applied sciences

MDPI

Article

Parallel Enhancement and Bandwidth Extension of Coded Speech

Jongwook Chae [1], Eunkyun Lee [2], Sooyoung Park [3,4] and Jong Won Shin [2,*]

[1] Department of AI Convergence, Gwangju Institute of Science and Technology, Gwangju 61005, Republic of Korea; whddnr97@gm.gist.ac.kr
[2] Department of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, Republic of Korea; lek940309@gm.gist.ac.kr
[3] Media Coding Research Section, Electronics and Telecommunications Research Institute, Daejeon 34129, Republic of Korea; sooyoung@etri.re.kr
[4] School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, Republic of Korea
* Correspondence: jwshin@gist.ac.kr

Abstract

An important use case of speech bandwidth extension (BWE) is generating high-frequency components from band-limited speech processed by a speech codec. Recent works on BWE have demonstrated remarkable capabilities in generating high-quality, high-band components using deep learning techniques. Among them, Streaming SEANet (StrmSEANet) has also been shown to be effective for BWE with reduced delay and computational complexity, making it suitable for real-time speech processing. However, the effect of the coding artifact in the lower band of the input signal has not been sufficiently considered in many deep learning-based BWE methods. In this work, we propose Parallel Enhancement and Bandwidth Extension of coded speech (PEBE), where two lightweight networks, referred to as Compact Streaming SEANet (CompSEANet), for coded speech enhancement (CSE) and BWE are configured in parallel. The CSE and BWE models are separately trained with the task-specific training settings, thereby effectively improving the reconstruction quality of the band-limited speech signals degraded by coding artifacts. Experimental results demonstrate that the proposed PEBE significantly outperforms the baseline AP-BWE, StrmSEANet, and standalone CompSEANet in reconstructing wideband (WB) and fullband speech from Opus-coded narrowband and WB signals. The proposed method achieves the highest scores in the subjective MUSHRA test while providing the fastest inference among all compared methods, with real-time factors (RTF) of 33.95× and 18.38× measured on a Samsung SM-F711 mobile device under single-thread execution.

Keywords: speech bandwidth extension; coded speech enhancement; speech coding; Streaming SEANet

Academic Editor: Douglas O'Shaughnessy

Received: 27 December 2025
Revised: 26 January 2026
Accepted: 28 January 2026
Published: 30 January 2026

1. Introduction

In voice communication systems, speech signals are degraded by bandwidth (BW) limitations introduced by speech codecs such as AMR-NB [1], AMR-WB [2], Opus [3], and EVS [4]. BW limitation is primarily due to standards for speech communication rather than the codecs themselves. Recently proposed neural network–based speech bandwidth extension (BWE) models [5–8] have shown remarkable capability in reconstructing high-frequency components of band-limited speech signals.

In ref. [5], an autoencoder convolutional neural network model [9] was adopted for BWE, trained with a time-frequency loss function to jointly optimize the model in both domains. To facilitate on-device deployment in communication systems, Streaming SEANet (StrmSEANet) [7] was proposed based on the original Sound EnhAncement Network (SEANet) [10], utilizing the causal convolutions and reducing the dimensions of the internal feature maps for real-time processing with low latency and complexity. In ref. [6], a BWE model operating on the complex spectrogram from the short-time Fourier transform (STFT) was proposed to better handle phase information. AP-BWE [8] adopted a ConvNeXt-based backbone [11] and introduced a dual-stream structure to separately process the amplitude and phase spectra of an input signal, effectively extending the frequency bandwidth and improving the speech quality. To further enhance the perceptual quality of the extended speech, a multi-resolution amplitude discriminator (MRAD) and a multi-resolution phase discriminator (MRPD) [8] were introduced, inspired by the multi-resolution spectrogram discriminator [12]. Although these approaches demonstrate strong reconstruction performance, their BWE performance on coded speech signals has not been extensively evaluated.

Several studies have aimed to address the restoration of high-frequency components while considering coding artifacts [13,14]. In ref. [13], a cascaded architecture of coded speech enhancement (CSE) and BWE modules was proposed, trained via supervision with a time-domain L2 loss and frequency-domain L2 losses on multi-resolution log spectral amplitude and log mel-spectrograms. A side-information-based approach to restoring original wideband (WB) speech from coded versions using a narrowband (NB) codec named Codec 2 [15] was proposed in [14]. Recently, numerous approaches [16–21] have been proposed as a result of the URGENT 2025 Challenge [22] to solve the universal speech enhancement problem, where input speech signals with diverse sampling rates are corrupted not only by bandwidth limitation and coding artifacts but also by additive noise, reverberation, clipping, packet loss, and wind noise.

Another approach to restoring high-band components, distinct from conventional BWE, was proposed in the name of Alias-and-Separate [23], where intentional aliasing is utilized to hide the high-frequency components of a wideband (WB) speech signal in the lower band region when the signal is downsampled to a narrowband (NB) one. After compression with an NB speech codec, the original signal can be reconstructed by separating the original lower band and aliased high band components. An improved version of the Alias-and-Separate was proposed in [24] to reduce algorithmic delay and enhance the quality of the reconstructed speech, as well as enable fullband (FB) speech coding tasks with a WB speech codec.

On the other hand, various approaches to coded speech enhancement (CSE) have been investigated [25–29] in recent years. Several studies have focused on exploiting codec-derived features such as LPC coefficients, pitch lags, and quantization gains, or side information transmitted from the encoder to improve the quality of the reconstructed signal [25–27]. In parallel, the authors of [28,29] have proposed approaches based on generative adversarial networks (GAN) [30] to effectively enhance the quality of coded speech.

In this work, we first propose a Compact Streaming SEANet (CompSEANet) to increase the computational efficiency of post-processing coded speech while minimizing performance degradation. We further propose a Parallel Enhancement and Bandwidth Extension of coded speech (PEBE) to effectively enhance coded speech signals and extend their spectral bandwidth. In the PEBE, two separate models for CSE and BWE individually produce the short-time Fourier transform (STFT) coefficients of the outputs from a band-limited input. The final output waveform is then reconstructed via an inverse STFT (iSTFT) of the combined outputs in the lower and upper bands.

Implemented with CompSEANet, the proposed PEBE demonstrates superior performance compared to the baseline StrmSEANet [7] and AP-BWE [8], achieving the highest subjective speech quality in the MUSHRA test and the fastest inference speed among all compared models in two tasks: reconstruction of wideband (WB) speech from coded narrowband (NB) speech (NB-to-WB) and reconstruction of fullband (FB) speech from coded wideband (WB) speech (WB-to-FB).

The remainder of this paper is organized as follows: Section 2 describes the baseline StrmSEANet architecture and training procedure. Section 3 presents the proposed CompSEANet and PEBE frameworks. Section 4 reports the experimental results. Finally, Section 5 concludes the paper and discusses future work.

2. Streaming SEANet

StrmSEANet [7] was proposed for real-time speech BWE, featuring a lighter network structure and lower algorithmic delay than the original version [10], while maintaining competitive reconstruction performance in terms of SI-SDR. As depicted in Figure 1, the coded input signal is upsampled with an upsampling ratio R (defined as the ratio between the target sampling rate RF_s and the input sampling rate F_s) and then processed by an initial 1D convolutional layer with the kernel size of seven, four encoder blocks, four decoder blocks [10] with residual skip connections between encoder and decoder feature maps of identical time scale, and a final 1D convolutional layer. Although the original StrmSEANet [7] does not explicitly describe the upsampling procedure of the input signals, the input coded signal should be upsampled to the same sampling rate as the target speech waveform. Therefore, we explicitly include the upsampling process in our implementation of the StrmSEANet.

The encoder blocks consist of a strided convolution and three residual units [10], where dilated convolutions with dilation factors of 1, 3, and 9 are sequentially applied to enlarge the receptive field without increasing model size. Each encoder block reduces temporal resolution according to the stride factor while doubling the channel dimension. Conversely, each decoder block mirrors this process using a transposed convolution for upsampling followed by the same residual units, progressively restoring temporal resolution and reducing channels. All convolutions in the StrmSEANet [7] are causal, and no normalization across the time axis is employed in the generator. In refs. [7,10], the strides (S_1, S_2, S_3, S_4) were set to (2, 2, 8, 8). The StrmSEANet is trained following the GAN-based training procedure [30] with multi-scale discriminators (MSD) [7], originally introduced in [31].

The loss function for adversarial training of the StrmSEANet is composed of a weighted sum of the hinge loss $\mathcal{L}_{adv}^{hg}$ [10] and feature-matching loss $\mathcal{L}_{fm}$ [10,31] as follows:

$$\mathcal{L}_G = \frac{1}{M}\sum_{m=1}^{M}[\mathcal{L}_{adv}^{hg}(\hat{x};D^m) + \lambda\mathcal{L}_{fm}(\hat{x},x;D^m)], \tag{1}$$

$$\mathcal{L}_{adv}^{hg}(\hat{x};D^m) = \frac{1}{T^m}\sum_{\tau=1}^{T^m}\max(0,1-D_\tau^m(\hat{x})), \tag{2}$$

$$\mathcal{L}_{fm}(\hat{x},x;D^m) = \frac{1}{L^m}\sum_{l}^{L^m}\frac{1}{T_l^m}\sum_{\tau=1}^{T_l^m}|D_{l,\tau}^m(\hat{x}) - D_{l,\tau}^m(x)|, \tag{3}$$

where x and $\hat{x}$ are the real and generated waveforms, M is the number of discriminators, D^m is the m-th discriminator, $D_\tau^m(\cdot)$ is the τ-th element in the output of D^m among T^m elements in total, L^m is the number of layers in D^m, and $D_{l,\tau}^m(\cdot)$ is the τ-th element in the

l-th internal feature map of D^m among T_l^m elements in total. In refs. [7,10], λ was set to 100. Each discriminator D^m is trained with a corresponding loss function to $\mathcal{L}_{adv}^{hg}$ defined as

$$\mathcal{L}_{D^m}^{hg}(\hat{x}, x) = \frac{1}{T^m} \sum_{\tau=1}^{T^m} [\max(0, 1 + D_\tau^m(\hat{x})) + \max(0, 1 - D_\tau^m(x))]. \quad (4)$$

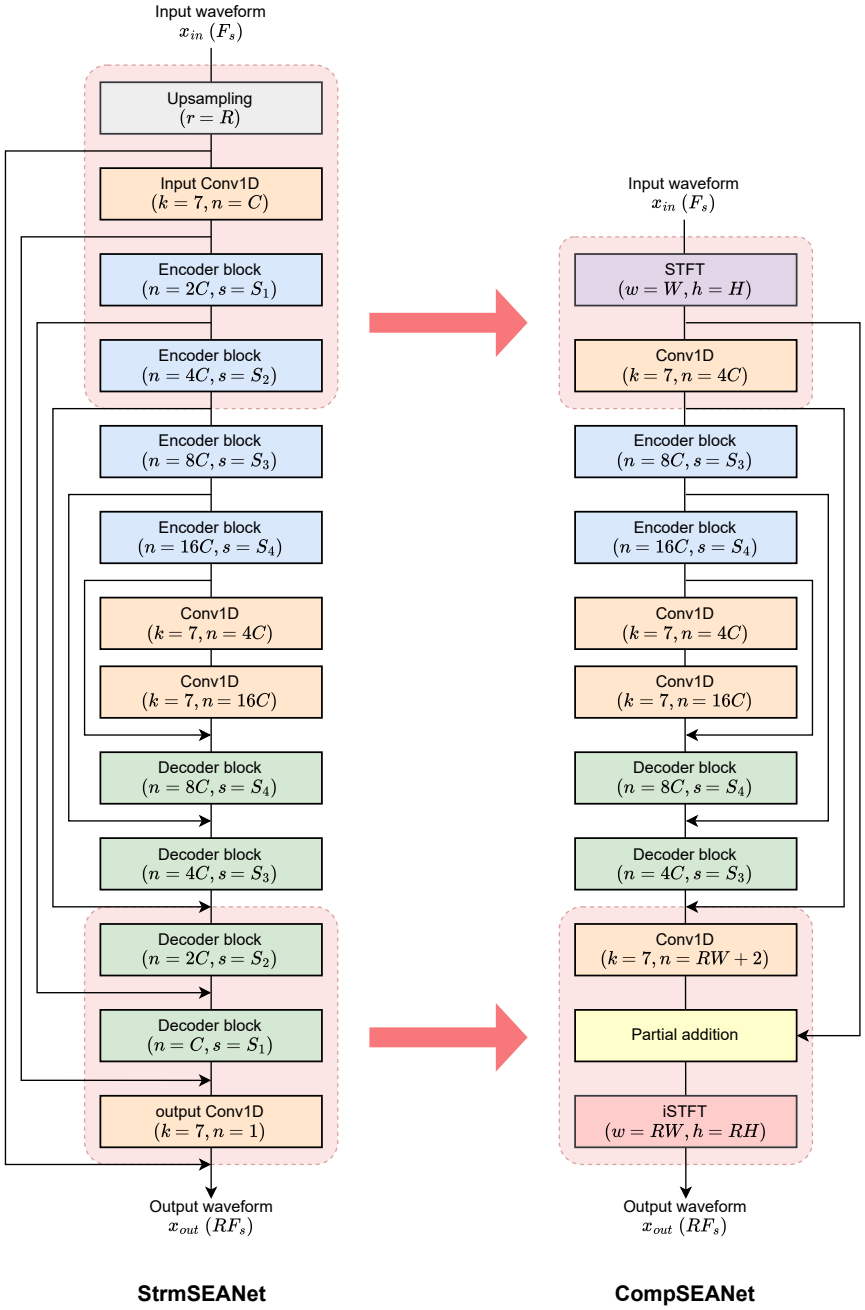

Figure 1. Architectures of StrmSEANet (**left**) and CompSEANet (**right**).

3. Proposed Methods

3.1. Compact Streaming SEANet

Firstly, we design a lightweight post-processing model based on the StrmSEANet to improve computational efficiency while minimizing performance degradation. Inspired by iSTFTNet [32], we propose CompSEANet, a compact version of the StrmSEANet where the first two encoder blocks and the last two decoder blocks are replaced with STFT and iSTFT operations, respectively. A detailed illustration of the CompSEANet architecture

is presented in Figure 1. We set the DFT sizes for the STFT and iSTFT to match their respective window sizes. The real and imaginary components in each frame are treated as separate channels and concatenated along the frequency axis to be processed by the first convolutional layer. The last convolutional layer outputs the real and imaginary components for the output waveform, concatenated along the frequency axis. Similarly to the StrmSEANet, residual skip connections are employed between feature maps to facilitate effective information flow across blocks. The STFT coefficients of the input are added to the corresponding lower-band part of the final output. Finally, the output waveform is reconstructed by the iSTFT with window and hop sizes R times larger than those of the STFT, where R corresponds to the upsampling ratio.

We used a loss function for adversarial training composed of the sum of the adversarial loss $\mathcal{L}_{adv}$, feature-matching loss $\mathcal{L}_{fm}$, and waveform reconstruction loss $\mathcal{L}_{rec}$ as follows:

$$\mathcal{L}_G = \frac{1}{M}\sum_{m=1}^{M}[\alpha^m \mathcal{L}_{adv}(\hat{x}, x; D^m) + \lambda^m \mathcal{L}_{fm}(\hat{x}, x; D^m)] + \mathcal{L}_{rec}(\hat{x}, x), \tag{5}$$

where x and $\hat{x}$ are the real and generated waveforms. In addition to the hinge loss [10], we also consider the least-squares GAN (LSGAN) loss function [33] for the adversarial loss $\mathcal{L}_{adv}$, as used in [34], to find a more suitable loss function for coded speech post-processing,

$$\mathcal{L}^{ls}_{adv} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{T^m}\sum_{\tau=1}^{T^m}(D^m_\tau(\hat{x}) - 1)^2. \tag{6}$$

The discriminators are trained with

$$\mathcal{L}^{ls}_{D^m}(\hat{x}, x) = \frac{1}{T^m}\sum_{\tau=1}^{T^m}[(D^m_\tau(\hat{x}))^2 + (D^m_\tau(x) - 1)^2]. \tag{7}$$

As utilized and shown to be effective for BWE in [6], we employ the multi-resolution STFT loss function [35], defined as

$$\mathcal{L}_{rec} = \frac{1}{P}\sum_{p=1}^{P}\mathcal{L}_{sc}(\hat{x}, x; \theta_p) + \mathcal{L}_{mag}(\hat{x}, x; \theta_p), \tag{8}$$

$$\mathcal{L}_{sc}(\hat{x}, x; \theta_p) = \frac{|||\hat{X}_p| - |X_p|||_F}{|||X_p|||_F}, \tag{9}$$

$$\mathcal{L}_{mag}(\hat{x}, x; \theta_p) = \frac{1}{T_p F_p}||\log(|\hat{X}_p|) - \log(|X_p|)||_{1,1}, \tag{10}$$

where P is the number of STFT configurations $\{\theta_p\}_{p=1}^{P}$, $\hat{X}_p$ and X_p are the STFT coefficients of $\hat{x}$ and x transformed with θ_p, and T_p and F_p are the number of time steps and frequency bins, respectively. The operator $|\cdot|$ denotes the magnitude of a complex-valued STFT coefficient, and $\|\cdot\|_F$ and $\|\cdot\|_{1,1}$ are the Frobenius norm and matrix norm, respectively. $\log(\cdot)$ denotes the natural logarithm applied to each element.

We present the results of our preliminary experiment for the StrmSEANet and CompSEANet in Table 1. Both models are trained with the loss function in Equation (5), as we confirmed that utilizing the waveform reconstruction loss improved speech quality. The strides for the StrmSEANet were set to (2, 4, 5, 8) for NB-to-WB and (4, 5, 6, 8) for WB-to-FB tasks. For CompSEANet, strides were set to (5, 8) for both tasks. We set $\alpha^m = 1$ and $\lambda^m = 100$ as defined in Equation (5) for this experiment.

All results indicate that using the LSGAN loss was beneficial for the reconstruction quality. Even though CompSEANet for the NB-to-WB task showed slightly degraded performance compared to StrmSEANet in terms of Perceptual Evaluation of Speech Quality (PESQ) [36], it dramatically reduced inference time by 70%. For the WB-to-FB task, CompSEANet achieved the best performance in terms of Virtual Speech Quality Objective Listener in audio mode (ViSQOL-A) [37,38], even with 80% reduced real-time factors (RTFs) compared to the StrmSEANet.

Table 1. Preliminary experimental results for StrmSEANet and CompSEANet trained with MSD, $\alpha^m = 1$, and $\lambda^m = 100$. PESQ and ViSQOL-A scores are measured for NB-to-WB and WB-to-FB tasks, respectively. The best performance for each task is highlighted in bold.

Task	Model	RTF	Adversarial Loss	Score
NB-to-WB	StrmSEANet ($C = 8$)	0.0384 (26.02×)	Hinge	2.71
	StrmSEANet ($C = 8$)	0.0384 (26.02×)	LSGAN	**2.88**
	CompSEANet ($C = 8$)	**0.0114 (87.77×)**	Hinge	2.51
	CompSEANet ($C = 8$)	**0.0114 (87.77×)**	LSGAN	2.72
WB-to-FB	StrmSEANet ($C = 12$)	0.1163 (8.60×)	Hinge	3.11
	StrmSEANet ($C = 12$)	0.1163 (8.60×)	LSGAN	3.12
	CompSEANet ($C = 12$)	**0.0242 (41.41×)**	Hinge	2.92
	CompSEANet ($C = 12$)	**0.0242 (41.41×)**	LSGAN	**3.16**

We also present experimental results in Table 2, where discriminators and weights (α^m, λ^m) were varied, along with C for CompSEANet in Figure 1. The use of MPD, MRAD, and MRPD [8] alongside MSD improved the PESQ score for both models compared to using MSD alone. Although the PESQ for CompSEANet with $C = 8$ is not yet close to the StrmSEANet, the gap is reduced. When the RTF of the StrmSEANet approaches that of the CompSEANet with $C = 16$, the PESQ for the CompSEANet exceeds that of the StrmSEANet.

Table 2. Second preliminary experimental results comparing PESQ scores for StrmSEANet and CompSEANet in the NB-to-WB task by varying discriminator combinations and loss weights. All models were trained with LSGAN loss. The best performance for each task is highlighted in bold.

Model	RTF	Discriminators	(α^m, λ^m)	PESQ
StrmSEANet ($C = 8$)	0.0384 (26.02×)	MSD	(1, 100)	2.88
		MSD, MPD, MRAD, MRPD	(1, 100) for MSD and MPD (0.1, 0.1) for MRAD and MRPD	2.83
			(1, 1) for MSD and MPD (0.1, 0.1) for MRAD and MRPD	**2.93**
CompSEANet ($C = 8$)	0.0114 (87.77×)	MSD	(1, 100)	2.72
		MSD, MPD, MRAD, MRPD	(1, 100) for MSD and MPD (0.1, 0.1) for MRAD and MRPD	2.73
			(1, 1) for MSD and MPD (0.1, 0.1) for MRAD and MRPD	**2.87**
CompSEANet ($C = 16$)	0.0393 (25.46×)	MSD, MPD, MRAD, MRPD	(1, 1) for MSD and MPD (0.1, 0.1) for MRAD and MRPD	**3.04**

3.2. Parallel Enhancement and Bandwidth Extension of Coded Speech

Based on the preliminary experiments presented in Section 3.1, we also propose PEBE to effectively enhance an input speech signal band-limited and coded during the coding procedure and extend the spectral bandwidth of the input signal by adopting separate parallelized models for CSE and BWE, each of which is optimized to the corresponding

task. Compared to a cascaded structure of the CSE and BWE models, the parallel structure is beneficial for the total latency and design flexibility of each model. The total latency becomes the sum of the latencies of the CSE and BWE when they are cascaded, while the latency becomes the larger of the two models when parallelized. The modularity between the CSE and BWE models facilitates finding the optimal training settings or architectures for the models and improves the computational scalability, as the complexity of the CSE and BWE models can be adjusted separately according to the computational resources.

Figure 2 describes the processing procedure of the PEBE implemented with CompSEANet for CSE and BWE. The input speech waveform $x_{in} \in \mathbb{R}^{N_{in}}$ is transformed into STFT coefficients $X_{in} \in \mathbb{C}^{F_{in} \times T}$, where N_{in} is the number of samples and F_{in} and T are the numbers of frequency bins and frames, respectively. The CSE and BWE models process X_{in} and output $X_{CSE} \in \mathbb{C}^{F_{CSE} \times T}$ and $X_{BWE} \in \mathbb{C}^{F_{BWE} \times T}$, respectively. X_{CSE} is added to the corresponding part of X_{in}, and the output spectra $X_{out} \in \mathbb{C}^{F_{out} \times T}$, where $F_{out} = F_{CSE} + F_{BWE}$, are composed as follows:

$$X_{out}(f,t) = \begin{cases} X_{CSE}(f,t) + X_{in}(f,t) & 0 \le f < F_{CSE} \\ X_{BWE}(f - F_{CSE}, t) & F_{CSE} \le f < F_{out} \end{cases}. \tag{11}$$

Finally, the iSTFT is applied to X_{out} to reconstruct the final output waveform x_{out}.

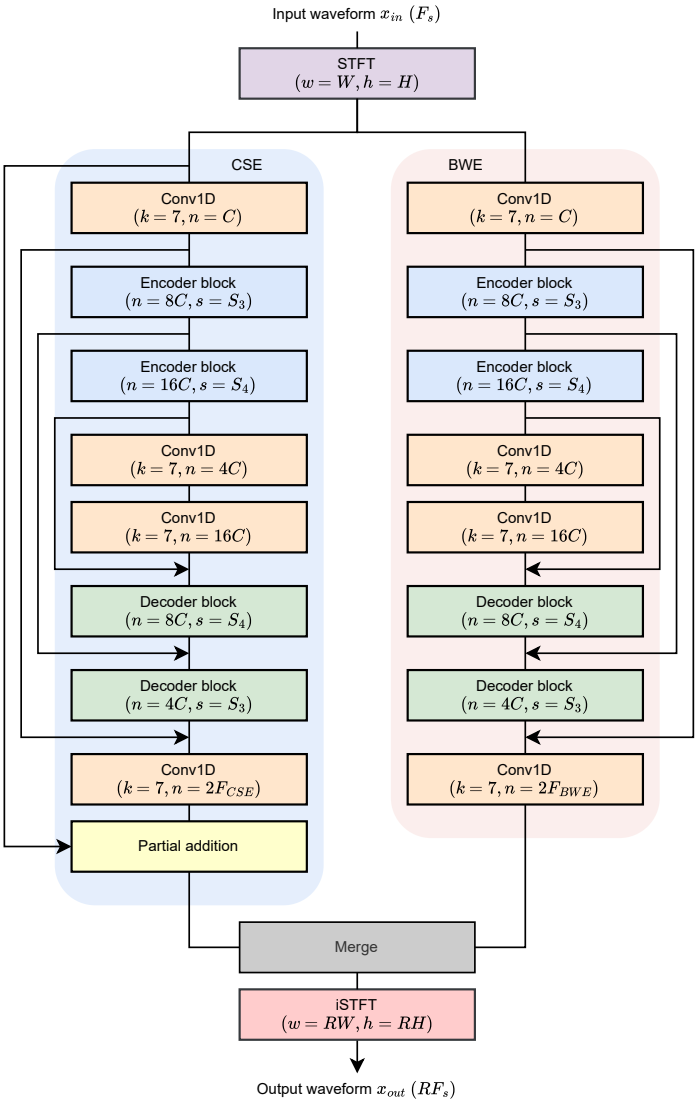

Figure 2. Architecture of PEBE adopting CompSEANet for CSE and BWE models.

Since PEBE utilizes CSE and BWE models in parallel, they can be trained separately with schemes suitable for each task. In this work, we trained the CSE and BWE models with individual loss functions and discriminator sets by applying iSTFT to each of the outputs of the CSE and BWE models, as given in Figure 3. To fill empty bands, the lower band for X_{BWE} and the high band for X_{CSE}, corresponding components of the clean signal X, are utilized. The waveforms to train the CSE and BWE models, $\tilde{x}_{CSE}$ and $\tilde{x}_{BWE}$, are obtained by applying iSTFT to their corresponding STFT coefficients $\tilde{X}_{CSE}$ and $\tilde{X}_{BWE}$, defined as:

$$\tilde{X}_{CSE}(f,t) = \begin{cases} X_{CSE}(f,t) & 0 \le f < F_{CSE} \\ X(f - F_{CSE}, t) & F_{CSE} \le f < F_{out} \end{cases}, \tag{12}$$

$$\tilde{X}_{BWE}(f,t) = \begin{cases} X(f,t) & 0 \le f < F_{CSE} \\ X_{BWE}(f - F_{CSE}, t) & F_{CSE} \le f < F_{out} \end{cases}. \tag{13}$$

With respective discriminator sets for CSE and BWE, $\{D^m_{CSE}\}_{m=1}^{M_{CSE}}$ and $\{D^m_{BWE}\}_{m=1}^{M_{BWE}}$, we define the respective loss functions as

$$\mathcal{L}_{CSE} = \frac{1}{M_{CSE}} \sum_{m=1}^{M_{CSE}} [\alpha^m_{CSE} \mathcal{L}_{adv}(x, \tilde{x}_{CSE}; D^m_{CSE}) + \lambda^m_{CSE} \mathcal{L}_{fm}(x, \tilde{x}_{CSE}; D^m_{CSE})] + \eta_{CSE} \mathcal{L}_{rec}(x, \tilde{x}_{CSE}), \tag{14}$$

$$\mathcal{L}_{BWE} = \frac{1}{M_{BWE}} \sum_{m=1}^{M_{BWE}} [\alpha^m_{BWE} \mathcal{L}_{adv}(x, \tilde{x}_{BWE}; D^m_{BWE}) + \lambda^m_{BWE} \mathcal{L}_{fm}(x, \tilde{x}_{BWE}; D^m_{BWE})] + \eta_{BWE} \mathcal{L}_{rec}(x, \tilde{x}_{BWE}). \tag{15}$$

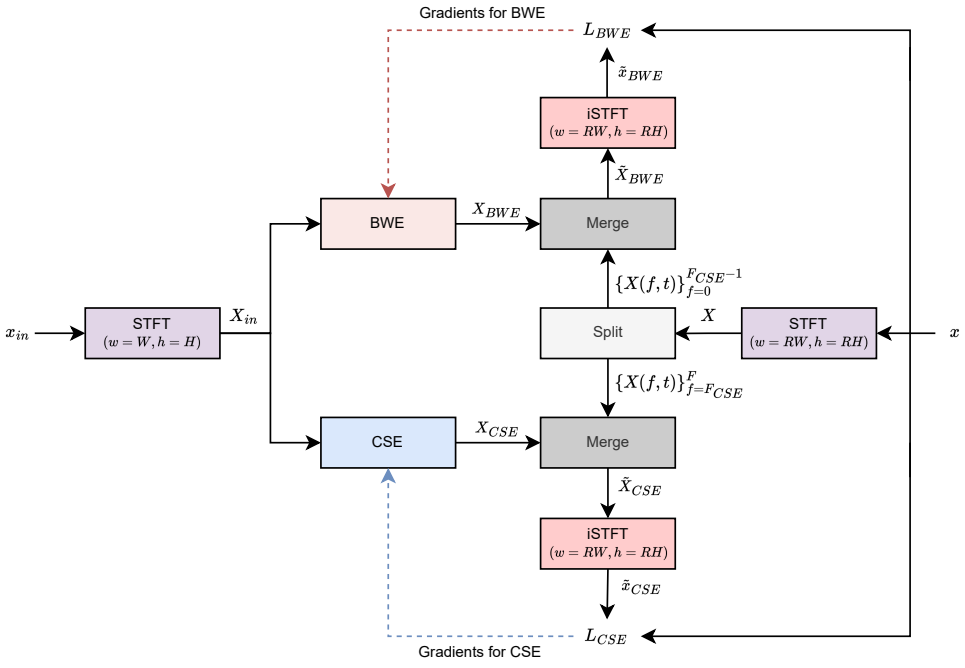

Figure 3. Block diagram describing the computation of loss functions for CSE and BWE models.

4. Experiments

4.1. Experimental Settings

We used the clean speech signals from the Valentini speech dataset [39] which is a subset of the VCTK corpus [40]. The dataset includes FB speech waveforms sampled at 48 kHz from 84 English speakers for training and from 2 speakers for evaluation. For the

NB-to-WB task, the FB speech waveforms were resampled to 8 kHz and coded using the Opus NB codec operating at 8 kbps in constant bitrate (CBR) mode. The corresponding target WB signals with a sampling rate of 16 kHz were obtained by resampling the original FB speech waveforms. For the WB-to-FB task, the resampled WB signals were coded using the Opus WB at 10 kbps in CBR mode. The frame sizes for both codecs were set to 20 ms.

For the StrmSEANet, strides (S_1, S_2, S_3, S_4) were set to (2, 4, 5, 8) for the NB-to-WB task and (4, 5, 6, 8) for the WB-to-FB task. Similarly, strides (S_3, S_4) for the CompSEANet were set to (5, 8). STFT window and hop sizes for CompSEANet were 8 and 4 for NB-to-WB, and 16 and 8 for WB-to-FB. Accordingly, iSTFT window and hop sizes were 16 and 8 for the NB-to-WB task, and 48 and 24 for the WB-to-FB task. With these configurations, both StrmSEANet and CompSEANet introduce the algorithmic delay of 20 ms for both tasks.

We set the dimension of the first feature map (C) in the StrmSEANet to 8 and 12 for NB-to-WB and WB-to-FB tasks, respectively. For the CompSEANet, C was set to 16 and 24. For the CompSEANets in the PEBE framework, C was set to 8 and 12. F_{CSE} and F_{BWE} were 4 and 5 for the NB-to-WB task, and 6 and 19 for the WB-to-FB task.

We used the Adam optimizer [41] with $\beta_1 = 0.5, \beta_2 = 0.9$, and a learning rate of 0.0001. For the PEBE, BWE models were trained with the MSD [7], MPD [34], and MRAD and MRPD in [8]. Only the first discriminator without downsampling in the MSD was adopted for BWE. CSE models were trained with MSD and MPD. For the StrmSEANet and CompSEANet, we set $\alpha^m = 1$ and $\lambda^m = 1$ for the MSD and MPD, and $\alpha^m = 0.1$ and $\lambda^m = 0.1$ for the MRAD and MRPD. For the PEBE, we set $\alpha^m_{CSE} = 1$, $\lambda^m_{CSE} = 2$, $\eta_{CSE} = 45$, and $\eta_{BWE} = 1$, while α^m_{BWE} and λ^m_{BWE} were set to 1 for the MSD and MPD and 0.1 for the MRAD and MRPD. We used the LSGAN loss [33] based on preliminary results.

In this work, we have adapted the AP-BWE source code from its open-source repository https://github.com/yxlu-0102/AP-BWE (accessed on 25 January 2026), modifying the internal convolutional layers in the ConvNeXt blocks [11] to be causal and replacing layer normalization (LN) [42] with cumulative layer normalization [43] for real-time operation. We set the ConvNeXt feature map dimension to 256. The window and hop sizes were 320 and 80 for NB-to-WB and 960 and 240 for WB-to-FB, respectively, with a DFT size of 1024 for both, which leads to 15 ms algorithmic delay for both tasks.

4.2. Experimental Results

We present spectrograms of clean and reconstructed WB speech signals in Figure 4. AP-BWE and PEBE reconstructed a clearer harmonic structure in the voiced region (green box) compared to the StrmSEANet and CompSEANet. PEBE also suppressed more artifacts in the high-band region (blue box). These tendencies are more prominent in Figure 5 for the WB-to-FB task. While the AP-BWE, CompSEANet, and PEBE produced fewer artifacts in the lower band than the StrmSEANet, only PEBE generated high-band components without severe over-smoothing.

Table 3 presents objective evaluation results. PESQ and ViSQOL-S were used for NB-to-WB, and ViSQOL-A and NISQA [44] for WB-to-FB. RTFs were measured on a Samsung SM-F711 mobile device (Samsung, Suwon, Republic of Korea) with single-thread operations. For the NB-to-WB task on the Valentini test set, the PEBE achieved the highest PESQ and ViSQOL-S scores, outperforming the AP-BWE, StrmSEANet, and CompSEANet. Even when the model size of the CompSEANet was adjusted to match the RTF of the PEBE, the PEBE consistently exhibited higher scores, yielding the lowest RTF, which demonstrates superior reconstruction quality with high computational efficiency. On the PTDB-TUG dataset, PEBE achieved the highest ViSQOL-S score, while its PESQ score, although slightly lower than that of AP-BWE, remained comparable to the other models, indicating stable generalization performance. In the WB-to-FB task on the Valentini test set, the PEBE

achieved the highest NISQA score of 4.38. Although it obtained a slightly lower ViSQOL-A score than other models, the difference may be marginal. Notably, the PEBE also recorded the lowest RTF, offering the fastest inference speed with superior or comparable objective quality scores. On the PTDB-TUG test set, PEBE achieved the best performance in terms of NISQA, while exhibiting comparable ViSQOL-A results to other models, with only marginal performance differences observed across models.

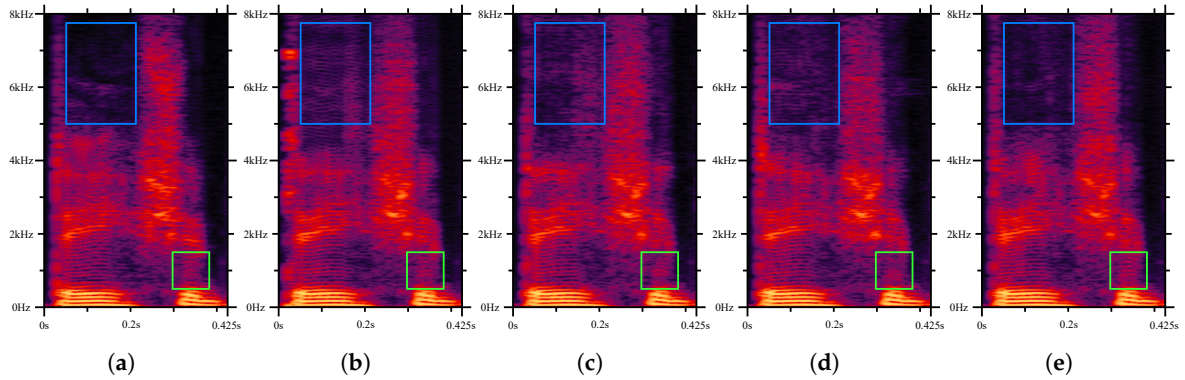

Figure 4. Spectrograms for (**a**) a clean WB speech and the reconstructed WB speech signals by (**b**) AP-BWE, (**c**) StrmSEANet, (**d**) CompSEANet, and (**e**) PEBE.

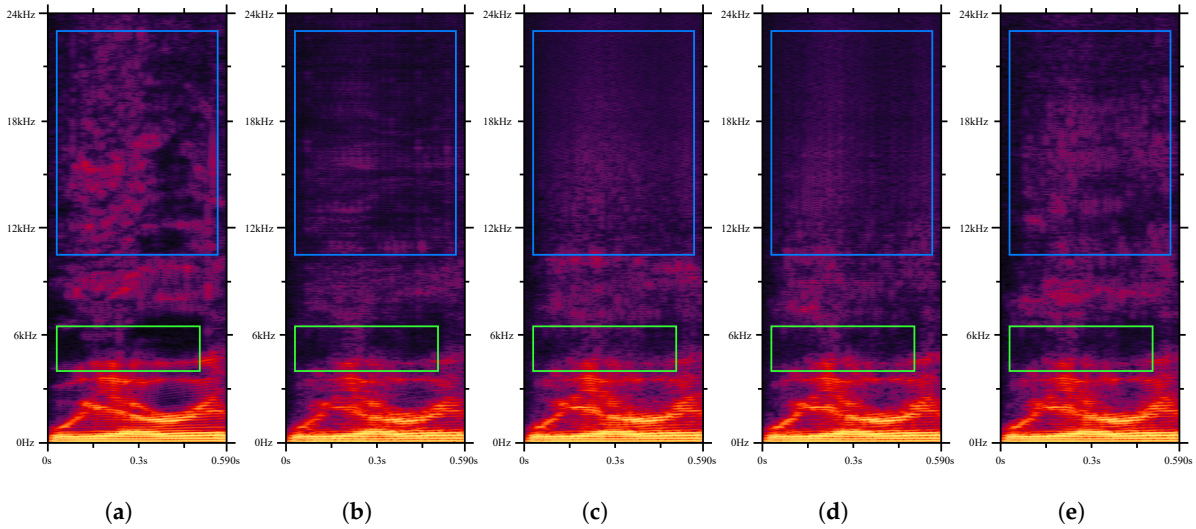

Figure 5. Spectrograms for (**a**) a clean FB speech and the reconstructed FB speech signals by (**b**) AP-BWE, (**c**) StrmSEANet, (**d**) CompSEANet, and (**e**) PEBE.

We present MUSHRA [45] test results for both tasks in Figure 6. 11 experienced listeners evaluated the models. In both sessions, the PEBE exhibited significant improvement in speech quality. CompSEANet, StrmSEANet, and AP-BWE showed no significant difference in the NB-to-WB task, as shown in Figure 6a. However, for the WB-to-FB task, a significant difference was observed between AP-BWE and the other models as presented in Figure 6b, while the CompSEANet and StrmSEANet performed comparably. Despite superior objective scores of the CompSEANet in Table 3, it showed slightly lower average MUSHRA scores than StrmSEANet, possibly due to weak artifacts near 6 kHz.

Table 3. PESQ, ViSQOL, and NISQA scores for the baseline and proposed models on the Valentini test set and whole excerpts from the PTDB-TUG dataset, along with their real-time factors (RTFs). The best performance for each task is highlighted in bold.

Task	Model	RTF	Valentini		PTDB-TUG	
			PESQ	ViSQOL-S	PESQ	ViSQOL-S
NB-to-WB	AP-BWE	0.1045 (9.57×)	2.94	4.26	**2.42**	3.67
	StrmSEANet ($C = 8$)	0.0384 (26.02×)	2.93	4.24	2.24	3.73
	CompSEANet ($C = 16$)	0.0393 (25.46×)	3.04	4.26	2.38	3.68
	PEBE ($C = 8$)	**0.0295 (33.95×)**	**3.07**	**4.27**	2.38	**3.78**
Task	**Model**	**RTF**	**Valentini**		**PTDB-TUG**	
			NISQA	**ViSQOL-A**	**NISQA**	**ViSQOL-A**
WB-to-FB	AP-BWE	0.1160 (8.62×)	4.07	3.22	3.95	**3.80**
	StrmSEANet ($C = 12$)	0.1163 (8.60×)	4.18	3.22	4.05	3.76
	CompSEANet ($C = 24$)	0.0775 (12.90×)	4.35	**3.24**	4.08	3.78
	PEBE ($C = 12$)	**0.0544 (18.38×)**	**4.38**	3.20	**4.16**	3.77

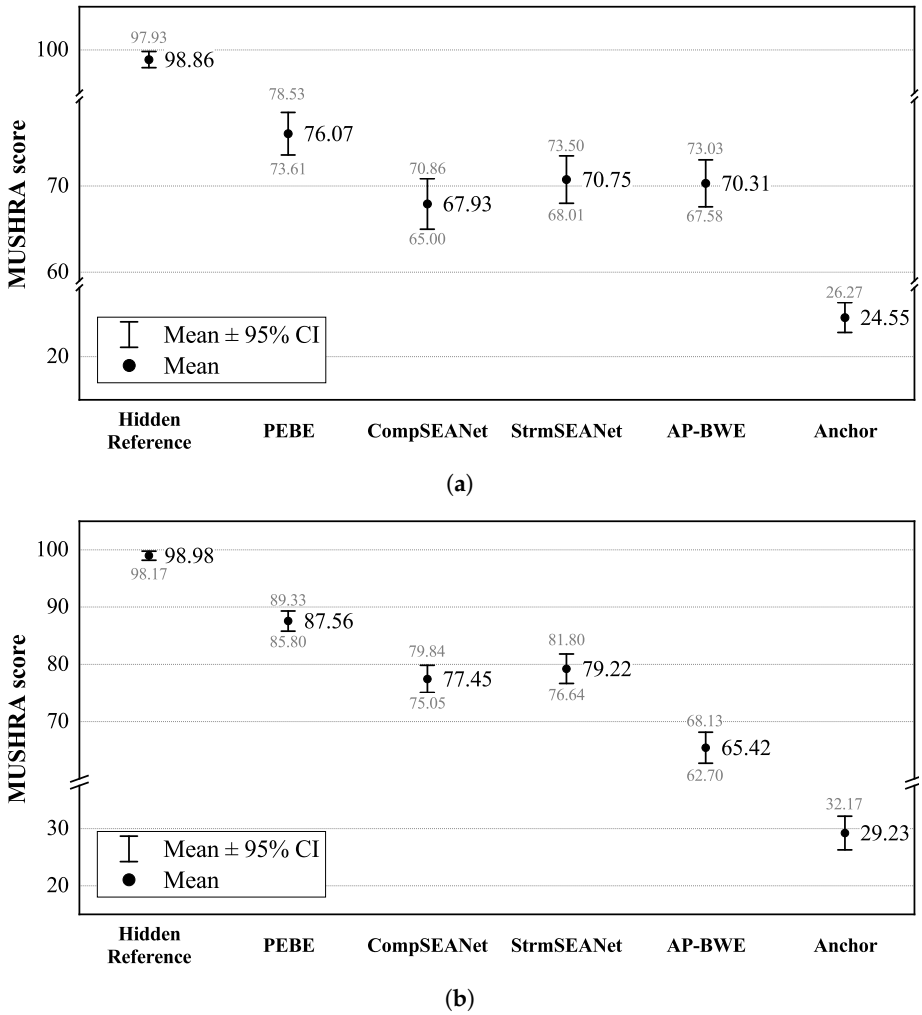

Figure 6. MUSHRA test results for (**a**) NB-to-WB and (**b**) WB-to-FB tasks.

5. Conclusions

In this work, we proposed a Parallel Enhancement and Bandwidth Extension (PEBE) framework to address the post-processing of coded speech. We also introduced Compact Streaming SEANet (CompSEANet) to improve computational efficiency, utilized within

PEBE for both coded speech enhancement and bandwidth extension. The proposed PEBE, implemented with the CompSEANet, outperformed the baseline StrmSEANet and AP-BWE models in two speech reconstruction tasks, NB-to-WB and WB-to-FB, while achieving faster inference speeds. While the current PEBE adopts independently trained CSE and BWE models, the two models may learn complementary features that are not explicitly shared in the current architecture. Future work will investigate feature-level interactions between the CSE and BWE models within the PEBE framework, while maintaining the advantages of the parallel architecture.

Author Contributions: Conceptualization, J.C. and E.L.; methodology, J.C. and E.L.; software, J.C. and E.L.; validation, J.C., E.L., S.P. and J.W.S.; formal analysis, J.C., E.L., S.P. and J.W.S.; investigation, J.C. and E.L.; resources, J.W.S.; data curation, J.C. and E.L.; writing—original draft preparation, J.C. and E.L.; writing—review and editing, S.P. and J.W.S.; visualization, J.C. and E.L.; supervision, J.W.S.; project administration, J.W.S.; funding acquisition, S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [25ZC1100, The research of the basic media contents technologies].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The speech dataset used in this study is available in https://datashare.ed.ac.uk/handle/10283/2791 (accessed on 25 January 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. *3GPP TS 26.090*; Mandatory Speech Codec Speech Processing Functions; Adaptive Multi-Rate (AMR) Speech Codec; Transcoding Functions. 3GPP: Valbonne, France, 1999.
2. *ITU-T Recommendation G.722.2*; Wideband Coding of Speech at Around 16 kbit/s Using Adaptive Multi-Rate Wideband (AMR-WB). ITU: Geneva, Switzerland, 2002.
3. Valin, J.M.; Vos, K.; Terriberry, T. *Definition of the Opus Audio Codec*; IETF RFC 6716; Internet Engineering Task Force (IETF): Fremont, CA, USA, 2012.
4. *3GPP TS 26.443 V17.0.0*; Codec for Enhanced Voice Services (EVS); ANSI C Code (Floating-Point). 3GPP: Valbonne, France, 2022.
5. Wang, H.; Wang, D. Time-Frequency Loss for CNN Based Speech Super-Resolution. In Proceedings of the IEEE ICASSP, Barcelona, Spain, 4–8 May 2020; pp. 861–865.
6. Mandel, M.; Tal, O.; Adi, Y. AERO: Audio Super Resolution in the Spectral Domain. In Proceedings of the IEEE ICASSP, Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
7. Li, Y.; Tagliasacchi, M.; Rybakov, O.; Ungureanu, V.; Roblek, D. Real-time speech frequency bandwidth extension. In Proceedings of the IEEE ICASSP, Toronto, ON, Canada, 6–11 June 2021; pp. 691–695.
8. Lu, Y.X.; Ai, Y.; Du, H.P.; Ling, Z.H. Towards High-Quality and Efficient Speech Bandwidth Extension with Parallel Amplitude and Phase Prediction. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2025**, *33*, 236–250. [CrossRef]
9. Pandey, A.; Wang, D. A New Framework for CNN-Based Speech Enhancement in the Time Domain. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1179–1188. [CrossRef] [PubMed]
10. Tagliasacchi, M.; Li, Y.; Misiunas, K.; Roblek, D. SEANet: A Multi-Modal Speech Enhancement Network. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 1126–1130.
11. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF CVPR, New Orleans, LA, USA, 18–24 June 2022; pp. 11966–11976.
12. Jang, W.; Lim, D.; Yoon, J.; Kim, B.; Kim, J. UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation. In Proceedings of the Interspeech 2021, Brno, Czech Republic, 30 August–3 September 2021; pp. 2207–2211.
13. Wen, L.; Wang, L.; Zheng, Y.; Shi, W.; Choi, K.P. FT-CSR: Cascaded Frequency-Time Method for Coded Speech Restoration. In Proceedings of the IEEE ICME, Niagara Falls, ON, Canada, 15–19 July 2024; pp. 1–6.

14. Lin, J.; Kalgaonkar, K.; He, Q.; Lei, X. Speech Enhancement for Low Bit Rate Speech Codec. In Proceedings of the IEEE ICASSP, Singapore, 22–27 May 2022; pp. 7777–7781.
15. Rowe, D. Codec 2-open source speech coding at 2400 bits/s and below. In Proceedings of the ARRL and TAPR Digital Communications Conference, Baltimore, MD, USA, 16–18 September 2011; pp. 80–84.
16. Rong, X.; Wang, D.; Hu, Q.; Wang, Y.; Hu, Y.; Lu, J. TS-URGENet: A Three-stage Universal Robust and Generalizable Speech Enhancement Network. In Proceedings of the Interspeech 2025, Rotterdam, The Netherlands, 17–21 August 2025; pp. 863–867.
17. Le, X.; Chen, Z.; Sun, S.; Xia, X.; Huang, C. Multistage Universal Speech Enhancement System for URGENT Challenge. In Proceedings of the Interspeech 2025, Rotterdam, The Netherlands, 17–21 August 2025; pp. 868–872.
18. Sun, Z.; Li, A.; Lei, T.; Chen, R.; Yu, M.; Zheng, C.; Zhou, Y.; Yu, D. Scaling beyond Denoising: Submitted System and Findings in URGENT Challenge 2025. In Proceedings of the Interspeech 2025, Rotterdam, The Netherlands, 17–21 August 2025; pp. 873–877.
19. Serbest, S.; Stojkovic, T.; Cernak, M.; Harper, A. DeepFilterGAN: A Full-band Real-time Speech Enhancement System with GAN-based Stochastic Regeneration. In Proceedings of the Interspeech 2025, Rotterdam, The Netherlands, 17–21 August 2025; pp. 878–882.
20. Goswami, N.; Harada, T. FUSE: Universal Speech Enhancement using Multi-Stage Fusion of Sparse Compression and Token Generation Models for the URGENT 2025 Challenge. In Proceedings of the Interspeech 2025, Rotterdam, The Netherlands, 17–21 August 2025; pp. 883–887.
21. Goswami, N.; Harada, T. Universal Speech Enhancement with Regression and Generative Mamba. In Proceedings of the Interspeech 2025, Rotterdam, The Netherlands, 17–21 August 2025; pp. 888–892.
22. Saijo, K.; Zhang, W.; Cornell, S.; Scheibler, R.; Li, C.; Ni, Z.; Kumar, A.; Sach, M.; Fu, Y.; Wang, W.; et al. Interspeech 2025 URGENT Speech Enhancement Challenge. In Proceedings of the Interspeech 2025, Rotterdam, The Netherlands, 17–21 August 2025; pp. 858–862.
23. Hwang, S.; Lee, E.; Jang, I.; Shin, J.W. Alias-and-Separate: Wideband Speech Coding Using Sub-Nyquist Sampling and Speech Separation. *IEEE Signal Process. Lett.* **2022**, *29*, 2003–2007. [CrossRef]
24. Lee, E.; Beack, S.; Shin, J.W. Improved Alias-and-Separate Speech Coding Framework with Minimal Algorithmic Delay. *IEEE J. Sel. Top. Signal Process.* **2024**, *18*, 1414–1426. [CrossRef]
25. Büthe, J.; Valin, J.M.; Mustafa, A. LACE: A Light-Weight, Causal Model for Enhancing Coded Speech Through Adaptive Convolutions. In Proceedings of the IEEE WASPAA, New Paltz, NY, USA, 22–25 October 2023; pp. 1–5.
26. Büthe, J.; Mustafa, A.; Valin, J.M.; Helwani, K.; Goodwin, M.M. NOLACE: Improving Low-Complexity Speech Codec Enhancement Through Adaptive Temporal Shaping. In Proceedings of the IEEE ICASSP, Seoul, Republic of Korea, 14–19 April 2024; pp. 476–480.
27. Hwang, S.; Cheon, Y.; Han, S.; Jang, I.; Shin, J.W. Enhancement of Coded Speech Using Neural Network-Based Side Information. *IEEE Access* **2021**, *9*, 121532–121540. [CrossRef]
28. Biswas, A.; Jia, D. Audio Codec Enhancement with Generative Adversarial Networks. In Proceedings of the IEEE ICASSP, Barcelona, Spain, 4–8 May 2020; pp. 356–360.
29. Korse, S.; Pia, N.; Gupta, K.; Fuchs, G. PostGAN: A GAN-Based Post-Processor to Enhance the Quality of Coded Speech. In Proceedings of the IEEE ICASSP, Singapore, 22–27 May 2022; pp. 831–835.
30. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2014), Montreal, QC, Canada, 8–13 December 2014; pp. 1–9.
31. Kumar, K.; Kumar, R.; De Boissiere, T.; Gestin, L.; Teoh, W.Z.; Sotelo, J.; de Brébisson, A.; Bengio, Y.; Courville, A.C. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; pp. 14910–14921.
32. Kaneko, T.; Tanaka, K.; Kameoka, H.; Seki, S. iSTFTNet: Fast and Lightweight Mel-Spectrogram Vocoder Incorporating Inverse Short-Time Fourier Transform. In Proceedings of the IEEE ICASSP, Singapore, 22–27 May 2022; pp. 6207–6211.
33. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.K.; Wang, Z.; Smolley, S.P. Least Squares Generative Adversarial Networks. In Proceedings of the IEEE ICCV, Venice, Italy, 22–29 October 2017; pp. 2813–2821.
34. Kong, J.; Kim, J.; Bae, J. HiFi-GAN: Generative Adversarial Networks for Efficient and High-Fidelity Speech Synthesis. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2020), Virtual, 6–12 December 2020; pp. 17022–17033.
35. Yamamoto, R.; Song, E.; Kim, J.M. Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram. In Proceedings of the IEEE ICASSP, Barcelona, Spain, 4–8 May 2020; pp. 6199–6203.
36. *ITU-T Recommendation P.862.2*; Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs. ITU: Geneva, Switzerland, 2007.
37. Hines, A.; Skoglund, J.; Kokaram, A.; Harte, N. ViSQOL: The Virtual Speech Quality Objective Listener. In Proceedings of the IWAENC, Aachen, Germany, 4–6 September 2012; pp. 1–4.

38. Chinen, M.; Lim, F.S.; Skoglund, J.; Gureev, N.; O'Gorman, F.; Hines, A. ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric. In Proceedings of the QoMEX, Athlone, Ireland, 26–28 May 2020; pp. 1–6.
39. Valentini-Botinhao, C.; Wang, X.; Takaki, S.; Yamagishi, J. Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System Using Deep Recurrent Neural Networks. In Proceedings of the Interspeech 2016, San Francisco, CA, USA, 8–12 September 2016; pp. 352–356.
40. Veaux, C.; Yamagishi, J.; King, S. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In Proceedings of the O-COCOSDA, Seoul, Republic of Korea, 1–3 November 2013.
41. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the ICLR, San Diego, CA, USA, 7–9 May 2015.
42. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**, arXiv:1607.06450. [CrossRef]
43. Luo, Y.; Mesgarani, N. Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1256–1266. [CrossRef] [PubMed]
44. Mittag, G.; Naderi, B.; Chehadi, A.; Möller, S. NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets. In Proceedings of the Interspeech 2021, Brno, Czech Republic, 30 August–3 September 2021; pp. 2127–2131.
45. *ITU-R Recommendation BS.1534-3*; Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems. ITU: Geneva, Switzerland, 2015.