Indonesian Journal of Electrical Engineering and Computer Science
Vol. 13, No. 2, February 2019, pp. 499~506
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v13.i2.pp499-506

Robust speaker verification in band-localized noise conditions

Ali O. Abid Noor
Department of Communication Engineering, University of Technology, Iraq

Article Info

Article history:

Received Sep 6, 2018
Revised Nov 23, 2018
Accepted Dec 15, 2018

Keywords:

Adaptive filtering
Noise cancellation
Speaker verification
Threshold control

ABSTRACT

This research paper presents a robust method for speaker verification in noisy environments. The noise is assumed to contaminate certain parts of the voice's frequency spectrum, therefore, the verification method based on splitting the noisy speech into subsidiary bands then using a threshold to sense the existence of noise in a specific part of the spectrum, hence activating an adaptive filter in that part to track changes in noise's characteristics and remove it. The decomposition is achieved using a non-uniform filter bank that resembles human hearing perceptual. Speaker recognition is performed using vector quantization VQ or template matching technique. Features are extracted from speaker's voice using the normalized power in a similar way to the Mel-frequency cepstral coefficients. The performance of the proposed system is evaluated using 60 speakers subjected to five levels of signal to noise ratio SNR. Total success rate TSR, false acceptance rate FAR, false rejection rate FRR and equal error rate are used as performance indicators. The proposed method showed higher recognition accuracy than existing literature techniques even in severe noise conditions.

Corresponding Author:

Ali O. Abid Noor,
Department of Communication Engineering,
University of Technology-Iraq,
Baghdad, Iraq.
Email: email2019@vitstudent.ac.in

1. INTRODUCTION

Speaker verification is a vital tool in many modern speakers based applications such as security authentication, mobile phones, mobile computing systems, smart car and smart houses to name a few. Speaker verification in the presence of environmental noise has drawn the attention of many researchers in recent speaker recognition researches [1-3]. This has been the result of using human sound in the same way as a finger print in modern technologies. In speaker recognition systems, the effect of the environmental noise has an adverse effect on the performance of the feature extracting techniques [4]. The performance of feature extracting methods such as the line spectral frequency LSF representation, the linear prediction cepstral coefficients LPCC and the Mel frequency cepstral coefficients MFCC representation are highly degraded by environmental noise [4-7].

When the noise type and the characteristics of noise are changing between time and time the problem becomes sounder and requires the use of sophisticated techniques to solve it. The spectral power of certain types of noise such as car's engine noise occupies lower parts of the voice's frequency spectrum [8]. Other types of noise can occupy a limited part of the spectrum in any frequency range such as colored noise which can be in any form of environmental noise [9-11]. White noise has a wideband spectrum and it is the easiest to remove using a simple adaptive filter such as the least mean square LMS algorithm, while colored noise is the hardest and it requires the use of more complex algorithms such as the reclusive least squares RLS [12]. Therefore, in speaker recognition systems, the linear prediction LP analysis can be performed in certain parts of the frequency band of the speaker's voice rather than the whole band. Spectrum splitting can be advantageous in cases where the noise's frequency is restricted to any part in the noise's frequency band using

subsidiary band decomposition [13]. Band decomposition and adaptive filtering have been used in literature to remove noise from corrupted signals such as those found in [14], [15]. However, in these literatures, the implementation of adaptive filtering process has been assigned to all subsidiary bands, which results in a waste of computational power and a degraded filtering performance especially when the noise is localized in a specific band of noise's spectrum. This can be the situation of environmental noise corrupting speaker's voice in a speaker recognition system.

Therefore, the objective of the current work is to develop a speaker verification technique that uses a threshold controlled filter in a non-uniform band decomposition of the noisy speech's spectrum, hence removing band-localized noise, so that features are extracted robustly from the target speaker's voice. In many applications of noise cancellation, the changes in signal characteristics can be fast which requires the use of adaptive algorithms that converge quickly [12], [16]. Based on this point of view, the normalized least mean square NLMS algorithm can be an appropriate choice to remove noise from noisy parts of the target signal. Performance of the proposed method in this paper is compared to two well-known techniques, namely the linear prediction cepstral coefficients LPCC and the Mel frequency cepstral coefficients MFCC.

2. RESEARCH METHOD

The first step in the proposed algorithm is to decompose the input signal, which represents speaker's voice into subsidiary bands using non-uniform decomposition procedure as shown in Figure 1. The reason behind choosing octave implantation rather than uniform decomposition is that the octave bands are close match to humans hearing perceptual ability. The speaker's signal is split into non-uniform bands by applying low pass $H_0(z)$ and high pass $H_1(z)$ filters repeatedly. These filters are designed using direct form finite impulse response FIR design procedure splitting the signal into two equal bands, low and high.

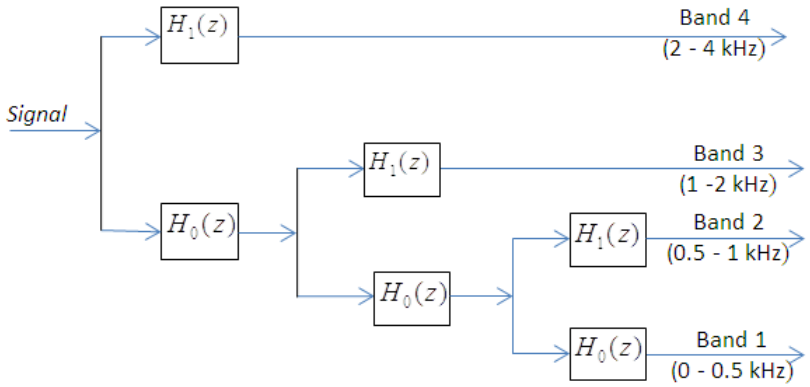

Figure 1. Band-decomposition of input signal.

To achieve low complexity implementation of the filters, all filters are based on designing a single prototype low pass filter in a quadrature mirror filter bank QMF configuration with normalized cutoff frequency. The cutoff frequencies are then adjusted in each stage of the split as appropriate, leading to four bands as shown in Table 1. More splitting of the original band is possible and more subsidiary bands can be created, but this will increase the computational power of the implementation. It is sufficient for the purpose of the current research to have four bands. In speech and speaker recognition systems the implementation complexity should be kept as minimum as possible in order to avoid processing delay in on-line applications. The speaker's signal is assumed to occupy a frequency range from 0-4 kHz which is a realistic assumption in voice telecommunication system as recommended by ITU-T [17].

Table 1. Band decomposition of the speaker's voice signal

Band No.	Frequency(kHz)
1	0-0.5
2	0.5-1
3	1.0-2.0
4	2.0-4.0

The two QMF filters responses are symmetric around normalized frequency of $\pi/2$. The use of QMF pair is commonly used in audio and voice digital signal processing for band splitting. The resulting high pass and low pass signals are reduced by 2 resulting in critically sampled signals [18]. After splitting the input signal, noise cancellation in the decomposed bands is performed using the NLMS subjected to a condition that the power of the input signal must exceed some predefined threshold. The NLMS adaptive algorithm is used to control a FIR filter. The choice of the NLMS algorithm is based upon its reduced computational costs as well as good tracking capabilities for changing noise characteristics. The normalized least mean squares NLMS algorithm is described mathematically by the following set of equations:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \hat{\mu}\mathbf{x}e(n) \tag{1}$$

$$e = s - y \tag{2}$$

$$y = \mathbf{x}\mathbf{w}^T \tag{3}$$

where, *s* is the desired signal, $\mathbf{w}$ is the filter weight coefficient vector, *T* is a transposition operator, $\mathbf{x}$ is the input noise vector at time *n*, *y* represents the output of the adaptive filter, *e* is the error signal which also is the clean output, and $\hat{\mu}$ is a variable step-size given by:

$$\hat{\mu} = (\mu / (\alpha + \| \mathrm{x} \|^2) \tag{4}$$

The fixed adaptation step size factor μ is t restricted to the range $0 < \mu < 2$ to avoid oscillation of the algorithm [19]. In the NLMS version of the direct LMS, the step-size factor μ is divided by the power of the input signal giving a variable step-size algorithm, where $\| \mathbf{x} \|$ is the norm or power of the input vector $\mathbf{x}$ at a certain point in time, α is a very small constant above zero that is used to avoid possible division by zero, hence securing the algorithm from shooting to infinity. For the speaker verification application considered in this research, the normalized version is implemented for better performance in dealing with non-stationary signals such as speech and environmental noise.

Threshold controlled noise cancellation is described as follows. Depending on the level of noise power in one or more of the decomposed bands, the adaptive filter is activated in response to a preset threshold in a specific band. In audio applications, ambient noise may not corrupt the whole speech spectrum; it rather confides to a certain part of signal's frequency range, so it is not necessary to run adaptation in all decomposed bands as in conventional way. Therefore, a new technique is devised in this paper based on threshold controlled noise cancellation. The threshold technique adopted here compares the average power of the input speaker's signal in a certain band with a predefined threshold, i.e. if the average power of the input signal to a certain band exceeds that of the one calculated for a clean signal, then noise in that band exists and therefore adaptive filter is activated accordingly. The flow chart shown in Figure 2 describes this mechanism. Mathematically, this can is expressed as follows. For an arbitrary subsidiary band of the decomposed speaker's signal spectrum, the average power is expressed by the following equation:

$$P_k = \frac{1}{N}\sum_{n=1}^{N} \left| s_k(n) \right|^2 \tag{5}$$

Where *k* is the band number, *N* is number of samples accumulated in band number *k* and *s* is input speaker's signal in band *k*. A threshold *T* is calculated for a clean template of the speaker's voice, which is assumed to be available beforehand, so that the comparison process is performed as follows.

$$\begin{aligned} &\text{if } P_k = \frac{1}{N}\sum_{n=1}^{N} \left| s_k(n) \right|^2 > T \text{ then turn adaptive filter 'ON' in this band} \\ &\text{Otherwise} \qquad \text{turn it off/or stay off} \end{aligned} \tag{6}$$

The algorithm described in the flow chart of Figure 2 restricts adaptive filtering to the noisy channel of the of the divided speech's spectrum. The aim of this possesses is to have a reduced noise in the target speaker's signal, hence obtaining fast and correct decisions in the proposed speaker verification system.

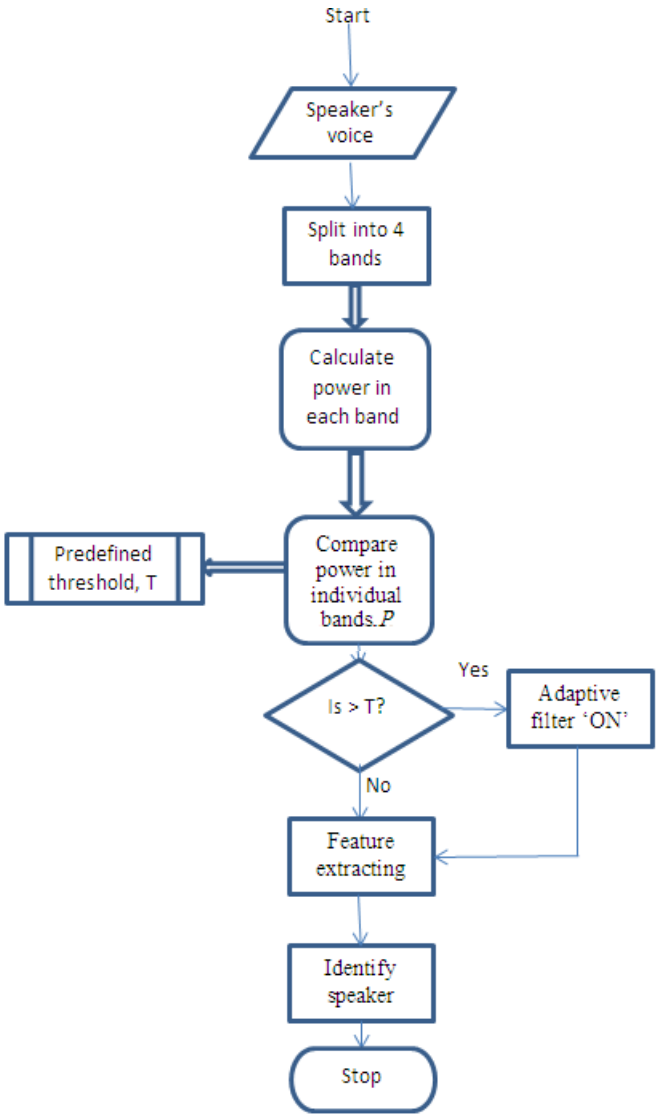

Figure 2. Expressing the processes in the proposed method as a flow-chart

Having cleaned the speaker's signal from band-localized noise, features of the speaker's voice is then extracted from individual bands using decomposition filters shown in Figure 1. The parameter that forms the feature vector in the proposed system is defined in a similar way to that used in the Mel Frequency Cpestral Coefficient MFCC [20]. The equation defining the feature vector, F in each of the channels in the non-uniform filter bank is expressed as follows.

$$F(m) = \sum_{k=1}^{4} \log(P_k) \cos\left(\frac{m(k-0.5)}{4}\pi\right) \quad (7)$$

$m = 1,2,3...12$. This equation can be extended to more bands; however, it is sufficient for the purpose of this research to restrict the number of bands to four. When the speaker is located in a noisy environment, the use of the proposed method provides clean features which leads to efficient verification of the target speaker. The presence of ambient noise reduces the performance of any feature extracting technique that is related to speaker

verification applications or any other speech recognition applications. Therefore, the procedure devised in this research is aimed to identify speakers located in places such as streets, markets and any other noisy environments.

3. RESULTS AND ANALYSIS

3.1. Experimental Setup

In this subsection, details of the simulation procedure and pameters are discussed. Signals of speakers' voice as well as input noise are sampled at 16 kHz sampling frequency. A sample of a speaker's voice signal and its spectrum is shown in Figure 3. In order to test the capability of the adaptive filter to be asssigned to a certain subsidiary band, colored noises ar generated by passing white noise through band-pass filters and added to the speaker's signal. Figure 4 shows the spectrums of the noise signals in each of the four bands representing environmental noise contaminating specific bands of the speaker's voice. The order of the prototype filter in the QMF is set 31. The adaptive filter in the NLMS algorithm is a FIR filter with 32 taps, this choice of filter's type is based on securing stability of the algorithm. The constant step size factor μ is set to 0.05 which has been obtained after testing the adaptive filter with several runs.

Template matching or vector quantization VQ technique is used for speaker recognition method proposed in this paper, in which an 8-bit codebook is used for this purpose [21]. Experiments were conducted for five SNR levels to prove the feasibility of the proposed method in severe noise conditions, numerically 0 dB, -10 dB, -20 dB, -30 dB and -40 dB. The utterances of 60 males and female's speakers are used in these tests. The whole recognition system is trained with 25 utterance samples belonging to 25 different speakers, while performance is evaluated with 35 speakers. A window size of 20 ms i.e. 320 samples is taken with an overlap of 10 ms i.e. 160 samples. The parameters of the proposed speaker verification algorithm are derived from (7) and speakers are identified according to the feature vector constructed from these parameters.

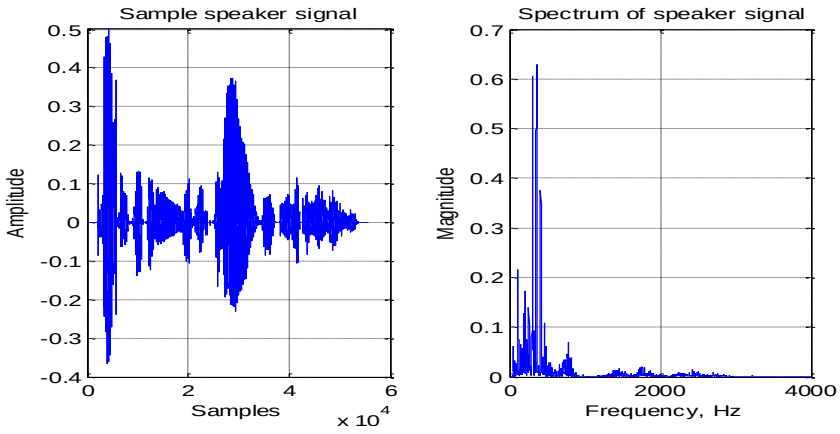

Figure 3. Speaker's signal and its frequency spectrum

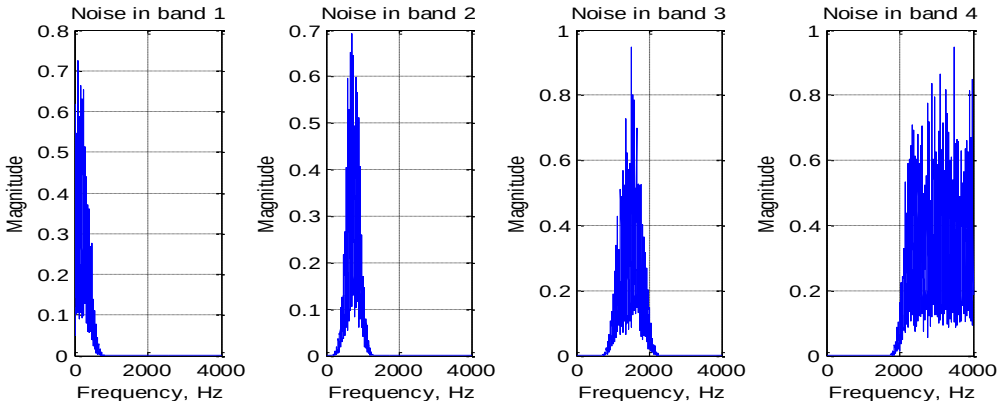

Figure 4. Noise spectrums corrupting certain bands in the voice's range

3.2. Evaluation and discussion

At the beginning of the evaluation process of the proposed speaker recognition method, the threshold controlled noise filter is tested by adding colored noise to the speaker's voice signal. The colored noise represents environmental noise corrupting a specific part of the input signal's spectrum. Mean square error MSE is plotted for band1 as an example and compared with a normal filter working on the whole band as shown in Figure 5. In the normal full band case, the input signal is also subjected to the same colored noise conditions. The results showed superior performance of the threshold controlled filter.

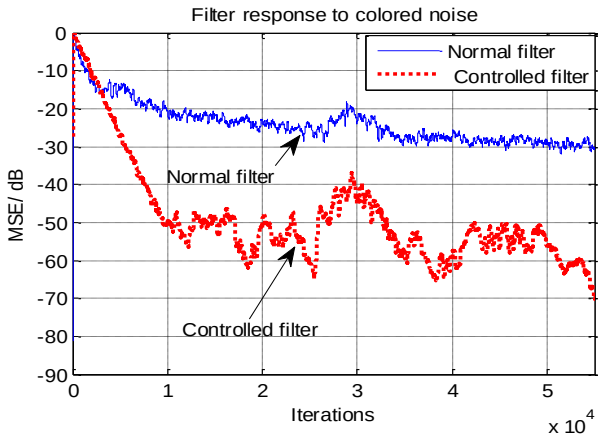

Figure 5. MSE plots for normal and controlled filters

The overall performance of the proposed speaker verification system is evaluated using the following indicators: The percentage of total success rates TSR in identifying sample speakers, the percentage of false acceptance rates FAR of the total number of speakers where the system fails to identify a certain number of speakers, false rejection rate FRR and equal error rate EER is when FAR equals to FRR. These indicators are compared to two existing and well-known techniques, namely the linear prediction cepstral coefficients LPCC and the Mel frequency cepstral coefficients MFCC. All tests cases are subjected to the same conditions with five levels of signal to noise ratios, numerically, $0,-10,-20,\ and\ -40\ dB$, with noise is in band 1. Band 1 noise case represents the most important one since most of speech contents are concentrated in this band and can affect the verification system seriously. Simulations are repeated for these levels of SNR starting from 0 dB which is the mild noise case to a severe noise of -4 0dB, results of these simulations are shown in Table 2, and Figures 6 and 7.

Table 2. Total success rate comparison of proposed method to existing algorithms.

TSR % SNR level/ (dB)	LPCC	MFCC	Proposed
0	97.2556	97.1944	98.7111
-10	95.1417	95.0083	97.0611
-20	93.6361	91.1722	95.8389
-30	91.2917	88.0667	93.8917
-40	89.8444	87.9361	92.9806

It can be seen from Table 2 that the TSR for the proposed method shows higher recognition percentages than existing techniques in all SNR levels. In the worst case where the SNR is -40 dB, the recognition is around 93% while with existing methods for the same -40 dB, the best performance is 88% for MFCC and below 90% is in the case of LPCC. When the SNR is higher than -40 dB, the performance of the proposed system shows remarkable recognition rates than other algorithms for all SNR level. The number of false hits in the proposed system is the lowest even when the noise level is severe i.e. -40 dB. It is clear that performance of LPCC and MFCC methods deteriorates when SNR level decreases i.e. the noise becomes stronger.

To confirm the success of the proposed system, Figures 6 and 7 display the averages of FRR and EER calculated over all SNR levels for the proposed method as well as those obtained using LPCC and MFCC

algorithms. It is clear from these figures that the method devised in this research shows the lowest false rejection rate (FRR) percentage and the least EER value compared to LPCC and MFCC algorithms. The average FRR for the proposed algorithm is as low as 11.5% while it is greater than 18% in the case of MFCC and 14.5% for LPCC. The proposed method shows the lowest average equal error rate EER of 0.31 as shown in Figure 7. These evidences prove the validity of the method developed in this research.

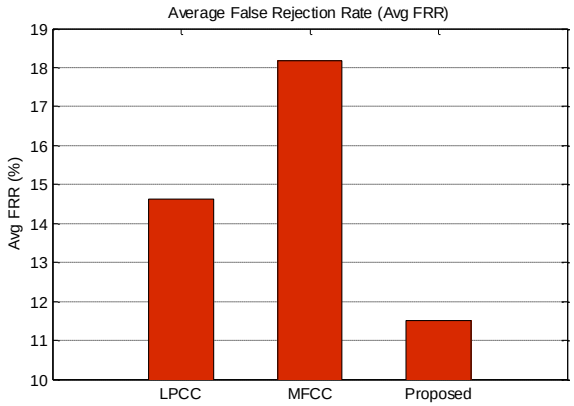

Figure 6. Average false rejection rate comparison.

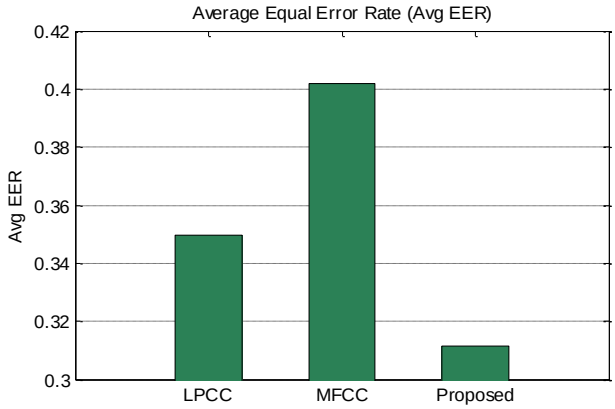

Figure 7. Comparison of average EER.

4. CONCLUSION

In this research, a robust speaker recognition method in the presence of in-band localized environmental noise is developed and evaluated. Speakers are verified robustly when the noise is restricted to certain part of the voices' spectrum. From the obtained results, it is experimentally proved that the proposed method gives the highest recognition rates for severe colored noise conditions. The average TSR reaches about 96%, while for other algorithms the best TSR is achieved using LPCC method was below 90% and it is around 88% for the MFCC method. The proposed system shows minimum EER of 0.31 and the best recognition performance even in severe noise conditions when SNR is -40 dB.

REFERENCES

[1] J. Chang, D. Wang, *Robust speaker recognition based on DNN/i-vectors and speech separation*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 5-9 March 2017. New Orleans, LA, USA.

[2] T. Yao, C. Meng, H. Liang, and L. Jia, Speaker recognition system based on deep neural networks and bottleneck features, *Journal of Tsinghua University (Science and Technology)*, 2016; vol. 56, no. 11, pp. 1143–1148.

[3] D. Matrouf, W. Ben. Kheder, P. Bousquet, M. Ajili, J. Bonastre. *Dealing with additive noise in speaker recognition systems based on i-vector approach*. IEEE Signal Processing Conference (EUSIPCO), 31 Aug. 2015. Nice, France.

[4] K. Hermus, P. Wambacq, and H. Van Hamme, A review of signal subspace speech enhancement and its application to noise robust speech recognition. *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 45821, 15 pages.
[5] M. Sahidullah & G. Saha. Design, analysis and experimental evaluation of block baed transformation in MFCC computation for speaker recognition. *Speech Communication*. 2012; 54(4): 543-565.
[6] U. Bhattacharjee, S. Gogoi, & R. Sharma, *A statistical analysis on the impact of noise on MFCC features for speech recognition.* Proceedings of IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), 2016 1-5.
[7] A. Zabidi, W. Mansor, Khuan Y. Lee. Optimal Feature Selection Technique for Mel Frequency Cepstral Coefficient Feature Extraction in Classifying Infant Cry with Asphyxia. *Indonesian Journal of Electrical Engineering and Computer Science*. 2017; 6(3): 646-655.
[8] E. Erzin and A.E. Cetin, Line spectral frequency representation of subbands for speech recognition, *Elesevier Signal Processing.* 1995*;* (44): 117-119.
[9] H. Kozou, Kujala, T. Shtyrov, Y. Toppila, E., Starck, J., Alku, P. & R. Näätänen. The effect of different noise types on the speech and non-speech elicited mismatch negativity. *Hearing research*. 2005; 199(1-2): 31-39.
[10] X. Zhao, Y. Wang, & D. Wang. Robust speaker identification in noisy and reverberant conditions. *IEEE/ACM Transactions on Audio, Speech and Language Processing*. 2014; 22(4): 836-845.
[11] J.S. Ashwin1, N. Manoharan. Audio Denoising Based on Short Time Fourier Transform. *Indonesian Journal of Electrical Engineering and Computer Science*. 2018; 9(1): 89-92.
[12] P. S. Diniz. Adaptive filtering: Algorithms and practical implementation. *Springer*. New York. USA, 2012.
[13] T. Jiang, R. Liang, Q. Wang, Y. Bao. Speech noise reduction algorithm in digital hearing aids based on improved sub-band SNR estimation. *Journal of circuits Systems and Signal Processing*. 2017.
[14] Z. Zheng, Z. Liu, H. Zhao, Y. Yu & L. Lu. Robust set-membership normalized subband adaptive filtering algorithms and their application to acoustic echo cancellation. *IEEE Transactions on Circuits and Systems–I Regular Papers*. 2017; 64(8): 2098-2111.
[15] J. Lorente, M. De. Ferrer, & A. González. GPU implementation of multichannel adaptive algorithms for local active noise control. *IEEE/ACM Transactions Audio Speech Language Processing*. 2014; 22(11): 1624-1635.
[16] A.H. Sayed. Adaptive filters. John Wiley & Sons, 2011.
[17] ITU-T., Artificial noise fields under laboratory conditions. Recommendation ITU-T, 6/2018.
[18] S. K. Agrawal and O. P. Sahu. Two-Channel Quadrature Mirror Filter Bank: An Overview. *Hindawi Publishing Corporation, ISRN Signal Processing*. Vol.2013, Article ID 815619, 10 pages.
[19] Paulo, S.D. Adaptive filtering: algorithms and practical implementation. *The International Series in Engineering and Computer Science*. 2008; 23-50.
[20] B. Karahoda, K. Pireva, A.S. Imran. *Mel frequency cepestral coefficients based similar Albanian phonemes recognition.* Int. Conference on human interface and the management of information. 2016; pp491-500.
[21] G. Nijihawan, M. K. Soni. Speaker recognition using MFCC and Vector Quantization. *Int. Journal on Recent Trends in Engineering and Technology*. 2014; 11(1): 211-218.

BIOGRAPHY OF AUTHOR

Ali O. Abid Noor received his Bacholor of Engineering. from Coventry/England, M.Sc. from the University of Technology, Iraq and PhD from national University of Malaysia (Universiti Kebangsaan Malaysia UKM), specialised in Electronic and Communication Engineering. He occupied several industrial as well as academic posts. He is currently the head of the wireless communication systems engineering branch in the Department of Communication Engineering at the University of Technology-Iraq. He is also serving as a member of the academic staff of the department. His main fields of interests are digital signal processing, adaptive filtering, multirate filter design and implementation, speech processing, noise cancellation, RF and microwave engineering.