International Journal of Innovative
Computing, Information and Control
Volume **16**, Number **5**, October **2020**

 ISSN 1349-4198
pp. 1711–1723

RESEARCH ON ENGLISH SPEECH ENHANCEMENT ALGORITHM BASED ON IMPROVED SPECTRAL SUBTRACTION AND DEEP NEURAL NETWORK

Qiaoling Zhou

International College
Fujian Agriculture and Forestry University
No. 15, Shangxiadian Road, Cangshan District, Fuzhou 350002, P. R. China
zhouqiaolinglw@163.com

Received March 2020; revised July 2020

Abstract. *In order to solve the introduced unstructured voiceless problems of conventional spectrum subtraction in English speech signals enhancement, this paper proposes a novel English speech signals enhancement algorithm. This algorithm uses an improved minimal controlled recursive averaging (IMCRA) method to estimate noise spectrum, and tracks the estimated noise spectrum in real time. Then, the deep neural network (DNN) is used to construct the nonlinear mapping function of log amplitude spectrum between speech with noises and ideal pure speech for English speech enhancement. To validate the feasibility and effectiveness of the proposed algorithm, the standard IEEE speech signals and Noise-91 noise signals are used for experiments. Experimental results have shown that the proposed IMCRA method has stronger ability to avoid noises in speech signals, and the DNN method can well recover the speech components and spectrum structure polluted by noises. To enhance English speech in daily international speech communication, the proposed combination method has strong robustness to various real noise environments, and brings significant improvement to interpersonal communication and human computer communication.*

Keywords: Improved spectrum subtraction, Deep neural network, Speech enhancement, Amplitude spectrum, English communication

1. **Introduction.** As one of the most widely used languages in the world, English is widely used in normal interpersonal communication. In the actual English communication, English themed pronunciation has become a key factor to communication. In fact, in the normal English communication environment, speech usually receives noise interference, which leads to the degradation of speech quality. In order to solve the problem of English speech communication caused by noises, the researchers have conducted in-depth researches on the enhancement of English speech. The purpose of speech enhancement is to suppress or separate noise. Then, the target speech will be recovered as undistorted as possible from the speech signals polluted by noises. The enhancement of speech signals will improve the perception quality, intelligibility, and recognition accuracy of speech, or as a front-end to improve the recognition accuracy. The speech enhancement has been widely used in speech communication systems, such as hearing aid equipment and automatic speech recognition system, and bringing huge improvement of inter-personal communication and human-computer communication.

In the past few decades, many speech enhancement methods have been proposed, which greatly promote the development of speech enhancement technology. However, in the actual scene, especially in the single channel and non-stationary noise environment, due to

DOI: 10.24507/ijicic.16.05.1711

the lack of time-spatial domain information in the multi-channels scene and the difficulty in modeling and using the structured information in an effective way [1], single channel speech enhancement is still a very challenging topic. According to whether the prior information of speech signals and noise signals is needed, the existing speech enhancement methods can be divided into two categories: supervised methods and unsupervised methods. Classical unsupervised enhancement methods include spectral subtraction, Wiener filtering, and statistical model based methods [2], which are generally based on the assumption that speeches and noise are not correlated and the spectrum coefficients obey Gaussian distribution. The enhancement performance depends on the accuracy of voice activity detection or noise power spectrum estimation. Generally, this kind of method can achieve better noise suppression effect in the stationary noise environment. However, for the non-stationary noise or low signal-noise-ratio (SNR) environment, such methods will become more difficult to track and accurately estimate the noise spectrum in real-time condition, which will seriously affect the performance of speech enhancement. As a data-driven method, supervised speech enhancement methods are divided into two categories: representative of dictionary based methods [3] and neural network based methods [4]. This kind of method starts from the data directly, obtains the speech model and the noise model by model's training procedure, or uses the prior information to learn the non-linear mapping between the noisy speech signals and the pure speech signals. Due to the fact that the assumption of signals distribution is no longer required, supervised enhancement methods often achieve better enhancement effect than conventional unsupervised methods in low SNR or non-stationary noises environment.

In the early researches of speech enhancement, Boll [5] proposed a spectrum subtraction (SS) method to reduce the noises from the estimated noises spectrum, and the enhanced speech power spectrum is obtained. This method is simple and easy to reduce noises, but it leads the phenomenon of "music noise". Berouti et al. [6] introduced over subtraction factor to balance speech distortion and noise residual effectively. Thomson [7] proposed a multi-window spectral estimation method to estimate the orthogonal power spectrum of each frame from speech signals, and such method has less error than the conventional periodogram method. The periodogram method only uses one sliding window to estimate the noisy speech power spectrum. Voice activity detector [8] is used to analyze the start and end points of speech signals. The noise spectrum of the leading silent segments is used as the estimated noise spectrum to extract useful information from speech segments.

In addition, deep learning has made a great contribution in the field of noise recognition and speech enhancement [9]. In [10], deep neural network (DNN) model is used to establish a nonlinear mapping function between the log energy spectrum of noisy speech and the log energy spectrum of pure speech, and the global equilibrium variance method is used to solve the problem of over-smoothed speech spectrum after the enhancement of DNN model. The results reported that such method has a good ability to suppress the noises that are not included in the training set and the non-stationary noises that are included in the real scene. In [11], DNN model estimation is used to calculate the ideal binary masking in CASA, which can effectively enhance the intelligibility of noisy speech. In addition, deep recurrent neural network (DRNN) and long short term memory (LSTM) are used to model the long-term and short-term correlations of speech signals using cyclic connection or storage and gate structure units, which further improve the performance of speech and noise signals separation [12]. In [13], the generative adversarial network (GAN), which has been successfully applied in the field of computer vision and image processing, has been applied to the fields of speech signals enhancement and has been achieved useful results. [14] proposes a speech enhancement method using reinforcement learning to self-optimized DNN model. The reward feedback training network is based on

the quantitative index of human auditory scores. Subjective and objective experimental results have shown that the method is effective for speech enhancement under limited sample data.

Considering the sparse features of speech signals in time domain and frequency domain and the spectrum preserving features from DNN model in speech enhancement application, this paper first estimates the noise spectrum with the help of the improved minimum control recursive average algorithm proposed by Cohen [15]. By smoothing the noisy speech signals of each frame in the first order, then the probabilities of the existence of speech signals in different frequency bands are estimated to obtain the ratio between the local energy value of the noisy speech signals. The ratio to the minimum value of each frame window is used to determine whether the frame is a speech segment or a silent segment. If the current frame is a speech segment, the noise spectrum in the frequency band is equal to that of the previous frame. Otherwise, the noise spectrum in the frequency band is searched and tracked in the real-time condition to obtain the minimum value and then multiplied by the offset compensation. Then, the nonlinear mapping function of log amplitude spectrum between speech with noises and ideal pure speech is separated by the training procedure of DNN model, and the missing components of speech structure are recovered by the learned nonlinear function. To validate the proposed method has feasibility and effectiveness in speech enhancement, autoregressive speech enhancement method based on DNN [4] (DNN in short), spectrum subtraction speech enhancement method based on minimum controlled recursive averaging (MCRA) [15], and supervised spectrum subtraction speech enhancement method based on improved minimum controlled recursive averaging (IMCRA) [24] are included in the experiments for the comparison for the proposed combination method. The experimental results have shown that the proposed method effectively suppresses the noise signals and retains the speech components. In addition, such method is superior to the baseline method in terms of the perception quality and log spectral distortion performance evaluation index.

The rest of the paper is organized as follows. Section 2 describes the preprocessing of speech signals based on improved spectrum subtraction. Section 3 introduces the proposed speech enhancement method based on improved spectrum subtraction and deep neural network. Section 4 presents the experimental results, comparisons and related discussing works. Section 5 concludes the paper and points out our important future work.

2. Speech Signal Preprocessing Based on Improved Spectrum Subtraction.

To estimate the noise spectrum from speech signals, the conventional spectrum subtraction often uses the statistical average value of mute frame to replace the noise spectrum for each frame in the speech signals. Using the statistical value let the noise spectrum present a series of random peaks and cause the "music noise" spectrum [16] when the speech signals have dramatic changes. Due to the fact that the "music noise" will influence the intelligibility of speech signals, the "music noise" spectrum reduction is the focused problem for the improved spectrum subtraction method. In this paper, a multiple window spectrum estimation with the adaptive spectrum reduction coefficient to reduce the "music noise" spectrum and enhance the robustness of spectrum subtraction method is applied.

2.1. Spectrum subtraction estimated by multi-window spectrums.

Based on the multi-window spectrums estimation (MWS) algorithm, the spectrum value of the same data sequence is calculated by adding multiple orthogonal data windows. Then the weighted average is more accurate than the periodogram method, which can effectively reduce the "music noise". The main steps of this method are as follows [17].

1) The speech signal is pre-weighted by a high-pass filter to improve the high frequency amplitude.

2) According to the short-time stationary characteristic of the speech signals, the noisy speech signal frames $y(n)$ together with Hamming windows achieve n segments, of which the overlapped parts among frames are called m, and the ith frame is $y_i(n)$.

3) The discrete short-time Fourier transform of $y_i(n)$ is used to get the spectrum value $Y_i(k)$ of the speech signal at the ith frame, k belongs to $k \in [0, N]$ and N is the length of a frame. The amplitude and phase angle, $|Y_i(k)|$ and $\theta(Y_i(k))$, are obtained respectively. The mean value of the amplitude spectrum between left frames and right frames at the ith frame is calculated as $\left|\bar{Y}_i(k)\right| = \frac{1}{2M+1}\sum_{j=-M}^{j=M}|Y_{i+j}(k)|$. The frame is smoothed by $2M+1$ centered at frame ith, which makes $M = 1$.

4) According to MATLAB's multi-window spectrums estimation function: $pmtm()$, the smoothed power spectrum after windowing is calculated as $P_y(k,i) = \frac{1}{2M+1}\sum_{j=-M}^{j=M} P_y(k, i+j)$.

5) If there is a period of silence IS before the speech starts, and the corresponding frame length is N_{IS}, then the noise power spectrum is uniformed as $D_{IS}(k) = \frac{1}{N_{IS}}\sum_{i=1}^{N_{IS}} P_y(k,i)$.

6) The gain function $G(k,i)$ is calculated from the spectral subtraction Formula (1). The over subtraction factor α mainly affects the distortion degree of the speech spectrum. The gain compensation factor β is used to control the residual noise signals, including "music noise". The formula is as follows:

$$G(k,i) = \begin{cases} \beta P_y(k,i) - \alpha D_{IS}(k)/P_y(k,i), \\ \beta P_y(k,i) - \alpha D_{IS}(k) \geq 0; \\ \beta D_{IS}(k)/P_y(k,i), \\ \beta P_y(k,i) - \alpha D_{IS}(k) < 0. \end{cases} \tag{1}$$

7) Amplitude spectrum after spectral subtraction is calculated as $\left|\hat{S}_i(k)\right| = G(k,i)\left|\bar{Y}_i(k)\right|$.

8) Finally, the IDFT method is performed on both $\hat{S}_i(k)$ and $\theta(Y_i(k))$ with phase angles to recover the enhanced speech signals:

$$\hat{s}(n) = IDFT\left[\left|\hat{S}_i(k)\right| \exp[j\theta(Y_i(k))]\right] \tag{2}$$

2.2. **Adaptive spectral reduction coefficient.** When the power spectrum of noisy speech is negative compared with the estimated noise spectrum, the over reduction factor α and the gain compensation factor β are introduced to compensate the "music noise" caused by the noise spectrum estimation [18]. Table 1 shows the effect of different spectral subtraction coefficients for signal-to-noise ratio (SNR) and intelligibility of speech signals.

TABLE 1. Effect of different spectral subtraction coefficients for SNR and intelligibility

α, β	$\alpha = 4$, $\beta = 0.001$		$\alpha = 4$, $\beta = 0.098$		$\alpha = 2.8$, $\beta = 0.098$	
background noise	Signal-to-noise ratio	Speech intelligibility	Signal-to-noise ratio	Speech intelligibility	Signal-to-noise ratio	Speech intelligibility
White	9.6703	1.6207	9.8036	1.8036	9.9731	1.2309
Babble	7.4219	1.4219	7.4956	1.4956	8.3965	1.0924
Volve	9.6808	1.6808	9.8268	1.8268	10.8402	1.2376
Machinegun	1.4212	1.5788	0.9479	0.9479	−0.8188	0.9879
F16	8.1684	1.1684	8.2416	1.2416	9.3815	1.0312

2.3. **IMCRA for noise power spectrum estimation.** To solve the problem of residual noise caused by conventional spectral subtraction, this paper uses the improved minimal controlled recursive averaging (IMCRA) method [19] to smooth multiple spectrums, and obtains the probability of speech existence and the threshold of speech start-stop decisions to update the noise spectrum. After the noise spectrum is updated, the current estimation Formula (3) of the noise spectrum at the ith frame is given as follows:

$$P_d(k,i) = \alpha_{isc}\bar{P}_d(k,i-1) + [1-\alpha_{isc}]P_y(k,i-1) \tag{3}$$

where $P_d(k,i)$ represents the spectrum value k at the first ith window and $P_y(k,i-1)$ denotes smoothing power spectrum of the noisy speech at the previous frame. According to [6,10], the ideal smoothing coefficient is set as $\alpha_{isc} = 0.95$, $\bar{P}_d(k,i-1)$ is divided by a deviation compensation factor β_m $(0 < \beta_m < 1)$, and the noise power spectrum after compensation is calculated as $\hat{P}_d(k,i) = \bar{P}_d(k,i)/\beta_m$. Compared with the conventional MCRA method to update noise, IMCRA method has better adaptability to the mutation of noise spectrum.

3. Speech Enhancement Based on Improved Subtraction and Deep Neural Network.

3.1. **Deep neural network structure.** The speech enhancement method based on IMCRA has a good ability to suppress noise signals. However, by observing the speech spectrum generated by the improved spectral subtraction, it can be found that the speech spectrum is also damaged while the noise is removed. There is a lack of block and spectrum components in the enhanced speech signals, which will cause the damage of the harmonic components of the speech and inevitably introduce the distortion structure of the speech signals. In order to improve the perceptual quality and intelligibility of the enhanced speech signals, considering the effective spectrum reconstruction characteristics of the speech enhancement method based on DNN model, this paper uses DNN model to post-process the speech signals that are enhanced by the IMCRA method. Firstly, the noisy speech signals can be suppressed by IMCRA method. Secondly, the enhanced speech signals as the input of DNN model may reduce the complexity of network during training procedure of DNN.

Figure 1 shows the specific DNN model structure that is designed for speech signals enhancement. The designed DNN model includes the following.

1) Input layer: the input of the DNN model has two characteristic matrices composed of multi-frames vector TR_Y and multi-frames vector TI_Y, respectively.

2) Convolutional layer: the DNN model used in the proposed method consists of three convolutional layers. Among them, the convolutional filter size of the first layer is 7×7, the filter size of the other two layers is 3×3, the numbers of all filters are 64, 128, 256, respectively, and the convolutional step size is set to 1×1. The ELU (experimental linear units) activation function is adopted to activate the nodes in each layer.

3) Pooling layer: after the convolutional layer with ELU activation, following are three Max pooling filter layers. The pooling filter size and step size are set to 3×3 and 2×2, respectively.

4) Fully connected layer: the number of nodes in two fully connected layers is 1024.

5) Output layer: the last part of the DNN model is two fully connected layers with each one containing 129 nodes. The nodes are corresponding to the 129 dimensional speech signals output of real part and virtual part, respectively.

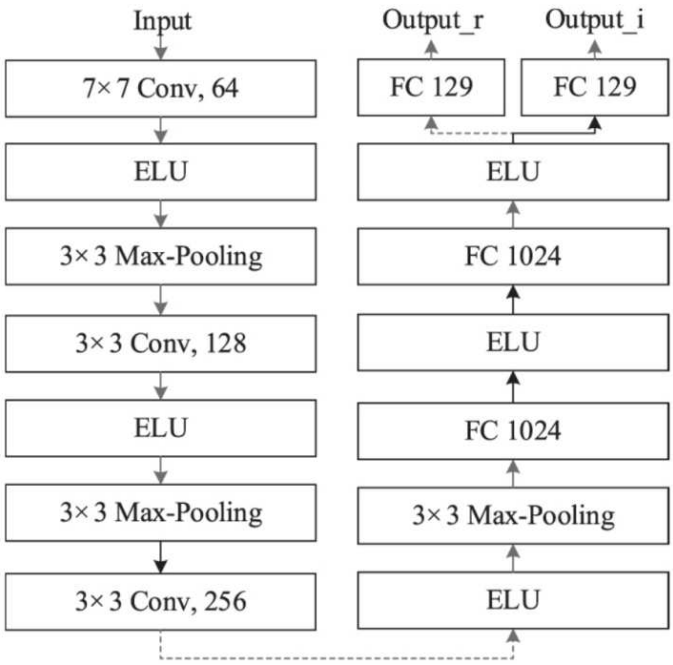


FIGURE 1. The designed structure of deep neural network model

3.2. **Speech enhancement process.** In the process of speech signals enhancement using DNN model, let s and d denote pure speech signals and additive noise signals, the noisy speech signals can be expressed as follows:

$$y = s + d \tag{4}$$

Speech signals enhancement is to calculate the estimated value $\hat{s}$ of s under the condition of given y. Assumed that the STFT forms of y, s, and $\hat{s}$ at the nth frame are $Y(n,k)$, $S(n,k)$, and $\hat{S}(n,k)$, respectively, with the band sequence number $k = 1, 2, \ldots, K$. Assumed that the phase of $S(n,k)$ is set to $\varphi(n,k)$, then the equation $S(n,k) = |S(n,k)|e^{j\times\varphi(n,k)}$ is satisfied. The amplitude and phase can be represented by the real part $S_r(n,k)$ and the virtual part $S_i(n,k)$:

$$|S(n,k)| = \sqrt{S_r(n,k)^2 + S_i(n,k)^2} \tag{5}$$

$$\varphi(n,k) = \tan^{-1}\frac{S_i(n,k)}{S_r(n,k)} \tag{6}$$

Therefore, the amplitude and phase of the real part and the virtual part of $S(n,k)$ can be estimated simultaneously. Then, for the signal at the nth frame, the speech enhancement task in STFT domain is to minimize the following error function:

$$Er = \sum_{k=1}^{K}\left[\left(\hat{S}_r(n,k) - S_r(n,k)\right)^2 + \left(\hat{S}_i(n,k) - S_i(n,k)\right)^2\right] \tag{7}$$

Let $S_r(n)$, $S_i(n)$, $\hat{S}_r(n)$, and $\hat{S}_i(n)$ denote the real part and virtual part vectors of the pure speech frame and their estimated values at the nth frame, respectively. The error function can be rewritten as:

$$Er = \left\|\hat{S}_r(n) - S_r(n)\right\|_2^2 + \left\|\hat{S}_i(n) - S_i(n)\right\|_2^2 \tag{8}$$

In order to train the DNN model which can estimate $\hat{S}_r(n)$ and $\hat{S}_i(n)$ synchronously, the real part and the virtual part of $Y(n,k)$ are used as the input of the DNN model, and the real part and the virtual part of $S(n,k)$ are used as the output of the DNN

model. In order to adapt to the training of DNN model and ensure the simplicity of speech reconstruction, hyperbolic tangent function is used to compress the real part and virtual part, respectively. Then, the TR (Tanh-compressed real component) and the TI (Tanh-compressed imaging component) components are obtained as the input and output characteristics of the DNN model:

$$TI_Z(n,k) = b\frac{1-e^{-a\times Z_i(n,k)}}{1+e^{-a\times Z_i(n,k)}} \tag{9}$$

$$TR_Z(n,k) = b\frac{1-e^{-a\times Z_r(n,k)}}{1+e^{-a\times Z_r(n,k)}} \tag{10}$$

where $a = 0.5$ and $b = 10$ are empirical parameters obtained from [19]. Z can be treated as the training characteristics and objectives from Y and S, respectively. Based on the above training characteristics and objectives, the basic idea of the proposed method can be described as follows: the error function is minimized by simultaneously training the DNN model to optimize two parameter sets, λ and θ, and such two parameter sets are used to construct two highly complex nonlinear function sums f_λ and f_θ, as shown in Equation (11):

$$Er = \|f_\lambda(X(n)) - TR_S(n)\|_2^2 + \|f_\theta(X(n)) - TI_S(n)\|_2^2 \tag{11}$$

Then, the enhanced speech signals output is obtained:

$$\hat{T}R_S(n) = f_\lambda(X(n)) \tag{12}$$

$$\hat{T}I_S(n) = f_\theta(X(n)) \tag{13}$$

The input feature $X(n)$ is composed as:

$$\begin{aligned} X(n) = [&TR_Y(n-N), TR_Y(n-N+1), \ldots, TR_Y(n), \ldots, TR_Y(n+N); \\ &TI_Y(n-N), TI_Y(n-N+1), \ldots, TI_Y(n), \ldots, TI_Y(n+N)] \end{aligned} \tag{14}$$

where $X(n)$ consists of the vector TR_Y from the frame $(2N+1)$ centered at the nth frame n and the vector TI_Y from the frame $(2N+1)$. The number of $(2N+1)$ is the input window length.

DNN model adopts multi-tasks learning mode to train λ and θ simultaneously, and the small batch gradient descent method is used for training. The following defined cost function is used for training DNN model:

$$C(\lambda,\theta) = \frac{1}{M}\sum_{n=1}^{M}\left[\|f_\lambda(X(n)) - TR_S(n)\|_2^2 + \|f_\theta(X(n)) - TI_S(n)\|_2^2\right] \tag{15}$$

where M is the size of mini-batch that is used in network training.

When speech signals enhancement is performed after the training of DNN model, the sum of the estimated values $\hat{T}R_S(n)$ and $\hat{T}I_S(n)$ for the training target is obtained for the noisy speech y_n at the nth frame, and then the real part and the virtual part of $\hat{S}(n)$ are calculated using the estimated values:

$$\hat{S}_r(n) = -\frac{1}{\alpha}\log\left(\frac{\beta - \hat{T}R_S(n)}{\beta + \hat{T}R_S(n)}\right) \tag{16}$$

$$\hat{S}_i(n) = -\frac{1}{\alpha}\log\left(\frac{\beta - \hat{T}I_S(n)}{\beta + \hat{T}I_S(n)}\right) \tag{17}$$

Finally, time domain reconstruction of speech signals enhancement $\hat{S}(n)$ is obtained by the inverse STFT (ISTFT) algorithm:

$$\hat{s}_n = ISTFT\left(\hat{S}_r(n) + j \times \hat{S}_i(n)\right) \tag{18}$$

Combined with the above IMCRA algorithm, the reconstructed enhanced speech signals are recovered, and the music noise generated by the spectrum subtraction is eliminated. Figure 2 shows the specific steps of the DNN model combined with the IMCRA algorithm.

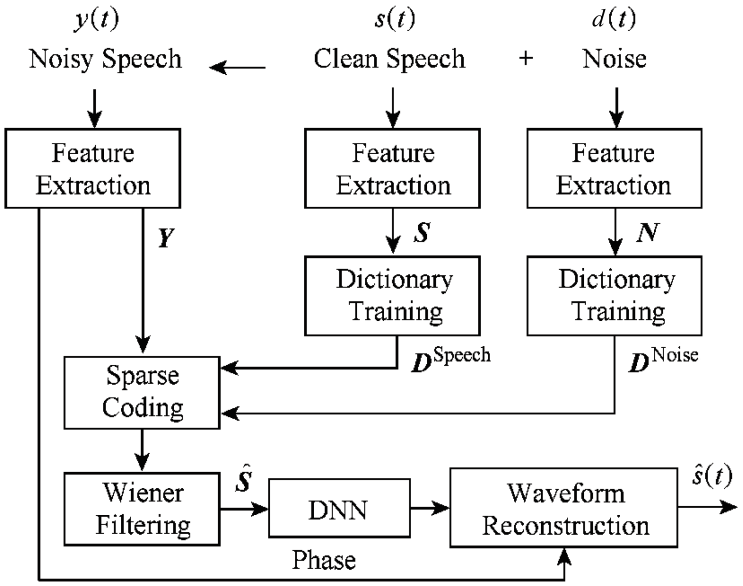


FIGURE 2. English speech signals enhancement based on IMCRA-DNN

In the enhancement stage, IMCRA method is used to process the speech amplitude spectrum. The amplitude spectrum feature normalization combines the amplitude spectrum feature of adjacent frames as the input of the well-trained DNN model, and the output of the network is the logarithmic amplitude spectrum of the enhanced speech signals. Based on the fact that speech signals are not sensitive to phase information, the speech signals in the time domain can be reconstructed by combining the phase information of noisy speech and ISTFT algorithm.

4. **Experimental Results and Analysis.**

4.1. **Data set and evaluation index.** In this experiment, pure English speech is selected from IEEE statement data set [21], and noise signal is taken from noise data set of NOISEX-92 standard. The IEEE pure English speech statements are composed of 720 male speakers; the noise database NOISEX-92 is composed of 15 kinds of real scene noises, mainly including a variety of military environment noises and some common environment noises. Such noises are non-stationary and each length is about 4 min.

In order to increase the diversity of data and avoid overfitting problem, each kind of noises is divided into the first two segments for training and validation [4]. In the training stage, 200 sentences of pure English speech signals are randomly selected from the IEEE pure English speech statement data set. The noise signals are mixed by seven SNR kinds of noises with −9 dB, −6 dB, −3 dB, 0 dB, 3 dB, 6 dB, and 9 dB, and the first half of the four kinds of noises in the NOISEX-92 noises data set. Such four kinds of noises are: F16

double cabin noise, factory noise, HF channel high frequency noise and Gaussian white noise. In this way, 5600 noisy speech signals are generated as training data set. 10% of the training set is selected as the validation set. After each epoch, the DNN model performance is tested on the validation set, and the parameters of DNN model with the best performance in the validation set are selected as the final model for DNN training. In the testing stage, another 20 pure English speech sentences are selected from the pure English speech statement data set and mixed with the second half of the four kinds of noises used in the training stage according to -5 dB, 0 dB, 5 dB, and 10 dB, respectively, and 320 noisy speech statements are generated as the test set.

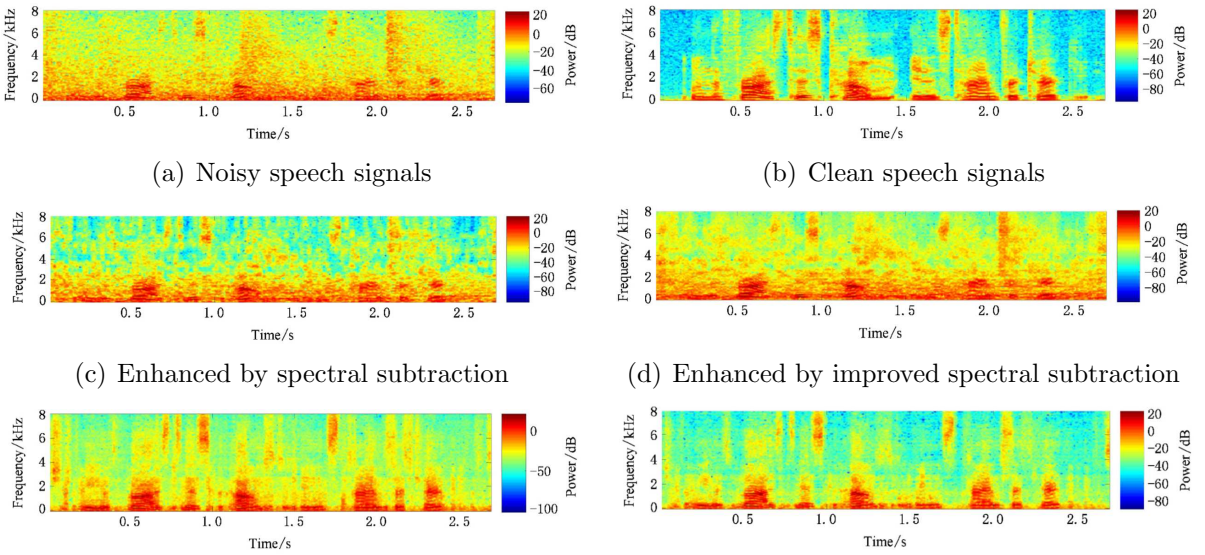
In this paper, the perceptual evaluation of speech quality (PESQ) [22] value and the log spectral distance (LSD) [23] value are selected as indicators to evaluate the performance of the proposed method. Among them, PESQ value is a widely used objective evaluation method, which focuses on the evaluation of the overall quality of speech processing, and its score is in the range of $[-0.5, 4.5]$. LSD value represents the short-term power spectrum difference between pure speech signals and enhanced speech signals. Smaller LSD value represents smaller spectrum distortion of the enhanced speech signals.

4.2. **Baseline methods and parameters setting.** In this paper, two methods are selected as the baseline: autoregressive speech enhancement method based on DNN [4] (DNN in short) and supervised spectral subtraction speech enhancement method based on improved minimum controlled recursive averaging (IMCRA) [24]. These two methods use the same training and testing data as the proposed method used in the experiments. In the DNN autoregressive method, the input of the model is the normalized log amplitude spectrum of noisy speech signals, and the output is the log amplitude spectrum of pure speech signals. The speech enhancement method based on IMCRA uses multiple spectrum smoothing method to obtain the probability of speech existence and the threshold of speech start-end decision to estimate the noise spectrum, which can dynamically describe the information of speech signals and noise signals. Speech dictionary and noise dictionary are obtained by pre-training. According to experience, the speech dictionary base is set to 100, and the noise dictionary base is set to 60. The 512 dimensional amplitude spectrum calculated by Hamming window is selected as the training feature of the speech dictionary. The window length is 32 ms and the frame shift is 8 ms. In the IMCRA method, the value of sparse constraint λ is set to 0.1 empirically, and the size of time-frequency atom is set to 8 according to experience [24].

In the proposed method, the long feature vector of 5 frames from the improved spectrum subtraction method is incorporated as the input of DNN model, and the output of DNN model is the logarithmic amplitude spectrum of the pure speech signals corresponding to the current frame. The input layer of the network is 257×5 nodes, and the output layer is 257 nodes. In order to extract the implicit features in the speech signals, three hidden layers are included in the DNN model, and each hidden layer contains 2048 nodes to extract implicit features by using convolutional kernels. Due to the fact that there are $2048 * 3$ nodes in the hidden layers, the dropout and batch normalization strategies are introduced to avoid gradient disappearance during DNN model training. The dropout parameter is set to 0.2, and the batch size is selected as 1024. After 300 epochs, the network parameters are no longer updated, and the best network parameters are selected as the well-trained DNN model. In the experiment, the modified linear activation function [25] is selected as the activation function of the hidden layers. Compared with the Sigmoid function and the hyperbolic tangent function, the activation function is more consistent with the excitation principle of neurons, and its output is sparse to the speech signals enhancement. The research shows that when the DNN model training adopts the rectified

linear unit (ReLU), it can achieve better results without the unsupervised pre-training of the DNN model in the large-scale training data set [4]. Because the output target is the log amplitude spectrum of pure speech signals, the linear activation function is selected in the output layer of DNN model.

4.3. **Experimental results and analysis.** In order to illustrate the effectiveness of the method and reflect the better details of noise suppression and speech spectrum information retention, Figure 3 shows a spectrum of noisy speech signals with a noise type of Factory and an input SNR of 0 dB enhanced by different supervised methods. It can be seen from Figure 3 that the noise suppression level of the MCRA based enhancement method is higher than that of IMCRA method and DNN model, but the spectrum components of the enhanced speech signals in the low-frequency part are missing compared with the pure speech signals. DNN model can recover the speech components and speech spectrum structure polluted by noises, but there are many redundant noises. We think the main reason is that the loss function based on MMSE has equal weight to each frequency band. However, for speech signals, the energy of low frequency components is much higher than that of high frequency components, so the speech signals enhanced by the DNN method especially in high frequency parts will have noise redundancy. It can be seen from Figure 3 that the proposed method can recover the spectrum structure of speech better while suppressing noise components.



(a) Noisy speech signals

(b) Clean speech signals

(c) Enhanced by spectral subtraction

(d) Enhanced by improved spectral subtraction

(e) Enhanced by the DNN model

(f) Enhanced by the proposed method

FIGURE 3. (color online) Spectrum results of signals before and after English speech enhancement

Table 2 and Table 3 show the PESQ value and LSD value results of speech polluted by four kinds of noises under different SNRs. It can be seen from Table 2 and Table 3 that except for −5 dB HF channel, the PESQ value and LSD value of the proposed method are slightly lower than that of DNN model baseline method, and the two evaluation values proposed in this paper are higher than the other four methods in other noise and SNR cases. It can be seen from Table 3 that under F16 noise type, the performance of MCRA baseline method or IMCRA baseline method is similar to that of DNN baseline method at low SNR condition, while under the noise type of Gaussian white, the performance of such methods is lower than that of DNN baseline method and the proposed method.

TABLE 2. PESQ values under different English speech enhancement algorithms and noises

Noise	SNR/dB	PESQ value				
		Original	MCRA	IMCRA	DNN	Proposed
F16	−5	1.325	1.551	1.644	1.719	1.732
	0	1.609	1.940	2.065	2.107	2.135
	5	1.945	2.326	2.425	2.454	2.482
	10	2.290	2.597	2.656	2.688	2.733
Factory	−5	1.307	1.487	1.576	1.671	1.686
	0	1.603	1.843	1.943	2.054	2.113
	5	1.924	2.230	2.316	2.393	2.469
	10	2.296	2.492	2.547	2.639	2.729
HF Channel	−5	1.351	1.425	1.521	1.808	1.768
	0	1.499	1.684	1.878	2.157	2.172
	5	1.701	2.041	2.247	2.432	2.491
	10	1.958	2.416	2.574	5.672	2.733
White	−5	1.269	1.536	1.654	1.921	1.954
	0	1.457	1.918	2.053	2.230	2.301
	5	1.749	2.302	2.428	2.478	2.593
	10	2.100	2.577	2.671	2.742	2.808

TABLE 3. LSD values under different English speech enhancement algorithms and noises

Noise	SNR/dB	LSD value				
		Original	MCRA	IMCRA	DNN	Proposed
F16	−5	2.485	1.982	1.901	1.598	1.515
	0	2.312	1.723	1.652	1.395	1.336
	5	2.030	1.449	1.419	1.251	1.205
	10	1.685	1.211	1.232	1.088	1.102
Factory	−5	2.752	2.148	2.106	1.741	1.721
	0	2.575	1.905	1.865	1.553	1.512
	5	2.280	1.657	1.639	1.388	1.310
	10	1.913	1.416	1.443	1.174	1.160
HF Channel	−5	2.330	1.848	1.687	1.371	1.382
	0	2.165	1.602	1.495	1.238	1.224
	5	1.902	1.354	1.320	1.184	1.118
	10	1.586	1.128	1.171	1.057	1.056
White	−5	3.632	2.382	2.303	1.736	1.639
	0	3.446	2.032	2.010	1.614	1.527
	5	3.131	1.773	1.761	1.464	1.394
	10	2.726	1.563	1.573	1.238	1.223

We think that the main reason is that the F16 noise type is in double cabin (such as equipment engine) atmosphere, which has a strong spectrum structure. This kind of noise can be represented by a small number of atom combinations in the dictionary. Therefore, the enhancement method based on MCRA has better modeling ability for this kind of structured noise, while Gaussian white noise type does not have obvious spectrum

structure information. Therefore, for the enhancement method based on MCRA and IMCRA, the type of the speech distortion noise is higher than that of other types.

In the experiment, the noise type selected by the training set and the validation set is the same, but the testing set and the training set use different statements in the speech data set, and add different parts of the same type of noises according to different SNR, for the sake of avoiding over-fitting problem, and testing the generalization performance of the proposed method for the type of mismatching. The experimental results validate the effectiveness of the proposed method in the case of mismatching in the stationary and non-stationary noise environments. Since the noise dictionary is only obtained under four types of noises, no further experiments have been made on the performance of the complete mismatched data set (the completely mismatched generally refers to the type of noise that has not been seen in the training stage, the SNR that has not been trained and the speaker that has not been seen). In fact, the generalization performance on the completely mismatched data set is a problem for supervised speech signals enhancement methods such as dictionary based methods or DNN model based methods. On the one hand, this problem can be solved by increasing the number of samples in dictionary training or DNN model training, that is, increasing the diversity of samples in the training set. In [4], 104 kinds of noise types and 625 h corpus are selected for training, which proves that the autoregressive speech enhancement method based on DNN model has a good denoising effect on the completely mismatched data set. In [26], the method of adding vibration to data in training set is used to increase data diversity and improve performance. On the other hand, we can use semi-supervised or adaptive dictionary learning method, that is, online learning dictionary from samples and data to increase the adaptability of the method, which is also another next research direction of mine.

5. **Conclusion.** English speech signals enhancement is the most commonly used during English environment based communication. This paper proposed an English speech enhancement algorithm combining improved spectral subtraction and DNN model to meet the requirement of English speech signals enhancement. In the proposed algorithm, IMCRA is used to process the amplitude spectrum of speech in the enhancement stage, and the amplitude spectrum features of the adjacent frames are normalized and combined as the input to train the DNN model. After training, the well-trained DNN model outputs the logarithmic amplitude spectrum of the enhanced speech signals, and then the time-domain English speech enhancement signals are reconstructed by the ISTFT algorithm. In different noise environments, experimental results have shown that the proposed method has brought significant improvement in the two key values of PESQ and LSD for the enhancement of English speech signals, making English speech signals get a qualitative improvement in communication. Future work will focus on the construction of a deeper and problem-oriented DNN model for the amplitude spectrum features, and the use of convolutional and pooling operations to bring more significant enhancement to the amplitude spectrum features of speech signals.

Acknowledgment. This work was supported by the Education and Scientific Research Project for Young and Middle-aged Teachers in Fujian Province (No. JZ170067).

REFERENCES

[1] K. M. Nayem and D. S. Williamson, Incorporating intra-spectral dependencies with a recurrent output layer for improved speech enhancement, *2019 IEEE the 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp.1-6, 2019.

[2] S. Wisdom, J. R. Hershey, K. Wilson et al., Differentiable consistency constraints for improved deep speech enhancement, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, pp.900-904, 2019.
[3] N. Mohammadiha, P. Smaragdis and A. Leijon, Supervised and unsupervised speech enhancement using nonnegative matrix factorization, *IEEE Trans. Audio, Speech, and Language Processing*, vol.21, no.10, pp.2140-2151, 2013.
[4] Y. Xu, J. Du, L.-R. Dai and C.-H. Lee, A regression approach to speech enhancement based on deep neural networks, *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol.23, no.1, pp.7-19, 2014.
[5] S. Boll, Suppression of acoustic noise in speech using spectral subtraction, *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol.27, no.2, pp.113-120, 1979.
[6] M. Berouti, R. Schwartz and J. Makhoul, Enhancement of speech corrupted by acoustic noise, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1979)*, vol.4, pp.208-211, 1979.
[7] D. J. Thomson, Spectrum estimation and harmonic analysis, *Proc. of the IEEE*, vol.70, no.9, pp.1055-1096, 1982.
[8] B. Liu, Z. Wang, S. Guo et al., An energy-efficient voice activity detector using deep neural networks and approximate computing, *Microelectronics Journal*, vol.87, pp.12-21, 2019.
[9] S. Sarakon, T. Phoka and K. Tamee, Robust noise for human activity recognition using convolutional neural network, *ICIC Express Letters, Part B: Applications*, vol.11, no.3, pp.229-236, 2020.
[10] O. Abdel-Hamid, A. Mohamed, H. Jiang et al., Convolutional neural networks for speech recognition, *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol.22, no.10, pp.1533-1545, 2014.
[11] J. Chen and D. L. Wang, DNN based mask estimation for supervised speech separation, in *Audio Source Separation. Signals and Communication Technology*, S. Makino (ed.), Cham, Springer, 2018.
[12] A. Murad and J. Y. Pyun, Deep recurrent neural networks for human activity recognition, *Sensors*, vol.17, no.11, 2017.
[13] T. Kim, M. Cha, H. Kim, J. K. Lee and J. Kim, Learning to discover cross-domain relations with generative adversarial networks, *arXiv Print, arXiv:1703.05192*, pp.1857-1865, 2017.
[14] A. B. Nassif, I. Shahin, I. Attili et al., Speech recognition using deep neural networks: A systematic review, *IEEE Access*, vol.7, pp.19143-19165, 2019.
[15] I. Cohen, Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging, *IEEE Trans. Speech and Audio Processing*, vol.11, no.5, pp.466-475, 2003.
[16] D. Brezeale and D. J. Cook, Automatic video classification: A survey of the literature, *IEEE Trans. Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol.38, no.3, pp.416-430, 2008.
[17] Y. Maleki, Optimal scale invariant wigner spectrum estimation of Gaussian locally self-similar processes using hermite functions, *Journal of Theoretical Probability*, vol.32, no.1, pp.202-215, 2019.
[18] H. Müsch, Subjective rating and PESQ prediction of listener echo and duplex impairments, *Journal of the Audio Engineering Society*, vol.67, no.3, pp.124-134, 2019.
[19] D. Imparato, P. J. G. Teunissen and C. Tiberius, Minimal detectable and identifiable biases for quality control, *Survey Review*, vol.51, no.367, pp.289-299, 2019.
[20] N. Mohammadiha, *Speech Enhancement Using Nonnegative Matrix Factorization and Hidden Markov Models*, Ph.D. Thesis, KTH Royal Institute of Technology, Stockholm, 2013.
[21] E. H. Rothauser, IEEE recommended practice for speech quality measurements, *IEEE Trans. Audio and Electroacoustics*, vol.17, pp.225-246, 1969.
[22] I. T. U. T. Recommendation, Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, *Rec. ITU-T*, p.862, 2001.
[23] A. Erell and M. Weintraub, Estimation using log-spectral-distance criterion for noise-robust speech recognition, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.853-856, 1990.
[24] P. Smaragdis, Convolutive speech bases and their application to supervised speech separation, *IEEE Trans. Audio, Speech, and Language Processing*, vol.15, no.1, pp.1-12, 2006.
[25] A. L. Maas, A. Y. Hannun and A. Y. Ng, Rectifier nonlinearities improve neural network acoustic models, *Proc. of International Conference on Machine Learning*, vol.30, no.1, 2013.
[26] J. Chen, Y. Wang and D. L. Wang, Noise perturbation for supervised speech separation, *Speech Communication*, vol.78, pp.1-10, 2016.