# Speaker recognition based on pre-processing approaches

Samia Abd El-Moneim[1] · El-Sayed M. EL-Rabaie[2] · M. A. Nassar[2] · Moawad I. Dessouky[2] · Nabil A. Ismail[3] · Adel S. El-Fishawy[2] · Fathi E. Abd El-Samie[2]

## Abstract

This paper presents two pre-processing methods that can be implemented for noise reduction in speaker recognition systems. These methods are adaptive noise canceller (ANC) and Savitzky-Golay (SG) filter. Also, discrete cosine transform (DCT), discrete wavelet transform (DWT) and discrete sine transform (DST) are considered for consistent feature extraction from noisy speech signals. A neural network with only one hidden layer is used as a classifier. The performances of the proposed noise reduction methods are compared with those of a hybrid method that comprises empirical mode decomposition (EMD) and spectral subtraction and also with spectral subtraction method only. Recognition rate is taken as a performance metric to evaluate the behavior of the system with these enhancement strategies. Simulation results prove that the DCT is the optimum transform with the suggested methods, while the DWT is the best one with the hybrid method and the spectral subtraction method.

**Keywords** Speaker identification · Adaptive noise canceller · Savitzky-Golay filter · EMD · Spectral subtraction

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

  Samia Abd El-Moneim
  semsemkab@yahoo.com

  El-Sayed M. EL-Rabaie
  srabie1@yahoo.com

  M. A. Nassar
  nassar54@yahoo.com

  Moawad I. Dessouky
  dr_moawad@yahoo.com

  Nabil A. Ismail
  Nabil.Ismail@el-eng.menofia.edu.eg

  Adel S. El-Fishawy
  aelfishawy@hotmail.com

[1] Department of Electrical Communications, Tanta High Institute of Engineering and Technology, Tanta, Egypt

[2] Department of Electronics and Electrical Communications Engineering, Faculty of Electronic Engineering, Menoufia University, Menouf 32952, Egypt

[3] Department of Computer Science and Engineering, Faculty of Electronic Engineering, Menoufia University, Menouf 32952, Egypt

## 1 Introduction

Speech is the earliest method of communication between persons, since it carries imperative information about the identity, gender, emotional state, language more than messages or words. The main objective of implementing artificial intelligence in the speech processing field is to create a system that is capable of identifying the person from his voice (Abimbola 2007). A speaker recognition (SR) system comprises two step: feature extraction and classification. Feature extraction is the procedure of finding out the essential characteristics of the speaker from his speech signals, while rejecting redundancy. It abbreviates the speech signal into a small feature vector that is sufficient for good speaker distinction (Reynolds 2002). There are various feature extraction methods that have been used such as linear prediction coefficients (LPCs), cepstral linear prediction coefficients (CLPCs), and mel frequency cepstral coefficients (MFCCs). The MFCCs method is the most prevalent feature extraction method. It is based on low-level features (Das 2014).

The classification stage includes two sub-classes: training and testing. In the training, after extracting features from the speech signals, a model for each speaker is made and kept in the database. In the testing phase, when an unknown

speaker is enrolled to the system, the corresponding model is matched to those stored in the database and a decision is made. An Artificial Neural Network (ANN) can be used for classification (Hossain et al. 2007).

Extraction of the features in the existence of noise is the basic problem in the SR systems, and it is still a challenging task. The circumstantial noise and channel mis-representations are the main causes of distortion in speech signals. A prior step in SR is introduced to lessen distortion to achieve better performance of the system. There are many speech enhancement methods that have been used. Spectral subtraction is a well-known method that was used by Kaladharan in 2014 for speech denoising (Kaladharan 2014). In 2015, El-Moneim et al. 2015 used a hybrid method based on EMD and spectral subtraction with some transform domains and achieved the best accuracy with the DWT.

In the framework presented in this paper, two proposed filtering methods: Adaptive Noise Canceller (ANC) and SG filter are used in a pre-processing step in the SR systems with features extracted from transform domains. The ANC (Shanmugam et al. 2013) is a filter that requests two inputs: a primary input containing the distorted version of the signal and a reference input containing noise correlated with the primary signal. The reference signal is adaptively filtered and subtracted from the primary signal to get the enhanced version of the signal. The obtained results show good performance of the SR system with the ANC, although it has some limitations such as the requirement of prior knowledge about the noise and the channel. The SG filter is another filter that is used for smoothing prior to the identification process. A comparison is presented to show how the system is affected by the different speech enhancement methods.

The paper is structured as follows. In Sect. 2, a summary of the SR system and its main steps is given. The proposed SR system with enhancement methods is presented in Sect. 3. The used pre-processing methods in the SR system are investigated in Sect. 4. The performance of the suggested methods with the SR system is substantially verified by experimental results and compared with those of the spectral subtraction and the hybrid methods with the SR system in Sect. 5. Finally, in Sect. 6, the conclusion is presented.

## 2 Speaker recognition system

As mentioned above, the SR system aims at recognizing the speaker based on discriminative attributes extracted from the speech signals (Campbell 1997). The SR has two stage: feature extraction stage and classification stage as in Fig. 1.

### 2.1 Feature extraction

Feature extraction is the main phase of the SR system. It retains the most important information, while discarding redundancy. The MFCCs are the features used in this paper with polynomial coefficients.

#### 2.1.1 Mel frequency cepstral coefficients (MFCCs)

The main procedure of estimating the MFCCs is shown in Fig. 2. The MFCCs are extracted with short-term analysis of speech (Muda et al. 2010; Abd El-Samie 2011). The speech signal is sub-divided into short segments called frames. At first, the speech is pre-emphasized to promote high-frequency components. Framing and windowing is then performed. After that, discrete fourier transform (DFT) is taken to get the speech spectrum. The speech signal spectrum is warped on Mel scale to approximate the human hearing perception for speech, which is linear below 1 kHz and logarithmic above 1 kHz.
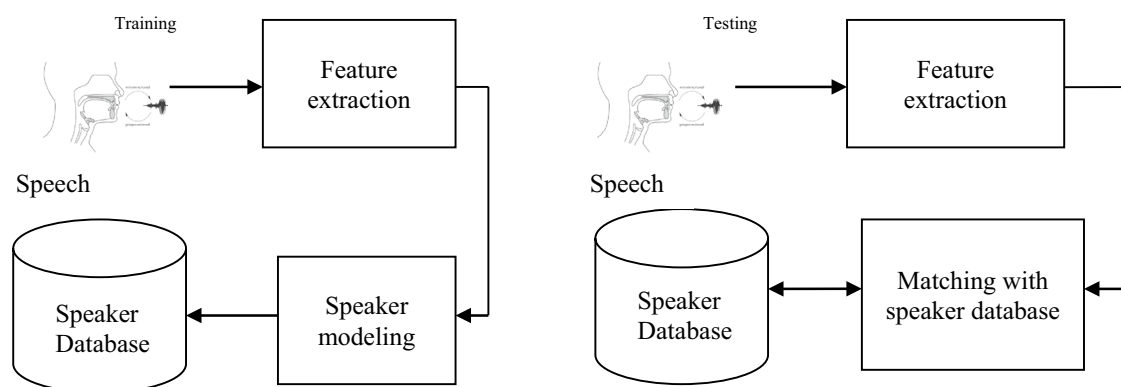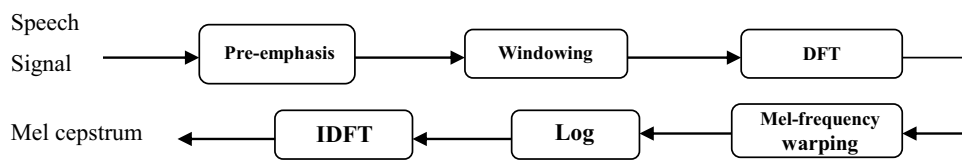


**Fig. 1** Speaker recognition system

**Fig. 2** MFCCs extraction procedure



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

Finally, the DCT is performed on the log of the warped spectrum to get the MFCCs:

$$c_g = \sqrt{\frac{2}{N_f}} \sum_{m=1}^{N_f} \log\left(\hat{S}(m)\right) \cos\left[\frac{g\pi}{N_f}(m - 0.5)\right] \tag{2}$$

where $c_g$ is the MFCC, $\hat{S}(m)$ is the $m$th Mel filter output, $g = 0,1,2,\ldots,G-1$, and $G$ is the number of MFCCs. It is chosen between 12 and 20. The $N_f$ is the number of Mel filters. The first few coefficients are taken as they give the most specific information about the speaker.

### 2.1.2 Polynomial coefficients

The MFCCs are not sufficiently robust in noisy environments, and hence polynomial coefficients are added to reduce sensitivity to mismatch amongst training and testing data. Polynomial coefficients are calculated to represent the slope and curvature of the time waveform for each cepstral coefficient. Polynomial coefficients are computed by expanding the time waveform by orthogonal polynomials. Two orthogonal polynomials are used (Furui 1981):

$$p_1(i) = i - 5 \tag{3}$$

$$p_2(i) = i^2 - 10i + 55/3. \tag{4}$$

The polynomial coefficients can be computed as;

$$a_g(t) = \frac{\sum_{i=1}^{9} p_1(i) c_g(t + i + 1)}{\sum_{i=1}^{9} p_i^2(i)} \tag{5}$$

$$b_g(t) = \frac{\sum_{i=1}^{9} p_2(i) c_g(t + i + 1)}{\sum_{i=1}^{9} p_2^2(i)} \tag{6}$$

where $a_g(t)$ and $b_g(t)$ are the slope and curvature of the $g$th cepstral coefficient in $t$th frame, respectively. The vectors containing $c_g$, $a_g$ and $b_g$ are concatenated together to shape one feature vector for each frame.

### 2.2 Classification

Classification has two stages: training and testing. In the training, a model for each registered speaker is made using the MFCCs, and then it is stored in the database. In testing, when a new speaker enters the system, similar features are extracted and associated with those already existing in database, and a decision is then taken. The used classifier here is an ANN as in Fig. 3. The ANN is an intelligent mathematical algorithm that entails three main parts: input layer, middle or hidden layer(s), and output layer. Firstly, the input layer receives the system input. The hidden layer acts as the innermost layer of the ANN. It consists of many units called neurons. Inside the neurons, the main mathematical calculations are performed to process the inputs and submit the appropriate outputs (Dreyfus 2005; Sharma et al. 2005).

A neuron in the hidden layer receives and sends a group of values from the previous layer to the subsequent layers. The ANN entails a huge number of simple processing units called neurons that influence each other via connections called weights. Every layer consists of cells, and these cells are linked by the weights for the information to flow from the input layer to the output layer through the hidden layer. The ANN is trained by adjusting the weights between neurons via the learning rules (Dreyfus 2005; Sharma et al. 2005; Galushkin 2007; Love et al. 2004).

In Fig. 3, if **x** is an $m$-element vector containing $x_1$, $x_2$, $x_3$, …., $x_n$ as the neuron inputs, and the weights are $w_{1,1}$, $w_{1,2}$, $w_{1,3}$, …..,$w_{1,n}$, the input of the net $v$ will be:

$$v_j = \sum_{i=1}^{m} \mathbf{x}_i \mathbf{w}_{ji} + b \tag{7}$$

where $b$ is the bias, and $w_{ji}$ is the weight from the unit $i$ to unit $j$. The input is then the argument of the activation function. Once the net input is computed, the output activation is computed as a function of $v_j$.

$$y_j = f\left(v_j\right) \tag{8}$$

where $f$ is the activation function, $y$ is the output of the neuron, $b$ is the bias that performs affine transformation of the output. The training algorithm sets the weights by minimizing the sum of the squared error between the desired output $d_n$ and the real output $y_n$ of the output neurons specified by:
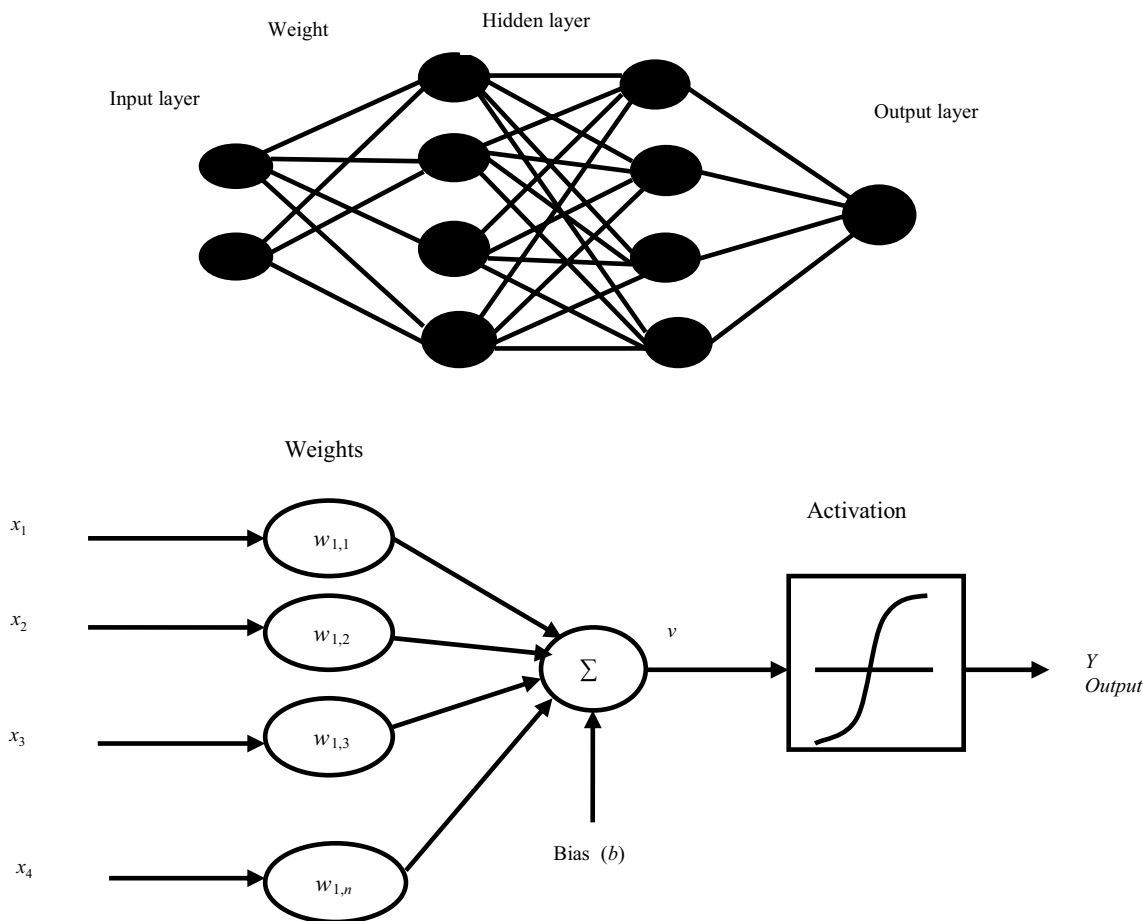
**Fig. 3** ANN structure

$$E = \frac{1}{2} \sum_{n=1}^{N} \left( d_n - y_n \right)^2 \qquad (9)$$

where $N$ is the number of output neurons. The error back-propagation algorithm is used for that purpose until the error is diminished via an iterative method.

## 3 Proposed SR system with pre-processing

Speech enhancement is a preliminary step in the SR system, since the speech signal is sometimes contaminated with noise. The enhancement process aims at reducing or removing this unpleasant noise. Different filters are used such as ANC and SG filters. Their performance is compared with the well-known spectral subtraction method and also with the hybrid method. For robust extraction of features, discrete transforms are used in the SR system. The DWT, DCT, and DST are investigated for this purpose. The DWT decomposes the speech into approximation and details. The effect of noise is less on the approximation. On the other hand, the DCT and the DST have a brilliant energy compaction property (Nasr et al. 2018). The block diagram of the proposed SR system with pre-processing is illustrated in Fig. 4.

**Fig. 4** Proposed SR system with pre-processing
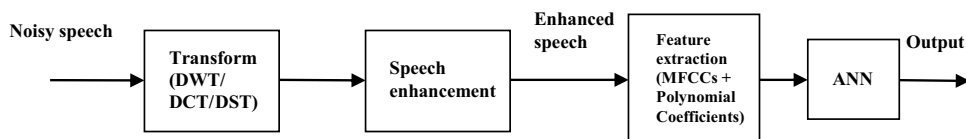
# 4 Enhancement methods

## 4.1 Spectral subtraction method

Spectral subtraction is a common enhancement method. Its general idea is the estimation of noise during the non-activity of speech and the subtraction of this estimated noise from the noisy speech to get an estimate of the original speech. If $s(k)$ is the original speech signal, $y(k)$ is the degraded speech signal and $v(k)$ is the noise, then (Karam 2014; Pawar 2013);

$$y(k) = s(k) + v(k) \tag{10}$$

Taking the fast Fourier transform (FFT), we get

$$Y(\omega) = S(\omega) + V(\omega) \tag{11}$$

where $Y(\omega)=|Y(\omega)|\, e^{j\theta_y}$ and $V(\omega)=|V(\omega)|\, e^{j\theta_y}$. Hence,

$$\hat{S}(\omega) = \{|Y(\omega)| - |V(\omega)|\}e^{j\theta_y(\omega)} \tag{12}$$

where $\hat{S}(\omega)$ is the expected spectrum of the clean speech signal, and $V(\omega)$ is estimated by averaging of spectra during non-active periods of speech. Taking the inverse FFT, an estimate of the original signal is obtained. This method is efficient, but it introduces musical noise (Evans 2006).

## 4.2 Hybrid method with EMD and spectral subtraction

The EMD is a signal decomposition method that decomposes the signal to a number of components, called intrinsic mode functions, which must satisfy the criteria of a symmetric envelope with zero mean (Rilling et al. 2003). The EMD is performed in time domain. The IMFs are deduced from the signal itself. The method of deriving the IMFs is called sifting, and its steps are as follows (Kim and Oh 2009; Alotaiby et al. 2014):

1. For a signal $s(t)$, all local extrema (minima, maxima) are specified.
2. Interpolation is performed between minima to form the lower envelope $l(t)$, and between maxima to form the upper envelope $u(t)$.
3. The local mean $m(t)$ of the envelopes, $m(t) = \frac{u(t)+l(t)}{2}$, is estimated.
4. The mean $m(t)$ is subtracted from the original signal $s(t)$ to get the primary IMF,

   $$h(t) = s(t) - m(t)$$

5. The primary IMF $h(t)$ must satisfy the condition that the numbers of extrema and zero crossings are the same or vary by one at most and the zero mean condition. If this is not achieved, steps 1 to 4 are repeated on $h(t)$ as the basis input till the resultant IMF satisfies the conditions.

6. The residual is estimated between the IMF component $c_i(t)$ and the signal as $r(t) = s(t) - c_i(t)$.
7. Steps 1 to 6 are repeated on the residue $r(t)$.
8. Sifting process stops, when the residual has only one extremum.

After obtaining a number of IMFs, spectral subtraction is performed on each individual component, and then the original or enhanced speech signal is reconstructed by the superposition of the enhanced IMFs and the residual as shown in Fig. 5. This method decreases the musical noise owing to the averaging effect (El-Moneim et al. 2015).

## 4.3 Proposed adaptive noise canceller (ANC)

Noise is any objectionable distortion in the speech signal that causes reduction in the signal strength, and difficulty in hearing and reduces intelligibility of the speech. Consequently the performance of the SR system will be reduced. Removal of the noise corrupting the speech signal is the greatest shared problem in speech processing techniques. The ANC is a noise reduction filter whose parameters are changed to meet the required performance. The used criterion is the minimization of the MSE (Shanmugam et al. 2013) by using a normalized least mean square (NLMS) algorithm. Two input signals are applied to the ANC simultaneously: the contaminated signal as the primary input and the noise as the reference signal. The reference signal is filtered to get an approximation of noise corrupting the speech signal (Jafari and Chambers 2003). This estimated noise is subtracted from the corrupted speech signal to obtain an error signal, which is fed back to the adaptive filter to adjust the filter coefficients through the NLMS algorithm to minimize total MSE (Dhubkarya 2012).

If $s$ is the original speech signal and $n_o$ is the noise corrupting the speech, then the degraded speech will be,
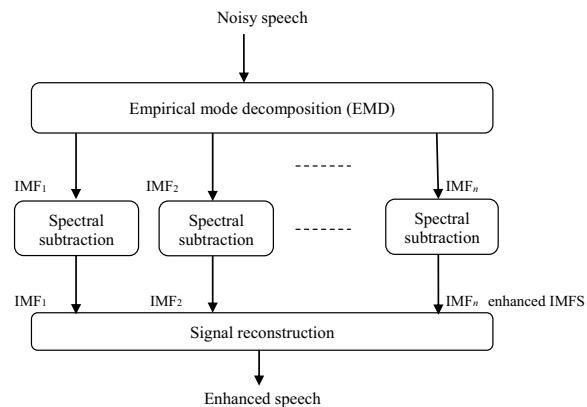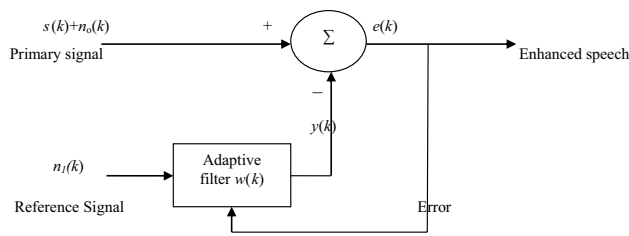
**Fig. 5** Flowchart of the hybrid method

**Fig. 6** Adaptive noise canceller

$$s_n = s + n_o \tag{13}$$

The error signal will be,

$$e = s_n - y \tag{14}$$

where $y$ is the filter output

The MSE will be,

$$E[e^2] = E\left[(s_n - y)^2\right] \tag{15}$$

$$E[e^2] = E[s^2] + E\left[(n_o - y)^2\right] + 2E[s(n_o - y)] \tag{16}$$

$n_o$ is uncorrelated with $s$,

$$E_{min}[e^2] = E[s^2] + E_{min}\left[(n_o - y)^2\right] \tag{17}$$

The term $(n_o - y)$ is minimized to lower the MSE (Anand 2013). The error is the improved version of the speech signal, which is used as the input to SR system as shown in Fig. 6.

Despite the ability of the ANC to remove noise from a noisy speech signal, it has some limitations such as the requirement of prior information about the signal and the noise characteristics.

### 4.4 Proposed Savitzky Golay (SG) filtering approach

The SG filter is a smoothing polynomial digital filter (Schafer 2011). For the signal samples, approximated coefficients are found during fitting with the help of the neighboring samples of the windowed signal. The SG filter is thought to be a generalization of the moving average filter that better preserves the features of the signal such as the height and width of peaks (Awal et al. 2011). The data samples are windowed with equal numbers of samples to the right and the left of the central sample. This central sample is replaced by the approximated coefficient, and then the window is lifted by one sample to the right to fit the next central sample. This process is repeated for the whole data samples (Guiñón et al. 2007; Shajeesh et al. 2012).

The output signal of the SG filter is calculated by:

$$g_m = \sum_{k=-n_L}^{n_R} c_{k+n_L} \cdot s_{m+k} \tag{18}$$

where $g_m$ is the output waveform, $c_{k+n_L}$ is the SG filter coefficients, $n_L$ is the number of samples to the left of the data sample $m$, and $n_R$ is number of the samples to the right of data sample $m$,

$$\mathbf{c}_{M+n_L} = \left[\left((\mathbf{A})^T.\mathbf{A}\right)^{-1}.\left((\mathbf{A})^T\right)\right]_{0,M+n_L} \tag{19}$$

where

$$M = -n_L, \cdots, n_R \tag{20}$$

The SG filtering is simple with high computational speed, but some of the first and last data samples cannot be smoothed.

## 5 Simulation results

A database of 8 speakers has been created. Each speaker repeated a certain sentence 10 times to obtain 80 samples of speech. Feature vectors consisting of 13 MFCCs and 26 polynomial coefficients have been created from these samples. 39 coefficients (13 MFCCs +26 polynomial coefficients) for each frame are fed to train the ANN. In the testing phase, each speaker is asked to say the same sentence, and additive white Gaussian (AWGN) is added to the speech signal with different signal-to-noise ratios (SNRs) from 0 to 30 dB. Features are extracted from the noisy speech signals and fed to the ANN for matching. Spectral subtraction is applied to reduce the noise and features are extracted from different transforms (DST, DCT, DWT) of the resultant enhanced speech signals. Moreover, the speech signals are also preprocessed with the hybrid method. Comparisons are presented between different scenarios: feature extraction from noisy speech, feature extraction after spectral subtraction, feature extraction after the hybrid method, feature extraction after the ANC and feature extraction after the SG filter. The recognition rate defined as the ratio between the number of success identifications and the total number of identification trials is estimated. The results are given in Figs. 7, 8, 9, 10 and 11. It is clear that the proposed methods achieve good results, and the DCT is a proper candidate with filtering methods in the cases of AWGN, while the DWT is the optimum transform for feature extraction, when the hybrid method and the spectral subtraction method are used.
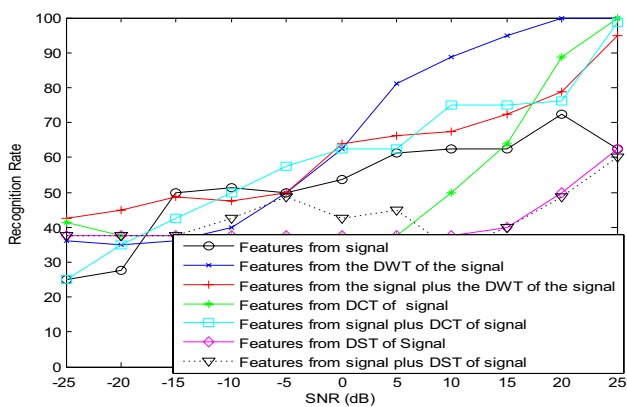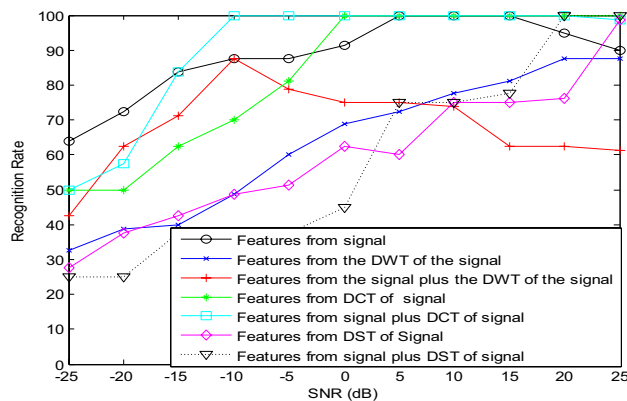
**Fig. 7** Recognition rate vs. SNR from different transforms of the noisy speech
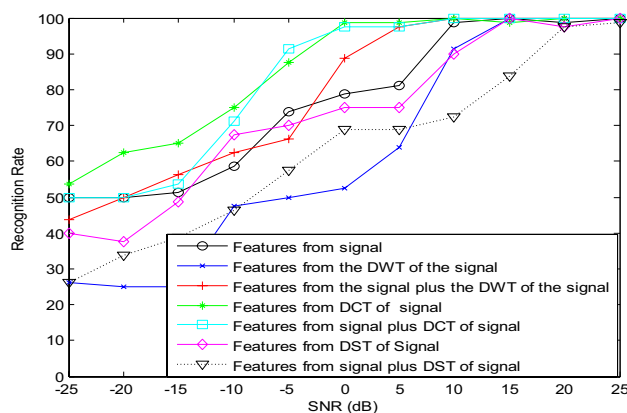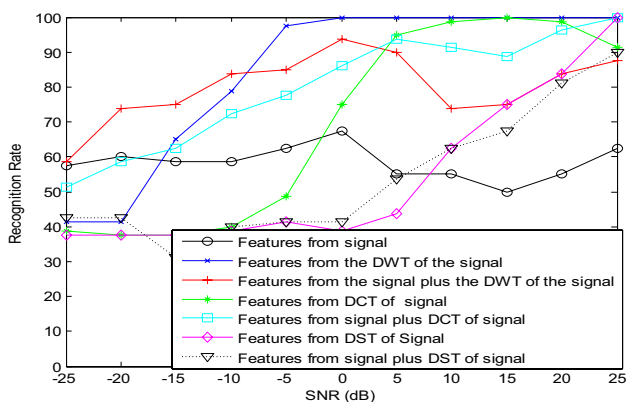


**Fig. 10** Recognition rate vs. SNR from different transforms after ANC method
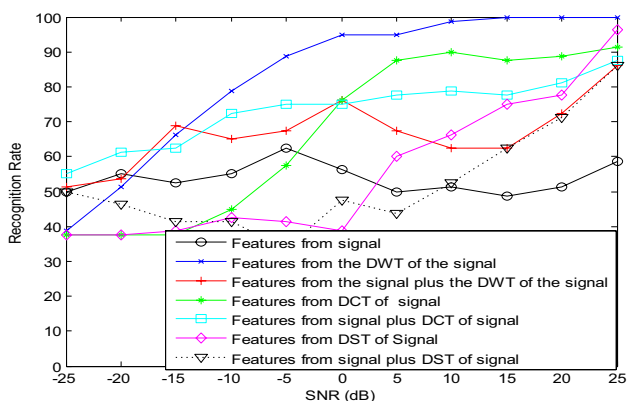


**Fig. 8** Recognition rate vs. SNR from different transforms after spectral subtraction method



**Fig. 11** Recognition rate vs. SNR from different transforms after SG filter

## 6 Conclusion

This paper presented some filtering methods for noisy speech to be used for robust SR systems. The suggested filters are the ANC and the SG. The MFCCs as features and the ANN as a classifier have been considered in the paper. Simulation results have shown a great success of the proposed filters compared to the spectral subtraction and the hybrid methods. In addition, the suggested methods, when applied prior to SR, lead to a noticeable improvement of performance. For the spectral subtraction method and the hybrid method, the DWT is the optimum transform for feature extraction, while the DCT is the optimum transform in the cases of ANC and SG filtering.



**Fig. 9** Recognition rate vs. SNR after the hybrid method

# References

Abd El-Samie, F.E. (2011). El-Samie 'information security for automatic speaker identification. Springer Briefs in Electrical and Computer Engineering.

Alotaiby, T., Alshebeili, S.A., Alshawi, T., Ahmad, I., Abd El-Samie, F.E. (2014). EEG seizure detection and prediction algorithms: A survey. *EURASIP Journal on Advances in Signal Processing*.

Anand, V., Shah, S., Kumar, S. (2013). Intelligent adaptive filtering for noise cancellation. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering* 2.

Awal, Md A, Mostafa, S. S., & Ahmad, M. (2011). Performance analysis of savitzky-golay smoothing filter using ECG signal. *IJCIT, 01,* 90–95.

Campbell, J.P. (1997). Speaker recognition: A tutorial, Senior Member, IEEE. In *Proceedings of IEEE*, Vol. 85.

Das, A., Jena, M. R., & Barik, K. K. (2014). Mel-frequency cepstral coefficient (MFCC)—A novel method for speaker recognition. *Digital Technologies, 1*(1), 1–3.

Dhubkarya, D. C., Katara, A., & Thenua, R. K. (2012). Simulation of adaptive noise canceller for an ECG signal analysis. *International Journal on Signal & Image Processing, 03*(01), 1–12.

Dreyfus, G. (2005). *Neural networks: Methodology and applications* (pp. 1–83). Berlin: Springer.

El-Moneim, S. A., Dessouky, M. I., Nassar, M. A., El-Naby, M. A., & Abd El-Samie, F. E. (2015). Hybrid speech enhancement with empirical mode decomposition and spectral subtraction for efficient speaker identification. *International Journal of Speech Technology, 2015*(18), 555–564. https://doi.org/10.10077 2-015-9293-5.

Evans, N.W.D., Mason, J.S., Liu, W.M. and Fauve, B. (2006). An assessment on the fundamental limitations of spectral subtraction. In *IEEE* (pp. 145–148).

Fisusi, A.A., Yesufu, T.K. (2007). Speaker recognition systems: A tutorial. *African Journal of Information and Communication Technology*, *3*(2).

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions of Acoustics, and Signal Processing, 29*(2), 254–272.

Galushkin, A. I. (2007). *Neural network theory*. Berlin: Springer.

Guiñón, J. L. et al. (2007). Moving average and savitzki-golay smoothing filters using mathcad. In *Proceedings of the International Conference on Engineering Education—ICEE*.

Hossain, M.M., Ahmed, B. and Asrafi, M. (2007). *A real time speaker identification using artificial neural network, conference: Computer and information technology, IEEE*.

Jafari, M.G., and Chambers, J.A. (2003). Adaptive noise cancellation and blind source separation. In *4th International symposium on independent component analysis and blind signal separation* (pp. 627–632).

Kaladharan, N. (2014). Speech enhancement by spectral subtraction method. *International Journal of Computer Applications, 96*(13), 45–48.

Karam, M., Khazaal, H. F., Aglan, H., & Cole, C. (2014). Noise removal in speech processing using spectral subtraction. *Journal of Signal and Information Processing, 5,* 32–41.

Kim, D., & Oh, H.-S. (2009). EMD: A package for empirical mode decomposition and hilbert spectrum. *The R Journal, 1*(1), 40–46.

Love, B.J., Vining, J., Sun, X. (2004). *Automatic speaker recognition using neural network, intro* (pp. 1–25). Neural Networks Electrical and Computer Engineering Department, The University of Texas at Austin, Spring.

Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *Journal of Computing, 2,* 138–143.

Nasr, Marwa A., et al. (2018). Speaker identification based on normalized pitch frequency and Mel frequency cepstral coefficients. *International Journal of Speech Technology*. https://doi.org/10.1007/s10772-018-9524-7.

Pawar, A.P., Choudhari, K.B. (2013). Enhancement of speech in noisy conditions. *The International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, *2*(7).

Reynolds, D.A. (2002). An overview of automatic speaker recognition technology. IEEE, pp. 4072–4075.

Rilling, G., Flandrin, P. and Goncalves, P. (2003). On empirical mode decomposition and its algorithms. In *IEEE-EURASIP workshop on nonlinear signal and image processing*.

Schafer, R.W. (2011). What is a Savitzky-Golay filter. *IEEE Signal Processing Magazine*.

Shajeesh, K. U., Sachin Kumar, S., & Soman, K. P. (2012). A two stage algorithm for denoising of speech signal. IOSR. *Journal of Computer Engineering (IOSRJCE), 8,* 48–53.

Shanmugam, A., et al. (2013a). Adaptive noise cancellation for speech processing in real time environment. *International Journal of Engineering Research and Applications (IJERA), 3*(2), 1102–1106.

Shanmugam, et al. (2013b). Adaptive noise cancellation for speech processing in real time environment. *International Journal of Engineering Research and Applications (IJERA), 3,* 1102–1106.

Sharma, A., Singh, S. P., & Kumar, V. (2005). *Text-independent speaker identification using backpropagation MLP network classifier for a closed set of speakers* (pp. 665–669). IEEE International Symposium on Signal Processing and Information Technology: INDIA.