**Effective Modelling of Voice Emotions through Optimally Tuned Mel-Frequency Spectral Coefficients and Probabilistic Neural Network**

Niharika Pattanaik1, a) and Hemanta Kumar Palo2, b)

1,*2Institute of Technical Education and Research, Siksha O Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India*

a)Corresponding author: [hemantapalo@soa.ac.in](mailto:hemantapalo@soa.ac.in),

b)[pattanaikniharika25@gmail.com](mailto:pattanaikniharika25@gmail.com)

**Abstract.** Human emotions from voice is an active field of research, catering to several vivid application domains such as psychology, criminal investigation, human-computer interfacing, banking and finance, safety and security, online tutoring, etc. Mostly discussed conventional techniques in the literature discuss periodic coefficients such as pitch, energy, speech rate, etc., that are inadequately generalized, leading to inaccurate results. This work addresses these issues by promising and integrating the paralinguistic log Mel-frequency spectral coefficients (MFSCs) through Particle Swarm (PS) optimization models popular in this field for improved voice emotions (VEs) modelling. The developed optimization model has been tested and compared for its efficacy with the statistical MFSC model, which is similarly conceptualized. Simulation through the model of the Probabilistic Neural Network reveals the superiority of the optimized models over state-of-the-art techniques in classifying the chosen VEs; hence, it is a novel effort in this direction.

**Keywords:** Voice emotions, mel-frequency spectral coefficients, optimization, probabilistic neural network, classification.

**INTRODUCTION**

Recognition of human voice emotions (VEs ) by machines remains an ongoing research topic due to its application in criminal investigation, banking and finance sectors, online tutoring, robotics, education system, psychology consultations, etc. However, it remains a complex task due to the variations in the acoustic properties that change among speakers, age, gender, cultural environment, and affective states [1-2]. Similarly, the change in formants, speaking rate, fundamental frequency, and acoustic correlations with emotional outbursts often challenge the community.

The derivation of discriminating voice coefficients constitutes a significant module to enhance the output of a recognition machine. Nevertheless, the community focuses on fetching prosodic or spectral coefficients and their variants to develop a suitable identification system model [3-5].The involvement of temporal coefficients enables the frame-level parameter extraction to be superior to the utterance-level extractions, as evidenced in these articles.Among the spectral coefficients, the Mel frequency cepstral coefficients (MFCC) have been widely employed in this field due to their effectiveness in logarithmically representing the human hearing mechanism. However, the MFCC provides a compact representation of a speech signal, decorrelates the features, and is suitable for modelling linear classifiers such as the Gaussian mixture model (GMM). On the contrary, the MFSC describes the spectral envelope more directly, showing a better result than the MFCCs in analyzing HVEs [6-7], and has been explored here.

However, the frame-level vectors obtained from a voice sample comprise thousands of elements, with irrelevant information. It results in larger memory requirements, increases the training time, and reduces the system efficacy, besides introducing system complexity. Representing these high-dimensional vectors in a compact platform arguably improves the system performance, hence creating scope in this field. Several feature selection and optimization techniques, such as Genetic algorithm, particle swarm, ant colony, cuckoo search, whale optimization, etc., have been suggested to alleviate these issues and for compelling emotional portrayal [8-12]. Among these, the PS is easy to implement, robust, flexible, and quicker to converge to the optimal solution; henceforth, it is chosen here.

Similarly, several machine learning models, such as the GMM, Hidden Markov Model (HMM)/ Support Vector Machine (SVM), artificial neural networks (NN), deep networks, decision trees, etc., have been effectively applied to model the HVEs earlier. Nevertheless, each classifier has limitations and benefits; hence, finding the optimal learner to fetch the best accuracy is difficult. In this regard, the NN-based classifiers have shown better processing quality, are simple to design, and can discriminate voice emotion adequately, thus remaining popular compared to the conventional HMM/GMM/SVM while modeling optimized vectors [13]. Among NN, the PNN is a non-parametric method, requiring less variable adjustment, henceforth experiencing a better outcome with higher accuracy and speed than the multilayer perceptron or the Radial basis function networks in this field [14]. Further, the absence of local minima, faster learning, adequate representation of statistical characteristics, optimization through Bayesian optimum criteria, least prone to over-fitting, etc., makes it a potential candidate to model the HVEs; henceforth, it is considered here.

The paper is structured as follows. Section 2 provides the proposed feature extraction and optimization techniques. The simulation results, including the characterization and classification accuracy, have been analyzed and discussed in Section 3. The conclusion and limitations of the proposed method are briefed in Section 4 with further future directions.

**THE PROPOSED METHOD**

Figure 1 provides the block diagram of the proposed approach with a brief of different blocks further down the line.

C:\Users\Lenovo\Desktop\AMSE_2025\Niharika\tiff\proposed approach.tif  
**FIGURE 1.** The proposed approach

**Database used**

The EMO-DB (Berlin) database is used for the proposed work. It comprises 127 anger, 62 sadness, 46 disgust, 71 joy, 81 boredom, and 79 neutral voice samples [15]. Out of the database, only six emotions are chosen, i.e., fear, happiness, neutral, sad, and anger. Ten actors—five men and five women—have mimicked these feelings by simulating ten German lines, five longer and five shorter. This database contains a selection of ordinary, semantically neutral utterances ranging from 1.5 to 4 seconds. A sampling frequency of 48 kHz down-sampled to 16 kHz with 16-bit precision is maintained to record the utterances in an anechoic chamber.

**Feature Extraction**

The selection of a suitable feature extraction technique is an important task for the recognition of HVEs. The proposed scheme is conceptualized with baseline MFSC features using the following steps.

* The raw signal is initially pre-processed through pre-emphasis, normalization, framing, and windowing blocks. Normalization eliminates finite accuracy effects, while the pre-emphasis block balances the voice signal's magnitude spectrum. The pre-emphasis filter with a constant pre-emphasis factor β= 0.98 is defined as

 (1)

Framing and windowing are done to represent the non-stationary voice signal in a periodic plane for improved analysis. This work considers a 30-ms frame size with 10-ms frame overlapping and the Hamming window wn to eliminate the edge effect.

 (2)

wheredenotes the number of samples in a signal.

* The FFT represents the signal in the frequency domain by computing the signal spectrum. For a signal frame, where  and  represent the number of samples per frame and number of frames, respectively, the Discrete FT of length  is given by

 (3)

This work considers a 512-point DFT for the intended analysis.

* The human ear's non-linear mechanism often masks the low-frequency signals with high-frequency ones. To alleviate this issue and approximate the signal in a logarithmic spectrum, the Mel-scale has been used as given by

 (4)

The Mel-scale often comprises 20 to 40 triangular filters, from which twenty-six filters have been considered in general.

**Optimization**

Kennedy and Eberhard were the first to coin the PS algorithm based on the movement of a group of birds in search of food randomly by following the closest chosen path. Particle with a fitness function represents each bird. A group of particles is called a swarm that searches for the local best solution, or  in every iteration, followed by another best solution called the global best or .However, without volume information, each particle is defined by its position  and speed. A particle’s  positionand speed vector correspond to an -layered vector space. In executing advancement, the particle's swarm begins from their positionsand searches iteratively in the -layered space with the coordinated speed, varying them once every cycle according to the set criteria. The  and positions in an individual ideal arrangement are defined as

,  (5)

The particle velocity  successive time is updated using the following equation.

 (6)

The first component of equation (1) provides the inertial, while the second and third components resemble the personal and social influence. The following are the significant parameters used to model the VEs using PS.

* : The position of theth particle at time .
* : The velocity of the th particle at time .
* : The local best or positional best solution of.
* : The global (swarm) best solution.
* ,: The acceleration constants for the cognitive and social components.
* ,: The stochastic random constant between 0 and 1 corresponds to the individual and social learning.
* : The scaled inertia weight of the earlier velocity at a particle time .

Here,  is the global best position of the particle at time and  is the scaled inertia weight of the earlier velocity at a particle time.The variables and  are the acceleration constants for the cognitive and social components, respectively. Similarly, the variables and denote the stochastic random constant corresponding to the individual and social learning, respectively.

The following equation can update the particle's position by finding the new velocity.

 (7)

In PS, random numbers ​ and  are the learning factors representing the weights, pulling a particle towards  and .They prevent the swarm from getting stuck too early in local minima by injecting controlled randomness to make the swarm search more effective and avoid premature convergence. They have a value of 2, 1.49445, 2.8, and 1.3, providing the optimal convergence. There are many variants of PS in which these values change with time and are dynamically adjusted corresponding to several fitness values, dispersion degree, evolutionary states, etc. However, there are no decisive conclusions on the asymmetric value of and  which provides an optimal convergence. This work trials several values of the parameters for the intended task. An inertia weight of 0.8, population size of 20, and  and of 1.5 (providing equal weight to the social and cognitive search) have experienced the optimum performance, and hence, have been retained. Figure 2 provides the PS algorithm.

C:\Users\Lenovo\Desktop\AMSE_2025\Niharika\tiff\ps.tif

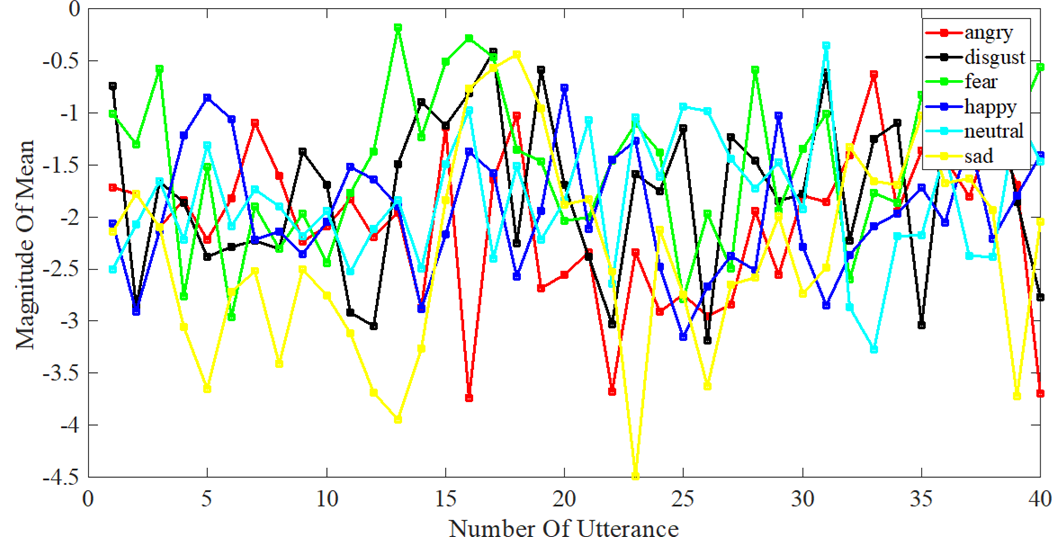
**FIGURE 2.** Steps of the PS algorithm.

**SIMULATION RESULTS AND DISCUSSION**

Experiments were conducted using the EMO-DB voice emotion databases to validate the effectiveness of the proposed MFSC-based global statistical feature extraction method for voice emotion recognition [15]. From each MFSC magnitude matrix of an utterance, eight statistical coefficients, as mean, standard deviation, skewness, kurtosis, maximum, entropy, median, and peak-to-mean ratio, are computed. The feature vectors are concatenated over 40 utterances per emotion for further classification.

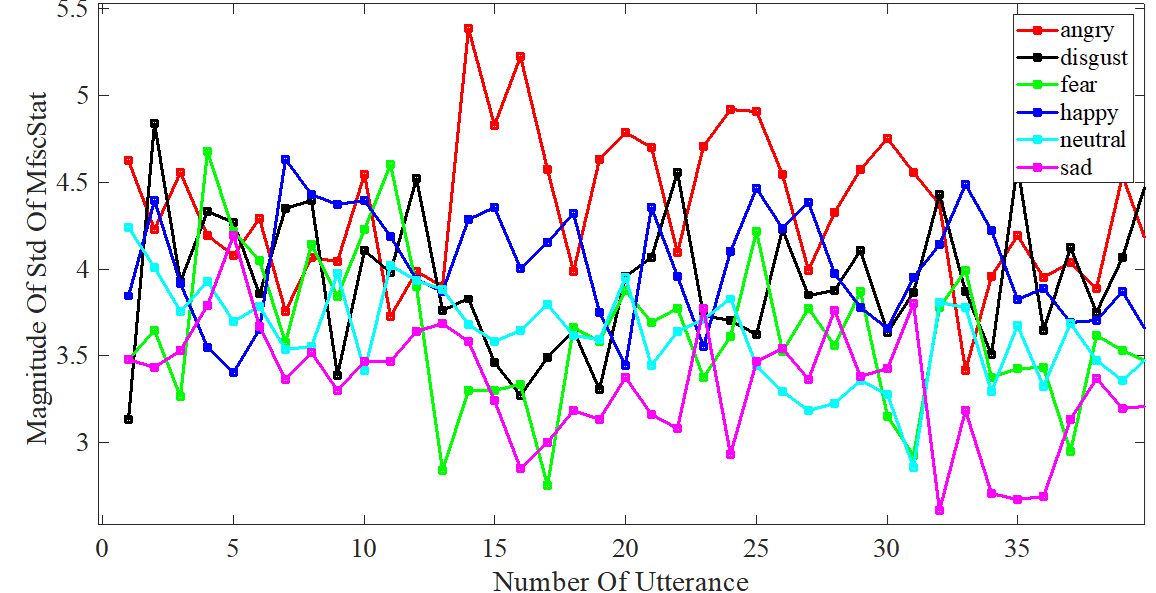
**Characterization using Statistical Coefficients**

Figure 3 illustrates the variation in the mean MFSC magnitudes across 40 utterances and six emotional states. The states such as anger, disgust, fear, and happiness possess a high arousal level, thus noticeably reflecting the higher mean values compared to the low arousal sad state as observed from this Figure. The increased energy levels and presence of high-frequency components are the possible reasons for this trend.



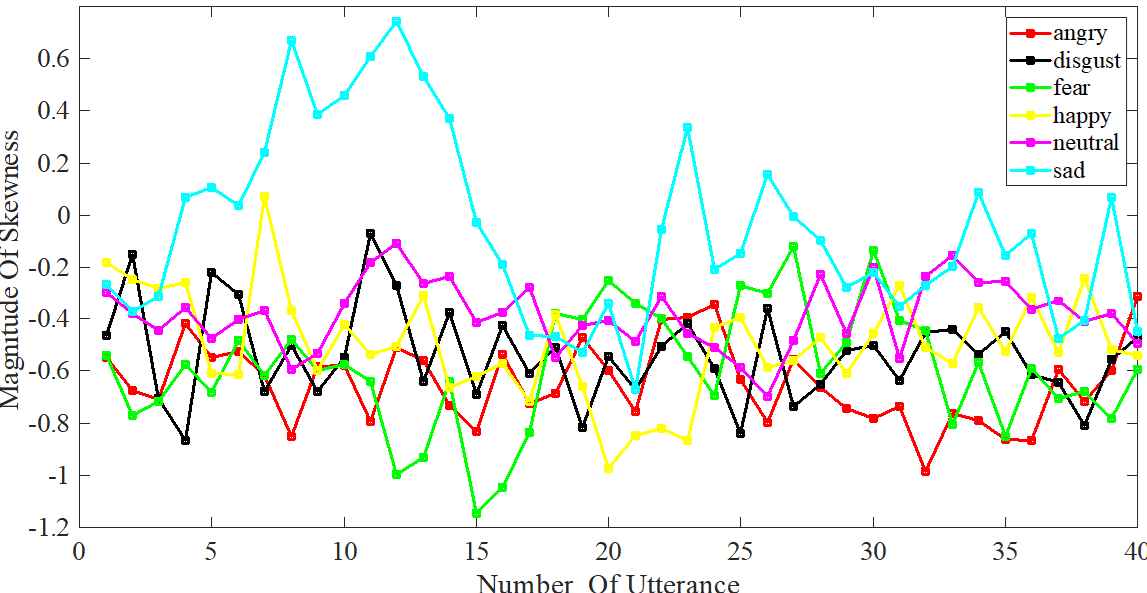
**FIGURE 3.**Mean MFSC variations among emotions

Figure 4 presents the standard deviation of the MFSC magnitude across 40 utterances corresponding to each emotion. It provides information about the variability of the emotional expressions. As earlier, the high arousal states have witnessed the higher variability, with the angry state showing the highest magnitude. In contrast, neutral and sad are associated with a lower variability, smoother, and more stable spectral energy, typical of a calm or subdued voice. Such analysis can demarcate between expressive and monotonic emotions, facilitating the development of an arousal-based recognition system.



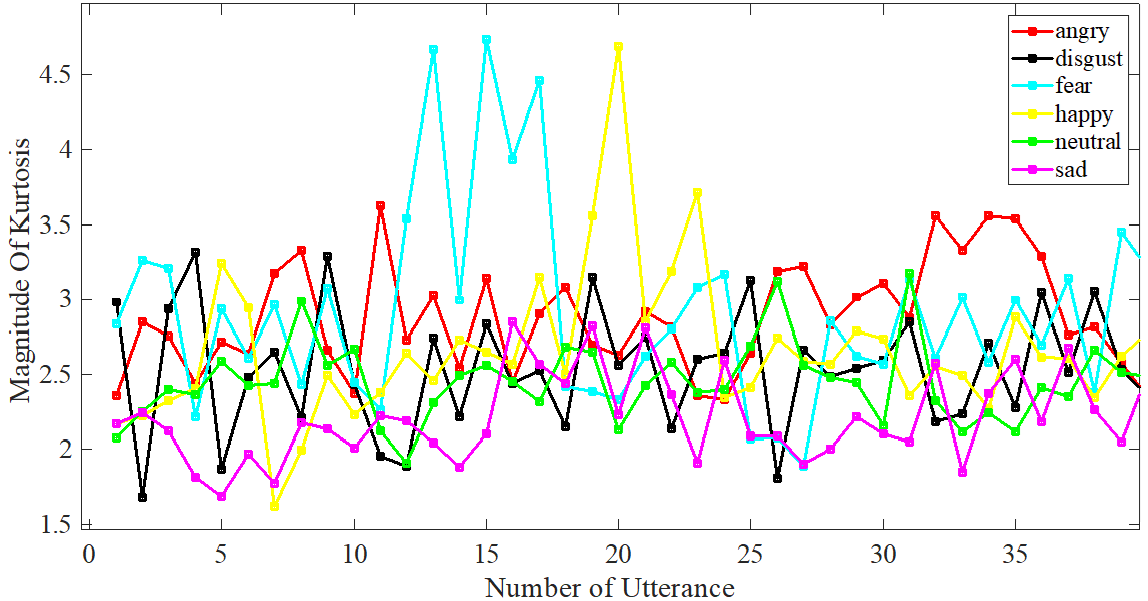
**FIGURE 4.** Variations of the MFSC standard deviation among emotions

Figure 5 plots the skewness of the MFSC magnitude across different emotional states. It illustrates the spectral energy distribution asymmetry, thus helping to characterize the chosen states. The neutral and sad voices have been experiencing a higher skewness, indicating more energy concentration in lower frequency bands. It denotes a slow or flat voice delivery compared to the high arousal state, such as anger, happiness, and disgust, with better-balanced or mildly skewed energy dispersion across frequencies.



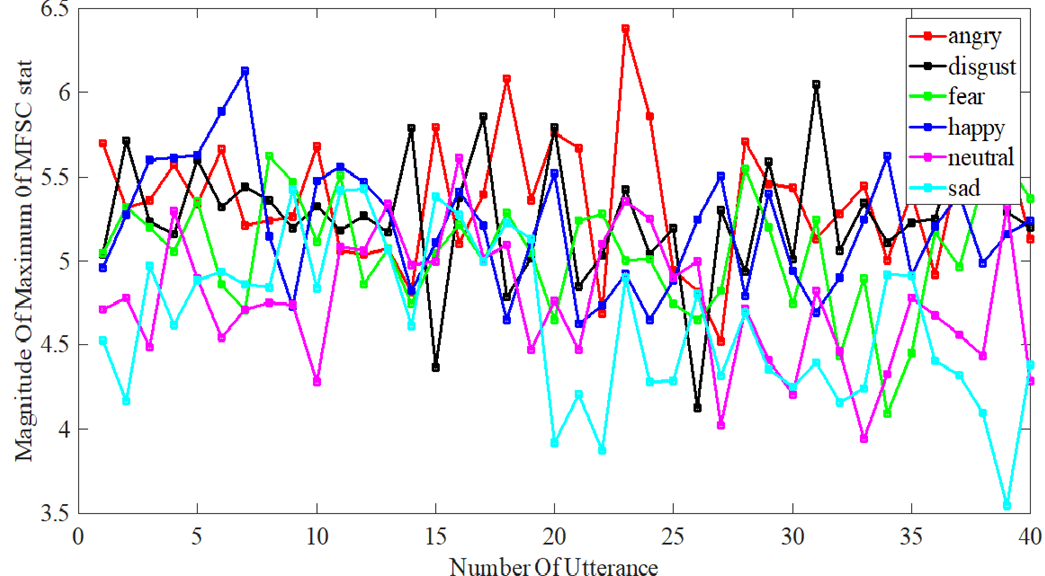
**FIGURE 5.** Variations of MFSC skewness across emotions

Figure 6 reflects the variation in the MFSC magnitude kurtosis across the chosen emotional states, indicating the peakedness or tailedness of the spectral energy distribution. The high arousal states exhibit high kurtosis, suggesting more pronounced variations in frequency content and vocal variation. In contrast, the sad, disgust, and neutral show lower kurtosis, which indicates less spread-out energy and expressive articulation. This feature effectively captures the contrast between compact vs. dynamic energy profiles, enhancing emotion differentiation in combination with other spectral descriptors.



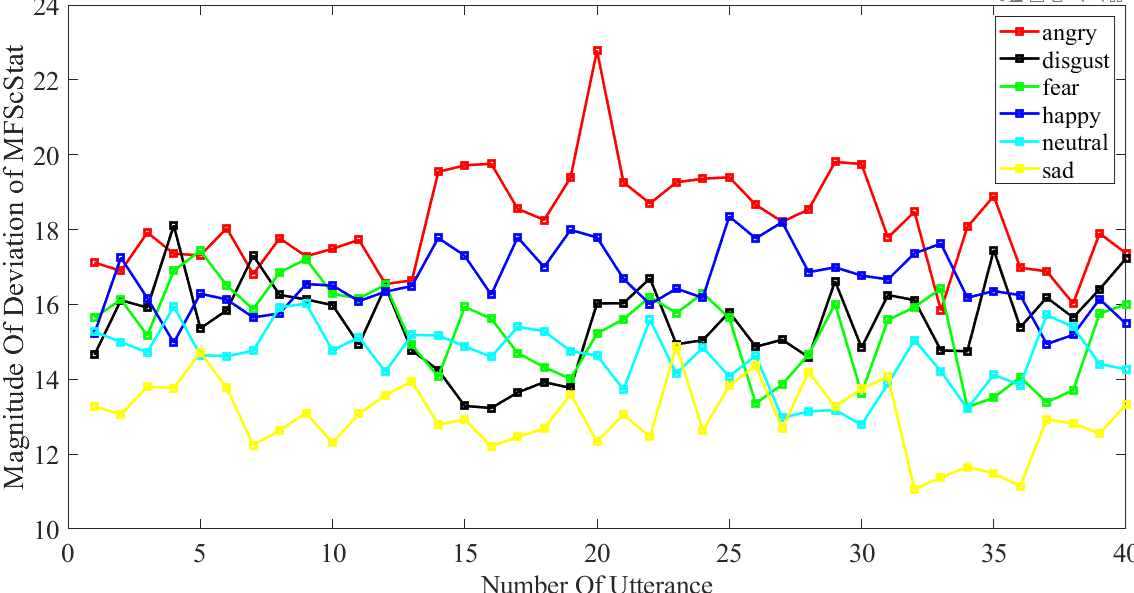
**FIGURE 6.** Variations of MFSC kurtosis across emotions

Figure 7 displays the maximum MFSC magnitude values across utterances, representing the strongest spectral component within each emotional sample. Emotions like anger, disgust, and, to some extent, happiness show higher peaks, reflecting the presence of sharp vocal bursts or emphatic utterance. In contrast, neutral and sad maintain lower peak values, with softer, more uniform voices. This feature captures momentary emotional intensity, making it valuable for identifying emotions with sudden or forceful energy patterns.



**FIGURE 7.** Variations of the MFSC maximum across emotions

In VE recognition, amplitude statistics are used to differentiate between emotions and observe how the loudness or intensity of voice changes with different states. Essentially, different VEs are associated with variations in voice samples' amplitude (loudness). Henceforth, examining the amplitude fluctuation range provides valuable insights into the expressed emotion. Figure 8 indicates that the MFSC amplitude range across emotions like anger and happiness has the highest maximum amplitudes and larger ranges, indicating louder and more fluctuating voices. The feelings like sadness, fear, and disgust have lower average amplitude changes with smaller ranges. The amplitude statistics range helps capture the intensity and dynamic variations in VEs that are often linked to different affective states.



**FIGURE 8.** Variations of MFSC range across emotions

**Classification using PNN**

An input consisting of the statistical feature streams of all the designated states has been used to train the PNN, and the output is associated with the state intended for classification. Table 1 compares the PNN recognition performance among the MFSC-statistics and MFSC-PS to validate the efficacy of the proposed VE models. The objective is to approximate the high-dimensional MFSC vectors in a condensed and optimized feature space for improved performance. The raw MFSC feature vectors are first retrieved at the frame level and subsequently subjected to the statistical and PS mechanisms to develop the low-dimensional optimized vectors to train the classifier. The proposed algorithms have shown better accuracy than the stand-alone raw algorithms cited in earlier literature. A training/ validation/ testing data division ratios of 70%:15%:15% have been used to simulate the PNN. Among the statistical and PS optimized vectors, the latter has outperformed the former due to better discriminating and relevant coefficients.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **TABLE 1.** Comparison of PNN accuracy using optimized MFSC-PS and MFSCs Statistics | | | | | | | |
| **Features** | **Emotions** | | | | | | |
| **Angry** | **Disgust** | **Fear** | **Happy** | **Neutral** | **sad** | **Average** |
| MFSC-Stat | 70.05% | 68.01% | 72.15% | 72.01% | 69.50% | 64.02% | 69.29% |
| MFSC-PS | 95.05% | 92.01% | 86.15% | 90.02% | 89.50% | 78.01% | 88.46% |

Table 2 represents the classification time required to model and classify the chosen emotions using the PNN. These factors entirely depend on the type and dimension of the feature vectors and, hence, vary accordingly. The PS vectors are of larger dimension than the statistical vectors, thus requiring a larger classification time, although they take less time to model the chosen states.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **TABLE 1.** Comparison of classification time in seconds using optimized MFSC-PS and MFSCs Statistics | | | | | | | |
| **Feature Vectors** | **Modeling-Time** | | | | | | **Classification Time** |
| **Angry** | **Disgust** | **Fear** | **Happy** | **Neutral** | **Sad** |  |
| MFSC-Stat | 6.4 | 3.2 | 2.7 | 3.78 | 3.54 | 2.14 | 74.12 |
| MFSC-PS | 2.87 | 1.6 | 1.62 | 2.05 | 1.72 | 2.02 | 82.77 |

**CONCLUSION**

This work conducts a thorough statistical analysis using several measures to demarcate the low and high-arousal states. The raw MFSC feature vectors are analyzed at the frame level and are then approximated in a reduced and optimized feature space using statistical and PS algorithms. The experimental results indicate that PS has improved the accuracy of HVE recognition better than previous works in this field. It also ensures the learning ability of PNN with a better generalization ability and integrated performance. Although the optimized feature vectors outperformed the stand-alone raw vectors, there is scope for further improvement. Exploring other optimization algorithms, feature extraction techniques, and more efficient classifiers can provide new inputs; hence, they are kept for future work.

# References

1. P. Laukka and H. A. Elfenbein, Emotion Review **13**, pp. 3-11 (2021).
2. H. K. Palo, “The effect of age, gender, and arousal level on categorizing human affective states,” in *Emotion and Information Processing: A Practical Approach*, edited by [Sachi Nandan Mohanty](https://link.springer.com/book/10.1007/978-3-030-48849-9#author-1-0) (Springer International Publishing, Cham, 2020), pp. 97–124.
3. G. H. Mohmad and R. Delhibabu, IEEE Access **12**, pp. 151122 – 151152(2024).
4. S. Sekkate, M. Khalil, and A. Adib*,* Multim. Tools Appl. **82**, pp.11443-11460 (2023).
5. N.Pattanaik and H. K. Palo, Speech Commun. **169**, 103187 (2025).
6. J. Ancilin and A. Milton, Appl. Acoust. **179**, pp.108046 (2021).
7. S. Mishra, N. Bhatnagar, P. P., and T. R. Sureshkumar, Multi. Tools Appl. **83**, pp. 37603-37620 (2024).
8. S. N. Panigrahi and H. K. Palo, “Analysis and recognition of emotions from voice samples using ant colony optimization algorithm,” in *Innovation in Electrical Power Engineering, Communication, and Computing Technology,* Proceedings Second IEPCCT Singapore (Springer Singapore, 2021), pp. 219-231.
9. S. N. Panigrahi and H. K. Palo, “Emotional speech recognition using particle swarm optimization algorithm,” in *Advances in Power, Signal, and Information Technology*,2021 International Conference(APSIT, IEEE), pp.1-5.
10. M. A. A. Albadr, S. Tiun, M. Ayob, F. T. AL-Dhief, K. Omar, and M. K. Maen, Multi. Tools Appl. **81**, pp. 23963-23989 (2022).
11. Z. Zhang, Alex. Eng. J. **60**, pp.1499-1507 (2021).
12. C. Deepika and S. Kuchibhotla, Measurement: Sensors **32**, pp.101030 (2024).
13. T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, IEEE access **9**, pp. 47795-47814 (2021).
14. M.N. Mohanty and H. K. Palo, Measurement **152**, pp.107369 (2020).
15. F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier and B. Weiss, Interspeech **5**, pp. 1517-1520 (2005).