Wireless Personal Communications (2019) 108:1055–1068
https://doi.org/10.1007/s11277-019-06453-4

A Novel Speech Enhancement Method Using Fourier Series Decomposition and Spectral Subtraction for Robust Speaker Identification

Ali I. Siam[1] · **Heba A. El-khobby**[2] · **Mustafa M. Abd Elnaby**[2] · **Hatem S. Abdelkader**[3] · **Fathi E. Abd El-Samie**[4]

Published online: 11 June 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract
This paper presents a novel speech enhancement approach by combining Fourier series expansion and spectral subtraction. This approach is implemented in speaker identification systems where degraded speech could result in high false speaker identifications. A Fourier series is estimated for the noisy speech signals, and then spectral subtraction is used to reduce the amount of noise in order to enhance quality of the speech signals before the speaker identification process. Experimental results presented to compare between the proposed approach and the traditional methods demonstrate the ability of the proposed approach to both enhance speech quality and improve speaker recognition rates.

Keywords Speech enhancement · Speaker identification · Voice authentication · Fourier series · Spectral subtraction

✉ Heba A. El-khobby
h_khobby@yahoo.com

Ali I. Siam
eng.ali.siam@gmail.com

Mustafa M. Abd Elnaby
mnaby45@gmail.com

Hatem S. Abdelkader
hatem6803@yahoo.com

Fathi E. Abd El-Samie
fathi_sayed@yahoo.com

1 Department of Smart Network Systems Technology, Faculty of Artificial Intelligence, Kafrelsheikh University, Kafrelsheikh, Egypt

2 Department of Electronics and Electrical Communications Engineering, Faculty of Engineering, Tanta University, Tanta, Egypt

3 Department of Information Systems, Faculty of Computers and Information, Menoufia University, Menoufia, Egypt

4 Department of Electronics and Electrical Communications, Faculty of Electronic Engineering, Menoufia University, Menouf 32952, Egypt

1 Introduction

Noise is a random, undesirable signal that does not convey any useful information. If it is superimposed on the speech signal, it leads to some distortion in the signal. This may lead to poor intelligibility and poor hearing of the speech. Therefore, the ability to communicate between speaker and listener is reduced in noisy environments. Noise may originate from crosstalk speech, interference from other sources of sound, or mismatch between media utilized in the operation. Presence of noise has a severe impact on Speaker Identification (SI) system performance leading to a dramatical reduction in the recognition rate. Therefore, speech enhancement is crucial for such systems, and it is usually utilized as a pre-processing step in such systems for performance enhancement, as shown in Fig. 1.

Various speech enhancement methods have been adopted for noise reduction in speech signals. Spectral subtraction is among the popular and commonly used methods for speech enhancement [1]. It depends on subtracting the magnitude of the spectrum of the noise from that of the noisy signal, while keeping the phase [2]. However, this method suffers from the so-called musical noise, which is difficult to be omitted [3]. Wiener filtering is another speech enhancement method that depends on minimizing the Mean Square Error (MSE) between the source and estimated speech signals. However, the Wiener filter requires prior estimation of the noise level in the signal before filtering [4], for which it is not suitable for real-time operation.

This paper presents a combination of Fourier series decomposition and spectral subtraction to enhance the speech signals and improve the SI process.

2 Spectral Subtraction

An estimation of clean speech signal spectrum can be obtained by subtracting an estimate of noise spectrum from the noisy speech spectrum [5]. An estimation of the noise spectrum can be perceived during silence periods, which contain only background noise generally found at the beginning and end of recording.

Let,

$$o(n) = s(n) + v(n) \tag{1}$$

where $o(n)$ is the non-clean speech signal, which is a combination of the noise $v(n)$ and the clean speech signal $s(n)$.

Taking FFT,

$$O(\omega) = S(\omega) + V(\omega) \tag{2}$$

Then $S(\omega)$ can be written as

$$S(\omega) = O(\omega) - V(\omega) \tag{3}$$

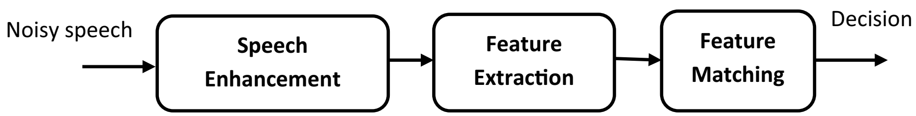

Fig. 1 Speaker identification system with a priori speech enhancement stage

$$S(\omega) = |O(\omega)|e^{j\theta_o} - |V(\omega)|e^{j\theta_v} \tag{4}$$

It is assumed that phase of the noise signal θ_v equals the phase of the noisy speech signal θ_o.

$$S(\omega) = [|O(\omega)| - |V(\omega)|]e^{j\theta_o} \tag{5}$$

$$\widehat{S}(\omega) = [|O(\omega)| - |\mu(\omega)|]e^{j\theta_o} \tag{6}$$

where $\widehat{S}(\omega)$ is the estimated spectrum of the clean signal and $\mu(\omega) = mean\{|V(\omega)|\}$ is the average value taken during a non-speech period.

By taking inverse FFT, we get an estimation of the clean speech signal $s(n)$.

The performance of the spectral subtraction is greatly dependent on the amount of estimated noise. If estimated noise is too low, residual noise can still be heard, and also if estimated noise is too high, some useful information might be lost.

The main drawback of spectral subtraction method is the presence of musical noise in the enhanced signal. Musical noise is difficult to be reduced since the musical noise spectrum is not stationary in short time frames [3].

3 Wiener Filter

Wiener filter is defined in the frequency domain. We can have [6]:

$$S(\omega) = H(\omega)X(\omega) \tag{7}$$

where $S(\omega)$ is the Discrete Fourier Transform (DFT) of the clean signal, $X(\omega)$ is the DFT of the noisy signal, and $H(\omega)$ is the transfer function of the Wiener filter.

The Wiener filter is given by:

$$H(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_v(\omega)} \tag{8}$$

where $P_v(\omega)$ is the power spectrum of the noise $v(n)$, and $P_s(\omega)$ is the power spectrum of the speech signal $s(n)$.

The Signal-to-Noise Ratio (SNR) can be expressed as:

$$SNR = \frac{P_s(\omega)}{P_v(\omega)} \tag{9}$$

Then, the transfer function $H(\omega)$ of the Wiener filter is obtained as:

$$H(\omega) = \left[1 + \frac{1}{SNR}\right]^{-1} \tag{10}$$

The main drawbacks of the Wiener filter are that it has a fixed frequency response at all frequencies, and also it requires estimation of the noise prior to filtering [3], for which it is not suitable for real-time operation. Wiener filtering has superior enhancement results compared to those of the spectral subtraction method, resulting in more acceptable hearing of the utterance with less noticeable noise. However, the speech signal spectrum processed by spectral subtraction is more like the clean signal spectrum than the Wiener filter output spectrum. That is why the spectral subtraction method gives better performance in speaker recognition systems, which are frequency-dependent.

4 The Fourier Series

Fourier series decomposes the signal into a (possibly infinite) number of simple harmonic functions called sines and cosines [7]. These harmonics have amplitudes and frequencies covering a wide range (possibly whole) of the spectrum; the frequency of one harmonic is higher than the frequency of last one. Fortunately, only a finite number of these harmonics could describe (approximate) the signal with possibly some distortion. The higher the number of harmonics taken to describe the signal, the lower the amount of distortion that appears in the signal (Fig. 2). The number of harmonics taken to approximate the signal is called the order of the Fourier series.

4.1 Fourier Series Expansion

The Fourier series of a discrete signal $y(k)$ is given by:

$$y \approx \frac{1}{2}a_0 + \sum_{m=1}^{M} a_m \cos\left(\frac{2\pi m}{L}y\right) + \sum_{m=1}^{M} b_m \sin\left(\frac{2\pi m}{L}y\right) \tag{11}$$

where M is the order of the Fourier series, $1 \leq M < \infty$, L is the length of the signal, and

$$a_0 = \frac{1}{L}\sum_{k=1}^{L} y(k) \tag{12}$$

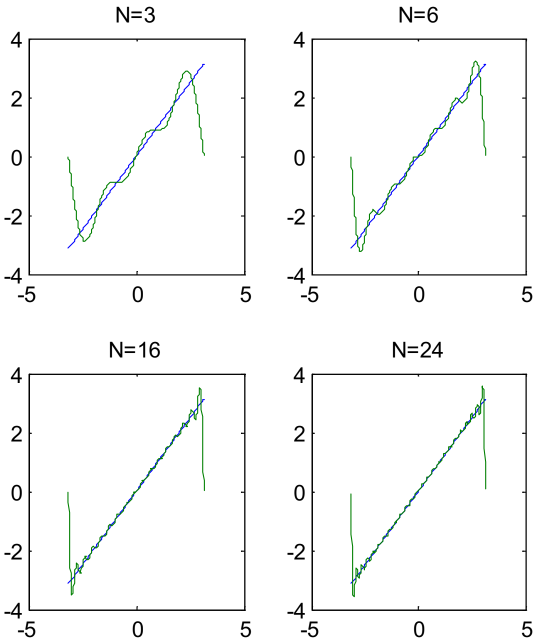

Fig. 2 Signal approximation with Fourier series with different orders

$$a_m = \frac{2}{L}\sum_{k=1}^{L} y(k)\ \cos\left(\frac{2\pi mk}{L}\right) \tag{13}$$

$$b_m = \frac{2}{L}\sum_{k=1}^{L} y(k)\ \sin\left(\frac{2\pi mk}{L}\right) \tag{14}$$

The term $a_m \cos\left(\frac{2\pi m}{L}y\right) + b_m \sin\left(\frac{2\pi m}{L}y\right)$ in Eq. (11) is called the mth harmonic of the Fourier series, thus the term

$a_1\ \cos\left(\frac{2\pi(1)}{L}y\right) + b_1\ \sin\left(\frac{2\pi(1)}{L}y\right)$ is called the 1st harmonic,
$a_2\ \cos\left(\frac{2\pi(2)}{L}y\right) + b_2\ \sin\left(\frac{2\pi(2)}{L}y\right)$ is called the 2nd harmonic, and so on.

4.2 Fourier Series for Noise Reduction

Fourier series decomposes the signal into simple harmonics (sines and cosines), each with a single frequency, covering all signal bandwidth. Each harmonic is weighted by amplitudes (a_n and b_n) to shape the waveform of the signal. Signals with low frequencies can be expressed by the first fewer harmonics, and the number of harmonics increases by increasing the frequency components within the signals. For a clean signal contaminated by an Additive Whaite Gaussian Noise (AWGN), the spectrum of the compound is extended to cover the high-frequency components contained within the noise signal such that most of the clean signal power occupies the lower part of the spectrum while the noise power

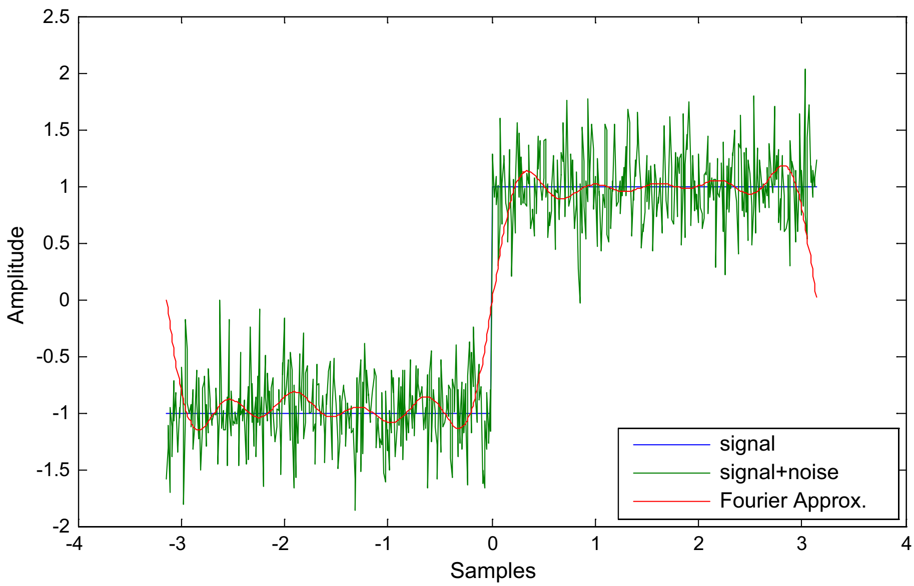

Fig. 3 Fourier approximation for a noisy signal ($N = 10$)

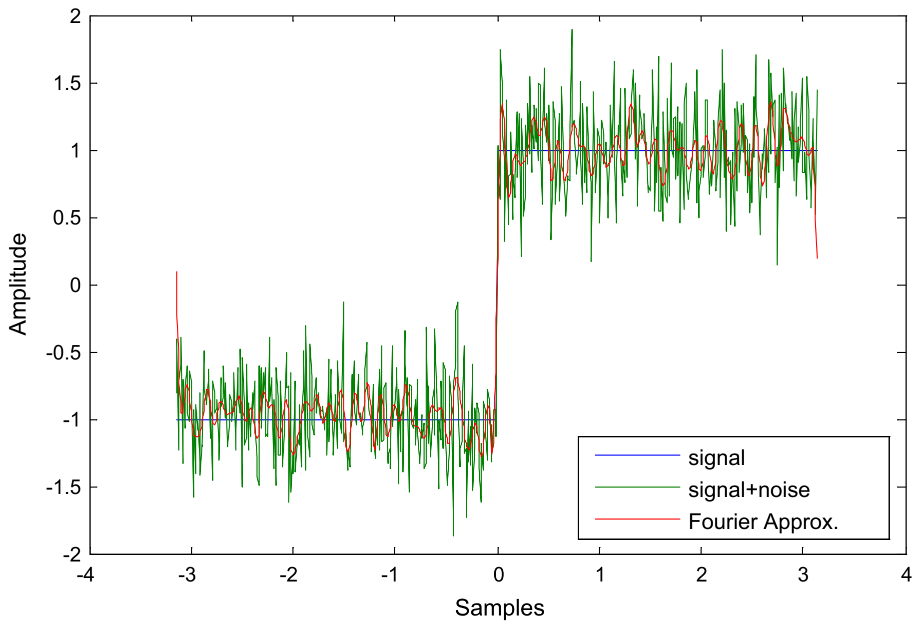

Fig. 4 Fourier approximation for a noisy signal ($N = 60$)

spans over whole spectrum. When applying Fourier series expansion to such compound, we can obtain an approximation for the clean signal by taking the first few harmonics only which can express most of the signal with some low frequency noise components. Higher harmonics representing most of the noise and some high frequency signal components are ignored. Figures 3 and 4 show a clean signal, the generated noisy signal after adding AWGN with $SNR = 10$ dB, and the Fourier approximation for the noisy signal with $N = 10$ and $N = 60$, respectively. These figures clarify that the much higher the order of the Fourier approximation for a noisy signal, the more the noise in the resultant waveform.

5 The Proposed Algorithm

The proposed algorithm for speech enhancement comprises both Fourier series expansion and spectral subtraction. The complete form of the proposed speech enhancement algorithm is shown in Fig. 5. Firstly, the noisy speech signal is segmented into small frames. Then, each frame is decomposed into N harmonics using Fourier series. Then, the frame is reconstructed again by summing up these harmonics to get an approximated frame which often has lower noise than the original one. This process continues till all frames are processed. After that, the spectral subtraction is applied to the reconstructed signal to obtain an enhanced speech signal.

The reason for framing of speech signals before Fourier expansion is to obtain approximated signals with more details as the smaller the scale of a signal, the more the details we get from the Fourier expansion. Figure 6 shows the enhanced speech signal with the proposed approach and a noisy speech signal with $SNR = 0$ dB for comparison.

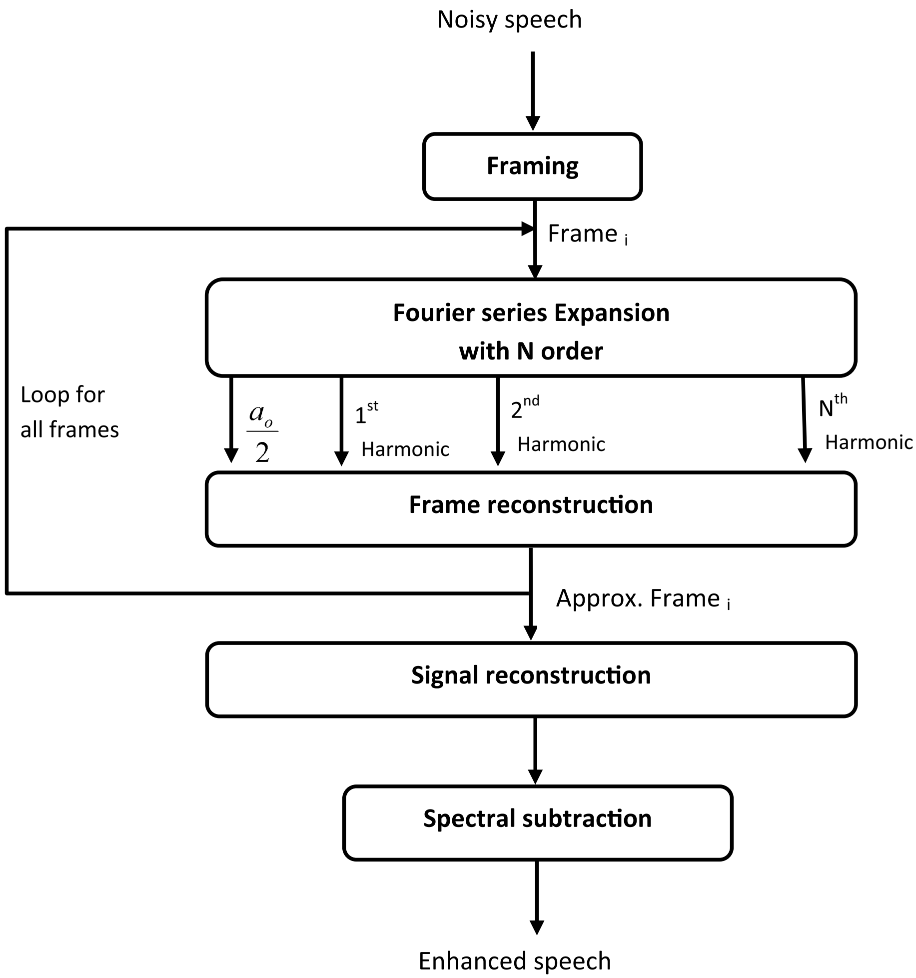

Fig. 5 Proposed speech enhancement approach

6 Speaker Identification

The speaker identification system comprises two phases: feature extraction and feature matching [8]. Figure 7 illustrates the two phases of the speaker identification system. In feature extraction phase, unique features (voice print) are extracted from the speaker utterance. The feature set extracted from authorized persons is stored for later use for discriminating between persons. Feature matching phase involves identification of a claiming speaker by comparing his voice print with pre-stored voice prints of authorized persons. If the speaker's voice print matches one of those of the authorized persons, the speaker is accepted, else the speaker is rejected.

There are various techniques to extract features form user utterance such as Mel Frequency Cepstral Coefficients (MFCCs), Dynamic Time Warping (DTW), Linear Predictive Coding (LPC), and Zero Crossings with Peak Amplitudes (ZCPA). Moreover, feature matching techniques include Vector Quantization (VQ), Hidden Markov Models (HMMs), Gaussian

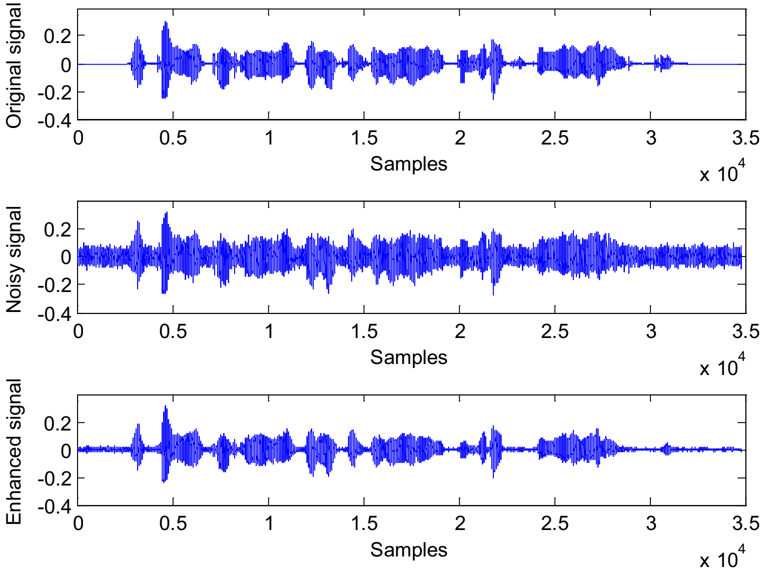

Fig. 6 Noisy and enhanced speech signal with the proposed approach

Mixture Models (GMMs), and Artificial Neural Networks (ANNs). In this paper, we will consider MFCCs as features with VQ for feature matching.

6.1 Feature Extraction

Human hearing organ can distinguish different speakers through extracting high-level perceptual features from utterance like dialect, speaking style, tone, and emotional state [3]. These features can discriminate between speakers effectively; however they are very complex to implement in a software or hardware system. Instead, low-level features of speech such as frequency, loudness, energy, and spectrum can discriminate between speakers with a recognition rate depending on the feature extraction technique and the amount of features extracted from the utterance. The MFCCs is an example of such low-level features.

The MFCCs are commonly used as features in speaker identification systems, because the basic principles of their extraction resemble the operation of the actual human auditory system [9].

The MFCCs work analogue to the human auditory perception system, which cannot perceive frequencies higher than 1 kHz, linearly. Thus, extraction of MFCCs requires two types of filters spaced linearly at low frequencies below 1 kHz and logarithmically beyond 1 kHz. The outputs of these filters are aligned with the Mel scale which can be described by Eq. (15).

$$Mel(f) = 2595 * \log_{10}\left(1 + \frac{f}{700}\right) \tag{15}$$

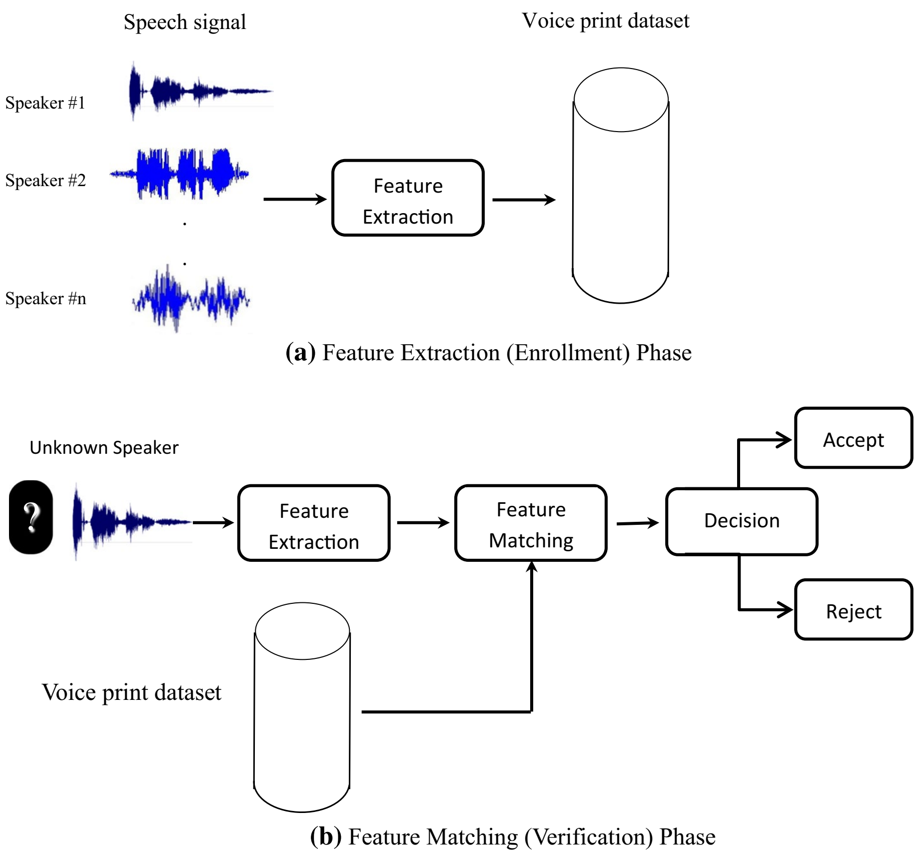

Fig. 7 Speaker verification system

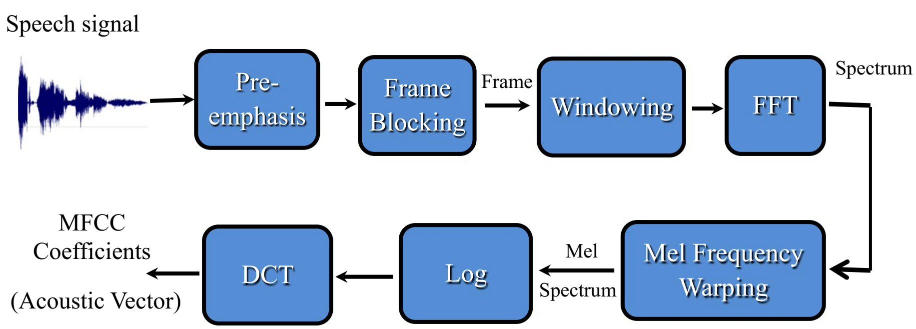

Fig. 8 Block diagram of MFCC extraction processor

where *Mel* is the Mel frequency and *f* is the linear frequency in Hz.

A block diagram of the structure of an MFCC extraction processor is given in Fig. 8. The operation of the MFCC extraction processor starts with capturing the input speech signal through a microphone with sampling frequency $F_s \geq 8$ KHz. This ensures that most of the energy contained in the baseband signal with frequency $300 \leq F_m \leq 3400$ Hz is captured. Then sampled signal is passed through seven computational steps till we get the MFCCs (voice print) from the last step.

6.2 Feature Matching

Vector Quantization (VQ) is a lossy data compression approach based on mapping vectors from a large vector space to a finite number of regions in that space. It works using the principle of the LBG algorithm which was originally proposed by Linde et al. [10]. In the VQ-based speaker identification algorithm, the speaker model is formed by clustering the speakers' feature vectors in K non-overlapping clusters. Each cluster is expressed by its center called a codeword, which is the centroid. The collection of all codewords is called a codebook. In the identification phase, the constructed codebook of the speaker is compared against stored codebooks of all speakers, and the distance is measured. The codebook with the least average distance is identified as that of the speaker of the input speech.

7 Experimental Results

The experiments were conducted on 50 speakers from ITU-T speech database [11]. ITU-T database is a collection of speech sentences with duration ranges from 4 to 12 s spoken by different males and females in different languages. Theses speech signals are contaminated by AWGN with different SNRs to test the speaker identification system when using the proposed speech enhancement approach as a pre-processing step as shown in Fig. 1. Different enhancement methods are adopted in the pre-processing step in the testing phase to evaluate the effect of each one on the speaker identification system performance. Two evaluation metrics are used: recognition rate and output SNR ($\text{SNR}_{\text{output}}$). The recognition rate is the ratio of the number of correcct identifications to the total number of identification trials. The output SNR is computed as [12]:

$$SNR_{output}\,(\text{dB}) = 10\ \log_{10}\left(\frac{\sum_{i=1}^{k} s^2(i)}{\sum_{i=1}^{k} (s(i) - y(i))^2}\right) \tag{16}$$

Table 1 Output SNR versus input SNR for speech enhancement methods

Input SNR (dB)	Output SNR (dB)		
	Wiener filter	Spectral subtraction	Proposed method
0	9.609	9.487	10.138
5	11.279	10.865	11.731
10	14.789	14.316	15.125
15	18.413	18.108	18.425
20	21.79	21.714	22.201
25	25.864	25.317	25.626

Table 2 Recognition rates of the speaker identification system with different speech enhancement methods

SNR	Wiener filter	Spectral subtraction	Proposed method
0	8	18	36
5	18	30	48
10	30	48	68
15	38	74	82
20	44	90	94
25	78	96	100

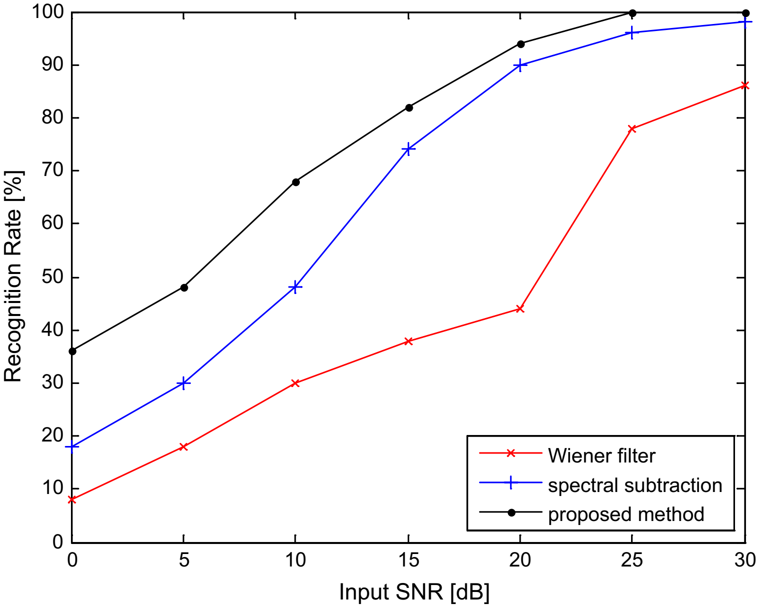

Fig. 9 Comparison of recognition rates for different enhancement methods

where $y(i)$ is the enhanced signal and $s(i)$ is the original speech signal.

Table 1 shows the output SNRs for different speech enhancement methods versus input SNRs for noisy speech signals when enhanced by different enhancement methods. Table 2 and Fig. 9 show the results of recognition rates for the speaker identification system when using different speech enhancement methods, versus different input SNRs.

8 Conclusion

This paper presented and evaluated a proposed speech enhancement algorithm using Fourier series expansion and spectral subtraction. This algorithm is to be used prior to the speaker identification process for noise reduction. The results showed that the proposed algorithm provides better results for noise reduction in speech signals than those obtained with the baseline speech enhancement algorithms. Furthermore, if it is used prior to the speaker identification process, the proposed method provides a robust speaker identification system from degraded speech.

References

1. Mavaddaty, S., Ahadi, S. M., & Seyedin, S. (2016). A novel speech enhancement method by learnable sparse and low-rank decomposition and domain adaptation. *Speech Communication, 76,* 42–60.
2. Kamath, S., & Loizou, P. (2002). A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *IEEE international conference on acoustics, speech and signal processing*, p. 4164.
3. El-Samie, F. E. A. (2011). *Information security for automatic speaker identification*. New York: Springer.
4. Scalart, P., & Filho, J. (1996). Speech enhancement based on a priori signal to noise estimation. In *IEEE international conference on acoustics, speech and signal processing,* pp. 629–632.
5. Lu, Y., & Loizou, P. (2008). A geometric approach to spectral subtraction. *Speech Communication, 50,* 453–466.
6. El-Fattah, M. A. A., Dessouky, M. I., Diab, S. M., & El-Samie, F. E. A. (2008). Speech enhancement using an adaptive wiener filtering approach. *Progress in Electromagnetics Research, 4,* 167–184.
7. Osgood, B. (2013). *Lecture notes for EE 261: The Fourier transform and its applications*. Stanford University.
8. Reynolds, D., & Rose, R. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing, 3,* 72–83.
9. Kurzekar, P., Deshmukh, R., Waghmare, V., & Shrishrimal, P. (2014). A comparative study of feature extraction techniques for speech recognition system. *IJIRSET, 3,* 18006–18016.
10. Linde, Y., Buzo, A., & Gray, R. M. (1980). An algorithm for vector quantizer design. *IEEE Transactions on Communications, 28,* 84–96.
11. ITU-T Test Signals for Telecommunication Systems. http://www.itu.int/net/itu-t/sigdb/genaudio/Pseries.htm.
12. Kondo, K. (2012). *"Subjective quality measurement of speech", signals and communication technology* (pp. 7–20). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-27506-7_2.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ali I. Siam was born in Egypt on September 09, 1987. He received the B.Sc. degree in Electronics and Electrical Communications Engineering from the Faculty of Engineering, Tanta University, Egypt, in 2009. He received the M.Sc. degree in Electrical Engineering from Tanta University, in 2016. Since 2009, he has been a freelancer engineer with intensive contributions in the field of software programming and electronics. He is now a teaching assistant in the Faculty of Artificial Intelligence, Kafrelsheikh University. His research interests include Cloud Computing, IoT, Systems Security, Software Engineering, and Signal Processing.

Heba A. El-Khobby received the M.Sc. and Ph.D. degrees in Electronics and Electrical Communications Engineering from Tanta University, Egypt, in 2003, and 2009, respectively. Her research interests include distributed computing, task allocation and scheduling, congestion control, QoS, multimedia networking, image enhancement, image restoration, image interpolation, super-resolution reconstruction of images, and medical image processing.

Mustafa M. Abd Elnaby is a Professor Emeritus at the Electronics and Communications Engineering Dept., College of Engineering, Tanta University, Egypt. Since 2001, he has been a Professor of electrical engineering at the University of Qatar, Qatar. Prof. Mustafa was a visiting Professor at Kyushu University-Japan, working in design and fabrication of MOS devices for flash memory and communication system applications. He has also served as a Visiting Professor at Aachen University.

Hatem S. Abdelkader obtained his B.Sc. and M.Sc. in Electrical Engineering from the Alexandria University, Faculty of Engineering, Egypt in 1990 and 1995, respectively. He obtained his Ph.D. degree in Electrical Engineering from the Alexandria University, Faculty of Engineering, Egypt in 2001 in neural networks and applications. He is currently a professor in Information Systems Department, Faculty of Computers and Information, Menoufia University, Egypt since 2004. He has worked on a number of organizations. He has contributed more than 100 technical papers in the areas of neural networks, database applications, information security and Internet applications.

Fathi E. Abd El-Samie received the B.Sc. (Hons.), M.Sc., and Ph.D. degrees from Menoufia University, Menouf, Egypt, in 1998, 2001, and 2005, respectively. Since 2005, he has been a Teaching Staff Member with the Department of Electronics and Electrical Communications, Faculty of Electronic Engineering, Menoufia University. He is currently a researcher at KACST-TIC in Radio Frequency and Photonics for the e-Society (RFTONICs). His current research interests include image enhancement, image restoration, image interpolation, super-resolution reconstruction of images, data hiding, multimedia communications, medical image processing, optical signal processing, and digital communications. He was a recipient of the Most Cited Paper Award from the Digital Signal Processing journal in 2008.