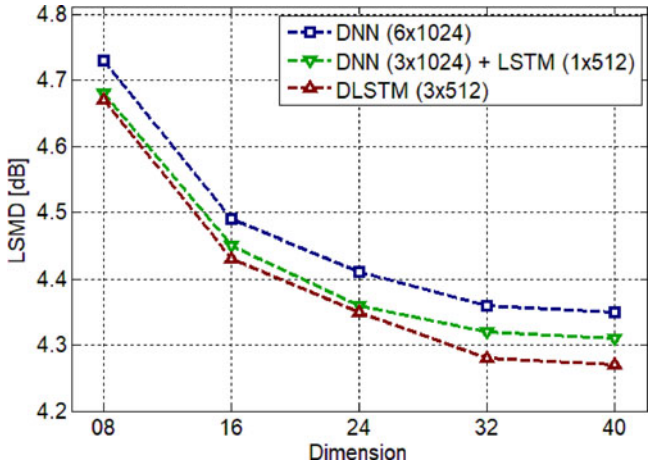

Fig. 2. Average LSMD results for the reconstructed SEW magnitude from the trained DNN or LSTMs (hybrid and DLSTM) with respect to the various dimensions of FB-DCT-SEW coefficients. The DNN consists of six feed-forward hidden layers with 1,024 units; the hybrid system consists of three feed-forward hidden layers with 1,024 units and one LSTM hidden layer with 512 memory blocks, respectively; the DLSTM consists of three LSTM hidden layers with 512 memory blocks. More setup details are shown in Section VI-A.

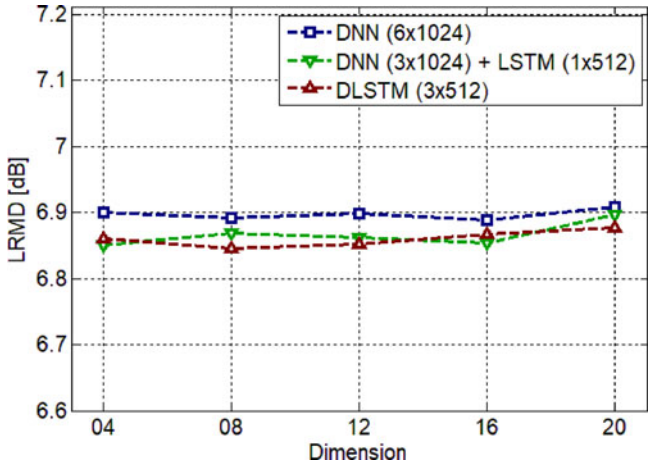

Fig. 3. Average LRMD results for the reconstructed SEW magnitude from the trained DNN or LSTMs with respect to the different dimensions of FB-DCT-REW coefficients. In contrast with the LSMD results shown in the Fig. 2, the reconstructed REW spectra were not highly affected by the DCT dimensions.

of DCT coefficients (i.e., 32-nd or 40-th order) was enough to represent the SEW spectra.

Similarly, we investigate the effect of the FB-DCT-REW dimension on reconstruction error by analyzing the log-REW magnitude distance (LRMD). Fig. 3 depicts the LRMD results for each system with respect to the various dimensions of FB-DCT-REW coefficients. As the REW represents the randomly distributed noise components, the LRMDs from the REW spectra were not highly affected by the DCT dimensions. Consequently, a few DCT coefficients (i.e., 4-th order) were sufficient to estimate the envelope of noise spectra.

V. Spectral Enhancements for Speech Synthesis

In the previous section, we showed the effectiveness of LSTM-based acoustic modeling methods in terms of trainability of ITFTE parameters. To further improve the synthesis quality of the entire framework, this section introduces spectral enhancement techniques, including the LP-BWE and LSF-S filters.

A. LP-BWE Filter

The LP-BWE filter is designed to prevent unnatural peaks in the LP spectral envelope. The high-order LP analysis/synthesis improves spectral intelligibility, i.e., 40-th or 64-th order for a 16 kHz or 48 kHz sampling rate, respectively [25]. However, as the bandwidth of the spectral peak becomes too narrow, the estimation errors in the excitation signal are sometimes boosted after the spectral filtering process. To avoid generating discontinuous speech segments, the BWE technique is applied to the LP analysis and synthesis filters. This technique broadens the bandwidth at the overly emphasized spectral peaks by radially shifting the spectral poles toward the origin as follows:

$$H_1(z) = 1 - \sum_{i=1}^{p} (a_i/\gamma)\, z^{-i}, \tag{10}$$

$$H_2(z) = \frac{1}{1 - \sum_{i=1}^{p} (a_i/\gamma)\, z^{-i}}, \tag{11}$$

where $H_1(z)$ and $H_2(z)$ represent bandwidth-expanded LP analysis and synthesis filters, respectively; a_i and γ denote the p-th order LP coefficients and the LP-BWE factor, respectively. To verify how the LP-BWE filter addresses unstable filter conditions, we analyze the unstable frame rates (UFR; %) in cases of both analysis/synthesis and DLSTM training. The number of unstable frames N_{uf}, which is counted when the consecutive LSFs are extremely close [35], is defined as follows:

$$N_{\text{uf}} = \sum_{n=1}^{N} f\left(\left\{\sum_{i=2}^{p} f\left(lsf_{n,i} - lsf_{n,i-1} < D\right)\right\} > 0\right), \tag{12}$$

where $l_{n,i}$ denotes the i-th LSF coefficient at the n-th frame; D denotes a threshold for minimum distance; $f(\bullet)$ denotes a logical function defined as follows:

$$f(\bullet) = \left\{ \begin{array}{ll} 1, & \text{if ``}\bullet\text{'' is true} \\ 0, & \text{otherwise.} \end{array} \right\}. \tag{13}$$

Fig. 4 depicts the UFR results for spectral parameters with respect to various distance thresholds ($D = 10, 20, ..., 80$ Hz). Without the LP-BWE filter, the figure shows many unstable spectral peaks, both extracted from recorded speech and generated from the trained DLSTM. Although the DLSTM generates smoothed spectra, resulting in a smaller UFR, it still contains a large number of unstable frames. In contrast, applying the LP-BWE filter significantly reduces the UFR in cases of both analysis/synthesis and trained DLSTM.

B. LSF-S Filter

The LSF-S filter is used to alleviate the overly smoothed spectral structure caused by the statistical training process. The main reason for its use is to sharpen the spectral peaks and valleys by adjusting the generated LSFs [28], [29].

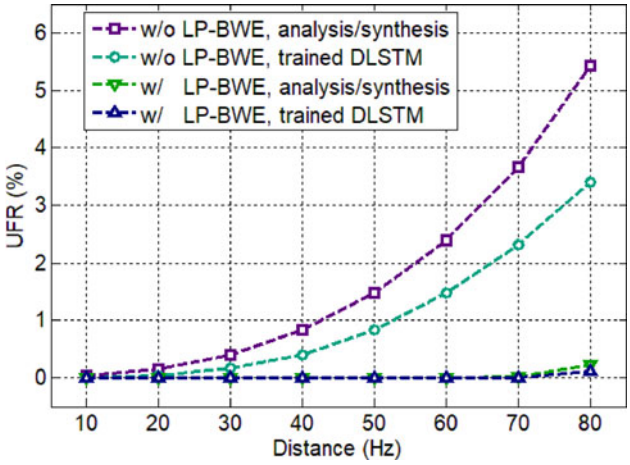

Fig. 4. UFR results for spectral parameters extracted from recorded speech and generated from the trained DLSTM, both with (w/) and without (w/o) the LP-BWE filter, with respect to various distance thresholds. The DLSTM consists of 3 LSTM hidden layers with 512 memory blocks.

Let $\widetilde{l}_i$ denote the i-th LSF coefficient adjusted by the relative distance between two adjacent LSFs as follows:

$$\widetilde{l}_i = \frac{|d_i|^2}{|d_{i-1}|^2 + |d_i|^2}\widehat{l}_{i-1} + \frac{|d_{i-1}|^2}{|d_{i-1}|^2 + |d_i|^2}\widehat{l}_{i+1}, \tag{14}$$

$$d_i = \widehat{l}_{i+1} - \widehat{l}_i, \tag{15}$$

$$1 < i < p, \tag{16}$$

where $\widehat{l}_i$ denotes the i-th LSF coefficient generated from the trained DNN/LSTMs. Then, the sharpened LSF is obtained as follows:

$$l_i^* = \alpha_i \widehat{l}_i + (1 - \alpha_i)\,\widetilde{l}_i, \tag{17}$$

where α_i denotes a frequency-dependent weighting factor.[1] As shown above, the sharpening process shifts the current LSF coefficient closer toward one between the adjacent left and right LSFs, which enables the reconstruction of sharper spectral peaks and valleys.

Fig. 5 depicts the all-pole spectrum obtained from recorded speech and generated by the trained DLSTM both with and without applying the LSF-S filter. As shown in Fig. 5(b), the formant structure was smoothed because of the statistical averaging of the DLSTM training process. However, the LSF-S filtering output makes the spectral peak/valleys much clearer, thereby not only alleviating over-smoothed spectral envelopes but also synthesizing more-natural speech signals, the results of which are confirmed in the next section.

VI. Experiments

A. Experimental Setup

The experiments used a phonetically and prosodically balanced speech corpus recorded by a Korean male professional speaker. The speech signals were sampled at 16 kHz, and each sample was quantized by 16 bits. In total, 3,300 utterances (about 10 hours) were used for training, 330 utterances (about 1 hour) were used for validation, and another 330 utterances not included in either the training or validation steps were used for testing.

In the analysis step, the frame length was set to 20 ms, and the spectral and excitation parameters were extracted at every 5 ms. The 40-dimensional LP coefficients were extracted and converted to the LSFs for spectral parameters. To prevent unnatural spectral peaks in the LP analysis/synthesis filter, each coefficient $(a_i, i = 1, ..., 40)$ was multiplied by the LP-BWE factor (0.981^i). In contrast, the 32-dimensional and 4-dimensional FB-DCT-SEW and FB-DCT-REW coefficients, respectively, were extracted for the excitation parameters. The fundamental frequency (F0), energy, and v/uv information were also extracted for the DNN and LSTM training processes.

In the training step, all these parameters, together with their time dynamics [36], consisted of 235-dimensional output feature vectors. The corresponding input feature vectors included 268-dimensional contextual information consisting of 261 binary features for categorical linguistic contexts and seven numerical features for numerical linguistic contexts. Before training, both the input and output features were normalized to have zero-mean and unit-variance.

Table I summarizes the architectures of the DNN, hybrid, and DLSTM-based acoustic modeling methods. In the DNN-based system, the hidden layers comprised six DNN layers with 1,024 units. The weights were first initialized using a layer-wise BP pre-training method [37], and then trained using the BP procedure based on the mini-batch stochastic gradient descent algorithm. The mini-batch size was 128; RMSProp was performed to determine the learning rate [38].

The hybrid system consisted of three DNN layers with 1,024 units and one unidirectional LSTM layer with 512 memory blocks. The DNNs and LSTM were connected to the linguistic input layer and the acoustic output layer, respectively. The weights were randomly initialized and 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 speed, tens of utterances were randomly selected for each update and used to update the weights simultaneously.

The architecture of the DLSTM-based system consisted of three unidirectional LSTM layers with 512 memory blocks. Note that their training procedures were the same as those in the hybrid system. For all methods, the training and test procedures were implemented using the computational network toolkit (CNTK) [39].

In the synthesis step, the mean vectors of all output feature vectors were first predicted by DNN or LSTMs. Then, with pre-computed global variances of output features from all the training data, the SPG algorithm was applied to generate smooth trajectories of acoustic parameters. 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 [16], whereas the REW phase was

[1]In this research, we exploited the exponentially decaying weighting factor $(\alpha_i = 0.8^{i-1}, 1 < i < p)$ to make the filter preserve the low-frequency spectrum itself and sharpen the higher frequency components. Note that this approach is advantageous because it avoids generating unnatural spectral peaks in the low frequency region, as discussed in the previous section.

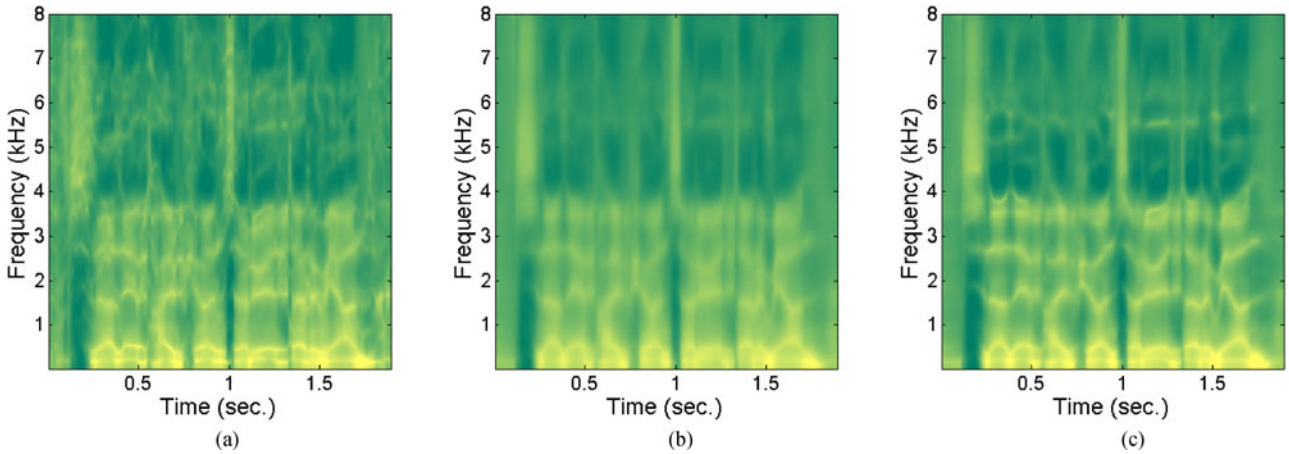

Fig. 5. Comparison between all-pole spectra (a) obtained from recorded speech (analysis/synthesis), (b) generated by trained DLSTM without the LSF-S filter, and (c) with the LSF-S filter. The DLSTM consists of three LSTM hidden layers with 512 memory blocks.

TABLE I
ARCHITECTURES OF DNN, HYBRID, AND DSLTM-BASED ACOUSTIC MODELING METHODS, AND THEIR MODEL PARAMETER SIZES

System	Type of layer(s)	Number of layers	Number of units (memory blocks)	Model size
DNN	DNN	6	1,024	5.76 M
Hybrid	DNN LSTM	3 1	1,024 512	5.64 M
DLSTM	LSTM	3	512	5.92 M

The hybrid system consists of three DNNs and one LSTM layer, connected to the input and output layers, respectively.

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, the LSF-S filter was also applied to the generated LSFs.

B. Objective and Subjective Test Results

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

Fig. 6 shows the test results for the architectures of various acoustic modeling methods, with respect to the various amount of data sets (2, 4, ..., 10 hours). The findings can be analyzed as follows: (1) In all systems, as the number of training hours increased, the overall estimation performances gradually improved. (2) The LSTM-based systems had a limitation to model the temporal sequences when the amount of training data set was small (2 hours), resulting in relatively large errors in the LSF and F0 parameters compared to those in the DNN-based system. (3) With the sufficient amount of training data set (more than 4 hours), it is clear that all the parameters generated by the LSTM-based systems had much smaller estimation errors than those generated by the DNN-based systems. As the accuracy of these parameters is closely related to the quality of synthesized speech, the perceived quality of synthesized speech from the proposed system is expected to be better than that from the baseline system.

To evaluate the perceptual quality of the proposed system, A-B preference and mean opinion score (MOS) listening tests were performed.[2] In the preference tests, 12 native Korean listeners were asked to rate the synthesized speech by quality preference. In total, 20 utterances were randomly selected from the test set, and then synthesized by the architectures of the various acoustic modeling methods. To verify vocoding performance, additional DLSTM-based systems using the STRAIGHT (DLSTM-STRAIGHT) and WORLD (DLSTM-WORLD) vocoders were also included [10], [11]. Note that only the excitation parameters (e.g., SEW and REW) were replaced with the BAPs; whereas all other parameters were kept the same as those of the ITFTE vocoder.

Table II shows the preference test results, whose trends can be analyzed as follows: (1) The perceptual quality of the synthesized speech from the LSTM-based acoustic modeling methods was significantly better than its DNN-based counterpart (Tests 1 and 2). This confirms that employing the recurrent layer effectively captured the temporal nature of the ITFTE parameters, enabling the accurate prediction of acoustic features. (2) The DLSTM-based method showed a higher perceptual quality than the hybrid method (Test 3), implying that the deeper architecture of the LSTM achieved better performance. (3) The modeling method based on ITFTE excitation provided better perceptual quality than BAP-based approaches[3] (Tests 4 and 5),

[2]The tests were performed in an acoustically isolated room using a Sennheiser HD650 headphone.

[3]In the BAP-based approach, an excitation signal at the synthesis step is constructed by a weighted sum of a pulse train and white noise depending on the BAP value. In contrasts, in the ITFTE-based approach, the harmonic and noise components are represented in the time-frequency domain (SEW and REW, respectively) and separately parameterized for the DNN/LSTM frameworks. As the SEW and REW have opposed properties (i.e. strongly correlated along the time axis and randomly distributed, respectively) it is beneficial to individually control those parameters to model the statistical characteristics.

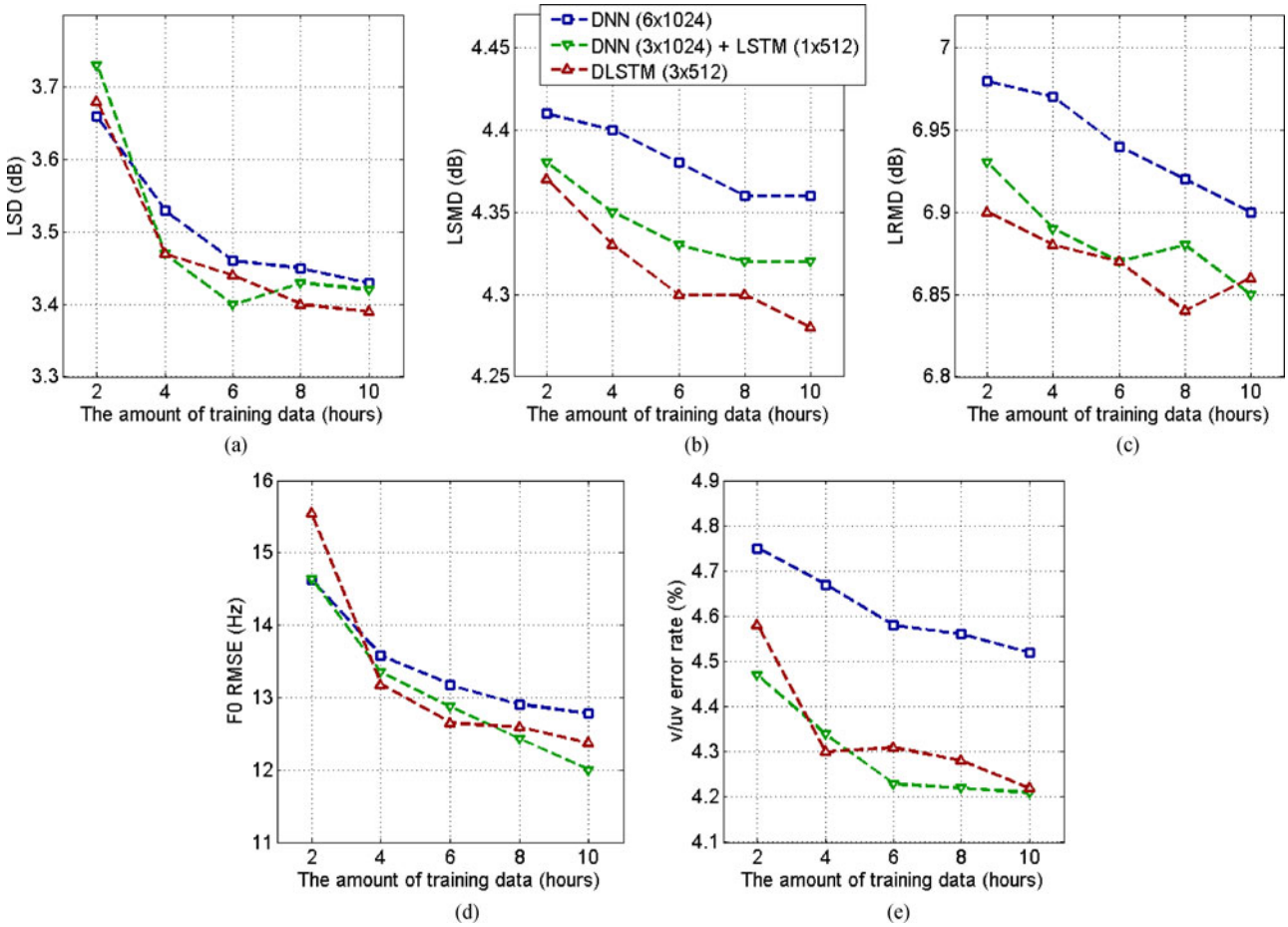

Fig. 6. Objective test results for the architectures of the various acoustic modeling methods, with respect to various amount of training data sets: (a) LSD (dB), (b) LSMD (dB), (c) LRMD (dB), (d) F0 RMSE (Hz), and (e) v/uv error rate (%).

TABLE II
SUBJECTIVE PREFERENCE TEST RESULTS (%) BETWEEN THE SYNTHESIZED SPEECH SAMPLES

Test index	DNN (5.76 M)	Hybrid (5.64 M)	DLSTM (5.92 M)	DLSTM-STRAIGHT	DLSTM-WORLD	Neutral	p-value
Test 1	11.7	**45.8**	–	–	–	42.5	$< 10^{-12}$
Test 2	7.5	–	**51.7**	–	–	40.8	$< 10^{-21}$
Test 3	–	17.5	**36.3**	–	–	46.3	$< 10^{-4}$
Test 4	–	–	**97.1**	1.7	–	1.3	$< 10^{-133}$
Test 5	–	–	**90.4**	–	3.8	5.8	$< 10^{-84}$

The systems that achieved significantly better preference at the $p < 0.01$ level are in bold font.

demonstrating that elaborate reconstruction of the excitation was also very important in the TTS applications.

The setups for the MOS test were the same as those for the preference test, except that listeners were asked to make quality judgments about the synthesized speech using the following possible responses: 1 = Bad, 2 = Poor, 3 = Fair, 4 = Good, 5 = Excellent. The test also evaluated the effectiveness of the proposed spectral enhancement filters, with speech samples being generated by the DLSTM-based system as follows: (1) without any spectral enhancement filter (DLSTM-a), (2) with an LP-BWE filter only (DLSTM-b), (3) with an LSF-S filter only (DLSTM-c), and (4) with both filters (DLSTM).

The test results shown in Fig. 7 confirm the effectiveness of each system in several ways: (1) The LSTM-based systems (hybrid and DLSTM) provided quality superior to that of the DNN-based systems (DNN), as was the case with the results of the preference test. (2) Employing the LP-BWE filter improved the perceptual quality of synthesized speech. The wider confidence interval of the systems without the LP-BWE filter (DLSTM-a and DLSTM-c compared to DLSTM-b and DLSTM, respectively) implies that the LP filter contained many

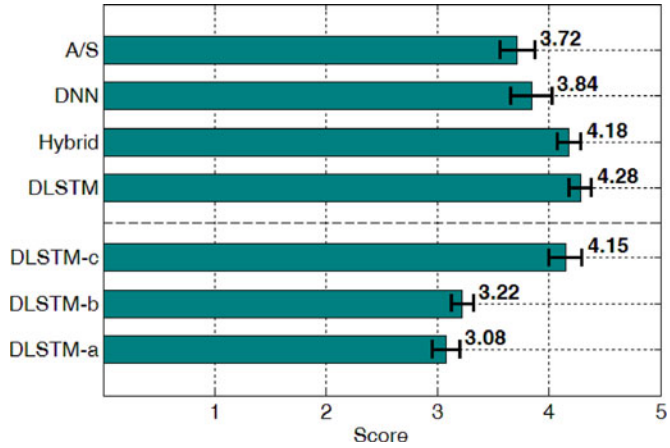

Fig. 7. Subjective MOS test results with 95% confidence interval for the architectures of the various acoustic-modeling methods (A/S = analysis/synthesis case). In the DLSTM-based architecture, four systems were evaluated: without any spectral enhancement filter (DLSTM-a), with an LP-BWE filter only (DLSTM-b), with an LSF-S filter only (DLSTM-c), and with both filters (DLSTM).

frames with unnatural spectral peaks, resulting in degradation of naturalness and inconsistent results. (3) Exploiting the LSF-S filter significantly improved the perceptual quality, such that most listeners were satisfied with its quality in terms of spectral clarity. In particular, combining the DNN/LSTM framework with the LSF-S filter (DNN, hybrid, DLSTM, and DLSTM-c) outperformed even the analysis/synthesis (A/S) case. This was because the unwanted artifacts produced by the analysis/synthesis process were statistically excluded during the generation process, and possible over-smoothing problems were alleviated by the LSF-S filter. Consequently, the proposed hybrid and DLSTM approaches achieved 4.18 and 4.28 MOS, respectively.

VII. Conclusion

This article proposed LSTM-RNN structures for modeling the acoustic parameters of a high-quality TTS system based on an ITFTE LPC vocoder. In particular, the research focused on modeling excitation parameters, i.e., ITFTE, in the LSTM framework. The intrinsic shortcomings imposed by the finite window length of a feed-forward DNN framework were shown to be dramatically alleviated with the new LSTMs. The spectral envelope parameters of the LPC filter were also trained with the same LSTM frameworks and further enhanced using LP-BWE and LSF-S filters. Listening tests using the new systems resulted in much higher measures of perceived quality and lower objective distortion. In addition, subjective A-B comparison listening tests confirmed that the ITFTE LPC vocoder performed significantly better than the STRAIGHT and WORLD vocoders when both were trained using the same LSTM-TTS configuration.

Future research includes reducing the synthesis speed. Although both the hybrid and DLSTM-based systems had better perceptual quality than the DNN-based system, they took 8 and 20 times longer, respectively, than the DNN-based system. This may be alleviated by other simplified types of RNN architectures that could not be included in the current framework. Future research will further consider how to effectively reduce the synthesis speed.

References

[1] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 5505–5509.
[5] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Delhi, India: Pearson Education India, 2006.
[6] 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.
[7] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Proc. INTERSPEECH*, 2001, pp. 2263–2266.
[8] 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.
[9] A. V. McCree and T. P. Barnwell III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 242–250, Jul. 1995.
[10] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1997, pp. 1303–1306.
[11] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.
[12] 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.
[13] 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.
[14] 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. 874–878.
[15] W. B. Kleijn, "Continuous representations in linear predictive coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1991, pp. 201–204.
[16] W. B. Kleijn and J. Haagen, "A speech coder based on decomposition of characteristic waveforms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1995, pp. 508–511.
[17] E. L. Choy, "Waveform interpolation speech coder at 4 kb/s," Ph.D. dissertation, McGill University Montreal, QC, Canada, 1998.
[18] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2000, pp. 1315–1318.
[19] E. Song and H.-G. Kang, "Multi-class learning algorithm for deep neural network-based statistical parametric speech synthesis," in *Proc. 24th Eur. Signal Process Conf.*, 2016, pp. 1951–1955.
[20] A. J. Robinson and F. Fallside, "Static and dynamic error propagation networks with application to speech coding," in *Proc. Neural Inf. Process. Syst.*, 1988, pp. 632–641.
[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[22] 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.
[23] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4470–4474.

[24] S. Achanta, T. Godambe, and S. V. Gangashetty, "An investigation of recurrent neural network architectures for statistical parametric speech synthesis," in *Proc. INTERSPEECH*, 2015, pp. 859–863.
[25] Y.-S. Joo, C.-S. Jung, and H.-G. Kang, "Enhancement of spectral clarity for HMM-based text-to-speech systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7840–7843.
[26] J.-H. Chen and A. Gersho, "Real-time vector APC speech coding at 4800 bps with adaptive postfiltering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1987, pp. 2185–2188, pp. 2185–2188.
[27] J. P. Campbell Jr, T. E. Tremain, and V. C. Welch, "The dod 4.8 kbps standard (proposed federal standard 1016)," in *Advances in Speech Coding*. New York, NY, USA: Springer, 1991, pp. 121–133.
[28] Y.-J. Wu, "Research on HMM-based speech synthesis," Ph.D. dissertation, Univ. Science Technol. China, Hefei, China, 2006.
[29] 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.
[30] F. Soong and B. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1984, pp. 37–40.
[31] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Commun.*, vol. 84, pp. 57–65, 2016.
[32] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," DTIC Document, Tech. Rep. ICS-8506, 1985.
[33] R. J. Williams and J. Peng, "An efficient gradient-based algorithm for on-line training of recurrent network trajectories," *Neural Computational*, vol. 2, no. 4, pp. 490–501, 1990.
[34] 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.
[35] N. Sugamura and F. Itakura, "Speech analysis and synthesis methods developed at ECL in NTT-From LPC to LSP," *Speech Commun.*, vol. 5, no. 2, pp. 199–215, 1986.
[36] 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, Feb. 1986.
[37] 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.
[38] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," in *Proc. COURSERA: Neural Netw. Mach. Learning*, 2012, pp. 26–31.
[39] D. Yu *et al.*, "An introduction to computational networks and the computational network toolkit," Microsoft Research, Tech. Rep. MSR-TR-2014-12, 2014.

Eunwoo Song received the B.S. degree in electrical and electronic engineering from the Yonsei University, Seoul, South Korea, in 2010. He is currently working toward the combined M.S. and Ph.D. degrees in electrical and electronic engineering at Yonsei University. Since March 2017, he has been joining with NAVER Corp., Seongnam, South Korea. He served his internships at Microsoft Research Asia, Beijing, China, from 2015 to 2016, and Qualcomm Technologies Inc., San-Diego, CA, USA, in 2016, respectively. His research interests include speech/audio signal processing, speech/audio coding, speech synthesis, and machine learning.

Frank K. Soong received the B.S., M.S., and Ph.D. degrees from National Taiwan University, University of Rhode Island, and Stanford University all in electrical engineering. He is a Principal Researcher and a Research Manager, Speech Group, Microsoft Research Asia (MSRA), Beijing, China, where he works on fundamental research on speech and its practical applications. His professional research career spans over 30 years, first with Bell Labs, USA, then with ATR, Japan, for 2 years, before joining MSRA in 2004. At Bell Labs, he worked on stochastic modeling of speech signals, optimal decoder algorithm, speech analysis and coding, speech and speaker recognition. He has served as a member of the Speech and Language Technical Committee, IEEE Signal Processing Society, and other society functions, including Associate Editor of the IEEE SPEECH AND AUDIO TRANSACTIONS and chairing IEEE Workshop. He published extensively with more than 250 papers and co-edited a widely used reference book, *Automatic Speech and Speech Recognition- Advanced Topics*, Kluwer, 1996. He is a Visiting Professor of the Chinese University of Hong Kong (CUHK) and a few other top-rated universities in China. He is also the co-Director of the National MSRA-CUHK Joint Research Lab. He is an IEEE Fellow "for contributions to digital processing of speech."

Hong-Goo Kang (M'94) received the B.S., M.S., and Ph.D. degrees in electronic engineering from Yonsei University, Seoul, South Korea, in 1989, 1991, and 1995, respectively. He was a Senior Member of Technical Staff at AT&T Labs—Research, from 1996 to 2002. In 2002, he joined the Department of Electrical and Electronic Engineering, Yonsei University, where he is currently a Professor. His research interests include speech signal processing, array signal processing, and machine learning.