Journal of Computers Vol. 37 No. 4, August 2026, pp. 293-310
https://doi.org/10.63367/19911599202608370401 7

Single-Channel Speech Enhancement Method Based on Deep Neural Networks

Wen Fan[1], Wei-Yu Liang[1*], Duo-Duo Han[1],
Wei Huang[2], and Han-Jiao Meng[1]

[1] Shijiazhuang Institute of Railway Technology,
Shijiazhuang City 050041, Hebei Province, China

[2] Hangzhou Hikvision Digital Technology Co., Ltd.,
Hangzhou City 310052, Zhejiang Province, China

{wyyx0012025, lwy123456789x, duoduosirt, 13821315956wei, 17331907926}@163.com

Received 10 July 2026; Revised 3 August 2026; Accepted 24 August 2026

Abstract. In practical voice communication and processing, speech signals are extremely sensitive to background noise, equipment noise, and interference from complex environments, leading to a decline in speech quality and clarity. This, in turn, affects the overall performance of downstream systems such as speech recognition, speech coding, and human-computer interaction. Therefore, researching efficient and robust single-channel speech signal enhancement methods has significant theoretical and practical implications. With the rapid development of deep learning technology, deep neural networks, with their superior nonlinear feature modeling capabilities and end-to-end training advantages, have become a core research area for single-channel speech enhancement. Given the limitations of traditional speech enhancement methods in complex, non-stationary, and noisy environments, this paper proposes a deep neural network-based speech enhancement method to address the challenges of single-channel speech enhancement. This method aims to extract a clearer, more natural, and more intelligible target signal from noisy speech. This paper first analyzes the basic principles of speech quality enhancement, including the characteristics of speech signals, types of background noise, time-domain and frequency-domain speech representation methods, and commonly used speech quality evaluation metrics. Based on this, this paper develops a deep neural network model for single-channel speech enhancement. This model effectively estimates the spectral information of the target speech by learning the spectral features of noisy speech.

Keywords: single-channel speech enhancement, speech noise suppression based on deep neural networks, time-frequency characteristics, deep learning

1 Introduction

1.1 Research Background and Significance

Voice signals serve as a vital medium for information exchange and emotional communication among humans, acting as carriers of information and representing the most fundamental, efficient, and convenient means of communication in human society [1]. With the rapid advancement of technology, speech signal processing techniques have emerged to facilitate broader applications of this expressive method across various domains of human life [2]. As a core application technology in emerging fields such as modern communication systems, artificial intelligence, and multimedia technology, speech signal processing primarily comprises four components: speech encoding, speech enhancement, speech recognition, and speech synthesis. During practical communication using speech signals, users often encounter unavoidable environmental noise interference, including background noise, device-generated noise, transmission medium noise, and interference from surrounding speakers. Such noise disturbances degrade transmission quality, resulting in received signals that are not pure original speech but rather contaminated noise signals [3]. In severe cases, interference noise may completely drown out useful speech signals, making them nearly unrecognizable. The deterioration of speech quality further impacts subsequent signal processing system performance, leading to more serious consequences.

* Corresponding Author

This necessitates preprocessing of noisy speech signals to effectively improve quality, mitigate noise effects, and enhance system performance. For receivers, the primary goal of speech enhancement is to improve speech quality, increase signal intelligibility, and reduce auditory fatigue [4]. For speech signal processing systems, the main objective is to enhance system recognition rates and improve interference resistance. Blind Source Separation (BSS) as a novel theory in signal processing, this technology has become a widely adopted method for multi-channel signal source separation [5]. Since its algorithms require no prior knowledge about the target signals, it demonstrates extensive applicability across various fields, including speech enhancement systems. The fundamental approach of ap-plying BSS to noise-contaminated speech signals involves modeling the signals as linear superpositions of mutually independent pure speech components and noise components [6]. Through BSS algorithms, the mixed signals are separated to isolate pure speech from noise, thereby achieving speech enhancement. Independent Com-ponent Analysis (ICA) remains the most prevalent method in BSS applications. The method directly employs the ICA approach to achieve separation of single-channel mixed signals [7]. In practice, however, only one observation channel is typically available. Therefore, for single-channel noisy speech signals, applying blind source separation techniques to enhance speech signals holds significant practical implications.

1.2 Review of Speech Enhancement Research

Currently, most speech signal processing systems urgently require performance optimization to meet the growing demands of practical applications [8]. As one of the key technologies for enhancing system performance, speech enhancement techniques hold significant importance due to their broad application prospects. Developing effective algorithms with superior noise reduction capabilities has become crucial [9]. However, real-world noise environments typically exhibit randomness, making it nearly impossible to completely eliminate noise from speech signals and extract pure original speech. Based on this reality, speech enhancement serves two primary objectives: First, improving speech quality by removing environmental noise to reduce listeners' auditory fatigue; second, enhancing speech intelligibility to facilitate understanding (an objective metric). These objectives often conflict, with trade-offs sometimes necessitated to prioritize one over the other. This demonstrates that speech enhancement technology extends beyond digital signal processing, involving interdisciplinary fields such as phonetics and psychology, making it a highly complex applied technology. Research on speech enhancement technology has garnered widespread attention from scholars across academic disciplines since the 1960s, with a development history spanning several decades [10]. By the 1970s, breakthroughs in foundational research propelled its advancement to a peak, establishing speech enhancement technology as a pivotal component in the field of speech signal processing. In recent years, the rapid advancement of Digital Signal Processing (DSP) and Very Large Scale Integration (VLSI) technologies has made real-time implementation of speech enhancement techniques feasible [11]. A series of innovative speech enhancement methods have emerged, progressively bringing these technologies closer to practical application. Representative approaches include wavelet transform-based techniques, human ear masking effect utilization, and microphone array-based speech enhancement methods. These methodologies have become mainstream research directions in speech enhancement studies, garnering widespread attention. In recent years, blind source separation technology has emerged as one of the most promising research topics in digital signal processing, distinguished by its robust theoretical foundations and broad application prospects. It has been widely applied and received significant attention in speech enhancement.

1.3 Current Application Status of Blind Source Separation Technology

Blind source separation technology, a novel technique rapidly developed in the late 20th century based on statistical methods and neural network theory, has become one of the most prominent research topics in both neural net-works and signal processing fields. It holds significant theoretical value and promising application prospects in areas such as speech signal processing, modern communication systems, remote sensing, and biomedical engineering [12]. In signal processing research, a common challenge arises where collected observation signals are solely derived from sensor outputs without knowledge of the original input signals or transmission systems [13]. The task involves separating independent source signals from a random mixture of unknown input signals to reconstruct the original signals, all without requiring prior knowledge about source characteristics or hybrid system models. This process is termed "blind source separation". Due to its ability to

operate without prior information about signal sources and communication channels, blind source separation technology demonstrates significant practical value in speech enhancement applications [14]. Recent research papers addressing signal blind source separation can be broadly categorized into two approaches: real-time mixed signal blind source separation and convolutional mixed signal blind source separation. While theoretical studies on instantaneous mixed signal blind source separation have achieved substantial progress with well-established methodologies, convolutional mixed signal scenarios remain particularly challenging [15]. However, existing blind source separation algorithms universally rely on specific assumptions about original input signals and channel propagation models. These assumptions often diverge significantly from actual signal characteristics and channel properties in real-world applications, thereby compromising algorithmic effectiveness. Therefore, developing new algorithms under more flexible signal model constraints has become imperative [16]. Current research focuses on addressing blind source separation problems with fewer observed signals than independent source signals, which represents a critical frontier in solving single-channel speech enhancement challenges. Since blind source separation algorithms are only applicable to separated complete signals, the primary challenge in single-channel speech enhancement applications lies in determining the optimal number of observed signals [17]. Current research primarily focuses on generating additional virtual observation signals from known single-channel mixed observations, which significantly compromises the mutual independence between the two observation channels and consequently impacts final enhancement performance. Moreover, most existing algorithms are developed under white noise assumptions, while enhancement techniques for colored noise remain under-researched and require further implementation.

2 Basic Theory of Speech Enhancement

2.1 Basic Knowledge of Voice Enhancement

1) Speech signal digital model, the development of digital voice signal models aims to enable computers to more effectively simulate and process speech signals by establishing mathematical representations that describe quantitative relationships under specific physical conditions [18]. As the foundation for speech signal processing, these digital models are grounded in theoretical principles derived from human vocal organ characteristics and speech generation mechanisms.

Fig. 1 illustrates the human vocal mechanism. The human vocal system consists of three components: the lungs, larynx, and vocal tract. During speech production, the lungs deliver airflow to the larynx as a power source. The larynx then modulates the incoming airflow into acoustic excitation signals that enter the vocal tract. The frequency spectrum of these acoustic signals undergoes reshaping through the oral cavity, nasal cavity, and pharyngeal cavity within the vocal tract, resulting in distinct timbres. Finally, lip movements generate pressure fluctuations that produce sound waves, thereby generating speech.

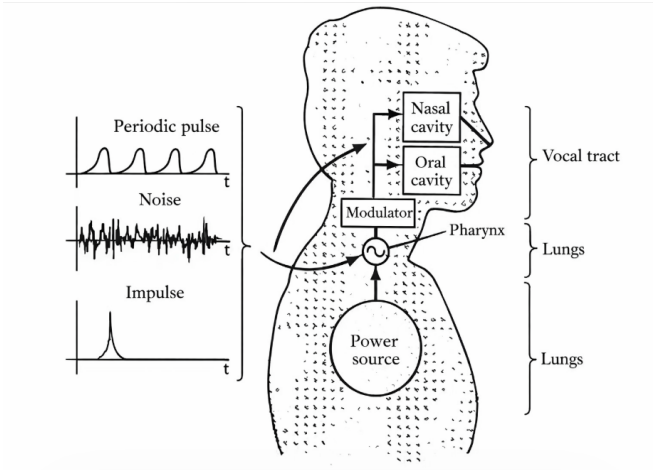

Fig. 1. Diagram of speech generation mechanism

As shown in the figure above, the digital model of speech signals primarily consists of three fundamental components:

This includes clear sound excitation and voiced sound excitation. The classification of clear and voiced sounds depends on the specific position of the clear/voiced sound switch. The clear sound excitation source signal is typically generated by a random noise generator, with its excitation model represented as:

$$U(z) = AN \tag{1}$$

The voiced excitation source signal is generated by a periodic pulse generator, with its excitation model expressed as:

$$U(z) = AvG(z)E(z) = \frac{Av}{1-z^{-1}} \cdot \frac{1}{\left(1-az^{-1}\right)^2} \tag{2}$$

Although the above incentive model classification is not comprehensive enough, it remains widely adopted due to its simplicity and superior performance [19]. The resonance peak model is currently the most widely used vocal tract model. Given that the average human vocal tract length is approximately 17 cm, five resonance peaks can be calculated at frequencies of 500 Hz, 1500 Hz, 2500 Hz, 3500 Hz, and 4500 Hz. Each resonance peak corresponds to an extreme value point in the system transfer function described by the vocal tract model. Therefore, the vocal tract model can be expressed by the following equation:

$$V(z) = \frac{1}{1-\sum_{k=1}^{N} a_k z^{-k}} \tag{3}$$

In this speech signal generation model, parameters of both the excitation source and vocal tract model vary over time, with their change rates constrained by the inherent inertia of the vocal organs. Typically within short time intervals of 10–30 milliseconds, vocal tract parameters can be approximated as remaining constant. The aforementioned discrete-time domain model is also referred to as the "excitation source/filter" model, indicating that within brief time intervals, speech signals can be regarded as outputs generated by a linear system composed of an excitation source signal and the system transfer function $H(z)$. Here, $H(z)$ represents the series connection of the three aforementioned models.

2) Speech Enhancement System, the principle block diagram of a speech enhancement system designed to improve speech signal quality is shown in Fig. 2.

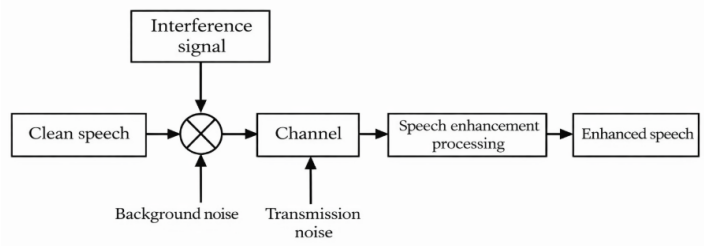

Fig. 2. Principle diagram of speech enhancement system

The mathematical model of typically acquired noisy speech signals can be expressed as:

$$x(n) = s(n) + d(n) \tag{4}$$

Here, $s(n)$ denotes the pure speech signal, while $d(n)$ represents the interference noise present in the speech. In general research work, the following assumptions also need to be made: Interference noise signals exhibit local stationarity, and their characteristics can be estimated from the pure noise segment preceding speech onset in noisy speech signals. Speech signals and noise are statistically independent or uncorrelated. Throughout the entire process, only noisy speech signals can be utilized, with no other reference signals available.

3) Speech and noise characteristics, research on speech enhancement methods is not only closely related to signal processing techniques, but also involves human phonetics, psychology, and the relevant characteristics of speech signals. Additionally, due to the diverse sources of interference noise. Characteristics vary significantly depending on application scenarios, making it challenging to identify a speech enhancement algorithm universally applicable to all noise conditions.

(a) Speech is a stochastic process with time-varying and non-stationary properties.

During human vocalization, the shape of vocal cords and vocal tract remains relatively stable over short time intervals, making their characteristics essentially invariable during this period. This property is commonly utilized in speech enhancement as the short-term stationarity of speech signals. The short-time speech signal can be expressed as:

$$x_m(n) = x(n + mH)w(n) \tag{5}$$

Where $x_m(n)$ denotes the m-th frame of the speech signal, $w(n)$ denotes the window function, H denotes the frame shift, and N denotes the frame length.

(b) Speech signal mainly consists of two parts: clear sound and voiced sound.

Clear voice signals and voiced voice signals exhibit distinct differences in both time domain and frequency do-main. Clear voices resemble white noise, showing no prominent characteristics in either domain. In contrast, voiced sounds demonstrate significant periodicity in the time domain, with their frequency domain energy predominantly concentrated in the low-frequency range and exhibiting a resonant peak structure.

(c) Speech can be described by its statistical properties.

As a time-varying stochastic process, speech signals exhibit non-stationarity and non-ergodic properties. Long-term time-domain statistical analysis has limited significance for speech enhancement algorithm research. In real-world environments, noise is omnipresent and exhibits diverse characteristics. Speech signals typically contain two major types of noise: additive noise and non-additive noise. Most non-additive noise can often be transformed into additive noise through various techniques (e.g., homomorphic transformations) for subsequent processing. Therefore, this study primarily focuses on additive noise scenarios. The additive noise model can be expressed as:

$$y(n) = x(n) + d(n) \tag{6}$$

Where $y(n)$, $x(n)$, and $d(n)$ denote the noisy speech signal, clean speech signal, and additive noise signal at time index n, respectively. In the frequency domain, the corresponding model can be written as:

$$Y(k,m) = X(k,m) + D(k,m) \tag{7}$$

Where $Y(k, m)$, $X(k, m)$, and $D(k, m)$ represent the short-time Fourier transform coefficients of noisy speech, clean speech, and noise at the k-th frequency bin of the m-th frame, respectively. Additive noise generally includes broadband noise, periodic noise, impulse noise, and interference from other speech signals within the same channel.

(d) Broadband noise

Broadband noise originates from diverse sources including respiratory sounds, wind noise, and other random noise sources. In practical applications, it is typically modeled as Gaussian noise or white noise. Since broadband noise spectra are distributed throughout the entire frequency spectrum of speech signals, this significantly increases the difficulty of noise cancellation. The situation becomes particularly complex when broadband noise exhibits non-stationary characteristics. Current industry practices generally employ specialized nonlinear processing methods to effectively mitigate such noise interference.

(e) Periodic noise

Periodic noise typically refers to the noise emitted by periodically operating equipment (e.g., fans, engines)

and other electrical interference, characterized primarily by a series of discrete narrow spectral peaks in the frequency spectrum. Since periodic noise can be detected through its power spectrum and removed using notch filters, it causes the least amount of interference. The power spectrum can be expressed as:

$$P_x(k) = |\ X(k)^2 \tag{8}$$

Where $P_x(k)$ denotes the power spectrum at the k-th frequency bin, and $X(k)$ denotes the Fourier transform coefficient of the signal.

2.2 Speech Enhancement Quality Evaluation

The evaluation of speech signal quality primarily encompasses two key aspects: clarity and intelligibility. Clarity measures the distinctiveness of individual words and syllables within speech, while intelligibility assesses the speaker's ability to be recognized. Through years of dedicated research multiple evaluation methods have been developed to effectively evaluate the noise suppression capabilities of various speech enhancement techniques. These methods can generally be categorized into two main types: subjective evaluation methods and objective evaluation methods. The signal-to-noise ratio can be expressed as:

$$SNR = 10log_{10}\frac{\sum_{n=1}^{N} x^2(n)}{\sum_{n=1}^{N} d^2(n)} \tag{9}$$

Where $X^2(n)$ denotes the clean speech signal, $d^2(n)$ denotes the noise signal, and N denotes the signal length.

Subjective evaluation methods serve as a key criterion for assessing speech enhancement quality through human subjectivity [20]. These approaches have gained widespread application due to their alignment with human auditory perception of speech. However, they present notable drawbacks: time and effort consumption, as well as susceptibility to human subjective biases that significantly compromise stability and reproducibility. Commonly employed testing methods include the Mean Opinion Score (MOS) and Diagnostic Rhyme Test (DRT). The specific procedure of this method involves: convening experimenters to conduct subjective audiometric testing of speech and rating the speech quality levels. The average score calculated from all participants' evaluations constitutes the MOS score for speech quality assessment. The MOS score can be calculated as:

$$MOS = \frac{1}{M}\sum_{i=1}^{M} r_i \tag{10}$$

Where M denotes the number of participants, and r_i denotes the score given by the i-th participant. The MOS score categorizes enhanced speech signal quality into five levels: 5 (excellent), 4 (good), 3 (moderate), 2 (poor), and 1 (bad). Due to various subjective and objective factors, the test results of MOS scores may exhibit some fluctuations. In addition to requiring sufficient participants and audio materials for testing, it is also essential to maintain consistency in the testing environment in Table 1.

Table 1. Score standard of MOS

MOS component	Quality level	Distortion level
5	Ample	Unperceived
4	Good people	Slight awareness without discomfort
3	Centre	Awareness accompanied by discomfort
2	Difference	The noticeable sensation is still tolerable
1	Bad	Intolerability

The grading system categorizes speech enhancement quality as follows: “Excellent” indicates minimal detail differences between enhanced and original speech that are imperceptible to humans; “Good” denotes slight distortion or distortion that remains unnoticed without attentive listening; “Medium” reflects noticeable distortion that can be perceived but does not significantly impair speech clarity or intelligibility; “Poor” signifies substantial distortion causing auditory fatigue yet still tolerable; “Bad” represents exceptionally poor speech quality that exceeds human auditory tolerance thresholds.

Objective evaluation methods assess the quality of speech signals by analyzing the differences between input and system output signals. These methods require minimal human resources, are easy to implement, exhibit strong repeatability and stability, and are immune to subjective human influence. Currently, the main objective evaluation methods include: Signal-to-noise ratio (SNR):

$$SNR = 10\log 10\frac{\sum_{n} s^2(n)}{\sum_{n}\left[s(n)-s'(n)\right]^2} \tag{11}$$

Here, $s(n)$ denotes pure speech, while $s'(n)$ represents enhanced speech. Segmented Signal-to-Noise Ratio (SEGSNR)is calculated by averaging the SNR of each speech frame across the entire speech segment, thereby providing a more accurate measurement of speech quality. Its definition formula is as follows:

$$E_{seg} = \frac{1}{L_s}\sum_{l=0}^{L_s} E_{frame}(l) \tag{12}$$

L denotes the number of speech frames, and $E_{frame}(l)$ represents the Signal-to-Noise Ratio (SNR) of each speech frame.

$$E_{frame}(l) = 10\log 10\left(\frac{\sum_{j=0}^{M_s-1} s^2(M_s l-j)}{\sum_{j=0}^{M_s-1}\left[s'(M_s l-j)-s(M_s l-j)\right]^2}\right) \tag{13}$$

The Sakagawa distance metric utilizes linear prediction analysis to evaluate the quality of enhanced speech signals. Its calculation formula is:

$$\sum_{i=1}^{L} 1 \Bigg/ \left\{2\left[\log\left(\frac{N}{2}\sum_{k=1}^{N/2}\frac{X_{(l,k)}}{X'_{(l,k)}}\times\frac{N}{2}\sum_{k=1}^{N/2}\frac{X'_{(l,k)}}{X_{(l,k)}}\right)\right]\right\} \tag{14}$$

In the formula, $X(n)$ and $X'(n)$ represent pure speech and enhanced speech, respectively. $X(A)$ and $X'(x)$ denote the first frame and k-th frame frequency components obtained from $X(n)$ and $X'(n)$ after short-time Fourier trans-form. PESQ (Perceptual Evaluation of Speech Quality) is a comprehensive method for evaluating speech quality perception. It establishes mathematical models to correlate physical speech characteristics with human psycho-logical perception, serving as an objective model-based approach to simulate subjective sensations. This evaluation method demonstrates broad applicability across diverse scenarios. The PESQ output ranges from-0.5 to 4.5. A score closer to 4.5 indicates improved speech signal quality after enhancement, while lower scores suggest poorer quality. Typically, PESQ results fall within the 1.0-4.5 range. Scores below 2 indicate suboptimal speech quality, making the enhanced audio difficult for human ears to comprehend.

2.3 Common Analysis Methods for Voice Enhancement

If $X(t, w)$ denotes the short-time Fourier transform of the signal $x(t)$, its definition can be expressed as:

$$X(t,w)=\sum_{t=-L/2}^{r=L/2} x(\tau)w^{*}(\tau-t)e^{-jwt}d\tau \tag{15}$$

The fundamental approach involves applying the direct method from classical spectral analysis theory to non-stationary signal processing. The implementation process entails first multiplying the signal $x(t)$ with a window function $w(\tau-t)$ centered at t, followed by Fourier transformation. Windowing is employed to extract and analyze information within specific time intervals. Short-time Fourier transform (STFT) utilizes the signal's short-term stationarity assumption. To achieve optimal time-frequency analysis performance, careful selection of window functions is essential.

Assuming the signal $x(t)$ belongs to L2(R)—meaning $x(t)$ is square-integrable—let $WT_z(a, b)$ denote the wavelet transform of $x(t)$, defined as:

$$\begin{aligned} WT_z(a,b) &= \frac{1}{\sqrt{a}}\int x(t)\psi^{*}\left(\frac{t-b}{a}\right)dt \\ &= \int x(t)\psi_{a,b}{}^{*}(t)dt \end{aligned} \tag{16}$$

Here, variables a, b, and t are all continuous, with a and b being constants where $a > 0$. The function $\psi_{a,b}(t)$ is referred to as the wavelet basis function. Since wavelet transform can concentrate most of the signal energy on a small number of wavelet coefficients, and then assign smaller weights or set the remaining coefficients to 0, that is, apply corresponding thresholds to them, it can suppress the noise contained in the signal to some extent.

3 Construction of a Single-Channel Speech Enhancement Model Based on Deep Neural Net-works

3.1 Overall Model Framework

To address the challenges of complex noise types, significant temporal dependencies, and limited generalization capabilities of traditional methods in single-channel speech enhancement, this paper constructs a single-channel speech enhancement model based on deep neural networks. This model uses the time-frequency representation of noisy speech as input and employs multi-objective learning as its basic strategy. By jointly learning spectral masking and spectral mapping targets, it establishes a nonlinear mapping relationship between noisy speech and clean speech, thereby achieving effective recovery of the target speech. This approach aligns with the research focus on "noise reduction performance, generalization, and computational efficiency" in the referenced papers. MS-TCN emphasizes efficient temporal modeling, PhaseDCN focuses on joint phase and amplitude enhancement, while MT-in-MS further integrates multi-objective and traditional statistical models. From the overall processing flow, the model mainly includes five stages: input feature construction, deep feature encoding, temporal information modeling, multi-objective output, and speech reconstruction. First, a short-time Fourier transform is performed on the single-channel noisy speech signal to extract features such as the logarithmic power spectrum and the real and imaginary parts of the complex spectrum. Second, these features are fed into a deep convolutional-temporal modeling network to extract local time-frequency structure and long-term contextual dependencies. Third, an intermediate target such as ideal ratio masking, logarithmic power spectrum, or complex spectrum is estimated through a multi-objective learning strategy. Finally, the enhanced amplitude spectrum or complex spectrum is combined with phase information and reconstructed into enhanced speech through inverse transform. The referenced paper shows that this joint learning approach of "masking + mapping" can fully utilize the complementarity between different targets and exhibits better robustness in mismatched testing environments.

3.2 Input Feature Representation and Preprocessing

The core premise of single-channel speech enhancement is to transform the temporal noisy speech into a stable feature representation suitable for neural network learning [21]. The referenced paper uses logarithmic power spectrum (LPS), real and imaginary parts of the complex spectrum (RI), and temporal waveform features as inputs in different chapters, with LPS and RI being the primary inputs. Logarithmic power spectrum effectively characterizes speech energy distribution, making it suitable for masking or amplitude mapping tasks; complex spectrum pre-serves both amplitude and phase information, providing a foundation for phase enhancement. In the preprocessing stage, all speech and noise signals are uniformly resampled to 16 kHz. In the basic model and most experimental settings, a Hamming window with a frame length of 32*ms* and a frame shift of 16*ms* is used for framing and short-time Fourier transform. For the MS-TCN model, 257-dimensional logarithmic power spectrum features are extracted as input; for the PhaseDCN model, 514-dimensional complex spectrum real and imaginary part features and 257-dimensional logarithmic power spectrum features are further extracted to support dual-path multi-objective learning. For the MT-in-MS model, which further pursues lightweight and real-time performance, the frame length and frame shift are further reduced to 20*ms*/ 10*ms* to achieve a better balance between noise reduction performance and computational cost. To improve the stability of model training, the input features need to be normalized. The referenced paper preserves the global mean and variance of the training set under a causal system setting and normalizes the input using the same statistics during the inference phase, thereby avoiding future in-formation leakage and enhancing the model's usability in real-world deployment scenarios. Therefore, the input layer in this paper is not a simple "noisy speech direct input network," but rather a feature engineering process built upon a unified sampling rate, standardized frame analysis, STFT transform, and statistical normalization.

3.3 Network Structure Design

The basic model in this paper adopts a structure design based on Forward Stacked Multi-Scale Temporal Convolutional Network (MS-TCN). Its basic idea is to stack multiple residual modules between two feedforward fully connected layers and gradually expand the temporal receptive field through convolutions with different dilation rates, thereby achieving the modeling of long-term temporal dependencies in speech. Compared with RNN or LSTM, TCN's advantage lies in being composed entirely of convolutional layers, allowing for parallel acceleration during training and inference, flexible temporal modeling, and higher computational efficiency [22]. The referenced paper points out that as the hole ratio increases, the network's perception range of context expands continuously. This characteristic gives TCN near-recursive network capabilities in temporal modeling, but with greater efficiency. In terms of specific module structure, the basic network consists of an input mapping layer, several multi-scale residual modules, and an output mapping layer. The input mapping layer projects the original features into a high-dimensional latent space; the intermediate residual modules handle the core feature extraction task, internally com-posed of convolutions, normalization, nonlinear activations, dropout layers, and residual connections, used to alleviate the gradient vanishing problem in deep network training; the output layer outputs masking values, enhanced spectra, or complex spectrum estimates according to different tasks. MS-TCN introduces multi-scale convolutional branches in the residual blocks for finer-grained analysis of speech structural changes at different time-frequency scales, thereby improving the ability to recover speech details under complex noise conditions. In further model optimization, the referenced paper proposes the PhaseDCN structure, which employs a dual-path learning mechanism within a dilated convolution framework: one path models the complex spectrum, and the other models the ideal ratio mask. It also utilizes intermediate features from the auxiliary path to generate attention gating factors, helping the main path more accurately distinguish between speech and non-speech components. This structure is essentially a more advanced multi-objective learning network, providing direct technical support for phase enhancement and dual-path complementarity in subsequent model optimization sections in Fig. 3.

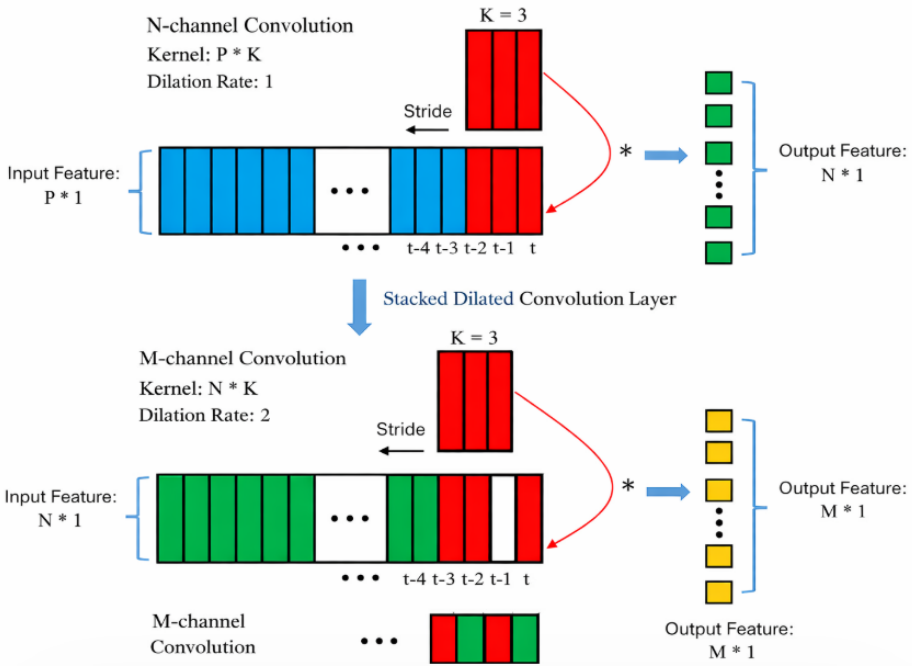

Fig. 3. Working principle of one-dimensional dilated convolution

3.4 Loss Function Design

Loss function design is crucial for improving the performance of single-channel speech enhancement models. The core idea of the referenced paper is not to regress a single objective, but to jointly optimize different enhancement objectives through multi-objective learning to leverage their complementary advantages. For the basic MS-TCN model, it primarily jointly learns two objectives: ideal ratio mask (IRM) and log power spectrum (LPS). The former emphasizes estimating the ratio of speech to noise components, suitable for noise suppression; the latter emphasizes approximating the enhanced spectral structure to the clean speech spectral structure, suitable for improving speech naturalness and fidelity. Combining these two approaches avoids the model focusing solely on "noise re-duction" while neglecting "distortion control." In subsequent models, the loss function is further extended to include both the complex spectrum objective and the prior signal-to-noise ratio (SNR) objective. PhaseDCN employs a joint learning strategy of "main objective + auxiliary objective," using complex spectrum mapping as the main objective and IRM as the auxiliary objective. Attention gating information is generated through intermediate features of the auxiliary path to improve the accuracy of complex spectrum estimation. This demonstrates that in the phase enhancement task, IRM is not discarded but continues to play a role in "speech/noise separation" as auxiliary supervision. In the MT-in-MS stage, the loss function is further extended to a two-stage joint optimization of the complex spectrum, masking objective, and prior SNR. The aim is to establish a bridge between the second-stage learning of the prior SNR and the traditional MMSE-LSA statistical model, combining the learning ability of the neural network model with the robustness of the statistical model. Therefore, the loss function design in this paper follows a progressive approach, moving from single-objective to multi-objective and from single-stage to two-stage.

3.5 Model Training Process

For model training, a relatively standardized data construction and parameter training process was established based on referenced papers. Training speech data was taken from the TIMIT dataset, with 4620 sentences used for training and 1680 for testing. Noise data was mainly taken from Noisex-92, using four types of noise during training: multi-person speaking noise, factory noise 1, destroyer engine noise, and destroyer control room noise. During training set construction, the first 60% of the noise samples were randomly mixed with the

training speech, with the mixed signal-to-noise ratio following a uniform distribution between -5 dB and 15 dB, ultimately constructing approximately 12.5 hours of noisy speech training set. The validation set was obtained by mixing the middle 20% of the noise with 280 test sentences. In terms of training strategy, model weights were randomly initialized using Xavier, and the optimizer was Adam. The PhaseDCN model is trained for 50 epochs with a batch size of 10,000 frames. The learning rate is set to 0.001 for the first 40 epochs and decays to 0.0001 for the last 10 epochs. Batch normalization and Dropout (dropout rate 0.2) are introduced to alleviate overfitting. The referenced paper also emphasizes the need to preserve the mean and variance of the training set features under causal inference settings for consistent normalization during the testing phase; this design is particularly important for real-time deployment scenarios.

4 Model Optimization and Enhancement Method Improvement

4.1 Temporal Feature Modeling Strategy

Speech signals naturally exhibit significant contextual dependencies; therefore, the ability to model temporal dependencies directly determines the model's enhancement performance [23]. The temporal dependency of speech signals can be generally expressed as:

$$h_t = f(x_t, x_{t-1}, \ldots, x_{t-\tau}) \tag{17}$$

Where x_t denotes the speech feature at time frame t and h_t denotes the temporal feature representation obtained by the model. Chapter 3 of the references specifically compares the advantages and disadvantages of different neural network temporal modeling schemes, pointing out that while traditional recursive models can model long-term dependencies, training and inference are difficult to parallelize efficiently. For recursive temporal modeling, the hidden state is usually calculated as:

$$h_t = f(x_t, h_{t-1}) \tag{18}$$

Where h_t indicates that the computation of the current frame depends on the previous hidden state, which leads to sequential computation and limits parallelization. TCN, on the other hand, continuously expands the receptive field through dilated convolutions, significantly improving computational efficiency while maintaining strong contextual modeling capabilities. The dilated convolution operation can be expressed as:

$$y_t = \sum_{i=0}^{k-1} w_i x_{t-d \cdot i} \tag{19}$$

Where K denotes the convolution kernel size, d denotes the dilation factor, w_i denotes the convolution kernel parameter, and y_t denotes the output feature at time frame t. The receptive field of stacked dilated convolution layers can be expressed as:

$$R = 1 + (K-1)\sum_{l=1}^{L} d_l \tag{20}$$

Where R denotes the receptive field, L denotes the number of convolutional layers, and d_l denotes the dilation factor of the l-th layer. Therefore, this paper adopts multi-scale TCN as the main temporal modeling strategy in model optimization. Furthermore, to improve the extraction capability of information at different temporal scales, the references do not use a single convolutional scale but introduce a multi-scale design in the residual module, enabling the model to perceive local changes in adjacent frames and capture long-distance temporal dependencies. The multi-scale temporal feature extraction process can be expressed as:

$$z^{(s)} = Conv_{K_s,d_s}(x) \tag{21}$$

Where S denotes the number of convolutional scales, respectively. The residual propagation process can be expressed as:

$$y = F(x) + x \tag{22}$$

Where $F(x)$ denotes the nonlinear transformation learned by the multi-scale TCN module, and x denotes the residual input. This improvement is particularly important for speech enhancement in non-stationary noise environments, because noise in real-world environments is characterized by suddenness, gradualness, and multi-source mixing, making it difficult to fully model using only single-scale convolutions. Therefore, the key points of temporal modeling optimization in this chapter can be summarized as: shifting from recursive modeling to convolutional modeling, from single-scale to multi-scale, and from single-path feature propagation to depth propagation with residuals.

4.2 Spectral Masking Estimation Methods

Spectral masking estimation is one of the most representative techniques in deep learning speech enhancement. The references employed joint learning of IRM+LPS in MS-TCN and further adopted joint learning of IRM+RI in PhaseDCN. The core logic is that IRM can explicitly represent the relative proportion of speech and noise at different frequency points, thus possessing strong noise suppression capabilities; the spectral mapping target (LPS or RI) is more conducive to recovering the structural information of the enhanced speech, thereby reducing the speech distortion that may be caused by using masking methods alone. In PhaseDCN, the masking target does not end immediately after independent output, but rather generates an attention gating factor through the intermediate features of the auxiliary path, further guiding the main path to learn complex spectra. In other words, masking estimation here is elevated to a kind of structural guidance information; it is not only the output target but also part of the feature selection mechanism. This improvement demonstrates that, compared to simply estimating mask values, this paper emphasizes the role of mask information in representation learning within the network. Therefore, in this chapter, the spectral mask estimation method should not be simply described as "outputting a mask matrix," but rather its three functions should be emphasized: first, as a noise suppression target; second, as a joint learning object complementary to the mapping target; and third, as an auxiliary information source for attention gating and feature selection. This is also a crucial transition from the basic MS-TCN to PhaseDCN and MT-in-MS in Fig. 4.

4.3 Parameter Settings and Training Optimization

To achieve a balance between performance improvement and generalization performance, the references significantly optimized the parameter settings and training strategies [24]. First, in the basic model stage, the input is a 257-dimensional LPS, the frame length is 32*ms*, and the frame offset is 16*ms*. In the PhaseDCN stage, to support complex spectral modeling, the input is expanded to 514-dimensional RI features and 257-dimensional LPS features. This parameter setting reflects the idea that the model's input representation capability gradually increases with the complexity of the objective. Secondly, regarding training process optimization, this paper consistently uses Adam for parameter updates and combines Xavier initialization, batch normalization, and Dropout to stabilize training. For PhaseDCN, a larger learning rate is used initially to achieve rapid convergence, followed by a reduction in the learning rate for fine-tuning. For MT-in-MS, by shortening the analysis frame length and frame offset, model parameters and computational load are further reduced, resulting in better real-time performance. This demonstrates that parameter settings are not isolated but closely related to the model structure and deployment objectives. Furthermore, regarding the configuration of training objectives, this paper emphasizes the significant complementarity between different objectives. Therefore, training optimization should not solely focus on the convergence of a single object.

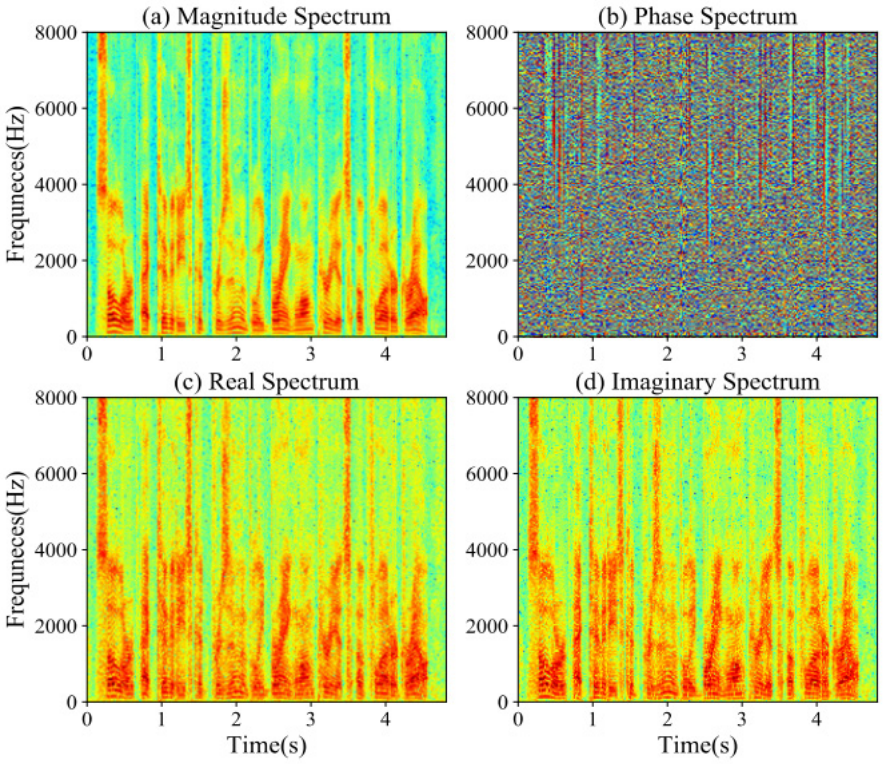

Fig. 4. Spectrogram of pure speech

5 Experimental Design and Results Analysis

5.1 Experimental Dataset and Noise Settings

To verify the effectiveness and generalization of the proposed model in single-channel speech enhancement, the experiment used TIMIT as the speech dataset and Noisex-92 as the main noise dataset. TIMIT contains 6300 sentences, of which 4620 were used for training and 1680 for testing. The speakers used for training and testing did not overlap, thus ensuring the effectiveness of the evaluation under speaker mismatch conditions. Regarding noise, four types of noise were selected during the training phase: multi-person speaking noise, factory noise 1, destroyer engine noise, and destroyer control room noise. In the testing phase, in addition to the trained noise, pink noise and factory noise 2 were additionally introduced to examine the generalization ability under noise type mismatch conditions. In the data construction process, the training set was generated by randomly mixing the first 60% of noisy samples with training speech, with the mixed signal-to-noise ratio (SNR) evenly distributed within the range of -5 to 15*dB*, for a total duration of approximately 12.5 hours. The validation set consisted of the middle 20% of the noisy samples with 280 test sentences. The test set consisted of another 320 test sentences mixed with both trained and untrained noise, and evaluated under five fixed SNR conditions: -5, 0, 5, 10, and 15*dB*. This setup covered both known noise conditions and scenarios with unknown noise and unknown speakers. Furthermore, to evaluate the more challenging channel mismatch generalization ability, the reference paper introduced the CHiME-4 data simulation tool, randomly generating noisy speech test samples from its real-world noise scenarios. Since the speech and noise in CHiME-4 were collected by different devices, a typical channel mismatch scenario was formed. This setup significantly improved the completeness of the experimental design and provided more convincing evidence for the “generalization analysis.”

5.2 Experimental Environment and Evaluation Metrics

Regarding the experimental environment, all speech and noise samples were resampled to 16 kHz before mixing. The basic model and most comparative experiments used a 32*ms* frame length and a 16*ms* frame shift for STFT analysis; in the lightweight and two-level model stages, a 20*ms*/10*ms* configuration was further used to reduce computational load. Input features included a 257-dimensional logarithmic power spectrum and 514-dimensional complex spectrum real and imaginary part features, selected according to the model structure. The Adam optimizer was used during training, with weight initialization using the Xavier method, combined with batch normalization and Dropout to stabilize the training process. Regarding evaluation metrics, this paper primarily uses PESQ and STOI to assess the quality and intelligibility of the enhanced speech. In the comparison between phase enhancement and higher-order models, SDR, model parameter count, and FLOPs were also introduced to jointly evaluate the noise reduction effect and model cost. The final comparison, further presents the Real-Time Frequency (RTF) on different deployment platforms, thus expanding the experimental evaluation from a "performance dimension" to an "efficiency dimension." Therefore, the experimental environment and indicator design in this chapter have three characteristics: first, it covers conventional speech quality and intelligibility evaluation; second, it covers generalization evaluation under mismatch conditions; and third, it covers engineering implementation indicators such as parameter scale, computational load, and real-time performance. This ensures that the experimental results not only prove the model's "effectiveness" but also its "usability."

5.3 Analysis of Enhancement Results under Different Noise Conditions

For the basic MS-TCN model, experimental results show that compared to the classic recursive structure model, it not only has advantages in noise reduction performance but also requires fewer parameters and less computation, indicating that multi-scale convolutional temporal modeling is cost-effective in general noise environments in Fig. 5. Furthermore, after introducing complex spectrum and phase enhancement, PhaseDCN significantly improves speech quality and intelligibility under low signal-to-noise ratio (SNR) noise conditions and exhibits stronger robustness under speaker, noise, and channel mismatch conditions. Finally, in the MT-in-MS model, because the first stage jointly learns the IRM and complex spectrum, and the second stage further learns the prior SNR and integrates it with the traditional statistical model, the model achieves the best overall performance under all three mismatch conditions. The referenced paper explicitly points out that MT-in-MS achieves a significant performance gain in average PESQ, with particularly outstanding advantages under the most challenging channel mismatch conditions in Table 2.

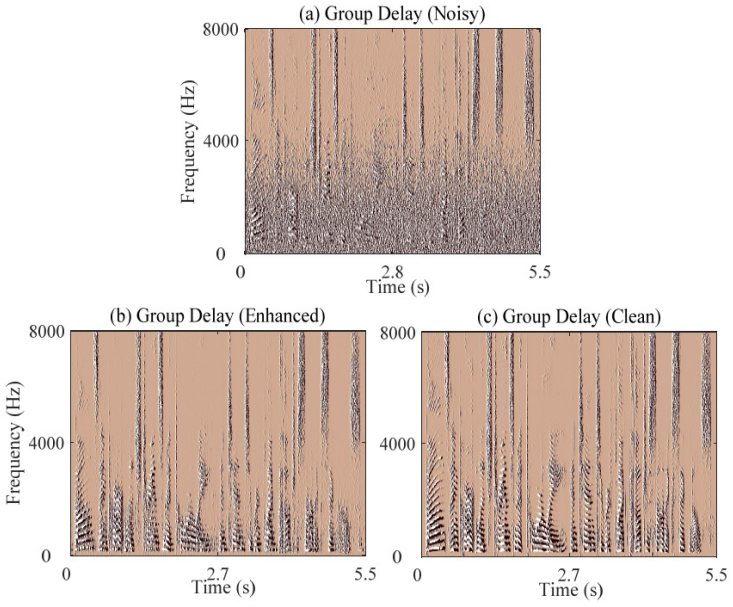

Fig. 5. The group delay of speech signal

Table 2. Averaged STOI results obtained for noisy and enhanced speech in seen and unseen noise cases under different SNR conditions

Method	-5 dB	0 dB	5 dB	10 dB	15 dB	Avg.
Noisy	0.547	0.663	0.772	0.861	0.924	0.753
FNN-SE	0.627	0.756	0.841	0.889	0.915	0.806
LSTM-SE	0.629	0.764	0.846	0.892	0.917	0.81
GRU-SE	0.645	0.767	0.845	0.891	0.915	0.813
TCN-SE	0.704	0.811	0.868	0.9	0.918	0.84
MSTCN-SE-1	0.713	0.821	0.88	0.915	0.935	0.853
MSTCN-SE-2	0.72	0.827	0.89	0.931	0.958	0.865
Noisy	0.592	0.704	0.807	0.887	0.941	0.786
FNN-SE	0.596	0.733	0.828	0.885	0.913	0.791
LSTM-SE	0.607	0.749	0.843	0.894	0.919	0.802
GRU-SE	0.638	0.764	0.845	0.892	0.916	0.811
TCN-SE	0.7	0.81	0.87	0.902	0.92	0.84
MSTCN-SE-1	0.727	0.827	0.885	0.918	0.938	0.859
MSTCN-SE-2	0.729	0.836	0.898	0.937	0.961	0.872

5.4 Comparative Experiments with Traditional Methods

Strictly speaking, Chapter 5 of the reference paper primarily compares the differences between advanced deep learning models, including GCRN, DDAEC, MS-TCN, PhaseDCN, Im-PhaseDCN, and MT-in-MS, rather than a point-by-point comparison of older traditional algorithms. Therefore, if you insist on retaining the current title, it is recommended to state in the text as “Comparative Experiments with Typical Baseline Methods and Advanced Models”; if minor adjustments are permissible, it is even more recommended to change it to “Comparative Experiments with Typical Baseline Models”. The results show that MT-in-MS achieves better PESQ and STOI performance than advanced models such as GCRN and DDAEC in all three mismatch scenarios; furthermore, compared to MS-TCN in Chapter 3 and PhaseDCN in Chapter 4, MT-in-MS further integrates the advantages of different objectives, demonstrating a higher overall energy efficiency ratio. The paper also points out that in terms of model size and complexity, MT-in-MS, with only 4.8M parameters and 9.6M FLOPs, significantly outperforms several high-complexity models. Therefore, this section can be written as follows: First, select statistically typical methods or representative deep learning baselines for comparison; then, analyze from four dimensions: speech quality, intelligibility, generalization, and computational complexity; finally, point out that the model in this paper does not achieve performance improvement simply by “building a larger model,” but rather achieves true optimization in target design, structural design, and two-stage fusion mechanism.

5.5 Ablation Experiments and Model Generalization Analysis

Ablation experiments are mainly used to verify the independent contributions of each component of the model. The conclusions show that using IRM, complex spectrum, or prior signal-to-noise ratio alone is insufficient to achieve optimal performance in all mismatch scenarios; only by reasonably integrating these three methods can stronger generalization ability be obtained in Fig. 6.

In terms of generalization analysis, this paper focuses on three types of mismatch, speaker mismatch, noise mismatch, and channel mismatch. Speaker mismatch tests the model’s adaptability to speaker features outside the training set; noise mismatch tests the model’s robustness to unseen noise types; and channel mismatch further tests the model’s generalization performance across devices and acquisition conditions. Experiments show that as the target design moves from single-target to multi-target and from single-stage to two-stage, the model’s performance gradually improves in the three mismatch scenarios, indicating that the fusion of multi-target learning and statistical models has a substantial effect on improving generalization. Therefore, this section should not simply list experimental tables, but should highlight a conclusion: multi-target joint learning is not about “adding more outputs,” but rather about significantly enhancing the model’s adaptability in complex real-world scenarios through the complementary relationships between different targets. This is also the fundamental logic behind the reference paper’s progression from MS-TCN to PhaseDCN, and then to MT-in-MS in Table 3.

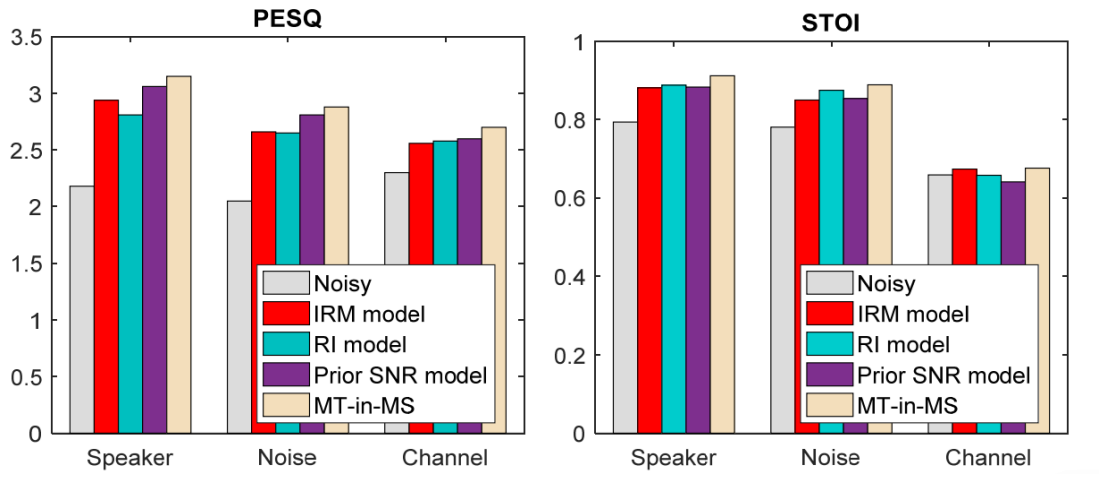

Fig. 6. Averaged PESQ and STOI results for different sub models and MT-in-MS model in speaker, noise and channel mismatch cases

Table 3. Performance comparison of different speech enhancement methods in terms of PESQ, STOI, and SDR

Method	PESQ	STOI	SDR
Noisy	1.72	0.605	-7.64
LSTM-PL	2.03	0.632	-5.11
LSTM-MT	2.01	0.627	-4.54
DNN-IFD	1.85	0.622	-4.64
TCNN	1.92	0.608	-4.25
GCRN	1.74	0.594	-4.29
PhaseDCN	2.05	0.639	-4.2

6 Conclusion

This paper investigates a method for improving single-channel speech signals using deep neural networks. Considering the challenges of complex noisy environments, insufficient model generalization ability, and high computational costs in practical applications, this paper systematically explores model construction, method optimization, and experimental verification from the perspectives of time-frequency representation, temporal modeling, multi-objective learning, and model efficiency. First, this paper constructs a framework for improving single-channel speech signals using deep neural networks. This framework establishes a nonlinear mapping from noisy speech signals to clean speech signals by converting noisy speech signals into time-frequency domain features. Based on this, the model can effectively learn the spectral structure of speech and noise, thus laying the foundation for subsequent improvement and reconstruction. Second, this paper introduces a multi-scale convolution strategy to enhance the network's temporal modeling capability. Compared with traditional recurrent structures, the convolution framework achieves a better balance between contextual modeling and computational efficiency. Furthermore, by combining spectral masking and spectral mapping, the model can fully utilize the complementary advantages of different objectives, thereby improving speech quality and intelligibility in noisy environments. To further enhance the model's robustness under mismatch conditions, phase-related information is introduced, and a multi-objective optimization strategy is applied. By jointly learning amplitude and phase-related features, this model improves the ability to recover speech details and suppress non-stationary noise. A two-stage optimization process combining neural network learning and statistical modeling further enhances the model's generalization ability while reducing the number of parameters and computational complexity. Finally, extensive experiments were conducted using standard speech and noise datasets under speaker, noise, and channel mismatch conditions. Results show that the proposed method outperforms various baseline and comparative methods in PESQ, STOI, and related target metrics. The optimized model achieves a better balance between performance improvement and efficiency, demonstrating significant application potential in practical speech communication systems. This paper validates the effectiveness of the deep neural network-based single-

channel speech signal enhancement method from three aspects: model design, method improvement, and experimental evaluation. The results indicate that multi-criteria learning, temporal convolutional modeling, and phase-aware enhancement are effective methods for improving the noise reduction performance, robustness, and ease of use of single-channel speech enhancement system.

References

[1] M. Tiwari, M. Tiwari, Voice - How humans communicate?, Journal of Natural Science, Biology, and Medicine 3(1) (2012) 3-11.
https://doi.org/10.4103/0976-9668.95933

[2] S. Latif, J. Qadir, A. Qayyum, M. Usama, S. Younis, Speech Technology for Healthcare: Opportunities, Challenges, and State of the Art, IEEE Reviews in Biomedical Engineering 14(2021) 342-356.
https://doi.org/10.1109/RBME.2020.3006860

[3] J.-Y. Li, L. Deng, Y.-F. Gong, R. Haeb-Umbach, An Overview of Noise-Robust Automatic Speech Recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing 22(4)(2014) 745-777.
https://doi.org/10.1109/TASLP.2014.2304637

[4] N. Das, S. Chakraborty, J. Chaki, N. Padhy, N. Dey, Fundamentals, present and future perspectives of speech enhancement, International Journal of Speech Technology 24(4)(2021) 883-901.
https://doi.org/10.1007/s10772-020-09674-2

[5] Y.-Y. Li, D.-A. Ramli, Advances in Time-Frequency Analysis for Blind Source Separation: Challenges, Contributions, and Emerging Trends, IEEE Access 11(2023) 137450-137474.
https://doi.org/10.1109/ACCESS.2023.3338024

[6] A. Kern, R. Stoop, Principles and Typical Computational Limitations of Sparse Speaker Separation Based on Deterministic Speech Features, Neural Computation 23(9)(2011) 2358-2389.
https://doi.org/10.1162/NECO_a_00165

[7] A.-K. Maddirala, R.-A. Shaik, Separation of Sources From Single-Channel EEG Signals Using Independent Component Analysis, IEEE Transactions on Instrumentation and Measurement 67(2)(2018) 382-393.
https://doi.org/10.1109/TIM.2017.2775358

[8] S. Basak, H. Agrawal, S. Jena, S. Gite, M. Bachute, B. Pradhan, M. Assiri, Challenges and Limitations in Speech Recognition Technology: A Critical Review of Speech Signal Processing Algorithms, Tools and Systems, CMES - Computer Modeling in Engineering and Sciences 135(2)(2023) 1053-1089.
https://doi.org/10.32604/cmes.2022.021755

[9] X.-P. Zhang, Thresholding neural network for adaptive noise reduction, IEEE Transactions on Neural Networks 12(3) (2001) 567-584.
https://doi.org/10.1109/72.925559

[10] J.-M. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, Developments and directions in speech recognition and understanding, Part 1 [DSP Education], IEEE Signal Processing Magazine 26(3)(2009) 75-80.
https://doi.org/10.1109/MSP.2009.932166

[11] T. Gerkmann, M. Krawczyk-Becker, J. Le Roux, Phase Processing for Single-Channel Speech Enhancement: History and recent advances, IEEE Signal Processing Magazine 32(2)(2015) 55-66.
https://doi.org/10.1109/MSP.2014.2369251

[12] S.-C. Li, L.-D. Xu, X.-H. Wang, Compressed Sensing Signal and Data Acquisition in Wireless Sensor Networks and Internet of Things, IEEE Transactions on Industrial Informatics 9(4)(2013) 2177-2186.
https://doi.org/10.1109/TII.2012.2189222

[13] F. Wang, J. Liu, Networked Wireless Sensor Data Collection: Issues, Challenges, and Approaches, IEEE Communications Surveys & Tutorials 13(4)(2011) 673-687.
https://doi.org/10.1109/SURV.2011.060710.00066

[14] Z.-Q. Luo, C.-J. Li, L.-D. Zhu, A Comprehensive Survey on Blind Source Separation for Wireless Adaptive Processing: Principles, Perspectives, Challenges and New Research Directions, IEEE Access 6(2018) 66685-66708.
https://doi.org/10.1109/ACCESS.2018.2879380

[15] Y.-Y. Li, J.-D. Wang, H.-Y. Zhao, C. Wang, Z.-C. Ma, Unraveling Mixtures: A Novel Underdetermined Blind Source Separation Approach via Sparse Component Analysis, IEEE Access 12(2024) 14949-14963.
https://doi.org/10.1109/ACCESS.2024.3349427

[16] G.-B. Giannakis, V. Kekatos, N. Gatsis, S.-J. Kim, H. Zhu, B.-F. Wollenberg, Monitoring and Optimization for Power Grids: A Signal Processing Perspective, IEEE Signal Processing Magazine 30(5)(2013) 107-128.
https://doi.org/10.1109/MSP.2013.2245726

[17] S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25(4) (2017) 692-730.

https://doi.org/10.1109/TASLP.2016.2647702
[18] L. Pan, The Importance of Deep Learning Models in Speech Signal Processing: Fundamentals, Strategies, and Future Research Directions, International Journal of Speech Technology 28(2)(2025) 443-459. https://doi.org/10.1007/s10772-025-10194-0
[19] B. Gerhart, Incentives and Pay for Performance in the Workplace, Advances in Motivation Science 4(2017) 91-140. https://doi.org/10.1016/bs.adms.2017.02.001
[20] Z.-B. Lin, L. Zhou, X.-J. Qiu, A Composite Objective Measure on Subjective Evaluation of Speech Enhancement Algorithms, Applied Acoustics 145(2019) 144-148. https://doi.org/10.1016/j.apacoust.2018.10.002
[21] C. Jannu, M. Burra, S.-D. Vanambathina, V. Parisae, C.-V.-M. Krishna, G.-L. Madhumati, Single Channel Speech Enhancement Using a Complex Dual-Path Multi Axial Transformer with Frequency Prompt, Circuits, Systems, and Signal Processing 44(6)(2025) 4224-4257. https://doi.org/10.1007/s00034-025-03010-2
[22] C.-H. Goay, N.-S. Ahmad, P. Goh, Temporal Convolutional Networks for Transient Simulation of High-Speed Channels, Alexandria Engineering Journal 74(2023) 643-663. https://doi.org/10.1016/j.aej.2023.05.059
[23] J. Wang, N. Saleem, T.-S. Gunawan, Towards Efficient Recurrent Architectures: A Deep LSTM Neural Network Applied to Speech Enhancement and Recognition, Cognitive Computation 16(3)(2024) 1221-1236. https://doi.org/10.1007/s12559-024-10288-y
[24] J.-C. Liu, Y.-L. Zhao, Improved Generalization Performance of Convolutional Neural Networks with LossDA, Applied Intelligence 53(11)(2023) 13852-13866. https://doi.org/10.1007/s10489-022-04208-6