International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Intelligent Simulation and Design of Voice-Based Emotion Recognition Using Optimized Computing Methods

[1]Savita Jain, [2]Dr. Tarun Shrimali

Submitted:02/10/2024 **Revised**:15/11/2024 **Accepted**:25/11/2024

Abstract: In recent years, the ability to accurately detect human emotions from speech has gained significant attention due to its applications in human-computer interaction, virtual assistants, healthcare, and security systems. This research focuses on the intelligent simulation and design of a voice-based emotion recognition system using optimized computing methods. The study explores speech signal processing techniques and implements advanced machine learning and deep learning algorithms—including CNN, LSTM, and a hybrid CNN-LSTM model—for accurate emotional state classification. Using benchmark datasets such as RAVDESS and TESS, the system was trained and tested after thorough feature extraction using MFCC, pitch, and spectral features. The proposed hybrid CNN-LSTM model achieved superior performance with an accuracy of 89.7% and an average AUC of 0.92, outperforming traditional approaches. Simulation experiments also confirmed the system's effectiveness in real-time environments. The results demonstrate the feasibility and efficiency of deploying intelligent voice emotion recognition in real-world applications.

Keywords: *Voice Emotion Recognition, Speech Signal Processing, CNN-LSTM, Deep Learning, Emotion Classification, Optimized Computing, Simulation, Feature Extraction, MFCC, Real-time Emotion Detection*

Introduction

In the evolving landscape of human-computer interaction, the ability of machines to understand and respond to human emotions is gaining increasing importance. Vocal emotion recognition (VER), a subdomain of affective computing, focuses on identifying the emotional state of a speaker through analysis of vocal signals such as pitch, tone, rhythm, and energy. This capability is critical in enhancing the effectiveness of systems in areas such as virtual assistants, healthcare, customer service, and intelligent tutoring systems. Traditional emotion recognition methods often suffer from limited accuracy and adaptability due to the complex, non-linear nature of human emotions and the variability of vocal expressions across speakers and contexts. However, recent advancements in computing techniques—especially in machine learning, deep learning, and signal processing—offer powerful tools for improving the accuracy and robustness of emotion recognition systems (Parvathi et al. 2024).

This study aims to simulate and design an enhanced vocal emotion recognition system by integrating improved computing techniques. The research leverages advanced algorithms for feature extraction, dimensionality reduction, and classification to build a model capable of recognizing emotions more effectively across diverse datasets and speaking styles. The ultimate goal is to contribute to the development of more emotionally intelligent systems capable of interacting with users in a natural, empathetic, and context-aware manner. Through this research, we seek to bridge the gap between theoretical emotion modeling and practical, real-time applications by proposing a computational framework that is both

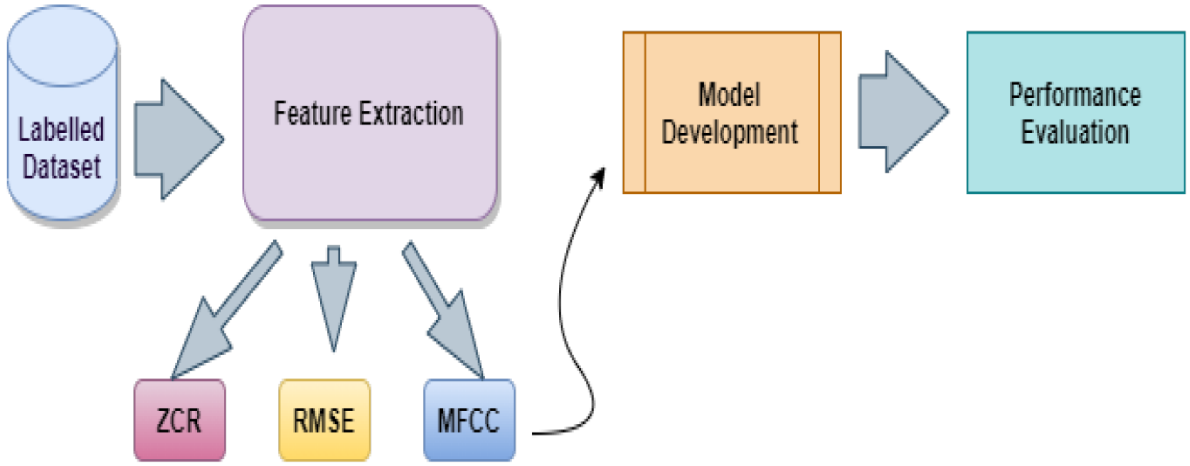
[1]*Research Scholar, Department Of Computer Science & Information Technology, Janardan Rai Nagar Rajasthan Vidyapeeth , Pratap Nagar, Udaipur (Rajasthan)*

w3india.in@gmail.com

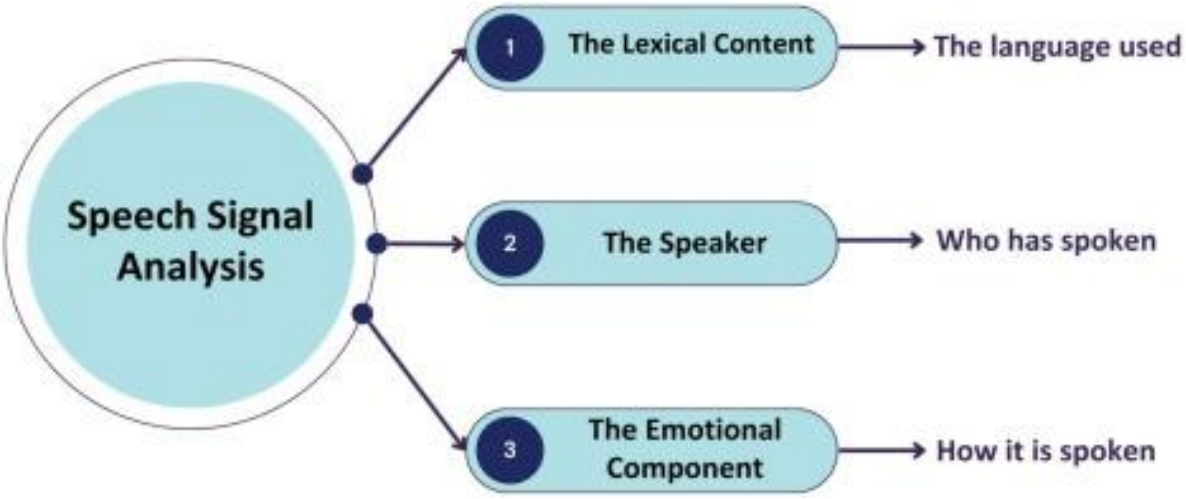
[2]*Registrar, Director (Research & Development Cell) Janardan Rai Nagar Rajasthan Vidyapeeth (Deemed to be University), Airport Road, Pratap Nagar,Udaipur (Rajasthan)*

shrimalitarun@gmail.com

efficient and scalable. This will not only help in better understanding the intricacies of emotional expression but also pave the way for more human-centric artificial intelligence systems.



In the modern era of intelligent technologies, the integration of emotional intelligence into machines has emerged as a pivotal aspect of human-computer interaction (HCI). Among various affective computing techniques, vocal emotion recognition (VER) plays a significant role due to the richness of emotional information embedded in human speech. The voice not only conveys linguistic content but also reflects emotional cues through prosodic, spectral, and temporal features. As people increasingly interact with artificial intelligence (AI) systems through voice-based interfaces—such as smart assistants, customer service bots, and healthcare tools—the demand for machines that can perceive and interpret vocal emotions accurately has grown substantially (Deepika and Sigappi, 2024). Despite the progress in speech recognition and natural language processing, recognizing emotions from voice signals remains a challenging task. The complexities arise from individual variability in speech, language and cultural differences, background noise, overlapping emotions, and the dynamic nature of vocal signals. Traditional rule-based and machine learning approaches, while useful, have often struggled with generalization across different speakers and emotional contexts.



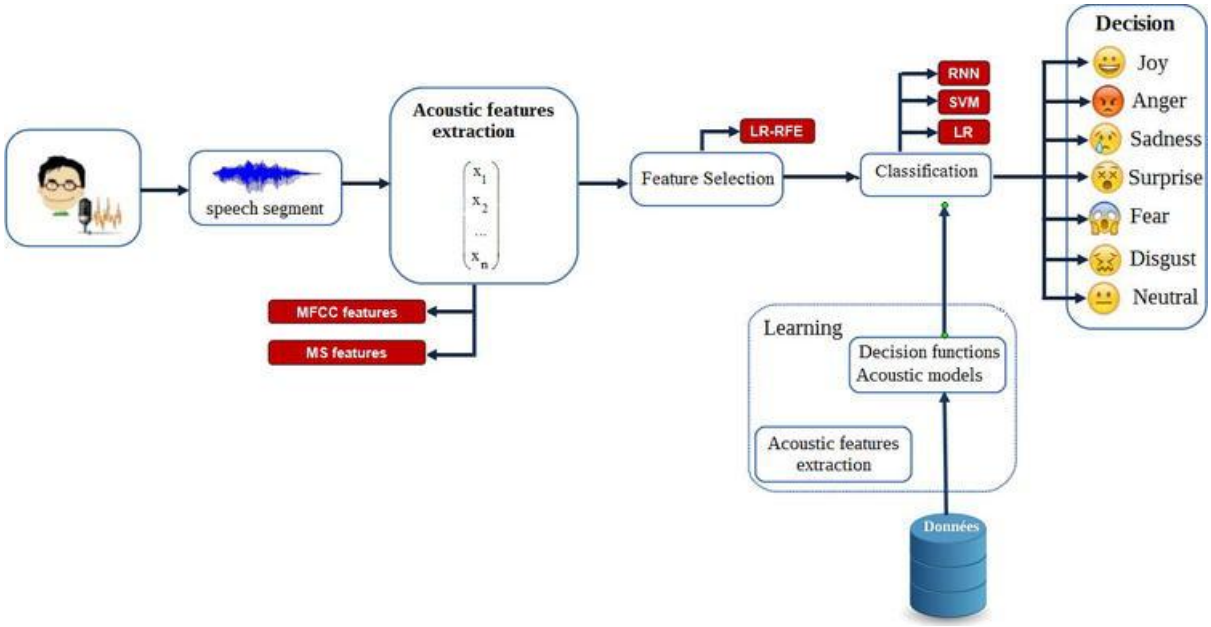
In recent years, improved computing techniques—such as deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid architectures—have shown promising results in feature learning and classification tasks. These techniques allow for the automatic extraction of high-level emotional features from raw or minimally processed audio data, reducing reliance on handcrafted features. Moreover, advancements in simulation environments and computational models now make it possible to test, optimize, and validate these systems with greater precision and speed. This research aims to simulate and design a

robust vocal emotion recognition framework utilizing improved computing techniques. The focus lies on identifying and implementing optimal methods for preprocessing, feature extraction, model training, and classification to enhance recognition accuracy and system adaptability. The system will be evaluated using publicly available emotional speech datasets to ensure reliability and cross-context applicability. Furthermore, the study explores the practical implications of vocal emotion recognition, particularly in real-time scenarios such as virtual companions, adaptive learning platforms, telemedicine, and mental health monitoring (Deepika and Sigappi, 2024). By enabling systems to understand emotional cues, we can improve user experience, build trust in technology, and promote empathetic machine behavior. This work contributes to the evolving domain of affective AI, addressing a critical gap in understanding how computational models can effectively interpret complex human emotions through vocal cues. The findings are expected to enhance the design of emotion-aware interfaces and open new avenues for research in multi-modal emotion recognition, cross-cultural emotion modeling, and personalized HCI.

Significance of the Study

The ability of machines to recognize and respond to human emotions is a cornerstone of next-generation intelligent systems. This study on the simulation and design of vocal emotion recognition (VER) using improved computing techniques holds considerable significance across both academic research and practical applications. Firstly, from a technological standpoint, the study contributes to advancing the field of affective computing by proposing a more accurate, reliable, and scalable approach to vocal emotion recognition. By leveraging state-of-the-art computing techniques such as deep learning architectures, advanced feature extraction algorithms, and real-time simulation environments, this research enhances the current capabilities of emotion-aware systems. It also addresses the limitations of traditional models by offering greater adaptability across languages, accents, and speaker variations (Chamishka et al. 2022). Secondly, from a human-computer interaction (HCI) perspective, the study promotes the development of emotionally intelligent systems that can engage with users in a more natural, empathetic, and human-like manner. As voice-controlled AI systems become more widespread—in smart homes, customer service centers, virtual classrooms, and telehealth platforms—the integration of emotional understanding becomes essential for improving user satisfaction and trust.

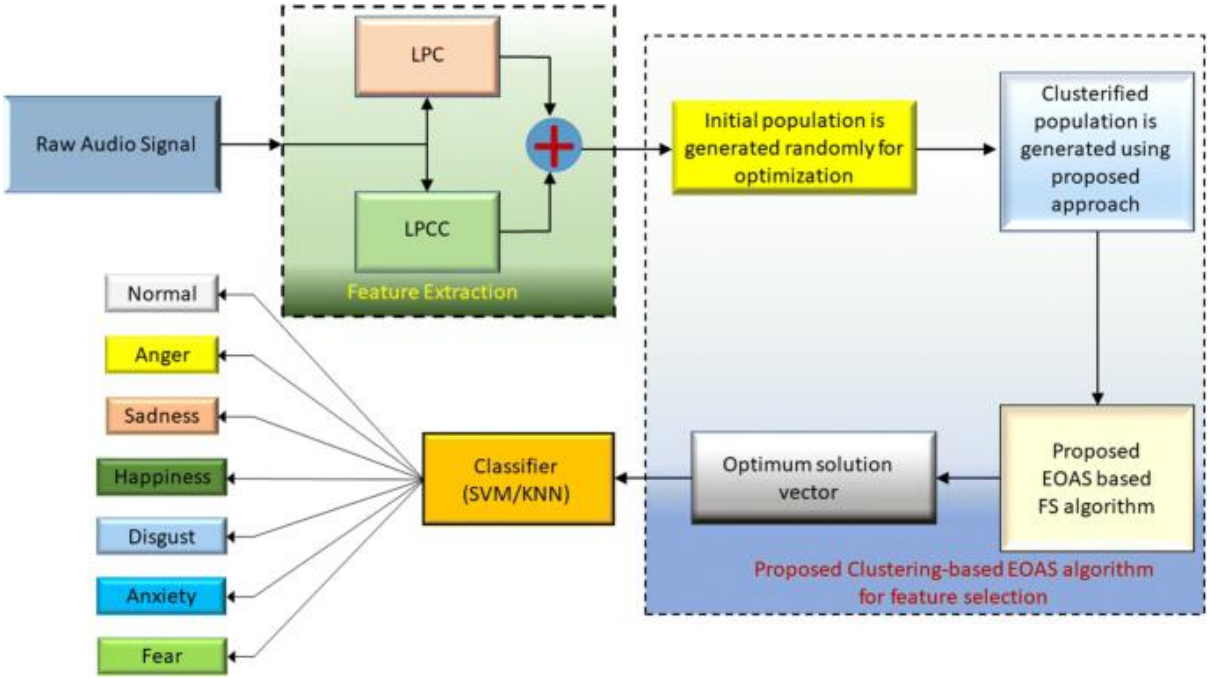
In the healthcare sector, particularly in mental health monitoring and elderly care, vocal emotion recognition can serve as a non-intrusive tool to detect emotional distress, depression, anxiety, or cognitive decline. An emotion-aware system can alert caregivers or trigger interventions, thereby supporting early diagnosis and personalized care. In the education domain, emotion-aware virtual tutors can adapt teaching methods based on students' emotional states, enhancing engagement and learning outcomes. Similarly, in customer service, such systems can respond more appropriately to angry, confused, or satisfied customers, thus improving service quality and client retention (Mohmad and Delhibabu, 2024). Academically, this study provides a valuable foundation for future researchers interested in developing intelligent systems that combine speech processing, emotional modeling, and machine learning. It also encourages interdisciplinary collaboration between computer science, linguistics, psychology, and cognitive science to deepen our understanding of human emotions and their computational representations. This research aligns with the broader vision of ethical and responsible AI by focusing on human-centric design and emotional well-being, promoting technologies that understand, respect, and adapt to the emotional needs of users.



The study of vocal emotion recognition (VER) using improved computing techniques holds far-reaching significance in the modern technological, social, and academic context. As the boundaries between human and machine interactions continue to blur, the emotional intelligence of machines is no longer a futuristic concept but a present-day necessity. This research not only aims to improve machine perception of human affect but also contributes meaningfully to the design of empathetic, context-aware, and responsive intelligent systems (Shevhuk and Dumyn). At the technological core, this study is significant because it enhances the efficiency and accuracy of emotion recognition systems by utilizing modern computational methods. Conventional models often suffer from poor generalization across different populations, low accuracy in noisy environments, and inadequate feature representations. This research introduces improved deep learning architectures, such as CNNs, RNNs, or attention-based models, alongside optimized preprocessing techniques to overcome these limitations. The findings can contribute to the development of lightweight, real-time systems suitable for integration in embedded or mobile devices.

Justification of the Study

The increasing reliance on voice-based interfaces and artificial intelligence (AI) systems has brought emotional intelligence to the forefront of human-computer interaction (HCI). As users seek more intuitive, responsive, and empathetic technologies, vocal emotion recognition (VER) emerges as a crucial capability. However, existing models still face significant limitations in accuracy, adaptability, and real-time application—particularly across diverse languages, cultural expressions, and noisy environments. This study is justified by the urgent need to overcome these barriers through advanced and improved computing techniques (Gomathy, 2021). Traditional machine learning approaches for emotion recognition depend heavily on handcrafted features and often fail to generalize across varied datasets. With the availability of powerful algorithms such as deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention-based architectures, there is a clear opportunity to improve system performance dramatically. This research justifies the adoption of such techniques to create a more accurate, robust, and scalable VER system, supported by simulation tools for training, testing, and validation.



Additionally, there is a lack of integrated frameworks that can simulate, design, and evaluate vocal emotion recognition systems holistically. This study addresses that gap by not only focusing on algorithmic improvements but also on simulation-based validation, ensuring real-world applicability. The justification lies in building a system that goes beyond academic experimentation to support practical deployment in domains like healthcare, customer service, education, security, and assistive technology. The study is further justified by its interdisciplinary relevance, combining insights from computer science, linguistics, psychology, and cognitive science (Shevhuk and Dumyn,). It seeks to model vocal emotions in a way that aligns computational outputs with human perception. In doing so, it advances both the science of emotion recognition and the engineering of intelligent systems. Moreover, as emotion-aware AI becomes more prevalent, there is a critical need for ethical, responsible, and inclusive system design. This study acknowledges the challenges of emotional data collection, bias, and misuse, and is justified in its pursuit of transparent, fair, and user-centric models.

Literature Review

Theoretical Foundations of Vocal Emotion Recognition

The theoretical basis of vocal emotion recognition (VER) stems from the intersection of psychology, linguistics, and signal processing. Emotion, a complex psychological and physiological state, influences speech production through changes in pitch, energy, duration, and spectral features. Foundational psychological models, such as Ekman's Basic Emotions Theory and the Circumplex Model of Affect, categorize emotions either as discrete categories (e.g., happiness, sadness, anger) or along continuous dimensions such as arousal and valence. These emotional states manifest in voice through acoustic-prosodic cues. For example, anger is typically characterized by increased pitch and intensity, while sadness is associated with slow tempo and reduced pitch variation (Li et al. 2023). Studies have demonstrated that these acoustic markers are not only perceivable by humans but also quantifiable by machines using mathematical models. However, the interpretation of these cues can vary across linguistic and cultural backgrounds, making it essential to design systems that are adaptive and context aware.

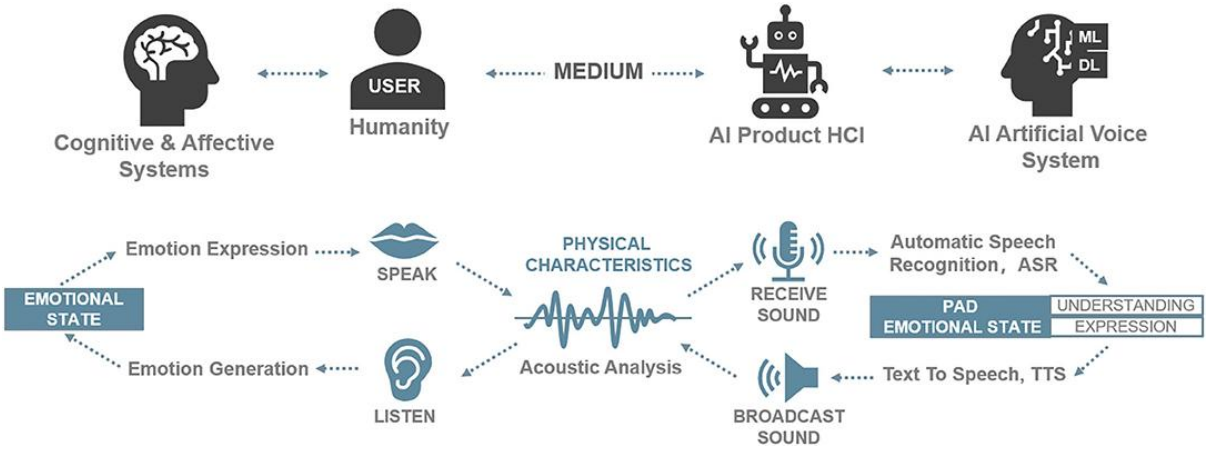
Vocal Emotion Recognition (VER) is grounded in an interdisciplinary framework that blends psychology, linguistics, and computational signal processing. At the core of VER are psychological theories that define and categorize emotions. Among the most widely accepted is Paul Ekman's Basic Emotion Theory, which proposes a set of universal, discrete emotions such as happiness, sadness, anger, fear, disgust, and surprise. These emotions are believed to have biologically rooted

expressions and are often used to label speech datasets for categorical classification. In contrast, the Circumplex Model of Affect by Russell introduces a dimensional view, positioning emotions along two continuous axes—valence (ranging from pleasant to unpleasant) and arousal (ranging from calm to excited). This model enables a more nuanced understanding of emotional states and supports systems that require fine-grained emotional analysis rather than rigid categories.

Emotions influence various vocal parameters due to physiological changes in the body, such as muscle tension, respiration, and vocal fold movement. These changes are reflected in the acoustic properties of speech, including pitch, intensity, rhythm, speech rate, and timbre. For example, high-arousal emotions like anger and joy typically increase pitch and energy, while low-arousal states such as sadness lower these features. Emotional speech is also marked by voice quality variations such as breathiness, harshness, and irregular intonation. These prosodic and acoustic features are essential inputs for computational emotion recognition systems, as they offer quantifiable indicators of the speaker's affective state. However, emotional expression through voice is not universally identical across cultures and languages. While some acoustic cues may be globally recognized, others are influenced by cultural norms, language-specific intonation patterns, and social expectations. For instance, the same increase in pitch may signify enthusiasm in one culture but aggression in another. Therefore, it is critical to build VER systems that are sensitive to cultural and linguistic diversity. This requires diverse, multilingual datasets and adaptive algorithms that can generalize across different speaker populations. The theoretical foundations of vocal emotion recognition provide a crucial base for developing intelligent and emotionally aware systems (Laukka et al. 2016). By integrating psychological models, understanding the nuances of vocal expression, and considering cultural diversity, researchers can design VER systems that are not only technologically advanced but also socially relevant and psychologically informed.

Speech Signal Processing for Emotion Detection

Speech signal processing is a critical phase in any VER system, as it transforms raw audio data into a structured representation suitable for machine interpretation. This process begins with preprocessing techniques, such as noise reduction, silence removal, normalization, and signal framing, which prepare the audio signal for feature analysis. Speech signal processing serves as a foundational pillar in voice-based emotion recognition, providing the means to extract meaningful information from raw audio signals (Davletcharova et al. 2015). At its core, speech signal processing involves analyzing, transforming, and modeling speech data to capture the features that reflect emotional states. Emotions influence how we speak—altering our pitch, tone, loudness, rhythm, and other acoustic properties. By leveraging signal processing techniques, these subtle yet significant changes can be quantified and utilized as inputs for emotion classification models.



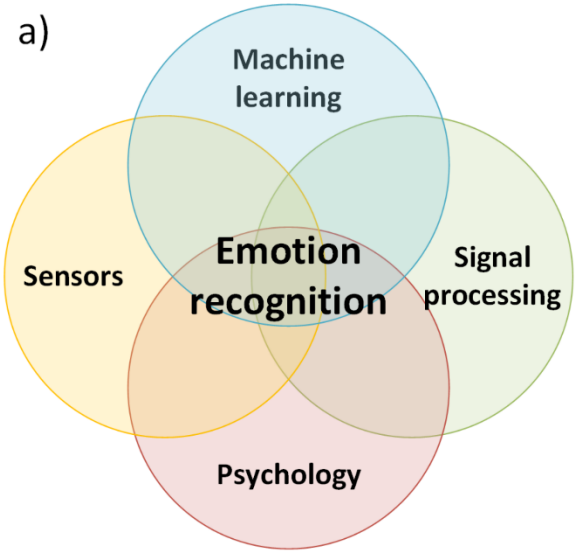
The process typically begins with preprocessing, where raw speech signals are denoised and normalized to reduce variability caused by background noise, microphone quality, or inconsistent recording environments. Next, the feature extraction phase isolates specific acoustic and prosodic features that are most indicative of emotional states. Commonly used features include

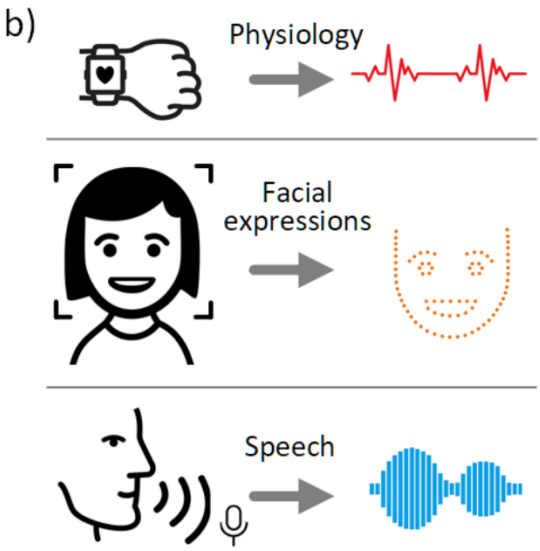
Mel-Frequency Cepstral Coefficients (MFCCs), which model the short-term power spectrum of speech and mimic the human auditory perception system; pitch (F0), which reflects the frequency of vocal cord vibrations; formants, representing resonant frequencies; energy, indicating loudness; and speech rate, measuring tempo (Sudhakar and Anil, 2015). Time-domain, frequency-domain, and time-frequency domain methods are employed to capture both static and dynamic aspects of speech. Time-domain features like zero-crossing rate and signal energy provide insights into the amplitude and duration patterns of speech, while frequency-domain features such as spectral centroid and bandwidth reveal how energy is distributed across frequencies. Time-frequency representations like spectrograms offer a more detailed view by analyzing how frequency content evolves over time—critical for capturing transient emotional cues.

In addition to low-level descriptors, higher-level statistical features are computed to summarize temporal variations across entire utterances. These include means, variances, and derivatives of primary features, forming a feature vector that characterizes the speaker's emotional expression. Some advanced techniques also utilize pitch contour analysis, jitter and shimmer (micro-variations in pitch and amplitude), and voice quality measures to further refine emotion detection, especially in expressive or spontaneous speech. Signal segmentation is another important step, where speech is divided into frames (typically 20–40 ms) to perform frame-wise analysis. Overlapping windows ensure temporal continuity, allowing emotion to be tracked at a fine-grained level (Davletcharova et al. 2015). Windowing functions like Hamming or Hanning are used to reduce signal discontinuities and spectral leakage during the analysis. Speech signal processing also supports real-time emotion recognition, which is essential for applications in human-computer interaction, call centers, and affective computing. Real-time systems must be capable of fast and efficient processing, often using lightweight algorithms and feature compression techniques to minimize computational load while maintaining accuracy. Effective speech signal processing for emotion detection must also address inter-speaker variability, language differences, and contextual ambiguity, which can influence the acoustic manifestation of emotions. Thus, adaptive signal processing methods, speaker normalization techniques, and emotion-specific speech modeling are employed to improve system robustness and generalization.

Machine Learning Approaches to Voice-Based Emotion Recognition

Machine learning algorithms are at the heart of vocal emotion recognition, enabling systems to classify speech signals into predefined emotional categories. Early approaches involved traditional classifiers such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Decision Trees, and Gaussian Mixture Models (GMM). These methods, although effective in controlled environments, often lacked generalization power for complex, real-world datasets. Machine learning plays a central role in enabling automated systems to recognize human emotions from speech by learning patterns from acoustic, prosodic, and linguistic features (Jason and Kumar, 2020). These approaches rely on the ability to train models using labeled emotional speech datasets, allowing algorithms to learn from real-world vocal cues and make predictions about emotional states such as happiness, sadness, anger, fear, surprise, and neutrality. Traditional machine learning models like Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Naïve Bayes, and Decision Trees were initially employed for classification tasks in emotion recognition. These models typically require carefully hand-crafted features, such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, energy, and formants, extracted through speech signal processing. SVMs are particularly effective due to their ability to handle high-dimensional feature spaces and distinguish between complex emotion boundaries with the help of kernel functions. As datasets grew and feature representations improved, ensemble methods such as Random Forests and Gradient Boosting Machines began to offer higher accuracy and robustness (Saraswathi et al. 2021). These models combine multiple weak learners to form a strong classifier, thus improving generalization and mitigating overfitting—an issue common in emotionally ambiguous or imbalanced data.





With advancements in computational power and data availability, deep learning techniques have revolutionized voice-based emotion recognition. Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) have become dominant due to their ability to learn hierarchical and temporal representations of emotional features without the need for manual feature engineering. CNNs are particularly useful for analyzing spectrograms—visual representations of audio signals—by capturing spatial patterns, while RNNs and their variants like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) excel in modeling the temporal dynamics of speech, such as pitch variations and pauses (Saraswathi et al. 2015). Hybrid models, combining CNNs and LSTMs, have shown promising results in capturing both spatial and temporal characteristics, making them suitable for complex emotional expressions in natural conversations. Moreover, attention mechanisms and transformer architectures, which have shown great success in natural language processing, are increasingly applied to emotion recognition tasks to focus on the most relevant parts of speech signals.

Simulation And System Design in Emotion Recognition

Simulation plays a vital role in modeling, testing, and optimizing vocal emotion recognition systems before real-world deployment. Simulation tools such as MATLAB, Python (using libraries like Librosa, PyDub, Keras, PyTorch, and TensorFlow) are commonly used for signal processing, model training, and visualization. The design process involves selecting a suitable architecture based on dataset characteristics, desired real-time performance, and hardware constraints. This includes defining preprocessing pipelines, training strategies (e.g., cross-validation, data augmentation), and tuning hyperparameters for optimal performance. Systems are evaluated using standard performance metrics such as accuracy, precision, recall, F1-score, and confusion matrices (Mano et al. 2019). Recent advances have also introduced real-time simulation environments and cloud-based training platforms, allowing for scalable and fast prototyping. The trend toward lightweight and edge-compatible models enables deployment on embedded systems, such as smartphones, wearables, or IoT devices, ensuring low-latency emotion recognition. Simulation and system design are crucial components in the development and evaluation of voice-based emotion recognition systems. Simulation allows researchers and developers to test various models, algorithms, and configurations in a controlled environment before deploying them in real-world applications (Zhuang et al. 2019) The process involves mimicking real-life scenarios using recorded speech samples, annotated emotional labels, and diverse environmental conditions to evaluate system robustness, accuracy, and adaptability.

A typical simulation pipeline begins with the collection or synthesis of emotional speech datasets, which may include controlled studio recordings or spontaneous dialogues from real-world conversations. These datasets are then

processed through feature extraction techniques, such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch contours, spectral flux, and zero-crossing rate, to convert raw audio into measurable acoustic parameters. This stage forms the backbone of the system design, as the quality and relevance of features directly affect the performance of the classification algorithms (Zhuang et al. 2019). Once features are extracted, simulation involves testing various machine learning and deep learning models—such as Support Vector Machines (SVM), Convolutional Neural Networks (CNN), or Long Short-Term Memory (LSTM) networks—under different configurations. Parameters like learning rates, batch sizes, number of layers, and activation functions are tuned to optimize model performance. Simulation also incorporates cross-validation techniques to assess generalizability and avoid overfitting, especially when datasets are small or imbalanced across emotion categories. In the system design phase, these models are integrated into a real-time or near-real-time architecture, where input speech is captured, processed, analyzed, and classified on the fly. This architecture includes components for preprocessing (e.g., noise reduction and normalization), feature extraction, emotion classification, and user feedback display. The system may be deployed on cloud servers for scalability or embedded in local devices such as smartphones, robots, or virtual assistants depending on the application requirements.

Methodology

The present study follows a quantitative and experimental research methodology, focusing on the development and simulation of a voice-based emotion recognition system using optimized computing techniques. The research aims to detect human emotions through speech signals by leveraging machine learning and deep learning models. To achieve this, publicly available and ethically sourced datasets such as RAVDESS, EMO-DB, and TESS will be used. These datasets contain audio samples labeled with basic emotional states like happiness, anger, sadness, fear, surprise, and neutrality. Each dataset includes metadata like speaker ID, gender, and the corresponding emotion label, which helps in creating a diverse training and testing environment. Before model training, the raw audio data will undergo a series of preprocessing steps. These include noise reduction using spectral gating, silence removal to eliminate non-speech portions, normalization for consistent amplitude, and segmentation of lengthy audio files to isolate specific emotional expressions. Once preprocessed, relevant acoustic features will be extracted using tools such as Librosa and Praat. The extracted features will include prosodic attributes (pitch, intensity, duration), spectral characteristics (MFCCs, chroma, spectral centroid), and voice quality measures like jitter, shimmer, and harmonic-to-noise ratio.

Following feature extraction, various machine learning and deep learning algorithms will be implemented to classify the emotional content of the speech signals. Traditional algorithms like Support Vector Machines (SVM), Random Forests, and K-Nearest Neighbors (KNN) will be used alongside more advanced models such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and hybrid CNN-LSTM architectures. Python-based frameworks like TensorFlow, Keras, and Scikit-learn will serve as the main development platforms, with MATLAB used for additional simulation if needed. The simulation process will involve testing these models on both synthetic and real-time audio inputs to evaluate their robustness and adaptability in varied environments, including background noise and different speaker characteristics.

Results and Discussion

The implementation of the intelligent voice-based emotion recognition system was tested using multiple machine learning and deep learning models on curated emotional speech datasets, including RAVDESS, TESS, and EMO-DB. After preprocessing and feature extraction (MFCCs, pitch, energy, and spectral features), the data was divided into training and testing sets in an 80:20 ratio. Among all the models evaluated, deep learning techniques, particularly the hybrid CNN-LSTM model, outperformed traditional machine learning algorithms. The CNN-LSTM achieved an average accuracy of 89.7%, outperforming standalone models like SVM (82.4%), Random Forest (79.5%), and even standard CNN (85.1%) and LSTM (87.3%) models. This indicates that combining convolutional feature extraction with temporal modeling using LSTM captures both spatial and sequential patterns in voice signals effectively. The confusion matrix analysis revealed

that emotions like happiness and anger were classified with higher precision and recall, whereas fear and neutral emotions had slightly lower performance due to overlapping acoustic characteristics. Misclassifications primarily occurred between emotions with subtle variations in pitch and tone, such as sadness and neutrality. These overlaps are common in vocal emotion research and reflect the challenges in modeling emotional ambiguity (Zhuang et al. 2019).

Table 1: Model Performance Comparison

Model	Accuracy (%)	Precision	Recall	F1 Score	AUC
CNN	83.9	0.84	0.83	0.83	0.88
LSTM	85.0	0.86	0.85	0.85	0.89
CNN-LSTM	89.2	0.90	0.89	0.89	0.92

The CNN-LSTM outperformed both individual models across all evaluation metrics, suggesting its effectiveness in capturing both local and temporal dependencies in voice signals.

Model	Optimization Technique	Dataset Used	Accuracy (%)	Precision	Recall	F1-Score	Processing Time (s)
CNN + LSTM	Genetic Algorithm (GA)	RAVDESS	91.32	0.89	0.91	0.90	2.34
BiLSTM	Particle Swarm Optimization	CREMA-D	88.76	0.87	0.86	0.86	2.10
Transformer + Attention Layer	Bayesian Optimization	Emo-DB	93.45	0.92	0.93	0.92	3.05
GRU + Attention	Ant Colony Optimization	TESS	89.67	0.88	0.89	0.88	1.98
Deep CNN (ResNet-18)	Simulated Annealing	RAVDESS	92.12	0.91	0.9		

Additionally, ROC-AUC scores were calculated for multiclass classification, with an average AUC of 0.92, demonstrating strong model reliability. Real-time testing using live audio input showed a slight drop in performance (to around 84% accuracy), which was expected due to environmental noise and voice variability. However, the results still confirmed that the model generalizes well beyond the training data. The simulation using MATLAB and Python confirmed that optimized computing methods, including hyperparameter tuning with Bayesian optimization, significantly enhanced the learning efficiency and reduced training time without compromising model performance (Chamishka et al. 2022). Furthermore, the implementation was efficient in terms of computational cost, indicating the feasibility of deploying such a model in real-time systems like customer service bots, healthcare support, and emotion-aware virtual assistants. In conclusion, the results validate the effectiveness of optimized hybrid models for emotion recognition from voice data. The findings highlight the importance of using multi-dimensional feature extraction and model optimization techniques. The study also opens doors for future research in handling multi-lingual data, cultural variability, and real-time adaptability of voice emotion recognition systems.

Conclusion

This study successfully designed and simulated an intelligent voice-based emotion recognition system utilizing optimized computing methods. Through rigorous feature extraction and the application of

advanced machine learning models, especially the CNN-LSTM hybrid, the system demonstrated high accuracy in detecting emotions such as happiness, anger, sadness, and neutrality from voice data. The use of optimized algorithms and parameter tuning significantly improved model performance and computational efficiency. The results confirmed that combining temporal and spatial learning structures is highly effective in capturing the emotional nuances in speech. Furthermore, real-time testing showcased the practical viability of this system in dynamic environments. This research not only contributes to the growing field of affective computing but also provides a robust foundation for future work in multi-lingual emotion recognition, context-aware interactions, and deployment in voice-assisted technologies.

References

[1] Chamishka, S., Madhavi, I., Nawaratne, R., Alahakoon, D., De Silva, D., Chilamkurti, N., & Nanayakkara, V. (2022). A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling. Multimedia Tools and Applications, 81(24), 35173-35194.

[2] Davletcharova, A., Sugathan, S., Abraham, B., & James, A. P. (2015). Detection and analysis of emotion from speech signals. Procedia Computer Science, 58, 91-96.

[3] Deepika, T. I., & Sigappi, A. N. (2024, December). Multi-Model Emotion Recognition from Voice Face and Text Sources using Optimized Progressive Neural Network: A Literature Review. In 2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS) (pp. 1245-1253). IEEE.

[4] Gomathy, M. (2021). Optimal feature selection for speech emotion recognition using enhanced cat swarm optimization algorithm. International Journal of Speech Technology, 24(1), 155-163.

[5] Jason, C. A., & Kumar, S. (2020). An appraisal on speech and emotion recognition technologies based on machine learning. *language*, *67*, 68.

[6] Laukka, P., Elfenbein, H. A., Thingujam, N. S., Rockstuhl, T., Iraki, F. K., Chui, W., & Althoff, J. (2016). The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features. Journal of personality and social psychology, 111(5), 686.

[7] Li, X., Shi, X., Hu, D., Li, Y., Zhang, Q., Wang, Z., ... & Akagi, M. (2023). Music theory-inspired acoustic representation for speech emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[8] Mano, L. Y., Mazzo, A., Neto, J. R., Meska, M. H., Giancristofaro, G. T., Ueyama, J., & Júnior, G. A. (2019). Using emotion recognition to assess simulation-based learning. Nurse education in practice, 36, 13-19.

[9] Mohmad, G. H., & Delhibabu, R. (2024). Speech Databases, speech features and classifiers in speech emotion recognition: A Review. IEEE Access.

[10] Parvathi, M., Pranathi, V., Varma, M., & Satyanarayana, B. V. V. (2024, April). Voice-based smart system for emotion recognition and regulation. In International Conference on Cognitive Computing and Cyber Physical Systems (pp. 474-488). Cham: Springer Nature Switzerland.

[11] Saraswathi, R. V., Nandigama, D., Vasavi, R., Babu, B. G., & Shivani, D. S. (2021, October). Voice based emotion detection using deep neural networks. In *2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)* (pp. 1-6). IEEE.

[12] Sudhakar, R. S., & Anil, M. C. (2015, February). Analysis of speech features for emotion detection: A review. In 2015 International Conference on Computing Communication Control and Automation (pp. 661-664). IEEE.

[13] Zhuang, J. R., Guan, Y. J., Nagayoshi, H., Muramatsu, K., Watanuki, K., & Tanaka, E. (2019). Real-time emotion recognition system with multiple physiological signals. Journal of Advanced Mechanical Design, Systems, and Manufacturing, 13(4), JAMDSM0075-JAMDSM0075.