# Automatic Emotion Recognition System using tinyML

John Tharian
*UG Student, Dept. of Electronics*
*Model Engineering College, Kochi*
Ernakulam, India
johntharian247@gmail.com

Nandakrishnan R
*UG Student, Dept. of Electronics*
*Model Engineering College, Kochi*
Cherthala, India
rnandakrishnan2001@gmail.com

Saurav Sajesh
*UG Student, Dept. of Electronics*
*Model Engineering College, Kochi*
Kozhikode, India
sauravnsajesh@gmail.com

Arun A V
*Research Scholar, Model Engineering College*
*APJ Abdul Kalam Technological University*
Trivandrum, India
arunav.aav@gmail.com

Jayadas C K
*Associate Professor, Model Engineering College*
*APJ Abdul Kalam Technological University*
Ernakulam, India
jck@mec.ac.in

***Abstract*—This paper proposes a system to detect and analyze emotions using tinyml. Automated emotion detection has attracted increasing interest from researchers in neuroscience, psychology, computer science, and associated disciplines for the past 3 decades. In this paper, an Automatic Emotion Recognition System implemented using tinyml (AERSUT) has been proposed. There are 3 levels/stages for the above-proposed method - 1) input data using tinyml, 2) feature extraction, 3) emotion recognition. The first section involves giving input raw speech data to analyze and detect through a tinyml board. Features like Mel Frequency Cepstral Coefficients(MFCC), root mean square energy, and zero crossing rate are extracted from this audio and the user's emotional state is detected. Here we classify emotions into 8 categories- surprise, neutral, disgust fear, sad, calm, happy, and anger. 2 models were trained using TensorFlow and the model with the best accuracy has been used for detection.**

***Index Terms*—Emotion Classification, Discrete Fourier Transform (DFT), Deep Learning, Tinyml**

,

## I. Introduction

Deep learning has been widely improved over the past years and this has contributed to significant development in the field of human-computer interaction. Although automatic emotion recognition has been widely growing, the difficulty in identifying the right emotion is still a challenging task. The proposed work is to find emotion in English speech. Audio samples are obtained from different individuals. It is possible to extract features from audio signals that can help identify the type of emotion the user is feeling. Due to its importance in human-machine interactions, speech-emotion recognition has recently been a popular research area in signal processing, pattern recognition, and artificial intelligence [1] - [2]. Loudness, pitch, and mel frequency coefficients are important features that are extracted from the audio signals. Deep neural networks have been used to determine the user's emotional state accurately. Deep Neural Network is one of the most widely used deep learning algorithms, with excellent results in extracting discriminative features. Deep learning methods can be used to automatically extract emotional features with the advancement of deep learning technology and learn the correlation between features for speech emotion recognition when in comparison to conventional methods, it's been tested to be extra effective[3]-[4].

## II. Dataset

Ravdess and Crema-D are the two datasets used for this project. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)determine the user's emotional state accurately male), vocalizing two lexically-matched statements in a neutral North American accent. CREMA-D is a data set of original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 from various races and ethnicities. Both these datasets were combined to produce a collection of around 20000 original clips that portray one of these six different emotions Anger, Disgust, Fear, Happy, Neutral, and Sad. The dataset is further processed by adding noise. The audio signals are stretched and shifted to further increase the authenticity of the dataset. The audio processing is done using the librosa library.

## III. Feature extraction

It is necessary to extract good features for any machine learning model to work accurately. Here mel frequency cepstral coefficients, mel-spectrograms, zero crossing rate and root mean squared energy are extracted from the audio signals.

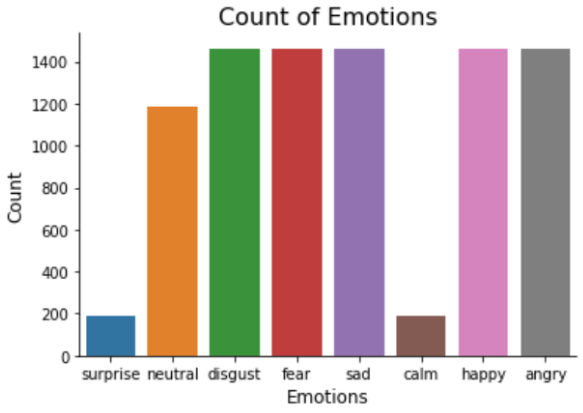


Fig. 1. Count of various emotions

### A. Mel Spectogran

A spectrogram is a visible depiction of speech's temporal and spectral changes in speech. 2D convolution filters may be used to extract 2D feature maps from any entry. Mel-spectrograms are vital attributes that are utilized for speech recognition, gender identification, and emotional states. A mel-spectrogram is a frequency-based picture created from a time-sensitive sound source, Time is addressed on the x-axis, and frequency is addressed on the y-axis. The signal strength is most elevated when the color is bright and least at the point when the color is dull. Such wealthy qualities can not be retrieved and utilized whilst speech is transformed into textual content and/or phonemes. Spectrograms, which include extra data not observed in the textual content by itself, are a useful resource in improving emotion recognition.

### B. Mel Frequency Cepstral Coeficients

Pre Emphasis is done to increase the magnitude of energy in high frequencies. Increasing the energy in high frequencies will help improve the accuracy of the model. For speech recognition, mel-frequency cepstral coefficients (MFCCs) are frequently used. The Frequency vs. Pitch size scale is denoted by the word Mel. Using the formula m = 2595 log10 (1 + (f/seven hundred)), the fee measured in frequency scale may be translated to Mel scale. The MFCCs are the spectrum's amplitudes. The MFCC feature extraction technique includes windowing the signal, applying the discrete Fourier transform, taking the magnitude log, and then warping the frequespeech's temporal and spectral changes inverse DCT.

### C. Zero cross rate

A very simple way for measuring the smoothness of a sign is to calculate the quantity of zero-crossing inside that signal. Zero cross rate is widely used in speech analysis and information retrieval. Zero. crossing rate is the rate at which a signal changes from positive to negative or back. It measures how many times the waveform crosses the zero axis. It can be perceived as a measure of the noisiness of the signal.

### D. Root Mean Squared Energy

RMS is a useful way of calculating the average of values over some time. With audio, the amplitude is squared and averaged over some time, then the square root of the result is calculated. The result is a value, that when squared, is proportional toarehe effective power of the signal.

All these features are extracted using the librosa library. Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems.

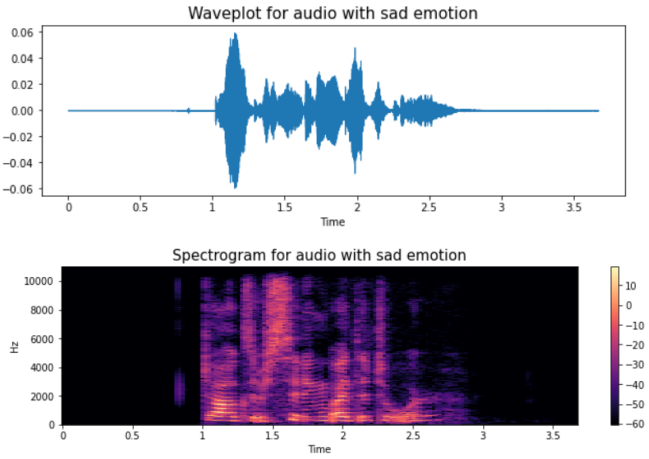


Fig. 2. An audio with sad emotion

fig2

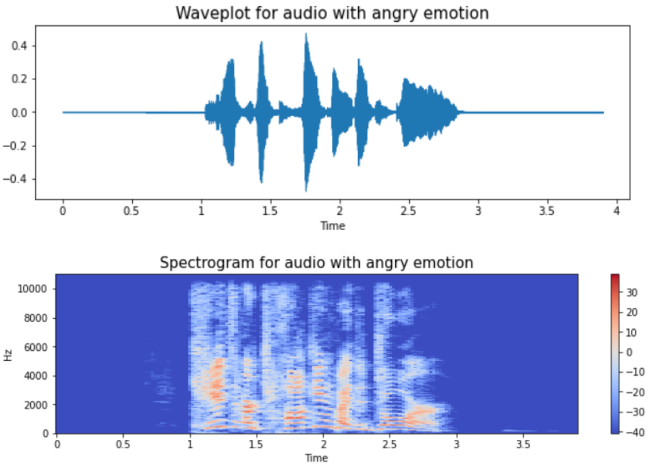


Fig. 3. An audio with anger emotion

## IV. Classification

Categorizing a given set of data into classes is known as classification. The process begins with predicting the

class of given points of data. It is performed on both organized or unstructured information. The classes are frequently alluded to as target, name or categories.

The classification predictive modeling deals with approximating the mapping function from input variables to discrete output variables. The primary objective is to distinguish which class the new predicted data will fall into.

The models were built using tensorflow. TensorFlow is an end-to-end open- source machine learning framework With a focus on deep neural networks. Massive amounts of unstructured data are analyzed through deep learning.

The training was done on google colab. Free GPU is currently supported by Google Colab, a cloud computing platform. Development of python programs while utilizing well-known packages like Keras, TensorFlow, PyTorch,etc is possible with colab.

### A. CNN Model

A convolutional neural network has been used to train the model for classification of emotions. CNN uses feedforward architecture to classify data. The cnn is built using tensorflow.keras. This can be converted into a tensorflow lite model in order to be uploaded to the tinyml board. The model uses a combination of convolutional layers and activation functions. The extracted features like mel frequency cepstral coefficients, zero crossing rate and root mean squared energy were fed into the model and trained for 120 epochs with a minimum learning rate of 0.00001. The training accuracy was observed to be at 99 percent while the test accuracy was around 67percent.

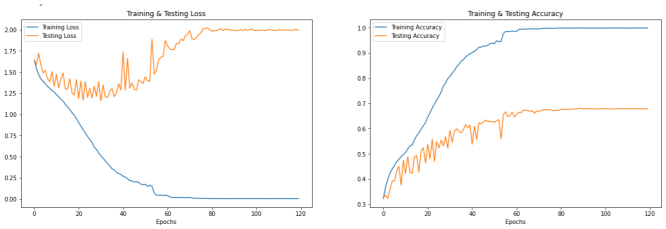


Fig. 4. Learning Curve of CNN

### B. CNN-LSTM Model

A convolutional neural network long short-term memory model is a type of lstm model primarily designed for sequence prediction problems like images and videos. The CNN is used for feature extraction with mel spectrogram as input[8]. LSTM is used to learn temporal dependencies between local invariant features at different time steps [9]. 70 percent of the dataset was chosen for training and 30 percent was chosen for testing. The model was trained for 120 epochs. A training accuracy of 96 percent and a test accuracy of 72 percent were achieved.

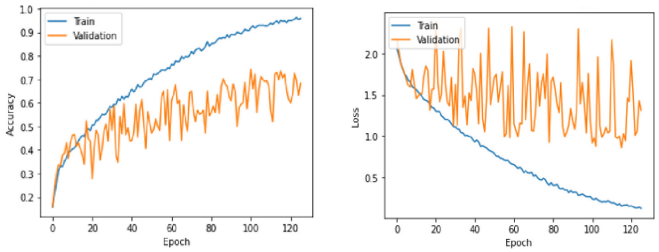


Fig. 5. Learning Curve of CNN-LSTM

## V. TINYML

Machine learning requires a huge amount of RAM and CPU cycles. The recent advancements in the field of chine learning have brought forth algorithms that are capable of running on several microcontrollers. It is possible to build models that are capable of running on devices with limited resources.

Tensorflow lite has an experimental l version that can run on micro control devices. The Adels built using Tens Flow can be converted to TensorFlow lite mode the s using it. lite.TFLiteConverter.

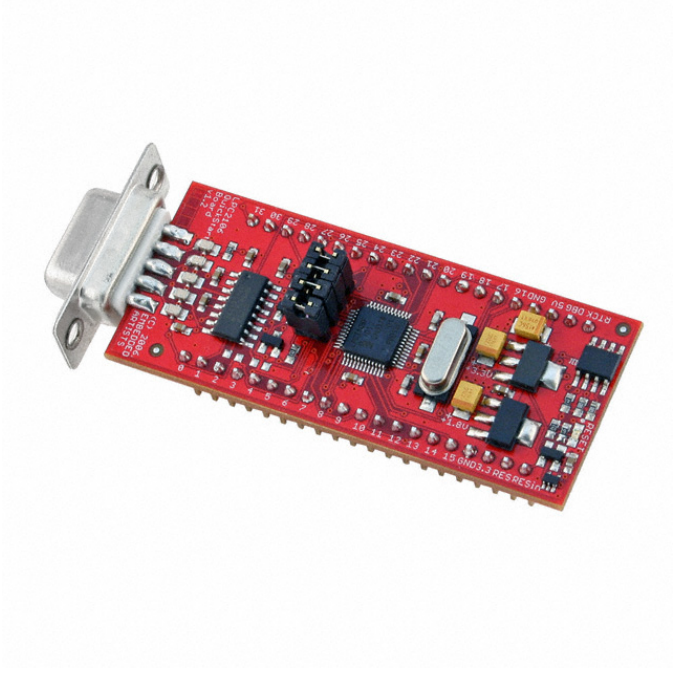

Fig. 6. Tinyml board

This paper proposed an approach to classify and detect emotions into eight categories. The dataset used was a combination of RAVDES and CREMA-D. A total of 20,000 audio files were collected. T, the audio files were processed by adding noise and by shifting and stretching. Features like mel frequency cepstral coefficients, mel spectrograms, zero crossing rate, and root mean squared energy. A deep learning model was trained on google

colab for 120 epochs. The trained model was converted to a TensorFlow lite model and uploaded to the tinyml board. Audio input was fed into the tinyml board using the onboard microphone and the necessary features were extracted and fed into the model and the output was obtained.

| Model | Training Accuracy | Test Accuracy |
|---|---|---|
| CNN | 99 percent | 67 percent |
| CNNLSTM | 96 percent | 72 percent |

TABLE I: Comparison of CNN and CNN-LSTM Models

## REFERENCES


[1] Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M.and Ambikairajah, E. (2021). A comprehensive review of speech emotion recognition systemstems. IEEE Access, 9, 47795-47814.

[2] Lee, K. H., Choi, H. K., and Jang, B. T. (2019, October). A study on speech emotion recognition using a deep neural network. In 2019 Internationaltional Conference on Information and Communication TechnConvergencegence (ICTC) pp. IEEE. 1162-1165.

[3] Issa, D., Demirci, M. F., Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59, 101894.

[4] Sandesara, A., Parikh, S., Sapovadiya, P., Rahevar, M. (2020). A Comparativeparative Study On Speech Emotion Recognition. International Journal of Research in Engineering, Science and Management, 3(11), 25-35.

[5] Shah A, Firoz R, Vimal Anto, Babu. (2009). Emotion Recognition From Malayalam Words Using Artificial Neural Networks.

[6] Aswin K.M, K. Vasudev, K. Shanty and Sreekutty I.K., "HERS:Human emotion recognition system," *2016 International Conference on Information Science (ICIS)*, 2016, pp. 176-179, doi: 10.1109/INFOSCI.2016.7845322.

[7] A. Winursito, R. Hidayat and A. Bejo, "Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition," *2018 International Conference on Information and Communications Technology (ICOIACT)* , 2018, pp. 379-383, doi: 10.1109/ICOIACT.2018.8350748.

[8] Lee, K. H. (2020, October). Design of a convolutional neural network for speech emotion recognition. In 2020 International Conference on Information and Communication Technology Convergence (ICTC) (pp. 1332-1335). IEEE.

[9] Xie, Y., Liang, R., Liang, Z., Huang, C., Zou, C., Schuller, B. (2019). Speech emotion classification using attention-based LSTM. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(11), 1675

[10] R. R. Subramanian, Y. Sireesha, Y. S. P. K. Reddy, T. Bindamrutha, M. Harika and R. R. Sudharsan, "Audio Emotion Recognition by Deep Neural Networks and Machine Learning Algorithms," *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, 2021, pp. 1-6, doi: 10.1109/ICAECA52838.2021.9675492.

[11] G. Wei, L. Jian and S. Mo, "Multimodal(Audio, Facial and Gesture) based Emotion Recognition challenge," *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 2020, pp. 908-911, doi: 10.1109/FG47880.2020.00142.

[12] R. Anusha, P. Subhashini, D. Jyothi, P. Harshitha, J. Sushma and N. Mukesh, "Speech Emotion Recognition using Machine Learning," *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2021, pp. 1608-1612, doi: 10.1109/ICOEI51242.2021.9453028.

[13] A. S. Nasim, R. H. Chowdory, A. Dey and A. Das, "Recognizing Speech Emotion Based on Acoustic Features Using Machine Learning," *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2021, pp. 1-7, doi: 10.1109/ICACSIS53237.2021.9631319.