

Audio-visual emotion recognition using FCBF feature selection method and particle swarm optimization for fuzzy ARTMAP neural networks

Davood Gharavian^{1,2} · Mehdi Bejani³ ·
Mansour Sheikhan¹

Received: 18 January 2015 / Revised: 28 October 2015 / Accepted: 18 December 2015 /
Published online: 14 January 2016
© Springer Science+Business Media New York 2016

Abstract Humans use many modalities such as face, speech and body gesture to express their feeling. So, to make emotional computers and make the human-computer interaction (HCI) more naturally and friendly, computers should be able to understand human feelings using speech and visual information. In this paper, we recognize the emotions from audio and visual information using fuzzy ARTMAP neural network (FAMNN). Audio and visual systems fuse at decision and feature levels. Finally, the particle swarm optimization (PSO) is employed to determine the optimum values of the choice parameter (α), the vigilance parameters (ρ), and the learning rate (β) of the FAMNN. Experimental results showed that the feature-level and decision-level fusions improve the outcome of unimodal systems. Also PSO improved the recognition rate. By using the PSO-optimized FAMNN at feature level fusion, the recognition rate was improved by about 57 % with respect to the audio system and by about 4.5 % with respect to the visual system. The final emotion recognition rate on the SAVEE database was reached to 98.25 % using audio and visual features by using optimized FAMNN.

Keywords Audio-visual emotion recognition · Particle swarm optimization · Fuzzy ARTMAP neural network

✉ Davood Gharavian
dgharavian@gmail.com

Mehdi Bejani
St_m_bejani@azad.ac.ir

Mansour Sheikhan
msheikhn@azad.ac.ir

¹ Department of Electrical Engineering, Islamic Azad University, South Tehran Branch, Tehran, Iran

² Department of Electrical Engineering, Shahid Beheshti University, Tehran, Iran

³ Islamic Azad University, South Tehran Branch, Tehran, Iran

1 Introduction

Humans communicate with each other far more naturally than they do with computers. One of the main problems in human–computer interaction (HCI) systems is the transmission of implicit information. To make HCI more naturally and friendly, computers must enjoy the ability to understand human’s emotional states the same way as human does.

In the recent years, emotion recognition has found many applications such as medical-emergency domain to detect stress and pain [15], interactions with robots [27, 41], computer games [26], and developing man–machine interfaces for helping weak and old people [36].

There are many modals such as face, body gesture and speech that people use to express their feelings. Combination of these modals depends on the place they occur and on the subjects themselves; therefore, there are a wide variety of patterns for combining [30].

Some studies in psychology and linguistics confirm the relation between affective displays and specific audio and visual signals [2, 17].

Mehrabian [33] states that there are basically three elements in any face-to-face communication. Facial expression and speech articulation in the visual channel are the most important affective cues (55 % and 38 %, respectively), and words contribute only 7 % of the overall impression.

There are some approaches for quantifying and measuring emotions such as discrete categories and dimensional description [40]. In this work, we used basic discrete emotion categories including happiness, fear, sadness, anger, surprise, neutral and disgust that are rooted in the language of daily life. This description was specially supported by the cross-cultural studies conducted by Ekman [16]. Most of the existing studies of automatic emotion recognition focus on recognizing these basic emotions. These seven emotional states are common and have been used in the majority of previous works [5, 7, 14, 21, 30, 31, 37, 38, 46]. Our method is general and can be extended to more emotional states. Using universal emotion models, it is easy to recognize emotional states [49].

Two main fusion approaches used in the literature are feature-level fusion and decision-level fusion. The goal of this paper is to simulate human perception of emotions by combining emotion-related information from facial expression and Audio. So, we used different approaches to fuse audio and facial expression information. The classifier type also affects emotion recognition rate significantly. Usually different classifiers such as artificial neural networks (ANNs), support vector machines (SVMs), decision trees, K-nearest neighbor (KNN), Gaussian mixture models (GMMs), hidden Markov models (HMMs), and Bayesian networks have been used for emotion recognition. Also researchers have proposed the hybrid and multi-classifier methods [49]. Here, we used fuzzy adaptive resonance theory mapping (ARTMAP) neural network [9] as our proposed classifier, and particle swarm optimization (PSO) was employed to determine the optimum values of the choice parameter (α), the vigilance parameters (ρ) and the learning rate (β) of the fuzzy ARTMAP neural network (FAMNN).

The remainder of this paper is organized as follows: Section 2 reviews the recent researches in this field. Section 3 presents our methodology for this problem. In this section, we first discuss about the SAVEE database that has been used in this work, and then on how audio and video features were extracted, as well as feature reduction and feature selection procedures. Also the FAMNN is introduced as a classifier, and finally, the particle warm optimization method is presented for optimizing FAMNN and improving the classification accuracy task. Section 4 contains the experimental results. In

Section 5, the influence of PSO-optimized FAMNN on the performance of emotion recognition is reported. Finally, conclusions are drawn in Section 6.

2 Background and related works

Recently, audio-visual based emotion recognition methods have attracted the attention of the research community. In the survey of Pantic and Rothkrantz [39], only four studies were found to focus on audio-visual affect recognition. Since then, affect recognition, using audio and visual information, has been the subject of many researches. The most updated survey on affect recognition methods for audio, visual and spontaneous expressions belongs to Zeng et al. [49]. Here, some main works in this field are mentioned in brief.

De Silva and Pei Chi [14] used a rule based method for decision level fusion of speech and visual based systems. In speech, pitch was extracted as the feature and used in the nearest-neighbor classification method. In video, they tracked facial points with optical flow, and hidden Markov model (HMM) was trained as the classifier. The decision level fusion improved the result of the individual systems.

Song et al. [46] used a tripled hidden Markov model (THMM) to model joint dynamics of the three signals perceived from the subject: a) pitch and energy as speech features, b) motion of eyebrow, eyelid, and cheek as facial expression features, and c) lips and jaw as visual speech signals. The proposed THMM architecture was tested for seven basic emotions (surprise, anger, joy, sadness, disgust, fear, and neutral), and its overall performance was 85 %.

Mansoorizadeh and Moghaddam Charkari [30] compared feature level and decision level fusions of speech and face information. They proposed an asynchronous feature-level fusion approach that improved the result of combination. For speech analysis, they used the features related to energy and pitch contour. For face analysis, the features representing the geometric characteristic of face area were used. The multimodal results showed an improvement over the individual systems.

Hoch et al. [24] developed an algorithm for bimodal emotion recognition. They used a weighted linear combination for the decision level fusion of speech and facial expression systems. They also applied a database of 840 audio-visual samples with 7 speakers and 3 emotions. Their system classified 3 emotions (positive, negative and neutral) with an average of 90.7 % recognition rate. By using a fusion model based on a weighted linear combination, the performance improvement became nearly 4 % compared to that of unimodal emotion recognition.

Paleari [38] presented a semantic affect-enhanced multimedia indexing (SAMMI) to extract real-time emotion appraisals from non-prototypical person independent facial expressions, and vocal prosody. Different probabilistic methods for fusion were compared and evaluated with a novel fusion technique called NNET. The results showed that NNET can improve the recognition score by about 19 % and the mean average precision by about 30 % with respect to the best unimodal system.

Haq and Jackson [21] used feature and decision level fusion for audio and visual features on the SAVEE database. 106 utterance-level audio features (fundamental frequency, energy, duration and spectral) and 240 visual features (marker locations on the face) were used for this system. The Gaussian classifier was employed to fuse the information in different levels. They used principal component analysis (PCA) and linear discriminate analysis (LDA) feature selection algorithms. Using PCA and LDA, 92.9 % and 97.5 % emotion classification rates for

audio-visual features, 50 % and 56 % for audio features and 91 % and 95.4 % for visual features were reported.

Bejani et al. [5] investigated a multi-classifier audio-visual system that combined the speech features (MFCC, pitch, energy and formants) and facial features (based on ITMI and QIM) on the eNterface'05 database. By using the multi-classifier system, the recognition rate was increased up to 22.7 % over the speech based system and up to 38 % over the facial expression based system.

In recent years, emotion recognition has had many applications in more generic mediated communications. Lopez-de-Ipina et al. [28] identified novel technologies and biomarkers or features for the early detection of Alzheimer's disease (AD) and its degree of severity. It concerns the Automatic Analysis of Emotional Response (AAER) in spontaneous speech based on Emotional Temperature and fractal dimension to validate tests and biomarkers for future diagnostic use. The AAER shows very promising results for the definition of features useful in the early diagnosis of AD. Harley et al. [22] presented a novel approach for measuring and synchronizing emotion data from three modalities (automatic facial expression recognition, self-report, electro dermal activity) and their consistency regarding learners' emotions. They found a high level of coherence between the facial recognition and self-report data (75.6 %), but low levels of consistency between them and electro dermal activation, suggesting that a tightly coupled relationship does not always exist between emotional response components. Weisgerber et al. [47] tested facial, vocal and musical emotion recognition capacities in schizophrenic patients. Dai et al. [13] proposed a computational method for emotion recognition on vocal social media to estimate complex emotion as well as its dynamic changes in a three-dimensional PAD (Position–Arousal–Dominance) space. They analyzed the propagation characteristics of emotions on the vocal social media site WeChat.

In recent years, the researchers are focused on finding reliable informative features and combining powerful classifiers in order to improve the performance of emotion recognition rate in real-life applications [37, 44]. In this way, developing optimal design methods for classification is an active research field. Here, we propose a PSO-optimized FAMNN that improves the emotion recognition results as compared with the audio, visual and audio-visual systems.

It is clear that emotional states influence audio and visual features of a person. In other words, audio and visual features maintain information about emotional states that synergistically influence the recognition process. Usage of data fusion method for audio and visual information and sequential process such as feature reduction, feature selection, classification and classifier optimization make a well-designed approach in this research. In this work, we examine different fusion approaches for audio-visual emotion recognition system, report the results, and finally, propose the most appropriate fusion method for such systems. To reduce computation cost and use the most effective features, feature reduction and feature selection algorithms were used for the audio and visual features.

3 Methodology

Various audio and visual information fusions were made by different setups of feature reduction and selection methods and classifiers in emotion recognition systems. In this setup, the audio features (Mel-frequency cepstral coefficient (MFCC), pitch, energy and formants) and the visual features (marker locations on the face) were extracted and the

features were reduced by the PCA feature reduction algorithm. Next, the FCBF feature selection method was applied to the reduced features. Then the FAMNN was used for various setups of the audio-visual emotion recognition systems. Finally, the PSO was employed to optimize the FAMNN to improve the experimental results.

The main goal of the present work is to quantify the performance of audio and visual systems, recognize the strengths and weaknesses of these systems setups, and compare the obtained setups to combine these two modalities for increasing the performance of the system.

To combine the visual and audio information, two different approaches were implemented: feature level fusion, in which a single classifier with features of both modalities is used, and decision level fusion, which uses a separate classifier for each modality, and the outputs are combined using stacked generalization method where the output of the ensemble serves as a feature vector to a meta-classifier. We used FAMNN as a meta-classifier to improve the generalization performance. Figure 1 shows an overview of the proposed recognition system.

FAMNN 1 shows the result of audio emotion recognition, and FAMNN 2 classifies the visual features after feature reduction and selection stages. The audio and visual features are also mixed together and pass the PCA and FCBF stages; then the selected features are fused to FAMNN 3.

The PCA-reduced audio and visual features were mixed together and then FCBF feature selection was applied to the mixed audio-visual reduced features. The selected features were used in FAMNN 4 for emotion recognition stage. FAMNN 5 used the selected separate audio and visual features and classified the emotion states accordingly.

The output of FAMNN 1 and FAMNN 2 serves as a feature vector to FAMNN 6. This experiment is a decision level fusion of audio and visual systems. In the following, the details are described.

3.1 Database

We used the Surrey Audio-Visual Expressed Emotion (SAVEE) database (<http://personal.ee.surrey.ac.uk/Personal/P.Jackson/SAVEE/Database.html>) that was recorded from four native

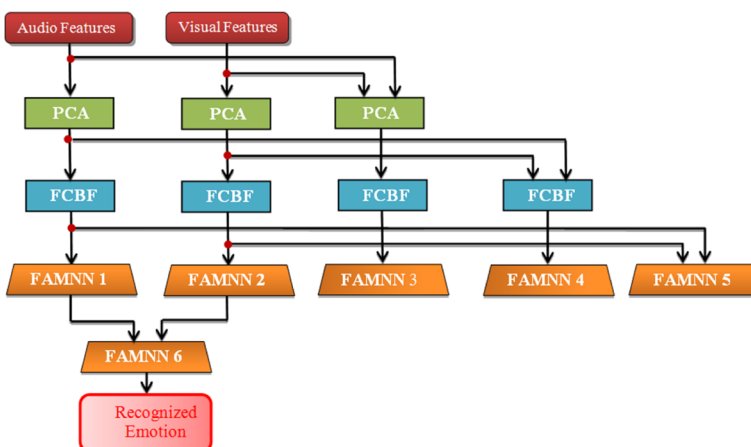


Fig. 1 Overview of the emotion recognition system

male English speakers (aged 27–31 years) with 60 markers painted on their face in CVSSP's 3D vision laboratory at Surrey University, UK. Figure 2 presents some examples of facial markers placed on four subjects with various emotions.

The sentences were recorded in seven emotional states: anger, disgust, fear, happiness, neutral, sadness and surprise. The recordings consisted of 15 phonetically-balanced TIMIT sentences per emotion: 3 common, 2 emotion specific and 10 generic sentences that were different for each emotion. The 3 common and 2 emotion specific sentences were recorded in neutral emotion, which resulted in 30 sentences for neutral emotion and 480 utterances in the database.

Emotion and sentence prompts were displayed on a monitor in front of actor during the recordings. The 3DMD dynamic face capture system [1] provided color video and Beyer dynamics microphone signals over several months during different periods of the year. The sampling rate was 44.1 kHz for audio and 60 fps for video. The 2D video of frontal face of the actor was recorded with one color camera.

3.2 Feature extraction

3.2.1 Audio features

Most of the existing approaches to audio emotion recognition used acoustic features as classification input. The popular features are prosodic features (e.g., pitch-related feature and energy-related features) and spectral features (e.g., MFCC and cepstral features). So, pitch, intensity, MFCC and formant features at the frame-level were used in this work for audio emotion recognition. Due to their popularity, descriptive power and bibliographical suggestions in related works [49], these features are used. 5-ms frames of the speech signal are analyzed every 10-ms using a Hanning window function in Praat speech processing software [6]. Because of large number of features at the frame-level, the statistical value of features over a specified sentence was used for training and testing of this system. Therefore, the mean, standard deviations, maximum and minimum values of the pitch, and energy were computed using Praat.

In addition MFCC was computed using Praat. MFCC's are a popular and powerful analytical tool in the field of speech recognition. In this work, we took the first 12 coefficients as the useful features. The mean, standard deviation, maximum, and minimum values of MFCC features were calculated, which produced a total number of 48 MFCC features.



Fig. 2 Facial markers placed on the subjects of SAVEE database with different emotions (from left): KL (anger), JK (happiness), JE (sadness) and DC (neutral)

Formant frequencies are the properties of the vocal tract system. In this paper, the first three formant frequencies and their bandwidths were calculated using Praat. The mean, standard deviation, maximum, and minimum values of formant features were calculated, which produced a total number of 24 formant features. In total, we extracted 80 features from speech signal and used them for emotion recognition.

3.2.2 Visual features

Visual features were created by painting 60 frontal markers on the face of the actor. The markers were painted on forehead, eyebrows, low eyes, cheeks, lips and jaw. After data capture, the markers were manually labeled for the first frame of a sequence and tracked for the remaining frames using a marker tracker. The tracked marker's x and y coordinates were normalized. Each marker's mean displacement from the bridge of the nose was subtracted. Finally, 480 visual features were obtained from the 2D marker coordinates, which consisted of mean, standard deviation, maximum, and minimum values of the adjusted marker coordinates.

In some previous works [7, 20, 21, 27], the facial markers on the face were used for facial expression recognition. Duo to excellent performance of these features in facial expression recognition and focus on other tasks (classification, optimization and fusion), we used them. To detect and extract facial points automatically in real world application, some techniques e.g., active appearance models (AAM) have been used [12]. And some software such as Luxand FaceSDK [29] provides the coordinates of facial feature points. It allows tracking and recognizing faces in live video.

3.3 Feature reduction

For dimension reduction and construction of a lower-size feature space, a statistical method was used to maximize the relevant information preserved. This can be done by applying a linear transformation, $y = Tx$, where y is a feature vector in the reduced feature space, x is the original feature vector, and T is the transformation matrix. PCA [45] is widely used to extract essential characteristics from high dimensional data sets and discard noise. PCA involves feature centering and whitening, covariance computation, and Eigen decomposition. We applied PCA as linear transformation technique for feature reduction.

3.4 Feature selection

In this study, the fast correlation-based filter (FCBF) [18] method was used for feature selection. This method selects the features, which are individually informative and two-by-two weakly dependent. It was noted that the Mutual Information (MI) of two vectors X and Y , $I(X, Y)$, computes their statistical dependency in the following way:

$$I(\mathbf{X}, \mathbf{Y}) = \sum_{y \in \mathbf{Y}} \sum_{x \in \mathbf{X}} p(\mathbf{X} = x, \mathbf{Y} = y) \log \left(\frac{p(\mathbf{X} = x, \mathbf{Y} = y)}{p(\mathbf{X} = x)p(\mathbf{Y} = y)} \right) \quad (1)$$

Where, p is the probability function. Obviously, $I(X, Y)$ is equal to 0, when X and Y are independent, ($p(X=x, Y=y) = p(X=x)p(Y=y)$), and increases when their dependency increases.

In FCBF method, Y is the vector of data labels, and X_i is the vector of i th feature value for all data. That is, when the number of features is N , there are $N + 1$ vectors. The FCBF selects features in two steps:

1. Removing the features (X_i), which are not dependent on the label vector Y :
 $I(X_i, Y) > \varepsilon$; where, ε is a positive threshold between 0 and 1. In this way, the FCBF selects the features that are individually informative. In this work, ε was set to 0.01.
2. Removing the remaining feature (X_i), which its dependency on the other remaining feature (X_j) is more than $I(X_i, Y)$, while $I(X_i, Y) > I(X_j, Y)$.
 In this way, the FCBF selects those individually informative features that are also two-by-two weakly dependent.

3.5 Classification

In the present study, the FAMNN was used as the emotion classifier. The theoretical foundation of Adaptive Resonance Theory (ART) has been introduced by Carpenter et al. [9]. The network has structural design for incremental supervised learning of recognition categories and multidimensional maps in reply to the random order of binary or analog input vectors. It gains a synthesis of fuzzy logic and ART neural networks by taking advantage of a close formal resemblance between the calculation of fuzzy method and the ART category choice, resonance, and learning.

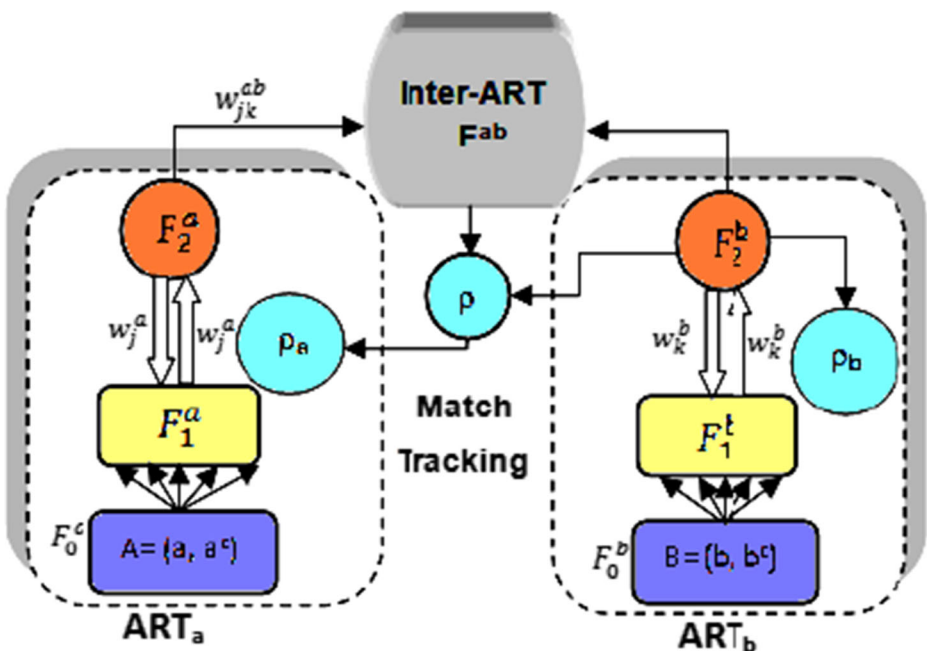


Fig. 3 Structure of the FAMNN

The FAMNN has been successfully used in many tasks, for e.g., remote sensing, data mining, and pattern recognition. The FAMNN is believed to be fast among the members of the ARTMAP family due to cheap mapping between the inputs and outputs.

The FAMNN networks has two fuzzy ART networks, ART_a and ART_b, interconnected by an inter-ART via an associative memory module (Fig. 3). The inter-ART module consists of a match tracking, and a self-regulatory mechanism whose purpose is to minimize the network error and maximize the generalization.

The input patterns of ART_a and ART_b are represented by the vector $a = [a_1, \dots, a_{M_a}]$ and $b = [b_1, \dots, b_{M_b}]$.

For ART_a, x^a denotes the F_1^a output vector, y^a denotes the F_2^a output vector, and w_j^a denotes the j th ART_a weight vector. Also for ART_b, x^b denotes the F_1^b output vector, y^b denotes the F_2^b output vector, and w_k^b denotes the k th ART_b weight vector. For the map field, x^{ab} denotes the F^{ab} output vector, and w_j^{ab} denotes the weight vector from the j th F_2^a note to F^{ab} .

After the resonance is confirmed in the networks, J gets the active category for the ART_a network, and K is the active category for the ART_b network. The next step is match-tracking to verify if the active category on ART_a corresponds to the desired output vector presented to ART_b. The vigilance criterion is given by [8]:

$$\rho_{ab} = \frac{|y^b \wedge W_{jk}^{ab}|}{y^b} \tag{2}$$

Once the resonance state is completed by vigilance criterion, the weight is updated according to the following equation [8]:

$$W_j^{new} = \beta (I \wedge W_j^{old}) + (1 - \beta) W_j^{old} \tag{3}$$

The performance of FAMNN is affected by three network parameters:

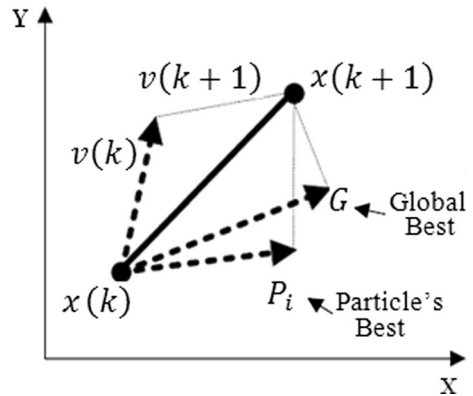
- 1- The choice parameter α ($\alpha > 0$), which acts on the category selection.
- 2- The baseline vigilance parameter, ρ (ρ_a, ρ_b , and ρ_{ab}) ($\rho_a \in [0,1]$) that controls the network resonance. The vigilance parameter is responsible for the number of formed categories.
- 3- The learning rate, ($\beta \in [0,1]$) that controls the velocity of network adaptation.

Table 1 shows the specifications of simulated FAMNN in this work.

Table 1 Specification of FAMNN in the base experiments

Specification	Value
Learning rate (β)	1
Vigilance parameter (ρ)	0.99
Choice parameter (α)	1
Number of classes	7
Number of training samples	384
Number of test samples	96

Fig. 4 PSO update of a particle's position $x(k)$, to $x(k+1)$ in a 2-dimensional space



3.6 Optimization

As mentioned before, the optimum values for FAMNN parameters were determined by PSO.

The PSO algorithm was first proposed by Kennedy and Eberhart in 1995 [25]. This algorithm is an evolutionary technique that was inspired by social behavior of bird flocking or fish schooling, and simulates the nature of the particles in a swarm. Figure 4 shows the examples of these patterns in nature. The PSO algorithm provides a population-based search procedure in which individuals, called particles, change their position (state) with time. In a PSO system, the particles fly around a multi-dimensional search space. During the flight, each particle adjusts its position according to its own experience and neighboring particle, making use of the best position encountered by itself and its neighbor. In this algorithm, each particle has a velocity and a position as follows [25]:

$$v_i(k+1) = v_i(k) + c_1 r_1 (P_i - x_i(k)) + c_2 r_2 (G - x_i(k)) \quad (4)$$

$$x_i(k+1) = x_i(k) + v_i(k+1) \quad (5)$$

Where, i is the particle index, k is the discrete time index, v_i is the velocity of the i th particle, x_i is position of the i th particle, P_i is the best position found by the i th particle (personal best), G is the best position found by swarm (global best), c_1 and c_2 are two positive constants called cognitive and social parameters ($c_1 = c_2 = 2$), and r_1 and r_2 are random numbers in the interval $[0,1]$ applied to the i th particle.

The PSO algorithm is similar to evolutionary computation (EC) techniques such as Genetic Algorithm (GA). These techniques are population-based stochastic optimization technique and utilize a fitness function to evaluate the population. They all update the population and search for the optimum with random techniques. The PSO unlike EC and GA techniques do not have genetic operators such as crossover and mutation. Particles update themselves with the internal velocity. Also, the information sharing mechanism in PSO is significantly different in comparison to other EC algorithms. In EC approaches, chromosomes share information with each other; thus, the whole population moves like one group towards an optimal area. But, in PSO, only the “best” particle gives out the



Fig. 5 Two PSO patterns in nature

information to others. PSO is designed very effective, in solving real valued global optimization problems, which makes it suitable for large scale studies. Figure 5 shows the update by PSO of a particle’s position from $x(k)$ to $x(k + 1)$.

4 Experimental

The audio-visual emotion recognition system was tested over the SAVEE audio-visual emotional database. All the experiments were person-independent. We used roughly 80 % of the data to train the classifiers and the remaining 20 % to test them. The emotion recognition was conducted through unimodal audio, unimodal visual, decision level, feature-level fusion

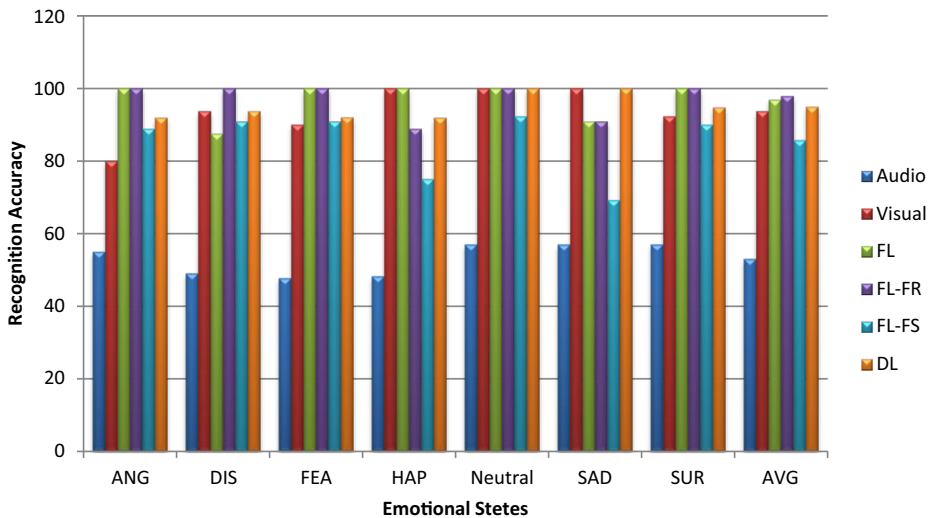


Fig. 6 Emotion recognition accuracy of the different systems. Each group of adjacent columns denote the classification accuracy of a single class. The first group contains the average recognition rate. The vertical axis is the recognition accuracy in percentage: Audio, Visual, FL(Feature-level fusion), FL-FR(Feature-level fusion after feature reduction), FL-FS(Feature-level fusion after feature selection), DL(Decision-level fusion). Class labels were abbreviated by their first three letters

(before feature reduction, after feature reduction and after feature selection). The results are presented in Fig. 6.

4.1 Audio experiments

In these experiments, 80 audio features were applied to PCA for feature reduction; 20 reduced features were applied to FCBF feature selection in the next stage, and 12 features were selected. The classification experiments were performed for seven emotional states using FAMNN. Figure 1 illustrates this setup using FAMNN1. The overall performance of this classifier was 53 %.

To show the good performance of our Audio recognition system, we examined it by eNterface'05 database [32]. The overall performance of this system was 63.1 %. The result was better than our previous work (55 %) [5]. This shows the good performance of our method for audio emotion recognition.

4.2 Visual experiments

In these experiments, 480 facial features were applied to PCA for feature reduction; 30 reduced features were applied to FCBF feature selection in the next stage, and 6 features were selected. The classification experiments were performed for seven emotional states using FAMNN. Figure. 1 shows this setup using FAMNN 2. The overall performance of this classifier was 93.75 %. The recognition accuracy in some states (e.g., happiness, neutral and sadness) was 100 %. Unfortunately, The SAVEE database is the only free public database that uses facial markers. So we could not to evaluate performance of our visual system.

4.3 Audio-visual experiments

The overall results of the unimodal systems showed that for accurate and reliable recognition of emotion classes, the modalities should be combined in a way that they benefit the interrelationships between the individual classes and the underlying modalities. In the following paragraphs, we present and compare different combination schemes. Two main fusion approaches used in the literature are feature level fusion and decision level fusion.

4.3.1 Decision level fusion

In this experiment, we used stacked generalization method for decision-level fusion. The output of the audio and visual ensembles serves as a feature vector to another FAMNN. The

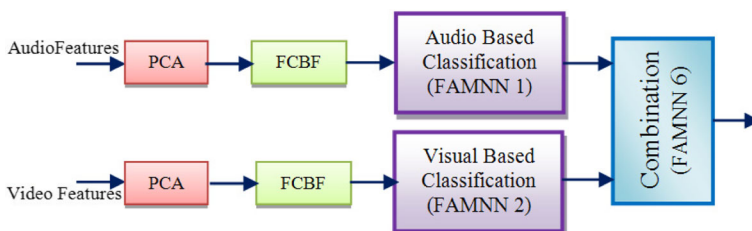


Fig. 7 Block diagram of the decision level fusion



Fig. 8 Block diagram of the feature level fusion

overall performance of this method was 95 %. As shown in Fig. 7, the output of FAMNN 1 and FAMNN 2 serves as a feature vector to FAMNN 6.

4.3.2 Feature level fusion

In this experiment, all audio and visual features were combined to get a total of 560 audio-visual features. Then these features were applied to PCA for feature reduction. Of which, 67 reduced features were applied to FCBF feature selection in the next stage and 6 features were selected. The classification experiments were performed for seven emotions with FAMNN. The overall performance of this emotion recognition system based on the audio-visual information at the feature level fusion classifier was 96.88 %. As shown in Fig. 8, this classifier has been performed by FAMNN 3.

4.3.3 Fusion after feature reduction

The 30 reduced audio and 20 reduced visual features were combined together and then FCBF was applied to the 50 reduced audio and visual features. In the next stage, 10 selected features were applied to the FAMNN classifier. The overall performance of this emotion recognition system based on Audio-visual data at feature level fusion after feature reduction was 97.92 %. Fig. 9 indicates this classifier with FAMNN 4. Table 2 shows the confusion matrix of the emotion recognition system based on the audio-visual data at feature level fusion after feature reduction. The recognition accuracy in some states (e.g., anger, disgust, fear, neutral and sadness) was 100 %. Also, some emotions are usually confused. Happiness is misclassified as surprise state by about 11.11 % and sadness is misclassified as fear by about 9.09 %.

4.3.4 Fusion after feature selection

The 6 selected audio features and the 12 selected visual features grouped together were applied to the FAMNN. The overall performance of this emotion recognition system based on the audio-visual data at feature level fusion after feature selection was 85.72 %. This classifier with FAMNN 5 is shown in Fig. 10.



Fig. 9 Block diagram of the fusion after feature reduction

Table 2 Confusion matrix of the emotion recognition system based on Fusion after feature reduction

	ANG	DIS	FEA	HAP	Neutral	SAD	SUR
ANG	100	0	0	0	0	0	0
DIS	0	100	0	0	0	0	0
FEA	0	0	100	0	0	0	0
HAP	0	0	0	88.88	0	0	11.11
SUR	0	0	0	0	100	0	0
SAD	0	0	9.09	0	0	90.90	0
SUR	0	0	0	0	0	0	100

Table 3 shows the emotion recognition results for the unimodal and different combining methods.

Combining of audio and visual information in different ways enhances the performance of unimodal systems. The results showed that the feature level fusion after feature reduction has better performance. The mean accuracy of this method is 97.92. Accordingly, this method improves the recognition rate by up to 45 % over the audio based system, and by up to 4 % over the visual based system.

5 Influence of FAMNN parameters optimization on emotion recognition accuracy

As mentioned before, PSO was used in this study to determine the optimum values for FAMNN parameters.

The operation of FAMNN is affected by three network parameters: the choice parameter, α , the baseline vigilance parameter, ρ (ρ_a , ρ_b , and ρ_{ab}), and the learning rate, β , which has a value between 0 and 1. The choice parameter takes values in the interval (0,1), while the baseline vigilance parameter assumes values in the interval [0,1].

In this study, the optimum values of the mentioned FAMNN parameters, which are corresponding to the minimum of fitness function, were determined by PSO. The fitness function in FAMNN simulation is was obtained by:

$$F = \frac{1}{pc}$$

Where, pc is the percentage of correct classification. The parameters' setting in PSO algorithm is listed in Table 4.

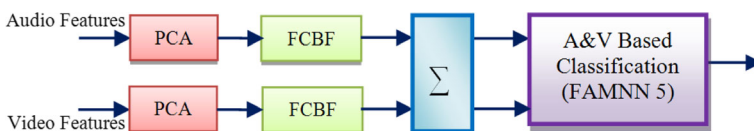


Fig. 10 Block diagram of the fusion after feature selection

Table 3 Recognition rate of emotional states for various proposed systems

	Audio	Visual	FL	FL-FR	FL-FS	DL
ANG	55	80	100	100	88.88	91.9
DIS	49	93.75	87.5	100	90.9	93.75
FEA	47.7	90	100	100	90.9	92
HAP	48.2	100	100	88.88	75	91.9
Neutral	57	100	100	100	92.30	100
SAD	57	100	90.9	90.90	69.23	100
SUR	57	92.30	100	100	90	94.75
AVG	53	93.75	96.88	97.92	85.72	95

The optimized parameters of the FAMNN for audio, visual, audio-visual data at decision level fusion and feature level fusion after feature reduction (the best result of our experiments) are reported in Table 5. The accuracies of emotion recognition when using the optimized FAMNN parameters are reported in Table 5 for mentioned modals. Table 5 further the accuracies of these modals when the parameters of the FAMNN are set by the user to $\alpha = 1$, $\beta = 1$ and $\rho_a = \rho_b = \rho_{ab} = 0.99$ as typical values.

As can be seen in Table 5, by using the optimized audio FAMNN parameters, the average audio emotion recognition accuracy improves by at least 10.5 %. The average audio emotion recognition is 63 %. Similar to the audio system, the average visual emotion recognition accuracy improves by at least 2 %, and the average visual emotion recognition is 95.83 %.

The best result of our experiments is for the audio-visual feature level fusion after feature reduction. By using the optimized FAMNN parameters for this system, the average emotion recognition accuracy improves by at least 0.33 %. So, the best result in this work is 98.25 %.

In previous work [19], we optimize audio emotion recognition FAMNN with GA. To compare performance of different methods of optimization, Specifications of the GA optimized FAMNN are reported in Table 6. Experimental results show that the two algorithms, almost the same results. But PSO, partly has better results and faster as well.

Emotion recognition performance for multimodal emotion recognition systems in other works and human evaluation may be helpful for analyzing the performance of the proposed work.

In SAVEE database, each actor's data were evaluated by 10 subjects at utterance level in three ways: audio, visual, and audio-visual. All of the evaluators were students at University of Surrey, UK with the age range of 21 to 29 years. To avoid gender biasing, half of the

Table 4 Specifications of the PSO algorithm for optimizing FAMNN

Specification	Value
Population size	30
C1	2
C2	2
Max number of iteration ρ_a	50
Max particle velocity	4

Table 5 Specifications of the optimized FAMNN

	α	β	ρ_a	ρ_{ab}	Non-optimized FAMNN CR	Optimized FAMNN CR
Audio	0.37	0.77	0.80	0.87	53	63.5
visual	0.44	0.99	0.97	0.91	93.75	95.83
Audio-visual	0.001	1	0.99	0.95	97.92	98.25
DL	0.46	0.79	0.99	0.96	95	97.37

evaluators were female. Also 5 of them were native, and the rest had lived in the UK for more than a year. The 120 clips from each actor were divided into 10 groups, resulting 12 clips per group. For each evaluator, a different data set was created, which resulted in 10 different sets for each of the audio, visual and audio-visual data per actor. The subjects were trained by using slides containing three facial expression pictures, two audio files, and a short movie clip for each of the emotions. The subjects were asked to play audio, visual and audio-visual clips, and select from one of the seven emotions on a paper sheet. The responses were averaged over 10 subjects for each actor. The average human classification accuracy is shown in Table 7. The mean was averaged over 4 subjects.

The performance of our work was lower than human evaluation for audio data. But our work showed higher classification performance than human evaluation for visual and audio-visual information. There are a few possible reasons for this. First, the difference in training data, i.e., machine was trained on a large part of data but humans were trained on a small amount of data, the task was discrete emotion classification, and the emotions may not be properly acted. Also, the quality of the human evaluators, they might not be very good at the job. But it is typical in the literature [10, 21, 23, 35]. The best visual system overall results was 95.83 % (for human is 88 %) and with audio-visual system, it was 98.25 % (for human is 91.8). The comparison of human perception and this work is shown in Fig. 11. Table 8 shows the performance of the proposed system and other multimodal emotion recognition systems.

6 Conclusion

The basics of most existing researches on emotion recognition can be summarized in three stages: feature extraction, feature selection and emotion recognition. A number of promising methods for audio, visual, and audio-visual feature extraction and feature selection have so far been proposed. So, feature extraction was not the goal of this paper. However, the good set of audio features was used in this paper. And, by using facial point markers on the face, visual

Table 6 Specifications of the GA optimized FAMNN

	α	β	ρ_a	ρ_{ab}	Non-optimized FAMNN CR	Optimized FAMNN CR
Audio	0.33	0.7	0.90	0.84	53	60
visual	0.31	0.94	0.59	0.53	93.75	95.83
Audio-visual	0.42	0.9	0.95	0.39	97.92	97.92
DL	0.02	0.99	0.76	0.47	95	97.37

Table 7 Average human classification accuracy

Human	KL	JE	JK	DC	Mean
Audio	53.2	67.7	71.2	73.7	66.5
Visual	89	89.8	88.6	84.7	88
Audio-visual	92.1	92.1	91.3	91.7	91.8

Mean is averaged over 4 subjects (KL-JE-JK-DC) data

emotion recognition was carried out appropriately. By using FCBF feature selection method, the efficient features are determined. Developing better methods and classifiers for emotion recognition and, fusion of different systems are two of the most important issues that need sufficient attention. We are focused on fusion of powerful classifiers in order to improve the performance of emotion recognition rate. Different classifiers such as ANNs, SVMs, KNN, and GMMs have been used for emotion recognition. Here, we used PSO optimized FAMNN as a powerful classifier. Also, a different fusion of audio and visual systems was tested. Experimental results confirm the excellent performance of our classifier and optimization by PSO. Also, they show that the performances of audio and visual systems were improved by using the different fusions.

This paper proposed particle swarm optimization-based FAMNN for audio-visual emotion recognition. FAMNN combines audio and visual information in feature and decision levels using stacked generalization approach. For this purpose, we employed the audio features such as MFCC, pitch, energy and formants, as well as marker locations on the face as visual features.

Experimental results showed that the performances of unimodal systems were improved by using the feature level and decision level fusions and PSO optimization-based FAMNN. The PSO algorithm was employed to determine the optimum value of the choice parameter (α), the vigilance parameters (ρ) and the learning rate (β) of FAMNN. As result, the recognition rate was improved by about 45.25 % with respect to the non-optimized audio unimodal system, and by about 5 % with respect to the visual system. In this study, we focused on a well proposed multi-modal fusion approach. The final emotion recognition rate on the SAVEE database was reached to 98.25 % using audio and visual features by using the optimized FAMNN.

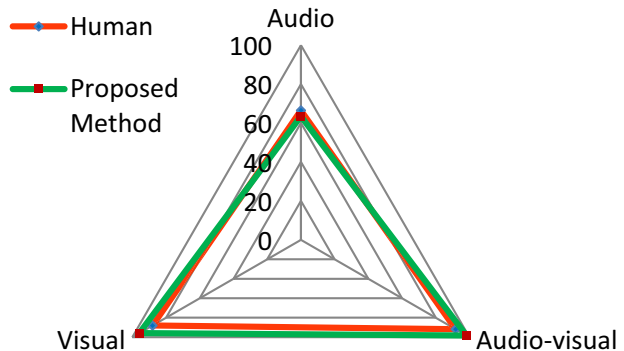
Fig. 11 Comparison of human perception and this work

Table 8 Performance of typical systems for multimodal emotion recognition in the recent decade

Reference	Classifier	Features	Fusion	Database	Acc	Acc (Fusion)
Paleari and Huet [38]	SVM, NN	Facial points, Formants, Prosody, MFCC, LPC	F,D	eNterface'05	Audio:25 Video:33	39
Busso et al. [7]	SVM	102 markers, Prosody	F,D	an actress	Audio:70.9 Video:85.1	89
Mansoozadeh et al. [31]	SVM	Facial points, Prosody	F,D,H	eNterface'05	Audio:33 Video:37	71
De Silva and Pei Chi [14]	HMM, Nearest neighbor	Facial points, Pitch	D	2 subjects, 144 clips	Audio:62 Video:32	72
Cheng-Yao et al. [11]	SVM	Facial points, Prosody	F	2 subjects, 350 clips	Audio:63 Video:75	84 ^{***}
Schuller et al. [43]	SVM	Face model, Formants, Prosody, MFCC	F	ABC database	Audio:74 Video:61	81
Zeng et al. [48]	SNoW and HMM	Facial points, Prosody	D	20 subjects, Eleven affect categories	Audio:58 Video:44	95
Bejani et al. [5]	Multi-classifier	ITMI, QIM	F,D,H	eNterface'05	Audio:54.99 Video:39.27	77.78
Haq et al. [21]	Gaussian (PCA)	Facial points, Prosody, formants, Duration, spectral	F,D	SAVEE	Audio:50 Video:91	92.9
Haq and Jackson [21]	Gaussian (LDA)	Facial points, Prosody, Formants, Duration, Spectral	F,D	SAVEE	Audio:56 Video:95.4	97.5
Banda et al. [4]	SVM	LBP, Prosody, Energy	D	SAVEE	Audio:79 Video:95	98
Haq et al. [20]	SVM	Facial points, 830 audio features	D	SAVEE	Audio:58 Video:66	83
This work	Optimized FAMNN	Facial points, Prosody, Formants, Duration, Spectral	F,D	SAVEE	Audio:63 Video:95.8	98.25

Future works will investigate new combining classification methods such as mixture of experts or new ways to the optimize FAMNN such as Cuckoo [42], Grey Wolf [34] and Imperialist Competitive Algorithm (ICA) [3].

Acknowledgments This work was supported by Islamic Azad University-South Tehran Branch under a research project entitled “Audio-Visual Emotion Modeling to Improve human-computer interaction”.

References

1. 3DMD 4D Capture System. Online: <http://www.3dmd.com>, accessed on 3 May, 2009
2. Ambady N, Rosenthal R (1992) Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. *Psychol Bull* 111(2):256–274
3. Atashpaz-Gargari E, Lucas C, (2007) Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition. *IEEE Congress on Evolutionary Computation* 4661–4667
4. Banda N, Robinson P (2011) Noise analysis in audio-visual emotion recognition. *International Conference on Multimodal Interaction*, Alicante, Spain
5. Bejani M, Gharavian D, Moghaddam Charkari N (2012) Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks. *Neural Comput & Applic* 24(2):399–412
6. Boersma P, Weenink D (2007) Praat: doing phonetics by computer (version 4.6.12) [computer program]
7. Busso C et al (2004) Analysis of emotion recognition using facial expressions, speech and multimodal information. In: *Proceedings of the sixth ACM International Conference on Multimodal Interfaces (ICMI '04)*, pp 205–211
8. Carpenter GA (2003) Default ARTMAP. In: *Proceedings of the International Joint Conference on Neural Networks*, Portland, Oregon, USA vol 2. pp 1396–1401
9. Carpenter GA, Grossberg S, Markuzon N, Reynolds JH, Rosen DB (1992) Fuzzy ARTMAP: a neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Trans Neural Netw* 3:698–713
10. Chen C, Huang Y, Cook P (2005) Visual/acoustic emotion recognition. *ICME 2005*:1468–1471
11. Cheng-Yao C, Yue-Kai H, Cook P (2005) Visual/acoustic emotion recognition, pp 1468–1471
12. Cootes TF, Edwards GJ, Taylor CJ (2001) Active appearance models. *IEEE Trans Pattern Anal Mach Intell* 23(6):681–685
13. Dai W, HanD DY, Xu D (2015) Emotion recognition and affective computing on vocal social media. *Inf Manag*. doi:10.1016/j.im.2015.02.003
14. De Silva LC, Pei Chi N (2000) Bimodal emotion recognition. In: *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, vol 1. pp 332–335
15. Devillers L, Vidrascu L (2006) Real-life emotions detection with lexical and paralinguistic cues on human human call center dialogs. In: *The proceedings of Interspeech*, pp 801–804
16. Ekman P (1971) Universals and cultural differences in facial expressions of emotion. *Proc Nebr Symp Motiv* 19:207–283
17. Ekman P, Rosenberg EL (2005) *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system*. Second ed. Oxford Univ Press
18. Fleuret F (2004) Fast binary feature selection with conditional mutual information. *J Mach Learn Res* 5:1531–1555
19. Gharavian D, Sheikhan M, Nazerieh AR, Garoucy S (2011) Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. *Neural Comput & Applic*. doi:10.1007/s00521-011-0643-1
20. Haq S, Asif M, Ali A, Jan T, Ahmad N, Khan Y (2015) Audio-visual emotion classification using filter and wrapper feature selection approaches. *Sindh Univ Res J (Sci Ser)* 47(1):67–72
21. Haq S, Jackson PJB (2009) Speaker-dependent audio-visual emotion recognition. In: *Proc. Int'l Conf. on Auditory-Visual Speech Processing*, pp 53–58
22. Harley Jason M et al (2015) A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system. *Comput Hum Behav* 48:615–625. doi:10.1016/j.chb.2015.02.013
23. Hassan A, Dampier R (2010) Multi-class and hierarchical SVMs for emotion recognition. *ISCA, INTERSPEECH*, pp 2354–2357
24. Hoch S, Althoff F, McGlaun G, Rigooll G (2005) Bimodal fusion of emotional data in an automotive environment. In: *The Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* vol 2. pp 1085–1088

25. Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of the IEEE International Conference on Neural Networks, Perth, Australia vol 4. pp 1942–1948
26. Klein J, Moon Y, Picard RW (2002) This computer responds to user frustration: theory, design and results. *Interact Comput* 14:119–140
27. Lee C-C, Mower E, Busso C, Lee S, Narayanan S (2009) Emotion recognition using a hierarchical binary decision tree approach. In: The proceedings of Interspeech, pp 320–323
28. López-de-Ipiña K, Alonso-Hernández JB et al (2015) Feature selection for automatic analysis of emotional response based on nonlinear speech modeling suitable for diagnosis of Alzheimer's disease. *Neurocomputing* 150:392–401. doi:10.1016/j.neucom.2014.05.083
29. Luxand FaceSDK 5.0.1 Face Detection and Recognition Library. online: <https://www.luxand.com/facesdk/index.php>
30. Mansoorizadeh M, Moghaddam Charkari N (2009) Hybrid feature and decision level fusion of face and speech information for bimodal emotion recognition. Proceedings of the 14th International CSI Computer Conference
31. Mansoorizadeh M, Moghaddam Charkari N (2009) Multimodal information fusion application to human emotion recognition from face and speech. *Multimed Tools Appl*
32. Martín O, Kotsia I, Macq B, Pitas I (2006) The interface'05 audio-visual emotion database. In: Proc. 22nd intl. conf. on data engineering workshops (ICDEW'06)
33. Mehrabian A (1968) Communication without words. In: *Psychology Today*, vol 2. pp 53–56
34. Mirjalili SA, Mirjalili SM, Lewis A (2014) Grey wolf optimizer. *Adv Eng Softw* 69:46–61. (<http://www.mathworks.com/matlabcentral/fileexchange/44974-grey-wolf-optimizer-gwo->)
35. Morrison D, Wang R, Silva D (2007) Ensemble methods for spoken emotion recognition in call-centres. *Speech Comm* 49(2):98–112
36. Oudeyer P-Y (2003) The production and recognition of emotions in speech: features and algorithms. *Int J Hum Comput Interact Stud* 59:157–183
37. Paleari M, Benmokhtar R, Huet B (2008) Evidence theory-based multimodal emotion recognition. In: *InMMM'09*, pp 435–446
38. Paleari M, Huet B (2008) Toward emotion indexing of multimedia excerpts. In: *CBMI*
39. Pantic M, Rothkrantz LJM (2000) Automatic analysis of facial expressions: the state of the art. *IEEE Trans Pattern Anal Mach Intell* 22:1424–1445
40. Picard RW (1997) *Affective computing*. MIT Press
41. Polzehl T, Sundaram S, Ketabdar H, Wagner M, Metze F (2009) Emotion classification in children's speech using fusion of acoustic and linguistic features. In: The proceedings of Interspeech, pp 340–343
42. Rajabioun R (2011) Cuckoo optimization algorithm. *Appl Soft Comput* 1.11:5508–5518. (<http://www.mathworks.com/matlabcentral/fileexchange/35635-cuckoo-optimization-algorithm>)
43. Schuller B, Arsic D, Rigoll G, Wimmer M, Radig B (2007) Audiovisual behavior modeling by combined feature spaces. In: *ICASSP*, pp 733–736
44. Sheikhan M, Bejani M, Gharavian D (2012) Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method. *Neural Comput Appl J*. doi:10.1007/s00521-012-0814-8
45. Shlens J (2005) A tutorial on principal component analysis. Systems Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla
46. Song M, You M, Li N, Chen C (2008) A robust multimodal approach for emotion recognition. *Neurocomputing*
47. Weisgerber A, Vermeulen N et al (2015) Facial, vocal and musical emotion recognition is altered in paranoid schizophrenic patients. *Psychiatry Res*. doi:10.1016/j.psychres.2015.07.042
48. Zeng Z, Hu Y, Roisman GI, Wen Z, Fu Y, Huang TS (2007) Audio-visual spontaneous emotion recognition. *Artif Intell Hum Comput* 4451:72–90
49. Zeng Z, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. *PAMI* 31:39–58



Davood Gharavian was born in Neyshabour, Iran, in 1973. He received the B.S. degree in electronic engineering from Amirkabir University, Tehran, Iran, in 1995 and M.S. in communication engineering from Tarbiat Modares University, Tehran, Iran in 1998 and Ph.D. degrees in electronic engineering from Amirkabir University, Tehran, Iran in 2004. He is currently an Assistant Professor in Electrical Engineering Department of Shahid Beheshti University. His research interests include digital signal processing, speech and image processing, digital signal processor, industrial networks and smart grid.

Dr. Gharavian has published more than 21 journal papers and more than 17 conference papers. He is the author of two books in the fields of communication systems and power line carrier.



Mehdi Bejani received the B.S. degrees in Electrical Engineering from Shahid Rajaei University, Tehran, Iran (2007), and the M.S. degrees in Electrical Engineering from Islamic Azad University, South Tehran Branch in 2010. His research interests include Human computer interaction, multimodal affective computing, Machine learning and Image & Speech Processing.



Mansour Sheikhan was born in Tehran, Iran, in 1966. He received the B.S. degree in electronic engineering from Ferdowsi University, Meshed, Iran, in 1988 and M.S. and Ph.D. degrees in communication engineering from Islamic Azad University, Tehran, Iran, in 1991 and 1997, respectively. He is currently an Associate Professor in Electrical Engineering Department of Islamic Azad University-South Tehran Branch. His research interests include security in communication networks, intelligent systems, signal processing, and neural networks.

Dr. Sheikhan has published more than 70 journal papers and more than 60 conference papers. He is the author of two books in the fields of optical signal processing and communication systems and has been selected as the outstanding researcher of IAU in 2003, 2008, and 2010–2013.