SERBIAN JOURNAL OF ELECTRICAL ENGINEERING
Vol. 20, No. 1, February 2023, 13-32

UDC: 159.942:[534.78:004.81 DOI: https://doi.org/10.2298/SJEE2301013S

Modeling of Human Mood States from Voice using Adaptively Tuned Neuro-Fuzzy Inference System

Biswajeet Sahu[1], Hemanta Kumar Palo[2], Mahesh Chandra[3]

Abstract: In this article, an attempt is made to model angry, happy, and neutral human mood states by adaptively tuning the Neuro-fuzzy Inference system for efficient characterization. The algorithm is self-tunable and can provide low-cost and robust solutions to many complex physical world problems. Such analysis can provide crucial inputs to many vivid application domains such as security organization, bio-medical engineering, computer tutors, call centers, banking and finance sectors, criminal investigations, etc. for effective functioning and control. The Surrey Audio-Visual Expressed Emotions (SAVEE) database has been chosen to procure the utterances corresponding to the chosen mood states. Initially, several feature vectors have been extracted that comprise Spectral Roll-off, Spectral Centroid, Spectral flux, Log Energy, Fundamental frequency, Jitter, and Shimmer to develop the desired models. The resultant Adaptive Neuro-Fuzzy Inference (ANFIS) algorithm can distinguish the chosen states based on the simulation models as revealed by the results. Performance measures such as the Root Mean Square Error at the start, convergence, minimal, checking, training, and testing have been investigated to validate the model performances.

Keywords: Mood states, Feature extraction, Adaptive neuro-fuzzy inference, Root mean square error.

1 Introduction

The study of human mood states remains a complex domain of research. These states are often ill-defined, some states are overlapping, and many subsidiary states confuse machine learners in identification and detection. The complexity further increases when the perceived states are reflected through voice conversation rather than easily perceived facial images or gestures. Furthermore, if the affected person is out of sight while speaking on phone, then their modeling warrants robust signal processing techniques for adequate portrayal. It requires the extraction of a suitable, reliable feature set from speech samples describing these states to develop efficient and viable modeling.

[1]Department of ECE, YBN University, Jharkhand, India; E-mail: sahubiswajeet1985@gmail.com
[2]Department of ECE, Siksha “O“ Anusandhan, Bhubaneswar, India; E-mail: hemantapalo@soa.ac.in
[3]School of ECE, REVA University, Bengaluru, India; E-mail: shrotriya69@gmail.com

Among several feature extraction techniques, the frequency contents of the signal remain more informative and can provide vital clues to a person's mood state [1 – 5]. These features are extracted at the frame level, thus remaining high-dimensional, requiring large storage space and computation time. The spectral contents of a signal extracted over the entire duration may mitigate these issues. However, it leads to a distorted power spectrum with the noise spread over the signal duration. Henceforth, it is advisable to focus on the sub-band frequencies of a signal that contains relevant information on human mood states [6 – 8]. This motivates the authors on the modulation of spectral contents containing long-term spectro-temporal information for the intended analysis. The investigated features are the Spectral Centroid (SC), Spectral Flux (SF), Spectral Roll-off (SR), jitter, shimmer, fundamental frequency, and signal energy. The pitch represents the tonal and rhythmic properties of the speech excitation source. It changes due to tension, sub-glottal air pressure, and vibration in the vocal folds when the speaker is in different mood states [5]. Similarly, the energy indicates the arousal level when the speaker is under stress, hence varying among human-affected states.

The development of an identification system model based on the extracted feature vector has been an integral part of any recognition system. Several modeling systems such as the statistical approach, structural technique, template matching, neural network, fuzzy logic, hybrid algorithms, Gaussian mixture model, hidden Markov model, etc. have been explored successfully in the field of affective speech [9 – 11]. Among these, the fuzzy-based approaches can provide viable and excellent models in a real-world environment that are fuzzy [12 – 13]. The models can perform flexibly in an unpredicted, dynamic, and unknown environment similar to artificial intelligence with information on several possibilities that may occur. The ANFIS is a variant of the Fuzzy algorithm that combines the artificial neural network and the fuzzy inference system. As compared to other NNs, it does not require frame length normalization and can adequately reflect the temporal dynamics of the base features using the state transition probability [14 – 16]. The structure achieves generalization-using approximation and learns rapidly from the experimental data. For better certainty and precision, the generalization is accomplished based on the data dimension, which differs from network structures [16]. The models developed using ANFIS remain more user transparent than the NNs with low errors during memorization and are adaptable to nonlinear signals. They do not rely solely on expert knowledge during training and contain both linguistic and numerical knowledge. The speech samples of the mood states considered in this work are non-stationary, thus, the ANFIS structure suits our purpose.

This paper aims to develop suitable ANFIS models of several human mood states based on a few effective and discriminating feature vectors. The objective is to compare the model performance errors corresponding to the extracted feature vectors for their efficacy. To develop the desired ANFIS models, two different approaches have been proposed:

- **Approach-1:** All the seven extracted feature sets of a chosen mood state are used as inputs to the ANFIS structure. The intelligent ANFIS model is used to synthesize by exploring the terms of a human mood state based on the extracted feature sets to approximate the decision-making. It does not consider the quantitative terms in a mood state while modeling. Thus, it conveniently maps the input feature space corresponding to a certain state into an output shape. The necessary rules are formed to obtain the desired output while training and checking an ANFIS model. The procedure is adopted to train all the chosen feature sets to explore a certain mood in a multi-environment platform. Several model performances errors such as the training, testing, checking, error at start, error at convergence, minimal error, etc. are computed to determine the efficient model.
- **Approach-2:** A particular feature vector is chosen from each mood state and is fed as input to develop the ANFIS model of that state. The computed mood errors of each mood are compared based on the different feature vectors.

The rest of the article has been organized as follows. Section 2 briefs the backgrounds of the chosen feature extraction techniques. Section 3 describes the ANFIS algorithm whereas the proposed methodology is explained in Section 4. The discussion on the simulation results is provided in Section 5, while Section 6 concludes the work with possible future directions.

2 Background

A brief discussion on the chosen feature extraction techniques has been provided below.

Let $S_m(n)$, $n = 1, 2, \ldots, N_L$, denotes the sequence of speech mood samples at the m^{th} frame, with frame length N_L. The term N_L is also the *N*-point DFT of the analysed frame while computing the frequency-domain features. Let us denote Nf_L as the number of DFT coefficients chosen for the computation of the feature to follow.

The Spectral Flux (SF) provides information on the rate at which the power spectrum of the speech sample varies between successive frames. It can be found using the equation below

$$SF_s = \sum_{k=1}^{Nf_L} \left(ES'_m[k] - ES'_{m-1}[k] \right)^2 , \tag{1}$$

where the $S_m(k)$, $k = 1, 2, \ldots, Nf_L$, is the spectrum magnitude, and $ES'_m[k] = \frac{s_m[k]}{\sum_{l=1}^{Nf_L} s_m(l)}$ denote the k^{th} normalized spectrum at frame *m*, respectively.

The Spectral Centroid feature extraction algorithm to represent the speech sample of the human mood state is given below.

$$SC_s = \frac{\sum_{k=1}^{Nf_L} k \times s_m[k]}{\sum_{k=1}^{Nf_L} s_m(k)} . \tag{2}$$

The Spectral roll-off (SR) of a speech sample indicates the certain percentage below which, the spectrum of the signal resides corresponding to a designated frame. SR of the m^{th} frame corresponding to F^{th} DFT must obey the relation given by

$$\sum_{k=1}^{F} S_m(k) = 0.85 \sum_{k=1}^{Nf_L} S_m(k) . \tag{3}$$

The energy of a speech sample provides the arousal level of an affected state. The log energy has been considered to approximate the human hearing mechanism that is logarithmic. For a speech sample $s(n)$ the log-energy in decibel (dB) can be computed as

$$E_s = \log \sum_{k=-\infty}^{\infty} \left| s^2(n) \right| . \tag{4}$$

The pitch (F_0) varies among gender, age, and the human mood state. The children and females have higher F_0 as compared to males. The angry state has a higher arousal level than the sad state. Thus, F_0 can provide important cues on different mood states. There have been many techniques to extract the F_0 from a speech sample. However, the cepstrum method of F_0 extraction is based on the logarithm of the spectrum by approximating the human ear mechanism which is considered here. The cepstrum of a speech sample can be estimated as

$$F^{-1}\{\log[S(\omega)]\} = F^{-1}\{\log[U(\omega)]\} + F^{-1}\{\log[H(\omega)]\} , \tag{5}$$

where $H(\omega)$ and $U(\omega)$ are the spectrum corresponding to the vocal tract $h(n)$ and the excitation signal $u(n)$. As observed from (5), the cepstrum can easily demarcate the vocal source and vocal tract parameters of a signal, thus can estimate the F_0 accurately.

The absolute jitter can be found by observing the variation of F_0 within the cycles. It is estimated as

$$s_{jitter} = \frac{\frac{1}{M-1}\sum_{i=1}^{M-1}\left|T_i - T_{i+1}\right|}{\frac{1}{M}\sum_{i=1}^{M}\left|T_i\right|}, \tag{6}$$

where M indicates the number of extracted periods of the fundamental frequency (F_0) and T_i is the wavelength of F_0 [17].

The shimmer provides information on the variation in the peak-to-peak amplitude of a signal. It is estimated using the relation [17]

$$s_{shimmer} = \frac{\frac{1}{M-1}\sum_{i=1}^{M-1}\left|A_{i+1} - A_i\right|}{\frac{1}{M}\sum_{i=1}^{M}\left|A_i\right|}, \tag{7}$$

where A_{i+1} and A_i are the peak-to-peak amplitudes of the signal in consecutive periods respectively.

3 The ANFIS Algorithm

The ANFIS learning has been formalized on each input feature vector $\boldsymbol{x}$ as follows.

If $x(v_1)$ is A_j, $x(v_2)$ is B_j and $x(v_l)$ is C_j, then

$$Rules_j = a_j x(v_1) + b_j x(v_2) + \cdots + c_j x(v_l) + q_j,$$

where:

$x(v_1), x(v_2), \ldots, x(v_l)$ are the input features;

$A_j, B_j, \ldots$ are the fuzzy sets and

$a_j, b_j, \ldots, q_j$ are the design parameters based on the training process.

The ANFIS structure comprises five layers the fuzzy, product, normalized, de-fuzzy, and the total output each performing a certain function [14 – 16]. In the structure, layer 1 and layer 4 have adaptive nodes while layer 2, layer 3, and layer 5 have fixed nodes.

Let us consider a two-input x and y ANFIS structure with the output as z. Every node j in the fuzzy layer (layer 1) denotes a square node function. The layer 1 output is represented as

$$O_{1,j} = \mu_{A_j}(x), \tag{8}$$

where $\mu_{A_j}(x)$ denotes the Membership Function (MF) corresponding A_j.

The MF specifies the degree of linguistic labels such as maximum or minimum or average to which the chosen input *x* must satisfy quantifier A_j. A bell-shaped MF with a range of 0 to 1 has been used in this work and is given by

$$\mu_{A_j}(x) = \frac{1}{1+\left|\frac{v-c_j}{a_j}\right|^{2b_f}} . \tag{9}$$

The bell-shaped functions provide different MFs for the fuzzy set with a change in the premise parameters *a*, *b* and *c*. The output $O_{2,j}$ of layer 2 comprising of fixed circle nodes. This layer multiplies the input features and the output is given by

$$O_{2,j} = \mu_{A_j}(x) \times \mu_{B_j}(y) = w_j, \quad j = 1,2 , \tag{10}$$

where each node of w_j denotes the firing strength of the rule.

The fixed circle nodes of layer 3 are labeled as *N* which computes the i^{th} rule's firing strength based on the sum of the firing strength of all the rules as given by

$$O_{3,j} = \bar{w}_j = \frac{w_j}{w_1 + w_2}, \quad j = 1,2 , \tag{11}$$

where $O_{3,j} = \bar{w}_j$ denotes the output of the 3rd layer and denotes the normalized firing strength.

The layer 4 square nodes are adaptive and their weighted output is computed as linear functions with Sugeno inference coefficients m_j, n_j and p_j. The parameters of layer 4 are called consequent parameters.

$$O_{4,j} = \bar{w}_j f_j = \bar{w}_j \left(m_j v_1 + n_j v_2 + p_j \right) . \tag{12}$$

The overall output is computed in layer 5 having a single circle node and is given by

$$O_{5,j} = \sum_j \bar{w}_j f_j = \frac{\sum_j w_j f_j}{\sum_j w_j} . \tag{13}$$

This is the overall output of the ANFIS or the estimated output of the Sugeno FIS model. The hybridization of ANN and FIS is formalized to estimate the consequent and premise parameters. The consequent parameters are computed in this hybridized learning system in the forward pass in which the information propagates up to the 4th layer. In the process, a least square regression algorithm has been used to optimize these consequent parameters. The premises parameters are updated using a gradient descent algorithm when the error is propagated during the backward pass.

4 The Proposed Method

The proposed modelling of human mood state from voice samples is provided in Fig. 1.

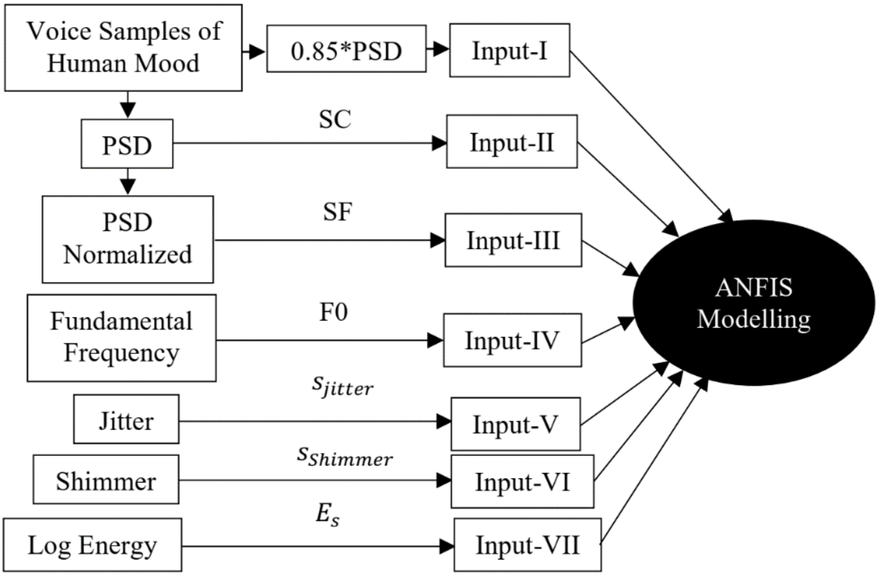


Fig. 1 – *The proposed ANFIS modeling of human mood states from voice samples.*

Initially, the voice samples of different mood states have been acquired from the SAVEE [18] dataset. Each sample is pre-processed using pre-emphasis filtering, normalization, and the mean subtraction stages. It helps to spectrally flatten and reduce the finite precision effects of the samples. The proposed ANFIS modeling of the human mood state is provided in Fig. 2.

Seven input feature sets are used to develop the desired ANFIS model of a chosen state. The algorithm is a universal estimator, which integrates both the ANN and Takagi–Sugeno-based FIS. Thus, it can capture the potential benefits of both in a single framework. The inference system follows a set of fuzzy

IF-THEN rules to learn and approximate nonlinear functions. For the chosen input feature sets, the membership functions (MFs) of the FIS are adjusted using the hybrid algorithm.

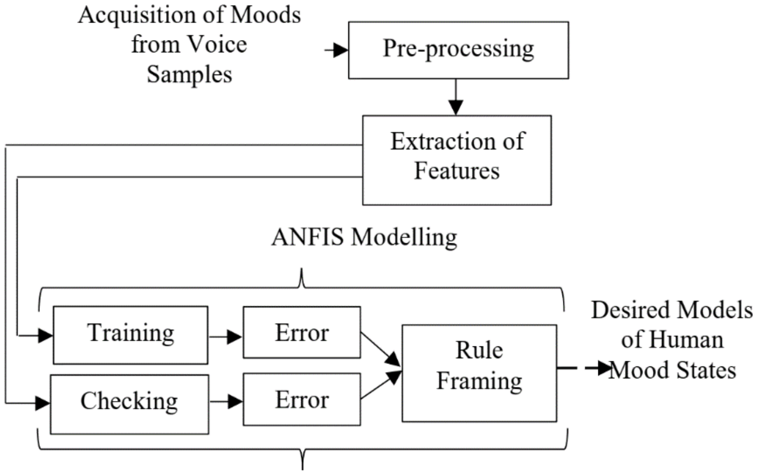


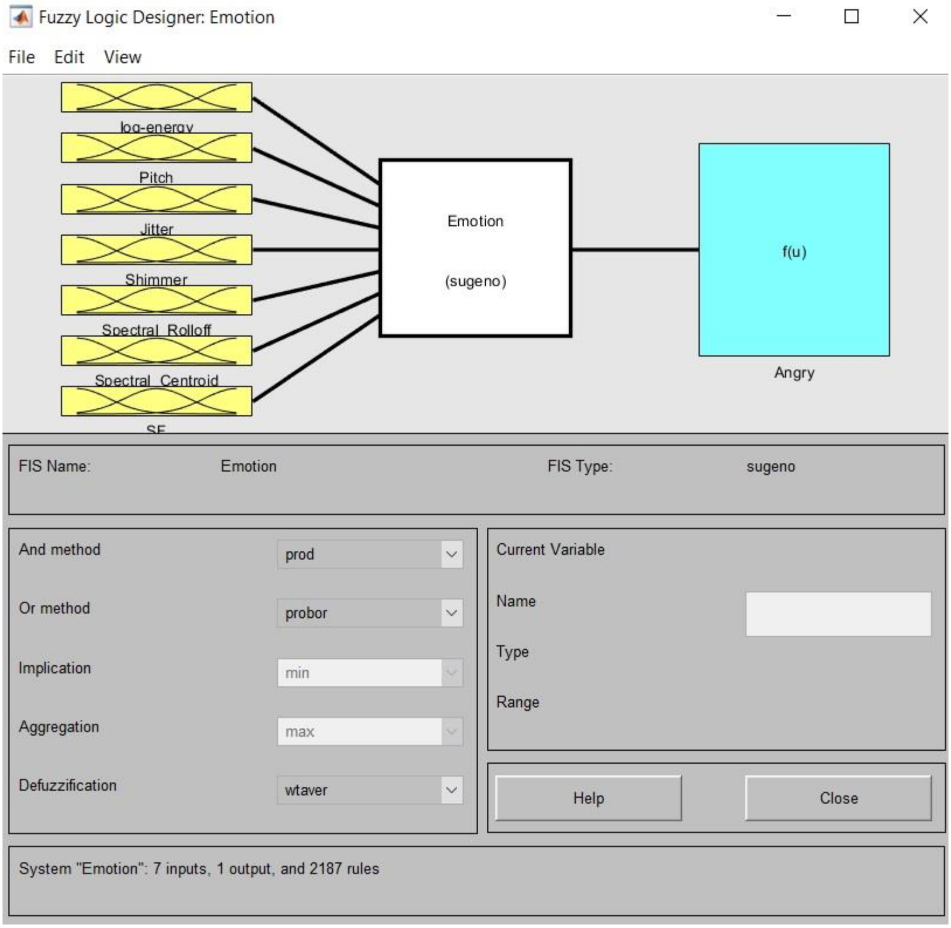
Fig. 2 – *Proposed ANFIS Models of Human Mood States.*

5 Results and Discussion

The SAVEE dataset chosen in this work is in the English language and comprises seven mood states. From, the dataset, sixty utterances of anger, happiness, and neutral states have been considered in this work. The samples are recorded at a sampling frequency of 44.1 kHz, which are down sampled to 16 kHz for convenience. To allow the ANFIS to learn from the input data and to obtain the consequent parameters, the structure of the ANFIS was designed initially. This has been done by determining the premise parameters and using subtractive clustering. Subsequently, the model is trained using a hybrid learning algorithm for 10 iterations and the training error is estimated. Finally, the system was tested to check the model performance.

The rows of the training data in an ANFIS model are the desired input-output pair of the given mood state. The output follows the input feature vector in each row of the training data. Thus, the number of rows contains the values corresponding to the number of chosen feature vectors. However, the number of training columns always becomes one more than the number of inputs.

Fig. 3 provides the ANFIS model of the angry state based on the seven chosen feature sets using the Sugeno FIS. The input feature vectors are SC, SF, SR, jitter, shimmer, log energy, and pitch. Each feature vector is designated using one of the three MFs as minimum, average, and maximum. There is only one model output representing the chosen state.



FIS Name: Emotion

FIS Type: sugeno

And method: prod

Or method: probor

Implication: min

Aggregation: max

Defuzzification: wtaver

Current Variable

Name

Type

Range

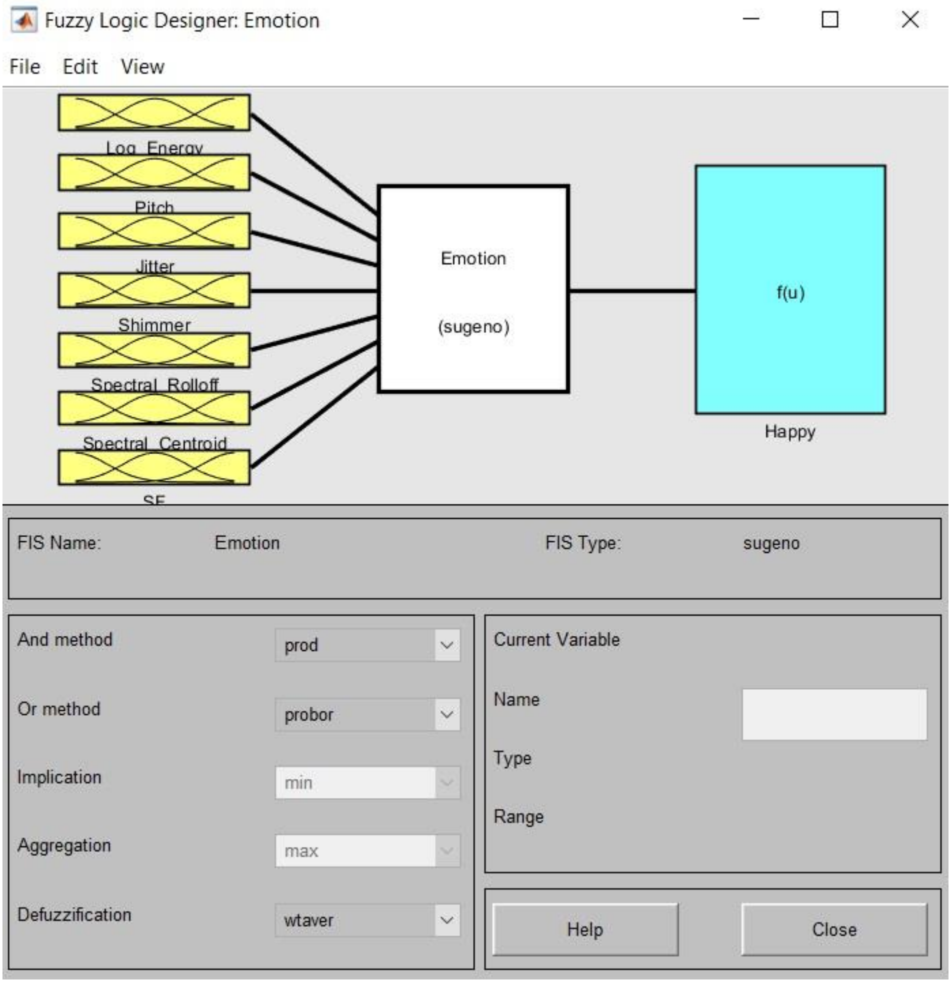
System "Emotion": 7 inputs, 1 output, and 2187 rules

Fig. 3 – *Modelling of the angry state using seven extracted feature sets.*

Similarly, Figs. 4 and 5 provide the ANFIS model of the Happy and Neutral states respectively. These figures validate the development of three chosen emotional models and are shown here for convenience.

The Root Mean Square Error (RMSE) is computed from these models during training, checking, and testing using the corresponding feature vectors as provided in Figs. 6, 7 and 8.

Fig. 4 compares the training Root Mean Square Error (RMSE) of the angry, state using different extracted feature sets based on the Sugeno FIS. The error is the difference between the output training data, and the FIS output corresponding to the same input training data i.e. one associated with the output training data. It is minimized for the training process accordingly at each epoch based on the error tolerance. The training stops using either a stopping criterion or when the network converges.



FIS Name: Emotion

FIS Type: sugeno

And method: prod

Or method: probor

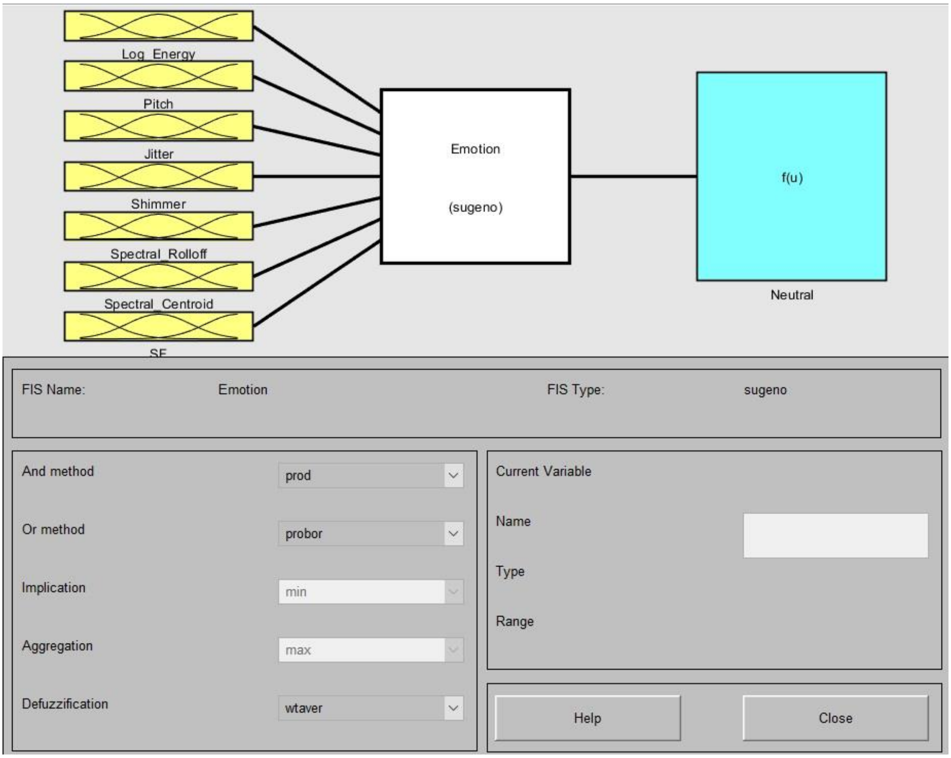
Implication: min

Aggregation: max

Defuzzification: wtaver

Fig. 4 – *Modelling of the happy state using seven extracted feature sets.*

Similarly, Figs. 7 and 8 compare the training Root Mean Square Error (RMSE) of the Happy and Neutral states using different extracted feature sets based on the Sugeno FIS for convenience.



FIS Name: Emotion

FIS Type: sugeno

And method: prod

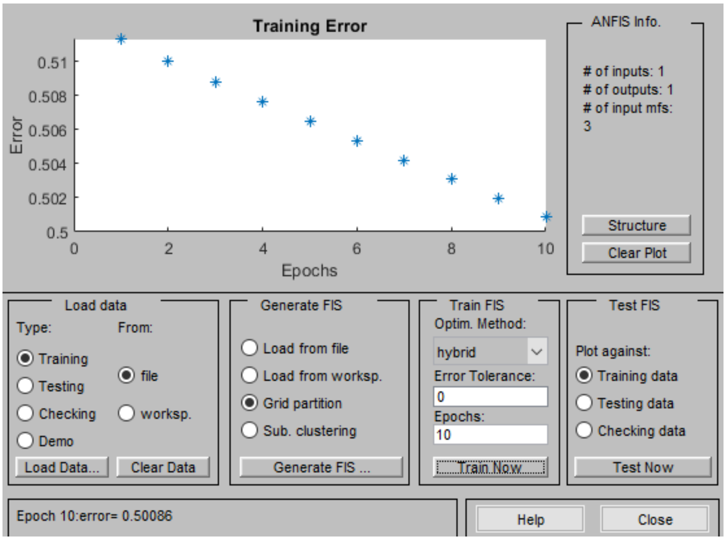
Or method: probor

Implication: min

Aggregation: max

Defuzzification: wtaver

Fig. 5 – *Modelling of the neutral states using seven extracted feature sets.*



# of inputs: 1
# of outputs: 1
# of input mfs:
3

Type: ● Training

From: ● file

● Grid partition

Optim. Method: hybrid

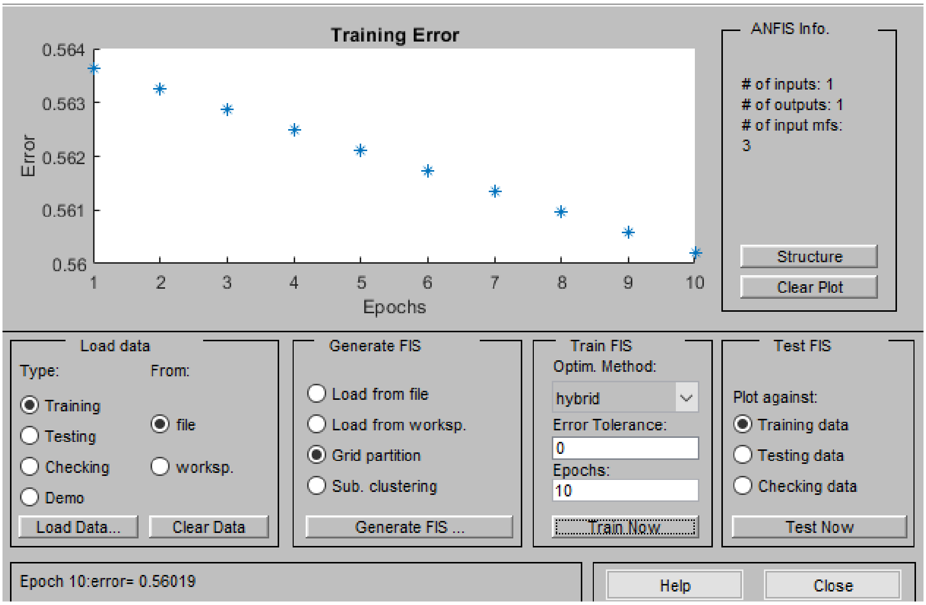
Error Tolerance: 0

Epochs: 10

Plot against: ● Training data

Epoch 10:error= 0.50086

Fig. 6 – *ANFIS training error (RMSE) corresponding to angry state using a single feature set.*



# of inputs: 1
# of outputs: 1
# of input mfs:
3

Type: ● Training ○ Testing ○ Checking ○ Demo

From: ● file ○ worksp.

○ Load from file ○ Load from worksp. ● Grid partition ○ Sub. clustering

Optim. Method: hybrid

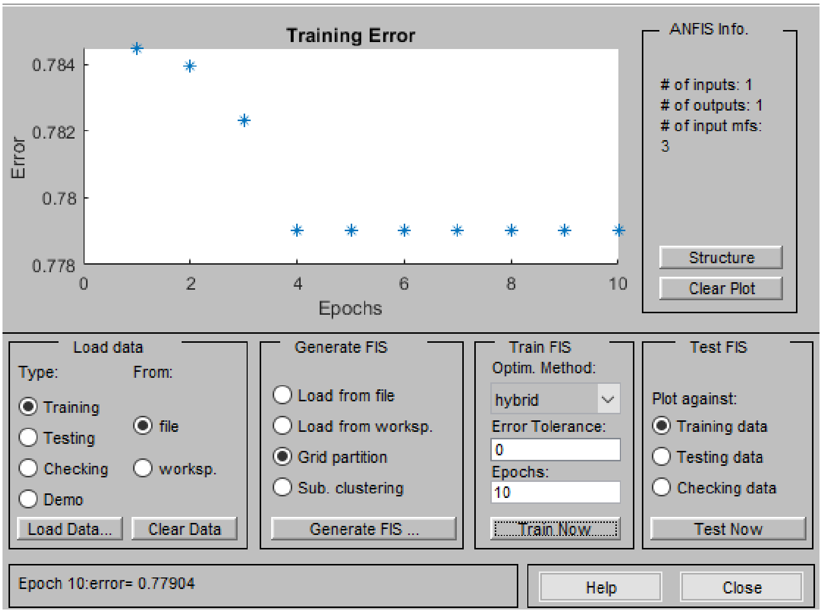
Error Tolerance: 0

Epochs: 10

Plot against: ● Training data ○ Testing data ○ Checking data

Epoch 10:error= 0.56019

Fig. 7 – *ANFIS training error (RMSE) corresponding to happy state using a single feature set.*



# of inputs: 1
# of outputs: 1
# of input mfs:
3

Type: ● Training ○ Testing ○ Checking ○ Demo

From: ● file ○ worksp.

○ Load from file ○ Load from worksp. ● Grid partition ○ Sub. clustering

Optim. Method: hybrid

Error Tolerance: 0

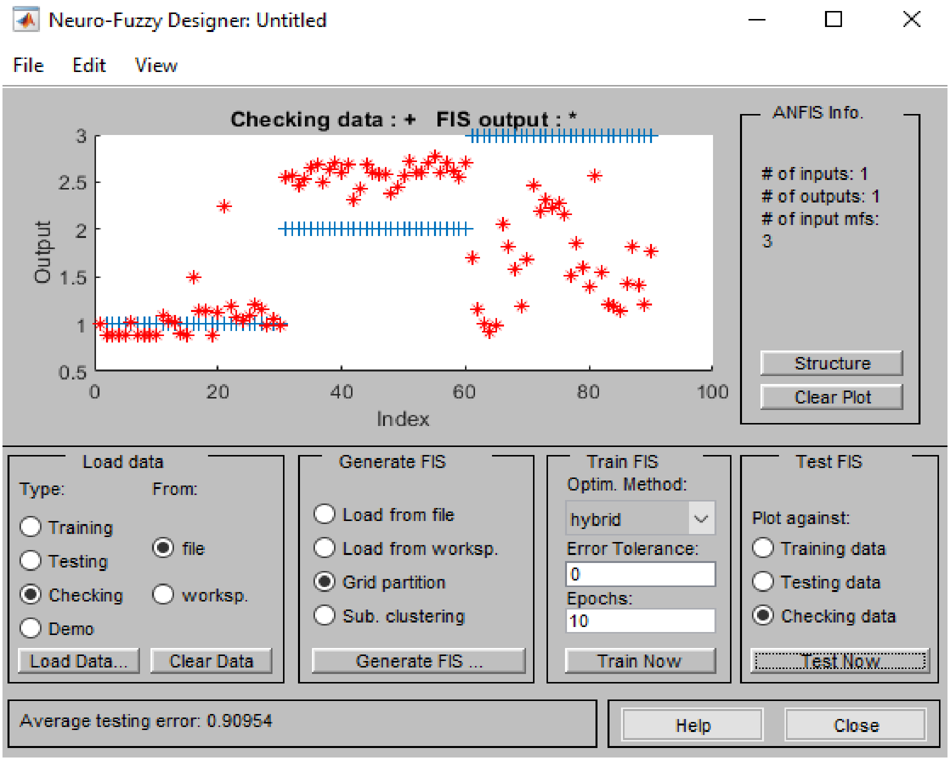
Epochs: 10

Plot against: ● Training data ○ Testing data ○ Checking data

Epoch 10:error= 0.77904

Fig. 8 – *ANFIS training error (RMSE) corresponding to neutral state using a single feature set.*

The ANFIS models provide fixed structures, hence may overfit the training data when the number of epochs is large. In this case, the structure cannot respond adequately to the independent feature sets chosen for the setup. To alleviate this issue, checking or validation of the data set is used. The checking RMSE error corresponding to extracted feature sets for the Angry state is shown in Fig. 9.



# of inputs: 1
# of outputs: 1
# of input mfs:
3

Type:
○ Training
○ Testing
● Checking
○ Demo

From:
● file
○ worksp.

○ Load from file
○ Load from worksp.
● Grid partition
○ Sub. clustering

Optim. Method:
hybrid

Error Tolerance:
0

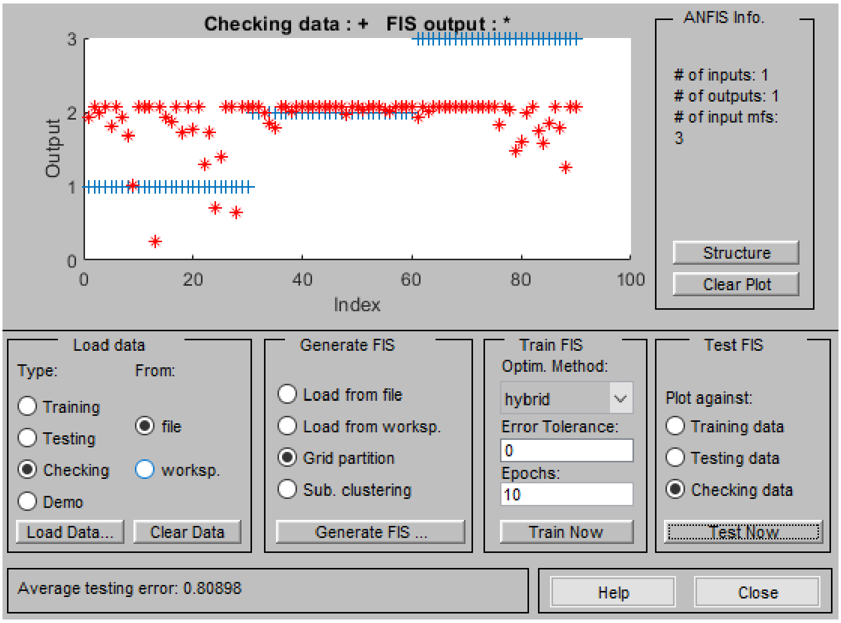
Epochs:
10

Plot against:
○ Training data
○ Testing data
● Checking data

Average testing error: 0.90954

Fig. 9 – *The ANFIS checking error (RMSE) using three input feature vectors for angry state.*

Similarly, the checking RMSE error corresponding to extracted feature sets for the Happy and Neutral states are shown in Figs. 10 and 11 respectively for convenience. From each mood state, approximately 25% of samples have been chosen randomly to check the ANFIS model. The error is the difference between the output of the checking data, and the output of the FIS corresponding to the same input checking data (the one associated with that output checking data). The RMSE is computed and minimized between the measured and modelled values at each epoch.



ANFIS Info.

# of inputs: 1
# of outputs: 1
# of input mfs:
3

Load data

Type: ○ Training ○ Testing ● Checking ○ Demo

From: ● file ○ worksp.

Generate FIS

○ Load from file
○ Load from worksp.
● Grid partition
○ Sub. clustering

Train FIS

Optim. Method: hybrid

Error Tolerance: 0

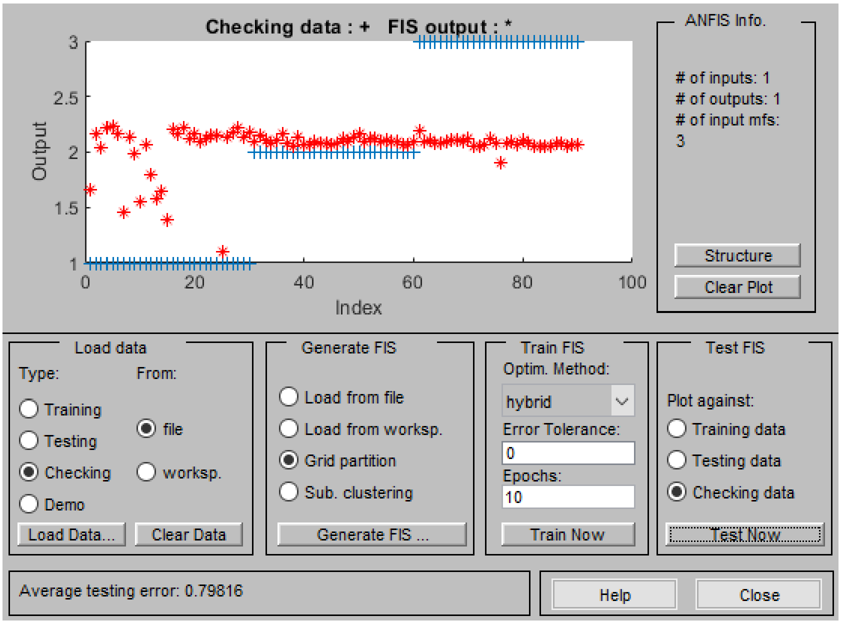
Epochs: 10

Test FIS

Plot against: ○ Training data ○ Testing data ● Checking data

Average testing error: 0.80898

Fig. 10 – *ANFIS checking error (RMSE) using three input feature vectors for happy state.*



ANFIS Info.

# of inputs: 1
# of outputs: 1
# of input mfs:
3

Load data

Type: ○ Training ○ Testing ● Checking ○ Demo

From: ● file ○ worksp.

Generate FIS

○ Load from file
○ Load from worksp.
● Grid partition
○ Sub. clustering

Train FIS

Optim. Method: hybrid

Error Tolerance: 0

Epochs: 10

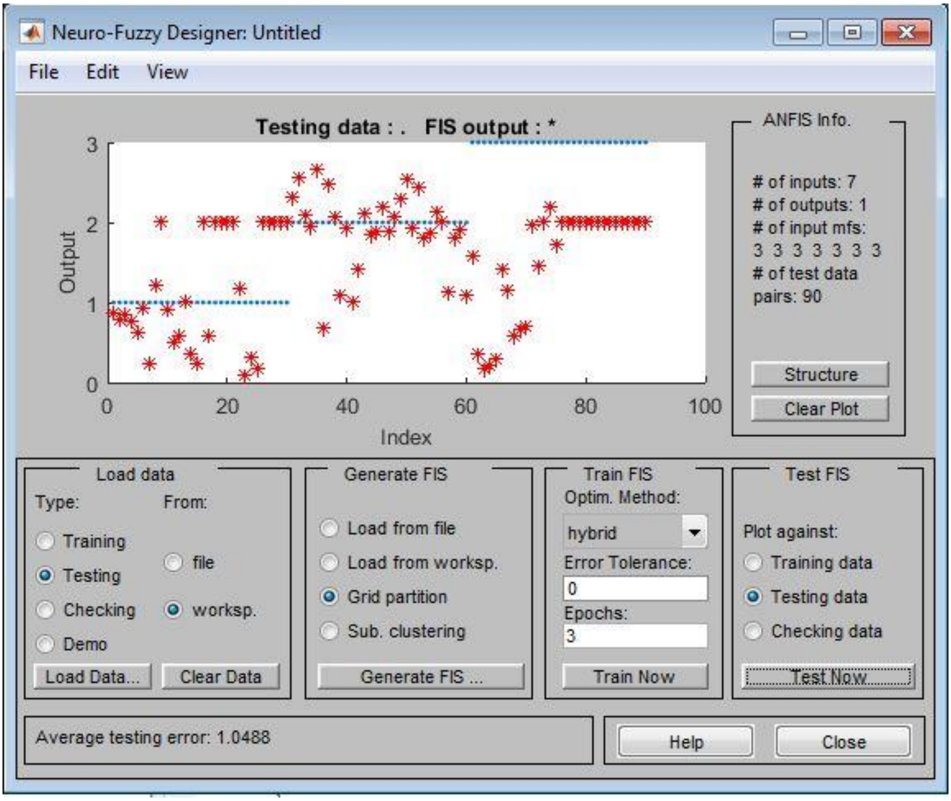
Test FIS

Plot against: ○ Training data ○ Testing data ● Checking data

Average testing error: 0.79816

Fig. 11 – *ANFIS checking error (RMSE) using three input feature vectors for neutral state.*

The testing RMSE error corresponding to three feature sets for the Angry state is shown in Fig. 12. It cross-validates the ANFIS structure to test the generalization ability of these models at each epoch. Ultimately, it indicates how efficiently the models can respond to the chosen checking data. Consequently, the MFs are optimized to minimize the checking error to tackle the overfitting issue.



Testing data : .   FIS output : *

ANFIS Info.

# of inputs: 7
# of outputs: 1
# of input mfs:
3 3 3 3 3 3 3
# of test data
pairs: 90

Load data

Type: ○ Training ● Testing ○ Checking ○ Demo

From: ○ file ● worksp.

Generate FIS

○ Load from file
○ Load from worksp.
● Grid partition
○ Sub. clustering

Train FIS

Optim. Method: hybrid

Error Tolerance: 0

Epochs: 3

Test FIS

Plot against: ○ Training data ● Testing data ○ Checking data

Average testing error: 1.0488

Fig. 12 – *ANFIS testing error (RMSE) for three input feature vectors for angry state.*

Fig. 13 provides the ANFIS structure of the angry mood state using seven input feature sets. Each input is represented by three MFs minimum, average, and maximum. The output is the desired mood state.

The rule viewer corresponding to the angry state using the seven-input feature set is shown in Fig. 14. The input and outputs of the rules can be viewed from this figure to investigate the crisp value of the chosen state. Similar rule viewers can be formed for other mood states.

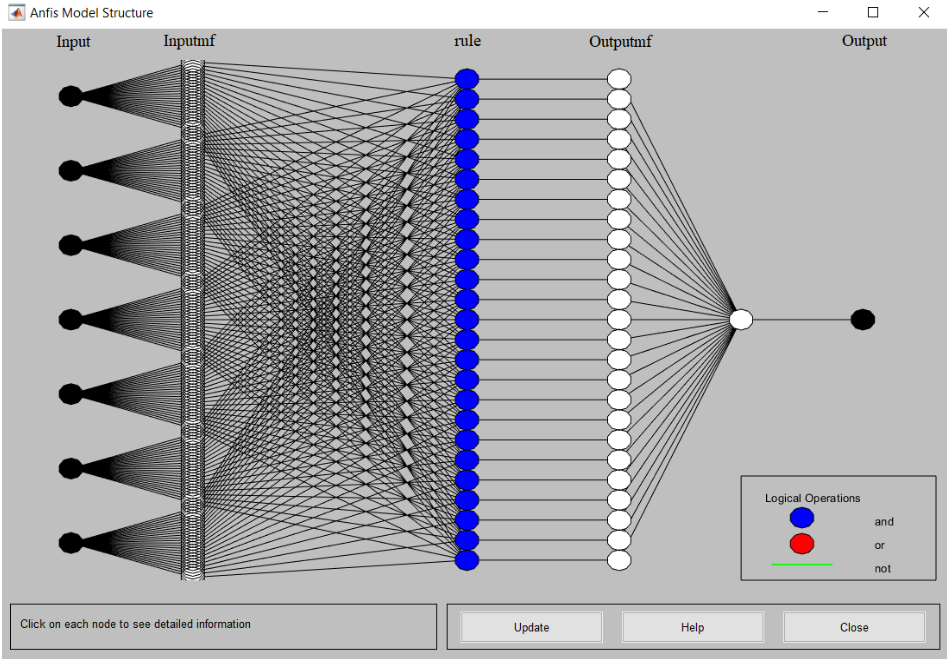


Fig. 13 – *ANFIS structure of the angry mood state using seven input feature sets.*

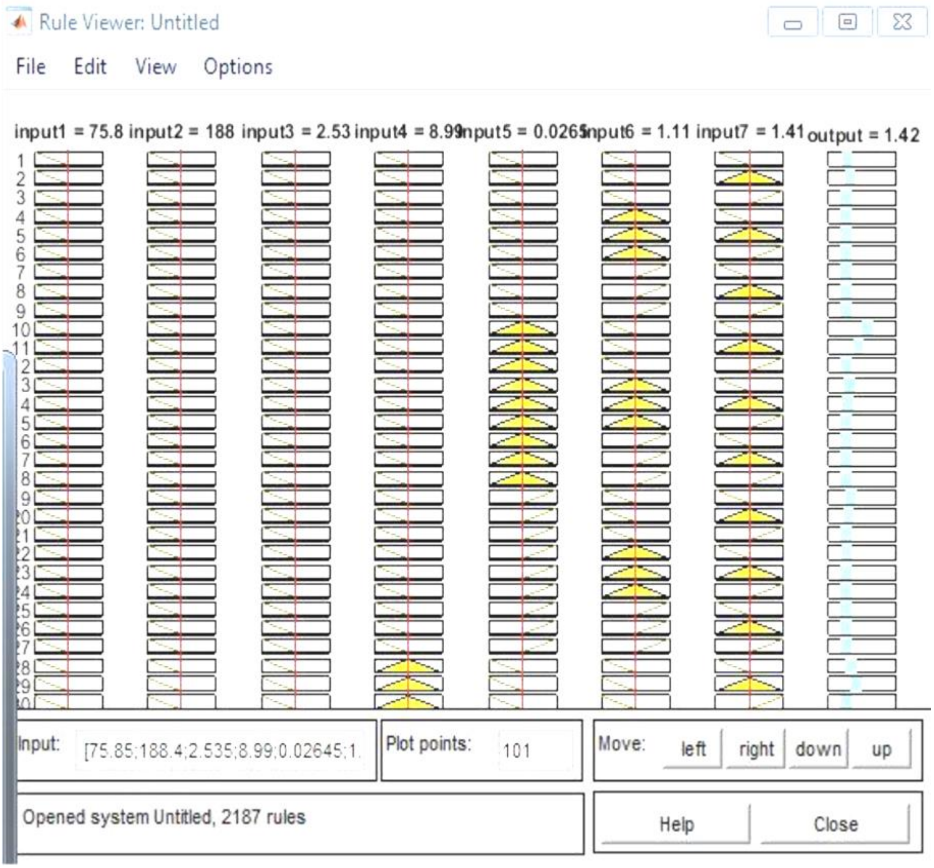
The ANFIS Simulation Parameters of a mood state using seven sets of Features are shown in **Table 1**.

Table 1

The ANFIS simulation parameters of angry state using seven feature sets.

Features	**Parameters**
Number of Nodes	4426
Number of Linear Parameters	2187
Number of Non-linear Parameters	63
Total Number of Parameters	2250
Number of Fuzzy Rules	2187

A comparison of the training and checking RMSE error among all the extracted feature sets has been made in **Table 2**. The training model error is lowest using the SC feature vector followed by the log-energy feature vector. The shimmer, jitter, and SF have experienced large training errors among the extracted feature vectors. Nevertheless, the mixture feature has shown the lowest model error, hence validating the efficacy of the proposed technique. From **Table 2** it can be observed that the training error is always less than the checking error irrespective of the chosen feature extraction techniques. It indicates the proposed ANFIS structure could model the desired mood states without overfitting.



input1 = 75.8 input2 = 188 input3 = 2.53 input4 = 8.99 input5 = 0.0265 input6 = 1.11 input7 = 1.41 output = 1.42

Input: [75.85;188.4;2.535;8.99;0.02645;1.

Plot points: 101

Opened system Untitled, 2187 rules

Fig. 14 – *ANFIS rule viewer with seven set of extracted input features corresponding to angry state.*

The following parameters have been trailed to develop the desired ANFIS models. The frequency to train the ANFIS model or the number of epochs considered is 5, 10, 14, 20, and 30. A higher number of epochs has resulted in larger training and checking time. There are two default learning algorithms such as hybrid and back-propagation to train the MFs while training. The hybrid method is a combination of least square and back-propagation which has provided the least RMSE, hence chosen for this work.

Table 2
Training errors of ANFIS models using the extracted feature sets.

Features	Training RMSE Error	Checking RMSE Error	Number of Epochs chosen	Number of Epochs taken to converge
SC	0.5005	0.9095	10	2
SF	0.8026	0.8035	10	2
SR	0.7871	0.8089	10	2
F0	0.7423	0.7981	10	2
Log-Energy	0.5598	0.7071	10	2
Jitter	0.8066	0.8496	10	2
Shimmer	0.8034	0.9329	10	4
Mixed Feature Vector	0.00010	1.0488	10	2

Table 3 compares the RMSE at start and convergence corresponding to different extracted feature sets. The rate of convergence is poor with SR features while it has been the best for the jitter features, indicating the modelling has been better with these techniques. Nevertheless, the RMSE has been the lowest when all the features are combined to model the chosen mood states. The reason may be due to the availability of more mood-relevant complementary information.

Table 3
ANFIS modelling information with different extracted feature sets.

Feature Extraction	ANFIS RMSE at Start	ANFIS RMSE at the convergence	ANFIS modelling Information
Log Energy	0.5605	0.5601	Number of nodes: 16 Optimization method: Hybrid Number of linear parameters: 3 Number of nonlinear parameters: 9 Number of fuzzy rules/ individual feature extraction techniques: 3
Pitch	0.7424	0.7423	
Jitter	0.8066	0.8066	
Shimmer	0.8035	0.8034	
SR	0.7790	0.7871	
SC	0.4996	0.5005	
SF	0.8028	0.8026	
Overall	0.00029	0.00010	

Table 4 provides different ANFIS modelling errors corresponding to the discussed human mood states. It also compares the RMSE at the start and convergence. It can be observed that the rate of convergence of the angry state is better followed by the happy state. These states have shown minimal RMSE and have high arousal levels with higher frequency components. Thus, the chosen

feature magnitudes of these states are more prominent, which makes the modelling easier as compared to the neutral state.

Table 4
Comparison of ANFIS Modelling Errors among the Different Mood States.

RMSE	**Mood States**		
	Angry	Neutral	Happy
ANFIS RMSE at Start	0.000013	0.000039	0.000034
ANFIS RMSE at the convergence	0.0000093	0.000031	0.000028
Minimal Training RMSE	0.000009	0.000032	0.000029
Average FIS output Error using Training Data	1	3	2
Average FIS output Error using Testing Data	0.94868	3	0
Average FIS output Error using Checking Data	0.63246	3	2

6 Conclusions

This piece of work attempts to investigate happy, angry, and neutral Mood states using an efficient soft computing approach. In this process, the ANFIS algorithm has been explored to model the chosen Mood states based on few spectral and speech quality features. The extracted feature sets have been compared based on the training and testing RMSE as well as the fuzzy-based rules for their efficient portrayal of the chosen Mood states. It can be inferred that the SC feature extraction technique has provided the lowest RMSE while that of the shimmer features has been the highest. However, the ANFIS simulation after combining all the features has indeed improved the modelling of the chosen states with the lowest RMSE as compared to any of the individual feature extraction techniques. Other efficient feature extraction techniques can be explored further for better modelling, which may provide a new future direction.

7 References

[1] S. Song, S. Jaiswal, L. Shen, M. Valstar: Spectral Representation of Behaviour Primitives for Depression Analysis, IEEE Transactions on Affective Computing, Vol. 13, No. 2, April-June 2022, pp. 829 – 844.

[2] K. Stsiampkouskaya, A. Joinson, L. Piwek, C.- P. Ahlbom: Emotional Responses to Likes and Comments Regulate Posting Frequency and Content Change Behaviour on Social Media: An Experimental Study and Mediation Model, Computers in Human Behavior, Vol. 124, November 2021, p. 106940.

[3] J. Zhang, Z. Yin, P. Chen, S. Nichele: Emotion Recognition Using Multi-Modal Data and Machine Learning Techniques: A Tutorial and Review, Information Fusion, Vol. 59, July 2020, pp. 103 – 126.

[4] B. Sahu, H. K. Palo, S. N. Mohanty: A Performance Evaluation of Machine Learning Algorithms for Emotion Recognition through Speech, Proceedings of the 8th International

Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, March 2021, pp. 13 – 17.

[5] H. K. Palo: The Effect of Age, Gender, and Arousal Level on Categorizing Human Affective States, Emotion and Information Processing – A Practical Approach, Edited by S. N. Mohanty, Springer, Cham, 2020.

[6] K. K. Paliwal: Spectral Subband Centroid Features for Speech Recognition, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seattle, USA, May 1998, pp. 617 – 620.

[7] S. Wu, T. H. Falk, W.- Y. Chan: Automatic Speech Emotion Recognition Using Modulation Spectral Features, Speech Communication, Vol. 53, No. 5, May-June 2011, pp. 768 – 785.

[8] J. Přibil, A. Přibilová: Evaluation of Influence of Spectral and Prosodic Features on GMM Classification of Czech and Slovak Emotional Speech, EURASIP Journal on Audio, Speech, and Music Processing, Vol. 2013, April 2013, p. 8.

[9] P. Mahajan: Applications of Pattern Recognition Algorithm in Health and Medicine: A Review, International Journal -of Engineering and Computer Science, Vol. 5, No. 5, May 2016, pp. 16580 – 16583.

[10] M. Deriche, A. H. Abo absa: A Two-Stage Hierarchical Bilingual Emotion Recognition System Using a Hidden Markov Model and Neural Networks, Arabian Journal for Science and Engineering, Vol. 42, No. 12, December 2017, pp. 5231 – 5249.

[11] H. K. Palo, M. Chandra, M. N. Mohanty: Emotion Recognition Using MLP and GMM for the Oriya Language, International Journal of Computational Vision and Robotics, Vol. 7, No. 4, July 2017, pp. 426 – 442.

[12] R. H. Abiyev, I. Günsel, N. Akkaya, E. Aytac, A. Çağman, S. Abizada: Robot Soccer Control Using Behaviour Trees and Fuzzy Logic, Procedia Computer Science, Vol. 102, 2016, pp. 477 – 484.

[13] R. Ram, H. K. Palo, M. N. Mohanty, L. Padma Suresh: Design of FIS-Based Model for Emotional Speech Recognition, Proceedings of the International Conference on Soft Computing Systems (ICSCS), Chennai, India, December 2015, pp. 77 – 88.

[14] W. Chen, X. Chen, J. Peng, M. Panahi, S. Lee: Landslide Susceptibility Modeling based on ANFIS with Teaching-Learning-Based Optimization and Satin Bowerbird Optimizer, Geoscience Frontiers, Vol. 12, No. 1, January 2021, pp. 93 – 107.

[15] S. Amid, T. M. Gundoshmian: Prediction of Output Energies for Broiler Production Using Linear Regression, ANN (MLP, RBF), and ANFIS Models, Environmental Progress & Sustainable Energy, Vol. 36, No. 2, March 2017, pp. 577 – 585.

[16] R. H. Abiyev, I. Günsel, N. Akkaya, E. Aytac, A. Çağman, S. Abizada: Robot Soccer Control Using Behaviour Trees and Fuzzy Logic, Procedia Computer Science, Vol. 102, 2016, pp. 477 – 484.

[17] S. Kanwal, S. Asghar, A. Hussain, A. Rafique: Identifying the Evidence of Speech Emotional Dialects Using Artificial Intelligence: A Cross-Cultural Study, PloS One, Vol. 17, No. 3, March 2022, p. e0265199.

[18] S. Haq, P. J. B. Jackson: Multimodal Emotion Recognition, Ch. 17, Machine Audition: Principles, Algorithms, and Systems, Edited by W. Wang, Information Science Reference, Hershey, New York, 2010.