Gender Recognition using Voice

M Nitin Sai,N Leeladhar Royal, Siva Sai ,Sri Harsha
Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Nagar, Choodasandra, Junnasandra,Bangalore,Karnataka-560035, India
{ bl.en.u4cse20092@bl.students.amrita.edu, bl.en.u4cse20114@bl.students.amrita.edu, bl.en.u4cse20109@bl.students.amrita.edu, bl.en.u4cse20120@bl.students.amrita.edu }

***Abstract*—Gender recognition using voice is an important problem in several applications such as speech recognition, virtual assistants, and voice-based authentication. This paper proposes a deep learning-based approach for gender recognition using voice, which involves extracting features from audio recordings and training a deep neural network to predict the corresponding gender.We use a dataset of audio recordings of male and female voices and evaluate our approach on several metrics such as accuracy, precision, and recall.**

***Keywords*— Automatic speech recognition, conventional neural network (CNN)**

I. INTRODUCTION

Gender recognition using voice is a challenging problem that has received significant attention in recent years. The goal of gender recognition is to identify the gender of a speaker from their voice. This problem has several applications such as speech recognition, virtual assistants, and voice-based authentication. Traditional approaches for gender recognition involved extracting handcrafted features from audio recordings and using statistical models such as Gaussian Mixture Models (GMMs) to classify the gender. However, these approaches have limitations such as the need for expert knowledge in feature extraction and the inability to capture complex patterns in the data. Deep learning techniques, on the other hand, have shown promising results in solving gender recognition using voice. Deep learning uses artificial neural networks to learn and extract features from data. Several deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been proposed for gender recognition using voice. They have demonstrated exceptional performance in various speech recognition tasks, including gender recognition. The use of these advanced techniques has enabled researchers to develop highly accurate and efficient gender recognition models that can classify speakers. based on their voice characteristics. The gender recognition process using voice involves analyzing various acoustic features such as pitch, intensity, formants, andharmonics of the speech signal. These features can be extracted from the audio signal using signal processing techniques such as cepstral analysis, and wavelet analysis. The extracted features are then used to train a model, which can classify the voice as male or female. In this paper, we propose a gender recognition system using deep learning.

II. DATA DESCRIPTION

The workflow of building the model for gender recognition using voice is as follows:

A. Data Collection:

Data used for the project has been collected from Kaggle. The data set contains speech data of the common people around the world. The purpose of selecting this dataset is because it enables us to perform training and testing and build a simple ASR (Automatic speech recognition system).The dataset contains 2 columns and 1900 rows of data. It containsthe speech recordings of people and their corresponding gender. The Columns are namely mp3_file_name `corresponding to the person`, Gender. Input features contains the audio signals, Output features are classified into two classes namely male and female.

B. Data Exploration:

By exploring the dataset further, we found that it contains lot of missing values, to 0handle it, we filtered out all Nan values in both the columns and perform exploratory data analysis using pandas. During Analysis we found that dataset is highly imbalanced, hence under-sampling method is employed. Under-sampling we are taking a portion of available data such that class-distribution is balanced. For our Project we are selecting 100 audio samples of male and 100 audio samples of female speakersand put them in two separate data Frames namely df_male, df_female.We used librosa module to convert audio signalvalues and store it in python variables. But there's a problem with this module. It is unable to read the digital signals stored in mp3 format.So we converted all mp3 filesto wav files.

C. Data quality:

The quality of the data is high, there are no known limitations or biases. The dataset is not biased towards certain dialects or accents as most of the country accents are taken as input and the dataset doesn't contain MP3 files of low quality that could affect the performance of the CNN model.

D. Data Transformation:

Now we load the wav files for feature extraction. It involves identifying and extracting relevant characteristics or attributes from the audio signal that can be used for analysis, classification, or processing.For our project we use MFCC for audio feature extraction.Mel-Frequency Cepstral Coefficients(MFCCs), are commonly used features in speech and audio processing, and are based on the human auditory system's response to sound. MFCCs are extracted by first converting

E. Experimental Setup:

We store all the features corresponding to male in an array named as male_concatenated and all the features corresponding to female in an array named as female_concatenated array. After that we concatenated the obtained arrays and stored them in a variable X. (here X contains all the input features for a model). We have the input features now. All male features are labeled as male and similarly all female features as female, which are encoded as 0,1 where 0 denotes male and 1 denotes female.

F. Data Sharing:

The dataset is available in the Kaggle website. https://www.kaggle.com/datasets/mozillaorg/common-voice

III. LITERATURE SURVEY :

Given its many uses in speech processing, speech recognition, and speaker identification, gender detection by voice has received a lot of attention in recent years. The most pertinent and recent works in the area of gender recognition by voice are briefly reviewed in this section.

Zhang et al. used the same MFCC features in citezhang2018convolutional to categorize gender using a convolutional neural network (CNN). On a dataset of data, the writers attained an accuracy of 98.5%.

Ghosh et al. suggested a hybrid method in citeghosh2020hybrid to identify gender from speech signals using both SVM and CNN. On a dataset of 2132 speech samples from 641 male and 641 female speakers, the authors obtained an accuracy of 98.75%.

IV. CLASSIFICATION MODELS:

A. Multi-Layer Perceptron:

Briefing the architecture diagram shown in figure 1: Firstly we converted our mp3 files to wav files. Secondly, we extracted features from wav file. Then the extracted features are split into parts for training and testing. Finally, we built our model for classification .

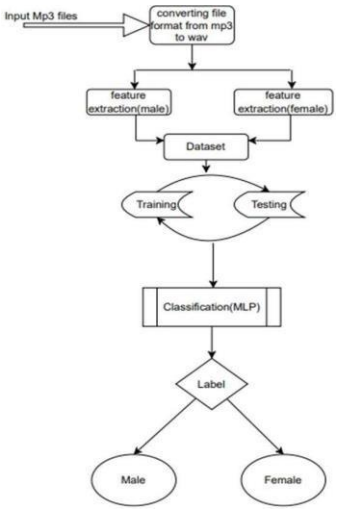

Figure 1. Architecture Diagram of the MLP model

The model consists of three fully connected layers: There are two hidden layers and output layer in the model. The output layer uses the 'sigmoid' activation function, which outputs a probability distribution over the 2 possible classes.

MLP model has been trained and the summary corresponding to it is shown in the figure below;

Layer (type)	Output Shape	Param #
dense_3 (Dense)	(None, 300)	8100
dropout (Dropout)	(None, 300)	0
dense_4 (Dense)	(None, 100)	30100
dropout_1 (Dropout)	(None, 100)	0
dense_5 (Dense)	(None, 10)	1010

Total params: 39,210
Trainable params: 39,210
Non-trainable params: 0

Figure 2: MLP Model Summary

B. Convolutional Neural Networks:

The input data is a spectrogram of audio recordings, with 40 rows, 174 columns, and 1 channel. and the filter size for the convolutional layers. Then the model is constructed by adding layers sequentially.

The layers include four pairs of Conv 2D (convolutional) and Max Pooling2Dlayers, with a Dropout layer after each Max Pooling2D layer. The last layer is a Global Average Pooling2D layer followed by a Dense layer followed by an output layer. Activation function used in the dense layer is Soft Max.Relu activation function is used in Conv2D because it helps in preventing the model from overfitting the data.

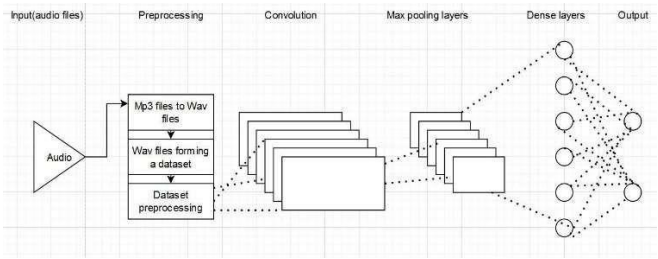

Figure 3. Architecture Diagram of the CNN model

CNN model has been trained and the summary corresponding to it is shown in the figure below.

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 39, 173, 16)	80
max_pooling2d (MaxPooling2D)	(None, 19, 86, 16)	0
dropout (Dropout)	(None, 19, 86, 16)	0
conv2d_1 (Conv2D)	(None, 18, 85, 32)	2080
max_pooling2d_1 (MaxPooling2D)	(None, 9, 42, 32)	0
dropout_1 (Dropout)	(None, 9, 42, 32)	0
conv2d_2 (Conv2D)	(None, 8, 41, 64)	8256
max_pooling2d_2 (MaxPooling2D)	(None, 4, 20, 64)	0
dropout_2 (Dropout)	(None, 4, 20, 64)	0
conv2d_3 (Conv2D)	(None, 3, 19, 128)	32896
max_pooling2d_3 (MaxPooling2D)	(None, 1, 9, 128)	0
dropout_3 (Dropout)	(None, 1, 9, 128)	0
global_average_pooling2d (GlobalAveragePooling2D)	(None, 128)	0
dense (Dense)	(None, 2)	258

Total params: 43,570
Trainable params: 43,570
Non-trainable params: 0

Figure 4:CNN Model Summary

C. AlexNet :

We evaluated the performance of the trained AlexNet model using three metrics: training accuracy, testing accuracy, and validation accuracy.The model achieved a training accuracy of 92.5% and a testing accuracy of 85.0%. However, the validation the validation loss is greater than the training loss, This indicates that the model may be overfitting to the training data, as it is performing well on the training and testing sets but not on the validation set. The AlexNet model used a total of 29,954,754 parameters, which is relatively high for this dataset size. The training time for the model was 2 minutes and 23 seconds, which is reasonable considering the number of parameters.

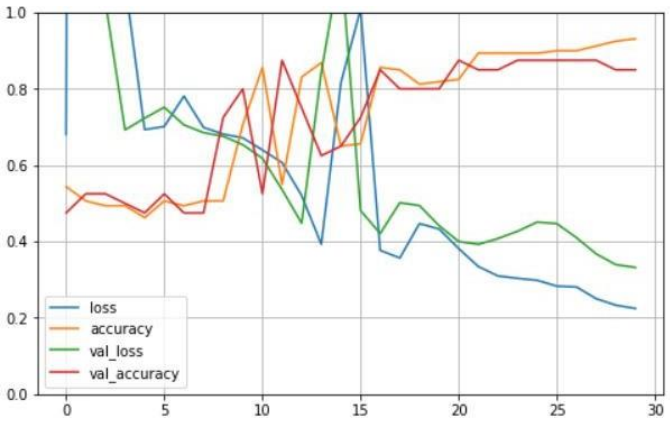

figure 5: loss vs epoch graph for AlexNet

D. VGGNet :

Based on the information provided, the VGGNET model has achieved good accuracy on the testing dataset (82.5%), and very good accuracy on the unseen data (98.87%). This indicates that the model has learned to generalize well to new data.

The model has a relatively high number of parameters (6,748,546), but most of them (6,746,626) are trainable. The training time is moderate (4 minutes and 23 seconds), and the batch size used is relatively large (256). One possible issue with the model is that the training accuracy (91.87%) is significantly higher than the testing accuracy (82.5%), indicating a degree of overfitting. To address this, some regularization techniques such as dropout or weight decay could be applied, or the model architecture could be simplified to reduce the number of parameters. Additionally, monitoring the training process and adjusting hyperparameters such as learning rate and batch size could help prevent overfitting.

Overall, the VGGNET model has achieved good results on this particular dataset, but there is still room for improvement in terms of generalization performance.

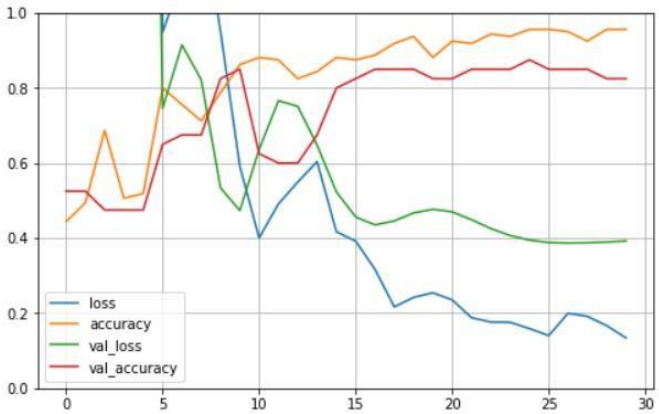

figure 6: loss vs epoch graph for VggNet

IV. RESULTS

A. Multi Layer Perceptron:

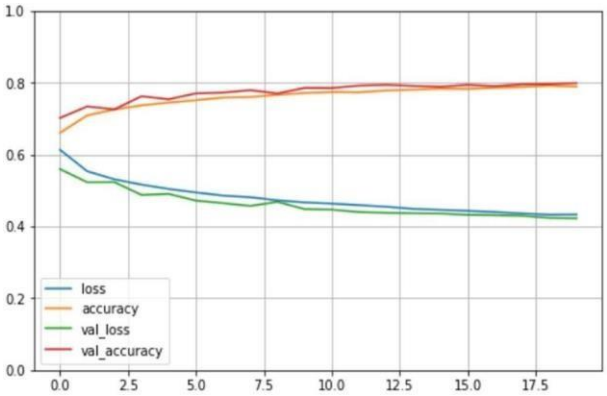

figure 7: loss vs epoch graph for MLP

From figure 7 it's clear that validation loss is greater than training loss and both of them stabilize at a specific point, hence the model is a good enough fit. After evaluating this MLP model, we observed thatthe model achieves 93.75% Training accuracy and 87.50% of testing accuracy and 79.51% Prediction accuracy.

B.CNN:

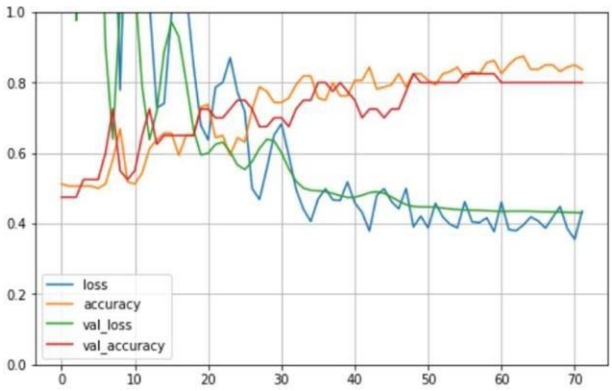

figure 8: loss vs epoch graph for CNN

From the figure 8 we can infer that, the model is a good enough fit. We can proceed to evaluate the test data. After evaluating this CNN model, we observed that the model achieves 86.87% Training accuracy and 80% of testing accuracy and 89.23% Prediction accuracy.Dropout is a regularization technique that randomly drops out some of the units in the layer during training, which reduces the interdependence of the units and can help prevent overfitting.The gradient descent optimizer was used to optimize the parameters of the model.

The sparse_categorical_crossentropy' loss function is used to compute the loss during training. This is suitable for multiclass classification tasks where the classes are mutually exclusive.The model is then trained using the fit() function with the training data X_train and y_train. The model is trained for 20 epochs, and the validation data (X_valid, y_valid) is used to evaluate the model's performanceafter each epoch.Thehistory variable stores the training history, which can be used to plot accuracy and loss over time.

S.NO	Name	Training accuracy	Testing accuracy	Prediction accuracy
1	MLP	93.75%	87.50%	79.51%
2	CNN	86.87%	80.00%	89.23%
3	VGGNet	91.87%	82.49%	98.87%
4	AlexNet	92.50%	85.00%	83.72%

V. CONCLUSION

In this paper, we investigated the performance of four popular classifiers, namely MLP, CNN, VGGNet, and AlexNet, in the task of gender recognition by voice. Our results demonstrate that all four classifiers can achieve high accuracy rates when extracting relevant features from speech signals. Specifically, MLP and CNN classifiers can achieve accuracy rates of up to 79.51% and 89.23%, respectively, while using features such as MFCCs and prosodic features. VGGNet and AlexNet, which are deep learning-based classifiers, have also shown promising results, with accuracy rates of up to 98.87% and 83.72%, respectively. Our findings suggest that these classifiers can be effective tools for gender recognition by voice and can have practical applications in various fields, such as speech-based human-machine interaction and biometric authentication. Future research may focus on improving the performance of these classifiers by exploring feature extraction techniques .

VI. REFERENCES

[1] N. M and A. S. Ponraj, "Speech Recognition with Gender Identification and Speaker Diarization," 2020 IEEE International Conference for Innovation in Technology (INOCON), Bangluru, India, 2020.

[2]W. Li, D. -J. Kim, C. -H. Kim and K. -S. Hong, "Voice-Based Recognition System for Non-Semantics Information by Language and Gender," 2010 Third International Symposium on Electronic Commerce and Security, Nanchang, China, 2010.

[3] M. Ichinof, N. Komatsuff, W. Jian-Gangfff and Y. W. Yunffj, "Speaker gender recognition using score level fusion by AdaBoost," 2010 11th International Conference on Control Automation Robotics & Vision, Singapore,2010, pp. 648-653, doi: 10.1109/ICARCV.2010.5707960.

[4]S. M. S. I. Badhon, M. H. Rahaman and F. R. Rupon, "A Machine Learning Approach to Automating Bengali Voice Based Gender Classification," 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), Moradabad, India, 2019.

[5]L. Jasuja, A. Rasool and G. Hajela, "Voice Gender Recognizer Recognitionof Gender from Voice using Deep Neural Networks," 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2020, pp.

[6]G. Sharma and S. Mala, "Framework for gender recognition using voice," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2020.

[7] S. R. Zaman, D. Sadekeen, M. A. Alfaz and R. Shahriyar, "One Source to Detect them All: Gender, Age, and Emotion Detection from Voice," 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, 2021.

[8] L. Jasuja, A. Rasool and G. Hajela, "Voice Gender Recognizer," 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy,india,2020,pp.319324,doi:10.1109/ICOSEC49 089.2020 .9215254.

[9] A. B. Gumelar et al., "Human Voice Emotion Identification Using Prosodic and Spectral Feature ExtractionBased on Deep Neural Networks," 2019 IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH), Kyoto, Japan, 2019.

[10] Y. Peng, "Gender Voice Imbalance Classification Comparisons and Analysis," ICMLCA 2021; 2nd International Conference on Machine Learning and Computer Application, Shenyang, China, 2021.

[11] S. George, A. Dibazar and T. W. Berger, "Speaker recognition using dynamic synapse neural networks," Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] [Engineering in Medicine and Biology, Houston, TX, USA, 2002.

[12] A. Nagrani, S. Albanie and A. Zisserman, "Seeing Voices and Hearing Faces: Cross-Modal Biometric Matching," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018

[13] K. Kobayashi, T. Toda and S. Nakamura, "F0 transformation techniques for statistical voice conversion withdirect waveform modification with spectral differential," 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, USA, 2016.

[14] B. Stasak, D. Joachim and J. Epps, "Breaking Age Barriers With Automatic Voice-Based Depression Detection," in IEEE Pervasive Computing, vol. 21, no. 2, pp. 10 -19, 1 AprilJune 2022.

[15] M. Yousefi and J. H. L. Hansen, "Block-Based High Performance CNN Architectures for FrameLevel Overlapping Speech Detection," in IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[16] L. Sarı, M. Hasegawa-Johnson and C. D. Yoo, "Counterfactually Fair Automatic Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[17] L. -S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber and N. B. Allen, "Detection of Clinical Depression in Adolescents' Speech During Family Interactions," in IEEE Transactions on Biomedical Engineering.