IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 1, 2018, accepted May 1, 2018, date of publication May 7, 2018, date of current version June 5, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2833890

Sequential Deep Neural Networks Ensemble for Speech Bandwidth Extension

BONG-KI LEE[1], KYOUNGJIN NOH[2], JOON-HYUK CHANG[2], (Senior Member, IEEE), KIHYUN CHOO[3], AND EUNMI OH[3]

[1]CTO Division, LG Electronics Co., Ltd., Seoul 06763, South Korea
[2]Hanyang University, Seoul 04763, South Korea
[3]Digital Media and Communication Research and Development Center, Samsung Electronics Co., Ltd., Seoul 06734, South Korea

Corresponding author: Joon-Hyuk Chang (jchang@hanyang.ac.kr)

This work was supported in part by the Institute for Information & Communications Technology Promotion through the Korea Government (MSIT) under Grant 2017-0-00474, and in part by the Intelligent Signal Processing for AI Speaker Voice Guardian.

ABSTRACT In this paper, we propose a subband-based ensemble of sequential deep neural networks (DNNs) for bandwidth extension (BWE). First, the narrow-band spectra are folded into the high-band (HB) region to generate the high-band spectra, and then the energy levels of the HB spectra are adjusted using the DNN-based on the log-power spectra feature. For this, we basically build the multiple DNNs, which is responsible for each subband of the HB and the DNN ensemble is sequentially connected from lower to higher subbands. This sequential structure for the DNN ensemble carries out the denoising and HB regression to better estimate the HB energy levels. In addition, we use the voiced/unvoiced (V/UV) classification to differently apply the DNN ensemble depending on either V/UV sounds. To demonstrate the performance of the proposed BWE algorithm, we compare it with a speech production model-based BWE system and a DNN-based BWE system in which the log-power spectra in the HB are estimated directly. The experimental results show that the proposed approach provides better speech quality than conventional approaches.

INDEX TERMS Bandwidth extension, sequential deep neural network, ensemble, log-power spectra, regression, voiced/unvoiced classification.

I. INTRODUCTION

In many digital speech transmission systems, the bandwidth of telephone speech remains limited to the narrow-band (NB), which has a frequency range from 300 Hz to 3.4 kHz, especially when terminals and part of the network have not been equipped with wide-band (WB) capability. However, users become aware of the limited intelligibility of NB speech when they try to understand unknown words or names. These restrictions can be overcome with an artificial bandwidth extension (BWE) algorithm, which extends the speech bandwidth using only information available from NB speech [1]. Originally, the BWE algorithms proposed in the literature can be realized in two different ways: with auxiliary transmissions and without transmitting side information [2]. A recent proposal for BWE using side information was standardized by 3rd generation partnership project (3GPP) enhanced voice service (EVS) codec [3], which allocates additional bits for a special structure on the encoder side. However, the most challenging application of BWE is improving NB telephone speech at the receiving end without transmitting any auxiliary information. Therefore, in this work, we focus on developing BWE without side information so that no modifications are necessary for the existing network infrastructure and so processing can be performed in the terminal device at the receiving end.

The BWE systems aiming at in this work can be basically classified into the algorithms with speech production models, also known as the source-filter model of human speech production, and without ones [4]. Many BWE algorithms have been developed based on the speech production model, motivated by previous studies of the human speech production system. Two steps are used for speech production model-based BWE system: estimation of the WB spectral envelope and extension of the excitation signal. Various methods have been presented in the literature to estimate the WB spectral envelope from the NB one. For instance, in [5], Pulakka *et al.* proposed Gaussian mixture model (GMM)-based approaches to model the joint distribution of WB and NB features, estimating the spectral envelope parameters of WB speech from the NB features using a Bayesian minimum

mean-square error (MMSE) estimate. The idea of using a codebook to recover WB spectral information was proposed in the work of Unno and McCree [6]. Another popular technique to model the joint distribution of features and retrieve the missing spectral components is based on the hidden Markov model (HMM) [7]; the BWE system being modeled is assumed to be a Markov process with unobserved states. Pulakka and Alku [8] devised a way to train a neural network to estimate the mel spectrum in the extension band based on features derived from the NB signal. Other techniques used to extend excitation, including spectral shifting and folding [9], modulation, function generator [10], and non-linear transformation [11] of NB excitation have been proposed in which the WB excitation signal is used as the input for the estimated WB filter when reconstructing the WB speech signal.

On the other hand, the BWE systems without the speech production model have been developed in different ways. In the extrapolation method or non-linear mapping [12], the high-band signal derived from a high pass filter passes through a shaping filter and is added to the original band-pass signal. For instance, Yasukawa [12] proposed a non-linear processing-based expansion method that uses rectification to produce the extension band of spectral components. Non-linear processing yields low computational costs, but poor extension quality, so it does not reproduce the high band well and also needs subjective power level adjustments. There has also been an attempt to use the spectral folding method followed by modification of the high frequency magnitude spectra using spline curves [13], where the spline control points are determined using the genetic algorithm. However, genetic algorithm-based spline control points have a limitation in that it is difficult to estimate the HB energy levels exactly, especially for sibilant sounds, which sometimes produces uncomfortable sounds. Also, Choo *et al*. [14] designed a way to use an advanced spectral envelope predictor in which the excitation signal of the WB is estimated using spectral double shifting, which is regarded as a simplified version of the adaptive spectral double shifting introduced in [15]. The spectral envelope of the NB is extended to the WB based on the spectral shape of the NB determined using a GMM-based classifier. However, the extension of the spectral envelope is processed in a heuristic manner and is not verified in noisy environments. Recently, Li and Lee [16] proposed a novel BWE algorithm using a deep neural network (DNN) that is widely used in popular classification and regression tasks, particularly in automatic speech recognition [17], voice activity detection [18], sound event classification [19], and packet loss concealment [20]. In this approach, the HB magnitude spectra are estimated directly from the NB magnitude spectra, which causes artifacts, including annoying sounds, when the regression of the HB spectra fails. Thus, the direct mapping method turns out inadequate for BWE systems. There are also previous studies that combine the speech production model with deep learning where the spectral envelope information of WB such as line spectral frequencies (LSFs) are estimated by various DNN structures, respectively [21]–[24]. However,

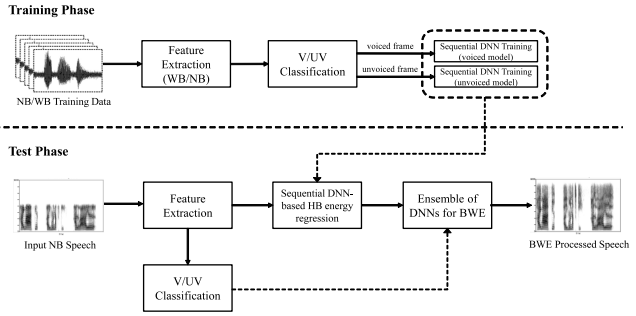

FIGURE 1. Flow chart of the proposed BWE algorithm.

speech model parameters such as LSFs are difficult to estimate with DNN because those are known to be sensitive to regression errors caused by DNN [20].

In this paper, we present a novel BWE algorithm that originally uses the DNN-based regression approach. Our study, for the first time as far as we know, proposes the DNN-based ensemble algorithm using voiced and unvoiced (V/UV) sounds classification to estimate the energies of the HB spectra. For this, We first apply spectral folding technique to the boundary between the NB and HB to maintain the spectral harmonics of the HB and then establish deep generative models of the log-power spectra features, which are widely used in regression tasks. The folded spectra of the NB to the HB are then smoothed to mitigate the sharpness of sounds. In practice, the HB is split into four subbands, and each subband is distinctly assigned to a separate DNN by which the log-power spectra of each subband are estimated in a sequential fashion. Specifically, the first subband's DNN model is fed with the log-power spectra of the NB, and the first DNN output is then fed into the second DNN. Note that this step is repeatedly accomplished up to the last DNN, which aims at estimating the subband energies. In addition, separate DNNs are designed for V/UV sounds classification, allowing us to refine DNN ensembles to V/UV conditions. In a test phase, the DNN being responsible for the V/UV classification offers the probability of voiced and unvoiced sounds at each frame and then uses that probability to combine the DNN ensembles on a frame-by-frame basis. We extensively evaluate the proposed BWE system in terms of objective and subjective measures and found it to produce better results than conventional BWE methods. The rest of this paper is organized as follows: Section II introduces the proposed BWE method based on DNNs, Section III presents simulation results, and Section IV presents our conclusions.

II. PROPOSED DNN-BASED BANDWIDTH EXTENSION ALGORITHM

In this section, we fully describe our proposed BWE system, which uses a subband energy level-based HB regression with a sequential DNN structure including both training and test phases. Furthermore, V/UV classification-based DNN ensemble is proposed as shown in Fig. 1, which exhibits the

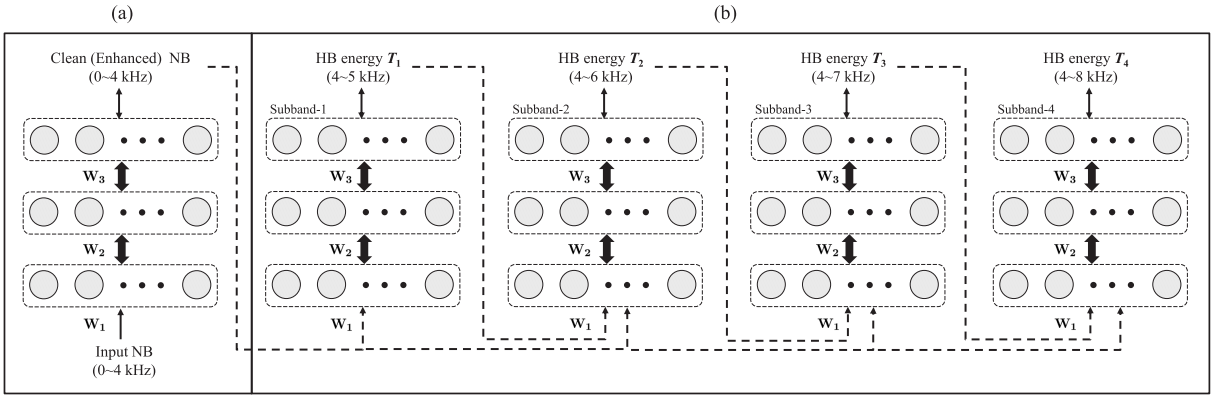

FIGURE 2. The proposed sequential DNN structure consists of DNNs for (a) denoising and (b) HB energy regression.

feature extraction, denoising, V/UV classification, sequential DNN training, the DNN ensemble, and signal synthesis.

A. FEATURE EXTRACTION

In the training phase of the proposed BWE system, the feature extraction used for the DNNs in both V/UV classification and BWE is processed. We use the log-power spectra in the discrete Fourier transform (DFT) domain, known to be well suited for DNN-based regression tasks, as the feature in this work. For feature extraction, we first perform the short-time Fourier transform (STFT) to obtain the DFT coefficients for each windowed frame such that

$$Y^f(k)=\sum_{m=0}^{M-1} y(m)h(m)e^{-j2\pi km/M}, \quad k=0,1,\ldots,M-1 \quad (1)$$

where k and M are the frequency bin index and window length, respectively, and $h(m)$ and f denote the window function and frequency domain, respectively. After the STFT, the log-power spectra are given as

$$Y^l(k) = \log|Y^f(k)|^2, \quad k = 0,1,\ldots,K-1 \quad (2)$$

where $K = M/2+1$ and l denotes the log-power spectra domain. For $k = K,\ldots,M-1$, $Y^l(k)$ is obtained using the symmetric property given by $Y^l(k) = Y^l(M-k)$; thus the dimension of the log-power spectra is given as $M/2+1$. As for the WB signal, $Y^l(k)$ is further separated into a low-frequency spectrum, $Y_L^l = [Y^l(0),\ldots,Y^l(M/4)]$ and a high-frequency spectrum, $Y_H^l = [Y^l(M/4+1),\ldots,Y^l(M/2)]$ where Y_H^l is to be recovered by the DNN-based BWE algorithm. Similar to the log-power spectra, the phase of the DFT domain can be defined as follows:

$$Y^p(k) \triangleq \angle Y^f(k), \quad k = 0,1,\ldots,K-1 \quad (3)$$

where p denotes the phase domain.

As for the WB signal, $Y^p(k)$ is separated into $Y_L^p(k)$ and $Y_H^p(k)$ in the same way like its corresponding magnitude $Y^l(k)$ do. The original WB signals (in the frequency range 0 Hz to 8 kHz) and the NB signals (decoded by the AMR-NB coder [25] after down-sampling) are used for the features. When setting the features, our BWE system attempts to extend the NB signal into the original WB one, which is limited to 8 kHz, unlike the AMR-WB coder limiting to 7 kHz [26].

B. SEQUENTIAL DNN TRAINING

We propose the subband-based sequential DNN for the BWE system as shown in Fig. 2, where the proposed sequential DNN module consists of five DNNs: one for denoising as proposed by Xu *et al.* [27] and four for the subband energy level regression of the HB. Subband processing splits speech into a number of different smaller frequency band and each band is processed independently for which local information is fully considered distinctly [28]. Four is chosen as the number of the subbands in this work to consider the trade-off between the computational complexity and regression performance.

First, when accomplishing denoising, clean and noisy NB features, decoded by the AMR-NB coder, are used for the first DNN input while the target is replaced by the clean NB features. Then, the first DNN output, the enhanced NB feature, is used as the next DNN input for the energy level regression at the HB. For the sequential training, the energy levels of the HB extracted from the WB signal are first used in the target features. Then, the first subband DNN output is then fed into the next DNN input, and that process is repeated until the last subband. Note that not only the previous DNN output but also the first denoising DNN output are conveyed into each subband DNN, which can be termed as multiple ensembles of serial modules. For this, the energy level of the HB is divided into t ($< M/4$) sub-levels, which have average values ($M/4t$) of consecutive frequency bins as follows:

$$y_n = \frac{\sum_k Y^l(k)}{M/4t}, \quad \frac{M}{4}+1 \le k \le \frac{M}{4}+\frac{M}{4t}\cdot n, \quad n = 1,2,\ldots,t. \quad (4)$$

Such y_n allows the target vector of the v-th subband energy level $\boldsymbol{T}_v$ to satisfy

$$\boldsymbol{T}_v = \{y_1, y_2, \ldots, y_{\frac{tv}{4}}\}, \quad v = 1,2,3,4. \quad (5)$$

In practice, we employ deep belief networks (DBNs) [29] for pre-training to initialize the weights and biases of the DNNs; each DNN is a feed-forward neural network with many hidden layers mapping the input features to output features where the features are normalized to zero mean and unit variance. Next, the pre-training of the DNN is carried out in an unsupervised manner that uses a contrastive divergence (CD) approximation as the objective criterion [30]. Once the pre-training is finished, the fine-tuning [31] is performed in a supervised manner. In the fine-tuning process, an MMSE-based back-propagation algorithm is used to minimize the error, which is widely used under the regression tasks [20]. When given an n-dimensional input vector $\boldsymbol{x}$ and model parameters $\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{b}\}$, the final output vector of the m-th subband through multiple nonlinear hidden layers is derived as follows:

$$\begin{aligned}\hat{\boldsymbol{T}}_v(\boldsymbol{x}, \boldsymbol{\theta}) &= \hat{\boldsymbol{T}}_v(\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{b}) \\ &= (y_1, y_2, \ldots, y_{\frac{tv}{4}}) \\ &= \boldsymbol{W}^{(L)}\phi^{(L)}(\boldsymbol{W}^{(L-1)}\phi^{(L-1)}(\cdots \boldsymbol{W}^{(1)}\phi^{(1)}(\boldsymbol{W}^{(0)}\boldsymbol{x} \\ &\quad + \boldsymbol{b}^{(0)}) + \boldsymbol{b}^{(1)}) + \boldsymbol{b}^{(L-1)}) + \boldsymbol{b}^{(L)} \end{aligned} \tag{6}$$

where $\hat{\boldsymbol{T}}_v$ denotes the estimated v-th subband energy level; $\boldsymbol{W}^{(l)}$ and $\boldsymbol{b}^{(l)}$ denote the weight and bias terms between two adjacent layers, the l-th and $(l-1)$-th layers, respectively; and, $\phi^{(l)}$ denotes the activation function of the l-th hidden layer. Note that all activation functions use the logistic function as stated in [18]. For the DNN training using mini-batches, the MMSE is used between the estimated and target subband energy levels for the objective criterion, as given by

$$E_v = \frac{1}{N}\sum_{n=1}^{N}\left(\hat{\boldsymbol{T}}_v^n(\boldsymbol{x}, \boldsymbol{\theta}) - \boldsymbol{T}_v^n(\boldsymbol{x}, \boldsymbol{\theta})\right)^2, \quad v = 1, 2, 3, 4 \tag{7}$$

where E_v is the mean squared error of the v-th subband energy level and N represents the mini-batch size. Then, the updated estimated weights $\boldsymbol{W}$ and bias $\boldsymbol{b}$ of each DNN, with a learning rate λ, can be computed iteratively, as follows:

$$(\boldsymbol{W}^l, \boldsymbol{b}^l) \leftarrow (\boldsymbol{W}^l, \boldsymbol{b}^l) - \lambda\frac{\partial E_m}{\partial(\boldsymbol{W}^l, \boldsymbol{b}^l)}, \quad 1 \le l \le L+1 \tag{8}$$

with L indicating the total number of hidden layers and $L+1$ representing the output layer. The proposed sequential DNN is used to estimate the HB spectral shape for BWE in a manner similar to that used in the training process. For example, in Fig. 2, the energy level of the estimated first subband, $\hat{\boldsymbol{T}}_1$, which is the second DNN output, is fed into the third DNN input with the enhanced NB feature to estimate the energy level of the second subband, $\hat{\boldsymbol{T}}_2$. Subsequently, all the energy levels of the HB are estimated until the last DNN in the sequential DNN structure, so that $\hat{\boldsymbol{T}}_4$ yields the final output of the sequential DNN. To prevent overfitting during the training phase, the denoising DNN output, namely, enhanced NB features are fed into all inputs of the other DNNs. The proposed BWE algorithm, which adopts the denoising and the sequential DNN structure, offers more exact outcomes in the energy level regression than a structure using a single DNN to improve the speech quality in the BWE system. The ensemble structure adopting the V/UV classification to the BWE system will be described in the next subsections.

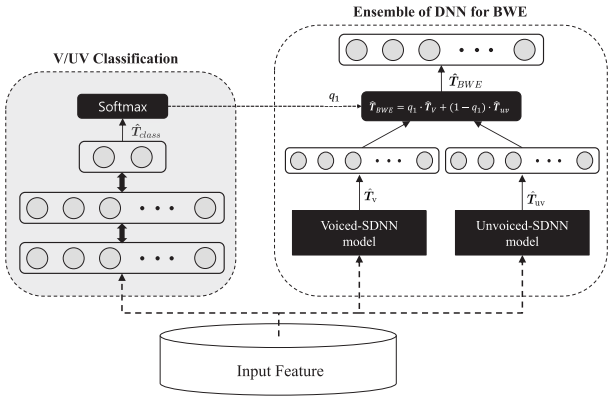

FIGURE 3. The proposed DNN ensemble structure using the V/UV classification.

C. V/UV CLASSIFICATION

In general, speech can be classified into voiced and unvoiced sounds in which voiced speech has relatively higher energy than unvoiced speech and contains periodicity, called the pitch, so that it has a large effect on speech quality. On the other hand, unvoiced speech looks like random noise without periodicity. Because each speech type is clearly distinct, our BWE algorithm is presented to work with V/UV classification. Accordingly, as shown in Fig. 3, the log-power spectra features extracted from the speech samples are first classified as voiced or unvoiced sounds using the V/UV classifier, which uses the DNN in a separated fashion. When training the DNN, the log-power spectra from the NB speech decoded by the AMR-NB coder are used as the input for the DNN that uses V/UV labels as the target output. Unlike sequential DNN training, the V/UV classification DNN training uses a conjugate gradient (CG)-based back-propagation algorithm to minimize a cross-entropy error [32]. The DNN-based V/UV classification test is performed in a similar manner to the training process by which the log-power spectra of noisy NB speech are fed into the DNN input. Given a binary classification problem, the estimated DNN output $\hat{\boldsymbol{T}}_{\text{class}}(\boldsymbol{x}, \boldsymbol{\theta}) = \{y_1, y_2\}$ is fed into the softmax function to obtain the probabilistic soft output q_j, as given by

$$q_j = \frac{\exp(y_j)}{\sum_{i=1}^{2}\exp(y_i)} \tag{9}$$

Finally, the probability of a voiced signal, q_1, and an unvoiced signal, $1 - q_1$, can be obtained and used for the DNN ensemble in the BWE system so that the characteristics of voiced and unvoiced speech can be fully considered.

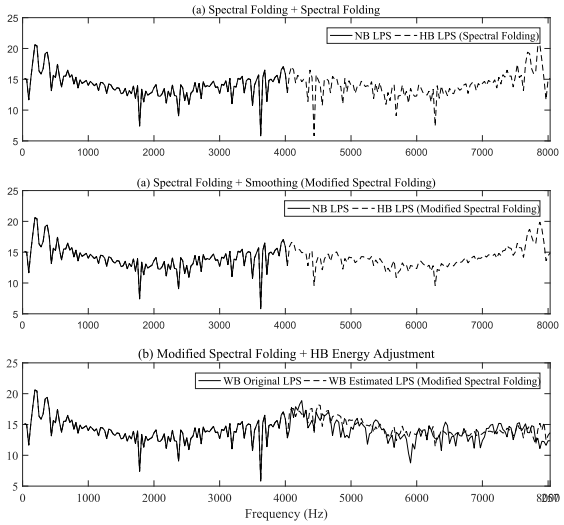

FIGURE 4. Examples of the log-power spectrum representation of (a) spectral folding of NB to HB and (b) smoothing of folded spectra, and (c) HB energy level adjustment.

D. ENSEMBLE OF SEQUENTIAL DNNS FOR BWE

The sequential DNN proposed in the previous subsection is generated for each voiced and unvoiced sequential DNN model: SDNN_v and SDNN_uv, where SDNN_v is trained using the voiced speech frames and SDNN_uv is trained using the unvoiced speech frames, as shown in Fig. 3. Then the final output of the sequential DNN ensemble is softly calculated with q_1 as follows:

$$\begin{aligned}\hat{\boldsymbol{T}}_{BWE}(\boldsymbol{x},\boldsymbol{\theta}) &= q_1 \cdot \hat{\boldsymbol{T}}_\text{v}(\boldsymbol{x},\boldsymbol{\theta}) + (1-q_1)\cdot \hat{\boldsymbol{T}}_\text{uv}(\boldsymbol{x},\boldsymbol{\theta}) \\ &= \{\hat{y_1}, \hat{y_2}, \ldots, \hat{y_t}\} \end{aligned} \quad (10)$$

where $\hat{\boldsymbol{T}}_\text{v}(\boldsymbol{x},\boldsymbol{\theta})$ and $\hat{\boldsymbol{T}}_\text{uv}(\boldsymbol{x},\boldsymbol{\theta})$ are the SDNN_v and SDNN_uv outputs, respectively. In this way, the DNN ensemble for the BWE system can somewhat diminish discontinuities while well representing the characteristics of voiced and unvoiced sounds.

E. SIGNAL SYNTHESIS

One strategy for signal synthesis is the spectral folding technique, by which the NB spectra are folded into the HB region and the HB energies are then adjusted using the sequential DNN ensemble. This technique is preferred because the direct feature mapping method can cause annoying artifacts when it fails to estimate the HB spectra directly. As shown in Fig. 4(a), the enhanced NB spectra are folded into the HB region so that the high frequency spectra are derived such that $\hat{Y}_H^l = \left[\hat{Y}^l(\frac{M}{4}), \hat{Y}^l(\frac{M}{4}-1), \ldots, \hat{Y}^l(0)\right]$. However, in some frequency bands, speech shows a harmonic structure, but, in some frequency bands it exhibits a noise-like feature. Thus, the conventional spectral folding leads to uncomfortable noise even if we use the spectral folding for the voiced segment only. This is why we employ the smoothing scheme to the folded spectra to mitigate the sounds sharpness, which is given by (11). In Fig. 4(b), the folded spectra are then smoothed to mitigate the sharpness of sounds such that

$$\mid \tilde{Y}_{Hs}^l(k) \mid = (1-\alpha) \mid \hat{Y}_H^l(k) \mid + \alpha \mid \tilde{Y}_{Hs}^l(k-1) \mid \quad (11)$$

where $\alpha(=0.4)$ is smoothing parameter. We believe that this method is justified because this algorithm turns out to have very low computational cost and memory requirement unlike correction of HB harmonic structure proposed in previous work [33], which would have made the algorithm much more complicated, was not obviously superior in terms of the perceived quality of the BWE processed speech.

To adjust the energy of the HB spectra, we define the level differences of the n-th sub-level, D_n, between an average of the subband energy in the folded NB spectra into the HB region and the estimated one using the sequential DNN model are defined as follows:

$$D_n = \frac{\sum_{k=1+\frac{M}{4t}(n-1)}^{\frac{M}{4t}n} \tilde{Y}_{Hs}^l(k)}{M/4t} - y_n, \quad n = 1, 2, \ldots, t. \quad (12)$$

Then, the values of the log-power spectra of the HB, $\hat{X}_H^l(k)$, can be obtained as follows:

$$\hat{X}_H^l(k) = \tilde{Y}_{Hs}^l(k) - D_n, \quad 1+\frac{M}{4t}(n-1) \le k \le \frac{M}{4t}n, \quad n = 1, 2, \ldots, t \quad (13)$$

where the log-power spectra of the HB, $Y_H^l(k)$, are subtracted by each level difference D_n, corresponding to the n-th sub-level. Next, the log-power spectra of the WB are derived such that $\hat{Y}_W^l = [Y_L^l, \hat{X}_H^l]$ where the NB spectra are not modified to prevent quality degradation. For example, the energies of the HB spectra are adjusted by the proposed algorithm to match the energies of the original WB spectrum as shown in Fig. 4(c).

As for the phase, an imaged phase of the NB is used for the HB phase as given by

$$\hat{Y}_H^p = \left[-Y_L^p(\frac{M}{4}-1), \quad -Y_L^p(\frac{M}{4}-2), \ldots, -Y_L^p(0)\right] \quad (14)$$

and the WB is then derived such that $\hat{Y}_W^p = [Y_L^p, \hat{Y}_H^p]$. Finally, the WB signals are reconstructed by applying inverse DFT (IDFT) to the reconstructed spectrum, $\hat{Y}_W^f(k) = e^{\hat{Y}_W^l(k)/2} e^{j\angle \hat{Y}_W^p(k)}$, as follows:

$$\hat{y}_w(m) = \frac{1}{M}\sum_{k=0}^{M-1} \hat{Y}_W^f(k) e^{j2\pi km/M} \quad (15)$$

where $\hat{y}_w$ denotes the time-domain signal in the proposed BWE algorithm.

III. EXPERIMENTS AND RESULTS

To assess the performance of the proposed algorithm, we used objective and subjective speech quality measures to compare it with the BWE algorithms in [14], [16], and [21]. For the tests, we evaluated with the standard TIMIT corpus consisting of 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. This speech samples were divided into 4,620 utterances (3.14 hours long) for the

TABLE 1. LSD results from the conventional methods and proposed algorithm.

Noise	SNR (dB)	Method									
		AMR-WB	AMR-NB	Choo [14]	Li [16]	Kang [21]	Proposed BWE	Proposed BWE w/o denoising	Proposed BWE w/o subband	Proposed BWE w/o ensemble	SDNN +direct mapping
clean	-	0.07	0.38	0.27	0.24	0.22	**0.16**	0.16	0.21	0.18	0.21
babble	5	0.28	0.87	0.82	0.77	0.85	**0.64**	0.72	0.74	0.66	0.75
	10	0.16	0.63	0.57	0.54	0.58	**0.46**	0.48	0.51	0.47	0.50
	15	0.11	0.51	0.46	0.38	0.37	**0.25**	0.27	0.28	0.27	0.37
office	5	0.25	0.91	0.85	0.77	0.73	**0.67**	0.72	0.75	0.70	0.76
	10	0.18	0.65	0.61	0.59	0.60	**0.50**	0.54	0.57	0.52	0.56
	15	0.13	0.48	0.42	0.40	0.38	**0.29**	0.31	0.33	0.31	0.38

training set and 1,680 utterances (0.97 hours long) for the test set. In the algorithm we implemented, the WB signals contain components up to 8 kHz, and the NB signals decoded by the AMR-NB codec are up-sampled to 16 kHz. Four types of noise (office, street, car, and white) were used for the training stage, and office and babble noises were used for the test stage to consider both seen and unseen environments, respectively. The noise signals were electrically added to the clean speech at various signal-to-noise ratios (SNRs): 5, 10, and 15 dB. To implement the DFT, we considered frame lengths of 20 ms with 50% overlap-add using the Hamming window and 512-point DFT in which 32 sub-levels ($M = 512$, $t = 32$) are used for the proposed BWE algorithm which were defined empirically. Also, the sequential DNNs and V/UV classification DNN each have three hidden layers, with 512 hidden nodes activated by the sigmoid function. We ran 100 epochs for the pre-training and fine-tuning while training each DNN model. The simulation performed on various experiments including comparison of the speech quality measures and graphical comparisons verified the superiority of the proposed algorithm.

A. SPEECH QUALITY MEASURES

First, we measured the performance by changing the number of the sub-level as 1, 2, 4, 6, and 8 to investigate how the performance changes depending on the number of the sub-levels. For this, objective quality measures such as the log-spectral distance (LSD) [34] and the perceptual evaluation of speech quality (PESQ) [35], which are known to be significantly correlated with perceptual speech quality are used. As shown in Fig. 5, the LSD and PESQ decrease as the number of the sub-levels increases and are saturated at 4, the number of the sub-levels was thus chosen as 4 in the subsequent tests.

Next, we compared the performance of the proposed BWE algorithm to that of the AMR-WB with 23.85 kbps, AMR-NB with 12.2 kbps, and conventional methods including Choo *et al.*'s [14], and Li and Lee's [16], Li and Kang's [21]

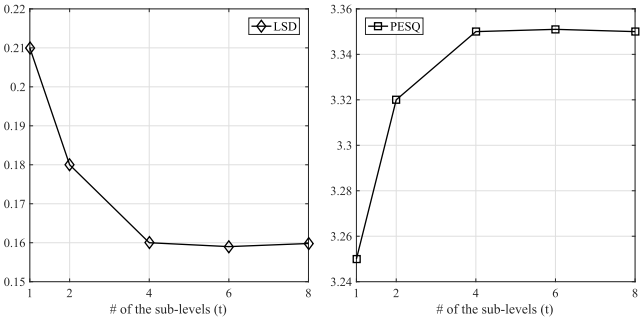

FIGURE 5. LSD and PESQ scores according to the number of the sub-levels (*t*).

algorithms via LSD and PESQ. In addition, we investigated that which part of the proposed BWE structure including denoising, subband-based sequential DNNs, and ensemble DNN using V/UV classifier parts contributes in performance gain. To compare the performance of the normal DNN and SDNNs, we also added a direct mapping of HB spectra using SDNNs (SDNN+direct mapping) like a Li's method. As in Table 1 showing the evaluation result, the LSD score of the proposed BWE method is the lowest among the methods, except for AMR-WB with 23.85 kbps, under both clean and noisy environments. In addition, the PESQ results, summarized in Table 2, were similar to the LSD results: the proposed BWE algorithm consistently outperformed the conventional BWE algorithms in terms of objective speech quality. For the SDNN+direct mapping method, LSD and PESQ performances are slightly better than the Li's method which uses vanilla DNN. As a result, it is noted that the SDNN yields only a slight improvement in performance in case of the direct mapping method. Based on the results of the comparison test of the proposed BWE structure including proposed BWE without denoising, subband, and ensemble, we point out that the subband-based sequential DNNs contributes more to the performance improvement than the ensemble DNN structure

TABLE 2. PESQ results from the conventional methods and proposed algorithm.

Noise	SNR (dB)	Method									
		AMR-WB	AMR-NB	Choo [14]	Li [16]	Kang [21]	Proposed BWE	Proposed BWE w/o denoising	Proposed BWE w/o subband	Proposed BWE w/o ensemble	SDNN +direct mapping
clean	-	4.15	2.66	2.79	2.98	3.09	**3.35**	3.35	3.25	3.27	3.05
babble	5	2.82	2.17	2.31	2.41	2.34	**2.74**	2.53	2.57	2.72	2.43
	10	3.37	2.36	2.46	2.72	2.67	**2.89**	2.76	2.82	2.84	2.72
	15	3.83	2.51	2.62	2.91	2.87	**3.28**	3.13	3.14	3.23	2.93
office	5	2.86	2.12	2.43	2.52	2.36	**2.77**	2.64	2.70	2.71	2.58
	10	3.41	2.38	2.50	2.83	2.64	**2.93**	2.86	2.88	2.88	2.85
	15	3.89	2.53	2.61	2.95	2.98	**3.28**	3.16	3.13	3.23	3.01

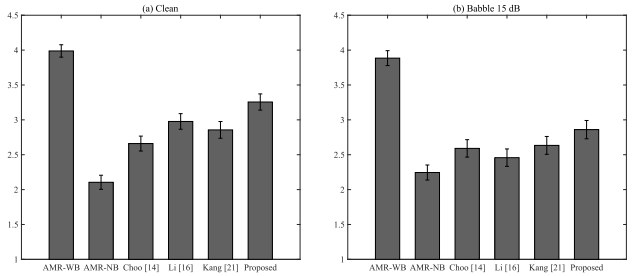

FIGURE 6. Overall DMOS test results under the (a) clean and (b) 15 dB babble environments (95% confidence intervals).

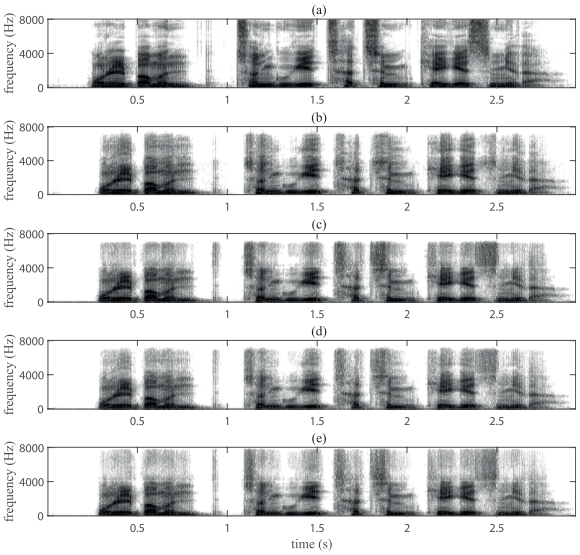

FIGURE 7. Spectrogram comparison of the speech signals processed by the (a) AMR-WB codec with 23.85 kbps, (b) Choo's method [14], (c) Li's method [16], (d) Kang's method [21], and (e) proposed BWE method under the clean environment.

by using the V/UV classifier. Note that the performance of the proposed BWE system without denoising is not degraded in the clean speech environment as given in Tables 1 and 2, which ensures that denoising DNN does not damage the BWE system in the clean speech environment.

Next, to verify the results of the objective quality tests, we conducted a degradation category rating (DCR) listening test [36]. The DCR test uses a degradation opinion scale, with a high-quality reference condition using the original WB speech preceding each condition being assessed. The test consisted of pairwise comparisons between the processing types. Specifically, one sentence, corresponding to the original WB speech, was presented to the listener in each test case, and then the listener was asked to evaluate the quality of the second sample in comparison with the quality of the first sample. Responses were given using the five-point degradation mean opinion score (DMOS) scale ranging from much worse (0) to much better (5). The results of the subjective speech quality test as shown in Fig. 6 represent that the DMOS results under both the clean and 15 dB babble environments are statistically significant; the mean score for each pair of processing types is shown on the horizontal axis together with the 95% confidence interval. Note that the performance of Li's method is lower than that of Choo's method at the 15 dB babble environment, in contrast to the result in clean environment. This is a different result from the objective measure result, which implies that the direct mapping of log-power spectra in a noisy environment may exhibit more unstable performance.

To summarize, the overall simulation results demonstrate that the proposed BWE algorithm improves speech quality compared to the reference BWE algorithms, Choo *et al.* [14] and Li and Lee [16].

B. GRAPHICAL COMPARISONS

We also evaluated the spectrograms of the reference WB speech signal and the speech signals processed using the Choo's method in [14], Li's method in [16], Kang's method [21], and the proposed BWE method under a clean environment. As shown in Fig. 7, the spectrograms of

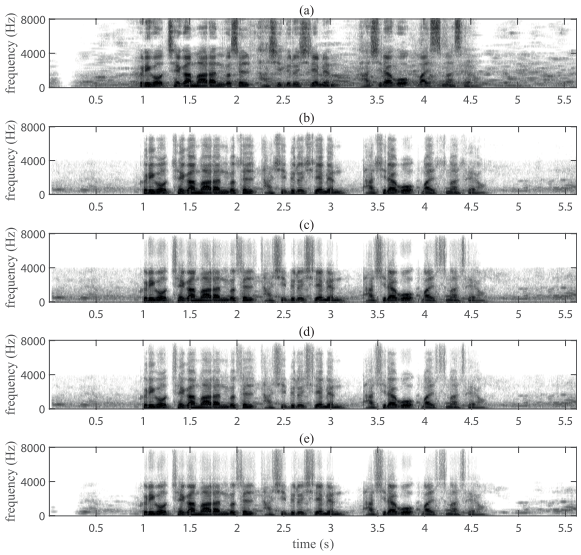

FIGURE 8. Spectrogram comparison of the speech signals processed by the (a) AMR-WB codec with 23.85 kbps, (b) Choo's method [14], (c) Li's method [16], (d) Kang's method [21], and (e) the proposed BWE method under the babble environment (SNR = 15 dB).

conventional methods do not represent up to 8 kHz; the spectrogram from the proposed method is most similar to the spectrogram in the WB original signal. The results from the 15 dB babble environment (Fig. 8) are similar to those in Fig. 7. Note that the spectral gap between 3.4 and 4 kHz are present in Figs. 6 and 7, but it is known to yield a negligible perceptual effect which has also been found by the previous work [37].

IV. CONCLUSIONS

In this paper, we have presented the subband-based sequential DNN ensemble for use as the BWE algorithm. To do this, we folded the NB spectra into the HB region and adjusted the energy levels of the HB using the sequential DNNs. In the sequential DNN model, the denoising DNN was first applied to prevent folding noisy components in the NB spectra, and the subband-based energy levels of the HB spectra were then sequentially estimated using the sequential DNN ensemble. The sequential DNNs were developed using the V/UV classification to better represent the characteristics of speech. In objective and subjective speech quality tests, the proposed approach (sequential DNN incorporating V/UV classification) outperformed the reference methods.

REFERENCES

[1] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Process.*, vol. 83, no. 8, pp. 1707–1719, Aug. 2003.
[2] P. Gajjar, N. Bhatt, and Y. Kosta, "Artificial bandwidth extension of speech & its applications in wireless communication systems: A review," in *Proc. Int. Conf. Commun. Sys. Netw. Technol.*, May 2012, pp. 563–568.
[3] M. Kaniewska *et al.*, "Enhanced AMR-WB bandwidth extension in 3GPP EVS codec," in *Proc. Global Conf. Signal Inf. Process.*, Dec. 2015, pp. 652–656.
[4] P. Jax and P. Vary, "Bandwidth extension of speech signals: A catalyst for the introduction of wideband speech coding?" *IEEE Commun. Mag.*, vol. 44, no. 5, pp. 106–111, May 2006.
[5] H. Pulakka, U. Remes, K. Palomäki, M. Kurimo, and P. Alku, "Speech bandwidth extension using Gaussian mixture model-based estimation of the highband mel spectrum," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2011, pp. 5100–5103.
[6] T. Unno and A. McCree, "A robust narrowband to wideband extension system featuring enhanced codebook mapping," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2005, pp. 805–808.
[7] P. Jax and P. Vary, "Artificial bandwidth extension of speech signals using MMSE estimation based on a hidden Markov model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2003, pp. 680–683.
[8] H. Pulakka and P. Alku, "Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband mel spectrum," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2170–2183, Sep. 2011.
[9] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1979, pp. 428–431.
[10] G. Miet, A. Gerrits, and J. C. Valiere, "Low-band extension of telephone-band speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Jun. 2000, pp. 1851–1854.
[11] U. Kornagel, "Improved artificial low-pass extension of telephone speech," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Sep. 2003, pp. 107–110.
[12] H. Yasukawa, "Enhancement of telephone speech quality by simple spectrum extrapolation method," in *Proc. Eurospeech*, Jan. 1995, pp. 1545–1548.
[13] A. Uncini, F. Gobbi, and F. Piazza, "Frequency recovery of narrow-band speech using adaptive spline neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 1999, pp. 997–1000.
[14] K. Choo, P. Anton, and E. Oh, "Blind bandwidth extension system utilizing advanced spectral envelope predictor," in *Proc. Audio Eng. Soc. Conv.*, May 2015, pp. 1–6.
[15] J. Jeon, Y. Li, S. Kang, K. Choo, E. Oh, and H. Sung, "Robust artificial bandwidth extension technique using enhanced parameter estimation," in *Proc. Audio Eng. Soc. Conv.*, Oct. 2014, pp. 1–6.
[16] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2015, pp. 4395–4399.
[17] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 7398–7402.
[18] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 697–710, Apr. 2013.
[19] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, "Robust sound event classification using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 3, pp. 540–552, Mar. 2015.
[20] B.-K. Lee and J.-H. Chang, "Packet loss concealment based on deep neural networks for digital speech transmission," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 2, pp. 378–387, Feb. 2016.
[21] Y. Li and S. Kang, "Artificial bandwidth extension using deep neural network-based spectral envelope estimation and enhanced excitation estimation," *IET Signal Process.*, vol. 10, no. 4, pp. 422–427, Jun. 2016.
[22] J. Abel and T. Fingscheidt, "Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 1, pp. 71–83, Jan. 2018.
[23] G. Yu and Z.-H. Ling, "Restoring high frequency spectral envelopes using neural networks for speech bandwidth extension," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jul. 2015, pp. 71–83.
[24] Y. Wang, S. Zhao, D. Qu, and J. Kuang, "Using conditional restricted boltzmann machines for spectral envelope modeling in speech bandwidth extension," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2016, pp. 5930–5934.
[25] K. Jarvinen, "Standardisation of the adaptive multi-rate codec," in *Proc. Eur. Signal Process, Conf.*, Sep. 2000, pp. 1–4.
[26] B. Bessette *et al.*, "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 8, pp. 620–636, Nov. 2002.
[27] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[28] Y. Ohtani, M. Tamura, M. Morita, and M. Akamine, "GMM-based bandwidth extension using sub-band basis spectrum model," in *Proc. Interspeech*, Sep. 2014, pp. 2489–2493.
[29] A.-R. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny, "Deep belief networks using discriminative features for phone recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2011, pp. 5060–5063.
[30] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.
[31] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 14–22, Jan. 2012.
[32] I. Hwang, H.-M. Park, and J.-H. Chang, "Ensemble of deep neural networks using acoustic environment classification for statistical model-based voice activity detection," *Comput. Speech Lang.*, vol. 38, no. 1, pp. 1–12, Jul. 2016.
[33] H. Pulakka, P. Alku, L. Laaksonen, and P. Valve, "The effect of high-band harmonic structure in the artificial bandwidth expansion of telephone speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Aug. 2007, pp. 2497–2500.
[34] A. Bayya and M. Vis, "Objective measures for speech quality assessment in wireless communications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 1996, pp. 495–498.
[35] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
[36] *Methods for Subjective Determination of Transmission Quality*, document ITU-T Rec. P.800, 1998.
[37] H. Pulakka, L. Laaksonen, M. Vainio, J. Pohjalainen, and P. Alku, "Evaluation of an artificial speech bandwidth extension method in three languages," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 6, pp. 1124–1137, Aug. 2008.

BONG-KI LEE received the B.S. degree in electrical and communication engineering and the M.S. and Ph.D. degrees in electronics and computer engineering from Hanyang University, South Korea, in 2010, 2012, and 2017, respectively. He is currently a Senior Research Engineer of CTO Division, LG Electronics. His areas of the interest are speech coding, speech enhancement, speech bandwidth extension, acoustic sound classification, and machine learning applied to speech/audio signal processing.

KYOUNGJIN NOH was born in Seoul, South Korea, in 1990. He received the B.S. degree in electronic engineering from Hanyang University, Seoul, in 2015, where he is currently pursuing the Ph.D. degree with the Department of Electronics and Computer Engineering. His research interests include speech/audio signal processing, speech detection and classification of acoustic scenes and events, speech recognition, and machine learning.

JOON-HYUK CHANG (M'03–SM'12) received the B.S. degree in electronics engineering from Kyungpook National University, Daegu, South Korea, in 1998, and the M.S. and Ph.D. degrees in electrical engineering from Seoul National University, South Korea, in 2000 and 2004, respectively. From 2000 to 2005, he was with Netdus Corp., Seoul, as a Chief Engineer. From 2004 to 2005, he was with the University of California, Santa Barbara, in a postdoctoral position involved on adaptive signal processing and audio coding. In 2005, he joined the Korea Institute of Science and Technology, Seoul, as a Research Scientist, where he involved on speech recognition. From 2005 to 2011, he was an Assistant Professor with the School of Electronic Engineering, Inha University, Incheon, South Korea. He is currently an Associate Professor with the School of Electronic Engineering, Hanyang University, Seoul. His research interests are speech coding, speech enhancement, speech recognition, audio coding, and adaptive signal processing. He was a recipient of the IEEE/IEEK IT Young Engineer Award of the year 2011. He is serving on the Editorial Board of *Digital Signal Processing*.

KIHYUN CHOO received the B.S.E.E. and M.S.E.E. degrees from Seoul National University, Seoul, South Korea, in 1998 and 2000, respectively. From 2000 to 2010, he was with the Samsung Advanced Institute of Technology. He was with the Digital Media and Communication Research and Development Center, Samsung Electronics, in 2010. Since 2017, he has been with the Samsung Research and involved in the area of speech and audio coding. His interests are in speech and audio Codec development and speech enhancement in the mobile communication. In this area, he developed speech and audio codec algorithms for standardization of speech and audio codec, MPEG-D Unified Speech and Audio Codec standardized in 2009, and 3GPP Enhanced Voice Service Codecs standardized in 2014. He is currently involved in speech and audio enhancement work.

EUNMI OH received the Ph.D. degree in psychology with an emphasis on psycho-acoustics from the University of Wisconsin-Madison in 1997. She has been with Samsung Electronics since 2000. She is currently a Master (Research VP) with Samsung Electronics. She has led researches on audio/speech coding and MPEG/3GPP Standard activities. Her recent researches include speech/audio quality enhancement and speech synthesis using deep neural network.