Emotional Cognition in Robotics Through Voice Tone Analysis

Hisham Iqbal Khokhar

***Abstract*—Enhancing human-robot interactions through emotional cognition is pivotal for creating intuitive and responsive robotic systems. This study investigates the feasibility of utilizing voice tone as a reliable indicator of emotional states, specifically distinguishing between angry and neutral emotions. Building on existing research in affective computing and benchmark machine learning models for real-time emotion detection, we employ key acoustic features such as Spectral Centroid Mean, Spectral Bandwidth Mean, RMS Mean, and Zero-Crossing Rate Mean to analyze vocal commands. Our experimental design involved analyzing two kinds of audio samples, angry and neutral toned, which were then normalized and examined through PCA and evaluated on effective meachine learning models. The results demonstrate significant differences in acoustic metrics, with the angry sample exhibiting higher spectral centroid and bandwidth values, as well as increased loudness and noisiness compared to the neutral sample. These findings indicate that distinct vocal features can effectively signal emotional states, enabling robots to adapt their responses accordingly. Such capabilities are crucial for applications in healthcare, customer service, and personal assistance, where empathetic and context-aware interactions enhance the overall user experience. This research contributes to the development of emotionally intelligent robotic systems, bridging the gap between functional support and meaningful engagement in human-robot collaborations.**

I. Introduction

Robots' abilities to detect and respond to human emotions have gained significant interest, especially in improving human-robot interaction. The robots' accuracy and ability to fully understand and analyze voice tones could transform fields like healthcare and customer service. Robots that can effectively identify emotional signals, such as distress or worry, can then adapt their behavior to offer both practical assistance and meaningful interaction with the user. For example, in a healthcare environment, robots could monitor patients' emotional states to detect any discomfort or distress, subsequently alerting medical staff. Additionally, in education, an emotionally intelligent robot can detect when a student's engagement or focus drops, allowing it to adapt and help the student with different teaching methods.

This paper explores the potential of voice tone analysis in enabling robots to precisely recognize emotional states. We specifically investigate the differences between neutral and angry voice tones. This research analyzes acoustic features obtained from audio recordings and applies them to emotion recognition in robotics. Focusing on angry and neutral tones provides a strong understanding of how emotionally intelligent robots can distinguish between emotionally charged and calm states. This capability is crucial in applications like customer service, where the robot must accurately detect and respond to varying emotional expressions.

II. Background Research

Emotional cognition in robotics has advanced significantly, with voice tone becoming a crucial factor for emotion recognition. This paper focuses on identifying angry and neutral emotions through key acoustic features, such as Spectral Centroid Mean (brightness of sound), RMS Mean (representing loudness), and Zero-Crossing Rate Mean (signal noisiness). These features were extracted from a dataset of 5-second audio clips sourced from platforms like YouTube, fictional performances, and public speeches. This approach ensured standardized analysis and labeling of the two tones, guaranteeing high-quality data.

Studies on emotion recognition during conversations highlight the potential of analyzing vocal characteristics, such as pitch and tone, to enhance robots' emotional understanding and responsiveness. Research has demonstrated the efficiency of convolutional neural networks (CNNs) in real-time emotion recognition, achieving high precision across various emotional states [1]. These findings underscore the feasibility of applying machine learning techniques to train robotic systems in emotional detection.

The experiment conducted in this study builds on these advancements. The dataset was analyzed using Python-based extraction methods to measure the brightness, intensity, and noise levels of the audio clips. The results revealed significant differences in these features between angry and neutral tones, confirming that emotional states can be consistently recognized through voice analysis. The higher brightness and intensity observed in angry tones imply that robots equipped with this understanding can recognize emotional shifts and respond accordingly. This ability could be further utilized in areas like negotiation and mental health support, where emotional states may constantly change.

The findings indicate that robots could leverage this capability to respond more efficiently, thereby improving interactions with individuals and enhancing their practical utility in real-world situations. However, emotionally intelligent robots may face challenges in detecting cultural differences, as variations in speech patterns, tone levels, and facial expressions can influence emotion recognition. Therefore, further research should address these aspects to ensure comprehensive emotional recognition across diverse cultural contexts.

III. Experiment

A. Data Collection/Labeling

For the data collection process, a diverse variety of audio recordings were gathered from online sources such as YouTube. These audio recordings encompassed a wide range of formats, including fictional performances, interviews, public recordings, speeches, and more. To ensure consistency and facilitate subsequent processing, each audio recording was segmented into 5-second clips. This segmentation allowed for standardized analysis and labeling. Afterward, Python was used for automated feature extraction.

B. Feature Extraction

To effectively analyze emotional states from voice clips, we extracted a set of acoustic features that capture various aspects of the audio signals. These features were selected based on their relevance in capturing emotional nuances and their effectiveness in distinguishing between different emotional tones. The following descriptions outline each feature, its purpose, and the rationale behind its selection:

Spectral Centroid Mean (Overall Brightness): For this dataset, the mean spectral centroid is the average position of the center of mass of the spectrum across the entire duration, reflecting how bright or dull the audio clip sounds overall.

Spectral Bandwidth Mean (Consistency in Frequency): The mean spectral bandwidth represents the average spread of frequencies over time, indicating how consistently the audio clip varies in frequency content.

RMS Mean (Loudness): The RMS mean for a 5-second clip represents the average energy level across the duration, providing an indication of the clip's perceived loudness.

Zero-Crossing Rate Mean (Smoothness): The ZCR mean is the average number of zero crossings per frame, commonly used for distinguishing tonal (smooth) from non-tonal (noisy) sounds.

Spectral Contrast Mean (Harmony): The mean spectral contrast averages the differences between peaks and valleys in the spectrum over time, providing insight into the audio's harmonic content.

Pitch Mean (Tonal Quality): The pitch mean is the average perceived frequency over the entire duration of the audio clip, representing the overall tonal quality.

Pitch Confidence Mean (Pitch Estimation): The mean pitch confidence reflects the average certainty of pitch estimation over time, with higher values suggesting clear, well-defined pitches.

Mel Spectrogram Mean (Frequency Distribution): The Mel spectrogram mean is the average energy across all Mel bins for a 5-second clip, providing an overall representation of the audio's frequency distribution.

Mel Spectrogram Variance (Consistency in Frequency): This captures the variability in the Mel spectrogram over time, indicating how dynamic or consistent the audio is in its frequency content.

Energy Mean (Intensity/Loudness): The energy mean is the average power across the duration of a 5-second clip, giving an overall measure of the clip's intensity or loudness.

C. Data Analysis

The objective of the data analysis was to see if the mean of the normalized numerical categories: Spectral Centroid Mean, Spectral Bandwidth, RMS Mean, Zero-Crossing Rate Mean, Spectral Contrast Mean, Pitch Mean, Pitch Confidence Mean, Mel Spectrogram Mean, Mel Spectrogram Variance, and Energy Mean would be different from their non-normalized counterparts.

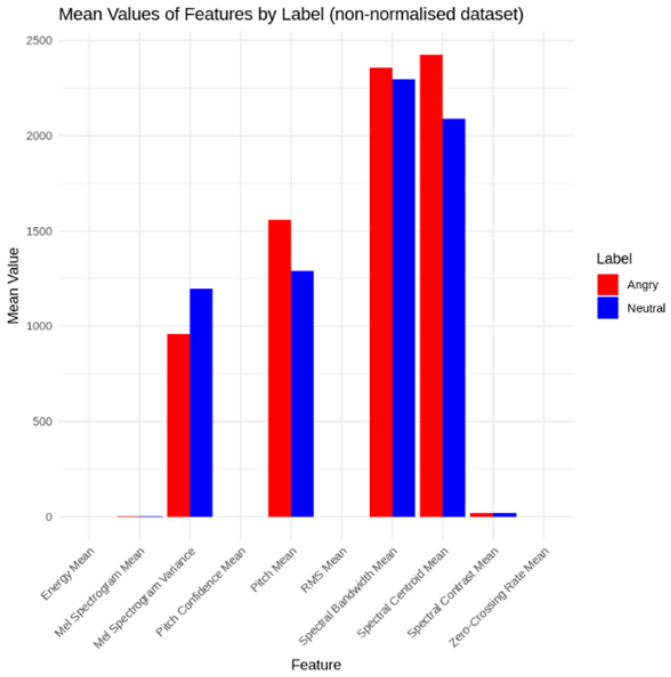

Fig. 1. The means of each numerical category in the non-normalized dataset.

Initially, the dataset was shuffled randomly to prevent any ordering effects before the graphs (Fig. 1 and Fig. 2) were graphed. Before graphing Fig. 2, the data was encoded using min-max normalization to scale their values between 0 and 1 [2].

Upon a thorough analysis of the correlation matrix, the following insights can be registered:

1) **Highly Correlated Features:**
   a) **Spectral Centroid Mean and Zero-Crossing Rate Mean (0.95):**
      - These features exhibit a very high positive correlation, indicating that as the spectral centroid mean increases, the zero-crossing rate mean also increases.
      - This is expected as the spectral centroid mean is a measure of the center of mass of the spectrum,

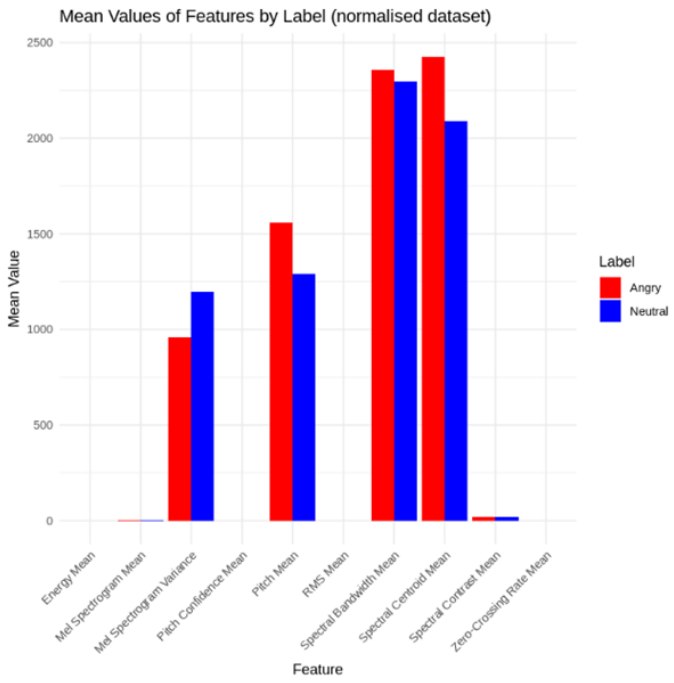

Fig. 2. The means of each numerical category of the normalized dataset.

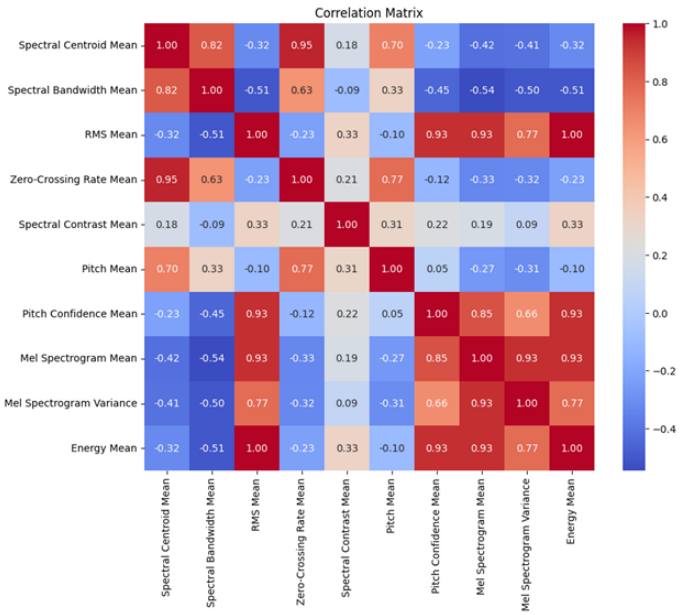

Fig. 3. Correlation matrix of the features

while the zero-crossing rate mean is a measure of the number of times the signal crosses the zero axis.

b) **RMS Mean and Pitch Confidence Mean (0.93):**

- This strong correlation suggests that energy-related features like RMS are closely tied to how confidently a pitch can be detected, likely due to strong harmonic content in the data.

c) **Energy-Based Features:**

- RMS Mean, Mel Spectrogram Mean, and Mel Spectrogram Variance are all highly interrelated, with correlations exceeding 0.9, which suggests these features capture similar underlying properties of the signal and might exhibit redundancy.

2) **Low or Negative Correlations:**

a) **Spectral Bandwidth Mean and Spectral Contrast Mean (-0.09):**

- These features have almost no correlation, indicating they capture independent characteristics of the data.

b) **Pitch Mean and Zero-Crossing Rate Mean (-0.12):**

- This slight negative correlation suggests a weak or inverse relationship between pitch and zero-crossing rate, demonstrating that they represent different aspects of the signal.

3) **Clusters of Relationships:**

a) **Energy-Based Features:**

- RMS Mean, Mel Spectrogram Mean, Mel Spectrogram Variance, and Energy Mean form a highly correlated cluster, which suggests that these features likely represent the same underlying property of the signal and may be candidates for dimensionality reduction techniques like PCA.

b) **Frequency-Related Features:**

- Spectral Centroid Mean, Spectral Bandwidth Mean, and Zero-Crossing Rate Mean show moderate to high correlations among themselves, implying they collectively capture spectral characteristics.

After analyzing the correlation matrix, we decided to perform PCA in order to optimize our workflow and reduce the computational complexity of running the models.

In analyzing the scree plot, we can observe a sharp decline in the explained variance for the first two principal components, which indicates that these two components capture the majority of the variance in the dataset. After PC3, the curve flattens considerably, suggesting diminishing returns in the variance explained by the subsequent components.

Regarding cumulative variance, the first two principal components account for approximately 80.33% of the total variance, which means they are sufficient to represent most of the information in the dataset. However, including the first four components brings the cumulative variance to 94.67%, enhancing representational efficiency with minimal loss of information. Beyond the fifth principal component, the addi-

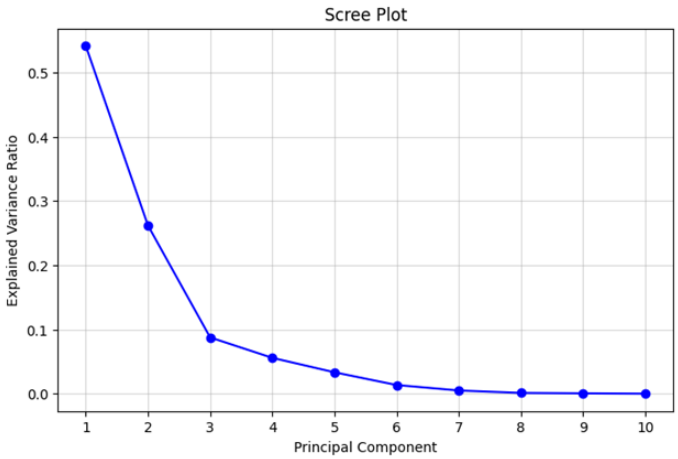

Fig. 4. Scree plot of the explained variance of the principal components

Principal Component (PC)	Cumulative Variance Explained (%)
PC1	54.15%
PC2	80.33%
PC3	89.09%
PC4	94.67%
PC5	97.99%
PC6	99.34%
PC7	99.84%
PC8	99.95%
PC9	100.00%
PC10	100.00%

Fig. 5. Table showing the explained cumulative variance of the principal components

tional components contribute marginally to the total variance (less than 3%), which suggests they can be excluded to reduce dimensionality without significantly losing information.

Based on the ”elbow” observed in the scree plot and the cumulative variance analysis, we decided that retaining the first 4 principal components appears to be an optimal choice for dimensionality reduction. This approach strikes a balance between computational efficiency and the retention of key information from the dataset.

Now that we have identified our principal components, we focused on selecting the most suitable model to achieve our goals. To do so, we first assessed the performance and interpretability of the major algorithms available.

Models like linear regression and decision trees are highly interpretable but offer moderate performance, making them ideal for applications requiring transparency. As we move towards more complex models like ensemble methods and neural networks, performance improves significantly, but interpretability declines due to their complexity and ”black-box”

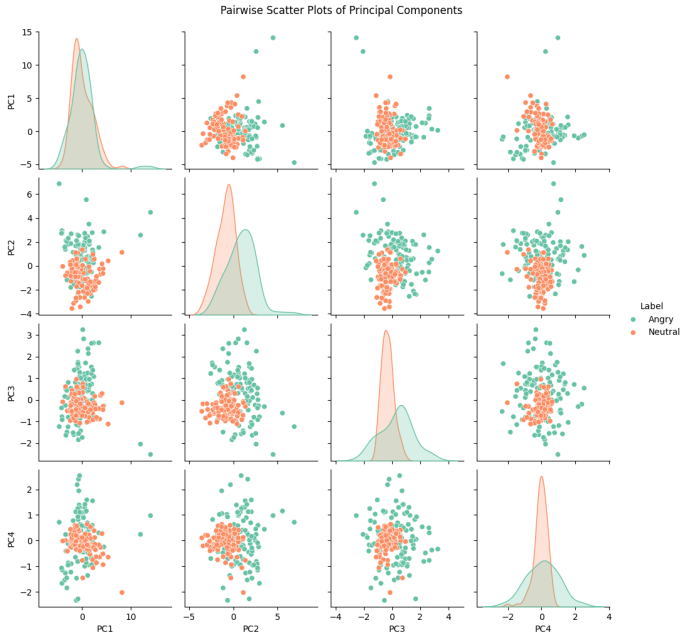

Fig. 6. Interrelation between each PC (PC1 - PC4)

nature [3].

Given that, we decided to evaluate our dataset containing audio features using four different models to assess their performance, keeping in mind their interpretability in order to understand the outcomes: K-Nearest Neighbors (KNN), Logistic Regression, Naïve Bayes, and Random Forest.

IV. Results

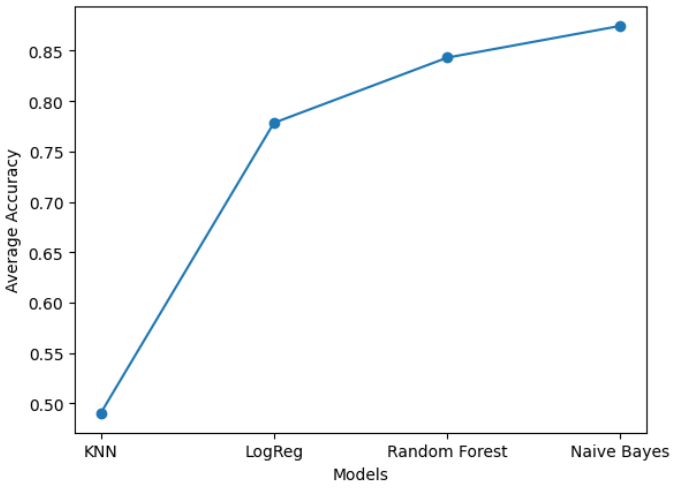

Fig. 7. Accuracy of different models on the audio feature dataset

From our analysis, we observed distinct performance differences across the four models tested on our audio feature dataset.

K-Nearest Neighbors (KNN): Performed the worst with an accuracy of 50%, likely due to the high dimensionality

of audio features and the lack of clear data clusters, which hindered its ability to classify effectively.

Logistic Regression: Showed a significant improvement with an accuracy of 75%, leveraging the dataset's linearly separable patterns. However, it struggled to model complex, non-linear relationships.

Random Forest: Achieved strong performance with an accuracy of 83%, demonstrating its ability to handle non-linear patterns and interactions. Its ensemble approach effectively captured the complexity of the audio features.

Naïve Bayes: Delivered the best performance at 85%, slightly outperforming Random Forest. This suggests the dataset aligns well with its conditional independence assumption, making it particularly effective for this task.

V. Conclusion

This study undertook a systematic approach to classify audio features by combining detailed exploratory data analysis, dimensionality reduction, and the evaluation of diverse machine learning models. The correlation matrix analysis revealed significant redundancies among energy-based and frequency-related features, motivating the application of Principal Component Analysis (PCA) to optimize the feature set. By retaining the first four principal components, which accounted for 94.67% of the total variance, we effectively reduced the dimensionality while preserving the critical information necessary for accurate classification.

The evaluation of models—K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, and Naïve Bayes—revealed distinct performance characteristics. Logistic Regression demonstrated its strength in identifying linearly separable patterns, achieving moderate accuracy, while Random Forest excelled in capturing non-linear relationships and complex feature interactions, with an accuracy of 83%. Notably, Naïve Bayes outperformed all other models with an accuracy of 85%, likely due to the dataset's favorable alignment with its conditional independence assumption. On the other hand, KNN struggled to perform effectively in the high-dimensional feature space, underscoring the importance of dimensionality reduction and appropriate model selection.

These outcomes highlight the critical role of feature engineering, dimensionality reduction, and model evaluation in the development of robust machine learning pipelines for audio feature classification. While Naïve Bayes proved to be the most effective model in this context, the interpretability and adaptability of other models like Logistic Regression and Random Forest suggest their utility in scenarios requiring specific trade-offs between transparency and predictive power.

Future research could explore optimization techniques, such as hyperparameter tuning, ensemble stacking, or the incorporation of deep learning architectures to further enhance performance. Additionally, applying this methodology to more diverse and larger audio datasets could provide broader insights into its generalizability and scalability in real-world applications.

VI. Source Code

- **Experimental Analysis**: View Code

References

[1] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," 2019. [Online]. Available: https://arxiv.org/abs/1905.02947

[2] D. Gunay, "Feature encoding," Aug 2023. [Online]. Available: https://medium.com/@denizgunay/feature-encoding-f099a6c1abe8

[3] S. Maarek, "Aws ai practitioner certified," n.d. [Online]. Available: https://www.udemy.com/course/aws-ai-practitioner-certified/?couponCode=ST19MT121224

***Index Terms*—Human-Robot Interaction, Emotional Cognition, Voice Tone Analysis, Acoustic Features, Machine Learning, Principal Component Analysis, Naïve Bayes, Random Forest.**