International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

A Review on Human Stress and Anxiety Detection using Speech Signals and Deep Learning Techniques

Swati Kumari

Department of CSE SRU Raipur, CG, India
kumariswati3894@gmail.com

Dr. Ranu Pandey

Department of CSE SRU Raipur, CG, India
ranu_pandey8@hotmail.com

https://doi.org/10.55041/ijsmt.v2i6.097

Cite this Article: Kumari, S. (2026). A Review on Human Stress and Anxiety Detection using Speech Signals and Deep Learning Techniques. International Journal of Science, Strategic Management and Technology, *02*(6). https://doi.org/10.55041/ijsmt.v2i6.097

Abstract—Mental health disorders such as stress and anxiety have become major healthcare concerns worldwide. Early iden-tification of stress-related conditions is essential for preventing severe psychological and physiological complications. In recent years, speech-based emotion recognition systems have gained significant attention due to their non-invasive and real-time mon-itoring capability. This review paper presents a comprehensive survey of machine learning and deep learning techniques used for stress and anxiety detection from speech signals. Various acoustic feature extraction methods including Mel Frequency Cepstral Coefficients (MFCC), spectral contrast, chroma features, and zero-crossing rate are discussed. In addition, different classifi-cation models such as Random Forest, XGBoost, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Residual Networks (ResNet) are reviewed and compared. Publicly available speech emotion datasets and recent advance-ments in deep learning-based emotional speech analysis are also summarized. The study highlights current challenges, research gaps, and future directions for intelligent speech-based mental health monitoring systems.

Index Terms—Stress Detection, Anxiety Detection, Speech Emotion Recognition, Deep Learning, Machine Learning, MFCC, CNN, LSTM, XGBoost

I. Introduction

Stress and anxiety are among the most common mental health disorders affecting millions of people globally. In-creasing work pressure, unhealthy lifestyles, social stress, and emotional instability have significantly contributed to the rise of psychological disorders [22], [32]. Prolonged stress can lead to depression, cardiovascular diseases, cognitive impairment, and sleep disorders. Therefore, early detection and continuous monitoring of mental health conditions have become critical research areas in healthcare and artificial intelligence.

Traditional mental health assessment methods mainly de-pend on clinical interviews and psychological evaluations, which are often subjective and time-consuming [23]. To over-come these limitations, researchers have explored automated systems capable of analyzing physiological and behavioral signals for emotional state detection.

Human speech contains important emotional and psycholog-ical information. Variations in pitch, speech intensity, articula-tion patterns, and spectral characteristics can reveal stress and anxiety conditions [14], [22]. Speech-based analysis provides a non-invasive and cost-effective approach for real-time mental health monitoring.

Recent advancements in machine learning and deep learn-ing have significantly improved speech emotion recognition systems [1], [6]. Deep neural networks can automatically learn complex speech representations and achieve superior classification performance compared to conventional methods. This review paper presents a detailed analysis of existing speech-based stress and anxiety detection techniques, acoustic feature extraction methods, publicly available datasets, ma-chine learning algorithms, deep learning architectures, and current research challenges.

II. Literature Review

Speech-based stress and anxiety detection has gained sig-nificant attention in recent years due to the increasing de-mand for intelligent mental healthcare systems. Researchers have explored multiple machine learning and deep learning approaches for analyzing emotional speech patterns and iden-tifying psychological disorders.

Amiriparian et al. [1] demonstrated that deep neural net-works significantly improve speech emotion recognition accu-racy compared to conventional machine learning approaches. Their work highlighted the effectiveness of deep feature ex-traction using spectrogram-based representations. Similarly, Goodfellow et al. [2] explained the theoretical foundations of deep learning architectures and their applications in speech and emotional signal analysis.

Chen and Guestrin [3] introduced the XGBoost algorithm, which became highly popular for speech emotion classification because of its efficient gradient boosting mechanism and strong performance with high-dimensional acoustic features. Breiman [4] proposed the Random Forest algorithm, which improved classification robustness by combining multiple de-cision trees and reducing overfitting problems.

Jurafsky and Martin [5] discussed several speech and lan-guage processing techniques that contribute to automatic emo-tional state recognition. Schuller et al. [6] investigated deep neural network architectures for emotional speech analysis and

demonstrated superior classification performance using deep learning methods.

Hochreiter and Schmidhuber [7] proposed the Long Short-Term Memory (LSTM) network, which became highly effective for modeling temporal dependencies in speech signals. Krizhevsky et al. [8] introduced Convolutional Neural Networks (CNN), which significantly improved automatic feature extraction capability from speech spectrograms and image-based representations.

Residual learning was introduced by He et al. [9] through the ResNet architecture, enabling the training of very deep neural networks without degradation problems. LeCun et al. [10] further emphasized the importance of deep learning in pattern recognition and intelligent systems.

Boersma [11] developed the Praat software, which be-came widely used for speech processing and acoustic feature extraction. Ellis [12] explored perceptual linear prediction and MFCC-based speech analysis methods. Davis and Mermelstein [13] proposed Mel Frequency Cepstral Coefficients (MFCC), which remain one of the most important acoustic features for speech emotion recognition.

Lugger and Yang [14] investigated voice quality features for speaker-independent emotion recognition and demonstrated that vocal variations effectively represent emotional states. Busso et al. [15] introduced the IEMOCAP dataset, which became a benchmark emotional speech database for speech emotion recognition research.

Livingstone and Russo [16] developed the RAVDESS emotional speech dataset containing audio-visual emotional recordings from multiple speakers. Cao et al. [17] proposed a hybrid CNN-LSTM architecture that combined spatial and temporal feature learning mechanisms for improved emotional speech classification.

Neumann and Vu [18] developed an attentive convolutional neural network for emotional speech analysis and demonstrated that attention mechanisms improve emotional feature learning capability. Eyben et al. [19] introduced the openSMILE toolkit, which provides a standardized framework for extracting large-scale acoustic speech features.

Graves et al. [20] investigated deep recurrent neural net-works for speech recognition and demonstrated the effective-ness of recurrent learning for sequential speech modeling. Fayek et al. [21] evaluated multiple deep learning architec-tures and concluded that deep neural networks outperform conventional machine learning approaches in speech emotion recognition tasks.

Cowie et al. [22] explored emotion recognition in human-computer interaction systems and highlighted the importance of speech signals as reliable emotional indicators. El Ayadi et al. [23] presented a comprehensive survey on speech emotion recognition techniques, datasets, and acoustic feature extraction methods.

Kim [24] demonstrated the effectiveness of convolutional neural networks for text and speech-related classification problems. Sainath et al. [25] applied deep convolutional neural networks for large vocabulary continuous speech recognition and achieved substantial performance improvements.

Deng et al. [26] introduced the ImageNet dataset, which played an important role in the development of transfer learning and deep learning architectures used in speech and image analysis applications. Schmitt et al. [27] proposed auto-matic speech emotion recognition systems using deep learning approaches and achieved high classification performance.

Satt et al. [28] developed an efficient deep learning frame-work for speech emotion recognition with low computational complexity. Tzirakis et al. [29] proposed an end-to-end multimodal emotional recognition framework combining speech and visual information for improved emotional analysis.

Zhang et al. [30] investigated feature fusion techniques using spectral and prosodic features and demonstrated im-proved speech emotion recognition accuracy. Mollahosseini et al. [31] explored deeper neural architectures for facial expres-sion recognition, which contributed to multimodal emotional analysis research.

Cummins et al. [32] reviewed depression and suicide risk assessment techniques using speech analysis and demonstrated the effectiveness of speech biomarkers for mental health monitoring. Han et al. [33] proposed spectral feature learning methods for emotional speech representation and improved emotional classification performance.

Hasan et al. [34] developed a machine learning-based human stress detection framework using voice signals and achieved promising stress classification results. Cho et al. [35] introduced recurrent neural encoder-decoder architectures that later contributed significantly to sequential emotional speech modeling and natural language understanding systems.

Although previous studies achieved promising performance in emotional speech analysis, challenges such as limited emotional datasets, speaker variability, environmental noise, and real-time deployment complexity still remain. Recent advancements in deep learning, transfer learning, transformer architectures, and multimodal emotional analysis continue to improve stress and anxiety detection systems for intelligent healthcare applications.

III. Speech-Based Emotional Analysis

Speech emotion recognition focuses on identifying emo-tional states from speech signals. Emotional speech analysis has become an important component of affective computing and intelligent healthcare systems.

Researchers have demonstrated that emotional states signif-icantly influence speech production mechanisms [23]. Stress and anxiety can alter speech rate, pitch, energy, and vocal tract characteristics. Acoustic analysis of these variations enables automatic emotional classification.

Speech-based systems are preferred because speech signals can be collected easily without specialized medical equipment. Moreover, speech-based analysis supports remote healthcare monitoring and telemedicine applications.

IV. Acoustic Feature Extraction Techniques

Feature extraction is one of the most critical stages in speech emotion recognition systems. Various acoustic features have been proposed for stress and anxiety detection.

A. Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCC) are the most widely used features for speech emotion recognition [13]. MFCC features represent the spectral properties of speech signals based on human auditory perception.
The MFCC computation is expressed as:

Ví.

$$MFCC = \sum_{n=1}^{\tilde{N}} x(n) \cos\left[\frac{\pi k(n-0.5)}{N}\right] \qquad (1)$$

MFCC features effectively capture emotional variations in speech signals and are extensively used in machine learning and deep learning models.

B. Chroma Features

Chroma features represent the energy distribution across different pitch classes and are useful for emotional speech analysis [30].

C. Spectral Contrast

Spectral contrast measures the spectral peak and valley differences in speech signals and helps distinguish emotional speech patterns [33].

D. Zero Crossing Rate

Zero Crossing Rate (ZCR) measures signal sign changes and provides useful information regarding speech activity and emotional intensity.

V. Machine Learning Techniques

Several machine learning algorithms have been applied for stress and anxiety classification.

A. Random Forest

Random Forest is an ensemble learning algorithm proposed by Breiman [4]. It combines multiple decision trees to improve classification robustness and reduce overfitting. Random Forest has shown strong performance in emotional speech classification tasks.

B. XGBoost

XGBoost is a gradient boosting algorithm introduced by Chen and Guestrin [3]. It provides efficient handling of high-dimensional acoustic features and achieves excellent classification accuracy in speech analysis applications.

C. Support Vector Machine

Support Vector Machine (SVM) is widely used in emotional speech analysis because of its ability to handle nonlinear feature spaces effectively.

Deep Learning Approaches

Deep learning has significantly improved speech emotion recognition systems due to its automatic feature learning capability.

A. Convolutional Neural Networks

CNN models automatically learn spatial representations from spectrogram images and acoustic features [8], [25]. CNNs are widely used for emotional speech classification tasks.

B. Long Short-Term Memory Networks

LSTM networks proposed by Hochreiter and Schmidhu-

ber [7] are capable of modeling temporal dependencies in speech signals and are extensively used for speech emotion recognition.

C. Residual Networks

Residual Networks (ResNet) introduced by He et al. [9] en-able deeper neural architectures without degradation problems. ResNet models achieve strong performance in spectrogram-based emotional analysis.

D. Hybrid Deep Learning Models

Hybrid CNN-LSTM models combine spatial and temporal learning mechanisms for improved speech emotion recognition performance [17].

VII. Speech Emotion Datasets

Several publicly available datasets are commonly used for stress and anxiety detection research.

A. RAVDESS Dataset

The RAVDESS dataset developed by Livingstone and Russo [16] contains emotional speech recordings from multi-ple speakers and is widely used in emotional speech analysis.

B. IEMOCAP Dataset

The IEMOCAP dataset proposed by Busso et al. [15] con-tains multimodal emotional interactions and supports speech emotion recognition research.

C. SAVEE Dataset

SAVEE is a speech emotion dataset containing emotional speech recordings from male speakers.

D. CREMA-D Dataset

CREMA-D is a crowd-sourced emotional speech dataset containing diverse emotional expressions from multiple speak-ers.

VIII. Performance Evaluation Metrics

The performance of speech emotion recognition systems is commonly evaluated using the following metrics:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC

Accuracy is computed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

where TP , TN , FP , and FN represent true positive, true negative, false positive, and false negative values respectively.

IX. Challenges in Speech-Based Stress Detection

Despite significant advancements, several challenges remain in speech-based stress and anxiety detection systems:

- Limited emotional speech datasets
- Environmental noise interference
- Speaker variability
- Language dependency
- Real-time deployment complexity
- Generalization issues

Speech signals are highly affected by recording conditions, accent variations, and emotional intensity differences, which can reduce classification accuracy.

X. Future Research Directions

Future research in speech-based stress detection may focus on:

- Transformer-based deep learning architectures
- Multilingual emotional speech analysis
- Real-time healthcare monitoring systems
- Explainable artificial intelligence
- Multimodal emotion recognition
- Wearable healthcare integration

The integration of speech analysis with physiological sig-nals such as ECG, EEG, and facial expressions may further improve stress detection accuracy.

XI. Conclusion

This review paper presented a comprehensive survey of speech-based stress and anxiety detection techniques using machine learning and deep learning approaches. Various acoustic feature extraction methods including MFCC, chroma features, spectral contrast, and zero-crossing rate were dis-cussed. Machine learning algorithms such as Random Forest and XGBoost along with deep learning architectures including CNN, LSTM, and ResNet were analyzed.

The reviewed studies demonstrate that speech signals con-tain valuable emotional information capable of identifying stress and anxiety conditions effectively. Deep learning ap-proaches have shown superior performance due to their auto-matic feature learning capability.

Although current systems achieve promising results, chal-lenges such as dataset limitations, environmental variability, and real-time implementation still exist. Future advancements in transformer-based models, multimodal analysis, and intelli-gent healthcare systems are expected to improve speech-based mental health monitoring significantly.

References

[1] S. Amiriparian et al., "Deep Learning for Speech Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 1–10, 2022.
[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
[3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of KDD*, 2016, pp. 785–794.
[4] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[5] D. Jurafsky and J. Martin, *Speech and Language Processing*. Pearson, 2021.
[6] B. Schuller et al., "Speech Emotion Recognition Using Deep Neural Networks," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 90–102, 2020.
[7] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[8] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *NIPS*, 2012.
[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, 2016.
[10] Y. Lecun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436–444, 2015.
[11] P. Boersma, "Praat: Doing Phonetics by Computer," *Glot International*, vol. 5, no. 9, pp. 341–345, 2001.
[12] D. Ellis, "PLP and RASTA and MFCC," Columbia University, Tech. Rep., 2005.
[13] S. Davis and P. Mermelstein, "Comparison of Parametric Represen-tations for Monosyllabic Word Recognition," *IEEE Transactions on Acoustics*, vol. 28, no. 4, pp. 357–366, 1980.
[14] M. Lugger and B. Yang, "The Relevance of Voice Quality Features in Speaker Independent Emotion Recognition," in *ICASSP*, 2007.
[15] C. Busso et al., "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.
[16] S. Livingstone and F. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," *PLoS ONE*, vol. 13, no. 5, 2018.
[17] J. Cao et al., "Speech Emotion Recognition Based on CNN and LSTM," *IEEE Access*, vol. 7, pp. 1771–1779, 2019.
[18] M. Neumann and N. Vu, "Attentive Convolutional Neural Network Based Speech Emotion Recognition," in *INTERSPEECH*, 2017.
[19] F. Eyben et al., "Recent Developments in openSMILE," in *ACM Multi-media*, 2013.
[20] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in *ICASSP*, 2013.
[21] H. Fayek, M. Lech, and L. Cavedon, "Evaluating Deep Learning Architectures for Speech Emotion Recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.
[22] R. Cowie et al., "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.
[23] M. El Ayadi, M. Kamel, and F. Karray, "Survey on Speech Emotion Recognition," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.
[24] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in *EMNLP*, 2014.
[25] T. Sainath et al., "Deep Convolutional Neural Networks for LVCSR," in *ICASSP*, 2013.
[26] J. Deng et al., "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR*, 2009.
[27] M. Schmitt et al., "Automatic Speech Emotion Recognition Using Deep Learning," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 1–12, 2021.
[28] A. Satt et al., "Efficient Emotion Recognition from Speech Using Deep Learning," in *INTERSPEECH*, 2017.

[29] P. Tzirakis et al., "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.
[30] S. Zhang et al., "Speech Emotion Recognition Using Fusion of Spectral and Prosodic Features," *Sensors*, vol. 19, no. 22, 2019.
[31] A. Mollahosseini, D. Chan, and M. Mahoor, "Going Deeper in Facial Expression Recognition," in *WACV*, 2016.
[32] N. Cummins et al., "A Review of Depression and Suicide Risk As-sessment Using Speech Analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.
[33] J. Han et al., "Learning Spectral Features for Speech Emotion Recog-nition," in *ICASSP*, 2014.
[34] M. R. Hasan et al., "Human Stress Detection from Voice Using Machine Learning," *International Journal of Advanced Computer Science*, vol. 12, no. 4, pp. 55–63, 2021.
[35] K. Cho et al., "Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation," in *EMNLP*, 2014.