

A Review of Speech Enhancement and Deep Learning Methods

Rongsheng Liu^{1*}

¹ School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu, Anhui, China

Email Address

Lrs20000109@gmail.com (Rongsheng Liu)

*Correspondence: Lrs20000109@gmail.com

Received: 14 October 2023; **Accepted:** 12 November 2023; **Published:** 25 December 2023

Abstract:

The speech processing system is inevitably interfered by various noises. The sources of noise are wide and difficult to avoid, including the surrounding environment, the introduction of transmission media, the electrical noise inside the equipment, and even the interference of other speakers. The noise not only reduces the quality and intelligibility of the speech, but also seriously affects the efficiency of the speech processing system. Speech enhancement is an important branch in the field of speech signal processing and one of the effective means to solve noise pollution. Its main purpose is to improve the quality of speech and improve the intelligibility of speech. As a front-end module, speech enhancement technology is widely used in speech recognition systems, speech coders, hearing aids and other fields. In recent years, deep learning has become a research hotspot in the field of machine learning. The successful application of deep neural network in the field of image classification and speech recognition provides an idea for solving the problem of speech enhancement in complex and changeable noise environment. Compared with other machine learning, deep learning has more powerful learning ability. Originally, through summarizing the research results of scholars in the field of speech enhancement, we systematically expounded the relevant basic knowledge of this discipline.

Keywords:

Speech Enhancement, Deep Learning, Deep Neural Network, Noisy Speech, Spectral Estimation, Signal to Noise Ratio

1. Introduction

In human social life, voice plays an important role. Communication between people and the realization of human-computer interaction are inseparable from voice signals. In daily life, we can communicate by voice through social software such as WeChat and QQ. However, during the voice communication process, we will inevitably be disturbed by some noise around us. These noises are everywhere and they affect our hearing experience all the time. , can easily cause our auditory fatigue, and even have a critical impact in some special situations, such as isolating important voice information from the black box of a crashed plane, filtering out vital key voices from

the sound of gunfire, etc. wait. Therefore, in order to improve the quality and intelligibility of speech, it is particularly important to extract relatively pure speech from noisy signals through speech enhancement technology. Speech enhancement can not only be used in the field of daily communication, but also plays an extremely important role in speech recognition and coding systems. In recent years, speech recognition technology has made breakthrough progress. It is becoming more and more popular as a natural and convenient way of human-computer interaction. People are paying more and more attention to it, especially with the success of Beijing's Olympic bid. Mobile speech recognition and translation systems will have a good market. However, most of the current speech recognition systems work in quiet environments. The introduction of background noise will seriously affect the performance of the recognition system. It is precisely because the inevitable noise in the actual environment greatly hinders speech recognition technology that the necessity of speech enhancement becomes apparent. In addition, voice enhancement is increasingly used in multimedia applications, smart home appliances, and national security and public security departments to improve interception effects and help solve cases. Nowadays, speech enhancement technology has been widely used in various speech signal processing fields [1].

There are two main purposes of speech enhancement. One is to improve the quality of speech and eliminate background noise so that the listener is willing to accept it and is less prone to fatigue. This is a subjective measurement. The other is to improve the intelligibility of speech. The purpose of improving the speech signal-to-noise ratio is an objective measurement. Generally speaking, the two cannot be satisfied at the same time. Some existing speech enhancement algorithms can significantly reduce background noise, increase signal-to-noise ratio, and improve speech quality, but generally cannot improve speech intelligibility. In many cases, although the signal-to-noise ratio is improved after processing, it will reduce the intelligibility of speech. Intelligibility, therefore different choices need to be made based on the specific occasion of speech enhancement.

2. Speech Enhancement

Speech enhancement means that when clean speech receives interference from various noises in real-life scenarios, it is necessary to filter out the noise through certain methods to improve the quality and intelligibility of the speech and make the listener less likely to feel tired. The purpose of speech enhancement is to separate clean speech from noisy speech. Noise and clean speech have large differences in the distribution of spectral characteristics, which also provides the possibility of realizing speech enhancement [2].

If divided according to the number of recording channels, speech enhancement is divided into monophonic speech enhancement and microphone array speech enhancement. Mono-channel speech enhancement only uses information in the time domain and frequency domain, while microphone array speech enhancement not only uses information in the time domain and frequency domain, but also uses information in the spatial domain. Therefore, the task of monophonic speech enhancement is more difficult, but the hardware cost requirements are relatively low. In addition, the microphone array speech enhancement method has strict requirements on the consistency of the physical characteristics of each microphone, and has high requirements on the time-varying spatial position of the sound source, that is, it does not allow large and rapid movement of the sound source. However, it must be

mentioned that the human auditory system is a dual-microphone structure. The human ear can perceive the spatial location distribution of different sound sources based on the difference in delay and sound intensity between different sound sources reaching the left and right ears, and then assist in detecting different sound sources. Sound source separation of sound sources. Interestingly, the human auditory system can easily distinguish the different sound sources in the mono noisy speech even when listening to the recorded mono noisy speech; or when different sound sources are in the same position, the human ear can also successfully distinguish and understand different sound sources.

If divided according to the speech enhancement method, it can be divided into unsupervised speech enhancement method and supervised speech enhancement method. The former is also called the traditional speech enhancement method. Although the traditional speech enhancement algorithm does not require offline training, the computing resources required are also small. Few, but due to the existence of many unreasonable assumptions, their performance upper limit is limited. Supervised speech enhancement methods have been proposed in recent years. They use existing speech data or noise data to train relevant statistical characteristic models. In fact, just like humans, first let the system learn and remember some speech and noise patterns. Use this guide to separate noise from noisy speech. Since the supervised speech enhancement method makes full use of existing data and masters some statistical characteristics of speech and noise, it will obtain better enhancement performance.

2.1. Speech signal characteristics

2.1.1. Speech characteristics

a) The speech signal is a time-varying, non-stationary random process.

There is a certain limit to the rate of change of the physiological structure of the human vocal system. The shape of the human vocal cords and vocal tract is relatively stable within a relatively short period of time. In speech enhancement, it is often necessary to take advantage of this relative stability of the speech signal.

b) Speech can be divided into two categories: unvoiced and voiced. Voiced sounds show obvious periodicity in the time domain: they have a formant structure in the frequency domain, and most of the energy is concentrated in the lower frequency band. Unvoiced sound is different. It has no obvious time domain and frequency domain characteristics, and is somewhat similar to white noise. In speech enhancement, the periodic characteristics of voiced sounds can be exploited.

c) As a random process, speech signals can be described by statistical analysis characteristics. Since speech is a non-stationary and non-ergodic random process, long-term time domain statistical models are of little significance. The vocalization process of speech can be modeled as a linear time-varying filter excitation source to form speech through this vocal tract model. In the case of voiced sounds, the excitation source is a periodic pulse generator, and in the case of unvoiced sounds, it is a random noise generator.

2.1.2. Noise characteristics

Noise comes from the objective environment, and the sources of noise are endless. Therefore, the characteristics of noise can be said to be endlessly changing.

Researchers roughly divide noise into additive noise and non-additive noise. Additive noise includes periodic noise, impulse noise, broadband noise and interference from other voices in the same channel.

a) Periodic noise. The biggest characteristic of periodic noise is that it has many discrete line spectra, which mainly originate from mechanical and electrical interference caused by periodic operation of engines, etc. In particular, power supply hum can also cause periodic noise. The periodic noise generated in the actual environment does not simply contain spectral line components, but is composed of many narrow-band spectra. Moreover, it is often time-varying and overlaps with the spectrum of the speech signal. An adaptive filtering method must be used to automatically identify and distinguish the noise components.

b) Impulse noise. Impulse noise represents the sudden appearance of narrow pulses in a time domain waveform. It comes from explosions, impacts and discharges. Generally, it can be eliminated by threshold judgment or smoothing filtering. The threshold is determined based on the average amplitude of the noisy speech signal. When the signal amplitude exceeds this threshold, it is judged to be impulse noise and then appropriately attenuated or even completely eliminated; it can also be interpolated based on adjacent signal samples. Impulse noise is smoothed in the time domain, which is a straightforward method of elimination.

c) Broadband noise. There are many sources of broadband noise, including thermal noise, airflow (such as wind, breathing) noise and various random noise sources. Quantization noise can also be considered as broadband noise. Broadband noise is the most difficult to eliminate because it completely overlaps with the speech signal in both the time and frequency domains. This noise only exists alone during speech intervals. For stationary broadband noise, it can usually be considered as white Gaussian noise. Noise that does not have a white spectrum can be whitened first.

d) Simultaneous speech interference. The human ear can distinguish the required sounds in a speech environment of more than two people. This discrimination ability is a perceptual ability of the human body's internal speech understanding mechanism and comes from the human binaural effect. But when multiple voices are stacked together and transmitted in a single channel, the binaural signals disappear due to merging. Therefore, the resolution ability is reduced. Independent component analysis ICA (playing dependent Component Analysis) is a feasible method.

e) Transmission noise. Transmission noise is also the circuit noise in traditional systems. In the time domain, it is the convolution of sound and noise. This kind of noise can be processed by homomorphic processing, by converting non-additive noise into additive noise.

All in all, noise destroys the original acoustic features and model parameters of speech signals, blurs the differences between different speech sounds, reduces speech quality and intelligibility. Strong noise can cause auditory fatigue and also affect the speaker, causing the speaker to change the way he speaks in a quiet environment or a low-noise environment, thereby changing the characteristic parameters of speech. This situation is called the Lombard effect. It can severely damage the efficiency and accuracy of the speech recognition process.

2.1.3. Perceptual characteristics of human ear

The human ear has the ability to distinguish the sounds that need to be heard in an environment where more than two people are speaking. The human ear has a masking effect, that is, strong signals mask and suppress weak signals. The degree of masking is a binary function of sound intensity and frequency, and the masking of components near frequencies is much more effective than components with large frequency differences. In addition, cutting off high-frequency components has a greater impact on consonants than on vowels, while removing low-frequency components has a greater impact on vowel clarity than on consonants.

The human ear's perception of speech is mainly obtained through the amplitude of the spectral component of the speech signal, and is not sensitive to the phase spectrum; the human ear's perception of the intensity of the spectral component is a binary function of frequency and energy spectrum, and loudness is proportional to the logarithm of the spectrum amplitude. ; Formants are very important to speech perception, especially the second formant is more important than the first formant. In other words, performing a certain degree of high-pass filtering on the speech signal will not affect intelligibility.

2.2. Signal-to-noise ratio and segmented signal-to-noise ratio

a. Signal-to-noise ratio (SNR)

Signal-to-noise ratio is a conventional method for measuring speech enhancement algorithms for broadband noise distortion. It refers to the ratio of signal to noise in an electronic device or electronic system. The signal here refers to the electronic signal from outside the device that needs to be processed by this device. The noise refers to the irregular additional signal (or information) that does not exist in the original signal generated after passing through the device, and this kind of The signal does not change as the original signal changes.

Assuming that $s(n)$ represents the pure speech signal and $\hat{s}(n)$ represents the enhanced signal, the signal-to-noise ratio is defined as:

$$SNR = 10 \log_{10} \frac{\sum_n s^2(n)}{\sum_n [s(n) - \hat{s}(n)]^2}$$

This method is mainly used in simulations of known pure speech, and its main advantage is simple calculation. But it has two main problems: first, this form of signal-to-noise ratio has no good connection with the subjective attributes of speech quality; second, it treats all errors in the time domain waveform equally. Because speech energy is time-varying, the signal-to-noise ratio between frames is different. If a pronunciation has a large energy accumulation in its voiced part, it is possible to obtain a deceptive signal-to-noise ratio value. Segmented signal-to-noise ratio can improve the above problems.

b. Segmented signal-to-noise ratio (SEGSNR)

Segmented signal-to-noise ratio, also known as average short-term SNR, refers to the ratio of the average energy of short-term pure speech and short-term noise interference in all frames (expressed in dB), which is defined as:

$$SEGSNR = \frac{10}{M} \sum_{i=1}^{M-1} \log_{10} \frac{\sum_{n=0}^{L-1} s^2(i, n)}{\sum_{n=0}^{L-1} [s(i, n) - \hat{s}(i, n)]^2}$$

Among them, $s(i, n)$ is the pure speech, $\hat{s}(i, n)$ is the processed speech, L is the frame length, and M is the total number of frames. That is to say, $s(i, n) - \hat{s}(i, n)$ represents noise and distortion. Generally, the greater the SEGSR fluctuates, the smaller the noise and distortion contained in the speech, and the closer the waveform is to pure speech.

The above method of estimating the signal-to-noise ratio has a prerequisite that the expression of pure speech must be known in advance. However, in practice, it is difficult to obtain the time domain waveform of pure speech. Therefore, when the pure speech is unknown, the signal-to-noise ratio formula needs to be modified:

$$SNR = 10 \log_{10} |(P_y - P_n)/P_n|$$

Among them, P_y is the total power of noisy speech, and P_n is the estimated total noise power.

3. Basic Theory of Speech Enhancement Methods

3.1. Spectral subtraction

Spectral subtraction is the most widely used speech enhancement algorithm today due to its simple structure and easy use on hardware platforms. This method first assumes that the noise signal is stable and stable, and uses the uncorrelated characteristics of the additive noise and the speech signal to estimate the power spectrum of the noise signal in the silent section of the noisy speech signal, and use it to replace the true value of the power spectrum of the noise signal. The estimated value of the power spectrum of the clean speech signal can be obtained by subtracting the estimated value of the power spectrum of the noise signal from the true value of the power spectrum of the noisy speech signal. However, spectral subtraction can easily introduce additional “musical noise”, causing severe distortion of the target speech. The construction framework diagram of spectral subtraction is shown in the Figure 1.

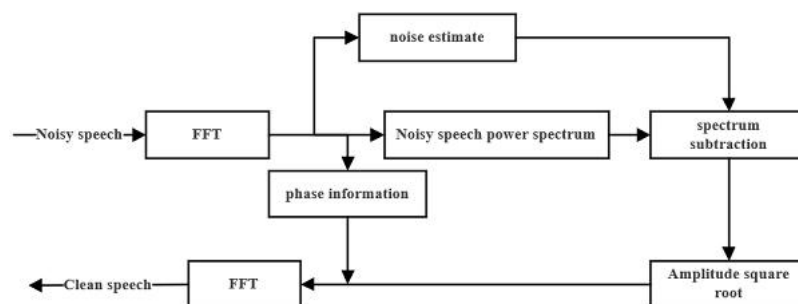


Figure 1. Structural block diagram of spectral subtraction.

It is assumed that the clean speech signal $x(n)$ is a short-term stationary signal and the noise signal $d(n)$ is additive, and the two are independent and uncorrelated. Then the noisy speech signal $y(n)$ can be expressed by the following formula:

$$y(n) = x(n) + d(n)$$

Among them, n represents the time index. Perform Fourier transform on both sides of the above equation to get:

$$Y(\omega) = X(\omega) + D(\omega)$$

Since the clean speech signal and the noise signal have been assumed to be independent and uncorrelated, the power spectrum is calculated on both sides of the equal sign, and we can get:

$$|Y(\omega)|^2 = |X(\omega)|^2 + |D(\omega)|^2$$

Among them, $|Y(\omega)|^2$, $|X(\omega)|^2$, $|D(\omega)|^2$ respectively represent the noisy speech $y(n)$, the clean speech $x(n)$ and the additive noise signal $d(n)$ power spectrum. The power spectrum of the additive noise signal is unknown, but the Voice Activity Detection (VAD) algorithm can be used to estimate the power spectrum of the noise signal in the silent section of the noisy speech signal and use it to replace the additive noise signal. The true value of the power spectrum, so the power spectrum estimate of the clean speech signal can be expressed as:

$$|\hat{X}(\omega)|^2 = |Y(\omega)|^2 - |\hat{D}(\omega)|^2$$

3.2. Wiener filter algorithm

The key to the Wiener filtering algorithm lies in the design of the linear filter $h(n)$. When the noisy speech $y(n)$ passes through the linear filter, the estimate of the original clean speech signal $\hat{x}(n)$ can be obtained, that is, $s(n) = \hat{x}(n) = h(n) * y(n)$. The mean square error of $x(n)$ and $\hat{x}(n)$ can be $E[|\hat{x}(n) - x(n)|^2]$, then $h(n)$ needs to make $E[|\hat{x}(n) - x(n)|^2]$ minimum. (Figure 2)

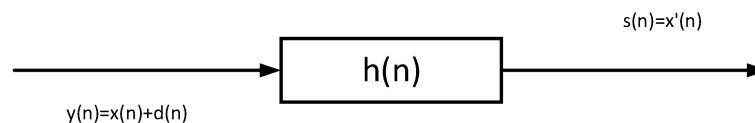


Figure 2. Time domain flow chart of Wiener filter algorithm.

The Wiener filter algorithm and spectral subtraction assume the same conditions for additive noise signals and clean speech signals. At frequency k_ω , we can get:

$$Y(k_\omega) = X(k_\omega) + D(k_\omega)$$

When $X(k_\omega)$ and $D(k_\omega)$ are uncorrelated and the random process is stationary, the above equation can be Fourier transformed to obtain the following equation:

$$\hat{Y}(k_\omega) = \hat{X}(k_\omega) + \hat{D}(k_\omega)$$

Suppose the frequency band response function of the Wiener filter function is $H(k_\omega)$, and the Fourier series transform of the best estimate of the signal $\hat{x}'(k_\omega)$ is $\hat{X}'(k_\omega)$ as follows:

$$\hat{X}'(k_\omega) = H(k_\omega) \cdot \hat{Y}(k_\omega)$$

In complex noise environments, traditional speech enhancement algorithms have the problem that the speech quality after enhancement is average and “music noise” still exists. To address this problem, wavelet packet transform is added to the adaptive filtering algorithm to divide the signal spectrum, and through adaptive The filtering method filters the wavelet packet coefficients of different scales. Experiments have shown that compared with traditional algorithms, this algorithm can not only preserve the spectral characteristics of speech, but also improve the quality of noisy speech [3].

3.3. Speech enhancement algorithm based on model statistics

The core idea of the Wiener filtering algorithm is to minimize the MSE of the estimated clean speech signal spectrum and the real clean speech signal spectrum. It is considered to be the optimal estimator of the complex spectrum, but it is not the optimal spectrum amplitude estimator. For now, the most widely used algorithm is the minimum mean square error estimator based on short-time spectrum amplitude derived by Y. Ephraim et al.

In speech enhancement, it is impossible to determine the distortion criterion in the auditory sense and the prior distribution of the speech spectrum under a given noise condition. Aiming at the specific distortion criterion and the insensitivity to the posterior probability, a condition under the minimum mean square error criterion is proposed. The mean estimation method is used to achieve the optimal estimation of the speech signal. The MMSE-STSA method assumes that both the noisy signal and the noise signal are Gaussian distribution, and uses the known noise power spectrum information to estimate the purer speech signal from the noisy speech signal, so that the estimated value of the short-time spectrum amplitude of the pure speech signal is consistent with that of the pure speech signal. The mean square value of the short-time spectrum of the speech signal is the smallest. According to the method proposed by Ephraim and Malah, the spectral gain can be obtained by calculating the prior signal-to-noise ratio and the posterior signal-to-noise ratio estimate through the estimated value of the noise power spectrum. The pure speech signal can be estimated from the spectral gain and the phase spectrum of the noisy signal. On the basis of MMSE-STSA, a widely used improved method MMSE-STSA is also proposed. This method is to make the estimated value of the logarithmic spectrum amplitude of the pure speech signal closest to the logarithmic spectrum of the pure speech signal under the minimum mean square error criterion. .

The enhanced algorithm for short-time spectrum estimation makes a compromise between noise reduction and improving speech intelligibility. It is applicable to a wide range of signal-to-noise ratios, but it requires a large amount of calculation, and the prior distribution of the speech spectrum is obtained to a large extent. It depends on the representativeness and reproducibility of statistical results, etc., so it has certain limitations in practical applications [4].

3.4. Single-channel speech enhancement algorithm based on deep learning

In traditional single-channel speech enhancement algorithms, certain assumptions are often made about clean speech signals, noise signals, or the probability density associated with them. These assumptions restrict its application in real life scenarios, resulting in a series of Problems, such as the introduction of “musical noise”, causing severe distortion of the estimated clean speech signal. Secondly, the above algorithm needs to accurately estimate the noise signal, and the algorithm that generally uses noise estimation will assume that the change rate of the noise signal is smaller than the change rate of the clean speech signal, that is, the noise signal is relatively stable, so the traditional single-channel speech enhancement algorithm The suppression effect on stationary noise signals is relatively good, but when processing non-stationary noise signals, the noise reduction effect will be significantly reduced. In our daily life, non-stationary noise is more common, so the effect of traditional single-channel speech enhancement algorithms in actual projects will be greatly reduced. In recent years, many researchers have witnessed the rapid development of deep learning algorithms and their practical applications in different industries. With the

enhancement of computer processing power, speech enhancement methods based on deep learning have attracted a large number of researchers to participate [5].

The basic idea of the DNN-based speech enhancement algorithm is to use DNN to conduct big data training on the characteristic parameters of clean speech signals and noise signals, and to continuously modify them through the forward propagation and back propagation (BP) algorithms of the neural network. The correlation parameters between neurons are used to model the mapping relationship between clean speech signals and noisy speech signals to obtain the optimal training model. Compared with traditional single-channel speech enhancement algorithms, DNN-based speech enhancement algorithms can obtain higher-quality enhanced speech, especially in non-stationary noise and low signal-to-noise ratio environments.

The process is mainly divided into two stages, one is the training stage and the other is the enhancement stage. In the training phase, the input speech signal is first framed. At present, most algorithms use the overlapping segmentation method in the framing phase. This method can make the recovered speech signal smoother and make people's voices sound smoother. Very comfortable. The short-time Fourier transform is then used to calculate the discrete Fourier transform coefficients in the speech signal. In the enhancement phase, the extracted features are the same as those used in the training phase. This feature is used as the input of the DNN, and the trained DNN model is used to decode it to obtain the LPS estimate of the clean speech signal, and finally reconstruct the waveform, and finally obtain the speech signal in the time domain [6].

3.5. Fixed beam forming algorithm

Fixed beam forming is one of the most basic algorithms in the field of multi-channel speech enhancement. It has a simple method and low computational complexity. It is generally used in conference call scenarios and has a good suppression effect on incoherent noise. Fixed beamforming is a basic technology for enhancing target speech signals with fixed microphone array weights. Fixed beamformers can be divided into delay-sum beamformers and super-directional beamformers. Among them, the delay summation beamformer is the most commonly used beamformer. It calculates the delay and compensation delay for each signal in the desired direction based on the time delay of each microphone relative to the reference microphone, so that each signal is in Aligned in the time domain, and finally weighted and summed for each channel to obtain the enhanced signal.

The gain of the fixed beamformer is determined by the number of microphones, and the number of microphones is proportional to SNR. Therefore, the fixed beamforming algorithm has great limitations and high cost, and is rarely used alone in practical applications.

3.6. Post-beam algorithm

Regardless of whether fixed beamforming or adaptive beamforming algorithms are used, it is impossible to completely remove background noise or interfering speech. The enhanced target speech will always contain some residual noise. The post-filtering algorithm can make up for the shortcomings of the previous structure and eliminate the residual noise in the enhanced target speech. The post-filtering algorithm is a type of Wiener filtering. The most far-reaching algorithm is the adaptive post-filter proposed by R. Zelinski in 1988. The core idea of this

method is to calculate the relative time delay between each channel. , and then align and add the signals of each channel according to the relative time delay between each channel, sum them up and take the average amplitude as the system output.

The Wiener filter calculates the autocorrelation function by estimating the power spectrum in the frequency domain and performs filtering in the time domain. In order to make the target speech signal and the noise signal uncorrelated with each other, the algorithm requires that the distance between each microphone in the microphone array is relatively large, and the distance between the target speech signal and the microphone is smaller than the distance between each noise signal and the microphone. distance [7].

3.7. Multi-channel speech enhancement algorithm based on deep learning

In the early 21st century, as deep learning was successfully applied in single-channel speech enhancement systems and achieved superior noise reduction effects, many scholars also began to use deep learning in multi-channel speech enhancement tasks. The multi-channel speech enhancement algorithm based on deep learning mainly adopts supervised training. By using a large number of multi-channel speech data sets for training, a non-linear mapping model between noisy speech signals and clean speech signals is obtained. In the enhancement stage, only the noisy speech signal needs to be input into the trained model, and the enhanced target speech signal can be obtained by decoding the model. Since DNN has good processing capabilities for nonlinear data, multi-channel speech enhancement algorithms based on deep learning can not only suppress a variety of different noise signals at the same time, but also effectively avoid speech distortion.

However, this method also has its shortcomings:

(a) It directly uses the multi-channel speech signal collected by the microphone array as the output of the DNN. It requires a very complex network to obtain the mapping relationship between noisy speech and clean speech, and the requirements for the network structure are relatively high; (b) In order to ensure the performance of the deep learning method, its training data set is large and the resulting model is less practical; (c) Training the network takes a long time and the cost of computer resources is high; (d) The network structure is too complex and difficult to implement on hardware devices, resulting in poor real-time performance.

4. Conclusions

Speech detection and enhancement technology is a very important branch in the field of speech signal processing, and its application prospects are also very promising. This article mainly studies excellent thesis works in the field of speech signal enhancement, and puts forward some evaluations based on its own understanding of the basic directions of speech enhancement. In general, single-channel speech enhancement algorithms cannot achieve satisfactory results. At the same time, in order to obtain high quality The voice signal often requires the speaker to be very close to the microphone, which also brings a lot of inconvenience. Multi-channel technology introduces the concept of spatial filtering, taking advantage of using spatial information to better process speech will be the focus of future research on speech signal processing.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this article.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

References

- [1] Xu, Y. Research on speech enhancement methods based on deep neural network. University of Science and Technology of China, 2015.
- [2] Saha, B.; Khan, S., et al. Configurable digital hearing aid system with reduction of noise or speech enhancement using spectral subtraction method and frequency dependent amplification[C]//ON 2018-2018 IEEE Region 10 Conference. JEJU: IEEE Press, 2019, 735-740.
- [3] Chen, J.D.; Benesty, J.; Huang, Y.T, et al. New insights into the noise reduction Wiener filter. *IEEE Transactions on audio, speech, and language processing*, 2006, 14(4), 1218-1234.
- [4] Cohen, I. Relaxed statistical model for speech enhancement and a priori SNR estimation. *IEEE Transactions on Speech and Audio Processing*, 2005, 13(5), 870-881.
- [5] Jan, E.; Flanagan, J. Microphone arrays for speech processing[C]// International Symposium on Signals, Systems and Electronics (ISSSE). *San Francisco: IEEE Press*, 1995, 373-376.
- [6] Y. Ephraim.; H.L. Van, T. A signal subspace approach for speech enhancement", *in IEEE Transactions on Speech and Audio Processing*, 1995, 3(4), 251-266, DOI: 10.1109/89.397090.
- [7] Y. Ephraim. Statistical-model-based speech enhancement systems, *in Proceedings of the IEEE*, 1992, 80(10), 1526-1555, Oct., DOI: 10.1109/5.168664.



© 2023 by the author(s); licensee International Technology and Science Publications (ITS), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (<http://creativecommons.org/licenses/by/4.0/>)