Multi-Model Audio-Visual Emotion Recognition

Project Report for EE 769, Introduction to Machine Learning, Spring 2023

Krishanu Das
Dept. of Energy Science and Engineering
IIT Bombay
20D170020@iitb.ac.in

Vedant Khokher
Dept. of Energy Science and Engineering
IIT Bombay
20D170043@iitb.ac.in

***Abstract*—In this paper we propose our multimodel deep learning model to perform Audio-Visual Emotion Classification. For this we used 2 separate models, a CNN model for image classification & multi-layered perception model for speech recognition. And for our final output we take the weighted average of the outputs from the respective models. The performance of the models were 60% and 99.12% (at the time of running) respectively. All the research papers we went through give only the theoretical aspect of this problem, but we tried to code it from scratch (taking help of some online sources), by estimating various features which are discussed in this paper.**

I. Introduction

Human communication is heavily influenced by their emotional states. Identification of the emotional states is a challenging task in many fields with multiple applications including lie detection, audio–visual surveillance, effective computing, online teaching–learning, online meeting, human–computer interaction (HCI), and many more. The study of emotional variability is an important factor for investigating psychological adaptation and well-being. Moreover, it is also important for machines to be able to recognize human emotions in order to make better decisions. It is because intelligent machines have become an indispensable part of our daily lives, hence developing methods to help them correctly classify users' emotions has become essential for the advancement of society.

Human emotion recognition can be accomplished through a variety of means, including speech data, facial expressions, body gestures, physiological parameters, and many others. Because each of these modalities is distinct, fusing their results in a rich representation of features capable of performing emotion recognition efficiently. Previous research works have shown that relying solely on a single modality for emotion recognition is inefficient. More specifically, some existing literature also demonstrates that using multiple modalities (audio, video, text, etc.) for emotion recognition yields significantly better results than using only one.

II. Background and Prior work

The traditional way of emotion recognition using machine learning involved the use of handcrafted features and classic machine learning algorithms.

- **Feature-Based methods** rely on the extraction of handcrafted features from the input signals (such as speech, facial expressions, or physiological signals) that are believed to be relevant to the emotional state. Examples of features commonly used in emotion recognition include Mel-frequency cepstral coefficients (MFCCs) for speech signals and local binary patterns (LBP) for facial expressions. The extracted features are then used as input to classic machine learning algorithms, such as support vector machines (SVM), k-nearest neighbors (KNN), or decision trees, to predict the emotional state.
- **Dimensionality reduction techniques**, such as principal component analysis (PCA) or linear discriminant analysis (LDA), can be used to reduce the dimensionality of the input signals and extract relevant features for emotion recognition. The reduced features are then used as input to classic machine learning algorithms.
- **Ensemble methods** combine the predictions of multiple models to improve the performance of emotion recognition. For example, bagging or boosting methods can be used to combine the predictions of multiple decision trees or SVM classifiers.

These traditional methods have limitations in capturing complex and subtle emotional cues, as they rely on pre-defined features that may not be sufficient to capture the full range of emotions. Moreover, these methods require manual feature engineering, which can be time-consuming and may not be robust across different datasets and tasks. Recent advancements in deep learning have shown promising results for automatic feature extraction and improved performance in audio-visual emotion recognition.

III. Dataset

Images

We used Kaggle dataset for images https://www.kaggle.com/datasets/jonathanoheix/face-expression-recognition-dataset which contains over 35,000+ images divided into 7 classes i,e Angry, Sad, Happy, Neutral, Surprise, Fear and Disgust. We used only these emotions because they are easily identifiable and also we need to have datasets of the same emotions for audio also, which can be difficult to find for other emotions.

Speech

For speech we used the famous RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset RAVDESS Emotional speech audio | Kaggle. This is a small portion of the complete dataset which contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalising two lexically-matched statements in a neutral North American accent. Speech emotions includes Neutral, Happy, Sad, Angry, Fearful, Surprise, and Disgust expressions. All these audio files are mono channel audio files.

Here is the waveform of all the sample classes from the dataset -

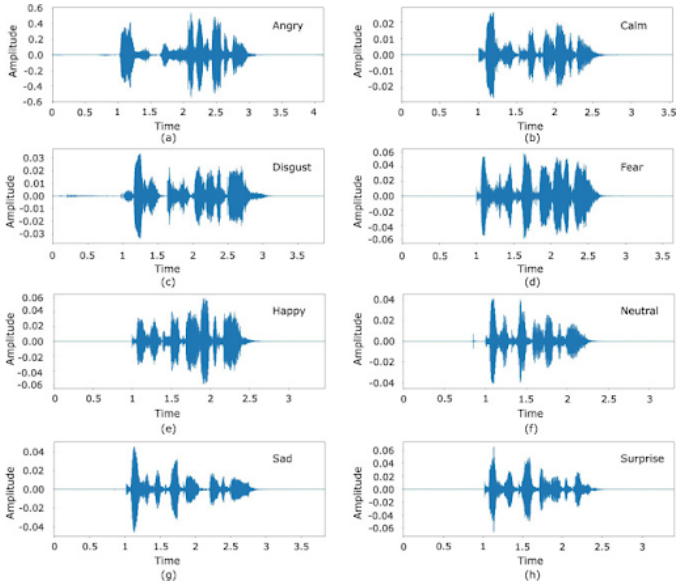

Fig. 1.
Images of waveform for various types of emotion file of RAVDEES dataset

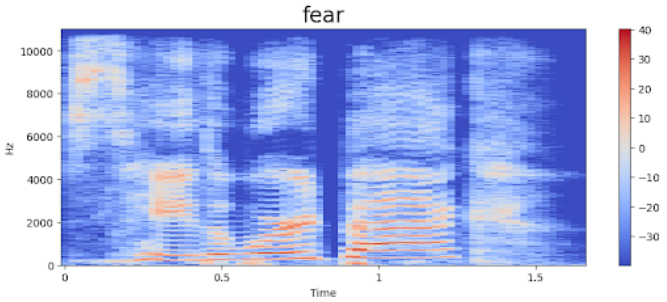

Fig. 2.
Spectrogram(aka Mel-spectrum) of a audio file classifies as fear

IV. Pre-Processing and Feature Extraction

Image based pre-processing

The images were of different sizes so we converted all the images into 48x48 matrix. We decolourised all the images from RGB to Grey so as to make the model training less expensive and less time consuming. Not all the images contain only face, some of them also have some background objects and directly applying machine learning model to the dataset can reduce the efficiency of the model so we used Haar Cascade Frontal Face to detect the face from images which act as our region of interest and in that roi we apply our model. The Haar cascade frontal face classifier uses Haar-like features to detect the face.

Audio based pre-processing

Speech recognition is a complex task and extracting important features have a major impact on your performance efficiency. Since our audio files are already in mono form we won't convert them (if they are in stereo form then you have to convert them into mono and then extract the following features).

The features which we will be extracting are:

- **MFCC:** MFCCs (Mel Frequency Cepstral Coefficients) are widely used in speech recognition due to their effectiveness in capturing the spectral characteristics of speech signals. The feature applies a Mel-scale filter bank to the speech signal, which is a set of overlapping triangular filters that mimic the non-linear frequency response of the human ear. The Mel-scale filter bank divides the speech signal into different frequency bands, and the energy in each band is then compressed using a logarithmic function. The resulting coefficients are then transformed using the discrete cosine transform (DCT), which produces a set of coefficients that are less sensitive to noise and are more compact than the original spectral information.

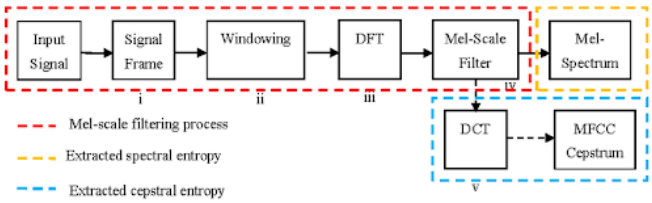

Fig. 3.
Process of extracting MFCCs coefficient from input audio file

- **Chroma:** Chroma is a feature extraction technique that is used in speech recognition to capture the pitch information of speech signals. The pitch information is important for speech recognition because it provides

information about the fundamental frequency of the speech signal, which is closely related to the speaker's identity and emotional state.To achieve this, the speech signal is first divided into short-term frames, and for each frame, the power spectrum is calculated using the fast Fourier transform (FFT).

- **Contrast and Tonnetz:** Contrast and Tonnetz are two feature extraction techniques that are used in music information retrieval (MIR) and have been applied to speech recognition as well. Contrast measures the differences between adjacent frequency bands in the power spectrum of the audio signal whereas Tonnetz represents the pitch relationships between musical notes as a geometric structure.

V. Proposed Model

For image classification we used multilayered CNN with 4 dense layers and 2 fully connected layers. The first dense connected layer is of (64,3,3), second (128,5,5), third (512,3,3) and fourth (512,3,3) for fully connected layers they were (256) and (512) resp. With this deep network the model was able to extract important features and was able to predict correctly.

For audio it was a bit challenging task for us. The first model which we used was a LSTM with 4 hidden layers, but the accuracy was very low, probably it wasn't able to capture some important features from the datasets. Then we used pre pre-trained MLP Classifier which had an accuracy of around 99% on the training dataset.

To form a combined output we took the output of both models and did weighted average for each emotion, for every file.

The overall pipeline is as follows:

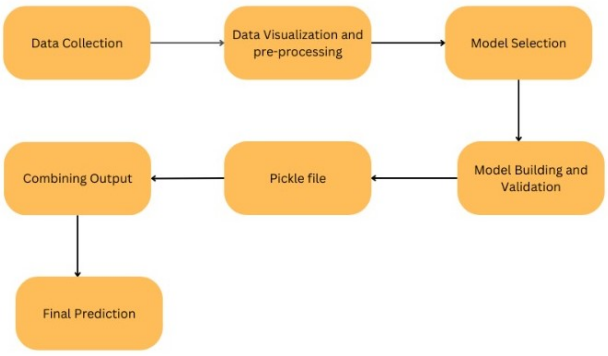

Fig. 4.
Pipe of the complete process

VI. Results and Difficulty faced

The accuracy of the CNN model was around 60% . As you can see for a video of a happy person the model detected happy with a probability of 46%.

Angry	0.072381
Disgust	0.007507
Fear	0.111848
Happy	0.467495
Neutral	0.107636
Sad	0.102065
Surprise	0.131068

Fig. 5.

Although for the audio of a happy person the speech model failed badly. One of the reasons why the model outperformed in training and had low accuracy in testing is overfitting. Maybe the amount of data collected was low or because of noise it was not able to predict well.Another reason might be that the features extracted from pre-processing of audio files might not include sufficient data points for a good accuracy.

References

[1] Emotion recognition using deep learning approach from audio–visual emotional big data - ScienceDirect

[2] Deep Learning for Audio Visual Emotion Recognition | IEEE Conference Publication | IEEE Xplore

[3] 1907.03196v1.pdf (arxiv.org)

[4] Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities - ScienceDirect

[5] https://www.sciencedirect.com/science/article/pii/S0950705122002593#b11

[6] https://www.sciencedirect.com/science/article/pii/S0950705122002593#b6

[7] https://www.sciencedirect.com/science/article/pii/S0950705122002593#b7

[8] https://towardsdatascience.com/audio-deep-learning-made-simple-part-1-state-of-the-art-techniques-da1d3dff2504

[9] https://towardsdatascience.com/audio-deep-learning-made-simple-part-2-why-mel-spectrograms-perform-better-aad889a93505

[10] https://towardsdatascience.com/speech-emotion-recognition-with-convolution-neural-network-1e6bb7130ce3

[11] https://towardsdatascience.com/classifying-emotions-using-audio-recordings-and-python-434e748a95eb

[12] https://towardsdatascience.com/building-a-speech-emotion-recognizer-using-python-4c1c7c89d713

[13] https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53