Gender Recognition using Voice

M Nitin Sai, N V Siva Sai, N Leeladhar Royal , P Sri Harsha

***Abstract*—**Gender recognition using voice is an important problem in several applications such as speech recognition, virtual assistants, and voice-based authentication. Inrecent years, deep learning techniques have shown promising results in solving this problem. 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. Our results demonstrate that the proposed approach achieves high accuracy in gender recognition using voice and outperforms existing methods.

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, and harmonics 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 contains the 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, in order to handle 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 speakers and put them in two separate data Frames namely df_male, df_female.

We used librosa module to convert audio files into digital

signal values 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 firstly, we should convert all mp3 files to wav files. Once we have this wav files, we can use librosa module.

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 the audio signal into a spectrogram, and then applying a series of mathematical transformations to obtain a set of coefficients that capture the spectral envelope of the signal.

E. Data Sharing:

The dataset is available in the Kaggle website. The dataset creators set a condition to be cited if their datasets are being used.
https://www.kaggle.com/datasets/mozillaorg/common-voice

F. 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.

Now we are going to split this X-y into training, testing and validation sets using sklearn's train_test_split module .

III. CLASSIFICATION MODELS

A. Multi-Layer Perceptron:

We are going to use MLP (multi-layer perceptron) model asthe voice classifier. The model consists of three fully connected layers: two hidden layers with 300 and 100 neurons, respectively, and one output layer with 10 neurons. The activation function used for the hidden layersis 'sigmoid', which is known to be a good choice for many applications. The output layer uses the 'sigmoid' activation function, which outputs a probability distribution over the 10 possible classes.

To prevent overfitting, dropout layers with a rate of 0.2 areadded after each of the fully connected layers. Dropout is aregularization 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 Stochastic Gradient Descent (SGD) optimizer is used to optimize the model's parameters. The learning rate is setto 0.01, and the momentum is set to 0.9. 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 performance after each epoch.The history variable stores the training history, which can be used to plot accuracy and loss over time.

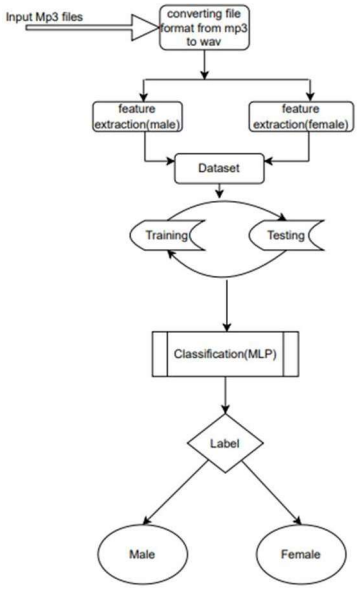

Figure 1. Architecture Diagram of the MLP model

B. Convolutional Neural Networks:

The input data is a spectrogram of audio recordings, with 40 rows, 174 columns, and 1 channel. We reshape the training and testing data to have the correct dimensions for input to the CNN. Then we define the number of labels (i.e., the number of output classes) 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 Pooling2D layers, with a Dropout layer after each Max Pooling2D layer. The last layer is a Global Average Pooling2D layer followed by a Dense layer with a SoftMax activation function. And we compile the model with a categorical cross-entropy loss function, Adam optimizer, and accuracy metric.

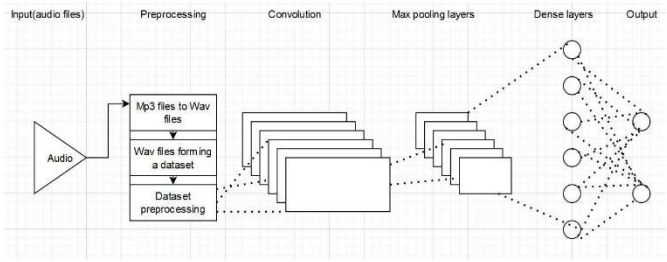

Figure 2. Architecture Diagram of the CNN model

IV. MODEL EVALUATION

A. MLP:

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: MLP Model Summary

Now plotting the validation and training loss to identify issues such as overfitting or underfitting and make decisions about adjusting the model's parameters.
Validation and training loss plots are commonly used to monitor the performance of a machine learning model during the training process.

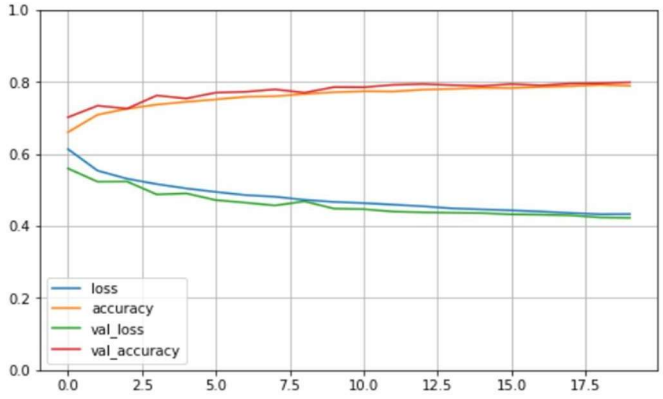

Figure: Plots

From the above figure it's clear that the training loss and validation loss both decrease and stabilize at a specific point, hence the model is a good enough fit.
As the model is a good fit, we can proceed to evaluate the test data. After evaluating this MLP model, we observed that the model achieves 93.75% Training accuracy and 87.50% of testing accuracy and 79.51% Prediction accuracy.

B. CNN:

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: CNN Model Summary

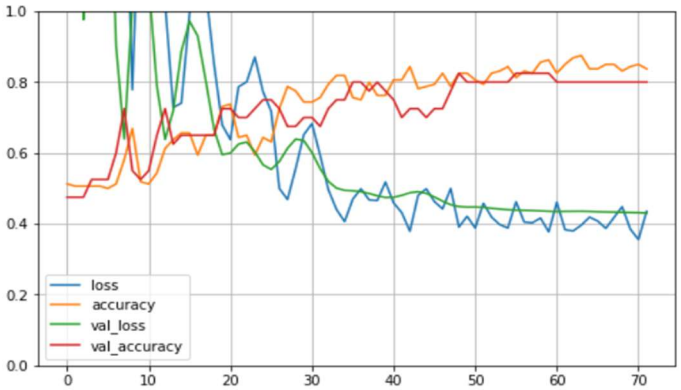

Figure: Plots

From the above figure it's clear that the training loss and validation loss both decrease and stabilize at a specific point, hence the model is a good enough fit.
As the model is a good 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.

V. RESULTS

S.no	Classifier	Training Accuracy	Testing Accuracy	Prediction Accuracy
1.	MLP	93.75%	87.50%	79.51%
2.	CNN	86.87%	80.00%	89.23%

VI. REFERENCES

[1] 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/ICOSEC49089.2020.9215254.

[2]M. A. Uddin, M. S. Hossain, R. K. Pathan and M. Biswas, "Gender Recognition from Human Voice using Multi-Layer Architecture," 2020 International Conference on INnovations in Intelligent SysTems and Applications(INISTA),Novi Sad, Sebriya, 2020 pp. 1-7, doi: 10.1109/INISTA49547.2020.9194654.

[3] 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,pp. 1-4, doi: 10.11 09/INOCON50539.2020.9298241.

[4]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 andSecurity*, Nanchang, China, 2010, pp. 84-88, doi: 10.1109/ISECS.2010.27.

[5] M. Ichinof, N. Komatsuff, W. Jian-Gangfff and Y. W. Yunffj, "Speakergender 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.

[6]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, pp.55-61, doi: 10.1109/SMART46866.2019.9117385.

[7]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. 319-324, doi: 10.1109/ICOSEC49089.2020.9215254.

[8]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, pp. 32-37, doi: 10.1109/Confluence47617.2020.9058146.

[9] 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, pp. 338-343, doi:10.1109/COMPSAC51774.2021.00055.

[10] A. B. Gumelar et al., "Human Voice Emotion Identification Using Prosodic and Spectral Feature Extraction Based on Deep Neural Networks," 2019 IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH), Kyoto, Japan, 2019, pp. 1- 8, doi: 10.1109/SeGAH.2019.8882461.

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

[12] A. A. M. Abushariah, T. S. Gunawan, J. Chebil and M. A. M. Abushariah, "Voice based automatic person identification system using Vector Quantization," 2012 International Conference on Computer and Communication Engineering (ICCCE), Kuala Lumpur, Malaysia, 2012, pp. 549-554, doi: 10.1109/ICCCE.2012.6271247.

[13] 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, pp. 151-152 vol.1, doi: 10.1109/IEMBS.2002.1134431.

[14] 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, pp. 8427-8436, doi:10.1109/CVPR.2018.00879.

[15] K. Kobayashi, T. Toda and S. Nakamura, "F0 transformation techniques for statistical voice conversion with direct waveform modification with spectral differential," 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, USA, 2016, pp. 693-700, doi: 10.1109/SLT.2016.7846338.

[16] 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 April-June 2022, doi:10.1109/MPRV.2022.3163656.

[17] M. Yousefi and J. H. L. Hansen, "Block-Based High Performance CNN Architectures for Frame-Level Overlapping Speech Detection," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 28-40, 2021, doi: 10.1109/TASLP.2020.3036237.

[18] L. Sarı, M. Hasegawa-Johnson and C. D. Yoo, "Counterfactually Fair Automatic Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3515- 3525, 2021, doi: 10.1109/TASLP.2021.3126949.

[19] 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, vol. 58, no. 3, pp. 574-586, March 2011, doi: 10.1109/TBME.2010.2091640.

[20] S. A. Almaghrabi et al., "The Reproducibility of Bio-Acoustic Features is Associated With SampleDuration, Speech Task, and Gender," in IEEE Transactions on Neural Systems and RehabilitationEngineering, vol. 30, pp. 167-175, 2022, doi:10.1109/TNSRE.2022.3143117.