Received: April 24, 2021. Revised: May 20, 2021. 422

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Enhanced Feature Extraction Based on Absolute Sort Delta Mean Algorithm and MFCC for Noise Robustness Speech Recognition

Anuruk Nosan[1]* **Suchada Sitjongsataporn[2]**

[1]The Electrical Engineering Graduate Program, Faculty of Engineering and Technology, Mahanakorn University of Technology, 140 Cheumsamphan Road, Bangkok 10530 Thailand
[2]Department of Electronic Engineering, Mahanakorn Institute of Innovation (MII), Faculty of Engineering and Technology, Mahanakorn University of Technology, 140 Cheumsamphan Road, Bangkok 10530 Thailand
* Corresponding author's Email: anuruk9348@gmail.com

Abstract: In this paper, a proposed absolute sort delta mean (ASDM) method obtaining the speech feature extraction for noise robustness is developed from mel-frequency cepstral coefficients (MFCC) named ASDM-MFCC, in order to increase robustness against the different types of environmental noises. This method is used to suppress the noise effects by finding a rearranging average of power spectrum magnitude combined with triangular bandpass filtering. Firstly, the spectral power magnitudes are sorted in each frequency band of the speech signal. Secondly, the absolute-delta values are arranged and then a mean value is determined in the last step. The purpose of proposed ASDM-MFCC algorithm is to require the noise robustness of the feature vector extracted from the speech signal with the characteristic coefficients. The NOIZEUS noisy speech corpus dataset is used to evaluate the performance of proposed ASDM-MFCC algorithm by Euclidean distance method with the low computation complexity. Experimental results show that the proposed method can provide significantly the improvement in terms of accuracy at low signal to noise ratio (SNR). In the case of car and station at SNR=5dB, the proposed approach can outperform in comparison with the conventional MFCC and gammatone frequency cepstral coefficient (GFCC) by 80% and 76.67%, respectively. Obviously, some experimental results of the proposed ASDM-MFCC algorithm are more robust than the traditional one.

Keywords: Noise robustness speech recognition, Feature extraction, Gammatone frequency cepstral coefficient (GFCC), Mel-frequency cepstral coefficients (MFCC).

1. Introduction

Currently, technology about Automatic Speech Recognition (ASR) is more attention, because it can be used in several areas including with the speech recognition and speaker recognition to identify as speaker identification for a security system.

For the feature extraction of speech characteristic, the mel-frequency cepstral coefficients (MFCC) technique is widely used as a standard model for the extraction of speech vector characteristics. That is similar to the human auditory perception.

In terms of speech enhancement, a proposed adaptive LMD (ALMD) algorithm for the energy threshold technique is adopted to update the effective rank value of each frame of the speech matrix. In [1], The ALMD algorithm can achieve acceptably the performance for low signal-to-noise ratio (SNR) levels without approximating the speech phase with the noisy phase compared with the ALMD algorithm in Gaussian white noise and non-Gaussian noise conditions. In [2], speech emotion recognition was proposed. A novel approach using a combination of prosody features, quality features and derived features for robust automatic recognition of the speaker's emotional states. According to previous researches, there are not yet improvements of feature extraction algorithm for THAI digits speech and THAI speaker recognition that will be faced a noisy environment.

To increase the accuracy rate of speech and speaker recognition, the characteristic speech signal represented more resistance to noise than another algorithm. The descend-delta-mean algorithm

(DDM) has been proposed [3] to against the interference and noise. The End-to-End noisy speech recognition using Fourier and Hilbert spectrum features [4] has been improved the noise robustness by adding components to the recognition system. The incorporating noise robustness in speech command recognition by noise augmentation of training data is presented [5]. This work thoroughly analyses the latest trends in speech recognition and evaluates the speech command dataset on different machine learning-based and deep learning-based techniques.

For robust speech recognition, a method has been evaluated in [6], respectively and finally in [7]. This paper presents a feature extraction algorithm called power normalized Cepstral coefficients (PNCC) that is motivated by auditory processing. A noise-suppression algorithm based on asymmetric filtering that suppresses background excitation, and a module that accomplishes temporal masking. The experimental results demonstrate that PNCC processing provides substantial improvements in recognition accuracy compared to MFCC and PLP processing for speech in the various types of additive noise and in reverberant environments. Moreover, PNCC and K-Means have been applied for Speech / Music Classification [8] and combining PNCC and robust Mel-log filter bank features for Bird sounds classification [9].

Typically, the MFCC and the perceptual linear predictive (PLP) [10] techniques are evaluated as the most widely used techniques in speech and speaker recognition systems. However, the PLP method relative spectral (RASTA) [10] filtering is combined with the feature extraction technique to remove channel noises compared to the speech signal. Recently, the enhanced automatic speech recognition system based on enhancing PNCC has been presented [10]. PNCC also proposes are estimated over a long duration that is commonly used for speech, as well as frequency smoothing. The experimental results demonstrate that PNCC processing provides substantial improvements in recognition accuracy compared to MFCC and PLP processing for speech in the various types of additive noise and in reverberant environments. The PNCC system is modified to increase robustness against the different types of environmental noises, where a new technique based on gammatone channel filtering combined with channel bias minimization is used to suppress the noise effects compared with the advanced noise robust feature extraction technique as gammatone frequency cepstral coefficient (GFCC). The results showed that the proposed method provides significantly improvements in the recognition accuracy at low signal to noise ratio (SNR) and the recognition rate of the method is higher than GFCC [10] and PNCC methods in the some case of noise.

According to enhance the conventional MFCC-based algorithm, the speaker recognition system based on a swarm intelligence algorithm called modified shuffle frog leaping algorithm (MSFLA) is proposed with cepstral analysis and the MFCC feature extraction approach [11]. By applying this algorithm, the process of speaker recognition is optimized by a fitness function by matching of voices being done on only the extracted optimized features produced by the MSFLA.

Several algorithms as LPC, HMM, and ANN are evaluated to identify a straightforward and effective method for voice signals. The viability of MFCC to extract features and DTW to compare the test patterns is proposed. The method for generating speech from filterbank MFCC was proposed in [12].

To improve the robustness of speech front-ends for noise-robust feature extraction, a new set of MFCC vector which is estimated through three steps. First, the relative higher-order autocorrelation coefficients are extracted. Then magnitude spectrum of the resultant speech signal is estimated through the FFT. Finally, the differentiated magnitude spectrum is transformed into MFCC-based coefficients. There are called MFCCs extracted from differentiated relative higher order autocorrelation sequence spectrum. In [13], the noise-robust speech feature extraction algorithm is proposed. That introduces the improving the feature extraction method of the MFCC algorithm for noise robustness when there is interference in the speech input signal. It was modified and developed from the DDM algorithm [3] as the descend-delta-mean and mel-frequency cepstral coefficients (DDM-MFCC). The purpose is to require the noise robustness of the feature vector extracted from the speech signal with the characteristic coefficients of the proposed algorithm.

In active sonar target classification with PNCC and convolutional neural network [14], the feature vectors are extracted with MFCC compared with the PNCC algorithm. The experiment results represented that the proposed algorithm has a higher classification rate than MFCC without affecting the target classification by the signal level.

Also, in [15] threshold-based noise detection and reduction for ASR system in human-robot interactions are proposed. Experimental results showed that the SNR values of the enhanced speech can exceed those of the received noisy speech by approximately 20 dB to 25 dB. The NOIZEUS noisy speech corpus dataset is widely discussed and used to analyze speech [16-18]. The NOIZEUS speech

dataset has been used to measure the efficiency of the proposed system under various environmental noisy conditions. In [19], the Thai voice interfaces has been proposed to control a car parking assist based on MFCC method to extract the features.

To overcome these problems above, in this paper to achieve satisfactory performance in speech recognition, a new method to obtain speech feature extraction for noise robustness is proposed. It will be modified and developed from the MFCC [3] and the DDM-MFCC [13] algorithm. This has been modified to provide the properties that can withstand noise at low SNR without affecting system performance. That is the absolute sort delta mean (ASDM) mel-frequency cepstral coefficients, called the ASDM-MFCC algorithm. It is modified to increase robustness against the different types of environmental noises.

A brief description of the process consists of the method as follows. First, the magnitude is sorting of the spectral power in each frequency band of the speech signal. Secondly, to find the absolute delta value in the arrangement and determine a value of mean in the last step. The purpose of the ASDM-MFCC algorithm is to require the noise robustness of the feature vector extracted from the speech signal with the characteristic coefficients.

This paper is organized as follows: Section 2 discusses the ASDM-MFCC algorithm for the feature extraction in detail. The experiments work and the results of each process and recognition rate are presented in Section 3. Finally, Section 4 concludes and summarizes the outcomes of the paper and future works.

2. Proposed absolute sort delta mean algorithm

In this section, two sub-parts of speech recognition are being considered. That is the processes of signal pre-processing and feature extraction. The speech signal, frame blocking and windowing are three solutions of pre-processing processes. And the feature extraction process is an objective for this work. This work concentrates to develop and improve the algorithm for the noise robustness. The MFCC algorithm is a method of feature extraction to select for more performance of recognition.

In order to discuss the speech recognition accuracy rate, the conventional MFCC and proposed ASDM-MFCC robustness algorithm of feature extraction are shown in Fig. 1, which divides into two parts at the right-hand side as the signal pre-processing process and feature extraction process.

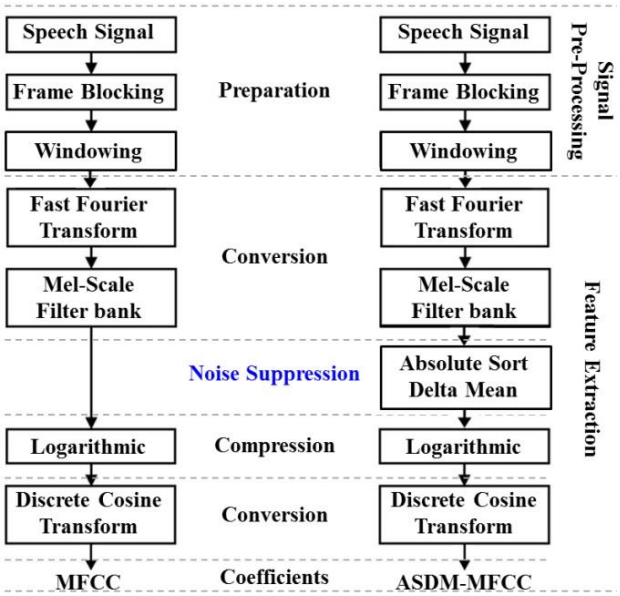


Figure. 1 Block diagram of MFCC and proposed ASDM-MFCC robustness algorithm of feature extraction

2.1 Signal pre-processing

Signal pre-processing consists of three sub-steps. Step 1 concludes the speech signal from speakers that explains in the next section. Steps 2 and 3 are frame blocking and windowing represented as follows.

2.1.1. Frame blocking

According to the speech signals, the signal has been characterized to a non-stationary system that is a statistical value changed over time. It is therefore necessary to divide the speech signal into short-time subsections, called frame blocking or framing of a speech signal. Each framing size of the short-time subsection is approximately 10-30 milliseconds. And the size of the frame overlapping is approximately 10 milliseconds. The effect of each speech frame is obtained the statistical value less changed over time. It can be considered that the statistical values of each speech frame are stationary or do not change over time. Thus, it can be processed by applying statistical values to the speech signals in each frame.

2.1.2. Windowing

The windowing or smoothing windows provides data of each speech signal for analyzing the autocorrelation. The autocorrelation is done by multiplying each signal value in the speech data frame by the window function value.

There are many types of window used such as rectangular window, Hamming window, Hanning window, Blackman window, Kaiser window and so on [18]. There are two kinds of results after windowing. Firstly, it is a slow reduction in

amplitude at each end of the speech signal to prevent suddenly changes at the end of the signal frame. Secondly, it is a convolution value for the Fourier transformation effect of the frame function and spectral of speech. The speech signal goes through this process will transform to the speech data for further use in digital signaling processing.

$$W_{(n)} = frame_{(n)} w_{(n)} \quad , \tag{1}$$

$$w_{(n)} = 0.54 - 0.46\cos\left(\frac{2\pi n}{N} - 1\right) \quad , \tag{2}$$

where $n = 0,1,2, \dots, N$; $W_{(n)}$ is windowed of framed of speech signal value. $frame_{(n)}$ is a speech signal value of frame data at n, where n is a sequence of frames of speech signal value and N is the amount of data in each frame of speech signal value. $w_{(n)}$ is Hamming window function.

2. 2 Feature extraction

In order to find the characteristic value used instead of the speech signal, this work uses the spectral envelope group of feature extraction as a feature extraction method.

2.2.1. Short time fourier transform

The discrete Fourier transform (DFT) is a way of mapping the signal from the speech, which is a signal varying continuously with time, into their frequency components. Fast Fourier transform (FFT) is shown as

$$F_{(m)} = \sum_{n=1}^{N} X_{(n)} e^{-j\frac{2\pi n}{N}n} \quad , m = 1,2,3, \dots, N \quad , \tag{3}$$

where $F_{(m)}$ is the data of FFT it is complex numbers, $X_{(n)}$ is an input speech signal data sequence at n. N is the amount of input data and the size-dependent on framing. The iteration of FFT is called nFFT, and then m is the index sequence of $nFFT/2$ [3, 13].

As considered the result of Eq. (1), the speech signal is divided into segments and overlapped over the entire speech signal range. Then, the multiple signals have been the short duration time characteristic. For this reason, the solution of FFT converting the time domain to the frequency domain of these signals is called the short time Fourier transform (STFT). The equation defines the STFT as

$$F_{(m,n)} = \sum_{n=1}^{N} W_{(m,n)} e^{-j\frac{2\pi n}{N}n} \quad , m = 1,2,3, \dots, N \, , \tag{4}$$

where $F_{(m,n)}$ is the data of STFT is complex numbers and the size is in two dimensions. $W_{(m,n)}$ is a windowed segment of framed input speech signal data and n will become is a sequence of frame data. And absolutely, the size of this STFT obtained is the mxn size of matrix. From the STFT conversion result of Eq. (4), the power spectrum magnitude of the speech signal is required.

The complex number $F_{(m,n)}$ is converted to the real number data as

$$P_{(m,n)} = |F_{(m,n)}|^2 \quad , \tag{5}$$

where $P_{(m,n)}$ is the power spectrum magnitudes of an input speech signal.

2.2.2. Mel Frequency cepstral coefficients (MFCC)

The analysis of mel-scale frequency cepstrum has been widely populated used in current ASR systems. The cepstrum is the discrete cosine transform of the logarithmic of the short-time signal of the power spectrum. The coefficient of mel-scale cepstrum is an enhanced technique adapted from cepstrum. The mel scale is a uniformly space of the triangular filter banks. Therefore, the bandwidth of the individual filter increases the logarithmic in the normal scale and also normalized frequency. Each triangular filter is of length in a frequency domain. Let the magnitude response of the sequence of a triangular filter is presented as the mel frequency, so the form involved of mel scale is mentioned below.

$$mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad , \tag{6}$$

where $mel(f)$ is the mel frequency scale and f is the actual frequency.

Substituting Eq. (5) into Eq. (6), it is obtained the mel frequency that converts from the actual frequency of power spectrum magnitudes of an input speech signal by mel-scale filter bank as

$$M_{(c,n)} = mel\left(P_{(m,n)}\right) \, , \tag{7}$$

where $M_{(c,n)}$ is converted of mel frequency coefficients and c is a sequence of the coefficients by mel-scale filter bank solution.

2.2.3. Proposed absolute sort delta mean MFCC (ASDM-MFCC) algorithm

The ASDM-MFCC algorithm is to find the average absolute value of a delta of a descendant as follows. The ASDM-MFCC algorithm was modified and developed from [3] and [13]. The modified algorithm is shown an acoustic property that can withstand noise at low SNR without affecting system performance and to increase robustness against the different types of environmental noises. It is based on finding a rearranging average of power spectrum magnitude and then combine with triangular bandpass filtering used to suppress the noise effects. The ASDM method is defined as

$$ASDM = \frac{\sum_{k=1}^{K-1}\left|S_{(k)} - S_{(k+1)}\right|}{K-1} \quad , \qquad (8)$$

$$S_{(k)} = D_{es}\{S_{(i)}\} \ , \qquad (9)$$

where $ASDM$ is the average of absolute delta of sorted result. $S_{(i)}$ is the data of S of a sequence number at sequence i. k is the amount of data and $D_{es}\{\bullet\}$ is an operator of sorting algorithm either the ascending or descending order. If the ascending or descending coefficients are set of the speech signal, the highest value is shifted to the left and then has been inserted into the appropriate point.

Moreover, this method is simple and appropriate for small data. The alignment can be shown in Fig. 2, it is a pseudocode of sorting of the ASDM- MFCC algorithm mentioned.

Therefore, an $M_{(c,n)}$ is converted of the mel frequency coefficients from Eq. (7) and substituting with Eq. (9), we get

$$M_{(c,k)} = D_{es}\{M_{(c,n)}\} \ , \qquad (10)$$

where $M_{(c,k)}$ is the sorted data of mel frequency coefficients of a sequence number at k.

An illustration of the sorting of the ASDM-MFCC algorithm is demonstrated in Fig. 3 (a) and Fig. 3 (b). It is noted that Fig. 3 (a) and Fig. 3 (b) are the next step of Fig. 3 (b) and Fig. 3 (b), respectively.

The second step is the process of the absolute delta. This process is to evaluate the difference value of each coefficient set of every frame of the speech data assorted. From Eq. (8) the absolute delta is a term of the $\left|S_{(k)} - S_{(k+1)}\right|$, which the $S_{(k)}$ is obtained from Eq. (9). Then, that means a term of the $M_{(c,k)}$ in Eq. (10) is equal to $S_{(k)}$. And, it establishes that $\left|S_{(k)} - S_{(k+1)}\right|$ is the absolute delta of the $\left|M_{(c,k)} - M_{(c,k+1)}\right|$. And the final step, the mean process is to evaluate the average of the absolute delta. From the all term obtained above, we substitute into the Eq. (8) as

$$ASDM_{(c)} = \frac{\sum_{k=1}^{K-1}\left|M_{(c,k)} - M_{(c,k+1)}\right|}{K-1} \quad , \qquad (11)$$

where $ASDM_{(c)}$ is the average result of the absolute delta of each coefficient set of c of every frame of the speech data. The size of this $ASDM_{(c)}$ is 40 x 1.

2.2.4. Logarithmic and discrete cosine transform

The two final steps of the standard technique of the MFCC are the logarithmic and discrete cosine transform processes. The process of discrete cosine transform is the coefficients of speech data converted to the cepstrum coefficients of MFCC. In fact, the discrete cosine transform (DCT) is the one technique of the inverse Fourier transform that is defined as

$$C_{(q)} = w_{(q)} \sum_{q=1}^{Q} \left[L_{(q)} cos \frac{\pi(2q-1)(q-1)}{2Q} \right], \qquad (12)$$

$$L_{(q)} = \log_{10}\left(Y_{(q)}\right) \ , \qquad (13)$$

$$w_{(q)} = \begin{cases} \dfrac{1}{\sqrt{Q}}, & q = 1 \\ \sqrt{\dfrac{2}{Q}}, & 2 \le q \le Q \end{cases} \quad , \qquad (14)$$

where $C_{(q)}$ is the cepstrum coefficients of discrete cosine transform. That is an output of the feature extraction of the MFCC method. $L_{(q)}$ is the result of the logarithmic compressed of $Y_{(q)}$. $Y_{(q)}$ is the coefficients as of data input. $w_{(q)}$ is the conditions of the cepstrum method. q is sequence of coefficient, by $q = 1,2,3,\dots,Q$; and Q is amount of q . Then, we substitute the $ASDM_{(c)}$ from Eq. (11) into Eq. (13) as

$$L_{(c)} = \log_{10}\left(ASDM_{(c)}\right) \ , \qquad (15)$$

where $L_{(c)}$ is the result compressed of coefficients of $ASDM_{(c)}$. And $L_{(c)}$ in Eq. (15) is instead of the coefficients of logarithmic of Eq. (12), we get

```
Algorithm: sorting of MFCC-ASDM
  Data: Input array Coefficients[]
  Result: Sorted Coefficients[]
  int i, j;
  N = length(Coefficients);
  for j = 1 to N do
  |  for i = 1 to N-1 do
  |  |  if Coefficients[i] > Coefficients[i+1] then
  |  |  |  temp = Coefficients[i];
  |  |  |  Coefficients[i] = Coefficients[i+1];
  |  |  |  Coefficients[i+1] = temp;
  |  |  End
  |  end
  end
```

Figure. 2 Pseudocode of a sorting of the ASDM- MFCC algorithm.

$$C_{(c)} = w_{(c)} \sum_{c=1}^{Q} \left[\log_{10}\left(ASDM_{(c)}\right) cos \frac{\pi(2c-1)(c-1)}{2Q} \right] , \quad (16)$$

$$w_{(c)} = \begin{cases} \frac{1}{\sqrt{Q}}, & c = 1 \\ \sqrt{\frac{2}{Q}}, & 2 \le c \le Q \end{cases} . \quad (17)$$

3. Experiments results and discussion

In this section, the representation and discussion of the results of the experiments are considered. The results consist of three sub-part as the speech recognition efficiency, noise environment accuracy, and resolution of nFFT as follows.

3.1 Speech corpus database

The NOIZEUS dataset [16-18] was used to evaluate the comparison of performance between the proposed and the traditional methods. The sentences of noisy database were produced by three male and three female speakers.

In addition, the noise database is corrupted by eight different places of real-world noises with four different SNRs from 0 dB to 15 dB with a step size of 5 dB. Noise signals are included the following recordings from eight different places such as airport, babble (crowd of people), car, exhibition hall, restaurant, street, train and train station noise which were taken from the AURORA database. The NOIZEUS database were originally sampling rate at 25 kHz and down sampled to 8 kHz. For all sentences were saved in wav format with 16 bit PCM by mono channel style.

Therefore, the sentences of noise speech signals is a training dataset was 120 sentences by 4 different SNRs of each different places of noises, while sentences of clean speech signal is a testing dataset was 30 sentences.

An illustration of the signal data of the NOIZEUS dataset noisy speech corpus shown in Fig. 4 as the example of speech signal data of Sp28. wav audio file sentences. That was established by the female gender speaker. They are comparing and should be listed as, Fig. 4 (a) is the clean speech signal for training speech data and Fig. 4 (b) is the airport noise environment of the speech signal with SNR 0 dB for testing speech data. Furthermore, Fig. 4 (c) and Fig. 4 (d) is the demonstration of the 3D model of power spectrum magnitude of the clean speech signal and airport noise environment of the speech signal with SNR 0 dB respectively.

3.2 Pattern matching and decision

Pattern matching aims to search the most similar pattern of coefficients set of proposed ASDM-MFCC algorithm. The matching method is an instance of pattern feature in a speech data at observation. This leads to a decision step of the recognition process.

For low-complexity of computation of pattern matching in the proposed algorithm, we design and develop the established fast and simple pattern matching method for the recognition process.

The pattern matching is the vector Euclidean distance [3, 13] which is used to evaluate the performance and to classify the results of the experiments. Additionally, this technique is also an appropriate method for finding the relationship between variables of vectors as well. The pattern matching is defined as

$$U_{(a)} \cong V_{(b)} \quad (18)$$

where $U_{(a)}$ is clean of speech training data of a of the NOIZEUS dataset. a is sequence of clean of speech training data, where $a = 1,2,3,\dots,A$. $V_{(b)}$ is the speech testing data of b of the NOIZEUS dataset. And b is the sequences of speech testing data; where $b = 1,2,3,\dots,B$.

The Euclidean distance vector defines as

$$d_{(a,b)} = \sum_{a=1}^{A} \sum_{b=1}^{B} \sqrt{(U_{(a)} - V_{(b)})^2} \quad , \quad (19)$$

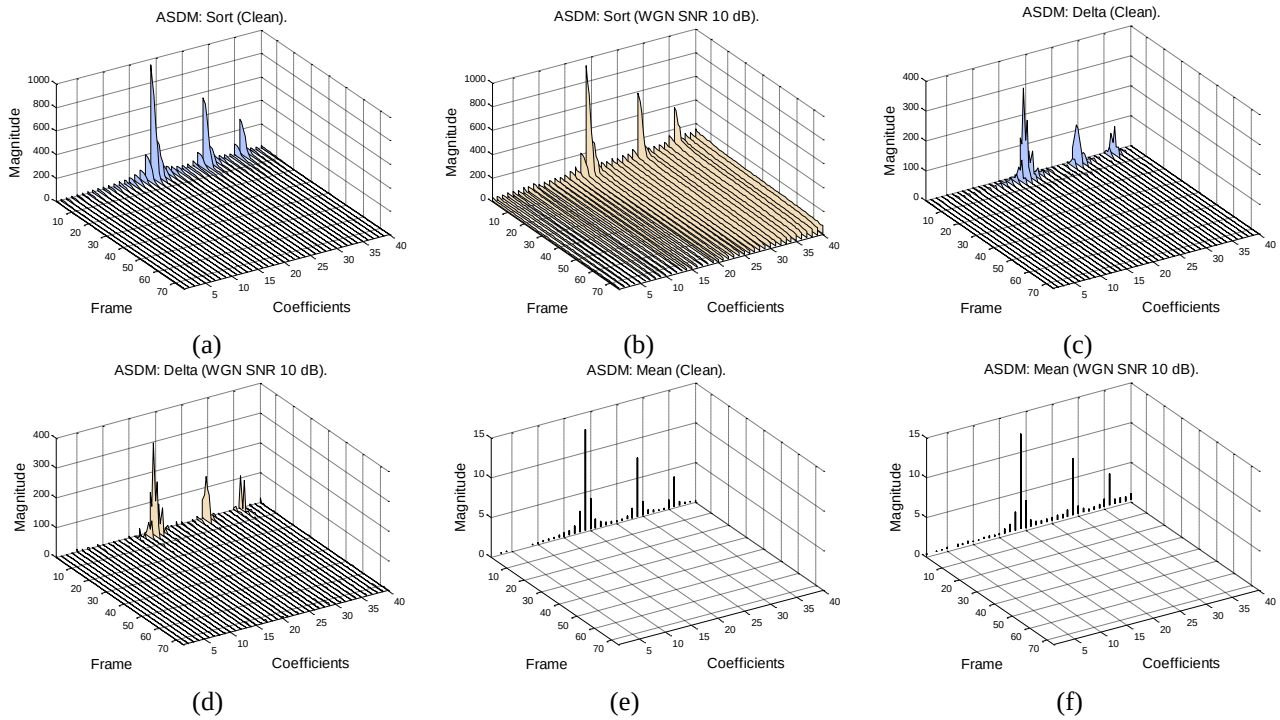


Figure. 3 The illustration of the comparison of 3D model of the ASDM-MFCC algorithm with clean and noise coefficient as: (a), (c) and (e) are the demonstration of the clean coefficients set as follows sorting of coefficients, taking an absolute delta, and evaluation of an average results, respectively, (b), (d) and (f) are the demonstration of a WGN of SNR 10dB noise coefficients for sorting, taking the absolute delta and evaluating of an average result, respectively

where $d_{(a,b)}$ is the vector distance value of coefficients of sequences a and b between the speech training and speech testing data of the NOIZEUS dataset. Therefore, the similarity method of the pattern matching is shown as

$$D_{(a)} = \min\left(d_{(a,:)}\right) \quad , \tag{20}$$

$$d_{(a,b)} = D_{(a)} \quad , \tag{21}$$

where $D_{(a)}$ is the minimum value of the vector distance of coefficients of a sequences a and all of sequences b. Finally, the decision of the recognition process is as follows.

If $d_{(a,b)} = D_{(a)} = U_{(a)}$ then it must have been the criterion of the scoring accuracy. Then, the conditions of b is establishing defined as

$$S = \begin{cases} 1, & if\ 1 \le b \le (E \cdot a) \\ 1, & if\ (E.a) - E < b \le (E \cdot a) \\ 0, & otherwise. \end{cases} \tag{22}$$

where S is the score of accuracy result obtained from recognition of the proposed ASDM-MFCC algorithm.

3.3 Results

The experimental results are established with the proposed ASDM-MFCC algorithm compared with the existing techniques as GFCC, RASTA-PLP PNCC, MFCC feature extraction methods. Fig. 5 illustrates the difference block diagram structures of the conventional GFCC, RASTA–PLP, PNCC, MFCC and proposed ASDM-MFCC methods. The block diagrams in Fig. 5 are divided into two stages. These stages are signal pre-processing, and feature extraction. All methods are obtained from the NOIZEUS speech corpus dataset.

3.3.1. Speech recognition efficiency

The efficiency of speech recognition results are shown in Fig. 6 about the recognition accuracy, which is presented graphically for the proposed ASDM-MFCC algorithm in comparison to the conventional MFCC algorithm system, as well as the most commonly used baseline systems, such as PNCC, RASTA-PLP, and GFCC.

Fig. 6 is the experimental result of the average accuracy percentage rate of speech recognition. By comparison of each algorithm using the noisy speech corpus of the NOIZEUS dataset, the result of the

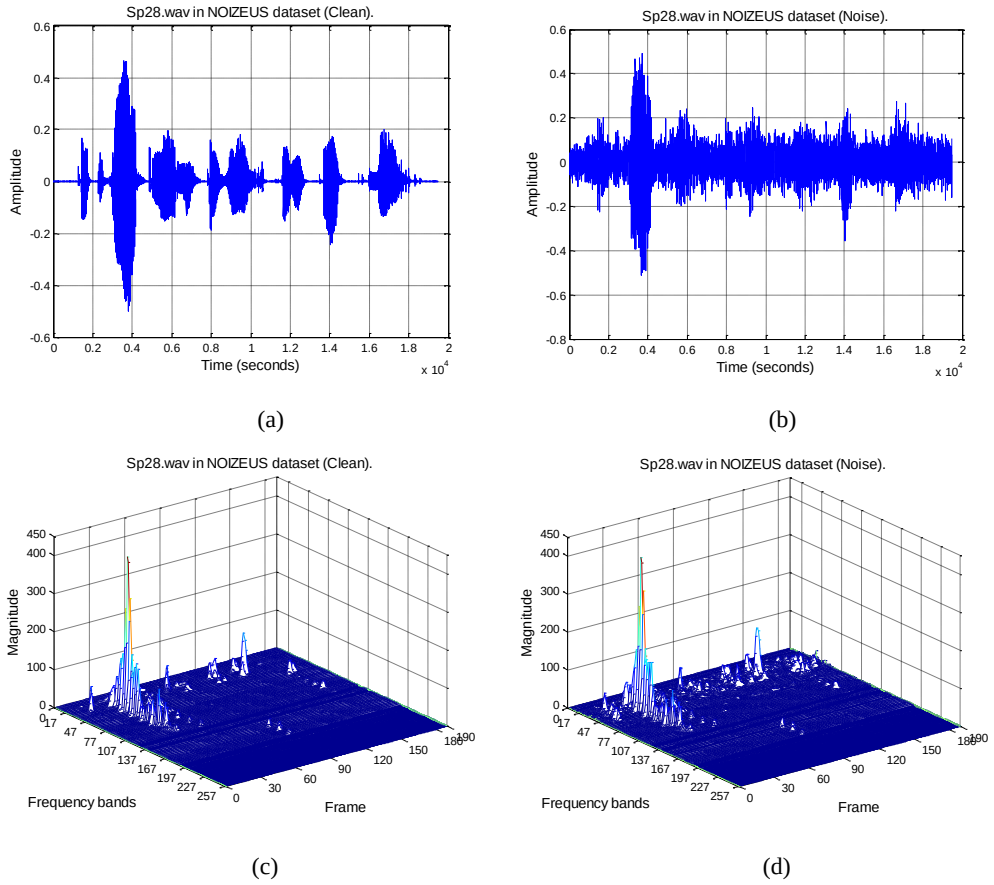


Figure. 4 Speech signal of Sp28.wav in NOIZEUS dataset as: (a) clean speech signal, (b) Airport noise with SNR= 0 dB, (c) 3D power spectrum magnitude of the clean speech signal, and (d) 3D power spectrum magnitude of airport noise

average accuracy percentage rate obtained. It is derived from the use of the NOIZEUS dataset with eight of the noise environments and was added with four levels: 0, 5, 10, and 15 dB of SNR values into the system.

It can be seen that, when considering the SNR at 0 dB, it was found that the algorithms ASDM-MFCC, PNCC, PLP, GFCC, and MFCC are given the average accuracy percentage rate of speech recognition are 61.24%, 38.33%, 20.83%, 13.74%, and 9.99%, respectively. The highest value was 61.24% and the second was 38.33% from the proposed ASDM-MFCC algorithm and the PNCC algorithm, respectively.

In the meantime, the smallest value is 9.99% and 13.74%, in the second order, which are derived from the traditional algorithms MFCC and GFCC respectively. In the consideration, the SNR level is 5 dB. Results show that each algorithm gives an average accuracy rate to increase. In other words, the ASDM-MFCC, PNCC, PLP, GFCC and MFCC algorithms provide the average accuracy rate are 92.91%, 89.99%, 49.99%, 39.99% and 28.74%, respectively.

And the highest value was the proposed ASDM-MFCC algorithm at 92.91%, and the smallest value was the MFCC algorithm at 28.74%.

Obviously, it is observed that at SNR levels of 10 and 15 dB, it seems that each algorithm still gives an average percentage accuracy rate of speech recognition is increasing. These values are equal to 99.16% and 100% at SNR levels of 10 and 15 dB, respectively. It is derived from the proposed ASDM-MFCC algorithm and the PNCC algorithm. This excludes the PLP, GFCC, and MFCC algorithms that yield 97.49%, 90.41%, and 97.91% at SNR levels of 15 dB.

From the results of the experiment, it can be seen that at the SNR level as 0 dB the traditional MFCC algorithms will be more effective when working with the ASDM technique that is the ASDM-MFCC algorithm. It gives the average accuracy percentage rate of speech recognition is increasing, from 9.99% to 61.24%, a 51.25% increase. The SNR level of 5 dB is increasing was from 28.74% to 92.91%, an 29.58%.

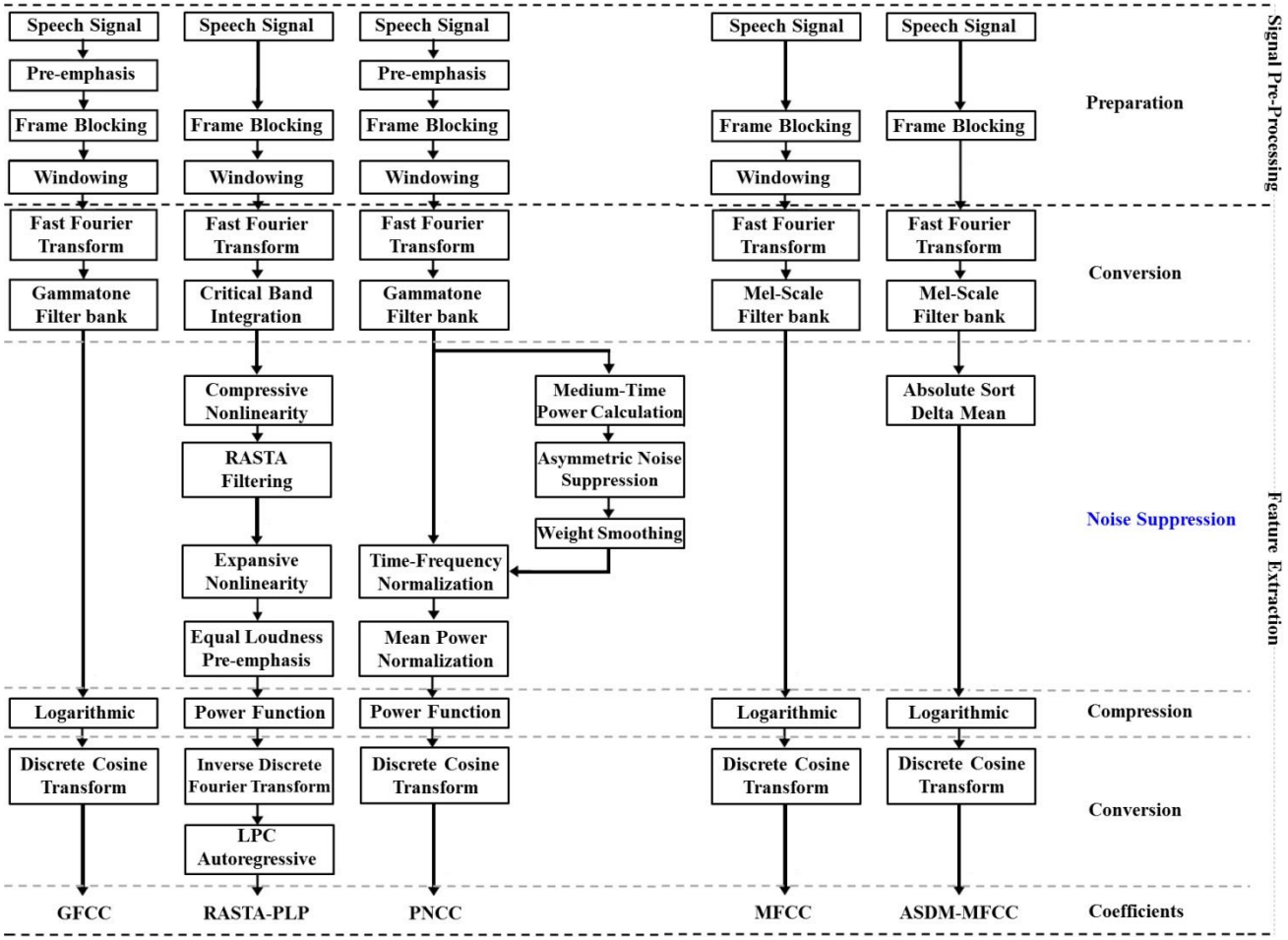


Figure. 5 Difference between the block diagram of GFCC, RASTA–PLP, PNCC, MFCC and proposed ASDM-MFCC feature extraction techniques

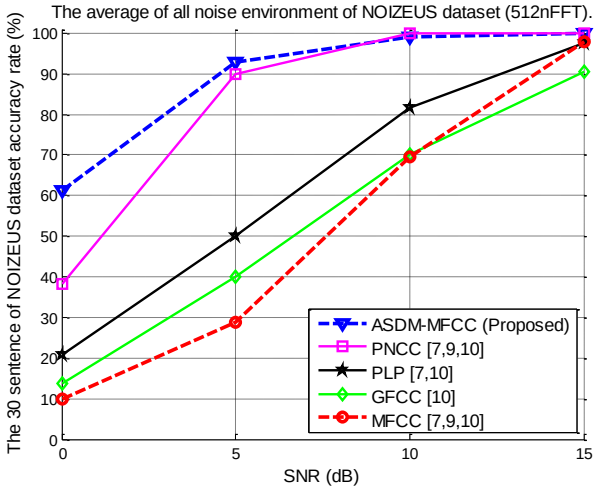


Figure 6 Dataset with eight of the noise environments added with SNR = 0, 5, 10, and 15 dB

increase of 64.17%. And at the SNR level of 10 dB increases was from 69.58% to 99.16%, an increase of Besides, the results of the experiment in Fig. 12 can be illustrated of the specific behavior of each of both algorithms a proposed and traditional. Of course, an affects the noise robustness is only possible to use the noisy speech corpus of the NOIZEUS dataset as follows. The ASDM-MFCC algorithm provides a good level of noise robustness characteristics, even at the SNR level of 0 dB. It is noticed that the PNCC algorithm is characterized by good noise robustness and increased quickly as the value of SNR level increases. The PLP and GFCC algorithms will increase at an increased SNR level, even though the SNR level is high at 15 dB. Noise robustness's still not as good as the previous both the ASDM-MFCC and the PNCC algorithms previously.

Fig. 7 is the experimental results of the average accuracy percentage rate of speech recognition in comparison with ASDM-MFCC, PNCC, PLP, GFCC and MFCC. The NOIZEUS dataset that with eight of the noise environments and was added at SNR= 0, 5, 10, and 15 dB. Obviously, for the NOIZEUS dataset of noise speech corpus, the proposed ASDM-MFCC algorithm achieves a high speech recognition

accuracy rate is more than the MFCC algorithm traditional previous.

Fig. 8 shows the performance of the proposed ASDM-MFCC algorithm in terms of average speech recognition accuracy rate. By comparing the ASDM-MFCC and the conventional MFCC algorithm, each environment of the NOIZEUS dataset is represented at SNR levels as 0, 5, 10, and 15 dB illustrated in Fig. 8 (a) – Fig. 8 (d), respectively. It can be seen that, the speech recognition accuracy rate of the ASDM-MFCC algorithm will be converge to a high rate is more than 45% all of noise environments and all of SNR levels of the NOIZEUS dataset. The proposed ASDM-MFCC algorithm achieves a high speech recognition accuracy rate is more than the MFCC algorithm traditional previous.

According to Table 1, Table 2, and Table 3, the results show the percentage of speech recognition accuracy rate in eight noise environments (airport, babble, car, exhibition, restaurant, station, street, and train) of the five algorithms using for feature extracting of the speech characteristics were ASDM-MFCC, PNCC, PLP, GFCC, and MFCC with added three SNR levels: 0, 5, and 10 dB.

In Table 1, at an SNR level of 0 dB the average accuracy percentage rate of ASDM-MFCC, PNCC, PLP, GFCC and MFCC algorithms are 61.24%, 38.33%, 20.83%, 13.74% and 9.99%, respectively. The MFCC algorithm needed to be improving the average accuracy percentage rate was 9.99%. Apparently, with the proposed ASDM-MFCC algorithm, the average accuracy percentage rate could be increased to 61.24%, a 51.258% increased.

In addition, at the SNR level of 0 dB, the ASDM-MFCC, PNCC, and PLP algorithms had the highest average accuracy percentage rate at the restaurant environment are 73.33%, 56.66% and 30.00%, respectively. At the same time, the GFCC and MFCC algorithms at the car and airport environments are 20.00% and 16.66%, respectively.

At the exhibition, car and car environments, the accuracy are 53.33% 6.66% and 3.33% which is the ASDM-MFCC PLP and MFCC algorithms, respectively. And the PNCC and GFCC algorithms at the train and station environments are 30.00% and 6.66%, respectively.

In Table 2, at an SNR level of 5 dB, the average accuracy percentage rate of the ASDM-MFCC, PNCC, PLP, GFCC and MFCC algorithms are 92.91%, 89.99%, 49.99%, 39.99% and 28.74%, respectively. By the MFCC algorithm of traditional, the average accuracy percentage rate was 28.74%, increased to 92.91%, an increase of 64.17%.

In Table 3, at SNR level of 10 dB, the average accuracy percentage rate of the ASDM-MFCC, PNCC, PLP, GFCC, and MFCC algorithms are 99.16%, 99.16%, 81.66%, 69.99%, and 69.58%, respectively. By the MFCC algorithm of traditional, to improve it previously, the average accuracy percentage rate was 69.58%, with the proposed ASDM-MFCC algorithm able to increase to 99.16%, an increase of 29.58%.

Table 4 is a summary of all 3 previous tables. From the experiment in eight noise environments and it has also been added to four levels of SNR is 0, 5, 10, and 15 dB as well. In order by descending is largest to smallest value, the details are as follows. The ASDM-MFCC, PNCC, RASTA-PLP, GFCC and MFCC algorithms are 88.32%, 81.87%, 62.49%, 53.53% and 51.55% respectively. Obviously, the ASDM-MFCC algorithm achieves the highest average accuracy percentage rate.

To calculate the difference rates, that increases the average accuracy when comparing with other algorithms. The difference rates are 6.45%, 25.83%, 34.79% and 36.77%. It is the minor PNCC algorithm, RASTA-PLP, GFCC algorithms, and the MFCC algorithms of traditional, respectively.

As mentioned, it can be seen that the speech recognition accuracy rate of all five algorithms is linear. It is proportional to the SNR level.

That is when the SNR level is increased from the values of 0, 5, 10, and 15 dB the higher the accuracy rate as well. The average accuracy percentage rate of speech recognition is increased from 61.24%, 92.91%, 99.16%, and finally to 100% at the SNR level of 15 dB.

3.3.3. Resolution of nFFT

Fig. 9 shows the behavior of each size of the nFFT as the four sizes of resolutions: 128, 256, 512 and 1024. And It is the average result of speech recognition accuracy rate. That is all of the noise environments of the NOIZEUS dataset of the proposed ASDM-MFCC and the conventional MFCC algorithm shown in Fig. 9 (a) and Fig. 9 (b), respectively. However, it is very difficult to define the size of the parameter, because it depended on the required work that either a short or a length of the speech data input.

4. Conclusions and future work

In this paper, we proposed a new technique of feature extraction method for obtaining speech feature extraction for noise robustness. The objective

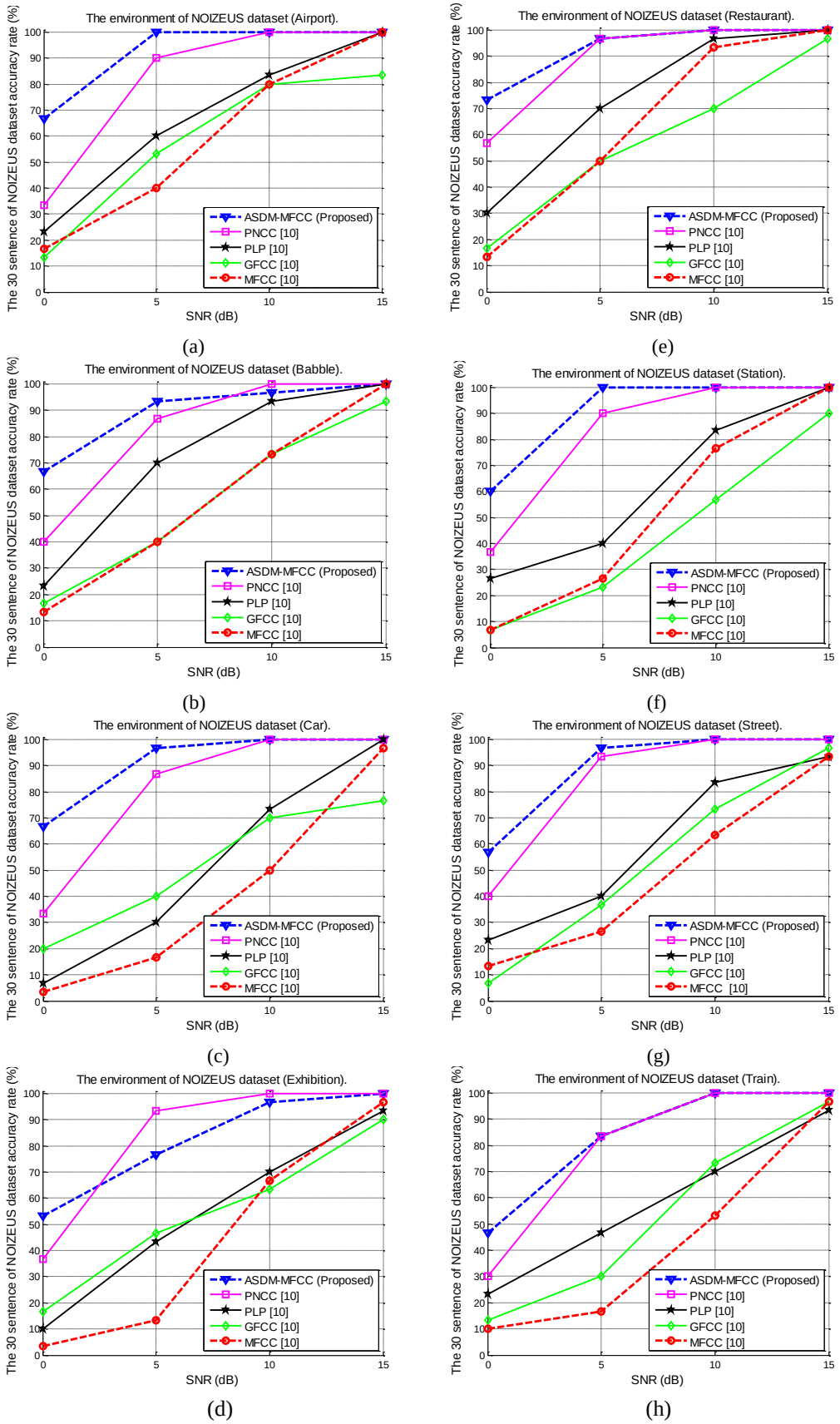


Figure. 7 Average accuracy percentage rate of speech recognition in comparison with ASDM-MFCC, PNCC, PLP, GFCC, and MFCC and the NOIZEUS database added with different SNR as: 0, 5, 10, and 15 dB at: (a) Airport, (b) Babble, (c) Car, (d) Exhibition, (e) Restaurant, (f) Station, (g) Street, and (h) Train

Received: April 24, 2021. Revised: May 20, 2021.

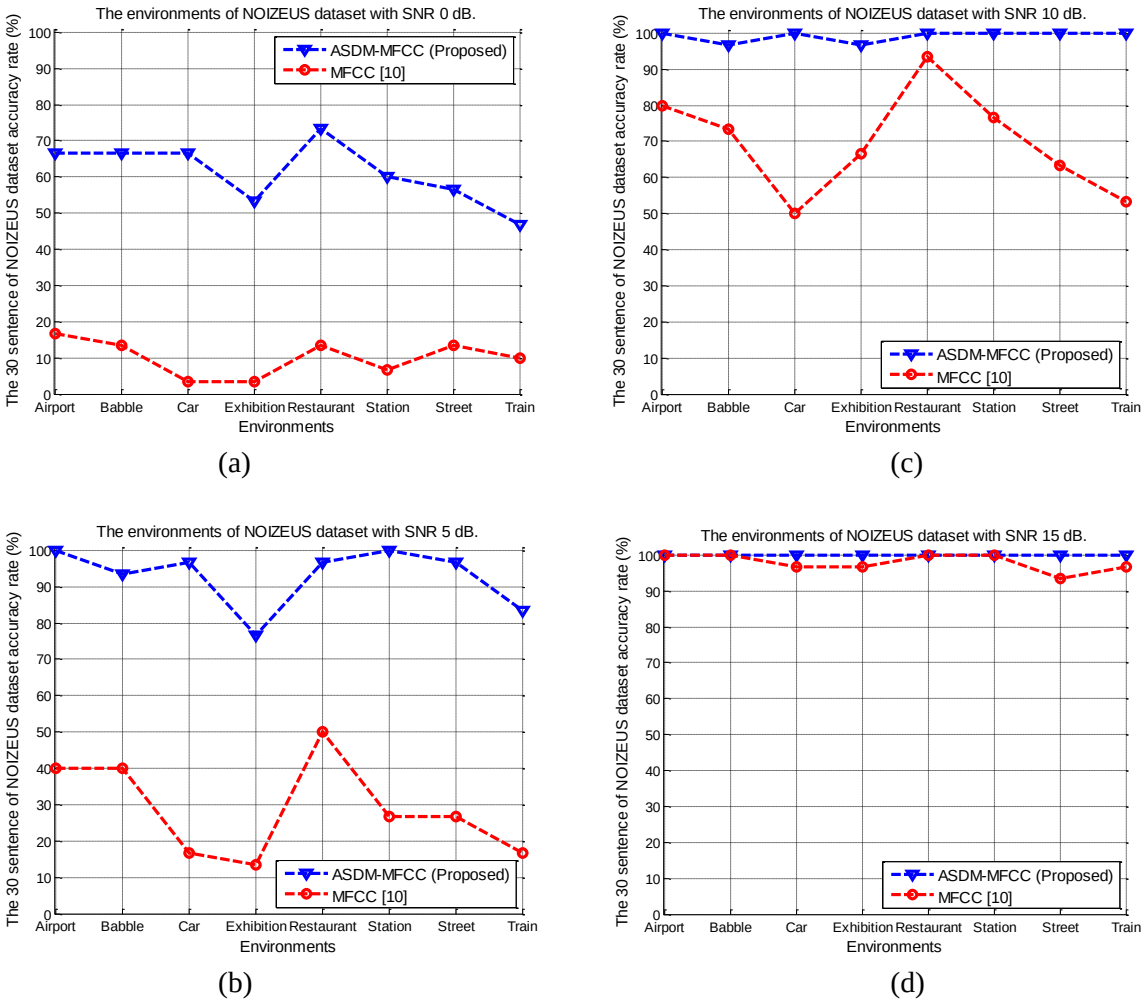


Figure. 8 Average resulted of speech recognition accuracy rate compared and should be listed as: (a) The SNR level is 0 dB, (b) The SNR level is 5 dB, (c) The SNR level is 10 dB, and (d) The SNR level is 15 dB

Table 1. The accuracy percentage of all noise environments with SNR 0 dB

Feature	Accuracy (%)								
	Airport	Babble	Car	Exhibition	Restaurant	Station	Street	Train	Average
ASDM-MFCC	66.66	66.66	66.66	53.33	73.33	60	56.66	46.66	**61.24**
PNCC [10]	33.33	40.00	33.33	36.66	56.66	36.66	40.00	30.00	**38.33**
RASTA-PLP [10]	23.33	23.33	6.66	10.00	30.00	26.66	23.33	23.33	**20.83**
GFCC [10]	13.33	16.66	20.00	16.66	16.66	6.66	6.66	13.33	**13.74**
MFCC [10]	16.66	13.33	3.33	3.33	13.33	6.66	13.33	10.00	**9.99**

Table 2. The accuracy percentage of all noise environments with SNR 5 dB

Feature	Accuracy (%)								
	Airport	Babble	Car	Exhibition	Restaurant	Station	Street	Train	Average
ASDM-MFCC	100	93.33	96.66	76.66	96.66	100	96.66	83.33	**92.91**
PNCC [10]	90.00	86.66	86.66	93.33	96.66	90.00	93.33	83.33	**89.99**
RASTA-PLP [10]	60.00	70.00	30.00	43.33	70.00	40.00	40.00	46.66	**49.99**
GFCC [10]	53.33	40.00	40.00	46.66	50.00	23.33	36.66	30.00	**39.99**
MFCC [10]	40.00	40.00	16.66	13.33	50.00	26.66	26.66	16.66	**28.74**

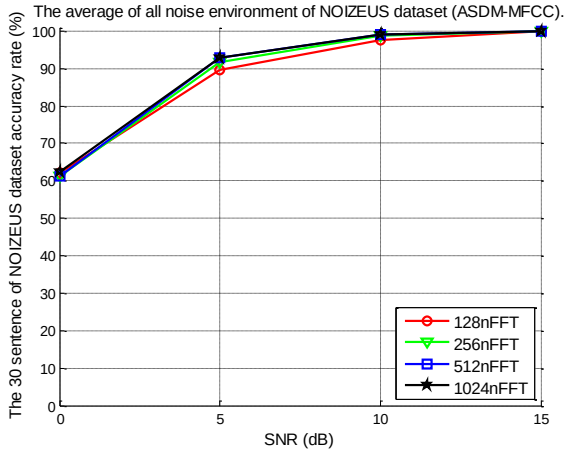
 DOI: 10.22266/ijies2021.0831.37

Table 3. The accuracy percentage of all noise environments with SNR 10 dB

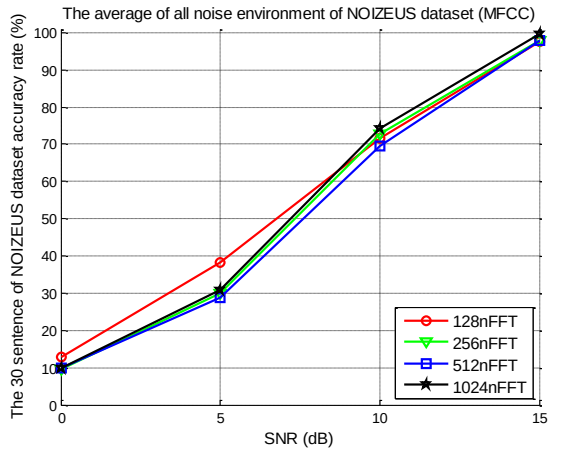
Feature	Accuracy (%)								
	Airport	Babble	Car	Exhibition	Restaurant	Station	Street	Train	Average
ASDM-MFCC	100	96.66	100	96.66	100	100	100	100	**99.16**
PNCC [10]	100	96.66	100	96.66	100	100	100	100	**99.16**
RASTA-PLP [10]	83.33	93.33	73.33	70.00	96.66	83.33	83.33	70.00	**81.66**
GFCC [10]	80.00	73.33	70.00	63.33	70.00	56.66	73.33	73.33	**69.99**
MFCC [10]	80.00	73.33	50.00	66.66	93.33	76.66	63.33	53.33	**69.58**

Table 4. The accuracy percentage of all algorithms

SNR (WGN)	Methods Accuracy (%)				
	ASDM-MFCC	PNCC	RASTA-PLP	GFCC	MFCC
15 dB	100	100	97.49	90.41	97.91
10 dB	99.16	99.16	81.66	69.99	69.58
5 dB	92.91	89.99	49.99	39.99	28.74
0 dB	61.24	38.33	20.83	13.74	9.99
Average	88.32	81.87	62.49	53.53	51.55



(a)



(b)

Figure. 9 The average result of speech recognition accuracy rate with different nFFT size used at 128, 256, 512 and 1024: (a) ASDM-MFCC algorithm and (b) conventional MFCC algorithm.

is to achieve satisfactory performance in speech recognition and speaker recognition. The main concept is known that one of the problems that can degrade the performance of the conventional MFCC algorithm is the property of extremely sensitive noise robustness conditions. We were developed from the MFCC to the DDM-MFCC algorithm previously. And now become the proposed algorithm. That is the absolute sort delta mean - Mel frequency cepstral coefficients called the ASDM-MFCC algorithm. It is modified to increase robustness against the different types of environmental noises.

The process consists of three methods: First, magnitude sorting of the spectral power in each frequency band of the speech signal. Second, finding the absolute delta value in the arrangement, and determining a value of mean is the last step. The purpose of the ASDM-MFCC algorithm is to require the noise robustness of the feature vector extracted Three main benefits of the proposed ASDM-MFCC algorithm is as: 1) a simple technique, 2) decreased the size of coefficients of the proposed algorithm and 3) increased the average accuracy percentage rate of a new algorithm of MFCC is a proposed ASDM-MFCC algorithm under the condition using of the NOIZEUS speech corpus dataset criterial.

The proposed ASDM algorithm can work with the MFCC algorithm and can be able to increase the capability and efficiency of the noise robustness. The ASDM-MFCC algorithm, that the SNR level as 0 dB it gives the average accuracy percentage rate ofspeech recognition is increasing, from 9.99% to 61.24%, a 51.25% increase.

In the case of car and station noise at SNR 5 dB, the proposed method outperforms the MFCC and gammatone frequency cepstral coefficient (GFCC) methods by 80% and 76.67%, respectively. Moreover, the average accuracy percentage rate of speech recognition of the proposed algorithm is higher than the relative spectral (RASTA)-perceptual linear predictive (PLP) and power-normalized cepstral coefficients (PNCC) methods in the case of airport, babble, car, station and street noise. It is enhanced by 64.17% in comparison to the MFCC method at SNR 5 dB, while it is improved by 22.91% in comparison to the PNCC method at SNR 0 dB.

The experimental results show that the proposed algorithm can provide significantly improvements in the recognition accuracy at low signal to noise ratios. Obviously, some speech recognition experiments of the proposed ASDM-MFCC algorithm are more robust than previously traditional ones in noise conditions of the NOIZEUS noisy speech corpus dataset. The proposed ASDM-MFCC algorithm achieves a high speech recognition accuracy rate is more than the MFCC algorithm traditional previous.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, Anuruk Nosan and Suchada Sitjongsataporn; methodology, Anuruk Nosan; software, Anuruk Nosan; validation, Anuruk Nosan and Suchada Sitjongsataporn; writing—original draft preparation, Anuruk Nosan; writing—review and editing, Suchada Sitjongsataporn; visualization, Suchada Sitjongsataporn; supervision, Suchada Sitjongsataporn.

References

[1] C. Li, T. Jiang, S. Wu, and J. Xie, "Single-channel speech enhancement based on adaptive low-rank matrix decomposition", *IEEE Access,* Vol. 8, pp. 37066-37076, 2020.

[2] A. Watile, V. Alagdeve, and S. Jain, "Emotion recognition in speech by MFCC and SVM", *International Journal of Science, Engineering and Technology Research (IJSETR)*, Vol. 6, No. 3, pp. 404-407, 2017.

[3] A. Nosan and S. Sitjongsataporn, "Descend-delta-mean algorithm for feature extraction of isolated Thai digit speech", In: *Proc. of the IEEE International Electrical Engineering Congress (iEECON)*, Cha-am, Thailand, pp. CIT13-CIT16, 2019.

[4] D. Vazhenina and K. Markov, "End-to-end noisy speech recognition using Fourier and Hilbert spectrum features", *Electronics Journal,* Vol. 9, No. 7, p. 1157, 2020.

[5] A. Pervaiz, F. Hussain, H. Israr, M. A. Tahir, F. R. Raja, N. K. Baloch, F. Ishmanov, and Y. B. Zikria, "Incorporating noise robustness in speech command recognition by noise augmentation of training data", *Sensors Journal*, Vol. 20, p. 2326, 2020.

[6] T. Fux and D. Jouvet, "Evaluation of PNCC and extended spectral subtraction methods for robust speech recognition", In: *Proc. of the European Signal Processing Conference (EUSIPCO)*, Nice, France, Aug, pp. 1416-1420, 2015.

[7] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for robust speech recognition", *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, Vol. 24, No. 7, pp. 1315-1329, 2016.

[8] R. Thiruvengatanadhan, "Speech/music classification using power normalized cepstral coefficients and K-means", *International Journal of Advanced Computer Science and Technology*, Vol. 6, No. 1, pp. 13-17, 2016.

[9] B. Alzahra, K. Kyungdeuk, and K. Hanseok, "Bird sounds classification by combining PNCC and robust Mel-log filter bank features", *Journal of the Acoustical Society of Korea*, Vol. 38, No. 1, pp. 39-46, 2019.

[10] M. Tamazin, A. Gouda, and M. Khedr, "Enhanced automatic speech recognition system based on enhancing power-normalized cepstral coefficients", *Applied Sciences Journal*, Vol. 9, No. 10, p. 2166, 2019.

[11] M. A. Abed. A. Ali, and H. Alasadi, "A hybrid model of MFCC/MSFLA for speaker recognition", *American Journal of Computer Science and Engineering*, Vol. *2*, pp. 32-37, 2015.

[12] L. Juvela, B. Bollepalli, X. Wang, H. Kameoka, M. Airaksinen, J. Yamagishi, and P. Alku, "Speech waveform synthesis from mfcc sequences with generative adversarial networks", In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5679-5683, 2018.

[13] A. Nosan and S. Sitjongsataporn, "Speech recognition approach using descend-delta-mean and MFCC algorithm", In: *Proc. of IEEE International Conference on Electrical Engineering/Electronics, Computer,*

Telecommunications and Information Technology (ECTI-CON), Pataya, Thailand, pp. 380-383, 2019.

[14] S. Lee, I. Seo, J. Seok, Y. Kim, and D. S. Han, "Active sonar target classification with power-normalized cepstral coefficients and convolutional neural network", *Applied Sciences Journal*, Vol. 10, No. 23, p. 8450, 2020.

[15] S. C. Lee, J. F. Wang, and M. H. Chen, "Threshold-based noise detection and reduction for automatic speech recognition system in human-robot interactions", *Sensors Journal*, Vol. 18, No. 7, pp. 1-12, 2018.

[16] K. Naithani and A. Semwal, "Various techniques used for English language speech recognition: A review", *International Journal of Scientific & Technology Research*, Vol. 8, pp. 3396-3405, 2019.

[17] L. Nahma, P. C. Yong, H. H Dam, and S. Nordholm, "An adaptive a priori SNR estimator for perceptual speech enhancement", *EURASIP Journal on Audio, Speech, and Music Processing*, Vo. l, No. 1, p. 7, 2019.

[18] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "ViSQOL: an objective speech quality model", *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 1, No. 13, pp. 1-18, 2015.

[19] S. Prongnuch and S. Sitjongsataporn, "Thai Voice-Controlled Analysis for Car Parking Assistance in System-on-Chip Architecture", *Advances in Technology Innovation (AITI)*, Vol. 5, No. 4, 2020.