**International Journal of Electrical and Computer Engineering (IJECE)**
Vol. 11, No. 2, April 2021, pp. 1259~1264
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i2.pp1259-1264 

# Speech emotion recognition based on SVM and KNN classifications fusion

**Mohammed Jawad Al Dujaili[1], Abbas Ebrahimi-Moghadam[2], Ahmed Fatlawi[3]**
[1]Department of Electronic and Communication, Faculty of Engineering, University of Kufa, Najaf, Iraq
[2]Electrical Engineering Department Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad, Iran
[3]Computer Engineering Department Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

**Article Info**

***Article history:***

Received Oct 2, 2019
Revised Aug 23, 2020
Accepted Oct 21, 2020

***Keywords:***

FEZ
FP
KNN
PCA
Speech emotion
SVM

**ABSTRACT**

Recognizing the sense of speech is one of the most active research topics in speech processing and in human-computer interaction programs. Despite a wide range of studies in this scope, there is still a long gap among the natural feelings of humans and the perception of the computer. In general, a sensory recognition system from speech can be divided into three main sections: attribute extraction, feature selection, and classification. In this paper, features of fundamental frequency (FEZ) (F0), energy (E), zero-crossing rate (ZCR), fourier parameter (FP), and various combinations of them are extracted from the data vector, Then, the principal component analysis (PCA) algorithm is used to reduce the number of features. To evaluate the system performance. The fusion of each emotional state will be performed later using support vector machine (SVM), K-nearest neighbor (KNN), In terms of comparison, similar experiments have been performed on the emotional speech of the German language, English language, and significant results were obtained by these comparisons.



***Corresponding Author:***

Mohammed Jawad Al Dujaili
Department of Electronic and Communication Engineering
Faculty of Engineering, University of Kufa
Najaf, Iraq
Email: Mohammed.challab@uokufa.edu.iq

## 1. INTRODUCTION

Detection of sensation from speech signal is one of the relatively new branches in speech processing, which can play an important role in human-robot interactions. Detecting the sense of speech is very useful for applications that Requires a natural interaction between humans and the machine, such as movies, web and computer training applications, where the response depends on the user's feelings. It can also be used as a useful medical device for the diagnosis of certain patients, such as autism, Parkinson's disease [1]. Speech recognition systems detect the emotional state of the speaker to analyze the sound characteristics. However, human voice has a combination of information including narrative traits and lexical, cultural, psychological, and emotional features. The presence of these communication dimensions cause the variables that affect the performance of the diagnostic system. Hence, the creation of emotional models requires their careful consideration to compensate the effects of these variables [2]. Human speech consists of two parts of the content and the tone. Human beings usually mean each other both by content and by using the tone. So, the same content that is expressed in two different tenses may have two different meanings and meanings. Investigations on the recognition of emotion from speech can be studied in terms of the features used and the classification algorithm. Many researchers in this area have focused their efforts on

choosing a strong classification algorithm. For example, in [3], a support vector machine (SVM) based retrieval method was used to extract the attribute and from the 2-class multi support vector machine (MSVM) and K-nearest neighboring. In research such as [4], the wavelet packet entropy and features such as Freund Frequency, Jitter, Schimer, Mel frequency cepstral coefficient, etc. have been used, and the data classification has been made using the SVM. Speech recognition systems can be divided into two essential parts of the feature extraction and classification. Figure 1 shows the block diagram of a speech recognition system. In the remainder of this paper, section 2 discusses the fusion of features, feature degradation with the main components (PCA) and the classifications used in the proposed method. Section 3 presents the proposed method. Section 4 discusses the results of the diagnostic tests by speech, and in the final section, the conclusion is drawn.

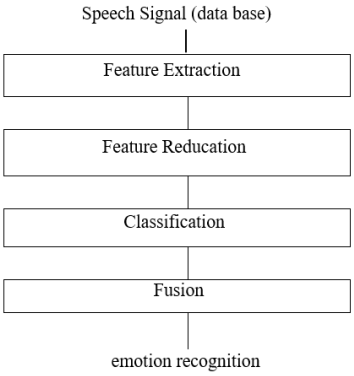


Figure 1. Block diagram general diagnosis of speech

## 2. STEPS TO DETECT FEELINGS BY SPEECH

As shown in Figure 1, the sensory recognition system consists of four parts. Initially, a speech signal enters the system. Then extraction and reduction functions are performed.

### 2.1. Database

We used three databases in German and English to evaluate the proposed method. The German Berlin database is publicly available on the Internet, and many researchers have used it in their work. This database contains 536 sentences (10 speakers, 7 feelings, 10 words including natural emotions, Boredom. disgust, Fear, happiness, sadness, Anger). The SAVEE database is in English and contains 480 sentences in 7 senses (natural emotions, disgust, Fear, happiness, Anger, sadness, surprise). In order to evaluate the proposed system in all experiments using the N-fold cross-validation was Used [2, 4].

### 2.2. Extraction and reduction of features

In this paper, tow feature groups have been investigated. The first group is based on fundamental frequency (F0), energy (E), zero-crossing rate (ZCR) (FEZ). The second group is based on the Fourier parameter (FP) model.

#### 2.2.1. FEZ continuous features

FEZ features are based on three fundamental functional groups (F0), energy (E), and zero-crossing rate (ZCR). These features are among the most commonly used features in recognizing sensation from speech, as part of the standard features of this field [5, 6]. The Speech Base Frequency (F0) is an important feature that provides the toneal and rhythmic properties of the speech. Energy strongly signals the speech signal, showing a pause and emphasis on speech, and is highly dependent on speech dialect. The zero crossing rate (ZCR) represents the moments that the adjacent samples of an acoustic signal are changing the sign. Research has shown that the set of F0, energy, and ZCR features are better than features such as Forment and Linear predictive cepstral coefficients (FPC) [7, 8]. In this study, FEZ features are extracted from the fundamental frequency (F0), energy (E), zero-crossing rate (ZCR) signals. For this purpose, the speech signal is first divided into frames of 20 milliseconds with an overlap of 10 milliseconds. Fundamental frequency (F0), energy (E), and zero-crossing rate (ZCR) are calculated for each frame. So, there are three curves for fundamental frequency (F0), energy (E), and zero-crossing rate (ZCR). Minimum, maximum, average, mean, and standard deviation of these curves will be used as FEZ attributes.

### 2.2.2. Fourier parameter (FP)

Fourier series analysis is one of the most important analytical methods in engineering. Fourier analysis has been widely used for signal processing, including filtering, correlation, encoding, integrating, and extracting features for pattern recognition. This Fourier analysis is a signal to decompose into its sinusoidal vibrations. From this perspective, a speech signal can be regarded as a result of an excitation signal (sonic audio output) passing through a time-varying linear time filter (phonological duct assembly) that models the resonance properties of the audio device. A speech signal (n) x divided into L frame can be represented by the combination of a FP model in the form (1) [9]:

$$X(n) = \sum_{K=1}^{M} \left( H_K^1(n) \left( \cos\left( 2\pi \frac{f_N^2}{Fs} n \right) \right) + \varphi_k^1 \right) \quad (1)$$

where: Fs is the sampling frequency of the speech signal (n) x, and the amplitude and the phase is the harmonic sinusoidal component k, l represents the frame, and M is the number of harmonic elements of speech. The harmonics of this model are a Fourier serial display of the alternating components of the speech signal. For each frame obtained from sound, we obtain the values of harmonic coefficients. The first 120 coefficients will be chosen in these coefficients. These 120 coefficients with first and second degrees differential values will form an elemental element array of 360. The maximum, minimum, average, mode, and standard deviation of an 1800 element array will be formed when placing those together [10].

### 2.3. Classification

In this work, we use a composite construct based on the two categories of backup vector machines (SVM), K-nearest neighbor (KNN).

### 2.3.1. Support vector machine (SVM)

Support vector machine (SVM) is one of the supervisory learning methods that it used for classification and regression [11]. The backup machine categorization is a linear data categorization that tries to select a line that has the highest margin of confidence. Although training at the fastest SVMs is very slow, this classification has a very high degree of accuracy. It has been widely used to recognize the feeling of speech [12, 13].

### 2.3.2. K-nearest neighbor's algorithm (KNN)

The K-nearest neighbor (KNN) algorithm is a classification technique that is based on comparative learning [14, 15]. For an experimental data, the algorithm looks for k samples from the nearest samples. In this method, it is decided which new class to fall into; which classifies the new instance belongs to a class that has the most votes in the K nearest neighbors [16].

### 2.3.3. Majority voting rule: MVR

In order to be able to use the MVR rule, the majority vote must always be correct [17]. This rule states that the input of x belongs to the class i if and only if the existing N classifier exists, the majority of them have chosen i class [18, 19]. To implement this method, each level word is firstly determined, then it is announced among the four existing levels of the winning surface, which has a majority of 5 words selected.

## 3. COMPONENT ANALYSIS ALGORIHM

The PCA technique is a suitable method for reducing data dimensions linearly. By eliminating the minuscule coefficients obtained from this conversion, the lost data are less than other methods. In this method, the axes of the new coordinates for the data are defined and the data is expressed based on these axes of the new coordinates. The first axis should be in the direction where the data variance is maximized (that is, in the direction in which the data is scattered). The second axis must be perpendicular to the first axis, so that the variance of the data is maximized. Similarly, the lateral axis is perpendicular to all of the previous axes so that the data is the most dispersed in that direction [20, 21].

## 4. THE PROPOSED METHOD

As mentioned, extracting the proper feature of speech, and an efficient and optimal classification system consisting of the combination of several classification algorithms play a central role in the function of a sensory recognition system from speech. In the method proposed in this paper, after extracting the properties (FEZ), (FP) the principal component analysis (PCA) algorithm is used as a feature reduction

feature, and then from the combination of support vector machine (SVM) and K-closest neighbor (KNN) have been used as a classification. Appropriate methods are used to increase the reliability of the emotion detection system using serial fusion of classifiers by majority voting law. Today, the multiple classifier systems (MCS) classifier is used to replace the very complex classes that require a lot of educational calculations [22]. There are three methods to combine classifications: combination, serial and parallel. In a hybrid method, in this method, the classification is regulated in a tree structure that is a collection of different classes [23]. In parallel, in this method, each classification is applied independently and the decision algorithm is applied to the outputs. In a serial method, the number of different classes for other classes is reduced. The proposed method was compared to conventional methods for performance accuracy and computational complexity [24, 25].

## 5. EVALUATION OF RESULTS

In this paper, we have extracted the characteristics of FEZ and FP from the German database sentences [2]. Berlin's German database is publicly available through the Internet and many researchers have used it in their work. This database contains 536 sentences (10 speakers, 7 feelings of 10 speeches including natural emotions, Boredom. disgust, Fear, happiness, sadness, Anger). The SAVEE database is in English and contains 480 sentences in seven feelings (natural emotions, disgust, Fear, happiness, Anger, sadness, surprise). In order to evaluate the proposed system, 10-fold-cross-validation technique was used in all experiments. Initially, we considered the fusion of the features together and then applied to the SVM and KNN classifications, and the results of the detection rate for each database according to the classifier, in order to better compare the database of German, English with class The SVM, KNN clauses show the average detection rate using the fusion of all features after applying the PCA to identify the senses in Tables 1 and 2. The property of the PCA is that it does not delete any important feature, as we can see, the best results for the German database in the SVM classification using FEZ feature in Table 3 for seven senses with an accuracy of 85.1% and an execution time of 0.05 seconds. And the best results for the German database and KNN classification are the fusion FP+FEZ features properties with a precision detection of 87.85% and a runtime execution time of 0.48 seconds, as shown in Table 4. Then, we performed similar experiments on the German language database and the classification of SVM, KNN in the English database for the classification of SVM and KNN, and the results are presented in Tables 5 and 6 as seen in these tables. The highest detection rate for the English database is from the fusion of FEZ properties in SVM classification with a resolution of 85.2% and a runtime execution time of 0.05 seconds, and the highest detection rate for the English language database in KNN classification by fusion of all features with a detection accuracy of 90.83% And the execution time of the algorithm is 0.35 seconds.

Table 1. Comparison of the rate of diagnosis of emotions in the German database with all the classifications

| German+KNN | German+SVM | | data base, classifiers | |
|---|---|---|---|---|
| Time Sec | Accu. % | Time Sec | Accu %. | features |
| FP | 49.2 | 0.7 | 87.66 | 0.45 |
| FEZ | 85.1 | 0.05 | 87.11 | 0.01 |
| FP+FEZ | 84 | 1.06 | 87.85 | 0.48 |

Table 2. Comparison of the rate of emotion detection in the English database with all the classifications

| English+KNN | English+SVM | | data base, classifiers | |
|---|---|---|---|---|
| Time Sec | Accu. % | Time Sec | Accu%. | features |
| FP | 53.1 | 0.65 | 91.04 | 0.37 |
| FEZ | 85.2 | 0.05 | 87.29 | 0.01 |
| FP+FEZ | 85 | 1.06 | 90.83 | 0.35 |

Table 3. Interaction matrix the best result of the German database with classification SVM

| FEZ Accurac =85.1 %, Time=0.05 Sec | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Neut. | Bored | Disg. | Fea | hap | Sad. | Ang. |
| Neut. | 97.6 | 0 | 0 | 0 | 0.7 | 0 | 1.6 |
| Bored. | 8.64 | 90.1 | 1.2 | 0 | 0 | 0 | 0 |
| Disg. | 8.69 | 0 | 89 | 0 | 2.1 | 0 | 0 |
| Fea. | 21.7 | 1.44 | 0 | 71 | 0 | 0 | 5.7 |
| Hpp. | 19.7 | 0 | 0 | 0 | 80 | 0 | 0 |
| Sad. | 9.67 | 1.61 | 0 | 1.6 | 0 | 85.4 | 1.6 |
| Ang. | 12.6 | 0 | 0 | 1.2 | 1.2 | 1.26 | 83 |

Neutral=Neut, Boredom=Bored, Disg=disgust, Fea=Fear, hap=happiness, sad=sadness, ang=Anger

Table 4. Interaction matrix the best result of the German database with the classification of KNN

| FEZ Accuracy=85.2% , Time=0.05Sec | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Neut. | Disg. | Fea | Hap | Ang. | Sad. | Surp. |
| Neut. | 98 | 0 | 0 | 0 | 1.9 | 0 | 0 |
| Disg. | 0 | 82 | 0 | 0 | 19 | 0 | 0 |
| Fear | 0 | 0 | 77 | 0 | 25 | 0 | 0 |
| Hap | 0 | 0 | 0 | 85 | 15 | 0 | 0 |
| Ang. | 17 | 0 | 0 | 0 | 84 | 0 | 0 |
| Sad. | 0 | 0 | 0 | 0 | 20 | 80 | 0 |
| Surp. | 0 | 0 | 0 | 0 | 22 | 0 | 78.4 |

Table 5. Interaction matrix the best result of the English database with the SVM classification

| FP+FEZ Accuracy = 87.85 % , Time = 0.48Sec | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Neut. | Bored | Disg. | Fea | hap | Sad. | Ang. |
| Neut. | 92.2 | 0 | 0.9 | 3 | 3.4 | 0 | 0.8 |
| Bored. | 0 | 88.9 | 1.2 | 3 | 0 | 1.3 | 6.2 |
| Disg. | 0 | 0 | 95 | 0 | 2.2 | 2.8 | 0 |
| Fea. | 0 | 0 | 7.3 | 77 | 7.3 | 2.9 | 5.6 |
| Hpp. | 9.9 | 0 | 1.5 | 2 | 85 | 0 | 2.9 |
| Sad. | 0 | 4.83 | 0 | 1.7 | 0 | 92 | 1.7 |
| Ang. | 0 | 5.06 | 0 | 1.3 | 0 | 1.3 | 93 |

Neutral=Neu, disgust=disg, Fear=Fea, happiness =hap, Anger=Ang, sad=sadness, surprise=surp

Table 6 . Interaction matrix the best result of the English database with the KNN classification

| FP+FEZ Accuracy=90.83 %, Time=0.35Sec | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Neut. | Disg. | Fea | Hap | Ang. | Sad. | Surp. |
| Neut. | 97 | 0.9 | 0.8 | 0 | 0 | 0.8 | 0 |
| Disg. | 0 | 88 | 3.4 | 0 | 5 | 3.5 | 0 |
| Fear | 0 | 3.3 | 89 | 0 | 0 | 1.6 | 6.66 |
| Hap | 3.4 | 1.7 | 1.7 | 93 | 0 | 0 | 1.7 |
| Ang. | 5 | 0 | 3.4 | 2 | 89 | 0 | 1.7 |
| Sad. | 0 | 3.4 | 1.7 | 0 | 3.3 | 92 | 0 |
| Surp. | 0 | 1.7 | 5 | 0 | 1.6 | 0 | 90.7 |

## 6. CONCLUSION

To solve the problem of detecting the direct feelings of the speaker, a sensation model is presented to classify seven senses. In this research, two feature groups have been investigated. The first group is based on fundamental frequency (F0), energy (E), zero-crossing rate (ZCR) (FEZ). The second group is based on the Fourier parameter (FP) model. These features are also considered as the input parameter for support vector machine, the closest KNN neighbor. In order to evaluate the proposed system, the principal component analysis (PCA) is proposed to reduce the feature and SVM, KNN closest neighbor. For this purpose, comparative tests, including the fusion of the characteristics together, and then the fusion of the characteristics is taken using PCA. Experimental results have proven good results in fusion of features and the best results with the German database on fusion of all features and PCA with KNN classification with a resolution of 87.85% and timing of execution of 0.48sec. The best results with the English database on fusion of all features and PCA with KNN classification with a resolution of 90.83% and run time of 0.35 seconds.

## REFERENCES


[1] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Transactions on Bio-medical Engineering*, vol. 47, vol. 7, pp. 829-837, 2007.

[2] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera Am Mittag German audio-visual emotional speech database," *2008 IEEE International Conference on Multimedia and Expo*, Hannover, 2008, pp. 865-868.

[3] Kerkeni, Leila, et al., "Automatic Speech Emotion Recognition Using Machine Learning," *Social Media and Machine Learning. IntechOpen*, pp. 1-16, 2019.

[4] Elyaderani, M. K., Mahmoodian, S. H., and Sheikhi, G., "Wavelet Packet Entropy in Speaker-Independent Emotional State Detection from Speech Signal," *Journal of Intelligent Procedures in Electrical Technology*, vol. 5, no. 20, pp. 67-74, 2015.

[5] Fu, Liqin, Xia Mao, and Lijiang Chen, "Speaker independent emotion recognition based on SVM/HMMs fusion system," *2008 International Conference on Audio, Language and Image Processing*, Shanghai, 2008, pp. 61-65.

[6] Lin, Y. L., and Wei, G., "Speech emotion recognition based on HMM and SVM," *2005 International Conference on Machine Learning and Cybernetics, Guangzhou*, China, vol. 8, 2005, pp. 4898-4901.

[7] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, S. Kollias, W. Fellenz, J., Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32-80, 2001.

[8] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011

[9] J.-H. Yeh, T.-L. Pao, C.-Y. Lin, Y.-W. Tsai, and Y.-T. Chen, "Segment-based emotion recognition from continuous Mandarin Chinese speech," *Computers in Human Behavior*, vol. 27, no. 5, pp. 1545-1552, 2011.

[10] J. S. Walker, "Fourier series," in *Encyclopedia of Physical Science and Technology*. New York, NY, USA: Academic, 2001.

[11] R. McAulay and T. Quatieri, "Speech analysis/Synthesis based on a sinusoidal representation," *in IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744-754, 1986.

[12] Altun, Halis, and Gökhan Polat, "Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection," *Expert Systems with Applications*, vol. 36, no. 4, pp. 8197-8203, 2009.

[13] B. Schuller, G. Rigoll, M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Que, vol. 1, 2004, pp. 577-580.

[14] Lee, Chul Min, and Shrikanth S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293-303, 2005.

[15] Lanjewar, Rahul B., Swarup Mathurkar, and Nilesh Patel, "Implementation and comparison of speech emotion recognition system using gaussian mixture model (gmm) and k-nearest neighbor (k-nn) techniques," *Procedia computer science*, vol. 49, pp. 50-57, 2015.

[16] Zhou, Jian, et al., "Speech emotion recognition based on rough set and SVM," *2006 5th IEEE International Conference on Cognitive Informatics*, Beijing, 2006, pp. 53-61.

[17] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787-800, 2007.
[18] Sheikhan, Mansour, Mahdi Bejani, and Davood Gharavian, "Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method," *Neural Computing and Applications*, vol. 23, no. 1, pp. 215-227, 2013.
[19] Giannakopoulos, Georgios, et al., "Ensemble majority voting classifier for speech emotion recognition and prediction," *Journal of Systems and Information Technology*, vol. 16, no. 3, pp. 222-232, 2014.
[20] Jolliffe, I. T., "Mathematical and statistical properties of population principal components," *Principal Component Analysis*, pp. 8-22, 2002.
[21] Shashidhar G. Koolagudi, K and Sreenivasa R., " Emotion recognition from speech: a review," *International Journal of Speech Technology*, vol. 15, pp. 99-117, 2012.
[22] P. Laukka, D. Neiberg, M. Forsell, I. Karlsson and K. Elenius "Expression of affect in spontaneous speech: Acoustic correlatesand automatic detection of irritation and resignation," *Computer Speech and Language*, vol. 25, no. 1, pp. 84-104, 2011.
[23] Oli F., "Fusion of multiple pattern classifiers," *Journal Article*, 2003
[24] Busso, Carlos, Sungbok Lee, and Shrikanth Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 582-596, 2009.
[25] Oster, A. M., and Arne Risberg, "The identification of the mood of a speaker by hearing impaired listeners," *SLT-Quarterly Progress Status Report*, vol. 27, no. 4, pp. 79-90, 1986.

## BIOGRAPHIES OF AUTHORS

**Mohammed Jawad Al-Dujaili Al-Khazraji** awarded B.S. degree in communication engineering from University of Al-Furat Al-Awsat Technical, Technical College of Engineering, Najaf, Iraq in 2008 and M.S. degree in communication system engineering from Ferdowsi university, Iran, in 2018. Currently, he is a member staff at the Department of Electronic and Communication, Faculty of Engineering, University of Kufa, Iraq. His research interest includes the development of Wireless communications and signal processing as well as image, speech processing and radar, 5G. Email : Mohammed.challab@uokufa.edu.iq

**Abbas Ebrahimi-Moghadam** received the B.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 1991, the M.Sc. degree in electrical engineering from the K. N. Toosi University of Technology, Tehran, in 1995, and the Ph.D. degree in electrical and computer engineering form McMaster University, Hamilton, ON, Canada. He has been an Assistant Professor with the Electrical Engineering Department, Ferdowsi University of Mashhad, Mashhad, Iran, since 2011. Email : a.ebrahimi.m@gmail.com

**Ahmed Fatlawi awarded** B.S. degree in technologies engineering of computer from Alkafeel university, Najaf , Iraq in 2015 and M.S. degree in computer engineering/artificial intelligence and robotics from Ferdowsi University of Mashhad, Mashhad, Iran, in 2018. Currently, he is a student staff Ph.D. Email:ah.fatlawi@mail.um.ac.ir