EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

# Application of Deep Neural Network Algorithm in Speech Enhancement of Online English Learning Platform

Haiyan Peng, Min Zhang*

Foreign Language Department, Guangzhou Huashang College, Guangzhou 511339, China

## Abstract

INTRODUCTION: In the online English learning platform, noise interference makes people unable to hear the content of English teaching clearly, which leads to a great reduction in the efficiency of English learning. In order to improve the voice quality of online English learning platform, the speech enhancement method of the online English learning platform based on deep neural network is studied.
OBJECTIVES: This paper proposes a deep neural network-based speech enhancement method for online English learning platform in order to obtain more desirable results in the application of speech quality optimization.
METHODS: The optimized VMD (Variable Modal Decomposition) algorithm is combined with the Moth-flame optimization algorithm to find the optimal solution to obtain the optimal value of the decomposition mode number and the penalty factor of the variational modal decomposition algorithm, and then the optimized variational modal decomposition algorithm is used to filter the noise information in the speech signal; Through the network speech enhancement method based on deep neural network learning, the denoised speech signal is taken as the enhancement target to achieve speech enhancement.
RESULTS: The research results show that the method not only has significant denoising ability for speech signal, but also after this method is used, PESQ value of speech quality perception evaluation of speech signal is greater than 4.0dB, the spectral features are prominent, and the speech quality is improved.
CONCLUSION: Through experiments from three perspectives: speech signal denoising, speech quality enhancement and speech spectrum information, the usability of the method in this paper is confirmed.

**Keywords:** deep neural network; online English learning; platform speech; enhancement; denoising; variational modal decomposition

Received on 22 August 2022, accepted on 15 October 2022, published on 26 October 2022



doi: 10.4108/eetsis.v10i1.2577

*Corresponding author. Email: minzhang3040@gdhsc.edu.cn

## 1. Introduction

Speech [1], as the most frequently used communication method by people, is both efficient and convenient at the same time. However, speech signals are inevitably disturbed by various noises in the process of acquisition and transmission. When acquiring speech signals, and the acquired speech signal will be mixed with other sounds. During transmission, the speech signal is affected by various circuit noises, making the quality of the speech signal at the receiving end degraded. These noises greatly affect the efficiency of information transmission and even lead to communication failure. At the same time, in the point-to-point English learning platform in a more complex environment, the application scenarios of speech teaching are more complex and bad. Moreover, in order to ensure the concealment, the signal is often weak, which is easily submerged in the noise of the environment and

affects the learning effect. Therefore, the application of speech enhancement technology in English learning platforms is particularly important.

NorezmiJamal et al. [2] proposed to separate the noise mixture signal from the noise background to predict the target mask. The noise in Malay speech is reduced by using the deep neural network method combined with the acoustic characteristics such as the power spectrum of the gamma pass filter bank to enhance the intelligibility of Malay speech. ShobaSivapatham et al. [3] proposed performance analysis methods for various training targets to improve speech quality and intelligibility. In order to improve the speech quality and intelligibility, the performance of binary and non binary training targets of the deep neural network is evaluated under different SNR and noise conditions.

Deep neural network method [4] is a hot application in the field of speech signal processing, therefore, this paper studies the problem of speech enhancement of online English learning platform, and proposes a deep neural network-based speech enhancement method for online English learning platform in order to obtain more desirable results in the application of speech quality optimization. The speech enhancement technology can suppress the interference of noise as much as possible, make the English speech clearer, and improve the user's learning experience.

## 2. Design and implementation of speech enhancement algorithm

### 2.1. Speech signal denoising based on optimized VMD algorithm

Variable fractional modal decomposition algorithm is referred to as VMD algorithm, and this paper optimizes the method, and the principle of optimizing VMD algorithm is as follows: firstly, this paper uses the Moth-flame optimization algorithm to optimize the VMD algorithm, find the number of decomposed modalities $H$ and penalty factor $\beta$ in the optimized combination $[H_0, \beta_0]$, and complete the decomposition of the speech signal [5]; Then, based on the principle of correlation coefficient filtering, the effective mode and the noise mode are selected for the decomposed modal components, and the noise mode is de-noising by wavelet threshold [6]; Finally, the de-noising modal components and effective modes are reconstructed to make the speech signal achieve the purpose of de-noising. The operation flow chart is shown in Figure 1.

#### 2.1.1 Signal decomposition based on variational modal decomposition algorithm

The adaptive, non-recursive signal decomposition method-variational modal decomposition algorithm [7] can be used to decompose the signal into a finite number of intrinsic modal components and the sum of these intrinsic modal components. These intrinsic modal components are characterized by the definition of intrinsic modes. The process of decomposing the signal by VMD is actually the process of solving the variational problem, and solving the variational first requires constructing the variational, so constructing the variational problem and solving the variational problem are the core of the algorithm.

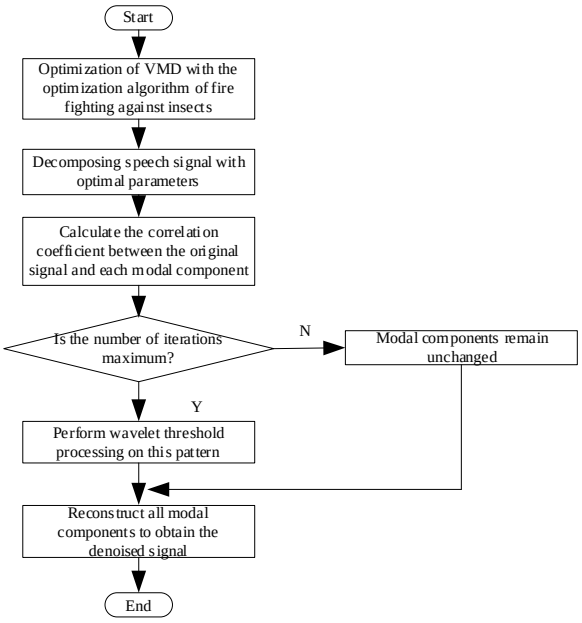



**Figure 1.** Speech signal denoising process of online English learning platform based on optimized VMD algorithm

The input signal is decomposed into $h$ modal categories $\alpha_h$ using the VMD decomposition algorithm, and the modal decomposition [8] results are assumed to contain a finite bandwidth of central frequencies. First, each modal component needs to be Hilbert transformed to obtain the analysis signal, i.e. one-sided spectrum; Secondly, the exponential term is added to the corresponding estimated center spectrum, and the spectrum of each modal component needs to be modulated to the corresponding fundamental frequency band; Finally, the gradient squared $Z_2$ norm is obtained to calculate the demodulated signal, and the estimated bandwidth corresponding to each modal component is obtained. Therefore, the constrained variational problem of speech signal decomposition can be constructed as follows:

$$\min_{\{\alpha_h\},\{\varpi_h\}} \left\{ \sum_h \left\| \partial_t \left[ \left( \varepsilon(t) + \frac{i}{\pi t} \right) \times \alpha_h(t) \right] e^{-i\varpi_t t} \right\|_2^2 \right\} \quad (1)$$

$$s.t. \quad \sum_h \alpha_h = g \tag{2}$$

where $\varpi_h$ is the set of central frequencies of all components of the speech signal; $g$ is the sum of the modal components of the speech signal; $\alpha_h(t)$ is the modal classification solution of the speech signal at time $t$ ; $\varepsilon(t)$ and $\partial_t$ are the Dirichlet functions, partial derivatives, $i \in h$ ; $e$ are the transcendental numbers. In solving the constructed speech signal constrained variational problem, the quadratic penalty function term $\delta$ and the Lagrange multiplier $\gamma$ are introduced, and the expression is $L(\{\alpha_h, \varpi_h, \gamma\})$ .

$$L(\{\alpha_h,\varpi_h,\gamma\}) = \delta\sum_h \left\| \partial_t \left[ \left( \varepsilon(t) + \frac{i}{\pi t} \right) \times \alpha_h(t) \right] e^{-i\varpi_h t} \right\|_2^2 + \left\| o(t) - \sum_h \alpha_h(t) \right\|_2^2 + \langle \gamma, o(t) - \gamma \rangle \tag{3}$$

where $o(t)$ represents the original speech signal.

The multiplicative operator alternating direction method is used to update the moments $t$ the $m+1$ th $\alpha_h^{m+1}$ , $\varpi_h^{m+1}$ to transform the minimization problem of Eq. (1) into the saddle point of Eq. (3) after the expansion of the iterative sub-optimization sequence, where the update equation of $\alpha_h^{m+1}$ is

$$\alpha_h^{m+1} = \arg\min_{\alpha_h \in o} \left\{ \sum_h \left\| \partial_t \left[ \left( \varepsilon(t) + \frac{i}{\pi t} \right) \times \alpha_h(t) \right] e^{-i\varpi_h t} \right\|_2^2 + \left\| o(t) - \sum_h \alpha_h(t) - \frac{\gamma}{2} \right\|_2^2 \right\} \tag{4}$$

Where $e$ is the error term; using the Fourier transform, Eq. (4) is converted from the time domain to the frequency domain, and the expressions of each modal component of the speech signal in the frequency domain are obtained as

$$\hat{\alpha}_h^{m+1}(\varpi) = \frac{\hat{o}(\varpi) - \sum_{j \neq h} \hat{\alpha}_j(\varpi) + \frac{\gamma}{2}}{1 + 2\gamma(\varpi - \varpi_h)^2} \tag{5}$$

where, $\hat{o}(\varpi)$ is the expression of the speech signal in the frequency domain; $\varpi$ is the mean value of the center frequencies of all components of the speech signal; $\hat{a}_j(\varpi)$ represents the expression decomposed into $j$ modal classifications $\hat{a}_j$ . For finding the center frequency $\varpi_h^{m+1}$ of each modal component of the update, the problem of solving the center frequency can be transformed to the frequency domain [9], and the expression of the update $\varpi_h^{m+1}$ can be obtained as

$$\varpi_h^{m+1} = \frac{\int_0^\infty \varpi \left| \hat{\alpha}_h(\varpi) \right|^2 d\varpi}{\int_0^\infty \varpi \left| \hat{\alpha}_h(\varpi) \right|^2 d\varpi} \tag{6}$$

where, $d$ is the derivative parameter. $\varpi_h^{m+1}$ denotes the center of the power spectrum of the $h$ th mode; $\hat{\alpha}_h(\varpi)$ denotes the Wiener filter of the current residual. The VMD algorithm continuously updates each mode in the frequency domain, and then converts it to the time domain by using the nonlinear Fourier transform [10]. The steps of updating the mode components are as follows:

(1) Initialize $\alpha_h^{m+1}$ , $\varpi_h^{m+1}$ , $\gamma$ .

(2) Update $\alpha_h$ and $\varpi_h$ .

(3) Update $\gamma$ .

(4) Given the discriminant error $e > 0$ , if $\frac{\sum_h \left\| \hat{\alpha}_h^{m+1} - \hat{\alpha}_h^m \right\|_2^2}{\left\| \hat{\alpha}_h^m \right\|_2^2} < e$ , stop the iteration; otherwise go to step (2) and continue the execution.

### 2.1.2 Screening of correlation coefficients

After the signal is decomposed by the variational modal decomposition algorithm, its correlation coefficient is filtered by the autocorrelation function.

First, each modal component $\alpha_h$ of the online English learning platform speech signal and the original speech signal using Eq. (6) to calculate $(h+1)^{th}$ autocorrelation function $S_y(n)$ :

$$S_y(t) = \frac{1}{M} \sum_{j=0}^{M-1} \left[ o(t) \cdot o(j+t) \right] \tag{6}$$

Then the autocorrelation function is normalized to obtain the correlation coefficients of the autocorrelation function. The calculation method is:

$$q(t) = \frac{\sum_{t=1}^{2M-1} S_y(t)}{\sqrt{\sum_{t=1}^{M-1} S_y(t)}} \tag{7}$$

where, $M$ denotes the number of speech signal points in the online English learning platform, and $\alpha_i$ denotes the $i$ modality of the speech signal.

According to the correlation coefficient screening principle, it is known that $q(t)$ is less than the threshold

value, then the modal component is considered to be well correlated with the original speech signal and needs to be retained; otherwise, the corresponding modal component is then subjected to wavelet denoising [11].

### 2.1.3 Optimization of VMD algorithm based on Moth-flame optimization algorithm

After screening the correlation coefficients of modal components, the VMD algorithm is optimized by moths flaring fire optimization algorithm. Suppose the moth is a candidate solution for solving the decomposition modulus $H$ and the penalty factor $\beta$, and the variable to be solved is the position of the moth in space. Thus, by changing its own position vector, the moth can fly in one, two, three, or even higher dimensions. Since the moth-flame optimization (MFO) algorithm is essentially a population intelligence optimization algorithm, the population of moths can be represented in the matrix as follows.

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix} (8)$$

where $H \in X$ , $\beta \in X$ ; $n$ is the number of moths; $m$ represents the dimensionality of the modal number $H$ and the penalty factor $\beta$ to be decomposed. For these moths, it is also assumed that there exists a vector of adaptation values corresponding to them, denoted as $P_n$ .

The MFO algorithm requires each moth to update its own position using only the unique flame corresponding to it, thus avoiding the algorithm from falling into local extremes and greatly enhancing the global search capability of the algorithm. Therefore, the moth position and the flame position in the search space are the same dimensional matrix of variables. In order to mathematically model the flight behavior of moths to a flame, the position update mechanism of each moth relative to the flame can be represented by the equation

$$X_j = R\left(X_j, G_i\right) (9)$$

where $X_j$ denotes the $j$ th moth that describes only the candidate solution for solving the decomposition modulus $H$ and penalty factor $\beta$ ; $G_i$ denotes the $i$ th flame; and $R$ denotes the spiral function. The function satisfies the following conditions.

(1) The initial point of the spiral function should start from the moth.

(2) The end point of the spiral is the position of the flame.

(3) The fluctuation range $R\left(X_j, G_i\right)$ of the spiral should not exceed its search space.

$$R\left(X_j, G_i\right) = E_j \cdot e^{ck} \cdot \cos\left(2\pi k\right) + G_i (10)$$

where $E_j$ denotes the distance between the $j$ th moth and the $i$ th flame; $c$ is the defined logarithmic spiral shape constant and the path coefficient $k$ is a random number in [-1,1].

$$E_j = \left|G_i - X_j\right| (11)$$

Equation (11) simulates the path of the moth's spiral flight, and it can be seen that the next position of the moth's update is determined by the flame it surrounds. The spiral equation shows that the moth can fly around the flames, not just in the space between them, thus guaranteeing the global search capability of the algorithm with local exploitation. If the fitness value of the updated moth position is better than that of the contemporary corresponding flame, its updated position will be selected as the position of the next generation of flames, and thus the moth has local exploitation capability. The model has the following characteristics when used.

(1) By modifying the parameter $k$ , a moth can converge to an arbitrary neighborhood of the flame.

(2) The smaller the $k$ , the closer the moth is to the flame.

(3) As the moth gets closer to the flame, it renews itself around the flame more and more frequently.

The flame position update mechanism described above ensures the local exploitation capability of the moth around the flame. To increase the probability of finding a better solution with the decomposition modulus $H$ and penalty factor $\beta$ , the currently found optimal solution is used as the location of the next generation of flames. Thus, the flame location matrix usually contains the currently found optimal solution. During the optimization process, each moth updates its position according to the matrix. The path coefficients present $k$ in the MFO algorithm are random numbers in a fixed interval, and by this treatment, the moths will converge more precisely to the flames in their corresponding sequence as the iterative process proceeds.

If each position update of $m$ moths is based on $m$ different positions in the search space, the local exploitation capability of the algorithm will be reduced. To solve this problem, an adaptive mechanism is proposed for the number of flames so that the number of flames can be reduced adaptively during the iterative process, thus balancing the global search capability of the

algorithm in the search space with the local exploitation capability, as follows.

$$gno = round\left( M - \xi \cdot \frac{M-1}{\xi_{\max}} \right) (12)$$

where $\xi$ and $\xi_{\max}$ are the current iteration number and the maximum iteration number, $M$ is the maximum number of flames. Due to the reduction of the flame, the moth corresponding to the reduced flame in the sequence in each generation updates its own position according to the flame with the worst current fitness value.

The general procedure for solving the problem using the MFO algorithm proposed in this paper is as follows.

(1) Initialization of the MFO algorithm, setting parameters such as the dimensionality of the input optimal decomposition modulus $H$ and penalty factor $\beta$, the moth population search size, the maximum number of iterations, and the logarithmic spiral shape constant.

(2) The variables to be solved are initialized, the moth positions are randomly generated in the search space, and the corresponding fitness value of each moth is evaluated.

(3) The spatial position of the moth is sorted in the order of increasing fitness value and assigned to the flame as the spatial position of the flame in the first generation.

(4) Use Equation (9) to update the position of the current generation of moths.

(5) Reorder the fitness values of the updated moth position and the flame position, and select the spatial position with the better fitness value to update as the position of the next generation flame

(6) Reduce the number of flames using the adaptive mechanism of equation (12).

(7) Return to step (4) to enter the next generation until the number of iterations meets the algorithm requirements.

(8) Output and display the optimization results of decomposing the modal number $H$ and penalty factor $\beta$, and the program ends.

### 2.1.4 Wavelet denoising

After obtaining the optimal value of the modal number and penalty factor, the optimization algorithm uses wavelet threshold to filter the noise information in the speech signal.

The selection of suitable wavelet bases and the determination of the number of decomposition layers in wavelet denoising are the prerequisites to achieve denoising, and then the noise containing speech signal is decomposed to obtain wavelet coefficients of different scales. After comparing these wavelet coefficients, it is found that the noise-containing wavelet coefficients are smaller than those of the actual signal, so a suitable threshold can be selected to compare with the coefficients obtained from wavelet decomposition. When the wavelet coefficients are higher than the threshold, it can be determined that the wavelet coefficients are mainly generated from the actual speech signal and retained; otherwise, it can be assumed that the wavelet coefficients are generated from noise and filtered out. Finally, the wavelet coefficients are wavelet inverse transformed and then reconstructed to achieve wavelet denoising [12].

According to the wavelet denoising principle, it is known that the wavelet denoising effect depends largely on the appropriate threshold and threshold function. Hard thresholding and soft thresholding are two commonly used thresholding functions, and the reconstructed signal after hard thresholding has disadvantages such as discontinuity, oscillation and distortion phenomenon; soft thresholding function [13], although continuous, often appears deviation will lead to the reconstructed signal with high frequency part information loss, edge blurring and other problems. Due to these defects, the traditional threshold function needs to be improved to construct a new threshold function. In addition to the selection of the threshold function, the selection of the wavelet denoising threshold is also very important. In the process of threshold selection, if the threshold is too small, the noise in the signal will not be filtered out; if the selected threshold is too large, the useful components may be filtered out, resulting in deviations in the data. The threshold function selected in this paper is.

$$\hat{\rho}_{i,h} = \begin{cases} sign\left(\rho_{i,h}\right) \cdot \left(\left|\rho_{i,h}\right| - \gamma\right), & \left|\rho_{i,h}\right| \geq \gamma \\ 0, & \left|\rho_{i,h}\right| < \gamma \end{cases} (13)$$

where, $\rho_{i,h}$ denotes the wavelet coefficients; $\hat{\rho}_{i,h}$ denotes the estimated wavelet coefficients. The threshold values selected in this paper are.

$$\ell = \frac{\eta\sqrt{2\lg M}}{\lg\left(i+1\right)} (14)$$

where, $\eta$ denotes the variance of the noise, $M$ denotes the length of the discrete sampled signal, and $i$ denotes the scale of the discrete wavelet transform [14-15].

## 2.2 Network speech enhancement method based on deep neural network learning

For the speech signal of the network English learning platform after denoising in Section 2.1, the network speech enhancement method based on deep neural network learning is used to achieve speech signal enhancement. This method mainly uses deep neural network [16]. This network initializes the network model by training restricted Boltzmann machine (RBM), and gradually optimizes a deep neural network (DNN)

through random gradient descent algorithm. This model can solve the problem that the network model falls into local optimization to a certain extent. The flow diagram of the whole algorithm is shown in Figure 2.

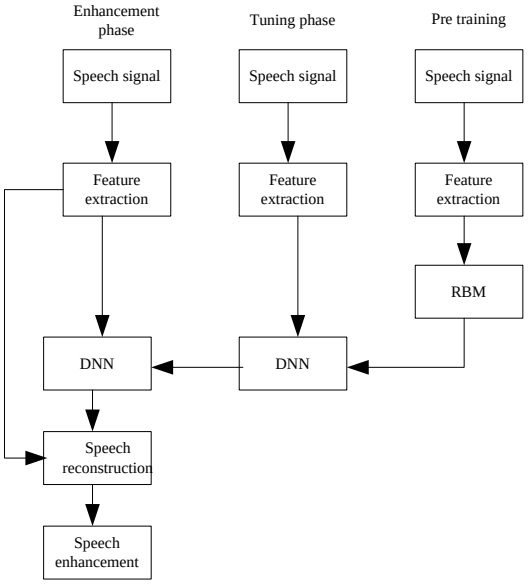



**Figure 2.** The theory of speech enhancement based on regressive DNN

### 2.2.1 Pre-training and fine tuning

(1) Pre-training

Pre-training uses the denoised network English learning platform speech signal $o'(t)$ to train the restricted Boltzmann machine [17], which is an energy-based model whose network is a bipartite graph. The first layer is the visible layer $\phi$ , and the second layer is the hidden layer $\varphi$ , connected in between by a sigmoid activation function, whose joint probability of the visible and hidden layers is defined as

$$q(\phi,\varphi)=\frac{\exp\{-\vartheta(\phi,\varphi)\}}{o'(t)\Gamma} \quad (15)$$

where $\vartheta$ is the energy function of the RBM; $\Gamma$ is the normalization constant. Since the speech signal is a real-valued distribution, the first RBM is usually a Gaussian Restricted Boltzmann Machine (GRBM), followed by a Bernoulli Restricted Boltzmann Machine (BBRBM) in superposition.

For GRBM, the energy function is defined as

$$\vartheta(\phi,\varphi)=\sum_j\frac{(\phi_j-\varphi_i)^2}{2\eta_j^2 o'(t)}-\sum_i\varphi_i-\sum_{j,i}\frac{\phi_j}{\eta_j}\varphi_i\varpi_{ji} \quad (16)$$

where, $j$ and $i$ are the visual layer and implicit layer codes, respectively. $\eta_j$ is the processing variance of the speech signal in the visual layer; $\varpi_{ji}$ is the connection weight between the visual layer and the implicit layer; the conditional probabilities of the visual and implicit layers of GRBM are as follows.

$$\begin{cases} q(\varphi_i=1|\phi)=sigmoid\left(\sum_{j,i}\varpi_{ji}\frac{\phi_j}{\eta_j}+\varphi_i\right) \\ q(\phi_j=\phi|\varphi)=M\left(\eta_j\sum_i\varpi_{ji}\varphi_i+\eta_j^2\right) \end{cases} \quad (17)$$

For BBRBM, the energy function is defined as

$$\vartheta(\phi,\varphi)=-\sum_j\phi_j-\sum_i\varphi_i-\sum_{j,i}\phi_j\varphi_i\varpi_{ji} \quad (18)$$

The conditional probabilities of the visible and implicit layers of the BBRBM are as follows.

$$\begin{cases} q(\varphi_i=1|\phi)=sigmoid\left(\sum_{j,i}\varpi_{ji}\phi_j\right) \\ q(\phi_j=1|\varphi)=sigmoid\left(\sum_{j,i}\varpi_{ji}\varphi_i\right) \end{cases} \quad (19)$$

Gibbs sampling is used for layer-by-layer training. The idea of Gibbs sampling is that given a training sample $\phi 1$ of a speech signal $o'(t)$ , the conditional probability of each node is found according to the formula $Q(\varphi_i=1|\phi)$ , and then the conditional probability of each node in $\phi_2$ is found according to the formula $q(\phi_j=1|\varphi)$ , and then iterated sequentially, at which time the probability of $q(\phi|\varphi)$ converges to the probability of $q(\phi)$ .

The contrast divergence (CD) algorithm is used to update RBM parameters. The output of the previous layer is the input of the next layer, and finally a stacked RBM network is formed.

(2) Fine tuning

Fine tuning is there adaptive learning [18] process, and fine tuning has three main phases: forward transfer, feedback conduction, and modification of weights.

a. Forward pass: input the minimum batch of speech features into the neural network and forward pass the activation values of each layer to the output layer to obtain the cost function based on the minimum mean square error criterion $Loss$ :

b.

$$Loss = \frac{1}{M}\sum_{m=1}^{M}\sum_{\Omega=1}^{\Omega}\left(\hat{Y}_m^{\Omega}\left(W^{\Omega},b^{\Omega}\right)-Y_m^{\Omega}\right)^2 \quad (20)$$

Where, $M$ is the passed batch size; $\Omega$ is the total dimensionality of the speech $o'(t)$ input feature vector; $\hat{Y}_m^{\Omega}\left(W^{\Omega},b^{\Omega}\right)$ , $Y_m^{\Omega}$ are the $m$ th sample $\Omega$ th dimension enhanced speech features [19] and expected speech features of $o'(t)$ in turn.

b. Feedback conduction: first calculate the residuals of each node in the output layer, then pass forward to obtain the residuals of other implied layers, and then find the partial derivatives of the weights of each layer based on the residuals.

c. Modifying weights: the stochastic gradient descent algorithm is used to adjust the network weights $W^{\Omega}$ :

$$W^{\Omega} = W^{\Omega} - \left[\left(\frac{1}{M}\Delta W^{\Omega}\right)+W^{\varpi}W^{\Omega}\right] \quad (21)$$

where, $W^{\varpi}$ is the weight decay parameter, which is used to control the pre-weight amplitude and prevent overfitting. The above steps are so repeatedly executed until the training is completed.

### 2.2.2 Enhancement phase

The speech is feedforward by the deep neural network, and the mean network is used to obtain the output of the hidden layer in the enhancement stage. After obtaining the enhanced speech features, the speech waveform is reconstructed [20], and the waveform reconstruction process is the inverse process of preprocessing. Assuming that the pure speech obtained is $o'(t)$ , the enhanced speech signal $o(l)$ is obtained from the Fourier transform as

$$o(l) = \frac{1}{M}\sum_{\Omega}^{M-1} o'(t) \quad (22)$$

## 3. Experimental analyses

The hardware environment processor of the experiment is Intel Core i7-9750H, with a memory capacity of 16GB and a maximum frequency of 4.5GHz. The software environment is MATLAB2020b. In MATLAB software, we test the effect of denoising and enhancing the speech of the online English learning platform by adding signal-to-noise ratio to the word "dark" in the TIMIT-Speech-Database speech database, and the time domain waveforms of the pure speech signal and the speech signal after adding noise are shown in Figures 3-4. The frequency domain waveforms are shown in Figures 5-6.

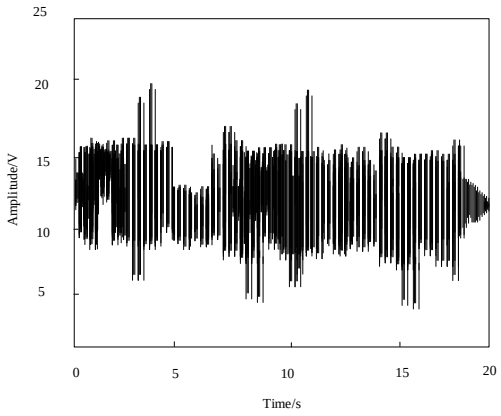



**Figure 3.** Time domain waveform of pure speech signal

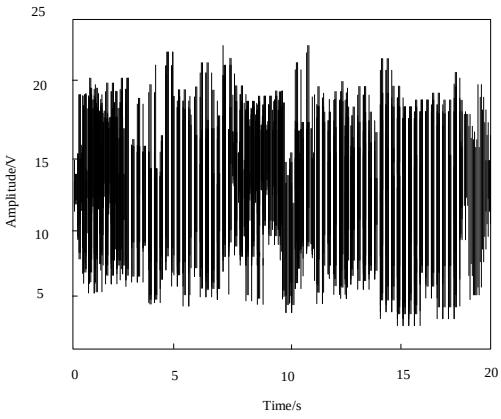



**Figure 4.** Time domain waveform of noisy speech signal

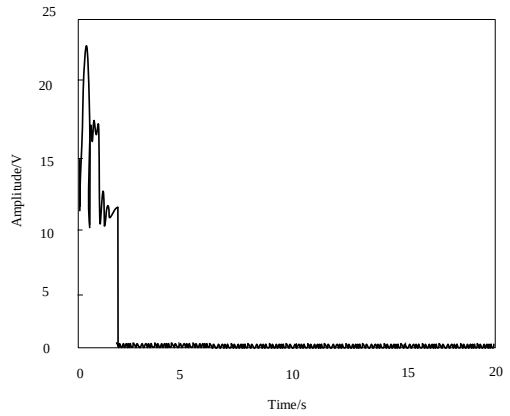



**Figure 5.** Frequency domain waveform of pure speech signal

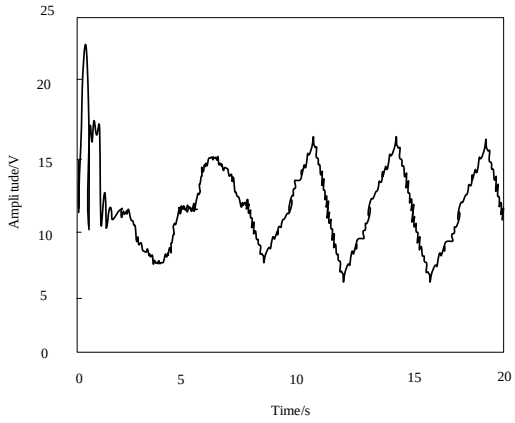



**Figure 6.** Frequency domain waveform of noisy speech signal

Comparing Figure 3 and Figure 4, it can be seen that the time-domain characteristics of the speech signal of the online English learning platform change obviously after the noise is introduced. The time-domain waveform can indicate the change of the signal with time, and the change of the time-domain characteristics leads to the distortion characteristics of the speech signal with the change of time. Comparing Figure 5 and Figure 6, it can be seen that after the noise is introduced into the speech signal of the online English learning platform, the frequency domain characteristics of the signal are homogenized, the frequency characteristics of the signal are not obvious, and the signal is distorted. After the method of this paper denoises the signal, its time-domain and frequency-domain features are shown in Figures 7-8.

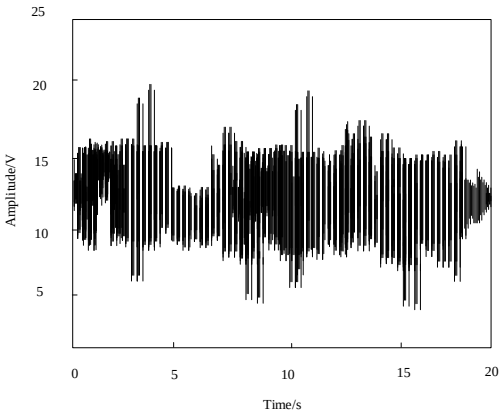



**Figure 7.** Time domain waveform of speech signal after denoising

According to the analysis of Figure 7 and Figure 8, the time-domain and frequency-domain characteristics of the signal are recovered after the denoising of the noisy network English learning platform speech by the method in this paper, and they are highly consistent with the original time-domain and frequency-domain characteristics, which proves that the method in this paper has a good denoising effect on the noisy speech signal. In order to test the enhancement effect of this method on the English speech signal, four types of unknown noise completely different from the original speech signal are used, namely A1 (white noise), A2 (pink noise), A3 (babble noise) and A4 (Leopard noise). These four types of noise are from noise_ 92 noise database. Then three different signal-to-noise ratios of -5dB, 0dB and 5dB and pure speech signals are synthesized. The enhancement effect of this paper on the noisy English speech signal is shown in Table 1, and the enhancement effect is mainly reflected by the speech quality perception assessment PESQ value, which takes the value range of -0.5~4.5, and a larger value indicates better speech quality.

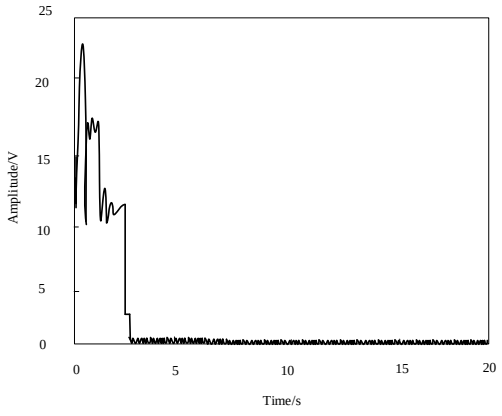
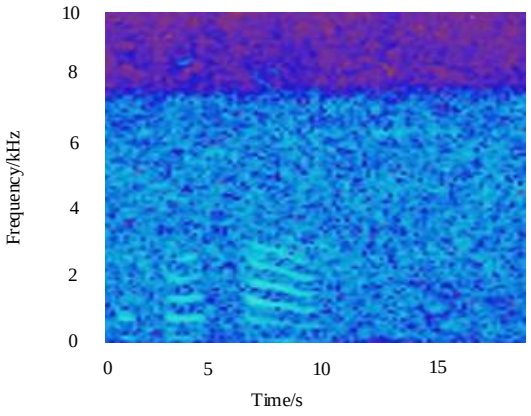



**Figure 8.** Frequency domain waveform of speech signal after denoising

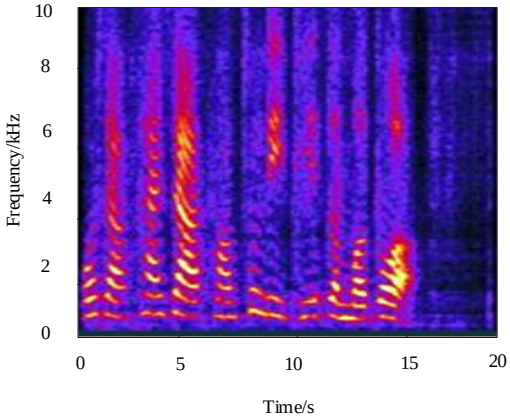
Table 1. enhancement effect of this method on noisy speech signal

| Noise type | Signal to noise ratio /dB | Before enhancement | After enhancement |
|---|---|---|---|
| | -5 | -0.2 | 4.3 |
| A1 | 0 | -0.1 | 4.2 |
| | 5 | -0.3 | 4.2 |
| | -5 | -0.3 | 4.4 |
| A2 | 0 | -0.3 | 4.4 |
| | 5 | -0.3 | 4.3 |
| | -5 | -0.1 | 4.3 |
| A3 | 0 | -0.3 | 4.4 |
| | 5 | -0.3 | 4.4 |
| | -5 | -0.1 | 4.3 |
| A4 | 0 | -0.3 | 4.3 |
| | 5 | -0.3 | 4.3 |
| Mean | - | -0.2 | 4.3 |

By analyzing the data in Table 1, it can be seen that in many scenarios, the PESQ values of the speech quality perception evaluation of noisy English speech signals are significantly different before and after the application of this method. Before the application, the PESQ values of the speech quality perception evaluation of noisy English speech signals are all negative. After the application of this method, the PESQ values of the speech quality perception evaluation of noisy English speech signals are higher than 4.0. It shows that after the use of this method, the PESQ value of speech quality perception evaluation of noisy English speech signals is improved, and the speech quality is improved. Taking 5dB noise of scene A1 as an example, the spectrogram of English speech signal before and after the enhancement of this method is shown in Figure 9.




(a) Before this method is enhanced




(b) After this method is enhanced

**Figure 9.** This method is used to enhance the spectrogram of speech signals before and after enhancement

It can be seen from Figure 9 that before using the method proposed in this paper to enhance the speech signal of the network English platform, the spectral characteristics of the speech signal are not obvious, indicating that the quality of the speech signal is not high. After using the method proposed in this paper to enhance the speech signal of the network English platform, the spectral characteristics of the speech signal are prominent and the quality of the speech signal is improved [21].

## 4. Conclusion

In view of the shortcomings of traditional speech enhancement methods, this paper proposes a speech enhancement method based on deep neural network for online English learning platform based on previous research. The difference between this method and other deep neural network based speech enhancement methods is that the special denoising method based on optimized VMD algorithm for online English learning platform speech signal denoising can effectively remove the noise information in the speech signal, optimize signal quality. The speech of the online English learning platform is effectively enhanced while the denoised speech signal is input into the deep neural network, and applied to the English learning application platform. Through experiments from three perspectives: speech signal denoising, speech quality enhancement and speech spectrum information, the usability of the method in this paper is confirmed.

## References

[1] Weisser, A. , Buchholz, J. M. (2019). Conversational speech levels and signal-to-noise ratios in realistic acoustic

conditions. Journal of the Acoustical Society of America, 145(1):349-360.
[2] Jamal, N. , Fuad, N . & Sha'Abani, M. (2021) . A Comparative Study of IBM and IRM Target Mask for Supervised Malay Speech Separation from Noisy Background. Procedia Computer Science, 179(4):153-160.
[3] Sivapatham, S., Kar, A. & Ramadoss, R. (2021). Performance analysis of various training targets for improving speech quality and intelligibility. Applied Acoustics, 175(12):107817.
[4] Liu, S., Li, Y. & Fu, W. (2022) Human-centered attention-aware networks for Action recognition, International Journal of Intelligent Systems, online first, doi: 10.1002/int.23029
[5] Sadasivan, J., Dhiman, J. K. & Seelamantula, C. S. (2020). Musical noise suppression using a low-rank and sparse matrix decomposition approach. Speech Communication, 125(2):41-52.
[6] Bayer, F. M., Kozakevicius, A. J. & Cintra, R. J. (2019). An Iterative Wavelet Threshold for Signal Denoising. Signal Processing, 162(SEP.):10-20.
[7] Liu, S., Wang, S., Liu, X., et al. (2022) Human Inertial Thinking Strategy: A Novel Fuzzy Reasoning Mechanism for IoT-Assisted Visual Monitoring, . IEEE Internet of Things Journal, online first, 2022, doi: 10.1109/JIOT.2022.3142115
[8] Demir, O. T., Bjornson, E. (2021). The Bussgang Decomposition of Nonlinear Systems: Basic Theory and MIMO Extensions [Lecture Notes]. IEEE Signal Processing Magazine, 38(1):131-136.
[9] Wakisaka, Y. , Iida, D. & Oshida, H.(2021). Fading Suppression of Φ-OTDR With the New Signal Processing Methodology of Complex Vectors Across Time and Frequency Domains. Journal of Lightwave Technology, 39(13): 4279-4293.
[10] Sedov, E. V., Chekhovskoy, I.S. & Prilepsky, J. E. (2021). Neural network for calculating direct and inverse nonlinear Fourier transform. Quantum Electronics, 51(12):1118-1121.
[11] Li, W. S., Xu, W. J. & Zhang, T. (2021).Improvement of Threshold Denoising Method Based on Wavelet Transform. Computer Simulation, 38(06):348-351,356.
[12] Bo, X. , Zxa, B. & Zw, C.(2020). Gamma spectrum denoising method based on improved wavelet threshold[J]. Nuclear Engineering and Technology, 52( 8):1771-1776.
[13] Zaeni, A. , Kasnalestari, T. & Khayam, U. (2019). Partial discharge signal denoising by using hard threshold and soft threshold methods and wavelet transformation. IOP Conference Series: Materials Science and Engineering, 602(1):012034.
[14] Hameed, A. S. (2021). Speech compression and encryption based on discrete wavelet transform and chaotic signals. Multimedia Tools and Applications, 80(9): 13663-13676.
[15] Barkalov, K., Lebedev, I., & Kozinov, E. (2021).Acceleration of Global Optimization Algorithm by Detecting Local Extrema Based on Machine Learning. Entropy, 23(10): 1272.
[16] Kwasny, D., Hemmerling, D. (2021). Gender and age estimation methods based on speech using deep neural networks. Sensors, 21(14): 4785.
[17] Lü, X., Meng, L. & Chen, C. (2020). Fuzzy Removing Redundancy Restricted Boltzmann Machine: Improving Learning Speed and Classification Accuracy. IEEE Transactions on Fuzzy Systems, 28(10):2495-2509.
[18] Saxena, D., Singh, A. K. (2022). Auto-adaptive learning-based workload forecasting in dynamic cloud environment. International Journal of Computers and Applications, 44(6): 541-551.
[19] Kim, G., Lee, H. & Kim, B. K.(2019). Unpaired Speech Enhancement by Acoustic and Adversarial Supervision for Speech Recognition. IEEE signal processing letters, 26(1):159-163.
[20] Jeeva, M., Nagarajan, T. & Vijayalakshmi, P. (2020). Adaptive multi-band filter structure-based far-end speech enhancement. IET Signal Processing, 14(5):288-299.
[21] Liu, S., Xu, X., Zhang, Y., et al. (2022). A Reliable Sample Selection Strategy for Weakly-supervised Visual Tracking, IEEE Transactions on Reliability, online first, doi: 10.1109/TR.2022.3162346