Research Article:

Speech emotion recognition system by deep learning

Anant Dev Pandey, Alisha Chhabra, Rohit Garg

Computer Science Department, Chandigarh University
NH-95, Ludhiana-Chandigarh State Highway Punjab, India.

[1]pandeyana8055@gmail.com

[2]alishachhabra1111@gmail.com

Abstract:

As of late, the significance of responding to the emotional condition of a client has been by and large acknowledged in the field of human-computer cooperation and particularly speech has gotten expanded center as a methodology from which to deduct data on feeling consequently. Up to this point, mostly scholarly furthermore, not very application-arranged disconnected investigations in light of recently recorded and clarified data sets with emotional speech were led. In any case, requests of online examination vary from that of disconnected examination, specifically, conditions are seriously difficult and less unsurprising. In the field of speech emotion recognition numerous strategies have been used to extricate emotions from signals, counting some deep rooted speech examination and grouping methods. In the customary method of speech emotion recognition highlights are removed from the speech signals and afterward the highlights are chosen which is all in all know as determination module and afterward the emotions are perceived this is an exceptionally extensive and time taking interaction so this paper gives an outline of the profound learning strategy which depends on a straightforward calculation in light of component extraction and model creation which perceives the emotion.

Keywords: Machine learning; speech emotion; MLP; python.

I. Introduction

Emotion assumes a critical part in everyday relational human communications. This is fundamental for our objective as well as keen choices. It assists us with coordinating and figure out the sensations of others by passing our sentiments and giving criticism on to other people. Research plays uncovered the strong part that emotion play in forming human social communication. Emotional presentations pass on impressive data about the psychological condition of a person. This has opened up another examination field called programmed emotion recognition, having essential objectives to comprehend and recover wanted emotions.

Investigations of programmed emotion recognition frameworks mean to make effective, ongoing techniques for recognizing the emotions of cell phone clients, call focus administrators and clients, vehicle drivers, pilots, and numerous other human-machine correspondence clients. Adding emotions to machines has been perceived as a basic consider causing machines to show up and act in a human-like way Robots equipped for understanding emotions could give fitting emotional reactions and display emotional characters. In certain conditions, people could be supplanted by computer-created characters being able to lead extremely normal and persuading discussions by engaging human emotions. Machines need to comprehend emotions conveyed by speech. Just with this ability, a totally significant discourse in light of common human-machine trust and understanding can be accomplished.

II. Overview

Speech handling typically works in a direct way on a sound sign [30]. It is considered critical and essential for different speech-based applications like SER,

speech denoising, and music arrangement. With late progressions, SER has acquired a lot importance. Be that as it may, it actually requires precise philosophies to impersonate human-like way of behaving for cooperation with people [31]. As examined before, a SER framework is comprised of different parts that incorporate element determination and extraction, highlight characterization, acoustic displaying, recognition per unit, and above all language-based demonstrating. The conventional SER frameworks commonly consolidate different arrangement models like GMMs and Well. The GMMs are used for outline of acoustic highlights of sound units, while, the Well are used for managing fleeting varieties event in speech signals.

2.1. Emotion Recognition: The Fundamentally, Emotion Recognition manages the review of inducing emotions, strategies utilized for surmising. Emotion can be perceived from looks, speech signals. Different methods have been created to see as the emotions, for example, signal handling, AI, neural networks, PC vision. Emotion examination, Emotion Recognition are being considered and fostered all around the world. Emotion Recognition is gaining its notoriety in research which is the way to take care of numerous issues too makes life more straightforward. The main need of Emotion Recognition from Speech is testing errands in Man-made reasoning where speech signals is separated from everyone else a contribution for the PC frameworks. Speech Emotion Recognition (SER) is additionally utilized in different fields like BPO Center and Call Center to distinguish the emotion helpful for recognizing the bliss of the client about the item, IVR Frameworks to improve the speech communication, to address different language ambiguities and adaption of PC frameworks as indicated by the mind-set and emotion of a person.

2.2. Speech emotion recognition: Speech Emotion Recognition is research region issue which attempts to induce the emotion from the speech signals.

Different review express that progression in emotion discovery will make part of frameworks simpler and subsequently making a world better spot to live. SER has its own application which is explained later. Emotion Recognition is the difficult issue in manners, for example, emotion might contrast in light of the climate, culture, individual face response prompts uncertain discoveries; speech corpus isn't sufficient to precisely surmise the emotion; absence of speech data set in numerous dialects.

2.3. How does Speech Emotion Recognition Work: Researchers apply different sound handling methods to catch this secret layer of data that can enhance and concentrate apparent and acoustic highlights from speech. Changing over sound signs into numeric or vector design isn't quite so direct as pictures. The change technique will decide how much urgent data is held when we forsake the "sound" design. In the event that a specific information change can't catch the delicateness and smoothness, it would be trying for the models to gain proficiency with the emotion and order the example. A few strategies to change sound information into numeric incorporate Mel Spectrograms that imagine sound signs in view of their recurrence parts which can be plotted as a sound wave and took care of to prepare a CNN as a picture classifier. We can catch this utilizing Mel-recurrence cepstral coefficients (MFCCs). Every one of these information designs has its advantages and weaknesses in light of the application.

2.4. Speech Recognition Applications: Uses of straightforward speech recognition are broad - YouTube auto-produced captions, live speech records, records for online courses, and astute voice-helped chatbots like Alexa and Siri. Along these lines, intensely committed research has yielded worthwhile and productive outcomes - YouTube auto-created captions work on every year. Be that as it may, utilizations of speech emotion recognition are more nuanced and add a fresher aspect to the utilization of man-made intelligence and how it can make our lives simpler to further develop them.

An exceptionally late use of SER has risen up out of the unexpected ascent in web based learning where teachers can notice an understudy's reaction in class and feature pointers that could end up being useful to them help the understudy's schooling. Another impending use is to assess competitors going after administrative jobs by dissecting their reactions during sound or video interviews. Their certainty or anxieties can be quantitatively estimated interestingly utilizing SER,

and hence recruiting supervisors can choose the competitor with the best fit.

2.5. Traditional techniques of SER: An emotion recognition framework in light of digitized speech is contained three central parts signal preprocessing highlight extraction and order. Acoustic preprocessing, for example, denoising as well as division is done to decide significant units of this sign. highlight extraction is used to distinguish the intriguing occasion include available in the sign. In conclusion, the planning of separated highlight vectors to pertinent emotion is completed by classifiers.

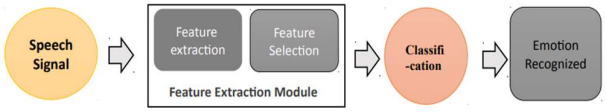

Figure. 1 Traditional Speech Emotion Recognition System

Figure 1 portrays a worked on framework used for speech-based emotion recognition. In the principal phase of speech-based signal handling, speech improvement is completed where the uproarious parts are eliminated. The subsequent stage includes two sections, highlight extraction, an element determination. The required highlights are removed from the preprocessed speech signal and the choice is made from the extricated highlights. Such component extraction and choice are normally founded on the investigation of speech signals in the time and recurrence domains. During the third stage, different classifiers such as GM Mand HMM, and so on are used for the characterization of these highlights. Finally, in light on highlight arrangement various emotions are perceived.

III. Problem Definition

These techniques required tremendous designing highlights and any variety in the elements would need re-demonstrating the general design of the method. All things considered, late advancement in deep learning applications and strategies for Search Emotion Recognition can be fluctuated too.

There are various writing and concentrates on the use of these calculations to comprehend emotions and perspective from human discourse. Furthermore, to deep learning, neural networks, what's more, utilization of upgrades of long momentary memory (LSTM) networks, generative ill-disposed models, and parts more, a wave in research on discourse emotion recognition and its application presently arises. Understanding its application and its job in emotion is fundamental. For this explanation, the target of the ongoing paper is to see deep learning procedures for discourse emotion recognition, from data sets to models. In the wake of applying the deep learning and highlight extraction strategies further, getting high precision in the model is truly challenging in light of the similitudes between the various emotions like cheerful and astonishing emotions have a similar sort of recurrence and tone. The length of the voice is likewise an issue since we all realize that the human emotions don't continue as before all through the sentence it keeps on changing so the framework needs to distinguish the pieces of the information to comprehend the full emotion of the voice.

3.1 Process

Planning & Requirements: As with most any upholding projects that we see in our daily life, the first step is to go through an initial planning stage to sort out the various documentaries, and establishing more connections so that each query can be sorted out.

Analysis & Design: After the completion of planning, an analysis is performed to keep down the preferred logic, different working models. The design stage is here which is doing its own working, establishing any requirement I technical case (languages, data layers, services, etc.) that we could use by this stage

Implementation: With planning and analysis now we will be doing the final implementation after this work. All planning, specification, and design docs up to the current point are coded and are embedded into this working of the project.

Testing: After embedding this iteration, next step is to travel through a series of testing step to trace and find any potential bugs or problems that have cropped up.

Evaluation: After completing all stages till up to now, it is time for a thorough evaluation of development up to this stage. It tells us to verify the items which are caught from other sourced files and documents.

3.2 Advantages

Inherent Versioning: It is rather obvious that most software creation teams needs the versioning by any case, indicating the release stage of the software at any particular stage. However, the iterative model makes our work easier and we think that the future iterations will surely improve our work by any further cases.

Easy Adaptability: By the use of adaptive culture, we will be using our model in Blockchain to gain adaptability in this modern world with technology.

Below is a brief overview of the solutions to satisfy these structures in online voting systems.

Privacy

Privacy with respect of social media means that no other than the authorized user of account can manipulate anything life comments, post, etc.

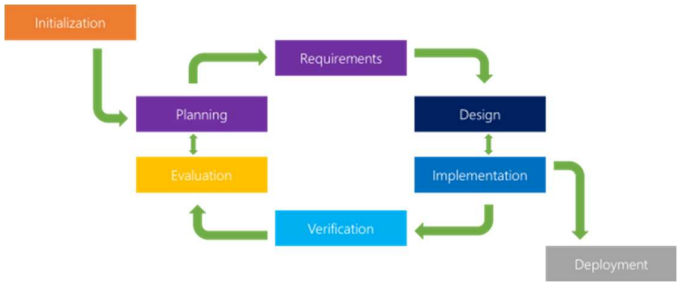

Fig. 2 Iterative model of software development

IV. DATASET FOR SPEECH EMOTIONS

In the field of affect discovery, a vital job is played by reasonable decision of speech data set. Three data sets are utilized for good emotion recognition the framework as given beneath.

4.1. Elicited emotional speech database:

For this situation emotional circumstance is made falsely by gathering information from the speaker.

- Advantage: This sort of information base is like a regular information base
- Issue: There is unavailability of all emotions and if the speaker is familiar with it that they are being recorded then fake emotion can be communicated by them.

4.2. Actor based speech database: Trained and proficient craftsmen gather this kind of speech dataset.

- Advantage: In this data set wide assortment of emotions are available and it is additionally extremely simple to gather it.
- Issue: It is especially fake and occasional in nature.

4.3. Natural speech database: Trained and proficient craftsmen gather this kind of speech dataset.

- Advantage: In this data set wide assortment of emotions are available and it is additionally extremely simple to gather it.
- Issue: It is especially fake and occasional in nature

V. FEATURE EXTRACTION FOR SPEECH EMOTION RECOGNITION

There are number of methods for feature extraction like Linear predictive cepstral coefficients (LPCC), Power spectral analysis (FFT), First order derivative (DELTA), Linear predictive analysis (LPC), Mel scale cepstral analysis (MEL), perceptual linear predictive coefficients (PLP) and Relative spectra filtering of log domain coefficients (RASTA).

5.1. Linear predictive coding (LPC): In encoding quality speech at a low piece rate LPC strategy is helpful that is one of the most impressive procedures of speech examination. At current time explicit speech test can be approximated as a direct mix of past speech tests is the essential thought behind straight prescient examination. It is a human speech creation base model that uses a customary source channel model. Vocal plot acoustics are reenacted by Lip radiation, vocal plot and that's what glottal exchange works are incorporated into one all shaft channel. Over a limited length the amount of squared contrasts between assessed and unique speech signal is limited involving LPC that aides in having exceptional arrangements of indicator coefficients.

In genuine recognition real indicator coefficients are not utilized as a high difference is shown by it. There is change of indicator coefficient to a cepstral coefficients more hearty arrangement of boundaries. Some of the sorts of LPC are lingering excitation, normal heartbeat energized, pitch excitation, voice excitation and coded energized LPC.

5.2. Mel frequency cepstral coefficients (MFCC): It is considered as one of the standard technique for include extraction and in ASR most normal is the utilization of 20 MFCC coefficients. Despite the fact that for coding speech utilization of 10-12 coefficients are adequate and it rely upon the ghostly structure because of which it is more delicate to commotion. This issue can defeat by utilize more data in speech signals periodicity despite the fact that aperiodic substance is additionally present in speech. Genuine cesptal of windowed brief time frame quick Fourier change (FFT) signal is address by MFCC. Non direct recurrence is use. The boundaries like people utilized for hearing speech are utilized to extricates boundaries utilizing sound element extraction MFCC method. Other data is deemphasizes and inconsistent number of tests contain time spans are used to partition speech signals. Covering from casing to outline is utilized to smooth the progress in most frameworks and afterward hamming window is utilized to wipe out the discontinuities from each time period.

5.3. Perceptual linear prediction (PLP): Hermansky fostered a PLP model that utilizes psychophysics idea of hearing to display a human speech. The speech recognition rate gets improved by disposing of superfluous data by PLP. Ghastly qualities are changed to human hear-able framework match is the main thing that makes PLP not quite the same as LPC. The force din power-regulation connection, equal loudness bend and basic band goal bends are three main perceptual viewpoints approximates by PLP.

5.4. Mel scale cepstral analysis (MEL): PLP examination also, MEL examination is like each other in which psychophysically based phantom changes is used to adjust the spectrum. As indicated by the scale of MEL a spectrum is enclosed by this technique on other hand as per bark scale a spectrum is distorted in PLP. So yield cepstral coefficients are the main different between scale cepstral investigation of PLP and MEL. The altered power spectrum is smooth utilizing all post model in PLP and afterward based on this model a result cepstral coefficients are processed. On other hand adjusted power spectrum is smooth utilizing cepstral smoothing in MEL scale cesptral examination. In this Discrete Fourier Change (DFT) is utilized to convert log power spectrum is straightforwardly change into capstal domain.

VI. COMMONLY USED ALGORITHM

The best elements come after highlights estimation is given to the classifier. In articulation of speaker's speech an emotion is perceives by classifier and for speech emotion recognition number of classifiers have been proposed by different researchers. In this segment audit of a portion of the classifier has been given.

6.1. K-Nearest Neighbours (KNN): Automated speech administrations like intelligent voice recognition frameworks have utilized speech based emotion recognition. In mental sorrow like clinical applications, lie indicators like insightful application utilization of speech administrations play incredible ramifications. Renjith S, et.al, (2017), have chipped away at Telugu and Tamil dialects to recognize emotions satisfaction, misery and outrage utilizing speech accounts [23]. In their work they have pre-handled to isolate unsettling influences from speech waveforms and crude speech signals. They have removed Hurst and Linear Prescient Cepstral Coefficients (LPCC) elements and afterward characterization is done based on measurable boundaries obtained from these highlights. Both KNN and ANN is utilized to recognize the responsive emotions and afterward precision, accuracy and review boundary is utilized to look at their exhibition for the two elements exclusively and in blend. When contrasted with LPCC when utilization of Hurst gives improved results when tried for person highlights with regards to review, accuracy and exactness.

6.2. Naive Bayes classifier: In human correspondence an significant job is played by emotion as sentiments can be effectively pass on through it. In speech handling domain, emotion recognition from speech has become a difficult and significant area of research. This undertaking has become significantly more testing due to arrangement of a few emotional classes from removed appropriate elements from speech. Naive Bayes classifier is utilized by Atreyee Khan, et.al, (2017), along with both phantom and prosodic elements for emotion location. As ghastly elements a Mel-Recurrence Cepstral Coefficients (MFCC) has been

utilized and pitch is utilized as prosodic component. Guileless Bayes Classifier is utilized to perform arrangement and they have considered seven emotional classes to create both orientation free and subordinate framework. Berlin Emotional db famous speech information base speech tests are utilized to test exactness of the framework later performing arrangement. Arrangement of sound sign into four essential emotional state is executed by S. K. Bhakre, et.al, (2016) by taking into account MFCC, picth, ZCR and energy measurable elements from 2000 expressions of the made sound sign data set. Normal greatness distinction technique (AMDF) is used to separate pitch highlights and greatness spectrum amount of square outright worth is utilized to compute energy. Energies spectrum of Discrete cosine change (DCT) is utilized to compute MFCC in which they have considered just 1-14 coefficients of DCT and rest is disposed of.

6.3. Support Vector Machine (SVM) classifier: Human
computer interaction (HCI) subset programmed emotion furthermore, speech recognition has become broadly researched point with the coming of digitization of each conceivable road. As we grasp machines, machine has additionally perceived us as humble positions are taken with machines. From given example plentifulness, pitch and MFCC highlights are separated and it run across developing and existing data set of training tests. Ashwini Rajasekhar, et.al, (2018), have recognized the given example utilizing SVM and speaker expression is identified utilizing MFCC. In the end SVM classifier separates between dread, outrage, trouble, bliss and updates the data set as needs be. Amiya Kumar, et.al, (2015), have presented a clever methodology by consolidating MFCC, LPCC inferred highlights, energy, ZCR, pitch prosody highlights, MEDC dynamic elements for programmed recognition of speaker's emotion state. Then, at that point, cheerful, shock, outrage, miserable, nausea, unbiased and dread are seven discrete emotional states distinguished utilizing staggered SVM classifier in five local assamese dialects. The proposed approach is assessed for blend of highlights regarding exactness that shows a decent outcome for speaker free cases when contrasted with individual highlights.

6.4. Convolution Neural network (CNN) classifier: Extraction of speech emotion highlights are major class of speech emotion recognition so Li Zheng, et.al, (2018), have proposed an irregular woods and CNN based new organization model (CNN-RF) (Zheng, L.,2018). From standardized spectrogram a speech emotion highlights are separated utilizing CNN and afterward speech emotion highlights are arrange utilizing RF order calculation. From results it has been anticipated that when contrasted with conventional CNN model utilization of CNN-RF model gives further developed results and it additionally further develops the Nao record sound order box. At last, Nao robot can "attempt to sort out" a human's brain science through speech emotion recognition and likewise have some familiarity with individuals' satisfaction, outrage, misery, furthermore, delight, accomplishing a more savvy human-PC collaboration.

6.5. Recurrent neural network: Move of each expression clear cut mark into a name grouping is the test which should be considered while demonstrating the all out speech emotion recognition errands in a consecutive methodology. Thus, Xiaomin Chen, et.al, (2018), need to make a speculation. In which both non-emotional and emotional portions comprise of expression then again (Chen, X., 2018). On the premise of that speculation, they have treated an expression name grouping as a chain of nulls indicating non-emotional casings and emotional states meaning emotional casings are two sorts of states. For naturally arrangement and mark an expression emotional sections with emotional marks a connectionist fleeting grouping based intermittent neural organization (CTC-RNN) is taken advantage of and the equivalent is utilized for non-emotional sections with non emotional names. The proposed strategy is tried on IEMOCAP corpus that shows the viability of it as contrasted with cutting edge emotion recognition algorithm.

S. No.	Author name	Classifier	Database
1	Renjith S, et.al, (2017),	kNN and ANN	Amritaemo
2	Steven A. Rieger Jr, et.al, (2014),	KNN	LDC emotional prosody speech database
3	Atreyee Khan, et.al, (2017),	Naive Bayes	Berlin Emo-db
4	Sagar K. Bhakre, et.al, (2016),	Naive Bayes	They have made dataset by considering 2000 sentences audio signal from 20 different speakers.
5	Ashwini Rajasekhar, et.al, (2018),	SVM	They have used computerized voice dataset
6	Amiya Kumar, et.al, (2015),	Multilevel SVM classifier	Utterances of "Multilingual Emotional Speech Database of North East India" (MESDNEI).
7	Li Zheng, et.al, (2018),	Convolution Neural Network combined with Random Forest (CNN-RF)	RECOLA natural emotion database
8	Weibkirchen, et.al, (2017),	CNN	we utilised three data sets Berlin Emotional Speech Database (EmoDB), eNTERFACE and Speech Under Simulated and Actual Stress (SUSAS)
9	Xiaomin Chen, et.al, (2018),	Connectionist temporal classification based recurrent neural network (CTC-RNN)	IEMOCAP corpus

Figure. 3 Comparison table of different classifier

VII. IMPLEMENTATION AND CODE

Step: 1 We import the necessary library.

```
import librosa
import soundfile
import os, glob, pickle
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import accuracy_score
```

Step: 2 Now we declare a function extract_feature to extract the chroma, mel and mfcc feature from sound. This capability takes 4 boundaries the document name and three Boolean boundaries for the three elements:

- **mfcc:** Mel Frequency Cepstral Coefficient, represents the short-term power spectrum of a sound
- **chroma:** Pertains to the 12 different pitch classes
- **mel:** Mel Spectrogram Frequency

```
#DataFlair - Extract features (mfcc, chroma, mel)
from a sound file
def extract_feature(file_name, mfcc, chroma, mel):
    with soundfile.SoundFile(file_name) as
sound_file:
        X = sound_file.read(dtype="float32")
        sample_rate=sound_file.samplerate
        if chroma:
            stft=np.abs(librosa.stft(X))
        result=np.array([])
        if mfcc:
            mfccs=np.mean(librosa.feature.mfcc(y=X,
sr=sample_rate, n_mfcc=40).T, axis=0)
            result=np.hstack((result, mfccs))
        if chroma:
            chroma=np.mean(librosa.fea-
ture.chroma_stft(S=stft, sr=sample_rate).T,axis=0)
            result=np.hstack((result, chroma))
    if mel:
            mel=np.mean(librosa.feature.melspectro-
gram(X, sr=sample_rate).T,axis=0)
            result=np.hstack((result, mel))
    return result
```

Step: 3 We define the dictionary to hold the digit & emotion present in RAVDESS datasets, and a list to carry – calm, happy, fearful, disgust.

```
#DataFlair - Emotions in the RAVDESS dataset
emotions={
  '01':'neutral',
  '02':'calm',
  '03':'happy',
  '04':'sad',
  '05':'angry',
  '06':'fearful',
  '07':'disgust',
  '08':'surprised'
}

#DataFlair - Emotions to observe
observed_emotions=['calm', 'happy', 'fearful', 'disgust']
```

Step: 4 Now we will stack the function load_data() – it takes in the overall size of test as parameter x and y are unfilled list; wee use glob() f(x) from the module to extract all pathnames for file. The patter use for this is ravdess data\\Actor_**.wav".

```
#DataFlair - Load the data and extract features for
each sound file
def load_data(test_size=0.2):
    x,y=[],[]
    for file in glob.glob("D:\\DataFlair\\ravdess
data\\Actor_*\\*.wav"):
        file_name=os.path.basename(file)
        emotion=emotions[file_name.split("-")[2]]
```

```
        if emotion not in observed_emotions:
            continue
        feature=extract_feature(file, mfcc=True,
chroma=True, mel=True)
        x.append(feature)
        y.append(emotion)
    return train_test_split(np.array(x), y,
test_size=test_size, random_state=9)
```

Step: 5 Time to split the dataset into training and testing sets! Let's keep the test set 25% of everything and use the load_data function for this.

```
#DataFlair - Split the dataset
x_train,x_test,y_train,y_test=load_data(test_size=0.25)
```

Step: 6 Observe the shape of the training and testing datasets:

```
#DataFlair - Get the shape of the training and testing
datasets
print((x_train.shape[0], x_test.shape[0]))
```

Step: 7 And get the number of features extracted.

```
#DataFlair - Get the number of features extracted
print(f'Features extracted: {x_train.shape[1]}')
```

Step: 8 Now, let's initialize an MLPClassifier. This is a Multi-layer Perceptron Classifier; it optimizes the log-loss function using LBFGS or stochastic gradient descent. Unlike SVM or Naïve bayes the MLPClassifier has an internal neural network for the purpose of classification. This is a feedforward ANN model.

```
#DataFlair - Initialize the Multi Layer Perceptron
Classifier
model=MLPClassifier(alpha=0.01, batch_size=256,
epsilon=1e-08, hidden_layer_sizes=(300,), learning_rate='adaptive', max_iter=500)
```

Step: 9 Fit/train the model.

```
#DataFlair - Train the model
model.fit(x_train,y_train)
```

Step: 10 Let's predict the values for the test set. This gives us y_pred (the predicted emotions for the features in the test set).

```
#DataFlair - Predict for the test set
y_pred=model.predict(x_test)
```

VIII. ADVANTAGES AND DISADVANTAGES

8.1. Advantages:

- Furnishes the adaptability to work with nonlinear qualities
- Less number of parameters required
- Can deal with missing qualities, model complex connections
- furthermore, support various information sources
- Better classification of parameters is shown.

8.2. Disadvantages:

- MLPs generally need fixed number of contributions to be given for fixed number of results, there is a decent planning capability between the data sources and the results in these feed-forward neural networks that represent an issue when a succession of inputs is given to the model.
- Network should be retrained when another emotion is added to the framework

IX. CONCLUSION

In this task we have attempted to break down certain examples of speech utilizing the deep learning method. First and foremost we stacked the datasets then we envisioned the different human emotions utilizing our capabilities waveshow and spectrogram utilizing the Librosa library. Then, at that point, we removed the acoustic highlights of every one of our examples utilizing the MFCC strategy and organized the successive information obtained in the 3D cluster structure as acknowledged by the LSTM model. Then, at that point, we fabricate the LSTM model and in the wake of training the model we envisioned the information into the graphical structure utilizing matplotlib library and after some continued testing utilizing various qualities the typical exactness of the model is viewed as 73%. Upgrade of the strength of emotion recognition framework is as yet conceivable by joining data sets and by combination of classifiers. The impact of training numerous emotion identifiers can be researched by melding these into a solitary recognition framework. We aim likewise to utilize other element choice techniques on the grounds that the nature of the component determination influences the emotion recognition rate: a decent emotion include determination strategy can choose highlights reflecting emotion state rapidly. The general aim of our work is to foster a framework that will be utilized in an educational connection in study halls, to assist the

educator with organizing his class. For accomplishing this objective, we aim to test the framework proposed in this work.

X. References

1. H. Cao, R. Verma, and A. Nenkova, “Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech,” Comput. Speech Lang., vol. 28, no. 1, pp. 186– 202, Jan. 2015.
2. L. Chen, X. Mao, Y. Xue, and L. L. Cheng, “Speech emotion recognition: Features and classification models,” Digit. Signal Process., vol. 22, no. 6, pp. 1154–1160, Dec. 2012
3. T. L. Nwe, S. W. Foo, and L. C. De Silva, “Speech emotion recognition using hidden Markov models,” Speech Commun., vol. 41, no. 4, pp. 603–623, Nov. 2003.
4. S. Wu, T. H. Falk, and W.-Y. Chan, “Automatic speech emotion recognition using modulation spectral features,” Speech Commun., vol. 53, no. 5, pp. 768–785, May 2011.
5. J. Rong, G. Li, and Y.-P. P. Chen, “Acoustic feature selection for automatic emotion recognition from speech,” Inf. Process. Manag., vol. 45, no. 3, pp. 315–328, May 2009.
6. C.-H. Wu and W.-B. Liang, “Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels,” IEEE Trans. Affect. Comput., vol. 2, no. 1, pp. 10–21, Jan. 2011.
7. S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” IEEE Trans. Speech Audio Process., vol. 13, no. 2, pp. 293–303, Mar. 2005.
8. B. Yang and M. Lugger, “Emotion recognition from speech signals using new harmony features,” Signal Processing, vol. 90, no. 5, pp. 1415–1423, May 2010
9. E. M. Albornoz, D. H. Milone, and H. L. Rufiner, “Spoken emotion recognition using hierarchical classifiers,” Comput. Speech Lang., vol. 25, no. 3, pp. 556–570, Jul. 2011.
10. C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, “Emotion recognition using a hierarchical binary decision tree approach,” Speech Commun., vol. 53, no. 9–10, pp. 1162–1171, Nov. 2011.
11. C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, “Emotion recognition using a hierarchical binary decision tree approach,” Interspeech, vol. 53, pp. 320–323, 2009.
12. S. Bjorn, S. Steidl, and A. Batliner, “The INTERSPEECH 2009 Emotion Challenge,” 2009.
13. J.-H. Yeh, T.-L. Pao, C.-Y. Lin, Y.-W. Tsai, and Y.-T. Chen, “Segment-based emotion recognition from continuous Mandarin Chinese speech,” Comput. Human Behav., vol. 27, no. 5, pp. 1545–1552, Sep. 2011.
14. W. Dai, D. Han, Y. Dai, and D. Xu, “Emotion Recognition and Affective Computing on Vocal Social Media,” Inf. Manag., Feb. 2015. 15. M. M. H. El Ayadi, M. S. Kamel, and F. Karray, “Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models,” in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP ’07, 2007, vol. 4, pp. IV–957–IV– 960. 16. J. P. Arias, C. Busso, and N. B. Yoma, “Shape-based modeling of the fundamental frequency contour for emotion detection in speech,” Comput. Speech Lang., vol. 28, no. 1, pp. 278– 294, Jan. 2014.
17. M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, “Primitives-based evaluation and estimation of emotions in speech,” Speech Commun., vol. 49, no. 10–11, pp. 787–800, Oct. 2007.
18. Sucksmith, E., Allison, C., Baron-Cohen, S., Chakrabarti, B., & Hoekstra, R. A. Empathy and emotion recognition in people with autism, first-degree relatives, and controls. Neuropsychologia, 51(1), 98-105,2013.
19. Hadhami Aouani et al. / Procedia Computer Science 176 (2020) 251–260.
20. M. Swain, A. Routray, and P. Kabisatpathy, ``Databases, features and classifiers for speech emotion recognition: A review," Int. J. Speech Technol., vol. 21, no. 1, pp. 93-120, 2018.
21. D. Ververidis and C. Kotropoulos, ``A state of the art review on emotional speech databases," in Proc. 1st Richmedia Conf., 2003, pp. 109-119.
22. P. Jackson and S. Haq, Surrey Audio-Visual Expressed Emotion (SAVEE) Database. Guildford, U.K.: Univ. Surrey, 2014.
23. F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, ``Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in Proc. IEEE 10th Int. Conf. Workshops Autom. Face Gesture Recognit. (FG), Apr. 2013, pp. 1-8.
24. R. Cowie, E. Douglas-Cowie, and C. Cox, ``Beyond emotion archetypes:Databases for emotion modelling using neural networks," Neural Netw.,vol. 18, no. 4, pp. 371-388, 2005.

25. T. Vogt and E. André, ``Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Jul. 2005, pp. 474-477.
26. C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, ``Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," Artif. Intell. Rev., vol. 43, no. 2, pp. 155-177, 2015
27. A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, and N. Amir, ``The automatic recognition of emotions in speech," in Emotion-Oriented Systems. Springer, 2011, pp. 71-99
28. E. Mower, M. J. Mataric, and S. Narayanan, ``A framework for automatic human emotion classification using emotion profiles," IEEE Trans. Audio, Speech, Language Process., vol. 19, no. 5, pp. 1057-1070, Jul. 2011.
29. J. Han, Z. Zhang, F. Ringeval, and B. Schuller, ``Prediction-based learning for continuous emotion recognition in speech," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2017, pp. 5005-5009.
30. Y. LeCun, Y. Bengio, and G. Hinton, ``Deep learning," Nature, vol. 521, no. 7553, p. 436, 2015.
31.W. Wang, Ed., Machine Audition: Principles, Algorithms and Systems. Hershey, PA, USA: IGI Global, 2010.