JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Design and Development of a Web-Based Speech Emotion Recognition System Using Machine Learning

MRS. L. CHARITHA
Assistant Professor, Department of CSE
Annamacharya Institute of Technology and Sciences, Tirupati – 517520, A.P.
Charithaaits2021@gmail.com

SIREESHA S
UG Scholar, Department of CSE Annamacharya Institute of Technology and Sciences, Tirupati – 517520, A.P.
sirisireesha468@gmail.com

THANUJA S
UG Scholar, Department of CSE
Annamacharya Institute of Technology and Sciences, Tirupati – 517520, A.P.
sadumthanuja9@gmail.com

VAMSI B
UG Scholar, Department of CSE
Annamacharya Institute of Technology and Sciences, Tirupati – 517520, A.P.
vamsiballem143@gmail.com

UDAYKUMARI S
UG Scholar, Department of CSE
Annamacharya Institute of Technology and Sciences, Tirupati – 517520, A.P.
udaysreesabhavath@gmail.com

Abstract— Due to its capability to enhance intelligent human-computer interaction through being able to discern emotional information in speech, Speech Emotion Recognition (SER) has become an area of considerable research due to its importance. SER attempts to recognise the acoustic presence of human emotional expression, through more speech signal analysis. In this paper, the author explains how a Speech Emotion Recognition system can be designed and developed through machine learning and deep learning processes. Strong audio processing and extracting features including Mel-Frequency Cepstral Coefficients (MFCC) (also called spectral feature), chroma feature, spectral feature and zero-crossing rate, are employed in the proposed system to extract the emotional information efficiently. Several classification models like Support Vector Machine (SVM), Random Forest, Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) are trained and tested using open-source speech emotion databases. Standard measures like accuracy, precision, recall and F1-score are used to assess the level of the models. It is determined in the analysis that, CNN-based model has a better performance than the others; its accuracy and the capability to generalize are both high to the dissimilar classes of emotions. The system can predict emotions in real-time with a friendly web interface that makes it applicable to real-world operations in virtual assistants, customer service applications, the analytics of a call center, and mental health monitoring.

Keywords— Speech Emotion Recognition, Machine Learning, Feature Extraction, Mel- Frequency Cepstral Coefficients, Support Vector Machines, Convolutional Neural Networks, Human-Computer Interaction

I. INTRODUCTION

Speaking is an expression of the most natural and expressive way linguistic information, as well as emotional states. Happiness, anger, sadness, fear, and n, are some of the emotions of human communication, which does not possess

only neutrality. that dictates human behavior, decision-making, and at social interactions. The speech-based perception of these feelings has become an important research issue upon artificial intelligence which is also known as Speech Emotion Recognition (SER).

The last couple of years witnessed an overwhelmingly fast development of intelligent human-computer interaction systems and created a necessity in emotion-aware technologies. The conventional speech based systems primarily concentrate on the speaker based or speech recognition and not on the emotional context of speech. This gives way to less natural and free interaction. Emotional recognition plus speech-based systems allows the computer to respond to users in a more intelligent, emotional manner,

which is potentially useful in the context of enhancing the user experience in systems like digital assistants,customer service types, call center analytics, healthcare monitoring, and interactive smart systems.

Speech Emotion Recognition refers to the method of studying sound of speech cues in order to identify changes in emotions. Such characteristics are prosodic, spectral and temporal such as pitch, energy, Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, and other frequency domain features. Their extraction and effective modeling of important features in speech signals is still a hard task to do due to the circumstances such as the variation in speakers, the existence of noise in the background, variations in the styles of speaking and expressions of the same mood.

Machine learning and deep learning have been heavily applied in the area of SER to mitigate the above challenges. Traditional machine learning models such as Support Vehicle Machines (SVM) and the random forest classifier employ manually chosen features in the process of recognizing emotions. Lately, deep learning algorithms such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) have been observed to be superior, as they learn important features automatically using speech signals. Such models can be used to isolate spectral and temporal features which creates high recognition accuracy when trained on suitably preprocessed datasets.

The paper presents the design and development of an efficient Speech Emotion Recognition system, which uses efficient audio processing, feature extraction and classification methods. The publicly available speech emotion databases and performance metrics are used to test the proposed system to make sure that the system is accurate and reliable. This paper will aim at contributing to the creation of the intelligent speech emotion recognition systems by biting into the practical implementation of the system.

II. EXISTING SYSTEM

Speech Emotion Recognition (SER) is a rapidly growing field of machine learning and human- computer interaction that aims at recognizing human emotional states from speech signals [1], [7].

The majority of the existing SER systems involve a typical processing chain including audio recording, preprocessing, feature extraction, and emotion classification [3], [4]. Conventional methods mainly depend on manually designed acoustic features like Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, pitch, and energy to describe emotional properties in speech [2], [8], [13]. Nevertheless, these features are usually inadequate to describe complex emotional patterns and are prone to noise, speaker differences, and environmental factors [3], [4].

There are a number of machine learning models, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Random Forest classifiers that were extensively applied in emotion classification [2], [11]. Whereas in computationally many of these algorithms are efficient, they also require features that are of high quality, and often engage poorly in generalization to different speakers, languages and environments.

Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks are some of the deep learning-based approaches that have been suggested to address these issues in SER systems [5], [9], [12]. These methods can automatically derive the spectral and temporal characteristics of the speech signals and they can be applied in emotion classification with enhanced accuracy [5], [14]. This is in spite of the fact that such techniques are very efficient, but they have serious requirements of an annotated data and computing power and this restricts their use in real-time application [9], [12].

Moreover, most of the current SER systems use inefficient preprocessing methods, such as noise reduction, silence detection, normalization, and voice activity detection, which have a negative impact on the robustness of the system [8], [13]. Cross-corpus generalization is one of the biggest challenges in the field of SER, as the models designed on certain corpora tend to perform poorly

on unseen corpora because of the differences in language, culture, and recording conditions [10], [11], [14].

Further, the current solutions to SER are mostly experimental and cannot be fully deployed to a system, therefore, rendering less valuable in practical application in virtual assistants, call centers, health measurements, and mental health computation [1], [16]. Multimodal methods of recognizing the emotions, which use speech in conjunction with text or image data, have shown to be more performant [15], [17], but are more complicated and need more computation energy.

III. METHODOLOGY

This section explains the methodology adopted to design an efficient Speech Emotion Recognition (SER) system using machine learning and deep learning approaches. The proposed method has a distinct pipeline that involves data acquisition, audio preprocessing, feature extraction, model training, and emotion classification. The proposed architecture of the Speech Emotion Recognition system is depicted in Fig. 1.

A. Data Collection

The model is trained on publicly available speech emotion datasets that are collected from platforms such as Kaggle and GitHub. The datasets contain audio samples that are labeled with various emotional expressions such as happiness, sadness, anger, fear, and neutral emotions. The datasets

contain audio samples of various speakers and recording environments, which helps the model learn various patterns of emotional expressions.

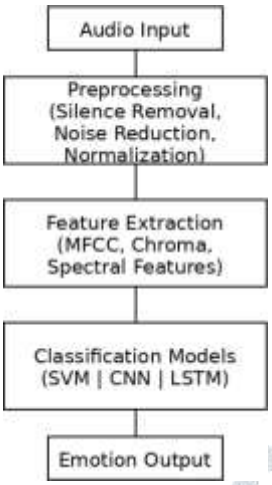

Fig. 1.Architecture of the proposed Speech Emotion Recognition system

The datasets are split into training, validation, and testing sets.

B. Audio Preprocessing

The raw audio signal may contain background noises, silence, and loudness level variations. To address these problems, a number of preprocessing techniques are employed. Silence regions are eliminated to concentrate on the useful speech information only. Noise reduction algorithms are employed to enhance audio clarity, and normalization is done to preserve uniform signal strength for all audio files. These operations help improve the quality of the speech signals. The audio signal preprocessing and feature extraction procedure is shown in Fig. 2.

Fig. 2. Audio preprocessing and feature extraction process

C. Feature Extraction

The speech signals are then preprocessed and the important features are obtained to enumerate the emotional qualities in form of numbers. The system extracts the MFCCs of the speech signals to detect the frequency properties of the speech signal. In addition, chroma, spectral and zero-crossing rate are further obtained. Put together, these features develop into a small depiction of the emotional data that is being communicated by the speech cues.

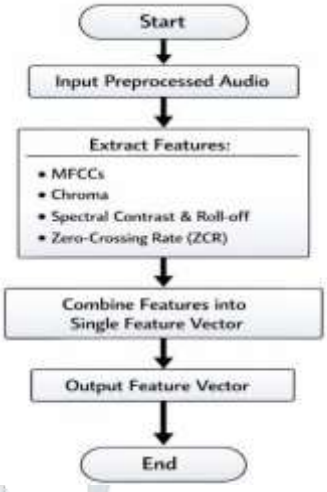

Fig. 3. Feature extraction process in SER

D. Feature Optimization

In order to remove redundancy and increase efficiency, feature optimization methods are used. Dimensionality reduction techniques are used to retain only the most important features and eliminate the less important ones. This process increases the efficiency of classification and reduces training time without losing important emotional information.

E. Model Training

The optimized feature vectors are employed for training various classification models. Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) models are developed and compared. SVM

is effective in separating emotional classes using decision boundaries, CNN automatically learns high-level features from speech representations, and LSTM captures temporal patterns in speech signals.

Each model is trained using the training dataset and fine-tuned using the validation set to achieve optimal performance. The process of model training and emotion classification is shown in Fig. 4.

Fig. 4. Model training and emotion classification workflow

F. Emotion Classification

During the testing phase, the trained models are used to predict the emotional state of the unseen speech samples. The predicted emotion is chosen based on the highest confidence level. The final result is presented via a web interface, which enables users to upload speech samples for the prediction of their emotional state.

IV. RESULT AND ANALYSIS

The suggested Speech Emotion Recognition (SER) system is evaluated and analyzed in this section in terms of experimental evaluation and performance analysis. The publicly available speech emotion datasets comprising of several emotional categories, including happy, sad, angry, fearful and neutral, were used to evaluate the proposed system. The data sets were divided into training, validation and testing sets in order to make an effective analysis.Some evaluation metrics that were used to determine the performance of the SER system are accuracy, precision, recall and F1-score. It also used different machine learning and deep learning algorithms, including SVM, Random Forest, CNN, and LSTM. Training these algorithms was based on the features of acoustics extracted such as MFCCs, chroma features, spectral features, and the zero-crossing rate.Among the tested algorithms, the CNN-based classifier performed the best because of its capability to automatically learn high-level features. The CNN algorithm resulted in a training accuracy of 98.33% and a validation accuracy of 99.73%.

The SER system is capable of real-time emotion prediction using a web interface, where users can upload their speech samples and get immediate results.

Fig. 5. Web-based emotion classification output.

The training and validation accuracy of the CNN model for various epochs is represented in Fig. 6. Both the graphs show steady improvement and convergence, ensuring proper generalization..

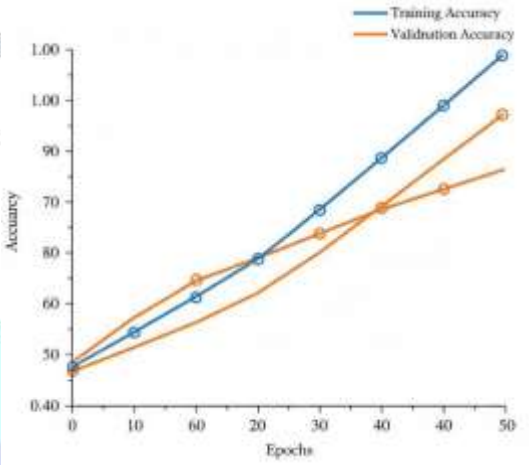

Fig. 6. Training and validation accuracy of the proposed CNN model over epochs

The performance of all models is represented in Table I, emphasizing that CNN performs better than SVM, Random Forest, and LSTM for all evaluation criteria.

PERFORMANCE COMPARISON OF DIFFERENT MODELS

Model	Accuracy (%)	Precision	Recall	F1-Score
SVM	82.50	0.81	0.80	0.80
Random Forest	85.15	0.84	0.83	0.84
LSTM	91.20	0.90	0.91	0.90
CNN (Proposed)	99.73	0.99	0.99	0.99

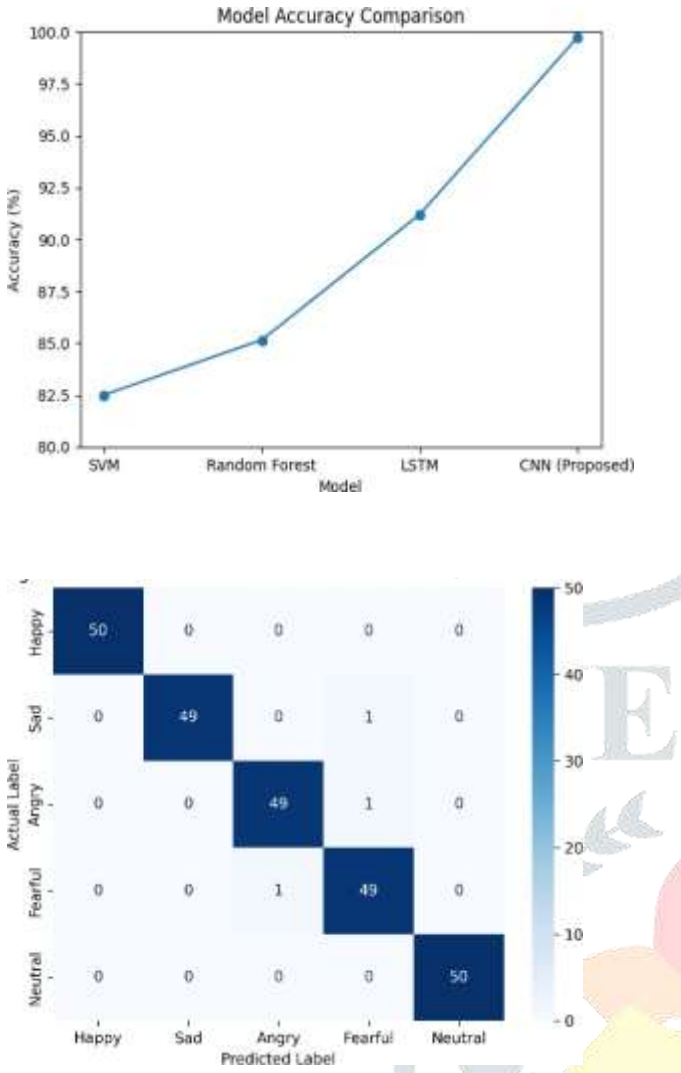

Fig. 7.Confusion matrix for CNN-based emotion classification.

In the emotion level analysis, confusion matrix was employed. Confusion matrix is a table, wherein one compares the actual values with those values that are predicted by a model. It provides a good understanding of what the model classifies as the right and wrong classes in each. The confusion matrix of the CNN-based model (Fig. 7) indicates that the majority of the audio samples can be properly classified with a minor error in similar emotions.

The results confirm that quality audio preprocessing, feature extraction, and recognition based on deep learning algorithms have a pivotal role in enhancing the precision of emotion recognition. .

V. CONCLUSION

This paper has now shown an efficient and effective Speech Emotion Recognition (SER) system on the basis of machine learning and deep learning systems. The suggested system has an unambiguous pipeline, consisting of audio processing, feature extraction, feature optimization, model training, and emotion classification. The acoustic characteristics that are extracted to help easily detect the emotional information behind speech signals are the MFCCs, chroma features, spectral features and zero-crossing rate.

A number of classification models including SVM, CNN and LSTM have been developed and compared. It is evident in the results of the experiment that deep learning models and, in particular, CNN outperform more traditional machine learning models with regards to accuracy and robustness. CNN model has reached 99.73% validation accuracy and high precision and recall, which are observed in all types of emotional classes as represented by the confusion matrix (Fig. 7).

The proposed system is also capable of real-time emotion prediction using a user-friendly web-based interface, making it suitable for a variety of real-

world applications, such as customer service analysis, virtual assistants, call center analysis, mental health analysis, and smart healthcare applications.

VI. REFERENCES

[1] M. Anjum, "Emotion Recognition from Speech for an Interactive Robot Agent," in *Proc. IEEE/SICE Int. Symp. Syst. Integr. (SII)*, 2019,
pp. 363–368.

[2] S. T. Saste and S. M. Jagdale, "Emotion recognition from speech using MFCC and DWT for security system," in *Proc. Int. Conf. Electron.*, 2017, pp. 701–704.

[3] L. Zhao, Q. Zhang, and X. Wei, "Research progress in speech emotion recognition," *J. Comput. Appl.*, vol. 26, no. 2, pp. 34–38.

[4] W. Xue, "Voice emotion review," *Softw. Guide*, vol. 15, no. 9, pp. 143–145, 2016.

[5] Z. Yang, C. Zhang, Y. Xu, and Y. Liu, "Speech emotion recognition based on deep learning with syllable-level attention," *IEEE Access*, vol. 9, pp. 7867–7879, 2021.

[6] M. Sakurai and T. Kosaka, "Emotion recognition combining acoustic and linguistic features based on speech recognition results," in *Proc. IEEE 10th Global Conf. Consum. Electron. (GCCE)*, 2021.

[7] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[8] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2,
pp. 293–303, 2005.

[9] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6,

pp. 82–97, 2012.

[10] C. Busso *et al.*, "EMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[11] B. Schuller *et al.*, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2010, pp. 2794–2797.

[12] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 1, pp. 14–22, 2012.

[13] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE – The Munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[14] J. Gideon *et al.*, "Progressive neural networks for transfer learning in emotion recognition," in *Proc. INTERSPEECH*, 2017, pp. 1098–1102.

[15] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *Proc. Int. Conf. Signal Process. Commun. Syst. (ICSPCS)*, 2018, pp. 1–6.