Neural Comb Filtering Using Sliding Window Attention Network for Speech Enhancement

Venkatesh Parvathala¹ · Sivaganesh Andhavarapu¹ · Giridhar Pamisetty¹ · K. Sri Rama Murty¹

Received: 3 October 2021 / Revised: 14 July 2022 / Accepted: 14 July 2022 / Published online: 5 August 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

In this paper, we demonstrate the significance of restoring harmonics of the fundamental frequency (pitch) in the deep neural network (DNN)-based speech enhancement. The parameters of the DNN can be estimated by minimizing the mask loss, but it does not restore the pitch harmonics, especially at higher frequencies. In this paper, we propose to restore the pitch harmonics in the spectral domain by minimizing cepstral loss around the pitch peak. Restoring the cepstral pitch peak, in turn, helps in restoring the pitch harmonics in the enhanced spectrum. The proposed cepstral pitch-peak loss acts as an adaptive comb filter on voiced segments and emphasizes the pitch harmonics in the speech spectrum. The network parameters are estimated using a combination of mask loss and cepstral pitch-peak loss. We show that this combination offers the complementary advantages of enhancing both the voiced and unvoiced regions. The DNN-based methods primarily rely on the network architecture, and hence, the prediction accuracy improves with the increasing complexity of the architecture. The lower complex models are essential for real-time processing systems. In this work, we propose a compact model using a sliding-window attention network (SWAN). The SWAN is trained to regress the spectral magnitude mask (SMM) from the noisy speech signal. Our experimental results demonstrate that the proposed approach achieves comparable performance with the state-of-the-art noncausal and causal speech enhancement.
methods with much lesser computational complexity. Our three-layered noncausal SW AN achieves 2.99 PESQ on the Valentini database with only $10^9$ floating-point operations (FLOPs).

**Keywords**  Spectral magnitude mask · Transformer · Pitch harmonics · Cepstral pitch peak · Production-related loss

1 **Introduction**

Prolonged exposure to noisy speech signals causes severe fatigue to the listeners [26]. Speech enhancement (SE) algorithms aim to improve the quality and intelligibility of the noisy speech signal, which is degraded by the background noise [28]. SE plays a vital role in human–human communication over mobile/radio channels, hearing aids, cochlear implants, etc. [65]. SE is a crucial first step for domestic voice assistants with multiple noise sources like television, microwave oven, competing speakers, etc. [15]. Even though several methods have been proposed, SE still remains a challenging problem, especially for unseen speakers and noises. SE algorithms can be broadly classified into two categories:

- Estimating the noise characteristics and suppressing them.
- Estimating the speech characteristics and emphasizing them.

Most of the statistical SE algorithms rely on the adaptive estimation of the noise component and subtracting it from the noisy signal [2, 9, 47]. The performance of these algorithms depends on their noise tracking ability. Although they have better performance in stationary noise environments, it degrades significantly in the presence of nonstationary noise environments [4]. Many advanced noise tracking algorithms were proposed to estimate the nonstationary noise characteristics [32, 33]. The algorithms based on Bayesian statistics are well received in the literature, and their performance primarily depends on the accuracy of the estimated a priori and a posteriori signal-to-noise ratio (SNR) [17]. The statistical methods typically involve explicit suppression of estimated noise, which introduces musical noise distortion in the enhanced speech signal [3]. Despite these drawbacks, statistical enhancement methods are widely used in commercial applications because of their low computational complexity and real-time performance. The DNN architectures are proposed to estimate a priori SNR, which is subsequently used in the minimum mean-squared error (MMSE) estimation of the clean speech [36].

DNN approaches, on the other hand, rely on learning the structure of the speech signal through a nonlinear mapping between noisy and clean speech signals [57]. As DNN approaches pose SE as a supervised learning task, various noises at different SNRs can be used during the training. Hence, DNN approaches perform considerably better than the statistical approaches in unseen noises and lower SNR conditions [12, 39]. The superior performance of DNNs comes at the expense of substantial computational complexity. There is a need to develop efficient low-complexity DNN architectures for low-power applications [46].
Network architecture and the objective function play an important role in efficiently capturing the signal characteristics. In the initial studies, feed-forward neural networks (FFNNs) were proposed for spectral regression-based SE [64]. However, FFNN architectures are memoryless networks that cannot capture the temporal dependencies across the frames. Subsequently, the convolutional and recurrent architectures were employed to exploit the long-term contextual information for improved SE [29, 40, 43]. However, the convolutional architectures require a deep stack of layers to achieve a higher receptive field [20]. Although the recurrent architectures offer longer contexts, they are not suitable for parallel processing [27]. Recent advances in the transformer architectures offer compact models for parallel processing of the frames while capturing long-term dependencies through explicit attention mechanism [51]. Kim et al. proposed a Gaussian weighted self-attention transformer (TGSA) for SE [22]. Subsequently, Nicolson et al. proposed a transformer-based architecture (MHANet) to estimate a priori SNR in a statistical enhancement framework. Later, Wang et al. proposed a two-stage transformer neural network (TSTNN) in the time-domain for SE [59]. Although TSTNN is a compact model, it requires enormous computation as it operates at the sample level in the time domain. In this paper, we propose a SW AN to estimate spectral masks in the frequency domain. As the proposed architecture operates at the frame level of the speech signal, it requires significantly lesser computation than TSTNN.

The frequency-domain approaches rely on estimating spectral mask from the noisy speech spectrum. The network parameters are updated using a combination of mask approximation loss and signal approximation loss [62]. State-of-the-art frequency-domain approaches for the SE use perception-related loss functions for signal approximation [14, 23, 60]. Although the perception-related loss functions operating in Mel/Bark domains are good at approximating the lower frequency bands, they do not perform well in the high-frequency bands. In this paper, we propose a production-related loss function for signal approximation.

Speech signal exhibits high SNR regions in both time and frequency domains. The region around the glottal-closure instants (GCIs) exhibits higher amplitudes in the time domain, and hence, they are less vulnerable to the degradation [66]. As the periodicity of GCIs transforms to pitch harmonics in the frequency domain, adaptive comb-filtering has been used in the literature to enhance them [30, 35]. The pitch harmonics in the frequency domain show up as a peak in the cepstral domain [38]. Hence, accurately estimating the pitch peak in the cepstral domain enhances the pitch harmonics in the frequency domain and quasi-periodic structure of GCIs in the time domain. Elshamy et al. proposed an approach to manipulate the excitation and envelope in the cepstral domain to estimate a priori SNR in a statistical framework [8]. The excitation manipulation restores the pitch harmonics. In this work, we propose a novel loss function to better approximate the pitch harmonics in the spectral mask estimated using a DNN. The estimated spectral mask resembles a comb-filter and enhances the noisy speech spectrum across the frequency bands.

The rest of the paper is organized as follows. Section 2 presents the proposed transformer architecture for the speech enhancement using a sliding window attention network. The production-related loss function to enhance the pitch harmonics is discussed in Sect. 3. In Sect. 4, we compare the proposed method with the other
state-of-the-art methods in terms of speech quality and intelligibility measures. Several ablation studies were conducted to understand the effect of various hyper-parameters. Finally, Sect. 5 summarizes the crucial contributions of this work.

2 Transformer for Spectral Mask Estimation

Let the desired speech signal of interest \( d[n] \) be corrupted by uncorrelated additive noise \( g[n] \), leading to a noisy speech observation \( s[n] \) given by

\[
s[n] = d[n] + g[n]. \quad (1)
\]

The discrete Fourier transform (DFT) representation is more pertinent to the stationary signals. As speech is a nonstationary signal, the spectral characteristics of the speech signal vary with time. Moreover, the effect of additive noise is not uniform across time and frequency in nonstationary environments [28]. As a result, the noisy speech observation exhibits varying segmental and subband SNRs across time and frequency bins, respectively. The noisy speech signal is typically processed in the STFT domain to deal with the varying levels of degradation in the time–frequency (TF) plane. The STFT of a signal \( s[n] \) is defined as

\[
S[t, k] = \sum_{l=0}^{L-1} s[l + tR] w[l] e^{-j2\pi tl/L}, \quad k = 0, 1, 2, \ldots, K - 1 \quad (2)
\]

where \( w[n] \) is a window of length \( L \), \( R \) is the frame shift in samples, \( t \) is the frame index, and \( k \) is the frequency bin. Since STFT is a linear transformation, STFT of the noisy speech signal \( s[n] \) can be expressed as

\[
S[t, k] = D[t, k] + G[t, k], \quad (3)
\]

where \( D[t, k] \) and \( G[t, k] \) are the STFTs of clean speech signal \( d[n] \) and additive noise \( g[n] \), respectively.

Because of the nonstationary spectral characteristics, noisy speech signals must be processed using a time-varying filter to improve the perceptual quality. DNN approaches realize this time-varying filter through spectral masking in the TF plane. The spectral mask is expected to attenuate the noise bins while retaining the speech bins in the TF plane. In this approach, the noisy speech spectrum is multiplied by the spectral mask to estimate the clean speech spectrum,

\[
\hat{D}[t, k] = M[t, k] S[t, k] \quad (4)
\]

where \( M[t, k] \) is the spectral mask, and \( \hat{D}[t, k] \) is the STFT of the enhanced speech signal. Both real and complex masks were explored in the literature. A real mask enhances only the magnitude spectrum \([10, 18, 49, 61]\), while a complex mask enhances both magnitude and phase spectra \([63]\). A detailed discussion on a variety of spectral masks
can be found in [57]. In this work, we use the SMM for speech enhancement. The SMM defined as

$$M[t, k] = \frac{|D[t, k]|}{|S[t, k]|},$$  

(5)

has been shown to perform better than other masks [57]. Another advantage of SMM is that the computation of the oracle mask does not require access to the noise. Since the SMM is a real mask, the noisy phase is retained during signal reconstruction from \(\hat{D}[t, k]\) through inverse STFT [5]. We use a transformer architecture with sliding window attention to estimate SMM from noisy speech signals.

### 2.1 Sliding Window Attention Network

A transformer is a neural network model that learns context information through an attention mechanism, with differential weights at each part of the input sequence [55]. Unlike recurrent neural networks (RNNs), the transformers need not process the input sequence in the temporal order of evolution. Instead, the attention mechanism extracts context for any position in the sequence. The proposed transformer architecture with sliding window attention for estimating \(M[t, k]\) from \(S[t, k]\) is shown in Fig. 1. We have used time-distributed dense layers and temporal attention layers as the basic building blocks to facilitate the parallel processing of the data. The dense layers capture correlations across the frequency, while the temporal attention layers capture correlations across the time.

The log-magnitude spectrum of the noisy speech signal is projected onto a \(d\)-dimensional space using a dense layer and batch-normalized to obtain context-independent nonlinear projections \(x[t]\), given by

$$x[t] = f_b(W \log(|s[t]|) + b)$$  

(6)

where \(s[t] \in \mathbb{R}^K\) is the vector of STFT coefficients at time \(t\), \(W \in \mathbb{R}^{d \times K}\), \(b \in \mathbb{R}^d\) are the weights and biases of the dense layer, \(f_b(.)\) denotes the nonlinear activation followed by batch normalization. As the dense layer is memoryless, it does not capture the context information from the sequence. These context-independent projections \(x[t]\) are processed using a transformer block, as shown in Fig. 1 to extract the context information.

As speech is a nonstationary signal, the context information can be extracted by performing short-term analysis on \(x[t]\). In this work, we propose to use sliding window attention for extracting time-varying context information from \(x[t]\) [1]. In the transformer block, \(x[t]\) is linearly projected to obtain query \(x_q[t]\), key \(x_k[t]\) and value \(x_v[t]\) projections, given by

$$x_q[t] = W_q x[t] \quad x_k[t] = W_k x[t] \quad x_v[t] = W_v x[t]$$

where \(W_q \in \mathbb{R}^{d_q \times d}\), \(W_k \in \mathbb{R}^{d_k \times d}\) \(W_v \in \mathbb{R}^{d_v \times d}\) are three learnable linear maps. The query and key projections are used to learn the attention required from the neighboring
frames to enhance the current frame. We use attention from $c$ frames on either side of the current frame. The scaled dot product attention between frame-$t$ and frame-$\tau$ is given by

$$A[t, \tau] = \text{softmax} \left( \frac{\langle x_q[t], x_k[\tau] \rangle}{\sqrt{d_k}} \right), \quad t - c \leq \tau \leq t + c$$  \hspace{1cm} (7)$$

where softmax function operates over the context of $2c + 1$ frames. The attention weights are used to linearly combine the value vectors $x_v[\tau]$ over a context of $2c + 1$ frames to obtain the context-vector $x_c[t]$, given by

$$x_c[t] = \sum_{\tau=t-c}^{t+c} A[t, \tau] x_v[\tau].$$ \hspace{1cm} (8)$$

In the case of multi-head attention, multiple context vectors $x_c^h[t]$ are estimated using parallel projections as shown in Fig. 1. The context vectors across the heads are concatenated and projected to a lower dimension to obtain the combined summary, given by
\( \mathbf{x}_c[t] = W_o[\mathbf{x}^1_c[t], \mathbf{x}^2_c[t], \ldots, \mathbf{x}^H_c[t]], \) 

(9)

where \( W_o \in \mathbb{R}^{d \times H_d} \) is a learnable linear projection, and \( H \) denotes the number of parallel heads. The context vector estimation in the transformer can be interpreted as kernel adaptive filtering [41]. The recent advances in transformer architectures offer a way to learn the optimum kernel from the data in a supervised manner.

The context-independent vector \( \mathbf{x}[t] \) is added to the context-vector \( \mathbf{x}_c[t] \) through a residual connection and the sum is batch-normalized. Assuming that the network is operating in the log-spectral domain, the residual connection can be interpreted as filtering \( \mathbf{x}[t] \) through a time-varying filter \( \mathbf{x}_c[t] \). As the attention block is intended to capture the correlations along the temporal axis, the batch-normalized output is processed using a couple of feed-forward layers to capture the correlations along the frequency axis. When multiple layers of transformer blocks are employed, the signal is filtered along the time and frequency domains to estimate the SMM in the TF domain.

The output of the transformer block is processed using the two time-distributed feed-forward layers to estimate the SMM. As SMM is nonnegative, an exponential activation function is used in the final layer to estimate \( \hat{M}[t, k] \).

The network parameters can be estimated to minimize the distance between the oracle and estimated masks \( \hat{M}[t, k] \)

\[
L_{\text{mask}} = \frac{1}{|\mathcal{B}|} \sum_{u \in \mathcal{B}} \sum_t \sum_k |M^u[t, k] - \hat{M}^u[t, k]|_p.
\]

(10)

where \( u \) denotes utterance number, \( |\mathcal{B}| \) denotes cardinality of the batch, and \( |.|_p \) denotes \( p \)-norm of the error. The mean square error (MSE) between the masks was used as the mask approximation objective [62]. However, closer observation of the oracle mask in Fig. 2a reveals the occasional large amplitude fluctuations, which need not be given a proportional weight in error. Hence, we preferred to use mean absolute error (MAE), i.e., \( p = 1 \), for the mask estimation.

In the earlier works, it has been reported that the mask approximation does not directly optimize the speech enhancement objective [62]. This occurs because all the finer fluctuations in the oracle SMM, shown in Fig. 2a, may not contribute equally to the perceptual improvement. The objective function should be chosen to prioritize the perceptually important regions of the spectrum. Hence, a spectral approximation objective is also used along with the mask approximation objective. Weninger et al. [62] used MSE in the Mel-spectral domain as a spectral approximation objective. The SMM estimated using a combination of mask-loss and Mel-spectral loss is shown in Fig. 2a. While the estimated SMM matches well with the oracle mask in the low-frequency bands (\(< 2 \text{ kHz}\)), the SMM approximation is not good in the high-frequency bands (\(> 2 \text{ kHz}\)). In the subsequent studies, perceptual evaluation of speech quality (PESQ) was used as the signal approximation objective [11, 23]. These signal approximation objectives were motivated from speech perception. Alternatively, the structure imposed by the speech production mechanism can be exploited for signal reconstruction. In this work, we propose to exploit the quasi-periodic structure imposed by the voice source for speech enhancement.

\( \odot \) Birkhäuser
3 Enhancing Pitch Peak in Cepstral Domain

During the production of voiced sounds, significant excitation to the vocal tract system is delivered at the GCIs [34]. Hence, the region around the GCIs is more robust to the additive noise within each pitch period [66]. The quasi-periodic sequence of GCIs in the time domain are manifested as pitch harmonics in the frequency domain and as pitch peak in the cepstral domain [38]. The real cepstrum of a clean speech signal can be computed from its spectrum $S[t, k]$ as

$$S_c[t, q] = (W_D \log(|S[t, k]|))$$  \hspace{1cm} (11)

where $W_D$ is the inverse discrete cosine transform (IDCT) matrix [44] and $q$ denotes the quefrency bin in the cepstral domain.

In the cepstral domain, the low-time components predominantly correspond to the vocal tract information, while the high-time components correspond to the voice source [42]. The most prominent peak at high-time cepstrum corresponds to the pitch period [38]. Figure 3a shows the cepstrum of a frame of clean speech signal for a female speaker. The cepstral peak at around 3.7 milliseconds (ms) corresponds to the pitch period of the speaker. Figure 3a also shows the cepstrum of the enhanced signal when the mel-spectral loss is used as a signal approximation objective. In this case, there is no reliable estimation of the pitch peak, as mel-frequency wrapping smooths the pitch harmonics.

The harmonic structure in the spectral domain can be enhanced by restoring the pitch peak in the cepstral domain. Hence, we propose to minimize the MSE between the pitch peaks of the clean and enhanced cepstra. During the training, the location of pitch peak is predetermined from clean speech using pitch tracking algorithm [21, 67]. To account for the minor deviations, we have considered a short window around the
Fig. 3 Real cepstrum of a voiced frame enhanced using (a) mel-spectral loss, (b) cepstral pitch-peak loss. Oracle cepstrum is shown for comparison.

estimated pitch peak, as indicated by dotted lines in Fig. 3a. The cepstral pitch-peak loss for a random batch \( B \) is computed as

\[
L_{\text{cep}} = \frac{1}{|B|} \sum_{u \in B} \sum_{t \in V} \sum_{q = Q_l}^{Q_u} \left| \hat{S}_c(u)[t, q] - \tilde{S}_c(u)[t, q] \right|^2
\]  

(12)

where \( \hat{S}_c(u)[t, q] \) is the real-cepstrum of the enhanced signal, \( V \) is the set of voiced frames in the utterance \( u \), \( Q_l \) and \( Q_u \) denote the lower and upper bounds of the pitch window, respectively. In this work, \( Q_l \) and \( Q_u \) are chosen to be seven samples each to account for a 1 ms deviation in the pitch period estimation. The cepstral pitch peak loss alone cannot be used for speech enhancement, as it is defined only for the voiced regions. Hence, we combined it with the mask loss to improve the quality of the enhanced signal in the voiced regions.

In the DNN approaches, the network weights \( \theta \) are updated using gradient descent, which involves computing gradients of the loss function \( L \) with respect to each network weight, that is \( \nabla_{\theta} L \). For spectral masking methods, the gradient of the loss function can be decomposed as

\[
\nabla_{\theta} L = \nabla_{\theta} \hat{M} \nabla_{\hat{M}} L
\]  

(13)

where \( \hat{M} \) is the estimated mask. The choice of the loss function influences the second term in (13), i.e., the gradient of the loss function with respect to the estimated mask \( \nabla_{\hat{M}} L \). In order to analyze the effect of different loss functions in training the network, we plotted \( \nabla_{\hat{M}} L \) for \( L_{\text{mask}} \), \( L_{\text{mel}} \) and \( L_{\text{cep}} \) in Fig. 4. The gradient of the mask loss \( \nabla_{\hat{M}} L_{\text{mask}} \) is shown in Fig. 4c. Since MAE is used for the mask loss, its gradients take only two values, \(+ \frac{1}{|B|}\) or \(- \frac{1}{|B|}\), depending on the sign of the error. It explains the
binary nature of the gradient plot in Fig. 4c. We can notice from the figure that the gradients captured the gross structure of the spectrogram. The gradients are higher in the unvoiced and silence regions to achieve noise suppression in those regions, which typically exhibit low segmental SNR. The gradients of the mask loss remain uniformly low in the voiced regions. Hence, it does not contribute to suppressing the noise present between the pitch harmonics.

Figure 4d shows the gradients of the mel-spectral loss, $\nabla_M L_{\text{mel}}$. In this case, the magnitude of the gradients is close to zero in most regions, which could be attributed
to the spectral wrapping using a mel-filter bank that smooths both the desired and estimated spectra. As the mel-filter bank offers higher resolution at low frequency, the first and second formants were emphasized in the gradients plot.

Figure 4e shows the gradients of cepstral pitch-peak loss, $\nabla_{\hat{M}} L_{\text{cep}}$. The gradients are zero in the unvoiced regions as the loss is evaluated only for the voiced regions by applying a voice activity flag. In the voiced regions, the gradients exhibit harmonic structure, negative values at pitch harmonics, and positive values in between the harmonics. Hence, the cepstral pitch-peak loss suppresses noise in the spectral valleys formed between the successive harmonics, emphasizing the harmonic structure.

The mask loss and the cepstral pitch-peak loss offer complementary advantages. The mask loss suppresses the noise in the unvoiced and silence regions, while the cepstral pitch-peak loss enhances the harmonic structure in the voiced regions. Hence, a weighted combination of mask loss and cepstral pitch-peak loss is optimized for updating the network parameters. The composite loss function used for training the network is given by

$$L = \alpha L_{\text{mask}} + (1 - \alpha) L_{\text{cep}}.$$  \hspace{1cm} (14)

where the constant $\alpha$ determines the relative importance given to the mask and cepstral pitch-peak losses. Since the gradient magnitudes of mask loss are ten times lesser than the gradients of cepstral pitch-peak loss, we choose $\alpha = 0.9$ to give equal importance to both the losses. Figure 3b illustrates the effectiveness of the proposed production-related loss function in retrieving the pitch peak. The pitch peak was accurately estimated with cepstral pitch-peak loss, whereas it is underestimated with mel-spectral loss. As a consequence, the SMM estimated using cepstral pitch-peak loss resembles a comb-filter as shown in Fig. 2b. The SMM estimated using the combination of mask loss and cepstral pitch-peak loss closely follows the oracle mask up to 5 kHz, demonstrating the effectiveness of the proposed composite loss function.

### 4 Experimental Evaluation

### 4.1 Database Description

The performance of the proposed SWAN is evaluated on a publicly available speech dataset created for evaluating the SE algorithms [54]. It contains clean speech recordings of 28 speakers with 14 male and 14 female taken from the Voice-Bank corpus [56] for training. Each speaker has about 400 sentences, creating a total of 11572 speech utterances. The noisy samples are generated by mixing ten different noises—eight real noises and two artificially generated noises. The real noise recordings were taken from diverse environments multi-channel acoustic noise (DEMAND) database [53], which include cafeteria, car, kitchen, meeting, metro, restaurant, station, and traffic noises. The two artificially generated noises are the speech-shaped noise (SSN) and the babble noise. These ten noises are mixed with the clean speech recordings at four different SNRs: 15 dB, 10 dB, 5 dB, and 0 dB. Hence, the DNNs were trained with 40 (10 noises $\times$ 4 SNR levels) different noisy conditions per speaker.
For testing, the clean speech recordings of two speakers (one male and one female), not included in the training set, were taken from the Voice-Bank speech corpus. Five noise types, viz. bus, cafe, living room, office, and public square that were not used for training, are taken from the DEMAND dataset. The noisy speech for testing was created at four different SNRs: 17.5 dB, 12.5 dB, 7.5 dB and 2.5 dB. Hence, 20 different noisy conditions are considered for each speaker. A total of 824 clean and noisy speech utterance pairs are available for evaluating the SE algorithms. The clean speech recordings of two speakers\(^1\) that are not included in either train or test dataset were taken from the Voice-Bank corpus to prepare the development dataset. To simulate the similar noisy conditions as the test dataset, the noisy signals are generated by mixing the clean speech signals with the same noises and SNR levels as the test dataset. All the speech utterances were downsampled from 48 to 16 kHz.

### 4.2 Feature Extraction and Model Training

Speech segments of the length of 1.6 s are randomly sliced from the available utterances, and STFT is computed with a frame size of 32 ms, shifted by 16 ms, resulting in 100 frames per segment. The network is trained to estimate SMM from the normalized 257-dimensional magnitude spectra. All the dense layers in the network are of 257 dimensions, i.e., \(d = 257\). We used swish activation for all the dense layers except for the final layer. In the final dense layer, the exponential activation function is used to estimate SMM. In the transformer block, we have used four heads \((H = 4)\) for extracting the context vector \(x_c[t]\). The dimension of the linear projections used to obtain query, key, and value vectors was set to 64, i.e., \(d_q = d_k = d_v = 64\). We observed that using different projections for keys and values did not offer a noticeable advantage in the performance. Hence, we have used the same linear projection for generating key and value vectors, i.e., \(W_K = W_V\). This choice significantly reduced the number of trainable parameters of the network.

The network is trained to minimize the composite loss function in Eq. (14). The Adam optimizer [24] with \(\beta_1 = 0.9, \beta_2 = 0.99\) is used to update the network parameters. The learning rate is exponentially decayed at the rate of 0.95 per every 15 epochs, starting from an initial value of 0.001. The network training was stopped if the training loss did not reduce for the 20 consecutive epochs. A batch size of 32 segments is used for training. The pitch period required for computing cepstral pitch-peak loss was estimated from the clean speech signal using yet another algorithm for pitch tracking (YAAAPT) algorithm [21].

### 4.3 Description of Evaluation Metrics

The performance of the SE algorithms is evaluated using the following objective quality and intelligibility measures:

- **PESQ** \([45]\) \([-0.5, 4.5]\): Perceptual evaluation of speech quality is an objective quality measure which quantifies the symmetric and asymmetric disturbances

\(^1\) p245(male) and p265(female) are considered for development dataset.
between the clean and enhanced speech. A higher PESQ score indicates better speech quality.
- STOI \([50] \in [0, 100]\): Short-time objective intelligibility is measured comparing the temporal envelopes of the clean and enhanced signals. A higher STOI score indicates better speech intelligibility.
- CSIG \([19] \in [0, 5]\): Composite signal distortion measure quantifies the level of distortion introduced by the SE algorithm in the enhanced signal. A higher CSIG score indicates lesser signal distortions.
- CBAK \([19] \in [0, 5]\): Composite background intrusiveness measure quantifies the level of leftover background noise in the enhanced signal. A higher CBAK score indicates better noise suppression.
- COVL \([19] \in [0, 5]\): Composite overall quality measure combines the above two composite measures to quantify the overall quality of the enhanced signal. A higher COVL score indicates better enhancement quality.
- SSNR \([16] \in [-10 \text{ dB}, 35 \text{ dB}]\): Segmental SNR is the average of the frame level SNR estimates.

In real-time applications, we often have limited memory and computing resources. Although the state-of-the-art methods achieve better speech quality, they may not be suitable for real-time operation because of the deeper models and high computational complexity. Hence, we compared the model size and the computational complexity of different algorithms to investigate their applicability in real time. The model size is quantified in terms of the number of network weights, while the computational complexity is measured in terms of the number of FLOPs \([52]\). We should note that a smaller model size need not necessarily result in lower computational complexity. Our experimental studies show that some smaller models require a much higher number of FLOPs than the deeper models.

### 4.4 Effect of Hyper Parameters

We conducted several experiments to study the effect of hyper-parameters on the performance of the proposed SWAN. A three-layered SWAN network is used for all these experiments. All the models are evaluated on the development data described in Sect. 4.1.

#### 4.4.1 Effect of Temporal Context

The information captured in the context vector of Eq. 8 critically depends on the width of the sliding window \((2c + 1)\). In order to understand the effect of sliding window size, we have conducted experiments with varying temporal contexts, ranging from \(c = 3\) to \(c = 15\), and the corresponding results are reported in Table 1.

For the longer temporal contexts, the attention layer averages the value vectors across the nonstationary regions of the signal and hence does not capture local characteristics. As a result, the performance is inferior for the longer contexts. In fact, the performance is poorest when the full-temporal (infinite) context is used for context vector estimation. However, a better SSNR has been achieved with infinite context.
### Table 1  Effect of temporal context on the proposed approach

| Context | PESQ  | STOI  | CSIG  | CBAK  | COVL  | SSNR  |
|---------|-------|-------|-------|-------|-------|-------|
| \(c = \infty\) | 2.95  | 83.47 | 3.73  | 3.37  | 3.33  | 8.07  |
| \(c = 15\)   | 3.06  | 83.87 | 3.84  | 3.41  | 3.45  | 7.81  |
| \(c = 10\)   | **3.08** | **83.89** | **3.85** | **3.42** | **3.46** | **7.90** |
| \(c = 5\)    | 3.05  | 83.61 | 3.84  | 3.41  | 3.44  | 7.88  |
| \(c = 3\)    | 3.05  | 83.65 | 3.84  | 3.40  | 3.44  | 7.82  |

Bold values indicate the best results in each metric among the different methods.

### Table 2  Performance comparison across loss functions

| Objective | PESQ  | STOI  | CSIG  | CBAK  | COVL  | SSNR  | V-SSNR | UV-SSNR |
|-----------|-------|-------|-------|-------|-------|-------|--------|---------|
| Mask      | 3.02  | 83.81 | 3.81  | 3.38  | 3.41  | 7.79  | 18.68  | 1.12    |
| Mask + Mel| 3.06  | 83.66 | 3.83  | 3.39  | 3.44  | 7.60  | **19.01** | 0.60    |
| Mask + Cep| **3.08** | **83.89** | **3.85** | **3.42** | **3.46** | **7.90** | 18.91  | **1.14** |

Bold values indicate the best results in each metric among the different methods.

Moreover, the computational complexity and latency increase with longer contexts. In contrast, shorter contexts offer lower latency. However, the context vectors could be noisy and may not capture the temporal dependencies. It is found that a ten-frame context on either side (\(c = 10\)) offers a good compromise between the performance and computational complexity. Hence, we used a sliding window of 21 frames (around 0.3 s) to extract the context vector for all the subsequent studies.

### 4.4.2 Significance of the Cepstral Pitch-Peak Loss

To demonstrate the importance of the proposed cepstral pitch-peak loss, we have conducted ablation studies with and without including it in training the model. Firstly, we trained our network with only mask loss, followed by the combination of mask loss and cepstral pitch-peak loss. We have also evaluated the model performance for the combination of mask loss and mel-spectral loss for comparison. The performance of a 3-layer SWAN trained with different loss functions is shown in Table 2.

The quality and intelligibility of the enhanced signal improved when mask loss is combined with either mel-spectral loss or cepstral-pitch peak loss. The system achieved the best performance when mask loss was combined with the cepstral pitch-peak loss. Since the cepstral pitch-peak loss is effective only in the voiced regions, we have evaluated the SSNR of voiced and unvoiced segments separately. The voiced and unvoiced segments are identified using a voice activity detector [21]. The segmental SNR of voiced regions (V-SSNR) improved significantly when the cepstral pitch-peak loss was combined with the mask loss. In contrast, the segmental SNR of unvoiced regions (UV-SSNR) remained almost the same because the gradient of cepstral pitch-peak loss was zero in the unvoiced regions. On the other hand, incorporating the mel-spectral loss adversely affected the unvoiced segments while performing relatively better in the voiced regions. This study demonstrates the significance of the proposed cepstral pitch peak loss in restoring pitch harmonics in the spectral domain.
Table 3 Performance of 3-layer SWAN across SNR levels

| i/p SNR | Signal | PESQ | STOI | CSIG | CBAK | COVL | SSNR |
|---------|--------|------|------|------|------|------|------|
| 2.5 dB  | N      | 1.49 | 76.61| 2.88 | 1.86 | 2.13 | –2.89|
|         | E      | 2.54 | 80.09| 3.46 | 2.94 | 2.98 | 4.87 |
| 7.5 dB  | N      | 1.81 | 79.60| 3.26 | 2.25 | 2.50 | 0.14 |
|         | E      | 2.96 | 82.22| 3.78 | 3.29 | 3.36 | 6.80 |
| 12.5 dB | N      | 2.23 | 83.63| 3.67 | 2.73 | 2.93 | 3.92 |
|         | E      | 3.27 | 85.67| 3.98 | 3.58 | 3.62 | 8.83 |
| 17.5 dB | N      | 2.62 | 86.31| 4.09 | 3.25 | 3.36 | 8.57 |
|         | E      | 3.54 | 87.63| 4.18 | 3.87 | 3.87 | 11.12 |

Signal—N denotes noisy signal and E denotes enhanced signal

4.4.3 Performance Evaluation Across SNR Levels

The performance of the proposed SWAN network is evaluated at different SNR levels, and the results are presented in Table 3. The evaluation metrics for the noisy signals are also provided in the table to analyze the relative improvement. The evaluation metrics across the SNR levels are consistently better for the enhanced signals than those of the noisy input signals (N). The proposed method achieves 70.47% absolute improvement in PESQ at 2.5 dB SNR, illustrating the effectiveness of the pitch restoration at lower SNR levels.

4.4.4 Performance Evaluation Across Genders

The effect of noise is not the same across the speakers from different genders [25, 31]. For the same SNR level, the perceived degradation can be significantly higher for female speakers compared to male speakers, which can be observed from the evaluation metrics of the noisy signals (N) of the male and female speakers in Table 4. This behavior can be attributed to the higher fundamental frequency of the female speakers, which results in widely-spaced pitch harmonics in the narrowband spectrogram. Hence, the frequency bins between the pitch harmonics are vulnerable to the additive noise, degrading the perceptual quality. The enhancement algorithms should sup-

Table 4 Performance of 3-layer SWAN across genders

| Gender | Signal | PESQ | STOI | CSIG | CBAK | COVL | SSNR |
|--------|--------|------|------|------|------|------|------|
| Male   | N      | 2.23 | 80.93| 3.64 | 2.65 | 2.92 | 2.63 |
|        | E      | 3.07 | 82.97| 3.96 | 3.41 | 3.51 | 7.59 |
| Female | N      | 1.83 | 82.17| 3.31 | 2.39 | 2.54 | 2.21 |
|        | E      | 3.08 | 84.83| 3.73 | 3.44 | 3.40 | 8.21 |

N denotes noisy signal and E denotes enhanced signal
press the noise between the pitch harmonics to achieve better performance on female speakers. The mask estimated using the proposed cepstral pitch-peak loss resembles a time-varying comb filter and helps in effectively suppressing the noise between the pitch harmonics. The performance of the proposed SE algorithm for both genders is reported in Table 4. Improvements in all the quality and intelligibility measures across the genders can be observed.

### 4.5 Performance Comparison with the State-of-the-Art Methods

The proposed SE algorithm\(^2\) is compared with several state-of-the-art methods evaluated on the Valentini test dataset. The methods chosen for comparison can be broadly classified into two categories—frequency domain methods aimed at estimating spectral gain/mask and time-domain methods that directly regress the clean waveform. A brief description of the methods chosen for the comparison is given below.

DeepMMSE [69] is based on a statistical enhancement framework estimating a priori SNR using a temporal convolutional network (TCN). Similarly, MHANet [37] also estimates a priori SNR using a multi-head attention network. The MMSE-GAN [48] is a generative adversarial network (GAN) that implicitly estimates a multiplicative mask. The generator and discriminator are composed of three dense layers each, and the mean squared error criterion is used to learn the parameters of the network. MetricGAN [11] and MetricGAN+ [14] are also based on the GAN framework in which the generator consists of bidirectional long short term memory (BLSTM) layers estimating the SMM, and the discriminator consists of convolutional layers evaluating the perceptual quality of the enhanced signal. The discriminator in these two networks was designed to mimic the behavior of the PESQ metric. SDR-PESQ [23] estimates a multiplicative mask using the combination of convolutional and BLSTM layers. In T-GSA [22], a transformer-based architecture is used to estimate the SMM. Both SDR-PESQ and T-GSA are trained to optimize the SDR and PESQ metrics. PhaseDCN [68] estimates the phase spectrum along with the ideal ratio mask using a dilated convolutional dual-path network.

DEMUCS [7] and TSTNN [68] are the time-domain approaches. DEMUCS uses a combination of convolutional and LSTM layers, while the TSTNN uses a combination of convolutional and transformer layers. However, the transformer layer of TSTNN consists of a gated recurrent unit (GRU) instead of a conventional dense layer. Both DEMUCS and TSTNN are trained to minimize a combination of waveform reconstruction loss and STFT loss.

### 4.5.1 Comparison with Noncausal SE Algorithms

Causality is one of the desirable features of SE algorithms for real-time processing. Noncausal models use frames from either side to extract context-rich representations and achieve better performance. However, noncausal models cannot be used in real time as they require future frames for processing the current frame. Although some of

---

\(^2\) Audio samples: [https://siplab-iith.github.io/SWAN](https://siplab-iith.github.io/SWAN).

\(^3\) Inference codes and pre-trained models: [https://github.com/SIPLab-IITH/SWAN-Neural-Comb-Filter](https://github.com/SIPLab-IITH/SWAN-Neural-Comb-Filter).
Table 5: Performance comparison with state-of-the-art noncausal SE algorithms on Valentini database

| Model       | PESQ  | STOI  | CSIG  | CBAK  | COVL  | SSNR  | Parameters (M) | FLOPS (B) |
|-------------|-------|-------|-------|-------|-------|-------|----------------|-----------|
| Noisy       | 1.97  | 92.00 | 3.35  | 2.44  | 2.63  | 1.68  | –              | –         |
| MMSE-GAN [48] | 2.53  | 93.00 | 3.80  | 3.12  | 3.14  | –     | –              | –         |
| MetricGAN [11] | 2.86  | –     | 3.99  | 3.18  | 3.42  | –     | 1.89           | –         |
| TSTNN $^4$ [58] | 2.94  | 94.68 | 4.30  | 3.54  | 3.64  | 9.93  | 0.92           | 5920.03   |
| Deep-MMSE [69] | 2.95  | 94.00 | 4.28  | 3.46  | 3.64  | –     | 2.00           | –         |
| SDR-PESQ [23] | 3.01  | –     | 4.09  | 3.54  | 3.55  | 10.44 | –              | –         |
| T-GSA [22]   | 3.06  | –     | 4.18  | 3.59  | 3.62  | 10.78 | –              | –         |
| DEMUCS$^4$ [6] | 3.04  | 95.30 | 4.36  | 3.52  | 3.72  | 9.29  | 33.50          | 8235.07   |
| MetricGAN$^e$ $^4$ [13] | 3.13  | 93.20 | 4.10  | 2.90  | 3.61  | 3.76  | 1.89           | 15.19     |
| SWAN (1L)    | 2.89  | 93.74 | 3.71  | 3.34  | 3.30  | 8.03  | 0.53           | 4.46      |
| SWAN (3L)    | 2.99  | 94.21 | 3.79  | 3.41  | 3.39  | 8.29  | 1.20           | 9.81      |
| SWAN (6L)    | 3.02  | 94.32 | 3.81  | 3.44  | 3.42  | 8.44  | 2.19           | 17.83     |

$^4$ The models are available online and are taken for detailed analysis in terms of computational complexity

The algorithms mentioned above can be restricted to use causal context, most of them were trained with noncausal contexts to achieve improved performance. Hence, we compare the performance of the proposed SWAN model with causal and noncausal contexts separately.

Table 5 compares the performance of SWAN with the state-of-the-art noncausal SE algorithms. Our model achieved comparable performance with the state-of-the-art methods like SDR-PESQ, T-GSA, and DEMUCS. Even though Metricgan+ has the best PESQ, its STOI and SSNR suffered because of the time-varying gain introduced while optimizing the PESQ metric. Figure 5 shows the denoised speech segment with different methods. In the case of Metricgan+, the enhanced signal does not match well with the clean signal because of the time-varying gain that cannot be compensated by a constant scaling factor. We evaluated the MSE between the clean and enhanced signals to quantify the distortion. Metricgan+ has higher signal distortion than other methods. It explains the reason for its lower STOI and SSNR, even though its PESQ is higher. As DEMUCS and TSTNN are time-domain approaches, they perform better on STOI and SSNR measures. However, their computational complexity is very high as they operate in the time-domain, i.e., on the speech samples. On the other hand, frequency-domain approaches like Metricgan+ and SWAN operate at the frame level and require much lesser computation.

In particular, SWAN offers a good balance between enhancement quality and computational complexity. Even a single-layer SWAN (1L) delivers impressive performance with just 0.53 million (M) parameters and 4.46 billion (B) FLOPs. Though the SWAN is not trained to optimize any of the metrics explicitly, the performance of SWAN is comparable to the approaches that explicitly optimize the speech quality metrics. It demonstrates the perceptual significance of restoring the pitch harmonics in the spectral domain using the cepstral pitch-peak loss.
4.5.2 Comparison with Causal SE Algorithms

The proposed sliding window attention can be restricted to include only the causal context for enhancing the streaming audio in real-time. The performance of the proposed SWAN trained with causal context is compared with state-of-the-art causal SE algorithms. The quality and intelligibility measures of the causal models are given in Table 6. It is observed that the performance of the causal algorithms is generally inferior to that of the noncausal counterparts. The performance of the SWAN (6L) is better than other frequency-domain methods and comparable with the DEMUCS, which is a time-domain method. However, as indicated earlier in Table 5, the computational complexity of the DEMUCS is much higher than the SWAN. Hence, the proposed SWAN architecture is better suited for real-time SE as it offers comparable performance with significantly lower computational complexity.

Table 6  Performance comparison of SWAN with the state-of-the-art causal algorithms

| Model          | PESQ | STOI | CSIG | CBAK | COVL | SSNR |
|----------------|------|------|------|------|------|------|
| Noisy          | 1.97 | 92.00| 3.35 | 2.44 | 2.63 | 1.68 |
| Weiner         | 2.22 | 93.00| 3.23 | 2.68 | 2.67 | 5.07 |
| Deep-MMSE [69] | 2.77 | 93.00| 4.14 | 3.32 | 3.46 | –    |
| PhaseDCN [68]  | 2.87 | 96.00| 4.37 | 3.75 | 3.64 | –    |
| MHANet [37]    | 2.88 | 93.60| 4.17 | 3.37 | 3.53 | –    |
| DEMUCS [7]     | 2.93 | 95.00| 4.22 | 3.25 | 3.52 | –    |
| SWAN (1L)      | 2.84 | 93.41| 3.66 | 3.31 | 3.24 | 7.98 |
| SWAN (3L)      | 2.90 | 93.78| 3.71 | 3.35 | 3.30 | 8.11 |
| SWAN (6L)      | 2.91 | 93.95| 3.72 | 3.37 | 3.31 | 8.18 |
5 Conclusions and Future Work

In this work, we proposed a sliding window attention-based transformer architecture for speech enhancement. The proposed network is trained to estimate the SMM from the noisy speech signal. We demonstrated the importance of incorporating the knowledge of the speech production process during the mask estimation, i.e., the quasi-periodic nature of speech signal, which transforms to harmonic structure in the frequency domain. In order to restore the pitch harmonics in the enhanced speech signal, we proposed to minimize cepstral pitch-peak loss, along with the mask loss. The spectral mask thus estimated resembles a time-varying comb-filter emphasizing the high SNR regions around the pitch harmonics. Our experimental results show that the proposed SWAN architecture achieves state-of-the-art performance with much lesser computational complexity.

The proposed approach aims to enhance the magnitude spectrum of the speech signal, and hence the noisy phase is reused during reconstruction. Our future efforts will be oriented toward developing a complex attention mechanism to incorporate the effect of phase in context vector generation. In addition, the cepstral pitch-peak loss proposed in this work is motivated by the speech production mechanism. In the future, we will combine the proposed production-related loss function with the perception-related loss functions.

References

1. I. Beltagy, M.E. Peters, A. Cohan, Longformer: the long-document transformer. arXiv preprint arXiv:2004.05150 (2020)
2. S. Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Speech Signal 27, 113–120 (1979)
3. O. Cappe, Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Trans. Speech Audio Process. 2(2), 345–349 (1994). https://doi.org/10.1109/89.279283
4. I. Cohen, B. Berdugo, Speech enhancement for non-stationary noise environments. Signal Process. (2001). https://doi.org/10.1016/S0165-1684(01)00128-1
5. R. Crochiere, A weighted overlap-add method of short-time Fourier analysis/synthesis. IEEE Trans. Acoust. Speech Signal Process. 28(1), 99–102 (1980)
6. A. Defossez, G. Synnaeve, Y. Adi, DEMUCS implementation codes and pre-trained models. https://github.com/facebookresearch/denoiser
7. A. Défossez, G. Synnaeve, Y. Adi, Real time speech enhancement in the waveform domain, in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH (2020). https://doi.org/10.21437/Interspeech.2020-2409
8. S. Elshamy, T. Fingscheidt, DNN-based cepstral excitation manipulation for speech enhancement. IEEE/ACM Trans. Audio Speech Lang Process 27(11), 1803–1814 (2019)
9. Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. (1984). https://doi.org/10.1109/TASSP.1984.1164453
10. H. Erdogan, J.R. Hershey, S. Watanabe, J. Le Roux, Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks, in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, 2015), pp. 708–712
11. S.W. Fu, C.F. Liao, Y. Tsao, S.D. Lin, Metricgan: generative adversarial networks based black-box metric scores optimization for speech enhancement, in International Conference on Machine Learning (PMLR, 2019), pp. 2031–2041
12. S.W. Fu, Y. Tsao, X. Lu, SNR-aware convolutional neural network modeling for speech enhancement, in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (2016). https://doi.org/10.21437/Interspeech.2016-211
13. S.W. Fu, C. Yu, T.A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, Y. Tsao, Metricgan+ pre-trained model and inference function. https://huggingface.co/speechbrain/metricgan-plus-voicebank
14. S.W. Fu, C. Yu, T.A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, Y. Tsao, Metricgan+: an improved version of Metricgan for speech enhancement. arXiv preprint arXiv:2104.03538 (2021)
15. J. Gnanamanickam, Y. Natarajan, S. Ramasamy, A hybrid speech enhancement algorithm for voice assistance application. Sensors 21, 7025 (2021). https://doi.org/10.3390/s21217025
16. J.H.L. Hansen, B.L. Pellom, An effective quality evaluation protocol for speech enhancement algorithms, in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia (1998)
17. R.C. Hendriks, R. Heusdens, J. Jensen, MMSE based noise PSD tracking with low complexity, in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, 2010), pp. 4266–4269
18. G. Hu, D. Wang, Speech segregation based on pitch tracking and amplitude modulation, in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics* (Cat. No. 01TH8575) (IEEE, 2001), pp. 79–82
19. Y. Hu, P.C. Loizou, Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang. Process. (2008). https://doi.org/10.1109/TASL.2007.911054
20. S. Jafarlou, S. Khorram, V. Kothapally, J.H.L. Hansen, Analyzing large receptive field convolutional networks for distant speech recognition, in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (2019), pp. 252–259
21. K. Kasi, S.A. Zahorian, Yet another algorithm for pitch tracking, in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1 (IEEE, 2002), pp. I–361
22. J. Kim, M. El-Khayam, J. Lee, T-GSA: transformer with Gaussian-weighted self-attention for speech enhancement, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020). https://doi.org/10.1109/ICASSP40776.2020.9053591
23. J. Kim, M. El-Khayam, End-to-end multi-task denoising for joint SDR and PESQ optimization. arXiv:1901.09146, J Lee - arXiv preprint (2019)
24. D.P. Kingma, J.L. Ba, Adam: a method for stochastic optimization, in *3rd International Conference on Learning Representations, ICLR—Conference Track Proceedings* (2015)
25. D.H. Klatt, L.C. Klatt, Analysis, synthesis, and perception of voice quality variations among female and male talkers. J. Acoust. Soc. Am. 87(2), 820–857 (1990)
26. M. Klatte, T. Lachmann, M. Meis, Effects of noise and reverberation on speech perception and listening comprehension of children and adults in a classroom-like setting. Noise Health (2010). https://doi.org/10.4103/1463-1741.70506
27. J.F. Kolen, S.C. Kremer, *Gradient Flow in Recurrent Nets: The Difficulty of Learning Long Term Dependencies* (Wiley-IEEE Press, New York, 2001), pp.237–243. https://doi.org/10.1109/9780470544037.ch14
28. P.C. Loizou, *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, 2007)
29. Y. Luo, N. Mesgarani, Conv-TasNet: surpassing ideal time-frequency magnitude masking for speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. (2019). https://doi.org/10.1109/TASLP.2019.2915167
30. D. Malah, R. Cox, A generalized comb filtering technique for speech enhancement, in *ICASSP’82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7 (IEEE, 1982), pp. 160–163
31. M.K. Marguiles, Male–female differences in speaker intelligibility: normal and hearing-impaired listeners. J. Acoust. Soc. Am. 65(S1), S99–S99 (1979)
32. R. Martin, Spectral subtraction based on minimum statistics, in *European Signal Processing Conference (EUSIPCO)—Proceedings*, pp. 1182–1185 (1994)
33. R. Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. 9(5), 504–512 (2001)
34. K.S.R. Murty, B. Yegnanarayana, Epoch extraction from speech signals. IEEE Trans. Audio Speech Lang. Process. 16(8), 1602–1613 (2008)
35. A. Nehorai, B. Porat, Adaptive comb filtering for harmonic signal enhancement. IEEE Trans. Acoust. Speech Signal Process. 34(5), 1124–1138 (1986)
36. A. Nicolson, K.K. Paliwal, Deep learning for minimum mean-square error approaches to speech enhancement. Speech Commun. 111, 44–55 (2019)
37. A. Nicolson, K.K. Paliwal, Masked multi-head self-attention for causal speech enhancement. Speech Commun. 125, 80–96 (2020)
38. A.M. Noll, Cepstrum pitch determination. J. Acoust. Soc. Am. (1967). https://doi.org/10.1121/1.1910339
39. A. Nicolson, K.K. Paliwal, Masked multi-head self-attention for causal speech enhancement. Speech Commun. 111, 44–55 (2019)
40. A.M. Noll, Cepstrum pitch determination. J. Acoust. Soc. Am. (1967). https://doi.org/10.1121/1.1910339
41. J.C. Príncipe, W. Liu, S. Haykin, Kernel Adaptive Filtering: A Comprehensive Introduction (Wiley, New York, 2011)
42. D. Terpstra, H. Jagode, H. You, J. Dongarra, Collecting performance data with PAPI-C, in Proceedings of the 3rd International Workshop on Parallel Tools for High Performance Computing 2009 (2010). https://doi.org/10.1007/978-3-642-11261-4_11
43. D. Wang, J. Chen, Supervised speech separation based on deep learning: an overview. IEEE/ACM Trans. Audio Speech Lang. Process. 26(10), 1702–1726 (2018)
44. K. Wang, B. He, W.P. Zhu, TSTNN implementation codes. https://github.com/key2miao/TSTNN
59. K. Wang, B. He, W.P. Zhu, TSTNN: two-stage transformer based neural network for speech enhancement in the time domain, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021). https://doi.org/10.1109/icassp39728.2021.9413740

60. Q. Wang, J. Du, L.R. Dai, C.H. Lee, A multiobjective learning and ensembling approach to high-performance speech enhancement with compact neural network architectures. *IEEE/ACM Trans. Audio Speech Lang. Process.* 26(7), 1185–1197 (2018)

61. Y. Wang, A. Narayanan, D. Wang, On training targets for supervised speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* 22(12), 1849–1858 (2014)

62. F. Weninger, J.R. Hershey, J. Le Roux, B. Schuller, Discriminatively trained recurrent neural networks for single-channel speech separation, in *IEEE Global Conference on Signal and Information Processing, GlobalSIP* (2014). https://doi.org/10.1109/GlobalSIP.2014.7032183

63. D.S. Williamson, Y. Wang, D. Wang, Complex ratio masking for monaural speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* 24(3), 483–492 (2015)

64. Y. Xu, J. Du, L.R. Dai, C.H. Lee, An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Process. Lett.* (2014). https://doi.org/10.1109/LSP.2013.2291240

65. L.P. Yang, Q.J. Fu, Spectral subtraction-based speech enhancement for cochlear implant patients in background noise. *J. Acoust. Soc. Am.* (2005). https://doi.org/10.1121/1.1852873

66. B. Yegnanarayana, C. Avendano, H. Hermansky, P.S. Murthy, Speech enhancement using linear prediction residual. *Speech Commun.* 28(1), 25–42 (1999)

67. S.A. Zahorian, H. Hu, A spectral/temporal method for robust fundamental frequency tracking. *J. Acoust. Soc. Am.* (2008). https://doi.org/10.1121/1.2916590

68. L. Zhang, M. Wang, Q. Zhang, X. Wang, M. Liu, PhaseDCN: a phase-enhanced dual-path dilated convolutional network for single-channel speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* 29, 2561–2574 (2021)

69. Q. Zhang, A. Nicolson, M. Wang, K.K. Paliwal, C. Wang, DeepMMSE: a deep learning approach to MMSE-based noise power spectral density estimation. *IEEE/ACM Trans. Audio Speech Lang. Process.* 28, 1404–1415 (2020)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.