Speech Emotion Recognition Using Machine Learning Approach

S. G. Shaila(✉), A. Sindhu, L. Monish, D. Shivamma, and B. Vaishali

Department of CSE (Data Science), Dayananda Sagar University, Bangalore, Karnataka, India
{shaila-cse,sindhua-cse,monishl-cse,shivammad-cse,
vaishalivb-cse}@dsu.edu.in

Abstract. Nowadays, emotion recognition and classification plays a vital role in the field of Human-Computer Interaction (HCI). Emotions are being recognized through behaviors of body such as facial expression, voice tone, and body movement. The present research considers Speech Emotion Recognition (SER) as one of the foremost used modality to identify emotions. SER dataset contains the four different datasets, Ravdess dataset is used in this project. This mechanism is used due to its high temporal resolution with no risks and less cost. Over the last decades, many researchers involved SER signals in sequence to cope up with Brain-Computer Interface (BCI) to detect emotions. It includes removing noises from audio signals, extracting temporal or spectral features from the audio signals, analysis on time or frequency domain respectively, and eventually, designing a multi-class classification strategy. The paper discusses the approach of identifying and classifying human emotions based on audio signals. The approach used machine learning technique such as Random Forest (RF), Multilayer Perceptron (MLP), Support Vector Machine (SVM), Convolution Network (CNN), and Decision Tree (DT) Models for classification. The obtained experimental result seems to be promising with good accuracy in the emotion classification.

Keywords: Emotions · Audio Signal · Random Forest (RF) · Multilayer Perceptron (MLP) · Support Vector Machine (SVM) Convolution Network (CNN) · Decision Tree (DT) · Ravdess Dataset · Classification

1 Introduction

Emotions play an essential role in human life in process of communication between people. Emotions are expressed in many ways, including facial expressions, move-ment of the body, and communication. Hence, nowadays, researchers tend to adopt the approach of acknowledging human emotions through audio signals. Emotions play a vital role in human life. It is one way to express One's feelings to others. Nowa-days, emotion recognition has become a very hot topic for researchers. Emotion has made effective and easy interaction between computers and people. Emotions can be recognized through different communication channels such as body language, facial expressions, voice recognition, etc. In some cases where there is a face-to-face conversation, the emotions of the person

S. Tamane et al. (Eds.): ICAMIDA 2022, ACSR 105, pp. 592–599, 2023.
https://doi.org/10.2991/978-94-6463-136-4_50

can be easily analyzed through his/her facial expression and body language, whereas the conversation is made through the medium. The person residing to expect from one another, the conversation and interaction, is made through the medium of the channel, and then it is hard to predict the emotion of the person. Here, speech emotion recognition (SER) is a method of ex-pressing One's emotional state through his/her speech. The main feature by which humans differ from other living beings is modulated vocal sounds. The voice of a human can be categorized into several attributes, such as loudness, pitch, vocal tone, and timbre. Through different vocal attributes, we can analyse human emotions easily. There are a few universal emotions like anger, sadness, happiness, surprise, fear, and neutrality that any system can be trained to identify easily. The feature extraction with the help of a human audio signal supports recognizing emotions. Emotion recognition has gained importance as it also supports physically disabled people who cannot express their emotions.

The paper compares and contrasts the approaches for identifying emotions using SVM, MLP, Random Forest, Decision Tree, and CNN models. SVM stands for Sup-port Vector Machines; it's a supervised learning algorithm. It works better for regression and classification problems. It is mainly used for classification problems. The main goal of SVM is to find the best boundary line that makes us easily classify the n-dimensional space into specified classes. This boundary line is known as a hyper-plane, so that we can easily add new data in the same category, which will be useful in the future. MLP stands for Multilayer Perceptron, and is a synthetic neural network feed-forward technique that generates a bunch of outputs from a few inputs. It con-sists of input, hidden, and output layers. It is one of the deep learning methods that use back propagation for training the model. MLP connects multiple layers into a single graph, which suggests the signal path to the nodes. Aside from the input node, each node features a nonlinear activation function. Random forest is a supervised learning technique based on the concept of ensemble learning. It is similar to the deci-sion tree. Ensemble learning is a method of combining many classifiers for better performance and to solve complex problems. The random forest contains the collection of decision trees. It predicts the output based on the majority voting instead of depending on a single decision tree. The decision tree is a type of supervised learning algorithm; here the dataset is represented as a tree. It works better for regression and classification problems and mainly for classification problems. As it is a tree-structured method, the dataset features are represented by internal nodes, decision rules are represented by branches, and the leaf node shows the output. CNN stands for "convolution neural network." It is an advanced neural network to classify emotions. CNN automatically detects the important features without any human intervention. It uses pooling and convolution operations for its computational efficiency. The rest of this paper is organized as follows. We review the literature in the Sect. 2, and the proposed work is presented in Sect. 3. In Sect. 4, we present the experimental results and conclude this paper in the Sect. 4.

2 Literature Review

This section discusses the research done in the field of speech emotion recognition. The various authors have used machine learning and deep learning techniques such as

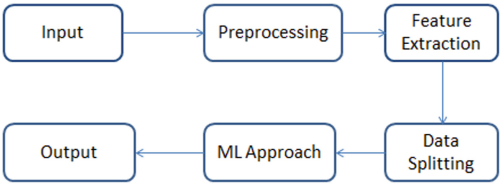

Fig. 1. Proposed Model for identifying Emotions

Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), long short-term memory (LSTM) and Bidirectional-LSTM, Support Vector Machine (SVM) for predicting human emotions. In the paper [1], the authors have analyzed the dataset using deep convolution recurrent network with LSTM in order to automatically learn the best representation of the speech signal. It uses recola dataset. The authors in [2] have used the Iemocap dataset, which contains both motion capture markers and audio data from five pairs of actors. It contains the 1D CNN LSTM and 2D CNN LSTM network algorithms. The authors in [3] have conferred a deep neural network for the dumb video, which does not contain any audio. Dump video detects 70% of the mood and condition of the person enacting. Similarly, from the voice of a person, 75% of the mood or condition could be detected. In the paper [4], the authors have owned the LDC and UGA datasets, which contain individual or groups of people's emotions. It uses support vector machines and it focuses mainly on gender datasets. The authors in [5] have presented a deep learning algorithm such as DCNN, to predict the accuracy of the model. For speaker-dependent, it gives an accuracy of 76.96% and speaker-independent accuracy of 65.32%. In the paper, the authors have [6] presented two popular machine learning techniques such as DNN and SVM that are applied on Iemocap dataset and results are compared on their performance. The accuracy is less than 55%. The authors have [7] used SVM classifier to recognize the emotion. Emotion recognition is analyzed in two phases; first, recognize the 42-dimensional features. Second, through the classification method using SVM, and achieved an accuracy of 74.62%. The main drawback of these papers is that as it is an audio signal-based dataset more features are used. This paper concentrates mainly on 5 features.

3 Proposed Work

This section represents the Proposed Model in Fig. 1. The approach used the Ravdess dataset as input for experimenting. The data set is pre-processed and further preceded with feature selection and extraction. The approach used cross-validation for splitting up of the Data into training and validation sets. Classification is done using various machine learning algorithms and performance is evaluated.

3.1 Dataset Description

The Ryerson Audiovisual Database of Emotional Speech and Song (RAVDESS) is a type of SER. It contains a total of 24 professional actors, with 12 male and 12 female

Table 1. Details of the Ravdess Dataset

Dataset	Actors	Instances	Emotions
RAVDESS	24	7356	8

actors recording voices. It contains a total of 7356 files. Speech includes emotions of happiness, calm, sad, angry, surprised, disgusted, and fear with two statements: "kids are talking by the door" and "dogs are sitting by the door". These two statements are expressed in all the above-listed emotions. The database contains full audio/video, only video, and only audio. As we are focusing on speech recognition, only audio is used in this project. It contains the emotional intensity of normal and strong. This is represented in below Table 1

3.2 Data Pre-processing

In preprocessing, data augmentation is done. It is a set of techniques to artificially increase the amount of data by generating new data from existing data points. This includes making small changes to the data to increase the performance of the model. It is noted that synthetic data generation of spoken MFCC can improve the recognition of a speaker from their utterances via a transfer learning method. Some ways of data augmentation are through noise injection into the dataset to check the performance of the database.

3.3 Feature Selection and Extraction

The extraction of features in the audio signal classification is a crucial method. The proposed approach is mainly focused on Chroma, MFCC, Mel features, contrast, and Tonnetz. Chroma is one of the powerful tools which is mainly used for analyzing pitch, Chroma is further categorized into 12 features. MFCC stands for Mel Frequency Cepstral Coefficient is one of the popular features used for recognizing the vocal tract mainly used to characterize speakers, for instance, it totally contains 39 inbuilt features to extract the audio of the speaker. Mel features are used to represent the short-term power spectrum. Contrast enhances the speech modulation and Tonnetz are used to fine-tune the tone. The proposed approach contains a total of 120 features, of which five are used. These are the five main features utilized in the proposed approach. The dataset is split into training, validation, and testing sets using a 70–20–10 split ratio. They are then split into input X and target Y of the respective categories for further processing. The given dataset is not identical across all of them with respect to the standard deviation of attribute values. Because of this issue, certain attributes end up being weighted over other attributes.

3.4 Classification

The dataset includes audio signal features such as Chroma, contrast, Mfcc, Mel-spectrum, and Tonnetz. The dataset contains a total of five features, and based on these

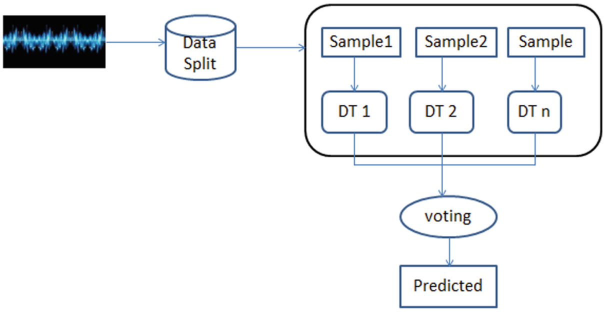

Fig. 2. RF Architecture

features, classification is done. Here, the models are trained to identify whether emotions are predicted or not. Initially, the proposed approach used the SVM model. Here, 70% of the data sample is used for developing the prediction model, and 30% is used for validating the developed model with a linear kernel function. In the next stage, the Random Forest model uses 80% of the data sample for training the model and 20% for validating the developed model without kernel function. Further, the proposed approach used a random forest with 100 decision trees for experimenting, and the results were with a ratio of 70:30, 70% for training the model and 30% for testing the model with 50 epochs. The next decision tree is used with three classes. A MLP classifier with 500 iterations and 300 iterations with a ratio of 80:20 is used. The CNN algorithm is used with two layers, three layers, and four layers. This is depicted in Fig. 2.

SVM stands for Support Vector Machines; it is mainly used for classification problems. The main goal of SVM is to find the best boundary line that makes us easily classify the n-dimensional space into specified classes. This boundary line is known as the hyperplane. The decision tree is a type of supervised learning algorithm; here the dataset is represented as a tree. As it is a tree-structured method, the dataset features are represented by internal nodes, decision rules are represented by branches, and the leaf node shows the output. The decision of the classifier is represented by the decision node and the outcome by the leaf node. Random forest is a supervised learning technique based on the concept of ensemble learning. It is similar to the decision tree. Ensemble learning is a method of combining many classifiers for better performance and to solve complex problems. The random forest contains the collection of decision trees. It predicts the output based on the majority voting instead of depending on a single decision tree.

MLP stands for Multilayer Perceptron, and is a synthetic neural network feed-forward technique that generates a bunch of outputs from a few inputs. It consists of input, hidden, and output layers. It's one of the deep learning methods that uses back propagation for training the model. MLP connects multiple layers into a single graph, which suggests the signal path to the nodes. Aside from the input node, each node features a nonlinear activation function. CNN stands for "convolution neural network." It is an advanced neural network to classify emotions. CNN automatically detects the important features without any human intervention. It uses pooling and convolution operations, and it's computationally efficient. It contains an input layer, a convolution layer, and an output

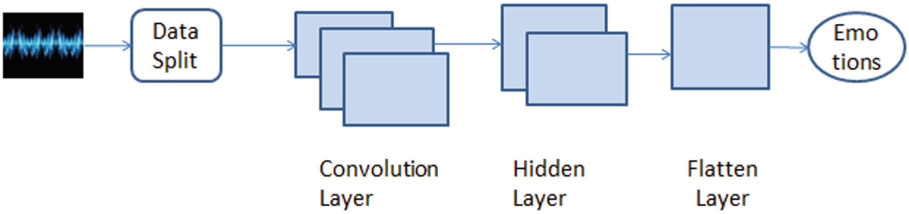

Fig. 3. CNN Architecture

Table 2. Performance evaluation on Testing Data

Algorithm	Accuracy	F1 Score	Precision	Recall
SVM	0.78	0.79	0.75	0.78
Decision Tree	0.78	0.79	0.80	0.84
Random Forest	0.85	0.88	0.87	0.89
MLP	0.81	0.82	0.83	0.82
CNN	0.82	0.84	0.82	0.84

layer. The hidden layer contains a convolution layer, a max pooling layer, and a flattening layer with dropout. This is depicted in Fig. 3 below

4 Results and Discussion

The experimentation evaluations are analyzed with five different models such as SVM, Random Forest, Decision Tree, MLP, and CNN. Out of 7356 samples, the data set has been divided into 70:30 (70% for training and 30% for testing), and then the dataset is divided into 80:20 and experimented for 50 epochs. Once the training data is modeled, the approach uses the confusion matrix as a performance metric to evaluate the performance of the algorithms used. The confusion matrix considers True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) values for evaluation. Using the confusion matrix, classification accuracy, precision, recall, and F1-Score are evaluated for both the classifiers. Classification Accuracy is evaluated for each of the models in which true labels and false labels are verified for correct classification. The best results were evaluated using five sets of features. The Table 2 depicts the performance measures of the proposed approach with respect to the SVM, Random Forest, Decision Tree, MLP, and CNN classifiers. The best results are obtained with the Random Forest model with 100 trees. The Random Forest achieved an accuracy of 0.83.

Thus, it is noticed that the performance of the machine learning approach in Emotion recognition based on audio signal has gained better results.

5 Conclusion and Future Work

In this paper, the proposed approach mainly focuses on emotion classification based on audio signals. The proposed system uses SVM, Decision Tree, Random Forest, MLP, and CNN models to identify the emotions based on the extracted 5 signals from an audio signal. The SVM model makes use of features to produce an accurate accuracy of 78.57%, the decision tree with 78.56%, the random forest with 85.71%, the MLP with 81.82%, and the 82.98% for CNN, for emotions of happy, sad, and neutral. Future work will be focused on compound emotions such as happily surprised, happily disgusted, sadly fearful, sadly angry, sadly surprised, sadly disgusted, and angrily fearful with the comparison of different algorithms.

References

1. Tzirakis, P., Zhang, J., & Schuller, B. W. (2018). End-to-end speech emotion recognition using a deep convolutional recurrent network. ICASSP.
2. Zhao, J., Mao, X., & Chen, L. (2018). Learning deep features to recognise speech emotion using merged deep CNN. IET Signal Processing, 12(6), 713-721..
3. Tarunika, K., Pradeeba, R. B., & Aruna, P. (2018, July). Applying machine learning techniques for speech emotion recognition. In 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1–5). IEEE.
4. Jain, M., Narayan, S., Balaji, P., Bhowmick, A., & Muthu, R. K. (2020). Speech emotion recognition using support vector machine. arXiv preprint arXiv:2002.07590.
5. Meng, H., Yan, T., Yuan, F., & Wei, H. (2019). Speech emotion recognition from 3D log-mel spectrograms with deep learning network. IEEE access, 7, 125868-125881
6. 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.
7. Aouani, H., & Ayed, Y. B. (2018, March). Emotion recognition in speech using MFCC with SVM, DSVM and auto-encoder. In 2018 4th International conference on advanced technologies for signal and image processing (ATSIP) (pp. 1–5). IEEE.
8. Mirsamadi, S., Barsoum, E., & Zhang, C. (2017, March). Automatic speech emotion recognition using recurrent neural networks with local attention. In 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP) (pp. 2227–2231). IEEE.
9. Aouani, H., & Ayed, Y. B. (2020). Speech emotion recognition with deep learning. Procedia Computer Science, 176, 251-260.
10. Sajjad, M., & Kwon, S. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE Access, 8, 79861-79875.