

# Effective modelling of human expressive states from voice by adaptively tuning the neuro-fuzzy inference system

Surjyo Narayana Panigrahi<sup>1</sup>, Niharika Pattanaik<sup>2</sup>, Hemanta Kumar Palo<sup>2</sup>

<sup>1</sup>Department of Electronics and Communication Engineering, YBN University, Jharkhand, India

<sup>2</sup>Department of Electronics and Communication Engineering, Institute of Technical Education and Research, Siksha O Anusandhan (Deemed to be University), Bhubaneswar, India

## Article Info

### Article history:

Received Jul 30, 2022

Revised Feb 12, 2023

Accepted Mar 10, 2023

### Keywords:

Adaptive neuro-fuzzy inference

Feature extraction

Human expressive states

Modeling

Root mean square error

## ABSTRACT

This paper aims to develop efficient speech-expressive models using the adaptively tuning neuro-fuzzy inference system (ANFIS). The developed models differentiate a high-arousal happiness state from a low-arousal sadness state from the benchmark Berlin (EMODB) database. The proposed low-cost flexible developed algorithms are self-tunable and can address several vivid real-world issues such as home tutoring, banking, and finance sectors, criminal investigations, psychological studies, call centers, cognitive and biomedical sciences. The work develops the proposed structures by formulating several novel feature vectors comprising both time and frequency information. The features considered are pitch (F0), the standard deviation of pitch (SDF0), autocorrelation coefficient (AC), log-energy (E), jitter, shimmer, harmonic to noise ratio (HNR), spectral centroid (SC), spectral roll-off (SR), spectral flux (SF), and zero-crossing rate (ZCR). To alleviate the issues of the curse of dimensionality associated with the frame-level extraction, the features are extracted at the utterance level. Several performance parameters have been computed to validate the individual time and frequency models. Further, the ANFIS models are tested for their efficacy in a combinational platform. The chosen features are complementary and the augmented vectors have indeed shown improved performance with more available information as revealed by our results.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Hemanta Kumar Palo

Department of Electronics and Communication Engineering

Institute of Technical Education and Research Siksha O Anusandhan (Deemed to be University)

Bhubaneswar, Odisha, India

Email: hemantapalo@soa.ac.in

## 1. INTRODUCTION

Human expressive states are highly unpredictable, vague, overlapping, and ill-defined. Human being via facial expressions, gestures, and through voice modalities often manifests these states. Trivial things, such as watching a movie, listening to songs, meeting an old-time friend, a pleasant scent, and seeing a funeral pyre can change our expressive states. The study remains a complex domain of research, particularly when these states are expressed via phone [1]. It requires effective signal-processing tools to adequately represent them that can benefit several fields such as artificial intelligence, cognitive sciences, psychological studies, criminal investigation and humanoid robotics. The tools and techniques must be capable of extracting discriminant and relevant voice parameters appropriately representing human expressive states for efficient modeling [2].

Among several techniques, the community often relies on the prominent features extracted either at the frame level or at the utterance level. The frame-level analysis facilitates studying the signal in a stationary

platform, however, results in high-dimensional data comprising redundant information, thus increasing the training time and memory space [3], [4]. The utterance level extraction of the speech features can alleviate these issues with improved accuracy [5]. Modeling algorithms often play a crucial role and remain an indispensable module to develop an effective recognition system. Earlier Learning algorithms applied in this field belong to neural networks, structural techniques, clustering approaches, Hidden Markov model, Gaussian mixture model and support vector machines (SVM). with excellent outcomes [6]–[13]. The spectrograms of speech emotions along with the squeeze and excitation residual neural network (ResNet) and a trainable discriminative ghost vector of locally aggregated descriptors (GhostVLAD) clustering layer extended convolution neural network (CNN) have been applied to extract a low-dimensional utterance-level feature vector [5]. Simulation on crowd sourced emotional multimodal actors dataset (CREMA-D), Ryerson audio-visual database of emotional speech, and song (RAVDESS) using the developed model has provided a global accuracy of 83.35% and 64.92% respectively. The combination of empirical mode decomposition and the Teager-Kaiser Energy Operator time-frequency and cepstral features has provided improved expressive state models than stand-alone feature vectors. The authors have reported an accuracy of 91.16% and 86.22% respectively using the recurrent neural network (RNN) and SVM in modelling the chosen speech expressive states [6]. The hybridization of the prosodic, cepstral, spectrum, and wavelet features has enhanced the modelling capability of SVM in recognizing Arabic expressive states [7]. These pieces of literature reveal that the hybridization of features and their effective selection often lead to enhanced models due to more available information, however not without limitations [1]–[4], [6]–[13]. However, the judicious selection of relevant features bearing complementary information challenges the community, hence motivating the authors.

Several issues that are inherently associated with hybrid models are the requirement of a large pool of samples, response time, selection of hyper-parameters, kernel function, the number of hidden layers, nodes, feature dimension, addressing the non-linear relationship among extracted parameters, regularization, generalization and issues of overfitting [12]. Henceforth, the application of fuzzy-based approaches along with global statistics seems a novel ideal to explore in real-world fuzzy environments. These models remain flexible, providing several feasible solutions besides performing in dynamic, unpredicted, and vague environments. Unlike other neural networks, the neuro-fuzzy inference systems when adaptively tuned lead to an outcome-based adaptive neuro-fuzzy inference system (ANFIS) structure that does not require frame length normalization. The application of state transition probability in ANFIS facilitates the representation of temporal dynamics associated with the baseline parameters. The structure rapidly learns from the experimental data with precision and certainty. The use of approximation while generalizing the network and lower errors than the conventional NNs during memorization makes it versatile [14]–[16]. User transparency, adaptability to nonlinear signals, faster training without expert knowledge, and representability of numerical and linguistic information make the algorithm superior [17], thus providing the desired platform for this work.

In this paper, the authors attempt to model a few chosen speech expressive states using effective features in section 2 and the ANFIS in section 3. The model can arguably limit the aforementioned issues by representing the expressive state using both numerical and linguistic knowledge. It makes the model more transparent and user-friendly due to low memorization errors. Finally, the adaptation capability, faster learning, and nonlinear ability, of the developed model add value to the recognition mechanism. Section 4 validates and discusses the developed models using the derived feature vectors considering three proposed instances whereas section 5 concludes the work with a few possible future directions.

## 2. THE PROPOSED APPROACH

The Berlin emotional speech database (EMODB) dataset chosen in this work comprises seven expressive states such as anger, happiness, boredom, anxiety, sadness, disgust, and neutral sampled at a rate of 48 kHz and are down-sampled to 16 kHz for convenience. It has been a widely accessed database used to analyze speech emotion (SE) states, which makes the comparing platform uniform [18]–[22]. From this database, this work compares the three expressive models based on the level of arousal. These states are happiness (high-arousal), sadness (low-arousal), and neutral. Forty-five utterances of each state are used to extract the chosen feature vectors and for further processing. Initially, the ANFIS structure is developed so that it can learn from the extracted input feature vector and compute the consequent parameters by estimating the premise parameters using subtractive clustering. The hybrid learning algorithm is used to train the ANFIS structure based on the premise parameters for 10 iterations. Finally, the developed structure is tested to validate its performance. The proposed ANFIS combination Model, shown in Figure 1 concatenates both the time and frequency-domain utterance-level statistical feature vectors to develop the desired ANFIS model for each expressive state.

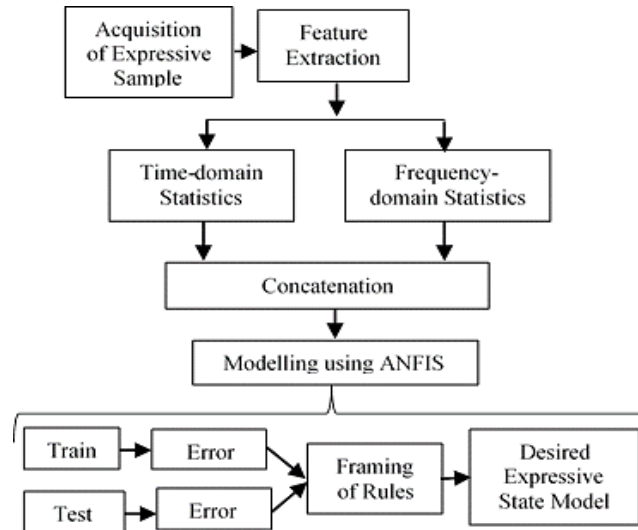


Figure 1. The proposed ANFIS modelling

The utterance-level statistics extracted from the frame-level features of a signal are the mean, range, standard deviation, skewness, and kurtosis. The proposed frequency-domain ANFIS model is shown in Figure 2. It considers five feature vectors spectral rolloff (SR), spectral flux (SF), spectral centroid (SC), fundamental frequency (F0), and standard deviation of F0 (SF0). Each speech sample corresponding to the chosen expressive state is pre-emphasized, normalized, and mean subtracted to spectrally flatten and reduce the finite precision effects [23]–[25]. The proposed time-domain model is shown in Figure 3. It considers six feature vectors such as the normalized log-energy, zero crossing rate (ZCR), jitter, Shimmer, auto-correlation coefficients (AC), and harmonic noise ratio (HNR). The necessary rule base is formed during the ANFIS training and testing to fetch the desired output. The objective is to develop an expressive model that can easily adapt to the multi-environment scenario.

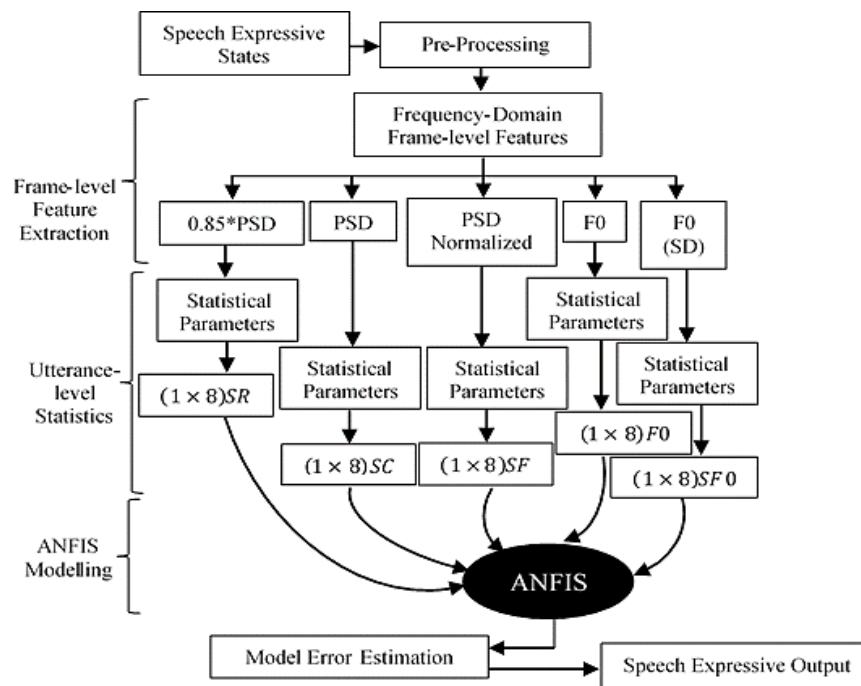


Figure 2. The proposed frequency-domain ANFIS model

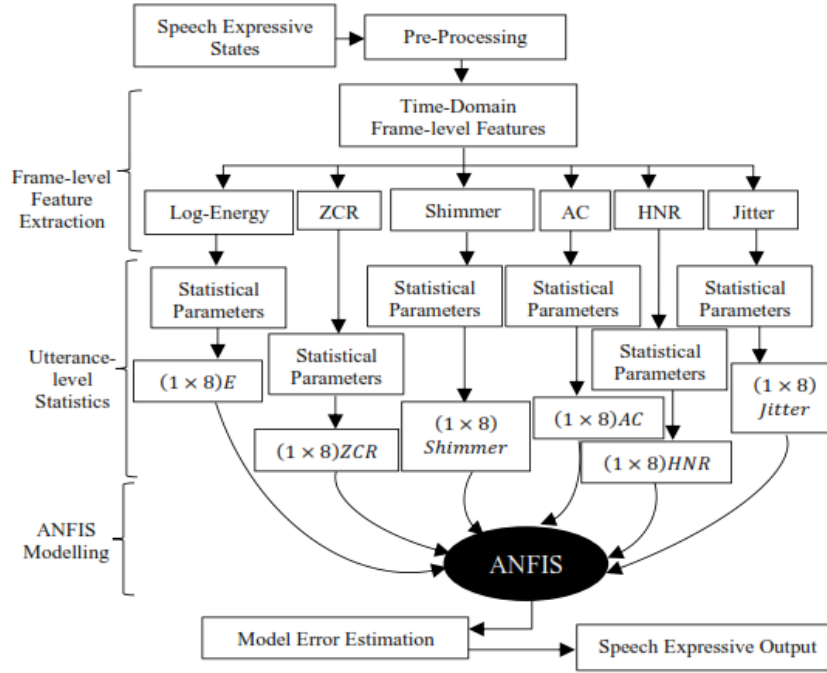


Figure 3. The proposed time-domain ANFIS model

### 3. THE ANFIS ALGORITHM

The ANFIS integrates the adaptive neural network (ANN) and fuzzy inference system (FIS) algorithms to determine the model parameters using fuzzy if-then rules and appropriate membership functions (MF) [26], [27]. There are five layers in this structure comprising adaptive nodes in layer-1 (fuzzy layer) and layer-4 (de-fuzzy layer) whereas layers-2 (product), 3 (normalized), and 5 (total output) have fixed nodes each staging a particular function. The rule can be formed corresponding to each extracted feature  $x$  as if  $x(u_1)$  is  $G_i$ ,  $x(u_2)$  is  $H_i$ , and  $x(u_l)$  is  $I_i$ , then  $Rules_i = p_i x(u_1) + q_i x(u_2) + \dots + r_i x(u_l) + a_i$ , where  $x(u_1), x(u_2), \dots, x(u_l)$  is the input features. The terms  $G_i, H_i, \dots$  represent the fuzzy sets, and the terms  $p_i, q_i, \dots$  are the design parameters estimated while training the structure.

The output of layer-1 considering  $x, y$  as inputs, and  $z$  as output is given by  $O_{1,i} = \mu_{G_i}(\cdot)$ , where  $\mu_{G_i}(\cdot)$  is the MF representing the inputs  $x$  or  $y$  corresponding to  $G_i$ . The MF assigns linguistic labels such as low or high or medium to specify the feature values of an input vector to quantify  $G_i$ . The popular bell-shaped MF having values between zero and one is chosen here and is represented for input  $x$  as

$$\mu_{G_i}(x) = \frac{1}{1 + |(u - r_i)/p_i|^{2q_i}} \quad (1)$$

By varying the premise parameters  $p, q$ , and  $r$ , it is possible to accommodate several MFs representing the fuzzy set. The layer-2 having fixed circle nodes multiplies the extracted input features of vectors  $x, y, \dots$ . The layer-3 fixed circle nodes estimate the  $i^{th}$  rule's firing strength using the firing strength of all the rules and provide an output  $O_{3,i}$  with normalized firing strength. The weights of adaptive layer-4 square nodes are estimated as linear functions with Sugeno inference coefficients  $m_i, n_i$ , and  $s_i$ . The output of layer 2,  $O_{2,i}$  and the output of layer 3,  $O_{3,i}$  with  $v_i$  as the firing strength of the rule are given in (2) and (3) respectively whereas the layer-4 provides the consequent parameters and its weighted output is described by (4). Similarly, the single circle layer-5 node provides the overall or the estimated Sugeno FIS model output and is given by (5). In this, the hybridized ANN and FIS compute the consequent parameters in the forward pass by propagating the information up to the fourth layer and optimizing the parameters using a least square regression algorithm. However, a gradient descent algorithm optimizes the parameters of the premises.

$$O_{2,i} = \mu_{G_i}(x) \times \mu_{H_i}(y) \times \dots = v_i, i = 1, 2, 3, \dots \quad (2)$$

$$O_{3,i} = \bar{v}_i = \frac{v_i}{v_1 + v_2 + \dots}, j = 1, 2, \dots \quad (3)$$

$$O_{4,i} = \bar{v}_i f_i = \bar{v}_i (m_i w_1 + n_i w_2 + \dots + s_i) \quad (4)$$

$$O_{5,i} = \sum_i \bar{v}_i f_i = \frac{\sum_i v_i f_i}{\sum_i v_i} \quad (5)$$

#### 4. THE RESULTS AND DISCUSSION

The simulation results using the extracted time and frequency domain features with the ANFIS structure is provided in this section to validate the proposed work. The work initially develops the time and frequency domain ANFIS structures. Further, the root mean square error (RMSE) while developing the ANFIS models for different states of emotions is graphically shown for comparison. Finally, the ANFIS model using both the time and frequency domain features has been developed and the error has been computed to validate the efficacy of the combined model.

Figure 4 provides the frequency domain ANFIS structure comprising five inputs such as SR, SF, SC, fundamental frequency (F0), the standard deviation of F0 (SF0), and one of the desired states as the output. The training rows constitute the desired input-output pair of an individual expressive state while developing the desired model of that statement using a set of chosen feature vectors. A similar time-domain ANFIS structure has been developed using six inputs such as log-energy, zero crossing rate (ZCR), jitter, Shimmer, auto-correlation coefficients, and harmonic noise ratio in Figure 5. The frequency and time-domain rules can be viewed from the ANFIS rule viewer for the chosen states. The Frequency-domain rule viewer in,

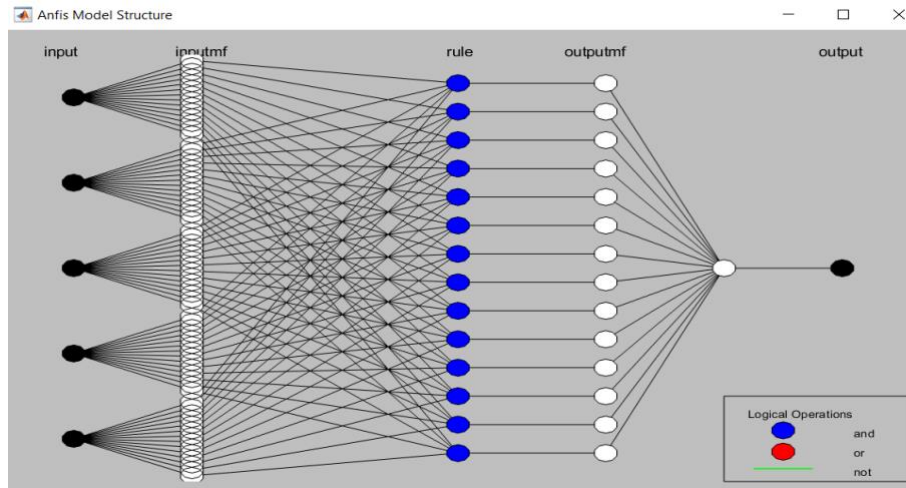


Figure 4. The frequency domain ANFIS structure

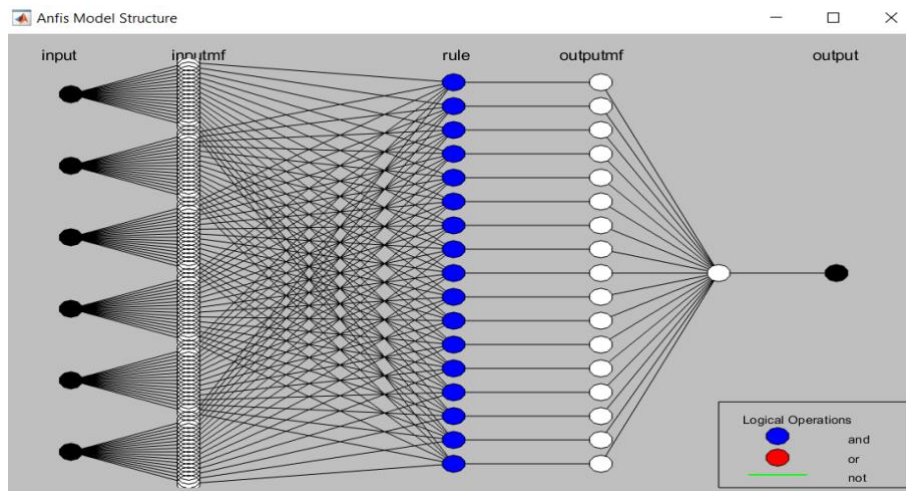


Figure 5. The time domain ANFIS structure

Figure 6 comprises the input and output rules for the state of happiness. The rule viewers can be developed similarly for the sadness and neutral states using the five frequency-domain inputs and six time-domain inputs. The rule viewer helps to investigate the crisp value of each state based on the inputs. Similarly, the time-domain rule viewer with six inputs is developed the intelligent model synthesizes all the crisp input terms describing the chosen expressive state while approximating the decision-making process. The model utilizes the pre-defined membership functions instead of the quantitative terms of the features to map the input feature vectors to the chosen output shape for such a purpose.

Figure 7 graphically analyses the training RMSE corresponding to the Happiness states using frequency-domain feature vectors. The RMSE is estimated using ten epochs and is the difference between the training output, and the FIS output. At each training epoch, the minimization of the error takes place to develop the desired ANFIS model. The training however stops when the network converges or uses a stopping criterion. At each epoch, the error between the measured and modeled values is estimated and minimized until the network converges. The RMSE compared to that of the low-arousal sadness and the neutral state using frequency-domain feature vectors can be observed similarly. It is found to be 0.64837, 0.67807, and 0.68863 corresponding to happiness, sadness, and neutral states respectively. It shows the suitability of ANFIS in modeling the high-arousal Happiness state as compared to the low-arousal Sadness state.

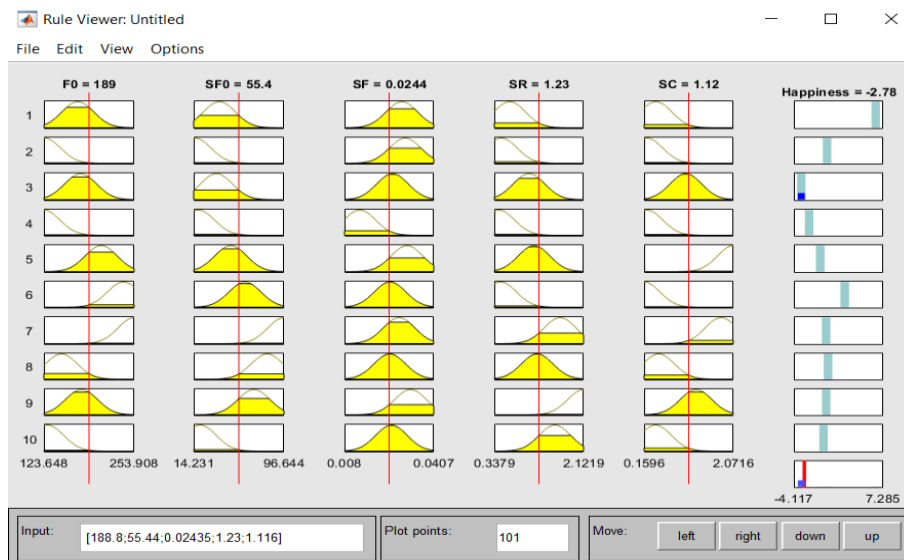


Figure 6. The frequency-domain rule viewer

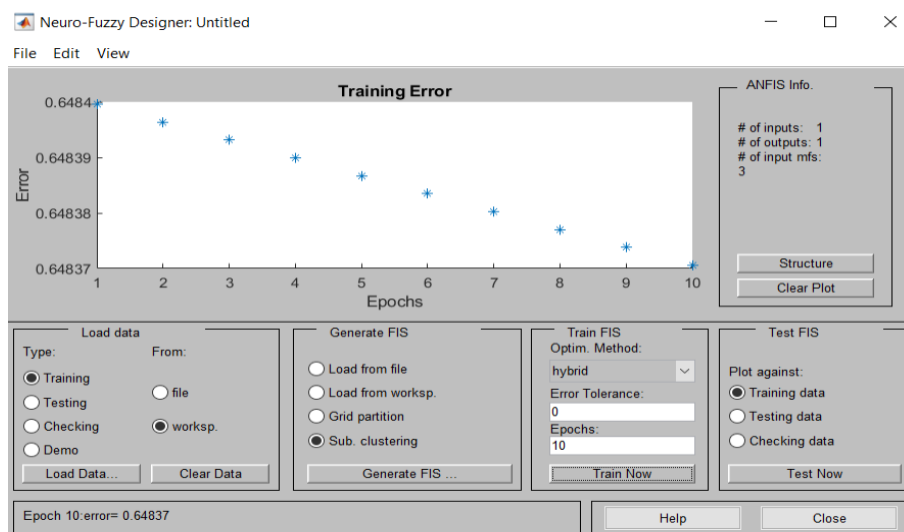


Figure 7. The ANFIS error (RMSE) for happiness using frequency-domain feature vector

Figure 8 graphically analyses the training RMSE corresponding to the Happiness states using time-domain feature vectors. The RMSE compared to that of the low-arousal sadness and the neutral state using can be observed similarly. It is found to be 1.02, 1.1522, and 1.5327 corresponding to happiness, sadness, and neutral states respectively. Figure 9 graphically analyses the training RMSE corresponding to the Happiness states using the combined time-frequency feature vectors. The RMSE is found to be 0.3868, 0.58327, and 0.7896 corresponding to happiness, sadness, and neutral states respectively. The frequency domain features are more informative, hence providing better modeling with lower RMSE than time-domain models. Nevertheless, the combinational model has outperformed the individual models due to more emotionally relevant available information as observed in Figure 7 through Figure 9. Figure 10 provides the testing RMSE for the Sadness state using frequency-domain feature vectors. It resolves the issues of overfitting by optimizing the MFs. The testing error also cross-validates the generated ANFIS models by testing their generalization ability at each epoch. It shows how effectively the ANFIS models behave to the extracted testing feature vectors.

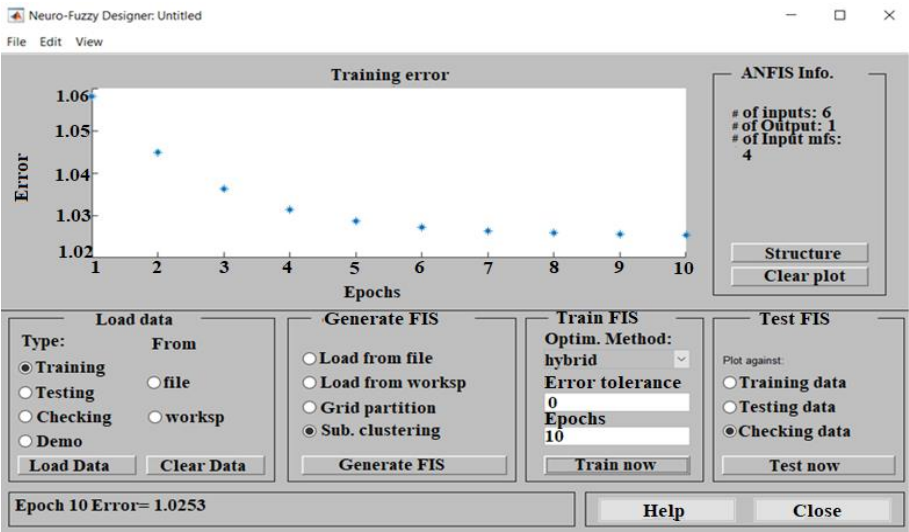


Figure 8. The ANFIS error (RMSE) for happiness using time-domain feature vector

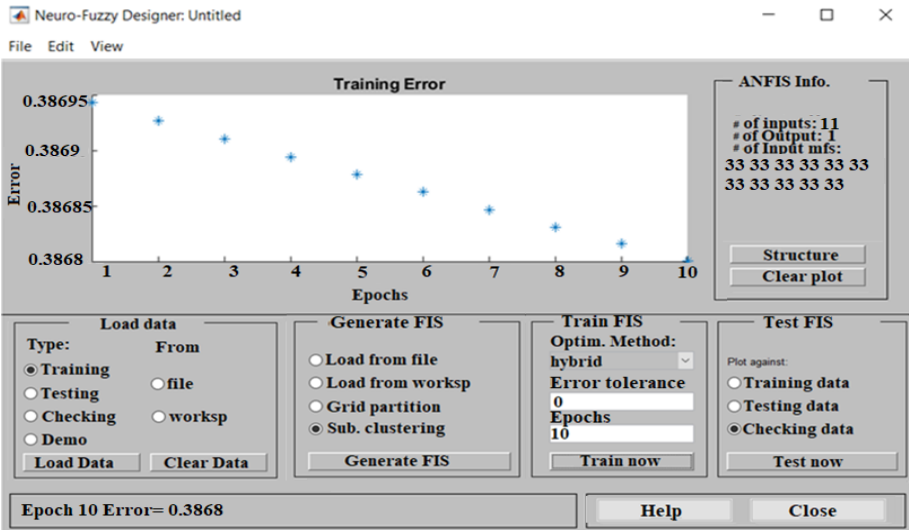


Figure 9. The ANFIS error (RMSE) for happiness using time-frequency-domain feature vector

A comparison of ANFIS performances with frequency and time-domain feature vectors is shown in Table 1. It shows that the checking and testing error is always higher than the training error, however, the

difference is meager, and the chosen expressive states can be materialized without overfitting. The time and frequency-domain ANFIS models are validated using several performance parameters including the RMSE at the start, at convergence, training, testing, and checking. The models of each expressive state have been trailed using four, eight, ten, fifteen, and twenty epochs to minimize the RMSE. With an increase in the number of epochs, the time to train, check and test the network has increased exponentially. The training error is reduced due to extensive learning; however, the testing error has increased as a trade-off due to overfitting with poor network generalization. On the contrary, with a small number of epochs, underfitting occurs due to inadequate learning. The network has provided the optimum performance with ten epochs, hence chosen here. Among the default FIS hybrid and back-propagation learning algorithms, the hybridization of the back-propagation and least square has witnessed the lowest RMSE in all the chosen cases, hence is considered. It has been observed that the ANFIS models of different expressive states using time-domain feature vectors have experienced higher RMSE in all the cases as compared to the frequency-domain feature vectors due to less relevant information as revealed in Table 1.

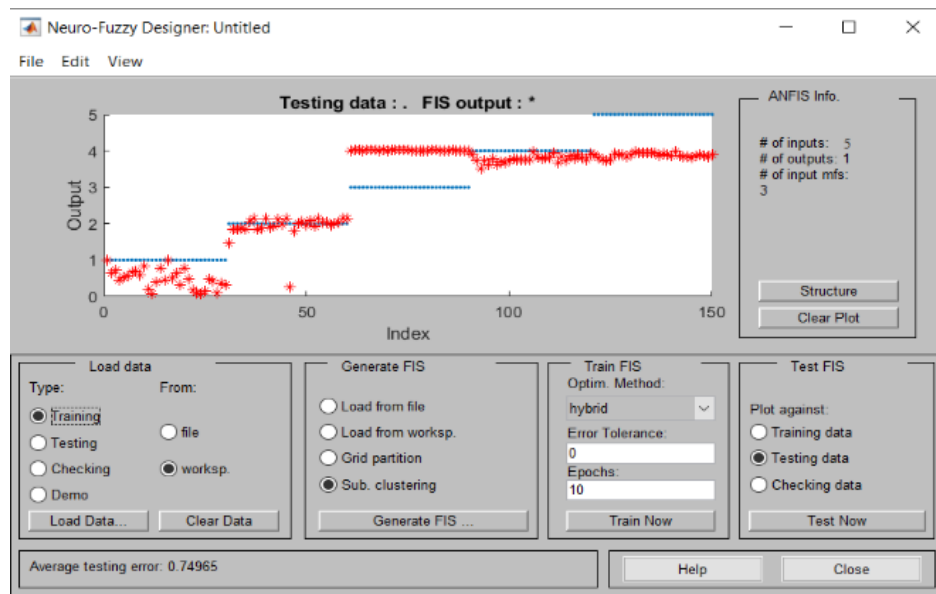


Figure 10. The testing RMSE for the sadness state using frequency-domain feature vectors

Table 1. Comparison of ANFIS performance parameters using frequency and time-domain feature vectors

Performance Parameters	Happiness		Sadness		Neutral	
	Frequency	Time	Frequency	Time	Frequency	Time
E1	0.64837	1.0253	0.67807	0.78967	0.68663	1.5327
E2	0.68441	1.2464	0.99422	0.14351	0.74127	1.6922
E3	0.74964	2.4167	0.99233	2.0197	0.73034	3.2346
E4	0.678071	1.27306	0.686631	1.15217	0.648371	8.62407
E5	0.686626	2.3666	0.686626	2.53514	0.648367	2.86288
R	10	17	14	21	13	25
N	176	247	128	303	164	359
L1	84	119	60	147	78	175
L2	140	204	100	252	130	300
L=L1+L2	224	323	160	399	208	475
I2	2	4	3	6	3	7
Clustering algorithms	Default Subtractive clustering parameters					
	– Range of Influence: 0.5					
	– Reject ratio: 0.15					
	– Squash factor: 1.25					
	Accept ratio: 0.5					
M	3 (High, Low, and Medium)					
I1	10					

E1: Average FIS training output error, E2: Average FIS checking output error, E3: Average FIS testing output error, E4: ANFIS error at start, E5: ANFIS error at the convergence, R: Number of rules, N: Number of nodes, L1: Number of linear parameters, L2: Number of nonlinear parameters, M: Number of inputs MFs, I1: Number of epochs considered, I2: Number of epochs for convergence

## 5. CONCLUSION

This piece of work attempts to investigate happiness, sadness, and neutral expressive states using an efficient soft computing approach. In this process, the ANFIS algorithm has been explored to model the chosen expressive states based on a few of the efficient time and frequency-domain utterance level features. Further, the ANFIS models are validated in a time and frequency combinational platform for better efficacy. Several performance parameters have been computed to test and check the developed models for their efficient portrayal of expressive states. It can be inferred that the feature combination indeed provides improved models due to the availability of more complementary information. It has witnessed the lowest training, testing, and checking RMSE as compared to either the frequency or time-domain feature vectors. Investigation and validation of other efficient feature extraction algorithms in combinational and reduction platforms may provide new insights into this field.





## REFERENCES

- [1] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168865–168878, 2020, doi: 10.1109/ACCESS.2020.3023871.
- [2] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47795–47814, 2021, doi: 10.1109/ACCESS.2021.3068045.
- [3] I. K. Fodor, "A survey of dimension reduction techniques," *Library*, vol. 18, no. 1, pp. 1–18, 2002, doi: 10.2172/15002155.
- [4] J. Yuan, L. Chen, T. Fan, and J. Jia, "Dimension reduction of speech emotion feature based on weighted linear discriminant analysis," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 8, no. 11, pp. 299–308, 2015, doi: 10.14257/ijsp.2015.8.11.27.
- [5] B. Mocanu, R. Tapu, and T. Zaharia, "Utterance level feature aggregation with deep metric learning for speech emotion recognition," *Sensors*, vol. 21, no. 12, 2021, doi: 10.3390/s21124233.
- [6] L. Kerkeni, Y. Serrestou, K. Raoof, M. Mbarki, M. A. Mahjoub, and C. Cleder, "Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO," *Speech Communication*, vol. 114, pp. 22–35, 2019, doi: 10.1016/j.specom.2019.09.002.
- [7] L. Abdel-Hamid, "Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features," *Speech Communication*, vol. 122, pp. 19–30, 2020, doi: 10.1016/j.specom.2020.04.005.
- [8] M. N. Mohanty and H. K. Palo, "Segment based emotion recognition using combined reduced features," *International Journal of Speech Technology*, vol. 22, no. 4, pp. 865–884, 2019, doi: 10.1007/s10772-019-09628-3.
- [9] D. Li, Y. Zhou, Z. Wang, and D. Gao, "Exploiting the potentialities of features for speech emotion recognition," *Information Sciences*, vol. 548, pp. 328–343, 2021, doi: 10.1016/j.ins.2020.09.047.
- [10] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020, doi: 10.1016/j.specom.2019.12.001.
- [11] S. G. Koolagudi, Y. V. S. Murthy, and S. P. Bhaskar, "Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 167–183, 2018, doi: 10.1007/s10772-018-9495-8.
- [12] M. Shah Fahad, A. Ranjan, J. Yadav, and A. Deepak, "A survey of speech emotion recognition in natural environment," *Digital Signal Processing: A Review Journal*, vol. 110, 2021, doi: 10.1016/j.dsp.2020.102951.
- [13] L. Sun, S. Fu, and F. Wang, "Decision tree SVM model with Fisher feature selection for speech emotion recognition," *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, 2019, doi: 10.1186/s13636-018-0145-5.
- [14] S. Lalitha, D. Geyasruti, R. Narayanan, and M. Shrivani, "Emotion detection using MFCC and cepstrum features," *Procedia Computer Science*, vol. 70, pp. 29–35, 2015, doi: 10.1016/j.procs.2015.10.020.
- [15] L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, "Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction," *Information Sciences*, vol. 509, pp. 150–163, 2020, doi: 10.1016/j.ins.2019.09.005.
- [16] R. Ram, H. K. Palo, M. N. Mohanty, and L. P. Suresh, "Design of FIS-based model for emotional speech recognition," *Advances in Intelligent Systems and Computing*, vol. 397, pp. 77–88, 2016, doi: 10.1007/978-81-322-2671-0\_8.
- [17] M. Asimuzzaman, P. D. Nath, F. Hossain, A. Hossain, and R. M. Rahman, "Sentiment analysis of bangla microblogs using adaptive neuro fuzzy system," *ICNC-FSKD 2017-13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, pp. 1631–1638, 2018, doi: 10.1109/FSKD.2017.8393010.
- [18] L. Tan et al., "Speech emotion recognition enhanced traffic efficiency solution for autonomous vehicles in a 5G-enabled space-air-ground integrated intelligent transportation system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 2830–2842, 2022, doi: 10.1109/TITS.2021.3119921.
- [19] S. Giripunje and N. Bawane, "ANFIS based emotions recognition in speech," in *Knowledge-Based Intelligent Information and Engineering Systems*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 77–84, doi: 10.1007/978-3-540-74819-9\_10.
- [20] M. Viswanathan, Z. X. Zhang, X. W. Tian, and J. S. Lim, "Emotional-speech recognition using the neuro-fuzzy network," *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication, ICUIMC'12*, 2012, doi: 10.1145/2184751.2184863.
- [21] K. Wang, G. Su, L. Liu, and S. Wang, "Wavelet packet analysis for speaker-independent emotion recognition," *Neurocomputing*, vol. 398, pp. 257–264, 2020, doi: 10.1016/j.neucom.2020.02.085.
- [22] H. Zhang, H. Huang, and H. Han, "A novel heterogeneous parallel convolution Bi-LSTM for speech emotion recognition," *Applied Sciences (Switzerland)*, vol. 11, no. 21, 2021, doi: 10.3390/app11219897.
- [23] M. Seo and M. Kim, "Fusing visual attention cnn and bag of visual words for cross-corpus speech emotion recognition," *Sensors (Switzerland)*, vol. 20, no. 19, pp. 1–21, 2020, doi: 10.3390/s20195559.
- [24] H. K. Palo and S. Sagar, "Comparison of neural network models for speech emotion recognition," *Proceedings-2nd International Conference on Data Science and Business Analytics, ICDSBA 2018*, pp. 127–131, 2018, doi: 10.1109/ICDSBA.2018.00030.
- [25] K. Chauhan, K. K. Sharma, and T. Varma, "Speech emotion recognition using convolution neural networks," *Proceedings-International Conference on Artificial Intelligence and Smart Systems, ICAIS 2021*, pp. 1176–1181, 2021, doi: 10.1109/ICAIS50930.2021.9395844.





- [26] V. Rezaie, A. Parnianifard, D. Z. Rodriguez, S. Mumtaz, and L. Wuttisittikulkij, "Speech emotion recognition using ANFIS and PSO-optimization with Word2Vec," *J Neuro Spine*, vol. 1, no. 1, pp. 41–56, 2023, doi: 10.21203/rs.3.rs-1237929/v1.
- [27] M. Dirik, "Optimized anfis model with hybrid metaheuristic algorithms for facial emotion recognition," *International Journal of Fuzzy Systems*, 2022, doi: 10.1007/s40815-022-01402-z.

## BIOGRAPHIES OF AUTHORS







**Surjyo Narayana Panigrahi**     is a Ph.D. research scholar at YBN University, Namkum, Ranchi, Jharkhand, India. His area of interest belongs to Signal and Image. He is involved in teaching and research for the last 18 years in different Engineering colleges in Odisha, India. He can be contacted at email: surjyo@gmail.com.



**Niharika Pattanaik**     is a Ph.D. research scholar at SOA University. Her area of interest belongs to Signal and Image. He is involved in teaching and research for the last 9 years in different Engineering colleges in Odisha, India. She can be contacted at email: pattanaikniharika25@gmail.com.



**Hemanta Kumar Palo**     received a Master of Engineering from Birla Institute of Technology, Mesra, Ranchi in 2011 and a Ph.D. in 2018 from the Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India. Currently, he is serving as an Associate Professor in the Department of Electronics and Communication Engineering at the Institute of Technical Education and Research, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India. His area of research includes signal processing, speech and emotion recognition, machine learning, and analysis of power quality disturbances. He can be contacted at email: hemantapalo@soa.ac.in.