Jurnal
Teknologi

Full Paper

THE COMPARATIVE PERFORMANCE EVALUATION OF WINDOW FUNCTIONS UNDER NOISY ENVIRONMENT FOR SPEECH RECOGNITION

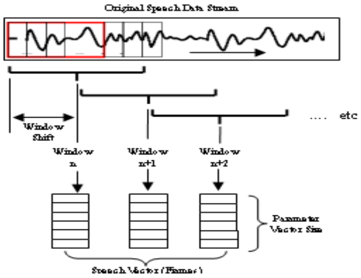
Article history
Received
14 June 2015
Received in revised form
10 November 2015
Accepted
3 Jan 2016

Syifaun Nafisah[a,b*], Oyas Wahyunggoro[a], Lukito Edi Nugroho[a]

[a]Department of Electrical Engineering, Universitas Gadjah Mada, Yogyakarta, Indonesia
[b]Department of Library and Information Science, UIN Sunan Kalijaga Yogyakarta, Indonesia

*Corresponding author
syifaun.nafisah@mail.ugm.ac.id

Graphical abstract



Abstract

The accuracy and user acceptance of speech recognition systems is increasing in the last few years especially for automated identification and biomedical applications. In implementation, it works based on the feature of utterance that will be recognized through a feature extraction process. One process in feature extraction is windowing that is done for minimizing the disruptions at the first and last of the frame. Basically, many window functions exist such as rectangular window, flat top window, hamming window, etc, but in the real application only hamming or Hanning function that are usually used as a function in the windowing. This article will analyzed the performance of all of window functions to prove the performance of those function. The method that was used are mel-frequencies cepstral coefficients (MFCCs) as feature extractor technique and back propagation neural networks (BPNNs) as classifier. The result shows that it can produce an accuracy at least 99%. The optimal accuracy up to 99.86% is achieved using rectangle window with the duration of process is 15.47 msec. This results show the superior performance of rectangle window as reference to recognize an isolated word based on speech.

Keywords: Speech recognition; MFCC; BPNNs; windowing, accuracy

1.0 INTRODUCTION

Automatic speech recognition (ASR) can be defined as the computer-driven transcription of spoken language into readable text in real time [1]. The goal of this system is to accurately and efficiently convert a speech signal into a text message transcription of the spoken words. The working principle of ASR functions is such as a pipeline that converts PCM (Pulse Code Modulation) digital audio from a sound card into recognized speech [2].

One example of ASR application is verification of Quranic verses. For this application, MFCC was used as feature extraction techniques and HMMs was used to modelling a language and dictionary and to avoid some system complexity in the framework. The result show that it is expected that the existing models will be enhanced and improved for more accuracy in recognition of Quran verses especially online [3]. ASR can also be used as an educational tool to solve the problem of lacking of teaching toolkits, especially for quranic teaching. To improve the performance of this tools, a few algorithms have been implemented in the experiments phase. Based on the experiments, the performance of the tools using Linear Predictive Coding Cepstrum (LPCC) achieved the accuracy of 100%, while the MFCC managed to get 99.92% [4].

In real applications, ASR is a complex system. It is embedded in large architecture that involves other modules. There are some related studies about the development method of ASR. The first method is the development of free model of sentence classifier for automatic extraction of topic sentences using corpus-based classification for constructing a sentence classifier as the model for finding optimal combination of features. These method was trained using Support Vector Machine (SVM). The performance of SVM achieved up to 80.68% [9]. In this study, an extreme learning machine is also used on feature extraction process and have a result in a higher accuracy up to 92.1% on average [6].

Kurniawan et al (2011) is also tried to improve the performance of ASR using contour analysis and neural network validation. This algorithm is performed to generate pre-segmentation by analyzing the vertical contours from right to left. Based on experiments that is conducted to 1,902 patterns, the average of recognition is 82.63% that is declared as a valid result [8]. The normalization process is one method that is ever used by researcher to improve the performance of ASR. One of the efforts is by utilizing fuzzy logic. The result shows that normalization using fuzzy logic has a promising and encouraging performance to improve the performance of the system until 86.36% [5].

Refer to the previous studies, each part of the process in the ASR has contributed to improve the performance of system. In the next study, the performance of ASR will be enhanced through the windowing with a consideration, speech signals is non-stationary signal where properties change quite rapidly over time. This phenomenom is natural but makes the transformation process using Discrete Fourier Transform (DFT) or autocorrelation to be impossible. To solve this problem, most of speech processing is taking short windows and processing them. In the windowing, there are seventeen type of window function: Bartlett, barthannwin, blackman, blackmanharris, bohmanwin, chebwin, flattopwin, gausswin, hamming, hanning, Kaiser, nuttallwin, parzenwin, rectwin. taylorwin, tukeywin and triangle. In the real application, the function that is ever used is hamming and hanning window. The study will be prove the performance of all functions to increasing the accuracy of system. The function that is produced the highest accuracy will be recommended as a reference that can be used in the ASR application.

2.0 RELATED WORKS

The position of this topic has been studied and revisited by many researchers. Favero [10] has explored the performance of ASR based on well understood of window functions using Hanning, Hamming, Blackman and Gaussian window by compares four modulated wavelets. The accuracy of each function is: Hanning is 67.65%, Hamming is 66.2%, Blackman is 68.5% and Gaussian is 68.5%. It is mean that Blackman and Gaussian window 1% better than Hanning window and 2.3% better than Hamming Window.

The efforts to improve the performance of ASR is also performed using non-standard window [11]. A non-standard windows that is referred in this study is digital filtering approach to the design of finite window sequences with linear and nonlinear phase response. The result shows that a non-standard window sequences can contribute to greater ASR robustness until 83.75% better than the Hamming window.

Asymmetric windows also has been used to boost the performance of ASR by combining the Hamming window with the other function [12]. The simulation results show that the filter designed using modified window function is more efficient than Hanning and Hamming window function. Based on the experiments, the equivalent noise bandwidth of Hamming is 1.3783, Hanning is 1.5238 and Blackman is 1.7542. It can be observed that the response of Blackman window are more smooth and perfect than that of the Hamming and Hanning window.

Actually, the windows functions that can be used in speech recognition applications not only Hamming, Hanning, Blackman and Gaussian, such as widely studied by many researchers. At least there are other types of window functions which can be utilized to improve the performance of this systems. To determine the influence of window functions, all of the types will be used in this study. Each functions will be analyzed to prove the effect of each function to the performance of ASR, then the result will be compared. It is hope that the performance of each function can be assessed objectively.

3.0 MATERIAL AND RESEARCH METHOD

3.1 Data Speech Selection

In this research , the speech data was taken from 45 speakers that were divided into three groups, the 1st group consisting 16 speakers, 9 male and 7 female speakers will used in the training phase, 2nd group consisting 14 speakers, 11 male and 3 female speakers is for testing phase and 3rd groups consisting 5 male and 10 female speakers. The speaker ranging from 15-22 years of age. All of the speakers will be taken their voice through the recording process using Cool Edit 2.0 that was performed in a sound treated audiometric booth using vocal microphone PG48-LC, mini mixer EuroRack UB1002FX. The microphone had a flat frequency response ranging from 10 Hz to 20 KHz. The recoded speech was loaded into the computer through an M-Audio 32-bit DIO 2448 input/ouput card. Throughout the experiment, the mouth to microphone distance was carefully maintained at 1 inch from the left hand corner of the mouth. The data was segregated and individually stored as *.wav files. The Speakers were asked to utter a set of 50 words in a normal manner which the utterance was repeated 12 times in a low-noise environment to reduce acoustic interference. From this process, there are 7008 wave files of data speech that will be stored in database such as described in Table 1, then the speech data were devided into three groups such as presented in

Table 2. The scenario of evaluation will be prepared using the combination such as shown in Table 3.

Table 1 Final files of data speech

The Speech Data	Utterances
Female	3942
Male	3066
Total	**7008**

Table 2 Group of dataset

The Speech Data	Training (Dataset I)	Testing 1 (Dataset II)	Testing 2 (Dataset III)
Male	3942	2190	3066
Female	3066	2044	4088
Total	**7008**	**4234**	**7154**

Table 3 Scenario of evaluation

Scenario	Data Training	Data Testing	Scenario	Data Training	Data Testing
1	Dataset I	Dataset I	6	Dataset II	Dataset III
2	Dataset I	Dataset II	7	Dataset III	Dataset I
3	Dataset I	Dataset III	8	Dataset III	Dataset II
4	Dataset II	Dataset I	9	Dataset III	Dataset III
5	Dataset II	Dataset II			

In this study, MATLAB was used to generate the spectogram of signal such as shown in Figure 1.

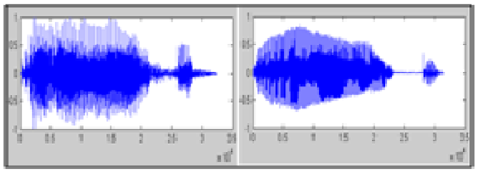


Figure 1 Spectogram of male and female speaker

3.2 Audio pre-processing and speech recognizer

The goal of a speech recognizer is to take a continuous speech waveform as an input to produce a transcription of the words being uttered. The steps are such as described in Figure 2.

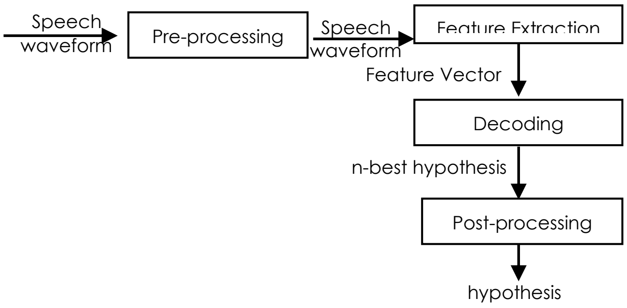


Figure 2 Schematic outline of a speech recognition process

Because of speech waveform is a non-stationary signal where properties change quite rapidly over time, windowing should be confronted using appropriate window function such as described in the following section.

3.3 Window function in speech recognition

The short time Fourier Transform (STFT) is used as a frequency analysis method where the signal was divided into short frames of N samples. Final windowed values *x(n)* in each frame are obtained by multiplying signal s(n) with a non-zero window sequence *w(n)* using Equation (1).

$$x(n) = s(n) * w(n), n = 0, \ldots, N-1 \qquad (1)$$

The frame length N must be short because of the rapidly changing spectrum of *s(n)*. The windowed spectrum of *X(e^jω)* is calculated as the frequency response of *x(n)*. *X(e^jω)* is also equal to the convolution integral of the Fourier Transform (FT) of the window sequence *W(e^jω)*and the FT of the original signal *S(e^jω)* such as presented in Equation (2).

$$X(e^{j\omega}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(e^{j\theta})W(e^{j(\omega-\theta)})d\theta \qquad (2)$$

The frequency response $X(e^{j\omega})$ is obviously influenced by *W(e^jω)*. In addition, it is typical for speech recognition that only the magnitude frequency response of signal samples in the frame is kept for further processing. In this study, a window function *w(n)* was selected to compute magnitude response I*X(e^jω)*I is as near as possible to the real magnitude response I*S(e^jω)*I. The overall process is illustrated in Figure 3 which shows the sampled waveform being converted into a sequence of parameter blocks.

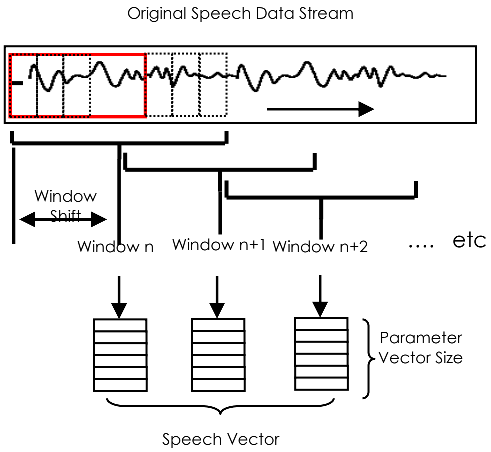


Figure 3 Speech encoding process

As mentioned above, there is no theoretical reason to believe about the best function of window that will produced the optimally speech recognition system. So, this research will explore the performance of each function of the window to get a window function with an optimal performance.

3.4 Feature Extractor

The feature extractor that is used in this study is mel-frequencies cepstral coefficients (MFCCs). MFCCs is one of the standard method for feature extraction because its sensitivity to noise due to its dependence on the spectral form [16]. This methods is utilize an information in the periodicity of speech signals that could be used to overcome this problem, although speech also contains a periodic content [18]. The step of MFCCs that was used in this study is presented in Figure 4.

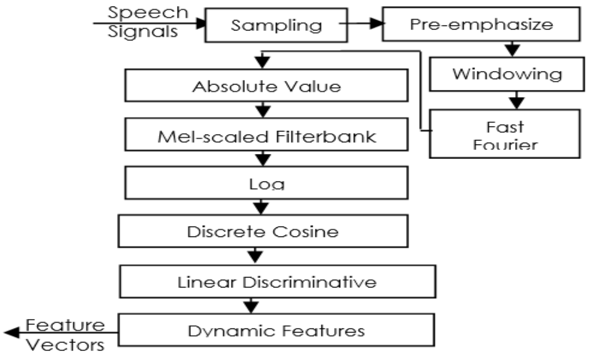


Figure 4 Complete pipeline MFCCs

3.6 Post-Processing

The post processing step of this study is send the features to the classifier method using 50 input nodes, 10 hidden nodes, and 50 output nodes. The architecture of classifier shown in Figure 5.

4.0 PERCEPTUAL EXPERIMENTS AND RESULT

The main purpose of this work is evaluating the contribution of window function using certain time-frequency properties to optimize the performance and inherent robustness in ASR. In this study, the performance analysis of window functions include spectral analysis, filter design, and beamforming that can be used to try to extract speech sources. The window functions that will be evaluated in this study are such as mentioned at the end of section 1.0.

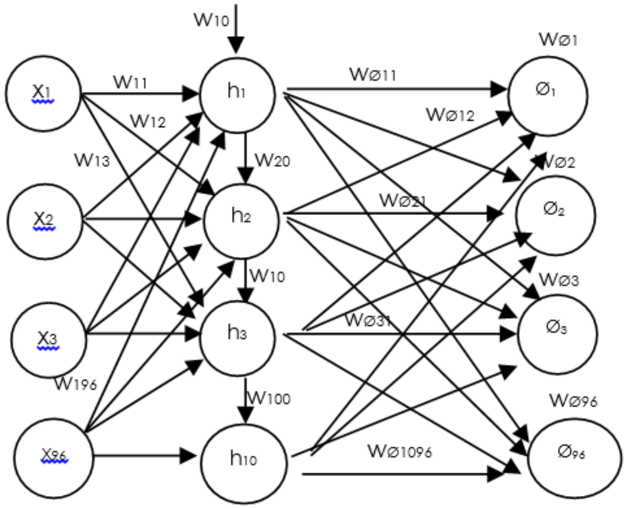


Figure 5 96-10-50 BPNNs Architecture

The analysis of window function's effect on the performance of ASR is conducted using MATLAB 2010a as software assistance. Before evaluating the effect of windows function on the performance of ASR, this software has been tested beforehand. Software design process begins with the speech recording. After the reference data is recorded, the next step is performed the segmentation process based on syllable, for example, the word 'Denpasar' is segmented into syllables 'Den' ‘Pa’ and 'Sar'. The list of words that is used in the software testing process using 50 words such as presented in Table 4.

Table 4 List of words

No	the variation of words				
	Family Members (1)	**Numbers (2)**	**The name of city (3)**	**Noun (4)**	**Human body (5)**
1	a-yah	sa-tu	Ja-kar-ta	ro-ti	ma-ta
2	i-bu	du-a	Ban-dung	na-si	gi-gi
3	a-dik	ti-ga	Se-ma-rang	ke-ju	pi-pi
4	ka-kak	em-pat	Yog-ya-kar-ta	bo-la	hi-dung
5	sa-ya	li-ma	Su-ra-ba-ya	to-pi	ta-ngan
6	ne-nek	e-nam	Den-pa-sar	bu-ku	ka-ki
7	ka-kek	tu-juh	Ma-ka-sar	me-ja	le-her
8	bi-bi	de-la-pan	Me-dan	kur-si	pung-gung
9	pa-man	sem-bi-lan	Pon-ti-a-nak	pin-tu	ping-gang
10	sau-da-ra	se-pu-luh	Jem-ber	sen-dok	bi-bir

After the segmentation process, the segmented file will be taken the feature based on the frequency domain using MFCCs algorithm and for classification of the pattern of speech signal, all of the data was trained using the BPNNs algorithm. The number of iterations use an epoch that was set as a variable. The variable was needed because the system iterated until all the errors were below the threshold of 0.5, or until the number of iterations reached 1,000,000. An epoch was trained with a fixed training set until all the errors produced by the data pairs in the training set were below a threshold. Every epoch comprised a variable number of backpropagation iterations.

In the experimental phase, this study has been conducting research on the effect of window functions on the accuracy to the 45 respondents. The entire respondent had been tested to measeure the performance of system using the variance of window functions. Of that numbers, 30 respondents are speakers who fill speech data as data reference in the database, while 15 respondents are the respondents who are not recognized by the database. The procedure test of these respondents is using the scenario such as already shown in Table 3. The step of the experimental phase is described in Figure 6.

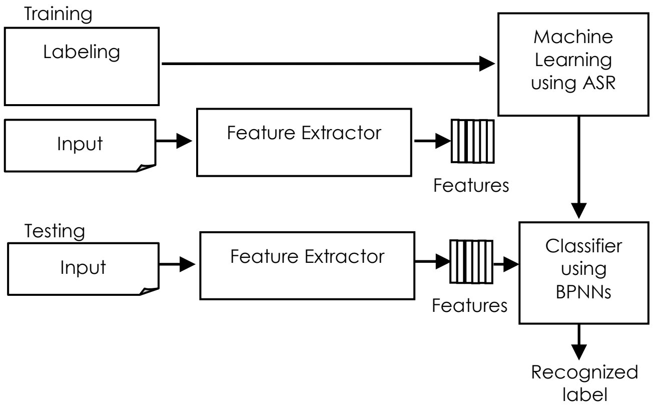


Figure 6 Experimental phase

Before the influence of window function has been tested, the performance of the combination algorithm between MFCCs and BPNNs was evaluated. Based on the result of the experiments, combination method between MFCCs and BPNNS algorithm using hamming window produces the 99% accuracy with duration of process is 27.4 msec on average. After the performance of the method can certainly work, the procedure of the evaluation is the data testing will be tested using a function of window and the system will be calculated the accuracy and then the function of window will be changed using another function and then all of the function is tested, this study will compare the accuracy that is produced by each function.

The function of window that is produced the optimum accuracy will recommanded as a function of window for isolated word recognition system. The tests that have been done repeatedly to 45 respondents using the scenarios such as set out in Table 3 and the result of the experiments is produce the average of accuracy such as shown in Table 5.

The performance of can also be seen from the curve formed. Based on experiments that have been carried out, every function of windows is produce different characteristic of curves. In this exposure, the curve that is presented in this study is in the amplitude-frequency such as described in Figure 7.

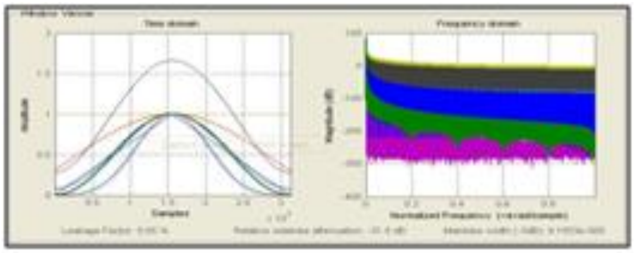


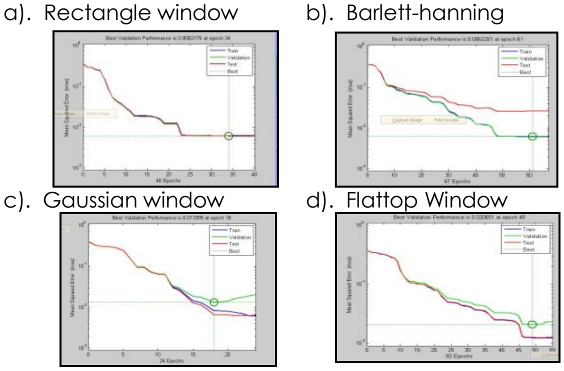
Figure 7 Emplitude-frequency characteristic curve of all of window function

Table 5 The performance based on function of window

Function of window	The average of accuracy (%)	Function of window	The average of accuracy (%)	Function of window	The average of accuracy (%)
Bartlett	99.07	Flattopwin	99.74	parzenwin	99.08
barthannwin	99	Gausswin	99	rectwin	99.86
blackman	99	Hamming	99	taylorwin	99
blackmanharris	99.11	Hanning	99	tukeywin	99
bohmanwin	99	Kaiser	99.40	triang	99.07
chebwin	99.12	nuttallwin	99.09		
Average		**99.15**			

5.0 DISCUSSION

Based on Table 5, after repeated random testing to 45 respondents, the average of accuracy the proposed method using various types of windows functions is 99.15%. An optimum accuracy was achieved using rectangle (rectwin) with the average of accuracy is 99.86% and the lowest accuracy is 99%. Based on the results, it was seen that rectwin function generates an optimum recognition than other functions. These results can also be seen from the graph that was formed in the process of speech recognition using BPNNs such as illustrated in Figure 8.


a). Rectangle window

b). Barlett-hanning

c). Gaussian window

d). Flattop Window

Figure 8 Performance of ASR

In Figure 8, It is clearly visible that the chart is formed by rectangle window are in one chart. It is differs from the graph that was formed by the other functions, that the graph are not in a one chart. The reason why rectwin can produces an optimum accuracy better than other function is this function produces the lowest noise at around 1:00 BINS. But unfortunately, this function gives lowest level of sidelobe than other function where the low level of sidelobe causes the amount of spectral leakage that occurs in the process of feature extraction. So, to overcome these leaks, this research using an overlapping windowing system method. The use of overlapping's method on a windowing process is also proven to reduce spectral leakage in the feature extraction that It can improve the accuracy of system. An optimum performance of rectwin is also can be proved by comparing the performance of the system with some research that has been done by other researchers before such as presented in Table 6.

Table 6 Comparison performance with other research

Method	The name of function	Parameter of measurement	
		Accuracy (%)	Noise Reduction (BINS)
Using default function of window [10]	1. four type of window function		
	a. Hanning Window	67.65	-
	b. Hamming Window	66.2%	-
	c. Blackman Window	68.5%	-
	d. Gaussian Window	68.5%	-
digital filtering [11]	2. Non-standard window	83.75%	-
Noise bandwidth reduction [12]	3. Asymmetric Window		
	a. Hanning Window	-	1.3783
	b. Hamming Window	-	1.5238
	c. Blackman Window	-	1.7542
Proposed method (Windowing + Feature extraction+ Classifier)	4. Default window function		
	a. Other window function	99.15 (on average)	1.38
	b. rectangle window	99.86	1.0

6.0 CONCLUSION

The conclusion that can be drawn from the explanation of method and the results above is the windowing function can be used to improve the performance of ASR, but this function can't work alone maximally without supported by other processes. Based on results, the rectwin function was proved have a better performance than the other functions. Furthermore, to reinforce the results of this study, we plan to test this model using sentences.

Acknowledgement

We would like to thank to Ministry of Religious Affairs of the Republic of Indonesia and Doctoral program Universitas Gadjah Mada.

References

[1] Stuckless, R. 1994. Developments in real-time speech-to-text communication for people with impaired hearing. In M. Ross, *Communication access for people with hearing loss*. 197-226. Baltimore, MD: York Press.

[2] Rabiner, R. L., & Juang, B. H. 2004. *Statistical Methods for the Recognition and Understanding of Speech.* Rutgers University and the University of California, Santa Barbara; Georgia Institute of Technology, Atlanta.

[3] Mohammed, A., Sunar, M. S., & Hj Salam, Md. S. 2015. Quranic Verses Verification using Speech Recognition Techniques. *Jurnal Teknologi (Sciences & Engineering)*. 73(2): 99–106.

[4] Sze, H. K., & Shaikh Salleh, S. H. 2004. Design of Educational Software for Automatic Speech Recognition (ASR) Techniques. *Jurnal Teknologi (Sciences & Engineering)*. 40 (D):133–144.

[5] Suyanto, & Putro, A. E. 2014. Automatic Segmentation of Indonesian Speech into Syllables using Fuzzy Smoothed Energy Contour with Local Normalization, Splitting, and Assimilation. *Journal of ICT Research and* Applications. 8(2): 97-112.

[6] Hardy, & Cheah, Y.-N. 2013. Question Classification Using Extreme Learning Machine on Semantic Features. *Journal of ICT Research and* Applications. 7(1): 36-58.

[7] Abu-Ain, T., Abdullah, S. N., Omar, K., Abu-Ein, A., Bataineh, B., & Abu-Ain, W. 2013. Text Normalization Method for Arabic Handwritten Script. *Journal of ICT Research and Applications.* 7(2): 164-175.

[8] Kurniawan, F., Mohd. Rahim, M. S., Sholihah, N., Rakhmadi, A., & Mohamad, D. 2011. Characters Segmentation of Cursive Handwritten Words based on Contour Analysis and Neural Network Validation. *Journal of ICT Research and Applications.* 5(1): 1-16.

[9] Khodra, M. L., Widyantoro, D. H., Aziz, E. A., & Trilaksono, B. R. 2011. Free Model of Sentence Classifier for Automatic Extraction of Topic Sentences. *Journal of ICT Research and Applications.* 5(1): 17-34.

[10] Favero, R. F. 1994. Comparison Of Mother Wavelets For Speech Recognition. *International Conference Speech Science and Technology*. 336-341.

[11] Rozman, R., & Kodek, D. M. 2003. Improving Speech Recognition Robustness Using Non-Standard Windows. *European Science Fiction Convention.* Ljubljana, Slovenia.

[12] Rozman, R., & Kodek, D. M. 2007. Using Asymmetric Windows In Automatic Speech Recognition. *Elsevier Speech Communication*. 268–276.

[13] Rajput, S. S., & Bhadauria, D. S. 2012. Comparison of Band-stop FIR Filter using Modified Hamming Window and Other Window functions and Its Application in Filtering a Mutitone Signal. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET).* 1(8): 325-328.

[14] Podder, P., Khan, T. Z., Khan, M. H., & Rahman, M. M. 2014. Comparative Performance Analysis of Hamming, Hanning and Blackman Window. *International Journal of Computer Applications.* 96(19): 1-7.

[15] Verma, A. R., & Kumar, S. A. 2012. A Comparative Study of Performance of Different Window Functions for Speech Enhancement. *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS)*. 993-1002. Jaipur, Rajasthan, India : Springer.

[16] Tiwari, V. 2010. MFCC And Its Applications In Speaker Recognition. *International Journal on Emerging Technologies*. 19-22.

[17] Furui, S. 2000. *Digital Speech Processing: Synthesis, and Recognition* (2nd ed.). CRC Press.

[18] Motlíček P. 2002. Feature Extraction in Speech Coding and Recognition, Report, Portland, to research, data, and theory. Belmont, CA: Thomson/Wadsworth, 2003 US, *Oregon Graduate Institute of Science and Technology*. 1-50.