Emotion-Based Music Recommendation

Akshata Kumble, Sriveda Medatati and Ashwini B

Dept of Electronics and Communication, PES University, Bangalore, India.

Contributing authors: akshatapraveenkumble@pesu.pes.edu; srivedamedatati@pesu.pes.edu; ashwinib@pes.edu;

Abstract

Music-related activities, whether both active and passive, activate various brain regions, boosting well-being and enhancing the quality of life. Through the utilization of rhythms and audio frequency, music therapy is a powerful tool for treating physical problems. According to research, listening to music stimulates the brain's linguistic, motor, and cognitive regions. An emotion-based music recommendation system is proposed to lift up the spirits of the individual with utmost ease. Our goal is to develop a deep learning model that uses neural network models to recommend certain songs based on the user's mood. In order to accomplish this, it is important to classify the speech based on the emotion. This paper presents a comparative analysis of emotion based speech recognition between four models.

Keywords: Deep Learning (DL), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), K Nearest Neighbour (KNN), Mel Frequency Cepstral Coeffecients (MFCC)

1 Introduction

The practice of examining audio recordings is known as audio classification or sound classification. In contrast to image processing and other classification methods, audio processing is one of the most challenging data science projects. One such application, emotion-based categorization, seeks to group audio recordings into different groups according to their emotional content. Humans are capable of expressing multiple emotions, these are mainly characterized into 4 types: Anger, Fear, Happy and Sad. Music is a form of art that enhances one's mood, calms and relaxes them. The project's goal is to develop a DL model which suggests music depending on one's feelings by recognizing emotions depending on pitch and other mfcc variables. The implementation deals with a multiclass classification problem with a dataset consisting of audio files (in .wav format) with various emotions such as anger, fear, sadness, surprise, joy and neutral. Our data set comprises 200 audio files of each emotion. Each file is around 3 seconds long and pre-processed and the labels are stored in a .json file. Four models were used for audio categorization using (i) basic artificial neural network (ii) convolutional neural network (iii) LSTM (long short term memory) based recurrent neural network and (iv) K nearest neighbors network. Based on the emotions predicted by the models the user is redirected to a youtube page recommending specific music.

2 Related Works

A variety of input formats, machine learning methods, song libraries, and Neural Network (NN) types have all been deployed with varying rates of success. A few of the ML models implemented by the authors in these papers classify music into its many musical genres and also incorporates the analysis of music categorization. The study uses a series of spectrograms generated from temporal slices of music that serve as inputs into the NN. [1–3].

The system model implemented in this paper has been trained and classified using a DL technique. Here, training and classification employ CNNs. The most important step in audio analysis is the feature extraction phase. With audio samples, the MFCCs are employed as a feature vector. The suggested technique uses feature vector extraction to categorize music into different genres. Results indicate that this method has an accuracy level of around 76%, which will substantially enhance and ease the automatic classification of musical genres [2, 12].

The feature sets can be categorised as dynamic, spectral, rhythmic, and harmonious. The central moments of each feature up to the fourth order, together with five other statistical characteristics, are taken into account as indicators of the features. In the end, the MRMR algorithm regulates a significant portion of representative attributes.
Expressing from a state of fear, rage, or excitement has a higher, broader range of pitch than speaking from a place of low pitch. Audio files may be used to identify emotions since they contain a variety of characteristics. The features include Tonnetz, MFCC, Mel, and Chroma. MFCC and PITCH can also be used to classify emotions based on their associated vocal speech signals. Here, they used the SVM Non-Linear Classifier and SVM Linear Classifier for classification [10]. The speech underwent undergo framing, and was then passed through Hamming window, after which fft was perform and the MFCC features were extracted. Mean value of MFCC and standard deviavation was

found, then thresholding was applied to distinguish between the emotions.

Various emotion recognition approaches that employ linguistic elements are based on a speech emotion recognition study that combines both acoustic and linguistic information. The acoustic characteristics of emotional speech change substantially depending on the kind and strength of the emotion. This research introduces a new emotional speech recognition approach that combines acoustic model and language model adaption to achieve good recognition performance on an emotional speech problem. The system's word recognition accuracy was 82.2%, and recognition mistakes were found. Regardless, the authors show that the combination of lexical and auditory elements is effective for emotion recognition [8, 10, 13, 15].

3 Methodology

The first step involves pre-processing the raw audio files into suitable waveforms. Then we transform the waveforms into Mel spectrograms and extract the features. In order to extract features, language input must be recognized and noise must be avoided. Three categories — high, mid and low-level characteristics are used to categorise audio information. 1) High-level elements like chords, rhythms, and melodies are connected to musical lyrics. 2) Pitch-like fluctuation patterns, MFCCs and beats level properties are examples of mid-level features. 3. Statistical measurements that are collected from sounds while extracting features includes energy and a zero-crossing rates.

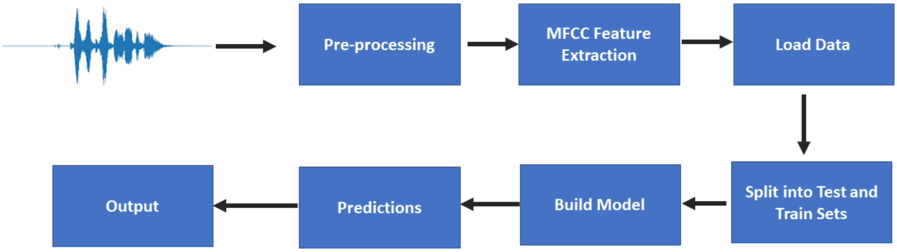
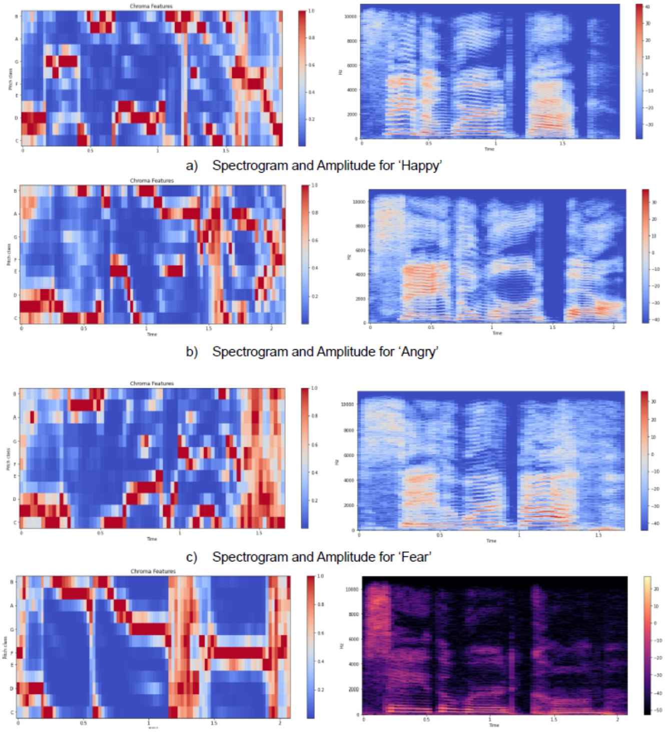


Fig. 1 Steps involved in the emotion-based speech recognition process.

So, to create these features, we employ a series of computations that are combined under a single name as MFCC, which aids in the extraction of mid-level and low-level audio characteristics. MFCC features are obtained by dividing the audio files into frames and then identifying and extracting different frequencies from each frame. These extracted features are stored in a json file which are later used for mapping during the training process.

We have created 4 models to perform a comparative analysis. These extracted features are fed into each of the models where the performance is measured. The basic steps involved in creating all the models are: a) Load



a) Spectrogram and Amplitude for 'Happy'

b) Spectrogram and Amplitude for 'Angry'

c) Spectrogram and Amplitude for 'Fear'

d) Spectrogram and Amplitude for 'Sad'

Fig. 2 Spectrograms and Amplitude plots for sample emotion

data, b) Training and Testing sets split, c) Build network architecture, d) Train network and e) Test cases We look into the methodology and architecture of each of these methods one by one.

3.1 Convolutional Neural Networks (CCN)

CNNs are ideal for audio classification as they can learn translation-invariant patterns with spatial hierarchies. A CCN is often comprised of several convolution and sub-sampling layers, followed by one or more fully connected layers at the end to aid with output prediction. Initially, a database of audio recordings depicting distinct emotions based on pitch was created, and feature extraction was conducted using the MFCC method. This sequential model is composed of five layers: an input layer, three hidden levels using the ReLu activation function, and an output layer utilising the Softmax layer. The ReLu activation function is applied to train the network faster, while the Softmax layer is employed to normalise the output and assist prediction.

Because CNNs outperform multilayer perceptrons (MLPs) despite having fewer parameters, we consider CNNs by perceiving audio files in an image format, i.e., a matrix of pixels, from which we extract features as horizontal and vertical bars, sense all the different components, and extract features using

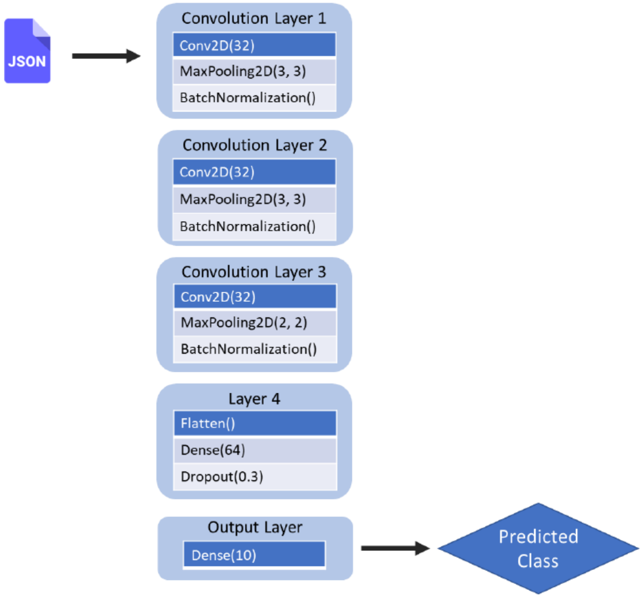


Fig. 3 Layers present in the CNN model

methods such as convolution and pooling. In the convolution procedure, a kernel (filter) associated with learning parameters (weights, bias) is applied to the input image, and an output image is created based on the shape and values of the applied filter in order to preserve certain information from a picture and discard the rest.

Considering that **x** is the original input with **n** elements, the convolution is performed as:

$$\mathbf{y} = \mathbf{x} * \mathbf{w} \rightarrow \mathbf{y}[\mathbf{i}] = \sum_{k=-\infty}^{\infty} \mathbf{x}[\mathbf{i} - \mathbf{k}]\mathbf{w}[\mathbf{k}], \quad (1)$$

While considering a padded input vector x^p with a vector size of $\mathbf{n} + \mathbf{2p}$ and a filter of m elements, the above equation is modified as:

$$\mathbf{y} = \mathbf{x} * \mathbf{w} \rightarrow \mathbf{y}[\mathbf{i}] = \sum_{k=0}^{m-1} \mathbf{x}^{\mathbf{P}}[\mathbf{i} + \mathbf{m} - \mathbf{k}]\mathbf{w}[\mathbf{k}], \quad (2)$$

Clusters of pixels are examined in the pooling procedure in order to execute an aggregation over them. Choosing the maximum value of the pixels in the cluster is one of the potential aggregations (known as Max Pooling). A 2-D max pooling layer downsamples the input by splitting it into rectangular pooling zones and then calculates the maximum of each region. Computing the average is another standard aggregate (known as Average Pooling). This approach decreases the amount of information in the audio while retaining and amplifying the valuable characteristics revealed by the filters while convolving the audio. We acquired a precision of 93.53% by using this technique.

3.2 LSTM-based Neural Networks

Recurrent neural networks identify the sequential properties of input and utilise patterns to forecast the next likely situation. The input is a sequence vector, and the output is a single vector. As a result, an LSTM model is implemented since it contains feedback connections and is capable of analysing not just single data points but also whole sequences of data. LSTMs assign data weights to help RNNs decide whether to accept new information, forget it, or give it enough weight to affect the output. The test accuracy is acheived is nearly 78%.

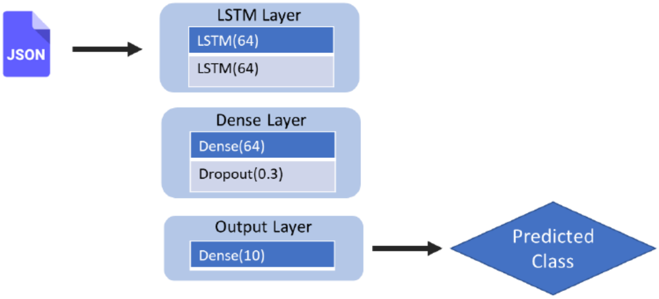


Fig. 4 Layers present in the RNN-LSTM model

3.3 K- Nearest Neighbours Network (K-NN)

Popularly known as the "lazy learner algorithm," KNN is a machine learning technique that is used for classification and regression. It simply applies a distance-based algorithm to determine the K number of identical neighbours to new data and the class to which the majority of neighbours belong, resulting in that class being returned as an output. First, we created a function that would receive training data, current instances, and the number of neighbours required. We calculated the distance between each point in the training data and every other point, then located the K nearest neighbours and returned all neighbours. We developed a dictionary that holds the class and the number of neighbours it has. After producing the frequency map, we sort it in decreasing order depending on the number of neighbours and return the first class. We then extracted the MFCC features and trained the model, using the numpy library to determine the distance between two locations. The data is then fed into the KNN algorithm, which makes predictions on the testing dataset, acheiving a remarkable accuracy of 99.25%.

3.4 Multi-Layer Perceptron Network (MLP)

Speech recognition with a multi-layer perceptron for emotion-based classification is used in this technique. A multilayer perceptron is created by connecting many single-layer perceptrons and analyzing their interactions parallelly. A perceptron consists of four basic components: the input value or input layer, the weight, the net summation, and the activation function. This sequential model allows us to develop models in a multilayer perceptron layer by layer, however it only works for stacks of layers with a single input and output.

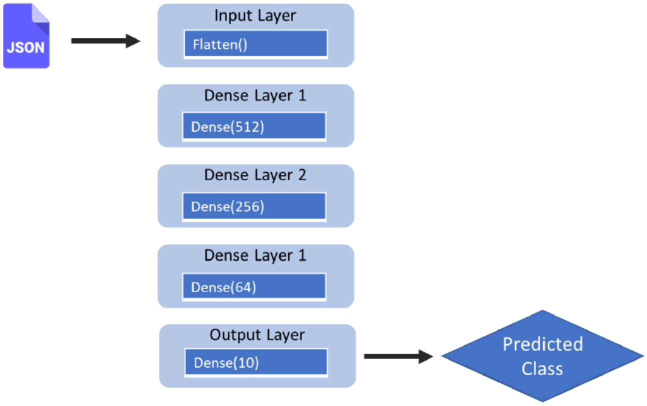


Fig. 5 Layers present in the Multi-Layer Perceptron model

The 'Flatten' function flattens the input without modifying the batch size. The activation stage employs the ReLu and Softmax activation functions. The first two thick layers are concealed and serve as the foundation for a fully linked model. The final thick layer, the output layer, has 10 neurons that determine which class the audio sample belongs to.

4 Results

Three deep learning methods and one machine learning model have been implemented for emotion based speech classification. A comparative analysis was performed where the results as depicted in the table were achieved. While keeping the batch size constant and varying the other training parameters, the observations show that the Multi Layer Perceptron Network and Convolutional Neural Network obtained comparable results and gave better accuracy over the Recurrent Neural Network. The models were also tested on real time audio files and it was observed that CNN works well with real time speech. Based on the emotion detected specific songs are recommended to lift the user's spirits by redirecting the user to a youtube page containing suitable music.

	CNN	RNN-LSTM	MLP
Accuracy	0.948	0.7800	0.9353
Loss	0.1496	0.5068	0.1395
Learning Rate	0.0001	0.0001	0.0001
Batch Size	32	32	32
Epochs	30	30	50
Layers	5	3	5

Table 1 Comparative analysis of the deep learning models using the parameters used for the simulation study.

5 Conclusion

Speech contains rich emotional factors and music therapy has been proven effective over the years. Emotion recognition based on speech is a fascinating subject to explore in the future and can be used to improve one's well-being. We have successfully performed a comparative study using 4 models and propose that CNN produces efficient results with an accuracy of almost 95%. This framework along with the others can further be extended to provide multilingual emotion and it's efficiency can be tested in noisy environments.

References

[1] N. Pelchat and C. M. Gelowitz, "Neural network music genre classification," Canadian Journal of Electrical and Computer Engineering, vol. 43,no. 3, pp. 170–173, 2020.

[2] S. Vishnupriya and K. Meenakshi, "Automatic music genre classification using convolution neural network," in 2018 International Conference on Computer Communication and Informatics (ICCCI), 2018, pp. 1–4.

[3] G. Deshmukh, A. Gaonkar, G. Golwalkar, and S. Kulkarni, "Speech based emotion recognition using machine learning," in 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 812–817.

[4] L. K. Puppala, S. S. R. Muvva, S. R. Chinige, and P. Rajendran, "A novel music genre classification using convolutional neural network," in 2021 6th International Conference on Communication and Electronics Systems (ICCES), 2021, pp. 1246–1249.

[5] A. Elbir, H. Bilal C¸ am, M. Emre Iyican, B. Ozt urk, and N. Aydin, "Music genre classification and recommendation by using machine learning techniques," in 2018 Innovations in Intelligent Systems and Applications Conference (ASYU), 2018, pp. 1–5.

[6] S. R. Siadat, I. M. Voronkov and A. A. Kharlamov, "Emotion recognition from Persian speech with 1D Convolution neural network," 2022 Fourth

International Conference Neurotechnologies and Neurointerfaces (CNN), Kaliningrad, Russian Federation, 2022, pp. 152-157.

[7] A. Yazdani, H. Simchi and Y. Shekofteh, "Emotion Recognition In Persian Speech Using Deep Neural Networks," 2021 11th International Conference on Computer Engineering and Knowledge (ICCKE), Mashhad, Iran, Islamic Republic of, 2021, pp. 374-378.

[8] Q. Yang, F. Xu, Z. Ling, X. Li, Y. Li and D. Fang, "Selecting and Analyzing Speech Features for the Screening of Mild Cognitive Impairment," 2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Mexico, 2021, pp. 1906-1910.

[9] K. V. Krishna, N. Sainath and A. M. Posonia, "Speech Emotion Recognition using Machine Learning," 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2022, pp. 1014-1018, doi: 10.1109/ICCMC53470.2022.9753976.

[10] M. Sakurai and T. Kosaka, "Emotion Recognition Combining Acoustic and Linguistic Features Based on Speech Recognition Results," 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), Kyoto, Japan, 2021, pp. 824-827.

[11] C. Jie, "Speech emotion recognition based on convolutional neural network," 2021 International Conference on Networking, Communications and Information Technology (NetCIT), Manchester, United Kingdom, 2021, pp. 106-109.

[12] L. Cai, J. Dong and M. Wei, "Multi-Modal Emotion Recognition From Speech and Facial Expression Based on Deep Learning," 2020 Chinese Automation Congress (CAC), Shanghai, China, 2020, pp. 5726-5729.

[13] Ainurrochman, I. I. Febriansyah and U. L. Yuhana, "SER: Speech Emotion Recognition Application Based on Extreme Learning Machine," 2021 13th International Conference on Information and Communication Technology and System (ICTS), Surabaya, Indonesia, 2021, pp. 179-183.

[14] J. Wang and Z. Han, "Research on Speech Emotion Recognition Technology based on Deep and Shallow Neural Network," 2019 Chinese Control Conference (CCC), Guangzhou, China, 2019, pp. 3555-3558.

[15] G. Assunção, P. Menezes and F. Perdigão, "Importance of speaker specific speech features for emotion recognition," 2019 5th Experiment International Conference (exp.at'19), Funchal, Portugal, 2019, pp. 266-267, doi: 10.1109/EXPAT.2019.8876534.

[16] J. Wang and Z. Han, ”Research on Speech Emotion Recognition Technology based on Deep and Shallow Neural Network,” 2019 Chinese Control Conference (CCC), Guangzhou, China, 2019, pp. 3555-3558.

[17] Ainurrochman, I. I. Febriansyah and U. L. Yuhana, ”SER: Speech Emotion Recognition Application Based on Extreme Learning Machine,” 2021 13th International Conference on Information and Communication Technology and System (ICTS), Surabaya, Indonesia, 2021, pp. 179-183.