Speeach Emotion Recognition

Abstract

Reviews speech emotion recognition based on classifiers for different emotions. Discusses features like energy, pitch, LPCC, MFCC for emotion recognition. Examines classification performance and limitations of speech emotion recognition system. Reviews speech emotion recognition literature, databases, features, and models. Discusses issues in emotional speech databases and classification models. Highlights research gaps and issues in speech emotion recognition Analyzed speech emotion recognition methods based on feature sets and classification. Evaluated performance, limitations, and future directions for emotion recognition systems. Reviews speech emotion recognition using deep learning and attention mechanisms. Examines the impact of attention mechanisms on SER performance. Compares system accuracies on the IEMOCAP benchmark database. Speech emotion recognition using spectrogram features and RNN classification. Investigates transfer learning from valence and activation to categorical emotions. Achieves performance comparable to state-of-the-art systems on USC-IEMOCAP dataset speech Emotion Recognition using spectral and prosodic features. Features include MFCC, pitch, loudness, and speech intensity. SVM for gender classification, Radial Basis Function for emotion recognition. Comparative study on speech emotion recognition systems using machine learning. Features extraction, classifiers, and emotion classification using different methods. Recurrent neural network, multivariate linear regression, and SVM techniques compared. Berlin and Spanish databases used for experimental data set analysis motion recognition using pitch, energy, formant, and MFCC features. SVM and QDA classifiers for emotion recognition performance evaluation. Achieved 96.3% accuracy for stressed/neutral style classification using Gaussian SVM. Identified pitch and energy as crucial factors in emotion recognition.

Key words: Speech emotion recognition, Classifiers, Deep learning, Spectrogram features,

RNN classification, Transfer learning, USC-IEMOCAP dataset, Machine learning, Recurrent neural network, Multivariate linear regression, Motion recognition, Stressed/neutral style classification, Gaussian SVM

1.Introduction

Emotions manifest themselves through language, behavior, and physiological mechanisms. Speech emotion recognition constitutes a multidisciplinary domain that encompasses psychology and linguistics.**[1]** Interpreting human emotions from speech signals poses a challenge for machines Speech plays a pivotal role in human societies by transmitting emotions and information. Systems for Speech Emotion Recognition (SER) are designed to detect emotions in speech.**[2]** These systems for SER analyze features extracted from speech signals to categorize emotions The inclusion of paralinguistic cues such as emotion, gender, and personality is crucial in speech analysis. Efficient human-computer interaction hinges on the comprehension of emotions conveyed through speech signals.**[3]** Utilizing speech signals as a means of communication is both rapid and innate. The integration of emotion recognition systems into human-machine communication enhances its efficacy. Various classifiers like KNN, HMM, and SVM are employed in this context. **[4]** Speech emotion recognition contributes to psychiatric assessments and lie detection.**[5]** The paper centers its attention on Speech Emotion Recognition (SER) within the realm of human-computer interaction. It delves into the discourse surrounding features, classifiers, databases, and applications pertinent to emotional speech recognition. Furthermore, it scrutinizes performance metrics such as precision, recall, F-measure, and recognition rate. **[6]** The paper conducts a comprehensive review of the literature on speech emotion recognition, encompassing databases, features, and models. It also explores emotion recognition utilizing the RAMSES speech system integrated with HMM technology.**[7]** The focal point lies in speakerdependent emotion recognition utilizing a Spanish corpus. Moreover, it emphasizes speech emotion recognition methods along with their corresponding evaluation parameters. The paper undertakes an analysis of feature sets, classification methods, and dataset preparation in the realm of research. Additionally, it evaluates the performance, limitations, and potential future directions for speech

emotion recognition.**[8]** It also reviews the advancements in speech emotion recognition (SER) and the impact of attention mechanisms. The research zeroes in on deep neural networks (DNNs) with attention mechanisms as applied to SER. System accuracies are juxtaposed using the IEMOCAP benchmark database. Emotion recognition through speech is explored utilizing spectrogram features and glottal flow signals. The study delves into representation learning and transfer learning for emotion classification. It also concentrates on enhancing RNN training for emotion recognition tasks**.[9]** The focus is on emotion recognition in speech through the utilization of pitch and formants. The paper compares the characteristics of male and female speech for emotion identification**.[10]** It delves into a comparison of speech emotion recognition systems employing diverse classifiers and features. The emphasis remains on emotion recognition from speech signals through the utilization of machine learning algorithms. Emotion recognition through speech signals is discussed in the context of human-computer interaction. The paper engages in a discussion regarding Convolutional Neural Networks (CNN) for emotion classification, which includes convolution layers, pooling layers, and fully connected layers. A deep CNN architecture is employed for emotion classification using spectrograms derived from speech signals. The research further focuses on speech emotion recognition utilizing a variety of classifiers for differentiation. Challenges in emotion recognition within speech are examined with regard to variability and cultural influences. The application areas encompass psychiatric diagnosis, lie detection, and call centers. Various classifiers such as KNN, HMM, SVM, ANN, and GMM are deliberated upon.

2.Related Work

Research is centered on databases, characteristics, and classifiers for the identification of emotions in speech. Systems for SER incorporate classifiers such as Convolutional Neural Networks (CNN), k-Nearest Neighbor (kNN), Decision Trees, and Support Vector Machines (SVM).**[11]** K-Nearest Neighbor (KNN) is a straightforward supervised algorithm. Previous research has utilized classifiers like Hidden Markov Models (HMM), SVM, and Artificial Neural Networks (ANN) for recognizing emotions**.[12]** The application of speech emotion recognition systems has been witnessed across diverse domains. The differentiation of emotions in speech poses a challenge due

to a variety of factors. **[13]** Emotions can be dissected into primary emotions to facilitate classification. The research is concentrated on emotion recognition through the utilization of hidden Markov models. Features encompass pitch, energy, spectral shape, as well as duration measurements. The database employed for experimentation is the INTER-FACE emotional speech database. The objective is to establish novel models for audio-video analysis. Reviews on the analysis of emotional speech lack comprehensive and up-to-date information**.[14]** Emotional speech databases play a crucial role in psychological investigations and recognition tasks. The framework for SER encompasses steps like feature selection and classifier evaluation. A review of literature on speech emotion recognition entails databases, features, and models. The importance of the excitation source and scope for further exploration in emotion recognition are discussed. There is a dearth of exhaustive review articles on speech emotion recognition post-2006. A review of speech emotion recognition systems, classifiers, features, and datasets is conducted. **[15]** The proposed techniques involve hierarchical structures, binary decision trees, and segment-based methodologies. The review encompasses SER systems, deep features, transfer learning, and techniques for generalization. The focus is on Deep Neural Network (DNN) models with an attention mechanism. Exclusion of papers on Electroencephalography (EEG), heart rate variability, and multimodal fusion is noted. **[16]** The primary consideration is on works related to speech for emotion recognition. Previous studies involved the use of deep learning for emotion recognition from speech. Autoencoders were employed to comprehend acoustic events from speech datasets. Recurrent neural networks were investigated for speech emotion recognition The exploration of multimodal deep learning for audiovisual emotion recognition was undertaken. Transfer learning from affective attributes to categorical emotions was examined. Existing work emphasized emotion recognition using spectral and prosodic features. Emotion recognition was based on the Mel Frequency Cepstral Coefficients (MFCC) approach utilizing the Radial Basis Function Network. An experimental study was carried out to ascertain emotion recognition in speech. SVM is extensively utilized in studies on audio emotion recognition. Physiological signals such as EEG are employed to objectively assess emotions. An experimental evaluation was performed on Berlin and Spanish databases using Multiple Linear Regression (MLR), SVM, and Recurrent Neural Networks (RNN). Emotion recognition was based on pitch, energy, and formant frequencies.

Feature extraction, selection, and classification were executed employing various classifiers. Pair-wise classification outcomes were obtained using Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Gaussian SVM (GSVM), and Linear SVM (LSVM).

Emotion Recognition Using CNN Deep Learning Model:

Data Preparation: Loading audio data and corresponding emotion labels from a specified directory. Preprocessing audio data, including adding noise, shifting, pitching, and stretching. Extracting features from audio data using Mel-frequency cepstral coefficients (MFCCs).

Data Augmentation: Augmenting the audio data three times for each file using random processing techniques.

Model Building: Implementing a convolutional neural network (CNN) with LSTM layers for speech emotion recognition. Compiling the model with appropriate optimizer and loss function.

Model Training: Training the model on the augmented and preprocessed data.

Evaluation and Visualization: Evaluating the model's performance using training and testing loss/accuracy metrics Visualizing training and testing loss/accuracy over epochs.

Model Saving: Saving the trained model for future use.

Confusion Matrix: Generating and visualizing the confusion matrix to analyze the model's performance in classifying different emotions.

the process of building, training, and evaluating a deep learning model for speech emotion recognition, along with necessary data preprocessing and analysis steps.

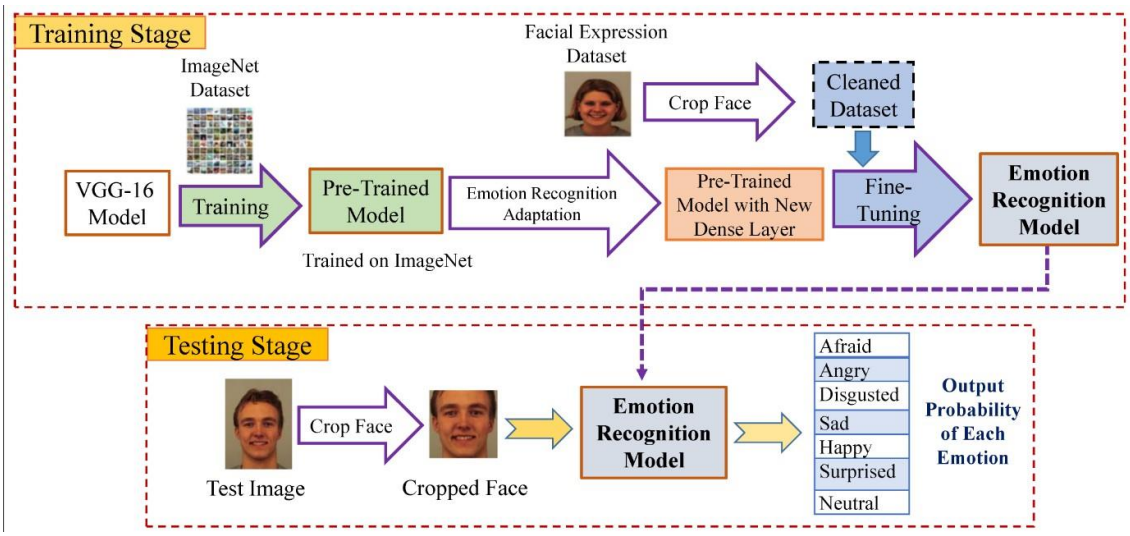


Fig. 1. Emotion Recognition Using CNN Deep Learning Model

3. Model Evaluation and Experimental Results

The number of samples in each category is detailed and summarized in **Table 1**, and **Figure 2** visualizes the sample distributions in training, validation, and testing in each class of the used dataset. The experimental and implementation results can be summarized as follows: **Figure 3** depicts the confusion matrix which presents true and false positives and true and false negatives values of the classification

TABLE 1. voice Dataset

Data Categories	Training Samples	Validation Samples	Testing Samples
neutral	451	60	63
clam	104	102	104
happy	104	18	20
sad	850	80	85
angry	283	99	100
fearful	105	25	25
disgust	719	159	159
surprised	862	183	184

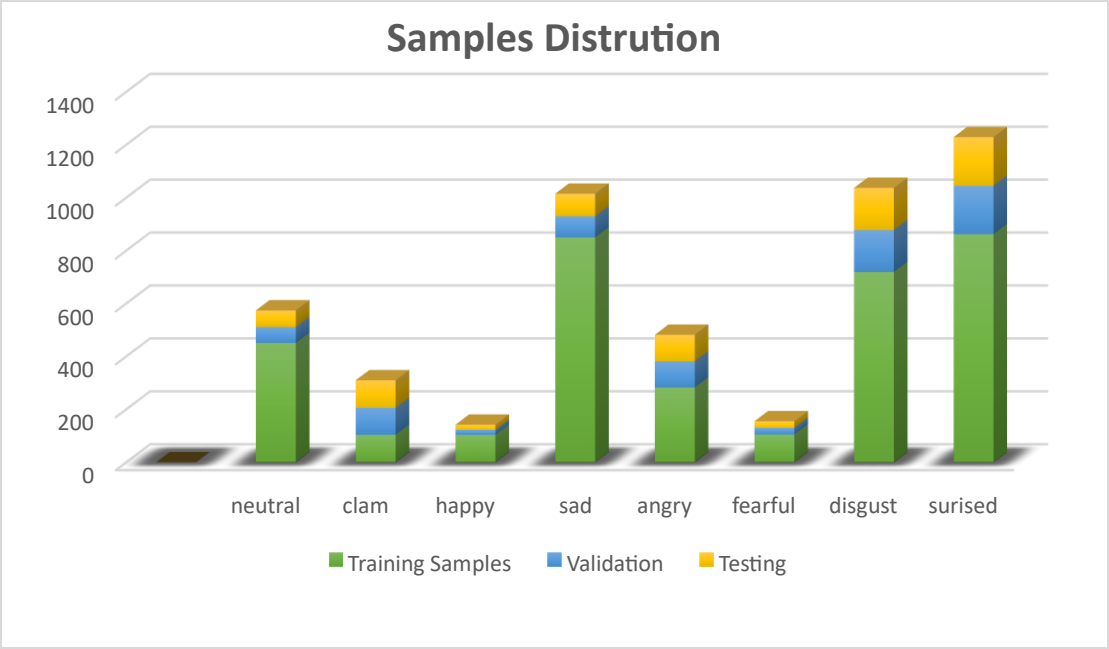


Fig. 2. Samples distributions in training, validation, and testing in the used dataset

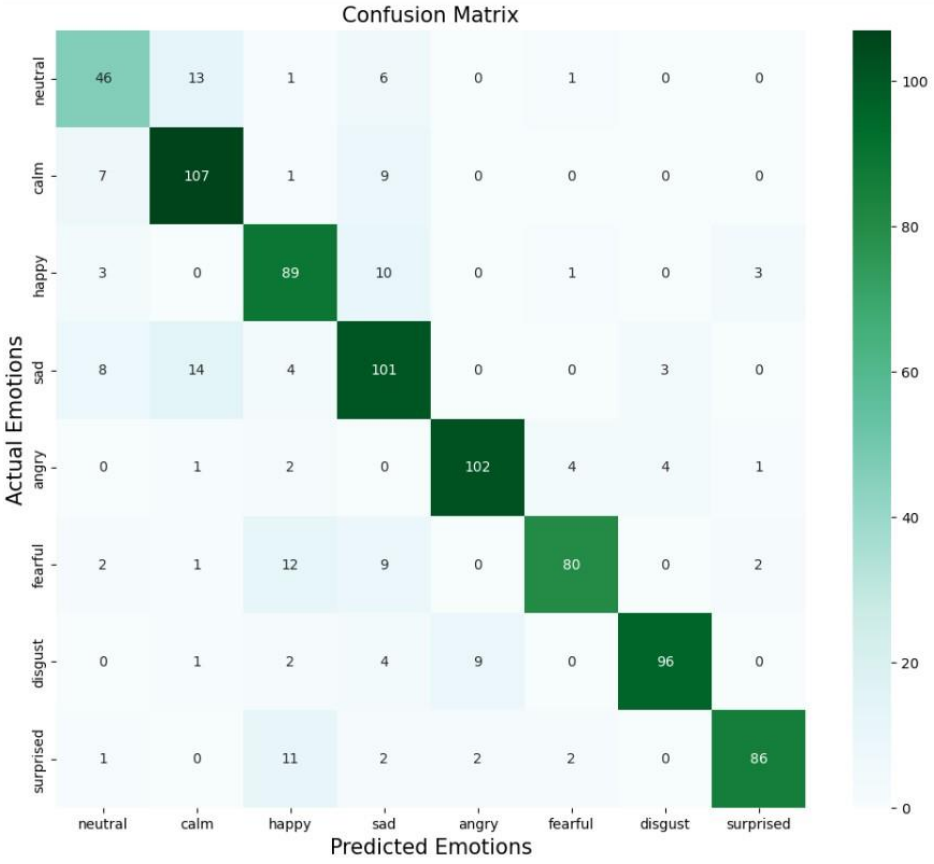


Fig.3. Confusion Matrix

Figure 4 depicts the training and validation loss and accuracy The results reasoned the efficiency of the proposed model, where it achieved training and validation accuracy of 92% and 94 % respectively. In addition, it achieved training and validation loss of 0.092 and 0.9 respectively.

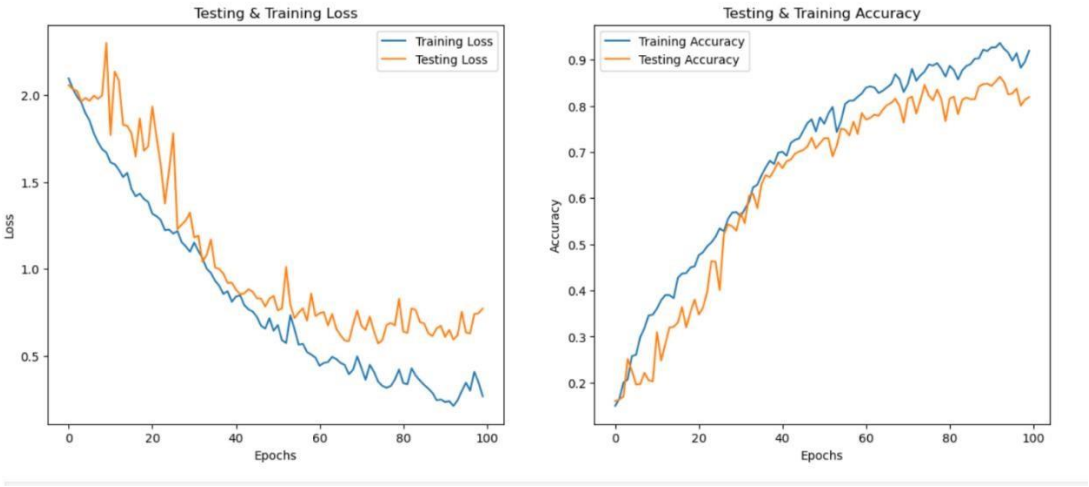


Fig.4. Classification performance: (a) Training and validation loss. (b) Training and validation Accuracy

4. Discussion

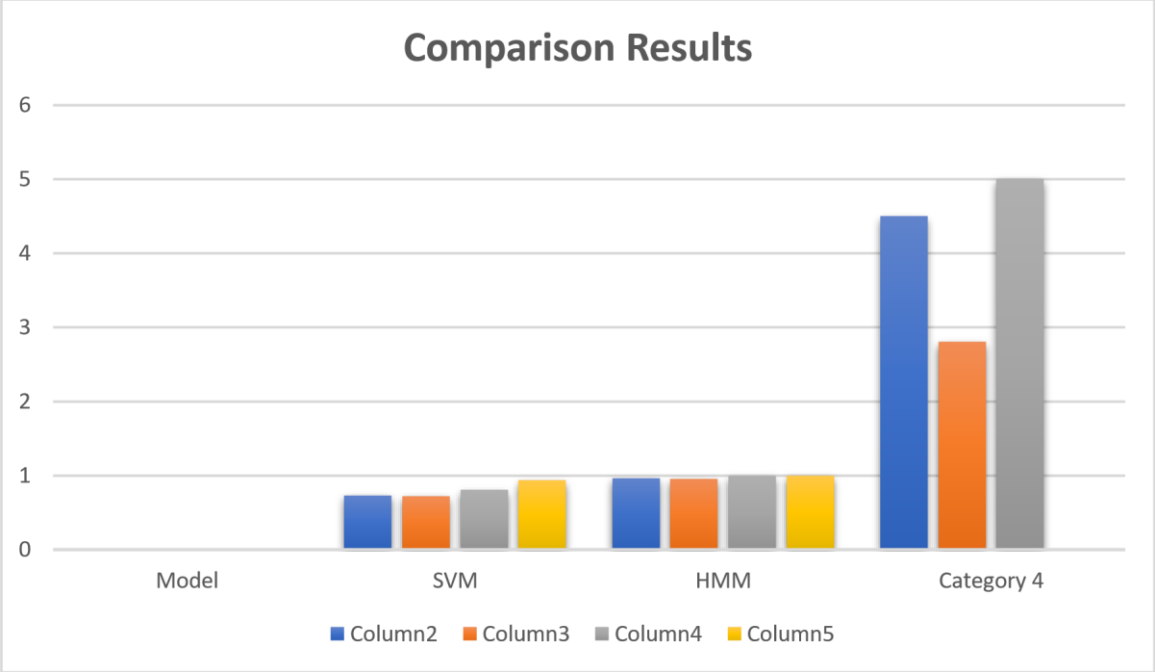
The discussion pertains to various databases and features employed in the identification of emotions in speech. It also mentions applications such as man-machine interactions and systems for classifying emotions. **[7-9]** Emphasis is placed on the significance of prosodic features within emotion recognition systems. Research in Speech Emotion Recognition (SER) is centered on developing robust methods for recognizing emotions through speech databases. Feature extraction encompasses prosodic and spectral acoustic features like MFCC. Both traditional and deep learning classifiers are utilized for recognizing emotions. Challenges in SER include the enhancement of accuracy rates and classification techniques. Speech emotion recognition systems heavily rely on signal processing and classifiers. [**10-12]** The average accuracy is comparatively lower for speaker-independent systems than for speaker-dependent systems. An HMM approach for emotion recognition has achieved an accuracy of over 80%. The results indicate potential for recognizing emotions across multiple speakers. Future endeavors involve exploring languageindependent conditions and conducting research on feature extraction. The discourse thoroughly examines emotional speech recognition features, databases, and applications.

Additionally, it underscores the importance of speech emotion recognition in interactions between humans and computers. Research gaps in speech emotion recognition are brought to light in the paper. The significance of emotional speech corpora and classification models is deliberated upon. Challenges associated with recording emotional speech databases and annotations are also addressed. An analysis is conducted on speech emotion recognition methods based on feature sets and classification techniques. **[13-15]** The evaluation assesses the performance and limitations of existing methods for emotion recognition. Moreover, promising directions for enhancing speech emotion recognition systems are highlighted. The discourse delves into the development of deep learning-based SER and neural architectures incorporating Attention Mechanisms (AM). Recent SER systems with various attention mechanisms are reviewed, and the accuracies of these systems on the IEMOCAP benchmark database are compared. **[16-18]** The USC-IEMOCAP dataset is utilized for experiments in speech emotion recognition, with the proposed approach achieving performance comparable to state-of-the-art systems. The classification results reveal confusion between the Happy and Angry classes. The discussion also explores gender identification and emotion recognition using speech features. A comparison is made between Radial Basis Function and Back Propagation Network for emotion recognition. Additionally, it delves into the discussion on recognition accuracy using MLR, SVM, and RNN classifiers. **[19-21]** A comparison of results for Berlin and Spanish databases with different classifiers is conducted to identify performance differences attributed to speaking styles between the SUSAS and AIBO databases. It is noted that feature extraction plays a critical role in emotion recognition compared to the selection of classifiers, with the HMM classifier found to be weak in modeling long-time temporal dynamics.

TABLE 3. Training and validation accuracy & loss: Comparison results

Model	Training Accuracy	Validation Accuracy	Training Loss	Validation Loss
SVM	73%	72%	80%	93%
HMM	96%	95%	99.5%	99.7%

GMM	75%	75.9%	87.9%	92.2%



5. conclusion

Speech emotion recognition encompasses a variety of databases, characteristics, and classifiers. Different linguistic contexts and emotional conditions are taken into account during the analysis. Diverse characteristics and classifiers play a crucial role in the process of speech emotion recognition. The utilization of Deep CNN for emotion categorization has resulted in an accuracy of 84.3%. Speech processing revolves around the manipulation of signals to extract essential data. Systems for speech emotion recognition employ a range of classifiers for the identification of emotions. The amalgamation of methodologies and the extraction of efficient speech characteristics contribute to the enhancement of accuracy in recognition. An approach based on Hidden Markov Models (HMM) has achieved an accuracy of over 80% in the recognition of seven distinct emotions. Spectral measures have been disregarded due to their complexity and reliance on phonetics. Features such as pitch and Mel-frequency cepstral coefficients (MFCC) prove to be effective in emotional recognition. The recognition of emotional speech is imperative for interfaces between humans and computers. Research literature on speech emotion recognition primarily

focuses on databases, characteristics, and models, while also shedding light on existing research gaps and challenges. There is a scarcity of studies on the utilization of multiple classifiers in the realm of speech emotion recognition. Hierarchical models have shown superior performance compared to conventional classifiers in the recognition of emotions. Proposed methodologies exhibit enhanced rates of recognition within the domain of speech emotion recognition. Spontaneous audio recordings may exhibit an imbalance in emotional categories. Attention mechanisms have a significant impact on the performance of Speech Emotion Recognition (SER) systems. The process of representation learning plays a pivotal role in the classification of emotions based on speech spectrograms. Transfer learning from valence and activation levels marginally enhances the recognition of emotions. The confusion between categories like Happy and Angry can be mitigated through inverse filtering. The classification performance reaches its peak for the Sad emotion category. Emotion recognition based on pitch and formant positions is thoroughly examined. Gender classification can be achieved through pitch extraction utilizing the autocorrelation method. Several classifiers have demonstrated an accuracy rate of 83% for the Berlin database. A Recurrent Neural Network (RNN) classifier has attained a remarkable accuracy of 94% for the Spanish database. The process of feature selection leads to improved recognition outcomes for the Berlin database. However, an RNN model with an excessive number of parameters may result in overfitting. The accuracy of feature combination is enhanced through the utilization of LR-RFE feature selection. The significance of pitch and energy in emotion recognition cannot be overstated. Gaussian Support Vector Machine (SVM) has exhibited superior performance in the classification of stressed and neutral styles. Feature extraction is deemed more crucial than the selection of classifiers. Further exploration is warranted for the introduction of new features and the enhancement of algorithms.

References

1. 1.Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: a review. International *Journal of Speech Technology*, *21*, 93120.
2. Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021). A comprehensive review of speech emotion recognition systems. *IEEE access*, *9*, 47795-47814.
3. Ingale, A. B., & Chaudhari, D. S. (2012). Speech emotion recognition. *International Journal of Soft Computing and Engineering (IJSCE)*, *2*(1), 235-238.

4. El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, *44*(3), 572-587. 5. Jain, M., Narayan, S., Balaji, P., Bhowmick, A., & Muthu, R. K. (2020). Speech emotion recognition using support vector machine. *arXiv preprint arXiv:2002.07590*.
6. Nogueiras, A., Moreno, A., Bonafonte, A., & Mariño, J. B. (2001). Speech emotion recognition using hidden Markov models. In *Seventh European conference on speech communication and technology*.
7. Han, K., Yu, D., & Tashev, I. (2014, September). Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech 2014*.
8. Ramakrishnan, S., & El Emary, I. M. (2013). Speech emotion recognition approaches in human computer interaction. *Telecommunication Systems*, *52*, 1467-1478.
9. Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. *International journal of speech technology*, *15*, 99-117.
10. Aouani, H., & Ayed, Y. B. (2020). Speech emotion recognition with deep learning. *Procedia Computer Science*, *176*, 251-260.
11. Basharirad, B., & Moradhaseli, M. (2017, October). Speech emotion recognition methods: A literature review. In *AIP conference proceedings* (Vol. 1891, No. 1). AIP Publishing.
12. Shen, P., Changjun, Z., & Chen, X. (2011, August). Automatic speech emotion recognition using support vector machine. In *Proceedings of 2011 international conference on electronic & mechanical engineering and information technology* (Vol. 2, pp. 621-625). IEEE.
13. 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.
14. Ghosh, S., Laksana, E., Morency, L. P., & Scherer, S. (2016, September). Representation learning for speech emotion recognition. In *Interspeech* (pp. 3603-3607).
15. Selvaraj, M., Bhuvana, R., & Padmaja, S. (2016). Human speech emotion recognition. *International Journal of Engineering & Technology*, *8*(1), 311-323.
16. Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M. A., & Cleder, C. (2019). Automatic speech emotion recognition using machine learning.
17. Deng, J., Xu, X., Zhang, Z., Frühholz, S., & Schuller, B. (2017). Semisupervised autoencoders for speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(1), 31-43.
18. Kwon, O. W., Chan, K., Hao, J., & Lee, T. W. (2003, September). Emotion recognition by speech signals. In *Interspeech* (pp. 125-128).
19. Peerzade, G. N., Deshmukh, R. R., & Waghmare, S. D. (2018). A review: Speech emotion recognition. *Int. J. Comput. Sci. Eng*, *6*(3), 400-402.
20. Zheng, W. Q., Yu, J. S., & Zou, Y. X. (2015, September). An experimental study of speech emotion recognition based on deep convolutional neural networks. In *2015 international conference on affective computing and intelligent interaction (ACII)* (pp. 827-831). IEEE.
21. Anagnostopoulos, C. N., Iliou, T., & Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, *43*, 155-177.