Statically Tuned Fractional Fourier Transform for Effective Representation of Voice Emotions

Suraj Kumar Naik1, a), Hemanta Kumar Palo2, b), Bibhuprasad Mohanty3, c)

1,2, 3 Department of Electronics and Communication Engineering, Institute of Technical Education and Research, Siksha O Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India

.   
*b) Corresponding author:* [*hemantapalo@soa.ac.in*](mailto:hemantapalo@soa.ac.in)

*a)* [*surajnaik@soa.ac.in*](mailto:surajnaik@soa.ac.in)

*c) bibhumohanty@soa.ac.in*

**Abstract.** This paper aims to characterize human voice emotions (HVEs) by computing several statistically significant performance parameters. The objective is to distinguish and classify the HVEs corresponding to their energy or arousal level. The relevance of the Fractional Fourier Transform (FrFT) has been explored to segment the weak and strong arousal states that include anger, sadness, happiness, neutral, and fear. The algorithm represents the analyzed signal through a suitable angle in the time-frequency (TF) plane, fetching emotionally specific information, hence considered here. The features are extracted from prominent intermediate TF sectors of interest and are statistically tuned for adequate characterization and classification using several machine learning algorithms. The strong arousal states have been experiencing a higher FrFT-magnitude across all statistical representations than the weak arousal states, providing an effective demarcation among the chosen HVEs. Simulation results reveal the superiority of the illustrated approach compared to spectral or wavelet transform (WT), hence a holistic approach in this direction.

# iNtroduction

Human emotions are complexly encoded in voice, revealing themselves through variations in pitch, energy, duration, and spectral content. Recognizing these emotional cues is crucial for developing intelligent human-computer interaction systems. It has gained considerable attention in affective computing, mental health monitoring, virtual assistants, and intelligent tutoring systems [1]. However, HVE Recognition remains a profoundly challenging task due to its subjective, context-dependent, and often ambiguous nature. It varies across age, culture, gender context, and environment [2-3].

Early approaches have emphasized low-level descriptors, including the Mel-Frequency Cepstral Coefficients (MFCCs), formant trajectories, pitch contours, and short-term energy [4]. These techniques capture local spectral properties but often fail to encapsulate the global time-frequency (TF) dynamism essential for differentiating emotional tones. TF analysis methods, such as the Short-Time Fourier Transform (STFT) and WT, have attempted to overcome this limitation by providing localized spectral information. However, STFT suffers from a fixed time-frequency resolution due to a constant window size. WT's performance highly depends on the mother wavelet selection, which may not generalize well across speakers and languages [5].

The Fractional Fourier Transform (FrFT) has emerged as a promising alternative to address these constraints. It extends the classical Fourier Transform by enabling a continuous rotation between the TF axes. It offers flexible analysis in intermediate domains that can effectively represent non-stationary signal characteristics of HVEs [6-7]. By appropriately tuning the fractional order, FrFT can emphasize emotion-specific modulations, including abrupt high-frequency bursts in anger or smooth low-frequency transitions in sadness, that are often obscured in FT or WT representations. Furthermore, FrFT-based representations can be statistically optimized to highlight emotionally salient sectors in the TF domain, improving interpretability and discriminative power for classification tasks, which motivates the authors.

Among feature extraction techniques, the frame-level or local features capture transient characteristics but are sensitive to noise and lack emotional context. In contrast, global statistical features, derived from entire utterances or significant segments, capture long-range prosodic and spectral trends essential for characterizing emotional content. These parameters, including mean, variance, skewness, entropy, and kurtosis, offer better generalization across speakers and are computationally efficient, making them ideal for real-time or embedded applications [8-9].

Classifying features into predefined emotional categories remains an essential module of any recognition system. Each classifier has limitations and benefits; hence, no universally accepted best classifier exists. The choice depends on the task, feature extraction technique, feature dimension, number of patterns to be classified, etc. Nevertheless, this work considers several classifiers to validate the derived feature vectors and compare them for their efficacy in classifying the chosen HVEs.

Motivated by these insights, this study introduces a novel FrFT-based HVE recognition framework that can characterize and classify emotions based on their level of arousal. It leverages the discriminative capacity of FrFT to extract global statistical features that represent emotionally significant TF regions, followed by classification using different classifiers. The proposed method focuses on distinguishing strong arousal emotions (e.g., anger, happiness) from weak arousal emotions (e.g., sadness, boredom), with the dual goals of enhancing recognition performance and improving the interpretability of feature representation. Experimental validation is conducted using the benchmark SAVEE datasets, which provide a mix of linguistic and speaker diversity. Results demonstrate superior accuracy and robustness of the proposed method compared to conventional FT and WT-based approaches, highlighting its potential for deployment in real-world affective applications.

The remainder of this work is arranged as follows. Section 2 reviews prior literature on HVE methods involving the Fourier, Wavelet, and Fractional Fourier transforms. Section 3 elaborates the proposed approach, including preprocessing, FrFT-based feature extraction, statistical tuning, and SVM classification. Section 4 explains the simulation results and the interpreted evaluation. Finally, Section 5 concludes the study by discussing findings and future directions.

# Related literature

HVE Recognition has undergone extensive research, with the earliest approaches found applying the FT-based acoustic cues, which decompose speech signals into their constituent frequency components [1-4]. These features have provided acceptable recognition accuracy in conjunction with traditional classifiers such as SVM, Gaussian Mixture Models (GMM), and KNN, especially in controlled environments. For example, log-spectrum-based FT features have shown exemplary performance in recognizing strong arousal emotions when training and testing conditions are consistent [10]. However, FT lacks temporal resolution, as it assumes signal stationarity and fails to capture dynamic emotional changes across Time. This makes it inadequate for processing spontaneous or conversational speech where emotional cues evolve rapidly [11]. The STFT alleviates this limitation by applying FT within short overlapping windows, but it suffers from a fixed resolution trade-off that limits adaptability to varying speech dynamics. The WT allows multi-resolution TF analysis using an adaptive window, thus remaining versatile in detecting non-stationary, transient emotional patterns such as sudden pitch modulations or prosodic shifts. It has demonstrated improved recognition of low-arousal emotions like sadness and fear, which are typically difficult to classify using FT-based features [12-13]. However, it suffers from critical drawbacks due to its dependence on selecting the mother wavelet, which is typically chosen empirically and may not generalize across datasets or languages. Additionally, WT's discrete and fixed decomposition structure limits its ability to adaptively align with emotion-specific signal characteristics.

More recently, the Fractional Fourier Transform (FrFT) has gained prominence in SVE analysis due to its ability to generalize the FT by introducing a tunable fractional order or rotation angle. This allows FrFT to analyze speech signals in domains between Time and frequency, offering greater flexibility in representing non-stationary behaviors that typify emotional speech [6-7]. Unlike FT or WT, FrFT enables continuous control over TF resolution, facilitating the extraction of emotionally discriminative patterns such as energy bursts in anger or slow spectral drifts in sadness. Studies have shown that statistically optimized fractional orders can significantly improve the separability between high- and low-arousal emotions [14]. The FrFT-based features represented using the statistical mean, skewness, kurtosis, entropy, and spectral flux with classifiers have achieved better performance than FT and WT in classifying HVEs across different databases. Although selecting optimal fractional orders can be computationally intensive, algorithmic advancements and parallelization have made real-time FrFT-based HVE systems feasible. Importantly, FrFT-derived features are compact and interpretable, which is beneficial for applications demanding understanding, such as healthcare diagnostics and forensic analysis.

This motivates the authors to traverse through the FT, WT, and then to FrFT in characterizing and classifying the HVEs, reflecting a shift toward more adaptable, TF aware techniques capable of capturing the nuanced and non-stationary nature of the signal. While FT and WT provided foundational insights and tools, their limitations have underscored the need for more flexible representations like those enabled by FrFT. This study builds upon these advancements by proposing a statistically tuned FrFT-based HVE framework incorporating global statistical features and robust classification, aiming to achieve higher accuracy, interpretability, and computational efficiency in real-world applications.

# proposed methodology

The proposed methodology is structured around four core stages: data acquisition, preprocessing and framing, FrFT-based global feature characterization, and classification using several machine learning algorithms. The system is designed to differentiate between intense arousal (e.g., anger, happiness) and weak arousal (e.g., sadness, boredom), emphasizing interpretability, robustness, and computational efficiency. Figure 1 provides the proposed scheme.

C:\Users\Lenovo\Desktop\AMSE_2025\Suraj\the proposed.tif

**Figure 1.** The Proposed Scheme

## Database Used

This work considers the Surrey Audio-Visual Expressed Emotion (SAVEE) to validate the reliability of the developed method. The database is in an understandable English language and is well-established in HVE research due to its diversity in emotional classes [15-16]. It comprises 480 utterances from four male English speakers, each expressing seven emotions: anger, fear, happiness, surprise, sadness, disgust, and neutral. The utterances are recorded at 44.1 kHz sampling frequency, and are further down-sampled to 16 kHz for our purpose and convenience. The utterances in SAVEE reflect a more spontaneous and naturalistic style, with phonetically balanced sentences and diverse prosodic structures due to the involvement of amateur subjects. Compared to other datasets, SAVEE introduces higher inter-speaker variability and linguistic diversity, making it suitable for testing model generalization in semi-realistic conditions.

## Preprocessing, framing, and Windowing

A structured preprocessing pipeline was applied to all voice samples to enhance signal quality and normalize feature characteristics across speakers. The pipeline includes amplitude normalization, silence removal, pre-emphasis filtering, framing, and windowing. The amplitude normalization was performed to scale all utterances within a uniform dynamic range, thereby reducing speaker-dependent energy variations. Next, silence segments were removed using an energy-based voice activity detection (VAD) algorithm that leverages short-time energy and zero-crossing rate to eliminate non-speech regions at the beginning and end of recordings. To emphasize high-frequency speech components critical for emotional cues, such as pitch transitions and glottal activity, a pre-emphasis filter with a transfer function has been used, as given by

 (1)

Each voice sample is further divided into overlapping short-time frames using a sliding window of 30 milliseconds in length with a 10-ms overlap. This frame size balances the need for local stationarity with sufficient resolution to capture emotional dynamics. To reduce spectral leakage and improve TF localization, each frame was multiplied by a Hamming window:

 (2)

Here, is the frame length in samples. This preprocessing framework ensures that input to the FrFT transform is clean, temporally stable, and emotionally relevant. The windowed signal is expressed as

 (3)

## Fractional Fourier Transform (FrFT) Feature Vector

The core innovation in this work lies in extracting global statistical features from speech utterances using the Fractional Fourier Transform (FrFT). FrFT generalizes the classical FT by rotating the signal representation within the TF plane according to a fractional order . This tunable rotation allows the signal to be analyzed in intermediate domains, revealing spectral patterns obscured in conventional time or frequency representations. The FrFT algorithm is briefed below.

Let denote the FrFT rotational angle where is the fractional order. The FrFT rotates the signal  along the -axis and is defined as:

 (4)

where the FrFT kernel is given by

 (5)

This transformation generalizes multiple known transforms as

: FrFT reduces to the identity operator (original time domain). 

: FrFT becomes the standard FT (frequency domain).

: the signal returns to its original state, i.e., identity operation again, due to periodicity, .

The inverse FrFT is defined as given by

 (6)

Each frame was transformed using the FrFT with an empirically optimized order. From these selected TF regions of a frame. Further, this work computes a set of global statistical descriptors comprising mean, standard deviation, skewness, kurtosis, maximum, Shannon entropy, median, and peak-to-mean ratio (PMF) to characterize and classify the chosen SEs. These features provide a holistic summary of the speech signal's energy distribution and spectral structure. The derived FrFT-based global features have revealed several advantages, including preserving long-range emotional trends, enhancing separability across arousal levels, and reducing computational complexity through compact representations.

## Classification

The final stage of the framework involves validating the statistically tuned FrFT features through several classification algorithms such as KNN, Discriminant analyzer (DA), Naïve Bayes, Random Forest, and a multi-class SVM. The SVM algorithm is briefed below for convenience.

The SVM is known for its high generalization performance, particularly in high-dimensional and non-linear spaces, and is well-suited for the handcrafted features derived in this study. The SVM decision function is given by:

 (4)

where  are the Lagrange multipliers,  are class labels, and is the kernel function. This work considers the Radial Basis Function (RBF) kernel due to its ability to handle non-linear feature distributions as given by

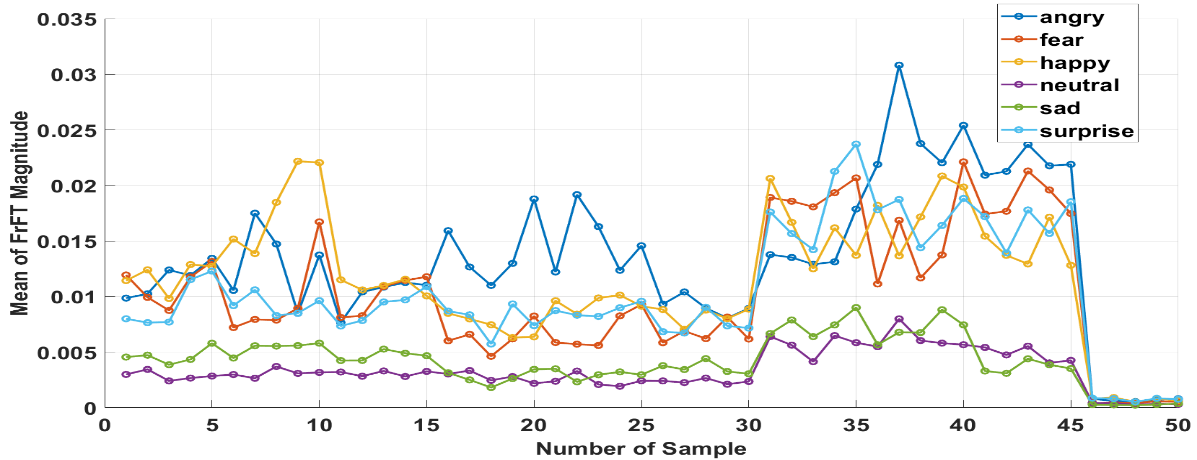
 (5)

The regularization parameter and the kernel parameter  were optimized through a 10-fold cross-validation protocol to ensure statistical reliability and speaker-independence. When the SVM is tuned via cross-validation on the training data, these parameters control the trade-off between margin maximization and classification error. The SVM classifier demonstrated strong performance due to its ability to handle compact but complex feature spaces and its robustness in small-to-medium dataset settings such as SAVEE.

# Simulation Results and analysis

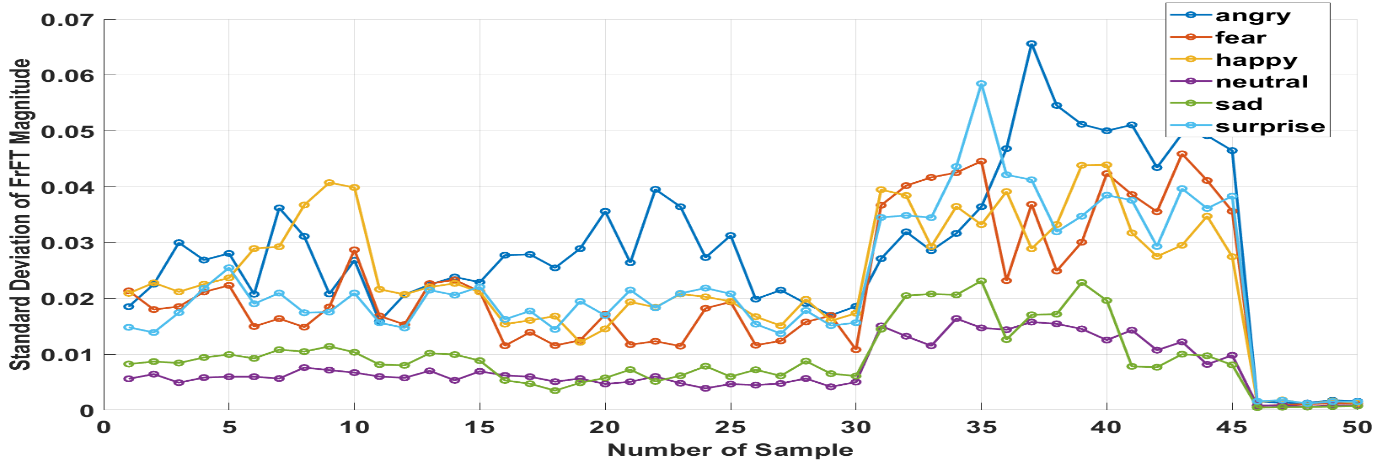
The statistical analysis of the low and high arousal affective states has been carried out to characterize and segment the chosen classes. When analyzed across the six emotions (angry, fear, happy, neutral, sad, surprise), a consistent pattern emerges that supports their use in distinguishing emotional states, particularly in arousal levels and spectral complexity.

Figure 2 plots the mean FrFT magnitude spectrum computed across 50 utterances of the SAVEE database across six different classes. Anger, happiness, and surprise exhibit higher mean values, reflecting increased energy levels and dynamic articulation typical of high-arousal emotions. In contrast, neutral and sad displays lower and flatter means, indicating steady and low-energy speech patterns associated with low arousal. The clear separation between these groups demonstrates that the mean FrFT magnitude effectively distinguishes emotional intensity.



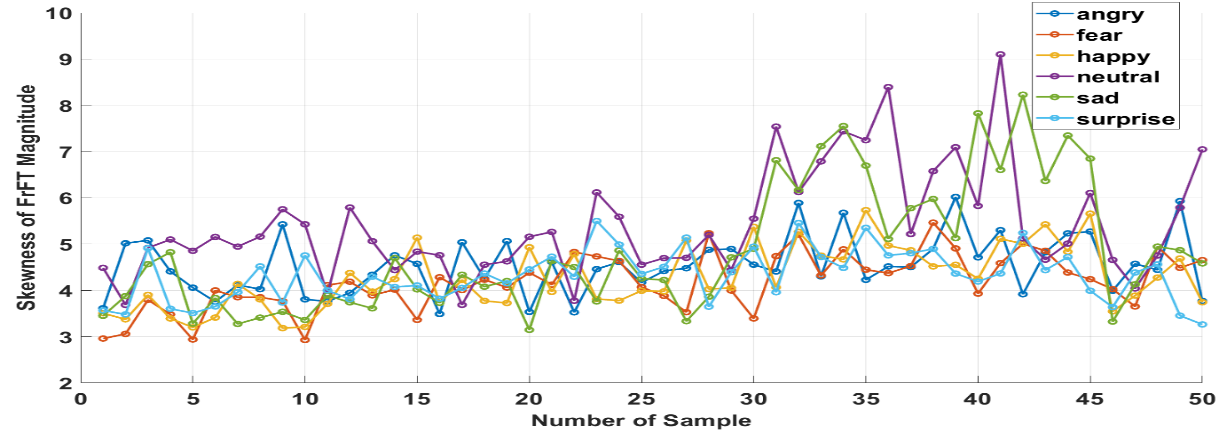
**Figure 2.** Mean FrFT magnitude across emotions

Figure 3 plots the standard deviation of the FrFT magnitude spectrum computed across 50 utterances of the SAVEE database across six different classes. This feature captures the variability of spectral energy, indicating how much the energy fluctuates across the fractional frequency domain. Emotions like anger, happiness, and surprise show higher standard deviation values, reflecting their expressive, dynamic, and variable vocal patterns. In contrast, neutral and sad exhibit lower variability, indicating smoother and more stable spectral energy typical of calm or subdued speech. The feature effectively differentiates between expressive and monotonic emotions, supporting its role in arousal-based classification.



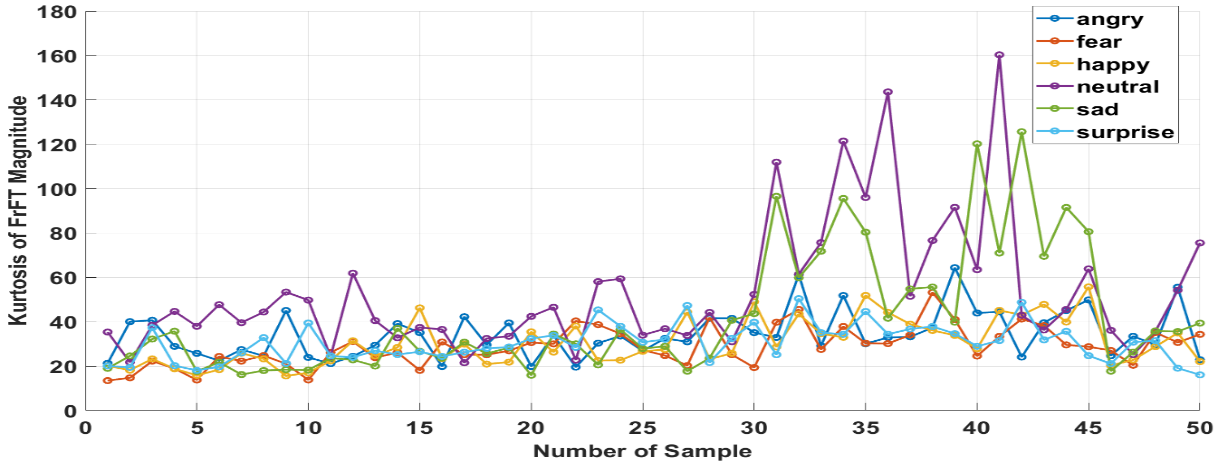
**Figure 3.** Standard Deviation of FrFT magnitude across emotions

Figure 4 shows the FrFT magnitude's skewness values, highlighting each emotional class's spectral energy distribution asymmetry. Higher skewness values, as observed in neutral and sad, suggest more energy concentration in lower frequency bands, characterizing flat or slow speech delivery. In contrast, anger, happiness, and surprise demonstrate more balanced or moderately skewed spectra, indicating a broader energy spread across frequencies. This helps capture the spectral bias of VEs and is particularly useful for identifying less expressive, low-arousal emotions.



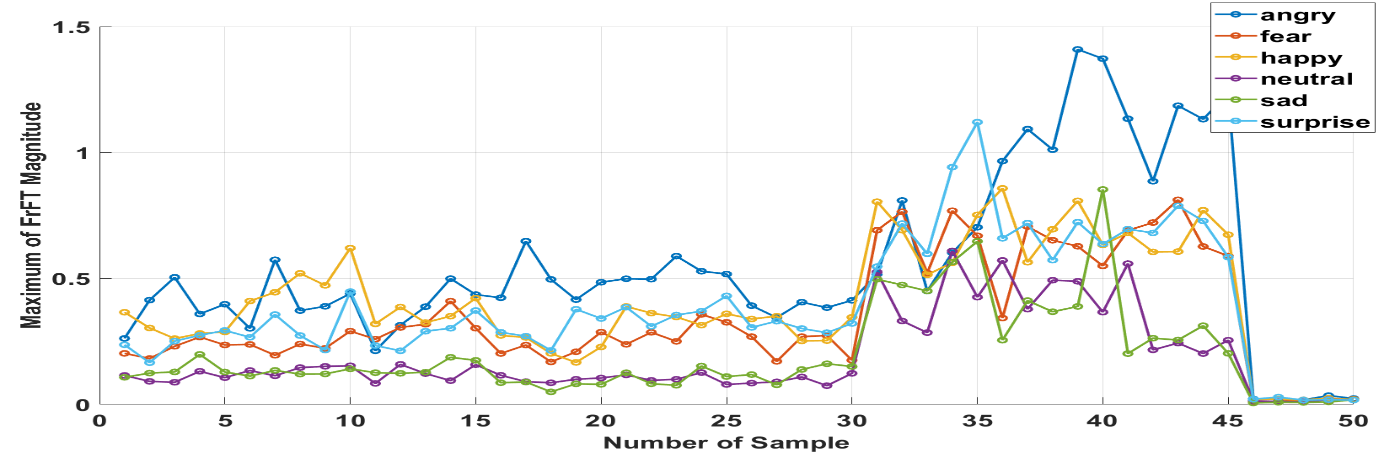
**Figure 4.** Skewness of FrFT magnitude across emotions

Figure 5 illustrates the kurtosis values of the FrFT magnitude across different emotional states using 50 samples per class. It indicates the peakedness or sharpness of the spectral energy distribution—emotions like neutral and sad exhibit high kurtosis, suggesting narrowly concentrated spectral peaks and limited vocal variation. On the contrary, the high arousal states, such as anger, happiness, and surprise, witness lower kurtosis, reflecting more 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.



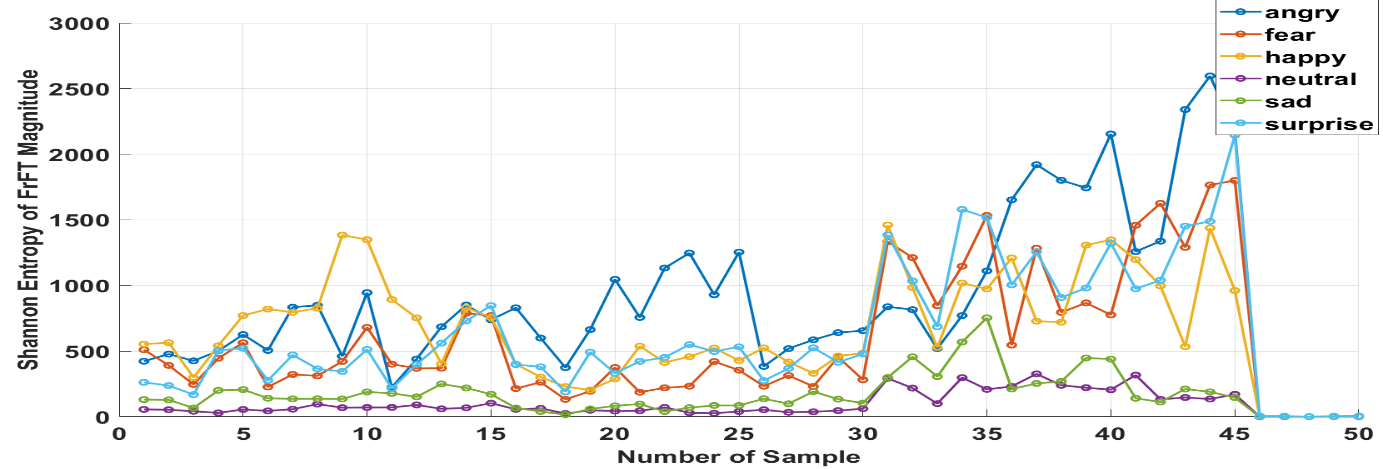
**Figure 5.** Kurtosis of FrFT magnitude across emotions

Figure 6 displays the maximum FrFT magnitude values across emotions, representing the strongest spectral component within each sample. Emotions like anger, surprise, and, to some extent, happiness, show higher peaks, reflecting the presence of sharp vocal bursts or intense articulation. In contrast, neutral and sad maintain lower peak values, consistent with softer, more uniform speech. This feature captures momentary emotional intensity, making it valuable for identifying emotions with sudden or forceful energy patterns.



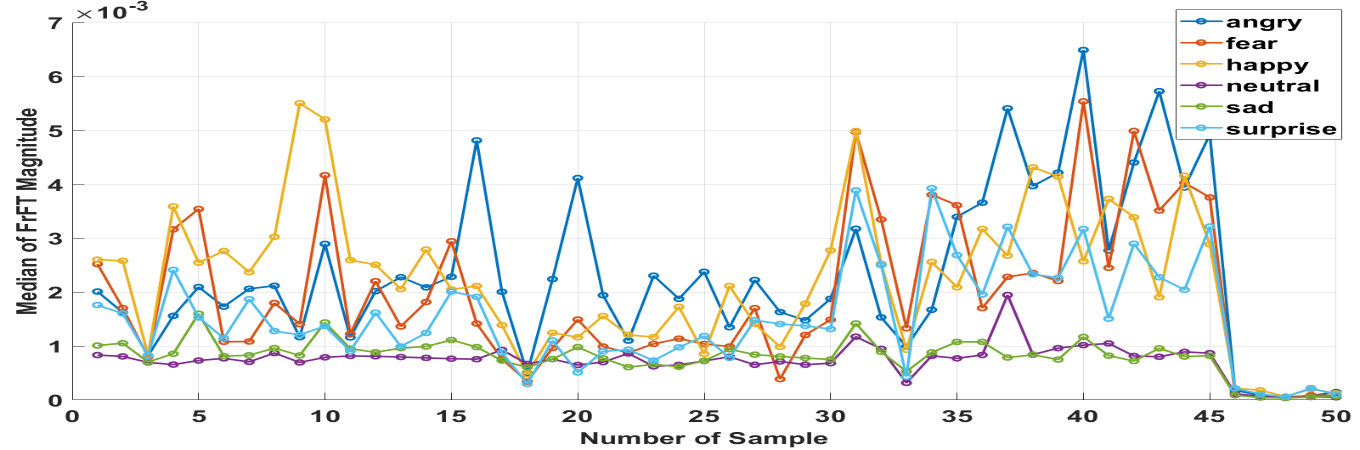
**Figure 6.** Maximum of FrFT magnitude across emotions

Figure 7 presents the Shannon entropy values derived from the FrFT magnitude, reflecting the spectral complexity and energy distribution uniformity. Emotions such as anger, happiness, and surprise exhibit higher entropy, indicating more prosperous and more diverse frequency content. In contrast, neutral and sad show lower entropy, suggesting more predictable and narrowly focused spectral patterns. This feature is effective in distinguishing emotionally rich expressions from calm or monotonous ones by quantifying the amount of information spread across the spectrum.



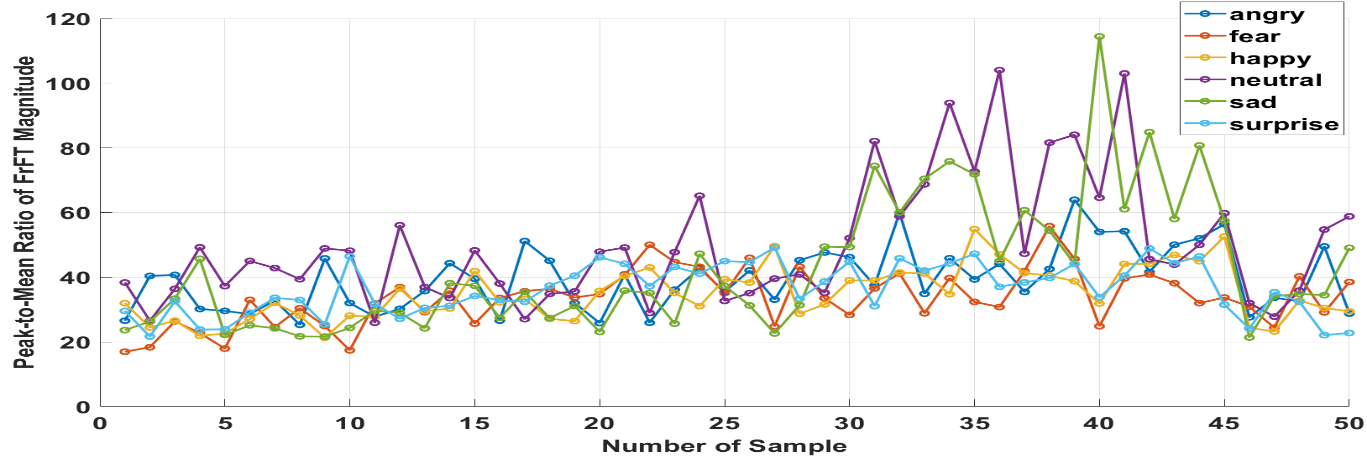
**Figure 7.** Shannon Entropy of FrFT magnitude across emotions

Figure 8 shows the median values of the FrFT magnitude, representing the central tendency of spectral energy in each utterance. Unlike the mean, extreme peaks have less influence on the median, making it helpful in highlighting typical energy levels across emotions. Angry, happy, and surprise generally exhibit higher median values, suggesting a consistently elevated energy floor, while neutral and sad display lower medians, reflecting more subdued speech. This feature supports robust emotion classification, especially when outliers or brief spikes may distort average-based measures.



**Figure 8.** Median of FrFT magnitude across emotions

Figure 9 illustrates the peak-to-mean ratio (PMR) of the FrFT magnitude, highlighting the dominance of the highest spectral peak relative to average energy. Higher PMR values, observed in anger and surprise, suggest the presence of sharp, high-energy bursts characteristic of intense emotional expression. In contrast, neutral and sad display lower PMR, indicating a flatter, more uniformly distributed spectral profile. This feature is particularly effective for detecting emotions marked by sudden energy surges or expressive articulation.



**Figure 9.** PMR of FrFT magnitude across emotions

A clear separation between high and low-arousal emotions is evident from the comparative plots in Figure 2 through Figure 9. Emotions such as anger, happiness, and surprise exhibit consistently higher values in features using mean (Figure 2), standard deviation (Figure 3), maximum (Figure 6), and entropy (Figure 7). This reflects their expressive and dynamic vocal patterns, which involve increased spectral activity and higher variability. These emotions also show elevated peak-to-mean ratios (Figure 9), indicating the presence of sharp energy bursts or pitch accents that rise significantly above the average spectral level.

Conversely, neutral and sad emotions are characterized by low values across most features, such as mean, standard deviation, and entropy, implying steady, monotonous, and energy-restricted articulation. These classes’ exhibit higher skewness (Figure 4) and kurtosis (Figure 5), which indicate narrow-band, low-frequency energy dominance with little variability or traits commonly found in subdued or low-arousal expressions.

The median values (Figure 8) follow trends similar to the mean but offer additional robustness by mitigating the effect of outliers or extreme bursts. This helps highlight the central tendency of speech energy in emotions like anger or surprise, which may include sporadic peaks. The emotion fear exhibits hybrid behavior across features, sometimes aligning with high-arousal traits (e.g., moderate-to-high entropy or PMR) and reflecting low-arousal features other times (e.g., higher kurtosis or skewness). This variability supports the notion that fear, unlike other clearly defined emotions, may be ambiguous in acoustic presentation, leading to potential overlaps in classification.

Table 1 compares several machine learning algorithms' classification accuracy and time using the developed FrFT-based statistical features. The RF demonstrated superior performance, indicating the ensemble methods' potential in HVE.

|  |  |  |
| --- | --- | --- |
| **TABLE 1.** Comparison of recognition accuracy among classifiers | | |
| **Classifier** | **Accuracy (%)** | **Elapsed Time (s)** |
| KNN | 35.28 | 0.93 |
| DA | 33.33 | 0.47 |
| NB | 33.89 | 0.45 |
| DT | 41.94 | 0.38 |
| SVM | 35.28 | 1.63 |
| RF | 46.39 | 1.76 |

Tables 2 to 7 provide the confusion matrix of different classifiers for comparison. A common trend observed across all classifiers is the substantial confusion between happy, fear, and surprise emotions, which often share overlapping acoustic features. For instance, in the KNN matrix, the class surprise is frequently misclassified as happy (17 cases) and fear (19 instances), indicating limited separability among these emotions. Similarly, fear is frequently misclassified as happy or angry across DA, NB, and SVM models. In contrast, neutral and sad emotions tend to be more distinctly classified, particularly in models like NB and RF, where neutral achieves a high actual positive rate with minimal confusion. The RF classifier, which outperformed others with the highest accuracy (46.39%), shows relatively better separation across all classes, especially for angry and sad, suggesting that ensemble methods offer greater robustness in handling inter-class variability. However, all classifiers still show notable confusion in distinguishing high-arousal emotions, highlighting the challenge in classifying subtle emotional nuances using traditional statistical features.

**TABLE 2.** The KNN Confusion Matrix

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Angry** | **Fear** | **Happy** | **Neutral** | **Sad** | **Surprise** |
| Angry | 33 | 12 | 8 | 1 | 0 | 6 |
| Fear | 18 | 24 | 9 | 1 | 1 | 7 |
| Happy | 17 | 18 | 11 | 0 | 1 | 13 |
| Neutral | 5 | 7 | 0 | 32 | 16 | 0 |
| Sad | 8 | 8 | 2 | 16 | 24 | 2 |
| Surprise | 19 | 19 | 17 | 1 | 1 | 3 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TABLE 3.** The DA Confusion Matrix | | | | | | |
|  | **Angry** | **Fear** | **Happy** | **Neutral** | **Sad** | **Surprise** |
| Angry | 29 | 6 | 11 | 1 | 0 | 13 |
| Fear | 5 | 20 | 17 | 5 | 1 | 12 |
| Happy | 12 | 16 | 14 | 1 | 0 | 17 |
| Neutral | 0 | 5 | 1 | 38 | 15 | 1 |
| Sad | 3 | 0 | 0 | 22 | 35 | 0 |
| Surprise | 14 | 15 | 13 | 1 | 2 | 15 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TABLE 4.** The NB Confusion Matrix | | | | | | |
|  | **Angry** | **Fear** | **Happy** | **Neutral** | **Sad** | **Surprise** |
| Angry | 15 | 5 | 7 | 10 | 5 | 18 |
| Fear | 13 | 4 | 6 | 12 | 12 | 13 |
| Happy | 15 | 7 | 10 | 9 | 8 | 11 |
| Neutral | 0 | 0 | 1 | 46 | 13 | 0 |
| Sad | 1 | 0 | 0 | 28 | 30 | 1 |
| Surprise | 14 | 6 | 4 | 11 | 8 | 17 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TABLE 5.** The DT Confusion Matrix | | | | | | |
|  | **Angry** | **Fear** | **Happy** | **Neutral** | **Sad** | **Surprise** |
| Angry | 29 | 6 | 11 | 1 | 0 | 13 |
| Fear | 5 | 20 | 17 | 5 | 1 | 12 |
| Happy | 12 | 16 | 14 | 1 | 0 | 17 |
| Neutral | 0 | 5 | 1 | 38 | 15 | 1 |
| Sad | 3 | 0 | 0 | 22 | 35 | 0 |
| Surprise | 14 | 15 | 13 | 1 | 2 | 15 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TABLE 6.** The SVM Confusion Matrix | | | | | | |
|  | **Angry** | **Fear** | **Happy** | **Neutral** | **Sad** | **Surprise** |
| Angry | 25 | 9 | 8 | 3 | 9 | 6 |
| Fear | 11 | 12 | 7 | 3 | 16 | 11 |
| Happy | 8 | 7 | 13 | 2 | 8 | 22 |
| Neutral | 0 | 2 | 0 | 39 | 18 | 1 |
| Sad | 5 | 6 | 0 | 17 | 30 | 2 |
| Surprise | 11 | 10 | 16 | 2 | 13 | 8 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TABLE 7.** The RF Confusion Matrix | | | | | | |
|  | **Angry** | **Fear** | **Happy** | **Neutral** | **Sad** | **Surprise** |
| Angry | 34 | 2 | 12 | 1 | 1 | 10 |
| Fear | 7 | 22 | 9 | 4 | 4 | 14 |
| Happy | 21 | 10 | 16 | 0 | 0 | 13 |
| Neutral | 0 | 4 | 1 | 40 | 14 | 1 |
| Sad | 2 | 2 | 0 | 19 | 36 | 1 |
| Surprise | 11 | 16 | 12 | 1 | 1 | 19 |

# CONCLUSION

This study proposed a novel framework for demarcating and classifying several emotional states from voice based on FrFT-statistical significances. By leveraging the tunable TF resolution of the FrFT, the method effectively captures emotionally salient spectral patterns often missed by traditional time or frequency domain analyses. The graphical statistical representation across six emotional classes (angry, fear, happy, neutral, sad, and surprise) revealed consistent patterns of emotional differentiation, particularly along the arousal dimension. It can be concluded that the features like entropy, mean, and PMR are particularly effective in differentiating expressive emotions (anger, happiness, surprise). At the same time, skewness and kurtosis highlight the subdued spectral patterns of calm emotions (sadness, neutrality). The analysis further demonstrated that the FrFT-based features provide strong class separability and are well suited for integration into machine learning classifiers such as multi-class SVMs. The features are compact, computationally efficient, and interpretable, enabling insight into how emotional expression maps onto time-frequency energy patterns. However, the study has several limitations, including inter-class overlap in emotions like fear, speaker variability, and the need for dynamic selection of the optimal FrFT order. Despite these constraints, the proposed method offers a robust, low-dimensional, and scalable solution for emotion recognition from voice. Future work will incorporate adaptive FrFT order selection, temporal modeling, and deep learning-based hybrid systems to improve emotion recognition performance in more complex, real-world environments.

# References

1. G. H. Mohmad and R. Delhibabu, IEEE Access **12**, pp. 151122 – 151152 (2024).
2. P. Laukka and H. A. Elfenbein, Emotion Review **13**, 3–11 (2021).
3. 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.
4. S. Aurobindo, R. Prakash, and M. Rajeshkumar, RINENG. **25**, 104561 (2025).
5. S.Heydari, and S. M. Shahidi, Multim.Tools Appl. 1-24 (2025).
6. L. Huang and X. Shen, Electronics **11**, 3393 (2022).
7. N. Pattanaik and H. K. Palo, Speech Commun. **169**, 103187 (2025).
8. B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, “Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing,” in *Affective Computing and Intelligent Interaction*, Lecture Notes in Computer Science Vol. 4738, edited by A. C. R. Paiva, R. Prada, and R. W. Picard (Springer, Berlin, Heidelberg, 2007), pp. 2-13.
9. K. S. Rao, S. G. Koolagudi, and R. R. Vempada, Int. J. Speech Technol. **16**, 143–160 (2013).
10. Rehman, Z. T. Liu, M. Wu, W. H. Cao, and C. S. Jiang, Appl. Acoust. **211**, 109444 (2023).
11. G. Tamulevičius, G. Korvel, A. B. Yayak, P. Treigys, J. Bernatavičienė, and B. Kostek, Electronics **9**, 1725 (2020).
12. H. K. Palo, S. Subudhiray, and N. Das, Speech Commun. **149**, 11–28 (2023).
13. K. Daqrouq, A. Balamesh, O. Alrusaini, A. Alkhateeb, and A. S. Balamash, Appl. Comput. Intell. Soft Comput. **2024**, Article ID 7184018, 12 pages.
14. W. A. Khan, H. ul Qudous, and A. A. Farhan, Multim. Tools Appl. **83**, 75557–75584 (2024).
15. M. N. Mohanty and H. K. Palo, *Int. J. Speech Technol.* **22**, pp. 865-884 (2019).
16. P. Jackson and S. Haq, *Surrey Audio-Visual Expressed Emotion (SAVEE) Database*, University of Surrey, Guildford, UK (2014).