International Journal of Engineering & Technology, 7 (4.11) (2018) 231-235

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Speech Enhancement based on 2D Gabor Filters for Arabic Phoneme Spoken by Malay Speakers

Ali Abd Almisreb[2], Nooritawati Md Tahir[1]*, Ahmad Farid Abidin[1], Norashidah Md Din[2]

[1]Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Selangor, Malaysia
[2]Institute of Energy Infrastructure, Universiti Tenaga Nasional (UNITEN), 43000 Kajang, Selangor, Malaysia
**Corresponding author E-mail: nooritawati@ieee.org*

Abstract

In this paper, a speech enhancement method using 2D Gabor filter is proposed. The proposed filter is used to enhance Arabic phoneme speech signals that have been recorded under control environment namely indoor room recording. All the phoneme signals are spoken by Malay speakers and considered as non-native Arabic speakers. Firstly, corrupted speech signals by noise must be enhanced before further processing. The effectiveness of the suggested approach is evaluated in compare with Wiener filter. It is proven that the proposed 2D Gabor filters performed appropriately for speech enhancement purpose at different wavelengths.

Keywords: *2-D Gabor filters; Arabic; Malay; Wiener filter.*

1. Introduction

Due to the importance of speech enhancement in many applications such as communications, coding systems, hearing aids, aircraft cockpits, automatic speech recognition systems and forensics are still considered as one of relevant and important research areas to be explored. As we know, speech enhancement aims to reduce or eliminate the unwanted data in a speech waveform with the aim of increasing the acceptability, clearance and intelligibility of the speech signal but without degrading the original signal. Speech signals noises can be classified into several types according to the characteristics of the time and frequency domain, narrow band noise, band limited white noise, colored noise, impulse noise and transient noise pulses. On the other hand, speech enhancement can be categorized into two classes: single-channel and multi-channel approaches. Single-channel enhances the intelligibility and quality of speech, whilst multi-channel shows the ability to improve the quality and intelligibility of the speech using spectral and spatial details of both speech and noise [1]. Thus, in order to solve speech enhancement challenges, many algorithms have been suggested. For instance, conventional method of speech enhancement using spectral subtraction was proposed by [2], followed by enhancement method based on Wiener filter [3, 4] and minimum-mean square error approach [5]. In addition, there are also speech enhancements algorithms that was suggested based on sub-band method. The main usage of sub-band adaptive filters is to identify the response of very long impulse, but these filters have low convergence. Also, it can be used to identify the linear systems depending on its impulse responses [6]. Furthermore, the main principle of speech enhancement approaches via Discrete Wavelet Transform coefficients (DTW) thresholding the noisy speech is the first estimation to determine the difference between DWT coefficients of noise as compared to pure speech. Then, DWT coefficients thresholding is implemented to reduce the noise in the speech wavefrom. Recently, many researchers have focused on using wavelet packet for speech enhancement as proposed by [7] using integration between perceptual filterbank and minimum mean square error short time spectral amplitude estimation. The filterbank was built according to undecimated wavelet packet decomposition tree. Another speech enhancement algorithm was also proposed by [8] that consist of two portions relayed on wavelet package. The first stage is accomplished using wavelet transform followed by the second stage for removal of wavelet coefficients of the noisy speech. Additionally, new enhancement method was also suggested by [9] using time-scale adaptation of wavelet thresholds specifically wavelet coefficients energy is used to represent time dependency along with scale dependency is represented by spreading the level dependent threshold principle into wavelet packet thresholding. On the other hand, in [10] suggested speech enhancement system that based on wavelet thresholding techniques to overcome basic wavelet thresholding algorithm limitations namely white Gaussian noise and bad auditory quality. As reported in [10], in order to solve the drawback of basic wavelet thresholding, a system of speech enhancement based on adaptive thresholding of the wavelet packets was proposed without voiced/unvoiced detection system as a different speech activity detector is setup as an alternative to update noise statistics for colored or non-stationary noise Furthermore, adaptive wavelet thresholding was developed for waveform enhancement as discussed in [11]. The Bionic Wavelet Transform was initially intended for voice coding, but later used for speech enhancement by deriving an adaptive wavelet transform from non-linear auditory model of the cochlea. Moreover, wavelet packet transform was also applied to remove additive white Gaussian noise from corrupted speech signal and sufficient results reported using soft thresholding function. On the other hand, as explained in [12], a critical-band decomposition was developed for waveform enhancement method. The method anticipated converting noisy background into wavelet coefficients, followed by enhancement of the coefficients by subtracting threshold from noisy coefficients. Thresholding was accomplished using segmental SNR and noise masking threshold. Another research on speech enhancement was

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

also conducted by [13] that introduced a gain factor derived from the noise masking threshold.

2. Database Acquisition

This section outlined the speech acquisition that acted as database in this study. A data corpus was acquired as reported in our previous research [14]. Recall that the corpus comprises of Arabic speech signals that were recorded in an unrestrained environment. These Arabic speech signals are recorded using a Logitech microphone and algorithms are developed in MATLAB to achieve the recording process. In this speech dataset, 11 KHz and 16-bits are allocated as the sampling rate and sample format respectively with one (1) channel (mono) assigned as a channel. All the involved speakers are non-native Arabic speaker namely the Malay subjects. The designed dataset comprises of 1400 samples, collected from 50 Malay individuals (25 males and 25 females). Each speaker is required to utter all twenty eight phonemes of Arabic language.

3. 2-D Gabor Filters

The 2-D Gabor filters were proposed by [15] to simulate the spatial summation functions of simple cells in the visual cortex. These filters are extensively implemented in image processing, computer vision, neuroscience and psychophysics. Hence, in this study the modified version of the 2-D Gabor is used as outline in (1).

$$\begin{aligned} &g_{\xi,\eta,\sigma,\gamma,\theta,\lambda,\varphi}(x,y) \\ &= e^{-(x'^2+\gamma^2 y'^2)2\sigma^2 cos\left(2\pi\frac{x'}{\lambda}+\varphi\right)} \\ &x' = (x-\xi)cos\theta - (y-\eta)sin\theta \\ &y' = (x-\xi)sin\theta + (y-\eta)cos\theta \end{aligned} \quad (1)$$

where the arguments x and y specify the position of a point in the visual field and the pair (ξ, η), which has the identical domain Ω as the pair (x, y), stipulates the center of a receptive field within the visual field. The parameter σ defines the linear size of the receptive field. Its eccentricity and herewith the eccentricity of the receptive field ellipse is represented by the parameter γ, called the special aspect ratio. The parameter λ is the wavelength and $1/\lambda$ is the spatial frequency of the harmonic factor $\cos\left(2\pi\frac{x'}{\lambda}+\varphi\right)$. The ratio σ/λ refers to the spatial frequency bandwidth of inhibitory stripe zones that can be perceived in the receptive fields. The angle parameter θ identifies the orientation of the normal to the parallel excitatory and inhibitory stripe zones. Lastly, the parameter φ that is phase offset in the argument of the harmonic factor $\cos\left(2\pi\frac{x'}{\lambda}+\varphi\right)$ determines the symmetry of the function $g_{\xi,\eta,\sigma,\gamma,\theta,\lambda,\varphi}(x,y)$.

4. Results and Discussion

By applying in (1) on the phoneme signals, the parameters are adjusted to give appropriate outcomes. The adoption and adjustment of the parameters are based on suggestions by [16, 17]. The parameter γ, known as spatial aspect ratio, is proven by [18] differs in the range of $0.23 < \gamma < 0.92$. This factor stipulates the Gabor function support ellipticity. For $\gamma = 1$, the support is circular. For $\gamma < 1$ the support is lengthened in orientation of the parallel stripes of the function whilst the default value is at $\gamma = 0.5$. The second parameter is the wavelength λ that represent the wavelength Gabor filter kernel cosine factor and bounded the selected wavelength of the corresponding filter. Its value is specified as scale. The validated numbers are real values, which are > = 2. Mostly, when $\lambda = 2$, it should not be used in combination with phase offset φ = -90 or φ = 90 since in such matters the Gabor function is sampled in its zero crossings. So, as to avoid the phoneme signal degrading, the wavelength value is supposed be lesser than $\frac{1}{5}$ of the input signal size. Next, the parameter θ identifies the orientation of the regular to the parallel stripes of a Gabor function. The parameter value is indicated in degrees. Whilst, the valid number are real values between 0 and 360. The phase offset φ in the argument of the cosine factor of the Gabor function is indicated in degrees. The valid numbers are real values in the range -180 and 180. The values 0 and 180 match up to center-symmetric 'center-on' and 'center-off' functions correspondingly, while -90 and 90 correspond to anti-symmetric functions. All other cases relate to asymmetric functions. The half-response spatial frequency bandwidth b (in octaves) of a Gabor filter is related to the ratio σ / λ, where σ and λ are the standard deviation of the Gaussian factor of the Gabor function and the selected wavelength separately as in (2):

$$b = \log_2 \frac{\frac{\sigma}{\lambda}\pi + \sqrt{\frac{\ln 2}{2}}}{\frac{\sigma}{\lambda}\pi - \sqrt{\frac{\ln 2}{2}}}, \frac{\sigma}{\lambda} = \frac{1}{\pi}\sqrt{\frac{\ln 2}{2}} \cdot \frac{2^b+1}{2^b-1} \quad (2)$$

The σ value is not computable in a straight line. In contrast, the σ value would be transformed via the bandwidth b. For the bandwidth value, it has to be identified as a real positive number. The default value is 1, and both of σ and λ are associated as follows: $\sigma = 0.56\ \lambda$. The smaller the bandwidth, the larger σ, the support of the Gabor function and the number of visible parallel excitatory and inhibitory stripe zones. The outcome of applying Gabor filters on Arabic phoneme signals is illustrated in Figure 1.

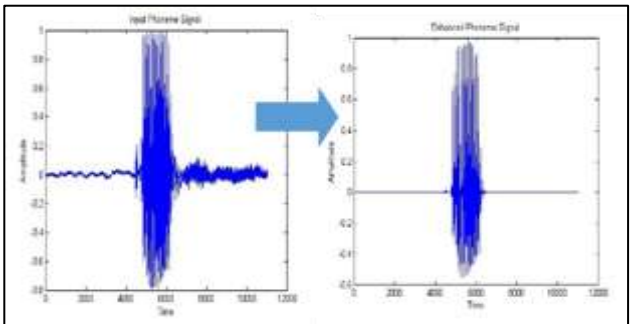

Fig 1: Input phoneme signals and the enhanced phoneme signals using Gabor filters

In order to show the efficiency of using Gabor filters in speech enhancement, Figure 2 depicted the spectrogram of both the noisy and enhanced phoneme signal.

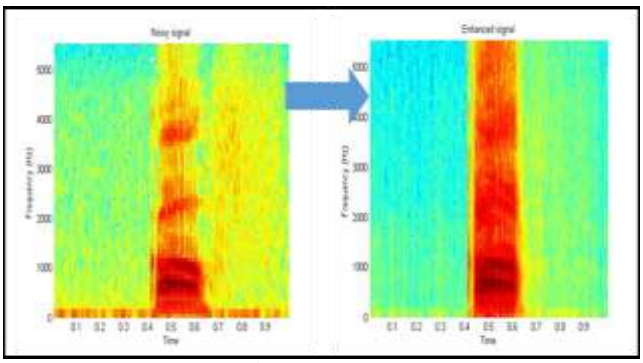

Fig 2: The spectrogram of the noisy signal and the enhanced signal

4.1 Signal-to-Noise Ratio

The results tabulated in Table 1 demonstrated the Signal-to-Noise Ratio of the suggested technique and Wiener filter. The Signal-to-noise Ratio (SNR) values are calculated as in (3).

$$SNR = 10 * \log_{10}\left(\frac{\sum_{n=1}^{N} S^2(n)}{\sum_{n=1}^{N}\left|S(n) - \hat{S}(n)\right|^2}\right) \quad (3)$$

In Table 1, all the speech waveform are degraded by ordinary room noise during the recording process. It is perceived that the SNR values of the suggested technique in Table 1 are lower than the SNR of Wiener filter. As for the phoneme signals with 10 dB, 5 dB, -5 dB and -10 dB noise signal, it showed that SNR of the anticipated approach are larger than the SNR values of the Wiener filter.

Table 1: SNR (dB) Performance on Both Noisy Signals (Left) and Phonemes Signals Corrupted by White Noise (Right) Using Proposed Method and Wiener Filter

Noisy Signal	Method		Input SNR White Noise	Method	
	Proposed	Wiener		Proposed	Wiener
1	2.00	7.88	10	2.28	0.76
2	6.14	13.23			
3	4.69	10.48	5	2.07	0.80
4	4.17	10.63			
5	3.67	7.27	-5	3.64	0.78
6	5.05	11.85			
7	6.16	12.25	-10	2.11	0.73
8	5.76	12.45			

4.2 Segmental Signal-to-Noise Ratio

The definition of Segmental Signal-to-Noise Ratio (SSNR) is the average SNR values computed from speech signals after divided into several frames. The definition of the Segmental Signal-to-Noise Ratio is stated as in (4).

$$SSNR = \frac{1}{N}\sum_{n=0}^{N-1} 10\log_{10}\sum_{k=0}^{K-1}\frac{|s(n,k)|^2}{|s(n,k) - \hat{S}(n,k)|^2} \quad (4)$$

where k is the frequency index and n is the segment index. For computational of SSNR values, the segment frame length was assigned to be 32 ms (512-point FFT). The larger the segmental SNR value, the better the recovery performance. As shown in Table 2, it is observed that the suggested technique contributed the highest Segmental SNR in every case and the speech signals that are corrupted with the recording room noises, the proposed method yield lowest Segmental SNR as compared to Wiener filter.

Table 2: Segmental SNR (dB) Performance on both Noisy Signals (Left) and Phonemes Signals Corrupted by White Noise (Right) Using Proposed Method and Wiener Filter

Noisy signal	Method		Input SNR White Noise	Method	
	Proposed	Wiener		Proposed	Wiener
1	0.42	2.48	10	2.30	0.74
2	1.13	2.71			
3	0.79	2.33	5	2.08	0.72
4	0.77	2.36			
5	1.14	2.37	-5	3.69	0.67
6	0.81	2.32			
7	1.19	2.59	-10	2.13	0.73
8	0.90	2.32			

4.3 SNR Analysis

The effect of the wavelength λ on the output SNR (dB) is shown in Figure 3.

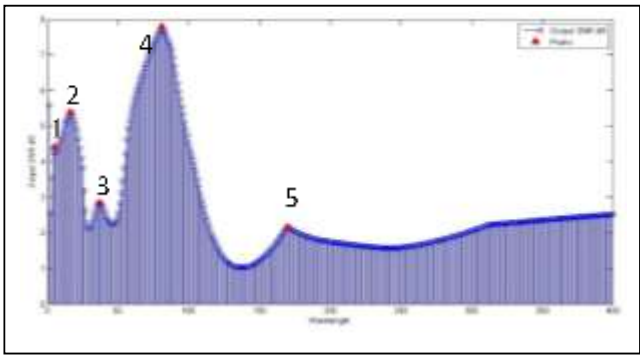

Fig. 3: SNR (dB) output versus wavelength

The highest SNR value is 7.79 dB, which corresponded to the wavelength λ = 81 that appears in the peak of number 4. Figure 4a showed the enhanced signal with λ = 81. Next, Figure 4(b) depicted the results of the enhanced speech signal that is corresponded to the peak of number 1 with wavelength λ = 5 and output of SNR as 4.41 dB. Further, Figure 4(c) shows the enhanced phoneme waveform which is matched to the peak of number 2 with wavelength λ = 16 and the output SNR is as 5.38dB. Moreover, Figure 4(d) demonstrates the enhanced speech signal that corresponds to the peak number 3 with wavelength λ = 37 and the output SNR is 2.84 dB. As shown in Figure 4(e), the enhanced phoneme waveform that is corresponding to the peak of number 5 with wavelength λ = 170 and the output SNR is 2.15 dB is plotted. Based on these results, it was found that the optimal value for λ is 81 that yield SNR at 7.79. Also, it is observed that all wavelengths greater than 170 contributed to similar enhanced signals with almost similar SNR.

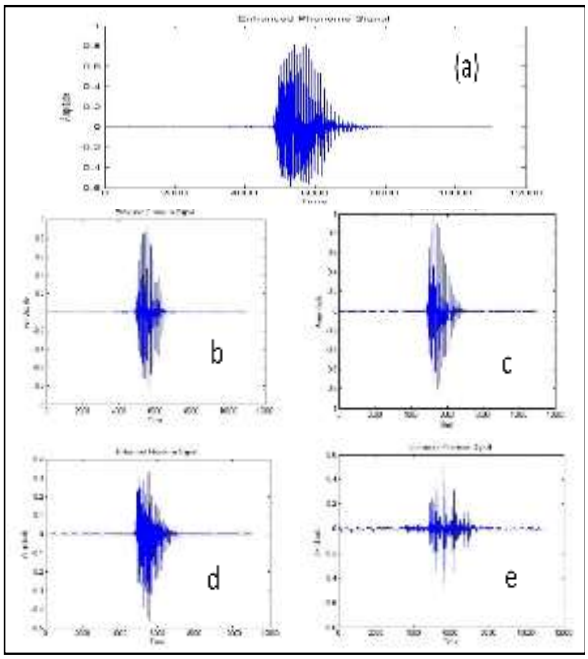

Fig. 4: (a) Speech Enhancement with $\boldsymbol{\lambda = 81}$, (b) $\boldsymbol{\lambda = 5}$, Output SNR = 4.41 dB, (c) $\boldsymbol{\lambda = 16}$, Output SNR=5.38 dB, (d) $\boldsymbol{\lambda = 37}$, Output SNR = 2.84 dB, (e) $\boldsymbol{\lambda = 170}$, Output SNR= 2.15dB

4.4 SSNR Analysis

The effects of the wavelength λ on the Segmental SNR are as shown in Figure 5. The highest Segmental SNR values shown as

peaks and the highest peak value is 3.75 dB, which is corresponding to the wavelength λ = 400 and appears at peak of number 9. Figure 6a showed the enhanced signal with λ = 400. In Figure 6a, at λ = 400, the noise of the signal is slightly removed.

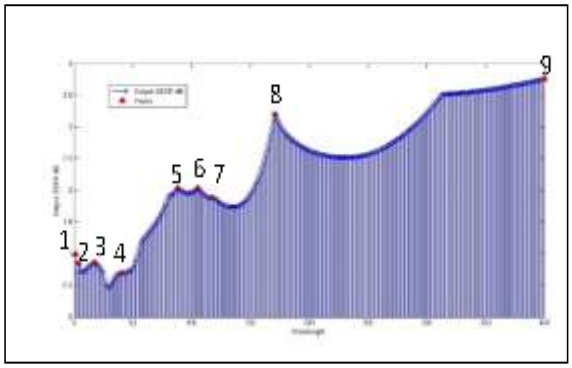

Fig. 5: Segmental SNR (dB) versus Gabor Filter wavelength

The presented results in Figure 6(b) showed the corresponded results to peak of number 1 with wavelength λ = 1 and the output Segmental SNR is 0.985 dB. In addition, in Figure 6(c) the enhanced phoneme signal is presented. These results are corresponded to the peak of number 2 with wavelength λ = 3 and the output Segmental SNR is 0.853 dB. By comparing the two segmental SNR ratio, it is observed that the segmental ratio is higher when λ = 1 than λ = 3 and the noise is lower in Figure 6a than in Figure 6c. Next, Figure 6(d) illustrates the enhanced speech signal which are corresponded to the peak of number 3 with wavelength λ = 17 with Segmental SNR at 0.857 dB. As for Figure 6(e), the enhanced phoneme signal and its related spectrogram respectively is presented too. These results are corresponded to the peak of number 4 with wavelength λ = 41 and the output Segmental SNR as 0.690 dB. Furthermore, Figure 6(f) depicted the enhanced speech signal. These results are corresponded to the peak of number 5 with wavelength λ = 88 and the output Segmental SNR is at 2.025 dB. As for Figure 6(g), the enhanced phoneme signal is plotted. These results are corresponded to the peak of number 6 with wavelength λ = 105, and Segmental SNR at 2.024 dB. Next, Figure 6(h) showed the enhanced speech signal. These results are corresponded to the peak of number 7 with wavelength λ = 117 and the output Segmental SNR at 1.875 dB. As for the enhanced version, Figure 6(i) shows the corresponded results to the peak of number 8 with the wavelength λ = 171 and the Segmental SNR as 3.188 dB.

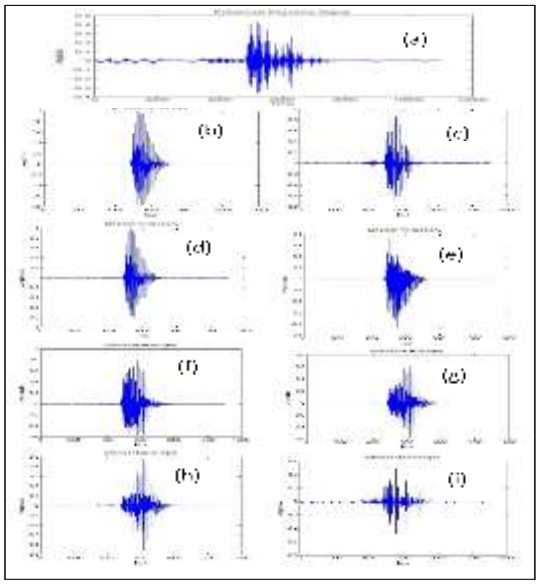

Fig. 6: (a) Enhanced Signal λ= 400, (b) λ= 1, (c) λ= 3, (d) λ= 17, (e) λ= 41, (f) λ= 88, (g) λ= 105, (h) λ= 117, (i) λ= 171

Further, Figure 7 illustrated the SNR and Segmental SNR of each peak. It was found that the both lowest value of SNR and segmental SNR are at the 7th peak specifically 2.132 dB with λ = 117 as the wavelength whilst the highest SNR value is on the 5th peak, which is equal to 7.147 dB and its corresponding wavelength is λ = 88.

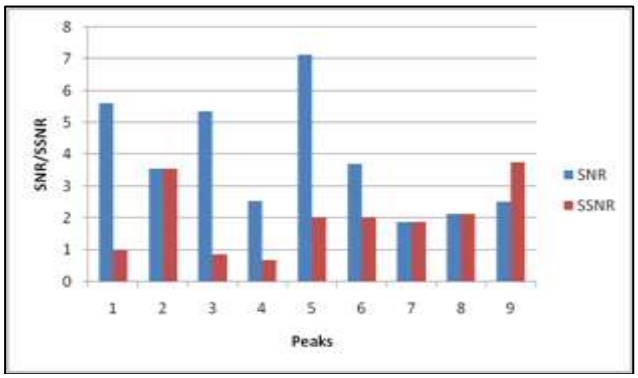

Fig. 7: Comparison between the SNR Ratio and the Segmental SNR Ratio at all Peaks

5. Conclusion

In conclusion, an approach for speech enhancement is implemented specifically for Arabic phonemes speech signals pronounced by Malay speakers. The proposed method is created by applying a 2-D Gabor filter on Arabic phonemes signals in order to eliminate noises in the voiced signals that is attained during the recording process. In addition, the wavelength of the filters is specified based on the output values of the SNR and Segmental SNR. Further, the performance of the proposed method is compared to Wiener filter and it was proven that the proposed method performed better. Hence it can be concluded that the proposed method using Gabor filters is indeed suitable to be used as speech enhancement method for speech signals that consists of one syllabus as owned by the recorded Arabic phoneme speech signals.

Acknowledgement

This research is funded by Institute of Research Management and Innovation (IRMI), Universiti Teknologi MARA (UiTM) Selangor, Malaysia under Grant No: 600-RMI/DANA 5/3/PSI (195/2013).

References

[1] Vaseghi, S. V. Advanced digital signal processing and noise reduction. JohnWiley and Sons, 2008.
[2] Boll, S. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing, 27, 1979; 113–120.
[3] El-Fattah, M.A.A., Dessouky, M.I., Diab, S.M., El-Samie, F.E.A. Speech enhancement using an adaptive wiener filtering approach. Prog. Electromagn. Res. 4, 2008; 167–184.
[4] Chen, J., Benesty, J., Member, S., Huang, Y.A., Doclo, S. New insights into the noise reduction Wiener filter. IEEE Trans. Audio. Speech. Lang. Processing. 14, 2006; 1218–1234.
[5] Sphraim, Y., Malah, D. Speech enhancement using a minimum mean- square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech, Signal Process. 32, 1984; 1109–1121.
[6] Djendi, M., Bendoumia, R. A new adaptive filtering subband algorithm for two-channel acoustic noise reduction and speech enhancement. Comput. Electr. Eng. 39, 2013; 2531–2550.
[7] Taşmaz, H., Erçelebi, E. Speech enhancement based on undecimated wavelet packet-perceptual filterbanks and MMSE–STSA estimation in various noise environments. Digit. Signal Process. 18, 2008; 797–812.

[8] Lun, D.P.-K., Shen, T.-W., Hsung, T.-C., Ho, D.K.C. Wavelet based speech presence probability estimator for speech enhancement. Digit. Signal Process. 22, 2012; 1161–1173.
[9] Bahoura, M., Rouat, J. Wavelet speech enhancement based on time–scale adaptation. Speech Commun. 48, 2006; 1620–1637.
[10] Ghanbari, Y., Karami-Mollaei, M.R. A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets. Speech Commun. 48, 2006; 927–940.
[11] Johnson, M.T., Yuan, X., Ren, Y. Speech signal enhancement through adaptive wavelet thresholding. Speech Commun. 49, 2007; 123–133.
[12] Lu, C.-T., Wang, H.-C. Enhancement of single channel speech based on masking property and wavelet transform. Speech Commun. 41, 2003; 409–427.
[13] Lu, C.-T., Wang, H.-C. Speech enhancement using hybrid gain factor in critical-band-wavelet-packet transform. Digit. Signal Process. 17, 2007; 172–188.
[14] Almisreb, A.A., Abidin, A.F., Tahir, N. Arabic letters corpus based Malay speaker-independent. Proceedings of the IEEE 3rd Int. Conf. Syst. Eng. Technol. 2013; pp. 19–20.
[15] Daugman, J.G. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. Opt. Soc. Am. 2, 1985; 1160–1169.
[16] Petkov, N., Kruizinga, P. Computational models of visual neurons specialised in the detection of periodic and aperiodic oriented visual stimuli: Bar and grating cells. Biol. Cybern. 76, 1997; 83–96.
[17] Petkov, N. Biologically motivated computationally intensive approaches to image pattern recognition. Futur. Gener. Comput. Syst. 11, 1995; 451–465.
[18] Jones, J.P., Palmer, L. a. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. J. Neurophysiol. 58, 1987; 123.