Practical Advances in Artificial Intelligence and Computing
Frontier Research Publishing • https://www.frpubs.org
Vol. 1, No. 2, 2025

A Lightweight Collaborative Attention Residual Network with Depthwise Convolutions for Visual Feature Representation

Fenna Marcelline
University of Windsor, Windsor, Canada
fennam432@gmail.com

Abstract:

This paper proposes a lightweight convolutional neural network model that integrates depthwise separable convolution and a collaborative attention mechanism to enhance classification accuracy and robustness in complex visual environments. The proposed CAM-DResNet model improves computational efficiency by replacing standard convolutions with depthwise convolutions and incorporates a Collaborative Channel-Spatial-Pixel Attention (CCSPA) module after each network stage to enhance global and local feature perception. This structure enables the network to effectively capture fine-grained texture features while maintaining strong generalization capability under diverse lighting, scale, and noise conditions. Extensive experiments on benchmark image datasets demonstrate that CAM-DResNet achieves superior performance compared to classical models such as ResNet50, DenseNet, and MobileNet, with improved accuracy and reduced model parameters. Specifically, the proposed model achieves over 91% classification accuracy while reducing computational complexity by one-third relative to ResNet50. The results indicate that the integration of multi-level attention and lightweight residual design provides a robust and efficient solution for high-precision image classification tasks in modern computer vision applications.

Keywords:

Image classification; Convolutional neural network (CNN); Depthwise separable convolution; Attention mechanism; Lightweight model

1. Introduction

Voice is the medium through which people communicate information during interaction. However, in daily living environments, signals are subject to various types of noise interference during transmission, which can have an impact on people's lives[1]. So speech enhancement has become a key research area in the field of speech signal processing, with the core task of improving the quality of noisy speech, which directly affects the availability of speech samples. In terms of speech denoising, the main goal is to obtain high-quality and pure speech signals to improve speech recognition capabilities. In recent years, speech signal processing has been widely applied in real life, such as speech recognition, speaker recognition, and gender recognition technologies, to assist patients with hearing impairments. Since the 21st century, wavelet transform has developed rapidly. It evolved from Fourier transform. Compared to Fourier transform, wavelet analysis can extract more useful information locally in the signal [3]. Speech signals containing noise can be decomposed into components of different scales, and wavelet coefficients can be used to represent information of various resolutions, thereby achieving the characteristics of multispectral resolution.

In 1998, Chinese American Norden E. Huang proposed Empirical Mode Decomposition (EMD), which essentially smooths signals by decomposing fluctuations at different scales step by step, generating a series of data sequences with different feature scales, known as Intrinsic Mode Functions (IMFs). Assuming that any complex signal is composed of different, simple, non sinusoidal intrinsic mode function (IMF) components. Each IMF can be linear or nonlinear [4]. Compared with wavelet transform, EMD is driven by the signal itself, avoiding the selection of basis functions. Based on EMD decomposition, threshold can be set for denoising, similar to the idea of wavelet soft threshold and hard threshold denoising. Select appropriate thresholds for each order of IMF components obtained, and use the obtained thresholds to process the corresponding IMF components [5]. The time-frequency properties, multi-resolution analysis, sparsity, reversibility, and effective algorithms of wavelet transform make it one of the widely researched and applied technologies in the field of speech signal processing. However, wavelet transform uses fixed wavelet bases and is not sensitive to sound in some cases, while EMD is an adaptive signal decomposition method that can automatically adjust the number and frequency band range of IMFs based on the local characteristics of the signal, adapting more flexibly to the characteristics of the signal. Therefore, based on the analysis of the theory of wavelet transform, this article proposes a speech enhancement method combining wavelet transform and EMD by utilizing the multi-resolution characteristics of wavelet and the adaptive characteristics of EMD. Using actual noisy speech samples, conduct simulation analysis on the MATLAB platform to compare the differences between traditional speech enhancement algorithms and improved methods. The experimental results show that the algorithm proposed in this paper significantly improves the output average signal-to-noise ratio of speech signals, and performs better than traditional methods in suppressing noise, improving speech signal clarity, and adaptability. Overall, the method proposed in this article demonstrates a relatively ideal performance in speech enhancement.

2. Application Analysis of Wavelet Transform in Speech Enhancement

2.1 Basic Principles of Wavelet Transform

Traditional signal analysis is based on Fourier transform, but Fourier analysis belongs to global transform, and its transformation process can only be performed globally in the time or frequency domain. Therefore, Fourier transform cannot capture the time-frequency local characteristics of speech signals, especially limited to the analysis of non-stationary signals. In order to better analyze non-stationary signals, various new signal analysis theories have been proposed based on Fourier analysis, including Gabor transform, short-time Fourier transform, wavelet transform, and time-frequency analysis.

In these new analysis theories, wavelet analysis is based on Fourier analysis and achieves local transformations of time and frequency. Wavelet can automatically adjust the size of time and frequency windows after multiple analyses to meet the requirements of actual signal analysis. Therefore, wavelet transform exhibits strong flexibility in signal analysis, overcoming the disadvantage of fixed resolution in Fourier analysis.

The basic principle of wavelet transform 11 is to perform translation and scaling operations on signals in the time domain to achieve local time-frequency domain analysis of signals. This enables wavelet transform to better adapt to non-stationary signals, providing a powerful and flexible tool for signal analysis.

In the regulation letter $^{L2}(R)$, a function or signal that satisfies the following conditions is $a(x)$ wavelet:

$$C_{\varphi} = \int_{R^*} \frac{|\psi(\omega)|^2}{|\omega|} d\omega < \infty$$

If there are real number pairs (a, b) and the parameters a are not zero, the function that satisfies formula (1) is called a $\varphi(\omega)$ continuous wavelet function generated by the wavelet mother function, abbreviated as wavelet, and the parameters of the real number pairs it depends on b a are called scaling factor and translation factor.

$$\varphi(a,b)(x)=\frac{1}{\sqrt{|a|}}\varphi(\frac{x-b}{a})$$

The $f(x)$ wavelet transform of continuous signals is

$$W_f(a,b)=\frac{1}{\sqrt{|a|}}\int_R f(x)\varphi(\frac{x-b}{a})dx=<f(x),\varphi_{a,b}(x)>$$

The inverse transformation is defined as

$$f(x)=\frac{1}{C_\varphi}\iint_{R\times R^*} W_f(a,b)\varphi(\frac{x-b}{a})dadb$$

The $f(x)$ wavelet transform of discrete signals is

$$W_f(2^j,2^jk)=2^{\frac{-j}{2}}\int_{-\infty}^{+\infty} f(x)\varphi(2^{-j}x-k)dx$$

The reconstructed signal obtained by its inverse transformation is

$$f(t)=C\sum_{j=-\infty}^{+\infty}\sum_{k=-\infty}^{+\infty} W_f(2^j,2^jk)\varphi_{<2^j,2^jk>}(x)$$

C represents a constant.

2.2 Analysis of Wavelet Denoising Principle

Wavelet analysis is a rapidly developing time-frequency analysis method in recent years [6]. Researchers can more effectively solve speech signal processing problems from the perspective of human auditory perception. In the low-frequency signal region, this method provides lower temporal resolution and higher frequency resolution; In the high-frequency signal region, the frequency resolution is low, but the time resolution is high. The multi-resolution property of wavelet analysis makes it more in line with the auditory characteristics of the human ear, making it suitable for more accurate analysis of non-stationary signals such as speech signals.

The basic principle of wavelet denoising is to perform multi-scale wavelet decomposition on noisy speech signals, extract as many useful signal wavelet coefficients as possible at each scale, remove the noisy wavelet coefficients, and finally reconstruct the wavelet coefficients at each scale to obtain the denoised useful signal [7]. The principle diagram of wavelet denoising is shown in Figure 1 [8].

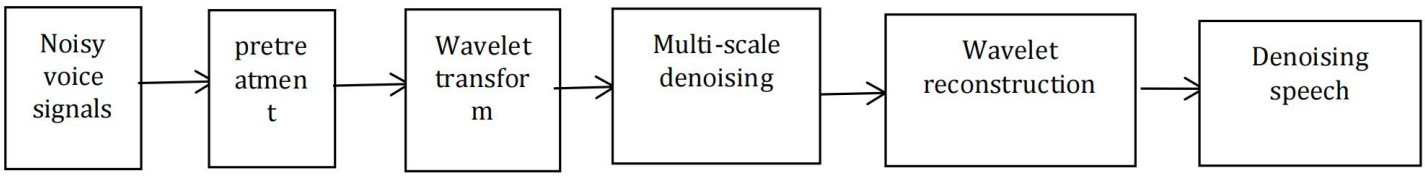

Figure1. Schematic diagram of wavelet denoising.

2.3 The method of wavelet denoising

There are various methods for wavelet denoising, including filtering denoising using wavelet decomposition and reconstruction, denoising using wavelet transform modulus maxima, signal-to-noise separation using spatial correlation after signal wavelet transform, nonlinear wavelet threshold denoising, translation invariant wavelet denoising, and multi wavelet denoising. Among them, threshold denoising is the most widely used method. Common threshold setting methods include hard threshold and soft threshold, which suppress noise while preserving important information of the signal as much as possible.

The idea of hard thresholding is to zero wavelet coefficients that are less than the set threshold, while retaining wavelet coefficients that are greater than the threshold, in order to suppress noise.

$$\omega_{new} = \begin{cases} \omega, |\omega| \geq T \\ 0, |\omega| \leq T \end{cases}$$

Soft thresholding is similar to hard thresholding, and it is also a thresholding process applied to the wavelet coefficients after wavelet transform. The purpose is to adjust the wavelet coefficients, retain the parts greater than the set threshold, and scale the parts smaller than the threshold to achieve signal denoising.

$$\omega_{new} = \begin{cases} 1 - \mathrm{sgn}(\omega)(|\omega| - T), |\omega| \geq T \\ 0, |\omega| < T \end{cases}$$

Note: ω represents the wavelet coefficient value, ω_{new} new represents the wavelet coefficient after applying the threshold, and T represents the final calculated threshold.

Let a noisy signal be:

$$f(t) = x(t) + y(t), t \subset (1, N)$$

Among them $f(t)$, $x(t)$ represents the actual signal detected, represents the effective signal, and

$y(t)$ represents the random noise. The steps for threshold denoising using sym5 wavelet are as follows:

Perform discrete wavelet transform (DWT) on $f(t)$;

Threshold processing of signals in the wavelet domain using hard and soft threshold functions

$$T = \delta\sqrt{2\log(N)}$$

Among them, δ is the noise standard deviation, and N is the number of signal sampling points

3. Analysis based on Empirical Mode Decomposition (EMD) theory

3.1 Basic principles of Empirical Mode Decomposition (EMD)

At the end of the 20th century, Huang E proposed the Hilbert Huang Transform (HHT) method, which includes two main components: Empirical Mode Decomposition (EMD) and Hilbert Transform (HT). EMD can decompose the original signal into a series of Intrinsic Mode Functions (IMFs). These IMF components are oscillatory functions with time-varying frequencies, which can effectively reflect the local characteristics of non-stationary signals. The instantaneous frequency obtained by performing Hilbert (HT) transform on the IMF component has practical physical significance. It can characterize the local characteristics of the signal, and the resulting Hilbert spectrum can accurately reflect the distribution of signal energy at various frequencies and times. The basic principle of EMD is shown in Figure 2:

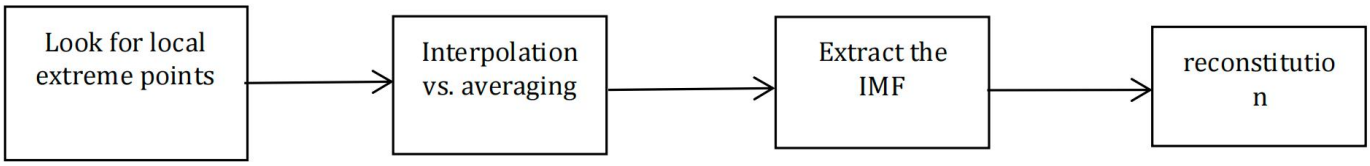

Figure 2. Basic schematic of EMD

The EMD algorithm adopts the idea of fitting the envelope with extreme points, which means that extreme values at the endpoints may contaminate the entire sequence, known as endpoint effects. Therefore, when applying EMD to speech enhancement, it is necessary to consider and handle the impact of endpoint effects. References [9-10] proposed several effective methods for suppressing EMD endpoint effects. The EMD method can theoretically be applied to the decomposition of any type of signal, and therefore has significant advantages in processing non- stationary and nonlinear data. It is suitable for analyzing nonlinear and non-stationary signal sequences and has a high signal-to-noise ratio.

3.2 Intrinsic Mode Function (IMF)

IMF (Intrinsic Mode Function) is a part of Empirical Mode Decomposition (EMD) method. EMD is a method of decomposing non-stationary and nonlinear signals into a set of intrinsic, local, and adaptive components [11].During the EMD decomposition process . The IMF is obtained through an iterative process called "screening", which gradually eliminates the asymmetry between the upper and lower envelopes of the original signal [12]. For each screening process, the local average of the signal needs to be calculated based on the upper and lower envelopes. The upper (lower) envelope is obtained by interpolating the local maximum (minimum) values of the signal through cubic spline interpolation [13].

IMF is one of the key outputs of EMD. Every IMF should meet the following two conditions:

In a given data segment, the number of extreme points is equal to the number of zero crossings, or the maximum difference is 1.

The average frequency of extreme points and zero crossing points should vary throughout the entire data segment.

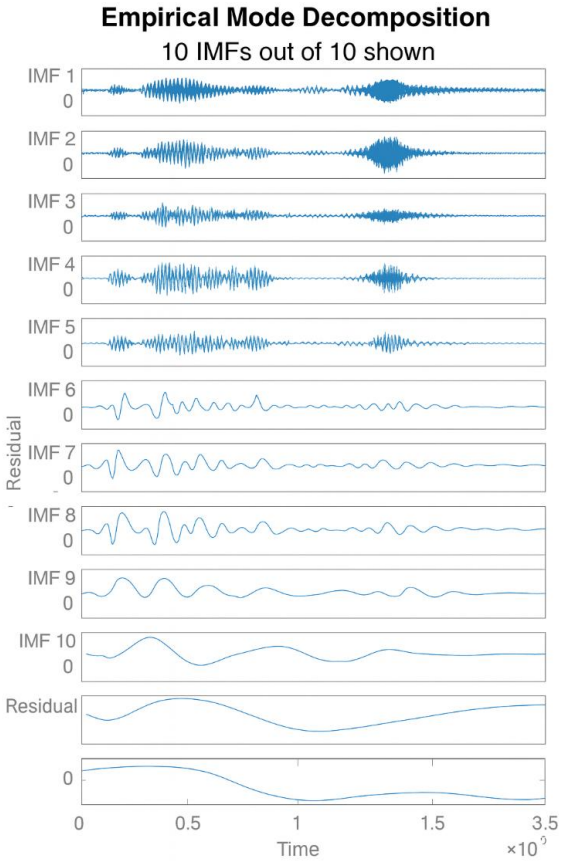

Figure 3. IMF components after EMD decomposition

Specifically, for a signal, the local extremum envelope formed by the extremum and zero crossing points is continuously extracted through iteration until the above two conditions are met. The first IMF obtained through this process is called the highest frequency IMF, the second is the second highest frequency IMF, and so on. Ultimately, the sum of these IMFs should be approximately equal to the original signal.

The characteristic of IMF is that they represent local vibrations or fluctuations of signals, with frequencies varying over time. This enables IMF to effectively capture local features in signals and has good adaptability to the decomposition of nonlinear and non-stationary signals. The IMF component can be represented by the following formula .

$$h_n(t) = h_{n-1}(t) - m_{n-1}(t)$$

$h_{n-1}(t)$ And $m_{n-1}(t)$ represent the initial signal and its mean signal at the $(n-1)$ th iteration. The iterative process of EMD involves continuously estimating the low-frequency components of the signal to obtain an estimate of the high-frequency components of the signal, known as IMF.

4. Wavelet Transform and EMD Joint Speech Enhancement

Combining time-domain and frequency-domain information, wavelet transform and EMD are used to enhance speech effects. This joint processing can improve the algorithm's ability to handle different

frequency components. Therefore, based on the multi-resolution characteristics of wavelet transform and the adaptability of EMD, a speech enhancement method combining wavelet transform and EMD is proposed.

Multi resolution refers to the analysis of signals at different scales in order to better understand the local features of the signal. Wavelet transform achieves multi-resolution analysis by decomposing signals into different frequency sub bands. Specifically, wavelet transform decomposes signals into low-frequency and high-frequency sub bands. The low-frequency sub bands contain the general trend of the signal, while the high-frequency sub bands contain the detailed information of the signal. By analyzing the low-frequency and high-frequency sub bands at different scales, we can better understand the local characteristics of the signal.

The following are the steps for enhancing speech signals:

(1)Preprocessing. Set parameters after reading the voice signal.

(2)Add Gaussian white noise.

(3)Wavelet transform. Perform wavelet transform on the signal and select the sym5 wavelet with good orthogonality and compactness to decompose the noisy signal into 5 layers.

(4)EMD decomposition. After performing wavelet transform on the speech signal, it is decomposed into EMD, and then each IMF is subjected to wavelet transform and thresholding again.

(5)Signal reconstruction. Finally, perform wavelet reconstruction and sum up each reconstructed IMF component. Obtain new voice signals.

This article chooses dynamic threshold to process speech signals, which has the advantage of being able to control speech signals by adjusting the threshold size. Obtain ideal results through this method.

Dynamic thresholding method:

$$W_{new}=\begin{cases}\omega, |\omega|>\varepsilon T\\ \operatorname{sgn}(\omega)(|\omega-T|)\times\dfrac{\varepsilon}{\varepsilon-T}, T<|\omega|<\varepsilon T\\ 0, |\omega|\le T\end{cases}$$

Defined ε as a constant with a value range of 0-1, its accurate value can be determined

experimentally.

The evaluation criteria for the quality of speech signals are the output average signal-to-noise ratio and mean square error value to assess their enhancement effect.

In order to better compare the denoising effect of the algorithm in this article, the average signal-to-noise ratio and mean square error value of the output under different input signal-to- noise ratios are calculated separately. The specific formula is as follows:

The calculation formula for signal-to-noise ratio:

$$SNR_{seq}=10\log\frac{\sum_t f^2(t)}{\sum_t (x(t)-f(t))^2}$$

The calculation formula for mean square error:

$$MSE = \frac{1}{n}\sum_{t}^{n}(x(t)-f(t))^2$$

$f(t)$, $x(t)$ Represents the original signal and the denoised speech signal. SNRseq The larger the signal, the better the denoising effect of the signal, and the noise has been effectively suppressed; SNRseq The smaller it is, the less ideal the denoising effect of the signal is. The smaller the output MSE value, the better the speech enhancement effect, and vice versa.

5. Simulation results and analysis

The dataset used in this experiment is from the GRID Audio Visual Sentence Corpus, with 1000 people each speaking 34 sentences in audio. The audio from the 25kHz endpoint audio is selected as the original signal, and Gaussian white noise is added to the original signal. By changing the noise intensity, noisy speech with signal-to-noise ratios of -5dB, 0dB, 5dB, 10dB, and 15dB is formed, and then wavelet transform is applied to the noisy speech signal. Among them, sym5 wavelet is selected with 5 decomposition layers; Afterwards, EMD decomposition is performed, followed by wavelet transform and thresholding for each IMF. Finally, reconstruct the signal and sum up the reconstructed IMF components.

Comparing different input signal-to-noise ratio conditions, wavelet transform, EMD, LMS, DCT, and the method proposed in this paper are used to enhance noisy speech signals. The experimental simulation results based on MATLAB are shown in Table 1 and Table 2; As shown in Figureures 4, 5, 6, and 7, Table 1 and Table 2 compare the output signal-to-noise ratio (SNR) and mean square error (MSE) of wavelet transform, EMD, LMS, DCT, and the proposed method in this paper. Figure4, 5, 6, and 7 represent the original signal, noisy speech signal, and processed speech signal waveforms under different input SNR conditions.

Table 1. Comparison of Average Signal to Noise Ratio

Inputsignal- to-noise ratio	-5dB	0dB	5dB	10dB	15dB
Wavelet transform method	1.14	3.04	6.3	10.37	15.14
EMD	-5.01	0.07	4.94	9.96	14.94
LMS	4.25	4.22	4.16	4.08	4.05
DCT	-31.03	-21.04	-21.03	-15.99	-10.99
Thisarticle's method	26.39	27.68	29.14	30.21	30.15

Table 2. Comparison of MSE Output Values

Inputsignal- to-noise ratio	-5dB	0dB	5dB	10dB	15dB
Wavelet transform method	0.0013	0.0004	0.0004	0.0003	0.0003

EMD	0.0079	0.0025	0.0008	0.0003	0.0001
LMS	0.0016	0.0017	0.0017	0.0015	0.0017
DCT	3.17	1	0.32	0.1	0.03
Thisarticle's method	0.0001	0.0001	0	0	0

From Tables 1 and 2, it can be seen that the combination of wavelet and EMD results in a much higher average signal-to-noise ratio compared to the other four methods. The average mean square error value is 0.00004, which is also the smallest among the five methods. Therefore, it can be seen that the new algorithm shows significant improvement in reducing speech signal noise and enhancing speech.

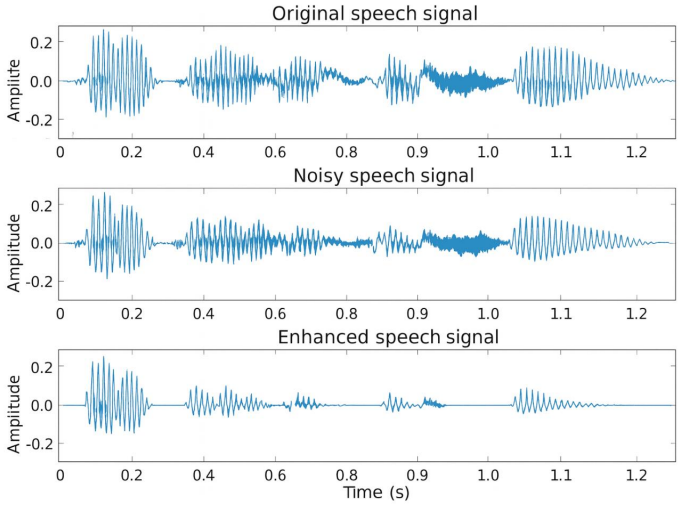

Figure 4. Wavelet Transform Simulation Results

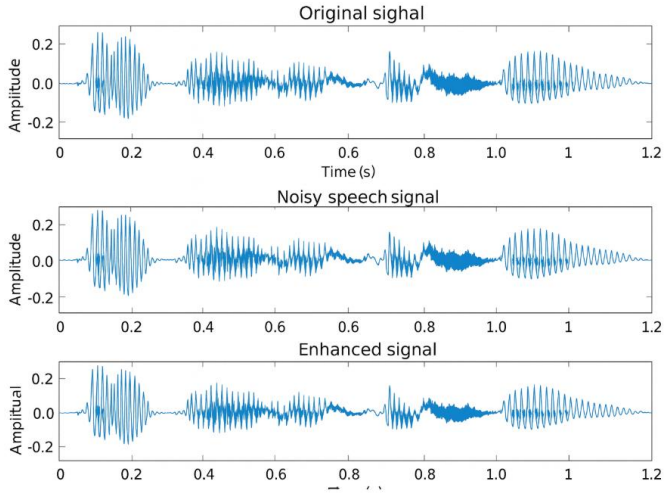

Figure 5. EMD simulation results

Add Gaussian noise to the original speech, set the signal-to-noise ratio of the noisy signal to - 5dB, 0dB, 5dB, 10dB, and 15dB, perform wavelet decomposition on the signal, select a dynamic threshold, and then apply a new algorithm for signal reconstruction to obtain the original speech signal, noisy signal, and reconstructed speech signal under the new algorithm. From the Figureure, it can be seen that the higher the signal-to-noise ratio of the input, the better the speech enhancement effect. Comparing the two, it is evident that the combination of wavelet transform and EMD not only preserves most of the original speech signal, but also makes flat areas smooth, maximizing the loss of original signal features. This indicates that the combination of wavelet transform and EMD has the best speech enhancement effect among the five methods. The relationship between SNR_seq and experimental values also demonstrate the superiority of the combination of wavelet transform and EMD in speech enhancement.

Figure 6, 7 respectively show the time-frequency plots of the four methods mentioned in the text. Here, we select the time-frequency plots after denoising with an input signal-to- noise ratio of 15dB.

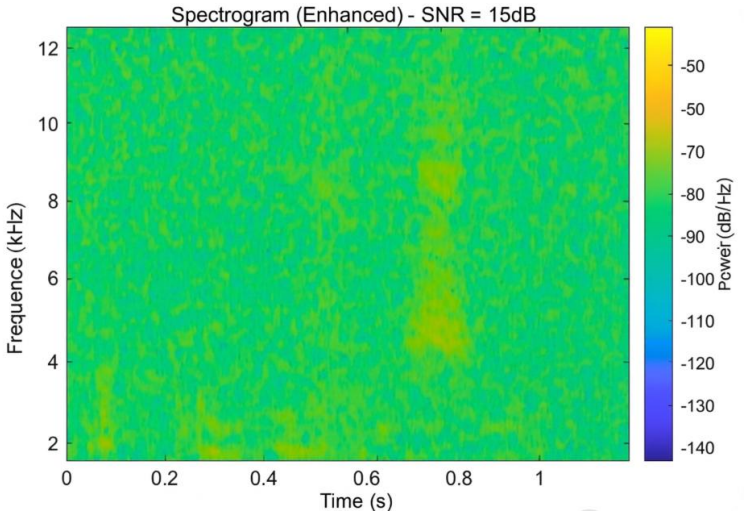

Figure 6. Time frequency diagram of wavelet transform

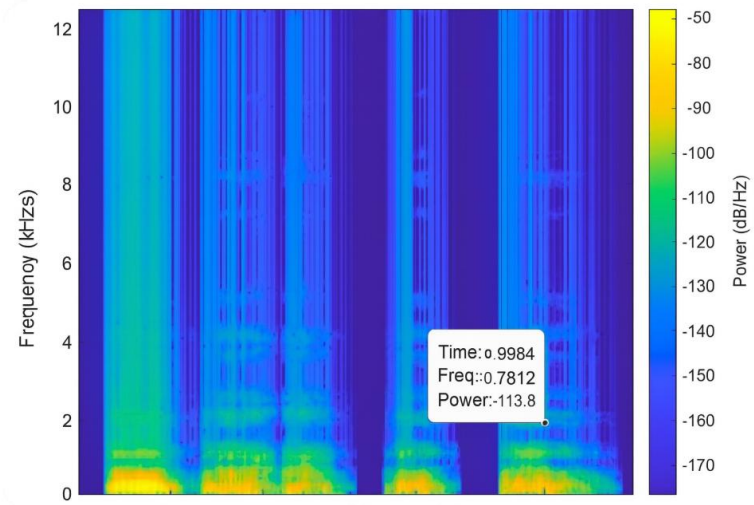

Figure 7. EMD time-frequency diagram

It is evident from the Figureure that the energy distribution of the enhanced time-frequency graph of LMS is significantly uneven and fluctuates greatly, indicating that the output signal strength is not strong. However, from Figure7, 8, and 11, as time goes from 0-1, the signal shows that the low-frequency band remains basically unchanged (0-1kHz), and other amplitudes are also approximately unchanged. However, in the high-frequency band, the frequency is not very stable (fluctuating between 2-3kHz), and then moves towards the high-frequency band, becoming even more unstable (fluctuating between 4-11kHz), but the amplitude is generally approximately unchanged (yellow). Looking at Figure 7, Figure8 alone, there is no subtle difference, but looking at the labeled numbers, the difference between the three is reflected in the power. Obviously, the method of combining wavelet and EMD outputs the highest power in the time-frequency graph.

Figure 9 is a graphical representation of the average signal-to-noise ratio gain as a function of signal-to-noise ratio after combining wavelet transform with EMD. It is evident from the graph that the higher the signal-to-noise ratio, the smaller the gain. This phenomenon is commonly referred to as the "Falkman Rachev effect" or "Pendragon effect".

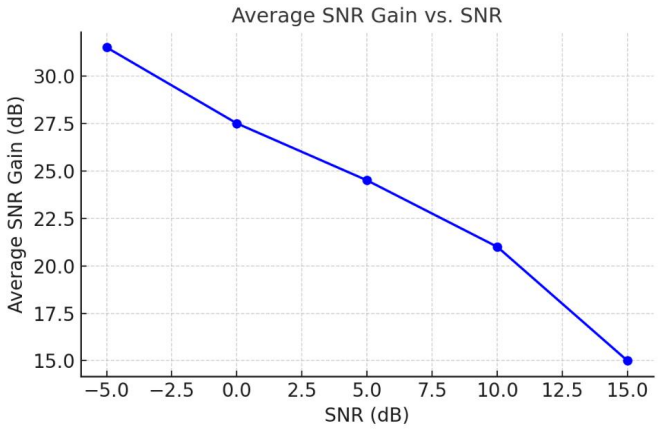

Figure 9. Average signal-to-noise ratio gain graph

6. Conclusion

The experimental results show that the combination of wavelet and EMD algorithm performs better than traditional methods in suppressing noise, improving speech signal clarity and adaptability. This indicates that by combining wavelet transform and EMD, different frequency components and non-stationary characteristics can be more effectively addressed, thereby improving the performance of speech enhancement algorithms and making them more practical and adaptable.

This improved wavelet and EMD method provides a new approach and method for research in the field of speech enhancement, offering a promising solution for more effective processing of speech signals in complex environments in practical speech processing applications.

References

[1] Li Yuanle, Duan Meijuan. Research on Speech Enhancement Algorithm Based on Wavelet Transform [J]. Information Recording Materials, 2023, 24(05): 117−120.

[2] Dessouky A M, Abbas, et al. Speech Enhancement with an Adaptive Wiener Filter [J]. International Journal of Speech Technology, 2014, 17(1): 53−64.

[3] Liu Wenqing. Research on Lung Sound Signal Reconstruction and Respiratory Cycle Segmentation Using EMD and Wavelet Transform [D]. Jianghan University, 2023.

[4] Wang J, Liu W, Zhang S, et al. An Approach to Eliminating End Effects of EMD Through Mirror Extension Coupled with Support Vector Machine Method [J]. Personal and Ubiquitous Computing, 2019, 23(3–4): 443–452.

[5] Yang Long, Chen Jianming. Speech Enhancement Algorithms and Their Progress [J]. Electroacoustic Technology, 2015, 39(07): 35–39+50.

[6] Li R, Bao C, Xia B, et al. Speech Enhancement Using the Combination of Adaptive Wavelet Threshold and Spectral Subtraction Based on Wavelet Packet Decomposition [C]. IEEE International Conference on Signal Processing, IEEE, 2013: 481–484.

[7] Gao Cheng, Dong Changhong, Guo Lei. Wavelet Analysis and Applications [M]. Beijing: National Defense Industry Press, 2007.

[8] Lu Jing, Zhao Fenghai. An Improved Speech Enhancement Algorithm Based on Wavelet Transform and Spectral Subtraction [J]. Electroacoustic Technology, 2018, 42(12): 8–12+69.

[9] Wang Ting. Research on EMD Algorithm and Its Application in Signal Denoising [D]. Harbin Engineering University, 2010.

[10] Lin Li, Zhou Ting, Yu Lun. Boundary Effect Processing Techniques in EMD Algorithm [J]. Computer Engineering, 2009, 35(23): 265–268.

[11] Yang Yongfeng, Wu Yafeng. Application of Empirical Mode Decomposition in Vibration Analysis [M]. Beijing: National Defense Industry Press, 2013.

[12] Wang Ping. Application of Empirical Mode Decomposition in Underwater Acoustic Signal Denoising and Feature Extraction [D]. Guilin University of Electronic Science and Technology, 2023.

[13] Qi Yanjie, Wang Liming, Yang Zehui, et al. Comparative Study on Several Methods for Improving EMD Endpoint Effects [J]. Modern Electronic Technology, 2013(22): 50–52.