

Unveiling Emotions in Speech: A Novel Machine Learning Framework for Vocal Sentiment Analysis

Moin Azam¹, Prof. S. H. Patil², Dr. Sunita Sachin Dhotre³

¹MTech (Computer Engineering) (Pursuing), Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune,

²Professor Department of Computer Engineering, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune

³Associate Professor Department of Computer Engineering, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune

Corresponding Author Email Id: moinazam7@gmail.com

Abstract

Emotion Recognition (SER) is a solution to the problem of extracting human emotional states from vocal signals, which is a key element in improving human-computer relations in such areas as healthcare, customer service, and entertainment. There are still many obstacles of SER, which are connected with variation of speech patterns, cultural differences, and environmental noise. Machine learning provides a rather strong alternative to address these challenges as they use the analysis of vocal characteristics, including tone, pitch, and intensity. This research work introduces the new structure for SER exploiting Toronto Emotional Speech Set (TESS) dataset, which has 2,800 quality bilingual audio sounds that express seven emotions. The proposed methodology combines old fashioned machine learning and the deep learning approach to extract and classify the emotional cues. Some of the extractable features include Mel-Frequency Cepstral Coefficients (MFCCs), chroma, and Mel-spectrogram for the purpose of reflecting timbral, harmonic, and temporal properties. Two old approaches, Logistic Regression and Decision Tree, set the baseline, and a hybrid deep learning model is presented, which is composed of Convolutional Neural Networks (CNN) and Deep Neural Networks (DNN). The hybrid model uses Conv1D layers for the localized extraction of features and DNN to higher-level abstractions with sparse categorical cross-entropy loss and the Adam optimizer. Assessed on the TESS dataset, the hybrid model outperforms the traditional methods by 98.39% in the accuracy. This framework shows a lot of promise for actual applications, such as an emotion-aware AI system and personalized music recommendation, and points to the direction of future research, in terms of noise robustness and cross-cultural generalization.

Keywords: Speech Emotion Recognition, Machine Learning, Deep Learning, Mel-Frequency Cepstral Coefficients, Convolutional Neural Networks, Deep Neural Networks, Toronto Emotional Speech Set, Emotion Classification

How to cite this article: Azam M, Patil SH, Dhotre SS. Unveiling Emotions in Speech: A Novel Machine Learning Framework for Vocal Sentiment Analysis. *Int J Drug Deliv Technol.* 2026;16(10s): 910-919. DOI: 10.25258/ijddt.16.10s.106

1. Introduction

Speech Emotion Recognition stands as an interdisciplinary research area which uses machine learning methods together with signal processing tools to identify emotional states from voice signals. Human communication depends on emotions because they affect both our decision-making and interactions between people as well as our mental functions. Speech Emotion Recognition tunes into emotions by evaluating vocal features such as tone and pitch intensity instead of using the textual content used in traditional sentiment analysis.

Escalating human-computer interaction (HCI) usage has initiated requirements for smart systems that can both recognize emotions and provide appropriate responses to them. SER brings emotional comprehension capabilities to machines operating in virtual assistants and call centers enabling them to respond to users based on detected emotional cues. SER remains difficult to solve even with current progress because speech pattern variations and cultural differences accompanied by environmental noise create significant challenges. Research for robust and efficient emotion recognition

models becomes vital to close current limitations while building emotionally intelligent artificial systems.

Emotion recognition systems serve important purposes in numerous different sectors. Healthcare utilizes SER to detect depression and anxiety early by detecting speech pattern changes. Through real-time analysis of call center interactions companies can pinpoint user dissatisfaction and deliver more sympathetic responses to enhance customer satisfaction.

SER provides entertainment users personalized music or movie recommendations through mood analysis within automated content delivery systems. Educational emotion-aware tutoring platforms evaluate engagements while learning systems modify teaching approaches based on gathered data. The value of SER becomes more pronounced through its application in security systems for stress and deception detection during forensic investigations. Because AI systems increasingly function in daily life, emotion recognition becomes necessary for developing digital user interactions that reflect human emotional behaviors.

2. Related Work

Human emotions derive from text analysis, speech interpretation and visual assessment which supports applications in robotics and human-computer interfaces. Different methods such as NLP and machine learning have been investigated as improvement strategies for system accuracy [1]. The survey examines currently available techniques plus datasets and outlines future research paths in detecting emotions from spoken words and written texts. The study examines feature sets from current methodologies while summarizing important achievements. The research presents inadequate multimodal investigation because it misses facial expression analysis and physiological signal integration which would improve accuracy. The research publication [2] introduces a machine learning algorithm for detecting emotions within speech segments as part of public emotion monitoring systems. The paper examines current research and compares both traditional learning algorithms with deep learning models to measure how dataset composition alongside text preprocessing and model design impacts performance. The research provides evidence of practical applications while analyzing the connection between human emotions and external societal occurrences. Drawback: The study does not address multimodal approaches nor examines potential biases in training data which might affect model generalisability.

The researchers in [3] introduced emotion detection from audio using databases from English to Italian through a multilingual framework. This research employs a Convolutional Neural Network model to classify emotions using extracted acoustic features such as MFCC and Chroma without utilizing any visual data. Speech-based emotion detection systems produce accurate results of 97.89% for multiple languages while demonstrating robust performance on both original and expanded data sets. This research fails to address noise robustness in realistic environments and makes no attempt to study generalization across more languages than those specifically analyzed. In paper [4] researchers examine how to recognize emotions in speech through algorithms such as SVM and MLP while utilizing MFCC and Chroma features. The system identifies calm emotional states alongside happy and angry ones with 86.5% accuracy levels. The development uses spoken language emotion recognition to facilitate smoother human-machine communication. The work does not address deep learning methods nor multimodal analysis while ignoring the effect of different accents and environmental noise on system performance.

In this research [5] speech emotion recognition (SER) methods from a machine learning perspective are systematically examined with attention to data processing steps and essential feature extraction and classification processes. The study identifies the low accuracy struggles of speaker-independent models but proposes contemporary solutions. The research presents evaluation guidelines and metrics which help both SER research development and model upgrades. This review fails to address deep learning developments together with multimodal fusion approaches and practical

deployment barriers such as noisy surroundings and cross-language system functionality.

Research in emotion recognition for speech and music now attracts substantial interest because scientists work on extracting features and developing classification methods. Numerous studies have developed emotional expression recognition frameworks which depend mainly on linear prediction cepstral coefficients (LPCC) and Mel-frequency cepstral coefficients (MFCC) along with pitch and energy characteristics (6). Different types of emotion classifiers including anger, joy, sadness, surprise and fear exist today although the performance evaluation of all these systems depends heavily on feature extraction. The three main approaches for speech and music emotion recognition systems include knowledge-based and acoustic-phonetic and pattern recognition methodologies. Recognition efficiency increases when analyzing speech-based emotions through principal component analysis (PCA) and MFCCs demonstrates the importance of dimensionality reduction techniques [6].

Recent studies have studied auditory gating methods to determine how little acoustic information is needed for identifying emotions in both recorded speech and musical tones. Research shows that longer gate periods enhance recognition performance until recognition reaches a stable level for different emotional categories. Research shows that emotions of anger along with happiness and neutrality together with sadness achieve high accuracy classifications in 100 milliseconds yet additional emotions need at least 250 milliseconds for successful identification. The studies show that both speech and music use similar acoustic signals to communicate emotions demonstrating the significance of basic acoustic patterns in quick emotion detection [7]. Research on music emotion recognition (MER) has gained significance because of its applications within recommendation systems while also serving music therapy needs and automatic composition tasks. MER tasks now benefit greatly from deep learning approaches that make extensive use of convolutional neural networks (CNNs). The research framework consists of three sections which include datasets coupled with feature extraction followed by classification algorithm selection. Entropy stands alongside zero-crossing rate and spectral centroid and spectral roll-off as common features in this research. The development of deep learning-based MER presents ongoing problems with restricted datasets and uninterpretable models and unclear emotion expression [8][41].

Academics have researched the way music-evoked emotions affect human mood perception through multiple experiments. Research on physiological-emotional correlations has enabled scientists to create sensor-based models which detect emotions. Achieving high accuracy in emotion detection faces obstacles from diverse issues related to how emotions get annotated together with the methods for training models and classification practices. Standardized datasets together with robust classifiers present a critical issue that researchers and developers face [9][42].

Research using 1D convolutional neural networks (1D CNNs) showed successful results in MER. Music signals benefit from 1D CNNs which extract important patterns from raw audio data through their ability to recognize temporal patterns. As a result these networks achieve superior accuracy with enhanced computational effectiveness. However, the lack of comparative analysis with alternative deep learning architectures, such as recurrent neural networks (RNNs) or transformer-based models, limits a comprehensive evaluation of its effectiveness. Real-time processing limitations along with challenges regarding the universal application of various datasets and the interpretability of learned system features need to be studied more thoroughly [10].

The research field of emotion detection in speech along with music continues to develop but still faces various knowledge gaps. The existing model generalization capabilities across different cultural datasets and datasets needs additional exploration to create cross-domain approaches. The exploration of how different features in multimodal annotation systems affect each other remains an unexplored field for researchers. Expanding explainable and transparent features in deep learning-based MER models is necessary to make them acceptable for practical implementation. There exists a necessary requirement for emotion recognition frameworks which provide real-time performance while maintaining an accurate computation balance. Research attention to these gaps will enable effective developments of scalable and robust emotion recognition systems which also provide interpretation capabilities for speech and music.

3. Dataset Information

The Toronto Emotional Speech Set (TESS) serves as a public dataset created to facilitate speech emotion recognition (SER) has been used for the following work.

TESS stands apart from most datasets by featuring bilingual high-quality recordings from two female participants ages 26 and 64 which promotes better class balance and generalization capabilities in analysis models.

TESS contains 2,800 WAV audio files, where 200 target words are spoken within the carrier phrase “Say the word ” while expressing seven emotions: Speakers in the TESS dataset express target words across seven emotional states which include anger and disgust with others being fear happiness pleasant surprise sadness and neutral. The arrangement of the dataset into speaker and emotion category sections enables researchers to retrieve data efficiently.

The superior audio recordings facilitate extraction of important audio features including MFCCs pitch and chroma features which serve as vital elements for building deep learning models. By incorporating TESS researchers achieved better emotion detection performance which makes this methodology ideal for use in emotionally intelligent AI applications.

4. Exploratory Data Analysis

From the figure 1, the dataset shows perfect label balance because 'fear' and 'angry' and 'disgust' and 'neutral' and 'sad' and 'ps' and 'happy' each contain exactly 400 examples. The consistent labeling distribution prevents specific emotions from dominating which creates essential conditions for developing unbiased machine learning models. The database presents a complete range of emotions by distributing instances between negative categories including fear, anger, disgust and sad emotions, neutral tones and positive feelings of happiness, and ps. The consistent distribution across emotional groups minimizes bias effects while improving predictive accuracy for various emotional classes.

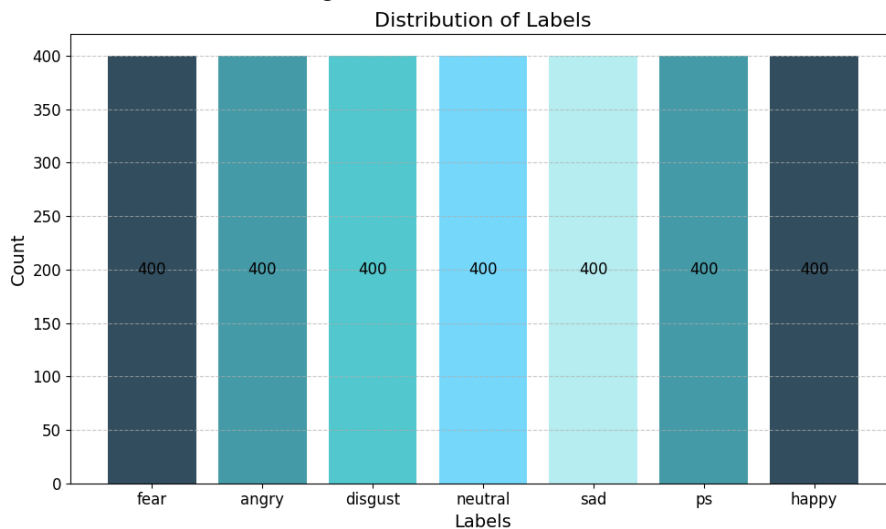


Figure 1 : Distribution of the Speech Emotion Labels

In figure 2, it can be clearly seen from the waveform that the "angry" emotion contains much amplitude variation, meaning that its vocal expression is high-energy and intense. Multiple peaks and bursts indicate a variable

energy pattern, corresponding to the fact that anger is a dynamic effect in speech. Also, pauses between high energy parts of the waveform are obvious, these are representative of typical breaks or tonal changes that

Unveiling Emotions in Speech: A Novel Machine Learning Framework for Vocal Sentiment Analysis
 occur when speaking passionately. The expressive and volatile flavor of anger is what comprises its vocal characteristics, which set it apart from other quieter or more neutral emotions in vocal analysis.

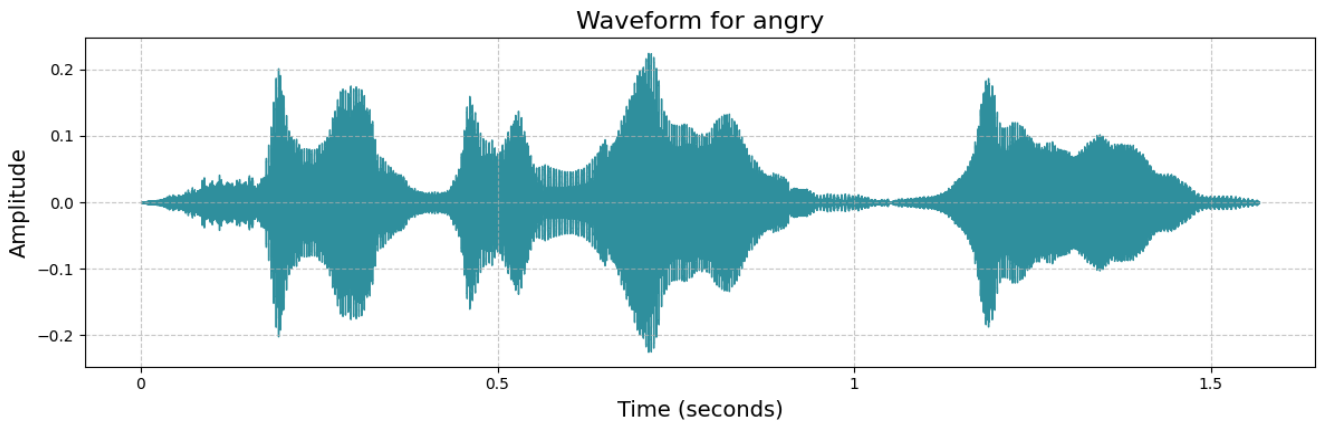


Figure 2: Analysis of Waveform of Angry Audio Speech

The spectrogram from figure 3 shows the analysis of "neutral" speech stating a balanced frequency distribution with a moderate intensity, meaning there is a stable energy level flow and also a lack of extreme emotional variations.

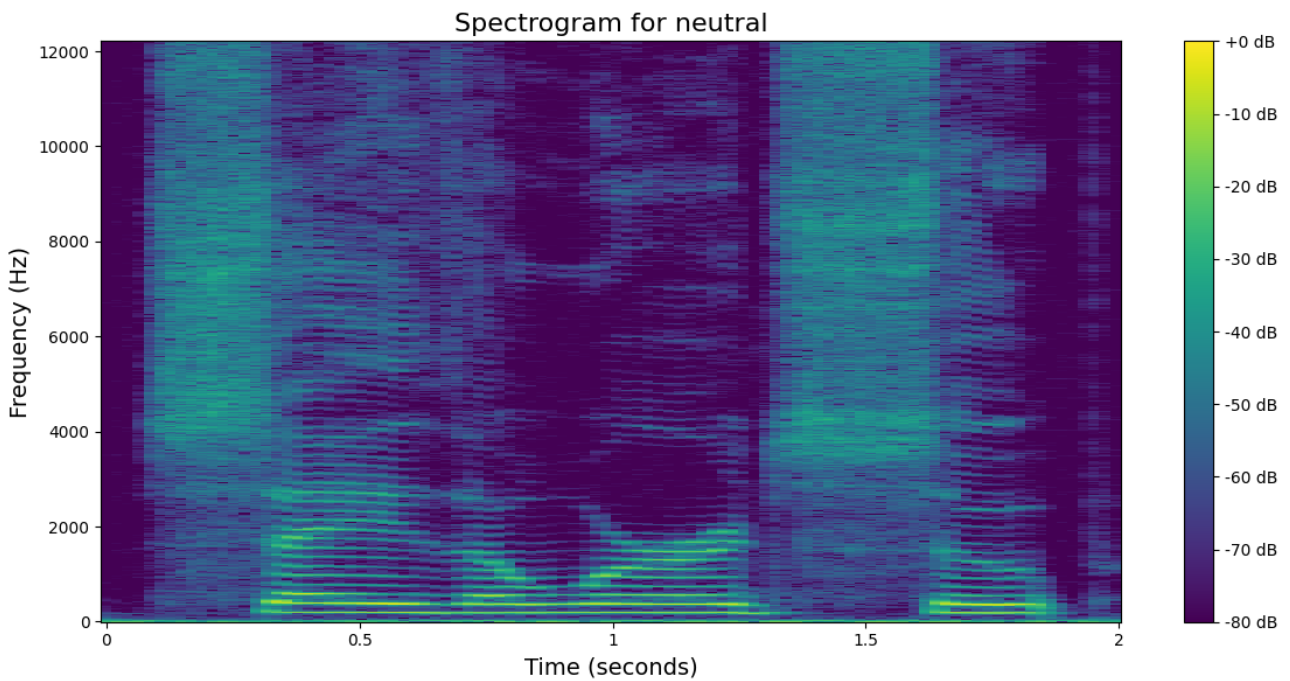


Figure 3: Analysis of Spectrogram of Natural Speech

The zero-crossing rate for the fearful speech as seen in figure 4 is found to be high and consistent, reflecting the rapid signal variations and a high frequency component, which is expected of fearful speech. These drops appear intermittent, virtually suggesting breaks or decreases in energy, perhaps relating to hesitation or breathiness.

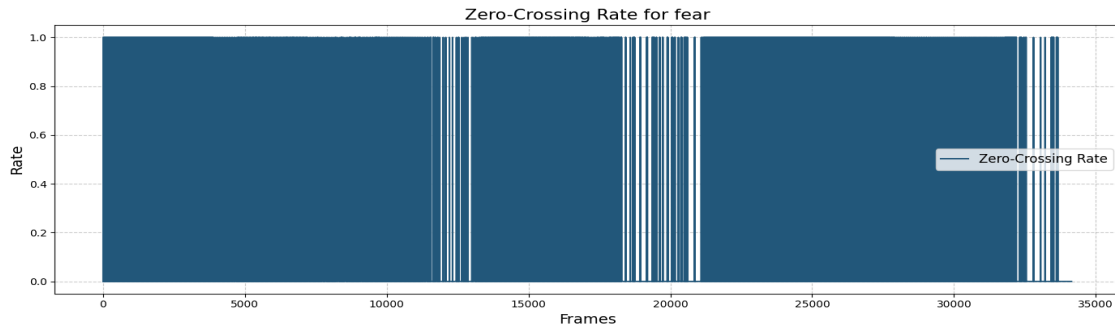


Figure 4: Analysis of Zero-Crossing Rate of FearSpeech

The spectral centroid of 'disgust in figure 5' consists of high frequency dominance at the beginning, and variability, which suggests shifting emphasis. The

pattern of disgust reflects the dynamic nature of the expression: tonal shifts, abrupt changes in vocal intensity abound.

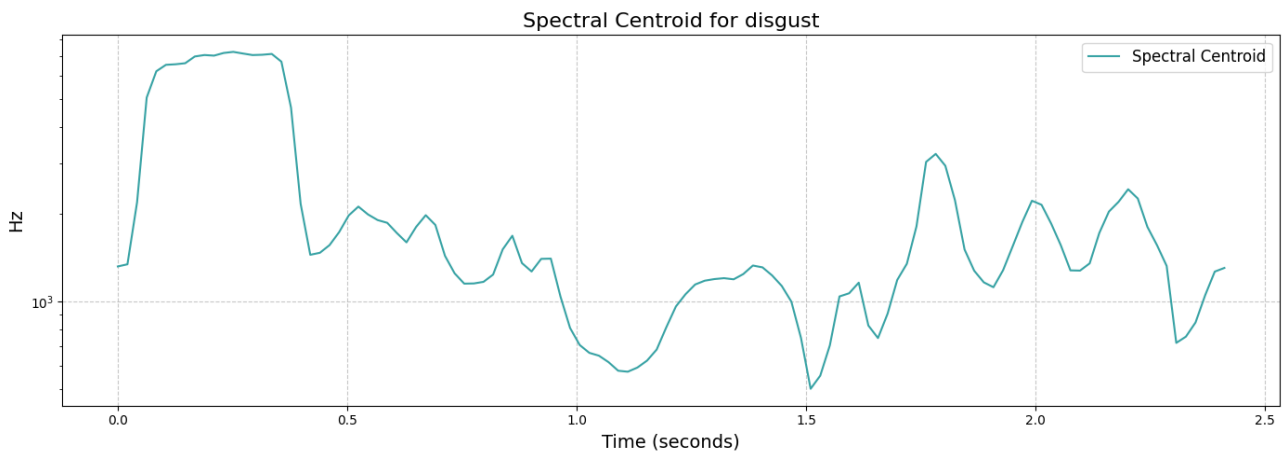


Figure 5: Analysis of Spectral Centroid of Disgust Speech

In sad speech (figure 6) low frequency abnormalities of irregular fluctuation and a lack of vocal energy with monotonicity appear in the pitch contour of 'sad.'

Infrequent sharp peaks suggest intonational variations on a brief scale, or as a result of emotional strain or hesitation.

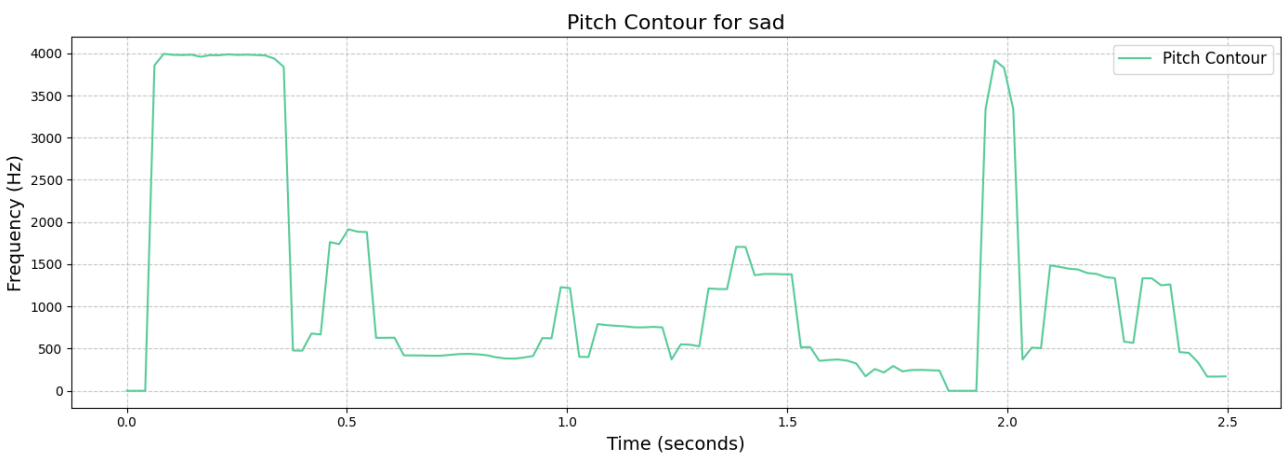


Figure 6: Analysis of Pitch Contour of Sad Speech

5.Methodology

5.1. Feature Extraction

For effective analysis and classification of emotions from audio signals, a multi faceted approach to feature extraction is used. The features extracted encapsulate the

important characteristics of human speech so that the emotion recognition is robust. The considered features include:

Unveiling Emotions in Speech: A Novel Machine Learning Framework for Vocal Sentiment Analysis

- **Mel-Frequency Cepstral Coefficients (MFCCs):** The mean of 13 MFCCs capture the timbral characteristics of speech. Because of the efficiency of MFCCs to model the short term power spectrum of sound they are widely used in speech and audio processing for discriminating between different emotions given cue of the tone and pitch variations.
- **Chroma Features:** Hidden representation, represent harmonic content of the signal important for tonal variation. Chroma features examine how energy is spread across different pitch classes to offer a view into the musical or harmonic content of speech to help diagnose emotional state.
- **Mel-Spectrogram Features:** It provides a time frequency representation of the audio considering both spectral and temporal variations. Audio intensity at different frequencies at different times is represented by something called the Mel-spectrogram (it gives a more detailed representation of how speech patterns change dynamically, which is useful for detecting emotional cues, for instance).

5.2. Traditional Machine Learning Models

To establish a baseline for emotion classification, two well-known machine learning models are implemented:

- **Logistic Regression Classifier:** A statistical model for binary and multi class classification task. A logistic function which estimates the probability that the given input should belong to a certain class. Logistic regression is simple, it is efficient, and it is interpretable; therefore, it can be applied in many real world problems such as medical diagnosis, spam detection and financial risk assessment. Although its nature is linear, it is effective on linearly separable data and can be extended with techniques like regularisation to achieve good generalisation.
- **Decision Tree Classifier:** A partition of the feature space into decision regions where a simpler and more interpretable model classifies. The hierarchical structure of this model makes it easy to interpret and understand, as data is split along the feature values at each node. However, decision trees are worth considering for emotion classification because they can easily tell you which features are most important to the classifier, but they might be overfitting to complex datasets.

5.3. Proposed Model

For better classification performance, a hybrid deep learning model, including Convolutional Neural Networks (CNN) and Deep Neural Networks (DNN), is introduced. With the use of ConvID layers, the input is

processed by the CNN branch as a sequence, enabling the task of localized feature extraction for a long time. The feed forward neural network DNN branch learns high level abstractions from the same feature set. Joint representations are extracted by cascading a dense layer over the outputs of the CNN and DNN branches after they have been concatenated in a fusion layer. Finally, a final classification layer is formed using a fully connected layer with a softmax activation function to predict the emotion class. It was to train the hybrid model with sparse categorical cross entropy loss and optimize with Adam optimizer. The training over epochs for 20 epochs with a batch size of 32, for which the training has converged to an optimal solution. The combined strengths of CNNs for spatial feature extraction and DNNs for high-level learning lead to improved generalization and robustness in emotion classification.

This approach takes a comprehensive approach to emotion recognition using both classical machine learning methods and more recent deep learning architectures to exploit all available information.

5.4. Song Recommendation Based on Emotion

A song recommendation system is integrated from Spotify's API to enhance the user experience. The predicted emotion of the audio sample is used by the system to suggest songs. The recommendation process involves the following steps:

- **Emotion-to-Genre Mapping:** Spotify database stores each detected emotion to a corresponding genre or a set of characteristics. For instance, upbeat pop songs are suitable for happy emotions and slow-tempo melancholic music for sad emotions.
- **Spotify API Querying:** For each emotion, a predefined query is constructed using the attributes such as genre, mood, and intensity and relevant to musical entity. We then query the API with the criteria we just matched and retrieve the songs.
- **Retrieving and Displaying Songs:** The API responds with a list of song recommendations and details about each – song title, artist, album and release – as listed in the table below. With these values, the user is shown a curated list of tracks that matches their detected emotion.

Using Spotify's library of music and smart searches, the system guarantees that a user's song recommendations are relevant, significant, and of the best quality for the emotional context.

6. Results and Discussion

6.1. Performance Metrics

The performance of the models was evaluated using standard metrics which can be seen in table 1.

Table 1: Performance Result of the Model on Test Data

Model	Accuracy	F1 Score	Precision	Recall
LR	94.64%	94.66%	94.76%	94.64%
DT	93.32%	92.40%	92.69%	92.32%
Proposed Model	98.39%	98.39%	98.39%	98.39%

6.2. Comparative Analysis

It can be clearly seen in the table that Hybrid Deep Learning (DL) Model is superior to both traditional models in all the metrics. As expected, the Logistic Regression model is better than the Decision Tree model since it can better generalize across different emotional variations. Both models however depend on manually extracted features and such features can be too simple to contain the intricacies of the expressed emotion in the speech.

Using the deep learning techniques, the Hybrid DL Model achieves an accuracy of 98.39% which is much better implying that it gets to read more meaningful

patterns from the audio data. Unlike traditional models, it learns feature representations automatically and hence relieves the need for manual feature engineering.

6.3. Error Analysis and Limitations

Despite this high accuracy, there are some errors left, mainly in distinguishing fear from surprise and anger from disgust since these emotions have similar acoustic features. Furthermore, this work is tested on background noise, variation in speech patterns, and imbalance in emotion distribution that can degrade performance.

6.4. Application Example

```

1/1 ██████████ 0s 21ms/step
Actual Emotion - Happy
The predicted emotion (Conv1D Model) for the audio file is: happy
-----
Songs for the emotion 'happy':
Song: The Other Side
Artist: Jason Derulo
Album: Upbeat Pop Hits
Release Date: 2022-07-08
---
Song: Green Light
Artist: Lorde
Album: Upbeat Happy Music
Release Date: 2021-02-19
    
```

Figure 7: Song Recommendation based on Emotion Detected

For an audio file labeled “Happy”, the model successfully predicts the emotion and recommends mood-based songs:

- "The Other Side" – Jason Derulo
- "Green Light" – Lorde

This result proves that the model is capable of personalized music recommendations, and can be used in the real world of emotion aware AI systems.

7. Conclusion and Future Work

This study investigates the evidence of using a traditional machine learning model (Logistic Regression, Decisions Tree) and a Hybrid Deep Learning (DL) Model in speech emotion recognition (SER). Performance evaluation shows that the Hybrid DL Model is significantly superior to traditional models

and has better performance as far as accuracy is concerned, having an accuracy value of 98.39 compared to Logistic Regression which has 94.64 and Decision Tree which has 92.32. The deep learning approach proved to be rather effective in capturing certain complex patterns from speech hence improving the classification accuracy.

Such emotion aware AI systems are shown to have real world application, particularly in the case of integrating SER with music recommendation. With high accuracy of emotion detection and intelligent suggestion of suitable songs, the model helps to deliver higher user experience in personalized entertainment systems. Even so, there still are some challenges like misclassifying similar emotions, variations in speakers and loud disturbing background noise.

Speech emotion recognition systems will be made more robust and applicable in the future research. Key improvement is expanding the dataset with more diverse voices i.e. male speakers, varied accents, and multilingual speech helps in obtaining an improved generalization across different demographics. Later will also prioritize real-time deployment of the model, such as to facilitate efficient processing on the edge devices (for example, voice assistants and mobile apps). Finally, personalization techniques will be devised to tune the system with respect to individual speech characteristics for more accurate and personalized emotion detection will be developed. These progressions will truly intensify SER's real world impact across numerous businesses.

8. Supplementary Materials:

The following supplementary material is provided to support the findings of this study and will be published alongside the article:

- Supplementary Dataset: Toronto Emotional Speech Set (TESS)

This dataset comprises 2,800 high-quality, labeled audio recordings expressing seven different emotional states (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral).

The dataset used for training and evaluating the emotion recognition models can be accessed at:

<https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess>

9. References

1. Sailunaz, K., Dhaliwal, M., Rokne, J. and Alhaji, R., 2018. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 8(1), p.28.
2. Kholodna, N., Vysotska, V. and Albota, S., 2021. A Machine Learning Model for Automatic Emotion Detection from Speech. In *MoMLeT+ DS* (pp. 699-713).
3. Bhattacharya, S., Borah, S., Mishra, B.K. and Mondal, A., 2022. Emotion detection from multilingual audio using deep analysis. *Multimedia Tools and Applications*, 81(28), pp.41309-41338
4. Krishna, K.V., Sainath, N. and Posonia, A.M., 2022, March. Speech emotion recognition using machine learning. In *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1014-1018). IEEE.
5. Madanian, S., Chen, T., Adeleye, O., Templeton, J.M., Poellabauer, C., Parry, D. and Schneider, S.L., 2023. Speech emotion recognition using machine learning—A systematic review. *Intelligent systems with applications*, p.200266.
6. Agarwal, G., Maheshkar, V., Maheshkar, S. and Gupta, S., 2018. Recognition of emotions of speech and mood of music: a review. In *International Conference on Wireless, Intelligent, and Distributed Environment for Communication: WIDECOM 2018* (pp. 181-197). Springer International Publishing.
7. Nordström, H. and Laukka, P., 2019. The time course of emotion recognition in speech and music. *The Journal of the Acoustical Society of America*, 145(5), pp.3058-3074.
8. Han, D., Kong, Y., Han, J. and Wang, G., 2022. A survey of music emotion recognition. *Frontiers of Computer Science*, 16(6), p.166335.
9. Chaturvedi, V., Kaur, A.B., Varshney, V., Garg, A., Chhabra, G.S. and Kumar, M., 2022. Music mood and human emotion recognition based on physiological signals: a systematic review. *Multimedia Systems*, 28(1), pp.21-44.
10. Shashidhar, R., Balivada, D., Shalini, D.N., Krishnappa, K.H. and Roopa, M., 2023, November. Music Emotion Recognition using Convolutional Neural Networks for Regional Languages. In *2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIIE)* (pp. 1-7). IEEE.
11. Sidhu, M.S., Latib, N.A.A. and Sidhu, K.K., 2024. MFCC in audio signal processing for voice disorder: a review. *Multimedia Tools and Applications*, pp.1-21.
12. Prabakaran, D. and Sriuppili, S., 2021. Speech processing: MFCC based feature extraction techniques-an investigation. In *Journal of Physics: Conference Series* (Vol. 1717, No. 1, p. 012009). IOP Publishing.
13. Zalkow, F. and Müller, M., 2021. CTC-based learning of chroma features for score-audio music retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, pp.2957-2971.
14. Chittaragi, N.B. and Koolagudi, S.G., 2021. Dialect identification using chroma-spectral shape features with ensemble technique. *Computer Speech & Language*, 70, p.101230.
15. Aldahdooh, A., Hamidouche, W., Fezza, S.A. and Déforges, O., 2022. Adversarial example detection for DNN models: A review and experimental comparison. *Artificial Intelligence Review*, 55(6), pp.4403-4462.
16. Hussain, H., Tamizharasan, P.S. and Rahul, C.S., 2022. Design possibilities and challenges of DNN models: a review on the perspective of end devices. *Artificial Intelligence Review*, pp.1-59.
17. Al-Fraihat, D., Sharrab, Y., Alzyoud, F., Qahmash, A., Tarawneh, M. and Maaita, A., 2024. Speech recognition utilizing deep learning: A systematic review of the latest developments. *Human-centric computing and information sciences*, 14.
18. Hickman, L., Langer, M., Saef, R.M. and Tay, L., 2024. Automated speech recognition bias in

- personnel selection: The case of automatically scored job interviews. *Journal of Applied Psychology*.
19. Kheddar, H., Hemis, M. and Himeur, Y., 2024. Automatic speech recognition using advanced deep learning approaches: A survey. *Information Fusion*, p.102422.
 20. Brahma, Z., Mahyoob, M., Al-Sarem, M., Algaraady, J., Bousselmi, K. and Alblwi, A., 2024. Exploring the role of machine learning in diagnosing and treating speech disorders: A systematic literature review. *Psychology Research and Behavior Management*, pp.2205-2232.
 21. Xue, H., Shao, Q., Huang, K., Chen, P., Liu, J. and Xie, L., 2024, July. SSHR: Leveraging self-supervised hierarchical representations for multilingual automatic speech recognition. In *2024 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE.
 22. O'Shaughnessy, D., 2024. Trends and developments in automatic speech recognition research. *Computer Speech & Language*, 83, p.101538.
 23. Chang, X., Yan, B., Choi, K., Jung, J.W., Lu, Y., Maiti, S., Sharma, R., Shi, J., Tian, J., Watanabe, S. and Fujita, Y., 2024, April. Exploring speech recognition, translation, and understanding with discrete speech units: A comparative study. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 11481-11485). IEEE.
 24. Mahum, R., El-Sherbeeney, A.M., Alkhaledi, K. and Hassan, H., 2024. Tran-DSR: A hybrid model for dysarthric speech recognition using transformer encoder and ensemble learning. *Applied Acoustics*, 222, p.110019.
 25. Fatehifar, M., Schlittenlacher, J., Almufarrij, I., Wong, D., Cootes, T. and Munro, K.J., 2024. Applications of automatic speech recognition and text-to-speech technologies for hearing assessment: a scoping review. *International Journal of Audiology*, pp.1-12.
 26. Dhanjal, A.S. and Singh, W., 2024. A comprehensive survey on automatic speech recognition using neural networks. *Multimedia Tools and Applications*, 83(8), pp.23367-23412.
 27. Dhanjal, A.S. and Singh, W., 2024. A comprehensive survey on automatic speech recognition using neural networks. *Multimedia Tools and Applications*, 83(8), pp.23367-23412.
 28. Miller, L., Patel, T.K., Uehara, G., Naik, S. and Spanias, A., 2024, October. Quantum Machine Learning and Spectrogram Fusion for Speech Recognition. In *2024 58th Asilomar Conference on Signals, Systems, and Computers* (pp. 674-678). IEEE.
 29. Arpita, H.D., Al Ryan, A., Hossain, M.F., Rahman, M.S., Sajjad, M. and Prova, N.N.I., 2025. Exploring Bengali speech for gender classification: machine learning and deep learning approaches. *Bulletin of Electrical Engineering and Informatics*, 14(1), pp.328-337.
 30. Mustavi Tasin, S., Chowdhury, M.E., Pedersen, S., Chabbouh, M., Bushnaq, D., Aljindi, R., Kabir, S. and Hasan, A., 2024. Ensemble Machine Learning Model for Inner Speech Recognition: A Subject-Specific Investigation. *arXiv e-prints*, pp.arXiv-2412.
 31. Zhang, Y., Yue, Z., Patel, T. and Scharenborg, O., 2024. Improving child speech recognition with augmented child-like speech. *arXiv preprint arXiv:2406.10284*.
 32. Bai, Y., Chen, J., Chen, J., Chen, W., Chen, Z., Ding, C., Dong, L., Dong, Q., Du, Y., Gao, K. and Gao, L., 2024. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *arXiv preprint arXiv:2407.04675*.
 33. Kanoujia, S. and Karuppanan, P., 2024, March. Depression Detection in Speech Using ML and DL Algorithm. In *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)* (Vol. 2, pp. 1-5). IEEE.
 34. Najmusher, H., Roopsagar, K., Sairamkumar, M. and Vanamuthu, V., 2025, January. A Review of Advancements in NLP and Speech Recognition for Enhanced Operating Systems. In *2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)* (pp. 693-698). IEEE.
 35. Moghe, B. and Kachhara, M., 2024, June. Multimodal Emotion Analysis for Depression Detection-Integrating Facial Expression and Speech Recognition. In *2024 Second International Conference on Inventive Computing and Informatics (ICICI)* (pp. 37-42). IEEE.
 36. Omoyemi, O.E., 2024. Machine learning for predictive AAC: Improving speech and gesture-based communication systems. *World Journal of Advanced Research and Reviews*, 24(01), pp.2569-2575.
 37. Lakomkin, E., Wu, C., Fathullah, Y., Kalinli, O., Seltzer, M.L. and Fuegen, C., 2024, April. End-to-end speech recognition contextualization with large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 12406-12410). IEEE.
 38. Kumar, Y., 2024. A comprehensive analysis of speech recognition systems in healthcare: current research challenges and future prospects. *SN Computer Science*, 5(1), p.137.
 39. Gupta, R., Gunjawate, D.R., Nguyen, D.D., Jin, C. and Madill, C., 2024. Voice disorder recognition using machine learning: a scoping review protocol. *BMJ open*, 14(2), p.e076998.

40. Kaviani, Pouria, and Sunita Dhotre. "Short survey on naive bayes algorithm." *International Journal of Advance Engineering and Research Development* 4.11 (2017): 607-611.
41. Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*, 8(4). <https://doi.org/10.1002/widm.1253>
42. Osmani, A., Mohasefi, J. B., & Gharehchopogh, F. S. (2020). Enriched latent Dirichlet allocation for sentiment analysis. *Expert Systems*, 37(4). <https://doi.org/10.1111/exsy.12527>