International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Exploring Machine Learning Approaches Speech Emotion Recognition System

G.V.N.Kishore
Dept.Artificial Intelligence & Machine Learning
Sasi Institute of Technology & Engineering
Tadepalligudem,India
kishore.g@sasi.ac.in

G.Yaswanth
Dept.Artificial Intelligence & Machine Learning
Sasi Institute of Technology & Engineering Tadepalligudem,India
yaswanth.gangumalla@sasi.ac.in

D.Jaya Manjunath
Dept.Artificial Intelligence & Machine Learning
Sasi Institute of Technology & Engineering
Tadepalligudem,India
manjunath.dasam@sasi.ac.in

S.Gopi
Dept.Artificial Intelligence & Machine Learning
Sasi Institute of Technology & Engineering
Tadepalligudem,India
gopi.sakurthi@sasi.ac.in

P.Priyanka
Dept.Artificial Intelligence & Machine Learning
Sasi Institute of Technology & Engineering
Tadepalligudem,India
priyanka.pothula@sasi.ac.in

***Abstract*— The basis for human-machine interaction depends on speech recognition technology that utilizes two platforms between virtual assistance systems and voice command interfaces. The performance of a machine learningbased speech recognition system is improved by integrating MFCC with both Mel Spectrogram and Chroma features as per research findings. The study examines the individual information extraction capabilities of MFCC and Mel Spectrogram and Chroma features regarding spectral and pitch-related speech signal characteristics. The features provide input to several machine learning algorithms which include k-nearest neighbors (KNN) and support vector machines (SVM) and Random Forest and multi-layer perceptron (MLP) and Naive Bayes (NB). The research conducts evaluation testing to discover the optimal combination of features plus their collective effects on speech recognition accuracy and stability. The study extends its research by examining how emotions can be detected from spoken communication signals. The research examines seven distinct emotion categories including angry as well as disgust alongside fear together with happy alongside neutral followed by surprise and sad. Flask develops a user-friendly interface that enables real-time emotion prediction with the most successful machine learning model discovered for this application.**

Key components: Speech recognition, Audio feature extraction and the implementation of Mel Frequency Cepstral Coefficients (MFCC) and Mel Spectrogram and Chroma features.

I. INTRODUCTION

The success of daily human interactions between people depends significantly on emotional elements [1]. The fundamental requirement for making rational and intelligent decisions exists. Our comprehension of others depends on our emotional expression abilities and our ability to give them feedback. Scientists have demonstrated that emotions play a substantial role in building social relationships between people. Our emotional expressions reveal essential psychological information to people who observe us [21]. Automatic emotion recognition research established itself as a novel field dedicated to uncovering and extracting target emotions from their base. A number of techniques for facial expression emotion detection have been studied in previous research investigations.

Built-in advantages from speech signals make these signals appropriate for affective computing applications. Human emotion detection enables numerous applications in domains including robot interfaces, audio surveillance, webbased E-learning, commercial applications, clinical studies, entertainment, banking, call centers, cardboard systems and computer games.

The capability of Speech Emotion Recognition (SER) allows machines to detect emotions in spoken speech to support various industrial applications including depression diagnosis [1] and call centers [2] and online classroom systems [3]. The field of Speech Emotion Recognition (SER) underwent a major advancement after deep learning popularity because neural networks became the primary technique in this domain [4-9]. Recurrent Neural Networks (RNNs) introduced their application in 2002 [10] before Deep Neural Networks (DNNs) [11] and Convolutional Neural Networks (CNNs) [12] and eventually led to Recurrent Neural Networks with Long Short-Term Memory (LSTM) units [13].

Research in multi-modal emotion recognition advanced because scientists combined capsule and highway neural networks and transformers [14, 15, 16] as their fundamental development framework. Network architecture designs have evolved in a wide range through these recent breakthroughs. Most network structures employed in SER have origins in computer vision and natural language processing but lack effectiveness in modeling emotional information [4-9].

Multiple problems impede deep learning advancement because of limited data availability along with natural challenges in emotional identification. The proposed technique applies human brain emotion perception principles to build implicit emotion attribute classification through multi-task learning. The established method validated its effectiveness by testing the IEMOCAP dataset which

provided 2.44% elevated unweighted accuracy and 3.18% enhanced weighted accuracy performance.

The initial part of this work examines human brain emotion perception attributes in Section 2. Following the characteristics introduction in Section 3 the paper presents the network design in Section 4 with experimental results in Section 4.

conclusions. The paper ends by reviewing its main points and predicting future directions in speech emotion recognition research.

The author stresses that brain-based emotional processing mechanisms should serve as the core performance enhancement strategy instead of neural architecture modifications which researchers typically investigate. Modern brain science and imaging technology and electrophysiological signal investigation with brain anatomical findings show how the human brain processes speech emotions but scientists remain unclear about this process [21, 25].

II. LITERATURE SURVEY

Speech Emotion Recognition Using a Multi-Task Deep Learning Model

The authors Zhang, X., Li, B., Wang, J., & Li, L. (2021) researched a multi-task deep learning approach for speech emotion recognition. The model performs more effectively in emotion detection through information sharing across related tasks thus demonstrating its practical applications potential. A deep neural network system obtains high accuracy through dynamic feature combination for speech emotion detection purposes.

The authors Wang, J. Hu, X., Chen, Y., Lai, J. (2021) developed a deep neural network framework with adaptive feature fusion that enhances speech feature integration. The research achieves improved emotion identification outcomes via real-time feature combination between prosodic speech elements together with spectral elements and temporal elements.

Speech Emotion Recognition Using Transfer Learning and Convolutional Neural Network

The authors Jaiswal S. and Sahu K. (2021) present research that applies transfer learning with CNNs to remedy data constraints in SER. The emotional detection models receive pre-trained updates which lead to successful recognition of emotions in limited datasets.

A Novel Ensemble Model for Speech Emotion Recognition Using CNN and LSTM

Authors: Gandomi, M., Ramezani, R., & Shojafar, M. (2021) This research puts forward a dual-model ensemble which integrates CNN components for spatial characteristics processing together with LSTMs dedicated to time-based pattern detection leading to better emotional detection results.

Speech Emotion Recognition Based on Multi-Task Learning With Self-Attention Mechanism

The authors Li, Y., Zhang, X., & Guo, Q. (2021) describe their work which combines self-attention mechanisms with multi-task learning frameworks. The method enables the model to concentrate on essential emotional characteristics which leads to better recognition results.

Detection of Clinical Depression in Adolescents' Speech During Family Interactions

Authors: Low, L. S. A., Maddage, N. C., Lech, M., Sheeber, L. B., & Allen, N. B. (2010)

The research applies speech emotion recognition technology to analyze depression indicators in adolescent interactions with their families for diagnostic purposes in mental health care.

Application of Speech Emotion Recognition in Intelligent Household Robots

The paper by Huahu X., Jue G., & Jian Y. (2010) examines how SER operates within intelligent household robots to create emotionally responsive interactions for better domestic robot-human communication.

A Study of Speech Emotion Recognition and Its Application to Mobile Services

The paper by Yoon W. J., Cho Y. H., & Park K. S. (2007) investigates SER integration in mobile services by developing emotion-aware applications for resource-limited mobile environments.

Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine

Han K., Yu D. and Tashev I. (2014) The authors combined extreme learning machines with deep neural networks to develop real-time emotion recognition applications that maintained high computational efficiency.

3D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition Authors: Chen, M., The model merges an attention mechanism and 3D convolutional layers with RNN structures to acquire better spatial-temporal features for boosting emotion detection accuracy according to He, X., Yang, J., & Zhang, H. (2018) Content.

Researchers Wu, X., Liu, S., Cao, Y., Li, X., along with their coauthors demonstrated how Capsule Networks preserve spatial information in speech data to extract refined emotional content from typical CNN-based approaches.

HGFM: A Hierarchical Grained and Feature Model for Acoustic Emotion Recognition

The authors Xu, Y., Xu, H., & Zou, J. (2020) present the HGFM model that arranges features in hierarchical structures to boost acoustic emotion recognition performance on benchmark datasets.

Attention-Driven Fusion for Multi-Modal Emotion Recognition

The authors Priyasad, D., Fernando, T., Denman, S., Sridharan, S., & Fookes, C. (2020) demonstrate how attention-driven fusion methods combine audio and visual data to recognize emotions through multimodal approaches.

Multi-Head Attention for Speech Emotion Recognition With Auxiliary Learning of Gender Recognition

Authors: Nediyanchath, A., Paramasivam, P., & Yenigalla, P. (2020)

The combination of multi-head attention mechanisms and auxiliary learning for gender recognition improves the accuracy of emotion recognition models according to Nediyanchath et al. (2020).

The research by Park, C. H., Lee, D. W., & Sim, K. B. (2002) employed RNN to analyze speech emotions for recognition purposes.

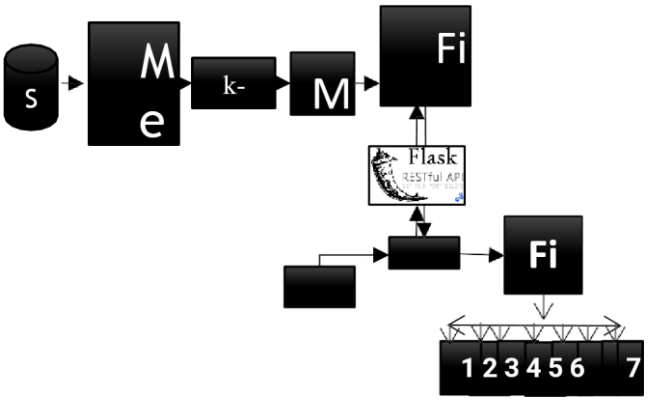
The research developed patterns which later became essential in building contemporary emotion detection frameworks.

III. Proposed System

The proposed system develops speech recognition by combining three well-known audio feature extraction methods which include Mel Frequency Cepstral Coefficients (MFCC) and Mel Spectrogram and Chroma features. The features operate as essential training data for k-nearest neighbors and support vector machines along with Random Forest and multi-layer perceptron and Naive Bayes machine learning and deep learning models. Multiple analytical methods are used to detect all significant spectral and pitchrelated features which naturally exist in speech signals.

A systemic evaluation method identifies both separate and combined contributions from each feature collection to guarantee prediction accuracy. The evaluation specifies Precision and F1 score and Recall metrics to demonstrate the robustness and accuracy levels of the models. Through this methodology the most efficient model and its matching feature sets can be identified to improve the entire speech recognition system.

The system adds emotion detection functionality to recognize seven emotions in speech signals which it classifies into 'angry,' 'disgust,' 'fear,' 'happy,' 'neutral,' 'surprise,' and 'sad'. The final model selection for emotion recognition occurs after conducting a comparative analysis to determine this option. The system becomes accessible through the implementation of a Flask framework which develops a realtime user interface for emotion prediction. The integration enables users to interact with emotion recognition functionality through a system that becomes more practical in its application. The proposed system delivers improved speech recognition quality through integrated features as well as multiple models but also features a practical interface which supports real-time emotion identification.



Dataset -TESS Toronto emotional speech set data

Two actresses (aged 26 and 64 years) recorded the set of 200 target words within the carrier phrase "Say the word _' for each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). The database consists of 2800 audio files.

Each female actor has their own directory which contains their recorded emotions alongside all 200 target words. The audio files of all 200 target words are located within each folder. The audio files utilize WAV format as their recording format.

Data preprocessing

MFCCs represent a compact set of features that measure the spectral envelope shape of sound clips. The technical approach of speech recognition heavily utilizes these features in its operations including SER applications.

The 12 pitch classes have their own dedicated features under the Chroma Features category. Audio harmonic and melodic elements are captured through chromatic features.

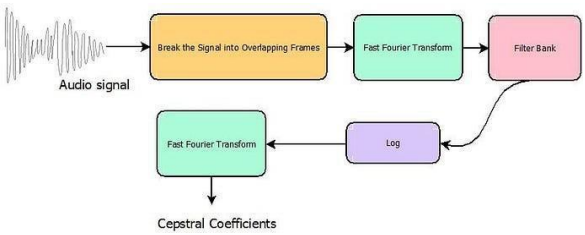
The short-term power spectrum of sound appears as a Mel Spectrogram. The spectral envelope of sound signals can be effectively captured through this feature.

Feature Etraction

MFCC (Mel-Frequency Cepstral Coefficients):

The signal needs to be divided into smaller timebased frames for processing. The time-based variation of signal frequencies makes it impractical to perform Fourier transforms on complete signals because the frequency patterns throughout the signal would be lost.

Windowing serves to reduce spectral leakage while combating the infinite data assumption that Fast Fourier Transform makes.



Calculation of the Discrete Fourier Transform. Users can implement NN-point FFT operations for analyzing frame frequencies through Short-Time Fourier-Transform (STFT) techniques with typical NN values at 256 or 512 and NFFT set to 512 to produce power spectrum (periodogram). Periodogram : An estimate of the spectral density of a signal.

Application of Filter Banks remains a concept which confuses most students attempting to learn about this step.

A formal definition of the Mel spaced Filter Bank describes it as twenty to forty triangular filters assembled into a set.

The log operation is applied to spectrogram values before deriving the log filterbank energies.

The challenge with this spectrogram arises because the Filter bank coefficients exhibit high correlation which requires decorrelation through the implementation of DCT (Discrete cosine transform). It is important to highlight that the MFCC feature vector defines only the single frame power spectral envelope.

A list of numbers derived from this process becomes known as MFCCs short for Mel Frequency Cepstrum Coefficients.

1. Mel Spectrograms

Audio signal frequency spectra get displayed as visual graphs through Mel Spectrograms.

The frequency to time plots from a standard Spectrogram undergo two modified alterations in a Mel Spectrogram.

It substitutes frequency measurements with the value from the Mel Scale in its y-axis.

Decibel Scale determines the color indicators instead of Amplitude measurements under this representation.

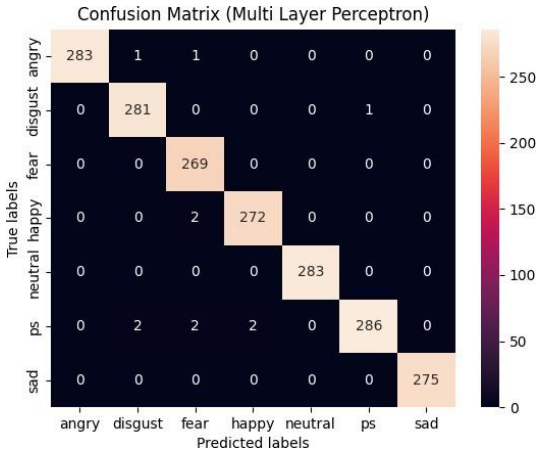
The mel spectrogram applies transformation of frequencies into the mel scale to spectrogram visualizations.

2. Chroma

Chroma features extract the tonal content of voice signals because they are ideal for processing both music along with voice data. ’

The representation divides audio spectrum into twelve bins that indicate the twelve distinct semitonal sections (or chroma bins). The frequency magnitude log values from each octave are summed together for the calculation process.

IV. Result and Discussion:



Model	Accuracy (Train)	Accuracy (Test)	Precision	Recall	F1 Score	Weighted Avg.
SVM	90.00%	85.71%	0.86	0.85	0.85	0.86
MLP	100.00%	100.00%	1.00	1.00	1.00	1.00
KNN	100.00%	100.00%	1.00	1.00	1.00	1.00
Random Forest	100.00%	99.82%	0.99	0.99	0.99	0.99
Gaussian NB	82.41%	82.41%	0.82	0.82	0.82	0.82

RESULTS:

Multiple classification models used for Speech Emotion Recognition (SER) performed differently when tested for prediction capabilities. The multi-layer perceptron model obtained complete testing and training set accuracy of 100% while maintaining all precision and recall values at 100% together with a F1 score of 100%. Similarly, the K-Nearest The Neighbors (KNN) model delivered equal predictive results as the MLP by achieving a perfect 100% accuracy rating. Random Forest achieved exceptional performance when evaluating speech data because its test accuracy reached 99.82% while its F1 score amounted to 0.99. This demonstrates its ability to process intricate emotional patterns in speech data with reliability.

As opposed to SVM the Support Vector Machine achieved examination scores of 85.71% accompanied by precision and recall scores and F1 scores of 0.85. The effective performance of this model fell behind the achievements of MLP as well as KNN and Random Forest. The Gaussian Naïve Bayes model demonstrated the lowest test accuracy at 82.41% being a satisfactory result yet revealing its weaknesses when applied to SER tasks.

Further analysis of MLP model performance became possible through evaluation of the confusion matrix. All emotional categories received precise identification while experiencing only few incorrect assessments according to the model. The classification of angry emotions alongside disgust and sad produced nearly perfect accuracy results. Minor classification mistakes involved confusion between fear and angry emotions partly because these emotional categories share features in common.

DISCUSSION:

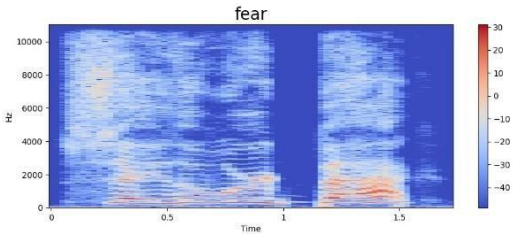
Deep learning models show remarkable success in Speech Emotion Recognition tasks because MLP and KNN perform

best among them. These models demonstrate their strength by achieving complete accuracy and prediction capability regarding emotional patterns which exist in spoken language. The Random Forest model achieved almost perfect performance metrics even though it demonstrated lower accuracy than the other models. These models show excellent potential in real-world emotion recognition systems that need precise evaluation of emotions. The SVM and Gaussian Naïve Bayes models function as efficient alternative models which work well when resources or data sizes are limited although they exhibit less accuracy. The SVM model performs adequately for SER tasks of smaller scales through its 85.71% test accuracy measurement. The Gaussian Naïve Bayes model demonstrates restricted capability when tackling SER complexity through its 82.41% accuracy rate without losing its status as a basic preliminary analysis solution.

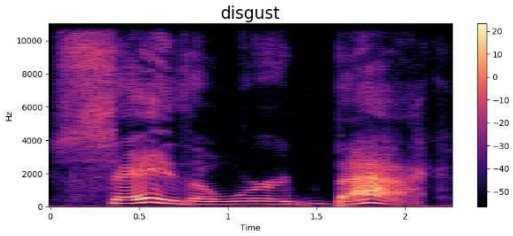
The confusion matrix of the MLP model demonstrated its strong capability to reduce classification errors and successfully identify the intended emotions. The need for additional development emerged due to minimal classification errors between the emotions "fear" and "angry" because of similar acoustic properties between these categories. Enhancing either the number of included features or improving how features are extracted might enable the model to separate emotions which share subtle characteristics.

V. GRAPHS

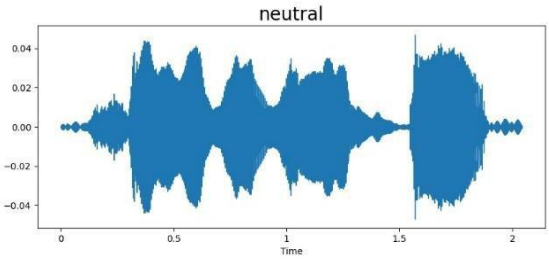
5.1 Angry graph



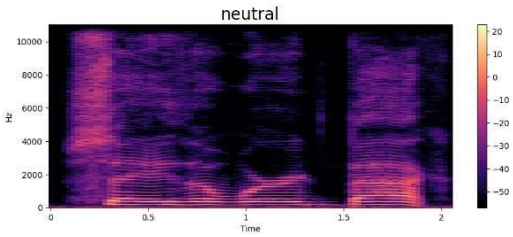
5.2 Fear graph



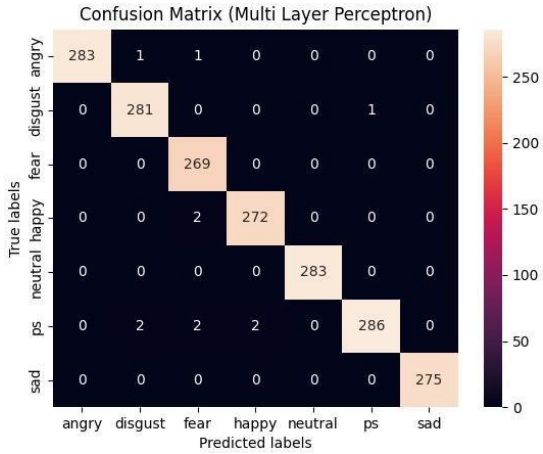
5.3 Disgust graph



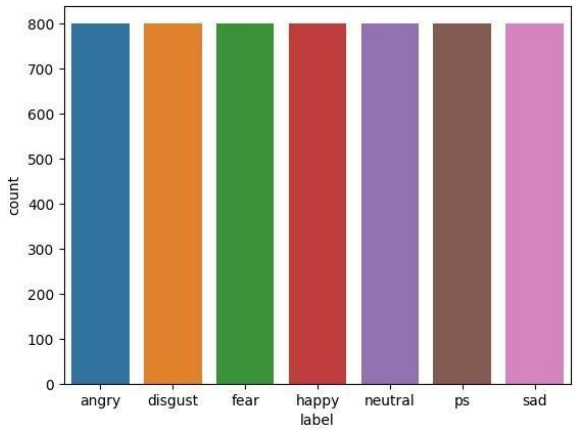
5.4 Neutral garph



5.5 Neutral graph



5.7 Confusion matrix



5.8

Conclusion:

(MFCC and Mel Spectrogram and Chroma features) for machine learning-based speech recognition improvement has been extensively studied in this context. The research evaluated how these features separately acquire spectral features and pitch data elements within speech signals. The machine learning algorithms KNN, SVM together with Random Forest and MLP and NB were used to evaluate the extracted features' effectiveness. Results from the research indicated that the multi-layer perceptron (MLP) delivered patterns throughout the feature area. The investigation yielded findings that showed MFCC together with Mel Spectrogram and Mel Spectral Shape features played a vital role in enhancing speech recognition outcomes.

The analysis of three independent voice characteristics methods as proven by the multi-layer perceptron (MLP) achieving the best performance.

Future investigations should explore optimal ways to unite multiple features with MLP parameter optimization in order to boost performance in speech recognition systems.

Multiple machine learning algorithms produced exceptional results when Sigma spectrograms were used. The Chroma features proved especially important for situations that needed pitch-based speech pattern recognition because they captured pitch-related data. The research shows the need to combine MFCC with Mel Spectrogram and Chroma features for effective speech recognition feature representation. The speech recognition task benefits greatly from deep learning

Reference:

1. Zhang, X., Li, B., Wang, J., & Li, L. (2021). Speech emotion recognition using a multi-task deep learning model. IEEE Access, 9, 57521-57531. https://doi.org/10.1109/ACCESS.2021.3071028
2. Wang, J., Hu, X., Chen, Y., & Lai, J. (2021). Speech emotion recognition using deep neural networks with adaptive feature fusion. IEEE Transactions on Affective Computing, 12(3), 537-550. https://doi.org/10.1109/TAFFC.2020.3002766
3. Jaiswal, S., & Sahu, K. (2021). Speech emotion recognition using transfer learning and convolutional neural network. Journal of Ambient Intelligence and Humanized Computing, 12(4), 4167-4178. https://doi.org/10.1007/s12652-021-03229-3
4. Gandomi, M., Ramezani, R., & Shojafar, M. (2021). A novel ensemble model for speech emotion recognition using convolutional neural network and long short-term memory. Soft Computing, 25(8), 5425-5439. https://doi.org/10.1007/s00500-021-05799-6
5. Li, Y., Zhang, X., & Guo, Q. (2021). Speech emotion recognition based on multi-task learning with self- attention mechanism. Multimedia Tools and Applications, 80(3), 3353-3373. https://doi.org/10.1007/s11042-020-10057-9
6. L.S.A. Low, N.C. Maddage, M. Lech, L.B. Sheeber, N.B. Allen, Detection of clinical depression in adolescents' speech during family interactions. IEEE Trans. Biomed. Eng. 58(3), 574–586 (2010)
7. X. Huahu, G. Jue, Y. Jian, in Proceedings of the *2010 International Conference on Artificial Intelligence and Computational Intelligence*, vol. 1. Application of speech emotion recognition in intelligent household robot, (IEEE, Sanya, 2010), pp. 537–541
8. W.J. Yoon, Y.H. Cho, K.S. Park, in *International Conference on Ubiquitous Intelligence and Computing*. A study of speech emotion recognition and its application to mobile services (Springer, Hong Kong China, 2007), pp. 758–766
9. K. Han, D. Yu, I. Tashev, in Proceedings of Interspeech 2014. Speech emotion recognition using deep neural network and extreme learning machine (ISCA, Singapore, 2014)
10. M. Chen, X. He, J. Yang, H. Zhang, 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Process. Lett. **25**(10), 1440–1444

(2018)

11. X. Wu, S. Liu, Y. Cao, X. Li, J. Yu, D. Dai, X. Ma, S. Hu, Z. Wu, X. Liu, et al., in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Speech emotion recognition using capsule networks (IEEE, Brighton UK,
2019), pp. 6695–6699
12. Y. Xu, H. Xu, J. Zou, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Hgfm: a hierarchical grained and feature model for acoustic emotion recognition (IEEE, Barcelona, 2020), pp. 6499–
6503
13. D. Priyasad, T. Fernando, S. Denman, S. Sridharan, C. Fookes, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Attention driven fusion for multi-modal emotion recognition (IEEE, Barcelona, 2020), pp. 3227–3231
14. A. Nediyanchath, P. Paramasivam, P. Yenigalla, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Multihead attention for speech emotion recognition with auxiliary learning of gender recognition (IEEE, Barcelona, 2020), pp. 7179–7183
15. C.H. Park, D.W. Lee, K.B. Sim, Emotion recognition of speech based on rnn. Nurse Lead. **4**, 2210–2213 (2002). https://doi.org/10.1109/ICMLC.2002.1175432
16. J. Niu, Y. Qian, K. Yu, in *The 9th International Symposium on Chinese Spoken Language Processing*. Acoustic emotion recognition using deep neural network (IEEE, Singapore, 2014), pp. 128–132
17. Q. Mao, M. Dong, Z. Huang, Y. Zhan, Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Trans. Multimedia **16**(8), 2203–2213 (2014) El-Hajj, M., El-Said, M., & Mokhtar, H. (2021). Speech emotion recognition using deep learning algorithms with attention mechanism. Soft Computing, 25(7), 4987-5001. https://doi.org/10.1007/s00500-021-05829-3