Intelligent Automation & Soft Computing

Tech Science Press

DOI: 10.32604/iasc.2023.037354
Article

Mobile Communication Voice Enhancement Under Convolutional Neural Networks and the Internet of Things

Jiajia Yu*

Department of Electronic and Computer Engineering, Southeast University Chengxian College, Nanjing, 210088, China
*Corresponding Author: Jiajia Yu. Email: yujiajia8@126.com
Received: 31 October 2022; Accepted: 07 February 2023

Abstract: This study aims to reduce the interference of ambient noise in mobile communication, improve the accuracy and authenticity of information transmitted by sound, and guarantee the accuracy of voice information delivered by mobile communication. First, the principles and techniques of speech enhancement are analyzed, and a fast lateral recursive least square method (FLRLS method) is adopted to process sound data. Then, the convolutional neural networks (CNNs)-based noise recognition CNN (NR-CNN) algorithm and speech enhancement model are proposed. Finally, related experiments are designed to verify the performance of the proposed algorithm and model. The experimental results show that the noise classification accuracy of the NR-CNN noise recognition algorithm is higher than 99.82%, and the recall rate and F1 value are also higher than 99.92. The proposed sound enhancement model can effectively enhance the original sound in the case of noise interference. After the CNN is incorporated, the average value of all noisy sound perception quality evaluation system values is improved by over 21% compared with that of the traditional noise reduction method. The proposed algorithm can adapt to a variety of voice environments and can simultaneously enhance and reduce noise processing on a variety of different types of voice signals, and the processing effect is better than that of traditional sound enhancement models. In addition, the sound distortion index of the proposed speech enhancement model is inferior to that of the control group, indicating that the addition of the CNN neural network is less likely to cause sound signal distortion in various sound environments and shows superior robustness. In summary, the proposed CNN-based speech enhancement model shows significant sound enhancement effects, stable performance, and strong adaptability. This study provides a reference and basis for research applying neural networks in speech enhancement.

Keywords: Convolutional neural networks; speech enhancement; noise recognition; deep learning; human-computer interaction; Internet of Things

1 Introduction

Currently, speech recognition is becoming increasingly mature in the artificial intelligence (AI) field. In 2006, deep learning took off again, with many scholars bringing it to the field of speech so that speech recognition could allow humans and machines to talk more accurately, just as humans do. Voice interaction has become an important interface for intelligent interaction in many fields. However, there are often various noises in the translation environment of speech interaction in real life. Once a speech system with good recognition performance is exposed to a noisy noise environment, the recognition performance will be reduced. With the continuous development of technology, AI technology has become increasingly mature. Research on information interaction is abundant. Lv et al. (2017) combined virtual reality (VR), geographic information systems, or geo-information systems (GISs) to achieve immersive human-computer interaction (HCI) in geography classrooms [1]. As an important technology for AI to interact with the outside world, speech recognition technology has also been studied by many people and has become the main interface for HCI. However, various background noises in real life seriously interfere with speech interaction. It filters the sound signal in speech interaction and affects the clarity of communication speech so that the interactive effect is affected and the accuracy of speech is greatly reduced [2,3].

Mobile communication refers to communication between moving bodies or between a moving body and a fixed body. One or both of the communication parties must be in motion, including land, sea, and air mobile communication. The moving body can be a person or a moving object such as a car, train, ship, or radio. The frequency bands used in mobile communications cover low frequency, intermediate frequency, high frequency, very high frequency, and ultrahigh frequency. The mobile communication system is composed of a mobile station, a base station, and a mobile switching office. Most of the daily long-distance speech communication process will be disturbed by different degrees of noise. In many cases, noise will be generated in the environment where speech is placed due to human or equipment shortcomings or faults, which will have a great impact on the collection of speech signals and the transmission of electronic communication equipment [4,5]. For example, interference from existing signals will affect the call quality in a public telephone environment at a transportation hub.

Since it is difficult to eliminate the strong background noise in the noise source from the environment, speech enhancement technology becomes imperative to reduce the noise of the sound signal. It aims to reduce the influence of noise in an environment with low speech distortion or low signal-to-noise ratio (SNR) and to acquire a relatively pure speech signal. The listener can basically understand the meaning of speech because it can make the speech signal clearer and reduce the interference of the sound. Some audio systems have excellent speech enhancement mechanisms that can be used to replace human work [6]. For example, improving the speech recognition interaction in cars can realize the operation of cars and reduce the driving fatigue of drivers. Robots can use audio systems with powerful antinoise functions to operate in dangerous environments, which are difficult for people to realize. In short, speech enhancement technology provides strong support and help for human speech and speech recognition systems [7]. Xu et al. (2020) emphasized this point of view in their research, suggesting that speech is easily affected by additive noise in a real environment [8]. To reduce the additive noise, the noisy speech based on the spectrograms was enhanced using the nonnegative matrix factorization (NMF) and sparse NMF (SNMF) algorithms to capture more information at a high sampling rate. Compared with the high sampling rate, the range of objective human vocal organs is limited to low frequency values. Therefore, describing low frequencies requires higher resolution. To address the issue that the short-time Fourier transform (STFT)-based traditional spectrogram may lack the frequency resolution of lower frequencies, the constant Q transform (CQT) is adopted to provide high resolution for low frequencies. Perceived speech quality assessment (PESQ)

and short-term target intelligibility (STOI) are used to evaluate the proposed method. Experimental results show that compared with the STFT baseline with a low signal-to-noise ratio, this method demonstrates better enhancement capabilities. Li et al. (2020) [9] used a new type of approximate message passing (AMP)-based speech enhancement algorithm to address the limitations of inaccurate noise estimation in traditional speech enhancement methods. AMP uses the difference between speech sparsity and noise sparsity to remove or silence noise from corrupted speech. The AMP algorithm is used to effectively reconstruct clean speech for speech enhancement. Regarding the correlation between the speech coefficients of adjacent frames, the k-nearest neighbor (k-NN) algorithm is used to learn sparsity, and computational simulations are used to verify the proposed algorithm, which achieves better speech enhancement performance than the traditional language enhancement methods. Islam et al. (2020) proposed a new single-channel speech enhancement algorithm that applies dual-domain transforms, including the dual-tree complex wavelet transform (DTCWT) and short-term Fourier transform (STFT) with SNMF [10]. The first domain belongs to the DTCWT, which is used for time-domain signals to deal with the weakness of signal distortion caused by discrete wavelet packet transform (DWPT) downsampling and transmitting a set of subband signals. The second domain refers to the STFT, which is used for each subband signal and constructs a complex spectrogram. SNMF is applied to the amplitude spectrogram to extract speech components. The proposed algorithm improves objective speech quality and intelligibility in all considered signal-to-noise ratios (SNR)-Generalized dictionary learning (GDL), STFT-CJSR, DTCWT-SNMF, and DWPT-STFT-SNMF. Wang et al. (2021) adopted deep learning models for speech enhancement [11], which is similar to this study. In his research, it was noted that most existing deep learning-based methods use fully convolutional neural networks (CNNs) to capture the time-frequency information of input features. Compared with CNN, it is more reasonable to use a long short-term memory (LSTM) network to capture contextual information on the time axis of features. However, the full LSTM structure has a heavy computational load. To reach a balance between the complexity of the model and the ability to capture time-frequency features, an LSTM-Convolutional-BLSTM encoder-decoder (LCLED) network is proposed for speech enhancement. Additionally, two LSTM parts and convolutional layers are used to model the context information and frequency dimension features, respectively, to obtain higher quality enhanced speech. Li et al. (2022) [12] proposed a two-stage speech enhancement method. First, a signal-to-noise ratio (SNR) classifier is used to classify speech frames into three categories according to different SNR thresholds. Second, three kinds of a dynamic neural network (DNN)-based estimators corresponding to three SNRs were trained to correct the amplitude spectrum. Finally, combined with the phase information of the noise speech, the discrete Fourier inverse transform of the corrected speech spectrum is carried out to realize speech enhancement. Bendoumia et al. (2022) [13] studied the enhancement of speech quality by using a subband form extended dual-sensor sparse adaptive algorithm on account of a forward blind source separation structure. It is proven that the algorithm has a great improvement in the convergence rate and steady-state value.

In summary, the existing research on speech enhancement processing has agreed that it is difficult to eliminate strong background noise from ambient noise sources, so speech enhancement technology is very important to reduce the noise of the sound signal. At present, the use of deep network technology to achieve SNR increase or speech separation for speech enhancement has certain limitations. One of the reasons is that speech enhancement has high requirements for real-time processing, that is, high requirements for the speed of translation. To achieve a more ideal state, the network processing time speed must be fast enough. It takes time to establish a mapping relationship for voice data directly. In addition, there are many types of noise in real life, and the SNR and other features of various audio are different, so the mapping relationships established are different. To learn

a relatively stable and effective model with strong adaptive ability, a huge amount of data training is needed, which is difficult to achieve for ordinary research institutions. Even if the organization has a sufficient data source, the training time of model stable establishment will be relatively long, and the research and development cost will be relatively high. Given the above objective reasons, there are few teams using deep network technology to study speech enhancement algorithms in China, and feasible deep network-based speech enhancement schemes in industry still need to be improved.

Traditional speech enhancement algorithms include spectral subtraction, Wiener filter, and wavelet transform, but they have the following problems during speech enhancement. 1. There is music noise. Although the partial speech enhancement algorithm filters out noise, it also causes partial distortion of the speech signal. 2. There is a limited processing capacity in a nonstationary noise environment. Many algorithms can filter transient static noise well, but they can't filter noise with no periodic characteristics. 3. The adaptive ability is not strong enough. Even for traditional algorithms with good nonstationary noise processing ability, parameter settings are generally empirical values, which can achieve good results in some noise environments, but not in others.

In view of the above shortcomings of traditional speech enhancement algorithms, the following research is implemented. 1. A novel CNN-based noise recognition algorithm is proposed. Noise recognition based on CNN is the first step of the speech enhancement model in this work. Through prenoise recognition work, the best processing parameters for each noise environment can be matched to provide better protection for speech enhancement in each environment, solving the problem that the traditional enhancement algorithm does not have strong adaptive ability. The proposed noise recognition algorithm is based on the different effects of different types of noise on speech signals and identifies the types of environmental noise parameters so that the enhanced model can be applied to different noise environments and improve the adaptive ability of the algorithm. 2. A speech enhancement model integrating CNN and SFTRLS is proposed.

The main contribution and innovation of this work lies in the proposed CNN and noise classification-based SFTRLS-CNN speech enhancement algorithm that can reduce noise and enhance speech information. Different from traditional language enhancement algorithms that are completely based on the CNN algorithm, the proposed algorithm adopts the CNN algorithm and uses a multiclass exponential loss function CNN (SFTRLS-CNN) to construct a speech enhancement model based on stage additive modeling. This algorithm is proven to be superior to the traditional language enhancement model. The objective of this work is to ensure the accuracy of voice transmission information and improve the efficiency of mobile communication.

Section 1 analyses the theory and technology of traditional speech enhancement and summarizes its shortcomings. In Section 2, a new noise recognition algorithm by noise recognition CNN (NR-CNN) is proposed, based on which a speech enhancement processing model is built to recognize and intensify the noise of the speech information. The experimental and control groups are designed to explore the performance. Section 3 analyses the experimental results of this paper. The experimental results prove that the model constructed in this work can adapt to a variety of speech environments and can simultaneously enhance and reduce the noise of a variety of different types of speech signals, and the processing effect is better than that of the traditional voice enhancement model. Section 4 summarizes the research results of this work and explains the deficiencies in the research and the future direction of work.

2 Related Theories

2.1 Related Theories and Analysis on Speech Enhancement

The main content of this study is speech enhancement technology. The development process, related concepts, and technical principles of speech enhancement technology are briefly introduced. From the perspective of the important development process of speech enhancement technology, speech enhancement algorithms are classified according to various processing methods and related optimization algorithms. There are three categories, namely speech enhancement technology based on spectral subtraction and related optimization algorithms, speech enhancement technology based on statistical technology, and speech enhancement technology based on deep learning [14].

The earliest speech enhancement technology refers to the speech enhancement optimization algorithm based on the expansion technology of spectrum subtraction. As the name suggests, spectrum subtraction is used to subtract the spectrum of the noise signal from the spectrum of the noise signal sound. It appeared in the 1960s and was first proposed and improved by Boll. Classical spectrum subtraction measures the noise spectrum estimate by using the spectrum value of the noise in the nonspeech environment and subtracts the frequency spectrum of the noise signal from the frequency spectrum of the noise signal to eliminate noise. Conventional spectral subtraction uses the estimated initial noise power spectrum and the phase spectrum of the noise speech signal to reconstruct the emphasized speech signal [15,16]. Spectral subtraction shows two major advantages. On the one hand, it is easy to implement, and on the other hand, the amount of calculation required to reduce noise using spectral subtraction is very small [17]. However, it still has deficiencies. First, two prerequisites must be met to use spectral subtraction: it is assumed that the noise environment is statistically stable, and it is assumed that the added noise signal and the audio signal are not correlated [18]. Second, when spectral subtraction is adopted to reduce noise, it is easy for followers to concentrate, leaving rhythmic waveform "musical noise", which makes it difficult to completely remove the noise. Many scientists all over the world have conducted research in this area.

In addition to spectral subtraction, another widely used speech enhancement method is the speech enhancement technique based on statistical techniques. There are many methods for such speech enhancement technology, but its theoretical basis uses statistical theory methods, so it is collectively referred to as speech enhancement technology based on statistical technology [19,20], such as the least mean square error (LMSE), short-time amplitude spectrum (SAS) algorithm, and Kalman filtering method. These algorithms are continuously used and optimized by researchers, making the noise reduction effect increasingly stronger and deriving many new conceptual terms. In general, statistical technology-based speech enhancement technology has a wide variety of features, powerful functions, strong adaptability, relatively complete development, and excellent noise reduction effects. Although there are some shortcomings, such as a large amount of calculation, complex extraction of raw signal data, and complex algorithms during its use, it is still widely used in many fields due to its better speech enhancement effect [21,22]. It is constantly being improved by the addition of new elements by researchers in related fields.

The third type of speech enhancement technology is an enhancement technology based on deep learning. With the rapid development of science and technology, deep learning and AI have been applied to many frontier fields, and deep learning-based speech enhancement methods have become increasingly mature [23,24]. The earliest speech enhancement technology based on deep learning was an improved speech enhancement algorithm based on a hidden Markov model (HMM) proposed by Epharim in 1989. The algorithm can convert audio signals and noise signals, construct different HMM models, and repeat iterative calculations using Wiener filtering. The filter performance of the HMM

noise reduction model is superior, which attracts many excellent scientists to study and improve on this basis. In addition to the HMM-based noise reduction technology, speech enhancement technology based on artificial neural networks (ANNs) has also begun to be studied by an increasing number of people. The earliest research on neural network-based noise reduction technology was also conducted in 1989 by Tamura to study the nonlinear relationship between neural networks learning loud sounds and clear sounds. Since then, an increasing number of speech enhancement processing techniques based on neural networks have been researched and utilized.

In the current field of speech noise reduction research, some well-known domestic research institutions and companies have introduced deep learning methods into noise reduction processing, made many contributions, and put them into practical use [25]. However, certain applications that use deep network technology to achieve an increased signal-to-noise ratio (SNR) or voice isolation to improve voice are limited due to the influence of objective factors and technology. One of the reasons is that voice transmission usually requires relatively high real-time processing and faster translation speed. The network processing time must be short to meet real-time work requirements. It takes some time to directly set the audio data mapping relationship. In addition, there are various types of noise that need to be processed in actual work, so it takes more time to process different audio [26]. In addition, learning a relatively stable and effective adaptive model requires training a large amount of data, which is limited by objective conditions and is difficult for general research institutions. Even if an organization has enough data sources, it takes time to establish a stable model, and the research and development cost is relatively high. Therefore, deep learning-based speech enhancement technology must be studied and further improved in many aspects.

2.2 Relevant Technologies

The application of deep learning-based CNNs in speech enhancement processing is analyzed in this work. Therefore, it is necessary to introduce and explain the relevant technical principles.

To build a neural network model of speech information, it must process the speech data first. The speech signal processing methods selected in this study include two types: preemphasis and speech endpoint detection [27].

Preemphasis is a signal processing method that can correct the high-frequency components of the input signal. The technical sources are as follows. There is radiation impedance in the signal field, that is, the reciprocal of the ratio of the sound velocity wave of the speech signal to the sound pressure wave, which is characterized by the influence of lip radiation. Radiation from the lips has a slight effect on the power spectrum of the speech signal. This is more pronounced in the higher frequency bands and less pronounced in the lower frequency bands. If the amplitude of the high-frequency part of the speech signal is small, signal distortion will occur [28]. The purpose of the preemphasis of the audio signal is to eliminate the influence of lip sound radiation and improve the high-frequency sound resolution. The preemphasis equation for sound is as follows:

$$\mathrm{f}(\mathrm{x}) = 1 - a\mathrm{x}^{-1}, 0.9 < a < 1.0 \tag{1}$$

In Eq. (1), α represents the preemphasis coefficient, and x represents the speech frequency signal. Eq. (1) can be transformed by the difference form, and the transformed expression is given as follows:

$$y(\mathrm{n}) = \mathrm{x}(\mathrm{n}) - a\mathrm{x}(\mathrm{n}-1) \tag{2}$$

In Eq. (2) above, n represents the number of intervals after the difference, and the value of α is 0.98.

When the speech signal is enhanced, the speech information should be preprocessed in advance. Speech endpoint detection is such a link. In speech enhancement processing, the speech that must be enhanced often appears in the form of segments. Moreover, the forms of the noise segment and the speech segment are different. Speech endpoint detection can determine the starting point and ending point of a certain segment of speech information [29,30]. The common method of speech endpoint detection is the double-threshold method, and the specific method is as follows.

Step 1: It has to be determined that the threshold T2 is greater than the speech threshold and the audio segment is greater than the T2 speech threshold, which is selected according to the speech energy envelope. The time point corresponding to the intersection of T2 and short-term energy is selected as the range of the starting and ending points of the external speech [31]. After that, the lower threshold T1 is determined regarding the average energy. The speech segment between the two intersection points of the threshold T1 and the energy envelope is selected as the speech segment.

Step 2: The search is started at the intersection of T1 and the envelope obtained in the previous step. If the short-term average zero-crossing value interval is less than the threshold T3, it is regarded as the final speech segment interval.

In this study, the data are processed by recursive least squares (RLS).

The Expectation Least Squares (ELS) algorithm is a fast recursive implementation of the least squares (LS). It is an adaptive filter that performs Wiener filtering by applying adaptive sampling methods and time updates [32]. For normal and smooth sound signals, the RLS filter has the same optimal solution as the Wiener filter. For unstable sound signals, the RLS filter has excellent convergence speed over time, so it has been widely used in speech enhancement [33,34].

However, the calculation speed of RLS method is slow and the calculation effect is unstable. Hence, the FLRLS method is introduced to optimize these two shortcomings, and the converted expression is as follows.

$$\left[\xi^{a}_{f_{\min}}(\mathrm{m},\mathrm{N})\right]^{-1}=\left[\alpha\xi^{a}_{f_{\min}}(\mathrm{m}-1,\mathrm{N})\right]^{-1}-\beta(m,N+1,1)\,\phi^{2}_{o}(m,N+1) \tag{3}$$

In Eq. (3), m represents the beginning time of a certain time period, N represents the end time of a certain time period, $\alpha\xi^{a}_{f_{\min}}(\mathrm{m}-1,\mathrm{N})$ and $\beta(m,N+1,1)$ represent the conversion factors of different moments, $\phi_o(m,N+1)$ represents the filter gain vector of the filter, and α and β are constants greater than 0 and less than 1.

In the above equation, the choice of $\beta(m,N+1,1)$ depends on the strategy to keep the pattern of the error system stable. The stable and fast horizontal RLS algorithm can be obtained by using the prior backward error expression with redundancy about the transformation factor.

In this study, the CNN is combined with the FLRLS method to learn and process speech data information. The CNN unit usually consists of three structures, namely a convolutional layer, a pooling layer, and a fully connected layer. CNN can effectively reduce the number of network parameters through convolution and aggregation operations.

The convolution operation can reduce parameters through local connection and weight sharing, and the local connection method is adopted to move other data groups. The weighted convolution kernel (filter) of the input matrix extracts data features from different positions. This process is called convolution, and the extracted features are called feature maps. Multiple convolution kernels can be used on convolution to generate multiple feature maps representing different input features. The structure of the CNN is shown in Fig. 1.

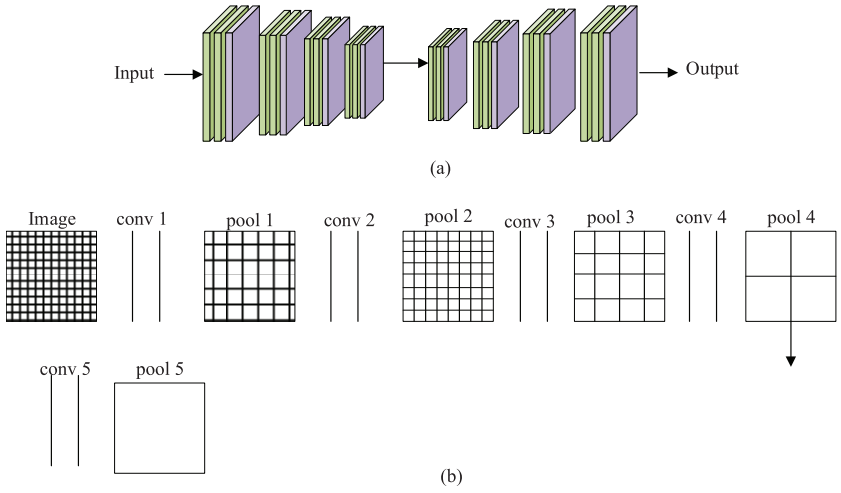

Figure 1: Structure of CNN (a: The structure principle, b: The working principle)

2.3 Construction of the CNN-Based Noise Recognition Algorithm

The noise recognition algorithm is applied in many aspects of audio signal processing. Different life backgrounds will cause different types of noise to have different effects on speech. This noise causes considerable interference in the recognition or coding of the language. The noise recognition algorithm can predetermine the frequency of noise and sound when performing language enhancement processing and use different solutions for different noises. It can effectively improve the efficiency of speech recognition and coding. In some speech enhancement models, it is necessary to preclassify the noise and use the best deep learning model for different types of noise so that the algorithm can process some noisy environments.

The Stagewise Additive Modelling using a Multiclass Exponential loss function (SAMME) based on the Back Propagation Neural Networks (BPNN) (SAMME-BPNN) algorithm is constructed and expanded to construct the process and structure of the NR-CNN noise recognition algorithm [35,36].

The SAMME algorithm is a multiclassification algorithm based on an adaptive enhancement algorithm. The efficiency of recognition can be improved through the classification and integration of target data. This classifier combines multiple BPNN weak classifiers and provides each weak classifier with a different impact index. Then, several weaker classifiers assign higher weights to more accurate classifiers with weighted voting to generate classification results. This method can provide a reference criterion for adjusting the error and prevent the result from being classified as a local optimal solution. Compared with the general BPNN, the SAMME-BPNN algorithm exerts a stronger generalization effect on the data. The calculation process of the SAMME-BPNN algorithm is shown in Fig. 2.

The proposed network combines several BP neural networks as weak classifiers, assigns influence factors to each weak classifier, and assigns higher weights to more accurate classifiers. Eventually, the classification results are generated by weighted voting of multiple weak classifiers to provide evaluation criteria for error adjustment and avoid the final results falling into the local optimal solution. The proposed network has better generalization ability compared with the single BP classification method.

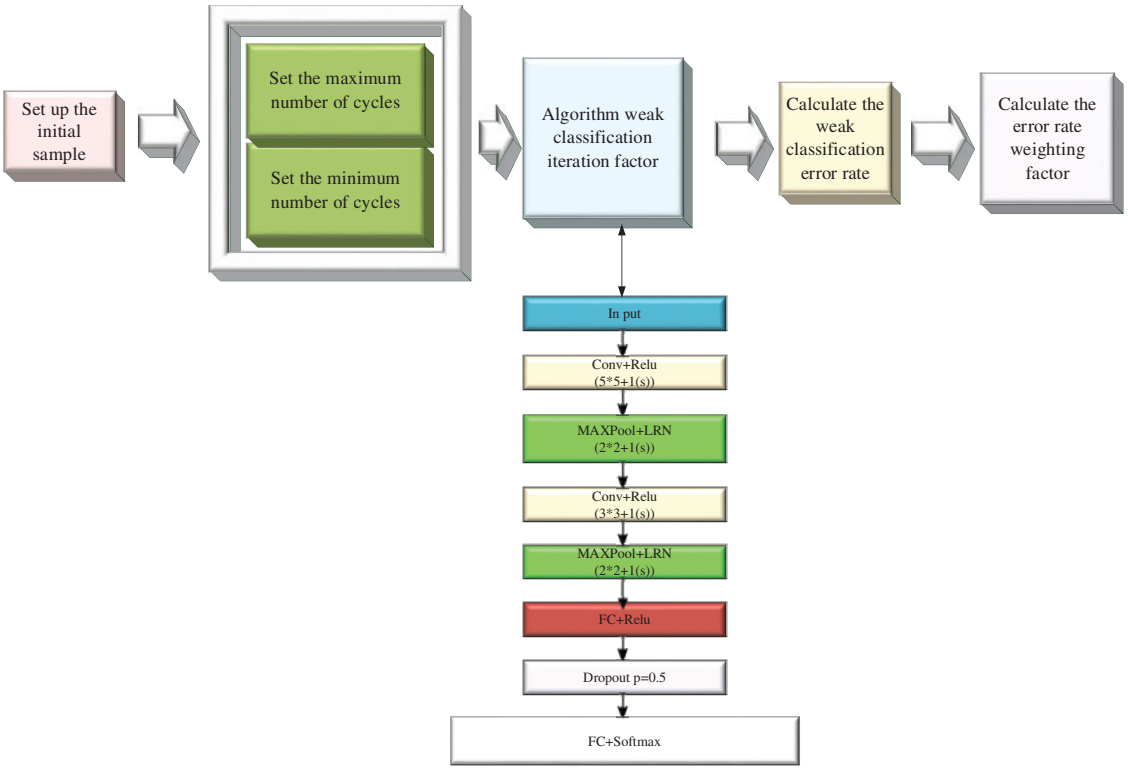

Figure 2: Calculation process of the SAMME-BPNN algorithm

In the algorithm flow given in Fig. 2, the calculation equation for the classification error rate of an individual sample can be expressed as follows:

$$\xi_n = \frac{\sum_{k=1}^{N} D_k^{(n)} I\left(y_n\left(x_k\right) \neq t_k\right)}{\sum_{k=1}^{N} D_k^{(n)}} \tag{4}$$

In Eq. (4) above, ξ_n represents the weighted metric of the classification error of the weak classification factor, and $I\left(y_n\left(x_k\right) \neq t_k\right)$ is a judgment function to judge whether the predicted category is consistent with the actual category. If $y_n\left(x_k\right) \neq t_k$, the value is 1; and if $I\left(y_n\left(x_k\right) = t_k\right)$, the value is 0. When the value is 1, it means that the predicted category is inconsistent with the actual category; otherwise, it is consistent. In addition, k represents the k-th noise sample, and $D_k^{(n)}$ refers to the weight coefficient. x_k refers to an input variable.

The calculation equation of the weight coefficient of the weak classification factor is given as follows:

$$\lambda_n = \ln\left\{\frac{1-\xi_n}{\xi_n}\right\} + \ln(k-1) \tag{5}$$

In Eq. (5), λ_n represents the weight coefficient, and k is a constant. Therefore, the size of the weight coefficient is only related to the weighted measure ξ_n of the classification error of the weak classification factor.

Given the limited learning ability of the BPNN, the accuracy and efficiency will decrease when a large number of learning and calculations are performed, which is a major defect of the SAMME-BPNN algorithm. When the simple noise source is classified, such as four or fewer noises, the SAMME-BPNN algorithm can be used to complete the classification. However, if the audio characteristics of the noise or the noise type increase, it is difficult to maintain a good classification effect. If the noise frequency type is extended to nine types of noise, the accuracy will be reduced. In another case, the sound of a car driving is similar to a windy sound if it is windy. At this time, it is difficult for BPNN to accurately recognize and classify when learning. Given this defect, the principle of the SAMME-BPNN algorithm is expanded, and the NR-CNN algorithm is built in this study. The specific flow of the algorithm is shown in Fig. 3.

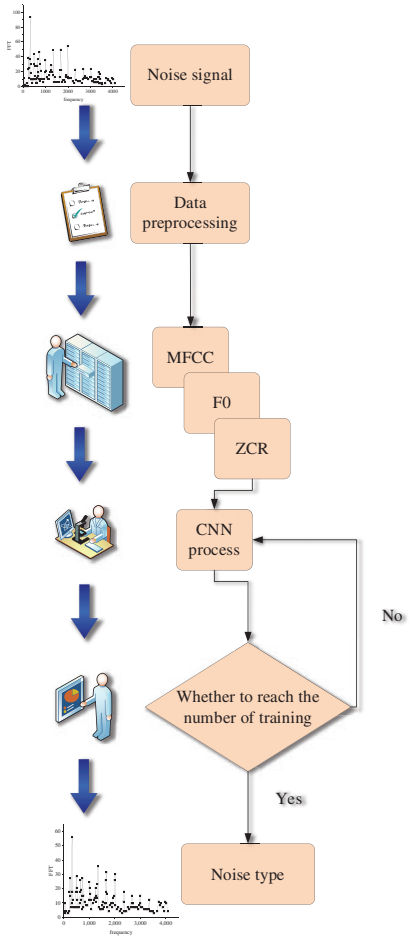

Figure 3: Specific flow of the NR-CNN algorithm

In Fig. 3, preprocessing of the noise signal data is required before speech information enhancement processing. The process of preprocessing includes four steps, namely, cleaning, cutting, windowing, and framing of noise data. Data cleaning refers to the preprocessing of data feature dimensions. During preprocessing, voice data of multiple dimensions are organized into a matrix of a certain size. The operations are performed on different network layers of the CNN, simple classification is performed on the noise signals extracted by the CNN, and the concentrated features are further

extracted to reduce the amount of calculation. The cutting of noise data refers to the segmentation of the entire speech signal during the framing process. Both framing and windowing are the preprocessing stages of the speech signal extraction feature. The framing is first, then the window is added, and then the fast Fourier transform is performed.

Framing means that the speech signal is unstable overall but can be regarded as locally stable. In subsequent speech processing, a stable signal needs to be input, so the entire speech signal needs to be framed, which should be divided into several fragments.

The signal can be considered stable in the range of 10–30 ms. Generally, no less than 20 milliseconds are used as the frame, and approximately 1/2 of the time is undertaken as the frame shift. The frame shift refers to the overlapping area between two adjacent frames to avoid the change of two adjacent frames.

Windowing means that after the framing is performed according to the above method, there will be discontinuities at the beginning and end of each frame. Therefore, the more frames that are divided, the greater the error with the original signal. Windowing solves this problem so that the framed signal becomes continuous, and each frame will show the characteristics of a periodic function. In speech signal processing, a Hamming window is usually added. Noise data preprocessing only needs to be carried out during neural network training. After the training is completed, the neural network can automatically identify the noise signal based on the extracted characteristic data. After the preprocessing is finished, data features must be extracted as the goal of neural network learning and training. The feature extraction in this study is for nine types of noise data. The extracted audio features are first subjected to the convolution, pooling, and full connection calculation of the CNN; then, the continuous iterative back propagation process similar to the BP neural network is performed until the error or the number of iterations meet the requirements, and finally, the type of noise is output. The extracted feature data have a total of 648 dimensions. These dimensional features are formed into a matrix and enter the CNN for training and iteration until the set target value is reached. The target value of this experiment is the threshold value of the noise signal. After reaching the threshold, the CNN training is completed, and then the noise signal can be diagnosed and recognized according to the trained CNN. The threshold of the neural network is related to its parameter settings. The parameter settings of the CNN in this study are shown in Table 1.

Table 1: Initial parameters of the neural network

Parameter options	Numerical value
Learning rate	0.01
Convolution kernel size	$5 * 5$
Learning frequency range	1500–1800
Number of convolution kernels	64-128-256-256-128
Number of convolutional layers	5

2.4 Design of the CNN-Based Speech Enhancement Model

A CNN-based noise recognition system is previously constructed. The constructed system is further improved to establish the speech enhancement model based on the FLRLS and CNN (FLRLS-CNN). The model algorithm process is shown in Fig. 4.

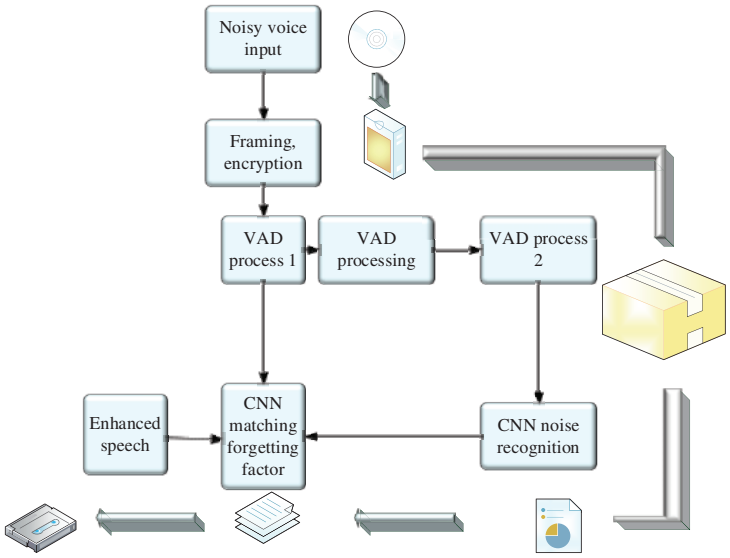

Figure 4: Calculation process of the speech enhancement model

In Fig. 4, the first step of the speech enhancement model and the noisy recognition model is similar, in which speech data with noise are input and preprocessed. The difference is that the CNN noise recognition system becomes a functional block of the speech enhancement operation model in this speech enhancement operation model. The data it receives are the nonspeech segment sound data after Voice Activity Detection (VAD) speech endpoint detection. These sound data are extracted from multiple dimensions of sound features to form feature data, subjected to the trained neural network model for recognition and detection, and inputted into the CNN matching the forgetting factor module. The CNN matching forgetting factor module contains an adaptive filter based on the FLRLS method to filter the noise frequency, and the speech signal is enhanced after the output is obtained. In this model, the VAD processing flow is a key part, and the structure of the flow is shown in Fig. 5.

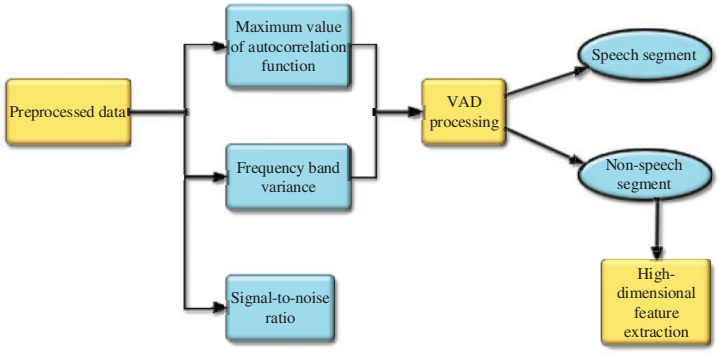

Figure 5: VAD processing flow

The calculation process in Fig. 5 is divided into the following two steps. (1) In the first stage, the value of the higher threshold T2 is determined, and the audio segment higher than T2 is the voice segment. The value is selected according to the speech energy envelope. Beyond the time point

corresponding to the intersection of T2 and short-time energy is the starting and ending point range of speech. Then, a lower threshold T1 is determined based on the average energy. (2) The second stage starts the retrieval from the intersection point of T1 and the envelope calculated in the previous stage. When the interval value of the short-term average zero crossing rate is less than the threshold value T3, it is identified as the interval of the final speech segment.

2.5 Design of the Simulation Experiment

Two different experiments are designed to test the performance of the proposed noise recognition algorithm and the CNN-based speech enhancement processing model.

Experiment 1 is the experimental design for the NR-CNN noise recognition algorithm. It includes nine kinds of audio, namely human voices, laughter, wind, bird calls, car engine sounds, raining sounds, clapping sounds, water flowing sounds, and factory machinery operating sounds. These data are preprocessed and coded to form a data set. The training set and data set parameters are shown in Table 2.

Table 2: Data set information

	Training set	Data set	Label
Vocal noise	2694	266	a
Wind noise	2410	267	b
The sound of rain	2498	252	c
Engine noise	2603	282	d
Laughter	2615	260	e
Machine sound	2532	261	f
Birdsong	2413	265	g
Sound of streams	2339	252	h
Applause	2487	267	i

The confusion matrix method is used to assess the performance of experimental noise recognition. The confusion matrix is shown in Table 3.

Table 3: Confusion matrix

	Positive	Negative
True	TP	TN
False	FP	FN

The confusion matrix method is used to assess the performance of experimental noise recognition. There are three judgment indicators of the confusion matrix, namely, the accuracy, recall rate, and F1 value, which are calculated with the following equations:

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

$$F_1 = 2\frac{P \times R}{P + R} \tag{8}$$

In the above three equations, P represents the accuracy; TP and FP refer to the number of true positives and false positives, respectively; R represents the recall rate; and FN is the number of false negatives. In addition, the support vector machine (SVM) and k-nearest neighbor (KNN) classification algorithms are introduced and compared with the proposed NR-CNN noise classification model regarding their respective performance.

Experiment 2: The test data of the speech enhancement model still use the data source in experiment 1, but the evaluation methods of the experimental results are different. There are two evaluation criteria for the speech enhancement model, namely, clarity and transmission efficiency. The clarity of the speech that has only been strengthened by the model is compared to the sound source. The transmission efficiency is the degree to which the processed sound can be understood and accepted by humans. The evaluation standard of transmission efficiency is the perceptual evaluation of speech quality (PESQ) proposed by the International Telecommunication Union (ITU). In addition, a control group is set to verify the PESQ values of different algorithms before and after the CNN is added.

3 Discussion and Analysis of Simulation Results

3.1 Simulation Experiment Results of the Noise Recognition Algorithm Based on NR-CNN

According to the classification effect of NR-CNN on different types of noise, the noise classification effect is shown in Fig. 6.

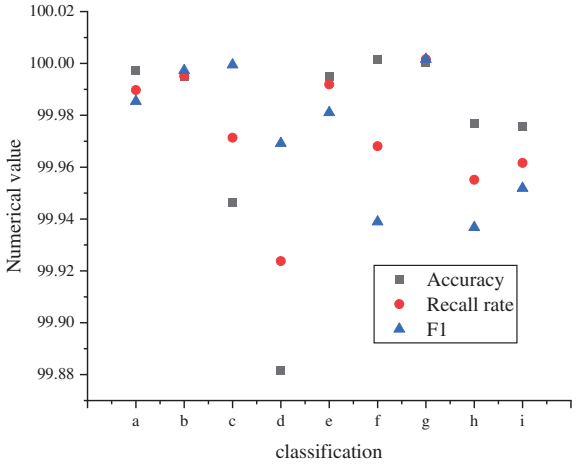

Figure 6: Noise classification effect (in the figure, a∼i refer to human voice noise, wind noise, rain noise, car engine noise, laughter, machine sound, bird song, sound of streams, and applause, respectively)

As shown in Fig. 6, the proposed NR-CNN noise recognition algorithm has an accuracy of more than 99% for different types of voice classification; among them, the classification and recognition accuracy of automobile engine noise is the lowest (99.88%), the recognition accuracy of human voice

noise is the highest (100%), and that of other types of noise all reach more than 99.92%. This shows that the proposed algorithm has excellent performance. The proposed NR-CNN noise recognition algorithm has the lowest recall rate for the classification of machine sounds, but it also reaches more than 99.92%; the high recall rate of the model indicates that the coverage of the algorithm is good, and most noises can be accurately identified. In addition, the F1 values of the model all reach more than 99.92%. Among them, the lower F1 recognition sound types are found in machine sound and stream sound, which are 99.95% and 99.92%, respectively. The highest F1 value is found in the recognition of bird songs, reaching 100%. In summary, the NR-CNN noise recognition algorithm proposed in this study shows superior classification performance, comprehensive sound type recognition, and stable performance.

To further highlight the superiority of the performance of the proposed noise recognition algorithm, several other algorithms are introduced for comparison, and the results are shown in Fig. 7.

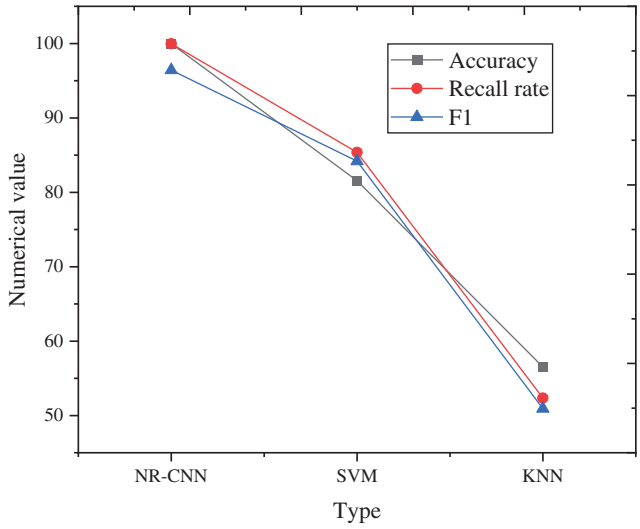

Figure 7: Comparison of the noise classification effects of three algorithms

As illustrated in Fig. 7, the proposed NR-CNN noise recognition algorithm is superior to the SVM and KNN algorithms in terms of classification accuracy, recall rate, and F1 value. In terms of specific values, the average accuracy rate and recall rate of the noise recognition algorithm based on NR-CNN reached 99.97%, while the F1 value was slightly lower, reaching 96.43%. For the other two algorithms, the accuracy of the SVM is only approximately 81%, and the accuracy of the KNN algorithm is even lower (only 50%–60%). In conclusion, the noise recognition algorithm based on NR-CNN shows excellent performance in terms of noise recognition and classification.

3.2 Experimental Results of the Speech Enhancement Model

The noise enhancement effect is shown in Fig. 8, which is drawn according to the experimental data.

As shown in Fig. 8, the enhanced sound spectrogram is very close to the original sound signal frequency, and the frequency trend is almost the same as the overall sound frequency trend. Regarding the frequency of a single sound, the enhanced sound spectrogram is also consistent with the original sound, and several noise and sound interference signals are successfully filtered out. The comparison

of the sound spectrogram shows that the proposed sound enhancement model can enhance the original sound under the condition of noise interference, and the enhancement effect is good.

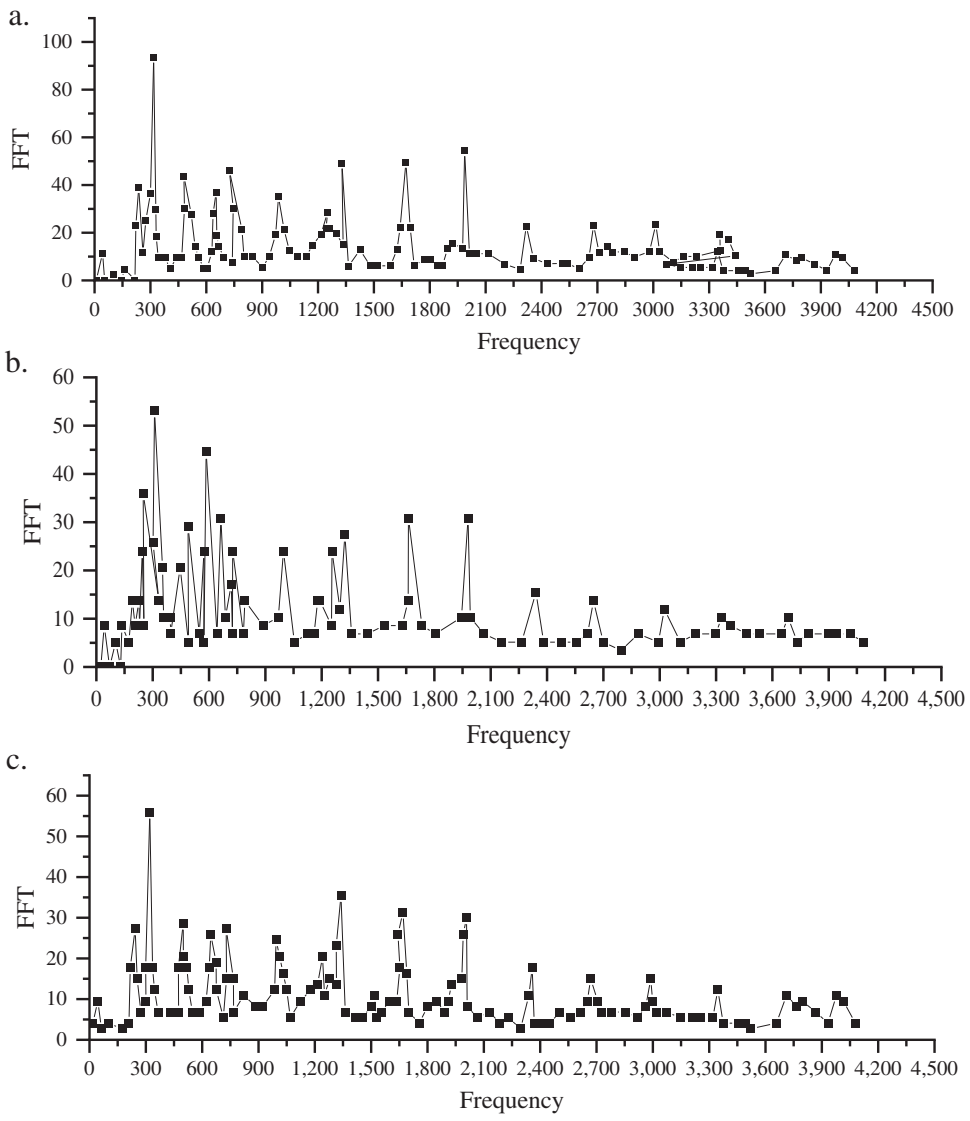

Figure 8: Comparison of the noise enhancement effect ((a) in the figure represents the original sound signal frequency, (b) represents the noise signal frequency, and (c) represents the enhanced sound frequency)

To explore the effect of the signal of the sound enhancement model proposed in this study, the comparison of the noisy signal before enhancement is presented in Fig. 9 according to the comparison of the signal-to-noise ratio in the interval of −4, 0, 4.

According to the comparison results of noisy signals before and after enhancement in Fig. 8, the sound enhancement model enhances the sound to a much higher SNR than before the enhancement in the range of −4 to 4. When the SNR is 0, the difference between the two is the largest, and then it starts to decrease slowly; when it approaches 4, the two coincide.

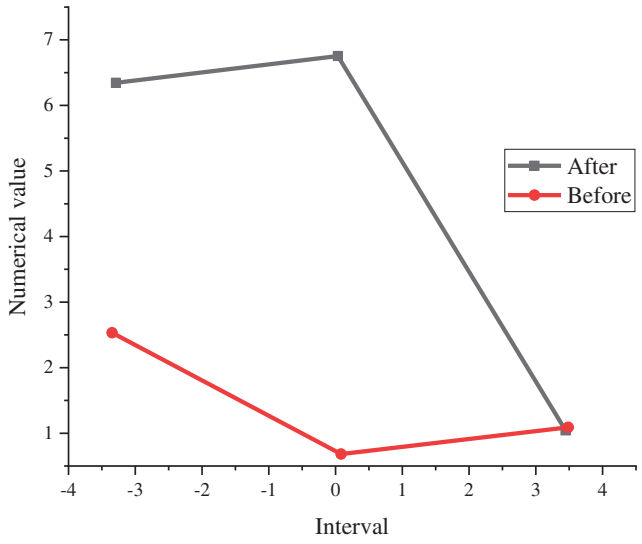

Figure 9: Comparison of SNR before and after enhancement

To further explore the performance of the sound enhancement model, the sound enhancement effect is compared based on the PESQ values of different algorithms before and after the CNN is added, and the results are illustrated in Fig. 10.

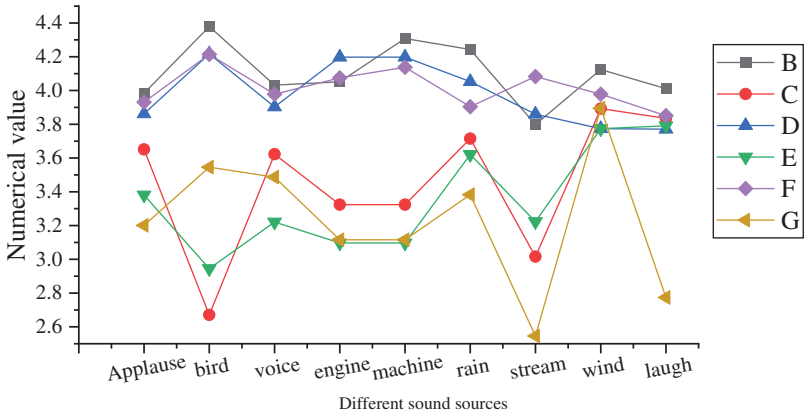

Figure 10: Comparison of the sound enhancement effect (in the above figure, B and C represent the data of experimental group B and control group C when the SNR is −4, respectively; D and E represent the data of experimental group D and control group E when the SNR is 0, respectively; F and G represent the data of experimental group F and control group G when the SNR is 4, respectively)

As illustrated in Fig. 10, under different sound types and different SNRs, the speech perception quality evaluation system values of the experimental group (the enhanced model proposed) exceeded 3.5. This shows that after the CNN is added, a very good noise reduction effect can be achieved in different sound environments. In addition, after the CNN is added, the average value of all the noise perception quality evaluation system values in the experiment is improved by more than 21% compared with the traditional noise reduction method, which shows that the algorithm has the powerful function of speech enhancement and noise reduction.

Based on the distortion measurement values of the two algorithms before and after the CNN is added, a distortion comparison chart is plotted in Fig. 11.

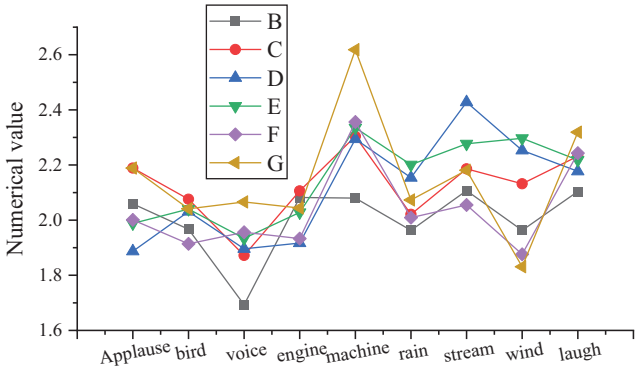

Figure 11: Comparison of speech distortion comparison (in the above figure, B and C represent the distortion index of experimental group B and control group C when the SNR is −4, respectively; D and E represent the distortion index of experimental group D and control group E when the SNR is 0, respectively; and F and G represent the distortion index of experimental group F and control group G when the SNR is 4, respectively)

As shown in Fig. 11, the sound distortion index of the speech enhancement model proposed is inferior to that of the control group, indicating that after the CNN is added, it is not easy to cause sound signal distortion in different sound environments, showing stronger robustness.

In summary, the proposed speech enhancement processing model based on CNN and SFTRLS can adapt to a variety of speech environments. In practical applications, it can simultaneously perform enhancement and noise reduction processing for multiple different types of speech signals, and the processing effect is better than that of traditional sound enhancement models. It can reduce the interference of external noise to mobile communication and ensure the accurate transmission of information.

4 Conclusion and Prospects

With the rapid development of multimedia technology and communication technology, people have entered the intelligent era. In this quick and convenient communication process, voice signals are the most direct and simple method and play an important role. However, due to the influence of the surrounding environment and noise in the communication process, the listener receives noisy speech rather than pure speech, which seriously affects people's hearing effect and daily life. Therefore, it is of great practical significance to extract the desired target signal from the complex observation signal with noise and interference and improve the communication quality. In this work, the principle and technology of speech enhancement are analysed, and the fast recursive least square method is proposed to process sound data. Combined with a CNN, the NR-CNN noise recognition algorithm and speech enhancement model are proposed. Then, some experiments are designed to verify the actual performance of the two. Compared with the audio features acquired in traditional algorithms that have relatively small dimensions and single attributes, this study involves 648 attribute dimensions. A variety of audio characteristics are used to form a large network input, making the research and description of audio characteristics in this article completer and more accurate [37,38]. Based on the high-dimensional learning and the dimensionality reduction learning ability of the CNN, it saves

calculation time and makes the experiment achieve a good classification effect. The speech information enhancement processing model proposed in this study can adapt to a variety of speech environments, can simultaneously enhance and reduce noise processing of multiple types of speech signals, and has achieved the expected results [39,40]. However, there are still some shortcomings in the research process, which are summarized as follows:

First, there are fewer types of sounds in this study, and only nine different sound sources have been tested, which is not enough to represent most sounds in life. Second, although a CNN is introduced to reduce the dimensionality of the data, which improves the computational efficiency to a certain extent, there is still a delay when the data set is relatively large. In view of the above two shortcomings, more comprehensive sound types will be considered, and other neural network algorithms will be introduced to improve calculation efficiency in future research so that the research results can be more valuable and meaningful.

Funding Statement: This work was supported by General Project of Philosophy and Social Science Research in Colleges and Universities in Jiangsu Province (2022SJYB0712), Research Development Fund for Young Teachers of Chengxian College of Southeast University (z0037) and Special Project of Ideological and Political Education Reform and Research Course (yjgsz2206).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Z. Lv, X. Li and W. Li, "Virtual reality geographical interactive scene semantics research for immersive geography learning," *Neurocomputing*, vol. 254, no. 2, pp. 71–78, 2017.

[2] B. Deng and R. Varatharajan, "Word order detection in English classroom teaching based on improved genetic algorithm of block coding," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 4, pp. 6901–6912, 2021.

[3] A. Randall, W. Toon and M. Marc, "Correction to: An integrated MVDR beamformer for speech enhancement using a local microphone array and external microphones," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 202, 2021.

[4] H. Li, Y. Xu, D. Ke and K. Su, "μ-law SGAN for generating spectra with more details in speech enhancement," *Neural Networks*, vol. 136, pp. 17–27, 2021.

[5] K. Hansol and S. J. Won, "Dual-mic speech enhancement based on TF-GSC with leakage suppression and signal recovery," *Applied Sciences*, vol. 11, no. 6, pp. 2816, 2021.

[6] Y. Zhou, H. Wang, Y. Chu and H. Liu, "A robust dual-microphone generalized sidelobe canceller using a bone-conduction sensor for speech enhancement," *Sensors*, vol. 21, no. 5, pp. 102, 2021.

[7] A. Karthik and J. L. MazherIqbal, "Efficient speech enhancement using recurrent convolution encoder and decoder," *Wireless Personal Communications*, vol. 21, no. 3, pp. 1–15, 2021.

[8] L. Xu, Z. Wei, S. Zaidi, B. Ren and J. Yang, "Speech enhancement based on nonnegative matrix factorization in constant-Q frequency domain," *Applied Acoustics*, vol. 174, no. 4, pp. 107732, 2020.

[9] C. Li, T. Jiang and S. Wu, "Speech enhancement based on approximate message passing," *China Communications*, vol. 17, no. 8, pp. 187–198, 2020.

[10] M. S. Islam, Y. Zhu, M. I. Hossain, R. Ullah and Z. Ye, "Supervised single channel dual domains speech enhancement using sparse non-negative matrix factorization," *Digital Signal Processing*, vol. 100, no. 2, pp. 102697, 2020.

[11] Z. Wang, T. Zhang, Y. Hao and B. Ding, "LSTM-convolutional-BLSTM encoder-decoder network for minimum mean-square error approach to speech enhancement," *Applied Acoustic*, vol. 172, no. 2, pp. 107647, 2021.

[12] X. Li, X. Wang, Y. Qin and J. Li, "SNR classification based multi-estimator IRM speech enhancement algorithm," *Journal of Physics: Conference Series*, vol. 2173, no. 1, pp. 012086, 2022.
[13] R. Bendoumia, A. Guessoum, I. Hassani, S. Bougheddaoui and R. Cherif, "New simplified sub-band sparse adaptive algorithm for speech enhancement in communication systems," *International Conference on Artificial Intelligence in Renewable Energetic Systems*, vol. 21, no. 1, pp. 12, 2022.
[14] H. Schrter, B. A. N. Escalante, T. Rosenkranz and A. Maier, "DeepFilterNet2: Towards real-time speech enhancement on embedded devices for full-band audio," *arXiv preprint arXiv*, vol. 12, no. 10, pp. 12, 2022.
[15] K. H. Yong, Y. J. Won, S. J. Cheon, W. H. Kang and N. S. Kim, "A multi-resolution approach to gan-based speech enhancement," *Applied Sciences*, vol. 11, no. 2, pp. 721, 2021.
[16] M. L. Rachel and K. Gopakumar, "Evaluation of speech enhancement algorithms applied to electrolaryngeal speech degraded by noise," *Applied Acoustics*, vol. 174, no. 1, pp. 107771, 2021.
[17] L. Jorge, R. Dayana, M. Antonio, L. Vicente, A. Ortega *et al.,* "Progressive loss functions for speech enhancement with deep neural networks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 201, 2021.
[18] S. Nasir, K. M. Irfan, A. H. Mu'ath and A. Jan, "Learning time-frequency mask for noisy speech enhancement using gaussian-bernoulli pre-trained deep neural networks," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 1, pp. 849–864, 2021.
[19] P. Ashutosh and D. Wang, "Dense CNN with self-attention for time-domain speech enhancement," *IEEE/ACM*, vol. 29, pp. 1270–1279, 2021.
[20] N. Soha, W. Julie, M. Mansour, C. Glackin and N. Cannings, "An experimental analysis of deep learning architectures for supervised speech enhancement," *Electronics*, vol. 10, no. 1, pp. 774–782, 2020.
[21] Y. Li, Y. Sun and S. Mohsen Naqvi, "Single-channel dereverberation and denoising based on lower band trained SA-LSTMs," *IET Signal Processing*, vol. 14, no. 10, pp. 17, 2020.
[22] S. Jishnu, K. D. Jitendra and S. C. Sekhar, "Musical noise suppression using a low-rank and sparse matrix decomposition approach," *Speech Communication*, vol. 125, no. 2, pp. 41–52, 2020.
[23] N. Aaron and K. P. Kuldip, "Masked multi-head self-attention for causal speech enhancement," *Speech Communication*, vol. 125, no. 3, pp. 80–96, 2020.
[24] H. Jia, W. Wang and S. Mei, "Combining adaptive sparse NMF feature extraction and soft mask to optimize DNN for speech enhancement," *Applied Acoustics*, vol. 171, no. 5, pp. 107666, 2020.
[25] W. Yuan, "A time-frequency smoothing neural network for speech enhancement," *Speech Communication*, vol. 124, no. 6, pp. 75–84, 2020.
[26] T. Hu, K. Mohammad, M. Mokhtar and T. A. Rashid, "Real-time COVID-19 diagnosis from X-Ray images using deep CNN and extreme learning machines stabilized by chimp optimization algorithm," *Biomed Signal Process Control*, vol. 68, no. 15, pp. 102764, 2021.
[27] Y. Liu, H. Pu and D. Sun, "Efficient extraction of deep image features using convolutional neural network (CNN) for applications in detecting and analysing complex food matrices," *Trends in Food Science & Technology*, vol. 113, no. 7, pp. 193–204, 2021.
[28] M. Tamilselvi and S. Karthikeyan, "Hybrid framework for a robust face recognition system using EVB_CNN," *Journal of Structural Engineering*, vol. 23, no. 3, pp. 43–57, 2021.
[29] M. B. Mukami, D. Sagnik and S. Rittika, "CEFEs: A CNN explainable framework for ECG signals," *Artificial Intelligence in Medicine*, vol. 115, no. 5, pp. 102059, 2021.
[30] M. Andrey, R. Gabriela, G. Bahaa, A. Badal and S. Glick, "Exploring CNN potential in discriminating benign and malignant calcifications in conventional and dual-energy FFDM: Simulations and experimental observations," *Journal of Medical Imaging*, vol. 8, no. 3, pp. 033501, 2021.
[31] H. Dong, D. Chen, L. Zhang, H. Ke and X. Li, "Subject sensitive EEG discrimination with fast reconstructable CNN driven by reinforcement learning: A case study of ASD evaluation," *Neurocomputing*, vol. 449, no. 1, pp. 136–145, 2021.
[32] G. Jia, H. Lam and Y. Xu, "Classification of COVID-19 chest X-Ray and CT images using a type of dynamic CNN modification method," *Computers in Biology and Medicine*, vol. 134, no. 8, pp. 104425, 2021.

[33] T. Shimamura, "Facial emotion recognition using transfer learning in the deep CNN," *Electronics*, vol. 10, no. 9, pp. 1036, 2021.
[34] J. B. Wu, Y. Zhang, C. W. Luo, L. F. Yuan and X. K. Shen, "A modification-free steganography algorithm based on image classification and CNN," *International Journal of Digital Crime and Forensics*, vol. 13, no. 3, pp. 47–58, 2021.
[35] S. Gao, "The application of agricultural resource management information system based on internet of things and data mining," *IEEE Access*, vol. 9, pp. 164837–164845, 2021.
[36] B. Choi, Y. Lee, Y. Kyung and E. Kim, "Albert with knowledge graph encoder utilizing semantic similarity for commonsense question answering," *Intelligent Automation & Soft Computing*, vol. 36, no. 1, pp. 71–82, 2023.
[37] Y. Kim, T. Kim, H. Choi, J. Park and Y. Kyung, "Reinforcement learning-based handover scheme with neighbor beacon frame transmission," *Intelligent Automation & Soft Computing*, vol. 36, no. 1, pp. 193–204, 2023.
[38] G. Elamparithi, "Resilient service authentication for smart city application using iot," *Intelligent Automation & Soft Computing*, vol. 36, no. 1, pp. 145–152, 2023.
[39] C. Nandagopal, P. S. Kumar, R. Rajalakshmi and S. Anandamurugan, "Mobility aware zone-based routing in vehicle ad hoc networks using hybrid metaheuristic algorithm," *Intelligent Automation & Soft Computing*, vol. 36, no. 1, pp. 113–126, 2023.
[40] C. Yu, Z. Li, D. Yang and H. Liu, "Liu A fast robotic arm gravity compensation updating approach for industrial application using sparse selection and reconstruction," *Robotics and Autonomous Systems*, vol. 149, no. 2, pp. 103971, 2022.