Temporal Convolutional Networks for Speech Emotion Recognition: A Benchmark Study against Deep Learning Models

Nitasha Rathore, Pratibha Barua, Arhaan Sood, Bandaru Yogesh Kumar and Ashutosh Singh
Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi, Delhi, India

Keywords: Speech Emotion Recognition (SER), Long Short-Term Memory (LSTM), Recurrent Convolutional Neural Networks (RCNN), Artificial Neural Network (ANN), Recurrent Neural Networks (RNN), Temporal Convolutional Networks (TCN).

Abstract: As technology becomes more and more human-centric, the ability to recognize and interpret emotions from speech is becoming more than just an innovation it is a necessity. Speech Emotion Recognition (SER) is a field that sits at the nexus of artificial intelligence and human communication. It offers perspectives on both our spoken words and our emotions. In addition to enhancing digital assistants, SER is revolutionizing mental health monitoring and how people interact with robots. This research investigates the intricate world of emotion-laden speech to learn how cutting- edge deep learning models, including Temporal Convolutional Networks (TCN), Artificial Neural Networks (ANN), Recurrent Convolutional Neural Networks (RCNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks, decode the smallest emotional cues. While some models employ spatial patterns in speech patterns, others are very adept at recognizing temporal links. Each model has its own merits. We analyze their performance in emotionally enriched environments and show a potential proof-of-concept how can they impact at the human-computer level. This study imagines a future of SER that is relevant to sympathetic AI, which is essential for the connections that exist between AI and humans, beyond the current measures of numbers and accuracy scores. So, the question is: How close are we really, as we strain those outer limits, to teaching machines the vocabulary of feelings?.

1 INTRODUCTION

Imagine living in a world where machines not only hear you but understand how you feel. The rapidly growing discipline of Speech Emotion Recognition (SER) which nests in the fields of artificial intelligence, psychology and human- computer interaction holds this promise. While SER will provide a new understanding of human emotions by analyzing speech elements (pitch, tone, intensity, etc.) to make digital conversations more intuitive, empathetic and smarter.

Along with the rise of accessible and lighter frameworks such as TensorFlow and Porch, SER has made an evolutionary leap from early, simplistic rule-based systems to complex deep learning architectures proficient at emotion detection with greater precision than ever. Early methods on the other hand utilized machine learning methods, such as support vector machine (SVM), decision trees, and Gaussian Mixture Models (GMMs) and features that were generated manually. But these models typically had difficulty resolving ambiguity, they were not contextually aware, they could also miss the fine emotional changes that occur in speech. Figure 1shows flow of SER Training.

In fact, the true leap in deep learning architectures were LSTM (Long Short-Term Memory) networks, which enable a more sophisticated understanding of feelings and the capacity to remember longer term dependencies to help understand speech. But what about million odal fusion with SER Technics That expands the limits. When combining speech with elements of non-verbal communication body language, facial expressions, etc. we get closer to communicating the full range of human emotion.

To examine the effectiveness of several deep learning models, we compare LSTMs, Temporal Convolutional Networks (TCNs), and Recurrent Convolutional Neural Networks (RCNNs) for SER. We also examine the role of transformer models and

Rathore, N., Barua, P., Sood, A., Kumar, B. Y. and Singh, A.
Temporal Convolutional Networks for Speech Emotion Recognition: A Benchmark Study against Deep Learning Models.
DOI: 10.5220/0013889500004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 2*, pages 753-759
ISBN: 978-989-758-777-1

attention mechanisms, which could revolutionize the profession by enhancing contextual awareness.

With the emergence of emotionally intelligent AI, SER is not just a research problem but the future of human-machine interaction. This essay examines the developments that have influenced SER and paved the way for a time when technology will be able to understand our feelings in addition to our words.

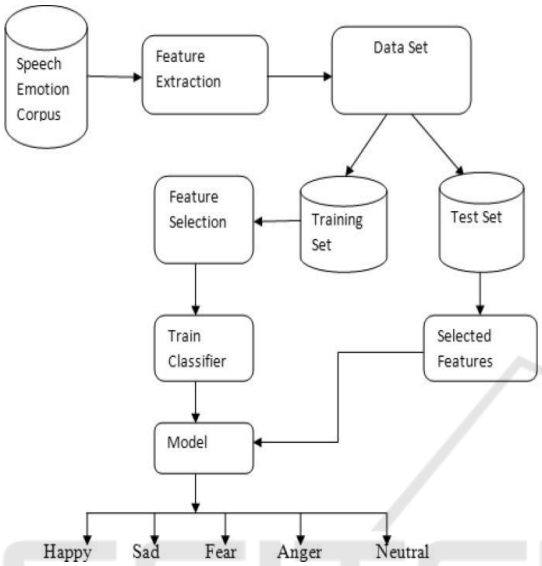

Figure 1: Flow of SER Training.

2 RELATED WORKS

The Research paper used the Gaussian Mixture Model to establish a general understanding of SER measurement achieving an accuracy rate of 89.12%. The method yielded positive findings however researchers recognized its dependence on recommended features since it failed to adapt to complex voice patterns.

The study authors mentioned in their work (Aouani, Hadhami & Benayed, Yassine 2020). that deep learning combined with SVM using all kernels produced 83.3% accuracy. Due to kernel-based processing requirements, the experimental set-up showed effective performance although it demanded strong computational power for functionality.

A noise-cleaning method led to an achievement of 71.75% accuracy when CNN was integrated with CNN+RNN for emotion analysis of audio and video content as described in the study (M. Singh and Y. Fang., 2020). Despite its non-nuisance to various audio types and reduction in performance with noisy datasets, the trial method achieved successful outcomes.

A research analysis (Mustaqeem, M. Sajjad and S. Kwon, 2020) proved that clustering-based SER achieved 95% accuracy through the integration of bidirectional connections with Belts. The model attained its final objective until it faced two significant challenges regarding processing extended datasets and running in real-time.

An auto encoder system at Patel, N., Patel, S. & Mankad obtained 90% accuracy by combining SVM and Decision Tree with CNN for reducing data dimensions. Limited dataset containment led to unsatisfactory performance because the system did not provide sufficient functionality when applied in real- time.

The authors behind (Lieskovská, E. et al., 2021) showed that attention mechanisms in deep learning models achieved an 85.2% success rate in SER tasks. The research evidence shows that attention approaches improve test outcomes but they demand additional processing power and their performance output changes based on the specifics of evaluated datasets.

The research in (Bagus Tris Atmaja et al., 2022) used SVM together with MLP and LSTM and handcrafted features to obtain a 78.8% accurate assessment of bimodal SER. This method faced difficulties in analyzing dynamic audio data because it used handcrafted features which restricted its robustness features.

The researchers in (Aggarwal et al., 2022) created a system that merged Reset VGG16 pre-trained architecture with DNN models to function in two directions achieving 96.26% accuracy for SER. High precision outcomes came from using pre-trained models yet achieving adaptations in this method required large extensive datasets because of its complexity.

The combination of CNN+LSTM with a stochastic fractal search optimization algorithm generated 97.38% accuracy levels according to data in the Study (A.A. Abdelhamid et al., 2022). The advanced optimization method faced implementation challenges because its deployment proved to be too complex.

The EEG-based SER evaluation carried out in (Houssein, E.H et al., 2022) demonstrated various ML approaches to assess brain activities through EEG testing resulting in 87.25% accuracy from examinations of SVM, ANN, Random Forest and Decision Tree and KNN, RNN, and CNN. This method became impractical for large-scale applications because researchers acquired EEG data.

A joint approach uniting NLP and DLSTA conducted deep semantic analysis of Service

Excellence big data to achieve 93.3% accuracy efficiency as documented in Study

(Guo, Jia., 2022). The model consisted of reliable functionality involving extensive data evaluation although its complex preprocessing requirements created obstacles for real-time system deployment because of extensive feature specifications.

Through the combination of CNN and multiple-head convolutional transformers researchers achieved 82.31% accuracy according to information found in the document

9 (Ullah et al., 2023) which utilized IEMOCAP and RAVDESS databases. The transformer architecture reached adequate accuracy benchmarks but this achievement came at the expense of system performance speed and resource-intensive memory consumption.

This study by (Samaneh Madanian et al., 2023) performed an organized review that showed how ML technology brought together SVM and Random Forest algorithms and noise reduction methods applied to MFCC extraction in reaching 91% precision. This approach resulted in improved speech data tolerance but caused a decreased response quality in noise-free situations because of overused data augmentation methods.

3 RESEARCH METHODOLOGY

The research employed deep learning models for systematic speech emotion recognition (SER), encompassing data collection, preprocessing, feature extraction, and classification. The dataset was first organized by labeling speech recordings, followed by visualization using waveform and spectrogram representations. Mel-Frequency Cepstral Coefficients (MFCC) captured key spectral features. Preprocessing involved dimension expansion and one-hot encoding for compatibility with deep learning architectures. Various models, including TCN, RNN, ANN, RCNN, and LSTM, were tested. The dataset was split into training, validation, and test sets, and model performance was assessed using confusion matrices, validation accuracy, and loss metrics. The Toronto Emotional Speech Set (TESS), featuring 2,800 recordings from two female actors expressing seven emotions (anger, contempt, fear, happiness, surprise, sadness, and neutrality), was used. TESS is widely utilized in affective computing and machine learning to enhance emotion-aware applications. Figure 2 shows research methodology.

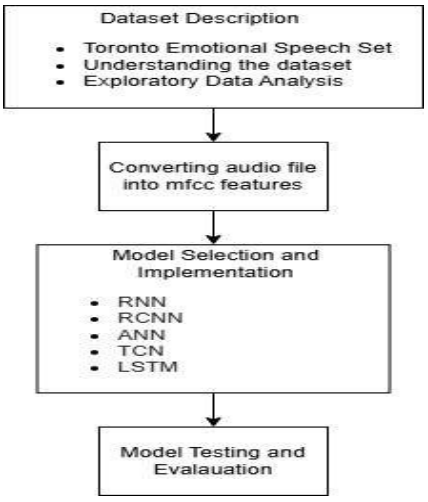

Figure 2: Research Methodology.

3.1 Data Preprocessing

The voice dataset was methodically ready for model training and analysis. For precise emotion identification, audio files were loaded and sorted by filename. To ensure consistent duration and offset across files, Librosa was utilized to clean signals, extract features, and minimize noise. While MFCC collected important spectral features for Speech Emotion Recognition (SER) and stored qualities in numerical form, waveform and spectrogram representations examined changes in pitch, tone, and frequency.

To ensure alignment for training, processed features were prepared for model input and categorical labels were one-hot encoded to fit deep learning models. This pipeline improved model performance in emotion classification, decreased variability, and optimized the dataset.

3.2 Exploratory Data Analysis

EDA (exploratory data analysis) helps to explore and understand the distribution of emotions in the dataset which is the most influencing factor for such types of analysis. The speech samples and their emotional properties were analyzed using a variety of statistical and graphical techniques. Wave plots displaying speech signals in the time domain allowed us to study amplitude and intensity variability between moods. Such variations provided insights into the way emotional expressions influence dynamics of speech.

Moreover, MFCC visualizations were also used in this analysis to capture the speech spectrum, considering how frequency components are affected in different emotional states. Figure 3 shows Bar Plot

for Audio Files in Dataset. Through tone, pitch and speech modulation analytics, distinguishing emotional pathways were identified and each differentiated emotion was tracked. Figure 4 shows Mel-Spectrograms for each emotion.

EDA played a significant role in optimizing feature selection, as well as guaranteeing that the deep learning models could accurately learn the data and classify emotions based on speech. It recognized emotions based on the fact that the underlying patterns were underlying trends. Figure 5 shows Wave Plots are obtained for each emotion.

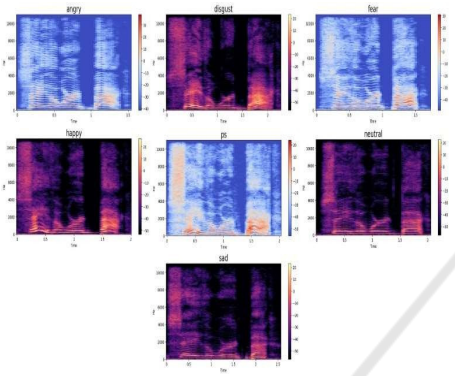

Figure 3: Bar Plot for Audio Files in Dataset.

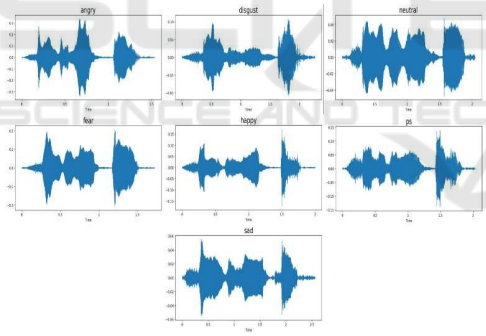

Figure 4: Mel-Spectrograms for Each Emotion.

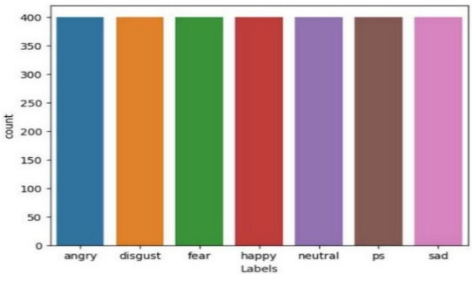

Figure 5: Wave Plots Are Obtained for Each Emotion.

Model Selection and Implementation: Models with deep learning capabilities are selected to process the temporal and sequential data for speech emotions recognition (SER). Other model was chosen based on its potential which is able to extract and classify speech signal features. Model performance on MFCC- extracted features was evaluated by overall confusion matrices, validation accuracy, and validation loss metrics in order to ensure robust emotion classification.

Long Short-Term Memory (LSTM): Long-term dependencies in sequential data allow LSTM networks to perform exceptionally well in speech-based emotion identification. Their gated architecture improves speech context retention by resolving the vanishing gradient issue that conventional RNNs face. To greatly increase recognition accuracy, MFCC characteristics were processed via several LSTM layers before going through a dense output layer for emotion categorization.

Recurrent Neural Networks (RNN): Since RNNs can analyses sequential input, they were used as a baseline model for time-sensitive speech analysis. Recurrent and fully connected layers were used to classify the MFCC features. However, they were less effective than LSTMs due to vanishing gradient restrictions, which hindered their capacity to learn long-term dependencies.

Artificial Neural Network (ANN): Using a fully linked feedforward architecture, ANNs served as a benchmark model to evaluate the efficacy of MFCC-based feature extraction. Compared to sequential models such as LSTMs and RNNs, they fared quite well, but their accuracy was restricted by their incapacity to grasp temporal correlations in speech.

Recurrent Convolutional Neural Network (RCNN): Recurrent and convolutional architectures were integrated in RCNNs to enhance feature learning. Recurrent layers found temporal patterns in speech, while convolutional layers collected spatial characteristics from MFCC representations. By utilizing both sequential and spatial learning capabilities, our hybrid model improved the accuracy of emotion classification.

Temporal Neural Network (TCN): TCNs successfully modelled long-range dependencies without the processing burden of LSTMs by using dilated causal convolutions rather than recurrence. By modelling the MFCC information with convolutional layer, TCNs managed to learn temporal linkages efficiently, whilst achieving comparable accuracy.

The comparative study of deep learning methods for SER was help by each model. Though hybrids such like RCNNs achieved strong performance, or effective architectures like TCNs were promising, LSTMs performed superior since they could hold long dependence. The findings provide critical

insights toward the selection of suitable deep learning architectures for speech emotion recognition use cases

4 RESULTS

Evaluation of deep learning models for Speech Emotion Recognition (SER) yielded varying degrees of accuracy. The performance of LSTM was remarkably accurate, with an index of 96.42% in distinction long-tem reliance. It is important to notice that with the 98.92% and 98.65% achievable accuracies, TCN and RCNN subjected methods were the best among all others, emphasizing the strength of temporal convolutional and hybrid convolutional approaches (Hybrid wins both, but at the cost of time and resources). Second, with a 98.57% success rate, ANN has been demonstrated efficiency when trained on carefully extracted MFCC features. RNN performed the worst, at 87.14%, presumably due to the vanishing gradient problem for learning over longer periods of input. Table 1 illustrate Results of deep learning architectures.

Overall, the top-performing models for SER were LSTM, RCNN, and TCN that yielded higher accuracy and alternatives to traditional recurrent architectures. Figure 6 shows Confusion Matrices obtained for each model.

Table 1: Results of Deep Learning Architectures.

Deep Learning Model	Accuracy Rate (%)
LSTM (Long Short-Term Memory)	96.42
RNN (Recurrent Neural Networks)	87.14
RCNN (Recurrent Convolutional Neural Networks)	98.65
ANN (Artificial Neural Networks)	98.57
TCN (Temporal Convolutional Networks)	98.92

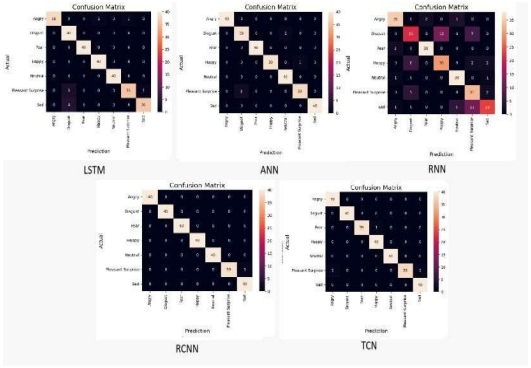

Figure 6: Confusion Matrices Obtained for Each Model.

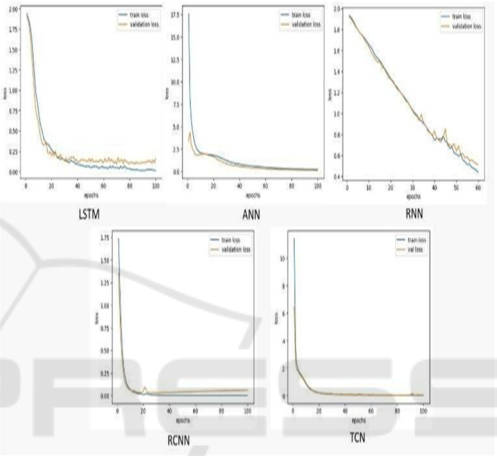

Figure 7: Training/Validation Loss V/S Epochs Plots for Each Model.

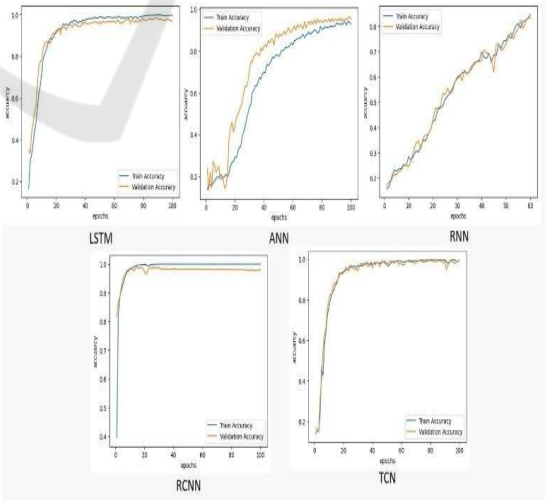

Figure 8: Testing/Validation Accuracy V/S Epochs Plots for Each Model.

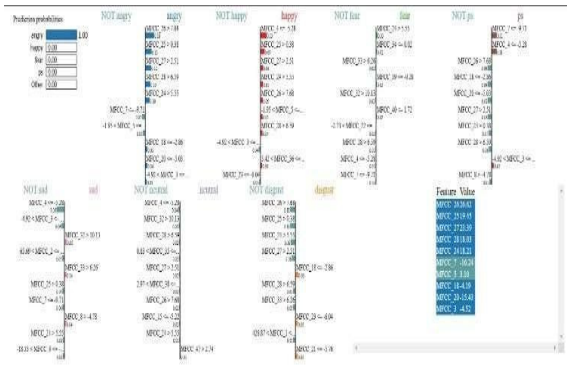

Figure 9: TCN Lime Chart.

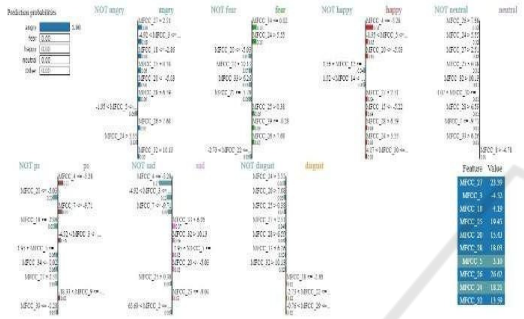

Figure 10: RCNN Lime Chart.

Figure 11: ANN Lime Chart.

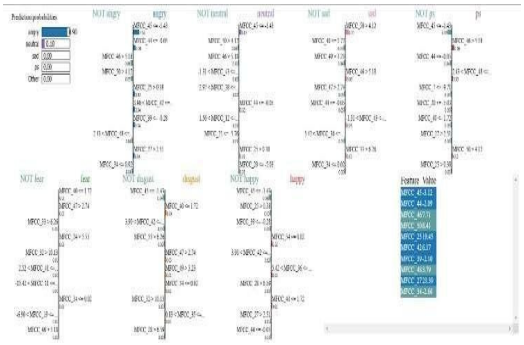

Figure 12: RNN Lime Chart.

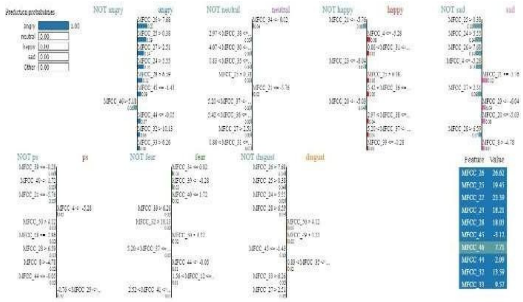

Figure 13: LSTM Lime Chart.

REFERENCES

"Speech emotion recognition." International Journal of Soft Computing and Engineering (IJSCE), vol. 2, no. 1, Mar. 2012, pp. 235–36.

A.A.Abdelhamid et al., "Robust Speech Emotion Recognition Using CNN+LSTM Based on Stochastic Fractal Search Optimization Algorithm," in IEEE Access, vol. 10, pp. 49265-49284, 2022, doi: 10.1109/ACCESS.2022.3172954.

Aggarwal, A., Srivastava, A., Agarwal, A., Chahal, N., Singh, D., Alnuaim, A. A., Alhadlaq, A., & Lee, H.-N. (2022). Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning. Sensors, 22(6), 2378.

Aouani, Hadhami & Benayed, Yassine. (2020). Speech Emotion Recognition with deep learning. Procedia Computer Science.176. 251-260. 10.1016/j.procs. 2020.08.027.

Bagus Tris Atmaja, Akira Sasou, Masato Akagi, Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion, Speech Communication, Volume 140,2022, Pages 11-28, ISSN 0167-6393,

Guo, Jia. "Deep learning approach to text analysis for human emotion detection from big data" Journal of Intelligent Systems, vol. 31, no. 1, 2022, pp. 113-126.

Houssein, E.H., Hammad, A. & Ali, A.A. Human emotion recognition from EEG-based brain–computer interface using machine learning: a comprehensive review. Neural Comput & Applic 34, 12527–12557 (2022).

Kogila, R., Sadanandam, M. & Bhukya, H. Deep Learning Algorithms for Speech Emotion Recognition with Hybrid Spectral Features. SN COMPUT. SCI. 5, 17 (2024).

Lieskovská, E., Jakubec, M., Jarina, R., & Chmulík, M. (2021). A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism. Electronics, 10(10), 1163.

M. Singh and Y. Fang, "Emotion Recognition in Audio and Video Using Deep Neural Networks," arXiv preprint arXiv:2006.08129, June 2020.

Mustaqeem, M. Sajjad and S. Kwon, "Clustering-Based Speech Emotion Recognition by Incorporating Learned

Features and Deep BiLSTM," in IEEE Access, vol. 8, pp. 79861-79875, 2020, doi: 10.1109/ACCESS.2020.2990405.

Patel, N., Patel, S. & Mankad, S.H. Impact of autoencoder based compact representation on emotion detection from audio. J Ambient Intell Human Comput 13, 867–885

Samaneh Madanian, Talen Chen, Olayinka Adeleye, John Michael Templeton, Christian Poellabauer, Dave Parry, Sandra L. Schneider, Speech emotion recognition using machine learning — A systematic review, Intelligent Systems with Applications,Volume 20,2023,200266, ISSN 2667-3053

Ullah, R.; Asif, M.; Shah, W.A.; Anjam, F.; Ullah, I.; Khurshaid, T.; Wuttisittikulkij, L.; Shah, S.; Ali, S.M.; Alibakhshikenari, M. Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer. Sensors 2023, 23, 6212.