International Journal of Speech Technology (2018) 21:563–569
https://doi.org/10.1007/s10772-018-9516-7

Study on processing of wavelet speech denoising in speech recognition system

Xinmei Zhong[1] · Yunzhong Dai[1] · Yong Dai[2] · Tao Jin[1]

Received: 21 January 2018 / Accepted: 25 April 2018 / Published online: 8 May 2018

Abstract
The development of society promotes the continuous progress of science and technology, and speech processing technology gradually occupies an increasingly important position in people's life and work, which puts forward higher requirements on the speech processing technology, especially in noisy environment. Due to the complexity of the real environment, denoising processing has great practical significance. In order to improve the level of speech denoising and increase the accuracy of the speech recognition system, wavelet denoising technology was used to analyze the de-noising requirements and hard and soft threshold functions in the speech recognition system, and an improved wavelet threshold denoising algorithm was put forward. Firstly, the signals were processed by wavelet decomposition according to primary function; then denoising was performed using the improved function; finally the denoised signals were reconstructed using inverse operation. The denoising effect of the algorithm was verified. The results showed that it was effective in denoising conventional speech signals. Besides, it was applied to the speech recognition system to denoise the noisy speech collected in the real environment, and finally high system self-assessment parameters were obtained. Thus it is concluded that wavelet denoising is effective in the speech denoising of the speech recognition system and can be put into practice.

Keywords Speech recognition · Wavelet denoising · Wavelet threshold

1 Introduction

Communication is an indispensable activity in people's life. However, noise interference becomes more and more severe with the development of the society, which can reduce the quality of speech information. In order to reduce the noise interference and obtain the speech information more accurately and efficiently, the speech recognition system is developed, which translates speech information into written information through machines and realizes human speech instructions (Lecouteux et al. 2016). Speech recognition system has a strong recognition capability in a quiet environment. But the noise environment can affect the performance of speech recognition system, reduce its recognition rate, and the noise can even mask speech information, which makes it impossible to identify speech information (Alissali et al. 2017; Sui et al. 2015). Therefore, many researchers have conducted a number of studies to reduce the impact of noise on speech recognition systems. Aicha and Jebara (2012) used the single-channel enhancement technique of the short-term spectrum amplitude to detect and restore the spectral peak with music characteristics. It was found that music peak could be detected when the masking threshold was exceeded. Mak and Yu (2014) used speech enhancement technology to study the characteristics of SRES and the difficulty of speech segmentation of interview audio files in NIST. It was found that the low signal-to-noise ratio of VAD based on energy was the life of noise suppression. The denoising of speech recognition system is generally realized by carrying out frequency domain filtering processing on signals based on Fourier changes, which has limitations (Gesell et al. 2009). Wavelet transform was born at the right moment. Wavelet transform as a simple and efficient signal analysis technology can greatly improve the recognition rate of speech signals (Srivastava et al. 2016). The current wavelet denoising methods include modulus maximum denoising, shielding denoising, wavelet threshold denosing and

✉ Xinmei Zhong
xinmzhongsw@163.com

[1] Southwest Petroleum University, Chengdu 610500, Sichuan, China

[2] Weatherford(Sichuan) Petrochemical Equipment Co., Ltd., Chengdu Branch, Chengdu 610500, Sichuan, China

translation invariant (Sun and Feng 2016). Wavelet threshold denosing has been extensively applied in various engineering for its convenience (Kumar and Agarwal 2015). Based on the wavelet denoising technique, this study proposed an improved wavelet threshold algorithm to improve the recognition of speech signals. This work lays a solid basis for the application of speech recognition.

2 Speech recognition system

Speech recognition system is an application system as well as a pattern recognition system built on certain hardware and software platform which includes feature extraction, pattern matching and reference pattern library (Torres-Carrasquillo et al. 2010). Due to the complex structure of human voice information and its very comprehensive content contained, the identification of speech information is far more difficult than ordinary pattern recognition (Bauer et al. 2001). During operation, a matching algorithm is selected according to the type of the recognition system, the relevant characteristic parameters are extracted by the relevant speech signal processing method, and the related learning, training and recognition are performed (Shanthi and Lingam 2015). In the process of identification, the characteristic parameters are compared with the templates generated in the training process, and the template closest to the speech signal is selected if the error is within the allowable error. According to the definition of the template, the corresponding computer identification result is found in the table (Rajam and Balakrishnan 2012).

3 Improved wavelet threshold algorithm

The traditional denoising method aims to remove the noisy signals with filter after Fourier transform. It has good performance in denoising stationary audio in quiet environment, but is difficult to denoise audio in real environment, especially when other secondary signals or similar frequency noises are contained. For such kind of audio, Fourier transform is difficult to work effectively. The speech mentioned here is non-stationary signal.

Wavelet transform which is significantly different with Fourier transform can be used for processing non-stationary signals. There are four kinds of wavelet denoising. The first one is modulus maximum. The characteristics of signals are reflected using the variation speed of the local maximum value of signals under different scales after wavelet transform. Generally the modulus maximum scale of effective signals is in direct proportion to the scale, and that of noise is inversely proportional to the scale. It is suitable for denoising of signals with multiple singular points. But the computation is slow during signal reconstruction after denoising, and scale selection for wavelet decomposition has large impact on denoising effect. The second method is shielding denosing which specifies denosing means according to the difference of wavelet coefficients of effective signals and noises at different scales. It is effective and stable in denoising and suitable for signals with high signal-to-noise ratio. But the computation load is heavy, and moreover the variance of noise signals needs to be calculated. The third method is wavelet threshold denoising. A proper value was taken from each scale as threshold. When the coefficient is smaller than the threshold, the coefficient was attributed to noise and abandoned; otherwise it is retained or collapse towards zero. Signals were reconstructed after the new coefficient is obtained. The method has gained the most extensive application because of the high calculation speed and complete signal characteristics, but its denoising effect is closely correlated to signal-to-noise ratio, and the selection of threshold will also produce decisive impacts on results. The last one is translation invariant. It is suitable for signals which are mixed with discontinuous signals. It can effectively prevent pseudo-Gibbs phenomena and improve signal-to-noise ratio. But the calculation is slow, especially in processing long signals.

Wavelet threshold denoising method was selected in this study, and moreover an improved method was put forward.

3.1 Wavelet threshold denoising algorithm

Wavelet threshold denoising algorithm is an important part of wavelet algorithm. The specific algorithm (Li and Zhenxing 2009) is as follows.

$$f(v) = s(v) + n(v) \tag{1}$$

Equation (1) is a noisy one-dimensional signal model, where $f(v)$ is the noisy signal, $s(v)$ is the useful signal and $n(v)$ is noise which complies with the Gaussian white noise of $N(0, \sigma^2)$ distribution.

When the algorithm is used, appropriate values of wavelet base and wavelet transform layer T should be selected firstly. Wavelet decomposition is carried out on $f(v)$ to obtain its average part and detail part. Then, the appropriate threshold value and threshold function are selected to deal with the threshold value of each layer wavelet coefficients, and the estimated wavelet coefficients of each layer are obtained. Afterwards, the scale factor and the wavelet coefficients of all layers were reconstructed to obtain the estimated value of the original useful signal $s(v)$.

Due to the wavelet transform basis, the choice of wavelet is not fixed and there is no wavelet basis function that can have a good de-noising effect on all types of signals. Hence, the quality of wavelet function is determined judged by the error between speech signal processing results of the wavelet

analysis method and the theoretical results; wavelet basis function is selected based on it. The selection of threshold value is an important reflection of the accuracy of threshold size estimation. If the estimated threshold value is too small, the noise in the signals cannot be totally removed; if the estimated threshold value is too large, useful speech information can be filtered, leading to the decline of the authenticity of the signals. The common threshold value selection methods include unbiased risk estimate, fixed threshold estimation, heuristic threshold estimation and extremum threshold estimation.

3.2 Threshold function

The traditional threshold function can be divided into hard threshold function and soft threshold function. The calculation formula of the hard threshold function is:

$$\hat{\omega}_{m,n} = \begin{cases} \omega_{m,n} & |\omega_{m,n}| \geq \mu \\ 0 & |\omega_{m,n}| < \mu \end{cases}, \tag{2}$$

where $\hat{\omega}_{m,n}$ is the estimated wavelet coefficient, $\omega_{m,n}$ is the wavelet decomposition coefficient of the original signal, and μ is the removed-noise threshold. The hard threshold function contrasts the wavelet decomposition coefficient of the noisy signals at different scales with the threshold value. If the wavelet decomposition coefficient is smaller than the threshold value, it is directly set to 0, and if it is larger than the threshold value, the point remains unchanged. The formula of soft threshold function is

$$\hat{\omega}_{m,n} = \begin{cases} \operatorname{sgn}(\omega_{m,n})\left(|\omega_{m,n}| - \mu\right) & |\omega_{m,n}| \geq \mu \\ 0 & |\omega_{m,n}| \leq \mu \end{cases}. \tag{3}$$

Equation (3) is applied to eliminate the effect of the discontinuity of the function on denoising. The wavelet decomposition coefficient of noise signal on different scales is compared with the threshold value. If the wavelet decomposition coefficient is smaller than the threshold value, it is directly set as 0, and if the threshold value is greater than the threshold value, the point remains unchanged.

In order to improve the recognition rate of the speech recognition system, the useful signals in the frequency range of the human body requires maximum retention. However, as can be seen from the above two formulas, the hard threshold function retains more useful signal features, but there is still a small loss of signals, which is also found in the soft threshold function. Therefore, this study put forward an improved threshold function.

3.3 Improvement of threshold function

Generally speaking, the value of the threshold decreases with the increase of the degree of decomposition, but the threshold function itself does not have the specific adjustment for speech signal. Therefore, if the threshold function is closer to the hard threshold function on a high scale, then the wavelet coefficients processed can be closer to the real coefficient, and the denoising effect can be further improved. Based on the original hard threshold function and soft threshold function, this paper adjusts the decomposition scale to promote the improvement of the whole threshold function.

$$\hat{\omega}_{m,n} = \begin{cases} \operatorname{gn}(\omega_{m,n})\left(|\omega_{m,n}|^m - \mu^m\right)^{\frac{1}{m}} & |\omega_{m,n}| \geq \mu \\ 0 & |\omega_{m,n}| < \mu \end{cases}, \tag{4}$$

where m is the number of layers of wavelet decomposition. If $n = 1$, then the function value is equivalent to that of the soft threshold function. When m tends to be positive infinity, then

$$\lim_{x\to\infty}\left[\operatorname{sgn}(\omega_{m,n})\left(|\omega_{m,n}|^m - \mu^m\right)^{\frac{1}{m}}\right] = \omega_{m,n}. \tag{5}$$

In this equation, the middle and low frequency bands of the threshold value is slowly approaching the hard threshold function when decomposition scale m increases, which is to say, $\hat{\omega}_{m,n}$ is gradually approaching $\omega_{m,n}$ and useful signals in speech signal are retained to the maximum extent. Besides, the continuity on μ is kept and concussion noise is avoided. However, as m is not a fixed value in actual situations, the improvement function's denoising effect cannot be well reflected. Therefore, the regulatory factor ρ is introduced.

$$\hat{\omega}_{m,n} = \begin{cases} \operatorname{sgn}(\omega_{m,n})\left(|\omega_{m,n}|^{\rho+m} - (\mu)^{\rho+m}\right)^{\frac{1}{\rho+m}} & |\omega_{m,n}| \geq \mu \\ 0 & |\omega_{m,n}| < \mu \end{cases}, \tag{6}$$

where $\rho \in \{1, 2, 3...\}$.

In the case of small decomposition scale, the negative effect of the reduction of denoising effect can be prevented by the adjustment factor. If $\omega_{\mathrm{m,n}} > 0$, then $\mu = 0.5$, $\rho = 1$, $m \in \{1, 2, ..., n\}$.

After the improvement, the continuity of the curve of the new function is ensured on μ. When the constant deviation between $\hat{\omega}_{m,n}$ and $\omega_{m,n}$ reduces with the increase of m, the middle and low-frequency voice information is well preserved and the reconstructed signal is closer to the real speech signal.

4 Establishment of speech recognition system

4.1 Speech acquisition module

Speech acquisition was performed in WM8731 on DE2 development board. I^2C bus was controlled, and the data

of working mode was allocated. After A/D conversion, the sampling frequency and bit wide of the input speech were set as 15 KH and 16. The microphone at the input end of the DE2 board sampled the input noisy speech at A/D mode and 16 kHz.

4.2 Register configuration module

WM8731 is a low-power and high-quality audio decoder that adjusts the volume at 1.5 db steps within the step range between + 12 and − 34.5. Through the high quality structure digital analog converter and the high-pass digital filter, the dc component in the input signal could be removed after A/D conversion. After powering on the chip, the internal register of the chip was configured, and the working mode of the chip was set. The register configuration data width was 16 bits, including the register's address and data. The audio data format was set as left-aligned.

5 Simulation experiments

Wavelet decomposition and coefficient reconstruction and processing were performed on speech signals using MATLAB. Moreover Blocks signals and Doppler signals were denoised. Firstly a segment of Blocks signal and Doppler signal generated respectively. Then Gaussian white noise was added to the signals. Next the noisy signals were denoised using wavelet threshold denoising. The denoising functions used included hard threshold function, software function and the improved function. db3 wavelet basis was selected as the basic function for wavelet decomposition of the noisy Blocks signals. The decomposition had five layers. The regulatory factor ρ of the improved function was set as 2, and the value of threshold μ was determined by the default option of MATLAB. sym8 wavelet basis was selected as the basic function for the wavelet decomposition of the noisy Doppler signals, and the other operations was the same with Blocks signal.

Then a speech (PCM, 8 bits, dual track and 55 kHz) transcribed in a job fair was denoised. Firstly abnormal Gaussian white noise was added into the speech. Then wavelet decomposition was performed taking db3 as the basic function. The decomposition had six layers. The regulatory factor was set as 1. The thresholds were the same and both determined by the default option of the software. Finally the speech was denoised using hard and threshold functions and the improved function.

6 Results

6.1 Conventional signal denoising

Figure 1 shows the threshold functions of the two kinds of signals before denoising, and Fig. 2 shows the signal waveform after denoising using the improved function. As shown in Figs. 1 and 2, the waveform appeared smoother and the

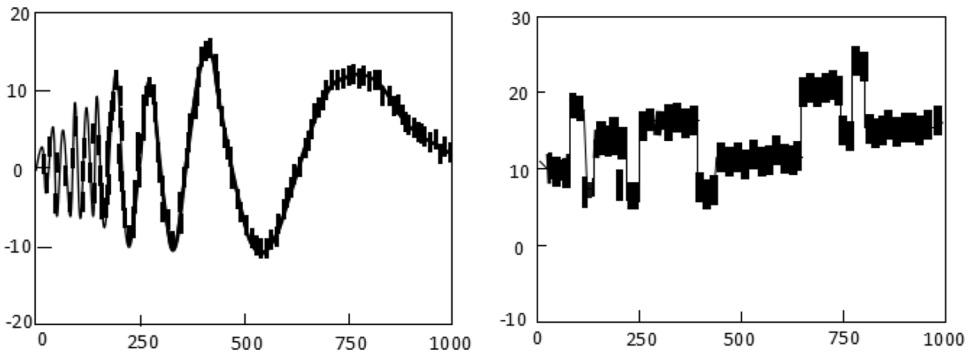

Fig. 1 The threshold function of the noisy Doppler signal (left) and Block signal (right) before denoising

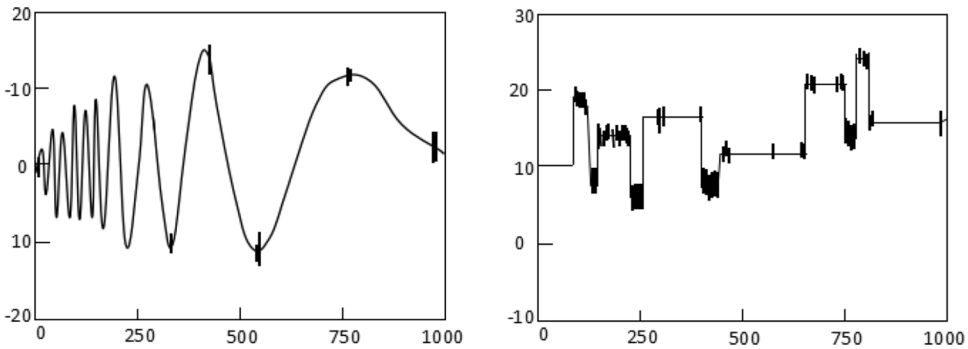

Fig. 2 The improved threshold function of the noisy Doppler signal (left) and Block signal (right)

Table 1 Parameters of Doppler signals before and after denoising

Parameters	Noisy signals	Soft threshold function	Hard threshold function	Improved threshold function
Doppler				
Signal-to-noise ratio	36.4330	51.6564	45.4816	52.9839
Mean square error	0.9833	0.4591	0.6254	0.4480
Blocks				
Signal-to-noise ratio	43.1328	53.1030	58.0101	61.0457
Mean square error	0.9453	0.7375	0.5771	0.5157

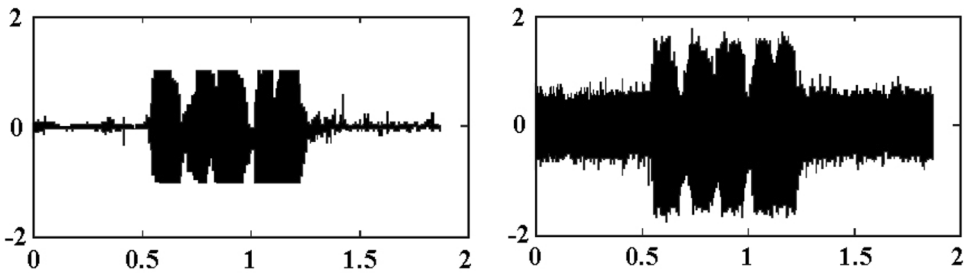

Fig. 3 The original speech signal (left) and noisy speech signal (right)

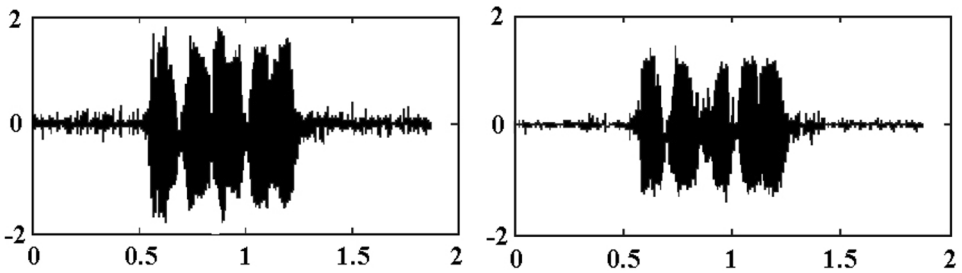

Fig. 4 The speech signal denoised using hard threshold function (left) and software threshold function (right)

glitches decreased after denoising, suggesting a good denoising effect. Table 1 shows the parameters before and after denoising.

As shown in Table 1, the signal-to-noise ratio of the traditional soft and hard threshold functions in Doppler signals was higher than that of non-denoised signals. The signal–noise ratios of the signals denoised using the improved function were higher than those of the signals denoised using the traditional hard and soft threshold functions, and the mean square error was lower than the mean square error before denoising. Blocks signal was the same. It indicated that the denoising effect of the improved function had favourable denoising effect.

6.2 Denoising of the speech signal

Figures 3, 4 and 5 suggested that the signal which was denoised using the improved function was more similar to the original signal. The comparison between the original signal and the signal denoised using the improved function indicated that the signal was more stable and smooth after

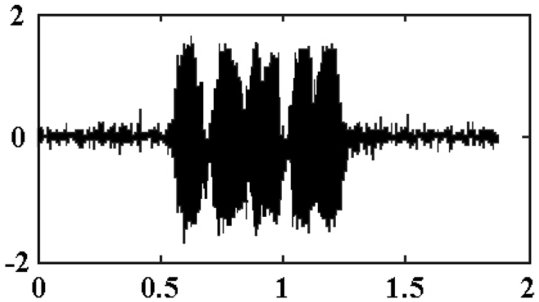

Fig. 5 The speech signal denoised using the improved function

denoising. The formation of the glitches contributed to the large noise in the environment, and the elimination of the glitches meant the decrease or disappearance of environmental noise. It indicated that the improved function could inhibit the environmental noise. For the quantitative comparison of the denoising results, the signal-to-noise ratio and mean square error of the above signals were estimated (Table 2).

Table 2 The index parameters of the signals denoised using different functions

Performance index	Noisy signal	Soft threshold function	Hard threshold function	The improved function
Signal-to-noise ratio	13.0663	14.2639	15.6590	16.0138
Mean square error	0.1999	0.1960	0.1939	0.1807

Table 2 demonstrated that the signal which was denoised using the improved function had larger signal-to-noise ratio and smaller mean square error compared to the signal which was denoised using soft and hard threshold functions. It indicated that the improved function was much better in reducing man-made noise.

6.3 Application of the improved function in speech recognition

To verify the denoising effect of the improved function in speech recognition, the improved algorithm was used in the denoising of an online speech recognition system which called Xunfei open platform whose core technology has reached the international leading level. It has an accuracy of 98% in recognizing speech and moreover supports the recognition of Chinese, English and dialect. The speech input speed is 180 words/min. It is capable of predicting Chinese punctuations. The system is composed of preprocessing module, speech recognition module and text translation module.

The system can be used after users log in and open related services. A self-evaluation parameter will be feed back to the backstage after audio recognition to reflect the self evaluation on the recognition accuracy.

To explore the effectiveness of the improved function in enhancing the accuracy of speech recognition, multiple groups of audio samples which were collected in the real environment before and after denoising were submitted to the system for recognition. The feedback self-assessment parameters were compared. The interface of upload is shown in Fig. 6.

The different audio samples collected from streets and bus stations were compared before and after denoising on the operating interface in Fig. 6. Due to the complexity of the actual environment, the street and bus station were selected as the collection locations. There were six groups of audio samples. The recognition accuracy of the speech recognition system was determined through the self-assessment parameters feed back by the system. Large self-evaluation parameter indicated better recognition result of the speech recognition system.

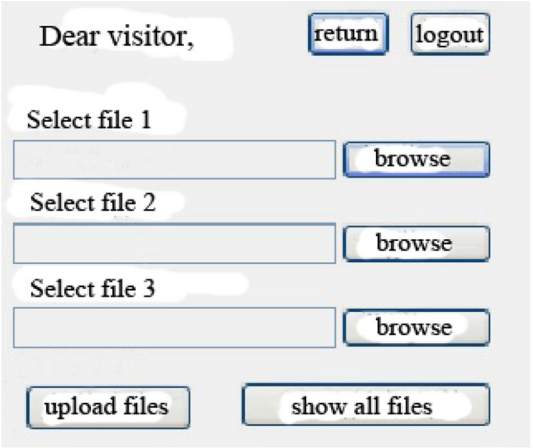

Fig. 6 The operation interface of the speech recognition system

As shown in Table 3, the recognition accuracy of the speech recognition system had obvious differences in different real environments. The system self-evaluation on the street audio and station audio was greatly improved after denoising. Therefore, the improved threshold function was considered as effective in promoting denoising in the speech recognition system. Moreover the recognition resolution was improved.

7 Discussion and conclusion

Speech recognition system plays an important role in communication. However, noise can interfere with the recognition of the speech recognition system, forming a barrier to communication (Tharwat et al. 2015). In speech recognition system, speech signal processing is based on linguistic and digital signal processing (Swaminathan et al. 2015). Digital signal processing is a technology that uses computers and other special equipment to enhance and compress discrete signals in the form of numbers (Halim et al. 2015). In order to improve the influence of noise pollution in speech recognition system, this study put forward an improved threshold function algorithm. Threshold function algorithm is a kind of wavelet denoising. The threshold function algorithm had obvious advantages over the other algorithms in the recognition of noisy audio frequency, suggesting that the threshold function algorithm has certain advantages in itself, which could reduce the glitches in the noisy signal and make the audio graph smoother. Then the improved threshold function was applied to the speech recognition system, and the identification effect was determined by the self-assessment parameter. The study found that the recognition rate of speech recognition system under different environments was different. However, no matter in what circumstances, the self-assessment parameters after denoising were greater

Table 3 Self-assessment of the speech denoising system

System self-assessment parameters	Street audio 1	Street audio 2	Street audio 3	Station audio 1	Station audio 2	Station audio 3
Before denoising	77.71	79.32	76.14	80.22	82.03	79.59
After denoising	82.47	84.15	81.05	86.02	87.49	85.71

than the self-assessment parameters before denoising, indicating that the improved threshold function was significant for denoising of the speech recognition system and could promote speech recognition. In conclusion, wavelet denoising has favourable denoising capability in the speech recognition system, which is worth application and promotion and has a good development prospect.

References

Aicha, A. B., & Jebara, S. B. (2012). Reduction of musical residual noise using perceptual tools with classic speech denoising techniques. *Signal Image & Video Processing, 6*(1), 85–97.

Alissali, D., Deleglise, P., & Rogozan, A. (2017). Asynchronous integration of visual information in an automatic speech recognition system. International conference on Spoken Language, 1996. Icslp 96. *Proceedings of the IEEE 1*, 34–37.

Bauer, B., & Kraiss, K. F. (2001). Towards an automatic sign language recognition system using subunits. Revised papers from the international gesture workshop on gesture and sign languages in human-computer interaction (pp. 64–75). New York: Springer-Verlag.

Gesell, G., Fischer, H., & König, T. (2009). Reduction of noise interference from METEOSAT water vapor image data by means of Fourier transform and frequency domain filtering. *Journal of Atmospheric Oceanic Technology, 1*(2), 147–151.

Halim, Z., Abbas, G. (2015). A kinect-based sign language hand gesture recognition system for hearing- and speech-impaired: A pilot study of Pakistani sign language. *Assistive Technology the Official Journal of Resna, 27*(1), 34.

Kumar, P., & Agarwal, S. K. (2015). Analysis of wavelet denoising of a colour image with different types of noises. *International Journal of Signal Processing Image Processing & Pattern Recognition*, 8, 125–134

Lecouteux, B., Linares, G., Esteve, Y., et al. (2016). Generalized driven decoding for speech recognition system combination. In IEEE international conference on acoustics, speech and signal processing. IEEE (pp. 1549–1552).

Mak, M. W., & Yu, H. B. (2014). A study of voice activity detection techniques for NIST speaker recognition evaluations. *Computer Speech & Language, 28*(1), 295–313.

Rajam, P. S., & Balakrishnan, G. (2012). Real time Indian sign language recognition system to aid deaf-dumb people. In IEEE, international conference on communication technology (pp. 737–742).

Shanthi, T. S., & Lingam, C. (2015). Speaker based Language Independent Isolated Speech Recognition System. In International conference on communication, information & computing technology. IEEE (pp. 1–7).

Srivastava, M., Anderson, C. L., & Freed, J. H. (2016). A New wavelet denoising method for selecting decomposition levels and noise thresholds. *IEEE Access Practical Innovations Open Solutions, 4*, 3862.

Sui, C., Bennamoun, M., & Togneri, R. (2015). Listening with your eyes: Towards a practical visual speech recognition system using deep boltzmann machines. In IEEE international conference on computer vision (pp. 154–162).

Sun, L., & Feng, Z. R. (2016). Classification of imagery motor EEG data with wavelet denoising and features selection. In IEEE international conference on wavelet analysis and pattern recognition (pp. 184–188).

Swaminathan, D., Kiruthika, S., Anton, A. L. N., et al. (2015). Video based indian sign language recognition system for single and double handed gestures with unique motion trace as feature. *International Journal of Tomography & Simulation, 28*(1), 71–88.

Tharwat, A., Gaber, T., Hassanien, A. E., et al. (2015). SIFT-based arabic sign language recognition system. In Afro-European conference for industrial advancement (pp. 359–370).

Torres-Carrasquillo, P. A., Singer, E., Gleason, T., et al. (2010). The MITLL NIST LRE 2009 language recognition system. In IEEE international conference on acoustics, speech, and signal processing, ICASSP 2010, 14–19 March 2010, Sheraton Dallas Hotel, Dallas, Texas, USA. DBLP, (pp. 4994–4997).

Zhenxing, L., & Hongzhou, X. (2009). A wavelet threshold de-noising algorithm based on empirical mode decomposition. *Computer Simulation, 26*(9), 192–325.