Speech Emotion Recognition

Yamini S

*Department of Computer Science and Engineering,*

*Panimalar Engineering College*

Chennai, India

yaminisundar2002@gmail.com

Dr TamilVizhi T, M.TECH., Ph.D.,

Associate Professor

*Panimalar Engineering College* Chennai, India

tamilvizhi.phd.it@gmail.comNelopher Nisha M

*Department of Mechatronics Engineering,*

*Panimalar Engineering College* Chennai, India

ashraafnisha3@gmail.com

***Abstract*— Emotions are the best indicators of the actions of humans in advance. It is of great advantage in the current smart world. Prediction of these emotions can be able to sense the current mood of the driver and control the smart automobile accordingly and can be used in case of chatting with the customers using AI devices,etc.This can be done by extracting the features(including MFCC) of the respective emotions and train the learning model using the classification algorithms(CNN) and eventually the model can predict the emotion by comparing the newly retrieved features and the features of the training dataset and classify them accordingly.**

***Keywords—Mel Frequency Cepstral Coefficients, Convolutional Neural Networks.***

# Introduction

The objective of the proposed system is to detect emotions from continuous and spontaneous speech. Speech emotion recognition can be of huge help to prevent cyber crimes and can regulate the smart automobiles according to the emotion assessed from the driver. Other applications can include the chatbots for better understanding of the customers on the other side, audio surveillance, online marketing, call centers, etc. MFCC is one among the features which are used to characterize an audio signal. MFCC is a crucial step to select the most relevant features out of it. A classifier model is built using Convolutional Neural Networks (CNN) algorithm which proves to be the better deep learning algorithm for processing and recognition tasks.

# Related Works

Bagus Tris Atmaja, Akira Sasou and Masato Akagi developed a speech emotion and naturalness recognition system[1] using Multi-Task Learning(MTL). The naturalness ratings are labeled on a five-point scale as dimensional emotion. Here multitask refers to dimensional emotion and naturalness score prediction. This work proceeds with the assumption that predicting both can improve the performance but it has been proved from their research that it fails to predict the low and extremely high scores and adopts MTL which involves more than one local minimum which can eventually lead to variable validation accuracy therefore many parameters and hidden layers has to be modified and MLP is not applicable if more number of features are included. Jennifer Santosoand Takeshi Yamada has developed Speech Emotion Recognition system based on Self-Attention Weight Correction for Acoustic and Text Features[2] which uses BLSTM and self attention mechanism that focuses on the incorrectly recognised terms of ASR but the main limitation of this system is Information loss. Liu Yunxiang;Zhang Kexin developed Design of Efficient Speech Emotion Recognition Based on Multi Task Learning[3] where the Feature space is divided into shared LSTM and private LSTM, which are used to extract shared features and private features respectively. Though this system enhances the high-frequency part of the speech it does not solve the problem of limited data and contains very few emotion types. Felicia Andayani, Lau Bee Theng, Mark Teekit Tsun and Caslon Chua developed Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files[4] where both the transformer and LSTM architectures are combined due to their individual limitations. The adoption of long-term dependencies could affect the features of the samples and it also requires improvement on the preprocessing methods.Changing input sequence cause various disadvantages to the system.

# Proposed Work

Proposed system includes inputting the audio signal, Feature extraction, Feature enhancement, Classifier training, Emotion detection. The audio signal input is preprocessed before feature extraction to remove unwanted noise signal.The features are extracted using MFCC technique .Other techniques are LPC and PLP.Classification is performed which maps the features to emotion. CNN is used for classifying the emotions which shows a larger learning rate .

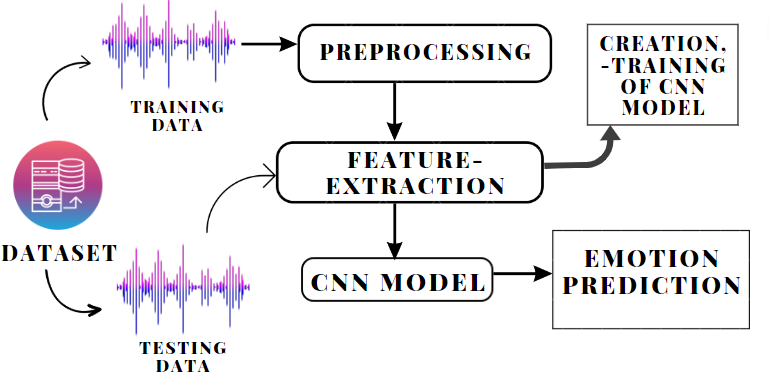


Fig.1 data flow diagram

Fig 1 depicts the outline of the sequential processes undertaken by the model in prediction of the emotion.

*Steps followed by CNN*

CNN is an algorithm which consists of many layers where the initial layers extract the features and the final layer makes the final prediction.The algorithm optimizes by itself and this optimization is done through the use of function gradient. This is built with the help of Keras library.

model = Sequential()

model.add(Conv1D(256,5,padding='same', input\_shape=(x\_traincnn.shape[1],x\_traincnn.shape[2])))

model.add(Activation('relu'))

model.add(Conv1D(128, 5,padding='same'))

model.add(Activation('relu'))

model.add(Dropout(0.1))

model.add(MaxPooling1D(pool\_size=(8)))

model.add(Conv1D(128, 5,padding='same',))

model.add(Activation('relu'))

model.add(Conv1D(128, 5,padding='same',))

model.add(Activation('relu'))

model.add(Flatten())

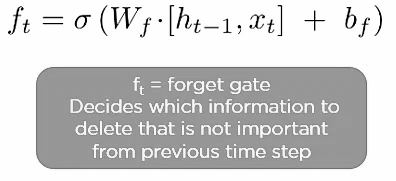
model.add(Dense(y\_train.shape[1]))

model.add(Activation('softmax'))

*Steps followed by LSTM*

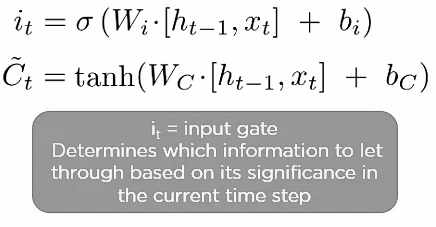
LSTMs is one among the RNN algorithms which learns long-term dependencies by remembering information for long periods is the default behavior. It defines the Network, compiles the Network, fits Network, evaluates Network and makes Predictions.

*Step 1: Decide How Much Past Data It Should Remember*



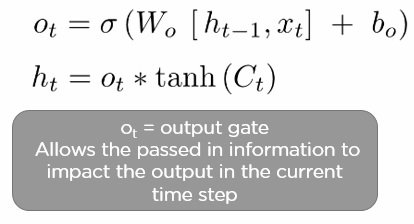
*Fig.2*

*Step 2: Decide How Much This Unit Adds to the Current State*

****

*Fig.3*

*Step 3: Decide What Part of the Current Cell State Makes It to the Output*

****

*Fig.4*

model.compile(loss='categorical\_crossentropy', optimizer=opt,metrics=['accuracy'])

Datasets used are RAVDESS and SAVEE dataset and the libraries used are librosa, matplotlib , numpy, pandas, torchvision, wave,Keras,etc. and the frameworks used are Flask and Pytorch.

Audio speech signals are inputted and Librosa library in Python is used to process and extract features from the audio files using MFCC which are widely used in automatic speech and speaker recognition.The signal is processed through the Fast Fourier Transform which uses less computational time and used to assess the frequency properties of a signal.FFT is just another version of DFT but more efficient and faster.The typical applications of FFT are compression in more complex processing of signals and filtering ,remodeling of signals,etc.

*1. Data Preprocessing*

Librosa and Pytorch are the libraries that are used for preparing the data before-hand.It is required for cleaning the data and making it suitable for a learning model which can enable it to predict the result with more accuracy.Importing libraries, dataset and their division into training and testing set are involved in this process.

import librosa

import librosa.display

import numpy as np

import pandas as pd

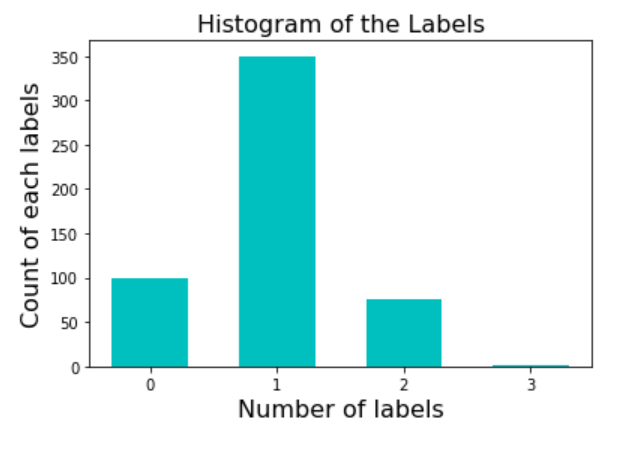
import matplotlib.pyplot

import IPython.display

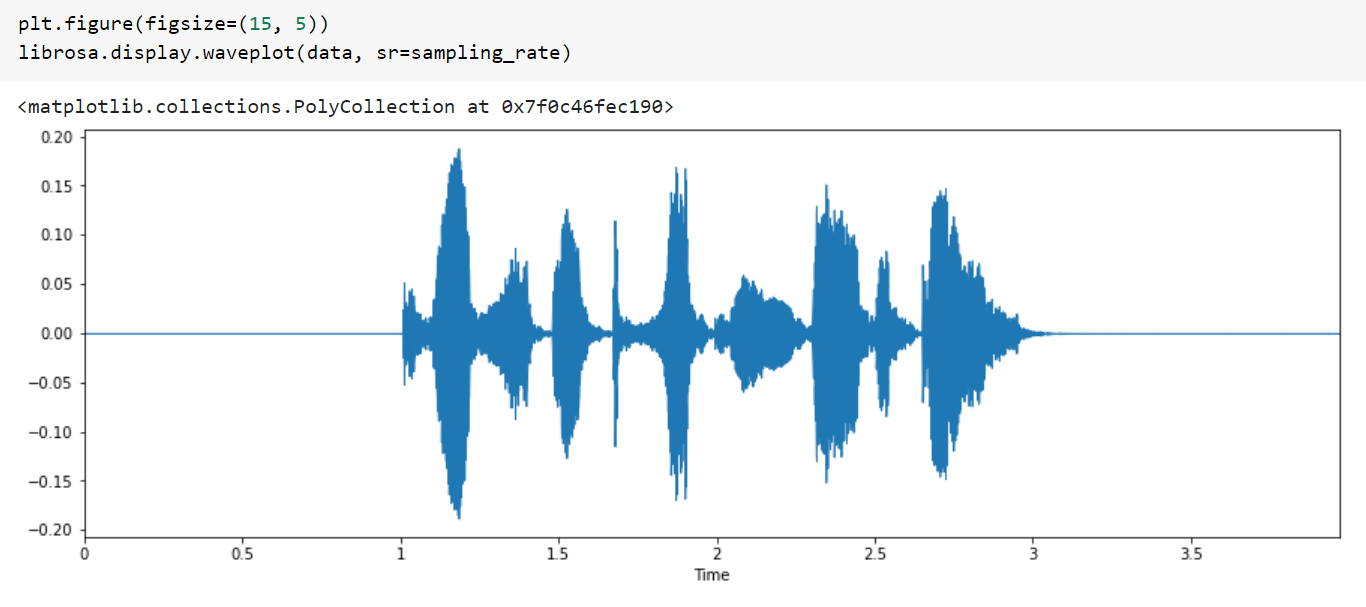
import keras

*2. Data Visualization*

Data Visualization Libraries generally used are [Matplotlib](https://www.geeksforgeeks.org/matplotlib-tutorial/) and [Plotly](https://www.geeksforgeeks.org/python-plotly-tutorial/). It is nothing but visual representation of data for easy analysis. It is typically done using waveform, table or graphs,etc. It is an easy method to retrieve the insights rather than going through all the data and comparing them sequentially. The following Spectrogram and histogram are few examples.



*Fig.5 Histogram representation*



*Fig.6 Spectrogram representation*

*3. Feature Extraction*

It is the process where the main features which characterizes the audio signal are retrieved and compared with the brand new audio file to assess the similarity and classify them as respective emotions.It is usually done by converting the waveform to parametric form and representing the signals as features. Other features extracted are Magnitude, Pitch and Chroma.

from utils.feature\_extraction import get\_features\_dataframe

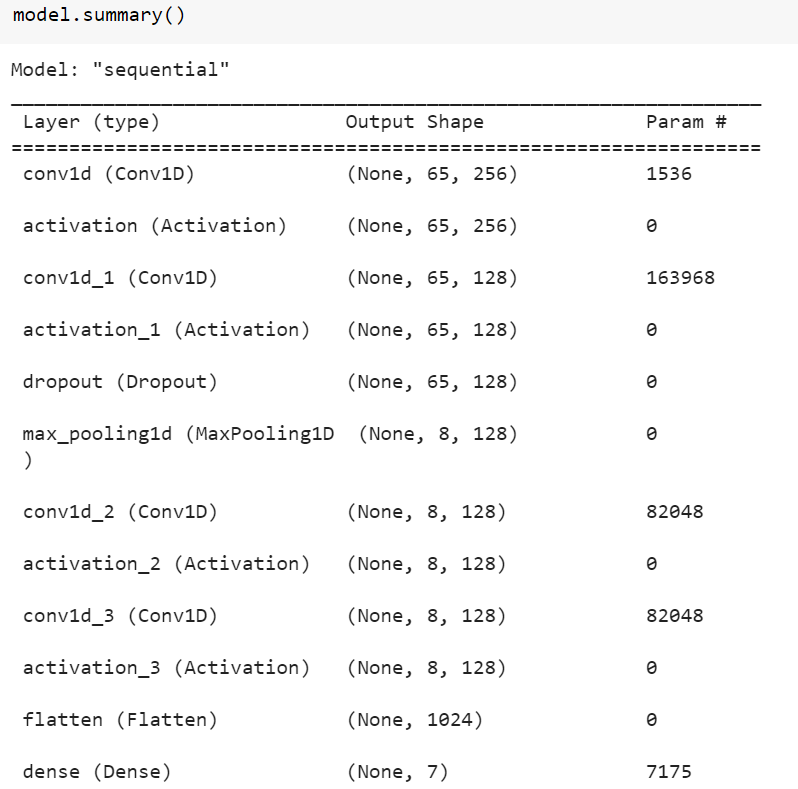
from utils.feature\_extraction import get\_audio\_features

trainfeatures, trainlabel = get\_features\_dataframe(train\_df, sampling\_rate)

testfeatures, testlabel = get\_features\_dataframe(test\_df, sampling\_rate)

*4. Model Creation*

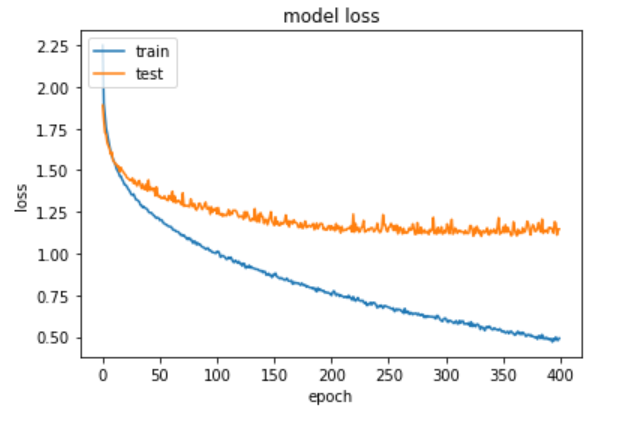
CNN is an algorithm which consists of many layers where the initial layers extract the features and the final layer makes the final prediction.The algorithm optimizes by itself and this optimization is done through the use of function gradient. This is built with the help of Keras library which is typically used for developing and training the deep models, specifically neural-network based models.It provides simple api and is user friendly.

**

*Fig.7 Depiction of CNN Model*

*5. Training and Evaluation*

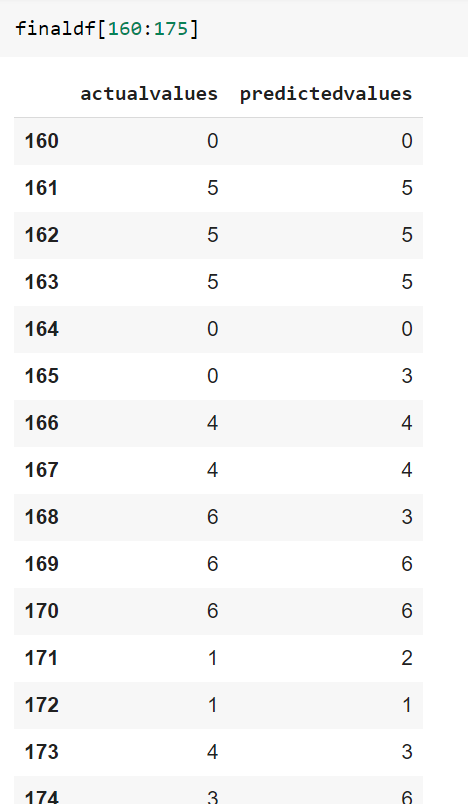
The built CNN Model is trained using the features extracted from the training data set and the loss of data is analyzed and revised.



*Fig.8 Loss vs Epoch*

*6. Testing*

The Model is predicted against the testing data to assess the accuracy of the model.



*Fig.9 Actual vs predicted emotions*

*7. Live Demonstration*

The User can upload the audio files and the model can predict the emotion of the speech by extracting the features and comparing those features against the emotion and displaying the respective emotion using the training acquired.

demo\_mfcc, demo\_pitch, demo\_mag, demo\_chrom = get\_audio\_features(demo\_audio\_path,sampling\_rate)

mfcc = pd.Series(demo\_mfcc)

pit = pd.Series(demo\_pitch)

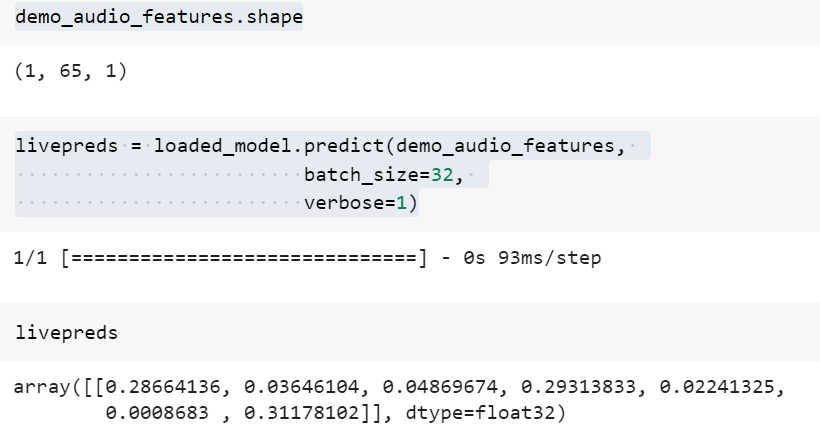
mag = pd.Series(demo\_mag)

C = pd.Series(demo\_chrom)

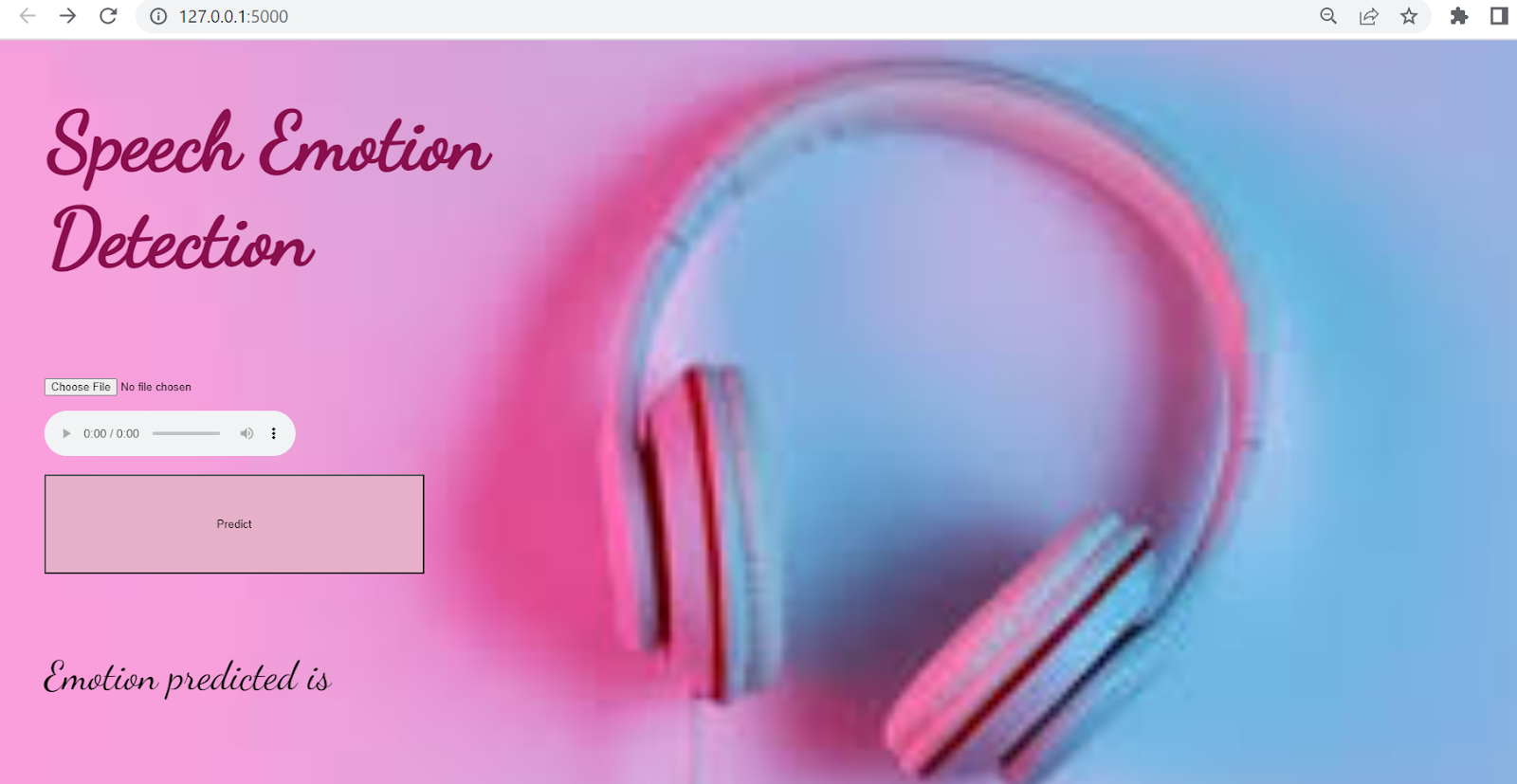
demo\_audio\_features=pd.concat([mfcc,pit,mag,C],ignore\_index=True)

demo\_audio\_features=np.expand\_dims(demo\_audio\_features, axis=0)

demo\_audio\_features=np.expand\_dims(demo\_audio\_features, axis=2)



*Fig. 10 Live Prediction*

****

*Fig.11 Website*

****

*Fig.12 Prediction of the CNN Model*

# Conclusion and Future direction

Thus the proposed system has the capability of understanding the human emotions which contains extra insights about human actions without the actual need for an actual presence and thereby can understand the motives of the people. This system can be further enhanced by combining the facial emotion recognition system with the speech emotion recognition system so that the emotions are predicted more accurately by analyzing both facial and audio-extracted features.

##### V. References

1. Bagus Tris Atmaja, Akira Sasou and Masato Akagi, “Speech emotion and naturalness recognition system”, vol.10, pp.72381-72387, July 2022.
2. Jennifer Santosoand Takeshi Yamada, “Speech Emotion Recognition system based on Self-Attention Weight Correction for Acoustic and Text Features”, vol.10, pp.115732-115743, November 2022.
3. Liu Yunxiang;Zhang Kexin, “Design of Efficient Speech Emotion Recognition Based on Multi Task Learning”, vol.11, pp.5528-5537, January 2023.
4. Felicia Andayani, Lau Bee Theng, Mark Teekit Tsun and Caslon Chua, “Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files”, vol.10, pp.36018-36027, March 2022.
5. Mohammad Reza Falahzadeh, Edris Zaman Farsa, “3D Convolutional Neural Network for Speech Emotion Recognition With Its Realization on Intel CPU and NVIDIA GPU”, vol.10, pp.112460-112471, October 2022.
6. Samuel Kakuba, Alwin Poulose, “Attention-Based Multi-Learning Approach for Speech Emotion Recognition With Dilated Convolution”, vol.10, pp.122302-122313, November 2022.
7. Samuel Kakuba, Alwin Poulose, “Deep Learning-Based Speech Emotion Recognition Using Multi-Level Fusion of Concurrent Features”, vol.10, pp.125538-125551, November 2022.
8. [Bagus Tris Atmaja](https://ieeexplore.ieee.org/author/37086950772); [Akira Sasou](https://ieeexplore.ieee.org/author/37351267800), “Evaluating Self-Supervised Speech Representations for Speech Emotion Recognition”, vol.10, pp.124396 - 124407, November 2022.
9. Haiyan Wang, Xiaohui Zhao, “Investigation of the Effect of Increased Dimension Levels in Speech Emotion Recognition”, vol.10, pp.78123-78134, July 2022.
10. Kusha Sridhar, Carlos Busso, “Unsupervised Personalization of an Emotion Recognition System: The Unique Properties of the Externalization of Valence in Speech”, vol.13, pp.1959-1972, December 2022.