Journal of Advanced Engineering Trends
ISSN : 2682 - 2091

Vol.43, No.1. January 2024

http://jaet.journals.ekb.eg

An efficient Automatic Speaker Identification System based Pitch Frequency Estimation in Degraded Environmental Conditions

Amira Shafik[1,2], Ashraf A. M. Khalaf[2], EL-Sayed M. El-Rabaie1[1], Fathi E. Abd El-Samie[1,3]

1 Department of Electronics and Electrical communications Engineering, Faculty of Electronic Engineering, Menoufia university, Menouf 32952, Egypt.

2 Department of Electrical Engineering, Electronics and Communications Engineering, Faculty of Engineering, Minia University, Minia 61111, Egypt.

3Department of Information Technology, College of Computer and Information Sciences, Princess NourahBint Abdulrahman University, Riyadh 21974, Saudi Arabia

A R T I C L E I N F O

Article history:
Received:
Accepted:
Online:

Keywords:
Speaker identification
Feature extraction
Normalized pitch frequency
Speech enhancement

A B S T R A C T

In this paper, we investigate the utilization of the Normalized Pitch Frequency (NPF) as an extracted feature from speech signals to be combined with the Mel Frequency Cepstral Coefficients (MFCCs) and polynomial coefficients. The objective is to compose more robust feature vectors to various forms of degradation such as Additive White Gaussian Noise (AWGN) and music interference. A matching process is performed to determine the identity of the unknown speaker, using a trained Artificial Neural Network (ANN) as a classifier. An Automatic Speaker Identification (ASI) system is presented in this paper comprising pre-processing methods based on Discrete Transforms (DTs) such as the Discrete Cosine Transform (DCT), the Discrete Sine Transform (DST), and the Discrete Wavelet Transform (DWT) for presenting robust features. Speech enhancement techniques such as Spectral Subtraction, Wiener filtering, adaptive Wiener filtering, and wavelet denoising are investigated to mitigate the impact of noise and improve the efficiency of the ASI system. Simulation results demonstrate that the utilization of the NPF with MFCCs as features extracted from both the speech signals and the DCTs of these signals increases the ASI system accuracy in the presence of noise and interference. The wavelet denoising enhances the proposed system effectiveness and gives high recognition rates even with very low Signal-to-Noise Ratios (SNRs).

1. Introduction

The ASI is designed to identify each speaker from his speech utterances through feature extraction and classification [1]. The feature extraction process eliminates the redundancy by extracting the essential speaker characteristics, and hence it is some sort of data reduction. The traditional types of features used for speaker identification include Linear Prediction Coefficients (LPCs), MFCCs, Linear Prediction Cepstral Coefficients (LPCCs), and Perceptual Cepstral Coefficients (PLPCs) [2]. The classification process includes speaker modeling and speaker matching stages. In the speaker modeling, features are extracted from the training data of each speaker and enrolled into the database. Using pattern matching, features from the input speech of an unknown speaker are mapped to a model that is compared to those of known speakers in the database through a selected classifier. Different types of classifiers can be considered in the matching task. These classifiers include Gaussian Mixture Models (GMMs), ANNs, Support Vector Machines (SVMs), Vector Quantization (VQ), Deep Neural Networks (DNNs) and Hidden Markov Models (HMMs) [3, 4].

Most ASI systems adopt MFCCs as features [5]. Unfortunately, MFCCs are not robust enough to noise. Hence, there is a need for additional features that can tolerate the noise effect. These features may include polynomial features and the NPF. The NPF will be considered in this paper for the task of speaker identification to enhance the matching accuracy in the presence of noise and interference.

Revised:30 May, 2022, Accepted:22 July , 2022

2. Literature Review

Speaker recognition is split into two categories: speaker verification and speaker identification [6]. Speaker verification task aims to determine the credibility of the person claim from his or her voice. A speech signal from an anonymous speaker is compared in the speaker identification process with those of the known speakers recorded in the database. The anonymous speaker is identified as a speaker, who gives the best match with a database model. The ASI can be categorized into two categories: closed-set and open-set speaker identification. In the first type, one of the speakers recorded in the database gives the test signal. In the second type, the test signal might be given by registered or unregistered speakers. The ASI operation may be text-dependent or text-independent.

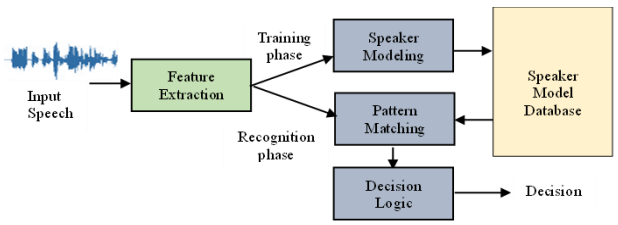


Figure 1: The ASI system structure.

The ASI includes two phases as shown in Figure 1. These are the training phase and the identification phase. During the training phase [7], speech signals are gathered from the speakers to construct their models, and hence the speaker database. In the identification phase, a comparison between an unknown speech signal and those stored in the database is performed to take a decision. Both phases involve feature extraction, in order to generate reliable feature vectors. In fact, the feature extraction is some sort of data reduction. In the enrollment phase, the features extracted from signals of known speakers are saved. On the other hand, in the recognition phase, features extracted from signals for an unknown speaker are compared to those stored after the enrollment phase. The result of this comparison allows speakers to be identified or declined.

Recently, a lot of research works adopted DTs such as DWT [5], DCT [8], and DST [9] for feature extraction. These DTs have an excellent energy compaction property, which is suitable for the elimination of the noise effect. Shafik et al. [8] proved that the identification of speakers with features extracted from DTs, such as DWT and DCT, gives better recognition rates. Li et al. [9] suggested the utilization of the DST in the speaker identification systems.

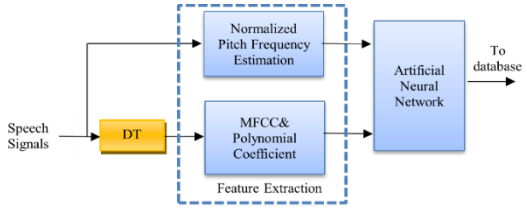
Modeling of the speech production system in the ASI process is very important to discriminate between speakers. One on the features that are important for this modeling process is the pitch frequency. It should be normalized to keep its value within a small dynamic range. Bai et al. [7] investigated the integration of the pitch frequency with MFCCs for ASI. They also studied Wiener filtering for noise reduction. Naser et al. [10] studied the NPF as an additional feature that can assist with MFCCs for ASI. Hassan et al. [11] adopted the maximum pitch frequency and the maximum cepstrum value for ASI with an SVM classifier.

L. Wang et al. [12] proposed a method that depends on pseudo-pitch synchronized phased information with maximum amplitude synchronization for speaker identification from telephone speech. In [13], Meng Ge et al. proposed a hybrid structure of pitch-synchronized relative phase data and MFCCs to minimize the noise effect on the ASI. A peak error detection scheme using an autocorrelation-based algorithm was also proposed. In [14], H. El-Kfafy et al. studied the influence of decoding and decompression on the ASI with techniques such as Compressive Sensing (CS) with features extracted from the DT domains. They proved the ability of that proposal to increase the system recognition rate, but at high SNRs without any kind of interference. In [15], S. A. El-Moneim proposed an Adaptive Noise Canceller (ANC) and Savitzky Golay (SG) filtering as pre-processing techniques in the ASI system to reduce the noise effect. The DCT, DST, and DWT have been used for feature extraction from noisy speech, but the authors did not consider any kind of interference.

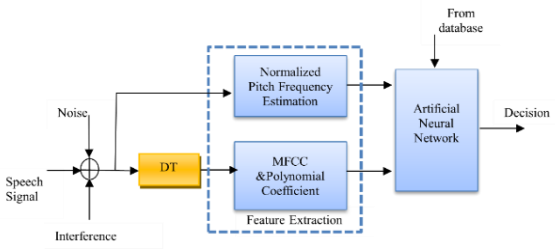
In our work, we present a robust ASI system that relies on a combination of MFCCs and NPF as extracted features and a neural network as a classifier. The system performance is investigated in different conditions including noise, and interference. Noise reduction techniques are investigated in this paper.

3. Methodology

The proposed ASI system is shown in Figure 2. The training is implemented with clean speech signals. The extracted features are cepstral features with polynomial coefficients in addition to the NFP. The testing phases are different. One of them depends on noisy signals with interference. The other one begins with speech enhancement to improve the performance. The motivation of all of these studies is to find a way that makes the features of speech signals more robust against various kinds of degradations.



(a)Training phase of the proposed ASI system



(b) Testing phase of the proposed ASI system in the presence of interference.

Figure 2: The proposed speaker identification system.

3.1 Discrete Transform

Discrete Transforms (DTs) such as the DWT, DCT and DST are investigated in this paper for more representative feature extraction as shown in Fig.2 (a, b). These transforms allow some sort of energy compaction.

3.1.1 DWT

The utilization of wavelet transform is an effective way to extract features from non-stationary speech signals as it is capable of extracting information about frequency and time. Furthermore, it can be used to reduce the noise effect on speech signals. Wavelet-based extraction of features depends on shifted, scaled versions of mother wavelets. The idea of the DWT depends on passing the input speech signal through a series of filters. The speech signal is decomposed by the usage of low-pass and high-pass filters with different scales as illustrated in Figure 7a. The outputs of the filters can be mathematically expressed as [5]:

$$y_A(k) = \sum_{n=-\infty}^{\infty} x(n) h_0(2k - n) \qquad (1)$$

$$y_D(k) = \sum_{n=-\infty}^{\infty} x(n) h_1(2k - n) \quad (2)$$

3.1.2 DCT

The DCT is a trigonometric transform that can be estimated for a speech signal *x* (*n*) as follows [8]:

$$X(k) = \propto (k) \sum_{n=0}^{N-1} x(n) cos\left[\frac{\pi(2n+1)k}{2N}\right] \quad (3) \ ,$$

$$0 \le k \le N-1$$

$$\propto (0) = \sqrt{\frac{1}{N}} \ , \text{ and } \quad \propto (k) = \sqrt{\frac{2}{N}}$$

3.1.3 DST

The mathematical representation of the DST is given by [9]:

$$X(k) = \sum_{n=0}^{N-1} x(n) sin\left[\frac{\pi}{N+1}(n-1)(K+1)\right]$$

$$, k = 0, \dots\dots\dots\dots, N-1 \quad (4)$$

***3.2** Normalized Pitch Frequency (NPF)*

The pitch frequency is defined as the frequency of vocal fold vibration caused by the air flow. The pitch frequency is related to the length, strength, hardness and articulation, which distinguish each speaker from the others. In this paper, we adopt a combination of NPF and MFCCs for speaker identification as shown in Figure 2. There are three categories for pitch frequency detection: waveform methods, transform-domain methods and hybrid methods. The waveform estimation is the most common in the processing of speech signals. Two of the popular methods are Auto-correlation Function (ACF) and Average Magnitude Difference Frequency (AMDF). We use a combination of these methods (ACF/ AMDF) in this paper [16].

3.2.1 Auto-Correlation Function (ACF)

A short-time ACF for a signal *x*(*n*) is computed as:

$$R(k) = \sum_{n=0}^{N-k-1} x(n) x(n+k) \quad (5)$$

where *N* is the signal length, and *k* is the time lag index (maximum delay). The benefits of this method are simplicity, and accuracy. The ACF method is suitable for noisy environments.

***3.2.2** Average Magnitude Difference Function (AMDF)*

The AMDF takes the absolute value of the difference between the original signal and the delayed version of it instead of the product of them to decrease the computational complexity, which makes the AMDF more suitable for the real-time applications. The AMDF is defined as:

$$D(k) = \frac{1}{N-k} \sum_{n=1}^{N-k} |x(n) - x(n+k)| \quad (6)$$

3.2.3 Combined ACR/ AMDF

This combination is represented as follows:

$$s(k) = \frac{R(k)}{1 + D(k)} \quad (7)$$

The NPF is defined as the pitch frequency divided by the maximum signal frequency.

4. Speech Enhancement

Speech enhancement is included to enhance the performance of the speaker identification process as shown in Figure 3.

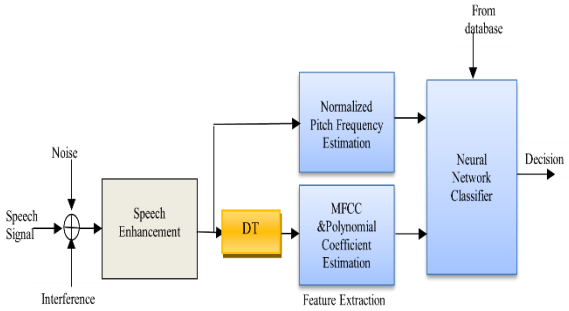


Figure 3: Testing phase of the proposed ASI system based on speech enhancement.

*4.1 Spectral Subtraction***:** It is aimed to suppress the effect of noise on speech signals [17]. A noisy speech signal is given as:

$$x(n) = s(n) + d(n) \quad (8)$$

where *s* (*n*) is the clean speech signal, and *d* (*n*) is the noise. Applying the Fourier transform gives:

$$X(e^{j\omega}) = S(e^{j\omega}) + D(e^{j\omega}) \quad (9)$$

The spectral process is represented with the following equations:

$$\hat{S}(e^{j\omega}) = \left[|X(e^{j\omega})| - \mu(e^{j\omega})\right] e^{j\Theta_x(e^{j\omega})} \quad (10)$$

$$\hat{S}(e^{j\omega}) = H(e^{j\omega}) X(e^{j\omega}) \quad (11)$$

$$H(e^{j\omega}) = 1 - \frac{\mu(e^{j\omega})}{|X(e^{j\omega})|} \quad (12)$$

$$\mu(e^{j\omega}) = E\{|D(e^{j\omega})|\} \quad (13)$$

4.2 Wiener Filter: This filter is defined with [18]:

$$\hat{S}(e^{j\omega}) = H(e^{j\omega})X(e^{j\omega}) \quad (14)$$

where

$$H(e^{j\omega}) = \frac{P_s(e^{j\omega})}{P_s(e^{j\omega}) + P_d(e^{j\omega})} \quad (15)$$

$P_s(e^{j\omega})$ is power spectrum of the speech signal and $P_d\ (e^{j\omega})$ is the power spectrum of noise

The *SNR* is defined as:

$$SNR = \frac{P_s(\omega)}{P_d(\omega)} \quad (16)$$

This leads to:

$$H(\omega) = \left[1 + \frac{1}{SNR}\right]^{-1} \quad (17)$$

4.3 Adaptive Wiener Filter:

For a stationary noise with zero mean and σ_d^2 variance, we can get [19]:

$$P_d(e^{j\omega}) = \sigma_d^2 \quad (18)$$

The speech signal can be represented according to its mean and variance as:

$$x(n) = m_x + \sigma_x w(n) \quad (19)$$

The Wiener filter transfer function can be approximated by:

$$H(e^{j\omega}) = \frac{P_s(e^{j\omega})}{P_s(e^{j\omega}) + P_d(e^{j\omega})} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_d^2} \quad (20)$$

Working on small speech segments leads to the following simplification of the Wiener filter impulse response:

$$h(n) = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_d^2}\delta(n) \quad (21)$$

Hence, the obtained enhanced speech signal $\hat{s}(n)$ can be expressed as [1]:

$$\hat{s}(n) = m_x + (x(n) - m_x * \frac{\sigma_s^2}{\sigma_s^2 + \sigma_d^2}\delta(n)) = m_x + \frac{\sigma_s^2}{\sigma_s^2 + \sigma_d^2}(x(n) - m_x) \quad (22)$$

4.4 Wavelet Denoising: Wavelet denoising or wavelet thresholding is intended to minimize noise effect on a noisy speech signal. The wavelet transform concentrates the signal power in approximation coefficients. The detail coefficients contain the effect of noise, and hence they can be thresholded with hard or soft thresholding as follows [19]:

$$f_{hard}(x) = \begin{cases} x, & |x| \geq T \\ 0, & |x| < T \end{cases} \quad (23)$$

$$f_{soft}(x) = \begin{cases} x & |x| \geq T \\ 2x - T & T/2 \leq x < T \\ T + 2x & -T < x \leq -T/2 \\ 0 & |x| < T/2 \end{cases} \quad (24)$$

where *T* represents the threshold value

5. Artificial Neural Network (ANN) as a classifier

The classification process is the matching between unknown speaker features and registered speaker features stored in the database. Multi-layer perceptron (MLP) network is one of the most popular neural network architectures. It is composed of an input layer and an output layer with one or more hidden layers between them. Machine learning strategy used with MLP is supervised learning, which is also called back-propagation algorithm. The neuron of the input and output layers have linear activation function, and the hidden neurons have a sigmoid activation function. The sigmoid activation function can

$$F(u) = \frac{1}{(1+e^{-u})} \quad (25)$$

$$E = \frac{1}{2}\sum_{o=1}^{O}(D_o - Y_o)^2 \quad (26)$$

where D_o and Y_o are the target and actual outputs of the Oth output neuron. The O refers to the number of output neurons. Several training iterations are executed to minimize E until a satisfactory small value is obtained, or a given number of epochs is reached. The error back-propagation algorithm can be used for this task

6. Experiments and Results

6.1 Dataset Preparation

A dataset containing recordings of ten speakers (male and female) was constructed [20]. Each speaker was asked to utter a sentence in Arabic language 10 times. Thus, 100 speech signals are used to generate NPF, MFCCs and polynomial coefficients to form the feature vectors of the database model. These features are used to train the back- propagation neural network which consists of three layers: input layer, hidden layer, and output layer. The hidden layer with 125 neurons and sigmoid activation function. In the testing phase, each one of these speakers is asked to say the sentence again and his/her speech signal is then degraded with AWGN with SNR from -25dB to 25 dB and high-power music interference. The features used in all experiments are 13 MFCCs and 26 polynomial coefficients forming a feature vector of 39 coefficients plus NPF for each frame of the speech signal. The system is implemented under the MATLAB R2018a.

6.2 Results and Discussion

In the simulation experiments, different types of degradation scenarios are considered, including noise, and interference. Arabic speech signals are considered for 10 speakers with 10 signals for each speaker. The SNRs range from -25 to 25 dB. Both MFCCs, polynomial coefficients and NPF are used as

features. The system is implemented with the Matlab (R2018a) on a machine with core i7 CPU, and 16 GB RAM. Different schemes for feature extraction are considered. Features are extracted from the signals or their transforms. Feature vectors generated from feature concatenation are also considered.

Three test scenarios are considered. In the first scenario, the features are extracted from noisy speech signals in the presence of music interference. The results of this scenario are given in Figure (4). In the second scenario, the features are extracted from the enhanced speech signals with different enhancement methods. The results of the ASI with spectral subtraction are given in Figure (5). On the other hand, the results with Wiener and adaptive Wiener filters are shown in Figures (6) and (7), respectively. Figures (8) to (15) give the ASI results with wavelet denoising as a preprocessing stage. Different types of wavelets are considered for single- and two-level decomposition. In addition, both soft and hard thresholding are considered. The recognition rate can be computed according to the following equation:

$$\textbf{\textit{Recognition Rate}} = \frac{\textit{the number of success identifications}}{\textit{total number of identification trials}} \quad (27)$$

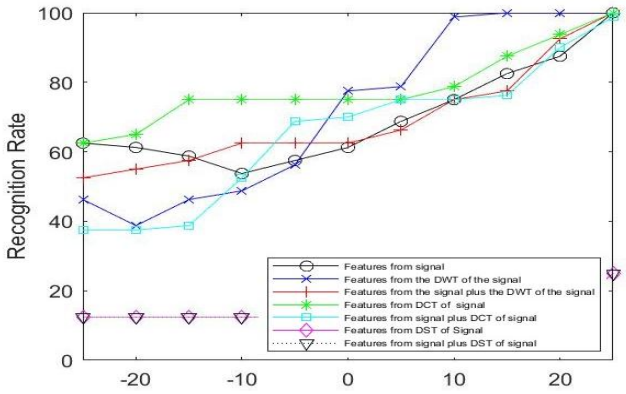
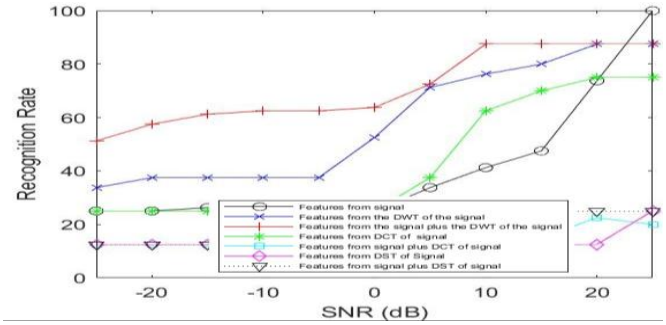


Figure 4: Recognition rate versus SNR for ~~the proposed methods~~the ASI system on speech signals contaminated by AWGN ~~with~~ and interference**.**

Figure 5: Recognition rate versus SNR for the ASI system on speech signals contaminated by AWGN and interference with a preprocessing spectral subtraction stage.



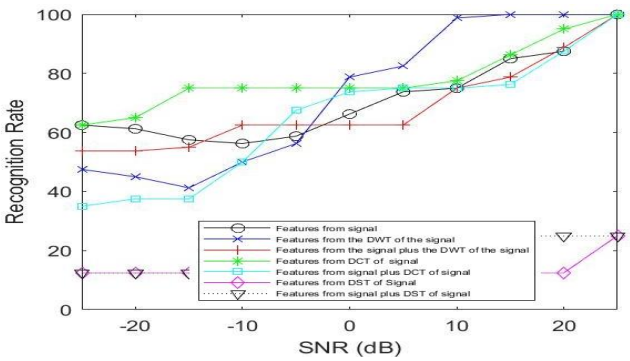


Figure 6: Recognition rate versus SNR for the ASI system on speech signals contaminated by AWGN and interference with a preprocessing Wiener filtering stage.

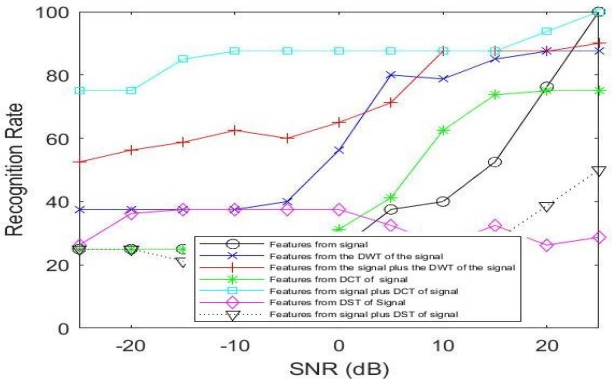


Figure 7: Recognition rate versus SNR for the ASI system on speech signals contaminated by AWGN and interference with a preprocessing adaptive Wiener filtering stage.

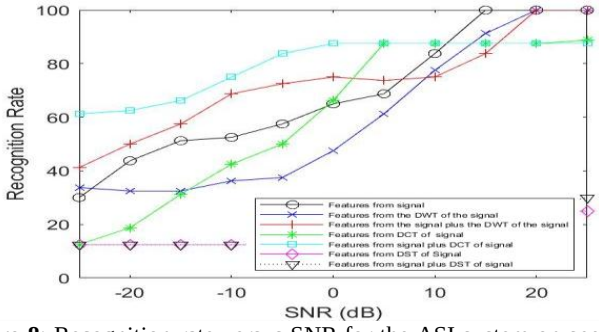


Figure 8: Recognition rate versus SNR for the ASI system on speech signals contaminated by AWGN and interference with a preprocessing wavelet soft thresholding stage (single-level Daubechies wavelet).

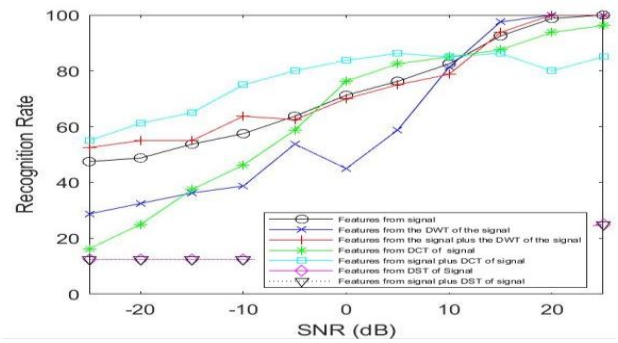


Figure 9: Recognition rate versus SNR for the ASI system on speech signals contaminated by AWGN and interference with a preprocessing wavelet hard thresholding stage (single-level Daubechies wavelet*)*

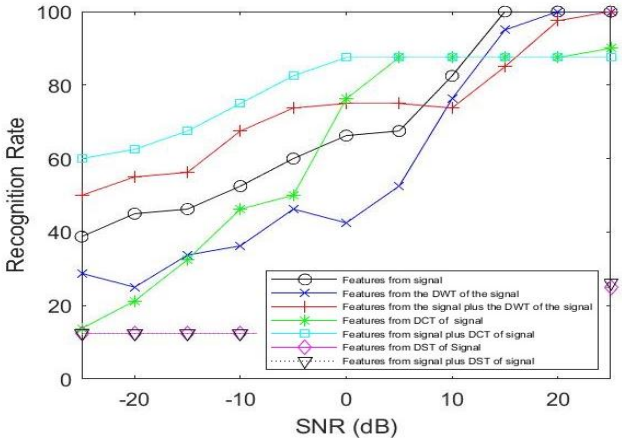


Figure 10: Recognition rate versus SNR for the ASI system on speech signals contaminated by AWGN and interference with a preprocessing wavelet soft thresholding stage (single-level Haar wavelet).

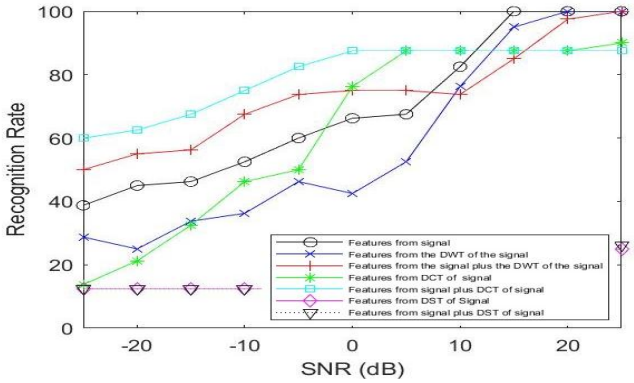


Figure 11: Recognition rate versus SNR for the ASI system on speech signals contaminated by AWGN and interference with a preprocessing wavelet hard thresholding stage (single-level Haar wavelet).

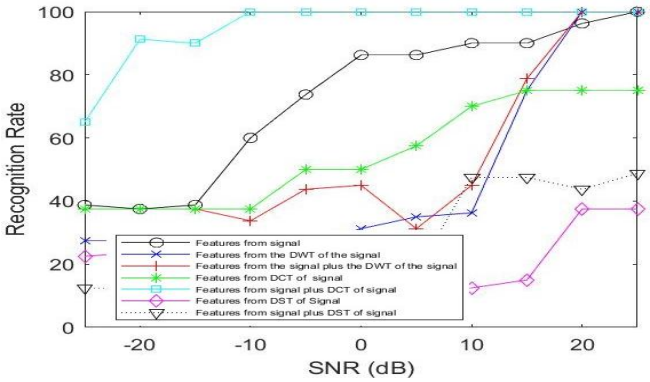
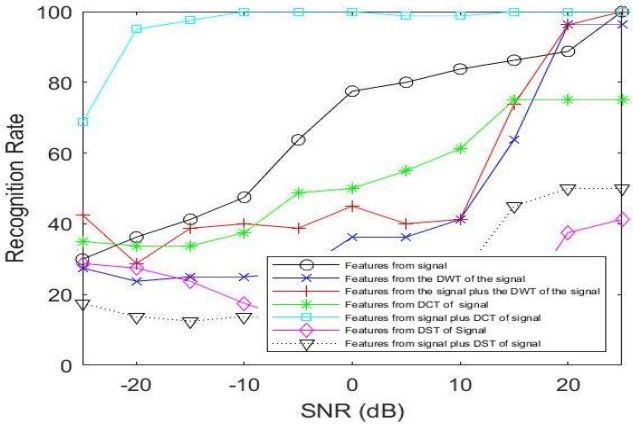


Figure 12: Recognition rate versus SNR for the ASI system on speech signals contaminated by AWGN and interference with a preprocessing wavelet soft thresholding stage (two-level Daubechies wavelet).

Figure13: Recognition rate versus SNR for the ASI system on



speech signals contaminated by AWGN and interference with a preprocessing wavelet hard thresholding stage (two-level Daubechies wavelet).

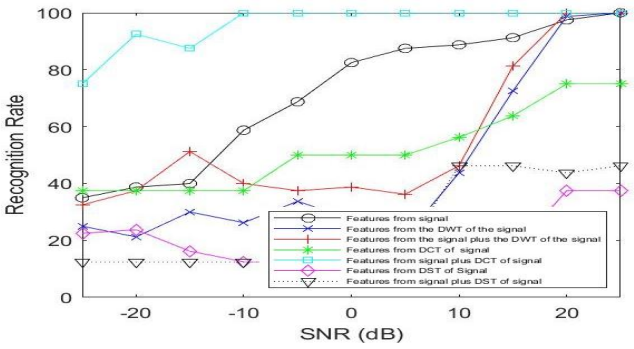


Figure 14: Recognition rate versus SNR for the ASI system on speech signals contaminated by AWGN and interference with a preprocessing wavelet soft thresholding stage (two-level Haar wavelet).

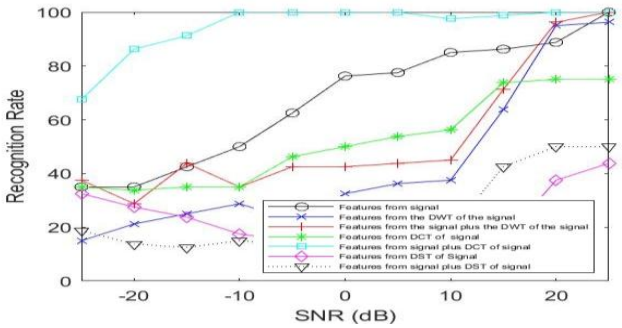


Figure.15: Recognition rate versus SNR for the ASI system on speech signals contaminated by AWGN and interference with a preprocessing wavelet hard thresholding stage (two-level Haar wavelet).

7. Comparison Study

A comparison study is presented between the proposed ASI system and the traditional ones at different SNRs ranging from -10 to 10 dB. The results of this comparison are given in Figures 16,17,18,19 and 20. From this comparison, we can come to a conclusion that it is possible to identify speakers with high recognition rates up to 100%, even with degradations including AWGN and interference. The interference effect was not considered in the related work. The best detection accuracy is achieved with the proposed system based on features from the signals and the DCTs of these signals concatenated together with a wavelet denoising enhancement stage.

In [21] N. A. Hindawi, I. Shahin, eta al., introduced a system based on modified Support Vector machine as a classifier to enhance the ASI performance under an extreme high-pitched condition in a neural taking environment. The performance equal 93.95%.

In our work we enhance the speaker features to be more robust against severe degradations such as AWGN and high-power music interference with ANN as classifier, the recognition rate reaches to 100% even with high SNR as illustrated in table (1).

Table 1: Recognition rates (%) of the proposed ASI system on speech signals contaminated of AWGN and interference using the wavelet hard thresholding with two-levels Daubechies wavelet.

SNR (dB)	Features from speech signal	Features from DWT of signal	Features from signal+ DWT of signal	Features from DCT of signal	Features from signal+ DCT of signal	Features from DST of signal	Features from signal+ DST of signal
-25	38.75	27.5	37.5	37.5	**65**	22.5	12.5
-20	37.5	27.5	37.5	37.5	**91.25**	23.75	12.5
-15	38.75	21.25	37.5	37.5	**90**	15	12.5
-10	60	27.5	33.75	37.5	**100**	12.5	12.5
-5	73.75	26.26	43.75	50	**100**	12.5	12.5
0	86.25	31.25	45	50	**100**	12.5	12.5
5	86.25	35	31.25	57.5	**100**	12.5	17.5
10	90	36.25	45	70	**100**	12.5	47.5
15	90	75	78.75	75	**100**	15	47.5
20	96.25	100	100	75	**100**	37.5	43.75
25	100	100	100	75	**100**	37.5	48.75

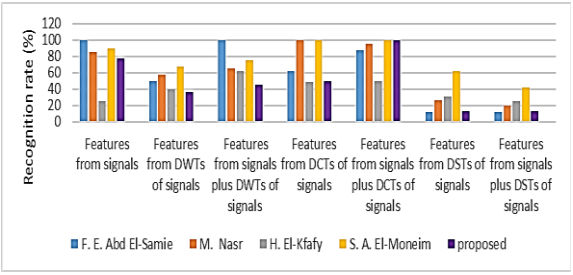


Figure 16: Comparison of the ASI systems at 0 dB.

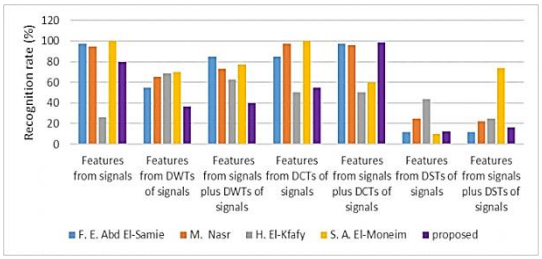


Figure 17: Comparison of the ASI systems at 5 dB.

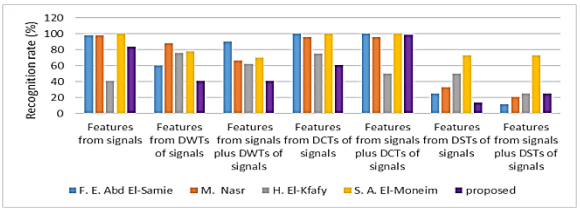


Figure 18: Comparison of the ASI systems at 10 dB.

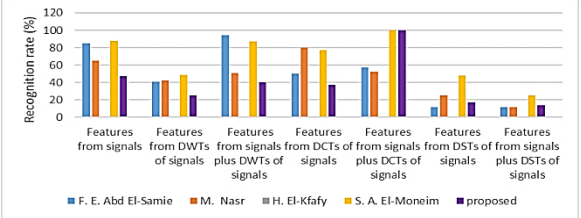
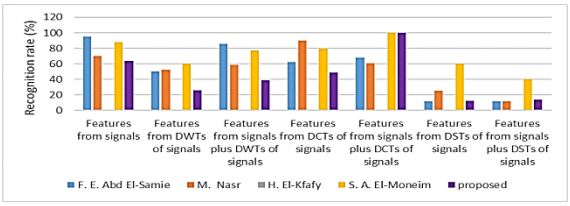


Figure. 19: Comparison of the ASI systems at -5 dB.



Fiure.20: Comparison of the ASI systems at -10 dB.

8. Conclusion and future work

This paper presented a robust ASI system to work on noisy speech signals in the presence of interference. The NPF and the MFCCs have been utilized as features based on DTs such as DCT, DST, and DWT. Features are extracted from degraded speech signals and discrete transforms of those signals. Enhancement techniques such as spectral subtraction, Wiener Filtering, adaptive Wiener filtering, and wavelet denoising are utilized to improve system performance and accuracy. Results show high recognition rates of the proposed ASI system even at low SNRs and high-power music interference with feature extraction from the speech signals and their DCTs, especially with wavelet denoising as an enhancement stage. In the future work, we will study the utilization of the wavelet denoising as a preprocessing stage to the a ASI system based on the deep learning Model and Radon transform.

Conflict of Interest

No conflict of interest.

Acknowledgements

I am extremely grateful to my supervisors, Prof. Dr. Fathi E. Abd El-Samie, Prof. Dr. S. El-Rabaie, Prof. Dr. Ashraf A. M. Khalaf for their invaluable advice, continuous support, and patience

References

[1] R. Jahangir, Y.W. Teh, N.A. Memon, G.Mujtaba, M .Zareei, U. Ishtiaq, "Text-independent speaker identification through feature fusion and deep neural network", IEEE Access 8:32187–32202, doi: 10.1109/ACCESS.2020.2973541

[2] H. Li et al., "The I4U system in NIST 2008 speaker recognition evaluation," in 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 19-24 April 2009, pp. 4201-4204, doi: 10.1109/ICASSP.2009.4960555.

[3] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," IEEE Transactions on Speech and Audio Processing, Vol. 3, 1995, pp. 72-83. http://dx.doi.org/10.1109/89.365379

[4] Z. Kozhirbayev, B. A. Erol, A. Sharipbay and M. Jamshidi, "Speaker Recognition for Robotic Control via an IoT Device", World Automation Congress (WAC),2018, pp.1-5,doi: 10.23919/WAC.2018.8430295.

[5] A. Shafik, S. M. Elhalafawy, S. M. Diab, B. M. Sallam and F. E. Abd El-samie," A wavelet-based approach for speaker identification from degraded speech", Int J Commun Netw Info Secur (IJCNIS) 1(3):53–60, DOI: https://doi.org/10.54039/ijcnis.v1i3.23

[6] D. A. Reynolds, "An overview of automatic speaker recognition technology," in 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, 13-17 May 2002, vol. 4, pp. IV-4072-IV-4075, doi: 10.1109/ICASSP.2002.5745552.

[7] B. Junmei, Z. Rong, X. Bo, and Z. Shuwu, "Robust speaker recognition integrating pitch and Wiener filter," in 2004 International Symposium on Chinese Spoken Language Processing, 15-18 Dec. 2004 2004, pp. 69-72, doi: 10.1109/CHINSL.2004.1409588.

[8] A. Shafik, S. Elhalafawy, S. M. Diab, and B. M. Sallam, "DCT assisted speaker identification in the presence of noise and channel degradation," in 2009 International Conference on Computer Engineering & Systems, 14-16 Dec. 2009, 2009, pp. 191-196, doi: 10.1109/ICCES.2009.5383285.

[9] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," IEEE/ACM Trans. Audio, Speech and Lang. Proc., vol. 22, no. 4, pp. 745–777, 2014, doi: 10.1109/taslp.2014.2304637.

[10] M. A. Nasr, M. Abd-Elnaby, A. S. El-Fishawy, S. El-Rabaie, and F. E. Abd El-Samie, "Speaker identification based on normalized pitch frequency and Mel Frequency Cepstral Coefficients," International Journal of Speech Technology, vol. 21, no. 4, pp. 941-951, 2018/12/01 2018, doi: 10.1007/s10772-018-9524-7.

[11] B. Hassan, R. Ahmed, B. Li, O. Hassan, and T. Hassan, "Autonomous Framework for Person Identification by Analyzing Vocal Sounds and Speech Patterns," in 2019 5th International Conference on Control, Automation and Robotics (ICCAR), 19-22 April 2019, pp. 649-653, doi: 10.1109/ICCAR.2019.8813463.

[12] L. Wang, S. Nakagawa, Z. Zhang, Y. Yoshida, and Y. Kawakami, "Spoofing Speech Detection Using Modified Relative Phase Information," IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 4, pp. 660-670, 2017, doi: 10.1109/JSTSP.2017.2694139

[13] M. Ge, L. Wang, S. Nakagawa, Y. Kawakami, J. Dang, and X. Li, "Pitch Synchronized Relative Phase with Peak Error Detection For Noise-robust Speaker Recognition," in 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), 26-29 Nov. 2018 2018, pp. 156-160, doi: 10.1109/ISCSLP.2018.8706701

[14] H. S. El-Kfafy et al., "Efficient remote access system based on decoded and decompressed speech signals," Multimedia Tools and Applications, vol. 79, no. 31, pp. 22293-22324, 2020/08/01 2020, doi: 10.1007/s11042-019-08150-7.

[15] S. A. El-Moneim et al., "Speaker recognition based on pre-processing approaches," International Journal of Speech Technology, vol. 23, no. 2, pp. 435-442, 2020/06/01 2020, doi: 10.1007/s10772-019-09659-w.

[16] L. Zhijun, H. Xuelong, and G. Na, "Judicial expertise of speaker identity based on improved pitch algorithm," in 2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI), 20-22 Oct. 2017 , pp. 401-405, doi: 10.1109/ICEMI.2017.8265832.

[17] B. Saha, S. Khan, C. Shahnaz, S. A. Fattah, M. T. Islam, and A. I. Khan, "Configurable Digital Hearing Aid System with Reduction of Noise for Speech Enhancement Using Spectral Subtraction Method and Frequency Dependent Amplification," in TENCON 2018 - 2018 IEEE Region 10 Conference, 28-31 Oct. 2018, pp. 0735-0740, doi: 10.1109/TENCON.2018.8650450.

[18] Widrow B, Stearns SD (1985) Adaptive Signal Processing. Prentice-Hall, Upper Saddle River.

[19] R. Hidayat, A. Bejo, S. Sumaryono, and A. Winursito, "Denoising Speech for MFCC Feature Extraction Using Wavelet Transformation in Speech Recognition System," in 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE), 24-26 July 2018 2018, pp. 280-284, doi: 10.1109/ICITEED.2018.8534807.

[20] A. Shafik et al., "Speaker identification based on Radon transform and CNNs in the presence of different types of interference for Robotic Applications," Applied Acoustics, vol. 177, p. 107665, 2021/06/01/ 2021, doi: https://doi.org/10.1016/j.apacoust.2020.107665.

[21] N. A. Hindawi, I. Shahin, and A. B. Nassif, "Speaker identification for disguised voices based on modified SVM classifier," in Proceedings of the Multi-Conference on Systems, Signals & Devices (SSD), pp. 687–691, IEEE, Monastir,Tunisia, March 2021.

Abbreviation and symbols

MFCC	Mel Frequency Cepstral Coefficients
AWGN	Additive White Gaussian Noise
ANN	Artificial Neural Network
ASI	Automatic Speaker Identification
DWT	Discrete Wavelet Transform
DCT	Discrete Cosine Transform
DST	Discrete Sin Transform
SS	Spectral Subtraction
SNR	Signal to Noise Ratio
NPF	Normalized Pitch Frequency
AMDF	Average Magnitude Difference Frequency
ACF	Auto-Correlation Function