Speech Emotion Recognition Using Convolutional Neural Networks on Spectrograms and Mel-frequency Cepstral Coefficients Images

Sambhavi Mukherjee, Shikha Mundra, and Ankit Mundra

Abstract A Speech Emotion Recognition (SER) system is a collection of methods for processing and classifying voice inputs to recognize emotions. This type of system could be beneficial in several sectors, including interactive voice-based assistants and caller-agent conversation analysis. We want to reveal underlying emotions in the recorded speech by analyzing the acoustic features of audio data. The majority of Emotion Recognition research has concentrated on the use of speech descriptors such as mel-frequency cepstral coefficients (MFCC), Linear Prediction Coefficient (LPC), energy, spectral flux, spectral centroid, spectral roll-off, and zero-crossing rate, followed by the application of machine learning classifiers such as SVM, Nave Bayes, and others, or an ensemble of a few such classifiers. In other research papers, the speech recognition problem was turned into an image recognition problem, and then convolutional neural network (CNN) architectures were used, only evaluating MFCC images of audio signals. In our technique, we gathered spectrogram images from audio samples to train our CNN architecture. Spectrograms are graphical representations of the signal strength, or 'loudness,' of a signal across time at various frequencies contained in a waveform. We also compared the results with the CNN model applied to this dataset's MFCC images. When compared to our spectrogram CNN model, the MFCC image CNN model improved by 3.75% (accuracy 82.5%). https://github.com/sambhavi10/Speech-Emotion-Recognition.

Keywords Emotion recognition · Mel-frequency cepstral coefficients (MFCC) · Spectrogram

S. Mukherjee · S. Mundra (✉)
Department of Computer Science and Engineering, Manipal University Jaipur, Jaipur, India
e-mail: a.shikha1990@gmail.com

A. Mundra
Department of Information Technology, Manipal University Jaipur, Jaipur, India

M. S. Kaiser et al. (eds.), *Information and Communication Technology for Competitive Strategies (ICTCS 2022)*, Lecture Notes in Networks and Systems 615,
https://doi.org/10.1007/978-981-19-9304-6_4

1 Introduction

One of the most natural ways for humans to express themselves is through speech. We rely on it so much that we notice its use in other channels of communication, such as emails and text messages, where we commonly use emoticons to express our emotions. Emotion detection and analysis are vital in the digital age of distant communication since emotions are so important in communication. Emotions are difficult to discern since they are subjective. There is no universally accepted method for quantifying or categorizing them.

Humans are emotional beings, and they express themselves using speech. The same sentence can have different meanings when spoken with different tones. Speeches can be sarcastic, which have meanings that contradict the linguistic meaning. If our computers were only trained on simple Natural Language, they would often misinterpret what's being said [1]. Thus, it becomes crucial to understand the emotional intent of the speaker, along with the content. Emotion detection has wide-ranging applications like teaching, security, medicine, and entertainment. It could be integrated with conversational AI such as Alexa or Siri, for the AI agent to be able to identify actual human sentiments and emotions [2]. This will prove to be a huge step in making our computers more 'human-like.' In this work, we have focused on popular signal features known as mel-frequency cepstral coefficients (MFCC).

Mel-frequency cepstral coefficients (MFCC) is a well-known feature for speech signals. The several applications of it in speech processing are speech recognition, speaker recognition, speech synthesis, speech coding, etc. The research paper [3] has used this feature, and it is described as a limited group of features (often 10–20) that represent the general shape of a spectral envelope clearly [4].

2 Related Work

In the modern world, voice recognition functions must be done by robots just as naturally as they are by humans. As a result, a significant portion of research has been conducted where the goal of SER was re-defined as an image classification problem and then accomplished using a pre-trained model [5]. We have read research where notable features were retrieved from voice data using MFCC. In addition to spectral (roll-off, flux, centroid, bandwidth), energy (root-mean-square energy), raw signal (zero-crossing rate), pitch (fundamental frequency), and chroma features, papers [5, 6] discuss the usage of MFCC features.

Several freely available speech datasets were used in the URDU dataset, including SAVEE, EMODB, and EMONO [5]. The EMODB dataset with seven emotion classes was utilized in the paper [7]. The RAVDESS dataset is used in papers [6, 8]. Reference [6] made use of the TESS dataset.

Researchers used both machine learning techniques, their ensembles, and deep learning models such as CNN models, semi-CNN models, and transfer learning models in their categorization method.

In [5] a comparison of decision tree (J48), random forest (RF), and sequential minimal optimization (SMO), machine learning techniques were presented, as well as an ensemble of these machine learning algorithms using majority voting. Paper [7] used MATLAB 2019a programming software and an HP Z440 Workstation with an Intel Xeon CPU, 2.1 GHz, and 128 GB RAM to deploy a transfer learning model called AlexNet to perform the task of SER. The researchers demonstrated the use of autoencoders for dimensionality reduction, followed by Support Vector Machines (SVM), decision tree classifiers, and convolutional neural networks (CNN), AlexNet, and ResNet50. The use of a deep transfer learning model to train and recognize emotions was demonstrated in the paper [8].

3 Proposed Methodology

To complete the SER goal, the focus of this paper is on merging the original speech attributes and employing images generated from speech signals.

Our model's design is mostly composed of two modules:

(1) SER utilizing audio features experimented using machine learning models.
(2) CNN models based on spectrogram and MFCC images.

We have converted the MFCC signal to images and used CNN to extract relevant features from MFCC images and spectrogram images. In addition, to compare its performance we have used audio features and classification using several machine learning models, including logistic regression, Random Forests, Naïve Bayes, and Support Vector Machine, which have been contrasted hhhh to one another.

Using CNN to analyze MFCC and spectrogram images produced accuracy levels of 82.5% and 86.25%, respectively.

Our main work has been focused on the following points:

- Using the Librosa package [9], extract feature vectors from audio (.wav) files. Applying feature extraction approaches to audio data namely extracting MFCC features from audio signals and extracting features from images. Then, to perform emotion classification on them, employ supervised machine learning algorithms and convolutional neural network architectures.
- To extract the spectrograms from the audio recordings and feed them into a convolutional neural network architecture to forecast the right output classes for the test dataset.
- To extract MFCC features from audio files and use them with a CNN architecture to predict the labels for test images.

- Compare the performance of the proposed MFCC image and spectrogram feature extraction method with existing audio features in terms of AUC score and accuracy.

4 Data Preprocessing

The initial step is to pre-process the audio files so that they may be used with machine learning methods. Librosa, a Python speech recognition package, was used to read the audio files for 2.5 s using a resample type of Kaiser fast, a sampling rate of 44,100 Hz, and an offset of 0.5 s. Following that, the 'feature. mfcc' technique of Librosa was used to transform the signals into feature vectors of 216 dimensions by employing the sample rates and time series collected from the signals. Mel-frequency cepstral coefficients (MFCC) are among the most commonly used speech and emotion identification features.

We experimented with numerous divides for the train and test datasets, and 80:20 proved to be the best possible split for this dataset. After doing a train–test split, our dataset is ready to be fed to machine learning algorithms, which will provide predictions. The spectrogram images for CNN architectures are extracted using the Open-Soundscape library's spectrogram module. Open-Soundscape [10] is a utility library for analyzing bioacoustic data. This produced graphics of 224 by 224 pixels for the audio files. The MFCC pictures were extracted using the Librosa Library and a 2-s speech signal for each audio file. Our CNN architectures may now use these images for image categorization.

5 Classifiers

For extraction and classification of the most relevant feature, we have experimented with machine learning and deep learning methods as described in below sections.

5.1 Machine Learning Classifiers

Our dataset is known as the Urdu dataset, and it comprises four output emotion types. In such supervised learning situations, machine learning algorithms are frequently extremely useful. Supervised learning is a sort of machine learning in which machines are trained with well-labeled training data and then predict the output [5]. Labeled data shows that some input data has already been assigned an output. As a result, our problem is a multi-class classification problem that has been solved with classifiers

such as logistic regression, Support Vector classifier, Random Forest classifier, and Nave Bayes Classifier [11]. Our dataset is known as the Urdu dataset, and it comprises four output emotion types as explained in Sect. 6.

5.2 Deep Learning Classifiers (CNN)

CNN is a neural network-based architecture popular in image categorization. This is critical for the task of SER using photos. It is useful for feature extraction and classification since it can pass values to the next layer while preserving spatial information and may be used in noisy images. The overall architecture of CNN is comprised of several layers that function as input, hidden, and output layers. The hidden layers are composed of feature maps, a fully connected layer containing convolutional neural networks, and pooling layers. The convolutional layer and the pooling layer collect essential properties from the input data, and the extracted value is mapped to the feature map [12, 13]. In this process, the characteristics of the MFCC and spectrogram images can be extracted, and then the fully connected layer focuses on the features extracted to perform classification.

6 Dataset Description

For our SER job, we used publicly available URDU data [14]. The URDU dataset is made up of emotional utterances from Urdu talk shows. It has 400 phrases that depict four fundamental emotions: angry, pleased, neutral, and emotional. There is a total of 38 speakers (27 male and 11 female). This information was collected from YouTube content. Speakers are selected at random. The nomenclature used to label the files in the dataset includes information on the speaker, gender, file number for that speaker, and overall numbering of the file in a certain emotion. The files have been renamed so that the first letter indicates the emotion: *S* for Sad, *H* for Happy, *A* for Angry, and *N* for Neutral, followed by a number to represent the file order. The dataset is divided into four emotion categories: angry, happy, neutral, and sad. Each lesson included 100 audio files with the .wav extension. We experimented with other splits for training the models, such as 75:25 and 80:20. The latter proved to be the preferable option, so we chose 80:20 for training and testing. As a result, the training dataset contained 320 speech files, and the remaining 80 speech files were used to test the performance of our models. Our dataset is also available on GitHub. (Link: https://github.com/siddiquelatif/URDU-Dataset).

Table 1 CNN architecture for the spectrogram model and the MFCC model

Layer	Output shape
Rescaling	(None, 180, 180, 3)
Conv2D	(None, 178, 178, 32)
MaxPooling2D	(None, 89, 89, 32)
Conv2D	(None, 87, 87, 32)
MaxPooling2D	(None, 43, 43, 32)
Conv2D	(None, 41, 41, 32)
MaxPooling2D	(None, 20, 20, 32)
Flatten	(None, 12,800)
Dense	(None, 128)
Dense	(None, 4)

7 Experimental Setup and Hyperparameter

To facilitate comparability, the machine learning models chosen were trained on identical training and testing datasets. Sklearn, a popular Python machine learning toolkit [15], was used to create these models. The model achieved a test accuracy of 48.75% by employing the hyperparameters solver = 'saga,' penalty = '*l*2,' and max iter = 80 in logistic regression. The accuracy of SVM was 56.25%. Nave Bayes, on the other hand, provided an accuracy of 58.75%. The Random Forest model obtained an accuracy of 60% on the test dataset after tweaking the hyperparameters: *n* estimators = 120, criteria = 'entropy.'

Apart from machine learning algorithms, convolutional neural networks are also applied to the generated MFCC and spectrogram images. The CNN model using MFCC was able to achieve a test accuracy of 82.5%, and the CNN model using spectrogram images was able to achieve a test accuracy of 86.25%. Table 1 shows the CNN architecture used.

8 Experimental Results and Conclusion

The classification metric used for contrasting the models is accuracy, AUC score, and area under AUC-ROC curve. Table 2 shows the AUC scores, and it is observed that CNN using spectrogram, and MFCC image has performed significantly better than traditional audio features. We have also plotted the AUC-ROC curve for the ML model with audio features and CNN with MFCC image and spectrogram image as shown in Fig. 1.

Table 3 depicts a comparative examination of the models based on accuracy. From Table 3, it has been observed that logistic regression with MFCC has performed lowest and CNN with spectrogram has performed best among all.

Table 2 AUC scores in tabular form

Model	AUC score
Logistic regression	0.7229
Random forest	0.8183
SVM	0.7821
Naïve Bayes	0.7558
CNN using MFCC	0.8416
CNN using spectrogram	**0.8499**

Best score is highlighted in bold

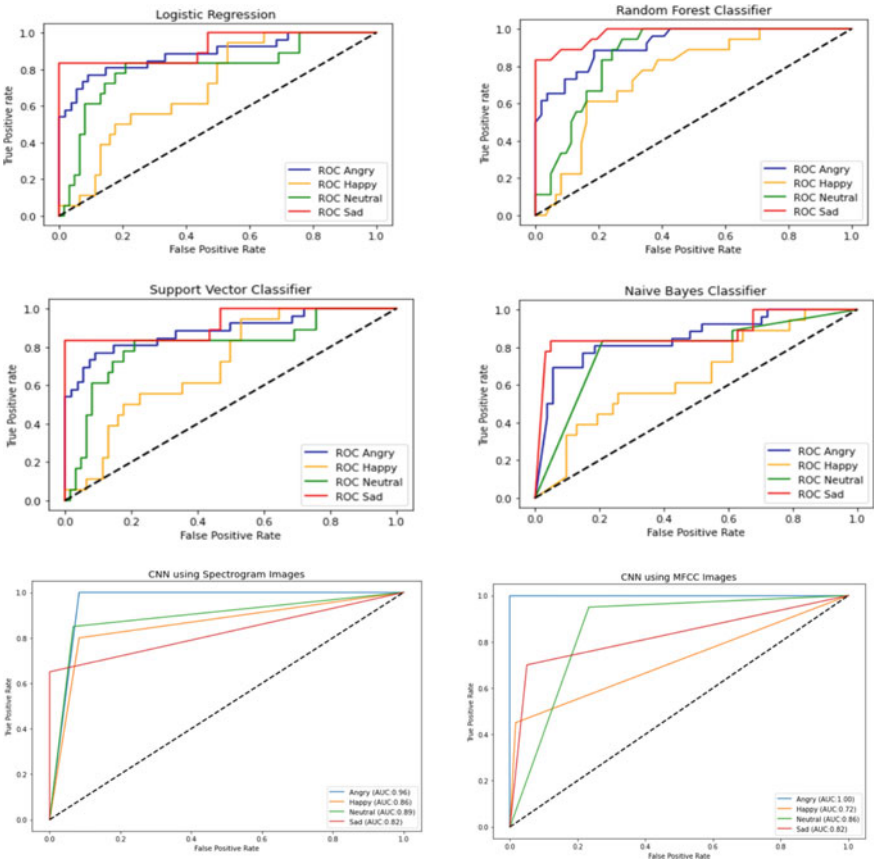

Fig. 1 AUC-ROC curves for all models

Table 3 Comparative performance in terms of accuracy (%)

Models	Test accuracy
Logistic regression	48.75
Random forest	60
SVM	56.25
Naïve Bayes	58.75
CNN using MFCC	82.5
CNN using spectrogram	**86.25**

Best score is highlighted in bold

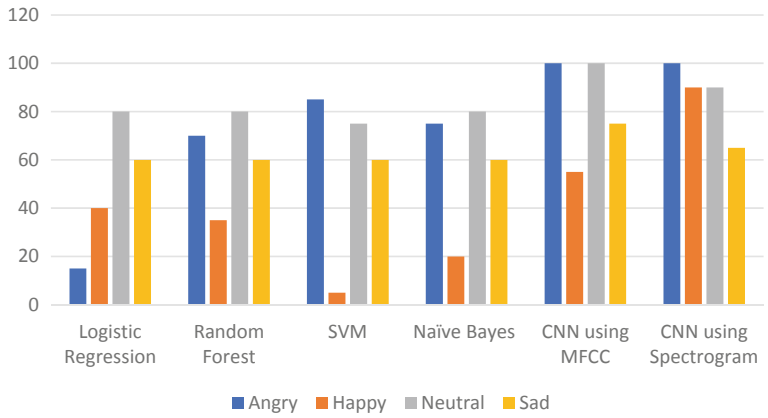

Fig. 2 Class-wise accuracies represented as bar plots

This research aims to compare machine learning and deep learning models in executing the SER task. As shown in Fig. 2, we made comparisons between various machine learning and deep learning models. Logistic regression (48.75%), Naïve Bayes (58.75%), SVM (56.25%), and Random Forests have the highest accuracies (60%). Following that, we used CNN on MFCC images to achieve an accuracy of 82.5% and on spectrogram images to achieve an accuracy of 86.25%. Here, we observed that image-based features can play a crucial role in the extraction of emotion from the speech signal. In the future, spectrogram image features can be combined with text-based features [16] to enhance the performance and improve the robustness of the model.

References

1. https://www.techtarget.com/searchenterpriseai/feature/How-emotion-analytics-will-impact-the-future-of-NLP
2. https://blog.cfte.education/conversational-ai-examples-how-siri-alexa-google-assistant-have-

human-like-conversations/
3. Tiwari V (2010) MFCC and its applications in speaker recognition. Int J Emerg Technol 1(1):19–22
4. https://musicinformationretrieval.com/mfcc.html
5. Zehra W, Javed AR, Jalil Z, Khan HU, Gadekallu TR (2021) Cross corpus multi-lingual speech emotion recognition using ensemble learning. Complex Intell Syst 7(4):1845–1854
6. Patel N, Patel S, Mankad SH (2022) Impact of autoencoder based compact representation on emotion detection from audio. J Ambient Intell Humaniz Comput 13(2):867–885
7. Lech M, Stolar M, Best C, Bolia R (2020) Real-time speech emotion recognition using a pre-trained image classification network: effects of bandwidth reduction and companding. Front Comput Sci 2:14
8. Togootogtokh E, Klasen C (2021) DeepEMO: deep learning for speech emotion recognition. arXiv preprint arXiv:2109.04081
9. McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E, Nieto O (2015) librosa: Audio and music signal analysis in python. In: Proceedings of the 14th python in science conference, pp 18-25
10. Darras K, Pérez N, Mauladi, Hanf-Dressler T (2020) BioSounds: an open-source, online platform for ecoacoustics. F1000Research 9:1224. https://doi.org/10.12688/f1000research.26369.1
11. Mundra S, Dhingra A, Kapur A, Joshi D (2019) Prediction of a movie's success using data mining techniques. In: Satapathy S, Joshi A (eds) Information and communication technology for intelligent systems. Smart innovation, systems and technologies, vol 106. Springer, Singapore. https://doi.org/10.1007/978-981-13-1742-2_22
12. Mundra S, Mundra A, Saigal A, Gupta P (2020) Text document representation and classification using convolution neural network. In: 2020 Sixth international conference on parallel, distributed and grid computing (PDGC), pp 202–205. https://doi.org/10.1109/PDGC50313.2020.9315752
13. Mundra S, Mittal N (2021) Evaluation of text representation method to detect cyber aggression in Hindi English code mixed social media text. In: 2021 Thirteenth international conference on contemporary computing (IC3-2021) (IC3 '21). Association for Computing Machinery, New York, pp 402–409. https://doi.org/10.1145/3474124.3474185
14. Latif S, Qayyum A, Usman M, Qadir J (2018) Cross lingual speech emotion recognition: Urdu versus western languages. In: 2018 International conference on frontiers of information technology (FIT). IEEE, pp 88–93
15. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, Varoquaux G et al. (2013) API design for machine learning software: experiences from the scikit-learn project. arXiv preprint arXiv:1309.0238
16. Mundra S, Mittal N (2022) FA-Net: fused attention-based network for Hindi English code-mixed offensive text classification. Soc Netw Anal Min 12(1):1–14