IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 5, 2021, accepted August 26, 2021, date of publication August 30, 2021, date of current version September 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3108784

Enhancement of Coded Speech Using Neural Network-Based Side Information

SOOJOONG HWANG[1], (Graduate Student Member, IEEE), YOUNGJU CHEON[1], SANGWOOK HAN[1], (Graduate Student Member, IEEE), INSEON JANG[2], AND JONG WON SHIN[1], (Member, IEEE)

[1]School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, Buk-gu 61005, South Korea
[2]Electronics and Telecommunications Research Institute, Daejeon, Yuseong-gu 34129, South Korea

Corresponding author: Jong Won Shin (jwshin@gist.ac.kr)

This work was supported by the Electronics and Telecommunications Research Institute (ETRI) Grant funded by the Korean Government (The research of the basic media contents technologies) under Grant 21ZH1200.

ABSTRACT Audio codecs generate notable artifacts when operating at low bitrates, which degrade the quality of the coded audio significantly. There have been several approaches to enhance the quality of decoded signals with and without side information. While pre- or post-processing approaches without side information can be applied directly to existing systems without modifying codecs, approaches utilizing side information can further enhance the performance while maintaining backward-compatibility with existing codecs. In this paper, we propose a method to improve decoded signals using neural network-based side information. A neural network in the transmitter side that generates the side information and another neural network in the receiver side that estimates the log power spectra (LPS) of the original signal from the decoded signal and the side information are jointly trained to accurately reconstruct the original signal. In the same line with the analysis-by-synthesis, the neural network that generates the side information in the transmitter side takes not only the LPS of the original signal but also the LPS of the decoded signal as the input by decoding the encoded bitstream at the transmitter side. Experimental results show that the proposed audio codec enhancement scheme using neural network-based side information outperformed the audio codec enhancement without side information for the same codec operating at higher bitrates.

INDEX TERMS Audio codec, speech codec, side information, deep neural network, decoded signal enhancement.

I. INTRODUCTION

Speech and audio codecs have been studied extensively to realize higher perceived quality with lower bitrate for either compressed storage or efficient transmission such as speech communication and broadcasting [1]–[4]. However, the quality of decoded signals degrades at low bitrates, due to issues such as quantization noise, pre-echo, and bandwidth limitations. There have been researches to enhance the quality of the decoded signal. Approaches leveraging preprocessing on the transmitter side have been proposed. In such approaches, the input signal is modified so that the codec output for the modified input is closer to the original input signal [5], [6]. Standardized speech codecs such as ITU-T Recommendation G.711 [7] and G.718 [8] have postfilters inside the decoder to reduce quantization noise and enhance low frequency pitch. There have also been approaches to reduce pre-echo with and without side information. A method to detect transient signals in decoded signals, suppress pre-transient signals, and amplify post-transient signals when transient signals are detected has been proposed in [9]. In [10], envelope flattening was applied to high pass signals as a preprocessor to reduce pre-echo, and the gain for envelope flattening was transmitted to the receiver side as side information, which is used to reconstruct the signal during post-processing. Companding [11] has been applied in the quadrature mirror filter (QMF) domain to achieve temporal noise shaping [12], [13] with a lower number of bits. Recently, deep learning-based approaches have been proposed to enhance the quality of decoded signals. Convolutional neural network (CNN)-based post-processing in the cepstral domain was proposed in [14], which enhances low order cepstral coefficients for decoded signals. In [15], long short-term memory recurrent neural networks (RNNs)

The associate editor coordinating the review of this manuscript and approving it for publication was Jonathan Rodriguez.

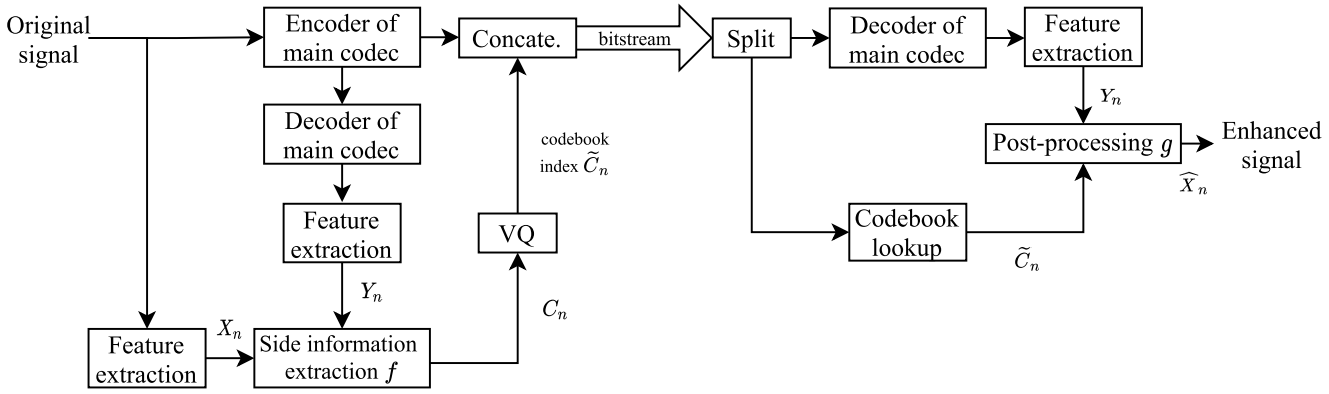


FIGURE 1. Block diagram of decoded signal enhancement using side information.

were employed to exploit temporal and spectral correlations to restore original music from that coded at a low bitrate. Signal restoration using generative adversarial networks was proposed in [16], by assuming that the generative model is capable of recovering components lost by low bitrate coding. Mask-based post-filter was also investigated by employing a convolutional encoder-decoder [17]. In [18], [19], and [20], decoders of parametric coders were constructed using neural speech synthesizers based on WaveNet [21], SampleRNN [22], and LPCNet [23], respectively. Moreover, in [24], the decoder of a waveform matching coder was replaced with LPCNet. The entire coding scheme was replaced by neural networks in [25], showing impressive performance. However, this model was prone to errors that the traditional coding scheme did not induce, such as phoneme mismatches or slurred speech [26].

In this paper, we propose a coded speech enhancement scheme using neural network-based side information. As the generated side information is appended to the bitstream from the encoder of the target codec, the proposed approach can ensure backward-compatibility with existing devices and content, while improving the quality of the decoded signal significantly. Examples of such side information can be seen in the MPEG high-efficiency advanced audio coding (HE-AAC) family [27], [28], where HE-AAC v1 appends the spectral band replication [29] as side information to the bitstream of AAC [30], and HE-AAC v2 adds the parametric stereo [31], [32] on top of the HE-AAC v1 bitstream. A neural network on the transmitter side generates side information from the signals, and another neural network on the receiver side estimates the log power spectra (LPS) of the original signal from the decoded signal and the quantized side information. The side information extraction network on the transmitter side and the post-processing on the receiver side are jointly trained to minimize the loss. Experimental results showed that the proposed scheme outperformed deep neural network (DNN)-based enhancement methods without side information applied to the same codec operating at higher bitrates.

II. ENHANCEMENT OF CODED SPEECH USING DNN-BASED SIDE INFORMATION

Fig. 1 shows a block diagram of the decoded signal enhancement scheme using side information. The coded side information is concatenated with the bitstream generated by the main part of the encoder, and then transmitted to the receiver side. The side information is reconstructed at the receiver side from the received bitstream and is used for post-processing with the decoded signal. The side information transmitted to the receiver side is in many cases hand-crafted features such as voice activities, transient signal detection, and soft gain at time-frequency bins. The postfilter at the receiver side should then be designed accordingly to enhance the audio quality using the side information. In this paper, we propose to employ neural networks for both side information extraction and post-processing. Instead of designing the extractor to estimate hand-crafted features and then constructing the model for the post-processing to incorporate the resultant side information, we trained the networks for side information generation and post-processing simultaneously. As a result, we were able to generate side information and process the decoded signal in a data-driven manner, with an objective function on the similarity between the original input and the final output. Vector quantization (VQ) for the side information was not considered during network training. The codebook for the VQ was constructed using k-means clustering [33] after training two neural networks. We expect that the proposed method would produce higher fidelity sound compared with the output of DNN-based post-processing without side information, even at the higher bitrates.

The neural network models for side information extraction and post-processing can be any deep learning model, such as feed-forward DNN (ffDNN), CNN, RNN, and adversarial loss-based models in the time- or frequency-domain [14]–[17]. In this study, we employed the simplest model to confirm if neural network-based side information generation and post-processing are effective. The neural network that generates side information takes not only the features for the original signal but also those

for the decoded signal obtained by the decoder equipped in the encoder, as they provide important information on what is missing in the decoded signal. The side information extraction was implemented as a CNN that takes the LPS for the original and decoded signals in the current and previous frames as inputs and produces a d-dimensional vector. The post-processing is a ffDNN that estimates the LPS for the original signal using the LPS for the decoded signal in the current and two previous frames and the received side information. In speech enhancement, most widely-used targets may be the LPS [34] and spectral masks [35]. Because audio codecs produce artifacts such as high-frequency cuts, band ruptures, holes, and isolated clusters on spectrograms [36], the LPS for the original audio signal may be more suitable as the target of the post-processing compared with spectral masks. Over-smoothing is a common problem in DNN-based regression approaches, but it may aid in smoothening the spectrogram to relieve the artifacts described above. Two neural networks were trained to minimize the loss function E as follows:

$$E = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\left(X_n^k - \widehat{X}_n^k\right)^2, \tag{1}$$

where N is the mini-batch size, K is the number of frequency bin, X_n^k denotes the LPS for the original signal at the k-th frequency bin in the n-th frame, and $\widehat{X}_n^k = g(\widetilde{\mathbf{C}}_n, \mathbf{Y}_n)$ is the LPS estimated by post-processing from both the LPS for the decoded signal $\mathbf{Y}_n = [Y_n^1, Y_n^2, \ldots, Y_n^K]$ and the transmitted side information $\widetilde{\mathbf{C}}_n$. It should be noted that post-processing considers the quantized side information $\widetilde{\mathbf{C}}_n$ instead of the $\mathbf{C}_n$ generated by the neural network, but the difference between $\widetilde{\mathbf{C}}_n$ and $\mathbf{C}_n$ was ignored in the training. The network f to generate the side information $\mathbf{C}_n$ was jointly trained with the post-processing g to minimize the loss in Eq. (1). Post-processing without side information was also learned to minimize the mean squared error (MSE) between the LPS of the original and estimated signals. The phases of the decoded signal were used along with the estimated LPS to reconstruct the time domain signal. The configuration is described in more detail in the next section.

The extracted side information, which is in the form of a d-dimensional vector, should be quantized appropriately for efficient transmission. In this study, the codebook for the side information was constructed using k-means clustering on the side information vectors for the training set obtained with the trained network f. The codebook index for the quantized side information $\widetilde{\mathbf{C}}_n$ was converted into a bitstream and concatenated with the bitstream from the encoder of the main codec. Although it is desirable to reflect the effect of the VQ of the side information during training as in [25] and [37], VQ did not degrade the performance in the experiments conducted in this study.

III. EXPERIMENTS

A. TARGET CODEC AND DATASET

To demonstrate the performance of the proposed approach, HE-AAC [27], [28] and adaptive multi-rate wideband speech codec (AMR-WB) [38] were used as the baseline codecs. While there are various implementations for HE-AAC v1, we adopted two different codecs, Nero AAC [39] and QAAC [40], in our experiments. The first is the Nero AAC operating at constant bitrates of 20, 21, and 24 kbps. The bitrates for QAAC, which is an open-source wrapper for Apple AAC, were set to constant bitrates of 20 and 24 kbps. As for the AMR-WB, the seven highest bitrates were used for the experiments.

For evaluation, we used speech data sampled at 16 kHz rather than music data because many speech databases are readily available, and adequate objective measures are available for speech quality assessments, such as the ITU-T Recommendation P.862.2 wideband perceptual evaluation of speech quality (PESQ) [41]. Speech signal is also easier to analyze, as a single frame of speech signal can be classified as voiced, unvoiced, or silence. In the experiments, the TIMIT [42], VCTK [43], and Wall Street Journal0 (WSJ0) [44] databases were used, which are monaural speech databases resampled at 16 kHz. The training set in the TIMIT dataset contains 4,620 utterances spoken by 326 male and 136 female speakers, and the test set contains 1,680 utterances spoken by 112 male and 56 female speakers. The VCTK dataset contains 12,396 utterances from 15 males and 15 females. The WSJ0 dataset includes 35,487 utterances spoken by 66 males and 65 females. We built the training set using the training set of the TIMIT database, 11,572 utterances from 14 male and 14 female speakers in the VCTK database, and all utterances from the WSJ0 database. The test set in the TIMIT database and 824 utterances from one male and one female speakers in the VCTK database were used as the test set.

B. COMPARED SYSTEMS AND MODEL CONFIGURATIONS

We analyzed the signals with a 32 ms long Hamming window with a 16 ms frame shift. A 512-point FFT was applied to retrieve the LPS features. The LPS features of the decoded and original signals were normalized to zero-mean and unit-variance. The estimated LPS was denormalized to reconstruct the signal in the time domain. The dimension of the side information $\mathbf{C}_n$ generated by f, d, was set to 3. For the VQ of the side information, 2^{10} centroids were obtained by the k-means algorithm using the side information vectors generated for a randomly selected 10% portion of the training set. As 10 bits of side information was generated for every frame of 16 ms shift, the bitrate for the side information was approximately 0.6 kbps. The proposed method with vector-quantized side information applied to the main codecs operating at 20 or 23.05 kbps is denoted as 20+0.6 kbps enhanced or 23.05+0.6 kbps enhanced in the experimental results. We also evaluated the performance of the proposed

method without VQ on the side information, which resulted in 3 kbps of the side information using three 16-bit numbers to verify the effect of the VQ for the side information on the performance. This system is denoted as 20+3 kbps.

The input size of the network for side information extraction was $2 \times (3 \times 257)$, including the LPS for the two previous frames and the current frame of the original and decoded signals. The input was passed through three convolutional layers without padding. The kernel sizes were (1×257), (3×1), and (1×1), and the parametric rectified linear unit (PReLU) and the sigmoid function were used as the activation functions for the first two layers and the last layer, respectively. The number of output channels were set to 64, 16, and d. The model for post-processing with the side information had an ffDNN structure of 1024-1024-1024 units with the PReLU activation function when the input dimension was 771+3, including the LPS for the current and two previous frames and 3-dimensional side information, and the output dimension was 257. The activation function for the output layer was a linear function. The network structure for side information extraction, f, and post-processing, g, is illustrated in Fig. 2. The filters in the first layer of f in the transmitter side extract features from the original and decoded signals in all frequencies of each of the individual frames, while the filters in the second layer summarize the features across the temporal dimension, which is further compressed in the last layer of f. The CNN was suitable to apply filtering in appropriate dimensions step-by-step. The ffDNN-based post-processing, g, in the receiver side was constructed using a simple structure that takes the transmitted side information along with the LPS from the three frames of the decoded signals to estimate the original LPS. The proposed pre-and post-processing required approximately 200 weighted million operations per second (WMOPS) [45] in addition to that of the main codec. It is noted that the current work focused on demonstrating the effectiveness of using neural network-based side information for coded speech enhancement, and computational complexity was not of primary concern. We compared the performance of the proposed method to that of the decoded signals for the same codec operating at higher bitrates enhanced with ffDNN-based post-processing without any side information and the mask-based post-filter proposed in [17]. The ffDNN structure for post-processing without side information was also 1024-1024-1024 units for three hidden layers, which was the same as that for the post-processing g in the proposed method, except for the input dimension. The mask-based post-filter [17] had a convolutional encoder-decoder structure that takes the log magnitude spectra for the current and previous frames to estimate the ideal ratio masks bounded by 4 and 2 for HE-AAC and AMR-WB, respectively. The ffDNN-based post-processing without side information and the mask-based post-filter required approximately 197 WMOPS and 1187 WMOPS, respectively. The loss functions were minimized using the adaptive moment estimation (Adam) optimizer [46] with a learning rate of 0.0001. The model was trained for 100 epochs with a batch size, N, of 128.

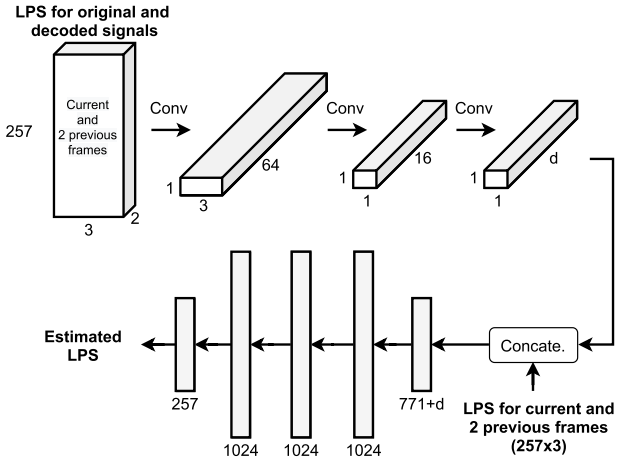
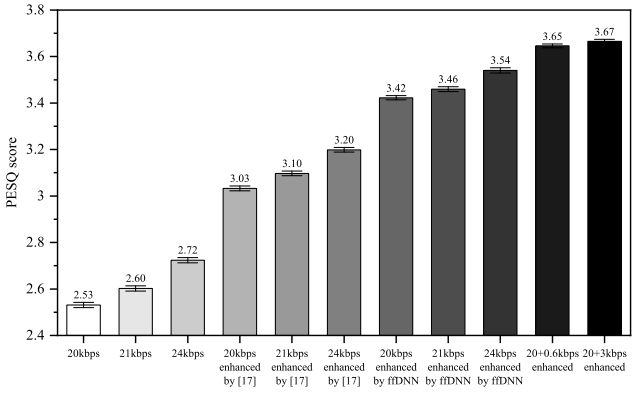


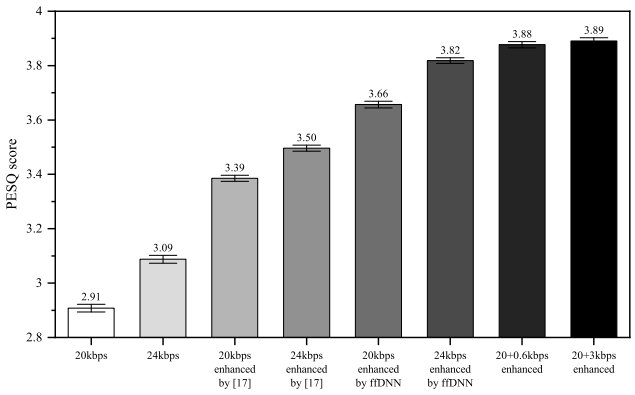
FIGURE 2. The structure of networks for side information extraction f and post-processing g.

IV. EXPERIMENTAL RESULTS

We evaluated the quality of the output signals using an objective measure of subjective quality and a subjective listening test. The objective measure used in the experiments was the wideband PESQ scores [41], which are designed to mimic the ITU-T Recommendation P.800 Absolute Category Rating (ACR) Mean Opinion Score (MOS) test scores [47]. The average PESQ scores for the decoded signals, outputs of the mask-based post-filter in [17], outputs of the ffDNN-based post-processing without side information, and outputs of the proposed method utilizing side information for the HE-AAC operating at various bitrates are shown in Fig. 3. The quality of the decoded signals for the Nero AAC was generally worse than that for the QAAC at the same bitrates. For both the Nero AAC and QAAC codecs, the experimental results show that the average PESQ scores for the decoded speech could be significantly improved by the mask-based post-filter [17] or the ffDNN-based post-processing without side information, and could be further improved by using side information. The average PESQ score for the proposed system with 0.6 kbps of side information applied to the Nero AAC operating at 20 kbps (denoted as *20+0.6 kbps enhanced*) was 0.23 higher than that for the signals enhanced by the ffDNN-based post-processing without side information. *20+0.6 kbps enhanced* outperformed the signals enhanced without side information by 0.11 when the bitrate of the Nero AAC codec was 24 kbps. As for QAAC, the performance improvement over the signals enhanced by the ffDNN-based post-processing without side information for bitrates of 20 and 24 kbps was 0.22 and 0.06 in terms of the PESQ scores, respectively. The 95 % confidence intervals are also shown in Fig. 3. We can conclude that the proposed system with 20+0.6 kbps outperformed the signal enhanced without side information even when the bitrate was 24 kbps

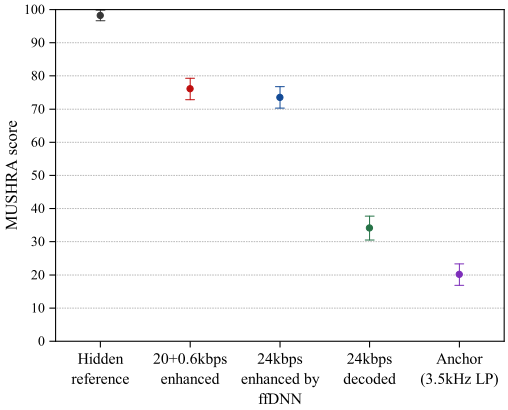


(a) Nero AAC codec

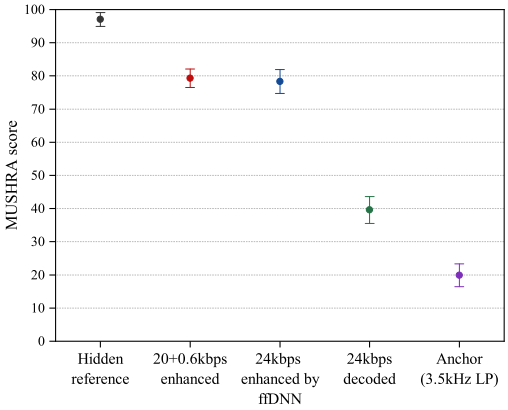


(b) QAAC codec

FIGURE 3. Average PESQ scores for the decoded signals (20, 21, and 24 kbps), the signals enhanced by the mask-based post-filter (20, 21, and 24 kbps enhanced by [17]), those enhanced by the ffDNN-based post-processing without side information (20, 21, and 24 kbps enhanced by ffDNN), and those enhanced by the proposed method with side information (20+0.6 kbps enhanced, 20+3 kbps enhanced) for (a) Nero AAC and (b) QAAC. Whiskers indicate 95 % confidence intervals.



(a) Nero AAC codec



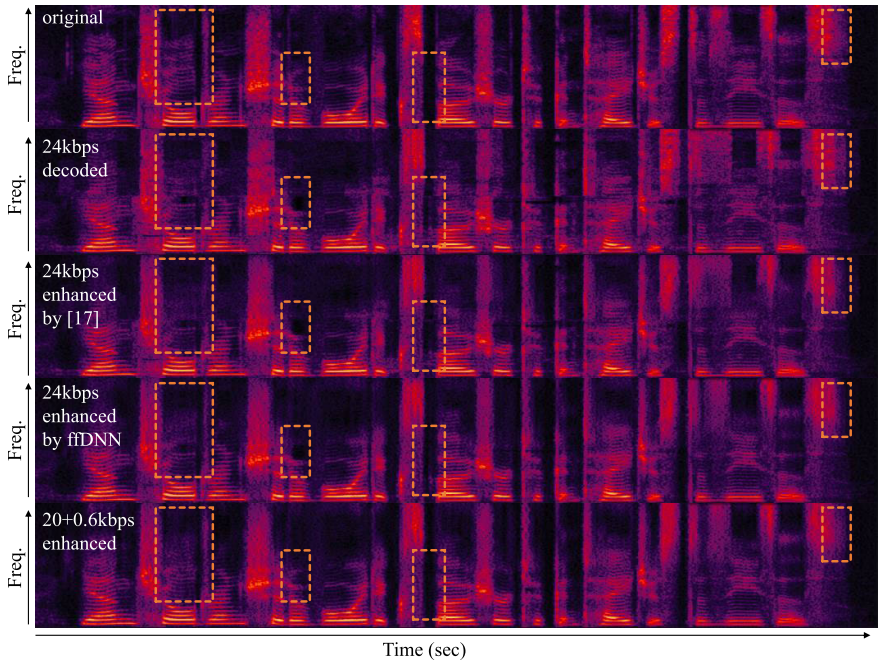
(b) QAAC codec

FIGURE 4. Average MUSHRA scores for 9 items evaluated by 11 listeners comparing the quality of the decoded signal (24 kbps decoded), the signal enhanced by a ffDNN-based post-processing without side information (24 kbps enhanced by ffDNN) for the HE-AAC operating at 24 kbps, and the output of the proposed approach with 0.6 kbps of the side information for the HE-AAC operating at 20 kbps (20+0.6 kbps enhanced) when the HE-AAC codec was (a) Nero AAC and (b) QAAC. Whiskers indicate 95 % confidence intervals.

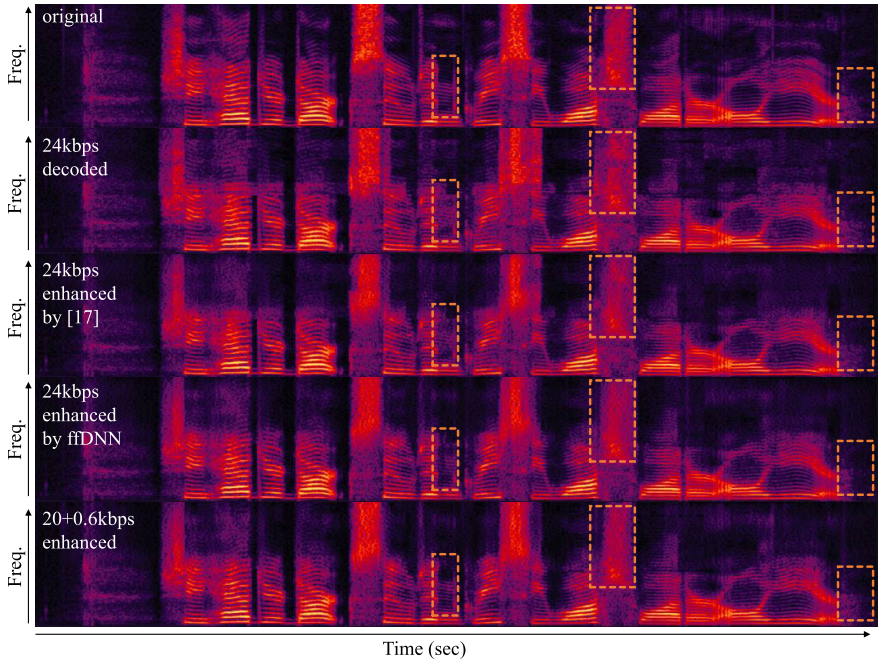
for both Nero AAC and QAAC. Additionally, we evaluated the quality of the signals enhanced by the proposed method without VQ, which resulted in 3 kbps of side information. As observed in Fig. 3, the average PESQ scores for 20+0.6 kbps and 20+3 kbps were very close, implying that 10 bits of vector-quantized side information was suitable to provide information to reconstruct the original signal. We can also speculate that incorporating the effect of the VQ into the neural network training may not improve the average PESQ scores by more than 0.02, as the performance for 20+3 kbps would be upper bound.

The subjective quality was assessed with MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) listening tests [48]. Fig. 4 shows the average MUSHRA scores for nine items evaluated by 11 listeners to compare five systems for each codec: the decoded signals at 24 kbps, the signals enhanced with side information operating at 20+0.6 kbps, and the signals enhanced without side information at 24 kbps, along with the hidden references and 3.5 kHz low-pass (LP) filtered anchors. We can confirm that both the enhanced signals were perceived significantly better than the decoded signals for 24 kbps, and the proposed method exhibited comparable perceptual quality to the 24 kbps enhanced by ffDNN although it required only 20.6 kbps.

Fig. 5 shows the spectrograms of the original signals, the decoded signals (denoted as *24 kbps decoded*), the enhanced signals with the mask-based post-filter (denoted as *24 kbps enhanced by* [17]) and the ffDNN-based post-processing without side information (denoted as *24 kbps enhanced by ffDNN*) for the HE-AAC operating at 24 kbps, and the outputs of the proposed system with 0.6 kbps of side information applied to the HE-AAC operating at 20 kbps (denoted as *20+0.6 kbps enhanced*) for Nero AAC and QAAC. For both codecs, artifacts such as pre-echo, high-frequency cuts, band ruptures, or holes were observed in the spectrograms of the decoded signals at 24 kbps, and these artifacts were mitigated to a certain extent in the enhanced signals. The signals enhanced with side information were more similar to the original signal than others, especially in the areas indicated by the dotted boxes. For the Nero AAC, the spectral holes in the first and the second boxes of *24 kbps decoded*, which are the time-frequency



(a) Nero AAC codec



(b) QAAC codec

FIGURE 5. Spectrograms for the original, decoded, and enhanced signals with the mask-based post-filter in [17] and the ffDNN-based post-processing without side information for the HE-AAC operating at 24 kbps, and the output of the proposed system with 0.6 kbps of side information applied to the HE-AAC operating at 20 kbps when the HE-AAC codec was (a) Nero AAC and (b) QAAC.

components quantized to zero resulting in sharp or instable sounds, were not recovered in *24 kbps enhanced by ffDNN* and *24 kbps enhanced by* [17], but were filled in *20+0.6 kbps enhanced*. In the third boxes, the pre-echo in *24 kbps decoded*, which is the spectral component before the onset of the signal generated by the codec due to the use of long windows, was reduced in *24 kbps enhanced by* [17] and *24 kbps enhanced by ffDNN*, and was completely removed in *20+0.6 kbps enhanced*. As for the last boxes, the reverberation in the original signal, which is a gradual diminishing of spectral components, was present in *24 kbps decoded* and *20+0.6 kbps enhanced*, but was suppressed

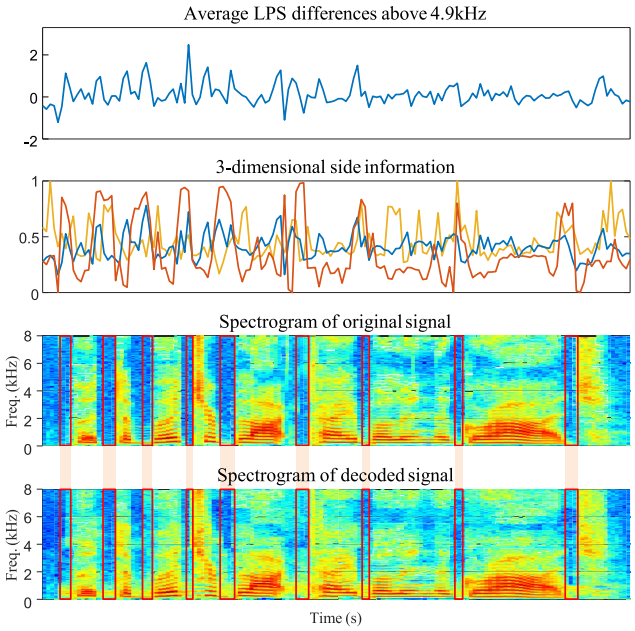


FIGURE 6. LPS differences between the original and decoded signals averaged for frequencies above 4.9 kHz, temporal evolution of the 3-dimensional side information for Nero AAC operating at 20 kbps, and spectrograms of the original and decoded signals.

in *24 kbps enhanced by ffDNN*. In Fig. 5(b), the harmonic structure in the first boxes was more evident in *20+0.6 kbps enhanced*, compared with others. The pre-echo in the second boxes was also removed most successfully in *20+0.6 kbps enhanced*. In the last boxes, the reverberation was preserved better in *20+0.6 kbps enhanced*, compared with *24 kbps enhanced by ffDNN*.

We examined how the extracted side information was related to the characteristics of the input signal. Fig. 6 shows the LPS differences between the original and decoded signals averaged over frequencies above 4.9 kHz for Nero AAC operating at 20 kbps, along with the spectrograms of the original and decoded signals. The side information marked with a blue line in the second plot resembled the average LPS differences in the high frequency shown in the top plot. The red boxes in the spectrograms indicate the frames with pre-echoes, which is the audible component before the onset of the signal that was not present in the original signal. This was introduced by the use of long windows. It can be seen that the red line in the second plot had high values for the regions with pre-echoes. The pattern of the last variable marked with yellow was not as clear as the others; this is possibly because it should consider all the residual information required to reconstruct the original signal.

To verify that the proposed method can enhance the quality of the decoded speech for the speech codec, we performed additional experiments on the AMR-WB codec operating at various bitrates in a similar manner to the experiments with HE-AAC. Fig. 7 shows the average PESQ scores for the decoded signals, signals enhanced with the ffDNN-based post-processing without side information, the mask-based post-filter [17], and the proposed method with 0.6 kbps of side information when the AMR-WB was operating with

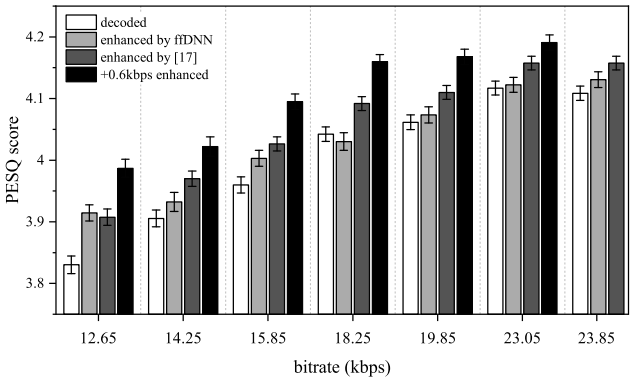


FIGURE 7. Average PESQ scores for the decoded signals, the signals enhanced by the post-processing without side information, the mask-based post-filter [17], and the proposed method with side information, when the baseline codec was AMR-WB operating at various bitrates. Whiskers indicate 95 % confidence intervals.

12.65 to 23.85 kbps. In contrast to the experiments with the HE-AAC codecs, the mask-based post-filter [17] outperformed the ffDNN-based post-processing in this experiment, which is possibly because each codec generates different types of artifacts. The average PESQ scores for the outputs of the proposed method with side information were similar to or higher than those for the signals coded and decoded with higher bitrates and enhanced by [17]. From the experimental results, we can confirm that the proposed method enhanced the quality of the decoded speech effectively using only 0.6 kbps of side information.

V. CONCLUSION

In this paper, we propose a codec enhancement scheme using neural network-based side information. Compared with designing an entire codec with new principles, the proposed approach appends side information to the bitstream generated by legacy coders to provide backward-compatibility for existing devices, while providing adequate performance improvement that cannot be achieved without side information. A CNN generates side information using the LPS for the original and decoded signals in the current and two previous frames, which is vector-quantized and transmitted with the bitstream from the original encoder. On the receiver side, an ffDNN-based post-processing takes the side information and the LPS for three frames of the decoded signal as inputs to estimate the LPS of the original signal. The networks were jointly trained to minimize the mean square error between the original and reconstructed signals. Our experiments on two different HE-AAC and AMR-WB codecs showed that the proposed method with neural network-based side information could achieve better perceptual quality of the output signal than DNN-based enhancement without side information applied to the same codec operating at higher bitrates.

REFERENCES

[1] T. Tremain, "Linear predictive coding systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, Apr. 1976, pp. 474–478.
[2] P. Kroon, E. Deprettere, and R. Sluyter, "Regular-pulse excitation—A novel approach to effective and efficient multipulse coding of speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 5, pp. 1054–1063, Oct. 1986.

[3] R. Salami, C. Laflamme, J.-P. Adoul, and D. Massaloux, "A toll quality 8 kb/s speech codec for the personal communications system (PCS)," *IEEE Trans. Veh. Technol.*, vol. 43, no. 3, pp. 808–816, Aug. 1994.
[4] W. B. Kleijn, R. P. Ramachandran, and P. Kroon, "Interpolation of the pitch-predictor parameters in analysis-by-synthesis speech coders," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 1, pp. 42–54, Jan. 1994.
[5] J.-H. Chang, J.-W. Shin, S. Y. Lee, and N. S. Kim, "A new structural pre-processor for low-bit rate speech coding," in *Proc. Interspeech*, Sep. 2005, pp. 2829–2832.
[6] J. W. Shin and N. S. Kim, "Signal modification for ADPCM based on analysis-by-synthesis framework," *IEEE Signal Process. Lett.*, vol. 13, no. 3, pp. 177–179, Mar. 2006.
[7] *G.711 Amendment 2: New Appendix III—Audio Quality Enhancement Toolbox*, document ITU-T G.711, 2009.
[8] *Frame Error Robust Narrowband and Wideband Embedded Variable Bit-Rate Coding of Speech and Audio From 8–32 kbit/s*, document ITU-T G.718, 2008.
[9] J. Lapierre and R. Lefebvre, "Pre-echo noise reduction in frequency-domain audio codecs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 686–690.
[10] F. Ghido, S. Disch, J. Herre, F. Reutelhuber, and A. Adami, "Coding of fine granular audio signals using high resolution envelope processing (HREP)," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 701–705.
[11] A. Biswas, P. Hedelin, L. Villemoes, and V. Melkote, "Temporal noise shaping with companding," in *Proc. Interspeech*, Sep. 2018, pp. 3548–3552.
[12] J. Herre and J. D. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS)," in *Proc. 101st AES Conv.*, 1996.
[13] J. Herre, "Temporal noise shaping, quantization and coding methods in perceptual audio coding: A tutorial introduction," presented at the AES 17th Conf. High Qual. Audio Coding, Florence, Italy, Sep. 1999.
[14] Z. Zhao, H. Liu, and T. Fingscheidt, "Convolutional neural networks to enhance coded speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 4, pp. 663–678, Apr. 2019.
[15] 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 Comput. Appl.*, vol. 32, no. 4, pp. 1095–1107, Feb. 2020.
[16] A. Biswas and D. Jia, "Audio codec enhancement with generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 356–360.
[17] S. Korse, K. Gupta, and G. Fuchs, "Enhancement of coded speech using a mask-based post-filter," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6764–6768.
[18] W. B. Kleijn, F. S. C. Lim, A. Luebs, J. Skoglund, F. Stimberg, Q. Wang, and T. C. Walters, "WaveNet based low rate speech coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 676–680.
[19] J. Klejsa, P. Hedelin, C. Zhou, R. Fejgin, and L. Villemoes, "High-quality speech coding with sample RNN," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 7155–7159.
[20] J.-M. Valin and J. Skoglund, "A real-time wideband neural vocoder at 1.6 kb/s using LPCNet," 2019, *arXiv:1903.12087*. [Online]. Available: https://arxiv.org/abs/1903.12087
[21] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, pp. 1–15, Sep. 2016.
[22] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, Apr. 2017.
[23] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2019, pp. 5891–5895.
[24] J. Skoglund and J.-M. Valin, "Improving opus low bit rate quality with neural speech synthesis," in *Proc. Interspeech*, Oct. 2020, pp. 2847–2851.
[25] C. Gârbacea, A. van den Oord, Y. Li, F. S. C. Lim, A. Luebs, O. Vinyals, and T. C. Walters, "Low bit-rate speech coding with VQ-VAE and a WaveNet decoder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 735–739.
[26] W. A. Jassim, J. Skoglund, M. Chinen, and A. Hines, "Speech quality factors for traditional and neural-based low bit rate vocoders," in *Proc. 12th Int. Conf. Quality Multimedia Exp. (QoMEX)*, May 2020, pp. 1–6.
[27] J. Herre and M. Dietz, "MPEG-4 high-efficiency AAC coding [standards in a nutshell]," *IEEE Signal Process. Mag.*, vol. 25, no. 3, pp. 137–142, May 2008.
[28] A. C. den Brinker, J. Breebaart, P. Ekstrand, J. Engdegård, F. Henn, K. Kjörling, W. Oomen, and H. Purnhagen, "An overview of the coding standard MPEG-4 audio amendments 1 and 2: HE-AAC, SSC, and HE-AAC v2," *EURASIP J. Audio, Speech, Music Process.*, vol. 2009, pp. 1–21, Dec. 2009.
[29] P. Ekstrand, "Bandwidth extension of audio signals by spectral band replication," in *Proc. 1st IEEE Benelux Workshop Model Process. Coding Audio (MPCA)*, Leuven, Belgium, Nov. 2002, pp. 73–79.
[30] *Coding of Audio-Visual Objects—Part 3: Audio, Subpart 4: General Audio Coding (GA)-AAC, TwinVQ, BSAC*, Standard ISO/IEC 14496-3:2008(E), 2008.
[31] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegård, "Low complexity parametric stereo coding," in *Proc. 116th AES Conv.*, Berlin, Germany, 2004.
[32] H. Purnhagen, "Low complexity parametric stereo coding in MPEG-4," in *Proc. 7th Int. Conf. Audio Effects (DAFX)*, Naples, Italy, Oct. 2004, pp. 163–168.
[33] J. A. Hartigan and M. A. Wong, "A K-means clustering algorithm," *Appl. Stat.*, vol. 28, no. 1, pp. 100–108, 1979.
[34] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.
[35] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.
[36] R. Hennequin, J. Royo-Letelier, and M. Moussallam, "Codec independent lossy audio compression detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 726–730.
[37] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," *CoRR*, vol. abs/1711.00937, pp. 1–11, Nov. 2017.
[38] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 8, pp. 620–636, Nov. 2002.
[39] *Nero AAC*. Accessed: Apr. 30, 2021. [Online]. Available: https://www.nero.com/eng/company/about-nero/commercial-license-vLang.php
[40] *QAAC*. Accessed: Apr. 30, 2021. [Online]. Available: https://sites.google.com/site/qaacpage/
[41] *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*, document ITU-T P.862.2, 2007.
[42] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "Darpa timit acoustic-phonetic continuous speech corpus CD-ROM TIMIT," NIST Interagency/Internal Report (NISTIR), Nat. Inst. Standards Technol., Gaithersburg, MD, USA, 1993.
[43] 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 *Proc. Interspeech*, Sep. 2016, pp. 352–356.
[44] J. Garofolo, D. Graff, P. Doug, and D. Pallett, *CSR-I (WSJ0) Complete LDC93S6A*. Philadelphia, PA, USA: Linguistic Data Consortium, 1993
[45] *Software Tools for Speech and Audio Coding Standardization*, document ITU-T G.191, 2010.
[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, May 2015, pp. 1–15.
[47] *Methods for Subjective Determination of Transmission Quality*, document ITU-T P.800, 1996.
[48] *Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)*, document ITU-R BS.1534-1, 2011.

SOOJOONG HWANG (Graduate Student Member, IEEE) received the B.S. degree from the School of Electronics and Computer Engineering, Chonnam National University, Gwangju, South Korea, in 2014, and the M.S. degree from the School of Electrical Engineering and Computer Science (EECS), Gwangju Institute of Science and Technology (GIST), Gwangju, in 2016, where he is currently pursuing the Ph.D. degree. His research interests include speech enhancement, voice activity detection, and coded signal enhancement.

YOUNGJU CHEON received the B.S. degree from the School of Electronics and Computer Engineering, Chonnam National University, Gwangju, South Korea, in 2019, and the M.S. degree from the School of Electrical Engineering and Computer Science (EECS), Gwangju Institute of Science and Technology (GIST), in 2021. His research interest includes codec post-processing.

SANGWOOK HAN (Graduate Student Member, IEEE) received the B.S. degree in computer engineering from the Kumoh National Institute of Technology, Kumi, South Korea, in 2018, and the M.S. degree from the School of Electrical Engineering and Computer Science (EECS), Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea, in 2020, where he is currently pursuing the Ph.D. degree. His research interests include audio codec enhancement, speaker recognition, speaker representation, and machine learning.

INSEON JANG received the B.S. degree in electrical and electronic engineering from Chungbuk National University, Cheongju, South Korea, in 2001, the M.S. degree in computer science and engineering from POSTECH, Pohang, South Korea, in 2004, and the Ph.D. degree in electronic engineering from Chungnam National University, Daejeon, South Korea, in 2018. Since 2004, she has been the Principal Member of the research staff with the Media Research Division, Electronics and Telecommunications Research Institute (ETRI), South Korea. Her research interests include audio coding and audio signal processing.

JONG WON SHIN (Member, IEEE) received the B.S. and Ph.D. degrees from the School of Electrical Engineering and Computer Science (EECS), Seoul National University, Seoul, South Korea, in 2002 and 2008, respectively. From 2008 to 2012, he was with Qualcomm Inc., San Diego, CA, USA. Since 2012, he has been with EECS, Gwangju Institute of Science and Technology, Gwangju, South Korea, where he is currently an Associate Professor. His research interests include vast areas of speech signal processing, such as speech enhancement, voice activity detection, source localization, acoustic echo cancellation, and speech emotion recognition.