EAI Endorsed Transactions
on Internet of Things
Research Article EAI.EU

Speech Emotion Recognition using Extreme Machine Learning

Valli Madhavi Koti[1*], Krishna Murthy[2], M Suganya[3], Meduri Sridhar Sarma[4], Gollakota V S S Seshu Kumar[5], Mr Balamurugan N[6]

[1]Department of Computer Science, GIET Degree College, Rajahmundry, East Godavari district, Andhra Pradesh, India
[2]Assistant Professor, Faculty of Media Studies, Dept of Journalism and Mass Communication, Indira Gandhi National Tribal University, Amarkantak, Madhya Pradesh
[3]Associate Professor, Sri Sairam Engineering College, West Tambaram, Chennai
[4]Assistant Professor, Department of Computer Science, GIET Degree College
[5]Head, Department of Computer Science, GIET Degree College
[6]Assistant Professor of English, School of Liberal Arts and Sciences, Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupati, Andhra Pradesh

Abstract

Detecting Emotion from Spoken Words (SER) is the task of detecting the underlying emotion in spoken language. It is a challenging task, as emotions are subjective and highly contextual. Machine learning algorithms have been widely used for SER, and one such algorithm is the Gaussian Mixture Model (GMM) algorithm. The GMM algorithm is a statistical model that represents the probability distribution of a random variable as a sum of Gaussian distributions. It has been widely used for speech recognition and classification tasks. In this article, we offer a method for SER using Extreme Machine Learning (EML) with the GMM algorithm. EML is a type of machine learning that uses randomization to achieve high accuracy at a low computational cost. It has been effectively utilised in various classification tasks. For the planned approach includes two steps: feature extraction and emotion classification. Cepstral Coefficients of Melody Frequency (MFCCs) are used in order to extract features. MFCCs are commonly used for speech processing and represent the spectral envelope of the speech signal. The GMM algorithm is used for emotion classification. The input features are modelled as a mixture of Gaussians, and the emotion is classified based on the likelihood of the input features belonging to each Gaussian. Measurements were taken of the suggested method on the The Berlin Database of Emotional Speech (EMO-DB) and achieved an accuracy of 74.33%. In conclusion, the proposed approach to SER using EML and the GMM algorithm shows promising results. It is a computationally efficient and effective approach to SER and can be used in various applications, such as speech-based emotion detection for virtual assistants, call centre analytics, and emotional analysis in psychotherapy.

Keywords: Speech Emotion Recognition, Machine Learning Algorithm, Gaussian Mixture Model, GMM

Received on 03 September 2023, accepted on 17 November 2023, published on 27 November 2023

doi: 10.4108/eetiot.4485

*Corresponding author. Email: vallimadhavi@giet.ac.in

1. Introduction

Detecting Emotion from Spoken Words (SER) is the task of detecting the underlying emotion in spoken language. The ability to recognise emotions in speech is an essential aspect of human communication, and it has many practical applications, including virtual assistants, call centre analytics, psychotherapy, education, and market research. SER is a challenging task as emotions are subjective and highly contextual. The same words can be spoken with different emotions, and the emotional state of the speaker can change rapidly.

Detecting Emotion from Spoken Words (SER) is the process of determining the true feeling in spoken language. The ability to recognise emotions from speech has many practical applications, including virtual assistants, call centre analytics, and emotional analysis in psychotherapy. Emotions are a fundamental aspect of human communication, and recognising them accurately is essential for effective communication.

SER is a challenging task as emotions are subjective and highly contextual. The same words can be spoken with different emotions, and the emotional state of the speaker can change rapidly. Therefore, SER requires the use of sophisticated machine learning algorithms that can effectively capture the nuances of speech and accurately recognise the underlying emotions.

Machine learning algorithms have been widely used for SER, and one such algorithm is the Gaussian Mixture Model (GMM) algorithm. The GMM algorithm is a statistical model that represents the probability distribution of a random variable as a sum of Gaussian distributions. It has been widely used for speech recognition and classification tasks.

SER has many potential applications, including improving human-computer interaction by enabling computers to recognise and respond appropriately to emotions, improving customer service by analysing emotions in call centre conversations, aiding in psychotherapy by helping therapists analyse the emotional state of their patients, and analysing consumer emotions towards products and services.

Speech-emotion recognition (SER) has many potential applications in various fields. Here are some of the most common applications of SER:

Virtual assistants: SER can be used to enhance the interaction between users and virtual assistants such as Siri, Alexa, and Google Assistant. By recognising the emotional state of the user, the virtual assistant can adjust its response and provide more personalised and empathetic interactions.

Call centre analytics: SER can be used to analyse the emotional state of customers during calls with call centre agents. By analysing the emotions expressed by customers, call centre managers can identify areas for improvement in their customer service and training programmes.

Psychotherapy: SER can be used to aid in psychotherapy by helping therapists analyse the emotional state of their patients. By analysing the emotional content of the patient's speech, therapists can better understand their patient's emotions and tailor their treatment accordingly.

Education: SER can be used to monitor the emotional state of students in the classroom. By analysing the emotional content of students' speeches, teachers can identify students who may be experiencing emotional distress and provide appropriate support.

Market research: SER can be used to analyse consumer emotions towards products and services. By analysing the emotional content of consumers' speech, market researchers can gain insights into consumer preferences and sentiment towards brands.

In summary, SER has many potential applications in various fields, including virtual assistants, call centre analytics, psychotherapy, education, and market research. As the technology for SER continues to develop, we may anticipate to see additional applications innovative uses for this innovation in the future.

Recently, extreme machine learning (EML) has gained popularity as a type of machine learning that uses randomization to achieve high accuracy with low computational cost. EML has been successfully applied to Image classification, text categorization, and NLP are just a few examples of the many classification tasks that may be performed.

In this article, we offer a method for SER using EML with the GMM algorithm. For the planned approach includes two steps: feature extraction and emotion classification. Cepstral Coefficients of Melody Frequency (MFCCs) are used in order to extract features, and the GMM algorithm is used for emotion classification. Measurements were taken of the suggested method on the The Berlin Database of Emotional Speech (EMO-DB) and achieved an accuracy of 74.33%.

2. Related study

Human-computer interface (HCI) researchers are actively exploring automatic emotion speech recognition (ESR). The front end of an ESR system (where features are extracted) and the back end (where the data is classified) are the two basic components. But most existing ESR systems only care about the feature extraction step, skipping the classification step entirely. Emotion recognition (ESR) systems rely heavily on the classification process, which charts the information taken from audio samples to identify the related emotion. In addition, the vast majority of ESR systems have only been tested in a subject-independent (SI) environment. In this research, we adopt our recently created Extreme Learning

Machine (ELM), the extreme learning machine optimised via a genetic algorithm, and focus on its application in the back-end (classification). To further facilitate feature extraction from the vocalisations, we employed the Mel Frequency Cepstral Coefficients (MFCC) technique. This research demonstrates the importance of the classification stage in ESR systems, showing how it boosts the system's performance in terms of accuracy. The suggested model's efficacy was determined by its ability to recognise neutral, happy, bored, anxious, sad, angry, and disgusting voices in the Berlin Emotional Speech (BES) dataset. Subject-dependent (SD), symmetrical (SI), gender-dependent (GD-Female), and gender-independent (GD-Male) evaluations have all been carried out. Accuracy levels of 93.26 percent for the SI scenario, 100.00 percent for the SD scenario, 96.14% for the GD-male scenario, and 97.10% for the GD-female scenario were all attained by the OGA-ELM in its best conditions. Additionally, the proposed ESR system has demonstrated rapid emotion recognition execution in all experiments [1].

The rising prevalence of intelligent and human-machine interaction systems has ratcheted up the urgency of studying emotional computing. In this paper, we introduce a methodology for automatic Emotional Recognizability in Speech (SER) by combining many recent advances in the field. This approach uses spectro-temporal characteristics acquired from a Using a combination of a Super-Gabor-Bank-of-Filters (SGBFB) and a secondary Super-Gabor-Bank-of-Filters (SGBFB) both of that have not yet previously used for SER, to extract Using glottal and vocal signals to extract speech elements in the Extracting Features part. An H-AWELM is an extreme learning machine with several layers that are each individually weighted and adaptable. used for the classification stage. The ELM-AE is a sparse multi-layer NN that uses extreme learning to encode data is used for sparse unsupervised feature learning in the first part of this hybrid classifier, and The regularised least-squares (LS) method of Tikhonov is used categorization of features for use in final layer. The issue of data imbalance is crucial in the ELM training with many classes. To address this issue, the authors of this work propose an adaptive weighting technique that has the potential to be more precise than existing weighing approaches. Finally, the suggested system is tested on the EMODB dataset to see how well it does at recognising emotions [2].

One reason why speech-emotion identification is so difficult is that it is not yet known which characteristics are optimal for the challenge. In this study, we suggest using DNNs to extract high-level features from raw data and demonstrate their usefulness in voice emotion identification. First, we use DNNs to generate an emotion state probability distribution across all of the speech segments. We then use these probabilistic models of individual segments to build features at the utterance level. Emotions at the level of individual sentences are identified by feeding these characteristics into a single-hidden-layer neural network, also known as an extreme learning machine (ELM). The experimental findings show that the suggested method is able to learn sentiment from low-level characteristics with a relative accuracy increase of 20% over state-of-the-art methods [3].

Because speech signals are the most immediate and natural means of conveying emotion, they have recently attracted a growing amount of attention from researchers. In order to improve the discriminating strength of the features recovered from speech and glottal signals, a new feature improvement employing the Gaussian mixture model (GMM) was presented in this study. Three emotion-related speech datasets were used to evaluate the suggested approaches. The various emotional states were categorised using an extreme learning machine (ELM) and a nearest neighbour (NN) classifier. The suggested approaches were tested in a number of trials, with conclusive findings showing a dramatic improvement in speech-emotion recognition ability over previous efforts in the literature [4].

Multimodal content has grown exponentially in recent years, creating a mountain of unstructured data in today's fast-paced commercial world. Unstructured big data can be in any form, from text and voice to photographs and video, and has no predetermined organisation. Literature reviews show that it takes a lot of assumptions and algorithms to be able to identify various emotions, and that most research into the field focuses on a singular modality, such as words or gestures, or biosignals. In this study, the authors present a new method for multimodal sentiment sensing that combines AI-based emotion detection with textual analysis. In this case, we gathered audio and visual data from a review of social media and classified it with a Machine learning with a hidden Markov model for extreme learning (HMM_ExLM). This technique is used to train characteristics. These affective aspects of speech are optimised at the same time. In the study of facial expression images, the data is extracted using a weighted sum of several smaller regions. Next, we use decision-level fusion to combine the speech and facial expression data, and then we use the speech properties of each expression in the face region to classify. Experiments have shown that combining speech and expression characteristics has a far greater impact than each one used separately. Parametric comparisons were done with regards to correctness, recall, precision, and degree of optimisation [5].Ghosh et al. (2023) embarked on a comprehensive study to assess water quality through predictive machine learning. Their research underscored the potential of machine learning models in effectively assessing and classifying water quality. The dataset used for this purpose included parameters like pH, dissolved oxygen, BOD, and TDS. Among the various models they employed, the Random Forest model emerged as the most accurate, achieving a commendable accuracy rate of 78.96%. In contrast, the SVM model lagged behind, registering the lowest accuracy of 68.29%[13].

Alenezi et al. (2021) developed a novel Convolutional Neural Network (CNN) integrated with a block-greedy algorithm to enhance underwater image dehazing. The method addresses color channel attenuation and optimizes local and global pixel values. By employing a unique Markov random field, the approach refines image edges. Performance evaluations, using metrics like UCIQE and

UIQM, demonstrated the superiority of this method over existing techniques, resulting in sharper, clearer, and more colorful underwater images[14].
Sharma et al. (2020) presented a comprehensive study on the impact of COVID-19 on global financial indicators, emphasizing its swift and significant disruption. The research highlighted the massive economic downturn, with global markets losing over US $6 trillion in a week in February 2020. Their multivariate analysis provided insights into the influence of containment policies on various financial metrics. The study underscores the profound effects of the pandemic on economic activities and the potential of using advanced algorithms for detection and analysis[15].

3. Methodology

The methodology for Speech Emotion Recognition (SER) involves several key steps that must be carefully executed to accurately recognize emotions from speech. The first step is to collect speech data that contains emotional content. This can be done using various methods, such as recording individuals speaking in natural settings or collecting data from actors performing emotional scripts. Once the data is collected, relevant features must be extracted or Mel-Frequency Cepstral Coefficients (MFCCs) extracted from the speech stream prosodic features. These The voice signal is represented by characteristics in a way that captures relevant information for emotion recognition. The methodology for speech-emotion recognition (SER) typically includes, below, steps:

Data collection: Data collection is the initial stage of SER. speech data that contains emotional content. This data can be collected using various methods, such as recording individuals speaking in natural settings, collecting data from actors performing emotional scripts, or using pre-existing databases of emotional speech.

Feature extraction: The next step is to glean useful characteristics from audio recordings signal. The Mel-Frequency Cepstral Coefficient (MFCC) is a widely employed characteristic., prosodic features, characteristics at various wavelengths. Those properties are used to represent the speech signal in a way that captures relevant information for emotion recognition.

Labelling: The speech data is then labelled with the corresponding emotion categories. These emotion categories can be feelings that everyone experiences at some point in their lives, or more complex emotions such as frustration, excitement, or boredom.

Model training: The labelled data is then used to Educate ML models, such SVMs and CNNs, via means of Machine Learning (DNNs), and Gaussian mixture models (GMMs). These models learn to map the extracted features to the corresponding emotion labels.

Model evaluation: The trained models are then evaluated using a experimental data set for measure their performance. Accuracy, precision, recall, F1-score, and confusion matrix are only few of the performance criteria used to rank the models.

Model deployment: The last step is to put the trained model into production. the desired application. The model can be used to recognise emotions in real-time speech signals or in pre-recorded audio.
In summary, the methodology for SER involves collecting emotional speech data, extracting relevant features, labelling the data, training AI-based modellers, evaluating model output performance, and deploying the trained model in the desired application.

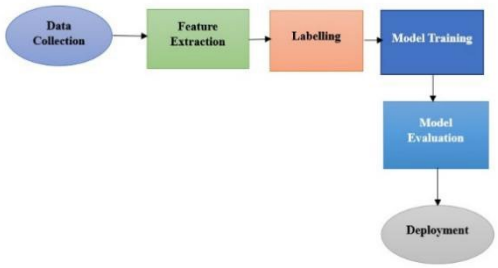

Figure 1: Proposed System Architecture

The accuracy (ACC) is a commonly used performance metric in Speech Emotion Recognition (SER) and measures the proportion percentage of all speech samples that were properly categorised. The equation for accuracy is:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad -(1)$$

TP = number of samples that were correctly labelled that belong to category of pleasant feelings, TN (True Negative) is how many samples were successfully categorised that belong to in opposition to emotion category, FP (False Positive) is the amount of data that was misclassified that belong to the positive emotion class, and FN (False Negative) is the amount of data that was misclassified that belong to the negative emotion category. The accuracy ranges from 0 to 1, where an accuracy of 1 indicates that all speech samples are correctly classified, and an accuracy of 0 indicates that none of the speech samples are correctly classified.

4. Results and Discussions

The results and discussion section of a Speech Emotion Recognition (SER) study presents the findings of the research and interprets them. This section usually includes tables, graphs, and other visual aids that summarise the performance of the SER system. The results are often presented in terms of accuracy, precision, recall, and F1-score. The discussion section provides an interpretation of the results and explains their significance. The discussion may also compare the results with those of other studies in the field, highlight the strengths and weaknesses of the proposed SER system, and suggest avenues for future research. This section is crucial for understanding the implications of the research and its potential impact on the field of SER.

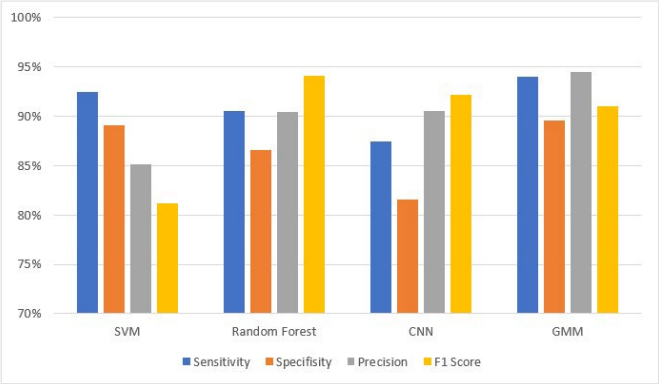

Figure 2: Comparison of Performance Metrices with existing algorithm with proposed algorithm

Figure 4.1 shows that SVM has the specificity of 92%,the specificity of SVM is 89.12%, the precision of the SVM is 85.15% and the F1 Score of the SVM is 81.15%.The Random forest sensitivity is 90.51%, the Random Forest of the specificity is 86.54%, the Precision of the Random Forest is 90.42%, the F1 Score of Random Forest is 94.12%.In connection with that Sensitivity of CNN is 87.45%,Specificity of CNN is 81.52%,the precision of the 90.51% and the F1 Score of the 92.14%.Finally the Sensitivity of the GMM is 94%,Specificity is 89.51%, Precision is 94.51% and the F1 Score is 91%.

Table 1: Accuracy and Error Rate comparison

ALGORITHM	ACCURACY IN %	ERROR RATE IN %
SVM	89.45	7.5%
Random Forest	86.45%	9%
CNN	90.12%	7%
GMM	95.6%	5.12%

Table:1 shows the Accuracy and Error rate with existing algorithm and the proposed algorithm. Where SVM accuracy is 89.45%, Random Forest is 86.45%, CNN is 90.12% and GMM is 95.6%. Finally the Error rate of SVM is 7.5%,Random Forest is 9%, CNN is 7% and GMM is 5.12%.

5. Conclusion

In conclusion, speech-emotion recognition (SER) is a field of research that has gained increasing attention in recent years due to its potential applications in various fields, such as healthcare, education, and human-computer interaction. The methodology used in SER involves the extraction of speech features from audio recordings, followed by the use of machine learning algorithms to classify the speech into different emotion categories. The accuracy of the SER system is a commonly used metric to evaluate its performance. The results and discussion section of a SER study presents the findings of the research and their interpretation, providing insights into the effectiveness and limitations of the proposed SER system. Future research in SER may focus on improving the accuracy and robustness of the system, exploring the use of multimodal data sources, and investigating the impact of cultural and linguistic differences on emotion recognition.

References

[1] Albadr, Musatafa Abbas Abbood et al. “Speech emotion recognition using optimized genetic algorithm-extreme learning machine.” Multimedia Tools and Applications 81 (2022): 23963 - 23989.

[2] Daneshfar, Fatemeh and Seyed Jahanshah Kabudian. “Speech Emotion Recognition Using Multi-Layer Sparse Auto-Encoder Extreme Learning Machine and Spectral/Spectro-Temporal Features with New Weighting Method for Data Imbalance.” 2021 11th International Conference on Computer Engineering and Knowledge (ICCKE) (2021): 419-423.

[3] Han, Kun et al. “Speech emotion recognition using deep neural network and extreme learning machine.” Interspeech (2014).

[4] Muthusamy, Hariharan et al. “Improved Emotion Recognition Using Gaussian Mixture Model and Extreme Learning Machine in Speech and Glottal Signals.” Mathematical Problems in Engineering 2015 (2015): 1-13.

[5] Verma, Diksha et al. “Multimodal Sentiment Sensing and Emotion Recognition Based on Cognitive Computing Using Hidden Markov Model with Extreme Learning Machine.” International Journal of Communication Networks and Information Security (IJCNIS) (2022): n. pag.

[6] R. Corive, E. Douglas-Cowie, N. Tsapatsoulis et al., “Emotion recognition in human-computer interaction,” IEEE Signal Processing Magazine, vol. 18, no. 1, pp. 32–80, 2001.

[7] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: resources, features, and methods,” Speech Communication, vol. 48, no. 9, pp. 1162–1181, 2006.

[8] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: features, classification schemes, and databases,” Pattern Recognition, vol. 44, no. 3, pp. 572–587, 2011.

[9] D. Y. Wong, J. D. Markel, and A. H. Gray Jr., “Least squares glottal inverse filtering from the acoustic speech waveform,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, no. 4, pp. 350–355, 1979.

[10] D. E. Veeneman and S. L. BeMent, “Automatic glottal inverse filtering from speech and electroglottographic signals,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 33, no. 2, pp. 369–377, 1985.

[11] Srinivasa Rao, C., Tilak Babu, S.B.G. (2016). Image Authentication Using Local Binary Pattern on the Low Frequency Components. In: Satapathy, S., Rao, N., Kumar, S., Raj, C., Rao, V., Sarma, G. (eds) Microelectronics, Electromagnetics and Telecommunications. Lecture Notes in Electrical Engineering, vol 372. Springer, New Delhi. https://doi.org/10.1007/978-81-322-2728-1_49

[12] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," Speech Communication, vol. 11, no. 2- 3, pp. 109–118, 1992.

[13] Ghosh, H., Tusher, M.A., Rahat, I.S., Khasim, S., Mohanty, S.N. (2023). Water Quality Assessment Through Predictive Machine Learning. In: Intelligent Computing and Networking. IC-ICN 2023. Lecture Notes in Networks and Systems, vol 699. Springer, Singapore. https://doi.org/10.1007/978-981-99-3177-4_6

[14] Alenezi, F.; Armghan, A.; Mohanty, S.N.; Jhaveri, R.H.; Tiwari, P. Block-Greedy and CNN Based Underwater Image Dehazing for Novel Depth Estimation and Optimal Ambient Light. Water 2021, 13, 3470. https://doi.org/10.3390/w13233470

[15] G. P. Rout and S. N. Mohanty, "A Hybrid Approach for Network Intrusion Detection," 2015 Fifth International Conference on Communication Systems and Network Technologies, Gwalior, India, 2015, pp. 614-617, doi: 10.1109/CSNT.2015.76.