MIJRD Multidisciplinary International Journal of Research and Development

Multidisciplinary International Journal of Research and Development

Volume: 01 | Issue: 03 | 2022 - Open Access - Website: www.mijrd.com - ISSN: 2583-0406

Enhancement of Speaker Recognition from Speech Signal Under Noisy Environment

Muhammad Usman Arshad[1], Zhenhua Tan[2], Muhammad Mursil[3], and Muhammad Shoaib[4]

[1,2,3]Software College, Northeastern University, Shenyang 110169, China

[4]Abdul Wali Khan University, Mardan 23200, Pakistan

Email: [1]usmanarshad5051@outlook.com

Abstract— Speech signal processing has become a challenging area in speaker separation and recognition under noisy conditions. Hereafter, new researchers and areas of development have led speaker identification as a speech processing subfield. The traditional speaker recognition method uses an autocorrelation detection algorithm. When disturbed by the background noise, the detection output signal-to-noise ratio is not high. Thus, an algorithm is proposed based on wavelet speech enhancement and text-related feature extraction. The speaker speech recognition system's overall design is done in a noisy environment by voice noise's feature matching to complete speech signal noise filtering processing. Wavelet adaptive feature decomposition is used to accomplish speech enhancement processing and to extract the relevant features. The extracted is put into the backpropagation (BP) neural network classifier to realize speaker recognition. The simulation results show that the algorithm for speech detection and analysis has high recognition accuracy, low probability of false detection, good performance of noise reduction, and superior indicators.

Keywords— Speech Recognition, Speaker Recognition, Speech Signal Processing, Detection Filer, Noisy Environment.

I. INTRODUCTION

The speech produced by human speech is an important acoustic signal. With the development of signal and information processing technology, signal processing, intelligent analysis, and processing of the speech generated by speech can effectively realize speech recognition and localization of the speaker. According to the target speaker, the speech recognition different technologies are divided into specific person voice recognition and non-specific person voice recognition, analyze the speaker's voice through desktop (PC) voice recognition, telephone voice recognition, and recognition of embedded devices (mobile phones, PDA, etc.) can be effectively used in suspect investigation, voice dialing, voice control intelligence, home services, hotels services, and other various fields. Therefore, the study of speech recognition methods is of great significance [1-3].

The speaker's speech recognition can be divided into isolated word recognition, keyword recognition, and continuous speech recognition according to the different recognition objects. On the basis of signal processing, in extreme environments, the speaker's voice signal collection is interfered with by strong

background noise and the output voice signal. The noise ratio is low, and the recognition is difficult. A reliable speech recognition technology is needed for speaker recognition. The voice signal features are not significant in a noisy environment, and the speech-related components between speakers are strongly coupled. Traditional methods mostly use speech recognition methods based on statistical signal analysis and basic speech recognition algorithm based on time-frequency analysis and the speech recognition algorithm based on wavelet analysis, etc. [4-5], the realization principle is to detect the collected speech signal measurement, filtering, feature extraction, and statistical analysis to realize the matching and recognition of speech and speaker. Literature [6] proposed a voice classification and recognition algorithm based on wavelet packet decomposition for feature extraction of ship radiated noise, using continuous voice recognition and keyword detection combined recognition method to achieve matching filter detection of voice feature points and improve recognition accuracy. However, the algorithm is not good for speech recognition in a strong noise environment, and its anti-interference ability is not strong.

Literature [7] proposed a speaker speech recognition algorithm based on sparse representation to achieve speaker recognition and speech detection; Speech signal is decomposed by wavelet multi-scale, combined with feature matching algorithm to realize speech enhancement, improve the stability and completeness of speech output, and improve recognition accuracy.

However, the algorithm has a high computational cost and does not perform well in real-time speech recognition; [8] Propose a voice endpoint detection method in a complex noise environment, using an autocorrelation detection algorithm, when it is interfered with by large background noise, the signal-to-noise ratio of the detection output is not high. The convergence in the classification and recognition process is not good.

Therefore, this paper proposes an algorithm for speech enhancement and text-related feature extraction based on wavelet enhancement to realize the speech recognition of the speaker.

Carry out noise reduction filter preprocessing, adopt wavelet adaptive feature decomposition, carry out speech enhancement processing, carry out text-related feature extraction on the enhanced speech signal, and input this feature into the BP neural network classifier to realize speaker recognition.

Finally, the performance test is carried out through simulation experiments to show the superior performance of the algorithm in this paper, an accurate speech recognition, and a valid conclusion is drawn.

II. SIGNAL NOISE REDUCTION FILTER PREPROCESSING IN NOISY ENVIRONMENT

A. Design of the Speech Recognition System

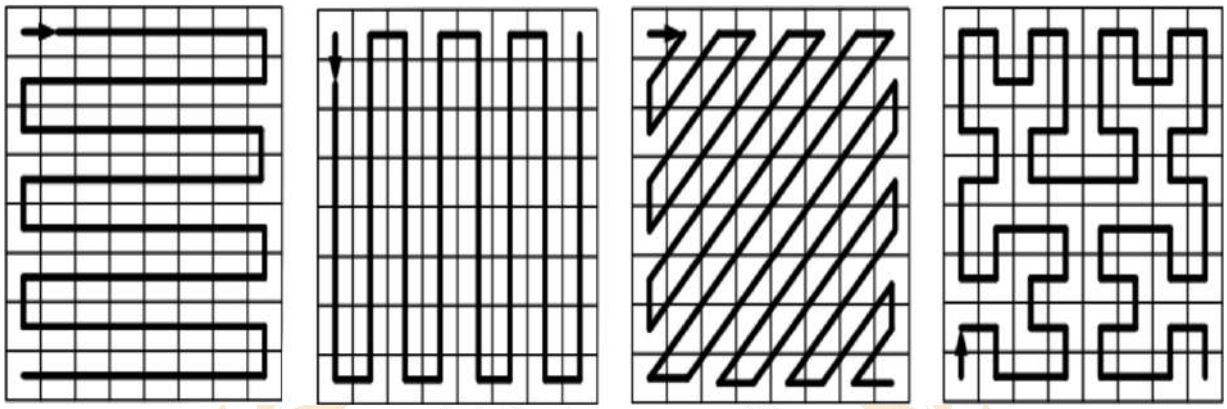
The voice signal is affected by various factors such as the environment and the transmission medium during the transmission process, resulting in the lack of obvious characteristics of the voice signal in the noisy environment.

MIJRD Multidisciplinary International Journal of Research and Development

Multidisciplinary International Journal of Research and Development

Volume: 01 | Issue: 03 | 2022 - Open Access - Website: www.mijrd.com - ISSN: 2583-0406

It is necessary to scan the voice of the text speaker in the noisy environment to collect and recognize the voice signal and use the speaker scale invariance, feature invariance, and text-related invariance of speech are characterized by keyword scanning and audio scanning. Different methods such as horizontal, vertical, oblique, and block are used to scan the speaker's voice features and signal acquisition [9 -12], as shown in Figure 1.



(a) Horizontal scan *(b) Vertical scan* *(c) Oblique scan* *(d) Block scan*

Figure 1: Speaker's speech scan acquisition model in noise environment

Based on the speaker's voice feature scanning and signal collection, for the speaker, it is assumed that the noisy model of the text speaker's voice signal in a noisy environment is f (x, y), and several known keywords are detected under background interference.

The speaker's audio characteristics are matched, assuming that the radial velocity of voice transmission is v, the initial distance between the voice acquisition system and the speaker is R0, and the broadband model for obtaining the speaker's voice echo in a noisy environment is:

$$w(t)=n(t)\times h\_w(t). \quad (1)$$

Among them: n(t) represents the background reverberation noise where the speaker is located, and h_w (t) represents the reflected echo of the speech equipment and channel.

According to the propagation loss of the voice equipment and the channel transmission medium, the strong correlation and non-stationarity of the reverberation and the signal are considered, and the speech enhancement is performed.

In the context of reverberation, in the wideband of the speaker's voice echo, the receiver distance R of the voice collection is obtained as the envelope amplitude function of time t:

$$R(t)=(R\_0-vt)wt. \quad (2)$$

The Wiener filter is used to analyze the envelope amplitude of the voice collection. Under the interference of fading noise, by eliminating the influence of environmental noise on the voice, the received voice signal is:

$$\gamma(t)=g(t)+\eta(t). \quad (3)$$

Equation (3) shows that the acoustic features have instantaneous envelope attenuation in the time domain and frequency domain. Among them, g(t) is the echo information of voice control, and η(t) is the interference signal.

Using noise and reverberation matching method, when measuring the information attenuation coefficient λ of voice control in a noisy environment, f(t) matches the speaker's target characteristics through the transmission channel and determines the starting point g(t) of the speech signal, which is:

$$g(t)=bf(t-\tau(t)). \quad (4)$$

Among them, b is the reflection gain of the starting point of the speech signal, which is related to the characteristics of the speaker's target and the channel loss of the speech transmission medium.

The BP neural network classifier is used to achieve speaker recognition by extracting statistical features. According to the above analysis, the overall design block diagram of the speaker's speech recognition system in a noisy environment is obtained, as shown in Figure 2.

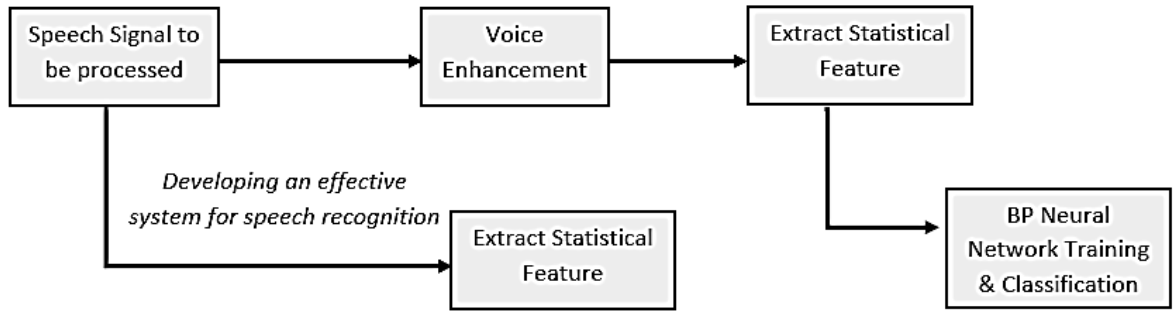


Figure 2: Design diagram of Speaker speech recognition system in noisy environment

B. Speech Signal Noise Reduction Filter Processing

In the signal duration, auto-correlation matched filter detection is used for noise reduction filtering processing and voice signal noise reduction processing, and the detection feature matching point g (x, y) of the voice signal is obtained.

Using the local detection window to perform feature point matching detection on g(x, y), the speaker is interfered by the additive noise item η(x, y) during the speech collection process, and the average energy of the speaker's speech detection window is examined to obtain the original speech An estimate of human speech feature band f`(x,y).

According to the extraction results of the acoustic features, the adaptive noise cancellation algorithm is used for the noise reduction filtering of the audio signal [13-15], and the structure block diagram of the speech signal noise reduction filtering is obtained, as shown in Figure 3.

Multidisciplinary International Journal of Research and Development

Volume: 01 | Issue: 03 | 2022 - Open Access - Website: www.mijrd.com - ISSN: 2583-0406

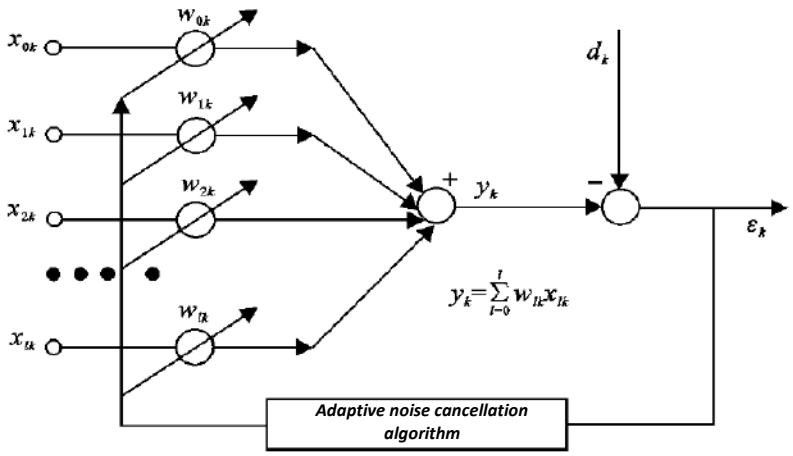


Figure 3: Structure block diagram of speech signal noise reduction filter

Combine Figure 3 to design the noise reduction filter algorithm. Firstly, the estimated value of filtering on a short speech signal is given as $\hat{f}(x,y) = \beta F(x, y) + (1-\beta)\, m_l$, where F(x, y) is the speaker's voice in a noisy environment Scan the time scale value at point (x, y); m_l is the embedding dimension of the filter, let δ_l^2 be the local variance of the speaker's voice scan; δ_η^2 is the noise variance; $\beta = max\left[\frac{\delta_l^2-\delta_\eta^2}{\delta_l^2}, 0\right]$ represents the acoustic feature.

The feedback coefficient of the line spectrum is detected by autocorrelation matched filter according to the acoustic characteristic line spectrum, and the noise variance δ_η^2 is obtained, and the transfer function form of the autocorrelation matched filter is obtained as:

$$\hat{f}(x,=\begin{cases} g(x,y)-1, if g(x,y)-\hat{f}_{Le}(x,y) \geqslant t, \\ g(x,y)+1, if g(x,y)-\hat{f}_{Lee}(x,y) < t, \\ g(x,y), else. \end{cases} \quad (5)$$

By obtaining the transfer function of the autocorrelation matched filter, using adaptive local noise reduction, the continuous noise is decomposed into words, phonemes, and other units, and the feature matching of speech noise:

$$g(x,y) = \hat{f}(x,y) + \eta_m(x,y) + f(x,y) \quad (6)$$

Among them, $f(x,y)$ represents the input voice signal; $g(x,y)$ represents the output voice signal of the noise reduction filter; $\eta_m(x,y)$ represents the noise between the sound line feature points of the voice signal, and the additive noise satisfies $\eta_m(x,y) \in \{-1,0,1\}$. Through the feature matching of speech noise, the noise reduction filter processing of speech signal is completed.

III. ALGORITHMS

Speech Enhancement Algorithm Based on Wavelet Analysis

MIJRD Multidisciplinary International Journal of Research and Development

Multidisciplinary International Journal of Research and Development

Volume: 01 | Issue: 03 | 2022 - Open Access - Website: www.mijrd.com - ISSN: 2583-0406

The traditional speaker speech recognition method uses an autocorrelation detection algorithm; when it is interfered with by large background noise, the signal-to-noise ratio of the detection output is not high.

To overcome the shortcomings of traditional methods, this paper proposes a speech enhancement and text-related feature extraction algorithm based on wavelet analysis to realize the speaker's speech recognition. Under the complex noise background, the speaker's speech signal analysis and feature extraction are used to obtain the speech. The discrete signal of human speech is x(n), and the orthogonal wavelet transform is constructed, which is :

$$p(\eta_m(x,y)) = \begin{cases} r/4, & \eta_m(x,y) = -1 \\ 1 - r/2, & \eta_m(x,y) = 0 \\ r/4, \eta_m(x,y) = 1 & \end{cases} \quad (7)$$

Among them: r is the speaker's speech word and phoneme resolution scale in a noisy environment, 0≤r≤1. The noise is additive noise, the mathematical expectation is 0, and the variance is r/2. On this basis, the hyperbolic FM wavelet of speaker speech feature decomposition is defined as:

$$\Phi^H(t) = A(t)\, exp[j\theta(t)]\, p(\eta_m(x,y)) = A(t)\, exp\left[-j2\pi\, Kln\left(1 - \frac{t}{t_0}\right)\right] p(\eta_m(x,y)), |t| \leqslant \frac{T}{2} \quad (8)$$

After obtaining the hyperbolic frequency modulation wavelet decomposed by the speaker's speech feature, the speech signal is discretely processed. Under the j-th decomposition scale, the high-frequency component coefficients of noise and reverberation at time k are obtained as $d_{j,k}$; speech in wavelet domain space, the attenuation parameter coefficient is $a_{j,k}$; the voice output frequency is fs.

Under different attenuation of the transmission medium, the energy of the speech feature detail signal of j = 0, 1,..., M is obtained as $E_j = \sum_k |C_j(k)|^2$, where the wavelet coefficient is $C_j(k) = [x(t), \phi_{j,k}(t)]$, the discrete processing result of the speech signal obtained by the wavelet scale projection method is :

$$E = \| x(t) \|^2 g(t) = \sum_j \sum_k |C_j(k)|^2 g(t) = \sum_j E_j g(t). \quad (9)$$

The total energy of the speech signal can be expressed as:

$$g(t) = \frac{1}{\sqrt{a_0 k}} f\left(\frac{t_0 - \tau_0}{a_0}\right) \quad (10)$$

Among them: $t_0 = f_0 T/B, K = T f_i(-T/2) f_i(T/2)/B$ the wavelet frequency center is f_0; the normalized relative wavelet energy Pj = Ej/E. Through discrete processing of speech signals, a speech enhancement algorithm based on wavelet analysis is realized.

IV. UNITS

Text-Related Feature Extraction

Assuming that the speaker's speech wavelet energy set {P1, P2,..., Pj} in the wavelet time-frequency space covers the entire frequency band of the speech signal, the two-dimensional wavelet time-frequency characteristic of the speaker's speech feature output is:

Multidisciplinary International Journal of Research and Development

Volume: 01 | Issue: 03 | 2022 - Open Access - Website: www.mijrd.com - ISSN: 2583-0406

$$W_{\Phi^H}\Phi^H(a,\tau) = \int \Phi^H(t)\Phi_a^{H*}(t-\tau)dt \quad (11)$$

Perform text-related feature scanning on the sub-signal Pj of voice features in each voice time period, decompose the voice signal of each voice time period into the wavelet feature scale Cj(k) of the j-th layer coefficient, and output the wavelet for the voice feature. The time-frequency two-dimensional characteristics are scanned, and the voice signal is divided into n cells by the sound ray equalization method; then there are:

$$E_{j,k} = \sum_{k}^{\frac{m}{n}} \left|C_j(k)\right|^2 W_{\Phi}^{H}\Phi^H(a,\tau) \quad (12)$$

By extracting the wavelet energy Ej,k of the output speech signal with noise reduction, calculate Ej and the wavelet energy Pj,k, where,

$$E_j = \sum_{k=1}^{n} E_{j,k} \quad (13)$$

$$P_{j,k} = \frac{E_{j,k}}{E_j} \quad (14)$$

In the process of calculating wavelet energy Ej,k, the energy loss in the k interval can be obtained by WEk, which is defined as:

$$WE_k = -\sum_{j} P_{j,k}\, ln(P_{j,k}) \quad (15)$$

By calculating the positive wavelet energy loss of the speech signal, the wavelet energy Ej and the wavelet energy Pj,k of the speech signal, the text-related features of the enhanced speech signal are extracted, and finally, the BP neural network classifier is used for feature classification to realize the speaker The structure of the BP neural network classifier is shown in Figure 4.

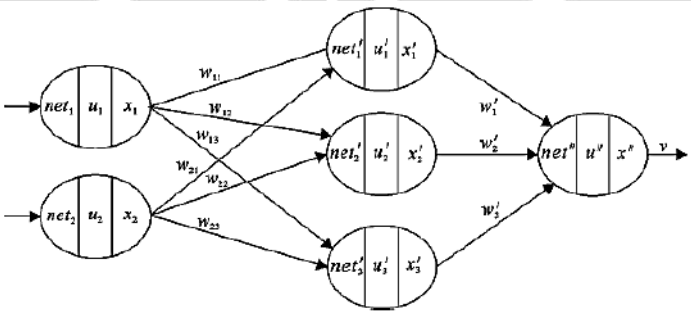


Figure 4: Neural network classifier

The BP neural network classifier uses three layers of forwarding neurons in a 2×3×1 structure, which are the input layer, middle layer, and output layer. Through the reliability recognition of the speech signal speaker, the control of the speaker's speech recognition is realized, the control function for speech recognition is:

Multidisciplinary International Journal of Research and Development

Volume: 01 | Issue: 03 | 2022 - Open Access - Website: www.mijrd.com - ISSN: 2583-0406

$$\begin{cases} \ddot{\varphi}_a = -(b_1 + \Delta b_1)\dot{\varphi}_a - (b_2 + \Delta b_2)\varphi_a - (b_3 + \Delta b_3)\delta_{\bar{\varphi}} + fd_1 \\ \ddot{\psi}_a = -(b_1 + \Delta b_1)\dot{\psi}_a - (b_2 + \Delta b_2)\psi_a - (b_3 + \Delta b_3)\delta_{\psi} + fd_2 \\ y = -(d_3 + \Delta d_3)\delta_{\gamma} + fd_3 + w_{1j} + w_{2j} + \cdots + w_{nj} \end{cases} \quad (16)$$

Among them: ¨γ is the text-related features of the input speaker in the input layer; w1j, w2j,...,wnj are the weight values; b1, b2, b3 are the input layer, the middle layer, and the output layer respectively; Δb1, Δb2, and Δb3 are the input respectively The error value of the speaker's speech recognition of the layer, the middle layer, and the output layer; φ_a is the correlation coefficient of speaker recognition; Ψ¨a is the classification recognition result obtained in the output layer; d1, d2, d3 are the input layer, middle layer, and The time interval of the output layer. By getting the control function of speech recognition, the improvement of the speaker speech recognition algorithm is completed.

IV. SIMULATION

The simulation is based on the Matlab Simulink software. The voice information collection parameters of text-related speakers in a noisy environment: Np = Ns = 20, f = 20.5 kHz, Cp = Cs = 0.57 μF. The bandwidth of the speaker's voice collection is 5kHz, the attenuation frequency of the voice transmission is 50 000Hz/s, the time width is T = 0.1s, and the intensity of noise interference is -12dB. The result of the noisy voice signal sampling is shown in Figure 5.

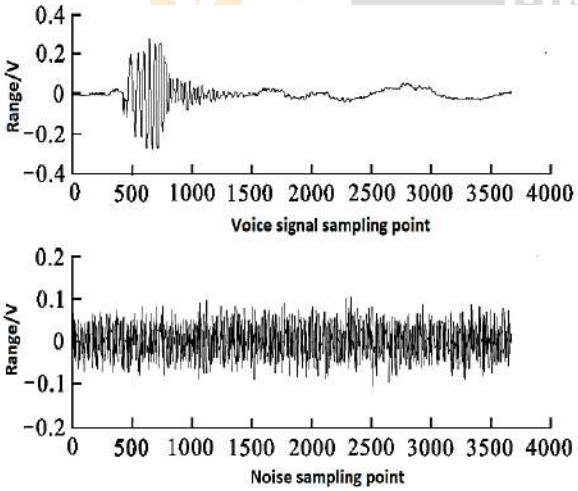


Figure 5: Speech signal sampling results and noise.

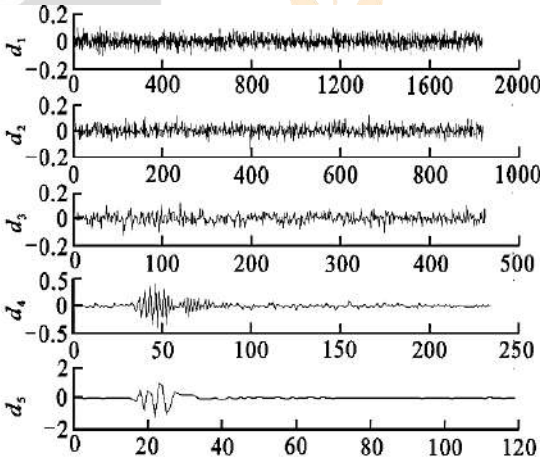


Figure 6: Wavelet enhancement of speech signal on the wavelet scale of each layer

Taking the above-sampled speech signal as the research object, the noise-containing speech signal collected is pre-processed by noise reduction filtering, and the speech enhancement processing is performed through wavelet adaptive feature decomposition. The wavelet enhancement result of the speech signal is obtained, as shown in Figure 6. Perform text-related feature extraction on the enhanced speech signal, and use this as a feature to input into the BP neural network classifier to realize speaker recognition. The results of speech recognition using this method and traditional methods are obtained, as shown in Figure 7. It can

Volume: 01 | Issue: 03 | 2022 - Open Access - Website: www.mijrd.com - ISSN: 2583-0406

be seen from Figure 7 that the method used in this paper for speaker speech recognition has better noise suppression performance. In order to quantitatively analyze the performance of the algorithm, the fifth issue of Tan Ping, et al.: The text-related speaker recognition method in a noisy environment improves the 643 recognition the output signal-to-noise ratio SNR is the test index, and the comparison result of the speaker's speech recognition performance is obtained, as shown in Figure 8. It can be seen from Figure 8 that the method used in this paper for speaker speech recognition has higher recognition accuracy, lower probability of false detection, and better speech noise reduction performance.

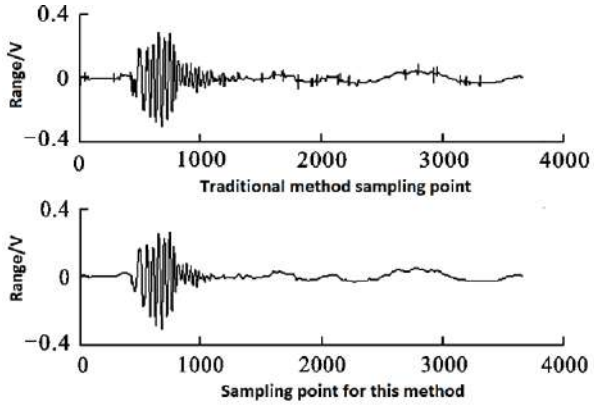


Figure 7: Speech recognition results

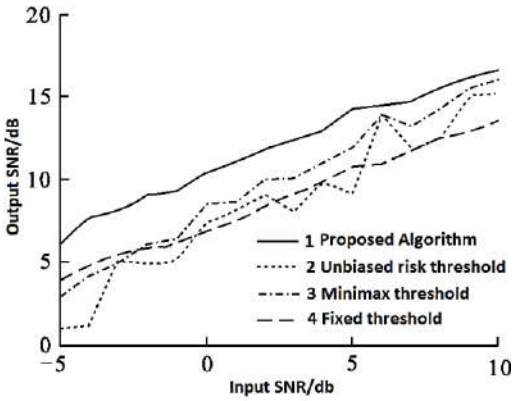


Figure 8: Speaker speech recognition *performance* and comparison

IV. CONCLUSION

Speaker's speech recognition can be divided into isolated word recognition, keyword recognition, and continuous speech recognition based on different recognition objects. Speech recognition is based on signal processing. In extreme environments, the speaker's voice signal acquisition is interfered with by strong background noise, and the output voice signal has a low signal-to-noise ratio, making it difficult to recognize. Therefore, an algorithm based on wavelet speech enhancement and text-related feature extraction is proposed to recognize the speaker's speech. First, scan the speech of the text speaker in a noisy environment and build a collection model, perform denoising filter preprocessing on the collected noisy speech signal, perform speech enhancement processing through wavelet adaptive feature decomposition, and perform text-related features on the enhanced speech signal Extract the features and input them into the BP neural network classifier to realize speaker recognition. The research shows that the use of the speaker speech recognition algorithm for speech detection and analysis has higher recognition accuracy, lower probability of false detection, and better speech noise reduction performance.

REFERENCES

[1] M. Hamza, T. Khodadadi, and S. Palaniappan, "A novel automatic voice recognition system based on text-independent in a noisy environment," Int. J. Electr. Comput. Eng., vol. 10, no. 4, pp. 3643–3650, 2020, doi: 10.11591/ijece.v10i4.pp3643-3650.

[2] K. J. Devi and K. Thongam, "Automatic speaker recognition with enhanced swallow swarm optimization and ensemble classification model from speech signals," J. Ambient Intell. Humaniz. Comput., 2019, doi: 10.1007/s12652-019-01414-y

[3] Javeed, Danish, Tianhan Gao, and Muhammad Taimoor Khan. "SDN-Enabled Hybrid DL-Driven Framework for the Detection of Emerging Cyber Threats in IoT." Electronics 10.8 (2021): 918.

[4] R. Peri, M. Pal, A. Jati, K. Somandepalli and S. Narayanan, "Robust Speaker Recognition Using Unsupervised Adversarial Invariance," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6614-6618, doi: 10.1109/ICASSP40776.2020.9054601.

[5] N. H. Tandel, H. B. Prajapati and V. K. Dabhi, "Voice Recognition and Voice Comparison using Machine Learning Techniques: A Survey," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 459-465, doi: 10.1109/ICACCS48705.2020.9074184.

[6] D. Cai, W. Cai and M. Li, "Within-Sample Variability-Invariant Loss for Robust Speaker Recognition Under Noisy Environments," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6469-6473, doi: 10.1109/ICASSP40776.2020.9053407.

[7] Javeed, D., MohammedBadamasi, U., Ndubuisi, C. O., Soomro, F., & Asif, M. (2020). Man in the Middle Attacks: Analysis, Motivation and Prevention. International Journal of Computer Networks and Communications Security, 8(7), 52-58.

[8] LU Yuanyao, ZHOU Ni, XIAO Ke, et al. Improved speech endpoint detection algorithm in strong noise environment[J]. Journal of Computer Applications, 2014, 34(5): 1386-1390.

[9] ZHANG Yong, GONG Dunwei, HU Ying, et al. A PSO-Based multi-Robot search method for odor source in indoor environment. Online [twenty] Acta "Electronica" Sinica, 2014, 42(1): 70-76.

[10] JIN Yan, DUAN Pingting, JI Hongbing. Paradise [J]. Journal of Electronics &Information Technology, 2014, 36(5): 1106-1112.

[11] MA Jinquan, GE Lingong, TONG Li. New time delay estimation algorithm of HF fading signal in symmetric stable distribution noise environments. Journal" of "Signal" Procedure, 2014, 30(5): 526-534.

[12] Javeed, Danish, et al. "A Hybrid Deep Learning-Driven SDN Enabled Mechanism for Secure Communication in Internet of Things (IoT)." Sensors 21.14 (2021): 4884.

[13] PEI Xiaozhong, ZHENG Tieran, Han Jiqing. An acoustic feature extraction approach based on frequency selectivity of human auditory under driving noisy environments [J]. Intelligent Computer and Applications, 2015, 5(3): 16-18.

[14] LEI Y, SCHEFFFER N, FERRER L, et al. A novel scheme for speaker recognition using a phonetically-aware deep neural network [C] // Proceedings-ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, 2014, 25(6): 1695-1699.

[15] MA Jinquan, GE Lingong, TONG Li. New time delay estimation algorithm of HF fading signal in symmetric stable distribution noise environments. Journal" of "Signal" Procedure, 2014, 30(5): 526-534.

[16] Khan, Tahir Ullah. "Internet of Things (IOT) systems and its security challenges." International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 8.12 (2019).

