Current Science
CS 5 (2), 1422 - 1431 (2025)
Received 08 **Dec** 2024 | **Revised 18 Jan** 2025 | **Accepted 15 Feb**. 2025 | **Online Available 6 March** 2025
https://doi.org/ 10.52845/CS/2025-5-2-10

Current Science

Current Science

OPEN ACCESS

ISSN (O) 2795-8639

CASE REPORT

A Study on a Two-Stage UAV Noise Removal Method Based on Deep Residual Neural Networks

Yang Lei[1], Tian Liqin[1], Wu Junyi[2]

[1]School of Computer, North China Institute of Science and Technology, Yan jiao, 065201, China

[2]School of Computer, Qinghai Normal University, Xining, 810016, China

[*]Corresponding Author: Yang Lei

Abstract

To improve speech intelligibility in UAV noise environments, this paper proposes a two-stage UAV noise removal method based on deep residual neural networks (DRNN). In the training stage, a DRNN with four residual blocks is employed to estimate the spectral gain function. In the enhancement stage, the estimated spectral gain function is applied to the noisy speech to obtain the enhanced speech signal. Comparative experimental results demonstrate that, on both the publicly available TIMIT speech dataset and a self-constructed UAV noise dataset, the proposed method consistently achieves higher average PESQ scores across all tested signal-to-noise ratio (SNR) conditions. This indicates that the enhanced speech produced by the proposed method is perceptually closer to the ideal quality as perceived by the human auditory system. Compared to traditional speech enhancement methods based on mask estimation and spectral mapping, the proposed approach more effectively removes UAV noise and improves speech quality.

Keywords: Deep Residual Neural Network; UAV Noise; Spectral Gain Function; Speech Enhancement

Introduction

In recent years, with the widespread application of unmanned aerial vehicles (UAVs) in emergency rescue operations, UAV noise has emerged as a critical interference factor during life-saving missions[1]. On the one hand, when UAV-mounted audio acquisition devices are used to capture distress calls from trapped individuals, UAV noise often masks or overwhelms these weak life signals. On the other hand, during air-to-ground communication between UAVs and trapped individuals, UAV noise can cause audio signal distortion, reduced speech intelligibility, or even complete loss of recognition. Therefore, researching speech enhancement methods for UAV noise suppression is of paramount importance.

Traditional monaural speech enhancement techniques rely on specific prior assumptions to model the separation of speech and noise signals. Spectral Subtraction[2] is prone to introducing "musical noise," and its performance heavily depends on the accuracy of noise estimation. Wiener Filtering[3] alleviates musical noise to some extent but still relies on an accurate noise model. The Decision-Directed (DD) approach[4] effectively suppresses musical noise; however, it introduces frame delays, which affect real-time performance. The Two-Step Noise Reduction (TSNR) method[5] reduces frame delay but struggles to effectively suppress high-intensity UAV noise in low-SNR environments.

Deep learning-based speech enhancement has garnered significant attention due to its strong capability in modeling complex nonlinear mappings[6]. Neural networks can extract more complex and higher-level features by increasing the depth of the network. However, when the network becomes excessively deep, it may suffer from gradient vanishing, gradient explosion, and

network degradation issues, which make training challenging[7]. Xu Y. et al. employed a deep neural network (DNN)[8] to extract speech features, using a five-layer architecture consisting of one input layer, three hidden layers, and one output layer. Kounovsky T. et al. utilized a convolutional neural network (CNN)[9] with a seven-layer architecture comprising one input layer, two convolutional layers, one pooling layer, two fully connected layers, and one output layer. However, both studies primarily focused on shallow neural networks for speech feature extraction.

This paper focuses on speech enhancement methods for UAV noise removal. To effectively eliminate UAV noise interference and extract richer speech features while addressing the challenges of gradient vanishing, gradient explosion, and network degradation as the network depth increases, we propose a two-stage UAV noise removal method based on deep residual neural network (DRNN).

1. Signal Model and Problem Formulation

Monaural speech enhancement refers to processing speech signals captured by a single microphone. The corresponding signal model can be expressed as:

$$y(t) = s(t) + v(t) \#(1)$$

where $y(t)$represents the noisy speech signal at time t, which is a mixture of the clean speech signal $s(t)$ and the noise signal $v(t)$ at the same time instant. It is assumed that $s(t)$ and $v(t)$ are mutually uncorrelated and both follow a zero-mean symmetric Gaussian distribution. By applying the Short-Time Fourier Transform (STFT), the time-domain signal can be transformed into the frequency domain, leading to the following signal model:

$$Y(p,k) = S(p,k) + V(p,k) \#(2)$$

where $Y(p,k)$, $S(p,k)$, and $V(p,k)$ represent the complex spectra of $y(t)$, $s(t)$, and $v(t)$ at the p-th frame and k-th frequency bin, respectively.

The goal of speech enhancement is to remove background noise from the speech signal. This is achieved by filtering out the noise spectrum from the complex spectrum of the noisy speech, obtaining the enhanced speech spectrum, and reconstructing the enhanced time-domain speech signal using the Inverse Short-Time Fourier Transform (ISTFT)[10]. The estimated clean speech spectrum $\hat{S}(p,k)$ is typically obtained by applying a nonlinear spectral gain function $G(p,k)$ to the noisy speech spectrum:

$$\hat{S}(p,k) = G(p,k)Y(p,k) \#(3)$$

where $G(p,k)$ attenuates the noise components in $Y(p,k)$ to extract $\hat{S}(p,k)$. In general, the spectral gain function $G(p,k)$ is a function of both the a priori and a posteriori Signal-to-Noise Ratios (SNRs)[11]:

$$G(p,k) = f\left(\xi_{(p,k)}, \gamma_{(p,k)}\right) \#(4)$$

where $\xi_{(p,k)}$ and $\gamma_{(p,k)}$ represent the a priori SNR and a posteriori SNR, respectively, and are defined as follows:

$$\xi_{(p,k)} = \frac{E\{|S(p,k)|^2\}}{E\{|V(p,k)|^2\}} = \frac{\lambda_S(p,k)}{\lambda_V(p,k)} \#(5)$$

$$\gamma_{(p,k)} = \frac{|Y(p,k)|^2}{\lambda_V(p,k)} \#(6)$$

where $|S(p,k)|^2$, $|V(p,k)|^2$, and $|Y(p,k)|^2$ denote the power spectra of the clean speech signal, noise

signal, and noisy speech signal, respectively. In the case of Wiener filtering, the spectral gain function depends only on the a priori SNR. Without loss of generality, it can be expressed as:

$$G(p,k) = \frac{\xi_{(p,k)}}{1+\xi_{(p,k)}} = \frac{\lambda_S(p,k)}{\lambda_S(p,k)+\lambda_V(p,k)} \#(7)$$

2. Calculation Process of the Spectral Gain Function

Let $\hat{S}(p-1,k)$ denote the estimated clean speech spectrum at the (p−1)-th frame and k-th frequency bin, where p represents the frame index. The estimated $\hat{S}(p-1,k)$ can be obtained by applying a spectral gain function $H(p,k)$ to the noisy speech spectrum $Y(p-1,k)$ of the previous frame:

$$\hat{S}(p-1,k) = H(p,k)Y(p-1,k) \#(8)$$

$Y(p-1,k)$ is derived from the time-domain noisy speech signal $y(t-1)$ via the STFT. When computing $H(p,k)$, it is assumed that the clean speech signal $s(t)$ is correlated with the noisy speech signal $v(t)$, and their phase difference is represented as $\cos(\theta^S - \theta^Y)$:

$$H(p,k) = \frac{S(p,k)}{Y(p,k)}\cos(\theta^S - \theta^Y) \#(9)$$

Substituting $H(p,k)$ into the a priori SNR estimation formula of the DD algorithm, we obtain two a priori SNR estimates:

$$\hat{\xi}^{DD_1}_{(p,k)} = a\frac{|H(p,k)Y(p-1,k)|^2}{\lambda_V(p,k)} + (1-a)\max\{\gamma_{(p,k)}-1\} \#(10)$$

$$\hat{\xi}^{DD_2}_{(p,k)} = b\frac{|H(p,k)Y(p-1,k)|^2}{\lambda_V(p,k)} + (1-b)\max\{\gamma_{(p,k)}-1\} \#(11)$$

where a and b are smoothing factors ranging within [0,1]. To mitigate the frame delay effect in the DD algorithm, $\hat{\xi}^{DD_1}_{(p,k)}$ is assumed to rely more on the a priori SNR estimation from the previous frame, while $\hat{\xi}^{DD_2}_{(p,k)}$ depends more on the estimation from the current frame. Thus, a is set to a higher value of 0.992, whereas b is set to a lower value of 0.6. In the second step, a smoothing factor μ is introduced to combine the results of $\hat{\xi}^{DD_1}_{(p,k)}$ and $\hat{\xi}^{DD_2}_{(p,k)}$, yielding the final a priori SNR estimate $\hat{\xi}^{\mu}_{(p,k)}$

$$\hat{\xi}^{\mu}_{(p,k)} = \mu\hat{\xi}^{DD_1}_{(p,k)} + (1-\mu)\hat{\xi}^{DD_2}_{(p,k)} \#(12)$$

The smoothing factor μ is also constrained within the range [0,1]. To determine μ, a cost function $f(\cdot)$ is constructed to measure the minimum mean square error (MMSE) between the true a priori SNR $\xi_{(p,k)}$ and its estimated value $\hat{\xi}^{\mu}_{(p,k)}$:

$$f\left(\xi_{(p,k)}, \hat{\xi}^{\mu}_{(p,k)}\right) = E\left\{\left(\hat{\xi}^{\mu}_{(p,k)} - \xi_{(p,k)}\right)^2\right\} \#(13)$$

where $E\{\cdot\}$ denotes the expectation operation. Taking the partial derivative of f and replacing $\xi_{(p,k)}$ with $\hat{\xi}^{DD_2}_{(p,k)}$ from Equation (11), we obtain the expression for μ:

$$\mu = \frac{(1-b)\{\max\{\gamma_{(p,k)}-1\}+1\}^2 - b\left\{\hat{\xi}_{(p-1,k)} - \max\{\gamma_{(p,k)}-1\}\right\}^2}{(a-b)\left[\left\{\hat{\xi}_{(p-1,k)} - \max\{\gamma_{(p,k)}-1\}\right\}^2 + \{\max\{\gamma_{(p,k)}-1\}+1\}^2\right]} \#(14)$$

By substituting μ into Equation (12), the final a priori SNR estimate can be obtained. This estimate is then incorporated into the Wiener filter-like gain function derived in Equation (7), resulting in the final spectral gain function:

$$G^F_{(p,k)} = \frac{\hat{\xi}^{\mu}_{(p,k)}}{1+\hat{\xi}^{\mu}_{(p,k)}} \#(15)$$

3. A Two-Stage UAV Noise Removal Method Based on Deep Residual Neural Networks

3.1 Design of the Deep Residual Neural Network

Increasing the depth of a neural network results in higher training difficulty and greater computational resource requirements. To balance computational efficiency and performance, this study adopts the concepts of residual learning and residual blocks from Residual Neural Networks (ResNet) to construct a Deep Residual Neural Network (DRNN) consisting of four residual blocks, as illustrated in Figure 1. The proposed network is composed of one input layer, one batch normalization layer, four residual blocks, and one output layer. Each residual block consists of two fully connected linear layers, which are residually connected. Specifically, a skip connection is applied within each residual block to add the input of the residual block directly to the output of the fully connected layers. The residual blocks are sequentially stacked, where the output of each residual block serves as the direct input to the next residual block.

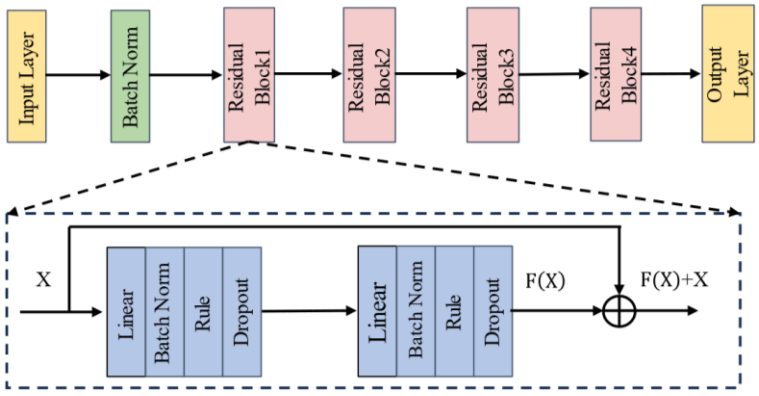

Figure 1: Spectral Gain Function Estimation Network F

3.2 Spectral Gain Function Estimation Using Deep Residual Neural Networks

In the spectral gain function estimation stage, a DRNN is employed to estimate the spectral gain function. The time-domain noisy speech signal is first transformed into the frequency domain using STFT. The transformed spectral features are then fed into the spectral gain function estimation network, which outputs the estimated spectral gain function. This estimation process follows a supervised learning approach, where the target label data corresponds to the power spectrum of clean speech signals. Once the spectral gain function is estimated by the DRNN, it is multiplied with the noisy speech spectrum to obtain the enhanced frequency-domain speech signal. Finally, the ISTFT is applied to reconstruct the enhanced time-domain speech signal. The overall framework of the two-stage UAV noise removal method based on DRNN is illustrated in Figure 2.

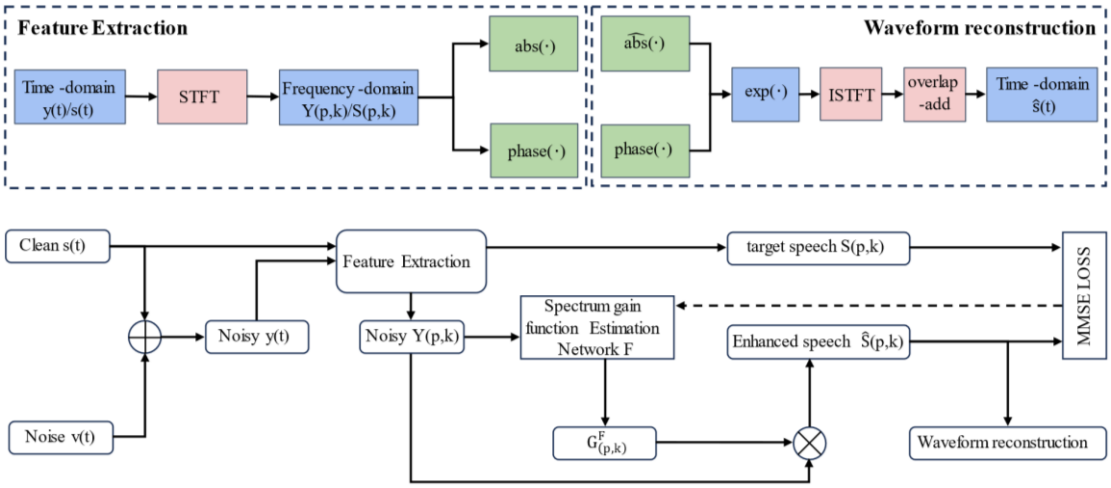

Figure 2: Overall framework diagram of the two-stage unmanned aerial vehicle noise removal method based on deep residual neural network

3.3 Detailed Steps of the Training and Enhancement Phases

The proposed method for UAV noise removal is introduced in detail from both the training phase and the enhancement phase.

Training Phase : First, the input time-domain noisy speech signal is transformed into the frequency domain using STFT, followed by frame concatenation along the time dimension. This frame expansion allows each short-time frame to access not only the current frame's information but also the contextual information from adjacent frames. Next, batch normalization is applied to the noisy speech signal to reduce data variance, accelerating model convergence. The spectral gain function estimation network is then trained, where the input data consists of noisy speech spectra, and the output data corresponds to the estimated spectral gain function. The clean speech signal undergoes STFT to extract spectral features, and its power spectrum is computed for each frame as the target label. To optimize training, a dynamic learning rate adjustment strategy is adopted. A learning rate that is too large may cause the model to fail to converge or oscillate, while a learning rate that is too small may lead to slow convergence and local optima. During the early stages of training, a larger learning rate is used to accelerate convergence, whereas a smaller learning rate is applied in later stages for fine-tuning. In this work, the learning rate is scaled down at fixed training intervals, with the learning rate adjustment cycle set to 30 epochs, the scaling factor set to 0.5, and the initial learning rate set to 0.01.

Enhancement Phase : The estimated spectral gain function $G^{F}_{(p,k)}$ is applied to the input noisy speech spectrum to obtain the enhanced clean speech spectrum. Then, ISTFT and waveform reconstruction are performed to obtain the enhanced time-domain speech signal. During waveform reconstruction, the magnitude spectrum of the enhanced clean speech signal is combined with the phase of the original noisy speech signal.

The detailed steps for different stages are shown in Table 1:

Table 1: Detailed Steps at Different Phases

Initialization Settings：
1. Define variables: $y(t)$ represents the noisy speech signal, $s(t)$ represents the clean speech signal, $v(t)$ represents the noise signal. 2. Resample the speech signal to 16 kHz. 3. Set learning rate (lr), number of training epochs (epoch), and batch size (batch_size). 4. Define the frame expansion range (n_expand).
Training Phase：

Step1. Transform the noisy speech signal from the time domain to the frequency domain using STFT $y(t) \rightarrow Y(p,k)$

Step2. Perform frame concatenation along the time dimension to provide contextual information $Y(p,k) = Y(p-1,k) + Y(p,k) + Y(p+1,k)$

Step3. Apply batch normalization, normalizing the data to a zero mean and unit variance distribution.

Step4. Train the spectral gain function estimation network

1. Initialize training iteration index $i = 1$

2. Perform iterative training

For $i <$ epoch do

$\widehat{G^{F}_{(p,k)}} = \text{Model}(X)$

$\text{loss} = \text{MSELoss}(\widehat{G^{F}_{(p,k)}}, Y)$

$\text{loss.backword}()$

$\text{optimizer.step}()$

lr_scheduler.StepLR()

End

3. Training completed. Save the best model weights as best.pth

Enhancement Phase：

Step1. Normalize the noisy speech signal in the test set to a zero mean and unit variance distribution.

Step2. Use the Noise Power Spectral Density Estimation Network

1. Convert the noisy speech signal from the time domain to the frequency domain using STFT $y(t) \rightarrow Y(p,k)$

2. Perform frame concatenation along the time dimension $Y(p,k) = Y(p-1,k) + Y(p,k) + Y(p+1,k)$

3. Estimate the spectral gain function using the trained model $\widehat{G^{F}_{(p,k)}} = \text{Model}\big(Y(p,k)\big)$

4. Compute the enhanced speech spectrum $\hat{S}(p,k) = G^{F}_{(p,k)} Y(p,k)$

5. Perform ISTFT and waveform reconstruction to obtain the enhanced time-domain speech signal $\hat{S}(t)$

4 Experiment and Result Analysis

4.1 Experimental Data and Settings

4.1.1 Experimental Data

The clean speech signals were taken from the TIMIT dataset, which includes recordings from male and female speakers with different accents in American English. All recordings were made with a 16 kHz sampling rate and 16-bit encoding[12]. This speech dataset contains samples from 630 speakers, with each speaker providing 10 sentences, resulting in a total of 6300 sentences and approximately 5 hours of audio. To construct the training and test sets, speech samples from 462 speakers, totaling 4620 sentences, were selected for training. The remaining 168 speakers, with 1680 sentences in total, were used for testing.

The noise signals were taken from a custom drone noise dataset, which includes various drone noises recorded under different flight conditions, such as takeoff, landing, hovering, rotating, left/right flight, vertical ascent/descent, and emergency stop. All recordings were made in mono-channel, with a 16 kHz sampling rate and 16-bit encoding.

The 4620 sentences from the TIMIT dataset used for training were mixed with drone noise from two randomly selected flight states at five different signal-to-noise ratios (SNRs) of -10 dB, -5 dB, 0 dB, 5 dB, and 10 dB, resulting in a total of 46,200 noisy speech samples for training. For testing, another set of 1680 sentences from the TIMIT dataset, unrelated to the training set, was mixed with drone noise from two randomly selected flight states at the same five SNRs, resulting in a total of 16,800 noisy speech samples for testing.

4.1.2 Experimental Setup

The experimental environment setup is shown in Table 2.

Table 2: Experimental Environment Setup

OS	Windows 11
CPU	Intel Core i5-13600KF
GPU	NVIDIA GeForce RTX 3060 Ti, 8192MiB
Python	3.9
Pytorch	1.12.0+cu113
CUDA	11.3

The experimental parameter settings are shown in Table 3.

Table 3: Experimental Parameter Settings

Training Epochs	300	Batch Size	16
Optimizer	Adam	Frame Length	256
Window Length	256	Frame Shift	128
Window Function	hamming	Initial Learning Rate	1e-2
Learning Rate Adjustment Period	30	Learning Rate Scaling Factor	0.5
Dropout Rate	0.1	Neurons in Hidden Layer	2048
Negative Slope	1e-1	Sampling Rate	16 KHZ
Frame Extension Range	3	Loss Function	MSELoss

A detailed description of the spectral gain function estimation network model parameters is shown in Table 4. The input format of the network is (T, D), where T represents the number of time frames and D represents the number of frequency bins, both obtained from the audio signal via STFT. The output format is $(T, D \div (2 * n\_expend + 1))$, where n_expend is the frame extension range, which means that for each current frame, it extends n_expend frames both forward and backward, and then concatenates them along the time dimension. The core idea of the residual neural network is $X + F(X)$, where X is the input and $F(X)$ is the nonlinear transformation in the main branch. The addition here is element-wise, meaning no increase in the feature dimension. If the dimensions of X and $F(X)$ are not the same, a linear layer is used to map X to the same dimensional space as $F(X)$. If the dimensions are the same, an identity mapping is applied. Therefore, the first residual block uses dimension transformation, and its output format is (T, 2048), while the remaining three residual blocks use identity mapping, and their output format is also (T, 2048).

Table 4: Spectral Gain Function Estimation Network Model Parameters

Network Structure	Input (T, D)	Output (T, D)
Input Laye	(T, D)	(T, D)
Batch Normalization Layer	(T, D)	(T, D)
First Residual Block	(T, D)	(T, 2048)
Second Residual Block	(T, 2048)	(T, 2048)
Third Residual Block	(T, 2048)	(T, 2048)
Fourth Residual Block	(T, 2048)	(T, 2048)
Output Layer	(T, 2048)	$(T, D \div (2 * n\_expend + 1))$

4.2 Evaluation Metrics

This study evaluates the performance of the proposed drone noise removal method from different dimensions. Objective evaluation metrics based on auditory perception, such as Perceptual Evaluation of Speech Quality (PESQ), are used. PESQ focuses more on assessing the overall

quality of the enhanced speech. Objective evaluation metrics based on speech intelligibility, such as Short-Time Objective Intelligibility (STOI), are also used. STOI focuses on evaluating the intelligibility of the enhanced speech. The evaluation dimensions, value ranges, and criteria for each metric are shown in Table 5.

Table 5: Evaluation Dimensions, Value Ranges, and Criteria for Each Metric

Metric	Evaluation Dimension	Value Range	Criteria
PESQ	Overall perceived speech quality	-0.5 to 4.5	The higher, the better
STOI	Speech intelligibility	0 to 1	The higher, the better

A higher PESQ indicates that the enhanced speech quality is closer to the ideal perceptible quality for the human ear. A higher STOI indicates that the enhanced speech is more similar to the reference speech in terms of intelligibility.

4.3 Experimental Results and Analysis

To verify the speech enhancement performance of the proposed method, comparative experiments were designed, using various types of speech enhancement methods for performance evaluation. The data used in the experiments were from the TIMIT speech dataset and the custom drone noise dataset. To facilitate intuitive comparison of the enhancement effects of different methods, the analysis of objective evaluation metrics is complemented with time-domain waveforms and spectrograms of the enhanced speech under -10 dB SNR conditions with drone noise.

4.3.1 Comparative Analysis of Objective Evaluation Metrics

Table 6 presents the results of the comparative experiments on the drone noise test set in a hovering state. The analysis shows that in this experimental environment, the proposed method's average STOI score under different SNR conditions (-10 dB, -5 dB, 0 dB, 5 dB, and 10 dB) is comparable to those of other methods, indicating that the enhanced speech's intelligibility is similar to that of other methods. However, the proposed method achieves higher average PESQ scores under all test SNR conditions, indicating that the enhanced speech is closer to the ideal perceptible quality for the human ear.

Table 6: Comparison of Objective Evaluation Metric Scores under Different SNR Conditions

	DNN		CNN		IBM		IRM		PSM		Our	
	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ
-10dB	0.54	1.34	0.55	1.27	0.54	1.54	0.56	1.35	0.53	1.46	0.56	1.68
-5dB	0.57	1.53	0.58	1.45	0.57	1.85	0.58	1.49	0.56	1.66	0.58	1.93
0dB	0.59	1.75	0.59	1.64	0.60	2.09	0.59	1.64	0.59	1.84	0.60	2.14
5dB	0.61	2.05	0.60	1.75	0.62	2.26	0.61	1.84	0.61	2.05	0.62	2.33
10dB	0.62	2.47	0.61	1.80	0.63	2.53	0.62	2.11	0.63	2.40	0.63	2.64

4.3.2 Comparative Analysis of Time-Domain Waveforms and Spectrograms

To more comprehensively analyze the speech denoising effect of the proposed method, a clean speech signal labeled as DR1 under FDAC1 from the TIMIT test set (FI1474) was selected. This clean signal was then mixed with drone noise in the hovering state at an SNR of -10 dB to generate a noisy speech signal, which was used as the input for processing. The time-domain waveforms and spectrograms of the clean speech, noisy speech, and enhanced speech were plotted using Python software, as shown in Figure 3. Figure 3(a) corresponds to the speech enhancement method based on deep residual neural networks for estimating the spectral gain function proposed in this study, while Figures 3(b) and 3(c) correspond to the spectral mapping methods based on DNN and CNN, respectively. Figures 3(d), 3(e), and 3(f) represent the mask estimation methods based on IBM[13], IRM[14], and PSM[14], respectively.

From the time-domain waveforms and spectrograms, it can be observed that, compared with the DNN- and CNN-based spectral mapping methods, the proposed method is able to more thoroughly eliminate drone noise. Furthermore, compared with the mask estimation methods based on IBM, IRM, and PSM, the proposed method is more effective in preserving the detailed information of the speech signal while simultaneously removing noise. The comparison of the time-domain waveforms and spectrograms further demonstrates that, under the same conditions, the proposed method not only better retains the speech signal's detailed information but also exhibits more significant advantages in drone noise suppression.

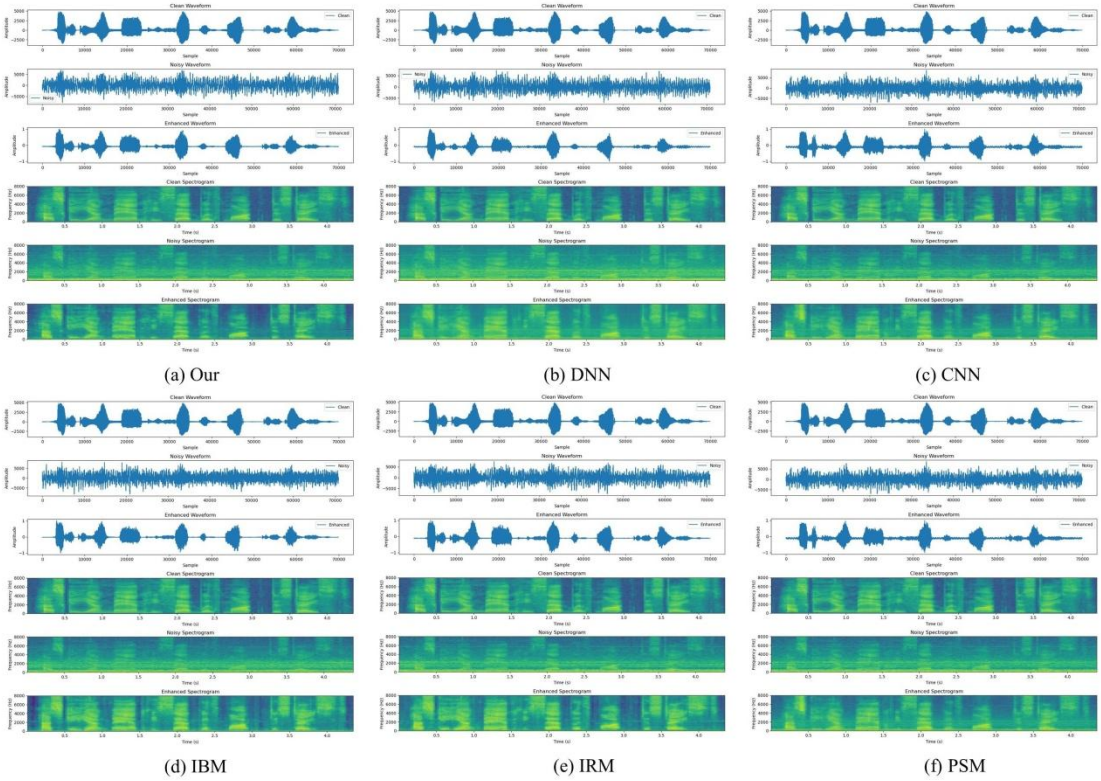

Figure 3: Comparison of Time-Domain Waveforms and Spectrograms

Conclusion

To address the issues of audio signal distortion, unclear speech, and even complete unintelligibility caused by drone noise, and to improve speech intelligibility by eliminating drone noise interference, a two-stage drone noise removal method based on deep residual neural networks is proposed. This method uses a deep residual neural network, consisting of four residual blocks, to estimate the spectral gain function. The obtained spectral gain function is then applied to the noisy speech to enhance it, resulting in a clean speech signal with reduced drone noise. Experimental results show that the proposed method outperforms traditional speech enhancement methods based on mask estimation and spectral mapping in terms of drone noise removal performance. Additionally, objective evaluation metrics (PESQ) confirm that the enhanced speech produced by the proposed method is closer to the ideal perceptible quality for the human ear. In the future, with the continuous development of deep learning and microphone array technologies, speech enhancement could integrate microphone arrays with deep learning techniques to achieve better denoising performance.

References

1. Xu X, Deng Q, Duan S, et al. A cooperative multi-drone priority coverage search algorithm [J]. Journal of System Simulation, 2024, 36 (04): 991-1000.
2. Zheng C, Hu X, Zhou Y, et al. Spectral subtraction based on the characteristics of noise spectrum structure [J]. Acta Acustica, 2022, 35(2): 215-222.
3. Pan J, Cao K, Ding J, et al. Design and implementation of a drone speech system

based on Wiener filtering [J]. Computer and Digital Engineering, 2021, 49(10): 2161-2167.
4. Ou S, Zhao Y, Song P, et al. A dual direct decision prior SNR estimation algorithm based on probabilistic coupling [J]. Journal of Electronics, 2020, 48(8): 1605-1614.
5. Plapous C, Marro C, Scalart P. Improved signal-to-noise ratio estimation for speech enhancement[J]. IEEE transactions on audio, speech, and language processing, 2006, 14 (6): 2098-2108.
6. Zhang X, Zhang T, Ge W, et al. Joint deep neural network and convex optimization based single-channel speech enhancement algorithm [J]. Acta Acustica, 2022, 46(3): 471-480.
7. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
8. Xu Y, Du J, Dai L R, et al. An experimental study on speech enhancement based on deep neural networks[J]. IEEE Signal processing letters, 2013, 21(1): 65-68.
9. Kounovsky T, Malek J. Single channel speech enhancement using convolutional neural network[C]//2017 IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM). IEEE, 2017: 1-5.
10. Zheng C, Zhang H, Liu W, et al. Sixty years of frequency-domain monaural speech enhancement: From traditional to deep learning methods[J]. Trends in Hearing, 2023, 27: 23312165231209913.
11. Shen S, Ou S, Wei J, et al. A priori SNR estimator based on a convex combination of two DD approaches for speech enhancement [C]//2016 IEEE International Conference on Signal and Image Processing (ICSIP). IEEE, 2016: 750-754.
12. Xu C, Wang L, Hu J, et al. U-Net speech enhancement method combined with residual and dual attention mechanisms [J]. Computer Engineering and Design, 2024, 45 (11): 3383-3389.
13. Wang D L. On ideal binary mask as the computational goal of auditory scene analysis[M]//Speech separation by humans and machines. Boston, MA: Springer US, 2005: 181-197.
14. Narayanan A, Wang D L. Ideal ratio mask estimation using deep neural networks for robust speech recognition[C]//2013 IEEE international conference on acoustics, speech and signal processing. IEEE, 2013: 7092-7096.
15. Erdogan H, Hershey J R, Watanabe S, et al. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015: 708-712.