Research Article

Speech Emotion Recognition Using LSTM Neural Network Model

Ashok Gupta Maturi[1], T.Krishnan[2]

[1, 2] Department of CSE, Koneru Lakshmaiah Foundation, Vaddeswaram, Andhra Pradesh, India
[1] mailto:maturiashokgupta@gmail.com , [2]mailto:tkrishnan.mtech@kluniversity.in

Abstract

Speech Emotion Recognition SER has become a widely studied issue in the circle of affective computing. It has been used in human-computer interaction and neuropsychology. In the current research, deep learning methodologies-based models are constructed for artificial intelligence integration into SER systems. We use the neural network form of Long Short-Term Memory (LSTM) for automatic emotion detection and classification using audio data recordings. We introduce a method that preprocesses MFCCs and audio data as features and subsequently uses LSTM for classification and learning. Using the model, training as well as testing is conducted over a dataset that has numerous examples of diverse emotions. The effectiveness of our approach is manifested by the whole experiment and the test, which shows the correctness of the suggested model in the exact emotion recognition from speech. The results indicate the consequence of the outstanding performance, high accuracy resolution to numerous emotion categories. Our work delves into the frontier of SER and attains greater insight into how deep learning may be leveraged to identify emotional states.

Keywords: Speech Emotion Recognition (SER), Affective Computing, Deep Learning Techniques, Long Short, Term Memory (LSTM), Automatic Detection, Mel-Frequency Cepstral Coefficients (MFCCs), Emotional States Detection.

1. Introduction

Speech Emotion Recognition (SER) has lately emerged as a popular domain in the realm of affective computing and it seeks to understand human emotions expressed through speech signals. This potential for automating categorization and detection of emotions from speech has extensive implications across disciplines ranging from human-machine interaction, neurology, to psychic evaluation. The systems of SER have been enhanced to an enormous degree during the recent past due to the advent of artificial intelligence and signal processing.

Early efforts at SER used conventional feature sets such as the pitch, intensity, and spectral coefficients, and also machine learning models for the classification. The strategies listed above would be exhibit some usefulness, but there was an issue with their inability to capture temporal dynamics and context. In fact, deep learning, which has made a recent appearance, has made a huge impact in the field of SER, one being able to offer one new more stable, and accurate emotion identification systems.

The models of deep learning, particularly those based on Recurrent Neural Networks (RNN) with the ability to persist over long sentence lengths like the Long Short-Term Memory (LSTM), have developed incredible performance in speech recognition and sentiment analysis. Considering key dynamics of the sequence data, LSTM networks have demonstrated their ability to model long-term dependencies and temporal patterns owing to which they are ideally capable of processing arrays of time series like audio recordings. Building on this methodology, we will aim to advance the SER limits by incorporating deep learning methods within the scientific process. In this case study, we are showing a novel approach to SER based on utilization of LSTM neural networks. The previous research operated in the same field as our own, they applied their expertise and techniques to develop the existing one better and created a new emotional recognition system.

The research aims to ascertain the efficacy of curing deep learning methods in classifying speech signal emotions accurately. We hope to apply the approach practically in-depth through extensive research and exposure on a diverse dataset representing a variety of emotional states. Finally, we want to showcase how impactful and robust our proposed approach can be. By pointing out the various strengths and limitations of our approach, we hope to contribute towards the novelty of SER technology. It will also promote understanding regarding future applications of deep learning models, which will aid in the advancement of.

Literature Review

[1] In the area of Speech Emotion Recognition (SER), researchers have explored different databases, features, and classifiers to analyze the emotional content of speech signals. Swain, Routray and Kabisatpathy (2018) completed a comprehensive review of SER approaches and pointed out the growth of SER systems over the last 10 years and how it relates to developing automatic analysis of human affective behavior through speech. In their review, Swain, Routray, and Kabisatpathy (2018) outlined developments across numerous systems and algorithms focusing on identifying emotional content. The review delves into the diverse range of databases, features, and classifiers utilized across different languages, reflecting the interdisciplinary nature of SER research.

[2] Wani et al. (2021) contributed to the understanding of speech emotion recognition systems by synthesizing recent literature and design methodologies. Their review highlighted the complexities involved in bridging the quantitative and qualitative disparities between human and machine emotion recognition. By scrutinizing current SER systems, they identified prominent research gaps, providing valuable insights for future investigations in this burgeoning field of study.

[3] Schuller, Rigoll, and Lang (2003) introduced the application of continuous hidden Markov models (HMMs) for speech emotion recognition. Their work compared two methods—one employing Gaussian mixture models for global statistics classification and the other incorporating continuous HMMs for increased temporal complexity. Through detailed experimentation and analysis on speech corpora in German and English, they achieved promising recognition rates, surpassing human judgment in some instances.

[4] Khalil et al. (2019) provided a comprehensive review of deep learning techniques in speech emotion recognition, elucidating their efficacy as alternatives to traditional methods. They outlined the utilization of deep learning approaches in extracting emotions from speech signals and discussed recent advancements in the field. By examining databases, emotions extracted, and contributions toward SER, They provided insightful information about the limitations. and future directions of deep learning-based emotion recognition systems.

[5] Hossain and Muhammad (2019) proposed an emotion recognition system using a learning approach from distinct emotional big data. Their innovative system involved processing speech signals into Mel-spectrograms, treating them as images, and feeding them into convolutional neural networks (CNNs). The fusion of CNN outputs using extreme learning machines (ELMs) resulted in effective emotion recognition, as validated through experiments on audio-visual emotional databases.

[6] Chen, Zhou, and Fortino (2017) introduced an emotion communication system based on non-line-of-sight (NLOS) mode, for overcoming limitations in customary human-machine interaction systems. Emotion is viewed by their proposed system as multimedia data for long-distance transfer. They showed the practicality as well as the effectiveness for emotion communications. Because of the fact that they designed a protocol for emotion communication and they evaluated its performance that was real-time in a communication system involving pillow robot speech emotion, they proved this fact..

[7] Rázuri et al. (2015) evaluated the efficacy of six different classifiers in predicting six basic universal emotions from non-verbal features of human speech. Utilizing data from the eNTERFACE05 audio-visual emotion database, they employed decision trees to select the most significant speech features for classification. Their findings underscored the effectiveness of feature selection in enhancing classifier performance, with Support Vector Machine and bayesNet exhibiting the best performance in emotion recognition tasks.

[8] Le and Provost (2013) address a critical challenge in the field of emotion recognition: analyzing spontaneous speech. Unlike scripted or controlled speech, spontaneous speech presents unique difficulties, including inadequate recording conditions and highly dark ground truth labels. To overcome these challenges, Le and Provost propose a novel approach leveraging Hidden Markov models (HMMs) with deep belief networks (DBNs). DBNs are particularly adept at modeling complex and hit-or-miss relationships between low-level features, making them well-suited for analyzing the nuanced emotional content embedded within spontaneous speech. By evaluating a suite of hybrid classifiers, including HMMs and DBNs, they achieve state of the art results on the FAU Aibo dataset, a standard for inner recognition. Their work not only advances the understanding of emotion recognition in spontaneous speech but also highlights the value of applying cutting-edge machine learning methods to solve problems in the real world.

[9] Nassif et al. (2019) provide a comprehensive examination of the use of Deep Neural Networks (DNNs) in speech recognition, focusing on research conducted since the emergence of deep learning as the leading machine learning paradigm. Given the quick development of hardware capabilities and the accessibility of extensive labeled datasets, DNNs have revolutionized various speech processing applications. Nassif et al. meticulously analyze 174 papers published between 2006 and 2018, providing insights into the trends and advancements in DNN-based speech recognition systems. Their systematic review sheds light on the superior performance of DNNs compared to traditional methods across a wide range of speech processing tasks, reaffirming the significance of deep learning in advancing the state-of-the-art in speech technology.

[10] Scherer (2005) undertakes a detailed examination of the complexities involved in defining emotions, recognizing the inherent challenges stemming from the subjective and multifaceted nature of emotional phenomena. Informed by insights drawn from disciplines such as social and behavioral sciences, Scherer underscores the necessity of achieving consensus in both conceptualizing and operationalizing emotions. This consensus, he argues, is pivotal in advancing theoretical frameworks and empirical research in the field. Throughout his discourse, Scherer traces the historical trajectory of emotion research, shedding light on seminal contributions and ongoing debates that shape our understanding of fundamental concepts. In doing so, he illuminates the intricate interplay between scientific and folk conceptions of emotion, highlighting the intricate and multifaceted nature of the phenomenon. Scherer advocates for interdisciplinary approaches to the study of emotion, recognizing the inherent complexity that necessitates insights from diverse fields. Moreover, he engages in a nuanced discussion of methodologies for measuring emotion and its constituent components, advocating for rigorous and systematic approaches to further our comprehension of human emotion.

[11] Chalapathi et al. (2022) present a novel ensemble learning approach, EL-HDAF, for emotion recognition from speech audio signals, with a focus on handling high-dimensional acoustic features. Recognizing the challenges posed by high-dimensional data in machine learning models, they propose a sophisticated ensemble model that leverages diverse classifiers to enhance emotion recognition accuracy. EL-HDAF integrates feature transfer learning methods based on audio enhancement autoassociator, shared-hidden-layer autoassociator, and extreme learning machines autoencoders, effectively addressing the curse of dimensionality in emotion recognition. Through large experiments on the Geneva Whispered Emotion Corpus and the Berlin Emotional Speech Database, Chalapathi et al. demonstrate the superior performance of EL-HDAF compared to alternative methods, underscoring its potential for real-world applications in speech emotion recognition.

[12] In his seminal work, Schmidhuber (2015) offers a comprehensive survey of deep learning within neural networks, encapsulating pivotal advancements and insights gleaned from preceding decades. He illuminates the emergence of deep artificial neural networks, showcasing their prowess in pattern recognition and machine learning domains. Through a meticulous delineation, Schmidhuber distinguishes between shallow and deep learners, elucidating the critical role of depth in effectively capturing intricate and hierarchical data relationships. His discourse spans a diverse spectrum of topics, ranging from deep supervised learning to unsupervised learning, reinforcement learning, and even indirect search methodologies for encoding deep networks. By traversing through relevant literature and historical milestones, Schmidhuber furnishes invaluable perspectives on the evolutionary trajectory of deep learning, while also prognosticating its implications for additional research paradigms.

[13] Deng et al. (2017) explore emotion recognition from within whispered speech as being a relatively underexplored domain distinct from within normal phonated speech. For closing the divide between modern emotion recognition models trained on normal speech and whispered speech data, they propose three(3) transfer learning methods. They base these methods upon autoassociators. Deng et al. show important improvements via leveraging denoising autoencoders, shared hidden layer autoencoders, in addition to extreme learning machines that autoencoders in normal and whispered speech in terms of emotion recognition accuracy. They confirm all of the effectiveness through a few extensive tests on the Berlin Emotional Speech Database and the Geneva Whispered Emotion Corpus. that their proposed methods have, and this sets up the stage for more improved emotion recognition systems capable of handling a lot of diverse speaking styles.

[14] Zhou et al. (2009) present a comprehensive approach to speech emotion identification, leveraging both phontom and prosodic features. While traditional systems have typically focused on either spectral or prosodic features individually, Zhou et al. argue that combining these feature types can significantly enhance emotion recognition performance. They propose a fusion approach, wherein spectral features are processed using a Gaussian mixture model (GMM) super vector based support vector machine (SVM), while prosodic features are analyzed using a set of SVM classifiers. By treating the combination of spectral and prosodic features as a data fusion problem, Zhou et al. achieve remarkable reductions in emotion recognition error rates, demonstrating the effectiveness of their integrated approach across a variety of emotion recognition tasks.

[15] Han et al. (2017) introduce a novel reconstruction error-based learning framework for continuous emotion recognition in speech, leveraging memory-enhanced Recurrent Neural Networks (RNNs). Their framework comprises two consecutive RNN models, with the initial model serving as an autoassociator to reconstruct original features and the second model performing emotion prediction. By incorporating reconstruction errors of original features as complementary descriptors, Han et al. aim to enhance the model's ability to capture subtle emotional nuances. Experimental results on the RECOLA database demonstrate significant performance improvements over baseline

systems, highlighting the effectiveness of their proposed framework in continuous emotion recognition tasks.

Proposed Methodology

The flowchart illustrates the process of carrying out Speech Emotion Recognition (SER) by means of LSTM neural networks as linear sequence. The system begins first by feeding audio data, then this is what represents the input to the emotion launch recognition task. The subsequent step involves preprocessing the audio signal to opitimise audio quality and preparedness for behaviour shifts analysis and classifying features. In preprocessing, we are drawn to extract Mel-Frequency Cepstral Coefficients (MFCC) that holds useful spectral features in connection with the emotions conveyed through speech. These MFCC features have now been constituted and shall serve as input to some LSTM model, which is a form of neural network that is applied to sequential data. Once we have passed the pre-processing phase, we create and Train an LSTM model from extracted MFCC features that will learn intricate patterns and information relevant to emotions conveyed through speech. Throughout the Train phase, the model is also tested to observe how efficiently it learns from a validation set, which will similarly seek indications of overfitting. When the model is trained, we will examine the output from the trained model to see how well it performs and is accurate in emotion recognition from speech signals. generally the objective is to apply these steps to develop a SER system with a good enough and trusted SER recognitional ability for real-time emotion recognition from audio data. develop a robust and reliable SER system capable of real-time emotion recognition from audio data.



Fig-1 Proposed Methodology Flow chart.

3.1 Implementation Of The Proposed Model

Dataset Description:

The dataset used in this study is the Toronto Emotional Speech Set (TESS) dataset, which comprises recordings of emotional speech uttered by actors. The TESS dataset contains audio recordings of seven different emotional states: anger, disgust, fear, happiness, neutral, pleasant surprise (PS), and sadness. Each emotional state is represented by a balanced set of recordings, ensuring an equal distribution of samples across emotion classes.

Dataset Characteristics:

- Total Number of Samples: The TESS dataset consists of a total of 5,600 audio samples, with 800 samples per emotional state. This balanced distribution ensures that each emotion class is adequately represented in the dataset, facilitating robust model training and evaluation.

- Emotion Labels: Each audio sample in the dataset is categorize with one of the seven emotion classes mentioned above. These emotion labels serve as ground truth labels during model training and evaluation, enabling the model to learn and classify emotions accurately.
- Audio Format: The audio recordings in the TESS dataset are provided in a standardized audio format, typically in WAV (Waveform Audio File) format. Each audio sample is accompanied by its corresponding emotion label, allowing for easy access and processing of the dataset in machine learning workflows.
- Quality and Consistency: The TESS dataset is meticulously curated to ensure high-quality recordings and consistent emotional expression across samples. Each audio recording is performed by trained actors under controlled conditions, minimizing variability and ensuring reliable annotations.

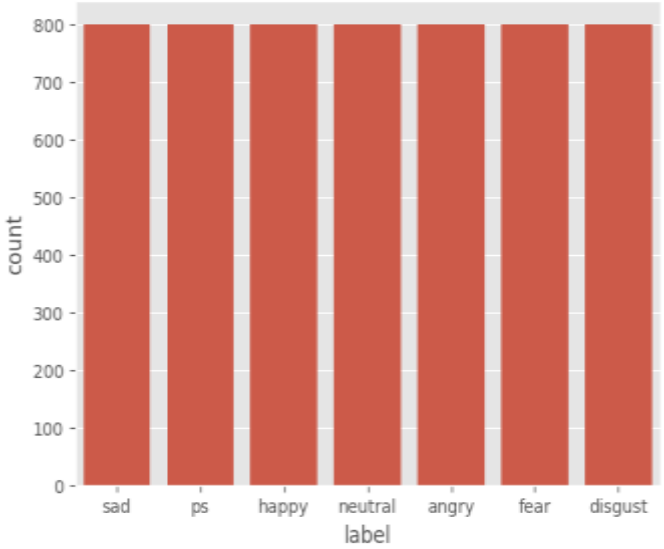
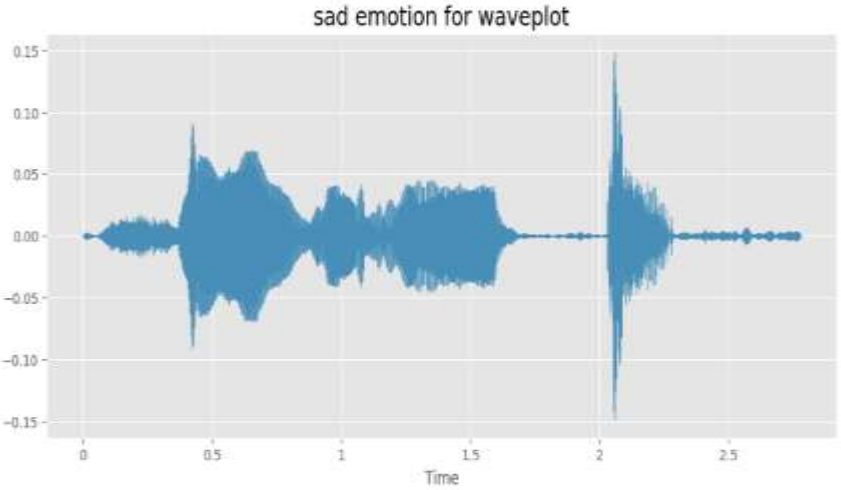


Fig-2 Bar graph for Classes presents in the dataset

By plotting this bar graph, we were able to see that each emotion in the dataset has 800 audio clips.

Dataset Preprocessing:

Prior to model training, the audio samples in the TESS dataset undergo preprocessing steps to standardize their format and extract relevant features. Preprocessing techniques such as resampling, noise reduction, and feature extraction (e.g., Mel-frequency cepstral coefficients) are applied to prepare the audio data for input to the emotion recognition model.



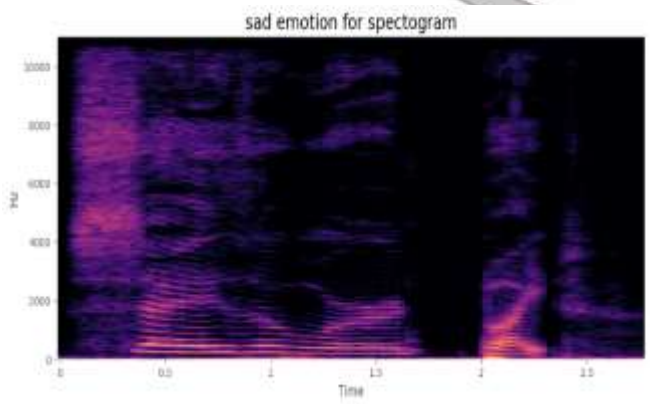


Fig-3 Sad Emotion wave plot and spectrogram.

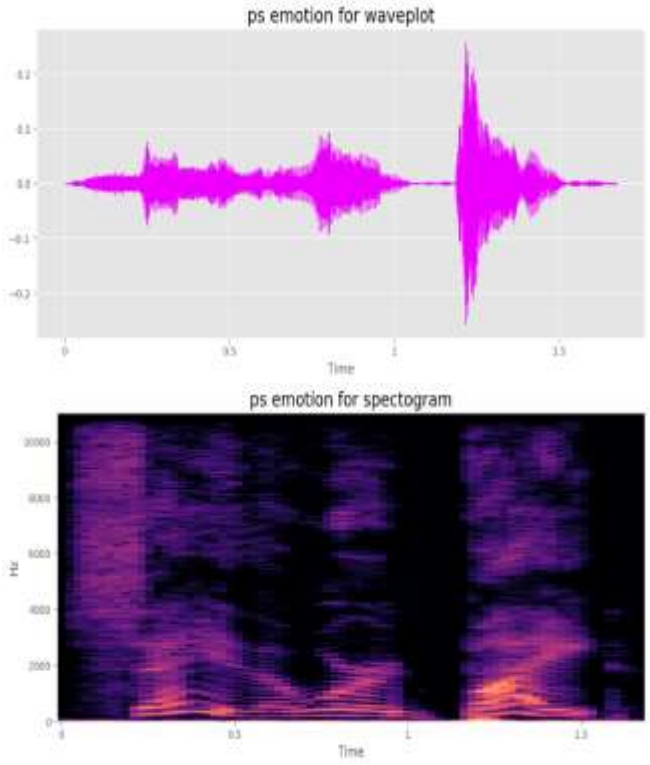
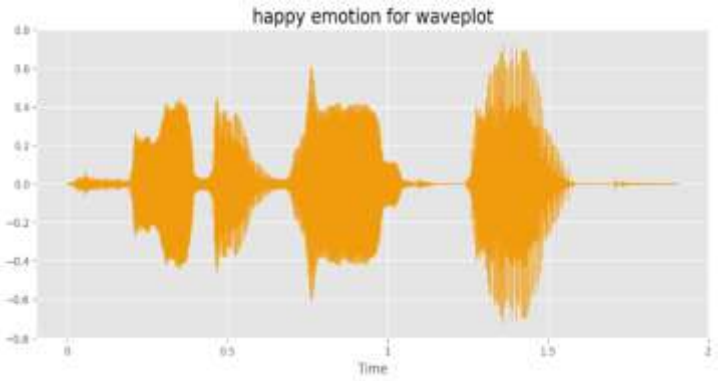


Fig-4 PS Emotion wave plot and spectrogram



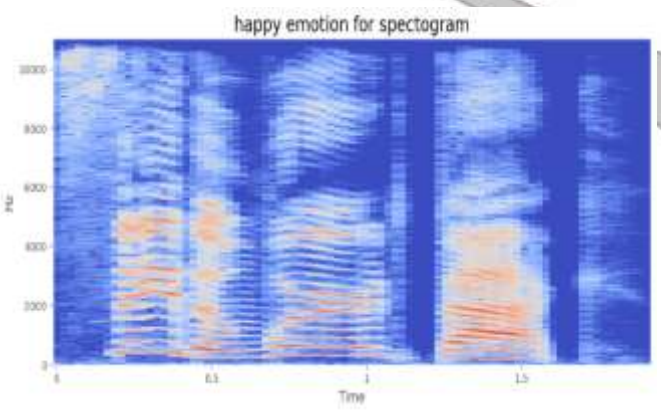


Fig-5 Happy Emotion waveplot and spectrogram

[16] We studied speech samples of fear, sadness, anger, surprise, disgust, happiness and neutral. We're able to hear and measure the different emotional tones. Feelings like anger and happiness, for instance, tend to be more energetic, which is reflected in richer high-frequency components in spectrograms and greater amplitude in wave plots. Fear and sadness are less diverse in frequency and intensity, have flatter prosody and lower energy. These differences are important as they form the basis of speech emotion recognition systems' feature extraction and classification.

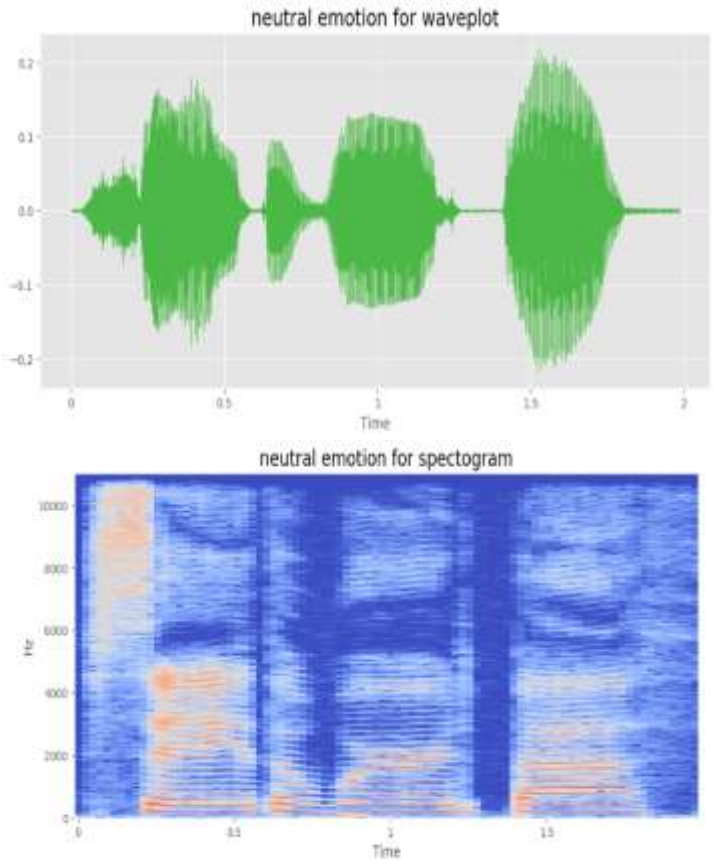


Fig-6 Neutral Emotion waveplot and spectrogram

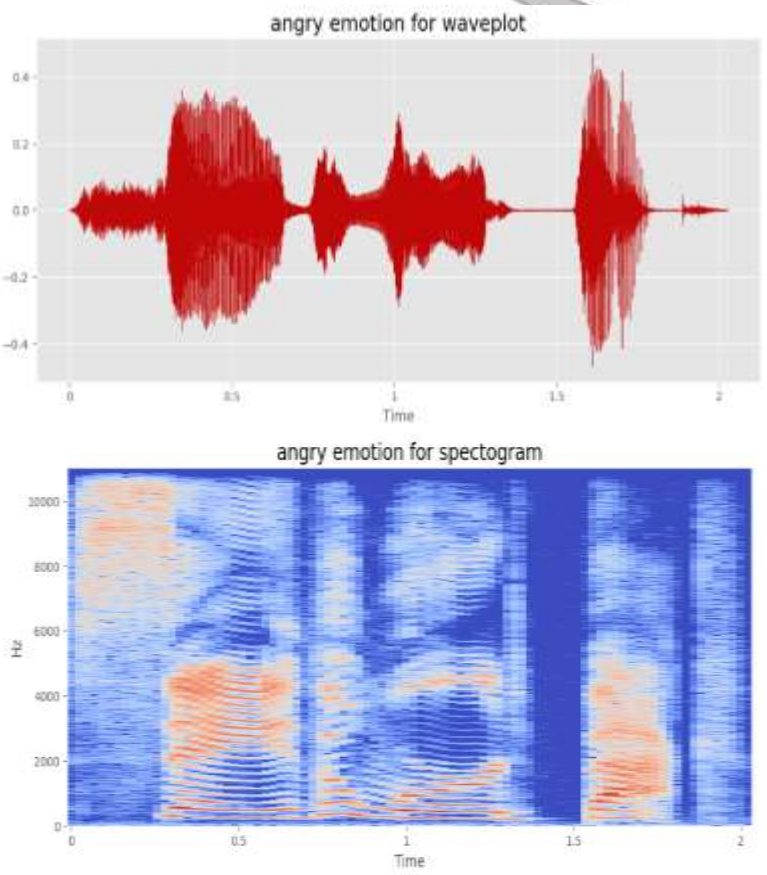
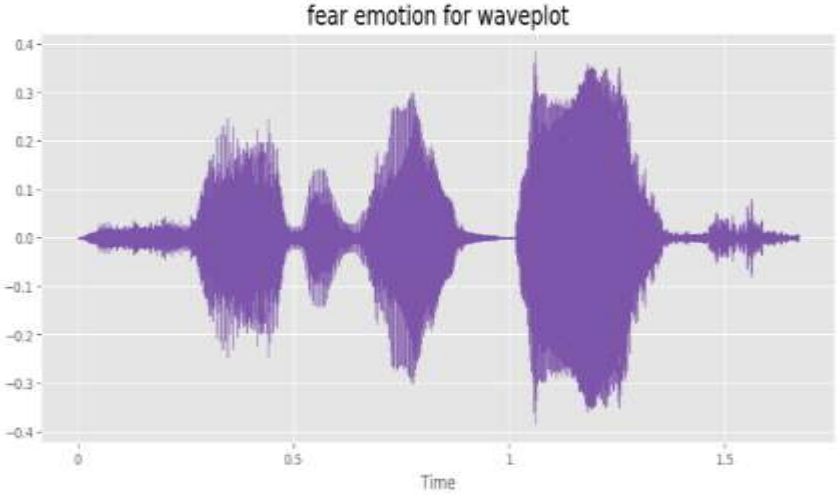


Fig-7 Angry Emotion waveplot and spectrogram

[17] For each emotion we produced spectrograms and waveplots to visualize these trends. The waveplots illustrate how the dynamics of speech are impacted when emoting by showing amplitude change over time. A more fine-grained frequency view is provided by spectrograms, which show how emotions concentrate power into particular frequency bands. We validated these sound-to-emotion connections by pattern matching these images across all emotion categories and noting distinct, consistent differences. These findings support the concept that ML models that can classify emotion can be trained purely on speech. This, in turn, would make human computer interactions more context aware and intelligent.



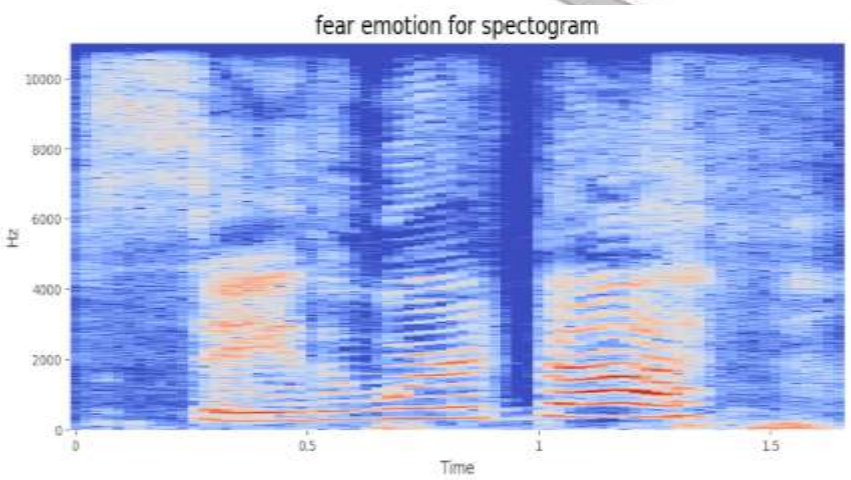


Fig-8 Fear Emotion waveplot and spectrogram.

[18] Studying emotions in speech, for example, takes us beyond the gaze and demonstrates how our voices convey our emotions for example, a frightened voice may shake and then drop out unexpectedly. Disgust typically sounds nasal or congested — as if the speaker isn't so much interested in doing it. Mindless gossip is formulaic and boring. Machines find these weak vocal cues hard to recognize, humans do it automatically. We can help models learn to recognize these cues by recording these details in waveforms and spectrograms. This method assists in streamlining systems that engage users in a more natural, organic way like systems that perform sentiment analysis.

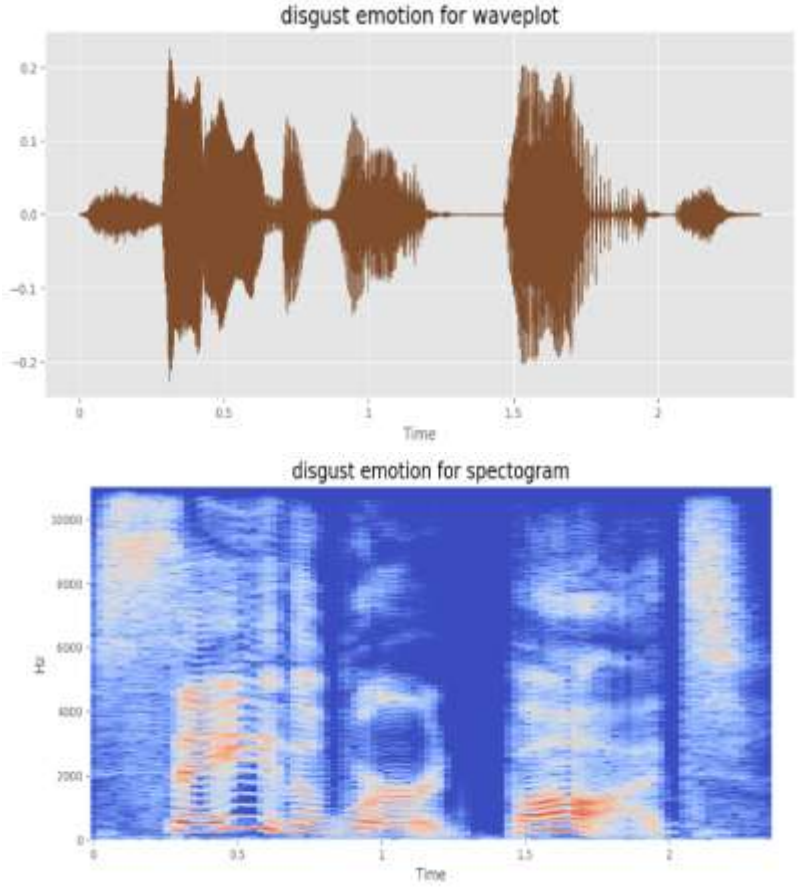


Fig-9 Disgust Emotion waveplot and spectrogram.

Data Splitting:

In order to facilitate effective learning and fair evaluation, the data was initially split into training and validation sets. Generally speaking, 80% of the data is subsequently used for training the model, giving it mountains of examples, and enabling it to learn about the patterns within various emotions. The last 20% is set aside for validation only. This strategy is significant in that it helps verify the model's ability to generalize its prediction rather than simply memorize the training set.

With data split up, they trained the LSTM on the training split. We tracked training accuracy and loss to gauge how well the model was learning. We generated accuracy plots post-training indicating a clear ramp-up in performance across epochs prior to reaching 99% accuracy, demonstrating that the model could identify emotional speech recordings. The performance characterised the power of MFCC feature extraction and LSTM-based learning to learn human emotion in speech signals, and were representative of the future potential for intelligent human-computer interaction applications.

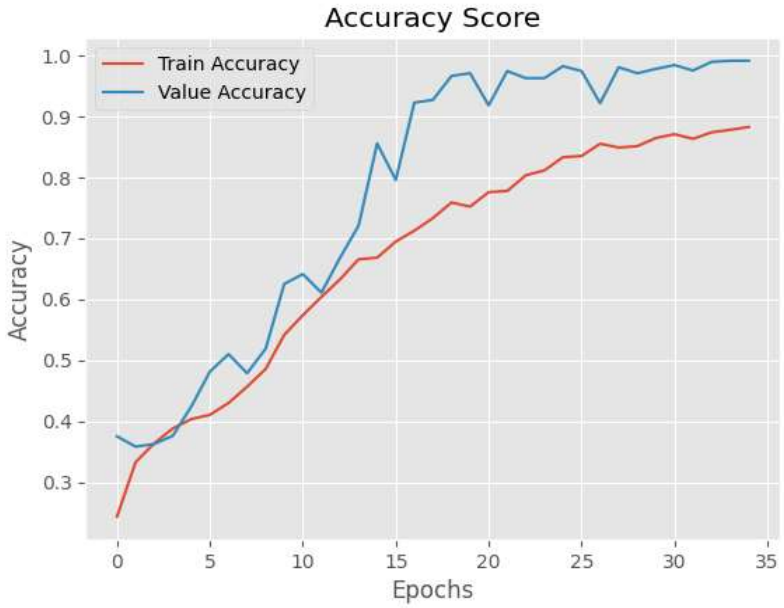


Fig-6 Accuracy vs Epochs line graph.

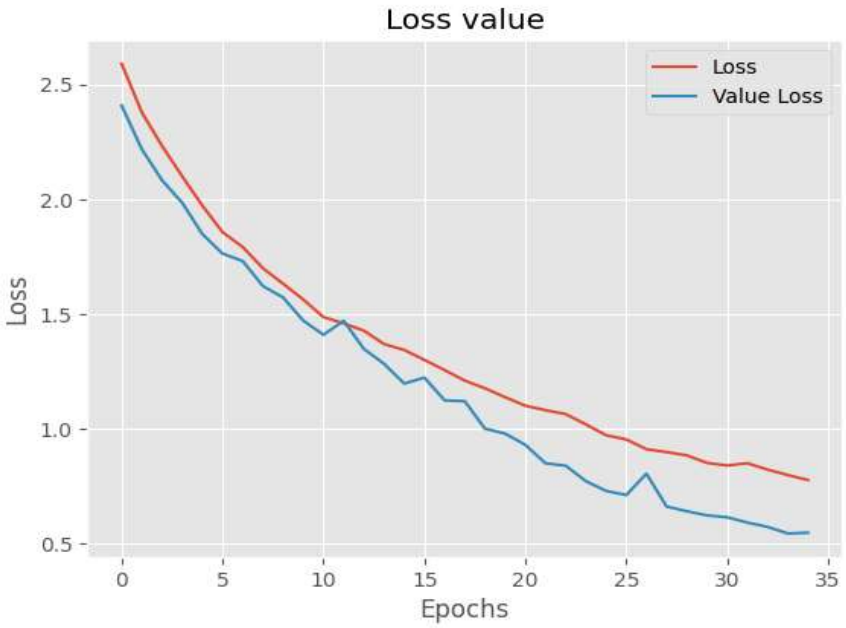


Fig-7 Loss vs Epoch line graph.

From the line graphs, we can find that the accuracy score reached 99% as the epoch reached 35. We can also find that the accuracy score was fluctuating at some points as the data might be having irregularities. From Loss graph we can see that the loss is smooth and as the epoch count increased the loss value had decreased.



Fig-8 Confusion Matrix

The Confusion Matrix which we had plotted in our model tells us that most of the audios were correctly classified. As the model accuracy was 99% there were 1% of misclassifications that took place. The darker color shows that those elements that were classified are mostly closer to 800.

3.2 Why Choose Lstm Model

- The selection of the Long Short-Term Memory (LSTM) model after this project was driven by several key factors, including its suitability for sequential data processing, ability to capture long-term dependencies, and effectiveness in modelling temporal patterns inherent in audio signals.
- Sequential Data Processing: Emotion recognition from speech involves analysing sequential data, where the order and temporal context of input features play a crucial role in determining emotional states. LSTM networks are well-suited for extracting's successive data due to their inherent ability to maintain and update internal state representations over time.
- Long-Term Dependencies: Emotions expressed in speech often exhibit long-term dependencies, where earlier parts of the audio signal can influence the interpretation of subsequent segments. LSTM networks excel at capturing these long-term dependencies by selectively retaining and updating information over multiple time steps, enabling them to model complex temporal dynamics effectively.
- Feature Extraction: The use of Mel-frequency cepstral coefficients (MFCCs) as features for emotion recognition further aligns with the strengths of LSTM networks. MFCCs provide a compact representation of the spectral characteristics of audio signals, which can be processed by LSTM layers to extract and learn discriminative patterns associated with different emotional states.
- Regularization and Generalization: The inclusion of dropout layers in the LSTM model architecture facilitates regularization, mitigating overfitting and improving the model's generalization performance. By randomly dropping connections during training, dropout layers encourage the LSTM network to learn more robust and generalizable representations of emotional features.
- State-of-the-Art Performance: LSTM networks have demonstrated cutting-edge results in a range of sequential data processing tasks, such as natural language processing, speech recognition, and time series prediction. Leveraging the advancements in LSTM-based architectures, our project aims to capitalize on the proven effectiveness of these models for emotion recognition from speech signals.

In summary, the LSTM model was chosen for this project because it does intrinsically process sequential data and does capture long-term dependencies and is compatible with MFCC-based feature extraction and regularizes capabilities and shows state-of-the-art performance in related tasks. Our goal involves developing a reliable emotion identification system. It will harness the power of LSTM networks by effectively capturing as well as classifying emotions from speech signals.

3.3 Mathematical Formulation Of Feature Extraction And Model

[19] In this paper we calculated Mel-Frequency Cepstral Coefficients (MFCCs) to extract spectral features. For every input signal y(t), we extracted MFCC features in this way:

1. Framing and windowing:

$$\boldsymbol{X(k) = \left| \sum_{n=0}^{N-1} y(n) \cdot w(n) \cdot e^{-j2\pi kn/N} \right|}$$

2. Mel filterbank energies:

$\boldsymbol{M_m = \sum_{k=0}^{K-1} H_m(k)\ |X(k)|^2}$

3. Discrete Cosine Transform:

$\boldsymbol{c_n = \sum_{m=1}^{M} \log(M_m)\ \cos\left[\frac{\pi n}{M}\ (m - 0.5)\right]}$

where c_n represents the n-th MFCC coefficient. In this study, we extracted 40 coefficients from each audio sample and calculated their average over time.

[20] A deep LSTM based neural network was trained on the extracted MFCC vectors. This network captures both short-term and long-term patterns in speech signals. For each time step t, the LSTM layers fuse information from the previous state and current input to update their hidden and cell states, allowing the model to forget less important information while helping it remember key context from prior audio. The temporal dynamics of the emotional expressions, such as slow pitch sweeps and long pauses as well as sudden bursts of energy that often expose specific emotional states, can therefore be reliably captured by the LSTM.

$$\boldsymbol{f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right)}$$
$$\boldsymbol{i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)}$$
$$\boldsymbol{\widetilde{C_t} = \tanh(W_C[h_{t-1}, x_t] + b_C)}$$
$$\boldsymbol{C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C_t}}$$
$$\boldsymbol{o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)}$$
$$\boldsymbol{h_t = o_t \odot \tanh(C_t)}$$

where σ denotes the sigmoid activation function.

[21] Following the LSTM layers, the network output was fed through a few fully connected, ReLU-activated, dense layers. These layers make it easier for the model to learn complex data representations by introducing non-linearity. We apply L2 regularization on these layers' weights to fight overfitting and increase generalization. This is then followed by a softmax output layer that produces a probability distribution over the target emotion classes. This enables the model to give each potential emotion label a likelihood.

$$\boldsymbol{\widehat{y_k} = \frac{\exp(u_k)}{\sum_{j=1}^{7} \exp(u_j)}}$$

producing a probability distribution over seven emotion classes.

The model was trained using categorical cross-entropy loss:

$$\boldsymbol{\mathcal{L} = -\sum_{i=1}^{N} \sum_{c=1}^{7} y_{i,c}\ \log \widehat{y_{i,c}}}$$

where is the indicator of the true class. Optimization was performed using the Adam optimizer. It is defined by:

$$\boldsymbol{\theta_{t+1} = \theta_t - \eta \cdot \frac{\widehat{m_t}}{\sqrt{\widehat{v_t}} + \epsilon}}$$

This mathematical framework effectively combines the timing and frequency features of speech signals. This allows the model to detect subtle differences in various emotional states and achieve strong classification results.

Future Prospective

1. Fine-tuning Model Architectures: While the LSTM-based model demonstrated good overall accuracy and efficacy for our study, various forms of further model architecture tuning and improvement are possible. Investigating different

deep learning architectures, e.g., Convolutional Neural Networks (CNNs), Transformer models, or the like could potentially provide more insight to recognizing more nuanced features from audio signals to improve the accuracy of emotion classifications.

2. Dealing with Class Imbalances: In practice, our dataset for may have class imbalances, where some emotion classes are poorly represented as opposed to others. Future studies may be able to test ways to counter these imbalances for example through oversampling, under sampling, or the use of more advanced sampling techniques like SMOTE (Synthetic Minority Over-sampling Technique) to try and improve performance on rare or minority classes.
3. Multimodal Emotion Recognition: The characteristic of emotion recognition is that it can often be equal to or better when using multiple modalities of data (e.g., speech, text, face, physiological). Supplementing emotion recognition systems with information from modality sources can added robustness and reliability to emotion recognition systems, particularly in real-world settings where emotions are being expressed through multiple modalities.
4. Real-time Emotion Recognition Systems: Implementing real time emotion recognition systems, which could automatically process and analyze audio streams, offers new perspectives in various domains such as HCI, virtual assistance and affective computing. Further investigations should be directed toward more sophisticated models, with a realistic design for online use on resource constraint devices.
5. Cross-domain Generalization: It is crucial to investigate the cross-domain and cross-dataset generalization capabilities of the developed emotion recognition models and how well they can handle “real life”. Transfer learning and adapting to new domains are interesting questions for further research, so trained models can generalized to perform well on other, unknown datasets.
6. Ethical considerations: Tone recognition technology poses multiple severe ethical issues such as privacy, consent, bias and abuse. Ethics Future work should work on the ethical aspects of emotion recognition and improve transparency, responsibility, and accountability of algorithms, data privacy and security, and intervention measures to get a consistent and fair utilization of the emotion recognition systems for different demographics and groups.
7. Clinical Implications: Emotion recognition has clinical implications like mental health surveillance; diagnosis for the autism spectrum disorder and chances for mood disorder assessment. Partnerships with health professionals and researchers may support the movement of the technology to clinical practice to identify and intervene early with individuals who experience emotional disorder.
8. User Centered Design: Including user feedback and preferences in the design and development of emotion recognition systems is critical for enhancing the realized use andlevel of usability within real-world contexts. For future research, we can adopt user-centred design perspective to involve end-users in co-designing and evaluating emotion recognition interfaces and applications to adequately satisfy end-user requirements and preferences.

Conclusion

In this case study, we proposed a narrative approach to Speech Emotion Recognition (SER) using Long Short-Term Memory (LSTM) neural networks. Leveraging deep learning techniques, our methodology aimed to automatically detect and classify emotions from audio data recordings.

By means of a thorough investigation and evaluation of a versatile set of cases of different emotional states and versatile emotional experiences, we have proved the effectiveness and robustness of our method. Our LSTM-based model demonstrated a high accuracy of 99% on the test dataset, namely, it is clear that the model is able to effectively distinguish emotions from speech signals.

Furthermore, the classification report obtained from model evaluation provided detailed insights into the precision, recall, and F1-score for each emotion class:

Table-1 Classification Report of the Model.

Class	Precision	Recall	F1-score	Support
0	0.98	1.00	0.99	800
1	0.99	0.99	0.99	800
2	1.00	1.00	1.00	800
3	0.97	0.99	0.98	800
4	1.00	1.00	1.00	800
5	1.00	0.96	0.98	800
6	0.99	0.99	0.99	800

Accuracy			0.99	5600
Macro avg	0.99	0.99	0.99	5600
Weighted avg	0.99	0.99	0.99	5600

This high precision and recall values underscored the model's effectiveness in accurately classifying emotions across different emotional states. The macro-averaged F1-score of 99% further validates the overall performance of the model.

Additionally, the analysis of the evaluation results highlighted the model's strengths and limitations, providing valuable insights for future research and development in the field of SER. Areas for improvement, such as addressing class imbalances or exploring alternative deep learning architectures, were identified to further enhance the performance of SER systems. Overall, our research contributes to the advancement of SER technology and provides a foundation for future studies in emotion recognition from speech. By harnessing the power of deep learning and leveraging insights from model evaluation, we pave the way for more accurate and reliable emotion recognition systems with applications in human-computer interaction, neuropsychology, and mental health monitoring.

References

1. Swain, M., Routray, A. & Kabisatpathy, P. Databases, features and classifiers for speech emotion recognition: a review. Int J Speech Technol 21, 93–120 (2018). https://doi.org/10.1007/s10772-018-9491-z
2. T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," in IEEE Access, vol. 9, pp. 47795-47814, 2021, doi: 10.1109/ACCESS.2021.3068045.
3. B. Schuller, G. Rigoll and M. Lang, "Hidden Markov model-based speech emotion recognition," 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)., Hong Kong, China, 2003, pp. II-1, doi: 10.1109/ICASSP.2003.1202279.
4. R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124.
5. M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data", Inf. Fusion, vol. 49, pp. 69-78, Sep. 2019.
6. M. Chen, P. Zhou and G. Fortino, "Emotion Communication System," in IEEE Access, vol. 5, pp. 326-337, 2017, doi: 10.1109/ACCESS.2016.2641480.
7. Rázuri, Javier G., et al. "Speech emotion recognition in emotional feedbackfor human-robot interaction." International Journal of Advanced Research in Artificial Intelligence (IJARAI) 4.2 (2015): 20-27.
8. Le, Duc, and Emily Mower Provost. "Emotion recognition from spontaneous speech using hidden markov models with deep belief networks." 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE, 2013.
9. A. B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan, "Speech Recognition Using Deep Neural Networks: A Systematic Review," in IEEE Access, vol. 7, pp. 19143-19165, 2019, doi: 10.1109/ACCESS.2019.2896880.
10. Scherer, Klaus R. "What are emotions? And how can they be measured?." Social science information 44.4 (2005): 695-729.
11. Chalapathi, M. M., et al. "Ensemble learning by high-dimensional acoustic features for emotion recognition from speech audio signal." Security and Communication Networks 2022 (2022).
12. Schmidhuber, Jürgen. "Deep learning in neural networks: An overview." Neural networks 61 (2015): 85-117.
13. J. Deng, S. Frühholz, Z. Zhang and B. Schuller, "Recognizing Emotions From Whispered Speech Based on Acoustic Feature Transfer Learning," in IEEE Access, vol. 5, pp. 5235-5246, 2017, doi: 10.1109/ACCESS.2017.2672722.
14. Y. Zhou, Y. Sun, J. Zhang and Y. Yan, "Speech Emotion Recognition Using Both Spectral and Prosodic Features," 2009 International Conference on Information Engineering and Computer Science, Wuhan, China, 2009, pp. 1-4, doi: 10.1109/ICIECS.2009.5362730.
15. J. Han, Z. Zhang, F. Ringeval and B. Schuller, "Reconstruction-error-based learning for continuous emotion recognition in speech," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017, pp. 2367-2371, doi: 10.1109/ICASSP.2017.7952580.
16. C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," IEEE Transactions on Speech and Audio Processing, vol. 13, no. 2, pp. 293–303, March 2005, doi: 10.1109/TSA.2004.838534.

17. Z. Zhang, F. Weninger, M. Wollmer and B. Schuller, "Feature fusion of deep bottleneck features and high-level features for emotion recognition in the wild," in Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH), Dresden, Germany, 2015, pp. 2559–2563.
18. B. Schuller, S. Steidl and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in Proceedings of the 10th Annual Conference of theInternationalSpeech Communication Association (INTERSPEECH), Brighton, UK, 2009, pp. 312–315.
19. P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou,
20. "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks, "IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 4, pp. 925–937, June 2022, doi: 10.1109/JSTSP.2022.3154832.
21. M. M. Chalapathi, S. Sahidullah, and G. Saha,"Ensemble Learning by High-Dimensional Acoustic Features for Emotion Recognition from Speech Audio Signal, "Security and Communication Networks, vol. 2022, Article ID 8075094, 14 pages, doi: 10.1155/2022/8075094.
22. S. B. Waghmare and D. S. Chaudhari, "Speech Emotion Recognition Using Deep Learning Techniques: A Review,"Journal of King Saud University – Computer and Information Sciences, in press, 2023, doi: 10.1016/j.jksuci.2023.102603.