Speech Emotion and Voice Gender Recognition Using Machine Learning

Disha Jayaprakash T, Diya Deb, Robin Kumar Mishra, Vinod B. Durdi, Ravikanti Swetha , Preeti Khanwalkar

Department of Electronics & Telecommunication Engineering, Dayananda Sagar College of Engineering, Bangalore-111, India
dishajayaprakash94@gmail.com

Department of Electronics & Telecommunication Engineering, Dayananda Sagar College of Engineering, Bangalore-111, India
diya.deb2506@gmail.com
Department of Electronics & Telecommunication Engineering, Dayananda Sagar College of Engineering, Bangalore-111, India
robinkmishra2016@gmail.com
Department of Electronics & Telecommunication Engineering, Dayananda Sagar College of Engineering, Bangalore-111, India
vinoddurdi-tce@dayanandasagar.edu

Department of Electronics & Telecommunication Engineering, Dayananda Sagar College of Engineering, Bangalore-111, India
swetha-tce@dayanandasagar.edu

Department of Electronics & Telecommunication Engineering, Dayananda Sagar College of Engineering, Bangalore-111, India
preeti-tce@dayanandasagar.edu

***Abstract*—** Speech recognition can be used for a variety of tasks, such as communication between humans and machines, labeling call by gender, tagging videos, and more. Speech emotion recognition is one component of voice processing that has steadily been growing in popularity. With the advancement voice-controlled interfaces, the significance of emotion recognition is growing. The development of speech-based emotion recognition systems has advantages for practical applications. Speech inputs are processed to extract Mel-frequency cepstrum coefficients (MFCC) and modulation spectral (MS) characteristics, and these parameters are used as inputs to a multilayer perceptron model. In this research work, we use Python's Tensor Flow framework to attempt to categorize gender based on speech. Many industries, including ASR, can benefit from gender recognition because it can raise the effectiveness of these systems. It can also be used to group calls according to the gender, or you can incorporate it as a function into a virtual assistant that can identify the gender of the talker. In order to try and identify a person's emotions and gender without directly asking them, we will be integrating these two into a single project.

***KEYWORDS*—** *Voice gender recognition, Multilayer perceptron networks, Emotion recognition, Neural network*

I. INTRODUCTION

Speech signals have evolved into a form of machine-human communication in the current digital era, made possible by a number of technological advances. Speech-to-Text (STT) technology, that is utilized in mobile phones as a means to facilitate human-computer interaction, is a result of the combination of speech recognition approaches and signal processing techniques [1]. The fastest-growing area of research in attempts to identify speech signals is speech recognition. Detecting the emotion of a speech signal, also called Speech Emotion Recognition (SER) is a growing study area as a result of high demand for processing of speech signals even if emotions are easily detected from visual aids . It has the potential to advance several fields, including automatic translation systems, machine-human interaction, and speech-to-text synthesis.Emotion identification from speech has developed from a specialized field to a crucial element in Human-Computer Interaction. SER systems aim to facilitate the natural communication with machines through direct voice interface rather than utilizing traditional devices as input to grasp what is being said and make it easier for listeners to reply. Voice User Interfaces are used in several applications, such as call center discussions, GPS systems on cars, use of patterns detected in speech in medicine etc. Human emotional state assessment is a unique task that be the starting point of all emotion recognition models. A discrete emotion detection approach makes use of a range of emotions, including neutrality, boredom, contempt, surprise, fear, joy, and melancholy.

In order for a speech emotion recognition system to be successful, there are three essential features: a suitable dataset to train our model with, extracting the correct features and selection of an appropriate machine learning model. But the biggest hurdle when it comes to SER is the extraction of appropriate features from the speech signal [1-2].

The human voice signal is analog in nature and in order for the machine to understand it better; it needs to converted into a digital signal before feature extraction [3]. The next step is to create a classification model since the gender of a voice sample is being detected. The soundness of features that are dependent on a training set using machine learning (ML) [4-5] techniques determine the robustness and efficacy of

classifiers. As a result, since the human voice is susceptible to producing unhelpful features, eliciting speech features is essential for increasing the effectiveness of classifiers. Numerous studies have been done on how to extract valuable features from voice, such as determining the verbal components of a voice signal and eliminating useless information like background noise, in order to increase the efficiency of the chosen machine learning model [6-7]. The gender of the voice can be determined using a collection of features like Tonnetz, Modulation spectral coefficients, Pitch, Mel-frequency cepstral coefficients (MFCCs) [8-10], Chroma, Contrast [11] etc. A model for detecting the gender of a voice sample is created using machine learning approaches with the help of extracted features which are labeled and used as the training set.

One of the main issues in the field of speech analysis nowadays is gender classification [12-15]. The ability to determine gender from the acoustic characteristics of the voice, such as mean, median, frequency, etc., is crucial. Because machine learning produces positive outcomes for classification techniques, it is employed to resolve this issue. The gender can be predicted using a variety of algorithms based on audio characteristics [16-20]. There are various applications where gender recognition can be useful. Some of these include:

- for more human sound recognition, such as male laughter and female singing
- adding tags to audio and video files, categorizing them, and condensing and narrowing the search space
- automated greetings
- can aid in the question's resolution for personal assistants like Siri and Google Assistant
- with results for generic male or female

The paper is divided into the following sections: Section II briefly describes the extensive literature survey conducted on the topic. In section III, the proposed methodology is discussed. In section IV, the intermediate results are discussed. Finally, the conclusions are drawn in section IV.

II. LITERATURE SURVEY

[1] This paper provides a comprehensive evaluation of various deep learning techniques for Speech Emotion Recognition, such as Deep Boltzmann Machine (DBM), Recursive Neural Network (RvNN), Recurrent Neural Network (RNN), Deep Belief Network (DBN), Convolutional Neural Network (CNN), and Auto Encoder (AE), in the context of identifying emotions such as happiness, joy, sadness, surprise, boredom, disgust, fear, and anger. The paper briefly explains the architecture of these techniques and how they make efficient use of weights assigned to each layer. However, deep learning techniques have some drawbacks, including large architecture that can lead to over-learning of information and inflexibility in processing dynamic input data. This research lays the foundation for assessing the effectiveness and limitations of existing deep learning approaches and proposes ways to improve SER systems.

[2] A method for implementing a Multilayer Perceptron (MLP) deep learning model is presented by Mucahit Buyukyilmaz and Ali Osman Cibikdiken and is used to identify voice gender. 3,168 samples of male and female voices were recorded in total. Gender-specific features are discovered using an MLP deep learning system. During the testing phase, the MLP model had an accuracy rate of 96.74%. The voice's acoustic analysis is influenced by factors such as intensity, duration, frequency, and filtering. The gender of the speaker can be determined by characteristics present in one's voice. Acoustic analysis has been done using the WarbleR R biacoustic program. This study can be used to acquire the data set that contains acoustic parameters.

[3] This paper has provided a detailed review of the different models for SER systems. Different learning models and unique feature extraction methods are developed and deployed. The main issues that need to be addressed before a SER system can be called successful are: (a) Choosing an appropriate database of voice samples which can be used to detect emotions, (b) Extraction of the correct features effectively and (c) Selection of an appropriate machine learning model to classify emotions. This paper describes various models to classify six basic emotions of anger, disgust, terror, happiness, sadness and surprise, and modulation spectral features (MSF) and Mel-frequency cepstrum coefficient (MFCC) are used to extract the features of emotions from the speech signal.

[4] Speech emotion recognition is one of the trending topics gaining popularity in the Machine Learning domain nowadays. This methodology helps the machine to recognize the emotion of the speaker. Several techniques involve the use of the applications of Deep Learning. Automatic Speech Recognition (ASR) with Artificial Intelligence (AI) can be called the principal entity in realizing an emotion detector. Learning about the sentiments of a user by a BPO center can be beneficial in providing the user with the required service based on feedback.

[5] This paper offers a thorough analysis of the various SER system models. Theoretical definition, effective categorization of emotions and different methods of expressing emotions are presented. MFCCs and MS features of the speech signal are extracted and Feature Selection (FS) is employed to select the most pertinent ones. Machine learning algorithms like recurrent neural network (RNN), multivariate linear regression (MLR), support vector machine (SVM) is deployed to check for their efficiency in the detection of seven basic emotions. The highest accuracy of 94% was achieved with the RNN model using the Berlin database of speech signals.

III. PROPOSED METHODOLOGY

The three main components of any SER system are signal pre-processing, extraction of relevant features and correct classification of the emotion. To extract pertinent parts of the signal, pre-processing techniques such as removal of background noise and segmentation of the signal are done. To extract relevant information from the pre-processed signal, feature extraction is performed. Finally, the classifiers do the job of matching the features with the corresponding emotion. The dissimilitude between unprompted and voluntary speech is also considered because they play an important role in the detection of gender and emotion [21].

Figure 1 presents a basic flow chart for Speech Emotion Recognition (SER). The first stage involves pre-processing the signal to eliminate all non-relevant features. The second stage comprises feature extraction and selection, where relevant features are extracted from the output of the pre-processing stage. This typically involves studying speech signals in the temporal and frequency domains. In the third stage, a Multi-Layer Perceptron (MLP) classifier is used to categorize these features. Finally, based on the classification of features, various emotions are identified.

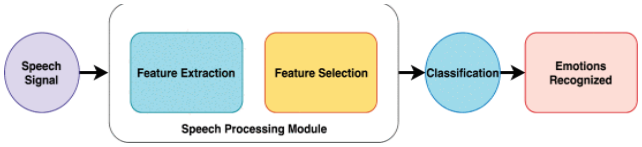


Fig. 1. Basic SER system

The whole machine learning pipeline is as follows:

1. Preparing the Dataset: The dataset is downloaded and data cleaning is done to prepare it for the next step.

2. Loading the Dataset: The dataset from the previous set is loaded into our feature extraction function in Python where features like MFCCs, Tonnetz, Pitch etc. are extracted

3. Training the Model: After the model has been loaded with the dataset, we train it on the MLP classifier.

4. Testing the Model: We will then use a separate testing dataset to see how well the model does on unlabelled data.

Our study used a limited portion of the complete dataset, where we aimed to minimize bias by ensuring that the number of audio files in the male and female categories was almost the same, as illustrated in Figure 2. To facilitate further analysis, we divided the .csv file into two columns, one for filenames and the other for gender. Specifically, our filtered audio collection contained 6995 male audio files and 5662 female audio files

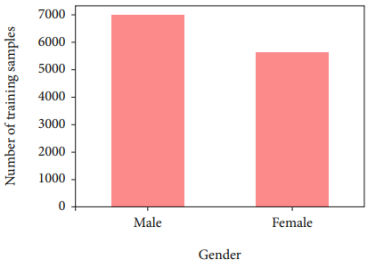


Fig 2: Statistics of voice samples in different genders.

IV. THEORY

A. Libraries used

- TensorFlow is one of the leading libraries and platforms for machine learning where full end to end services are provided right from loading data into the model to train and then test it to deploying it. Keras is one of the libraries used by TensorFlow to enable users to create their own models quite easily. There is no provision for JavaScript interoperability on the APIs present on the which are quite stable.
- Scikit-learn or sklearn is a pre-existing library in Python and uses NumPy for high-level mathematical computations like classification, regression etc. The functions in Scikit-learn are based off of previously existing libraries in Python called SciPy, NumPy and Matplotlib. Sklearn is particularly focused on the modeling aspect of ML. It is an open source library. To enhance performance, sklearn's algorithms are written in highly efficient Cython language.

B. Multi-Layer Perceptron Model (MLP Classifier)

A MLP classifier is a machine learning model which is a type of binary feedforward neural network. A MLP classifier's input, single hidden layer, and output layers are its foundational three layers. In MLP, the neurons are called perceptrons and are stacked on top of each other and hence the name, multi-layer. All the nodes are interconnected, between layers and within layers.

C. RAVDESS Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7,356 files. It has 24 actors who vocalize two textually similar statements in neutral North American accent. The emotions recorded are calm, happy, sad, angry, terror, surprise and disgust and each of the emotions are recorded at two levels of intensity: normal and strong.

D. Mathematical Equations

1. Mathematical definition of STFT: The usual mathematical definition of STFT is [16-17]:

$$X_m(w) = \sum_{n=-\infty}^{\infty} x(n)w(n - mR)e^{-j\omega n}$$

$$= DTFT_{\omega}(x \,.\, SHIFT_{mR}(w))$$

Where,

x(n)= *input signal at time n*

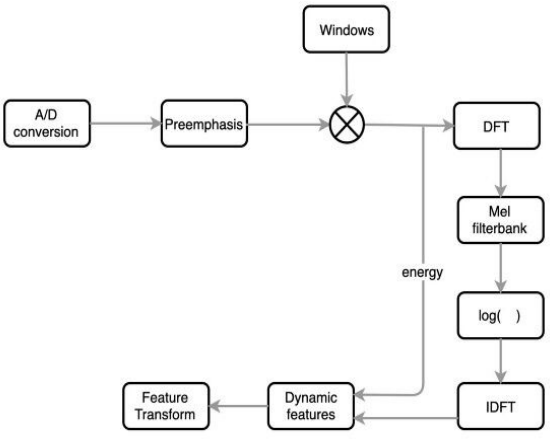
w(n)= length M window function (e.g., Hamming)

$X_{m(w)}$= DTFT of windowed data centered about time mR.

R= Hop size in sample between successive DTFTs.

2. Mathematical equation for conversion of Hertz scale to Mel scale:

$$m=1127 \times \log [1+(f/100)]$$



Where,
m = Frequency in Mel scale
f = Frequency in Hertz scale

V. IMPLEMENTATION

A. MFCC:

Fig 3: Algorithm for extraction of MFCC

The road map of the MFCC technique is given below:

According to Fig. 3, the MFCC precisely portrays the vocal plot, a sifted state of the human voice, as well as how it shows up in the envelope of a brief time frame power range. Following a series of sequential steps is necessary to compute MFCCs [9]:

(1) Shortening the Signal's Frames. Sound stream is partitioned into frames of 20-40 milliseconds long (25 milliseconds is considered normal) to compensate for sample shifts over lengthy periods of time that occur often.

(2) A power spectrum periodogram. This determines the frequencies in the frame by calculating the periodogram assignment of power range for each snippet of the signal.

(3) Using the Power Spectra with the Mel Filter bank. Due to extra information in periodogram spectral estimation, a filter is needed to estimate the energies in variety of frequency ranges that show up in a collection of accumulated periodogram canisters. Because there is less concern for variances, the Mel Filter bank assesses the energy close to zero Hz and afterward for higher frequencies.

(4) All Filter bank Energies Logarithm. Huge changes of energies are scaled involving a logarithmic scale as there are no recognizable sounds with enormous energies. For cepstral mean deduction, the logarithmic scale is a channel standardization procedure that is likewise utilized.

(5) Log Filter bank energies DCT. It is utilized to decorrelate energies because the Filterbank energies exhibit correlation that causes overlapping. This produces features in the form of diagonal covariance matrices.

B. Proposed methodology

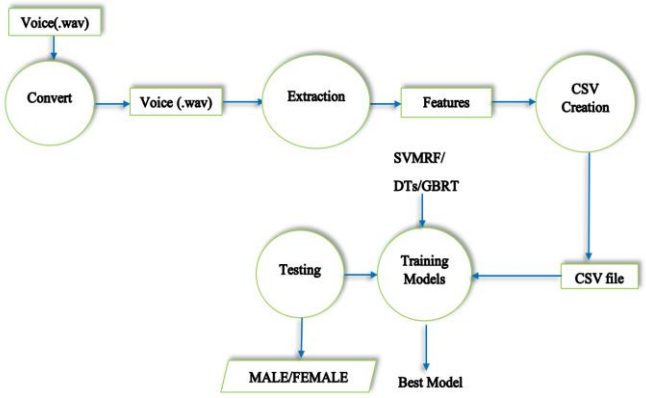


Fig. 4: Algorithm to recognize gender of a speaker

The explanation for fig 4 is as follows:

Step-1: The dataset is created by taking several different voices as input and converted to a .wav file

Step-2: Features from the voice sample are extracted.

Step-3: The several .wav files are put into a .csv file for easy access.

Step-4: The dataset is fed into a training model which is a feedforward neural network.

Step-5 Two subsets are defined from the dataset where the one used for testing will determine accuracy of our model.

Step-6: The model then predicts any individual sample as Male or Female depending upon parameters.

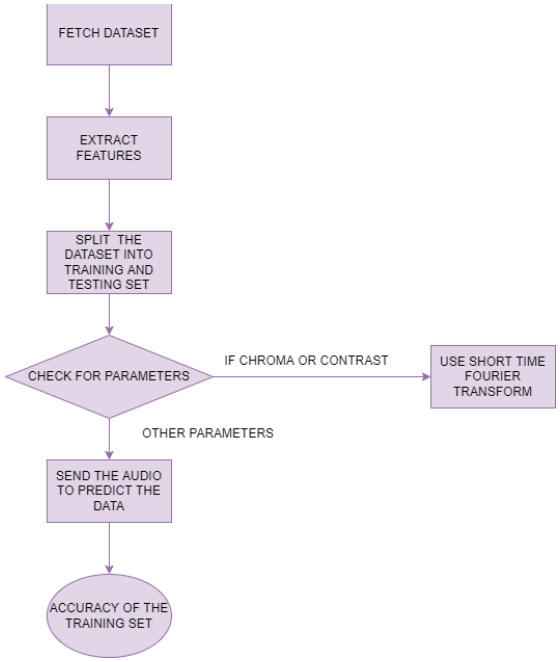


Fig. 5: Algorithm to determine accuracy of a model

The explanation for fig 5 is as follows:

*Step-1:*Dataset used for training and testing is fetched.

Step-2: Processing of dataset takes place to extract relevant features.

Step-3: Dataset is split into Training and Testing sets.

Step-4: If the parameters are chroma or contrast STFT is applied to it otherwise the audio is sent to predict the data.

Step-5: Accuracy of the training set is calculated.

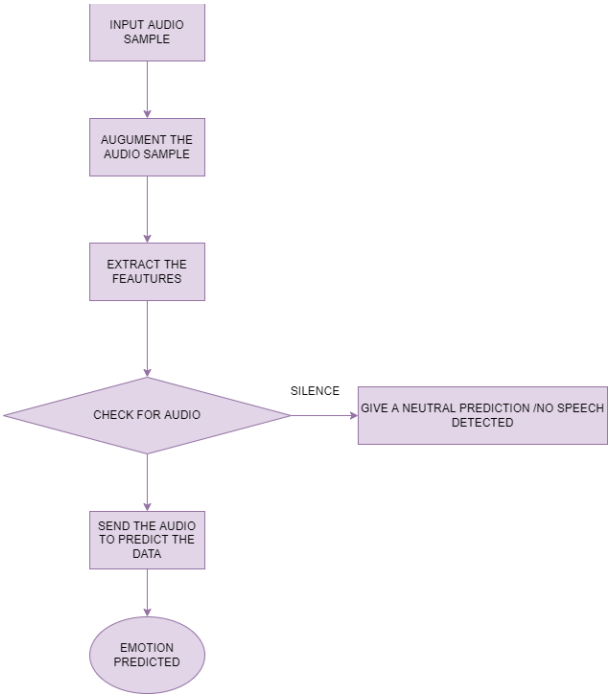


Fig. 6: Algorithm to predict emotion

The explanation for fig 6 is as follows:

Step-1-: Audio sample is given as input.

Step-2:The audio sample is augmented.

Step-3: The features are extracted.

Step-4: The audio is checked. If there is silence (no audio), a neutral prediction is given. Else the audio is sent to predict the data.

Step-5: Emotion is predicted and result is displayed.

C. Features used for classification of emotions

Emotion	Pitch	Intensity	Speaking
Anger	Abrupt on stress	Much higher	Marginally faster
Happiness	Much wider,	Higher	Faster/ slower
Sadness	Slightly narrow	Downward inflections	Lower
Neutral	Straight line	Not a lot of variations	Subdued

VI. RESULTS

A. Voice Gender Recognition

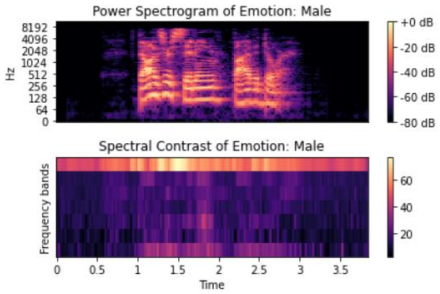
Fig 7 and Fig. 8 depicts the output we have obtained after we training our model to recognize the gender of the speaker along with the probability of it being a male or female voice.

```
In [15]: import argparse
parser = argparse.ArgumentParser(description="""Gender recognition script, this will load the model you trained,
                                                and perform inference on a sample you provide (either using your voice or a file)""")
parser.add_argument("-f", "--file", help="The path to the file, preferred to be in WAV format")
args = parser.parse_args()
file = args.file
# construct the model
model22 = create_model()
# load the saved/trained weights
model22.load_weights("results/model4k.h5")
features = extract_feature("emotion2.wav", mel=True).reshape(1, -1)
# predict the gender!
male_prob = model.predict(features)[0][0]
female_prob = 1 - male_prob
gender = "male" if male_prob > female_prob else "female"
# show the result!
print("Result:", gender)
print(f"Probabilities::: Male: {male_prob*100:.2f}%    Female: {female_prob*100:.2f}%")

<ipython-input-1-9e04a9e0d0cc>:37: FutureWarning: Pass y=[ 0.0000000e+00  0.0000000e+00  0.0000000e+00 ... -1.8908099e-07
 5.5059752e-07  0.0000000e+00] as keyword args. From version 0.10 passing these as positional arguments will result in an erro
r
  mel = np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)

Result: male
Probabilities::: Male: 93.35%    Female: 6.65%
```

Fig 7: Male voice being predicted

Fig 8: Female voice being predicted

B.



Graphs of Extracted Parameters

Fig. 9 depicts the difference in amplitude of a male voice and female voice. Females tend to have a higher pitch which makes the amplitude of their voice higher than that of a male.

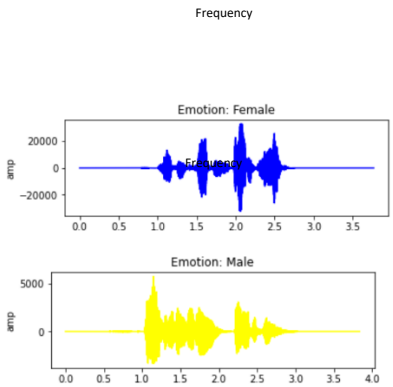


Fig 9: Graph of Voice Amplitude Vs. Frequency

Fig 10 depicts the power spectrogram and spectral contrast of a male voice. A spectrogram is a graphical method of representing the energy level or strength of a signal at different frequencies over time in a particular signal. It aids in visualizing the signal and one can see the varying levels of energy in a signal with respect to time [22-23].

Time

Fig. 10 Power Spectrogram and Spectral Contrast of Male Voice

In speech processing, the Mel-Frequency Cepstrum (MFC) is a way to visualize the power band of a signal which is done using a cosine transform on a non-linear mel frequency scale. The MFC is made of different components which are all collectively called the MFCCs. Fig 11 depicts the MFCC coefficient representation of a male and female voice.

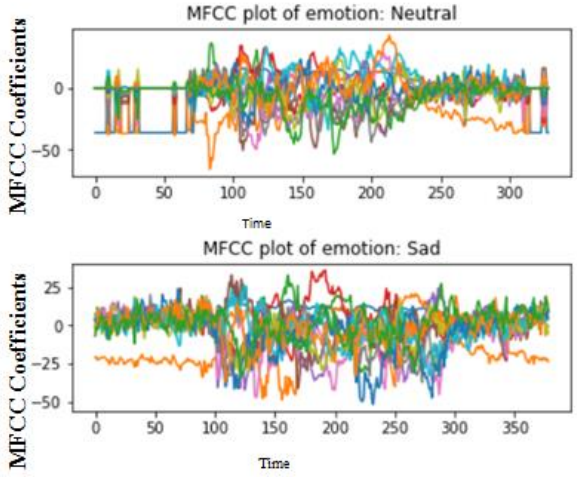


Fig. 11. MFCC Plot of Male and Female Voice

In sound processing, the Tonnetz is a theoretical representation of the tone of a speech signal. To demonstrate conventional harmonic relationships in speech samples, various Tonnetz graphic representations can be employed like the one of a female voice as depicted in Fig 12.

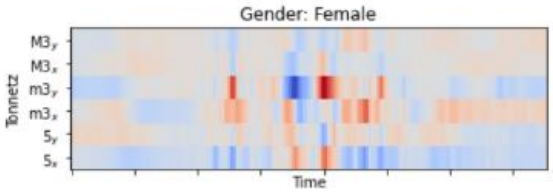


Fig. 12. Tonnetz Plot of Female Voice

C. Speech Emotion Recognition

In this project, we are attempting to recognize the emotion of a voice sample by means of training and testing our machine learning model. After the model has been trained with the training dataset, we give a voice sample as input to the model which will predict the emotion. Fig 13 and 14 show the prediction of neutral, angry, and happy emotions.

```
if __name__ == "__main__":
    # load the saved model (after training)
    model = pickle.load(open("result/mlp_classifier.model", "rb"))
    filename ="emotion3.wav"
    # extract features and reshape it
    features = extract_feature(filename, mfcc=True, chroma=True, mel=True).reshape(1, -1)
    # predict
    result = model.predict(features)[0]
    # show the result !
    print("result:", result)
```

```
result: neutral
```

Fig. 13 Neutral emotion prediction

```
if __name__ == "__main__":
    # load the saved model (after training)
    model = pickle.load(open("result/mlp_classifier.model", "rb"))
    filename ="emotion1.wav"
    # extract features and reshape it
    features = extract_feature(filename, mfcc=True, chroma=True, mel=True).reshape(1, -1)
    # predict
    result = model.predict(features)[0]
    # show the result !
    print("result:", result)
```

```
result: angry
```

Fig. 14. Angry emotion prediction

D. Graphs of Extracted Parameters

Fig 15 depicts the variations in voice samples' emotions with respect to amplitude and frequency. For example, an angry voice has a high amplitude owing to its loudness and a sad one has a comparatively low amplitude.

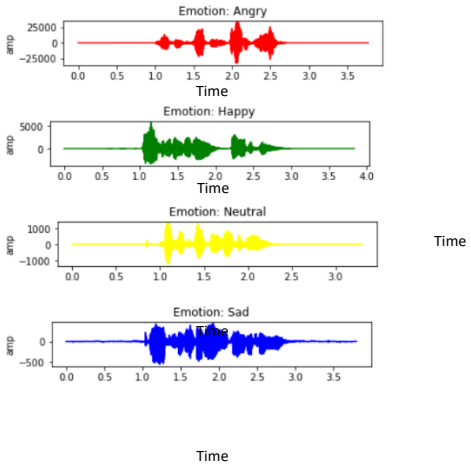


Fig 15 Graph of Voice Amplitude Vs. Frequency

Fig 16 and 17 depicts the power spectrogram and spectral contrast of a voice when happy or sad emotions are being conveyed. depicts the power spectrogram and spectral contrast of a male voice. A spectrogram is a graphical method of representing the energy level or strength of a signal at different frequencies over time in a particular signal. It aids in visualizing the signal and one can see the varying levels of energy in a signal with respect to time [20].

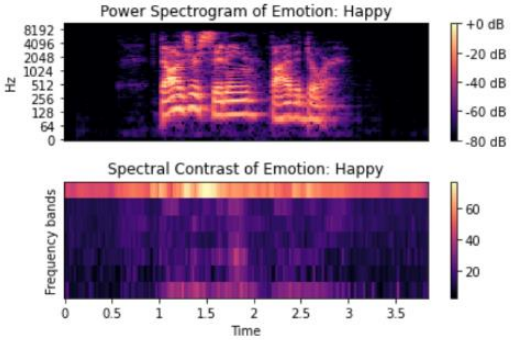


Fig 16: Power Spectrogram and Spectral Contrast Emotion: Happy

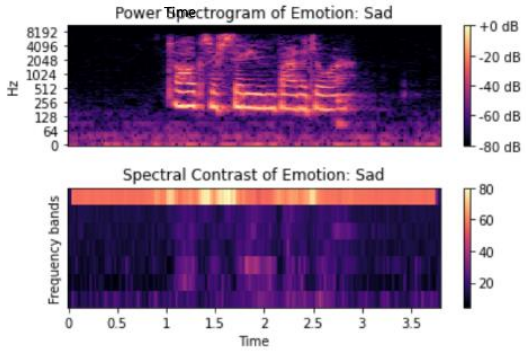


Fig 17: Power Spectrogram and Spectral Contrast Emotion: Sad

In speech processing, the Mel-Frequency Cepstrum (MFC) is a way to visualize the power band of a signal which is done using a cosine transform on a non-linear mel frequency scale. The MFC is made of different components which are all collectively called the MFCCs. Fig 18 depicts the MFCC coefficient representation of a voice when neutral and sad emotions are being conveyed.

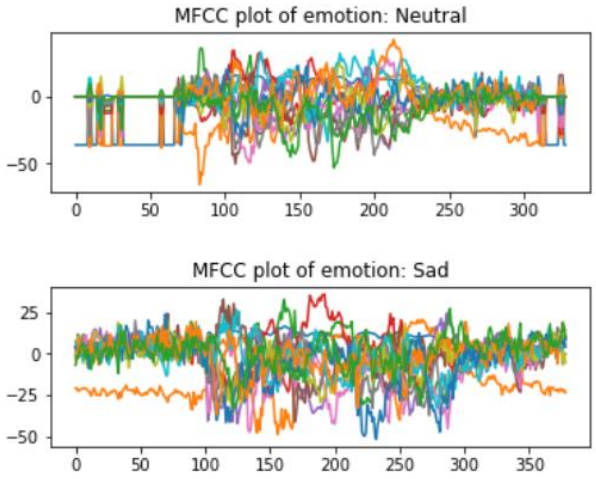


Fig 18: MFCC Plot of Neutral and Sad Emotion

In sound processing, the Tonnetz is a theoretical representation of the tone of a speech signal. To demonstrate conventional harmonic relationships in speech samples, various Tonnetz graphic representations can be employed like the one of a female voice as depicted in Fig 19.

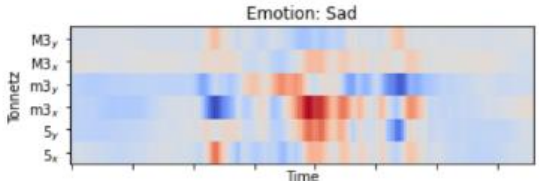


Fig 19: Tonnetz Plot of Sad Emotion

VII. CONCLUSION

Speech recognition technology has become increasingly popular in recent years and is being applied to a variety of different areas, from voice-controlled virtual assistants like Alexa and Siri and automatic captioning for videos to transcriptions for online courses. Such applications have made it much easier for people to interact with computers and access information quickly. AI can be used to make our lives easier and more efficient research has been conducted to create new applications of speech emotion recognition that can further improve the use of AI. Thanks to this, thorough analysis has yielded profitable and informative results and YouTube's captions are getting better every year. However, voice emotion recognition applications are more nuanced and add a new dimension to the use of AI and provide an easier way to improve our lives through it. Recognition of gender from the speech signal has been useful in the area of information-based multimedia cataloging.

A recent application of SER comes as a result of the rapid growth of online learning, where teachers can observe a student's actions and help them to become better at learning. Another promising application is evaluating candidates applying for leadership positions by analyzing responses during audio or video interviews. Previously inestimable quantities like their emotional reactions can now be remedied, thanks to SER.

SERs can also be used to evaluate the performance of current personnel, particularly in the call center sector where poor customer communication can seriously damage a company's reputation. These systems can also handle customer complaints in a more efficient and automated way. Similarly, you can monitor and take care of your employees' emotional health.

VIII. ACKNOWLEDGEMENT

We are truly grateful to all my colloquies and friends for their immense support in making of this research successful.

IX. REFERENCES

[1].R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019, doi:10.1109/ACCESS.2019.2936124.

[2]. L.-M. Zhang, Y. Li, Y.-T. Zhang, G. W. Ng, Y.-B. Leau, and H. Yan, "A Deep Learning Method Using Gender-Specific Features for Emotion Recognition," *Sensors*, vol. 23, no. 3, p. 1355, Jan. 2023, doi: 10.3390/s23031355

[3].M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data", *Inf. Fusion*, vol. 49, pp. 69-78, Sep. 2019.

[4]. Kerkeni, Leila & Serrestou, Youssef & Raoof, Kosai & Cléder, Catherine & Mahjoub, Mohamed & Mbarki, Mohamed. (2019). "Automatic Speech Emotion Recognition Using Machine Learning". 10.5772/intechopen.84856.

[5]. J. Oliveira and I. Praça, "On the Usage of Pre-Trained Speech Recognition Deep Layers to Detect Emotions," in IEEE Access, vol. 9, pp. 9699-9705, 2021, doi: 10.1109/ACCESS.2021.3051083.

[6].J. Deng, S. Frühholz, Z. Zhang and B. Schuller, "Recognizing emotions from whispered speech based on

acoustic feature transfer learning", *IEEE Access*, vol. 5, pp. 5235-5246, 2017.
[7]. A.P. Vogel, P. Maruff, P. J. Snyder, J.C. Mundt, "Standardization of pitch-range settings in voice acoustic analysis, Behavior Research Methods", v.41, n.2, p.318-324, 2019
[8]. I. Livieris, E. Pintelas, and P. Pintelas, "Gender recognition by voice using an improved self-labeled algorithm," Mach. Learn. Knowl. Extr., vol. 1, no. 1, pp. 492–503, 2019.
[9]. T. -W. Sun, "End-to-End Speech Emotion Recognition with Gender Information," in IEEE Access, vol. 8, pp. 152423-152438, 2020, doi: 10.1109/ACCESS.2020.3017462.
[10]. Bishop J and Keating P 2012 "Perception of pitch location within a speaker's range: Fundamental Frequency, voice quality and speaker sex", The Journal of the Acoustical Society of America 32-2 1100-1112 [6] Smith D R and Patterson R D 2005 The interaction of glottal-pulse rate and vocal-tract length in judgments of speaker size, sex, and age in The Journal of the Acoustical Society of America, 118-5 3177-3186
[11]. 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.
[12]. Büyükyılmaz, Mücahit & Çıbıkdiken, Ali. (2016). "Voice Gender Recognition Using Deep Learning". 10.2991/msota-16.2016.90.
[13]. A. B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan, "Speech recognition using deep neural networks: A systematic review", *IEEE Access*, vol. 7, pp. 19143-19165, 2019.
[14]. Muhammad G, AlSulaiman M, Mahmood A and Ali Z 2011 "Automatic voice disorder classification using vowel formants", 2011 Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '11) 1–6
[15]. Ftoon Abu Shaqra, Rehab Duwairi, Mahmoud Al-Ayyoub, "Recognizing Emotion from Speech Based on Age and Gender Using Hierarchical Models", Procedia Computer Science, Volume 151, 2019, Pages 37-44, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.04.009.
[16].Gaikwad S, Gawali B, and Mehrotra S C 2012 "Gender identification using SVM with combination of MFCC", Advances in Computational Research 4 69–73
[17.].Zeng Y M, Wu Z Y, Falk T and Chan W Y 2006 "Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech", Proceedings of the International Conference on Machine Learning and Cybernetics 3376–3379
[18]Z. Zeng, M. Pantic, G. I. Roisman and T. S. Huang, "A survey of affect recognition methods: Audio visual and spontaneous expressions", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39-58, Jan. 2009.
[19]. M. Golfer and V. Mikes (2018), "The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels", Journal of Voice, vol. 19, no. 4, pp. 544-554
[20]. W. Fei, X. Ye, Z. Sun, Y. Huang, X. Zhang and S. Shang, "Research on speech emotion recognition based on deep auto-encoder", *Proc. IEEE Int. Conf. Cyber Technol. Automat. Control Intell. Syst. (CYBER)*, pp. 308-312, Jun. 2016.
[21].J. Han, Z. Zhang, F. Ringeval and B. Schuller, "Prediction-based learning for continuous emotion recognition in speech", *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 5005-5009, Mar. 2017.
[22]. A. Satt, S. Rozenberg and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms", *Proc. INTERSPEECH*, pp. 1089-1093, 2017.
[23].Tursunov A, Mustaqeem, Choeh JY, Kwon S. "Age and Gender Recognition Using a Convolutional Neural Network with a Specially Designed Multi-Attention Module through Speech Spectrograms." Sensors (Basel). 2021 Sep 1;21(17):5892. doi: 10.3390/s21175892. PMID: 34502785; PMCID: PMC8434188.