AUDIO CODING BASED ON SPECTRAL RECOVERY BY CONVOLUTIONAL NEURAL NETWORK

Seong-Hyeon Shin[1], *Seung Kwon Beack*[2], *Taejin Lee*[2] *and Hochong Park*[1]

[1]Kwangwoon University, Seoul, Korea
[2]Electronics and Telecommunication Research Institute, Daejeon, Korea

ABSTRACT

This study proposes a new method of audio coding based on spectral recovery, which can enhance the performance of transform audio coding. An encoder represents spectral information of an input in a time-frequency domain and transmits only a portion of it so that the remaining spectral information can be recovered based on the transmitted information. A decoder recovers the magnitudes of missing spectral information using a convolutional neural network. The signs of missing spectral information are either transmitted or randomly assigned, according to their importance. By combining transmission and recovery of spectral information, the proposed method can enhance the coding performance, compared with conventional transform coding. The subjective performance evaluation shows that, for mono coding at 39.4 kbps, the proposed method provides higher sound quality than the USAC, by an average MUSHRA score of 8.5.

Index Terms— audio coding, convolutional neural network, spectral recovery, transform coding

1. INTRODUCTION

Since MPEG-1 Audio Layer III audio coding and Dolby AC-3 were deployed in the early 1990s, a transform coding has been a major structure of audio coding [1, 2]. The transform coding quantizes the spectral coefficients based on human psycho-acoustic model. It can reconstruct the output waveform of sound quality that is proportional to the coding bit rate, and is usually used for high-quality coding at high bit rate.

In the early 2000s, a new strategy known as parametric coding was developed to solve the problem of transform coding at low bit rate [3, 4]. It represents audio information in a parametric domain and determines a few parameters that can generate sound perception similar to the original sound. For a low bit-rate coding, the parametric coding is more efficient than the transform coding, by processing parameterized spectral information instead of individual coefficients. However, it has limitations in performing high-quality coding because, once the original spectrum is lost, it is almost impossible to reconstruct the fine spectral structure required to generate a high-quality waveform with only the given parameters.

In this study, a new method of audio coding is proposed that can enhance the performance of transform coding. The proposed method is based on spectral recovery by a neural network. Spectral information of an input is expressed in a two-dimensional (2D) time-frequency (T/F) domain after applying the transform on a sub-frame basis. In an encoder, only a portion of spectral coefficients selected in a 2D check pattern are quantized and coded by a normal transform coding. In addition, a small number of important signs in missing spectral coefficients are transmitted. In a decoder, the magnitudes of missing spectral coefficients are recovered based on the transmitted data by a convolutional neural network (CNN) [5] and the signs are assigned by the transmitted signs or randomly, according to their importance. The proposed method utilizes both transmission and recovery of spectral information, and it can reduce the quantization error by transmitting a reduced number of data, thus enabling enhanced coding performance.

Many studies on information recovery and generation for speech and audio signals have been carried out [3, 4, 6-15]. Most methods of spectral recovery operate on a block basis, where high-band blocks are recovered based on low band as a sort of blind bandwidth extension or super resolution [3, 6-11]. The performance of block-based spectral recovery cannot yet reach that of transform coding with current technologies, even when some additional parameters are used to help the correct recovery, as in the spectral band replication (SBR) [3]. On the other hand, the proposed method recovers individual spectral coefficients independently based on neighbor information, as opposed to block-based recovery, because data to be recovered are arranged in a 2D check pattern. The intelligent gap filling (IGF) supports a method of recovering individual spectral coefficients [12], but it fills the spectral gap by copying the spectral magnitude from low band, unlike the estimation from neighbor information in the proposed method. Hence, the proposed method of spectral recovery can provide better performance than the conventional spectral recovery methods.

A parametric speech coding in [13] generates speech waveform using a WaveNet [15], where speech parameters are used as the conditioning variables in WaveNet. Because of its parametric coding structure, however, it may not be directly applied to the transform audio coding with acceptable performance. A generative model in [14] reconstructs phase information from spectral magnitudes, but it does not provide a method of recovering missing spectral magnitudes. Therefore, the contribution of this study is the development of a new spectral recovery method applicable to transform coding, and an overall coding method equipped with the proposed spectral recovery can

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. 2017-0-00072002, Development of audio/video coding and light field media fundamental technologies for ultra realistic tera-media).

become a new transform coding scheme with potential performance improvement.

The performance of the proposed method is compared to that of the MPEG unified speech and audio coding (USAC) [16]. It is confirmed by subjective tests that, for mono coding at 39.4 kbps, the proposed method provides significantly better performance than the USAC.

2. PROPOSED METHOD OF AUDIO CODING

2.1. Recovery of MDCT magnitudes

The key concept of the proposed coding method is to transmit only a portion of spectral coefficients to the decoder in such a way that the decoder can recover the missing spectral coefficients in acceptable performance. For this concept to work, high correlation among spectral coefficients is required. However, the signs of spectral coefficients have very small correlation. The proposed method is therefore designed to recover only the spectral magnitudes and to process the signs separately. To further increase the data correlation, this method adopts a sub-frame-based transform. As a result, the spectral coefficients are expressed in a 2D T/F domain, which enables the spectral recovery to utilize data correlation in both time and spectral directions.

In the proposed method, a frame length and a sub-frame length are of 2048 samples and 1024 samples, respectively. A modified discrete cosine transform (MDCT) is applied to each sub-frame with 50% overlap. For each frame, then, the resulting MDCT coefficients are represented by a 1024 × 2 matrix in a T/F domain, denoted by $X[k][m]$, where $0 \le k < 1024$ is a frequency index and m is a sub-frame index. $m = 0, 1$ corresponds to the current frame and $m \le -1$ corresponds to the past frames.

Since human perception in low band is very sensitive even to small spectral distortion, it is not feasible to implement the spectral recovery in low band to the desired accuracy unless further studies and advancements are made. Hence, the spectral recovery in this study is applied only to the band above L_x Hz, corresponding to a frequency index $k = k_L$, and the band below L_x Hz follows the normal transform coding. After intensive experiments with signals of various characteristics, it was found that, for high-quality coding, the MDCT magnitudes below 3 kHz ~ 4 kHz require a quality level beyond the accuracy of spectral recovery. Accordingly, a varying L_x for each frame around 3 kHz ~ 4 kHz is used. In addition, it was found that the energy distribution can provide a good clue to determine an appropriate L_x. Based on this investigation, the encoder computes the energy below 4 kHz, EN_{low}, and the energy above 4 kHz, EN_{high}, using $X[k][0]$ and $X[k][1]$, and calculates $EN_{ratio} = (EN_{low} / EN_{high})$. Then, in frames with EN_{ratio} larger than a fixed threshold, L_x is set to 4 kHz; in other frames, L_x is set to 3 kHz. It was measured that, for a dataset for performance evaluation, the selection ratio between $L_x = 3$ kHz and $L_x = 4$ kHz is about 0.4:0.6.

Figure 1 shows $X[k][m]$, $-6 \le m \le 1$, with varying k_L, used in the proposed coding method, where $k = k_U$ defines a coding bandwidth, U_x Hz, and the bold line shows varying k_L for each frame. $X[k][m]$ is grouped into four exclusive sets. Set $\boldsymbol{R}$ consists of $X[k][m]$ in a 2D check pattern above L_x Hz that is not transmitted to the decoder in the current frame. This 2D check pattern ensures full utilization of 2D correlation in T/F domain when recovering $\boldsymbol{R}$. Set $\boldsymbol{Q}_0$ consists of $X[k][m]$ that is transmitted to the decoder in the current frame, and set $\boldsymbol{Q}_1$ consists of $X[k][m]$ that was transmitted to the decoder in the past frames. Set $\boldsymbol{P}$ consist of non-transmitted $X[k][m]$ in the past frames.

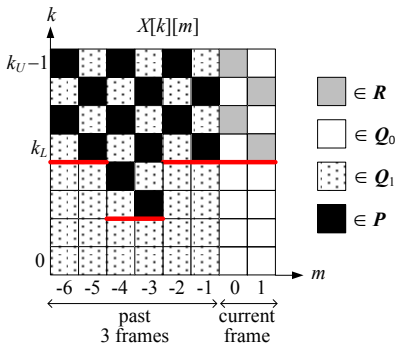

Fig. 1. The structure of 2D MDCT coefficients used in the proposed coding method.

For each frame, the encoder quantizes $\boldsymbol{Q}_0$ as in the general transform coding and transmits it to the decoder, and the decoder recovers $|X[k][m]| \in \boldsymbol{R}$ based on the quantized $\boldsymbol{Q}_0 \cup \boldsymbol{Q}_1$ by the CNN. $\boldsymbol{P}$ is not used when recovering $|X[k][m]| \in \boldsymbol{R}$, because $\boldsymbol{P}$ is recursively deteriorated by the recovery error and causes poor operation of CNN, if the CNN was trained using correct values of $\boldsymbol{P}$.

An experiment on 1D spectral recovery was conducted to confirm the validity of using 2D spectral recovery in the proposed method. In the 1D spectral recovery, for $k \ge k_L$, $|X[k][m]|$ of odd k is recovered based on $|X[k][m]|$ of even k. An informal subjective evaluation shows that the 1D recovery provides significantly lower sound quality than the 2D recovery, which confirms that the data correlation in both time and spectral directions plays a key role in enhancing the performance of spectral recovery.

2.2. Convolutional neural network

The proposed method uses a CNN of a basic structure [5]. To prepare an appropriate input to the CNN, $X[k][m]$ is converted to a new 2D matrix $Y[k][m]$ by shrinking it down in a k axis by half; for $0 \le k < k_U/2$, $Y[k][m] = X[2k][m]$ for even m and $Y[k][m] = X[2k+1][m]$ for odd m. Referring to Fig. 1, $Y[k][m]$ contains only the elements of $\boldsymbol{Q}_0 \cup \boldsymbol{Q}_1$, after half of the elements of $\boldsymbol{Q}_0 \cup \boldsymbol{Q}_1$ below L_x Hz are deleted. Despite that some useful information in $\boldsymbol{Q}_0 \cup \boldsymbol{Q}_1$ below L_x Hz is lost, $Y[k][m]$ has an advantage of having the same structure below and above L_x Hz and the same local relation in the entire region, which is a crucial condition required for local convolution operations in the CNN. This $Y[k][m]$ also makes the size of the neural network small.

The CNN inputs $|Y[k][m]|$, $0 \le k < k_U/2$, $-6 \le m \le 1$, and outputs the recovered $|X[k][m]|$, $0 \le k < k_U$, $m = 0, 1$, including the elements of $\boldsymbol{Q}_0$. Among the recovered $|X[k][m]|$, only the elements of $\boldsymbol{R}$ are used as the final recovered data, and those of $\boldsymbol{Q}_0$ are replaced by the transmitted $\boldsymbol{Q}_0$. In this input-output configuration of CNN, different coding structures of $L_x = 3$ kHz and $L_x = 4$ kHz can be handled using a single CNN.

When $U_x = 14.25$ kHz and the sampling rate is 48 kHz, which will be used in the performance evaluation, the CNN for spectral recovery is designed as summarized in Table 1. The input to the CNN is $|Y[k][m]|$, $0 \le k < 304$, $-6 \le m \le 1$, with a size of 304 × 8. The CNN consists of two parts: encoding and decoding networks. The encoding network comprises of five layers and encodes the

Table 1. The structure of convolutional neural network.

	layer	output shape	no. of filters	filter size	stride
encoding network	1	[152, 8]	32	[5, 5]	[2, 1]
	2	[152, 4]	64	[5, 5]	[1, 2]
	3	[76, 4]	128	[5, 5]	[2, 1]
	4	[76, 2]	256	[5, 5]	[1, 2]
	5	[38, 2]	512	[5, 5]	[2, 1]
decoding network	1	[76, 2]	256	[5, 3]	[2, 1]
	2	[152, 2]	128	[5, 3]	[2, 1]
	3	[304, 2]	64	[5, 3]	[2, 1]
	4	[608, 2]	32	[5, 3]	[2, 1]
	5	[608, 2]	1	[5, 3]	[1, 1]

input to 38 × 2 latent variables, by performing 2D convolution and rectified linear unit (ReLU) activation function with a decreasing output size. The decoding network also comprises of five layers and recovers the output from the latent variables. The first four layers perform 2D transpose convolution and ReLU activation function with an increasing output size, and the last output layer performs tanh activation function and outputs 608 × 2 data corresponding to $|X[k][m]|$, $0 \le k < 608$, $m = 0, 1$.

The training of CNN is done without the quantization in $\boldsymbol{Q}_0$ and $\boldsymbol{Q}_1$. This might result in mismatch between the training and testing of neural network, but the mismatch effect is not significant and the dependency of neural network on the bit rate and the specific operation of individual encoder can be eliminated. The CNN is trained by ADAM [17] using L1 cost function and a mini-match size of 1024.

2.3. Coding of MDCT signs

For reducing the coding bit rate, the number of transmitted signs of $\boldsymbol{R}$ needs to be as small as possible, subject to acceptable performance. The proposed method adopts a scheme of selective sign transmission, where a given number of signs of $X[k][m] \in \boldsymbol{R}$ that are expected to be more important than other signs are selected and transmitted in a descending order of sign importance, using one bit for each sign. This scheme works properly as long as the encoder and decoder can compute the same order of sign importance of $\boldsymbol{R}$ without any side information.

It was decided that the sign importance of $X[k][m]$ is estimated simply by $|X[k][m]|$. However, the recovered $|X[k][m]| \in \boldsymbol{R}$ cannot be used for this purpose, because the encoder and decoder may have different values of $|X[k][m]|$ due to numerical error in floating-point operation. Rather, in the proposed method, the quantized values of $X[k][m] \in \boldsymbol{Q}_0$ adjacent to $X[k][m] \in \boldsymbol{R}$ are used in replace of $X[k][m] \in \boldsymbol{R}$. Fig. 2 shows a block diagram of selective sign transmission. For each $X[k][m] \in \boldsymbol{R}$, $S[k][m] = |X[k-1][m]| + |X[k+1][m]|$ is computed after quantization, where $X[k-1][m] \in \boldsymbol{Q}_0$ and $X[k+1][m] \in \boldsymbol{Q}_0$. $S[k][m]$'s are sorted in a descending order; in case of the same $S[k][m]$'s, $S[k][m]$ of smaller k and smaller m comes first. Let N_{sign} be the number of signs to be transmitted. Then, the N_{sign} signs of $X[k][m]$'s in $\boldsymbol{R}$ corresponding to the N_{sign} largest $S[k][m]$'s are transmitted in the same order of $S[k][m]$. Due to using quantized $\boldsymbol{Q}_0$, the encoder and decoder can determine the same order as in Fig. 2, assuming no bit transmission error. The non-transmitted signs are randomly set in the decoder, which is motivated by the IGF, where the missing MDCT signs are copied from other MDCT signs without causing severe quality degradation [12]. In this way, the number of sign bits can be reduced, while focusing on transmitting more important signs.

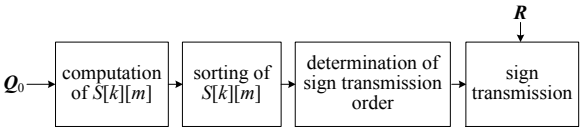

Fig. 2. Block diagram of selective sign transmission of $\boldsymbol{R}$.

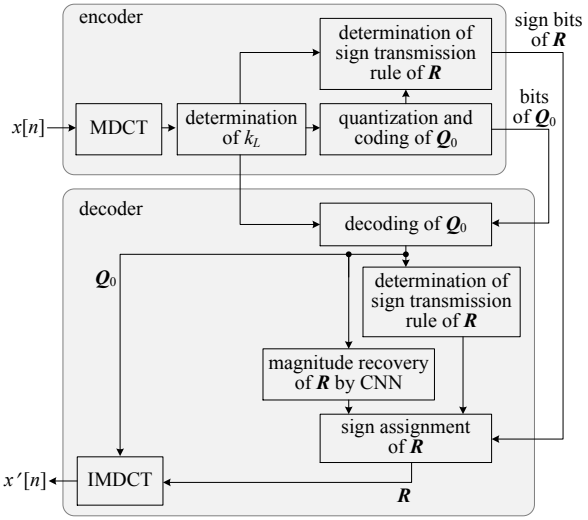

Fig. 3. Block diagram of the proposed coding method based on spectral recovery.

In the proposed method, $N_{sign} = 40$ is used when $L_x = 4$ kHz and $N_{sign} = 100$ is used when $L_x = 3$ kHz. In case of $L_x = 4$ kHz, the correct signs in 3 kHz ~ 4 kHz are transmitted as the elements of $\boldsymbol{Q}_0$ and a small N_{sign} can be used, compared with the case of $L_x = 3$ kHz.

2.4. Overall operation of proposed coding method

A block diagram of the proposed coding method, called an Audio Coding based on Spectral Recovery (ACSR), is shown in Fig. 3. $X[k][0]$ and $X[k][1]$ are computed by applying two MDCTs to $x[n]$ of 2048 samples. For each frame, k_L is determined based on the band energy ratio. The quantization of $\boldsymbol{Q}_0$ follows the general transform coding on a sub-frame basis by computing scale-factors for each sub-frame. Due to the independent quantization of $\boldsymbol{Q}_0$ for each sub-frame, the ACSR has the same spectral and temporal resolution to the USAC with a frame length of 1024 samples, but with an extra processing delay of 1024 samples.

An entropy coding is applied to the quantization indices of $\boldsymbol{Q}_0$ and the resulting bits are transmitted. Instead of developing a new entropy coder optimized for $\boldsymbol{Q}_0$, in this study, it was decided to use an arithmetic coder of the USAC, without considering the difference in structure between a 2D T/F domain in this study and a 1D spectral domain in the USAC. Since the arithmetic coding in the USAC deals with 1D spectral data, $X[k][m] \in \boldsymbol{Q}_0$ is converted to 1D data by two scanning patterns up to U_x Hz as shown in Fig 4 and the resulting 1D quantization indices are input to the USAC arithmetic coder. A scanning pattern having the smaller number of bits is then selected. A state of arithmetic coder is reset every frame

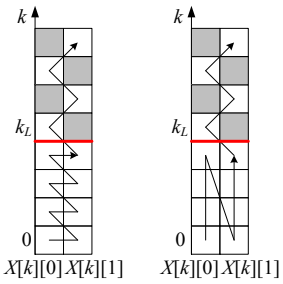

Fig. 4. Two scanning patterns for converting 2D MDCT coefficients into 1D data for entropy coding.

to implement the switching of scanning pattern and k_L. Using the quantized $\boldsymbol{Q}_0$, a rule of transmitting signs of $X[k][m] \in \boldsymbol{R}$ is determined and the signs are transmitted to the decoder accordingly.

In the decoder, with the transmitted k_L, $\boldsymbol{Q}_0$ is reconstructed as in the conventional transform coding. Based on $\boldsymbol{Q}_0$, a rule that maps each transmitted sign to $X[k][m] \in \boldsymbol{R}$ is determined. $|X[k][m]| \in \boldsymbol{R}$ is recovered by the CNN after setting $|Y[k][m]|$ to an input. The signs of recovered $|X[k][m]| \in \boldsymbol{R}$ are assigned either by the transmitted sign bits or randomly. $\boldsymbol{Q}_0$ and $\boldsymbol{R}$ together are transformed into $x'[n]$ of 2048 samples.

3. PERFORMACNE EVALUATION

In performance evaluation of the ACSR, a training dataset and a validation dataset of about 57-hours in total, both collected from audio CDs, RWC music database [18], and speech database [19] are used. The training is run by 100 epochs. A testing dataset consists of 12 audio clips in three categories – speech, music, and speech over music (SoM), and has a total length of 165 seconds [20]. All waveforms have a sampling frequency of 48 kHz. Only mono coding without the temporal noise shaping and window switching is considered for evaluation; operations for short window and multi-channels will be added at the next developing stage.

The performance of ACSR is compared to that of the USAC frequency-domain mode in terms of subjective sound quality. For a fare comparison between the two coding methods, they should use the same psycho-acoustic model and the same process of computing scale-factors for quantizer, in order to make the difference in coding performance be determined only by the difference in coding structure, excluding the difference in quantizer performance. Subsequently, for each sub-frame in the ACSR, the scale-factors are determined in exactly the same way as in the USAC encoder [21], and are applied to $X[k][0]$ and $X[k][1]$.

When using the spectral quantizer of USAC at 48 kbps, where the bandwidth is set to U_x = 14.25 kHz [21], the average number of coding bits for MDCT coefficients in the ACSR and USAC for the testing dataset is measured and summarized as shown in Table 2. For each 2048-sample-long input, the USAC has 608 × 2 frames = 1216 MDCT coefficients and the ACSR has 736 and 778 MDCT coefficients in $\boldsymbol{Q}_0$ for L_x = 3 kHz and 4 kHz, respectively. The ACSR codes 61% ~ 64% of the 1216 MDCT coefficients in the USAC, and the number of coding bits is reduced by 20.4%. Side information, such as scale-factors, is processed on a frame basis and its bit count is almost same for two coding methods. Finally, the overall bit rate is reduced by 18.1%; using the same spectral quantizer, 48.1 kbps in the USAC is reduced to 39.4 kbps in the ACSR.

Table 2. Bit counts when using the spectral quantizer of USAC at 48 kbps.

		Bit rate (kbps)		Reduction rate (%)
		ACSR	USAC	
MDCT coeffs.	$\boldsymbol{Q}_0$	32.4	42.6	20.4
	Signs in $\boldsymbol{R}$	1.5		
Side information		5.5	5.4	-
Total		39.4	48.1	18.1

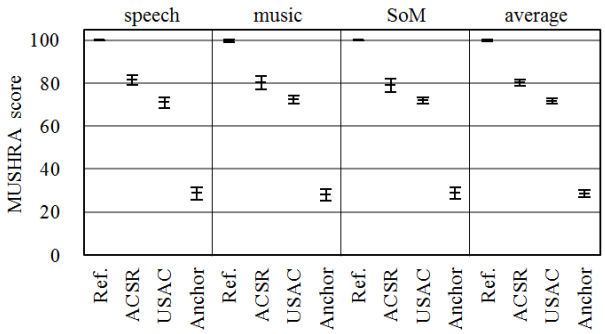

Fig. 5. The MUSHRA scores with 95% confidence interval.

To verify the performance enhancement by the ACSR, an informal subjective performance evaluation was conducted by MUSHRA [22] for the ACSR at 39.4 kbps, the USAC at 39.4 kbps, reference, and 3.5 kHz anchor. Seven subjects participated in the evaluation. Fig. 5 shows the MUSHRA scores with a 95% confidence interval (CI) for each of three categories. At the same bit rate of 39.4 kbps, the ACSR produces sound of significantly higher quality than the USAC in all categories, by an average MUSHRA score of 8.5; the enhancement is most significant in speech signals.

The results of subjective evaluation corroborate that, although limited to the case of bit rate of 39.4 kbps, a combination of both quantization and recovery of MDCT coefficients in the ACSR can provide better coding performance than quantization only as in conventional transform coding, as long as acceptable performance in spectral recovery is guaranteed. Therefore, the proposed ACSR can be a new way of performance enhancement for transform audio coding.

4. CONCLUSION

The aim of this study is to propose a new audio coding method based on 2D spectral recovery by a convolutional neural network. The proposed method represents spectral information in a 2D T/F domain by a sub-frame-based MDCT and only transmits a portion of MDCT coefficients, whereas the remaining MDCT coefficients are recovered in the decoder by the convolutional neural network. The signs of missing MDCT coefficients are either transmitted or randomly assigned, according to their importance. The subjective performance evaluation shows that the proposed coding method at 39.4 kbps provides significantly better sound quality than the USAC at the same bit rate. Therefore, the proposed coding method can be a new way of enhancing the performance of transform audio coding.

5. REFERENCES

[1] ISO/IEC 11172-3, "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s - Part 3," 1993.

[2] Advanced Television Systems Committee (ATSC), "Digital audio compression standard (AC-3)," 1994.

[3] M. Dietz, L. Liljeryd, K. Kjörling, and O. Kunz, "Spectral band replication, a novel approach in audio coding," *112th Conv. Audio Eng. Soc.*, May 2002.

[4] J. Breebaart, *et al.*, "MPEG spatial audio coding / MPEG surround: overview and current status," *119th Conv. Audio Eng. Soc.*, Oct. 2005.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, 521.7553, pp. 436-444, 2015.

[6] K. Li, and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing,* pp. 4395-4399, 2015.

[7] Y. Wang, S. Zhao, W. Liu, M. Li, and J. Kuang, "Speech bandwidth expansion based on deep neural networks," in *Proc. Annual Conf. of Int. Speech Communication Association*, 2015.

[8] V. Kuleshov, S. Zayd Enam, and S. Ermon, "Audio super resolution using neural networks*," arXiv:1708.00853*, 2017.

[9] B.K. Lee, *et al.*, "Sequential deep neural networks ensemble for speech bandwidth extension," *IEEE Access*, vol. 6, pp. 27039-27047, 2018.

[10] T.Y. Lim, *et al*., "Time-frequency networks for audio super-resolution," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 646-650, 2018.

[11] K. Schmidt, and B. Edler, "Blind bandwidth extension based on convolutional and recurrent deep neural networks," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 5444-5448, 2018.

[12] C.R. Helmrich, *et al*., "Spectral envelope reconstruction via IGF for audio transform coding," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 389-393, 2015.

[13] W.B. Kleijn, *et al*., "WaveNet based low rate speech coding," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing,* pp. 676-680, 2018.

[14] K. Oyamada, *et al*., "Generative adversarial network-based approach to signal reconstruction from magnitude spectrograms," *arXiv:1804.02181*, 2018.

[15] A. van den Oord, *et al*., "WaveNet: A generative model for raw audio," *arXiv:1609.03499*, 2016.

[16] ISO/IEC 23003-3, "MPEG audio technologies—Part 3: Unified speech and audio coding," 2012.

[17] D.P. Kingma, and J.L. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. on Learning Representation (ICLP)*, 2015.

[18] M. Goto, "Development of the RWC music database," in *Proc. Int. Congress on Acoustics (ICA)*, pp. I-553-556, April 2004.

[19] C. Veaux, *et al*., "Superseded-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2016.

[20] ISO/IEC JTC1/SC29/WG11 N9927, "Workplan for subjective testing of Unified Speech and Audio Coding proposals," April 2008.

[21] S. Beack, *et al*., "Single-mode-based Unified Speech and Audio Coding by extending the linear prediction domain coding mode," *ETRI Journal*, vol. 39, no. 3, pp. 310-318, 2017.

[22] ITU-R BS.1534-3, "Method for the subjective assessment of intermediate quality level of audio systems," 2015.