Received: May 20, 2020. Revised: August 10, 2020.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Modified MFCC for Improved Wavelet-Based Denoising on Robust Speech Recognition

Risanuri Hidayat[1]* **Anggun Winursito[1]**

[1]*Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Indonesia*
* Corresponding author's Email: risanuri@ugm.ac.id

Abstract: Research on the current speech recognition system leads to the creation of a noise-resistant system. The Mel Frequency Cepstral Coefficients (MFCC) extraction method becomes a popular method in the speech recognition system. In this paper, the MFCC's weakness of noise interference is the main reason underlies the accomplishment of a robust speech recognition system. Development was carried out by improving the denoising performance using a wavelet transform. Modifications were carried out by analyzing the weakness of the wavelet denoising process on the recognition system using the MFCC method. The analysis was conducted at one of the MFCC stages, the Fast Fourier Transform (FFT) stage. The proposed method was conducted by performing the denoising process using Wavelet only on the noise-related data based on the FFT process' analysis results. The study utilized speech data in the form of eleven isolated words in English added with noise with several different characteristics. Results showed that the proposed method was capable of generating a better accuracy than conventional wavelet denoising methods on the signal to noise ratio (SNR) of 10dB, 15dB, and 20dB using a Fejer Korovkin 6 wavelet type. The highest accuracy increase of the proposed method was in signal to noise ratio (SNR) of 15dB with a rise of 4.63%, followed by a 3.96% increase at 20dB intensity, and 2.3% at 10dB intensity. The performance of the proposed method is then compared with other methods. The results show that the proposed method has the best performance on clean speech and noisy speech at SNR intensities of 10dB, 15dB, and 20dB.

Keywords: Robust speech recognition, MFCC, Wavelet denoising, FFT.

1. Introduction

The speech recognition system technology is growing expeditiously in recent years. Many technological applications in everyday life utilize speech recognition systems. Even the speech recognition system has been researched to detect the cow's voice intended to control the animal's health [1]. Speech recognition systems can also be used for security systems in homes or industrial estates. Research by using a speech recognition system for security has been carried out [2]. The development of a speech recognition system is consecutively developed because it still has weaknesses. The systems' computing and recognition accuracy have become the main topic in the study of speech recognition systems [3]. Currently, the existing speech recognition system has been able to produce high accuracy for quiet environments, but the performance will decrease for the noise environment [4]. The research on the current speech recognition system leads to the creation of a noise-resistant system. Several methods have been developed to produce high recognition accuracy. One of the most popular feature extraction methods in speech recognition is Mel Frequency Cepstral Coefficients (MFCC). Some researches on speech recognition systems utilize the MFCC as a feature extraction method [1,5-10]. MFCC performs well in feature extraction processes with high accuracy results. Nevertheless, the problem is that MFCC has a weakness of not being able to be resistant to noise disturbance [8]. When the speech signal to be recognized contains noise, the recognition accuracy decreases.

Several pieces of research have been carried out to create a noise-resistant recognition system with the

MFCC extraction method. Research on noise-resistant speech recognition systems has been performed by comparing the MFCC's development method with other methods. The Bark Wavelet MFCC method is tested in the recognition system and claimed to produce better performance than the conventional MFCC method [5]. Research on speech denoising using Wavelet has also been done by other researchers [11]. The Discrete Wavelet Packet Transform was used to minimize the noise previously added to the speech data used for research experiments. The results show that the Wavelet transform is able to provide an approach of an almost optimal signal estimation from the speech data interrupted by noise. Meanwhile, Jangjit and Ketcham [12] researched the new wavelet denoising method for the noise threshold. The proposed method called Adaptive Thresholding with Mean for hybrid Denoising the hard and soft function (ATMDe) method has a better performance compared to other threshold methods that were tested. Another denoising speech research has been performed using a discrete wavelet packet decomposition [11]. The experiment was performed using speech data added with Gaussian noise of 0dB to 15dB. The research was conducted to observe the denoising process performance in comparing soft thresholding and hard thresholding schemes. The results show that wavelet packet decomposition gives an optimum signal estimation of the speech signals affected by noise disturbance. The application of the Wavelet transform to overcome speech signal noise in the MFCC was also performed by identifying the best parameters of the Wavelet transform [10]. The research was carried out by identifying the wavelet type, decomposition level, and the utilized thresholding method type. The used speech data contains noise with an intensity of 0-15dB. Results show that the Fejer Korovkin 6 type Wavelet produces the best performance in the intensity of signal to noise ratio (SNR) of 5-15dB, while the Daubechies 5 has the best performance in the SNR of 3dB with a level 10 decomposition.

There are many studies on a robust speech recognition system that selecting Wavelet transform to be utilized in the denoising speech process [10,11,13-17]. The Wavelet transform can capture time and frequency information at the same time on the speech signal; therefore, the wavelet denoising algorithm makes the system more robust against noise [16]. However, the performance resulted from the Wavelet transform in the denoising speech process has not been able to estimate the original speech signal ideally. As in research [10], which shows results that the system's accuracy in noise is still far from clear data signal accuracy, although it has shown an increase in accuracy compared without using wavelet denoising. Seeing that there are these problems, the development of denoising methods using wavelets still needs to be developed further. This paper proposes the improvement of the denoising wavelet method by modifying the MFCC feature extraction method to create a robust speech recognition system. Modifications were made by analyzing the weaknesses of the denoising process using Wavelet Transform on the recognition system using the MFCC method. In the conventional wavelet method, all data with any frequency is carried out the denoising process, so that the accuracy of the noise removal process becomes less than optimal. Whereas in the proposed method, a specific frequency section is analyzed to determine the right part for the denoising process. The analysis was conducted at one of the MFCC stages, the Fast Fourier Transform (FFT) stage. One of the MFCC stages is FFT serves to change signal from a time domain into a frequency domain [18]. Speech signal data changed in the frequency domain in the FTT process shows the part of data that as affected and not affected by the noise. The proposed method in this paper is to do the denoising process using Wavelet only on the noise-affected data and ignore the denoising process on noise-unaffected data. By precisely analyzing the noise-affected and noise-unaffected speech data, the denoising performance using Wavelet will be maximal. The performance of the proposed method will be compared with the conventional wavelet denoising method using the Singular Vector Machine (SVM) classification method.

The paper is organized as follows. Section 2 contains the research methodology which includes an explanation of the proposed method. Section 3 describes the experimental setup and results. Section 4 contains the conclusions of this paper.

2. Methodology

A speech recognition system has several stages in the speech signals recognition process, starting from the speech signals recording process, pre-processing, feature extraction, and the last is classification [10]. In the initial processing, data was cut to separate silent signals and information signals. Besides that, the filtering process was added by using a 2th-order median filter to smooth the signal at the initial processing. One of the most critical processes is feature extraction. Feature extraction is carried out to parse the speech signals characteristics so that they can be used as a differentiator between speech signals from one another. In this research, the utilized feature

Received: May 20, 2020. Revised: August 10, 2020.

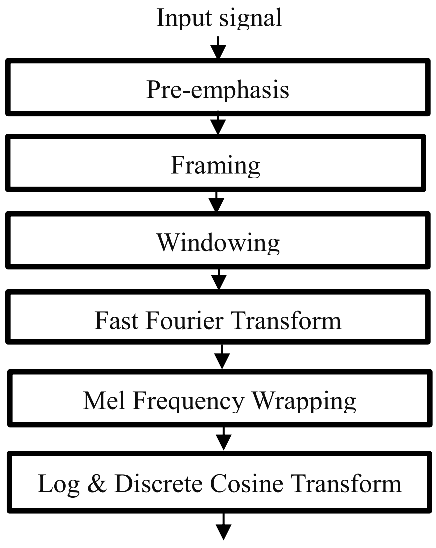


Figure. 1 Stages of a conventional MFCC

extraction process used was the Mel Frequency Cepstral Coefficient (MFCC) method.

The MFCC feature extraction process in the speech recognition system generally has the steps, as shown in Fig. 1. In MFCC feature extraction, the acoustic feature is extracted from the speech signal to represent the characteristics of the speech signal. Whereas to create a more noise-resistant system, researches have been conducted by adding a denoising process to the MFCC input. A popular method for the speech denoising process is to use the Wavelet transform. The Wavelet transform is conducted before the MFCC feature extraction. The proposed method in this research was carried out by analyzing the Fast Fourier Transform (FFT) output feature data in the MFCC process. Since the FFT process was carried out to convert speech signal data into the frequency domain, the analysis process was carried out in the frequency domain on n data of the FFT point. The stages of the proposed method's process in this research can be seen in Fig. 2.

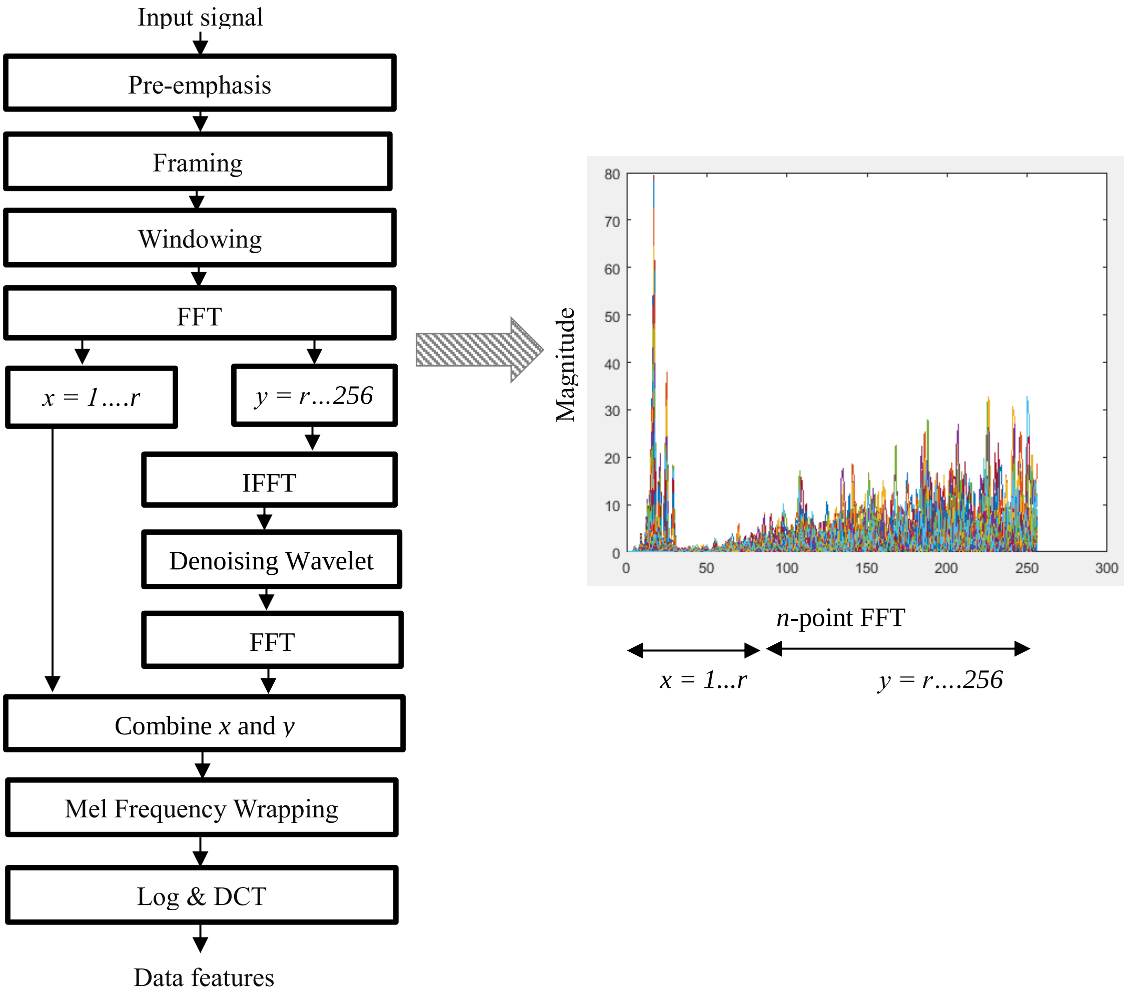


Figure. 2 Proposed modified MFCC

The conducted modification was by analyzing the speech data signals (*x*) which were not affected by noise interference, so there was no need to do a denoising process. It was carried out so that the characteristics of noise-unaffected speech data were maintained. The analysis was performed in the FFT output by determining the threshold of (*r*) data separating the noise-affected (*x*) and noise-unaffected (*y*) speech data. In determining the *r* threshold, there were two stages carried out. The first was to visually observe the FFT output comparison between clean data and noisy data. Observations were made by estimating the significant change in FFT data points between clean (*x*) and noisy (*y*) data. After estimating the value of *r* roughly, the subsequent analysis was carried out by trying out some predetermined *r*-value in the recognition system. The trial results were then analyzed based on the recognition results accuracy then the exact *r*-value was determined based on the highest recognition accuracy.

2.1 Preemphasis

This is a filtering process aiming to obtain a smoother spectral form of the speech signals frequency and to reduce noise during speech retrieval [19]. The mathematical equation of the preemphasis can be seen in Eq. (1).

$$y\,[n] = x\,[n] - a\,x\,[n - 1] \qquad (1)$$

where $y[n]$ is the output signal, $x[n]$ is the input signal, and a is a filter constant with the characteristics $0.9 < a < 1.0$. In this study, a was worth 0.95.

2.2 Framing and windowing

Framing is carried out to cut the signal into small parts in an overlapping manner. In this study, the speech signal was cut along 25ms with overlapping along 5ms. The windowing process was carried out to reduce spectral leakage and to minimize signal discontinuity per frame [9]. The utilized windowing function was the Hamming window according to Eq. (2)

$$(n) = 0.54 - 0.46.\ cos\left(\frac{2\pi n}{N-1}\right),\ 0 \le n \le N-1 \qquad (2)$$

so as the windowing process output is:

$$y1(n) = x1(n)\ w(n), \qquad 0 \le n \le N - 1 \qquad (3)$$

where $x1(n)$ is the input signal, $y1\ (n)$ is the output signal, $w(n)$ is the Hamming window equation, N is the number of signal samples in each frame, and values 0.54 and 0.56 are fixed coefficients.

2.3 Fast fourier transform (FFT)

FFT is commonly used in digital signal processing serves to change the signal from the time domain into the frequency domain [20]. The mathematical equation from FFT is as in Eq. (4).

$$X_n = \sum_{k=0}^{N-1} x_k e^{-2\pi jkn/N} \qquad (4)$$

where X_n is the FFT output, X_k is the input signal while the value of N = 0,1,2,....., ..., *N-1*.

FFT was performed on speech data that has been cut into several frames. The total FFT points in this study were 512, and one shift was taken (positive part) so that the total of utilized points were 256 FFT points. The FFT output display, along with the division of two parts separated by *r* threshold, can be seen in Fig. 3. The FFT results were then analyzed to find the *r* threshold separating the noise-unaffected data (*x*) and noise-affected data (*y*).

The noise-affected data (*y*) was then denoised using the Wavelet transform. Whereas the noise-unaffected data (*x*) did not experience the denoising process. Before the denoising process was performed, the *y* data was first converted back into the time domain using the Inverse Fast Fourier Transform (IFFT) according to Eq. (5)

$$y_k = \sum_{k=0}^{N-1} y_n e^{2\pi jkn/N} \qquad (5)$$

where FFT process input was y_n data and output was y_k data that had been transformed in the time domain. Next, y_k was denoised with the wave transformation.

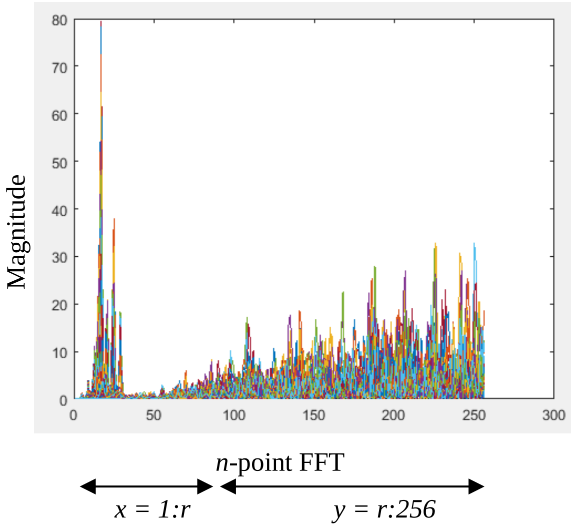


Figure. 3 The data output of the FFT process

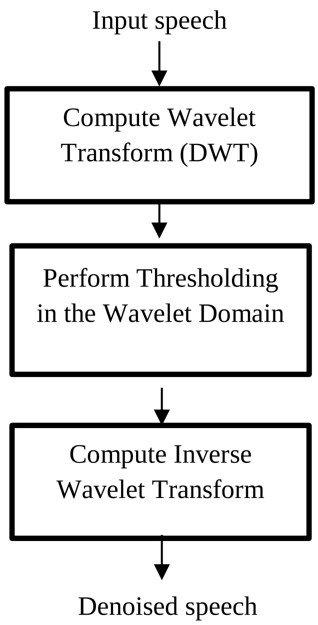

Input speech
Compute Wavelet Transform (DWT)
Perform Thresholding in the Wavelet Domain
Compute Inverse Wavelet Transform
Denoised speech

Figure. 4 Denoising with wavelet transform

The wavelet type utilized here was the type with the best performance in the previously conducted research [10], namely Fejer Korovkin 6. The utilized Wavelet transform parameters followed the previous researches' parameters [10]. Denoising by thresholding is based on the fact that the limited number of wavelet coefficients in the lower frequency band is sufficient to reconstruct the original signal [21]. The denoising process using the Wavelet transform can be seen in Fig. 4.

After going through the denoising process, the y_k output was then transformed back into the frequency domain using FFT following Eq. (4). Furthermore, the results of y_k data transformation that had become y_n were then combined with the noise-unaffected data (data *x*). Data *x* was equal to *r* (from 1st data to *r*), while *y* data as a result of transformation was equal to 256-*r* (from data number *r* to 256). The merging of data *x* and data *y* following Eq. (6)

$$x_{total} = [x_{(1 \dots r)}\, , y_{(r \dots 256)}] \qquad (6)$$

where x_{total} is the merging result of noise-unaffected data (*x*), and data resulted from the Wavelet denoising process (*y*).

2.4 Mel frequency wrapping

The Mel Frequency Wrapping process is based on the bandpass filter process. The filter is designed in the form of a triangle which has linear characteristics below 1000 Hz frequency and logarithmic above 1000 Hz frequency. The triangle filter is used to calculate the number of spectral filter components so that the process output approaches the Mel scale. The filter output is the sum of filtered spectral components [22]. The filter that has been designed is then applied to the x_{total} data. The number of triangular filters designed in this study was 26 so that the total MFCC output data features were 26 features per frame.

2.5 Discrete cosine transform

The last process of feature extraction is the Discrete Cosine Transform (DCT). The DCT process is the following Eq. (7)

$$\Sigma_{k=1}^{N} log\big(Y(i)\big)\, co\, s[mx(k-0.5)x\pi \div N] \qquad (7)$$

where *N* is the number of bandpass filter triangle. The value of m is between 1 to L, while for L is the sum of the output coefficients. Y is the input of data signals that will be processed on the DCT. The output used in the MFCC feature extraction process was 13 coefficients per frame of a total of 26.

3. Experimental setup and results

This study utilized speech data in the form of eleven isolated words in English ("zero" to "nine" and "oh"). The data were obtained from the TIDIGITS corpus. The utilized data was limited to only male voice data totaling 111 speakers with two repetitions. Therefore, the total utilized voice data was 2442 utterances (11 digits x 111 speakers x 2 times), and it was divided into training data and test data. The author then added noise obtained from the AURORA database into the test data. The noise added consisted of various types of environments, namely subway, babble, car, street, and restaurant with an intensity of signal to noise ratio (SNR) of 5dB, 10dB, 15dB, and 20dB. The Wavelet type utilized in the denoising process in this study referred to the best wavelet types in research that have been carried out before, namely Fejer Korovkin 6. The classification process of the system testing used the Support Vector Machine (SVM) classification method.

3.1 MFCC conventional

In this research, MFCC produced an accuracy of 95.37% for clean speech data. However, if the speech data contained noise, the recognition system accuracy decreased depending on the amount of noise intensity in the speech data. In this study, the speech recognition system was tested using the MFCC feature extraction method for speech data containing various environmental noise types. The recognition results using the MFCC method on the speech data

Table 1. The recognition accuracy (%) using the MFCC method for noise-containing speech data

Types	5dB	10dB	15dB	20dB
Subway	46.10	65.66	80.68	91.15
Babble	45.45	66.64	80.92	90.25
Car	46.75	66.88	81.49	90.16
Street	46.67	65.66	82.14	90.90
Restaurant	46.75	65.74	81.49	90.90

Table 2. Recognition accuracy (%) using MFCC method +Wavelet on the speech data with noise

Types	5dB	10dB	15dB	20dB
Subway	72.89	80.35	87.01	90.50
Babble	71.18	80.35	86.36	89.93
Car	72.32	80.19	86.52	90.50
Street	72.72	80.52	87.37	90.17
Restaurant	72.08	80.60	86.85	90.90

with noise can be seen in Table 1.

Table 1 shows the results of speech recognition using the MFCC method for various noise environments. It becomes apparent that recognition accuracy has decreased compared to the recognition of clear speech data (95.37%). The smaller the ratio of a speech signal to noise, the recognition accuracy decreases.

3.2 MFCC + wavelet denoising

The utilized Wavelet type was Fejer Korovkin 6. The recognition results using the MFCC method plus the denoising process using Wavelet Transform can be seen in Table 2.

Table 2 shows that the denoising method using wavelet transforms proven to be able to improve the accuracy of noise-containing data recognition. Accuracy improvement of the wavelet transform was carried out by removing noise from speech data in the denoising process before performing the MFCC feature extraction. However, the recognition results generated using the additional Wavelet method in Table 2 have not been able to approach the accuracy of the clear speech data recognition.

3.3 Proposed method

The purpose of designing this proposed method is to improve the performance of the wavelet-based denoising method. The proposed method in this study is to modify the MFCC feature extraction process and Wavelet transform. The wavelet denoising process was performed on a segment of the will-be-recognized speech data that was determined based on the characteristics of the Fourier Transform (FFT) process results or called as *n*-point FFT. The total FFT point data utilized in this study were 512 taken in one shift, so the total of analyzed FFT point data was 256. The display of Fourier speech data transformation process results on the MFCC feature extraction method can be seen in Fig. 5.

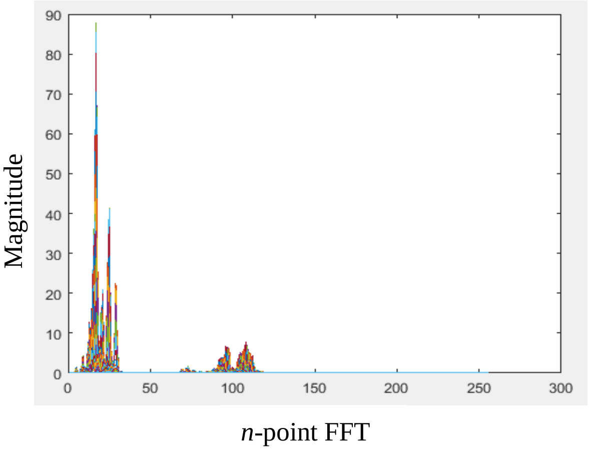


Figure. 5 The data output of the FFT process (clean)

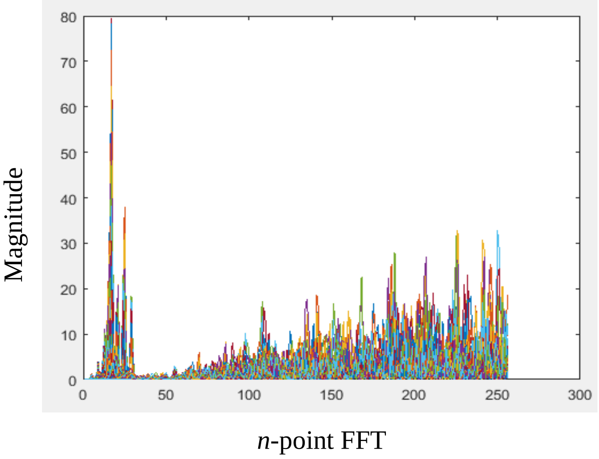


Figure. 6 The data output of the FFT process (noisy)

The accuracy in determining the *r*-value would significantly affect the recognition system's performance in the proposed method. Observations by comparing the FFT output data between clean speech and noisy speech were carried out to determine the *r*-value. The analysis was carried out by observing what data still has the same data characteristics between the FFT clean speech and noisy speech outputs. The comparison of FFT output between clean speech and noisy speech can be seen in Figs. 5 and 6.

From Figs. 5 and 6, it can be approximately determined that the data from 0 to 25th still have the same data character between clean speech (Fig. 5) and noisy speech (Fig. 6). As for the 25th to the 120th data, the difference is still visible. As for the 125th to 256th data, the differences in the data pattern characteristics are seen. Therefore, it can be determined that the threshold of the *r*-value is below

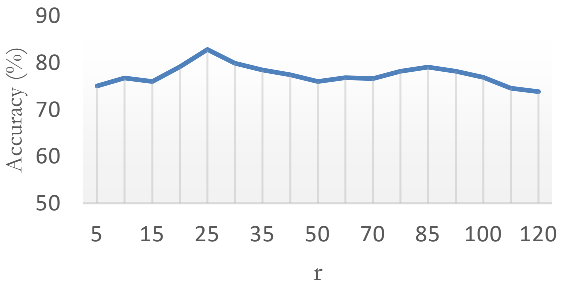


Figure. 7 The effect of *n*-point FTT (*r*) on recognition accuracy

the value of 120. To determine the threshold of *r*-value, further experiments were carried out on the recognition system. The experiment was carried out by adjusting the variation of the *r*-value or *n*-point FFT in the proposed method and observing the recognition accuracy results. To determine the value of the *r*, the conducted experiments employed Wavelet Fejer Korovkin 6 with speech data containing noise with an SNR of 10dB. The effect of *r*-value on the recognition system accuracy can be seen in Fig. 7.

From Fig. 7, it can be seen that the value of *n*-point FFT greatly affects the recognition accuracy. The highest accuracy is achieved at a value of 25 FFT points with an accuracy of 82.87%. The exact value of *n*-point FFT greatly affects the recognition accuracy because the proposed method is designed based on the analysis result of the first *n*-point FFT value unaffected by the noise signal. The series of the *n*-point FTT unaffected by the noise signal was stored in a variable, and no denoising process was performed. Therefore, if too many *n*-points FTT were taken, the noise signal would participate in the data features, thereby it reduced the recognition system's performance. Likewise, if too many *n*-points FTT were taken, the important information signals that should be stored would go through a denoising process and were susceptible to changing characteristics, so that it would affect the recognition system's performance.

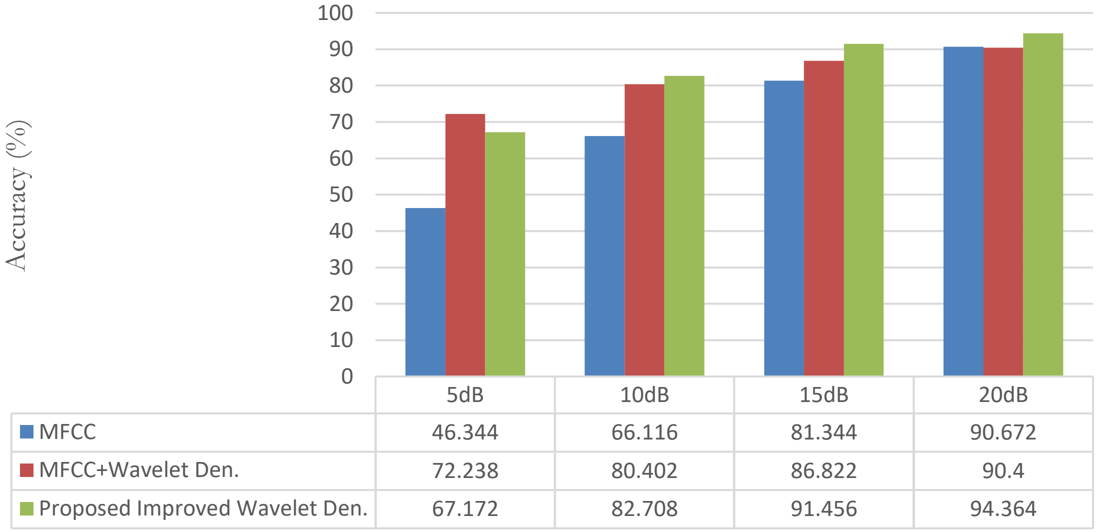
After obtaining the right FFT point value, then the value was utilized to test system performance against speech data containing several noise types. The signal to noise ratio (SNR) intensity was regulated in several values, namely 5dB, 10dB, 15dB, and 20dB. The recognition results using the proposed method can be seen in Table 3.

Table 3 shows the recognition accuracy using the proposed method of data containing noise with several environment types. The results show that the method, however, shows better accuracy than conventional wavelet denoising methods.

Fig. 8 shows that the proposed method can improve the accuracy of the conventional method (wavelet-based denoising) on SNR 10dB, 15dB, and 20dB. The highest accuracy increase in the proposed method is at an intensity of 15dB with an increase of

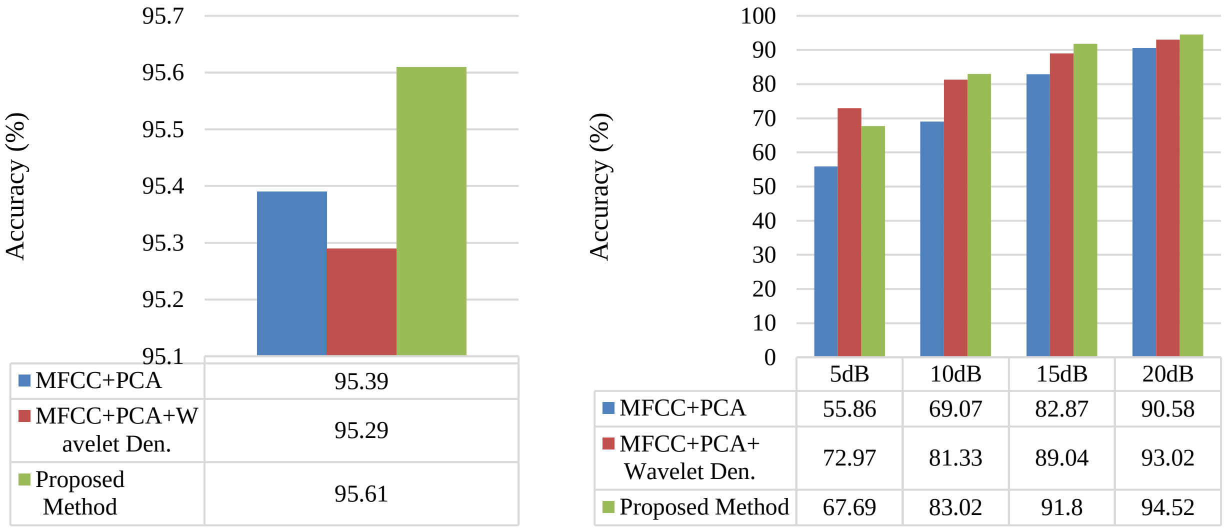
Table 3. Recognition accuracy (%) using proposed modified MFCC on the noisy speech data

Types	5dB	10dB	15dB	20dB
Subway	65.74	82.87	91.39	94.64
Babble	67.53	82.30	91.31	93.99
Car	67.53	82.87	91.55	94.40
Street	67.61	82.79	91.39	94.23
Restaurant	67.45	82.71	91.64	94.56



	5dB	10dB	15dB	20dB
🟦 MFCC	46.344	66.116	81.344	90.672
🟥 MFCC+Wavelet Den.	72.238	80.402	86.822	90.4
🟩 Proposed Improved Wavelet Den.	67.172	82.708	91.456	94.364

Figure. 8 Comparison of recognition accuracy of all methods at various SNR of 20dB, 15dB, 10dB, and 5dB



(a) (b)

Figure. 9 Comparison of recognition accuracy of all methods: (a) clean speech and (b) noisy speech

4.63%, followed by an increase of 3.96% at an intensity of 20dB, and 2.3% at an intensity of 10dB. Whereas the recognition accuracy of 5dB noise still shows lower performance compared to conventional Wavelet methods, but still shows higher accuracy compared to conventional MFCC methods. None of the proposed methods has been able to overcome the noise problem ideally, but overall, the proposed method shows the best performance.

The performance of the proposed method is then compared with the previous development method. One of the latest feature extraction method developments is MFCC + PCA. Several studies have been carried out by combining the MFCC and PCA methods and producing good performance [18], [23], [24]. Wavelet-based denoising methods are also often added as anti-noise in several studies [17], [25]. The results of the comparison of all methods can be seen in Fig. 9.

Fig. 9 shows that the proposed method has the best performance on clean speech and noisy speech at SNR intensities of 10dB, 15dB, and 20dB. Whereas the 5dB SNR of the proposed method produces lower performance compared to MFCC+PCA+Wavelet denoising methods, but still shows higher accuracy compared to MFCC+PCA methods.

4. Conclusion

The proposed method can relatively improve the denoising process using conventional wavelet to accomplish a noise-resistant speech recognition system. The results of speech data in the frequency domain show that not all data are affected by noise interference so that it can be sorted between the noise interference-unaffected data and the noise interference-affected data. The accuracy in determining the n-point FFT data range showing how much noise interference-unaffected data in the frequency domain that significantly determines the system recognition's performance. After an analysis of all 256 FFT points, 25 data points were selected as data that were not affected by noise interference and were ignored in the denoising process using wavelet transform. The results showed that the proposed method by modifying the MFCC could improve the accuracy of the conventional wavelet denoising method at 10dB, 15dB, and 20dB noise intensity. The highest accuracy increase in the proposed method is at an intensity of 15dB with an increase of 4.63%, followed by an increase of 3.96% at an intensity of 20dB, and 2.3% at an intensity of 10dB. The performance of the proposed method is then compared with MFCC+PCA and MFCC+PCA+ wavelet denoising method. The results show that the proposed method has the best performance on clean speech and noisy speech at SNR intensities of 10dB, 15dB, and 20dB. Future work will be carried out research using a larger dataset and using a more varied noise environment.

Conflicts of Interest

The authors declare that we do not have any circumstances or interest that may affect the results discussed in this manuscript.

Author Contributions

Conceptualization, Risanuri Hidayat; methodology, Risanuri Hidayat; software, Anggun Winursito; validation, Risanuri Hidayat and Anggun Winursito; formal analysis, Risanuri Hidayat and Anggun Winursito; investigation, Risanuri Hidayat; resources, Risanuri Hidayat; data curation, Risanuri Hidayat; writing—original draft preparation, Anggun Winursito; writing—review and editing, Risanuri Hidayat; visualization, Anggun Winursito; supervision, Risanuri Hidayat; project management, Risanuri Hidayat; funding acquisition, Risanuri Hidayat.

References

[1] N. Ding, X. Cheng, and Z. Cui, "Design of Ruminant Sound Detection for Dairy Cows Based on DWT-MFCC", In: *Proc. of 2018 5th International Conf. on Systems and Informatics (ICSAI)*, Nanjing, pp. 856–860, 2018.

[2] B. Rao and H. Chandrakanth, "Secured Speech Industrial Automation Based on Raspberry Pi and IoT", *International Journal of Intelligent Engineering and Systems*, Vol. 11, No. 6, pp. 234–243, 2018.

[3] Y. Guo, T. Li, Y. Si, J. Pan, and Y. Yan, "Optimized large vocabulary WFST speech recognition system", In: *Proc. of 2012 9th International Conf. on Fuzzy Systems and Knowledge Discovery*, Chongqing, Sichuan, China, pp. 1243–1247, 2012.

[4] C. Y. Fook, M. Hariharan, S. Yaacob, and A. Adom, "A review: Malay speech recognition and audio visual speech recognition", In: *Proc. of 2012 International Conf. on Biomedical Engineering (ICoBE)*, Penang, Malaysia, pp. 479–484, 2012.

[5] R. H. Tohidypour, S. A. Seyyedsalehi, and H. Behbood, "Comparison between wavelet packet transform, Bark Wavelet and MFCC for robust speech recognition tasks", In: *Proc. of 2010 2nd International Conf. on Industrial Mechatronics and Automation (ICIMA 2010)*, Wuhan, pp. 329–332, 2010.

[6] O. Eltiraifi, E. Elbasheer, and M. Nawari, "A Comparative Study of MFCC and LPCC Features For Speech Activity Detection Using Deep Belief Network", In: *Proc. of 2018 International Conf. on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*, Khartoum, pp. 1–5, 2018.

[7] Q. Li, H. Zhu, F. Qiao, Q. Wei, X. Liu, and H. Yang, "Energy-efficient MFCC extraction architecture in mixed-signal domain for automatic speech recognition", In: *Proc. of the 14th IEEE/ACM International Symposium on Nanoscale Architectures - NANOARCH '18*, Athens, Greece, pp. 138–140, 2018.

[8] K. K. Tomchuk, "Spectral Masking in MFCC Calculation for Noisy Speech", In: *2018 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF)*, St. Petersburg, pp. 1–4, 2018.

[9] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech based human emotion recognition using MFCC", In: *Proc. of 2017 International Conf. on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Chennai, pp. 2257–2260, 2017.

[10] R. Hidayat, A. Bejo, S. Sumaryono, and A. Winursito, "Denoising Speech for MFCC Feature Extraction Using Wavelet transform in Speech Recognition System", In: *Proc. of 2018 10th International Conf. on Information Technology and Electrical Engineering (ICITEE)*, Kuta, pp. 280–284, 2018.

[11] Z. Wang and S. Li, "Discrete Fourier Transform and Discrete Wavelet Packet Transform in speech denoising", In: *2012 5th International Congress on Image and Signal Processing*, Chongqing, Sichuan, China, pp. 1588–1591, 2012.

[12] S. Jangjit and M. Ketcham, "A New Wavelet Denoising Method for Noise Threshold", *Engineering Journal*, Vol. 21, No. 7, pp. 141–155, 2017.

[13] M. A. Oktar, M. Nibouche, and Y. Baltaci, "Speech denoising using discrete wavelet packet decomposition technique", In: *Proc. of 2016 24th Signal Processing and Communication Application Conf. (SIU)*, Zonguldak, Turkey, pp. 817–820, 2016.

[14] X. Liu, "A New Wavelet Threshold Denoising Algorithm in Speech Recognition", In: *Proc. of 2009 Asia-Pacific Conf. on Information Processing*, Shenzhen, China, pp. 310–313, 2009.

[15] M. A. Oktar, M. Nibouche, and Y. Baltaci, "Denoising speech by notch filter and wavelet thresholding in real time", In: *Proc. of 2016 24th Signal Processing and Communication Application Conf. (SIU)*, Zonguldak, Turkey, pp. 813–816, 2016.

[16] S. Li, "Speech Denoising Based on Improved Discrete Wavelet Packet Decomposition", In: *Proc. of 2011 International Conf. on Network Computing and Information Security*, Guilin, China, pp. 415–419, 2011.

Received: May 20, 2020. Revised: August 10, 2020.

[17] N. Soe, R. Hidayat, R. Hartanto, and Y. Miyanaga, "Discrete Wavelet Denoising into MFCC for Noise Suppressive in Automatic Speech Recognition System", *International Journal of Intelligent Engineering and Systems*, Vol. 13, No. 2, pp. 74–82, 2020.

[18] A. Winursito, R. Hidayat, A. Bejo, and M. N. Y. Utomo, "Feature Data Reduction of MFCC Using PCA and SVD in Speech Recognition System", In: *Proc. of 2018 International Conf. on Smart Computing and Electronic Enterprise (ICSCEE)*, Shah Alam, pp. 1–6, 2018.

[19] L. Marlina, C. Wardoyo, W. S. M. Sanjaya, D. Anggraeni, S. F. Dewi, A. Roziqin, and S. Maryanti, "Makhraj recognition of Hijaiyah letter for children based on Mel-Frequency Cepstrum Coefficients (MFCC) and Support Vector Machines (SVM) method", In: *Proc. of 2018 International Conf. on Information and Communications Technology (ICOIACT)*, Yogyakarta, pp. 935–940, 2018.

[20] B. P. Evans, *A Review of Automatic Music Transcription Low Level Processing Techniques and the Evaluation and Optimisation of Multiresolution FFT Parameters*, M. S. thesis, University of Huddersfield, Huddersfield, United Kingdom, 2012.

[21] S. Ayat, M. T. Manzuri, and R. Dianat, "Wavelet based speech enhancement using a new thresholding algorithm", In: *Proc. of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.*, Hong Kong, China, pp. 238–241, 2004.

[22] L. Muda, M. Begam, and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", *Journal of Computing*, Vol. 2, No. 3, pp. 138-143, 2010.

[23] M. B. Aithal, P. R. Gaikwad, and S. L. Sahare, "Speech Enhancement Using PCA for Speech and Emotion Recognition", *Global Journal of Engineering, Design, & Technology,* Vol. 4, No. 3, pp. 6-12, 2015.

[24] Adiwijaya, M. N. Aulia, M. S. Mubarok, W. U. Novia, and F. Nhita, "A comparative study of MFCC-KNN and LPC-KNN for hijaiyyah letters pronounciation classification system", In: *Proc. of 2017 5th International Conf. on Information and Communication Technology (ICoIC7),* Melaka, Malaysia, pp. 1–5, 2017.

[25] A. Niwatkar and Y. K. Kanse, "Feature Extraction using Wavelet Transform and Euclidean Distance for speaker recognition system", In: *Proc. of 2020 International Conf. on Industry 4.0 Technology (I4Tech),* Pune, India, pp. 145–147, 2020.