The Application of Deep Neural Network in Speech Enhancement Processing

Chen Jian-ming
Department of Information and Communication
Army Academy of Armored Forces
Beijing, China
cchjm@163.com

Liang Zhi-cheng
Department of Information and Communication
Army Academy of Armored Forces
Beijing, China
liangzhicheng93@163.com

***Abstract*—To solve the problem that Non-stationary noise is difficult to remove during speech enhancement process when using Fourier transform, this essay will put forward a speech enhancement algorithm based on the combination of Ensemble Empirical Mode Decomposition (EEMD) and Deep Neural Network (DNN). Firstly, preprocessing the original signal by EEMD, and decomposing a series of time-frequency information of the IMF component to meet the time-variation requirement better; Secondly, adjusting the weight of the IMF component by DNN and then synthesize it to enhanced the speech; Finally, comparing the differences of speech enhancement performance between using EEMD alone, using Fourier transform and EEMD as a preprocessing. The results show that the enhanced algorithm using EEMD as a preprocessing improves the scores of PESQ and STOI by 0.745 and 0.169 respectively, effectively improving the speech quality and intelligibility.**

Keywords- Time-frequency analysis; Speech enhancement algorithm; Ensemble Empirical Mode Decomposition; Deep Neural Network

I. INTRODUCTION

With the rapid development of information technology and artificial intelligence, people have higher demand for speech quality. Speech signal processing has become a research focus, in which speech enhancement plays an important role. Since the 1970s, in order to improve the speech quality and intelligibility, researchers have proposed traditional speech enhancement algorithms according to the short-term stability and linear assumption of the speech, and in combination with three dimensions of time, space and the spectral characteristics. These algorithms include spectral subtraction, adaptive filter method, statistic-based model and subspace method, etc. [1]. These algorithms use short-time Fourier transform to extract the frequency spectrum information of the windowed signal. The effect of processing the stationary signal is better, and the effect of processing the non-stationary signal is general. In the 1980s, the wavelet theory [2] developed rapidly, and its multi-frequency resolution enabled it to have band-pass filtering characteristics, which provided a new idea for non-stationary signal denoising. But its essence was a window-tunable Fourier transform. It still is not free from the limitations of Fourier analysis. In 1998, Chinese-American scientist N.E. Huang and others proposed a new signal processing method—Empirical Mode Decomposition (EMD) [3]. In order to overcome the end-effects and modal aliasing problems that easily occur in EMD, Wu and Huang proposed an Ensemble Empirical Mode Decomposition (EEMD) method in 2009 [4, 5]. The method decomposes the different scale fluctuations actually existing in the signal step by step, and generates a series of data sequences with different characteristic scales. Each sequence is called an Intrinsic Mode Function (IMF) [3]. The frequency of non-stationary signals changes with time. Using EEMD can obtain instantaneous frequency of each IMF, and more truly reflect the actual physical meaning represented by the signal, thereby retaining the feature of non-stationary and non-linear of speech in speech enhancement. In recent years, Deep Neural Network (DNN) has made great achievements in image classification and provided new ideas for speech signal processing. As a mapping model, DNN takes the signal's characteristic parameters as input and the clean speech signal as output, learns the complex nonlinear relationship between input and output. DNN has a good inhibitory effect on non-stationary noise [6]. However, during the DNN processing, since the feature parameters extracted by the frame signal and the Fourier transform of speech signal are approximately stationary, the time variation of speech signal cannot be accurately reflected.

In this paper, EEMD and DNN are combined. After the noisy speech is decomposed by the EEMD, each IMF component is taken as input. The clean speech signal is used as the expected output. DNN is used to find the mapping relationship between the input and output. The experimental results show that the proposed method overcomes the problems that the optimal time-frequency resolution cannot be obtained and time-frequency analysis error is increased due to frame processing and Fourier transform. It has obvious effect on the improvement of speech quality and intelligibility.

II. SPEECH ENHANCEMENT BASED ON EEMD AND DNN

When noisy speech is decomposed into different time scales by EEMD, the vibration modal of the whole signal is presented on different IMFs in descending order of frequency. Although the energy distributions of different IMFs decomposed by different noisy speeches are different, the main information of the speech signals are mostly distributed on a large time scale and are concentrated on a limited number of IMFs. The signal may be reconstructed to remove the noise by enhancing the weights of the speech signal components, eliminating or reducing the weight of the noise component. For how to choose proper weights, this

978-1-5386-5500-9/18/$31.00 ©2018 IEEE
DOI 10.1109/ICISCE.2018.00257

paper uses DNN to complete. Theoretically, it has been proved that DNN can implement arbitrary nonlinear mapping by adjusting the number of hidden layers and the number of nodes at each layer [7]. Therefore, the complex mapping relationship between the input IMF components and the output clean speech can be obtained through training. And then the network parameters are adjusted to complete the task of speech enhancement.

A. Processing by EEMD

EEMD is an improved algorithm that introduce white noise to EMD [8]. A section of noisy speech is selected and decomposed EEMD to obtain n IMF components. Calculate the total energy E_j of each IMF component, and use energy as the element to form a feature vector P and normalize it as network input:

$$P' = P / E = \left[P_1', ...P_i', ...P_n' \right] = \left[E_1 / E, ...E_i / E, ...E_n / E \right]$$

$(i = 1, 2, ...n)$, $E = (\sum_{j=1}^{n} \left| E_j \right|^2)^{1/2}$.

B. Network model construction

1) Network structure: In addition to the input and output layers, three hidden layers are set. The number of hidden layer nodes l is 128, the number of input layer nodes n is the number of IMFs, the number of output layer nodes t is 1, the structure is shown in "Fig.1".

2) Network initialization: Initialize connection weights and activate functions. The proper initial weight can shorten the network training time and make the parameters reach the global optimal solution. The initial values are generally given at random, which can easily lead to unstable results. Therefore, the non-zero random value between the empirical values $(-2.4 / F, 2.4 / F)$ are used as the initial weight. F are the number of connected neurons at the weight input [9].

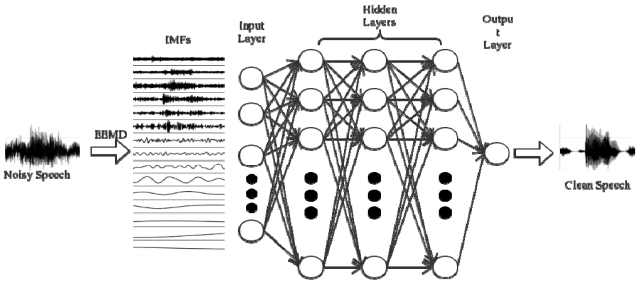

Figure 1. Network structure

C. Network training and testing

Based on actual needs, the noise of "factory" in the NOISEX data set for four minutes was selected to simulate the speech enhancement environment in the factory. The noise of the first two minutes is randomly cut and used as a training set for the network after mixing with 500 segments of clean speech in accordance with -2db. Randomly cut the noise after two minutes and mix it with 100 segments of clean speech according to -2db, as the test set of the network. The training steps are as follows:

1) Calculate the hidden layer output:

The first hidden layer output:

$$H_j = f(\sum_{i=1}^{n} w_{ij}^{12} P_i') \quad j = 1, 2, \ldots l \tag{1}$$

The second hidden layer output:

$$G_k = f(\sum_{j=1}^{n} w_{jk}^{23} H_j) \quad k = 1, 2, \ldots l \tag{2}$$

The second hidden layer output:

$$J_m = f(\sum_{k=1}^{l} w_{km}^{34} G_k) \quad m = 1, 2, \ldots l \tag{3}$$

In the formula, w_{ij}^{12} , w_{jk}^{23} , w_{km}^{34} are the weight of each layer, l is the number of hidden layers, and the hidden layer excitation function f uses ReLU.

2) Calculate the output layer output:

$$O = f(\sum_{m=1}^{l} w_m^{45} J_m) \quad m = 1, 2, \ldots l \tag{4}$$

In the formula, f is the output layer activation function.

3) Calculate errors:

The actual output O of the network and the expected output T are brought into the loss function to calculate the network error e. Since speech enhancement is a regression problem, Mean-Square Error (MSE) is chosen as the loss function [9]. In order to avoid overfitting of the model, $L2$ regularization is added to the loss function to limit the complexity of the model.

$$e = MSE(T, O) = \frac{\sum_{i=1}^{N} (T_i - O_i)^2}{N} \tag{5}$$

$$R(W) = \left\| W \right\|_2^2 = \sum \left| w^2 \right| \tag{6}$$

$$E(W) = e + \frac{\lambda}{2} R(W) \quad \lambda = 0.5 \tag{7}$$

In the formula, N is the sample number, T is the clean speech amplitude value, $R(w)$ is the function embodying the model complexity, W is the weight parameter of the network, $E(W)$ is the network optimization objective function, and λ is the model complexity loss in the total loss proportion.

4) Update weight:

Start from the weights between the third hidden layer and the output layer, adjusting along the direction of the fastest gradient drop. Then the error propagates forward and the weights of the previous layers are adjusted.

The third hidden layer weight adjustment:

$$w_m^{45} = (1 - \eta\lambda) w_m^{45} + \eta \frac{\partial e}{\partial f} J_m \quad m = 1, 2, \ldots l \tag{8}$$

The second hidden layer weight adjustment:

$$w_{km}^{34} = (1 - \eta\lambda) w_{km}^{34} + \eta \frac{\partial e}{\partial f} G_k \quad k = 1, 2, \ldots l \tag{9}$$

The first hidden layer weight adjustment:

$$w_{jk}^{23} = (1-\eta\lambda)w_{jk}^{23} + \eta\frac{\partial e}{\partial f}H_j \quad j = 1,2,\ldots l \qquad (10)$$

Input layer weight adjustment:

$$w_{ij}^{12} = (1-\eta\lambda)w_{ij}^{12} + \eta\frac{\partial e}{\partial f}P_i^{'} \qquad i = 1,2,\ldots n \qquad (11)$$

In the formula, η is the Learning Rate, η is set to 0.01 to define the amplitude of each parameter update.

The above is a round of weight adjustment. The rules are summarized as: Weight adjustment = learning rate * local gradient * upper output signal

In the remaining samples, a new set of data is sent to the network for training until all the samples have been trained. Parameters are saved and updated after the training, and the DNN has the ability to predict. By inputting the test set data into the trained DNN, background noise can be removed to achieve speech enhancement.

III. EXPERIMENTAL RESULTS AND ANALYSIS

Randomly select the sample "There is a strong chance that will happen once more." in the test for experimental verification. The time-frequency analysis of its clean speech, noise, and noisy speech are shown in “Fig. 2” “Fig. 3” and “Fig. 4”. It can be known from the spectrum diagram that the frequency of the speech signal is mainly distributed between 100 Hz and 3 kHz, mainly in the low-middle frequency range; the noise frequency is mostly concentrated within 1 kHz and dominated by low frequencies. The following are the enhancement effects of using EEMD alone, not using EEMD as a pretreatment, and using EEMD as a pretreatment. The experimental results are as follows:

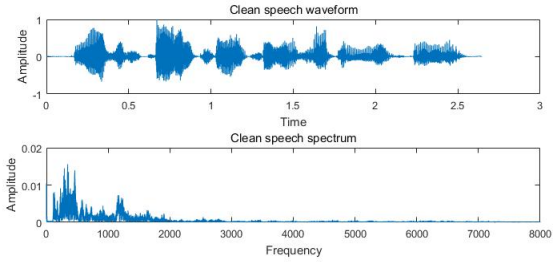

Figure 2. Clean speech waveform and spectrogram

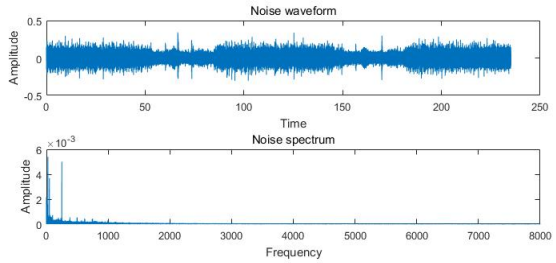

Figure 3. Noise waveform and spectrogram

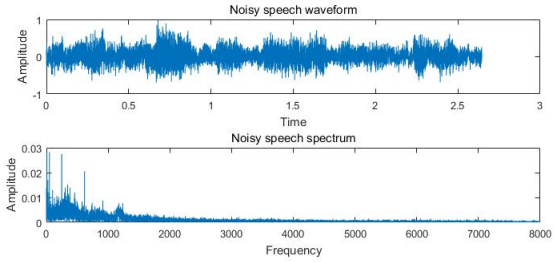

Figure 4. Noisy speech waveform and spectrogram

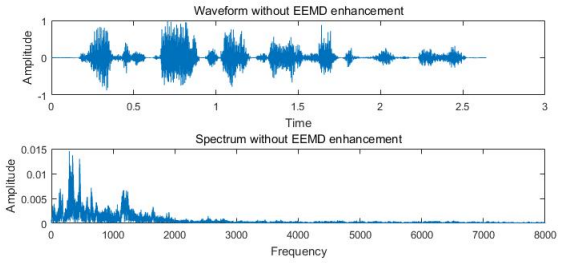

Figure 5. Waveform without EEMD enhancement

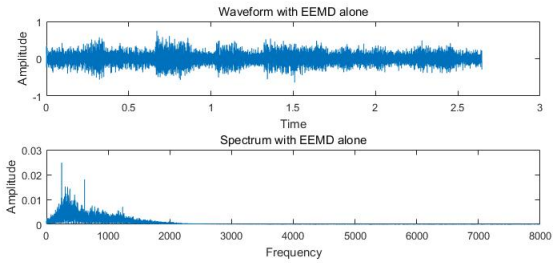

Figure 6. Waveform with EEMD alone

A. *Use of EEMD alone for speech enhancement*

The time-frequency analysis of EEMD alone is shown in “Fig.6”. From the waveform diagram, it can be seen that EEMD has no significant effect on the suppression of noise. Analysis of the spectrum shows that the noise with high amplitude within 200Hz is completely removed, and the noise with high amplitude near 250Hz is not removed. Because there are many voice signal components near 250Hz, and EEMD mainly removes noise components according to different frequencies. When the frequency of the two overlaps, the enhancement effect is not strong. At the same time, while removing the noise, the speech signal with the same frequency range as the noise frequency is also removed, causing speech distortion and seriously affecting the auditory effect.

B. *No EEMD was used for preprocessing*

The time-frequency analysis after using DNN alone is shown in “Fig.5”. Compared with the time-frequency analysis chart of clean speech: Observed from the waveform

diagram, the method of using DNN alone removes most of the noise, and the enhancement effect is obvious. However, from the frequency spectrum observation, low-frequency noise within 100Hz and noise components with high amplitude around 300Hz are not removed, and the speech components around 400Hz are weakened, resulting in speech distortion. This causes the speech signal to be doped with noise during the auditory process, reducing speech quality and intelligibility. This is because the speech is approximately stationary between 20 ms and 30 ms, and spectral leakage and aliasing occur when the frame is windowed, causing errors.

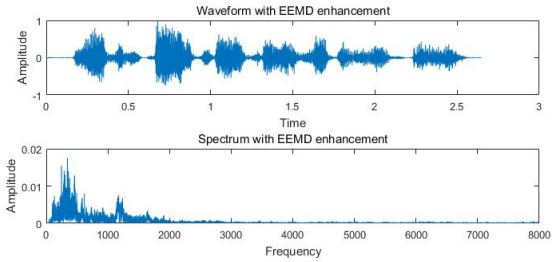

Figure 7. Waveform with EEMD enhancement

C. *Use EEMD for preprocessing*

Using the method proposed in this paper, first use EEMD to preprocess the noisy speech, and then the time-frequency analysis after DNN processing is shown in "Fig. 6". Observed from the waveform and spectrogram, the enhancement effect using EEMD pretreatment is more obvious. Compared with signals not processed by EEMD, it not only removes noise components within 100 Hz, but also does not significantly weaken the amplitude of speech signals. And it retains high-frequency speech components with lower amplitudes. These high-frequency components can improve speech quality.

In order to further evaluate the performance of the algorithm, this paper uses the most accurate speech quality assessment algorithm – Perceptual Evaluation of Speech Quality (PESQ) [10] and Short-Time Objective Intelligibility (STOI) [11]. And evaluate the two methods. PESQ is an objective expression of subjective evaluation. STOI is an objective evaluation method of intelligibility. The higher the score, the better the speech quality and intelligibility.

In this paper, 20 segments of speech are randomly selected from the test set to be enhanced by three methods respectively, and then the average scores are calculated after each segment is scored. The results are shown in "Tab. 1". From the data in the table, the PESQ scores of the three algorithms are 0.1, 0.628 and 0.745 higher than those of the original noisy speech, and the STOI scores are increased by 0.039, 0.131 and 0.169 respectively. After the EEMD pretreatment, the PESQ and STOI are increased by 0.117 and 0.038, respectively. Enhance the effect better.

TABLE I. SPEECH ENHANCEMENT PERFORMANCE ASSESSMENT

Evaluation Indicators	**Score**			
	Noisy speech	***Only EEMD***	***No EEMD***	***Use EMMD***
PESQ	1.381	1.481	2.009	2.126
STOI	0.652	0.691	0.783	0.821

IV. CONCLUSION

This essay demonstrates a speech enhancement algorithm based on the combination of EEMD and DNN. Firstly, using EEMD to preprocess the speech to ensure the correlation and completeness of the speech signal in the time and frequency domain. Secondly, inputting the feature vector into the trained DNN to enhance. Finally, evaluating the speech enhancement performance using PESQ and STOI. Through experimental verification, comparing with the original noisy speech, the speech enhancement effect of using EEMD as a preprocessing improves the scores of PESQ and STOI by 0.745 and 0.169 respectively. In summary, the neural network shows stronger and stronger performance in speech enhancement and if the input features are more fully extracted and the network generalization ability is improved, the neural network will be more and more applied to speech enhancement.

REFERENCES

[1] Loizou, Philipos C. Speech Enhancement: Theory and Practice. CRC Press, Inc. 2007.

[2] Daubechies, Ingrid, and C. Heil. Ten Lectures on Wavelets. Society for Industrial and Applied Mathematics, 1992.

[3] Huang, Norden E., et al. "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis." Proceedings of the Royal Society of London A 454.1971(1998):903-995.

[4] Huang, Norden E., and Z. Wu. "A review on Hilbert Huang transform: Method and its applications to geophysical studies." Reviews of Geophysics 46.2(2008).

[5] ZHAOHUA WU, and NORDEN E. HUANG. "ENSEMBLE EMPIRICAL MODE DECOMPOSITION: A NOISE-ASSISTED DATA ANALYSIS METHOD." Advances in Adaptive Data Analysis 1.01(2011).

[6] Xu Yong. Research on speech enhancement method based on deep neural network.. Diss. China university of science and technology., 2015.

[7] Lecun, Y, Y. Bengio, and G. Hinton. "Deep learning. " Nature 521.7553(2015):436.

[8] Zhang Meijun, Tang Jian, He Xiaohui. EEMD method and its application in mechanical fault diagnosis. National Defence Industry Press, 2015.

[9] Chen Ming. MATLAB neural network principle and the example refined solution. Tsinghua University Press, 2013.

[10] Rix, A. W., et al. "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs." IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings IEEE, 2001:749-752 vol.2.

[11] Taal, Cees H., et al. "A short-time objective intelligibility measure for time-frequency weighted noisy speech." IEEE International Conference on Acoustics Speech and Signal Processing IEEE, 2010:4214-4217.