INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Coded Speech Enhancement Using Neural Network-Based Vector-Quantized Residual Features

Youngju Cheon[1], *Soojoong Hwang*[1], *Sangwook Han*[1], *Inseon Jang*[2], *and Jong Won Shin*[1]

[1]Gwangju Institute of Science and Technology, Gwangju, Korea
[2]Electronics and Telecommunications Research Institute, Daejeon, Korea

juiscoming@gm.gist.ac.kr, sjhwang@gist.ac.kr, swhan9873@gm.gist.ac.kr, jinsn@etri.re.kr, jwshin@gist.ac.kr

Abstract

Various approaches have been proposed to improve the quality of the speech coded at low bitrates. Recently, deep neural networks have also been used for speech coding, providing a high quality of speech with low bitrates. Although designing an entire codec with neural networks may be more effective, backward compatibility with the existing codecs can be desirable so that the systems with the legacy codec can still decode the coded bitstream. In this paper, we propose to generate side information based on neural networks for an existing codec and enhance the decoded speech with another neural networks using the side information. The vector-quantization variational autoencoder (VQ-VAE) is applied to generate vector-quantized side information and reconstruct the residual features, which are the difference between the features extracted from the original and decoded signals. The post-processor in the decoder side, which is another neural network, takes the decoded signal of the main codec and the reconstructed residual features to estimate the features for the original signal. Experimental results show that the proposed method can significantly improve the quality of the enhanced signals with additional bitrate of 0.6 kbps for two of the implementations of the high-efficiency advanced audio coding (HE-AAC) v1.

Index Terms: Deep Neural Network, Speech Coding, Coded Speech Enhancement, Side Information, VQ-VAE

1. Introduction

Audio signals are compressed by codecs for efficient transmission and storing. As the coding artifacts such as pre-echoes and quantization noise are perceivable in the decoded signals when the codec operates at low bitrates, various pre/post-processor or coding schemes have been proposed. Pre-processing approaches modify the input signal to make the codec output close to the original signal [1, 2]. In the decoder side, postfilters are utilized to enhance the quality of the decoded signals. The postfilters standardized in G.711 [3] and G.718 [4] are designed to reduce quantization noise and emphasize the low-frequency pitch structures. [5] proposed to detect transient signals and then adjust powers of pre- and post-transient part of the signals to reduce the pre-echoes. Some approaches append side information to the bitstream of the original codec to improve the performance. Representative examples are the high-efficiency advanced audio coding (HE-AAC) family [6], in which the HE-AAC v1 is constructed by adding the bitstream for the spectral band replication (SBR) [7] to the bitstream of the AAC Low Complexity (AAC-LC) as side information, and the HE-AAC v2 is formed by appending the bitstream for the parametric stereo [8] to the bitstream of the HE-AAC v1. In this way, the HE-AAC family satisfies the backward compatibility so that the legacy decoder designed for the old standard codecs can decode the bitstream generated by the new coders. [9] improves the quality of the decoded signal by flattening the envelope of the high frequency components of the signal, for which the gain to reconstruct the dynamics of the high frequency envelope is transmitted to the decoder as side information.

Recently, deep learning approaches have been adopted for coded speech enhancement. Convolutional neural networks (CNN) based post-processing approaches were proposed in [10, 11, 12], while the long short-term memory (LSTM) recurrent neural networks (RNN) based method to exploit the temporal and spectral correlations was also proposed in [13]. Generative adversarial networks (GAN) [14] was also applied as a post-processing, which showed improved perceptual quality for speech and applause signals [15]. Neural speech codecs have also been proposed. WaveNet [16] based neural speech codec [17, 18] generates high-quality speech signals, but requires a high computational cost. SampleRNN [19] based speech coding [20] was also proposed. [21] proposed the LPCNet [22] based speech coding. [23] proposed a lightweight and scalable waveform neural codec employing collaborative quantization. Although the fully neural network-based codec may be more effective, backward compatibility with the existing codecs may be desirable so that the billions of devices equipping the legacy codecs can still decode the coded bitstream.

In this paper, we propose to generate side information from the residual error of the decoded signal, and enhance the decoded speech using the quantized side information by neural networks when the conventional codec operates at low bitrates. The decoder of the target codec equipped inside of the encoder is used to obtain the residual feature which is the difference between the original and decoded features, which is coded, quantized, and decoded by the vector-quantization variational autoencoder (VQ-VAE) [24, 25]. The decoded residual feature is used in the post-processor along with the signal decoded by the target codec to estimate the original signal. Experimental results showed that the proposed method enhances the perceived quality of the decoded speech in terms of the objective metric and the subjective test score with an additional bitrate of 0.6 kbps, while satisfying the backward compatibility.

2. Neural representation of the quantized side information for coded speech enhancement

The overall structure of the proposed coded speech enhancement system is described in Figure 1. For backward compatibility, the bitstream generated by the encoder of the main codec is kept intact, and the bitstream for the coded speech enhance-

http://dx.doi.org/10.21437/Interspeech.2021-1204

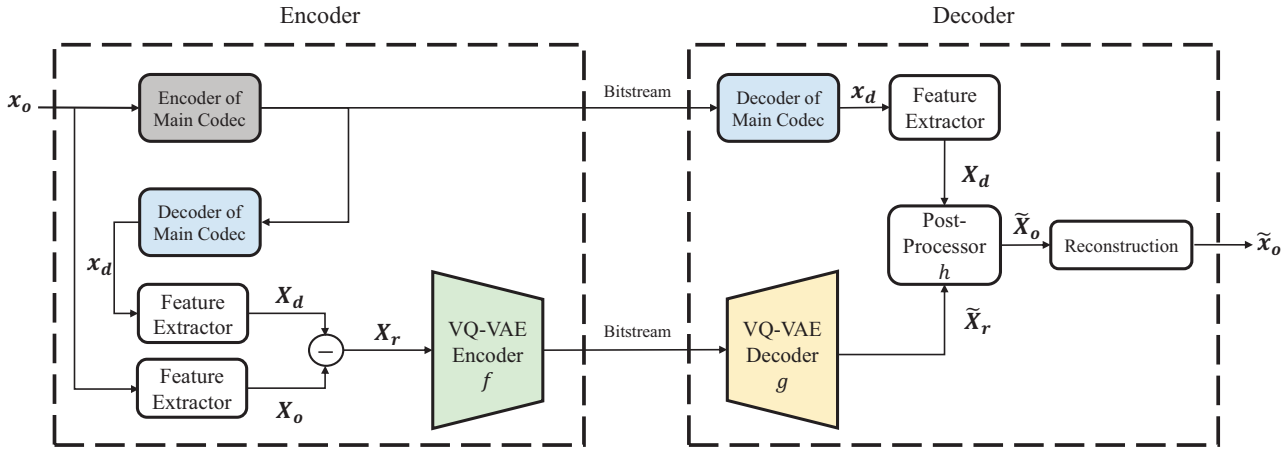

Figure 1: *Overall structure of the proposed method which generates the neural network-based quantized side information on the residual features and reconstruct the original signal using a post-processor with the decoded signal and the side information.*

ment is added to it as side information. There are three neural networks in the proposed system: the VQ-VAE encoder f, the VQ-VAE decoder g, and the post-processor (PP) h. The residual feature that represents the residual error of the coded speech is computed using the decoder of the main codec at the encoder side, which is coded and decoded with the VQ-VAE encoder and decoder, respectively. The post-processor takes the features extracted from the speech decoded by the main codec and the residual feature estimated by the VQ-VAE decoder to estimate the feature for the original signal, which is finally converted to the time domain signal. The detailed description of the blocks are as follows.

2.1. Residual feature

Using the decoder of the main codec equipped inside of the encoder, we can compute the difference of the original and the decoded features. The log power spectra (LPS) are used as features used to enhance the output for the decoder of the main codec. Let $\boldsymbol{x}_o$ and $\boldsymbol{x}_d$ denote the original and the decoded signals in the time domain, while $\boldsymbol{X}_o = [X_o(1), \cdots, X_o(F)]$ and $\boldsymbol{X}_d = [X_d(1), \cdots, X_d(F)]$ represent the LPS for $\boldsymbol{x}_o$ and $\boldsymbol{x}_d$, respectively, in which F is the number of frequency bins. With the assumption that the coding artifacts may be described in finite patterns in the latent domain, the residual feature $\boldsymbol{X}_r$, which is defined as $\boldsymbol{X}_r = \boldsymbol{X}_o - \boldsymbol{X}_d$, is encoded by the VQ-VAE encoder, quantized by the vector quantizer (VQ), and then reconstructed by the VQ-VAE decoder. The reconstructed residual feature $\tilde{\boldsymbol{X}}_r$ is then used with $\boldsymbol{X}_d$ obtained from the decoder of the main codec to reconstruct the signal in the post-processor.

2.2. Compression of the side information using VQ-VAE

The residual features can be efficiently coded using the VQ-VAE [24, 25]. The structure of the VQ-VAE is shown in Figure 2, which consists of an encoder f, a VQ, and a decoder g. To alleviate the difficulty that the gradient cannot propagate through the VQ, the VQ-VAE ignores the effect of the VQ when computing the reconstruction loss. Also, the stop gradient operator $sg(\cdot)$ is introduced to separate the updates of the codebook and the encoder.

The input to the VQ-VAE encoder is $\boldsymbol{X}_r$, and the output of it becomes $f(\boldsymbol{X}_r)$. The encoded residual feature $f(\boldsymbol{X}_r)$ is then

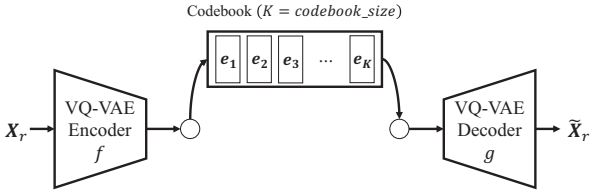

Figure 2: *The vector-quantization variational autoencoder (VQ-VAE) which consists of an encoder, a vector quantizer, and a decoder.*

quantized by the VQ to produce the quantized residual feature $\boldsymbol{e}$, which is the code vector closest to $f(\boldsymbol{X}_r)$ among K vectors in the codebook, $\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_K$, i.e.,

$$\boldsymbol{e} = \boldsymbol{e}_k, \text{ where } k = \arg\min_j \|f(\boldsymbol{X}_r) - \boldsymbol{e}_j\|. \quad (1)$$

The codebook index k becomes the side information transmitted to the decoder side with an additional bitrate of $\lceil \log_2 K \rceil \times f_s/L$, where f_s is the sampling rate and L denotes the frame shift in sample. In the decoder side, the quantized residual feature $\boldsymbol{e} = \boldsymbol{e}_k$ is fed into the VQ-VAE decoder g to reconstruct $\boldsymbol{X}_r$ as $\tilde{\boldsymbol{X}}_r = g(\boldsymbol{e}_k)$. The loss function for the VQ-VAE in [25] is given as

$$\begin{aligned}\mathcal{L}_1 = &\left\|\boldsymbol{X}_r - \tilde{\boldsymbol{X}}_r\right\|_2^2 \\ &+ \left\|sg(f(\boldsymbol{X}_r)) - \boldsymbol{e}\right\|_2^2 + \beta\left\|sg(\boldsymbol{e}) - f(\boldsymbol{X}_r)\right\|_2^2 \end{aligned} \quad (2)$$

where β is a weighting factor. The first term is called the reconstruction loss, the second term is the codebook loss to train the code vectors, and the last term is called commitment loss to guide the encoder to produce latent vectors close to one of the code vectors. This loss function can be adopted for the proposed method, but separate training of the VQ-VAE and the post-processor h may not be optimal.

2.3. Post-processor and joint training with VQ-VAE

The output of the VQ-VAE, $\tilde{\boldsymbol{X}}_r$, is concatenated with the feature vector extracted from the decoded signal, $\boldsymbol{X}_d$, and then fed

Table 1: *The structure of the VQ-VAE encoder and decoder used in the experiments. PReLU stands for parametric rectified linear unit.*

Network	No.	No. of filters	Activation function
Encoder	1	128	PReLU
	2	64	PReLU
	3	32	Linear
Decoder	1	64	PReLU
	2	128	PReLU
	3	257	Linear

into the post-processor h to estimate the feature vector for the original signal. The estimated feature vector $\tilde{\boldsymbol{X}}_o$ is given by

$$\tilde{\boldsymbol{X}}_o = h(\tilde{\boldsymbol{X}}_r, \boldsymbol{X}_d). \tag{3}$$

Instead of training the post-processor h separately, we train the post-processor and the VQ-VAE jointly using the loss function with a modified reconstruction loss given by

$$\begin{aligned}\mathcal{L}_2 = &\left\|\boldsymbol{X}_o - \tilde{\boldsymbol{X}}_o\right\|_2^2 \\ &+ \left\|sg(f(\boldsymbol{X}_r)) - \boldsymbol{e}\right\|_2^2 + \beta\left\|sg(\boldsymbol{e}) - f(\boldsymbol{X}_r)\right\|_2^2\end{aligned} \tag{4}$$

in which the first term is now dependent on f, g, and h. The output of the post-processor is converted back to the time domain signal by the inverse STFT using the phase of the signal decoded by the main codec.

3. Experiments

3.1. Target codec and dataset

To evaluate the performance of the proposed approach, we used the HE-AAC v1 [6] as the main codec for which we append the side information to enhance the quality of the decoded signal. HE-AAC v1 is an extension of the AAC-LC which improves compression efficiency in the frequency domain by using the SBR [26], and is used in billions of devices that provide broadcasting or streaming media. Among the various implementations of the HE-AAC, NeroAAC [27] and QAAC [28] are used in the experiments. We conducted the experiments for five bitrates of the HE-AAC v1, which were 10, 12, 16, 20, and 24 kbps.

The speech databases used for the experiments were the clean VCTK corpus [29] from the Voice Bank database [30] and the TIMIT corpus [31], which are monaural corpora resampled at 16 kHz. 11,572 utterances spoken by 26 speakers from the VCTK dataset and the 4,620 utterances from 462 speakers in the TIMIT corpus were used as the training data, while 1,819 utterances spoken by 4 speakers from the VCTK corpus and the 240 utterances spoken by 24 speakers from the TIMIT database were used as the validation set. The test set consisted of 1,680 utterances from the TIMIT corpus spoken by 168 speakers and 824 utterances from the VCTK corpus spoken by 2 speakers.

3.2. Model configurations

The feature extractor computed LPS using the 512-point short-time Fourier transform (STFT) for a 32 ms window with 50% overlap. The square root Hann window was used as the analysis and synthesis windows. LPS extracted from the original and decoded signals were normalized to have zero mean and unit variance. The codebook size K was set to 512, so $\lceil \log_2 512 \rceil = 9$ bits of the side information was generated for every 16ms, resulting in approximately 0.6 kbps of the additional bitrate. The VQ-VAE encoder and decoder consisted of stacked layers of 1D-convolutions with stride 1, for which the details are summarized in Table 1. The β in the loss function of Eq. (4) was set to 0.25. The post-processor consists two fully connected layers with 1024 and 257 units, respectively, and the activation function was the parametric rectified linear unit (PReLU). To investigate the effectiveness of neural network-based side information, we compared the performance of the proposed method with that of the decoded signal enhanced without side information, i.e., using the post-processor alone. The structure of the post-processor without the side information was the same with h of the proposed method except the input dimension. More number of layers in the post-processor did not improve the performance. For all experiments, we use adaptive moment estimator (Adam) as the optimizer with the learning rate initialized as 0.0001 and decreased by a factor of 0.97 for every epoch. The models were trained for 100 epochs with the batch size of 512 for each bitrate. The training is early-stopped if there is no improvement in the reconstruction loss on the validation set for five consecutive epochs.

Table 2: *Average wideband PESQ scores for the decoded signal, post-processed output without side information (+PP only), and the output of the proposed system with VQ-VAE-based side information and post-processor (Prop.) for various bitrates and the baseline codecs of the NeroAAC and the QAAC.*

Bitrate	NeroAAC	+PP only	Prop. (+0.6 kbps)
10 kbps	1.92	2.31	3.13
12 kbps	2.06	2.56	3.32
16 kbps	2.34	2.89	3.55
20 kbps	2.53	3.11	3.68
24 kbps	2.72	3.30	3.84

Bitrate	QAAC	+PP only	Prop. (+0.6 kbps)
10 kbps	1.89	2.46	2.92
12 kbps	2.03	2.69	3.18
16 kbps	2.53	3.04	3.62
20 kbps	2.89	3.39	3.85
24 kbps	3.09	3.54	4.00

3.3. Experimental Results

For objective assessment of the perceptual speech quality, we computed the ITU-T Recommendation P. 862.2 wideband perceptual evaluation of speech quality (PESQ) [32] scores. Table 2 shows the average PESQ scores for the signal decoded by the main codec, the post-processed output without side information (+PP only), and the output of the proposed system with VQ-VAE-based side information and post-processor (Prop.) for various bitrates when the main codec was NeroAAC and QAAC, respectively. We can see that for both of the codecs, the neural post-processor could enhance the quality of the decoded speech and the proposed system can further improve it. Even though the side information in the proposed method requires additional bitrate of about 0.6 kbps, the average PESQ scores for the proposed method were higher than those for the output of the post-processor without side information for much higher bitrate. For example, the proposed method applied to the NeroAAC oper-

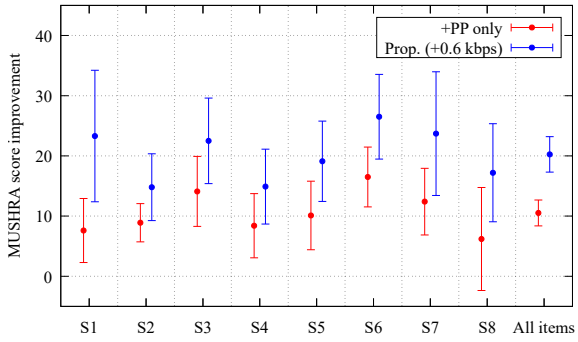

Figure 3: *MUSHRA score improvement over signal decoded by NeroAAC operating at 16 kbps for the post-processing only and the proposed system. Error bars indicate 95% confidence intervals.*

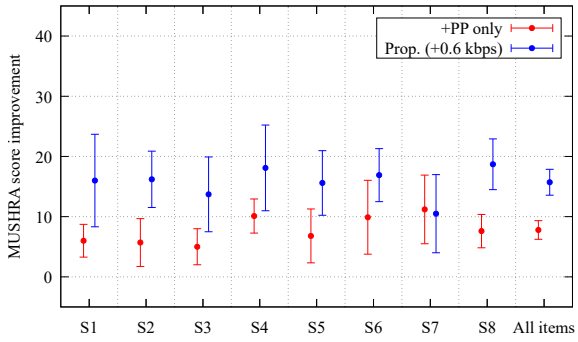

Figure 4: *MUSHRA score improvement over signal decoded by QAAC operating at 16 kbps for the post-processing only and the proposed system. Error bars indicate 95% confidence intervals.*

ating at 10 kbps, which requires 10.6 kbps in total, exhibited comparable performance to the enhanced signal without side information for 20 kbps.

In addition to the objective evaluation, we have conducted MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) tests [33, 34] to assess the subjective quality. Two MUSHRA sessions were configured for the NeroAAC and QAAC operating at 16 kbps, respectively. Each session includes eight trials comparing the quality of the decoded speech, decoded speech enhanced without side information, and the output of the proposed method along with hidden reference and the 3.5 kHz low-pass anchor. 10 listeners participated in the test. Figure 3 and 4 show the average MUSHRA score improvement over the speech decoded by HE-AAC v1 for the enhanced speech without (+PP only) and with the side information (Prop. (+0.6 kbps)). On average, the proposed method outperformed the post-processor without side information by 9.73 points for the NeroAAC and 7.93 points for the QAAC, respectively, which are considered to be statistically significant.

Figure 5 and Figure 6 show the spectrograms for the original speech, enhanced speech with and without side information, and decoded signal for NeroAAC and QAAC operating at 16 kbps, respectively. It can be seen from the area indicated by the dotted box that the proposed method recovers the harmonic structure in the high frequency band well, which was shown to

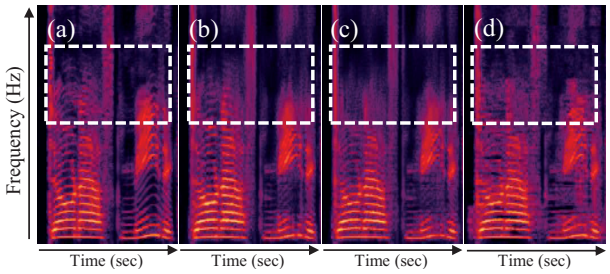

Figure 5: *Spectrograms of a test sample for (a) original signal, (b) output of the proposed method, (c) enhanced signal using post-processor only, and (d) signal decoded by NeroAAC operating at 16 kbps.*

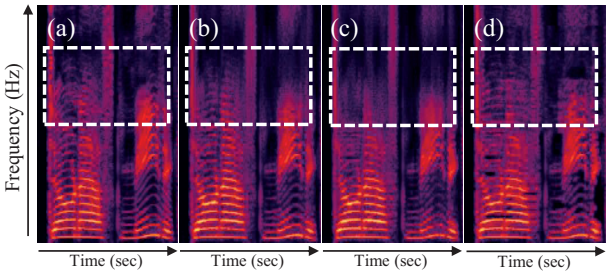

Figure 6: *Spectrograms of a test sample for (a) original signal, (b) output of the proposed method, (c) enhanced signal using post-processor only, and (d) signal decoded by QAAC operating at 16 kbps.*

be difficult without side information.

4. Conclusions

In this paper, we propose to encode the residual error in the decoded signal as a quantized side information and reconstruct the original signal using neural networks to enhance the quality of the decoded signal when the conventional codec operates at low bitrates. The VQ-VAE loss function is adopted with the modification to jointly optimize the VQ-VAE and the post-processor. The objective and subjective evaluation results demonstrated that the proposed coded speech enhancement utilizing 0.6 kbps of side information outperformed the speech enhanced without side information even when the speech was coded at much higher bitrates for two implementations of the high-efficiency advanced audio coding (HE-AAC), NeroAAC and QAAC, while maintaining the backward compatibility with the target codecs.

5. Acknowledgements

This work was supported in part by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [21ZH1200, The research of the basic media·contents technologies] and the National Research Foundation of Korea grant number NRF-2019R1A2C2089324.

6. References

[1] J.-H. Chang, J.-W. Shin, S. Y. Lee, and N. S. Kim, "A new structural preprocessor for low-bit rate speech coding," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[2] J. W. Shin and N. S. Kim, "Signal modification for adpcm based on analysis-by-synthesis framework," *IEEE Signal Processing Letters*, vol. 13, no. 3, pp. 177–179, 2006.

[3] C. Recommendation, "Pulse code modulation (pcm) of voice frequencies," in *ITU*, 1988.

[4] M. Jelinek, T. Vaillancourt, and J. Gibbs, "G. 718: A new embedded speech and audio coding standard with high resilience to error-prone transmission channels," *IEEE Communications Magazine*, vol. 47, no. 10, pp. 117–123, 2009.

[5] J. Lapierre and R. Lefebvre, "Pre-echo noise reduction in frequency-domain audio codecs," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 686–690.

[6] J. Herre and M. Dietz, "Mpeg-4 high-efficiency aac coding [standards in a nutshell]," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 137–142, 2008.

[7] P. Ekstrand, "Bandwidth extension of audio signals by spectral band replication," in *in Proceedings of the 1st IEEE Benelux Workshop on Model Based Processing and Coding of Audio (MPCA'02*. Citeseer, 2002.

[8] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegard, "Low complexity parametric stereo coding," in *Audio Engineering Society Convention 116*. Audio Engineering Society, 2004.

[9] F. Ghido, S. Disch, J. Herre, F. Reutelhuber, and A. Adami, "Coding of fine granular audio signals using high resolution envelope processing (hrep)," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 701–705.

[10] Z. Zhao, H. Liu, and T. Fingscheidt, "Convolutional neural networks to enhance coded speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 663–678, 2018.

[11] S.-H. Shin, S. K. Beack, W. Lim, and H. Park, "Enhanced method of audio coding using cnn-based spectral recovery with adaptive structure," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 351–355.

[12] S. Korse, K. Gupta, and G. Fuchs, "Enhancement of coded speech using a mask-based post-filter," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6764–6768.

[13] J. Deng, B. Schuller, F. Eyben, D. Schuller, Z. Zhang, H. Francois, and E. Oh, "Exploiting time-frequency patterns with lstm-rnns for low-bitrate audio restoration," *Neural Computing and Applications*, vol. 32, no. 4, pp. 1095–1107, 2020.

[14] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.

[15] A. Biswas and D. Jia, "Audio codec enhancement with generative adversarial networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 356–360.

[16] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[17] W. B. Kleijn, F. S. Lim, A. Luebs, J. Skoglund, F. Stimberg, Q. Wang, and T. C. Walters, "Wavenet based low rate speech coding," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 676–680.

[18] C. Gârbacea, A. van den Oord, Y. Li, F. S. Lim, A. Luebs, O. Vinyals, and T. C. Walters, "Low bit-rate speech coding with vq-vae and a wavenet decoder," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 735–739.

[19] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. C. Courville, and Y. Bengio, "Samplernn: An unconditional end-to-end neural audio generation model," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=SkxKPDv5xl

[20] J. Klejsa, P. Hedelin, C. Zhou, R. Fejgin, and L. Villemoes, "High-quality speech coding with sample rnn," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7155–7159.

[21] J.-M. Valin and J. Skoglund, "A real-time wideband neural vocoder at 1.6 kb/s using lpcnet," *arXiv preprint arXiv:1903.12087*, 2019.

[22] ——, "Lpcnet: Improving neural speech synthesis through linear prediction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.

[23] K. Zhen, M. S. Lee, J. Sung, S. Beack, and M. Kim, "Efficient and scalable neural residual waveform coding with collaborative quantization," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 361–365.

[24] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *NIPS*, 2017.

[25] A. Razavi, A. v. d. Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," *arXiv preprint arXiv:1906.00446*, 2019.

[26] M. Dietz, L. Liljeryd, K. Kjorling, and O. Kunz, "Spectral band replication, a novel approach in audio coding," in *Audio Engineering Society Convention 112*. Audio Engineering Society, 2002.

[27] "Neroaac," http://www.nero.com/eng/technologies-aac-codec.html, accessed: March 10, 2021.

[28] "Qaac," https://sites.google.com/site/qaacpage/, accessed: March 10, 2021.

[29] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks." in *Interspeech*, 2016, pp. 352–356.

[30] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*. IEEE, 2013, pp. 1–4.

[31] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[32] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[33] B. Series, "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.

[34] B. ITU-R, "Method for the subjective assessment of intermediate quality level of coding systems (mushra)," 2001.