Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

HF-SE: End-to-End Speech Enhancement for HF Speech Communications

Ke Xiao
Shanghai University
Kai Liu
liukai@shu.edu.cn

Shanghai University

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11642152/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

HF-SE: End-to-End Speech Enhancement for HF Speech Communications

Ke Xiao[1] and Kai Liu[1*†]

[1] School of Communication and Information Engineering, Shanghai University, Shanghai, China .

*Corresponding author(s). E-mail(s): liukai@shu.edu.cn;
Contributing authors: Xk01@shu.edu.cn;
†These authors contributed equally to this work.

Abstract

High-frequency (HF) speech communication is highly susceptible to channel fading and additive noise, whereas conventional speech enhancement (SE) methods primarily focus on noise suppression. To address this problem, this paper proposes HF-SE, an end-to-end speech enhancement method for HF speech communication based on a Multi-Scale Temporal Dilated Convolutional Network (MS-TDCN). The proposed method uses the in-phase and quadrature components of the received complex baseband IQ signal as dual-channel inputs and extracts speech and signal features to mitigate channel distortions and suppress additive noise, thereby reconstructing clean real-valued speech. MS-TDCN employs multi-scale convolutions to capture local temporal features across different multipath delay windows and exponentially dilated convolutions to model long-range temporal dependencies in the received signal. In addition, a time-Mel (TM) joint loss combines waveform and log-Mel spectral losses to constrain waveform reconstruction and perceptually relevant spectral details. An HF speech dataset is constructed from the TIMIT corpus by simulating HF channel impairments to evaluate the proposed method. Across all nine combinations of channel conditions and SNRs, HF-SE consistently improves speech intelligibility and perceptual quality over the unprocessed signal, increasing the average STOI from 0.583 to 0.773 and PESQ-NB from 1.424 to 2.196.

Keywords: HF speech communication, HF channel, anti-fading, speech enhancement, end-to-end

1 Introduction

High frequency (HF) speech communication relies on ionospheric propagation to achieve long-distance speech transmission. It is a widely adopted communication method in radio broadcasting, emergency relief, military operations, aviation, and amateur radio [1–4]. However, the stochastic dynamics of the ionosphere induce severe doubly-selective fading within the HF channel, resulting in complex signal distortion [5–7]. Furthermore, additive noise introduced during transmission and reception degrades the quality of the received speech. Consequently, HF speech enhancement (SE) is critical for improving speech intelligibility and perceptual quality.

Deep learning has significantly advanced SE by learning mappings from degraded signals to clean speech. While early models such as deep neural networks (DNNs) [8] established the foundation, recent advancements have focused on capturing temporal dependencies and recovering phase information. For instance, multi-stage time-domain convolution networks (TCNs) [9] and complex-valued TCNs [10] have demonstrated effectiveness in modeling time-domain features. Furthermore, architectures incorporating multi-scale convolutions have proven effective in handling diverse noise distributions [11]. However, these methods mainly focus on acoustic noise suppression without considering HF channel impairments.

In HF speech communication, channel effects cannot be ignored. Speech quality is severely degraded by additive noise and channel attenuation and time-frequency dispersion [12, 13]. Ganapathy et al. [14] proposed a frequency offset estimation and correction method based on an autoregressive model. However, this method primarily addresses carrier frequency mismatch in single sideband (SSB) modulation and demodulation. To address fading in HF speech communication, Chen et al. [15] proposed a multi-channel automatic switching method to mitigate speech interruptions caused by deep fading through optimal branch selection under different channel conditions. Chen et al. [16] proposed a two-stage hybrid convolutional neural network (CNN) method to improve the quality of HF SSB speech, demonstrating the effectiveness of CNNs in mitigating channel effects.

Although existing HF signal processing strategies can effectively mitigate channel fading and improve the quality of recovered speech, these methods primarily focus on channel compensation from a communication signal processing perspective, without fully exploiting the temporal and spectral characteristics of speech signals. Moreover, intermediate signal reconstruction does not necessarily lead to optimal speech enhancement performance, as signal recovery and speech enhancement have different optimization objectives. This motivates the investigation of an end-to-end enhancement method that directly maps received complex baseband speech IQ signals to clean speech.

This paper proposes HF-SE, an end-to-end speech enhancement method based on a Multi-Scale Temporal Dilated Convolutional Network (MS-TDCN). The main contributions are summarized as follows:

1. We propose HF-SE. The proposed method uses the in-phase and quadrature components of the received complex baseband IQ signal as dual-channel inputs and

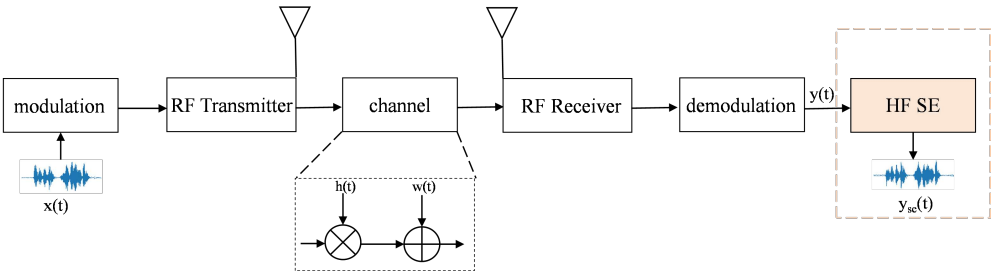


Fig. 1 Simplified block diagram of the HF speech communication system.

extracts speech and signal features to mitigate fading-induced distortions and suppress additive noise.

2. We design a Multi-Scale Temporal Dilated Convolutional Network (MS-TDCN) for HF speech enhancement. The network employs multi-scale convolutions to capture temporal features associated with different multipath delay scales [17] and combines exponentially dilated convolutions to model long-range temporal dependencies in the received HF speech.
3. We employ a time-Mel(TM) joint loss to optimize HF-SE. The loss combines waveform and log-Mel spectral losses to jointly constrain speech waveform reconstruction and perceptually relevant spectral detail recovery.
4. We construct an HF baseband IQ simulation dataset. Based on the TIMIT [18] speech corpus, clean speech is converted into IQ signals, and degraded signals are generated by incorporating pilot signals, Watterson channels [19], and complex Gaussian white noise. The dataset uses speaker-exclusive partitioning and covers three channel conditions (good, moderate, and poor), with seven SNR levels used for training.

The remainder of this paper is organized as follows. Section 2 discusses the HF speech enhancement problem. Section 3 describes the proposed HF-SE method. Section 4 presents the experimental results, and Section 5 concludes the paper.

2 Problem Analysis

In practical HF communication systems, such as those used for emergency rescue, military, and aviation applications, speech signals are commonly transmitted using single sideband (SSB) modulation over an HF ionospheric channel. A simplified block diagram of an HF speech communication system is illustrated in Figure 1. During propagation, the transmitted signal is affected by multipath propagation, Doppler spread, time-varying fading, and additive noise. The approximate altitude ranges of the ionospheric layers are illustrated in Figure 2.

Therefore, the HF speech communication process can be modeled as follows. Let $x[n]$ denote the real-valued clean speech signal. The speech signal is converted into a

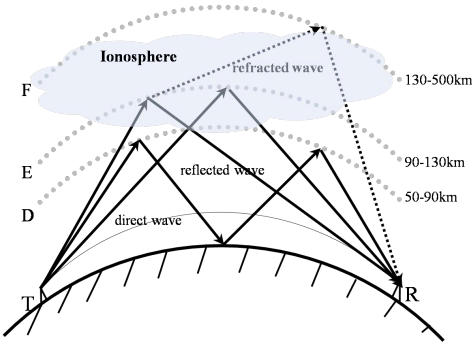


Fig. 2 The approximate altitude ranges of the ionospheric D, E, and F layers (50–90 km, 90–130 km, and 130–500 km, respectively).

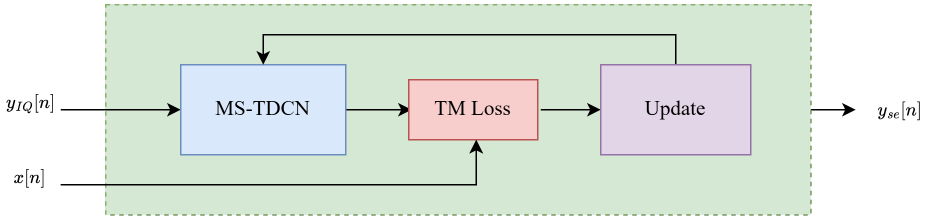


Fig. 3 Overall architecture and training procedure of the proposed HF-SE method.

IQ representation through a modulation process:

$$x_{\mathrm{IQ}}[n] = \mathrm{Modulation}\{x[n]\}, \tag{1}$$

where Modulation$\{\cdot\}$ denotes the analog modulation process.

After propagation through the HF channel, the received IQ signal can be expressed as:

$$y_{\mathrm{IQ}}[n] = \sum_{i=1}^{N} h_i[n] x_{\mathrm{IQ}}[n - d_i] + w[n], \tag{2}$$

where N is the number of propagation paths, $h_i[n]$ and d_i denote the channel response and discrete delay of the i-th path, respectively,and $w[n]$ denotes complex additive white Gaussian noise. Consequently, the received IQ signal is degraded by multiplicative interference and additive noise.

3 Proposed Method

The proposed HF-SE employs MS-TDCN to map the received complex baseband IQ signal to a clean real-valued speech waveform. The enhancement process is formulated as:

$$y_{\mathrm{se}}[n] = f_{\boldsymbol{\theta}}\left([\mathrm{Re}(y_{\mathrm{IQ}}), \mathrm{Im}(y_{\mathrm{IQ}})]\right), \tag{3}$$

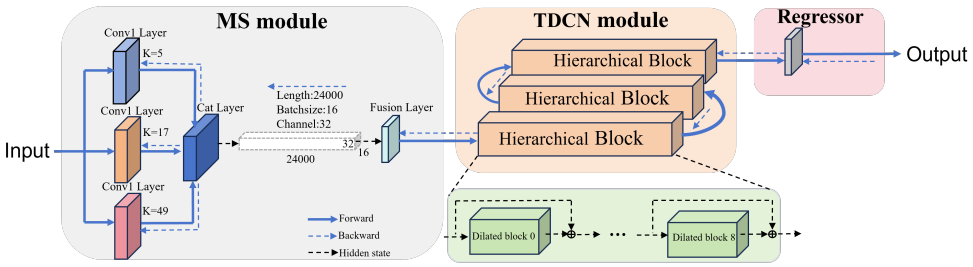


Fig. 4 Architecture of the proposed MS-TDCN model: (i) Multi-scale Feature Extractor; (ii) Temporal Dilated Convolution Network; (iii) Regressor.

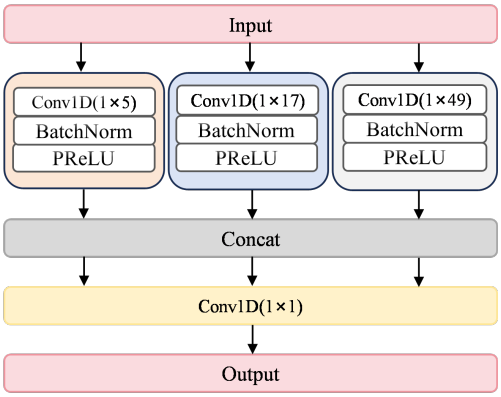


Fig. 5 The structure of MS module.

where $f_{\boldsymbol{\theta}}(\cdot)$ denotes the proposed MS-TDCN network and $y_{\text{se}}[n]$ represents the enhanced real-valued speech signal.

HF-SE employs MS-TDCN as its backbone network. Figure 3 presents the overall end-to-end method, while Figure 4 illustrates the structure of MS-TDCN. The network adopts a convolution-based design to efficiently capture temporal dependencies while benefiting from weight sharing and parallel computation [20–22].

Subsection 3.1 introduces the multi-scale feature extraction module, Subsection 3.2 describes the temporal dilated convolution network, and Subsection 3.3 presents the time-Mel joint loss.

3.1 Multi-Scale Feature Extraction Module

As illustrated in Figure 5, the MS module consists of three parallel convolutional branches with kernel sizes of 1×5, 1×17, and 1×49, respectively. Under the 8-kHz sampling rate used in this work, the temporal spans between the first and last samples of these convolutional kernels are 0.5, 2, and 6 ms, respectively. These kernel sizes are selected to cover the representative multipath delays of the three simulated HF channel conditions, allowing the MS module to capture local speech-related temporal

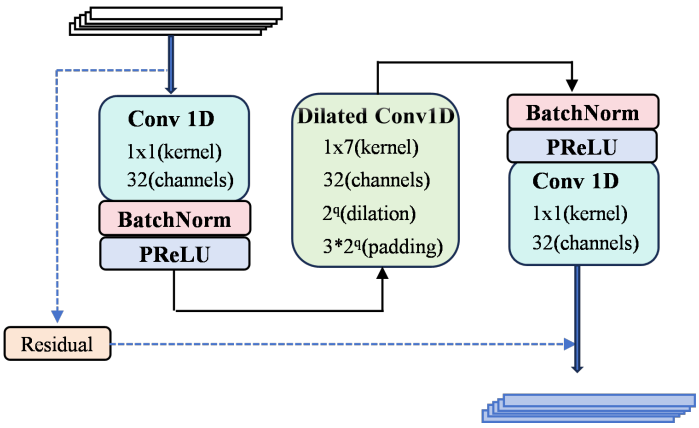


Fig. 6 The structure of dilated block.

variations within different delay windows and extract complementary features across multiple temporal scales.

Each branch extracts temporal features at a different scale through a convolutional block consisting of one-dimensional(1D) convolution, BatchNorm and a PReLU activation function. The resulting multi-scale features are concatenated along the channel dimension and then fused through a 1×1 convolution. This fusion operation integrates complementary short, medium, and long scale information and maps the concatenated features into a unified representation for the subsequent TDCN module.

3.2 Temporal Dilated Convolution Network

Based on the fusion features output by the MS feature extraction module, this paper designs a TDCN module to further capture the speech long-range temporal dependence. This module expands the receptive field layer by layer through the cascade structure to strengthen the modeling ability of the deep characteristics of the HF speech.

3.2.1 Dilated block

Deep fading in HF channels can cause interruptions in continuous speech segments, resulting in the degradation of speech information. Since the MS module primarily extracts local temporal features associated with different multipath delay scales, its limited receptive field may be insufficient to reconstruct severely degraded speech segments. Therefore, we introduce the TDCN to capture long-range temporal dependencies in speech-related features extracted from the received IQ signal, while modeling temporal variations associated with HF channel impairments through an exponentially expanding receptive field.

The architecture of the TDCN is illustrated in Figure 6. Each dilated block consists of a pointwise 1D convolution followed by a dilated 1D convolution. Each convolutional layer is followed by BatchNorm and a PReLU activation function to enhance the nonlinear representation capability of the network and facilitate training. We adopt a

constant channel width of 32 throughout the TDCN to avoid frequent channel expansion and compression, thereby controlling computational complexity. Specifically, the dilated convolution kernel size is set to 1×7, with the dilation rate d increasing exponentially with the layer depth q, i.e., $d = 2^q$. The kernel size is empirically selected based on experimental comparisons.

Additionally, residual connections are incorporated within each dilated block to facilitate cross-layer feature propagation and improve gradient flow. These connections help preserve previously extracted speech-related features during temporal modeling, supporting stable reconstruction of speech waveforms affected by HF channel impairments.

3.2.2 Hierarchical block

Within each hierarchical block, an exponentially increasing dilation factor is adopted to progressively expand the receptive field. Subsequently, the dilation factor undergoes a periodic reset between blocks. The specific process is formulated as follows:

$$d_q^{(b)} = 2^q, \quad q = 0, \ldots, L-1, \tag{4}$$

where 2^q denotes the dilation rate for each dilated block, L represents the total number of layers within a hierarchical block, and b indicates the index of the hierarchical block.

This segmented dilation strategy prevents the excessive expansion of the receptive field caused by indefinitely increasing dilation rates, thereby preserving the sensitivity to local details. More importantly, by cyclically resetting the dilation rates across different levels, it endows the model with a dynamic receptive field, enabling the simultaneous capture of both short-term and long-term features.

3.3 Time-Mel Joint Loss

The proposed HF-SE is optimized using a TM joint loss to directly reconstruct clean speech from the received HF IQ signal. The loss jointly constrains speech reconstruction in the time and Mel domains. The loss function is defined as follows:

$$\mathcal{L}_{\mathrm{TM}} = \alpha \mathcal{L}_{\mathrm{Wav}} + \gamma \mathcal{L}_{\mathrm{Mel}}, \tag{5}$$

where α and γ are the weighting factors for the waveform loss and Mel loss, respectively. The values of α and γ are set to 0.7 and 1.0, respectively.

The time-domain loss $\mathcal{L}_{\mathrm{Wav}}$ directly calculates the error between the enhanced waveform and the clean waveform. We employ the mean squared error (MSE) to constrain waveform reconstruction.

$$\mathcal{L}_{\mathrm{Wav}} = \frac{1}{T} \sum_{n=1}^{T} (y_{se}[n] - x[n])^2, \tag{6}$$

where $y_{se}[n]$ and $x[n]$ represent the samples of the enhanced and clean speech, respectively, and T denotes the number of waveform samples.

Table 1 The typical Watterson channel parameters for two paths in low latitudes.

Channel conditions	Delay (ms)	Doppler spread (Hz)
good	0.5	0.5
moderate	2	1.5
poor	6	10

Although the time-domain loss constrains waveform reconstruction, relying solely on MSE may lead to insufficient recovery of spectral details. To complement the time-domain loss, we introduce the Mel-domain loss $\mathcal{L}_{\text{Mel}}$. The Mel scale reflects the nonuniform frequency resolution of human auditory perception, providing finer frequency resolution at lower frequencies. Therefore, the Mel loss encourages the network to preserve perceptually important spectral details during speech reconstruction. The Mel loss is defined as follows:

$$\mathcal{L}_{\text{Mel}} = \frac{1}{M \cdot L} \sum_{k=1}^{M} \sum_{l=1}^{L} \left| \Phi(k,l) - \hat{\Phi}(k,l) \right|, \tag{7}$$
$$\text{where} \quad \Phi(k,l) = \log(\text{Mel}(x[n]) + \epsilon),$$

where $\Phi(k,l)$ and $\hat{\Phi}(k,l)$ denote the log-Mel spectra of the clean and enhanced speech obtained via the Mel spectral transformation, respectively. M represents the number of Mel frequency bands, L denotes the number of time frames, and $\epsilon = 10^{-5}$ is a small constant for numerical stability.

4 Experimental evaluation

4.1 Dataset generation

The Watterson model is a classical narrowband HF channel model. It generally simulates two independent fading paths, and ITU-R specifies the parameters of the path under different ionospheric regions and channel conditions, as shown in Table 1.

The TIMIT corpus was used to construct the HF speech dataset. The dataset is generated under the assumption of ideal carrier synchronization, without additional carrier frequency offsets or phase synchronization errors. The training, validation, and test sets contained 4,160, 460, and 1,680 clean utterances from 416, 46, and 168 speakers, respectively. The original 16-kHz speech signals were downsampled to 8 kHz, band-pass filtered to 300-3400 Hz, and level-normalized. Training utterances were cropped or zero-padded to 3 s, whereas validation and test utterances retained their full lengths. Each preprocessed waveform was converted into an complex baseband IQ signal, and two continuous pilot tones at 125 Hz and 3750 Hz were added. The signal was transmitted through a time-varying Watterson channel, and complex additive white Gaussian noise was added. The dataset contains three channel conditions(good, moderate, and poor). The training and validation sets used SNRs from -2 to 10 dB

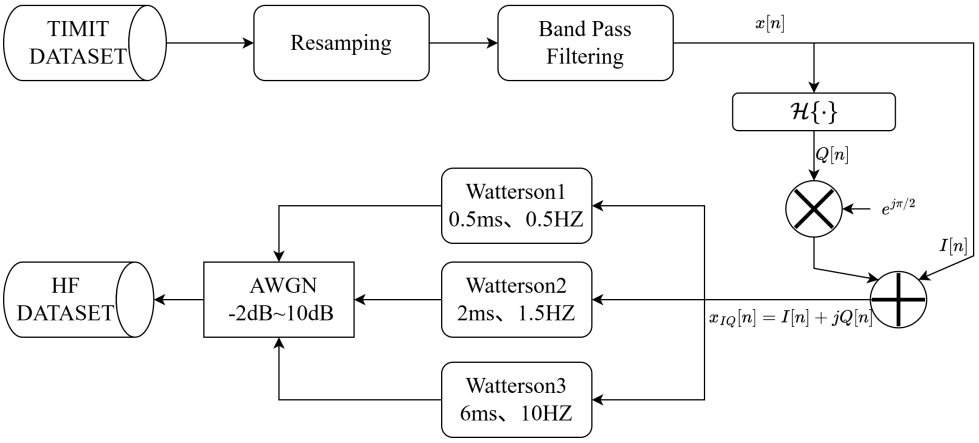


Fig. 7 The diagram of the HF speech dataset generation.

with a 2 dB interval, while the test set used 0, 5, and 10 dB; the 5 dB condition was not seen during training. The data generation process is shown in Figure 7.

4.2 Experimental setup

During training, we employed a distributed data parallel (DDP) strategy using a single-node dual-GPU configuration. The experimental environment is detailed in Table 2. The proposed HF-SE was trained end-to-end using paired received IQ signals and clean speech waveforms. The network was optimized using AdamW with an initial learning rate of 1×10^{-3}, a weight decay of 1×10^{-4}, and a maximum of 100 epochs. Gradient clipping, a cosine learning-rate schedule, mixed-precision training, and early stopping were employed. The batch size was set to 32.

Table 2 Software and hardware environment configuration.

Experimental Environment	Relevant Parameters
Operating system	Ubuntu 22.04
CPU	Intel(R) Xeon(R) CPU E5-2683 v3 @ 2.00GHz
GPU0	NVIDIA GeForce RTX 4080 SUPER (16G)
GPU1	NVIDIA GeForce RTX 4080 SUPER (16G)
CUDA	CUDA 11.8
Programming language	Python 3.10
Deep learning framework	PyTorch 2.7.1

Table 3 End-to-end reference-based speech quality performance of the proposed HF-SE under different channel conditions and SNRs. Δ denotes the performance gain.

Channel	SNR (dB)	STOI			PESQ-NB		
		Unprocessed	HF-SE	Δ	Unprocessed	HF-SE	Δ
Good	0	0.531	0.735	+0.204	1.287	1.911	+0.624
	5	0.604	0.800	+0.196	1.416	2.234	+0.818
	10	0.674	0.845	+0.171	1.608	2.559	+0.951
Moderate	0	0.516	0.713	+0.197	1.284	1.890	+0.606
	5	0.587	0.780	+0.193	1.397	2.202	+0.805
	10	0.652	0.827	+0.175	1.602	2.561	+0.959
Poor	0	0.503	0.702	+0.199	1.278	1.867	+0.589
	5	0.570	0.761	+0.191	1.384	2.138	+0.754
	10	0.614	0.790	+0.176	1.564	2.399	+0.835

4.3 Experimental Results

To evaluate the proposed HF-SE, Perceptual Evaluation of Speech Quality in narrowband mode (PESQ-NB) [23] and Short-Time Objective Intelligibility (STOI) [24] were adopted as reference-based objective metrics. PESQ-NB measures perceptual speech quality, with scores typically ranging from -0.5 to 4.5, whereas STOI measures speech intelligibility on a scale from 0 to 1. Although PESQ-NB and STOI provide reference-based evaluations when clean speech is available, practical HF receivers do not have access to the transmitted clean waveform. Therefore, we further adopt the local non-personalized DNSMOS P.835 model [25] for non-intrusive perceptual quality assessment. Higher scores indicate better speech intelligibility or perceptual quality.

4.3.1 Overall performance

The overall performance of the proposed HF-SE is reported in Table 3. In this comparison, Unprocessed refers to the real-valued speech signal recovered after minimum mean square error (MMSE) channel equalization. HF-SE consistently improves both STOI and PESQ-NB compared with the unprocessed baseline across all nine combinations of channel conditions and SNRs. Averaged over all conditions, STOI increases from 0.583 to 0.773, corresponding to an absolute gain of 0.189, while PESQ-NB increases from 1.424 to 2.196, corresponding to an average gain of 0.771. These results demonstrate the effectiveness of HF-SE in improving HF speech intelligibility and perceptual quality under different HF channel conditions.

As shown in Table 4, the proposed HF-SE consistently improves the DNSMOS P.835 background noise quality (BAK) and overall quality (OVRL) scores over the unprocessed baseline across all channel conditions and SNRs. Averaged over all conditions, BAK increases from 1.758 to 3.570, while OVRL increases from 1.645 to 2.421. The improvement is particularly pronounced in BAK, indicating the effectiveness of HF-SE in suppressing background noise. For speech signal quality (SIG), the average score increases from 2.519 to 2.799, with improvements observed under all Poor channel conditions and at 0 dB under the Good and Moderate channel conditions.

Table 4 Non-intrusive DNSMOS P.835 evaluation of the unprocessed baseline and the enhanced speech produced by HF-SE. Δ denotes the performance gain.

Channel	SNR (dB)	SIG			BAK			OVRL		
		Unprocessed	HF-SE	Δ	Unprocessed	HF-SE	Δ	Unprocessed	HF-SE	Δ
Good	0	2.401	2.784	+0.383	1.568	3.462	+1.894	1.503	2.361	+0.858
	5	2.997	2.959	-0.038	1.928	3.611	+1.683	1.850	2.557	+0.707
	10	3.236	3.123	-0.113	2.191	3.739	+1.548	2.094	2.750	+0.656
Moderate	0	2.240	2.695	+0.455	1.505	3.407	+1.902	1.448	2.269	+0.821
	5	2.913	2.900	-0.013	1.831	3.576	+1.745	1.763	2.495	+0.732
	10	3.141	3.069	-0.072	2.168	3.714	+1.546	2.034	2.693	+0.659
Poor	0	1.590	2.381	+0.791	1.281	3.439	+2.158	1.231	2.051	+0.820
	5	2.071	2.599	+0.528	1.576	3.546	+1.970	1.420	2.259	+0.839
	10	2.079	2.683	+0.604	1.772	3.633	+1.861	1.465	2.356	+0.891

However, slight decreases occur at higher SNRs under the Good and Moderate channel conditions. These results indicate that HF-SE can improve speech signal quality under severe channel fading and low-SNR conditions, while consistently enhancing overall perceptual quality across all evaluated HF channel conditions.

4.3.2 Comparison with SE Baselines

To evaluate the effectiveness of the proposed HF-SE under HF channel impairments, we compare it with three real-valued speech enhancement methods: SE-only, GTCRN [26], and AECNN [27]. All baseline models are trained using the MMSE equalized signal as input and the corresponding clean speech as the target, with their respective optimized training configurations. SE-only employs the same MS-TDCN backbone as HF-SE, while GTCRN and AECNN are existing deep learning-based speech enhancement models. In contrast, HF-SE directly utilizes the in-phase and quadrature components of the received IQ signal to reconstruct clean real-valued speech without MMSE channel equalization.

As shown in Figure 8, although all three real-valued speech enhancement methods improve STOI and PESQ-NB, the proposed HF-SE consistently outperforms them in both metrics across all nine combinations of HF channel conditions and SNRs. Averaged over all conditions, HF-SE achieves STOI and PESQ-NB scores of 0.773 and 2.196, respectively, compared with 0.724 and 1.987 for SE-only, 0.667 and 1.953 for GTCRN, and 0.679 and 1.793 for AECNN. These results demonstrate that, compared with the evaluated real-valued speech enhancement methods, HF-SE can mitigate channel-induced distortions and additive noise without requiring conventional explicit channel equalization, thereby significantly improving the intelligibility and quality of HF speech.

4.3.3 Comparison of SE Strategies

To further evaluate the proposed end-to-end enhancement strategy, we compare HF-SE with two extended processing methods: MMSE+HF-SE and HF-SE+SE-only. In MMSE+HF-SE, the received IQ signal is first processed by MMSE channel equalization, and the resulting complex-valued IQ signal is used to train an HF-SE model with the corresponding clean speech as the target. In HF-SE+SE-only, the pretrained

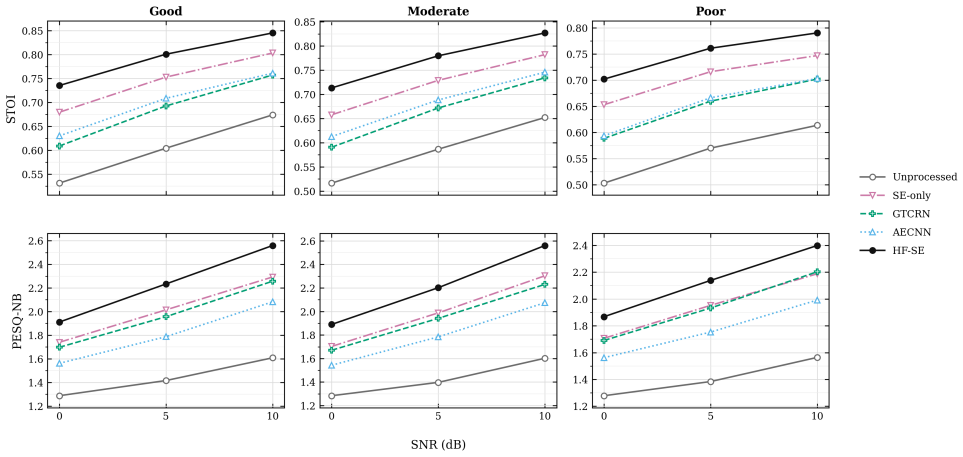


Fig. 8 Comparison of the unprocessed baseline, SE-only, GTCRN, AECNN, and the proposed HF-SE in terms of STOI and PESQ-NB under different HF channel conditions and SNRs.

Table 5 Effect of additional processing stages on the proposed HF-SE.Overall represents the macro-average across all nine channel and SNR combinations.

Method	$\overline{\text{STOI}}$				$\overline{\text{PESQ-NB}}$			
	Good	Moderate	Poor	Overall	Good	Moderate	Poor	Overall
HF-SE	**0.794**	**0.773**	**0.751**	**0.773**	**2.234**	**2.217**	**2.135**	**2.196**
MMSE+HF-SE	0.759	0.742	0.723	0.742	2.003	2.042	1.953	2.000
HF-SE+SE-only	0.778	0.759	0.739	0.759	2.228	2.205	2.126	2.186

HF-SE model first reconstructs a real-valued speech waveform from the received IQ signal, which is subsequently used as the input to train the SE-only model. During this stage, the parameters of HF-SE are frozen, while SE-only is optimized using the corresponding clean speech as the target. These comparisons investigate whether additional processing stages can further improve speech reconstruction under HF channel impairments.

As shown in Table 5, HF-SE achieves the highest average STOI and PESQ-NB scores across all three HF channel conditions. Compared with MMSE+HF-SE, HF-SE improves the overall STOI from 0.742 to 0.773 and PESQ-NB from 2.000 to 2.196, corresponding to absolute gains of 0.031 and 0.196, respectively. Furthermore, HF-SE outperforms HF-SE+SE-only in both metrics across all three channel conditions. These results suggest that HF-SE can effectively mitigate channel-induced distortions and additive noise through direct end-to-end speech enhancement from received HF IQ signals, achieving the best performance among the evaluated strategies without requiring additional processing stages.

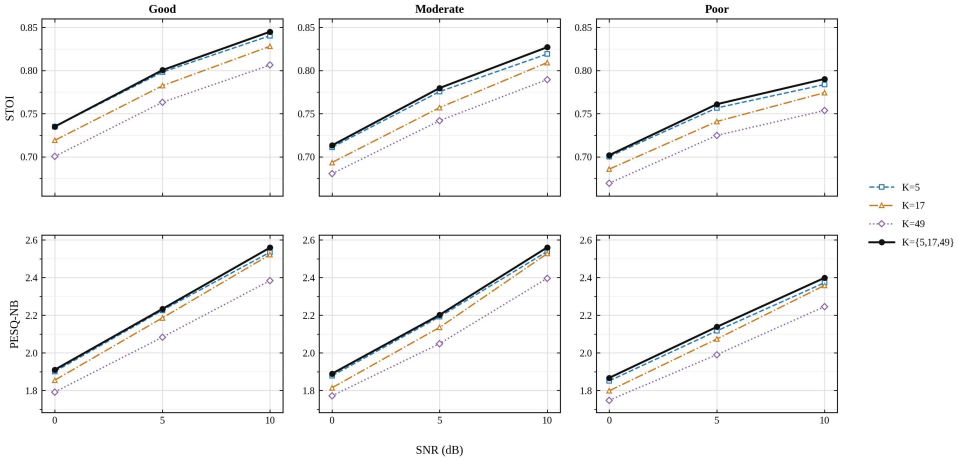


Fig. 9 Ablation results of different convolution kernel configurations under three HF channel conditions.

Table 6 Ablation study of the proposed TM joint loss under different HF channel conditions. Results are averaged over three SNR levels (0, 5, and 10 dB).

Method	$\overline{\text{STOI}}$			$\overline{\text{PESQ-NB}}$		
	Good	Moderate	Poor	Good	Moderate	Poor
Unprocessed	0.603	0.585	0.562	1.437	1.428	1.409
$\mathcal{L}_{\text{Wav}}$	0.626	0.611	0.584	1.765	1.730	1.569
$\mathcal{L}_{\text{Wav}} + \mathcal{L}_{\text{Mel}}$	**0.794**	**0.773**	**0.751**	**2.234**	**2.217**	**2.135**

4.3.4 Ablation Study

To evaluate the effectiveness of the proposed MS module in capturing temporal features associated with different multipath delay scales, an ablation study is conducted using three single-scale convolutional configurations with kernel sizes of $K = 5$, $K = 17$, and $K = 49$.

As shown in Figure 9, the multi-scale configuration consistently achieves higher STOI and PESQ-NB scores than all three single-scale configurations across the nine combinations of channel condition and SNR.These results demonstrate the effectiveness of the proposed MS module in extracting temporal features associated with different multipath delay scales and integrating complementary information from different temporal receptive fields.

To evaluate the effectiveness of the proposed TM joint loss, an ablation study is conducted by comparing two loss configurations: the waveform loss $\mathcal{L}_{\text{Wav}}$ alone and its combination with the Mel-domain loss $\mathcal{L}_{\text{Mel}}$. The unprocessed baseline is included as a reference.

Table 7 Computational complexity of SE-only, GTCRN, AECNN, and HF-SE for a 3-s, 8-kHz utterance with batch size 1.

Method	Parameters (M)	MACs (G)
SE-only	0.254	6.015
GTCRN	**0.048**	**0.040**
AECNN	6.312	43.776
HF-SE	0.255	6.032

As shown in Table 6, using $\mathcal{L}_{\text{Wav}}$ alone improves the average STOI and PESQ-NB scores over the unprocessed baseline under all three HF channel conditions. Incorporating $\mathcal{L}_{\text{Mel}}$ further improves both metrics across all evaluated channel conditions. Averaged over all nine combinations of channel conditions and SNRs, STOI increases from 0.607 to 0.773, corresponding to an absolute gain of 0.166, while PESQ-NB increases from 1.688 to 2.196, corresponding to a gain of 0.508. These results demonstrate that the proposed TM joint loss effectively constrains speech reconstruction in both the waveform and Mel spectral domains, thereby improving the intelligibility and perceptual quality of the HF speech.

4.3.5 Computational complexity

Table 7 compares the computational complexity of SE-only, GTCRN, AECNN, and the proposed HF-SE. Among these methods, GTCRN has the lowest parameter count and MACs, while AECNN has the highest computational complexity. HF-SE requires 0.255 M parameters and 6.032 G MACs. Despite its higher computational cost than GTCRN, HF-SE achieves the best STOI and PESQ-NB performance across all evaluated HF channel conditions. Future work will focus on reducing the computational complexity of HF-SE while maintaining its speech enhancement performance.

5 Conclusion

This paper proposed HF-SE, an end-to-end speech enhancement method to address speech degradation caused by channel fading and additive noise in HF speech communications. HF-SE directly processes the in-phase and quadrature components of the received complex baseband IQ signal to reconstruct clean real-valued speech. The proposed method employs MS-TDCN as the backbone network, where multi-scale convolutions extract local temporal features associated with different multipath delay scales, while exponentially dilated convolutions capture long-range temporal dependencies in speech-related features. A time-Mel (TM) joint loss is employed to guide speech reconstruction in both the waveform and Mel spectral domains. Experimental results across nine combinations of HF channel conditions and SNRs show that HF-SE improves the average STOI from 0.583 to 0.773 and PESQ-NB from 1.424 to 2.196. HF-SE also consistently outperforms the evaluated real-valued speech enhancement methods in both metrics. These results demonstrate the effectiveness of directly

utilizing received HF IQ signals to mitigate channel-induced distortions and additive noise for end-to-end speech enhancement under the evaluated HF channel conditions.

Declarations

- No funding was received for conducting this study.
- Data will be made available on request

References

[1] Kallberg, J., Hamilton, S.: Resiliency by retrograded communication-the revival of shortwave as a military communication channel it professional. IT Professional **22**, 47–51 (2020)

[2] Shen, C., Lyu, T., Li, Y., Lin, T., Liu, Y.: Shortwave signal modulation recognition method using adaptive time-frequency threshold denoising and feature fusion. Physical Communication **75**, 102994 (2026)

[3] Bingjian, S., Yuming, B., Xiaojun, X., Jicheng, W., Jiaping, Q.: Visual analysis of emergency communication research based on scientific knowledge graph. In: 2024 International Applied Computational Electromagnetics Society Symposium (ACES-China), pp. 1–6 (2024)

[4] Wang, J., Shi, Y., Yang, C., Feng, F.: A review and prospects of operational frequency selecting techniques for hf radio communication. Advances in Space Research **69**(8), 2989–2999 (2022)

[5] Feng, B., Si, L., Yin, H.: Research on single carrier frequency domain equalization in wideband high frequency communication system. In: 2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP), pp. 1956–1960 (2023)

[6] LI, Y., XU, Y., WANG, H., DING, G., Gu, J.: Hf skywave communication for air-to-ground networks: A cooperative transmission framework. Chinese Journal of Aeronautics **38**(5), 103269 (2025)

[7] Wang, J., Shi, Y., Yang, C., Zhang, Z., Zhao, L.: A short-term forecast method of maximum usable frequency for hf communication. IEEE Transactions on Antennas and Propagation **71**(6), 5189–5198 (2023)

[8] Xu, Y., Du, J., Dai, L.-R., Lee, C.-H.: An experimental study on speech enhancement based on deep neural networks. IEEE Signal Processing Letters **21**(1), 65–68 (2014)

[9] Lin, J., Wijngaarden, A.J.d.L., Wang, K.-C., Smith, M.C.: Speech enhancement using multi-stage self-attentive temporal convolutional networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3440–3450 (2021)

[10] Song, J., Zou, L., Zhou, L., Liu, Z., Fan, C., Wang, B.: Complex-valued temporal convolutional network for speech enhancement. International Journal of Wavelets, Multiresolution and Information Processing **22**(05), 2450015 (2024)

[11] Parisae, V., Bhavanam, D., Midasala, V.: Progressive learning framework for speech enhancement using multi-scale convolution and s-tcn, pp. 83–89 (2024)

[12] Hervás, M., Bergadà, P., Pages, R.M.A.: Ionospheric narrowband and wideband hf soundings for communications purposes: A review. Sensors (Basel, Switzerland) **20** (2020)

[13] Uysal, M., Heidarpour, M.R.: Cooperative communication techniques for future-generation hf radios. IEEE Communications Magazine **50**(10), 56–63 (2012)

[14] Ganapathy, S., Pelecanos, J.: Enhancing frequency shifted speech signals in single side-band communication. IEEE Signal Processing Letters **20**(12), 1231–1234 (2013)

[15] Chen, Y., Wang, Y., Yang, J., Gu, Y.: Automatic switching technique for speech enhancement in hf communications. In: 2018 8th International Conference on Electronics Information and Emergency Communication (ICEIEC), pp. 1–6 (2018)

[16] Chen, Y., Dong, B., Zhang, X., Gao, P., Li, S.: A hybrid deep-learning approach for single channel hf-ssb speech enhancement. IEEE Wireless Communications Letters **10**(10), 2165–2169 (2021)

[17] ITU-R Recommendation F.1487: Testing of HF modems with bandwidths of up to about 12 khz using ionospheric channel simulators. Recommendation, International Telecommunication Union, Geneva (2000)

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

[19] Watterson, C., Juroshek, J., Bensema, W.: Experimental confirmation of an hf channel model. IEEE Transactions on Communication Technology **18**(6), 792–803 (1970)

[20] Bai, S., Kolter, J.Z., Koltun, V.: An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling (2018)

[21] Strake, M., Defraene, B., Fluyt, K., Tirry, W., Fingscheidt, T.: Fully convolutional recurrent networks for speech enhancement. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6674–6678 (2020)

[22] Gu, L., Chen, T., Liu, K.: Hf-yolo: A model for high frequency signal detection and recognition in wideband: L. gu et al. Signal, Image and Video Processing **20**(3), 138 (2026)

[23] Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P.: Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), vol. 2, pp. 749–7522 (2001)

[24] Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J.: An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Transactions on Audio, Speech, and Language Processing **19**(7), 2125–2136 (2011)

[25] Reddy, C.K.A., Gopal, V., Cutler, R.: DNSMOS P.835: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors (2022). https://arxiv.org/abs/2110.01763

[26] Rong, X., Sun, T., Zhang, X., Hu, Y., Zhu, C., Lu, J.: Gtcrn: A speech enhancement model requiring ultralow computational resources. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 971–975 (2024). https://doi.org/10.1109/ICASSP48485.2024.10448310

[27] Pandey, A., Wang, D.: A new framework for cnn-based speech enhancement in the time domain. IEEE/ACM Trans. Audio, Speech and Lang. Proc. **27**(7), 1179–1188 (2019) https://doi.org/10.1109/TASLP.2019.2913512