KLE TECH

KLE Technological University

Creating Value
Leveraging Knowledge

SPEECH EMOTION RECOGNITION

School of Electronics and Communication Engineering

Under the guidance of : Prof. Satish & Prof. Nirmala

Overview ;

- Introduction
- Literature Survey
- Problem statement
- Objectives
- Functional block diagram
- Proposed Methodology
- Implementation
- Demonstration of results
- Optimization
- Conclusion
- Future Scope
- References

Introduction ;

- Emotions are important part of understanding human interactions.
- Research is going into finding methods that can recognise emotions displayed in the form of changes in tone while speaking ,Speech Emotion Recognition (SER) is one of such fields .
- Using deep learning and machine learning algorithms, we aim to design an automatic emotion recognition system.

Problem statement ;

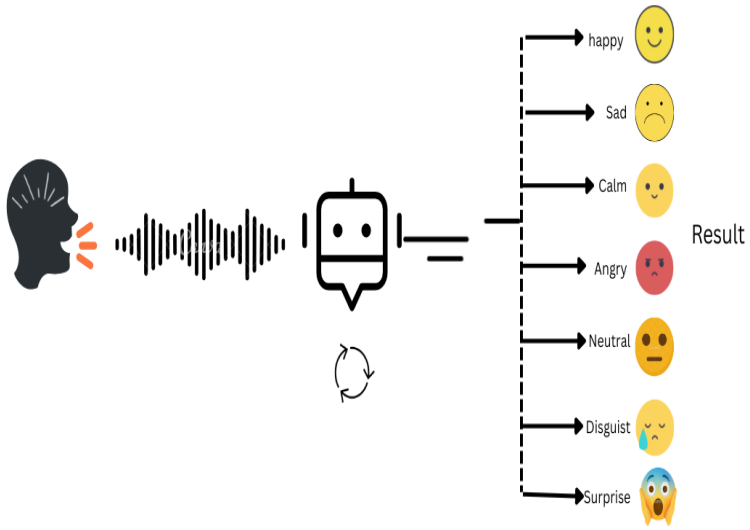
Speech Emotion Recognition using Machine learning and Deep learning Algorithms.

Literature survey ;

KLE Technological University
Creating Value
Leveraging Knowledge

What is Speech Emotion Recognition?

- Human speech contains several features that the listener interprets to unpack the rich information transmitted by the speaker.
- The speaker also inadvertently shares tone, energy, speed, and other acoustic properties, which helps capture the subtext or intention and literal words.



LINK FOR LITERATURE SURVEY=https://www.overleaf.com/1213337642pcqdjgvxrjzv

How are you feeling?

Data Set Information;

(The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS));

Description ;

- The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7356 files (total size: 24.8 GB).
- The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent.
- Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.
- All conditions are available in three modality formats: Audio-only (16bit, 48kHz .wav), Audio-Video (720p H.264, AAC 48kHz, .mp4), and Video-only (no sound). Note, there are no song files for Actor_18.

- Speech file (Audio_Speech_Actors_01-24.zip, 215 MB) contains 1440 files: 60 trials per actor x 24 actors = 1440.

- Video files are provided as separate zip downloads for each actor (01-24, ~500 MB each), and are split into separate speech and song downloads:

File naming convention

➢ Each of the 7356 RAVDESS files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 02-01-06-01-02-01-12.mp4). These identifiers define the stimulus characteristics:

- 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).

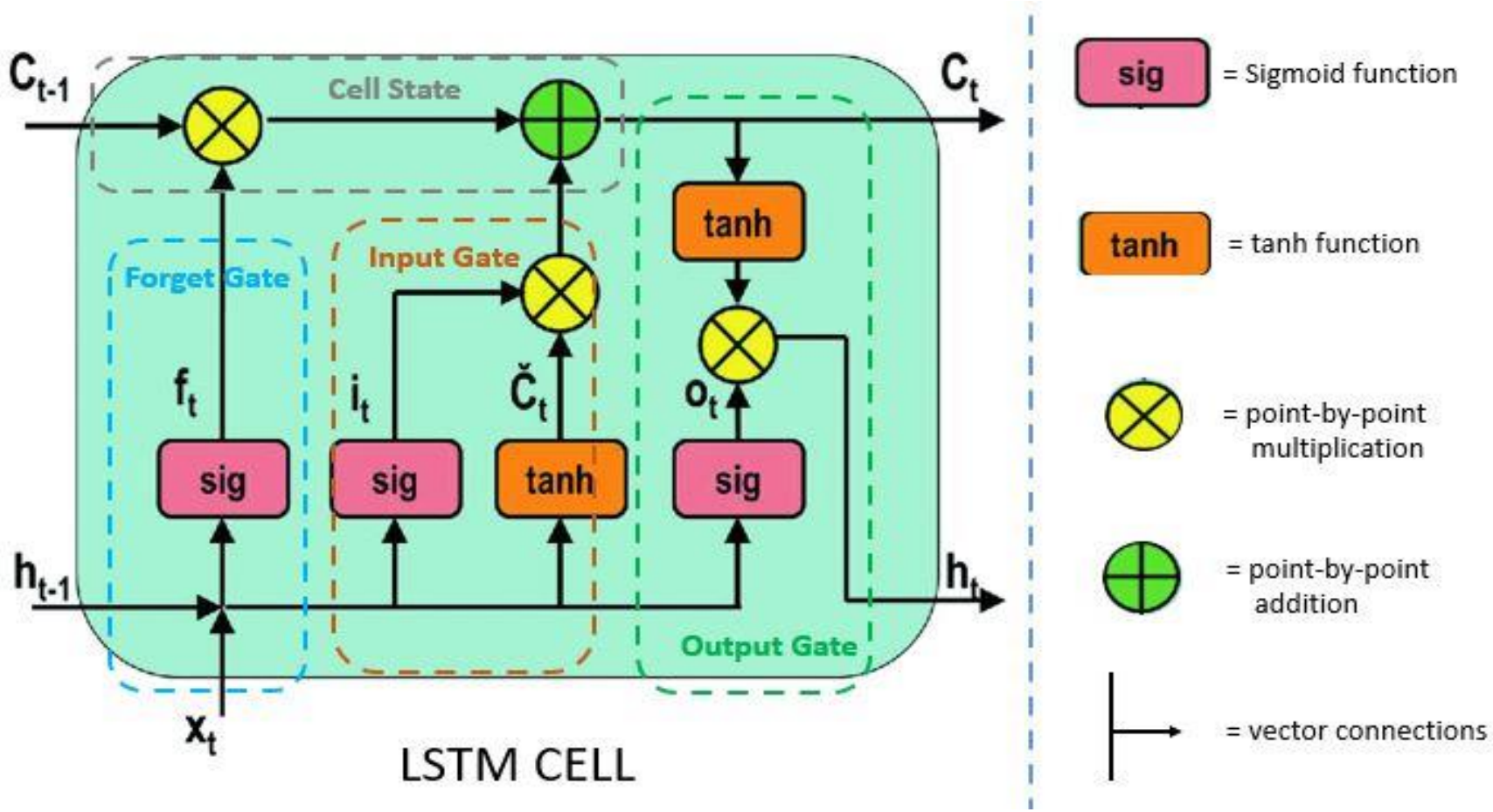
Provided link(Ravdess Dataset);

https://zenodo.org/record/1188976#.Y1EP53ZBxPZ

What is LSTM model (Long Short-Term Memory)?

- An LSTM module has a cell state and three gates which provides them with the power to selectively learn, unlearn or retain information from each of the units.
- The cell state in LSTM helps the information to flow through the units without being altered by allowing only a few linear interactions.
- The LSTM network has three layers, an input layer, a single hidden layer followed by a standard feedforward output layer.

LSTM Model block diagram ;



Purpose of using LSTM model;

- LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series.
- LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs.

LSTM model Applications;

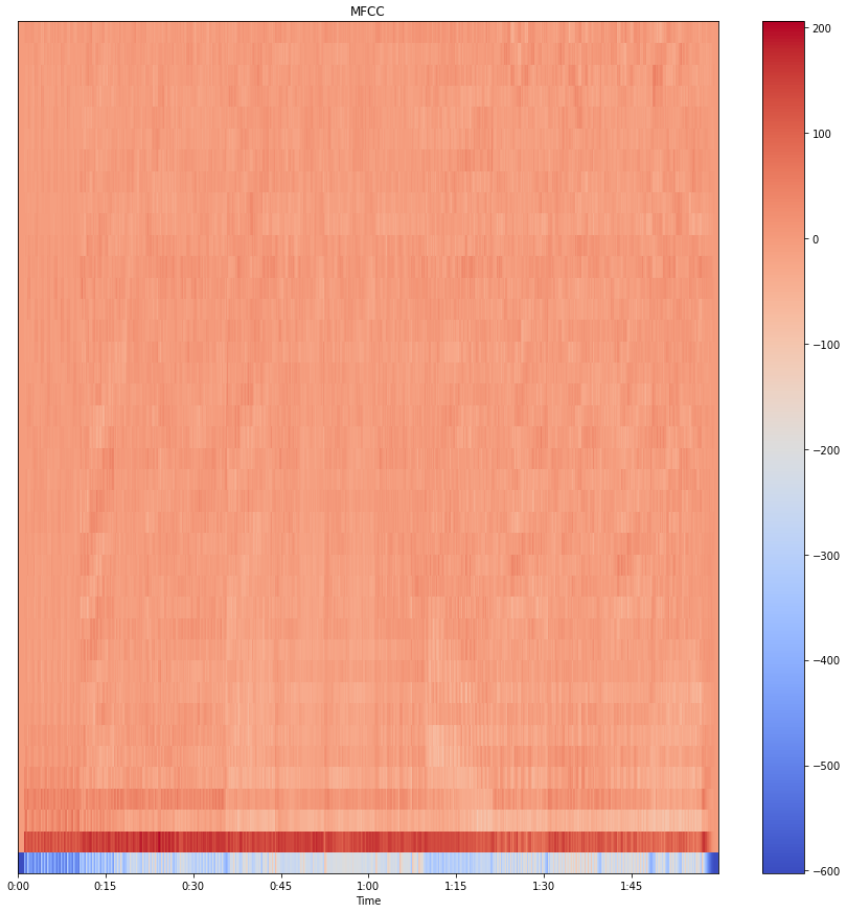
- Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems.
- This(LSTM) is a behavior required in complex problem domains like machine translation, speech recognition etc.
- LSTMs are a complex area of deep learning.

Features Extracted

- Mfccs
- Chroma
- Contrast

Mel-Frequency Cepstral Coefficients (MFCCs):

- It can be derived from a type of inverse Fourier transform(cepstral) representation
- Short term power spectrum of any sound represented by the Mel frequency cepstral (MFC) and combination of MFCC makes the MFC
- It can be derived by mapping the Fourier transformed signal onto the male scale using triangle or cosine overlapping windows
- It can be derived by mapping the Fourier transformed signal onto the male scale using triangle or cosine overlapping windows.
- Where after taking the logs of the powers at each of the Mel frequencies and after discrete cosine transform of the Mel log powers give the amplitude of a spectrum. The amplitude list is MFCC.

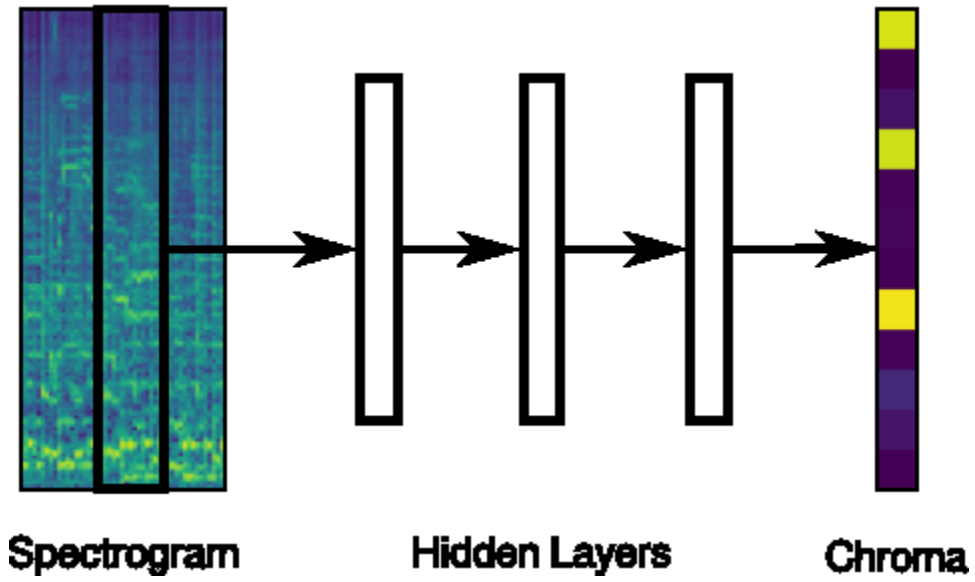


In the above image, we can see the MFCCs

Chroma

- The chroma feature is a descriptor, which represents the tonal content of a musical audio signal in a condensed form.
- In audio file analysis, an audio file can consist of 12 different pitch classes.
- These pitch class profiles are very useful tools for analyzing audio files. The term chromogram represents the pitches under an audio file, in one place so that we can understand the classification of the pitches in the audio files.
- Short Time Fourier Transforms and Constant Q Transforms are used for chroma feature extraction.
- One main property of chroma features is that they capture harmonic and melodic characteristics of music, while being robust to changes in timbre
- The feature vector is extracted from the magnitude spectrum by using a short time fourier transform(STFT),Constant-Q transform(CQT),Chroma Energy Normalized (CENS)

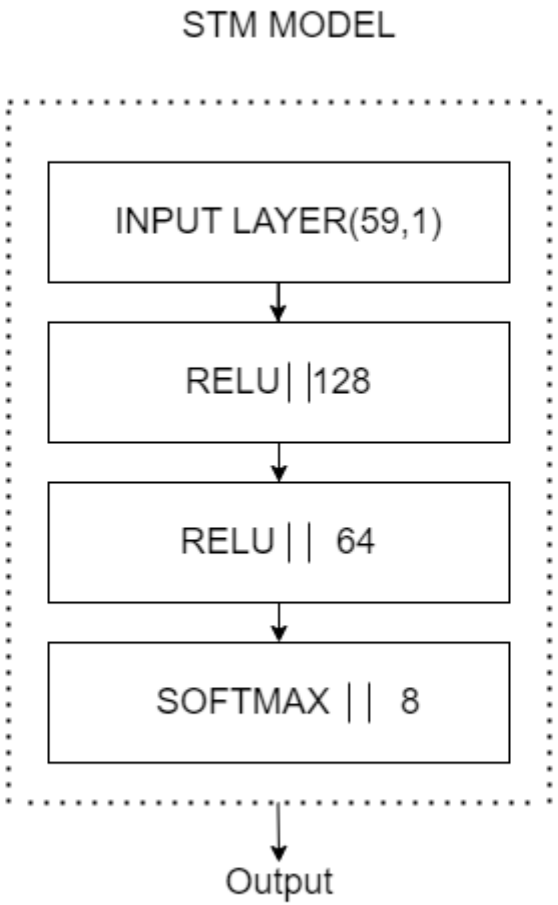
- It is some kind of measurement of the quality of the sound which helps in judging the sound as higher, lower, and medium



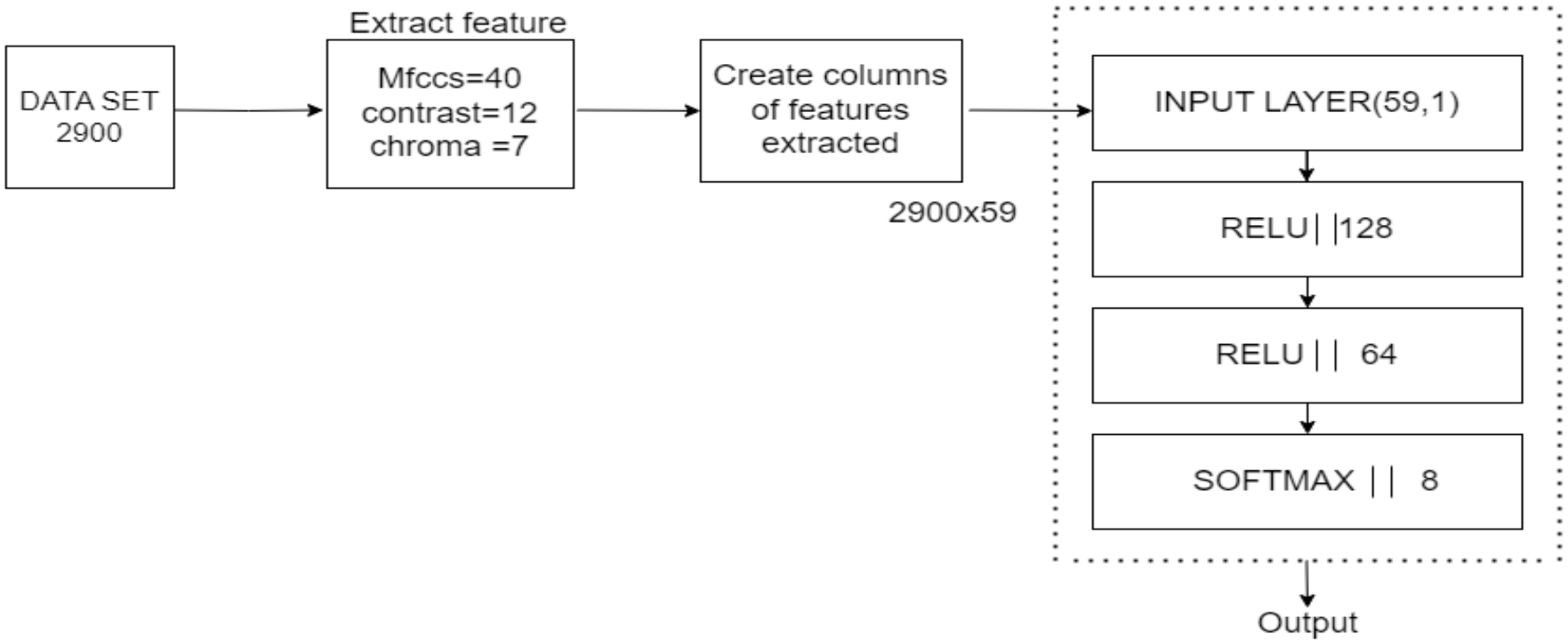
CONTRAST;

- Contrast is the difference in luminance or colour that distinguishes an object (or that item's depiction in an image or display).
- The distinction between the visual characteristics of objects in a composition is known as contrast. There are many different methods to use contrast, including through colour, line, shape, form, and context.

Proposed Methodology



Functional block diagram



THANK YOU