Multi-band Approach to Deep Learning-Based Artificial Stereo Extension

Kwang Myung Jeon, Su Yeon Park, Chan Jun Chun, Nam In Park, and Hong Kook Kim

In this paper, an artificial stereo extension method that creates stereophonic sound from a mono sound source is proposed. The proposed method first trains deep neural networks (DNNs) that model the nonlinear relationship between the dominant and residual signals of the stereo channel. In the training stage, the band-wise log spectral magnitude and unwrapped phase of both the dominant and residual signals are utilized to model the nonlinearities of each sub-band through deep architecture. From that point, stereo extension is conducted by estimating the residual signal that corresponds to the input mono channel signal with the trained DNN model in a sub-band domain. The performance of the proposed method was evaluated using a log spectral distortion (LSD) measure and multiple stimuli with a hidden reference and anchor (MUSHRA) test. The results showed that the proposed method provided a lower LSD and higher MUSHRA score than conventional methods that use hidden Markov models and DNN with full-band processing.

Keywords: Artificial stereo extension, Deep neural network (DNN), Mid/side coding, Multi-band, Quadrature mirror filter.

Manuscript received Oct. 28, 2016; revised Feb. 21, 2017; accepted Mar. 28, 2017. This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the government of Korea (MSIP) (No. 2015R1A2A1A05001687), by the ICT R&D program of MSIP/IITP (R01261510340002003, Development of hybrid audio contents production and representation technology for supporting channel and object based audio), and by the National Forensic Service (NFS2017DTB04), Ministry of the Interior, Rep. of Korea.

Kwang Myung Jeon (kmjeon@gist.ac.kr), Su Yeon Park (stellasy0213@gist.ac.kr), Chan Jun Chun (cjchun@gist.ac.kr), and Hong Kook Kim (corresponding author, hongkook@gist.ac.kr) are with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Rep. of Korea.

Nam In Park (naminpark@korea.kr) is with the Digital Technology and Biometry Division, National Forensic Service, Wonju, Rep. of Korea.

This is an Open Access article distributed under the term of Korea Open Government License (KOGL) Type 4: Source Indication + Commercial Use Prohibition + Change Prohibition (http://www.kogl.or.kr/news/dataView.do?dataIdx=97).

I. Introduction

It is well known that stereophonic sound provides a more pleasant and natural experience than monaural (monophonic) sound on account of the presence of spatial information containing both ambience and/or the distinguished relative positions of objects and events [1]. Thus, a monaural listening experience can be greatly improved if the corresponding spatial information is provided.

In general, the spatial cues that produce stereophonic effects comprise the inter-channel intensity difference (IID), inter-channel phase difference (IPD), and inter-channel coherence (ICC) [2]. IID and IPD relate to sound localization factors, such as the relative position, while ICC characterizes the wideness of the auditory image [3]. The aim of this study was to regenerate stereophonic effects for a given monaural sound, as shown in Fig. 1. Assuming that a sound source moves around a dotted circle, as indicated in Fig. 1, the sound localization parameters, such as the IID and IPD, are unobtainable with a single-channel microphone [4]. Therefore, this study focused on reproducing the wideness of the stereophonic effect.

Several works have reported the generating of a stereophonic signal from a monophonic signal [5]–[7]. Among them,

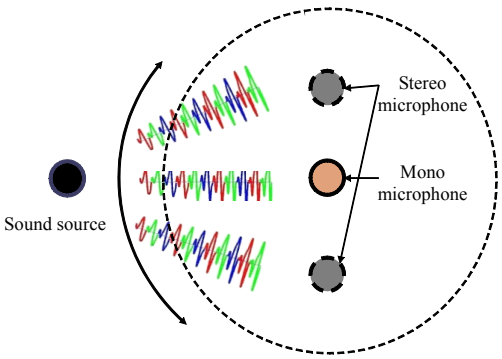


Fig. 1. Goal of artificial stereo extension.

© 2017 ETRI

parametric stereo methods [2]–[5] and a Gaussian mixture model (GMM) with a hidden Markov model (HMM)-based artificial stereo extension method [6], [7] have been successfully applied to this task. However, the parametric stereo method described in [2] uses the additional spatial information of the target stereo channel that is coded in extra bits. Thus, it is not a general solution for the stereo extension of arbitrary input signals. Instead, the parametric stereo method was realized in more common use by controlling the ICC parameter without additional bits [5]. Nevertheless, such ICC-based artificial stereo extension methods usually fail to estimate the desired stereophonic image because the manually chosen ICC parameter hardly tracks the ICC of real stereo signals, which often change over time [6]. This is because the ICC of real stereo signals varies over time. As an alternative, a statistical approach for estimating the spectra of stereo signals was proposed using HMM [7]. It was reported in previous works that the HMM-based stereo extension method outperformed ICC or GMM-based methods because the HMM was more suitable for modeling the time-varying characteristics of the spatial information over frames [7].

Recently, deep neural network (DNN)-based speech and/or audio processing applications, including speech recognition and speech synthesis, have exceeded the HMM counterparts [8]. DNN also showed its effectiveness in artificial stereo extension with a more accurate generation of spatial information than the HMM-based method [9]. Nonetheless, this DNN-based method oversimplified the spatial information of the full-band audio by representing it with a low-dimensional feature set, which is a line spectral frequency (LSF) of the 30th order. Thus, a more complex model that suitably represents audio signals is required for further improvement of the DNN-based stereo extension.

Multi-band or sub-band representation of audio has been utilized in a variety of applications, such as audio coding [10], [11], audio enhancement [12], audio upmixing [13], and automatic speech recognition (ASR) [14]. For example, several audio coding standards, including MPEG Surround [10] and MPEG-H [11], have incorporated a quadrature mirror filter (QMF) to obtain uniformly distributed and oversampled frequency representations of audio signals. In addition, multi-band representation has been applied to audio enhancement and has improved the enhancement quality by variously performing noise attenuation according to the given sub-bands [12]. Moreover, a DNN was realized in a sub-band manner for ASR, which resulted in the reduction of the average word error rate compared to that in a full-band manner [14].

This kind of sub-band approach was also applied to audio upmixing [13], whereby the rear and center channels in a 5.1-channel audio playback system were modeled by using a single DNN model. The input feature vector for the DNN in the model was constructed by concatenating all the sub-band spectral features. In other words, the actually trained DNN model was a full-band approach using sub-band features.

In this paper, a multi-band DNN approach is proposed for extending a mono audio signal into a stereo one. As previously mentioned, the proposed method is intended to model a DNN for each sub-band for stereo extension, while only one DNN was used together over all the sub-bands in [13]. To this end, the proposed method represents the stereo channel as a set of the band-wise log-spectral magnitude and the unwrapped phase of mid/side signals, which comprise the dominant and residual portions of the stereo channel [2]. Specifically, a 32-channel QMF [15] is applied in both the DNN training and the stereo extension stages. Unlike the conventional DNN-based method, the proposed method trains multiple DNNs for each sub-band, which models the band-wise nonlinearity between the mid and side signals. Once the multi-band DNNs are prepared, the log spectral magnitude and unwrapped phase of each side signal band are estimated via feed-forward decoding at the stereo extension stage. The estimated sub-band signals are then combined into a full-band signal via QMF synthesis. Finally, artificially extended stereo signals are obtained by adding (or subtracting) the input mono signal and the estimated full-band side signal.

The performance of the proposed stereo extension method was compared with those of conventional full-band stereo extension methods, including ICC [5], HMM [7], and DNN with LSF features [9]. Moreover, the proposed method was then compared with a multi-band DNN-based audio upmixing method [13]. In addition, to compare the proposed method with a multi-band HMM method, the full-band HMM-based method in [7] was modified into a multi-band HMM-based method.

The remainder of this paper is organized as follows. Section II describes conventional stereo extension methods. In Section III, the multi-band DNN-based stereo extension method is proposed. In Section IV, the performance of the proposed stereo extension method is evaluated and compared with those of conventional full-band and sub-band extension methods. Section V concludes the paper.

II. Conventional Stereo Extension Methods

In this section, the conventional methods that provide stereophonic effects, such as ICC-based and HMM-based stereo extension methods, are introduced.

1. ICC-Based Stereo Extension Method

Figure 2 shows a signal flow graph of the conventional

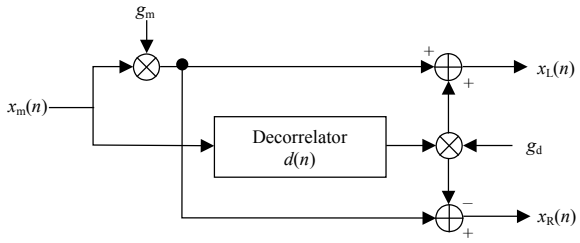


Fig. 2. Block diagram of a conventional ICC-based stereo extension.

stereo extension method based on ICC [5]. As shown in the figure, the extended stereo-channel signals, $x_L(n)$ and $x_R(n)$, can be obtained as [4], [5]

$$x_L(n) = g_m\left(x_m(n) + g_d\left(x_m(n) * d(n)\right)\right), \quad (1)$$

$$x_R(n) = g_m x_m(n) - g_d\left(x_m(n) * d(n)\right), \quad (2)$$

where $x_m(n)$ is a mono signal, and $d(n)$ is the impulse response of a decorrelator. In addition, $*$ denotes a convolution operation.

Next, the mono signal $x_m(n)$ and the decorrelated signal $x_m(n) * d(n)$ are weighted by scale factors g_m and g_d, which are defined as

$$g_m = \cos\left(0.5\arccos(ICC)\right), \quad (3)$$

$$g_d = \sin\left(0.5\arccos(ICC)\right). \quad (4)$$

Here, it is required that the two scale factors should be correlated with the ICC, defined as [4], [5]

$$ICC = \frac{\sum_{n=0}^{N-1}\left(x_L(n)x_R(n)\right)}{\sqrt{\sum_{n=0}^{N-1}x_L^2(n)\sum_{n=0}^{N-1}x_R^2(n)}}, \quad (5)$$

where N is the number of samples used for the ICC computation. In this paper, the ICC is set to zero so that the left and right signals of the stereo signals can be at a maximal distance apart for each loud speaker.

However, since the ICC of real stereo signals sometimes varies, the stereo signals generated by the ICC-based stereo extension method differ from real stereo signals. As an alternative, a statistical method that uses HMM is proposed to obtain stereo signals by estimating the side signal for a given mono input signal. The following subsection briefly describes the HMM-based stereo extension method.

2. HMM-Based Stereo Extension Method

Figure 3 shows the procedure of the conventional HMM-based stereo extension method that generates artificial stereo signals from mono signals [7]. As shown in the figure, this method is applied in the modified discrete cosine transform

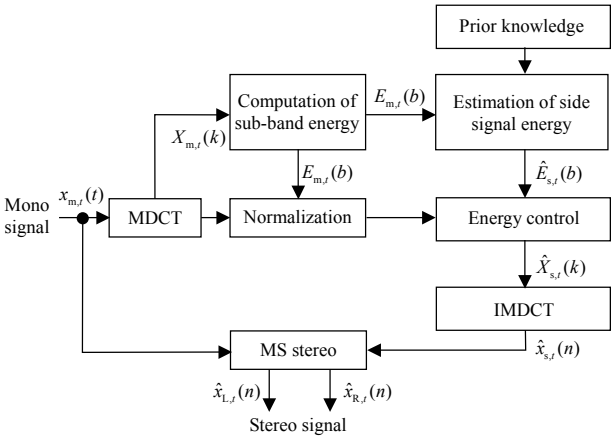


Fig. 3. Block diagram of a conventional HMM-based stereo extension method.

(MDCT) domain. This means that the mono signal, $x_{m,t}(n)$, is segmented into a consecutive sequence of frames and the t-th frame mono signal. Then, $x_{m,t}(n)$ is transformed into the frequency domain using a $2N$-point MDCT. The MDCT coefficients of the mono signal, $X_{m,t}(k)$, are grouped into 15 sub-bands, where each sub-band includes $N/8$ MDCT coefficients and is overlapped by $N/16$ MDCT coefficients. Next, the sub-band energy, $E_{m,t}(b)$, is extracted from $X_{m,t}(k)$, and $X_{m,t}(k)$ is normalized by $E_{m,t}(b)$, where b is the sub-band index and $b = 0, \ldots, 14$.

Basically, this method uses a mid-side stereo approach for the extension. That is, the incoming mono signal is considered the mid signal. Thus, a side signal should be estimated using HMM. To this end, the sub-band energy of the side signal $\hat{E}_{s,t}(b)$ is estimated from $E_{m,t}(b)$ using HMM under a minimum mean-squared error (MMSE) criterion [7].

Next, the MDCT coefficients of the side signal, $\hat{X}_{s,t}(k)$, are estimated based on both $X_{m,t}(k)$ and $\hat{E}_{s,t}(b)$. Then, the side signal $\hat{x}_{s,t}(n)$ in the time domain is obtained by applying an N-point inverse MDCT (IMDCT). Finally, the stereo signal is obtained by adding $\hat{x}_{s,t}(n)$ to (or subtracting it from) the mono signal, $x_{m,t}(n)$.

HMM-based stereo extension performs better for both speech and music signals than the ICC-based method. However, a single HMM with limited parameters is insufficient for modeling the nonlinear relationship between the mono and side signal [9]. Recent approaches have used DNN to estimate the side signal for the mono signal input to cope with this issue. The following subsection summarizes the DNN-based stereo extension method.

3. DNN-Based Stereo Extension Method

Figure 4 illustrates a block diagram of the DNN-based

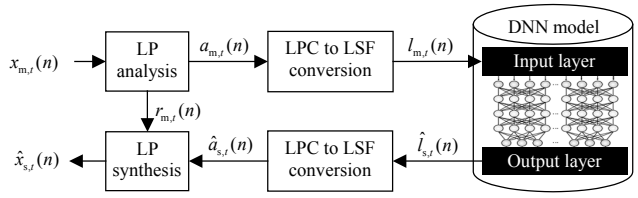


Fig. 4. Block diagram of a conventional DNN-based stereo extension method.

mono-to-stereo extension method. Similar to the HMM-based method, the monaural signals are assumed to be mid signal $x_{m,t}(n)$ for extended stereo signals. Then, side signals $x_{s,t}(n)$ are estimated by feed-forwarding $x_{m,t}(n)$ into the DNN, which consists of the features of the mid and side signals as input and output layers, respectively. Specifically, the residual signal for $x_{m,t}(n)$ is first obtained by performing M-th order linear prediction (LP) analysis [16], such as

$$r_{m,t}(n) = x_{m,t}(n) - \sum_{l=1}^{M-1} a_{m,t}(l) x_{m,t}(n-l), \quad (6)$$

where $a_{m,t}(l)$ and $r_{m,t}(n)$ are the LP coefficients of the mid signals and the residual signals, respectively. From that point, the LP coefficients of the mid signals are converted into LSF coefficients [17]. The LSF coefficients of both the mid and side signals of the training set are then used to train the DNN model.

Unsupervised pre-training is first conducted by stacking multiple restricted Boltzmann machines (RBMs) to train the DNN model [18]. Specifically, the input layer connected to a hidden layer is a Gaussian–Bernoulli RBM. Moreover, a pile of Bernoulli–Bernoulli RBMs is stacked behind the Gaussian–Bernoulli RBM. Next, supervised fine-tuning is conducted for the back-propagation method with the MMSE cost function for the target, that is, referencing the LSF features of the side signals [19].

After the DNN has been trained, the LSF coefficients of the side signals are estimated from the DNN and are converted into LP coefficients. Next, the estimated side signal $\hat{x}_{s,t}(n)$ is reconstructed using the residual signals for the mid signals and the estimated LP coefficients as

$$\hat{x}_{s,t}(n) = r_{m,t}(n) - \sum_{l=1}^{M-1} \hat{a}_{s,t}(l) r_{m,t}(n-l), \quad (7)$$

where $\hat{a}_{s,t}(l)$ is the l-th LP coefficient estimated from the DNN model. Finally, the stereophonic signals are obtained by adding and subtracting the mid and side signals.

The DNN-based method with LSF features obtained slightly better results than the HMM-based method. As described thus far, the conventional method operates with full-band spectral features. However, it is known that the stereo characteristics of audio signals are localized for some frequency bands [20], [21]. Subjective listening tests in [21] indicate that the differences between the left and right channels are hardly noticeable for a low frequency band below 250 Hz and a high frequency band above 5 kHz. On the other hand, their differences are easily noticeable for a frequency band around 1 kHz, which enables differentiation of the left channel audio from the right one [21]. Thus, the proposed multi-band stereo extension method is motivated by this frequency-dependent similarity and difference between each channel of stereo audio. This is because it is expected to further improve the performance of the full-band DNN-based stereo extension method if the DNN is modeled for each sub-band.

III. Proposed Stereo Extension Method

The DNN-based stereo extension method that incorporates multi-band processing is proposed in this section. Figure 5 depicts a block diagram of the proposed method. As shown in the figure, the proposed method estimates side-signals $x_{s,t}(n)$ using multi-band DNNs, which act as a mapping function between the sub-band of the mid and side signals. Similar to the conventional stereo extension methods based on HMM, as well as the DNN with LSF features described in Section II, the proposed method consists of training and stereo extension stages. Each stage is detailed in the following subsections.

1. Multi-band DNN Training

The set of stereo audio signals is first prepared in a mid/side

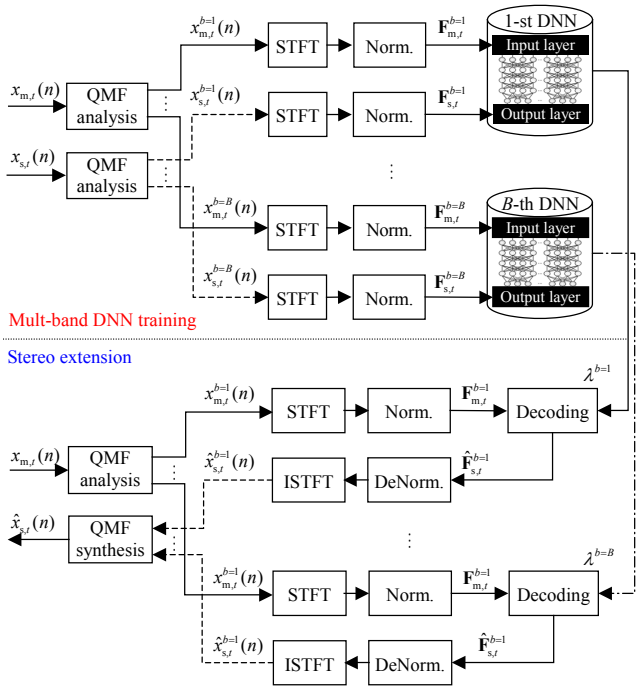


Fig. 5. Block diagram of the proposed multi-band DNN-based stereo extension.

form to train the multi-band DNNs. Next, pairs of mid and side signals are divided into B sub-band signals, $x^b_{\mathrm{m},t}(n)$ and $x^b_{\mathrm{s},t}(n)$, respectively, using QMF analysis [15]. Then, both $x^b_{\mathrm{m},t}(n)$ and $x^b_{\mathrm{s},t}(n)$ are transformed into complex spectra, $X^b_{\mathrm{m},t}(k)$ and $X^b_{\mathrm{s},t}(k)$, using a short-time Fourier transform (STFT) with a K-point fast Fourier transform (FFT). From that point, their respective log spectral magnitudes and unwrapped phases $\log\left|X^b_{\mathrm{m},t}(k)\right|, \log\left|X^b_{\mathrm{s},t}(k)\right|, U\left[\angle X^b_{\mathrm{m},t}(k)\right]$, and $U\left[\angle X^b_{\mathrm{s},t}(k)\right]$, are extracted. Here, $U[\cdot]$ indicates a phase unwrap function [22]. Prior to DNN training, these log spectral magnitudes and unwrapped phases are scaled to have values between zero and one, as

$$L^b_{\mathrm{m},t}(k) = \frac{\log\left|X^b_{\mathrm{m},t}(k)\right|}{\alpha}, \tag{8}$$

$$P^b_{\mathrm{m},t}(k) = \frac{U\left[\angle X^b_{\mathrm{m},t}(k)\right] + \beta}{2\beta}, \tag{9}$$

where α is the maximum log spectral magnitude, which can be defined as $\log(2^{15} \times K)$ [23]. Moreover, β is the maximum unwrapped phase value, which is empirically found for the abundant unwrapped phase values of the training dataset. Note that $L^b_{\mathrm{s},t}(k)$ and $P^b_{\mathrm{s},t}(k)$, which are the normalized version of $\log\left|X^b_{\mathrm{s},t}(k)\right|$ and $U\left[\angle X^b_{\mathrm{s},t}(k)\right]$, are also obtained in a manner similar to (8) and (9). The scaled features for the mid and side signals consist of feature vectors of the b-th sub-band DNN as

$$\mathbf{f}^b_{\mathrm{m},t} = \left\{L^b_{\mathrm{m},t}(k), P^b_{\mathrm{m},t}(k)\right\}, \tag{10}$$

$$\mathbf{f}^b_{\mathrm{s},t} = \left\{L^b_{\mathrm{s},t}(k), P^b_{\mathrm{s},t}(k)\right\}, \tag{11}$$

where the size of both $\mathbf{f}^b_{\mathrm{m},t}$ and $\mathbf{f}^b_{\mathrm{s},t}$ are $K + 2$. Next, the multi-band DNNs are trained for each b-th sub-band using a sequence of multiple feature vectors, $\mathbf{F}^b_{\mathrm{m},t} = [\mathbf{f}^b_{\mathrm{m},t-S} \cdots \mathbf{f}^b_{\mathrm{m},t} \cdots \mathbf{f}^b_{\mathrm{m},t+S}]$, and $\mathbf{f}^b_{\mathrm{s},t}$ as a pair of input and output layers for the networks. Specifically, multi-band DNNs are first initialized as a deep generative model by stacking multiple restricted Boltzmann machines (RBMs) [18]. Similar to the DNN-LSF method introduced in Section II.3, the input layer of linear variables is represented as a Gaussian–Bernoulli RBM. From that point, a pile of Bernoulli–Bernoulli RBMs is stacked behind the Gaussian–Bernoulli RBM. After the initialization of the multi-band DNNs has finished, supervised fine-tuning is conducted using the back-propagation method with the MMSE cost function between the oracle and estimations of $\mathbf{f}^b_{\mathrm{s},t}$. Note that the fine-tuning is iteratively conducted in a pair of feed-forward decoding and back-propagation with the development dataset until the MMSE is reduced to the predefined threshold.

2. Stereo Extension with Multi-band DNN

In the stereo extension stage, input mono channel audio signal $x_{\mathrm{m},t}(n)$ is divided into B-sub-band signals $x^b_{\mathrm{m},t}(n)$ via QMF analysis [15]. The sub-band signals of each channel are transformed into a log spectra magnitude and an unwrapped phase. They are then scaled to the range of zero to one, thus obtaining $\mathbf{f}^b_{\mathrm{m},t} = \left\{L^b_{\mathrm{m},t}(k), P^b_{\mathrm{m},t}(k)\right\}$. Then, a multiple feature vector, $\mathbf{F}^b_{\mathrm{m},t} = [\mathbf{f}^b_{\mathrm{m},t-S} \cdots \mathbf{f}^b_{\mathrm{m},t} \cdots \mathbf{f}^b_{\mathrm{m},t+S}]$, is assigned to the input layer of the b-th sub-band DNN model. In other words, $\mathbf{F}^b_{\mathrm{m},t}$ is applied to the feed-forward decoding on the b-th sub-band DNN model to estimate the features for the side signal $\hat{\mathbf{f}}^b_{\mathrm{s},t} = \left\{\hat{L}^b_{\mathrm{s},t}(k), \hat{P}^b_{\mathrm{s},t}(k)\right\}$ at the last layer of the networks. After decoding is finished, each $\hat{L}^b_{\mathrm{s},t}(k)$ and $\hat{P}^b_{\mathrm{s},t}(k)$ is denormalized as

$$\log\left|\hat{X}^b_{\mathrm{s},t}(k)\right| = \alpha \hat{L}^b_{\mathrm{s},t}(k), \tag{12}$$

$$U\left[\angle \hat{X}^b_{\mathrm{s},t}(k)\right] = 2\beta \hat{P}^b_{\mathrm{s},t}(k) - \beta. \tag{13}$$

Next, $\log\left|\hat{X}^b_{\mathrm{s},t}(k)\right|$ and $U\left[\angle \hat{X}^b_{\mathrm{s},t}(k)\right]$ are each applied to the exponential and phase wrap functions to obtain $\hat{X}^b_{\mathrm{s},t}(k) = \left|\hat{X}^b_{\mathrm{s},t}(k)\right| \angle \hat{X}^b_{\mathrm{s},t}(k)$. $\hat{X}^b_{\mathrm{s},t}(k)$ is then applied to the inverse STFT, thus obtaining $\hat{x}^b_{\mathrm{s},t}(n)$. Consequently, the full-band side signal $\hat{x}_{\mathrm{s},t}(n)$ is estimated by conducting QMF synthesis [15] on $\hat{x}^b_{\mathrm{s},t}(n)$ for every B sub-band. Finally, stereo extension is conducted by converting the mid/side format into a left/right format with $\hat{x}_{\mathrm{s},t}(n)$ as

$$\hat{x}_{\mathrm{L},t}(n) = x_{\mathrm{m},t}(n) + \hat{x}_{\mathrm{s},t}(n), \tag{14}$$

$$\hat{x}_{\mathrm{R},t}(n) = x_{\mathrm{m},t}(n) - \hat{x}_{\mathrm{s},t}(n). \tag{15}$$

IV. Performance Evaluation

In this section, the performance of the proposed stereo extension method is evaluated in terms of both objective and subjective qualities by measuring the log spectral distortion (LSD) [24] and multiple stimuli with a hidden reference and anchor (MUSHRA) [25]. In addition, the performance of the proposed stereo extension method is first compared with those of conventional full-band stereo extension methods, including ICC [5], HMM [7], and DNN with LSF features (DNN-LSF) [9]. Then, the proposed method is compared with a multi-band DNN-based audio upmixing (DNN-AU) approach [13]. To compare the proposed method with a multi-band HMM method, the full-band HMM-based method in [7] is modified into a multi-band HMM-based method by replacing the sub-

ETRI Journal, Volume 39, Number 3, June 2017
https://doi.org/10.4218/etrij.17.0116.0773

band-DNNs with sub-band HMMs, as shown in Fig. 5. In other words, an MDCT is applied once every sub-band and the MDCT coefficients for each sub-band are used to train an HMM of the corresponding sub-band, which is herein referenced as MB-HMM.

1. Experimental Setup

To prepare the stereo extension methods, 90 min of speech and 2 h of music data recorded in a stereo format were used. The speech databases used in training consisted of 20 min of Sound Quality Assessment Material (SQAM) [25], 50 min of the ETRI SWB Korean speech corpus [26], and 20 min of the TSP speech DB [27]. The music databases used in the training consisted of 40 min of SQAM, 20 min of orchestra, 30 min of popular music, and 30 min of audio-form user-created content (UCC). The total 3.5-h set was then split into training, development, and test sets at a ratio of 7:2:1. There was no overlap among the split sets, and every database clip was down-sampled to 32 kHz with 16-bit resolution.

The audio signal was segmented into a consecutive number of frames. Each frame length was 32 ms and the overlapped size was 16 ms. Thus, a 2,048-point MDCT (N = 1,024 in Section II.2) was used to transform the time-domain signal into the frequency domain one. These MDCT coefficients were brought to the HMM-based stereo extension method as feature vectors. To train a DNN for DNN-LSF, the LSF feature extraction method was applied to each audio frame, where the order of LSF was set to $M = 30$. The DNN had an input layer with 330 nodes, and it had five hidden layers with 2,048 nodes each. In addition, the output layer had 30 nodes.

The multi-band stereo extension methods in the experiments commonly employed a 32-channel QMF (B = 32 in Section III). To implement the DNN-AU in [13], two DNNs were constructed to estimate the respective left and right channel signals from the mono signal, respectively. As mentioned in Section I, the feature vectors for DNNs in DNN-AU were constructed by concatenating all the sub-band spectral features. Additionally, the size of the combined feature vector for a frame was 1,056 because each sub-band consisted of 33 spectral magnitudes. The combined feature vector was then spliced across 11 neighboring frames. Thus, the DNNs in DNN-AU had an input layer with 11,616 nodes, and it had five hidden layers with 2,048 nodes each.

In the proposed stereo extension method, each sub-band DNN was trained using feature vectors that were obtained by applying a 64-point FFT to each channel signal after QMF analysis. This 64-dimensional spectral feature vector was spliced across 11 neighboring frames ($S = 5$ in Section III), and they were used for the input layer of each DNN. In addition, the number of hidden layers with 256 nodes each was set to three. The learning rate and number of iterations were set to 0.008 and 100, respectively, for training DNNs for DNN-LSF, DNN-AU, and the proposed method.

2. Quality Evaluation

The artificial stereo extension methods, including the proposed method, were compared in terms of their objective and subjective qualities. The LSD between the true stereo signal and extended counterpart was measured in a manner given by [24] to measure the objective quality.

$$d_{\mathrm{LSD}} = \frac{1}{2T}\sum_{t=0}^{T-1}\sum_{c=\mathrm{L,R}}\sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\left(10\log_{10}\frac{\left|X_{c,t}(k)\right|}{\left|\hat{X}_{c,t}(k)\right|}\right)^2}, \qquad (16)$$

where $\left|X_{c,t}(k)\right|$ and $\left|\hat{X}_{c,t}(k)\right|$ are the k-th spectral magnitudes of $x_{c,t}(n)$ and $\left|\hat{x}_{c,t}(n)\right|$ for the c-th channel, respectively. Moreover, T indicates the total number of frames. Figure 6 compares the average LSD between the reference stereo spectra by the proposed multi-band DNN-based method and those by the conventional methods, such as ICC, HMM, DNN-LSF, MB-HMM, and DNN-AU. As shown in the figure, the proposed stereo extension method has a lower LSD value than conventional methods.

For the subjective evaluation, MUSHRA [25] was performed on ten listeners (seven males and three females) who had no auditory diseases. The classes of the MUSHRA test were as follows: 1) a hidden reference (true stereo audio signal); 2) anchor signals processed with a low-pass filter of 7 kHz and 3) 14 kHz; 4) a mono audio signal; 5) artificial stereo audio signals extended by ICC [5]; 6) HMM [7]; 7) MB-HMM; 8) DNN-LSF [9]; 9) DNN-AU [13]; and 10) the proposed method using multi-band DNN. All audio clips were hidden and randomly selected.

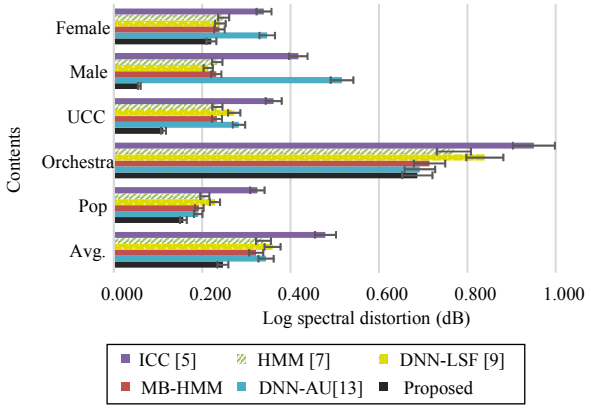


Fig. 6. Comparison of the LSD between true and artificial stereo signals.

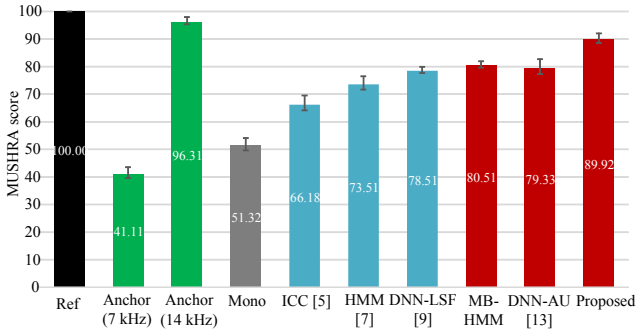


Fig. 7. Comparison of the MUSHRA scores.

Figure 7 shows the results of the MUSHRA listening test. As shown in the figure, the multi-band approaches were better than full-band approaches. That is, the average MUSHRA score of the multi-band HMM (MB-HMM) was higher than that of the full-band HMM. The proposed multi-band extension method and DNN-AU produced higher average MUSHRA scores than DNN-LSF (that is, the full-band extension method). However, DNN-AU had an average MUSHRA score similar to MB-HMM. A comparison of the proposed method with other sub-band methods, such as MB-HMM and DNN-AU, showed that the proposed method provided significantly higher average MUSHRA scores.

From the results of the objective and subjective evaluations, it is concluded that the proposed multi-band DNN-based stereo extension method can extend mono audio into stereo with a higher quality than conventional methods, including full-band HMM, sub-band HMM, and full-band DNN methods.

V. Conclusion

In this paper, a stereo extension method that applies multi-band DNNs was proposed. The method utilizes QMF analysis to train the DNN of each sub-band to estimate a more realistic side signal for the extension. Its sub-band signals are decoded by DNNs for the extension of an input mono signal to estimate the corresponding side signal of each sub-band. After the sub-band side signals are merged by QMF synthesis, artificial stereo signals are finally obtained by adding or subtracting the estimated side signals to the mono signal. Respective objective and subjective evaluations were conducted to demonstrate the performance of the proposed method. The results of the LSD and MUSHRA evaluations showed that the proposed stereo extension method significantly outperformed the conventional stereo extension methods.

References

[1] J. Lapierre and C. Faller, "Spatial Audio Processing," *Proc. AES Convention*, Paris, France, May 20–23, 2006, Preprint 6804.

[2] E. Schuijers et al., "Low Complexity Parametric Stereo Coding," *Proc. AES Convention*, Berlin, Germany, May 8–11, 2004, Preprint 6073.

[3] H. Purnhagen et al., "Synthetic Ambience in Parametric Stereo Coding," *Proc. AES Convention*, Berlin, Germany, May 8–11, 2004, Preprint 6074.

[4] C.J. Chun et al., "Real-Time Conversion of Stereo Audio to 5.1 Channel Audio for Providing Realistic Sounds," *Int. J. Signal Process. Image Process. Pattern Recogn.*, vol. 2, no. 4, Dec. 2009, pp. 85–94.

[5] N.I. Park and H.K. Kim, "Artificial Stereo Extension of Speech Based on Inter-Channel Coherence," *Adv. Sci. Technol. Lett.*, vol. 14, no. 1, Aug. 2012, pp. 168–171.

[6] N.I. Park et al., "Artificial Stereo Extension Based on Gaussian Mixture Model," *Proc. AES Convention*, Rome, Italy, May 4–7, 2013, Preprint 8877.

[7] N.I. Park et al., "Artificial Stereo Extension Based on Hidden Markov Model for the Incorporation of Non-stationary Energy Trajectory," *Proc. AES Convention*, New York, USA, Oct. 17–20, 2013, Preprint 8980.

[8] G. Hinton et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, Nov. 2012, pp. 82–97.

[9] C.J. Chun et al., "Extension of Monaural to Stereophonic Sound Based on Deep Neural Networks," *Proc. AES Convention*, New York, USA, Oct. 29–Nov. 1, 2015, Preprint 9400.

[10] J. Herre et al., "MPEG Surround—the ISO/MPEG Standard for Efficient and Compatible Multichannel Audio Coding," *J. Audio Eng. Soc.*, vol. 56, no. 11, Nov. 2008, pp. 932–955.

[11] J. Herre et al., "MPEG-H 3D Audio—the New Standard for Coding of Immersive Spatial Audio," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, Aug. 2015, pp. 770–779.

[12] K.M. Jeon et al., "An MDCT-Domain Audio Denoising Method with a Block Switching Scheme," *IEEE Trans. Consum. Electron.*, vol. 59, no. 4, Nov. 2013, pp. 818–824.

[13] S.Y. Park, C.J. Chun, and H.K. Kim, "Sub-band-based Upmixing of Stereo to 5.1-Channel Audio Signals Using Deep Neural Networks," *Int. Conf. Inform. Commun. Technol. Convergence*, Jeju, Rep. of Korea, Oct. 19–21, 2016, pp. 377–380.

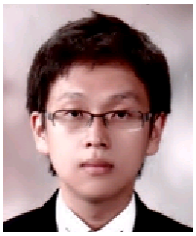
[14] G. Kovács, L. Tóth, and T. Grósz, "Robust Multi-band ASR Using Deep Neural Nets and Spectro-Temporal Features," *Proc. Int. Conf. Speech Comput. (SPECOM)*, Novi Sad, Serbia, Oct. 5–9, 2014, pp. 386–393.

[15] ISO/IEC 23008-3:2015, *Information Technology – High Efficiency Coding and Media Delivery in Heterogeneous Environments – Part 3: 3D Audio*, Oct. 2015.

[16] A. Spanias, T. Painter, and V. Atti, *Audio Signal Processing and Coding*, Hoboken, NJ, USA: John & Wiley & Sons, Inc., Jan. 2007.

[17] X. Mei and S. Sun, "An Efficient Method to Compute LSFs from

LPC Coefficients," *Int. Conf. Signal Process. Proc.*, Beijing, China, Aug. 21–25, 2000, pp. 655–658.
[18] Y. Bengio, "Learning Deep Architectures for AI," *Found. Trends® Mach. Learn.*, vol. 2, no. 1, Jan. 2009, pp. 1–127.
[19] Y. Xu et al., "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, Jan. 2014, pp. 65–68.
[20] G.S. Kendall, "Directional Sound Processing in Stereo Reproduction," *Int. Comput. Music Conf.*, San Jose, CA, Oct. 14–18, 1992, pp. 261–264.
[21] C. Shuixian et al., "Frequency Dependence of Spatial Cues and Its Implication in Spatial Stereo Coding," *Proc. Int. Conf. Comput. Sci. Softw. Eng.*, Wuhan, China, Dec. 12–14, 2008, pp. 1066–1069.
[22] A.V. Oppenheim and R.W. Schafer, *Discrete-Time Signal Processing*, Englewood Cliffs, NJ, USA: Prentice-Hall, 1989.
[23] A.H. Gray and J.D. Markel, "Distance Measures for Speech Processing," *IEEE Trans. Acoust. Speech Signal Process.,* vol. 24, no. 5, Oct. 1976, pp. 380–391.
[24] ITU-R BS.1534-1, *Method for the Subjective Assessment of Intermediate Quality Levels of Coding System*, Jan. 2003.
[25] EBU Technical Document 3253, *Sound Quality Assessment Material Recordings for Subjective Tests–Users' Handbook for the EBU-SQAM Compact Disc*, Apr. 1988.
[26] http://slrdb.etri.re.kr/
[27] P. Kabal, *TSP Speech Database*, Department of Electrical & Computer Engineering, McGill University, Montreal, Canada, Tech. Rep. TR-2002-09-04, Sept. 2002.



Kwang Myung Jeon received his BS degree in information and communications engineering from Sejong University, Seoul, Rep. of Korea in 2010, and his MS degree in information and communications engineering from Gwangju Institute of Science and Technology (GIST), Rep. of Korea in 2012. He is currently pursuing a PhD degree at the School of Electrical Engineering and Computer Science, GIST. His current research interests are machine-learning-based audio/speech signal processing, including speech enhancement, source separation, speech recognition, speech synthesis, acoustic near-field communication, and acoustic event detection.

Su Yeon Park received her BS degree in computer science from Daegu University, Kyungsan, Rep. of Korea in 2015 and her MS degree in electrical engineering and computer science from Gwangju Institute of Science and Technology, Rep. of Korea in 2017. Her research interests are deep-learning based audio signal processing.

Chan Jun Chun received his BS degree in electronics engineering from Korea University of Technology and Education, Seoul, Rep. of Korea in 2009. He received MS and PhD degrees in information and communications engineering, and electrical engineering and computer science from Gwangju Institute of Science and Technology, Rep. of Korea in 2011 and 2017, respectively. He is currently a research specialist at Korea Institute of Civil Engineering and Building Technology. His current research interests include signal processing techniques and embedded algorithms and solutions for signal processing.

Nam In Park received his BS degree in electronics and communications engineering from Kwang Woon University, Seoul, Rep. of Korea in 2007. He earned MS and PhD degrees in electrical engineering from the Gwangju Institute of Science and Technology, Rep. of Korea in 2009 and 2013, respectively. Since August 2014, he has been with the Digital Technology and Biometry Division of the National Forensic Service, Rep. of Korea as an audio forensic engineer. His current research interests include speaker recognition, speech/audio coding, and machine learning.

Hong Kook Kim holds a BS in control and instrumentation engineering from Seoul National University, Rep. of Korea (1988), and an MS and PhD in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, Rep. of Korea (1990; 1994). He was a senior researcher at Samsung Advanced Institute of Technology, Suwon, Rep. of Korea (1990–1998) and a senior technical staff member at AT&T Labs Research, Middletow, NJ, USA (1998–2003). Since 2003, he has been a professor at the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Rep. of Korea. He was a visiting professor at City University of New York, USA (2014–2015) and is an IEEE Senior Member and IEEE Speech and Language Technical Committee affiliate member. He is an APSIPA Speech, Language, and Audio Technical Committee member and distinguished lecturer from 2017 to 2018. Since 2012, he served as a *Digital Signal Processing* editorial committee member and area editor. His current research interests include large vocabulary speech recognition, audio coding, speech/audio source separation, and handheld-device speech/audio processing embedded algorithms/solutions.