Speech Signal Enhancement Using Progressive Learning and Dense Connected LSTM Networks

Nian Liu
Jiangsu College of Safety Technology, the Department of Basic Education, Xuzhou 221000, China
E-mail: liunian1616@163.com

Keywords: progressive learning, dense connection, long short-term memory, applied language, speech enhancement

Received: February 8, 2025

Aiming at the deficiencies of traditional speech signal enhancement models in dealing with long-term dependencies and noise filtering, an application speech signal enhancement model based on progressive learning and dense connection strategies is proposed. This method takes the long short-term memory network structure as the core and realizes the gradual enhancement of noisy speech through layer-by-layer learning and processing. The experimental results showed that this model exhibited excellent enhancement performance in different signal-to-noise ratio environments. In a -5dB signal-to-noise ratio environment, the short-term objective clarity of the research method reached 0.930, which was 4.1% higher than that of delayed neural networks. Moreover, under the 10dB condition, the short-term objective clarity score further increased to 0.957. The distortion signal ratio of the source signal has increased from 2.31 at -5dB to 14.81 at 10dB, indicating the model's ability in noise suppression and signal reconstruction. The assessment score of speech quality perception increased from 1.86 at -5dB to 3.13 at 10dB, and the word error rate decreased to 27.31%, which was 2.47% lower than that of the classical long short-term memory network. The research results show that the proposed model has strong robustness and a good speech enhancement effect when dealing with speech signals with a low signal-to-noise ratio, providing a new solution for the field of applied language processing.

Povzetek: Predlagan je pristop za zmanjšanje hrupa v govoru, ki združuje postopno učenje z gosto povezanimi rekurentnimi mrežami dolgoročnega spomina; model po plasteh čisti signal, ohranja dolge odvisnosti in krepi razumljivost za jezikovne sisteme.

1 Introduction

The popularity of mobile devices and smart homes has made the quality of speech enhancement signals more important for user experience. At the same time, people's dependence on applied language is becoming increasingly severe. However, there are many types of noise in real life, and different types of noise inevitably affect the clarity of speech signals, including noise from automobiles and industries [1]. This noise affects the clarity of speech signals and poses a challenge to the long-term dependence of speech recognition systems. The long-term dependence problem makes it difficult for the model to effectively extract useful information when dealing with speech signals disturbed by noise, thereby reducing the accuracy and user experience of the final speech interaction. High-quality speech signals are crucial for ensuring the accuracy of speech recognition and the effectiveness of user interaction. Therefore, efficient speech enhancement technology is one of the key technologies for the development of applied languages [2-3]. Ochieng P reviewed the Deep Neural Network (DNN) techniques currently used for speech enhancement and separation and conducted a comprehensive analysis model training. DNN had feasibility in speech signal enhancement [4]. Richter J et al. proposed a diffusion process based on stochastic differential equations and reversed the process from a mixture of noisy speech and Gaussian noise. Then, they made adjustments to the network architecture to improve speech enhancement performance. Finally, the experiment verified that the method has a good speech enhancement effect [5]. Zhang Q et al. used time frame attention and frequency channel attention to explicitly generate two-dimensional attention maps with significant T-F speech distributions based on positional information. The effectiveness of this model as a front-end for downstream speech recognition tasks has been demonstrated, and it significantly improved the system's robustness to noise conditions [6]. Bie X et al. designed an unsupervised speech enhancement algorithm. This algorithm combined the prior training of DVAE speech based on non-negative matrix factorization with a noise model, and derived a Variational Expectation Maximization (VEM) algorithm for speech enhancement, achieving good results [7]. The specific summary of the above research is shown in Table 1.

Table 1: Literature summary table

References	Research method	Research advantages	Research disadvantages
Reference [4]	DNN technology is used for speech enhancement and separation	Focus on a comprehensive review and applicability assessment of deep learning technologies	It is mainly a retrospective study, lacking specific experimental verification and performance evaluation
Reference [5]	Diffusion process method based on stochastic differential equations	Improving the performance of voice enhancement by enhancing the network architecture is innovative	Complex mathematical models may lead to difficulties in the implementation and understanding of the models
Reference [6]	The combination of attention to the time frame and the frequency channel	The generation of two-dimensional attention maps using position information improves the robustness under noisy conditions	The need for a complex attention mechanism may make the training process of the model complicated
Reference [7]	Unsupervised speech enhancement algorithm, combining non-negative matrix factorization and VEM algorithm	It is applicable to various noisy environments and has a good enhancement effect	The accuracy of unsupervised learning is limited by the type of noise and may not be able to handle all noise situations

Although there has been some progress in speech signal enhancement research in recent years, there are still many shortcomings, especially in terms of robustness and long-term dependence when dealing with noisy environments. Some of the current research focuses on the use of DNNs or specific mathematical models for speech signal enhancement. However, when confronted with extremely high noise environments, such as low Signal-To-Noise Ratios (SNR) or complex noise types, the robustness of these models appears insufficient, resulting in unstable performance in dynamic scenes. Although attention mechanisms can enhance the effectiveness of speech signals in certain situations, their ability to capture long-term dependencies still has limitations when processing long-term sequence data. This poses difficulties for speech processing tasks that require long-term contextual information. In addition, although the developed unsupervised learning methods have shown some effectiveness, the training process may not guarantee the reliability of the enhancement effect due to the lack of labeled data. Especially when encountering new types of noise, the generalization ability of the model will also be limited. Although various studies have demonstrated different enhancement effects, the performance of the system is relatively lacking, making the quality of the results unclear and making it difficult to comprehensively evaluate the actual effectiveness of existing methods. In response to the aforementioned research gaps, this study proposes an applied speech signal enhancement model based on Long Short-Term Memory and Progressive Learning and Dense Connection strategy (LSTM-PLDC). The study assumes that this new model can effectively improve the quality of speech signals in complex noise environments, thereby enhancing the clarity and comprehensibility of speech signals. The main purpose of the research is to verify the enhanced performance of the LSTM-PLDC model under different SNR conditions and evaluate its robustness when dealing with extreme noise environments. This method can effectively enhance useful speech features in the data and suppress background noise by layer by layer strengthening of the speech signal. Furthermore, by adopting the LSTM structure in combination with the dense connection strategy, the context information for a long time can be better retained. This will enhance the model's processing ability for long-term dependencies, thereby further improving the speech enhancement effect. Another important contribution of the research is the improvement of the model’s robustness. Compared with the existing methods, the LSTM-PLDC model shows stronger anti-noise interference ability in various SNR environments, which can effectively improve the speech quality and avoid signal distortion caused by excessive noise reduction.

2 Methods and materials

2.1 Construction of a speech enhancement signal enhancement model based on LSTM

Applied language processing is an important direction in natural language processing, which focuses on applying linguistic theories and techniques to practical problems and application scenarios. Speech enhancement processing covers a wide range, including speech recognition, speech synthesis, machine translation, speech enhancement, and other directions. Among them, speech signal enhancement is a core component of applied language processing, with the main task of improving the

quality of speech signals and ensuring the accuracy and effectiveness of subsequent processing [8]. The primary task of speech signal enhancement is to remove background noise, echo, and other interfering factors, ensuring the purity and clarity of speech signals, which is fundamental and critical for all speech-based applications. Currently, the commonly used speech signal enhancement model is based on LSTM, which has advantages such as long-term dependency processing, noise removal enhancement, and temporal data processing [9-10]. In LSTM, the core part mainly includes input, output, and hidden layers. Unlike traditional Recurrent Neural Networks (RNNs), the hidden layer structure of LSTM is more complex, adding a memory unit called "cell state". This unit performs traditional neural computation and also manages and maintains long-term information through cell state management, as shown in Figure 1.

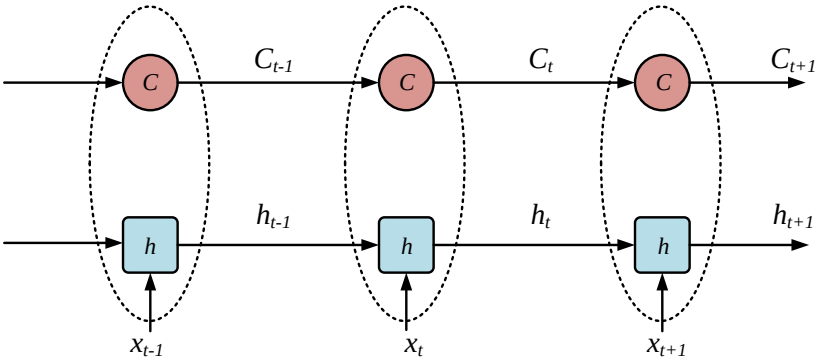

Figure 1: Hidden layer structure of LSTM

In Figure1, h is the calculation module of the model, and C is the internal state of the model. LSTM has the advantage of processing long time series data. Because of the existence of h and C, these modules work together at every moment of model operation to process and save key variables such as current input values, previous output values, previous cell states, current output, and current cell states. This mechanism ensures that information can be stored and transmitted in the model for a long time, effectively solving the problem of long-term dependencies. In the LSTM hidden layer, there are also input gates (I), forget gates (f), and output gates (o). The mathematical expression of the LSTM's f is shown in formula (1) [11].

$$g(x) = \sigma Wx + b \quad (1)$$

In formula (1), W is the weight vector, b is the bias term, and σ is the sigmoid function. In the architecture of LSTM, the output gate mainly extracts and selects key information from the cell state as the current output value, ensuring that only useful information for the current task is transmitted while shielding irrelevant or noisy information. During the speech enhancement process, the output gate can dynamically adjust the intensity and importance of the output information. The core mechanism of LSTM covers the forward propagation of information, the backpropagation of errors, and the process of optimizing network parameters through gradient descent. The forward propagation of information is similar to that of traditional neural networks, using the interaction between neurons for calculation and passing the results. The input gate calculation of the LSTM hidden layer is represented as shown in formula (2).

$$i_t = \sigma W_i \times [h_{t-1}, x_t] + b_i \quad (2)$$

In formula (2), h_{t-1} is the output value of the upper layer. Formula (2) can dynamically adjust the model's response to input information in speech enhancement, selectively introducing useful speech features and suppressing unnecessary noise. The input unit state at this moment can be obtained through formula (2), and its expression is shown in formula (3).

$$c_t = \tanh W_c \times [h_{t-1}, x_t] + b_c \quad (3)$$

In formula (3), c_t is the element state of the previous layer model, and $\tanh$ is the hyperbolic tangent function. Formula (3) ensures that the model can effectively store and manage speech feature information in long time series, enhancing its ability to handle long-term dependencies. The state of the current layer model can be represented by formula (4).

$$c_t = g_t \cdot c_{t-1} + c_t \cdot i_t \quad (4)$$

Formula (4) selectively outputs key information to ensure that only the information useful for the current task is conveyed. The output gate expression of the LSTM model is shown in formula (5).

$$o_t = \sigma W_o \times [h_{t-1}, x_t] + b_0 \quad (5)$$

In formula (5), o_t represents the model's output gate. Based on the above model construction, the specific structure of LSTM is shown in Figure 2.

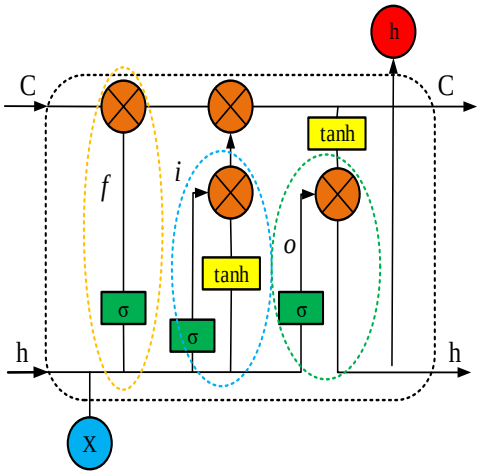

Figure 2: Unit Structure of LSTM

In Figure 2, the input gate (i) plays a crucial role in selectively introducing the current input information into the cell state. It can effectively filter out unnecessary noise information and only retain signals that are useful for speech enhancement. The main function of the forget gate (f) in the model is to determine which information in the cell state needs to be forgotten or retained. For processing speech signals with long time series data, it can prevent old information from interfering with the current processing and ensure that the model focuses on important current information. The model can represent the mathematical formula of the final result of LSTM based on the output gate and the current unit state, as shown in formula (6).

$$h_t = o_t \cdot \tanh c_t \quad (6)$$

In formula (6), h_t is the final result of the model. Formula (6) can ensure the purity and clarity of the output signal by generating an enhanced speech signal. In the training process of LSTM, error backpropagation, and gradient descent are key steps to correct and optimize network weights. Error backpropagation can calculate the gradient of each parameter in the model. For LSTM, this process unfolds over time, calculating gradients at each past time step to ensure that the model can learn and adjust parameters to minimize errors. In speech enhancement models, this means that the model can gradually learn how to effectively filter noise and enhance speech signals. Gradient descent adjusts the parameters of the model based on the gradient information obtained from error backpropagation. By continuously iterating and updating parameters, the gradient descent algorithm gradually approaches the optimal solution, minimizing the error function of the model. The principle of error transmission can be found in formula (7) [12].

$$\delta_k^T = \prod_{j=k}^{t-1} \delta_{o,j}^T W_{oh} + \delta_{f,h}^T W_{fh} + \delta_{i,h}^T W_{ih} + \delta_{c,j}^T W_{ch}. \quad (7)$$

In formula (7), δ_k^T is the propagation error. The calculation of its gradient descent is determined based on the sum of the gradients of the "input gate", "forget gate" and "output gate" in the model at this moment. The weight gradient formulas of each gate are shown in formula (8).

$$\frac{\partial E}{\partial W_{ox}} = \delta_{o,t} x_t^T \quad \frac{\partial E}{\partial W_{ix}} = \delta_{i,t} x_t^T \quad \frac{\partial E}{\partial W_{fx}} = \delta_{f,t} x_t^T \quad \frac{\partial E}{\partial W_{cx}} = \delta_{\tilde{c},t} x_t^T \quad (8)$$

In formula (8), E is the loss function. Therefore, after obtaining the overall error, the network will perform the backpropagation step of the error. During this process, errors will be dispersed to various neural units based on existing weights and thresholds, and then the gradient descent strategy will be used to adjust the weights and perform forward propagation to generate outputs.

2.2 Construction of progressive language enhancement model based on LSTM

The LSTM-based speech signal enhancement model constructed above has significant advantages in dealing with long-term dependencies and noise filtering, but the model still has some shortcomings. For example, the computational complexity of the model is high, resource consumption is high, and gradients are prone to vanishing and exploding [13-15]. To address the limitations of the above model, this study introduces a progressive strategy into LSTM, with the main objective of enhancing speech signals layer by layer. Figure 3 shows the basic idea of progressive speech enhancement.

In Figure 3, the kernel of progressive speech enhancement starts with simple tasks and gradually learns and solves more complex problems. This method is particularly suitable for the task of converting noisy speech signals into clear speech. The specific implementation method is to decompose the entire problem into multiple small steps by increasing the SNR. Each small step focuses on improving the SNR of the input speech. Specifically, under low SNR conditions, the model first focuses on removing background noise, improving the basic structure of speech signals, and gradually enhancing the comprehensibility of signals at various stages. With the gradual improvement of SNR, the model can identify and enhance speech features more accurately in the subsequent processing stage, thereby achieving higher-quality speech signal output. This study introduces a progressive strategy into the speech enhancement model, and its structure is shown in Figure 4.

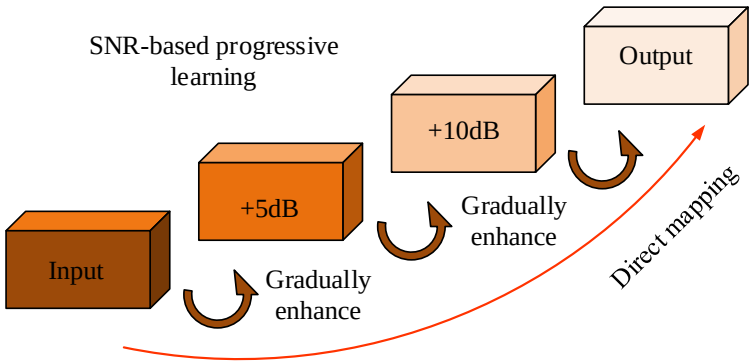

Figure 3 Basic Strategy Diagram of Progressive Speech Enhancement

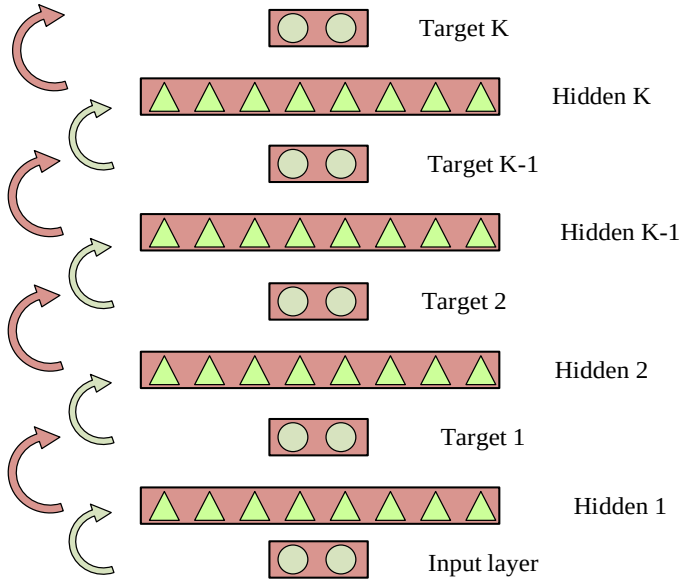

Figure 4: Speech enhancement structure based on progressive strategy

In Figure 4, the model achieves step-by-step optimization and enhancement of noisy speech signals through a multi-level and gradual enhancement method. The input layer receives noisy speech signals and performs preliminary preprocessing. The hidden layer extracts and processes speech features layer by layer, and enhances feature representation ability through nonlinear activation functions. The intermediate target layer learns an intermediate target with a higher SNR than the output of the last layer, gradually increasing the SNR. The target layer uses a linear activation function to generate the final enhanced speech signal. This study uses a weighted multi-objective learning objective function to train the network, as shown in formula (9) [16-17].

$$\begin{cases} e^{PL} = \sum_{k=1}^{K} \beta_k e_k^{PL} \\ e_k^{PL} = \frac{1}{U} \sum_{u=1}^{U} \left\| \mu_k^{PL}(\hat{x}_u^{k-1}, \Lambda_k^{PL}) - x_u^k \right\|_2^2 \end{cases} \tag{9}$$

In formula (9), e is the objective function. K represents the number of hidden layers. $\hat{x}$ denotes the estimated value of the target. x_u^k is the learning objective. U is the number of samples for network structure updates. β is the error weight coefficient. x_u^0 is a noisy speech feature. $\mu_k^{PL}(\hat{x}_u^{k-1}, \Lambda_k^{PL})$ is the target layer network function. Λ_k^{PL} is a paranoid vector.

In the above model structure, if the amount of intermediate target layers gradually increases, the performance of the model may be negatively affected. Meanwhile, the quality of effective information output by the model will decrease as the number of learning objectives increases. Therefore,

this study further adopts the DC approach to improve the model and obtained the LSTM-PLDC. Figure 5 shows the LSTM-PLDC model structure.

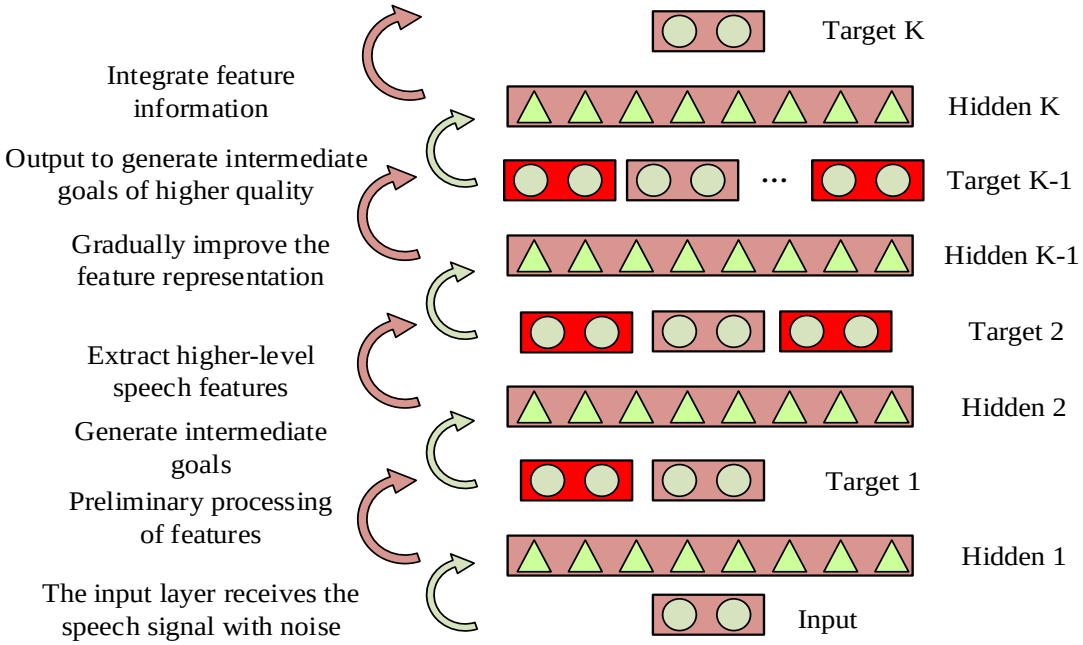

Figure 5: Speech enhancement structure based on LSTM-PLDC

In Figure 5, the design of the target layer aims to gradually optimize the SNR by combining the features output by the input layer and each intermediate layer. At each learning stage, the target layer extracts key information from the output of the previous layer and learns to generate intermediate targets with higher SNR. This mechanism enables the model to gradually improve the quality of speech signals in stages, making each target layer a performance measure and a key link in achieving effective information transmission and gradual enhancement, ultimately improving the clarity and comprehensibility of the overall speech signal. When learning intermediate targets, the model concatenates the original input and the estimated results of each target together, and then inputs them into a sub network. This method enables the sub network to simultaneously obtain the features of the initial noisy speech and the estimated features of speech with different SNRs. Due to the involvement of multiple learning objectives in DC's progressive learning, this study adopts a weighted Minimum Mean Square Error (MMSE) criterion as the objective function. The training and updating of the model are shown in formula (10).

$$\begin{cases} e^{PLD} = \sum_{k=1}^{K} \beta_k e_k^{PLD} \\ e_k^{PLD} = \frac{1}{U} \sum_{u=1}^{U} \left\| \mu_k^{PLD}(\hat{x}_u^0, \hat{x}_u^1, ..., \hat{x}_u^{k-1}, \Lambda_k^{PLD}) - x_u^k \right\|_2^2 \end{cases} \tag{10}$$

Formula (10) uses weighted MMSE criterion for multi-objective optimization to ensure that the model can effectively enhance speech signals under different SNR conditions. When the model processes different SNRs, speech distortion may occur due to excessive noise reduction. Other intermediate targets close to the target layer have lower SNR and can better preserve speech. This study uses formula (11) to solve the above phenomenon.

$$\hat{x}_n = \begin{cases} \frac{\hat{x}_u^K + \hat{x}_u^{K-1}}{2} & K = 2 \\ \frac{\hat{x}_u^K + \hat{x}_u^{K-1} + \hat{x}_u^{K-2}}{3} & K \geq 3 \end{cases} \tag{11}$$

By using formula (11), the model can learn speech features with different SNRs through multiple learning objectives, thereby better balancing denoising and speech feature preservation. The implementation process based on the LSTM-PLDC model proposed mainly includes the input layer receiving noisy speech signals and performing preliminary preprocessing operations for subsequent feature extraction. The hidden layer extracts and processes speech features layer by layer, and enhances feature representation ability through nonlinear activation functions. Each intermediate target layer learns an intermediate target with a higher SNR than the output of the previous layer, gradually increasing the SNR. By training and updating the model, the loss of each intermediate objective is calculated, and the total objective function is weighted and summed based on the weight

coefficients. According to the overall objective function, to backpropagate and update the model parameters to gradually optimize the model and achieve the best speech enhancement effect.

3 Results

3.1 Training analysis based on progressive language enhancement model

To verify the model's performance, the study adopts the TIMIT corpus for model training and testing. The TIMIT speech dataset has diverse speech samples and high-quality annotations, making it an ideal choice for evaluating speech enhancement effects. During the data preprocessing process, the audio file is first loaded and converted into a unified sampling rate. Then, the audio signal is subjected to frame segmentation and feature extraction, and the MEL frequency cepstral coefficient is adopted as the input feature. When simulating noise conditions in a real environment, various background noises are mixed with the target speech based on different SNRs. The processed data are divided into the training set, the validation set, and the test set to ensure the generalization ability of the model. Model training involves designing an architecture based on LSTM, combining dense connections and progressive learning mechanisms, using weighted MMES criterion as the loss function, and updating weights using Adam optimizer. During the training process, the early stop strategy is applied to prevent overfitting and ensure the performance optimization of the model on the validation set. 1660 sentences are randomly selected from the TIMIT corpus as the target speech, and the target speech is divided into training speech and validation speech in an 8:2 ratio. Then, the target speech is mixed with various types of noise based on different SNRs, including -5dB, 0dB, 5dB, and 10dB. The selection of mixed noise comes from five types of noise in noise -92 dB, including noisy noise, factory noise, spectral noise, vehicle noise, and horn noise. This study uses Perceived Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI), Source-to-Distortion Ratio (SDR), and Word Error Rate (WER) as evaluation metrics for the model.

To ensure the optimal performance of the LSTM-PLDC model, a series of strategies are adopted in the selection and tuning process of hyperparameters. Firstly, the initial setting of the learning rate is 0.001, and the learning rate attenuation method is adopted during the training process. Specifically, every 10 training cycles (Epochs), the learning rate is reduced to 90% of the original to promote the model's convergence. The Dropout rate is set to 0.5. This is to effectively prevent the overfitting phenomenon of the model and ensure the generalization ability of the network by randomly discarding some neurons. The number of layers of the model is set to 6 LSTM units to balance the model depth and computational complexity, while maintaining a good capture ability for long-term dependencies. The batch size is selected as 64 to enable the effective utilization of diverse data in each iteration while ensuring the stability of the training process. The dense connection part adopts the "DenseNet" structure, which specifically connects the output features of each layer with the features of the previous layer, effectively enhancing information flow and feature reuse. The initialization method used is He initialization, which can effectively avoid the problems of gradient vanishing and explosion. In LSTM-PLDC, different learning objectives have a certain impact, as displayed in Figure 6.

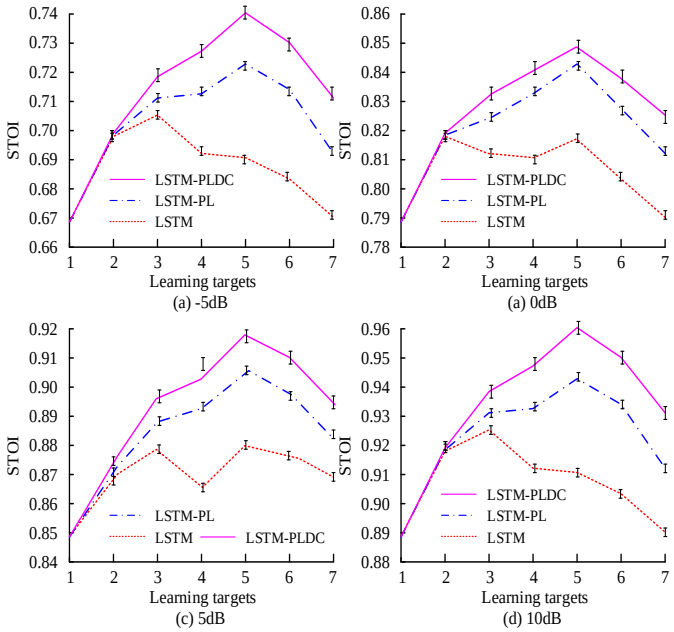

Figure 6: The impact of the number of target learning layers on the model

Figs. 6 (a) - (d) show the impact of the learning target layer on the model under SNR environments of -5dB, 0dB, 5dB, and 10dB. In Figure 6 (a), under the condition of -5dB, when the Learning target layer is 2, the differences among LSTM, Long Short-Term Memory-Progressive Learning (LSTM-PL), and LSTM-PLDC are relatively small. The STOI values of the three models are all around 0.715. When the learning objective layer gradually increases to 4, the STOI value of LSTM-PLDC reaches 0.740, the STOI value of LSTM-PL reaches 0.722, and the STOI value of LSTM is 0.693. In Figure 6 (b), under the condition of 0dB, the three algorithms achieve the best STOI value when the learning target layer is 5. Among them, the STOI value of LSTM-PLDC is 0.853, LSTM-PL is 0.846, and LSTM is 0.828. In Figure 6 (c), under the condition of 5db, the optimal STOI value for LSTM-PLDC is 0.918, LSTM-PL reaches 0.915, and LSTM is 0.903. In Figure 6 (d), the optimal STOI values under 10db conditions are 0.956 for LSTM-PLDC, 0.949 for LSTM-PL, and 0.948 for LSTM. This indicates that increasing the number of learning target layers can improve the short-term objective clarity of the model, and LSTM-PLDC exhibits the best clarity improvement effect under various SNR conditions. This also indicates that the LSTM-PLDC model has stronger robustness and higher speech quality when dealing with noisy speech. Similarly, the hidden layer structure of a model can also affect its performance. This study analyzes the number of different hidden layers and nodes, as shown in Figure 7.

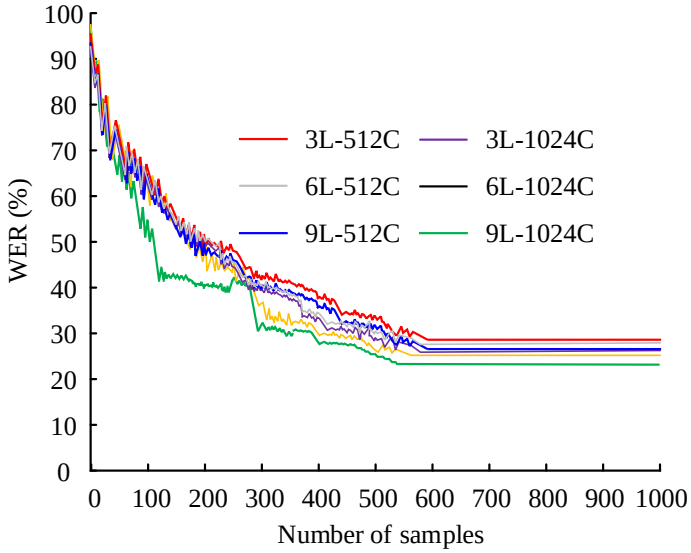

Figure 7: Performance impact of model structure

In Figure 7, the quantity of hidden layers is divided into 3, 6, and 9 layers, with 513 and 1024 nodes respectively. Under the same number of nodes and the same number of hidden layers, the higher the number of hidden layers and nodes, the lower the WER value of the model. Among them, when the number of hidden layers is 9 and the number of nodes is 1024, the WER of the model is about 24.82%. Compared to the model structure with 3 hidden layers and 513 nodes, its WER decreased by 3.09%. Figure 7 shows that increasing the number of hidden layers and nodes has a positive impact on model performance, and the combination of the two has a more significant effect. When designing a speech model, increasing the number of hidden layers and nodes appropriately can significantly improve the performance, reduce WER, and thus improve the accuracy of speech recognition. To further analyze the model's performance, this study trained different models on speech and obtained comparative results of the performance of different models, as shown in Table 2.

Table 2: Performance of models with different structures under training speech

Model structure	Model size (M)	WER (%)	Running time (s)
3L-512C	19.1	28.02	372
6L-512C	28.5	26.97	985
9L-512C	37.2	25.85	936
3L-1024C	46.3	25.88	948
6L-1024C	73.3	25.34	1769
9L-1024C	112.1	24.87	2051

In Table 2, the "L" in the model structure represents the number of hidden layers. "C" represents the number of nodes. When the hidden layers increase from 3 to 9, with 512 nodes, WER decreases from 28.02% to 25.85%, and runtime increases from 372 s to 936 s. Under 1,024 nodes, WER decreases from 25.88% to 24.87%, and runtime increases from 948 s to 2,051 s. When the nodes increase from 512 to 1024: under 3 hidden layers, the model size increases from 19.1 M to 46.3 M, and under 9 hidden layers, it increases from 37.2 M to 112.1 M. This indicates that increasing the number of hidden layers helps to reduce WER and improve model performance, but also increases computation time and resource requirements. To evaluate the impact of each model component on performance, an ablation study is conducted to compare three different models: standard LSTM, LSTM-PL, and LSTM-PLDC. The study evaluates the model through performance indicators (STOI, SDR and PESQ) under multiple SNR conditions, and the results are shown in Table 3.

Table 3: Ablation experiments of the model

Evaluation index	LSTM	LSTM-PL	LSTM-PLDC
-5dB STOI	0.678	0.811	0.930
0dB STOI	0.812	0.846	0.957
5dB STOI	0.885	0.915	0.979
10dB STOI	0.914	0.949	0.986
-5dB SDR	2.31	3.24	4.12
0dB SDR	5.14	6.88	8.67
5dB SDR	7.89	9.22	11.45
10dB SDR	10.67	13.24	14.81
-5dB PESQ	1.77	1.80	1.86
0dB PESQ	2.41	2.55	2.85
5dB PESQ	2.85	2.92	3.10
10dB PESQ	3.01	3.06	3.13

Table 3 presents the results of the ablation study, which highlights the performance enhancements achieved by integrating progressive learning and dense connection strategies into the model. From the standard LSTM model to LSTM-PL and ultimately to LSTM-PLDC, the evaluation metrics (STOI, SDR, and PESQ) show a clear improvement trend under all SNR conditions. For instance, in the -5dB SNR condition, STOI scores improves from 0.678 with the standard LSTM to 0.930 with the LSTM-PLDC model, indicating a significant enhancement in speech intelligibility. Similar trends are observed in SDR and PESQ scores, with LSTM-PLDC achieving a maximum SDR of 4.12 and a PESQ score of 1.86 under the same -5dB condition. These results underscore the importance of the added components in refining the model's ability to enhance speech quality and intelligibility in noisy environments, ultimately showing that each enhancement contributes substantially to overall performance.

3.2 Performance testing based on progressive language enhancement model

The research model obtains a good model structure in training speech, and now the performance of the model is analyzed by verifying the speech. This study uses Time-Delay Neural Network (TDNN) for comparative analysis [18]. Figure 8 shows the average STOI of each model on five types of noise in the test set.

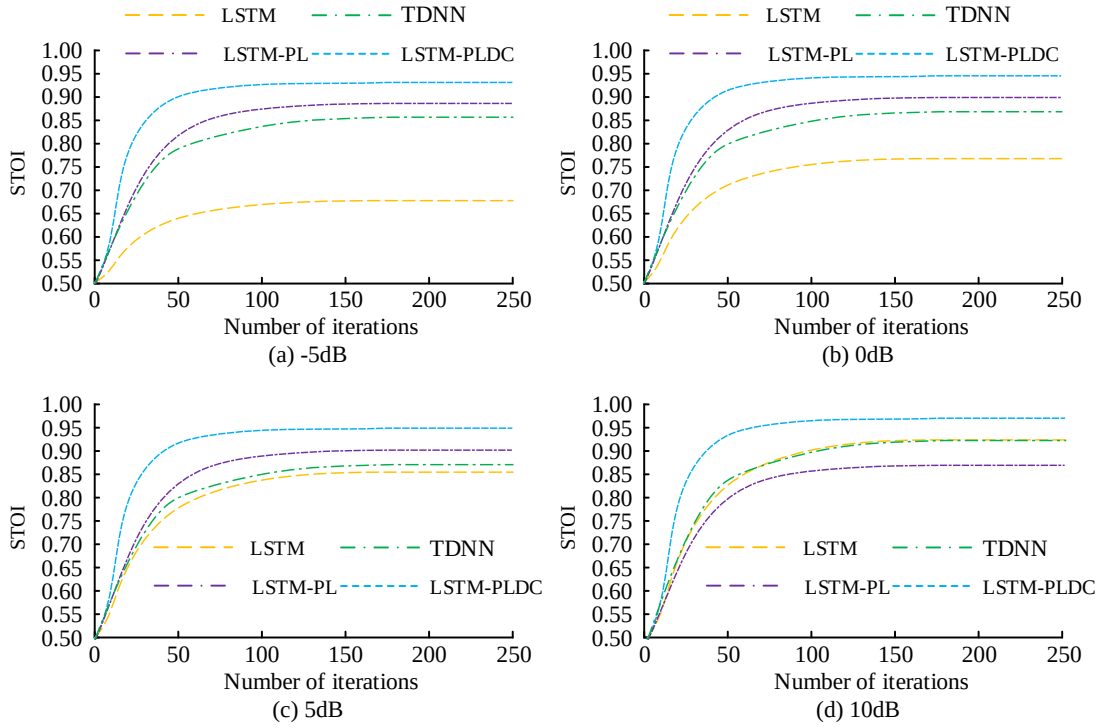

Figure 8: STOI results of each model in noise

Figs. 8 (a) -8 (d) show the average STOI values in SNR environments of -5dB, 0dB, 5dB, and 10dB. Regarding the results of the average STOI from -5dB to 10dB: LSTM increases from 0.678 to 0.930, LSTM-PL increases from 0.811 to 0.851, TDNN increases from 0.889 to 0.919, and LSTM-PLDC increases from -0.930 to 0.957. LSTM-PLDC has a 4.1% improvement compared to TDNN in the lowest SNR environment. Therefore, LSTM-PLDC has more effective noise suppression and speech enhancement effects in the model. Figure 9 shows the SDR of the model.

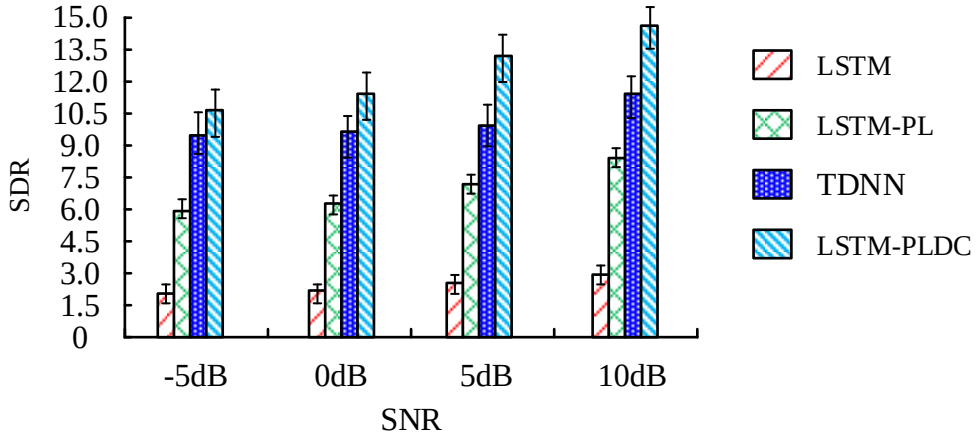

Figure 9: SDR results of four models in noise

Figure 9 shows the average SDR values in SNR environments of -5 dB, 0 dB, 5 dB, and 10 dB. The performance of LSTM gradually improves throughout the entire noise level range, from 2.31 at -5 dB to 10.67 at 10 dB. TDNN further increases the SDR value from 3.24 under low noise to 13.24 under high noise, indicating that TDNN may have specific advantages in extracting and preserving speech signals. LSTM-PLDC is 0.28 higher than TDNN, 1.57 higher at 10 dB, and has the highest SDR value at all noise levels. Overall, LSTM-PLDC still exhibits good performance in low SNR environments, indicating that the model has strong resistance to noise interference. Moreover, LSTM-PLDC can more accurately reconstruct speech signals and reduce noise components. Figure 10 shows the PESQ of each model on different types of noise in the test set.

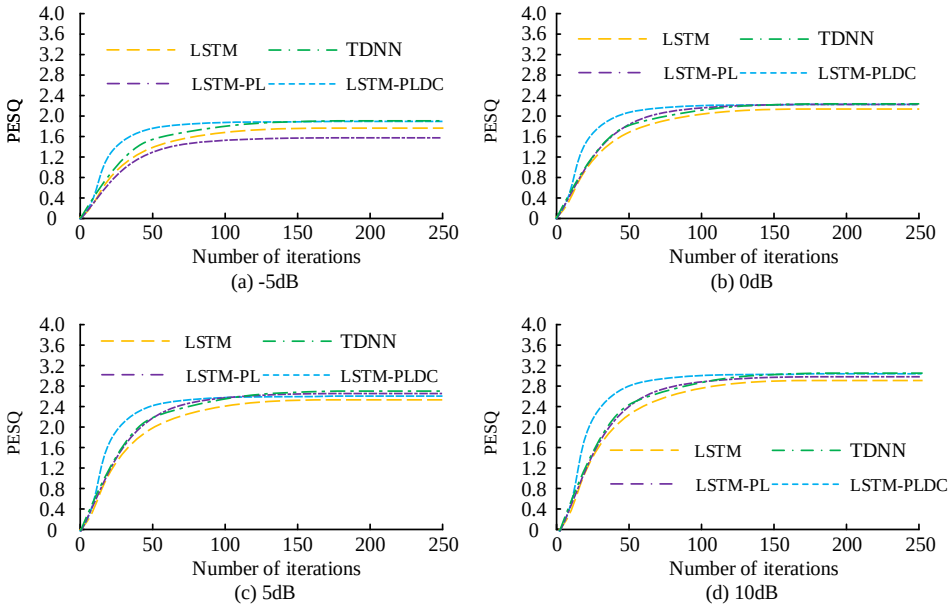

Figure 10: PESQ results of each model in noise

In Figure 10, the PESQ score increases from 1.77 to 3.01 in LSTM, from 1.70 to 3.06 in LSTM-PL, from 1.76 to 3.10 in TDNN, and from -1.86 to 3.13 in LSTM-PLDC in a noisy environment ranging from -5dB to 10dB. The performance of the LSTM-PLDC is superior to other models under -5dB noise conditions, indicating that the model effectively improves signal quality and reduces distortion. In the validation set, the WERs of each model are shown in Figure 11.

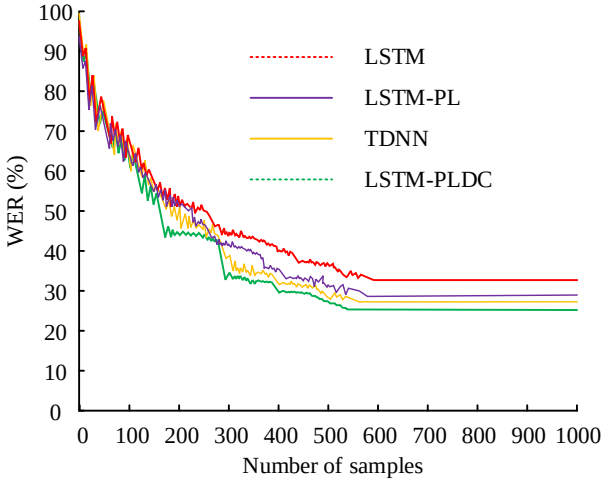

Figure 11: WERs of various models in the validation set

In Figure 11, the WER of LSTM is 36.45%, and that of the LSTM-PL is 30.33%. Compared to LSTM, LSTM-PL shows a significant decrease in WER, indicating that the progressive improvement method can effectively enhance speech enhancement performance. The WER of TDNN is 29.78%, while that of LSTM-PLDC is 27.31%. The WER of LSTM-PLDC further decreased by 3.02% compared to LSTM-PL, indicating that the progressive model using DC has better speech enhancement effect. Compared with TDNN model, LSTM-PLDC still has some advantages, which indicates that the improved model is feasible and progressiveness. To evaluate the robustness of the LSTM-PLDC model under real-world noise conditions, the study compares it with multiple baseline models to verify the model's validity. The noise datasets selected are CHiME and Aurora. Among them, the CHiME dataset is a noise dataset used for speech recognition, containing speech samples in different environments, such as homes, coffee shops, etc. The Aurora dataset is specifically designed to test the performance of speech recognition systems under various noise conditions, including different types of white noise and other environmental noises. The baseline models for comparison include Statistical Model-based Enhanced Noise Filtering (ENF), Wiener Filter (WF), DNN, Convolutional Neural Network (CNN), TDNN, and standard LSTM. The

experimental indicators adopted are STOI, SDR, PESQ, and MMSE. The performance comparison results of the model under real-world noise conditions are shown in Table 4.

Table 4: Performance verification of the model under real-world noise conditions

Model	LSTM	LSTM-PL	LSTM-PLDC	DNN	CNN	TDNN	WF	ENF
-5dB STOI	0.670	0.810	0.930	0.700	0.720	0.710	0.650	0.680
0dB STOI	0.790	0.826	0.950	0.783	0.795	0.810	0.760	0.775
5dB STOI	0.870	0.895	0.980	0.890	0.860	0.900	0.830	0.855
10dB STOI	0.900	0.940	0.990	0.910	0.880	0.920	0.850	0.880
-5dB SDR	2.100	3.100	4.120	2.550	2.800	2.650	1.900	2.200
0dB SDR	4.500	5.790	8.100	5.120	5.400	5.900	4.000	4.400
5dB SDR	6.810	8.100	11.300	7.670	8.000	8.200	6.200	7.300
10dB SDR	9.500	12.000	14.000	10.500	11.000	11.800	8.800	9.900
-5dB PESQ	1.700	1.800	1.870	1.750	1.760	1.790	1.600	1.730
0dB PESQ	2.200	2.400	2.900	2.300	2.340	2.360	2.100	2.250
5dB PESQ	2.600	2.850	3.100	2.750	2.800	2.880	2.500	2.700
10dB PESQ	2.850	3.010	3.120	3.000	3.050	3.020	2.700	2.950
-5dB MMSE	0.045	0.042	0.038	0.043	0.041	0.044	0.049	0.046
0dB MMSE	0.035	0.032	0.027	0.037	0.033	0.036	0.040	0.038
5dB MMSE	0.028	0.025	0.020	0.026	0.028	0.024	0.030	0.029
10dB MMSE	0.020	0.018	0.015	0.022	0.019	0.021	0.027	0.024

Table 4 presents the performance verification results of the LSTM-PLDC model under real-world noise conditions, with multiple baseline models for comparison. Under all the tested SNR conditions, the LSTM-PLDC model performs well in the STOI, SDR, and PESQ indicators. Especially under the condition of -5dB, its STOI reaches 0.930, and the result is higher than that of other models. Meanwhile, the SDR and PESQ scores of LSTM-PLDC are 4.120 and 1.870 under the condition of -5dB, both demonstrating superior noise reduction and speech quality performance. Compared with the traditional model, LSTM-PLDC reduces the MMSE value compared with other baseline models under all SNRs, indicating its higher accuracy in signal reconstruction. These results indicate that LSTM-PLDC has stronger robustness and effectiveness in adapting to real-world environmental noise, fully verifying its successful application in speech enhancement tasks. To enhance the statistical validation of the results, the study conducts a statistical significance test and clarifies the trade-off between model complexity and performance. The specific results are shown in Table 5.

Table 5: Trade-offs between model complexity and performance

Model structure	LSTM	LSTM-PL	LSTM-PLDC	DNN	TDNN
-5dB STOI	0.67	0.81	0.93	0.7	0.71
p value	/	0.003	<0.001	0.005	0.004
95% confidence interval (lower limit)	/	0.794	0.916	0.679	0.689
95% confidence interval (upper limit)	/	0.826	0.944	0.721	0.731
Number of hidden layers	2	4	6	3	5
Computational complexity (number of parameters)	1500	3000	5500	2200	4800
Performance improvement (%)	/	20.9	14.8	4.5	5.5

Table 5 shows the trade-off results of different model structures in terms of performance and complexity. Under the condition of a -5dB SNR, the STOI of the LSTM-PL model reaches 0.81, which is significantly increased by 20.9% compared with the standard LSTM, and its *p*-value is 0.003, indicating that this improvement is statistically significant. The LSTM-PLDC further increases the STOI to 0.93, and at the same time shows a *p*-value of <0.001. Moreover, the lower and upper limits of the confidence interval are 0.916 and 0.944, demonstrating the robustness of the model under noisy conditions. It is worth noting that the hidden layer number of LSTM-PLDC is 6 layers and the number of parameters reaches 5,500, showing a relatively high computational complexity. The number of parameters for DNN and TDNN models is 2200 and 4800, and the performance improvement is relatively limited. This result highlights the need for a balance between model complexity and performance. Although adding a hidden layer enhances the speech enhancement effect, it also brings higher demands for computing resources.

4 Discussion

The LSTM-PLDC model performed well in the speech signal enhancement task, especially with certain improvements in robustness and speech quality. The comparison with the relevant literature summary table clearly showed the advantages and disadvantages of the model. Firstly, in terms of STOI, the LSTM-PLDC model achieved 0.930 SNR at -5 dB, which was significantly higher than the result reported in Reference [6]. However, the STOI performance of the unsupervised learning method in Reference [7] under the same conditions was more limited. The main difference lies in the combination of progressive learning strategy and dense connection architecture in the LSTM-PLDC model, which effectively enhances the ability to extract useful features from speech signals and filters out background noise well. This design enables the model to maintain the clarity of speech in a high-noise background and has stronger adaptability in dynamic scenes, showing higher robustness compared to traditional methods. Furthermore, SDR is an important indicator for evaluating the effect of speech enhancement. In the experiments of the research, the SDR of the LSTM-PLDC model increased from 2.31 at -5 dB to 14.81 at 10 dB, which was superior to the relevant results in Reference [5]. This indicates that LSTM-PLDC performs outstandingly in the ability to effectively suppress noise and reconstruct clear signals. However, due to the characteristics of unsupervised learning, the method in Reference [7] may not be able to fully capture the signal uncertainty under certain noise conditions, resulting in the restoration effect not reaching the best. This result highlights the competitive advantage of LSTM-PLDC in complex noise environments. Priyanka S S et al. achieved speech enhancement through DNNs and CNNs. These models typically have a fixed feedforward structure and can also be trained to classify speech with different SNRs. LSTM-PLDC has the characteristics of progressive learning and dense connections, making it more suitable for sequential data and dynamically changing background noise [19]. Garg A optimized the structure of the LSTM model by using the attention mechanism and conducted speech recognition analysis on the optimized model, achieving certain results in the results [20]. However, compared with the LSTM-PLDC model, there is still a certain gap in its speech enhancement performance.

Although the LSTM-PLDC model performs well in multiple parameters, it also has some limitations at the same time. Firstly, due to its high model complexity, large computational resources, and high time overhead, its practical application is limited to a certain extent, especially under resource constraints. Although a higher number of hidden layers and nodes can reduce the power, it also brings a certain computational burden. Furthermore, although the model performs well in most noisy environments, it still faces challenges at extreme noise levels, such as extremely low SNRs or very complex noise backgrounds. This is because when the background noise intensity is too high, the model may encounter information loss, leading to difficulties in extracting and restoring speech features, thereby affecting the final enhancement effect.

5 Conclusion

In response to the limitations of traditional LSTM models in processing long time series speech signals, this study constructed a progressive language signal enhancement model based on LSTM and obtained the LSTM-PLDC model by introducing DC. Through comparative experiments, the proposed LSTM-PLDC had significant advantages in speech enhancement performance compared to LSTM, LSTM-PL, and TDNN. Experiments have shown that LSTM-PLDC had significant robustness in processing noisy speech in low SNR environments and could more accurately reconstruct speech signals, reducing the impact of noise interference. Although the LSTM-PLDC model has demonstrated superior speech enhancement performance in various noise environments, there are still some limitations. Firstly, the complexity of the model leads to a high demand for computing resources and training time, which may limit its application in a resource-constrained environment. Furthermore, although it performs well in most common noise environments, the robustness and generalization ability of the model still need to be further verified under extreme noise conditions or in the absence of noise types. Future research can focus on optimizing the model architecture to reduce computational complexity. Meanwhile, transfer learning and unsupervised learning techniques can be utilized to enhance the model's adaptability to new noisy environments. Furthermore, exploring the combination of context information and

higher-level audio feature extraction techniques may further enhance the speech enhancement effect and promote progress in this field.

References

[1] Tesch K, Gerkmann T. Insights into deep non-linear filters for improved multi-channel speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 31(1): 563-575. https://doi.org/10.1109/TASLP.2022.3221046

[2] Fang Y, Wang Y. Cross modal sentiment analysis of image text fusion based on Bi LSTM and B-CNN. Informatica, 2024, 48(21): 95-111. https://doi.org/10.31449/inf.v48i21.6767

[3] Yan X, Yao L, Zhou D. Optimizing Tourism Service Quality in 5G Multimedia Environments Using Deep Learning: A Model-Based Empirical Study. Informatica, 2024, 48(22): 147-161. https://doi.org/10.31449/inf.v48i22.6806

[4] Ochieng P. Deep neural network techniques for monaural speech enhancement and separation: state of the art analysis. Artificial Intelligence Review, 2023, 56(Suppl 3): 3651-3703. https://doi.org/10.48550/arXiv.2212.00369

[5] Richter J, Welker S, Lemercier J M, Lay B, Gerkmann T. Speech enhancement and dereverberation with diffusion-based generative models. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31(1): 2351-2364. https://doi.org/10.1109/TASLP.2023.3285241

[6] Zhang Q, Qian X, Ni Z, Nicolson A, Ambikairajah E, Li H. A time-frequency attention module for neural speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 31(1): 462-475. https://doi.org/10.1109/TASLP.2022.3225649

[7] Bie X, Leglaive S, Alameda-Pineda X, Girin L. Unsupervised speech enhancement using dynamical variational autoencoders. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30(3): 2993-3007. https://doi.org/10.48550/arXiv.2106.12271

[8] Wang J, Saleem N, Gunawan T S. Towards efficient recurrent architectures: A deep LSTM neural network applied to speech enhancement and recognition. Cognitive Computation, 2024, 16(3): 1221-1236. https://doi.org/10.1007/s12559-024-10288-y

[9] Huang P, Wu Y. Teacher-student training approach using an adaptive gain mask for lstm-based speech enhancement in the airborne noise environment. Chinese Journal of Electronics, 2023, 32(4): 882-895. https://doi.org/10.23919/cje.2022.00.307

[10] Pandey A, Wang D L. Self-attending RNN for speech enhancement to improve cross-corpus generalization. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30(1): 1374-1385. https://doi.org/10.48550/arXiv.2105.12831

[11] Zhu Q S, Zhang J, Zhang Z Q, et al. A joint speech enhancement and self-supervised representation learning framework for noise-robust speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31(1): 1927-1939. https://doi.org/10.1109/TASLP.2023.3275033

[12] Gupta A, Purwar A. Speech refinement using Bi-LSTM and improved spectral clustering in speaker diarization. Multimedia Tools and Applications, 2024, 83(18): 54433-54448. https://doi.org/10.1007/s11042-023-17017-x

[13] Chan D Y, Wang J F, Chin H T. A new speaker-diarization technology with denoising spectral-LSTM for online automatic multi-dialogue recording. Multimedia Tools and Applications, 2024, 83(15): 45407-45422. https://doi.org/10.1007/s11042-023-17283-9

[14] Pashaian M, Seyedin S, Ahadi S M. A novel jointly optimized cooperative DAE-DNN approach based on a new multi-target step-wise learning for speech enhancement. IEEE Access, 2023, 11(1): 21669-21685. https://doi.org/10.1109/ACCESS.2023.3250820

[15] Abdelhamid A A, El-Kenawy E S M, Alotaibi B, Amer G, Abdelkader M, Ibrahim A, Eid M. Robust speech emotion recognition using CNN+ LSTM based on stochastic fractal search optimization algorithm. IEEE Access, 2022, 10: 49265-49284. https://doi.org/10.1109/ACCESS.2022.3172954

[16] Parvathala V, Andhavarapu S, Pamisetty G, et al. Neural comb filtering using sliding window attention network for speech enhancement. Circuits, Systems, and Signal Processing, 2023, 42(1): 322-343. https://doi.org/10.1007/s00034-022-02123-2

[17] Wang H, Zhang X, Wang D L. Fusing bone-conduction and air-conduction sensors for complex-domain speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30(1): 3134-3143. https://doi.org/10.1109/TASLP.2022.3209943

[18] Tiwari M, Verma D K. Gender recognition in text-independent speaker identification using MFCC, spectrogram, Bi-LSTM, and rat swarm evolutionary algorithm optimization. International Journal of Speech Technology, 2025, 28(1): 245-260. https://doi.org/10.1007/s10772-025-10176-2

[19] Priyanka S S, Kumar T K. Multi-channel speech enhancement using early and late fusion convolutional neural networks. Signal, Image and Video Processing, 2023, 17(4): 973-979. https://doi.org/10.1007/s11760-022-02301-4

[20] Garg A. Speech enhancement using long short term memory with trained speech features and adaptive wiener filter. Multimedia Tools and Applications, 2023, 82(3): 3647-3675. https://doi.org/10.1007/s11042-022-13302-3