School of Computer Science and Engineering

KLE Society's
KLE Technological University

EIS Course Project Report

Topic - Audio Based Emotion Recognition

Team No. : 02

Name	USN	Roll No.
Shivtej Narake	01FE18BCS124	305
Pooja Majali	01FE18BCS143	324
Saloni Shah	01FE18BCS183	363
Siddarth Vankudre	01FE18BCS248	508

Course Instructor

Uday Kulkarni

School of Computer Science and Engineering

INDEX

1. INTRODUCTION

The goal of speech emotion recognition is to automatically detect a person's emotional or physical condition based on his voice. The emotional state of a person's speech is an essential component in human communication and engagement since it gives feedback in communication without changing the linguistic content. The discrete emotion classification method is used to recognise spoken emotion. In most situations, literature focuses solely on six emotion labels: joyful, sad, angry, disgusted, fear, and surprise. In actual life, however, the emotion categories are more numerous and complicated.

Speech is a rich, dense mode of communication that is capable of efficiently conveying information. There are two sorts of data in it: linguistic and paralinguistic. The former refers to the explicit information such as body language, gestures, facial expressions, tone, pitch, emotion, and so on, while the latter refers to the implicit information such as body language, gestures, facial expressions, tone, pitch, emotion, and so on.

An emotion identification system's performance is solely dependent on features/representations derived from audio. They are divided into two categories: time-based and frequency-based characteristics. Extensive study has been conducted to assess the benefits and drawbacks of these characteristics. There is no one sound characteristic that is capable of performing effectively across all sound signal processing applications. In addition, features are custom-made to meet the needs of the situation at hand. We have been successful in extracting the hierarchical representation of the speech from these characteristics and detecting the underlying emotion in the speech using deep learning techniques. As a result, the model's performance in a specific voice recognition task is far more dependent on the feature selection than on the model design.

2. PROBLEM STATEMENT

Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch. This is also the phenomenon that animals like dogs and horses employ to be able to understand human emotion. Emotions are one of the most fundamental aspects of human beings and we recognize the emotion of an individual from their speech using audio signals by extracting the features and classifying them. We will classify speech into 8 emotions: Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).

3. DATASET

Choosing an emotional speech database, feature selection from audio data, and classifiers to identify emotion are the three primary components in designing a SER. The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Music) dataset is a multi-modal database of emotional speech and song that has been verified. “This gender-balanced database consists of 24 professional actors, each performing 104 distinct vocalisations with emotions such as joyful, sad, angry, scared, surprise, disgust, calm, and neutral,” according to the website. For each mood, each actor performed two statements: "Kids are chatting by the door" and "Dogs are sitting by the door."

Except for neutral (normal only), these comments were also recorded in two distinct emotional intensities, normal and strong, for each emotion. Each vocalisation was performed twice by the actors. There are 1440 spoken utterances and 1012 music utterances altogether. The RAVDESS dataset is extremely diverse in nature, since it is free of gender bias and contains a wide

spectrum of emotions at various levels of emotional intensity. We also see that the RAVDESS dataset is evenly distributed across all emotion classes (15%), indicating that it is free of class-imbalance issues.

Additionally, the developers of the RAVDESS dataset conducted rigorous validation and reliability testing. 247 naïve individuals were asked to give three judgments on three classes: "category of the feeling, strength of the emotion, and authenticity of the emotion" from a "pseudo-randomly generated collection of 298 stimuli, consisting of 174 speech and 124 music presentations."

Figure 1 shows that the performers performed successfully for roughly 73 percent of the rater-selected emotions, confirming the accuracy of the emotion classification and audio content. Furthermore, we discovered that human raters had difficulty distinguishing between neutral and tranquil emotions. We personally listened to the RAVDESS audio files and noticed that the emotions calm and neutral sounded quite similar. As a result, we decided to combine these two feelings into one lesson. Figure 2 depicts the data distribution by gender and emotion class.

		Actor intended emotion							
		Neutral	Calm	Happy	Sad	Angry	Fearful	Disgust	Surprise
Rater chosen emotion	Neutral/Calm	86.6	69.9	14.25	17.12	4.03	4.5	4.36	7.03
	Happy	0.63	17.27	68.44	1.48	0.23	0.59	0.59	6.56
	Sad	4.65	6.06	2.29	60.85	1.02	6.58	8.65	0.76
	Angry	3.82	1.02	1.79	2.9	81.32	4.79	6.48	2.78
	Fearful	0.63	0.66	1.67	9.64	1.39	70.71	2.31	2.22
	Disgust	1.15	1.46	0.78	3.09	8.37	1.81	69.77	3.28
	Surprise	0.28	0.33	7.88	0.69	1.2	7.76	4.13	72.29
	None	2.26	3.3	2.9	4.24	2.45	3.26	3.72	5.07

Figure 1. Dataset Validation

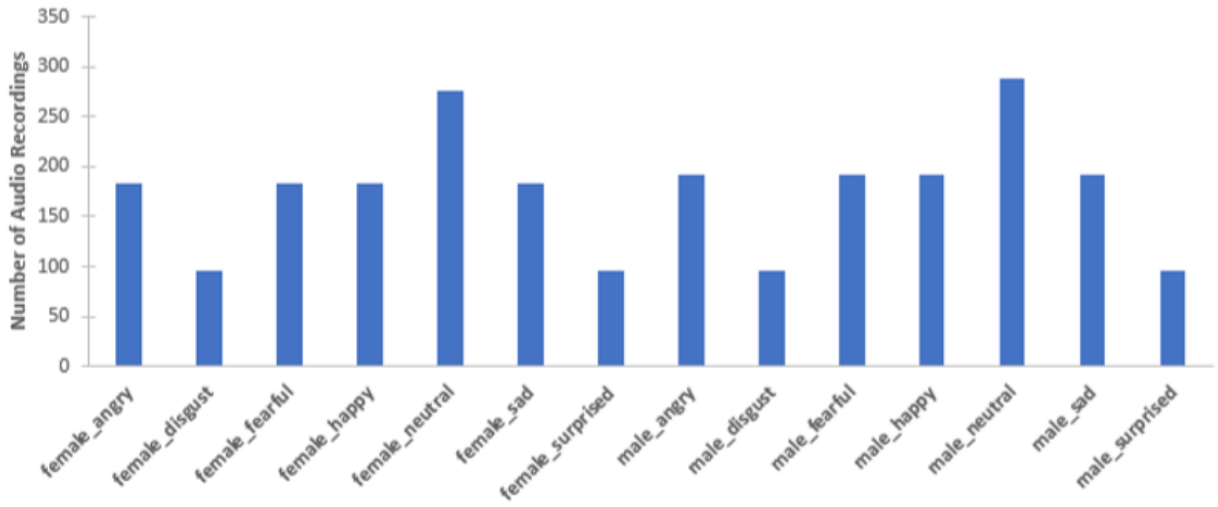


Figure 2. Data distribution across gender and emotion

Their naming convention gives the details about the audio as following:

- ➔ Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
- ➔ Vocal channel (01 = speech, 02 = song).
- ➔ Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
- ➔ Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.
- ➔ Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- ➔ Repetition (01 = 1st repetition, 02 = 2nd repetition).
- ➔ Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

4. PREPROCESSING

A. SVM

Pre-emphasis filter

To enhance all the high frequencies in the audio signal, apply a pre-emphasis filter.

Framing

After the pre-emphasis filter, we must divide the audio stream into frames, which are short-term windows. Window sizes for speech processing typically range from 20 to 50 milliseconds, with 40 to 50 percent overlap between two consecutive windows. Because audio transmissions are non-stationary by nature, the major purpose for this step is to avoid the loss of frequency contours of an audio signal over time. Because the frequency characteristics of a signal vary with time, applying the Discrete Fourier Transform over the whole sample isn't practical.

Hamming

We multiply each frame by a Hamming window function after dividing the signal into several frames. It improves signal clarity by minimising spectral leakage and any signal discontinuities. If the beginning and end of a frame do not match, the signal will appear to be discontinuous, and the Discrete Fourier Transform will display gibberish. While smoothing the signal, the Hamming function ensures that the beginning and finish match up.

Discrete Fourier Transform

Because it permits transforming a sequence from the time domain to the frequency domain, the Discrete Fourier Transform is one of the most commonly utilised transforms in all areas of digital signal processing. DCT is a useful tool for displaying the frequency content of an audio signal's dispersion.

Feature Extraction

- **Zero Crossing Rate**: The rate of sign-changes of the signal during the duration of a particular frame.
- **Energy**: The sum of squares of the signal values, normalized by the respective frame length.
- **Entropy of Energy**: The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
- **Spectral Centroid**: The center of gravity of the spectrum.
- **Spectral Spread**: The second central moment of the spectrum.
- **Spectral Entropy**: Entropy of the normalized spectral energies for a set of sub-frames.
- **Spectral Flux**: The squared difference between the normalized magnitudes of the spectra of the two successive frames.
- **Spectral Rolloff**: The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
- **MFCCS**: Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.

Global Statistics are then computed on upper features:
mean, std, med, kurt, skew, q1, q99, min, max and range.

B. Time-Distributed CNNs

- **Signal discretization**

  The analogue signal time-discretization is made by sampling the signal in equal time intervals. These samples have continuous-amplitudes. The amplitude discretization is made by the approximation of these samples to a discrete number of levels.

- **Audio Data Augmentation**

  Data augmentation is the process by which we create new synthetic data samples by adding small perturbations on our initial training set. To generate syntactic data for audio, we can apply noise injection, shifting time, changing pitch and speed.

- **Log-mel-spectrogram Extraction**

  Mel Spectrograms are spectrograms that visualize sounds on the Mel scale as opposed to the frequency domain.

C. CNN Model

- ### Data Augmentation

  Data augmentation is the process of creating fresh synthetic data samples from our initial training set by adding minor disturbances. We can use noise injection, time shifts, pitch changes, and speed changes to produce syntactic data for audio. The goal is to make our model insensitive to those perturbations and improve its generalisation capabilities. Adding perturbations must preserve the same label as the original training sample for this to operate.

  Augmentation is done by following methods:
  Noise Injection
  Stretching
  Shifting
  Pitching

- ### Feature Extraction

  The extraction of features is a critical step in evaluating and discovering relationships between various objects. We already know that the audio data supplied by the models cannot be comprehended directly by the models, so we need to transform it into a format that the models can understand, which is where feature extraction comes in. The audio signal is a three-dimensional signal with time, amplitude, and frequency represented on three axes.

Some valuable features that can be extracted from audio are-

Zero Crossing Rate : The rate of sign-changes of the signal during the duration of a particular frame.
Energy : The sum of squares of the signal values, normalized by the respective frame length.
Entropy of Energy : The entropy of sub-frames' normalized energies. It

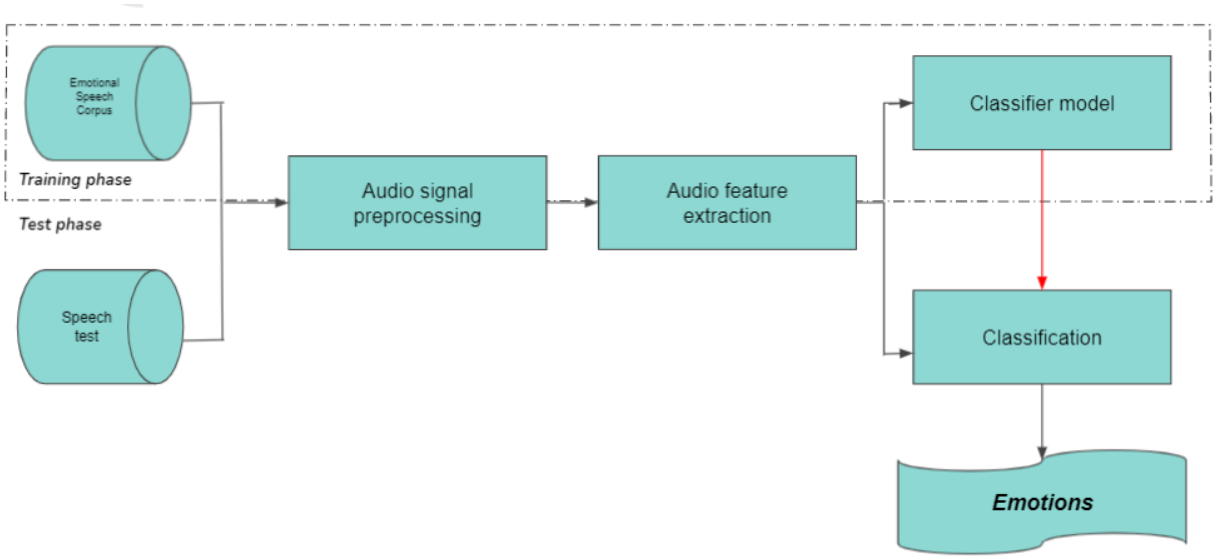
can be interpreted as a measure of abrupt changes.
Spectral Centroid : The center of gravity of the spectrum.
Spectral Spread : The second central moment of the spectrum.
Spectral Entropy : Entropy of the normalized spectral energies for a set of sub-frames.
Spectral Flux : The squared difference between the normalized magnitudes of the spectra of the two successive frames.
Spectral Rolloff : The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
MFCCs : Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
Chroma Vector : A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
Chroma Deviation : The standard deviation of the 12 chroma coefficients.

In this project we have extracted 5 features, namely:

- Zero Crossing Rate
- Chroma_stft
- MFCC
- RMS(root mean square) value
- Mel Spectrogram to train our model.

5. METHODOLOGY

PIPELINE:



- **CNN MODEL**

  RNN networks are known to work well for speech recognition tasks. However, there's a strong body of research that proves that CNN networks can outperform RNN networks in a lot of cases. In this case, we decided to go for CNN networks using 1-dimensional convolution layers and 1-dimensional pooling layers (as our training data is made of 3 dimensions).

  Our CNN network consisted of two blocks, each built of a 1-dimensional convolution layer, activation function ('ReLu'), 1-dimensional pooling layer and dropout. The two blocks were followed by two fully connected dense layers and a 'SoftMax' activation function, as we are dealing with a multi-class problem.

6. RESULTS

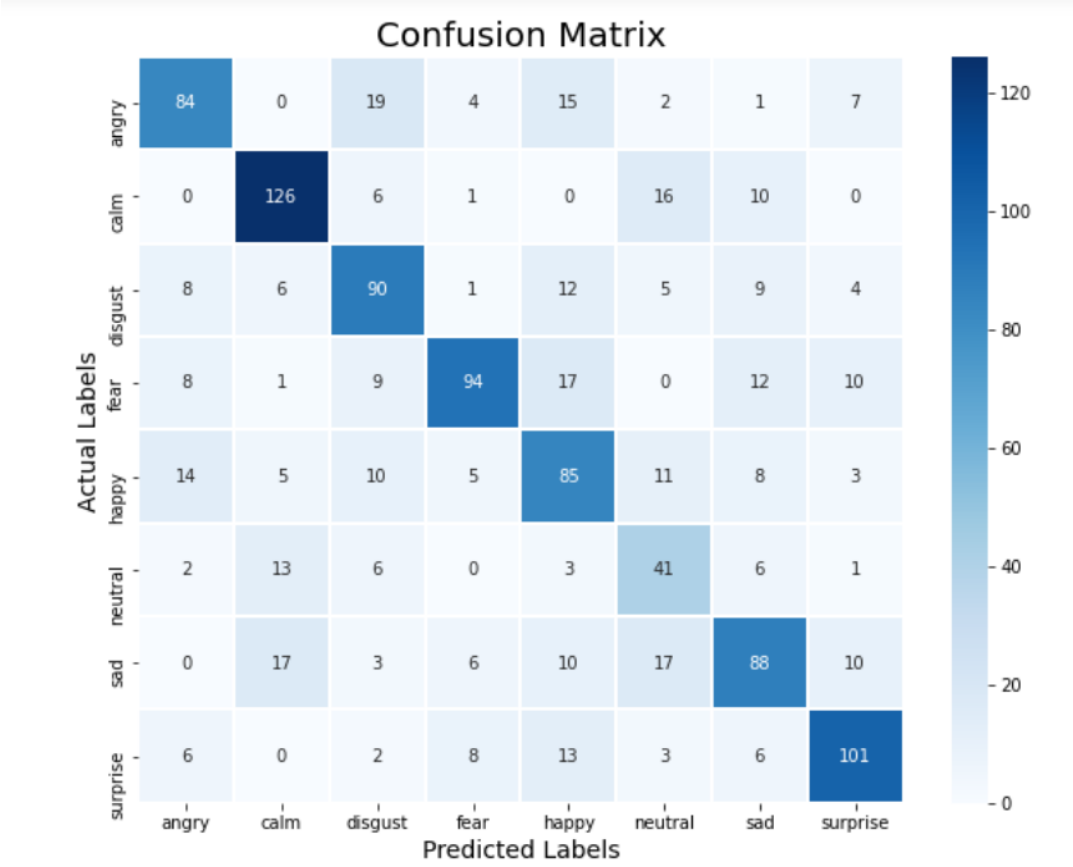
- ## CNN model

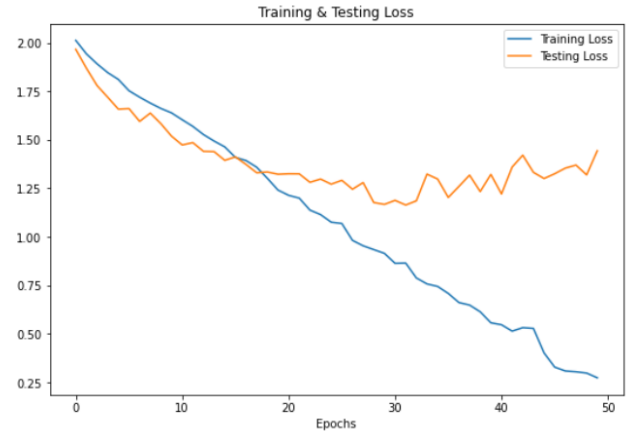
We managed to reach relatively good results, reaching 65.65% test accuracy.

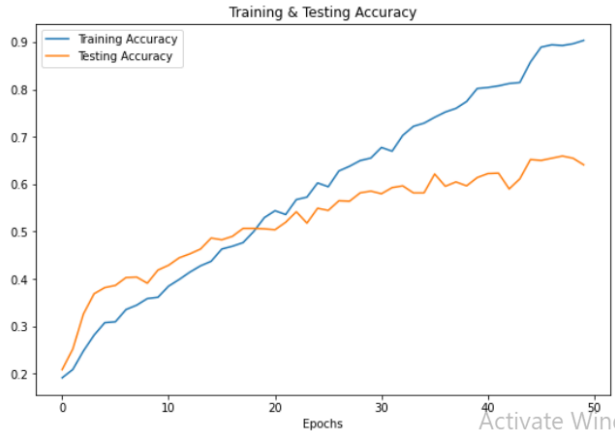
	precision	recall	f1-score	support
angry	0.69	0.64	0.66	132
calm	0.75	0.79	0.77	159
disgust	0.62	0.67	0.64	135
fear	0.79	0.62	0.70	151
happy	0.55	0.60	0.57	141
neutral	0.43	0.57	0.49	72
sad	0.63	0.58	0.60	151
surprise	0.74	0.73	0.73	139
accuracy			0.66	1080
macro avg	0.65	0.65	0.65	1080
weighted avg	0.67	0.66	0.66	1080

As expected, this classification seems relatively balanced, and with pretty good results as well. We can also observe the Confusion matrix, which shows that the vast majority of our samples were classified correctly.

It's interesting to see that quite a lot of samples that were predicted as 'neutral' turned out to be 'calm' (16 samples). This could possibly be explained by the subtle characteristics of calm when it is being expressed by voice, characteristics that could very well be mistaken as being associated with a neutral emotion instead.







7. APPLICATION

Through this project, we showed how we can leverage Machine learning to obtain the underlying emotion from speech audio data and some insights on the human expression of emotion through voice. This system can be employed in a variety of setups like Call Centre for complaints or marketing, in voice-based virtual assistants or chatbots, in linguistic research, etc. It can also be used to detect the mental health of an individual person and help the society. SER(Speech Emotion Recognition) is used in call centers for classifying calls according to emotions and can be used as the performance parameter for conversational analysis thus identifying the unsatisfied customer, customer satisfaction and so on.. for helping companies improve their services. It can also be used in-car board systems based on information of the mental state of the driver can be provided to the system to initiate his/her safety preventing accidents from happening.

8. REFERENCES

https://librosa.org/doc/latest/index.html

https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/CNN_ASLPTrans2-14.pdf-

https://www.izotope.com/en/learn/understanding-spectrograms.html-

https://arxiv.org/pdf/2007.11154.pdf--

https://www.youtube.com/watch?v=4_SH2nfbQZ8&t=1357s