Hindawi
Scientific Programming
Volume 2021, Article ID 8677043, 10 pages
https://doi.org/10.1155/2021/8677043

Research Article

Denoising Speech Based on Deep Learning and Wavelet Decomposition

Li Wang,[1] Weiguang Zheng,[2] Xiaojun Ma,[3] and Shiming Lin[4,5]

[1]*College of Chinese Literature and Media, Hubei University of Arts and Science, Xiangyang 441000, China*
[2]*School of Mechanical and Electrical Engineering, Guilin University of Electronic Technology, Guilin 541004, China*
[3]*Qinghai GLI Technology Limited, Xining 810001, China*
[4]*School of Informatics (National Demonstrative Software School), Xiamen University, Xiamen 361005, China*
[5]*Department of Computer Engineering, Changji University, Changji 831100, China*

Correspondence should be addressed to Weiguang Zheng; weiguang.zheng@foxmail.com and Shiming Lin; xmulsm@xmu.edu.cn

Received 7 May 2021; Revised 21 June 2021; Accepted 8 July 2021; Published 16 July 2021

Academic Editor: Yi-Zhang Jiang

The work proposed a denoising speech method using deep learning. The predictor and target network signals were the amplitude spectra of the wavelet-decomposition vectors of the noisy audio signal and clean audio signal, respectively. The output of the network was the amplitude spectrum of the denoised signal. Besides, the regression network used the input of the predictor to minimize the mean square error between its output and input targets. The denoised wavelet-decomposition vector was transformed back to the time domain by the output amplitude spectrum and the phase of the wavelet-decomposition vector. Then, the denoised speech was obtained by the inverse wavelet transform. This method overcame the problem that the frequency and time resolution of the short-time Fourier transform could not be adjusted. The noise reduction effect in each frequency band was improved due to the gradual reduction of the noise energy in the wavelet-decomposition process. The experimental results showed that the method has a good denoising effect in the whole frequency band.

1. Introduction

In the actual environment, speech signals are inevitably affected by the noises from the surrounding environment, transmission media, and electrical noise inside the communication equipment. These interferences greatly degrade the performance of the speech processing system and affect the quality of speech. Speech denoising aims to reproduce clean speech from noise-polluted signals, which is crucial for various applications, such as automatic speech recognition (ASR) and hearing aids. Several speech-denoising and speech-enhancement methods have been proposed based on the statistical difference between the speech and noise characteristics, including spectral subtraction [1], based estimation [2], Wiener filtering [3], subspace method [4], nonnegative matrix factorization (NMF) [5], and minimum mean square error (MMSE) [6].

Most of the filtering methods are limited to window-adding or masking operation in the frequency domain or time domain due to the strong time-frequency coupling between speech signals and noises. It is difficult for these filtering methods to achieve effective signal-noise separation. As a nonlinear filter, the neural network was applied to this problem in the past, such as the early use of the shallow neural network (SNN) for speech-denoising study. However, the constraints on computing power and the size of training data lead to the implementations of relatively small neural networks, limiting denoising performance.

By learning a deep nonlinear network structure, deep learning has the following advantages: achieving the approximation of complex functions, representing the distributed representation of input data, and demonstrating its powerful ability to learn data and essential characteristics from a few sample sets. Meanwhile, it emphasizes the deep structure of the learning model. The current learning framework usually adopts a multilevel model. In this way, the training of the model relies on a large number of data sets, highlighting the importance of big data for a complete

and complex model. Deep learning also focuses on feature learning. Deep neural networks (DNNs) contain multiple nonlinear hiding layers, showing great potential to capture the complex relationship between noises and clean speeches. Many training algorithms have been proposed to train a deep network. DNNs have been applied to speech recognition [7], speech denoising [8], and speech separation [9].

Recently, Zhao et al. [10] used both convolutional and recurrent neural network architectures to exploit local structures in both the frequency and temporal domains for speech enhancement. Tan and Wang [11] combined the convolutional code-decoder (CED) and long short-term memory (LSTM) into the convolutional recurrent network (CRN) to achieve real-time monophonic speech enhancement. The proposed model is independent of noise and speaker. Moreover, the trainable parameters of CRN are much smaller. The full connection layer involved in deep neural networks (DNN) and convolutional neural networks (CNN) may not accurately describe the local information of the speech signal, especially for the high-frequency component. Therefore, Fu et al. [12] proposed an enhancement model of a full convolutional network (FCN) based on the original waveform. The system performs speech enhancement in an end-to-end manner, different from most existing denoising methods only dealing with amplitude spectrum.

Speech is a time-varying signal, in which usually changes occur at syllabic rates of 10 times/sec and exceeds the fixed time intervals of 10–30 msec. Short-time Fourier transform (STFT) is often used to analyze the speech on a time-frequency range [8, 9]. However, the window length of the STFT is fixed, that is, the time-domain resolution is fixed. According to the Heisenberg uncertainty principle, the frequency-domain resolution is also fixed. For a low-frequency signal, the time interval should be wider to determine the frequency better; however, for high-frequency signals, the time domain should be narrower to locate them better in the time domain. The resolution of STFT is not adjustable in the time domain and frequency domain, so it is not suitable for broadband analysis.

Wavelet analysis, developed in the 1980s, plays an important role in signal processing [13]. Wavelet transform (WT) has multiresolution and can adjust the window function adaptively according to the signal frequency. For low-frequency signals, WT provides low time-domain resolution and high-frequency domain resolution. For high-frequency signals, it provides high resolution in the time domain and low resolution in the frequency domain [14]. The wavelet transform coefficient reaches a maximum value in a certain region, and this point is called the modulus maximum of the wavelet transform in the region. The modulus maxima of useful signals in the multiresolution analysis increase with the decreased resolution; however, the modulus maxima of noisy signals in the multiresolution analysis decrease with the decreased resolution [15]. Threshold values are set according to the characteristics of useful signals and noise, and the wavelet coefficient is analyzed using this threshold value. When the wavelet coefficient is lower than this threshold value, the wavelet coefficient corresponds to a noise signal. In the wavelet domain, the threshold is used to distinguish the useful signal from the noise signal. Finally, processed wavelet coefficients are reconstructed to obtain denoised signals [16].

The work proposed a speech denoising method based on deep learning. The predictor and target network signals were the amplitude spectrum of the wavelet-decomposition vector of the noisy audio signal and clean audio signal, respectively. The output of the network was the amplitude spectrum of the denoising signal. The output spectrum and the phase of the wavelet-decomposition vector were used to transform the denoised wavelet-decomposition vector back to the time domain. Then, the denoised speech was obtained by the inverse-wavelet transform. This method overcame the problem that the frequency and time resolution of STFT could not be adjusted.

2. General Theory

2.1. Short-Time Fourier Transform. STFT is widely used in speech analysis and processing, suitable for slow signal and time-varying signal spectrum analysis. In this method, the speech signal is first divided into frames, and then, the Fourier transform is carried out for each frame. Each frame of the speech signal can be intercepted from a variety of stationary signal waveforms, and the short-time spectrum of each frame of speech is an approximation of the spectrum value of the smooth signal waveform. Since the signal of each frame is short and stable, the Fourier transform of the frame signal is calculated to obtain the STFT:

$$\mathrm{STFT}_x(t, f) = \int_{-\infty}^{\infty} x(t)h(t-\tau)e^{-j2\pi f\tau}d\tau, \tag{1}$$

where $\mathrm{STFT}_x(t, f)$ is the coefficient of STFT. STFT is a function of time t and frequency f, which shows how the frequency of the speech signal changes with time.

According to the above STFT transformation, its inverse transformation can be defined as

$$x_t = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{STFT}_x(t', f')w(t-t')e^{-j2\pi f' t'}dt'df', \tag{2}$$

where $w(t)$ is a window function. The longer window length means higher spectral resolution; however, the time resolution of the long window decreases correspondingly. Due to the contradiction between the time resolution and the frequency resolution, the practical operation should be based on the STFT analysis, and the appropriate window length should be determined.

2.2. Wavelet Transform. STFT is a windowed FT transform. FT is based on sinusoidal functions of different frequencies, so the signal is often decomposed into the superposition sum of sinusoidal waves of different frequencies. The wavelet transform replaces the infinitesimal trigonometric basis function with the wavelet basis of finite length and attenuation, thus locating frequency and time. The continuous wavelet transform (CWT) is the inner product of wavelet function $\phi(t)$ and square-integrable function $x(t)$ with good local properties in the time-frequency domain:

$$\begin{aligned} \mathrm{CWT}_x(a,b) &= f, \\ \phi_{a,b} &= \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\phi^*\left(\frac{t-b}{a}\right)dt \ , \end{aligned} \tag{3}$$

where $a > 0$ is the scale factor and b the displacement factor. The scale factor plays an important role in wavelet transform. When it is very small, it will show the details of the signal changing rapidly. When it is large, the wavelet is extended to show the coarse features of the signal. When $\phi(t)$ meets the admissibility condition, the inverse continuous wavelet transform (ICWT) is

$$x(t) = \frac{1}{C_\phi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathrm{CWT}_x(a,b) \frac{1}{\sqrt{a}} \widetilde{\phi}\left(\frac{t-b}{a}\right) \frac{1}{a^2} dt \ da, \tag{4}$$

where $\widetilde{\phi}(t)$ is a dual function of $\phi(t)$ and C_ϕ an admissible constant. The data from CWT has large redundancy, which may not be suitable for DNN training for denoising speech. Discrete WT (DWT) uses filter banks to implement the Mallat algorithm. Figure 1 shows the three-level DWT, where cA1, cA2, and CA3 are approximate coefficients containing low-frequency information of the signal. cd1, cd2, and cd3 are detail coefficients and contain high-frequency information of the signal. c is the wavelet-decomposition vector; l is the bookkeeping vector containing the number of coefficients of each level.

2.3. Convolution Neural Networks for Deep Learning. A convolution neural network of deep learning is a deep-learning network generated on the theoretical basis of a neural network. The neural network is a fully connected network, that is, each neuron in the upper layer is connected to a neuron in the next layer. In this case, for multidimensional input information such as sound or image, the amount of information contained is relatively large; for the hidden layer, the traditional BP algorithm requires more weight parameters. The resulting slow training speed leads to more samples required for training. Overfitting is more likely to occur with insufficient training. In this way, the parameters learned are not universal, so they cannot represent and restore the input signal.

Ordinary neural network structure does not consider the characteristics of the input data. Even for a little change in the original data, the neural network does not take into account the data characteristics for optimized training. The neural network is fully connected, and all input data need to be considered; thus, it is impossible to identify and train the local regional features in the data.

Given the problems existing in the above ordinary neural network structure, the convolutional neural network transforms the ordinary neural network through local connection to feel the field of vision, weight sharing, and subsampling process through a local connection. It is used to learn features. Figure 2 shows the convolutional neural network model.

The total core operation of convolution in the convolution layer is as follows:

$$x_j^l = f\left(\sum_{i \in M_j} x_i^{l-1} \times k_{ij}^l + b_j^l\right), \tag{5}$$

where k is the convolution kernel (filter), l is the number of layers, M is the j^{th} feature map, b is the corresponding bias, and f is the activation function. The result of the convolution layer output goes to the downsampling layer, and down sampling is performed on each feature of the output in the convolution layer.

3. Proposed Method

Wavelet-decomposition vector c can be denoted as

$$c = [\ \mathrm{cA}_n\ \ \mathrm{cD}_n\ \ \ldots\ \ \mathrm{cD}_i\ \ \ldots\ \ \mathrm{cD}_1\ ]. \tag{6}$$

Assuming that the length of the signal is L and the frequency is Fs, the highest frequency of the signal is Fs/2. The frequency range of the lowest layer $\mathrm{cA_n}$ is $(0, \mathrm{Fs}/2^{n+1})$, with the size of $\mathrm{L}/2^n$. The frequency range of $\mathrm{cD_i}$ is $(\mathrm{Fs}/2^{i+1}, \mathrm{Fs}/2^i)$, with the size of $\mathrm{L}/2^i$. If we do STFT for c and select the window width as n_w, the sampling of cA_n is equivalent to the window width of about $2^{n-1} * n_w$ for the original signal, and the window width of cD_i is equivalent to that of the original signal $2^{i-1} * n_w$. In other words, if the frequency drops by one time, the window width increases by one time. Thus, we realize the effect of wavelet transform of large time windows at low frequency and small-time windows at high frequencies, almost without data redundancy.

Figure 3 shows the proposed deep-learning training. The predictor and target network signals are the magnitude spectra of the wavelet-decomposition vector of the noisy and clean audio signals, respectively. The network's output is the magnitude spectrum of the denoised signal. The regression network uses the predictor input to minimize the mean square error between its output and the input target. The denoised wavelet-decomposition vector is converted back to the time domain using the output magnitude spectrum and the phase of the noisy wavelet-decomposition vector. Then, the denoised speech can be obtained from the inverse wavelet transform.

4. Experiments and Discussion

The work used the Chinese Common Voice Corpus 6.1 subset of the Mozilla Common Voice dataset [17] to train and test our proposed method. Vehicle noise (Volvo) from the NOISEX-92 database [18] was taken as the noise source. The speech and noise were resampled at 16 kHz. The signal-to-noise ratios (SNR) of 5, 0, and -5 dB were set to compare the denoising effect.

Morse wavelet function was used in DWT. Another DNN method used STFT and convolution neural network for comparison [19]. The window length of 64 of STFT was adopted for our proposed method and those of 64 and 256 were adopted for the compared method. Hamming window with an overlap of 75% was used in all cases.

Figure 4 shows the clean speech and the noisy speech with different SNRs in the time domain and spectrogram.

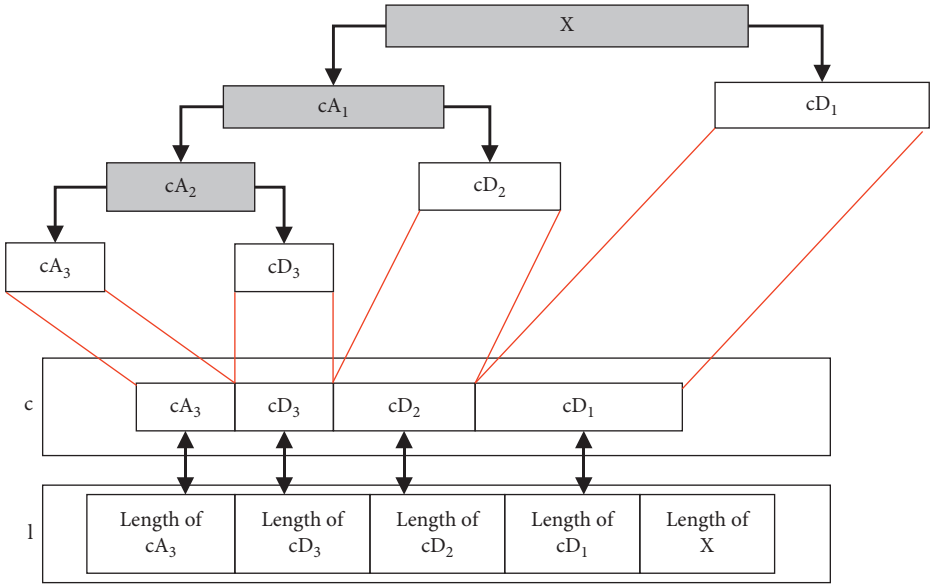

Figure 1: The diagram of wavelet decomposition.

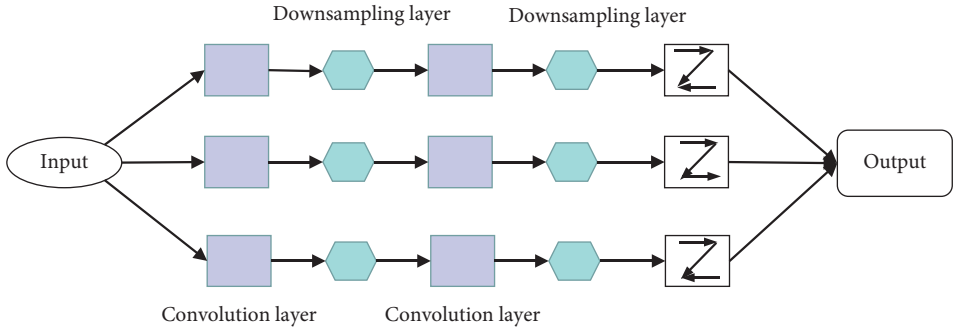

Figure 2: The diagram of the convolutional neural network model.

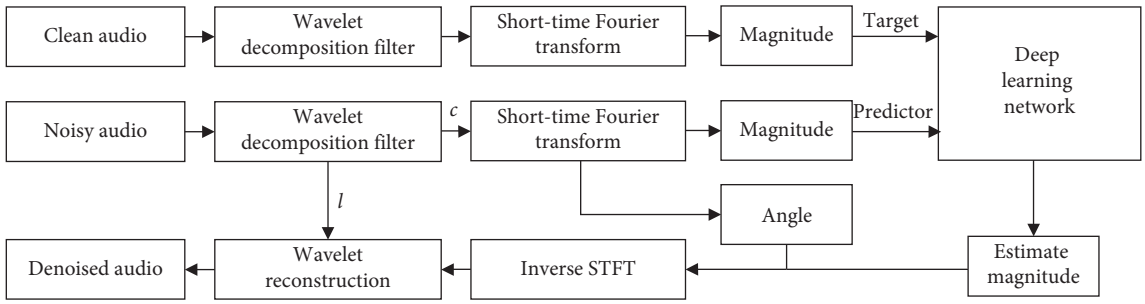

Figure 3: Block diagram of our proposed method.

The noise pollutes the noise in the broadband frequency. As the SNR decreases, more speech information is drowned out.

Figure 5 shows the speech signal enhanced by subtracting amplitude spectra. The noise has been reduced partly. The spectrogram shows that the rough points of the original noisy speech have been reduced to a large extent. Due to the half-wave rectification of negative values, small, independent peaks appear on the random frequency of the multiframe spectrum. Transformed to the time domain, these peaks sound like multiple vibratos with random frequency changes between frames, which is commonly referred to as music noise.

In Figure 6, after Wiener filtering, the speech signal polluted by noise has been improved to a certain extent. However, there are still some noises after Wiener filtering, related to the filtering characteristics of the Wiener filter.

Figures 7–9 show the denoising results using the proposed method and the compared method, respectively. The results of the proposed method show a better denoising effect from high to low SNRs in the whole frequency range. The compared

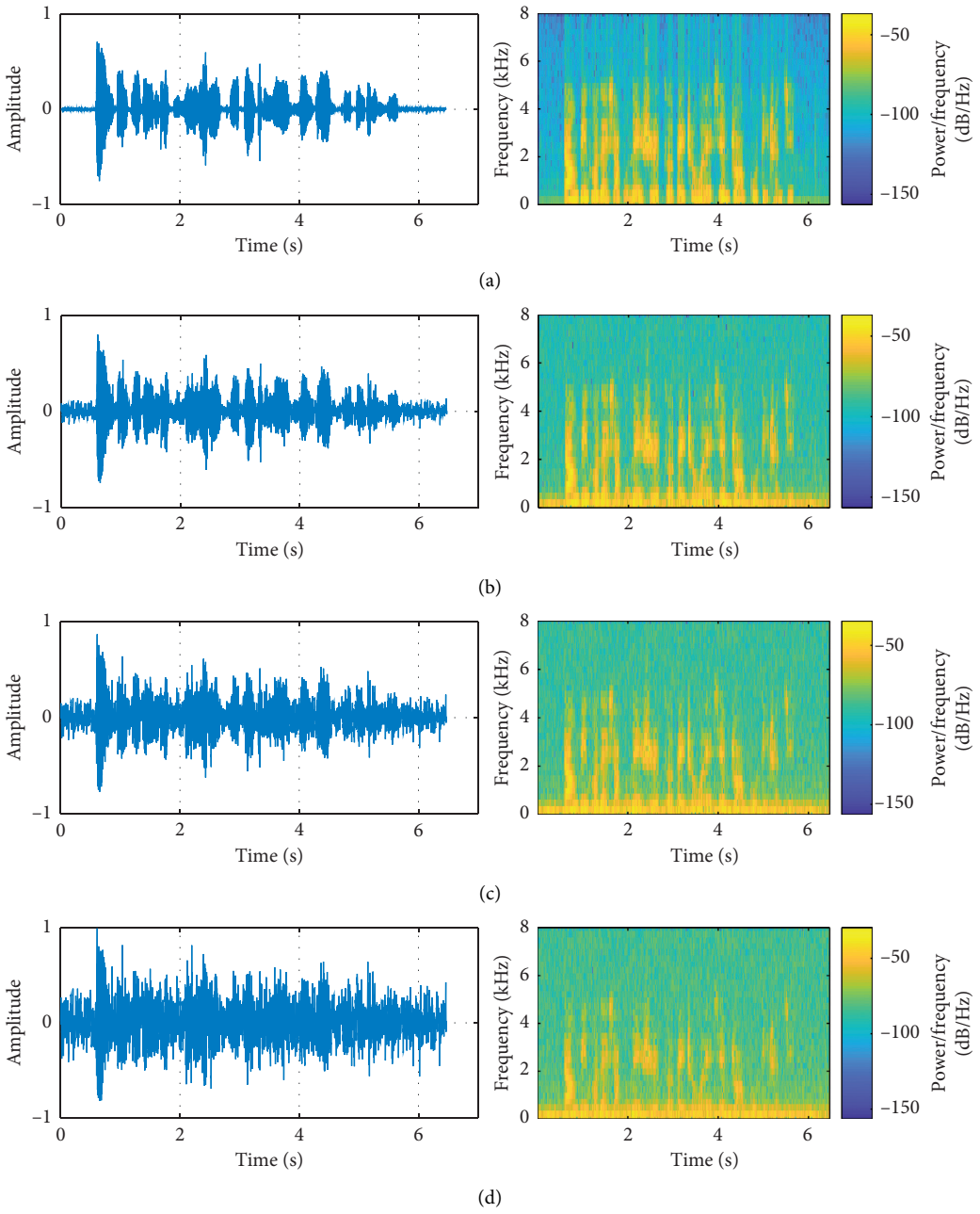

Figure 4: Clean speech and noise speech. (a) clean speech; (b) noisy speech (SNR 5 dB); (c) noisy speech (SNR 0 dB); (d) noisy speech (SNR-5 dB). Left: time domain. Right: spectrogram.

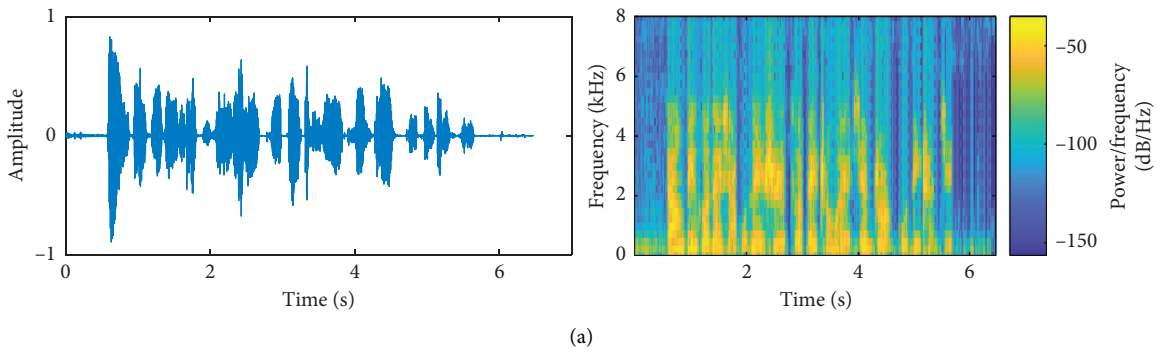

Figure 5: Continued.

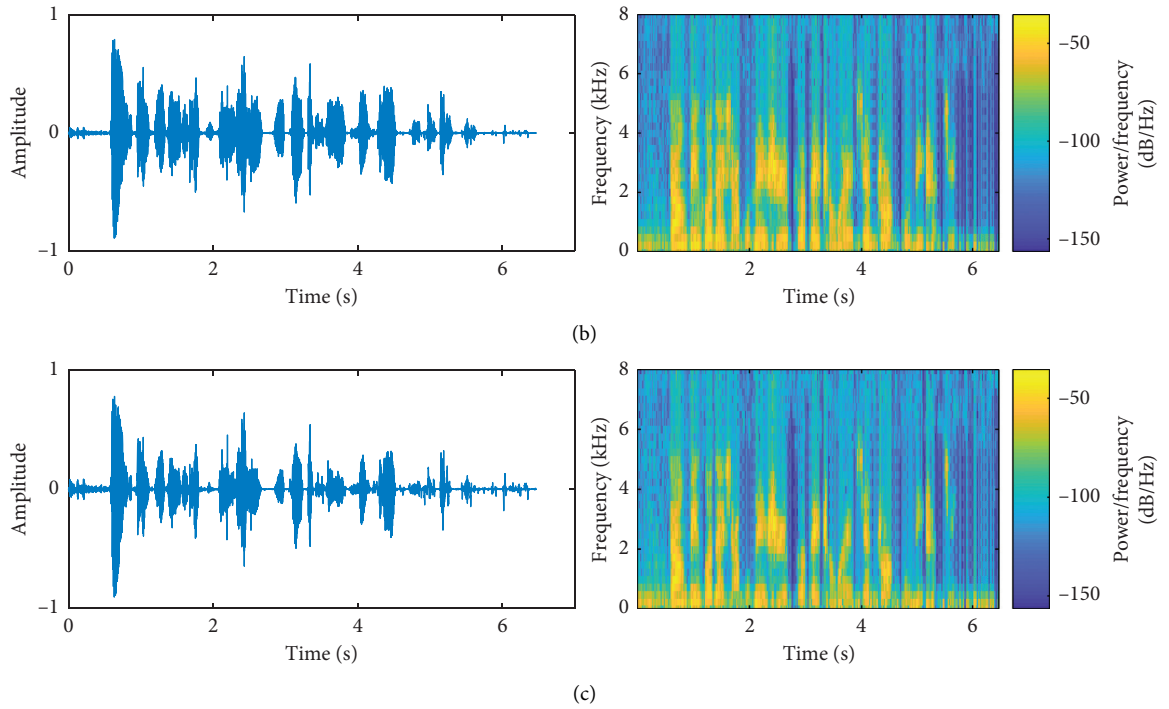

Figure 5: Enhanced speech using spectral subtraction. (a) SNR = 5Db; (b) SNR = 0 dB; (c) SNR = −5 dB. Left: time domain. Right: spectrogram.

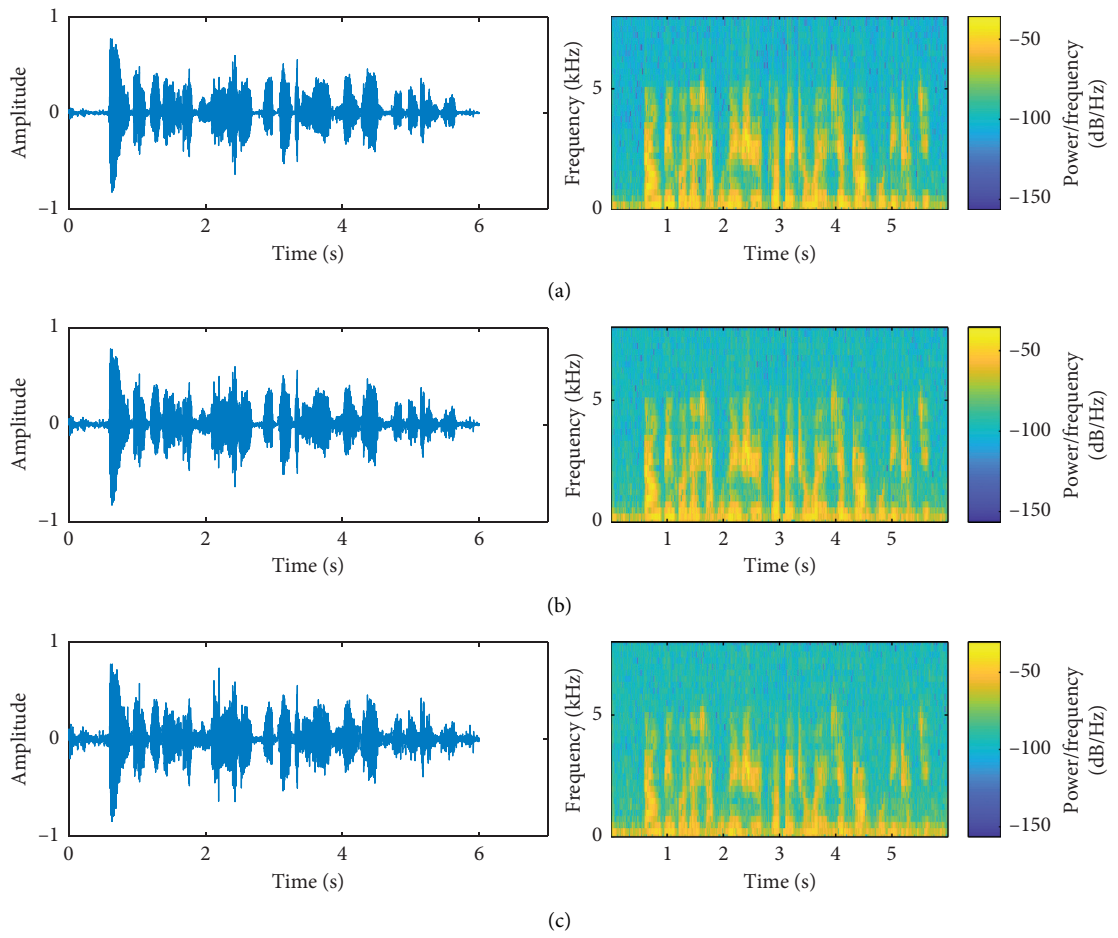

Figure 6: Enhanced speech using Wiener filtering. (a) SNR = 5 dB; (b) SNR = 0 dB; (c) SNR = −5 dB. Left: time domain. Right: spectrogram.

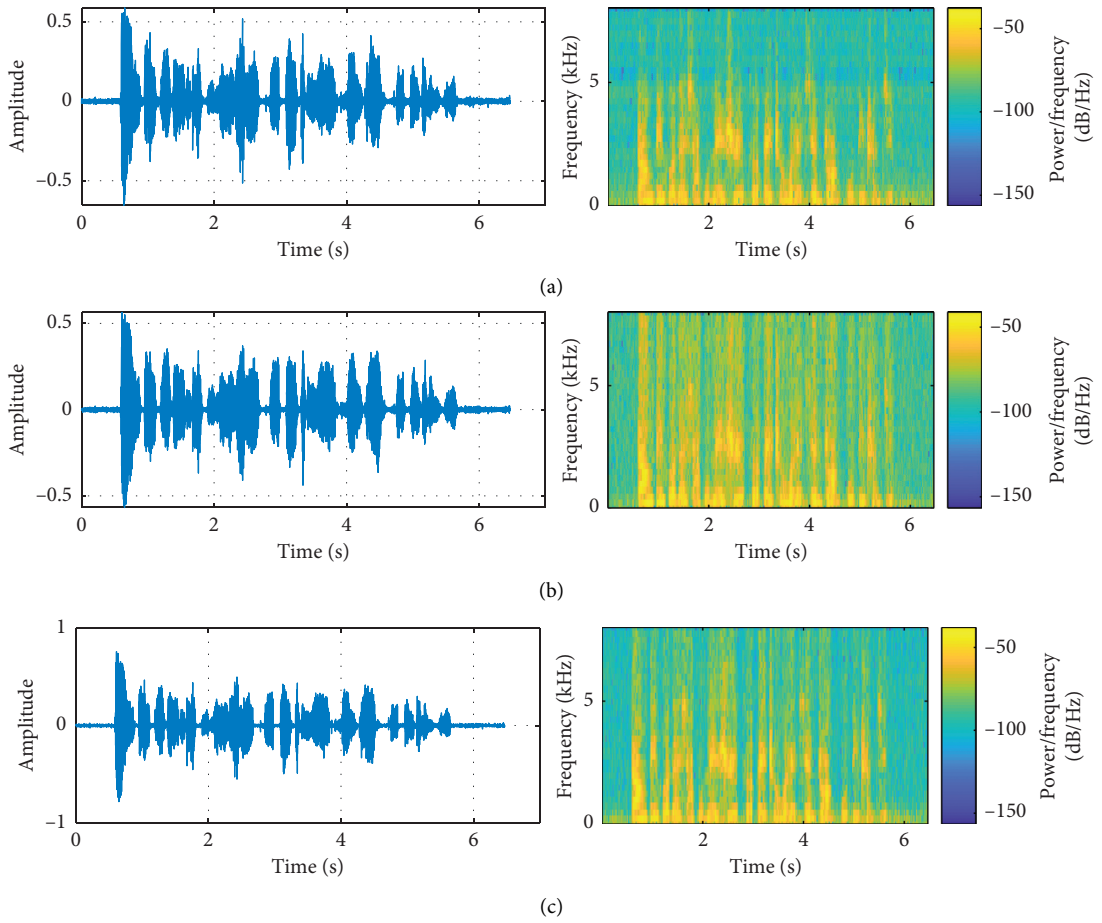

Figure 7: Enhanced speech (SNR = 5 dB). (a) The proposed method. (b) The comparison method with 256 window lengths. (c) The comparison method with 64 window lengths. Left: time domain. Right: spectrogram.

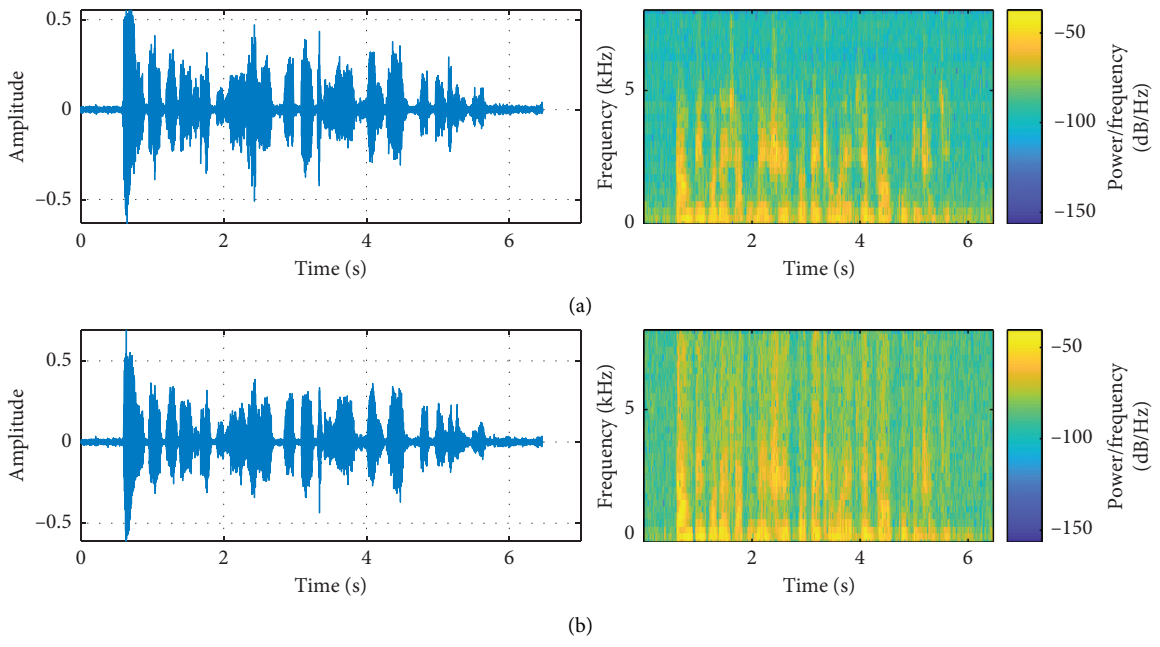

Figure 8: Continued.

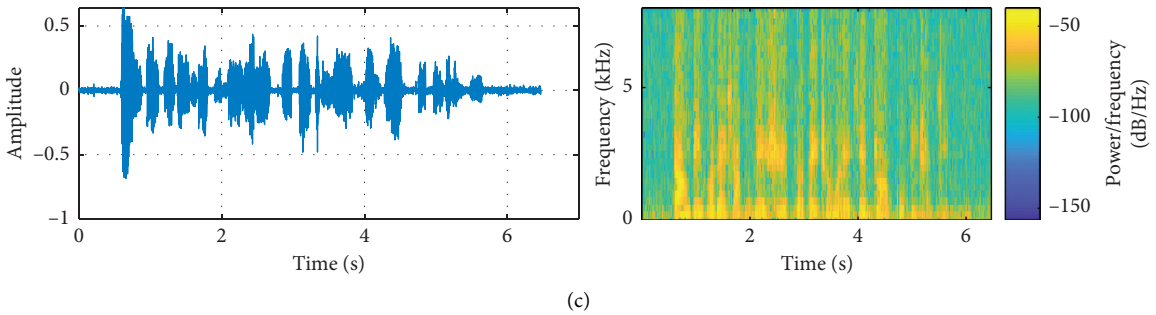

Figure 8: Enhanced speech (SNR = 0 dB). (a) The proposed method. (b) The comparison method with 256 window lengths. (c) The comparison method with 64 window lengths. Left: time domain. Right: spectrogram.

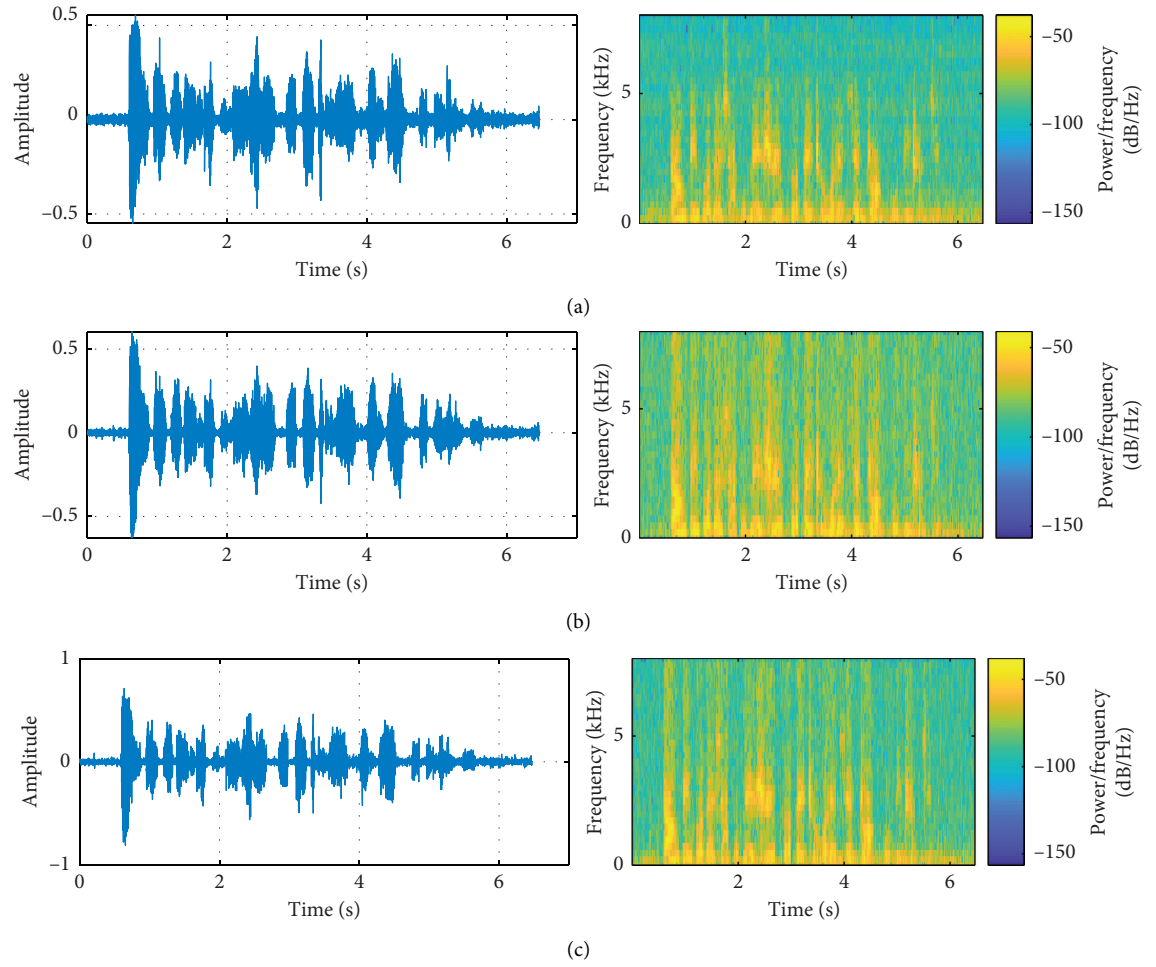

Figure 9: Enhanced speech (SNR = −5 dB). (a) The proposed method. (b) The comparison method with 256 window lengths. (c) The comparison method with 64 window lengths. Left: time domain. Right: spectrogram.

method with the window length of 256 achieves some noise reduction effect, but it performs poorly in the high-frequency band. The compared method with 64 window lengths performs some superiority in the high-frequency band, but is still inferior to the proposed method. In the process of the wavelet transform, the signal energy in the frequency band remains the same with the reduced noise energy, which improves the SNR in the frequency band and denoising effect. Table 1 shows the SNR of the denoising speech, indicating the proposed method is an improvement of the compared method.

Table 1: SNR for denoising speech.

Noisy speech	Proposed method	Comparing method (256 window length)	Comparing method (64 window length)
5	16.7	15.5	12.6
0	14.3	13.7	12.3
−5	13.2	12.5	9.5

5. Conclusions

For the proposed method in the work, the predictor and the target network signals were the amplitude spectra of the wavelet-decomposition vector of the noisy audio signal and the clean audio signal, respectively. The output of the network was the amplitude spectrum of the denoising signal. The regression network used the input of the predictor to minimize the mean square error between its output and input targets. The denoised wavelet-decomposition vector was transformed back to the time domain using the output amplitude spectrum and the phase of the denoised wavelet-decomposition vector. Then, the denoised speech was obtained by the inverse wavelet transform.

The proposed method overcame the problem that the frequency and time resolution of STFT could not be adjusted. Besides, since the noise energy was gradually reduced during wavelet decomposition, the noise reduction effect of each frequency band was improved. The experimental results showed that the proposed method has a good denoising effect in the whole frequency band.

Data Availability

The datasets and codes of this paper for the simulation are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was supported by the project of Hubei University of Arts and Science (XK2020021), Natural Science Foundation of Guangxi (No. 2018GXNSFAA281276), and Liudong Science and Technology Project (20200108).

References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, & Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, & Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[3] P. Scalart and J.V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proceedings of the Acoustics, Speech, and Signal Processing, 1996 ICASSP-96*, pp. 629–632, IEEE, Atlanta, Georgia, USA, May 1996.

[4] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[5] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4029–4032, IEEE, Las Vegas, NV, USA, March 2008.

[6] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Florida, United States, 2007.

[7] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[8] D. Liu, P. Smaragdis, and M. Kim, "Experiments on Deep Learning for Speech denoising," in *Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association*, pp. 1–5, (ISCA), Singapore, September 2014.

[9] Y. Wang, J. Du, L.-R. Dai, and C.-H. Lee, "A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1535–1546, 2017.

[10] H. Zhao, S. Zarar, I. Tashev, and C.-H. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2401–2405, IEEE, Calgary, AB, Canada, April 2018.

[11] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proceedings of the Proc. Interspeech*, pp. 3229–3233, Hyderabad, India, June 2018.

[12] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 6–12, IEEE, Kuala Lumpur, Malaysia, December 2017.

[13] I. Daubechies, "Where do wavelets come from? A personal point of view," *Proceedings of the IEEE*, vol. 84, no. 4, pp. 510–513, 1992.

[14] S. He, J. Chen, Z. Zhou, Y. Zi, Y. Wang, and X. Wang, "Multifractal entropy based adaptive multiwavelet construction and its application for mechanical compound-fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 76-77, pp. 742–758, 2016.

[15] A. Cohen, I. Daubechies, and J.-C. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, vol. 45, no. 5, pp. 485–560, 1992.

[16] X. Ma, C. Zhou, and I. J. Kemp, "Automated wavelet selection and thresholding for PD detection," *IEEE Electrical Insulation Magazine*, vol. 18, no. 2, pp. 37–45, 2002.

[17] https://voice.mozilla.org/en.

[18] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.
[19] https://ww2.mathworks.cn/help/deeplearning/ug/denoise-speech-using-deep-learning-networks.html.