Enhancement of speech signal denoising based on MFCC and Robust Principal Component Analysis RPCA

Sonia Moussa[1], Zied Hajaiej[1], Ali Garsallah[2]

[1]Laboratory of Signal, Image and Information Technology (LSITI) National Engineering School of Tunis (ENIT) BP 37, the Belvedere, 1002, Tunisia
[2]Laboratory of High Frequency Electronic Circuits and Systems, Faculty of Mathematical, Physical and Natural Sciences Of Tunis (FST) University of Tunis El Manar Tunis University Campus PB 94 - Rommana 1068, Tunisia

Summary
In the automatic speech recognition system, several techniques of feature extraction have been studied at different values of signal-to-noise ratio. This paper suggests to develop a new approach of the speech signal such as MFCC-RPCA in order to obtain a higher recognition rate. Thus, MFCC is one of the most commonly used features for speech recognition systems. In previous years, the research on robust principal component analysis (RPCA) has been attracting much attention, in many domains, such as image processing, separation of music/voice, etc. The purpose of this paper is based on the separation speech/noise. In the experimental part, isolated words were chosen from TIMIT database, under additive impulsive and convolutive noise conditions, with SNR (signal to-noise-ratio) ranges from -3 to more than +9 db using the HTK platform (Hidden Markov Model Toolkit). Experimental results have shown that the proposed method has enhanced the quality of signals by reducing the noise level.

Key words:
Speech Recognition; MFCC RPCA; MFCC-RPCA; TIMIT database.

1. Introduction

Speech denoising aims to improve the quality and intelligibility of the speech signal and therefore improve the performance of related applications.
Recently, several techniques of signal analysis have been applied in audio and speech denoising with relatively good results in controlled conditions. However, the signal may be corrupted by a wide variety of sources in such environments including: additive noise, linear and non-linear distortion, transmission and coding effects, and other phenomena. Thus, a number of works have explored the use of auditory models for building robust speech recognition system. However, a common approach to recover speech signals from noisy observations is a speech enhancement technique which estimates and removes the noise from the spectrum of the input speech signal [2]. Furthermore, Automatic Speech Recognition (ASR) has always been a scientific challenge. Many research efforts have been made over recent years to offer solutions and aiding systems in order to enhance the speech signal denoising.
Candès has proposed a new theory, in 2011, it is called RPCA to remedy the deficiency of PCA, on the one hand. This is an unsupervised technique which consists in decomposing a matrix into low-rank and sparse structures, using a convex optimization [13].
On the other hand, RPCA is used in face recognition into separating singing voices from music accompaniment (Huang, 2012) [9], which takes advantage of low-rank, i.e. repetition of music sound, and sparsity of speech signal in the spectral domain. Actually, a variety of noises present a similar repeating structure to music.
Furthermore, (Minghe Wang, 2016) [10], exploited in his article the RPCA method into the TVS modeled speaker verification system, called RPCA-TVS, which improved the robustness of speaker verification under additive noisy environment, especially in non-stationary and unseen noise conditions.
In addition, (Chengli Sun, 2014) [7], conducted a study in his article on the benefit of the RPCA method in speech / noise separation in which he experimentally proved that the RPCA based speech enhancement method can steadily obtain higher noise suppression performance in noisy conditions, compared to many traditional methods.
This paper considers the possibility of improving the performance of a noise robust automatic speech recognition (ASR) system by the integration of MFCC-Robust Principal Component Analysis (RPCA) algorithm for noise suppression. Furthermore, RPCA has the potential to recover clean speech from distorted speech under various types of noises and conditions, in Short Time Fourier Transform (STFT) domain.
The following sections, suggests a description on how a speech signal is converted into sequences of MFCC feature vectors and how these vectors are then processed by the RPCA model applying thereafter a HMM-based recognition system in order to estimate the sequences of spoken words. This paper is organized as follows: Section 2 explains the implementation of our proposed method. Section 3, describes the architecture of our provided model, the experimental results and their improvements in

Manuscript received March 5, 2019
Manuscript revised March 20, 2019

different scenarios are presented. Finally, some concluding remarks with suggestions for future directions of this work are provided.

2. The proposed model (MFCC-RPCA)

2.1 Mel frequency cepstral coefficients cepstrum MFCC

Mel-frequency cepstral coefficients (MFCC) are used as the primary audio features. They have been widely used in audio signal processing problems, for example, speech recognition, audio retrieval, and emotion recognition in speech, especially with HMM classifiers. The process of traditional MFCC is shown in figure 1 [5].

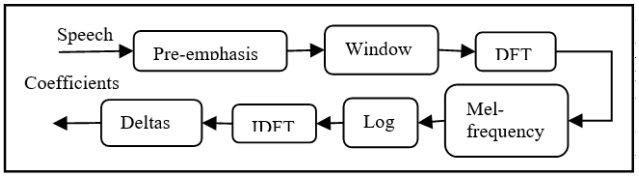


Fig. 1 Blok diagram of the traditional's MFCC process

Figure 1 illustrates the extraction of the MFCCs from a continuous speech signal. The speech signal is pre-emphasized to remove the effects of glottal and lip radiation. Pre-emphasis is performed with a first order Finite Impulse Response (FIR) filter of the form.

$$H(z) = 1 - az^{-1}\,, where\ 0.9 \le a \le 0.99 \qquad (1)$$

In the MFCC extracting stage, the log operation leads the additive noise in the spectral domain to be very complex in the cepstral domain. Therefore, the measurement sequence *y(m)* is preprocessed to obtain $y_i(m)$, where i is the number of frame, then each frame speech signal transition, to the frequency domain transforms from time domain using FFT or DCT, and it can be expressed as [5]:

$$Y_i(\omega) = FFT[y_i(m)] \qquad (2)$$

And the energy of each frame can be expressed as :

$$E_i(\omega) = [Y_i(\omega)]^2 \qquad (3)$$

The mel frequency, which is a unit of pitch defined so that pairs of sounds which are perceptually equidistant in pitch are separated by an equal number of mels, can be computed from
the raw acoustic frequency as follows [15]:

$$mel(f) = 2595 * \ln(1 + \frac{f}{700}) \qquad (4)$$

At the stage of extracting MFCCs, the first 12 cepstral values are chosen. These 12 coefficients will represent information solely about the vocal tract filter, cleanly separated from information about the glottal source.
We do this by adding for each of the 13 features (12 cepstral features plus energy) a delta or velocity feature, and a double delta or acceleration feature [20].

2.2 Principal Components Analysis PCA

Recently, several authors have proposed a new theory called RPCA, which can remedy the deficiency of PCA, which have talked about this approach in separation of voice/music.
Firstly, the properties and their drawbacks of the Principal Component Analysis PCA approach to achieve our goal of our proposed method are going to be mentioned.
PCA is a useful statistical technique that has found application in fields such as face recognition and image compression, and is a common technique for finding patterns in data of high dimension [7] [16]. The PCA is mainly used for: describe and visualize data; decorrelate them, whose the new basis consists of axes that are not correlated with each other; denoise them considering that the axes which one decides to forget are noisy axes.
Suppose that the single-channel noisy speech signal observation vector $x(t) \in \mathbb{R}^L$ can be constructed by the sum of the clean speech vector *s*(*t*) and the noise vector *n*(*t*) [17]:

$$x(t) = s(t) + n(t) \qquad (5)$$

We can represent on a two-dimensional plane, the available points according to the joint law of $x_1\ and\ x_2$. The determination of the two axes that explain the dispersion of the available points represents the result of a PCA.
The PCA method determines always the axes that show the dispersion of the cloud of available points and orders them by inertia. We can project the cloud of dimension n on a plane, and visualize it, if the decision is to retain only the first two axes of the PCA.
This approach is used, especially for visualizing data. It also allows:

- to decorrelate them in the new base made up of the new axes. In this case, the points have zero correlation;
- to reduce the noise that affects them, considering that the axes that are eliminated are noisy axes;
- to classify them into correlated clusters.

The principle of the PCA approach is to find an axis u. This axis is obtained from a linear combination of variables X_n, such that the variance of the cloud around this axis is maximal

The covariance matrix will be diagonalized, and its elements represent the eigenvalues λ_i. The first eigenvalue λ_1 represents the empirical variance on the first axis u of the PCA. We continue then, looking for the second axis of projection that we will call w on the same principle, w is orthogonal to u. Therefore, the PCA method represents the diagonalization of the correlation matrix. We consider a set of L centered observations [21]

$$x(t) = [x_1, x_2, \dots, x_L]^T \; and \; \sum_{k=1}^{L} x_k = 0 \qquad (6)$$

From the vector x, we compute the covariance matrix C. We use the PCA tool to diagonalize the matrix [17] :

$$C = \frac{1}{L}\sum_{k=1}^{L} x_k \, x_k^T \qquad (7)$$

The empirical covariance matrix of the observations is then defined by:

$$C = \frac{1}{N} X X^T \qquad (8)$$

It then seems to diagonalize the covariance matrix C and so we can write:

$$CUDU^T \qquad (9)$$

Where U is an orthogonal matrix, $UUT = I$, I is an identity matrix and D is a diagonal matrix.

The principal objective of PCA technique, is firstly, to filter out the noise and reduce a multidimensional speech to lower dimensions by avoiding redundant data and secondly, re-express a noisy speech set [20].

Therefore, the disadvantage of PCA, in modern applications such as image processing, bioinformatics, and in the speech signal processing, where big errors are now appearing and some measurements may be arbitrarily corrupted or simply insignificant to the low-dimensional structure to be necessary identified, more precisely the principal components are usually linear combinations of all input variables [13].

2.3 Robust Principal Component Analysis RPCA

This method assumes that the background music have a low-rank structure and the vocal components have sparse structures, in the time-frequency domain. RPCA proves to be a very effective tool for extraction of the vocal section from a sample containing mixture of vocal and music [1] [12]. For the final acoustic modeling, the MFCC representation with the new approach RPCA was extended. This will be detailed in this section.

Robust principal component analysis (RPCA) via decomposition in low-rank and sparse matrices, is a modification of the widely used statistical procedure principal component analysis (PCA), which works well with respect to grossly corrupted observations.

In this paper, the RPCA approach is applied to speech and noise separation problem and a speech enhancement method based on this approach is proposed. Robust Principal Component Analysis is one of the recent methods used in vocal separation from a mixture of speech and noise. This method assumes that the background speech have low-rank structure and noise components have sparse structures, in the time-frequency domain [19].

Denote by $Y \in \mathbb{R}^{m\times n}$ the original data matrix, by $L \in \mathbb{R}^{m\times n}$ the low-rank component and by $E \in \mathbb{R}^{m\times n}$ the sparse component, RPCA can be mathematically described as the following convex optimization problem [12] [13] [19]:

$$min_{L,E}\|L\|_* + \lambda\|E\|_1 \quad s.t \; Y = L + E \qquad (10)$$

Where $\|L\|_* = \sum_r \sigma_r(L)$ denotes the nuclear norm of $L, \sigma_r(L)$ $(r = 1,2, \dots, \min(m,n)) is \; the \; r^{th}$ singular value of L, $\|E\|_1 = \sum_{i,j}|e_{i,j}|$ denotes the L_1-norm of E and $e_{i,j}$ is the element in the i^{th} row and j^{th} column of E [19].

Therefore, The RPCA algorithm assumes that the noise signal should be converted in time-frequency domain and the noise is treated as a low-rank component while the human speech is analyzed as a sparse component [6].

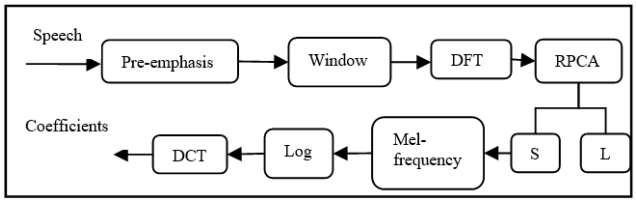


Fig. 2 Bloc diagram of the proposed method MFCC-RPCA

The figure illustrated above shows the proposed method and its different stages.

The most of traditional noise robust speech recognition approaches, was compared, which addressed to reduce for the noise impact after MFCC or i-vector extracting, for the reason not to deal with additive noise. This method insert RPCA based denoising algorithm into MFCC extraction phase, as shown in Fig 2. Based on the algorithm, the spectrum of the noisy speech, is firstly, generated by preprocessing and STFT, then RPCA decomposes the noisy speech spectrum into two matrices: low-rank matrix and sparse matrix [10].

If, as before, the clean speech signal is denoted with s(t) and the noise signal with n(t), it can be said that the speech signal y(t) is [11]:

$$y(t) = s(t) + n(t) \qquad (11)$$

Using short-time Fourier Analysis (STFT), the speech signal can be written as below:

$$Y(m,k) = \sum_{i=-\infty}^{+\infty} y(i)w(m-i)e^{\frac{-j2\pi ki}{L}} \quad (12)$$

Where $k \in \{1, \dots, L\}$, denotes the index of the discrete acoustic frequency, L is the length of the frequency analysis and m is the index of time-frame, w(m) is an analysis window function [11].

The normalisation of the magnitude $|Y(m,k)|$ is needed in each frame by averaging it with the magnitude values from adjacent three frames, therefore the enhanced speech signal Fourier transform detailed below is obtained by :

$$\hat{S}(m,k) = |S(m,k)|^{1/2} e^{j/Y(m,k)} \quad (13)$$

The enhanced speech signal that will result $\hat{s}(t)$ will be the inverse Fourier transform of the $\hat{S}(m,k)$ function.

3. Comparison and Analysis

Speech synthesis based on Hidden Markov models (HMM) has become a good choice for Text-To-Speech (TTS) due to its flexibility, small footprint and relatively high performance compared to concatenative speech synthesis.

HMMs realised in automatic speech recognition (ASR) typically use only spectral parameters, which are modeled by continuous distributions [8].

A voice analysis is done after taking an input through microphone from a user. At different levels, different operations are performed on the input signal such as Pre-emphasis, Framing, Windowing, Mel Cepstrum analysis and Recognition (Matching) of the spoken word.

The speech recognition algorithm consists of two distinguished phases. The first one is training sessions, whilst, the second one is referred to as operation session or testing phase [18].

we employed a subset of the well-known NOISEX-92 database for testing.

As input speech, the isolated words of TIMIT database were used to evaluate the performance of the proposed estimators. In this database 6132 words are used, which were composed of 21 words repeated, 292 times, 36 speakers (18 males and 18 females) for training uniformly divided on 8 American dialects. For the test phase of recognition we used 2201 words, 26 speakers (13 males and 13 females) repeated 104 times uniformly divided on 8 American dialects. These clean speech files were contaminated with additive impulsive and convolutive noise conditions.

The isolated words chosen were corrupted by F16, explosion, door and glass at -3, 0, 3, 6, 9 dB using the HTK platform (Hidden Markov Model Toolkit). The sampling frequency used is 16 khz. The signal to noise ratio SNR is defined as follows:

$$SNR = 10\log_{10}(\frac{P_{signal}}{P_{noise}}) \quad (14)$$

Where P_{signal} and P_{noise} represent respectively the power signal and the noise.

In this section, the comparison between two methods of analysis DFT and ST, is well defined. In addition, the difference in recognition rate is well observed using these two methods.

The S transform allows having a time-frequency representation of the signal. It combines only a dependent frequency resolution with a simultaneous location of the real and imaginary part of the spectrum. It was first published in 1996 by Stokwell and his team.

The basic idea of this time-frequency distribution is similar to the sliding-window Fourier transform, except that the amplitude and width of the analysis window are variable as a function of frequency, as is the case in wavelet analysis.

The X (f) spectrum of a signal x (t) by standard Fourier analysis is given by:

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-i2\pi ft}\,dt \quad (15)$$

If the signal x (t) is multiplied point by point by a window function g (t), the resulting spectrum is equal to:

$$X_g(f) = \int_{-\infty}^{\infty} x(t)g(t)e^{-i2\pi ft}\,df \quad (16)$$

The S-transform can be found by defining a normalized Gaussian window function:

$$g(t) = \frac{1}{\sigma\sqrt{2\pi}} exp^{\frac{-t^2}{2\sigma^2}} \quad (17)$$

Then allowing this Gaussian to translate and expand by τ et σ respectively:

$$S^*(\tau,f,\sigma) = \int_{-\infty}^{\infty} x(t)\frac{1}{\sigma\sqrt{2\pi}} exp^{\frac{-(t-\tau)^2}{2\sigma^2}} exp^{-i2\pi ft}dt \quad (18)$$

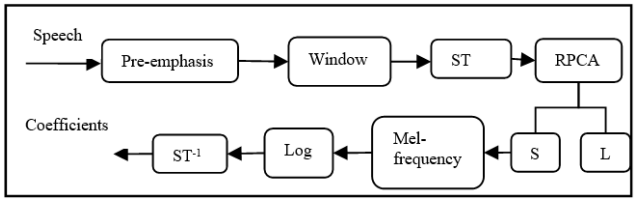


Fig. 3 MFCC-ST-RPCA

The following tables show the difference in the results of the two methods MFCC-DFT-RPCA and MFCC-ST-

RPCA by interpreting their performance at the recognition rate level at SNR= -3 dB and 9 dB.

Table 1: The recognition rate on TIMIT corpus by various types of noises at SNR= -3 dB using DFT

Noise type	MFCC	MFCC_E_D_A	MFCC-RPCA
Explosion	92.45	93.21	93.10
F16	73.91	76.13	77.71
Door	89.95	86.23	90.33
Glass	88.00	91.43	83.69

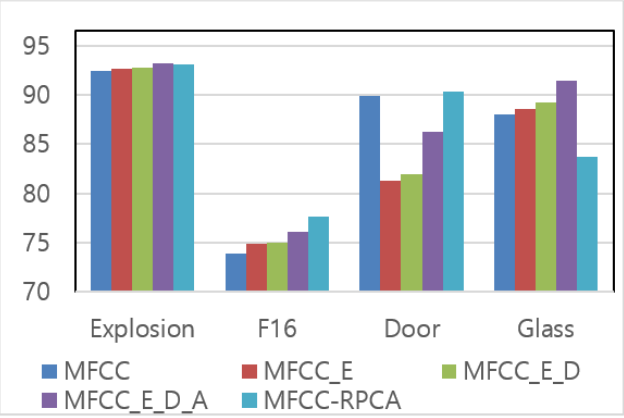
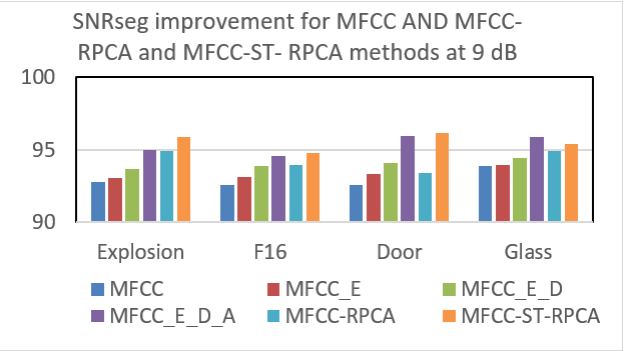


Table 2: The recognition rate on TIMIT corpus by various types of noises at SNR= 9 dB using ST

	MFCC	MFCC_E_D_A	MFCC-RPCA	MFCC-ST-RPCA
Explosion	92,77	95,02	94,92	95,9
F16	92,55	94,56	93,96	94,8
Door	92,55	95,96	93,42	96,2
Glass	93,88	95,87	94,92	95,4



4. Conclusion

In previous years, the research on Robust Principal Component Analysis (RPCA) has been attracting much attention. In this paper, a RPCA based speech enhancement approach has been presented. The advantage of this method is that it can directly estimate enhanced speech and do not need voice activity detector for noise estimation. Moreover, it can obtain high noise suppression performance in low SNR levels. Recently developed robust principal component analysis (RPCA) has been proven effective in separating the speech components from background noise. This approach decomposes the spectrogram matrix as the sum of a sparse matrix and a low-rank matrix representing speech and noise, respectively. Because the activation of the low-rank components can be temporally variable, this unsupervised decomposition can accommodate non stationary noise.
The Evaluation of this suggested recognition system, can be tested in the future, with others database such as Aurora database in the presence of other types of convolutive noise. It should also be tested on hybrid recognition systems such as HMM / SVM, HMM / RN.

References

[1] Tejus, R., et al. "Role of source separation using combined RPCA and block thresholding for effective speaker identification in multi source environment." Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2017 2nd IEEE International Conference on. IEEE, 2017.
[2] MARTÍNEZ, César E., GODDARD, J., DI PERSIA, Leandro E., et al. Denoising sound signals in a bioinspired non-negative spectro-temporal domain. Digital Signal Processing, 2015, vol. 38, p. 22-31.
[3] Gemmeke, Jort F., Tuomas Virtanen, and Antti Hurmalainen. "Exemplar-based sparse representations for noise robust automatic speech recognition." IEEE Transactions on Audio, Speech, and Language Processing 19.7 (2011): 2067-2080.
[4] Santosh, Kumar S., and S. H. Bharathi. "Non-negative matrix factorization algorithms for blind source separation in speech recognition." Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2017 2nd IEEE International Conference on. IEEE, 2017.
[5] Zheng, Guofei, et al. "Speech classification based on compressive sensing measurement sequence." Robotics and Biomimetics (ROBIO), 2017 IEEE International Conference on. IEEE, 2017.
[6] Gavrilescu, Mihai. "Noise robust automatic speech recognition system by integrating robust principal component analysis (RPCA) and exemplar-based sparse representation." Electronics, Computers and Artificial Intelligence (ECAI), 2015 7th International Conference on. IEEE, 2015.
[7] Sun, Chengli, et al. "Noise reduction based on robust principal component analysis." Journal of Computational Information Systems 10.10 (2014): 4403-4410.
[8] Tokuda, Keiichi, et al. "Speech synthesis based on hidden Markov models." Proceedings of the IEEE 101.5 (2013): 1234-1252.
[9] Huang, P.S., Chen, S.D., Smaragdis, P., et al.: Singing-voice separation from monaural recordings using robust principal component analysis. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, pp. 57–60 (2012)
[10] Wang, Minghe, Erhua Zhang, and Zhenmin Tang. "Robust Principal Component Analysis Based Speaker Verification Under Additive Noise Conditions." Chinese Conference on Pattern Recognition. Springer, Singapore, 2016.

[11] Gavrilescu, Mihai. "Noise robust automatic speech recognition system by integrating robust principal component analysis (RPCA) and exemplar-based sparse representation." Electronics, Computers and Artificial Intelligence (ECAI), 2015 7th International Conference on. IEEE, 2015.
[12] Zhao, Qian, et al. "Robust principal component analysis with complex noise." International conference on machine learning. 2014.
[13] Candès, Emmanuel J., et al. "Robust principal component analysis?." Journal of the ACM (JACM) 58.3 (2011): 11.
[14] Huang, Jianjun, et al. "Speech Denoising via Low-Rank and Sparse Matrix Decomposition." ETRI Journal 36.1 (2014): 167-170.
[15] Chauhan, Paresh M., and Nikita P. Desai. "Mel frequency cepstral coefficients (mfcc) based speaker identification in noisy environment using wiener filter." Green Computing Communication and Electrical Engineering (ICGCCEE), 2014 International Conference on. IEEE, 2014.
[16] Partridge, Matthew, and Marwan Jabri. "Robust principal component analysis." Neural Networks for Signal Processing X, 2000. Proceedings of the 2000 IEEE Signal Processing Society Workshop. Vol. 1. IEEE, 2000.
[17] Bouzid, Aïcha, and Noureddine Ellouze. "Speech enhancement based on wavelet packet of an improved principal component analysis." Computer Speech & Language 35 (2016): 58-72.
[18] Muda, Lindasalwa, Mumtaj Begam, and Irraivan Elamvazuthi. "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques." arXiv preprint arXiv:1003.4083 (2010).
[19] Sun, Pengfei, and Jun Qin. "Low-rank and sparsity analysis applied to speech enhancement via online estimated dictionary." IEEE Signal Processing Letters 23.12 (2016): 1862-1866.
[20] Winursito, Anggun, Risanuri Hidayat, and Agus Bejo. "Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition." 2018 International Conference on Information and Communications Technology (ICOIACT). IEEE, 2018.
[21] Benabderrahmane, Yasmina, Sid-Ahmed Selouani, and D. O'Shaughnessy. "Blind speech separation for convolutive mixtures using an oriented principal components analysis method." 2010 18th European Signal Processing Conference. IEEE, 2010.

Sonia Moussa received the diploma of Master Degree in automatic and signal processing (ATS) from the National Engineers School of Tunis (ENIT), in 2014. Currently, she prepares his doctorate thesis focused on the Analysis and Recognition of speech signal using auditory model.

Zied Hajaiej received the MS degree in electrical engineering (signal processing) in 2004, from National Engineering School of Tunis (ENIT). He is currently working towards the Ph.D. degree in electrical engineering (signal processing) in ENIT. His research involved speech recognition. Since September 2006, he has been an Assistant in the Physics Department at Faculty of Sciences of Bizerte, Tunisia, where he teaches electronics, VHDL.

Prof. Ali Gharsallah is a professor and a director of research laboratory of circuits and high frequency electronic systems at the Faculty of Sciences of Tunis-University of Tunis El Manar (FST).