Optimizing Fuzzy Inference Systems for Improving Speech Emotion Recognition

Reda Elbarougy [1,2] and Masato Akagi [1]

[1] Japan Advanced Institute of Science and Technology (JAIST), Japan
[2] Department of Math., Faculty of Science, Damietta University, New Damietta, Egypt
elbarougy@du.edu.eg,akagi@jaist.ac.jp

Abstract. Fuzzy Inference System (FIS) is used for pattern recognition and classification purposes in many fields such as emotion recognition. However, the performance of FIS is highly dependent on the radius of clusters which has a very important role for its recognition accuracy. Although many researcher initialize this parameter randomly which does not grantee the best performance of their systems. The purpose of this paper is to optimize FIS parameters in order to construct a high efficient system for speech emotion recognition. Therefore, a novel optimization algorithm based on particle swarm optimization technique is proposed for finding the best parameters of FIS classifier. In order to evaluate the proposed system it was tested using two emotional speech databases; Fujitsu and Berlin database. The simulation results show that the optimized system has high recognition accuracy for both languages with 97% recogintion acuracy for Japanese and 80% for German database.

Keywords: Fuzzy Inference System (FIS), Particle swarm optimization, Speech emotion recognition, Optimum clusters radius

1 Introduction

Fuzzy Inference System (FIS) is used for pattern recognition and classification purposes such as emotion classification [2]. FIS can be constructed using expert knowledge by human or from data. FIS based on expert knowledge only for complex systems such as emotion recognition system may suffer from a loss of accuracy. The relationship between the multivariable input acoustic features and output emotional state in emotion recognition system is fuzzy and complex [3]. Therefore generated partitioning is more difucult and meaningless for human experts. Therefore the main approach for constructing FIS is by using fuzzy rules inferred from data, such as the data-driven FIS constructed in [6]. In order to construct a FIS from data can be implemented using two main stages: automatic rule generation or initial system and the other is fine-tuning of the initial system parameters. In the first stage rule generation leads to an initial system with a

given space partitioning and the corresponding set of rules and specific structure. The Adaptive Neuro-Fuzzy Inference System (ANFIS) is used as a tool for fine-tuning the parameters of initial FIS [6].

Generating the initial FIS is a very important step in constructing a classification system using ANFIS [8]. In the initialization step the structure of the FIS is determined i.e the number of nodes, the number of membership functions for each input and output, and the number of rules [9]. This structure will be fixed through the training classification process. Therefore, this paper investigates how to optimize the structure of initial FIS for automatic rule generation. The function 'genfis2' in Fuzzy Logic Toolbox generates initial FIS model from the training data using subtractive clustering algorithm, and it requires the user to specify parameter cluster radius. The cluster radius indicates the range of influence of a cluster when the data space is considered as a unit hypercube. Specifying a small cluster radius will usually yield many small clusters in the data, resulting in many rules and vice versa [10].

Although, the estimation results of FIS are sensitive to the initialized FIS which is determined by selecting cluster radii. However, most of the previous studies neglect this effect by randomly selecting values such as 0.5 as in [4] or 0.95 as in [5] without any investigation of the effect of these values on the final results of their systems. In order to avoid this problem it is very important to prevent subjectivity of choosing this parameter. This can be done by optimizing this parameter. The purpose of this paper is to improve the estimation accuracy of the FIS in order to accurately estimate the emotional state from the speech signal, by finding the best cluster radii which determine the influence of effect of input and output data.

This paper proposed a method for optimizing the cluster radii for the following reasons: (1) the vector of radius has very important role for the recognition accuracy. (2) The radii determine the structure of the FIS: the number of membership functions (MFs) for each input and output variable and consequently the number of fuzzy rules (3) to reduce the computation complexity by selecting the appropriate radii (small radii leads to large number of MFs very large number of parameters which will need too much time to be estimated). (4) to avoid subjectivity of choosing this parameter. Therefore, a method for selecting the optimal radii is required to improve the performance of FIS as well as avoid the above drawbacks.

This paper investigates the design of a high efficient system for emotion recognition from the speech signal. In the literature the emotional states can be represented by the categorical approach such as happy, anger or can be represented as in a two-dimensional space spanned by the two basic dimensions valence (negative-positive axis), and activation (calm-excited axis) [1]. The emotional state in this paper is represented by hybrid model which can estimate emotion dimensions valence-activation as well as emotion category. Three FIS were used to detect the emotional state: two of them to estimate the two basic dimensions valence and activation and the third FIS was used to map the estimated emotion dimensions into emotion categories. Therefore, a novel optimization algorithm is

proposed for finding the best cluster radii parameter for the used FIS classifier based on Particle Swarm Optimization (PSO) technique.

2 Speech Material

In order to validate the proposed method, two emotional speech databases were used, one in the Japanese language and the other in the German language. The Japanese database is the multi-emotion single speaker Fujitsu database produced and recorded by Fujitsu Laboratory, it contains five emotional states: neutral, joy, cold anger, sad, and hot anger as described in [11].

The German database is the Berlin database. It comprises seven emotional states: anger, boredom, disgust, anxiety, happiness, sadness, and neutral speech. An equal distribution of the four similar emotional states (neutral, happy, angry, and sad) as follows: 50 happy, 50 angry, 50 sad, and 50 neutral; in total, 200 utterances were selected from the Berlin database.

For constructing a speech emotion recognition system based on the proposed method using the dimensional approach, many acoustic features must be extracted and the two emotion dimensions must be evaluated using human subjects for each utterance in the two databases. Therefore, a listening test was used to evaluate valence and activation as explained in [11], for the two databases. Then, an initial set of 21 acoustic features were extracted for each database. Moreover, the feature selection method proposed by Elbarougy and Akagi based on a three-layer model of a human perception model was used to select the most related acoustic features for each emotion dimensions [11]. Finally, 11 and 10 acoustic features were selected for Japanese and German databases, respectively.

3 The proposed optimization method

Previous studies show that the cluster radii have a great effect on the accuracy of the FIS. Therefore, this study tries to find the optimal cluster radii. Our proposed radii selection method is based on the following assumption: the smallest root mean squared error (RMSE) for initial FIS has a greater potential for achieving a lower RMSE when applying the training using FIS. Thus, we assume that the cluster radii for the initial FIS which correspond to the minimum RMSE is the optimal radii. Therefor, initializing FIS using this optimal radii will have a large impact for predicting values of emotion dimensions. The next subsection introduces the proposed method for finding the optimal radii. Finally, a FIS will be constructed using the obtained optimal radii in order to accurately estimate valence and activation form the extracted acoustic features.

3.1 Particle swarm optimization method for cluster radii

Particle swarm optimization is a stochastic, population based swarm intelligence algorithm developed by Kennedy and Eberhart [12]. The PSO algorithm is similar to evolutionary computation in producing a random population initially and

generating the next population based on current cost. Thus, PSO is faster in finding solutions compared to other evolutionary computation techniques. In this paper we proposed PSO for optimizing the radii. This method widely used for parameter optimization in many fields for example, it has been applied to optimize premise and consequent parameters in order to make the ANFIS output fit the training data [13, 14]. However, in this paper this method will be used to optimize radii in order to improve the accuracy of FIS.

In order to implement the PSO method for optimizing radii, the following steps will be done. (1) finding the suitable range for searching the optimal radii. (2) applying the PSO in the determined range to find the optimal radii for valence and activation. In this study, only the effect of radii parameter has been investigated. Since this parameter has the highest effect in changing the resulting clusters number and consequently the structure of FIS. The radius of each cluster specifies the range of influence of the cluster center. Specifying a smaller cluster radius will yield more smaller clusters in the data, consequently more rules. Therefore, in order to find the suitable range of radii, the relationship between the radii and the number of rules is investigated. Figure 1 shows the relationship between the number of rules for different radii in the range from 0.1 to 2.

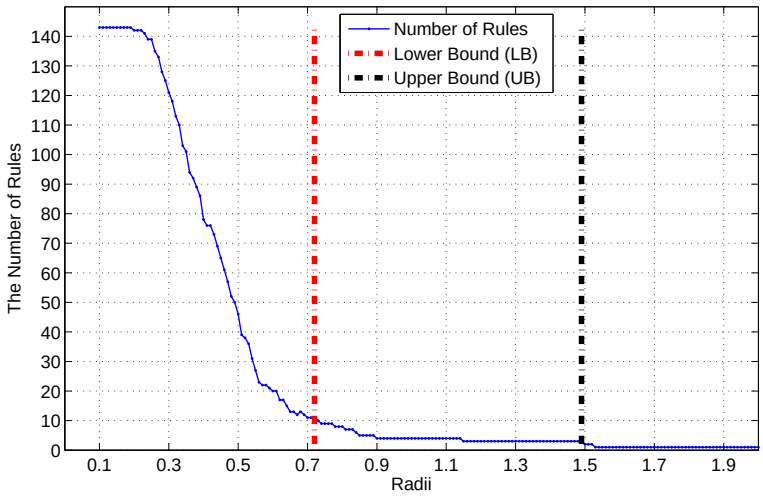


Fig. 1. Radius of clusters versus number of rules for initial FIS generated by *'genfis2'*.

From this figure we can easily notice that a smaller value of the cluster radius results in a large number of rules, and vice versa. Larger numbers of rules usually take longer for calculations for example, using radii=0.5 for calculating valence form the selected 10 acoustic features leads to 46 fuzzy rules as shown in Fig. 2.

This figure shows the structure of the initial FIS generated for estimating valence by subtractive cluster algorithm using Matlab function 'genfis2' at radii=0.5. The number of membership functions will be 46 for each input-output which need too much time in order to calculate the parameters of each membership function Therefore, it is better to use the suitable range which correspond to at least 11 fuzzy rules which correspond to the left vertical line in Fig. 2 at radii=0.72 which was considered as the Lower Bound (LB) for the searching range. To construct FIS it needs at least two rules to start training using ANFIS. Therefore, the radii which correspond to less than 2 rules must be removed for searching range. Hence, the radii correspond to 3 fuzzy rules is selected as the Upper Bound (UB) for the searching range of radii.

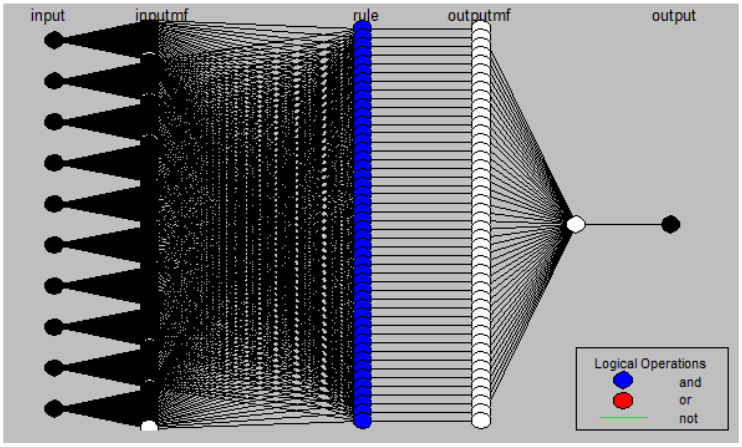


Fig. 2. The structure of the initial FIS generated by subtractive cluster at Radii=0.5.

The second step is to apply the PSO algorithm to find the optimal radii in the selected range. The mathematical representation of PSO is given as

$$x_{id_{new}} = x_{id} + v_{id_{new}} \tag{1}$$

$$v_{id_{new}} = w * v_{id} + c_1 * rand_1 * (P_{id} - x_{id}) + c_2 * rand_2 * (G_d - x_{id}) \tag{2}$$

where x_{id}, v_{id} represent the position vector and the velocity vector of the i^{th} particle in the d-dimensional search space, respectively. In addition, c_1,c_2 are the acceleration constants, $rand_1$, $rand_2$ are uniformly generated random numbers between 0 and 1; w is the inertia weight which decreased linearly from 0.9 to 0.4 during the run [13]. The first part of equation 2 represents the inertia of the previous velocity. The second part is the cognition part and it provides the best own position for the particles, where P_{id} is the best previous position. The third part is the social component which represent the collaborative effect of

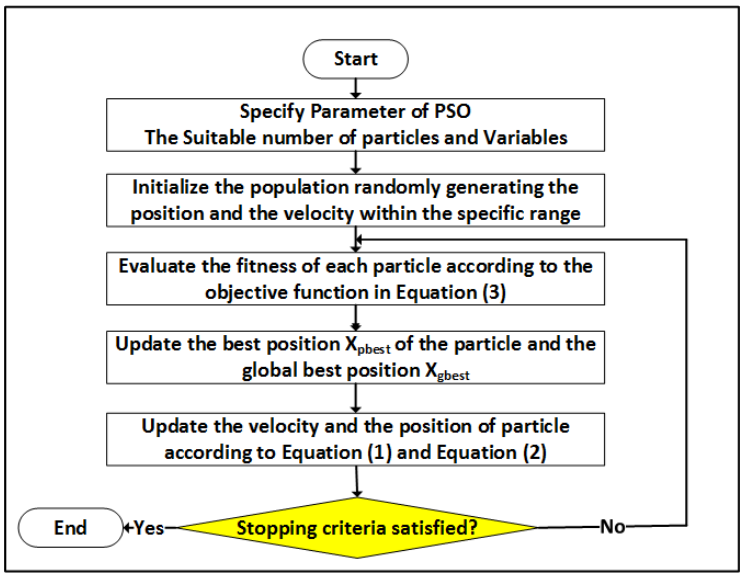


Fig. 3. The flowchart of PSO algorithm for Optimizing Radii.

all participles for finding the global optimal solution, where G_d is the global best position. The third component pulls the particles towards the global best position.

In this study, the PSO algorithm is used to find the optimum radii where $radii = (r_1, r_2, ..., r_d)$, where d represents the number of input-output. For example, in order to estimate valence as shown in Fig. 2 the number of input is 10 and the number of output is one i.e., $d = 11$ in this case. The flowchart shown in Fig. 3is used to optimize the radii where the fitness function here is the RMSE between the actual output and the desired output which can be described by:

$$fitness = RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - y_i)^2} \tag{3}$$

where y_i is the actual output evaluated by human subjects and $\widehat{y}_i$ is the desired output using the system, N represents the total number of data examples in the training dataset. In multidimensional data, different radii may be specified for each input-output dimension. If the same value is applied to all data dimensions, each cluster center will have a spherical neighborhood of influence with the given radius. For the propose of comparison we use the spherical neighborhood of influence i.e., the same value is applied for all the dimensions $radii = (r_1 = r, r_2 = r, ..., r_d = r)$.

4 Speech Emotion Recognition System

In this section we evaluate the performance of the proposed PSO-FIS for estimating valence and activation. Two databases were used to train and test the proposed system. The proposed emotion recognition system is shown in Fig. 4 which used to estimate the emotional state expressed in the speech signal. This system is consisted of two stages; the first stage the extracted acoustic features were used as inputs to FIS to estimate emotion dimensions valence and activation. In the second stage the estimated emotion dimensions were used as inputs to FIS to detect the emotion categories. FIS is multiple input, single output system therefore 3 FIS were required as follows: two FISs were used to estimate emotion dimension valence and activation from acoustic features, and one FIS was used to map the estimated emotion dimensions into emotion category neutral, happy, anger, and sad. The two stages of the proposed system will be explained in details in the next sections.

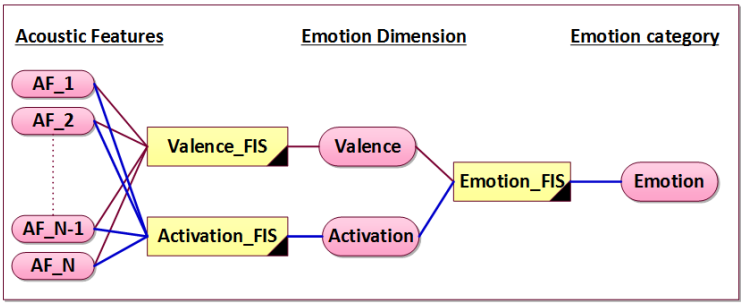


Fig. 4. The proposed speech emotion recognition system.

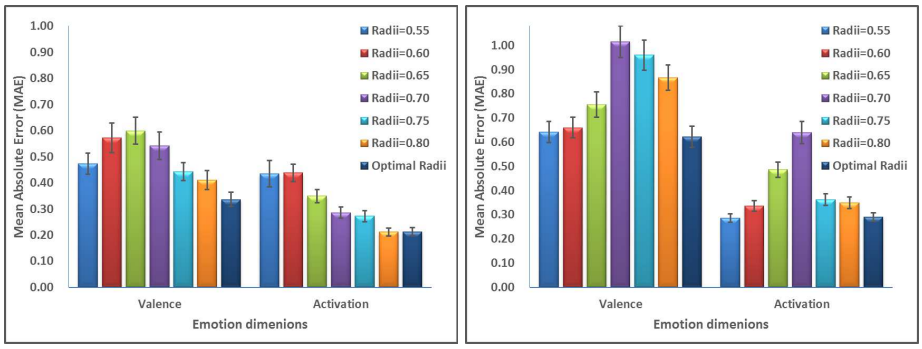
4.1 Emotion dimension estimation

Fujitsu and Berlin database were used to evaluate the proposed system performances. The PSO algorithm is implemented to find the optimal radii for each emotion dimension individually. Then, the optimal radii is used to generate the initial FIS for the investigated dimension. In addition, the initial FIS is trained using the two database individually. The mean absolute error (MAE) between the estimated values of emotion dimensions and the corresponding average value given by human subjects is used as a metric of the discrimination associated with each case. The MAE is calculated according to the following equation:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\widehat{y}_i - y_i| \tag{4}$$

where y_i is the actual output and $\widehat{y}_i$ is the desired output, N is number of the output.

The result of PSO-FIS for each emotion dimension generated using the optimal radii was compared with that generated using six different radius of cluster (0.55 ,0.60, 0.65, 0.70, 0.75, and 0.80). Figures 5(a) and 5 (b) show the MAEs for emotion dimensions for Japanese and German database, respectively.



(a) Japanese database. (b) German database.

Fig. 5. The estimation results for emotion dimensions using different radii.

For Japanese database the optimal radii for valence and activation were 0.89 and 0.78, respectively. In addition, for German database the optimal radii were 0.96 and 0.91, respectively. From these figures, we can see that the best results for each emotion dimensions were obtained form the optimal radii. Table 1 also reveals the more detailed view of different results with various radii for FIS. The correlation between the actual output y_i which evaluated by human subjects and the desired output $\widehat{y}_i$ using the system were calculated as listed in the third column of this table.

Increasing radius causes decreasing of the number of the clusters and therefore causes the decreasing of the number of the rules. FIS has gotten 27 rules for radius 0.55, 8 rules for radius 0.80. As it is seen in Table 1 depicting the results of valence dimension for different radii, where the star in the last row denotes the optimal radii. Decreasing the cluster radius increases complexity, however our results show that increasing radius causes performance improvement in terms of small MAE and high correlations between the estimated values and the human evaluation. Larger cluster radius assures well defined rules that helps to reduce the redundancy of data and determine the rules and membership functions, which is one of the benefits of the clustering method. The size of the cluster centers determines the number of rules associated with the fuzzy inference system.

The results of optimization steps for determining the optimal radii are reported in Table 2. It shows that using radii= correspond to 3 number of membership function is the optimal result.

Table 1. The results of valence for Japanese Database usinf different values of radii

Radii	No Of Rules	Correlation	MAE
0.55	27	0.82	0.47
0.60	20	0.72	0.57
0.65	13	0.71	0.60
0.70	11	0.73	0.54
0.75	9	0.85	0.44
0.80	8	0.86	0.41
Optimal= 0.89	**5**	**0.91**	**0.34**

These results for both languages revel that our proposed PSO-FIS for emotion dimension estimation has the ability to accurately estimate the emotion dimensions using the optimal radii, with a small MAE and high correlations.

4.2 Emotional Classification

The categorical and dimensional approaches are closely related, i.e. by detecting the emotional content using one of these two schemes, we can infer its equivalents in the other scheme. In this section, we want to strengthen our findings in this study by demonstrating that the dimensional approach can actually help us to improve the automatic emotion classification. So, the estimated values of emotion dimensions (valence and activation) were used as inputs for the FIS to predict the corresponding emotional category. The classifications results using the estimated values of emotion dimensions as shown in Tables 2(a) and 2(b) for the Japanese and German databases, respectively. These results for both languages revel that our proposed PSO-FIS for emotion classification has the ability to accurately predict the emotion category using the optimal radii, with a very high emotion classification rate for Japanese database with 97% and high classification rate for German database with 80%. The difference of the recognition rate between the two languages is due to Japanese database is a single speaker database however German database is multi-speaker database. The reason for the high classification accuracy is due the best estimation for emotion dimensions.

The Simulation results show that our proposed optimized system can accurately estimate emotion dimensions valence and activation compared with randomly selecting radii, using two different emotional databases. Finally the estimated values of valence and activation were used to preidect the emotional state.

5 Conclusion

The aim of this paper is to improve the emotion recognition accuracy using FIS by improving estimation accuracy for emotion dimensions; valence, and activation from acoustic features. In order to accomplish this task, first, a PSO method is used to find the optimal radii for each emotion dimensions, then, the PSO-FIS system was constructed based on an initial FIS generated using the optimal

Table 2. Classification results using FIS classifier:

(a) Japanese Database

Radii	MAE of dimensions		Classification
	Valence	Activation	Recognotion Rate %
0.55	0.47	0.43	84%
0.60	0.57	0.44	79%
0.65	0.60	0.35	83%
0.70	0.54	0.29	85%
0.75	0.44	0.27	92%
0.80	0.41	0.21	91%
Optimal	**0.34**	**0.21**	**97%**

(b) German Database

Radii	MAE of dimensions		Classification
	Valence	Activation	Recognotion Rate %
0.55	0.64	0.29	75%
0.60	0.66	0.34	76%
0.65	0.76	0.49	71%
0.70	1.02	0.64	60%
0.75	0.96	0.36	69%
0.80	0.87	0.35	67%
Optimal	**0.62**	**0.29**	**80%**

radii. For estimating emotion dimensions, the proposed PSO-FIS emotion recognition system was trained and testing using two different languages. The results for both language revel that our PSO-FIS emotion recognition system initialized using the optimal radii has the ability to accurately estimate the emotion dimensions, with a small errors as well as to improve the final recognition rate. The most important result is that increasing radius causes performance improvement, as well as decreases the complexity of the proposed system. The most important contribution of this study is that the proposed method can automatically select the optimal parameter and prevent subjectivity selection which may lead to different non-optimal results.

References

1. M. Grimm, and K. Kroschel, "Emotion Estimation in Speech Using a 3D Emotion Space Concept," *in Robust Speech Recognition and Understanding, M. Grimm and K. Kroschel (Eds.)*, June 2007.
2. Q. Zhang, S. Jeong, M. Lee, "Autonomous emotion development using incremental modified adaptive neuro-fuzzy inference system," *Neurocomputing*, **86**, pp. 33-44, (2012).
3. C. Huang, and M. Akagi, "A three-layered model for expressive speech perception," *Speech Communication*, **50(10)**, pp. 810–828, October, (2008).
4. E. Entchev, and L. Yang. "Application of adaptive neuro-fuzzy inference system techniques and artificial neural networks to predict solid oxide fuel cell performance

in residential microgeneration installation," *Journal of Power Sources* 1.170 (2007): 122-129.
5. Deregeh, F., Karimian, M., and Nezmabadi-Pour, H. (2013). "A New Method of Earlier Kick Assessment Using ANFIS," *Iranian Journal of Oil and Gas Science and Technology*, 2(1), 33-41.
6. C. Lee and S. Narayanan, "Emotion Recognition Using a Data-Driven Fuzzy Inference System," Proc. Eighth European Conf. Speech Comm. and Technology (EUROSPEECH ',03), pp. 157-160, 2003.
7. M. Grimm, and K. Kroschel, and E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, **49**, pp. 787–800, (2007).
8. Wei, M. Bai, B. Sung, AH. Liu, Q. Wang, J. Cather, M.E. "Predicting injection profiles using ANFIS," *Information Sciences*, 2007.
9. S. Guillaume, "Designing fuzzy inference systems from data: An interpretability-oriented review," in *IEEE Transactions on Fuzzy Systems*, vol. 9, no. 3, pp. 426-443, Jun 2001.
10. Zoveidavianpoor, M. Samsuri, A. and Shadizadeh, S.R. "Adaptive neuro fuzzy inference system for compressional wave velocity prediction in a carbonate reservoir," *J. Appl. Geophys* 89, 96-107, 2013.
11. Elbarougy, R. and Akagi, M. "Speech Emotion Recognition System Based on a Dimensional Approach Using a Three-Layered Model," *Proc. Int. Conf. APSIPA*, 2012.
12. Kennedy, J. Eberhart, R. "Particle Swarm Optimization," Proc. of IEEE Int. Conf. on Neural Network, Perth,Australia, pp. 1942–1948, 1995.
13. Toha, S. F. and Tokhi, M. O. "ANFIS modelling of a twin rotor system using particle swarm optimisation and RLS," Proc. of IEEE Int. Conf. on Cybernetic Intelligent Systems, Birmingham, United Kingdom, 2009.
14. Liu, P. Leng, W. and Fang, W. "Training anfis model with an improved quantum-behaved particle swarm optimization algorithm," Mathematical Problems in Engineering, pp. 1–10, 2013.
15. D. Wu, and T.D. Parsons, and S. Narayanan, "Acoustic Feature Analysis in Speech Emotion Primitives Estimation," *Proc. InterSpeech 2010*, pp. 785–788, 2010.
16. M. Schroder, and R. Cowie, and E.D.-cowie, M. Westerdijk, and S. Gielen, "Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis," *Proc. Eurospeech 2001*, pp. 87–90, 2001.
17. I. Kanluan, M. Grimm, and K. Kroschel, "Audio-Visual Emotion Recognition Using An Emotion Space Concept," *Proc. EUSIPCO 2008*, 2008.
18. R. Cowie, "Describing the emotional states that are expressed in speech," *Proc. ISCA Workshop on Speech and Emotion*, pp. 11-18, 2000.
19. C. Yu, P.M. Aoki, and A. Woodruff, "Detecting User Engagement in Everyday Conversations," *Proc. Eighth Intl Conf. Spoken Language Processing*, 2004.
20. M. Nicolaou, and H. Gunes, "Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space," *IEEE Transactions on Affective Computing*, vol. 2(2), pp. 92-105, 2011.
21. D. Wu, T. Parsons, E. Mower and S. Narayanan, "Speech Emotion Estimation in 3D Space," *Proc. ICME 2010*, 2010.