# Voice-based Gender and Age Recognition System

**Conference Paper** · June 2023

DOI: 10.1109/InCACCT57535.2023.10141801



**5 authors**, including:

Vinayak Sudhakar Kone
KLE Technological University
**5** PUBLICATIONS **17** CITATIONS
SEE PROFILE

Pranali Jadhav
KLE Technological University
**1** PUBLICATION **11** CITATIONS
SEE PROFILE

Atrey Mahadev Anagal
KLE Technological University
**5** PUBLICATIONS **18** CITATIONS
SEE PROFILE

Uday Kulkarni
B.V. Bhoomaraddi College of Engineering and Technology (BVBCET)
**62** PUBLICATIONS **382** CITATIONS
SEE PROFILE

2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT) | 979-8-3503-9648-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/InCACCT57535.2023.10141801

# *Voice-based Gender and Age Recognition System*

Vinayak Sudhakar Kone
Dept. of CSE
KLE Technological University
Hubballi, India
vskonenpn@gmail.com

Atrey Anagal
Dept. of CSE
KLE Technological University
Hubballi, India
atreyanagal@gmail.com

Swaroop Anegundi
Dept. of CSE
KLE Technological University
Hubballi, India
swaroopra2001@gmail.com

Pranali Jadhav
Dept. of CSE
KLE Technological University
Hubballi, India
pranalipjadhav2002@gmail.com

Uday Kulkarni
Asst. Professor, Dept of CSE
KLE Technological University
Hubballi, India
uday_kulkarni@kletech.ac.in

Meena S M
Head, Dept. of CSE
KLE Technological University
Hubballi, India
msm@kletech.ac.in

***Abstract*— The ability to detect gender and age from voice is a valuable tool in a variety of applications, like voice-based biometric identification, natural language processing, and speech recognition. Recent advances in Deep Learning have enabled the development of highly accurate gender and age detection models. In this paper, the discussion is about the Machine Learning based gender and age detection model using voice. The various approaches used to extract features from speech, and the data-set used for model evaluation and classification are obtained using different Machine Learning algorithms. The discussion is about the opportunities and challenges in this area of research. It is concluded by highlighting some of the open challenges and future directions in this field. Age prediction from voice using a grid search pipeline is a Machine Learning technique that uses a range of algorithms to detect the age of a person using their voice. In the proposed model, RobustScalar, Principal component analysis (PCA), and Logistic Regression algorithms are used. The grid search pipeline uses a combination of models to identify the best age prediction algorithm for a given data-set. For Gender prediction sequential model with 5 hidden layers has been used. The results were obtained based on the trained model for the common voice data-set with an accuracy of around 91% for gender and 59% for age.**

***Keywords*— *Convolutional Neural Network, MFCC, Mel-Spectrogram, PCA, FFT, STFT, Tonnetz.***

## I. INTRODUCTION

Communication plays an integral part in human interaction. Without the ability to communicate, it would be difficult to express and deliver thoughts and ideas to others. Speech is a primary form of communication and is considered the most fundamental way of communicating with others. Speech allows us to convey feelings, share opinions, and express ideas. It also helps to connect with people, build relationships, and resolve conflicts. Furthermore, it allows the creation of an understanding between two or more people and gets the point across more effectively. In conclusion, communication is a key factor in connecting people and is essential for creating and maintaining relationships.

Using speech to convey thoughts is an essential part of human interaction. The voice carries many different characteristics which vary from person to person, providing insight into emotional and mental states, age, and gender. This multidimensional information can be used in many different applications, including voice-based security systems, automated voice systems, and speech-based AI assistants [10], Automatic Speaker Verification [11], emotion [21] based AI systems. Voice-based input systems are also being used to reduce the search space in databases. Moreover, voice-based gender and age detection can be used in Artificial Intelligence-based security systems [1], crime-solving, and victim protection. The multidimensional features of speech [2] are invaluable in many modern technologies. The male audio signal as a waveform is shown in Fig. 1. The female audio signal as a waveform is mentioned in Fig. 2.

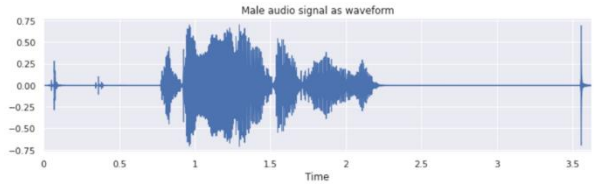



Fig. 1. Male Audio Signal as Waveform

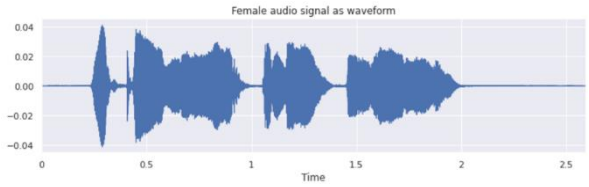



Fig. 2. Female Audio Signal as Waveform

A spectrogram [3], a waveform of the human's voice, can vary from one person to another, gender to gender, and age to age. As the age of a speaker increases, the pitch and frequency of voiced sounds [16] tend to lower [4]. With labelled training data, supervised Machine Learning algorithms can effectively make predictions that would otherwise go undetected. Speech analysis looks at features like gender, age, and emotion to make predictions. Detecting these features in speech could greatly aid the interaction between humans and machines. For instance, audio files can be categorized based on the speaker's age and gender, allowing companies like telecommunication services to predict customer demographics, age, gender, and emotions, and make suitable offers.

Gender prediction using voice in Deep Learning is a powerful tool for predicting a person's gender based on the



*ISBN: 979-8-3503-9648-5/23/$31.00 ©2023 IEEE*

sound of their voice. Depending on the complexity of the model, it can involve convolutional [6] [8], recurrent [7], temporal CNN [30], or fully connected layers. Deep Learning models are composed of layers that extract features from the audio file such as tone and pitch [9]. Once trained, the model can be used to make predictions on new audio files. Age prediction from the voice in Machine Learning is a research area that uses Machine Learning Algorithms to identify age from a voice recording. This involves identifying characteristics like pitch, timbre [12], amplitude, and other acoustic features [23]. Principal component analysis (PCA) [15] can be used to identify patterns in the data-set and identify significant features which can be used to accurately predict age.

The paper is organized as Section II of the paper presents an overview of existing models and studies related to the topic. This section provides valuable insights into the problem. It helps to identify potential areas of improvement, identify gaps in existing knowledge, and suggest alternative approaches. It also provides evidence of the effectiveness of a proposed solution or evidence of potential difficulties. Section III outlines the model that has been implemented. In Section IV, information regarding the data-set used for both testing and training is given. The findings of this report are outlined in Section V, which is followed by a finding summary and possible future direction in Section VI.

## II. Related work

The author of this paper [17] explored the potential of using Deep Neural Network based embedded architectures like x-vector, d-vector, and y-vector for the estimation of age and classification of gender tasks. A Vox-Celeb1 data-set pre-trained embedded network was used to develop a transfer learning-based training model for joint gender classification and age estimation tasks. The findings of this research demonstrate the capability of these models to capture various features of signals to make meaningful comparisons about how these systems learn about age estimation and gender recognition. The benefit of this model is that it leveraged the existing information of pre-trained models, while still being able to adapt to the specific characteristics of the data in the database used for this research. This demonstrates the potential of transfer learning to effectively incorporate the diversity of the human voice.

The author of this paper [18] presents a system for the detection of gender, emotion, and age from speech. Audio clips are analyzed through frequency spectrum analysis (FSA) to generate data-set for building predictive models. Ten different Machine Learning models are assessed across four distinct performance metrics such as F1 scores, accuracy and precision to identify the optimal solution for each task. CatBoost is found to be most accurate for gender detection, Random Forest for age detection, and XGBoost for emotion detection. The limitations of this model are identified to be a lack of accuracy when audio clips contain voices from multiple people or a high level of noise.

The author of this paper [19] presents a gender and age recognition system utilizing incoherent generative models, which are trained using sparse non-negative matrix factorization and post-processed with steps involving atom correction. Mel-frequency cepstral coefficients are used to extract the features for a good depiction of voice structure. The experimental results demonstrate that the proposed algorithm achieves better recognition rates than the other related algorithms. Additionally, evaluations in presence of background white Gaussian noise yielded the same results as those obtained in the absence of noise.

The author of this paper [22] discusses the use of deep feed-forward networks, or multi-layer perceptron (MLP), for supervised learning has been well established. In particular, these networks detect the gender of a speaker by leveraging the acoustic properties of their voice has become a popular area of research. To create the model, an MLP network is employed which is made of an input layer, the hidden layer of computation nodes, and the output layer. The back-propagation method is then used to train the network. One key component of such a model is the Specan Function which is used to measure 22 acoustic parameters [23] on acoustic signals.

## III. Proposed work

By utilizing the librosa library [24], it is possible to pre-process voice samples and generate their respective spectrograms [3].

1) Voice Pre Processing - Voice preprocessing is a process of enhancing and modifying the audio signal to improve the quality of the signal before it is further processed by a speech recognition system. Preprocessing can help in reducing the effect of environmental noise, enhance the accuracy of speech recognition systems, and reduce the computational complexity of processing.

2) A/D Signal Conversion - By using standard sampling and quantization techniques, A/D signal conversion converts raw analog audio spectrogram [3] to understand the digital signals that can be processed by a machine.

3) Pre-emphasis process - Pre-emphasis [25] is a process used in voice pre-processing to increase the relative importance of higher frequency components in an audio signal. It is achieved by adding a high-pass filter to the signal, which amplifies the higher frequencies and attenuates the lower frequencies. Pre-emphasis [25] is used to reduce the amount of noise introduced by the recording or transmission medium, and can also be used to improve the intelligibility of speech. And the equation is given by

$$H(z) = 1 - \mu z^{-1} \tag{1}$$

where $z$ shows the filter and $\mu$ is the pre-emphasis filter coefficient with a value lying commonly between [0.9, 1].

4) Frame Blocking and Hamming Window - Frame Blocking [26] is a process used in voice pre-processing to divide the audio signal into smaller frames. Each frame is then processed separately. This process helps to reduce noise and improves the quality of the audio signal. It is called Frame Blocking. Hamming Window [26] is a windowing technique used in voice pre-processing. It is used to reduce noise and improve the quality of the audio signal. It works by multiplying the audio signal by a window function, which reduces the energy at the start and end of each frame. This helps to reduce any artifacts caused by abrupt starts and endpoints. The equation is given by

$$\omega(n) = 0.54 - 4.4cos\left(\frac{2\pi n}{(N-1)}\right), 0 \le n \le N-1 \tag{2}$$

here $\omega(n)$ represents window operation, the number of individual samples is represented by $n$, and the total number of speech samples is denoted by $N$.

5) Fast Fourier Transform (FFT) - The Fast Fourier Transform (FFT) [27] is a powerful tool used in voice processing. It can be used to analyze the frequency spectrum of a signal and to detect the presence of specific frequencies. The FFT is used in varieties of applications including sound synthesis, speech recognition, and audio compression. In speech recognition, the FFT is used to identify the fundamental frequency [13] of a spoken word or phrase, as well as to detect patterns of speech. The equation is given by

$$X_k = \sum_{n=0}^{N-1} n e^{-(j2\pi kn/N)}\ k = 0,1,2,\dots\dots N-1 \qquad (3)$$

where $X_k$ denotes a complex number as a frequency magnitude or modulus. $X_k$ is sequence shown as follows: positive frequency $0 \le f < (1/2)Fs$ relates to values $0 \le n \le (1/2)(N-1)$, whereas, negative frequency $-(1/2)Fs < f < 0$ relates to $(1/2)(N+1) \le n \le N-1$. $Fs$ depicts the sampling rate. The results obtained are known as frequency spectrum of the voice signal.

6) Mel-Spectrogram - Mel-Spectrogram is a type of audio data representation that is used to represent the spectral content of audio signals [28]. It is a 2D representation of frequency versus time. It is used in speech and music analysis and is often used to extract features for Machine Learning algorithms. It is based on the Mel scale [28], which assigns to each frequency bin, allowing for a more natural reproof of audio data. The Mel-frequency scale's equation is given by

$$\text{mel} = 2595 * \log_{10}\big((1 + hertz)/700\big) \qquad (4)$$

7) MFCC - Mel-Frequency Cepstral Coefficients (MFCCs) are a set of features used to represent a voice signal. They are calculated from the frequency spectrum of a signal and are used in automatic speech recognition and other speech-related applications. MFCCs capture the characteristics of a voice signal, such as the shape of the vocal tract [14], and can be used to distinguish between different speakers.

### *A. Proposed Architecture for Gender Detection}*

To detect the gender of a voice sample from a speaker, the Sequential model [5] with 5 hidden layers has been used. Deploying the Common Voice data-set of Mozilla, a collection of speech data collected by users on the website of Common Voice. Its purpose is to enhance automated speech recognition by allowing it to be trained and tested. The following things have been done in order to prepare the data-set.

- Initially, only filtration of the genre field for labeled samples is done.
- In order to avoid over-fitting, balance the data-set such that there are equal numbers of female and male samples. After balancing the data-set there were total samples of 66938 voice samples of which 33469 were males and females. Gender distribution is shown in Fig.3. This will ensure that the Neural Network does not over-fit a particular gender.

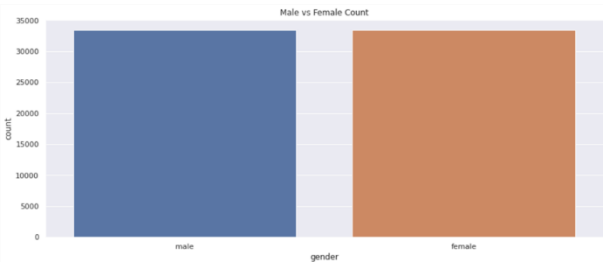
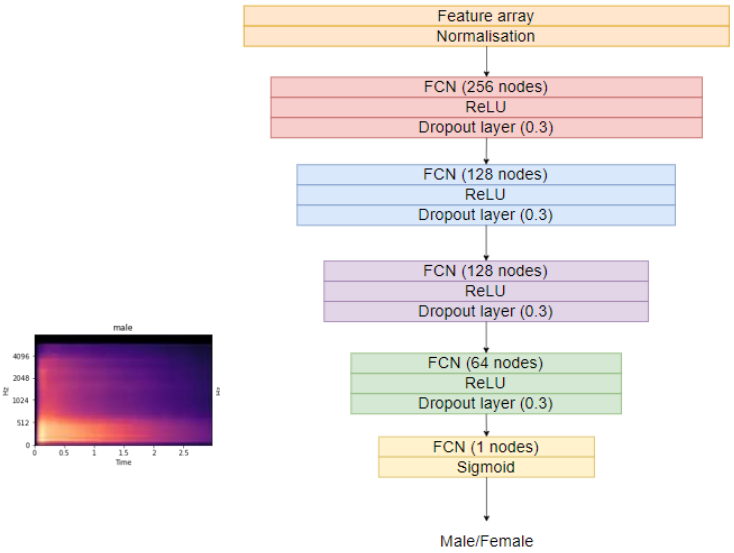



Fig. 3. Female Audio Signal as Waveform




The load_data() function reads the CSV file and combines all audio samples into a single array. As a summary, the label2int dictionary simply maps string labels to integer values; the need for the load_data() function is to translate string labels to integer labels and to split the data-set into training, testing, and validation data-set. As shows in Fig.4. The voice signal is normalized to minimize the huge variable differences between all the samples.

The data-set is shuffled and divided into testing and training sets, and rerun the algorithm on the training set is to get the validation set using Sklearn's train_test_split() function.

Fig. 4. Value distributions of the extracted features after normalization

#### *1) Building the Model*

A deep feed-forward Neural Network is a type of Artificial Neural Network with 5 hidden layers which are made of multiple different layers of neurons that are connected in a manner that passes information from the input layer to the output layer. The 5 hidden layers are composed of neurons arranged in an interconnected network, and each layer is responsible for a specific task. The neurons in each layer are activated by an activation function, like the Rectified Linear Unit (ReLU) or sigmoid activation functions, which determine the output of the layer.

The ReLU activation function is a type of activation function that gives a value of 0 if the input is negative and gives an input value if the input is positive. This activation function is used in applications of Deep Learning, as it tends to get better performance than other activation functions, like the sigmoid activation function. The sigmoid activation function is a type of activation function that gives a value between 0 and 1, depending on the input. This activation function is mostly used in Deep Learning applications where the output needs to be in a range between 0 and 1, like in classification problems.

In Fig.5 , initially, normalization is performed, later in the first hidden layer where ReLU is used as an activation function, consisting of 256 units with 0.3 dropout, succeeded by two hidden layers where ReLU is used as an activation function, consists of 128 units with 0.3 dropout and 64 units with 0.3 dropout in the next layer. Finally, there is the sigmoid activation function which helps to classify as male or female.

Fig. 5. Neural Network Model to detect Gender

This type of regularization attempts to avoid over-fitting on the training data-set, where each fully connected layer is followed by a 30% dropout rate. The output layer is composed of one neuron with a sigmoid activation function, the model outputs 1 when the speaker is male (or close to 1), and 0 when the speaker is female. Furthermore, binary cross-entropy has been chosen as the loss function, since it is a specific instance of a categorical cross-entropy when there are only two classes to predict. The neural Network Model to detect Gender is shown in Fig. 5.

*2) Training the Model*

A five-second patience will be specified so that the training will end when the model stops improving. This restores the best weights and assigns the optimal weights calculated during training. On epoch 21, the validation accuracy was almost 91 %. The model training ended at epoch 30. Fig. 6. depicts Training and Validation Accuracy.

*3) Testing the Model*

Recording the voice and saving it to a file, which is then fed to the model as features to return results (can speak in any language).


Fig. 6. Training and Validation Accuracy

*B. Proposed Architecture for Age Detection*

There are many different frequencies, and it depends on what the sound is. A low frequency, like 60 Hz, might be the sound of a bass guitar, while a high frequency, like 8000 Hz, might come from a bird song and it changes for a song by a human [29]. The frequency of human speech is usually somewhere in the middle.

The understanding of how quickly a signal needs to be interpreted depends on both the sampling rate and the sampling rate per second. In this case, the sampling rate per second is 16kHz. To detect the age of a voice sample from the speaker, grid searchCV which consists of Robustscalar, PCA [15], and Logisticregression arranged in form of a pipeline has been used.

*1) Feature extraction :*

*a) Onset Detection :*

Librosa is able to reasonably well identify the onset of new spoken words based on the waveform of a signal. Fig. 7. shows words in particular Voice.

*b) Tempo :*

Human speech is very melodic, and each person speaks in very different ways and at different speeds. Thus, an audio signal is the tempo of the speech, that is, how many beats are present in it.

*c) Fundamental frequency :*

The fundamental frequency [13] (or f0) is the lowest bright horizontal band in the image at which a periodic sound appears. While the repetition of the strip pattern above this fundamental is called harmonics. The fundamental frequency [13] can be seen in Fig. 8.


Fig. 7. Words in particular Voice

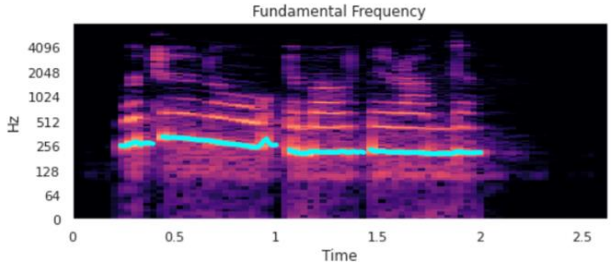



Fig. 8. Fundamental Frequency

*2) Exploratory data analysis (EDA) on the audio data-set :*

After gaining an understanding of audio data and how to process it, now proceed to an EDA. The data-set for this paper is taken from voice.mozilla.org. Since the data-set is huge, a much smaller subsample of roughly 10000 audio files is taken into consideration.

*a) Investigation of features distribution:*

The bar graph depicts the distribution of counts of age categorized as the eighties, seventies, sixties, fifties, forties, thirties, twenties, and teens. The plot of age distribution is shown in Fig. 9.

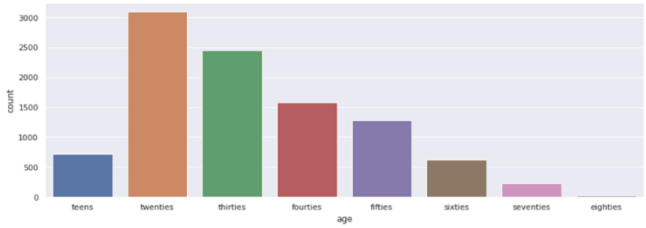



Fig. 9. Age distribution

*b) Extracted features:*

Apart from words_per_second, the majority of these feature distributions are skewed to the right and would benefit from a log transformation. Fig. 10. shows the value distribution of extracted features after normalization.

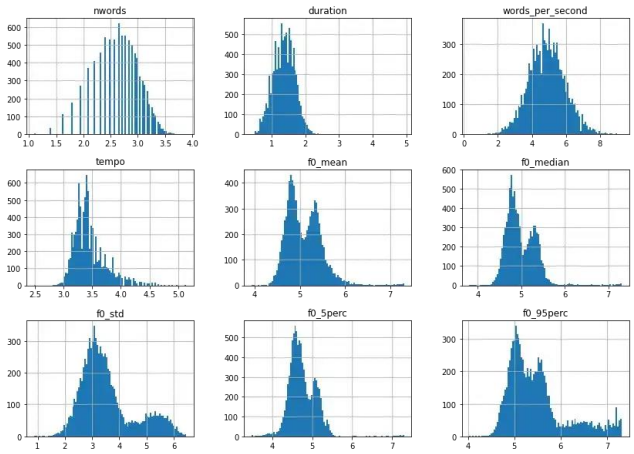



Fig. 10. Value distributions of the extracted features after normalization

*c) Feature correlation:*

Taking a look at the correlation between all features, now, encode the non-numerical target features as well. Note that encoding the non-numerical features could disrupt the age feature's correct order. So there is a need to perform a manual mapping. The feature correlation is depicted in Fig 11.

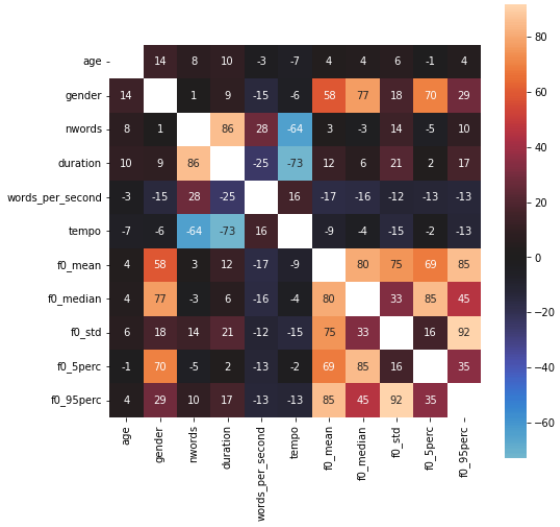



Fig. 11. Feature Correlation

*d) Spectrogram features:*

As a result, the spectrograms [3] will also be different in length. Therefore, in order to normalize all recordings, cut them exactly at 5 seconds in order to convert them into audio samples with the same length. For this reason, all recordings will be cut to exactly 5 seconds in length to normalize them, so that samples that are too short will be truncated, while samples that are too long will be shortened.

By combining the comma-separated values (CSV) data with the Mel strength of spectrograms, a tabular data-set was developed for the data from the CSV file containing the Mel Strength of each audio file as 128 parameters of Mel Strength. Above data-set which is used for model building.

*3) Building the models*

PCA [15] is often used as a pre-processing step in a Machine Learning pipeline for feature selection, feature engineering, and feature extraction [20]. PCA [15] can be combined with object scalar to further improve the performance of predictive models. Object scalar is a powerful unsupervised learning technique used to extract meaningful information from unlabeled data-sets. By combining PCA [15] and object scalar, the data-set can be reduced to a smaller number of features while still preserving the most important information. This can be beneficial for reducing the dimensionality of the data-set and improving the accuracy of predictive models.

After the following processes, the Simple Logistic Regression model is used to predict the age of a speaker. For this, there is a need to use a Pipeline object, and explore the benefits of certain pre-processing routines. Thus Pipeline is built and fed into GridSearchCV, which is used to explore and cross-validate different combinations of hyperparameters. The Data frame output table of hyperparameters can be seen in Fig. 12.

Additionally, the plot of the performance score for each combination of hyperparameters is based on the DataFrame output. However, given that there are multiple scalers and PCA [15] approaches, each plot will need to be created separately. The graph plot for the combination of hyperparameters is given in Fig. 13.

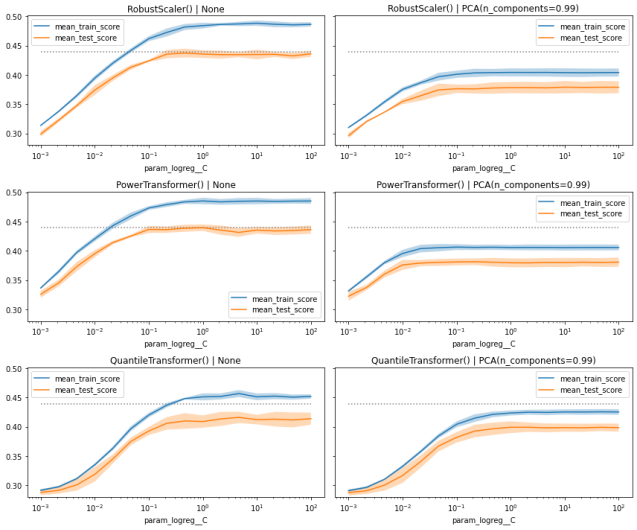
| | param_scaler | param_pca | param_logreg__C | mean_test_score | mean_train_score | std_test_score | std_train_score |
|---|---|---|---|---|---|---|---|
| 55 | PowerTransformer() | None | 1.0 | 0.439508 | 0.485124 | 0.005489 | 0.005539 |
| 49 | PowerTransformer() | None | 0.464159 | 0.438499 | 0.483538 | 0.005958 | 0.003447 |
| 48 | RobustScaler() | None | 0.464159 | 0.437203 | 0.481663 | 0.007420 | 0.005240 |
| 37 | PowerTransformer() | None | 0.1 | 0.436482 | 0.473059 | 0.005968 | 0.003246 |
| 43 | PowerTransformer() | None | 0.215443 | 0.436192 | 0.478923 | 0.005446 | 0.004047 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 3 | RobustScaler() | PCA(n_components=0.99) | 0.001 | 0.296178 | 0.310118 | 0.004719 | 0.001384 |
| 8 | QuantileTransformer() | None | 0.002154 | 0.291420 | 0.297573 | 0.005419 | 0.001818 |
| 11 | QuantileTransformer() | PCA(n_components=0.99) | 0.002154 | 0.290699 | 0.296563 | 0.005288 | 0.002046 |
| 2 | QuantileTransformer() | None | 0.001 | 0.287959 | 0.291613 | 0.004569 | 0.001804 |
| 5 | QuantileTransformer() | PCA(n_components=0.99) | 0.001 | 0.287670 | 0.290988 | 0.005001 | 0.001787 |

Fig. 12. Data frame output table of hyperparameters

Fig. 13. Graph plot for the combination of hyperparameters

It is often helpful to visualize the performance metrics as curves because the curves provided additional information that is otherwise unavailable when only looking at the pandas Data Frame.

There is a significant difference between train and test performance, especially when using models that drop off more quickly. The model demonstrated a 59 % accuracy rate when evaluated on test data.




## IV. Data-set

The data-set comprises data of users which is available on voice.mozilla.org and it is based on the text collected from the

public. The sources from which the voice samples have been collected are blog posts, movies, and old books. This data-set is mainly used to train and test automatic speech recognition (ASR) systems.

### A. *Data-set Structure*

The data-set is divided into different sections for the sake of easiness. Audio file with valid in their names is the audio files that have been listened to by at least 2 people.

### B. *Data-set Grouping*

The valid file is divided into 3 groups.

- dev - this folder contains the audio files which are used for development and experimentation. It includes 1397 voice samples.
- train - this folder contains the audio files which are used for the purpose to train speech recognition. It includes 64220 voice samples.
- test - to test the rate of accuracy of the model. It includes 1321 voice samples.

Each row of a CSV file depicts an individual clip of audio, and carries the below information:

- filename - the path of the audio file directory.
- text - the purported transcription of the audio.
- up_votes - number of people who indicated that the audio aligned with the text.
- down_votes - number of people who indicated that the audio is not aligned with the text.
- age - the speaker's reported age. (Table I. lists the age variation among speakers)

TABLE I. SPEAKER'S AGE RANGES

| **Speaker's Age Ranges** | |
|---|---|
| **Label** | **Range** |
| Teens | < 19 |
| The twenties | 20 - 29 |
| The Thirties | 30 - 39 |
| The forties | 40 - 49 |
| The fifties | 50 - 59 |
| The Sixties | 60 – 69 |
| The seventies | 70 – 79 |
| The eighties | 80 - 89 |
| The nineties | > 90 |

## V. EXPERIMENTAL RESULTS

By looking at the values in the confusion matrix, we can calculate several performance metrics such as F1 scores, accuracy, recall and precision. These metrics are used to assess the total model's performance in accurately predicting gender.

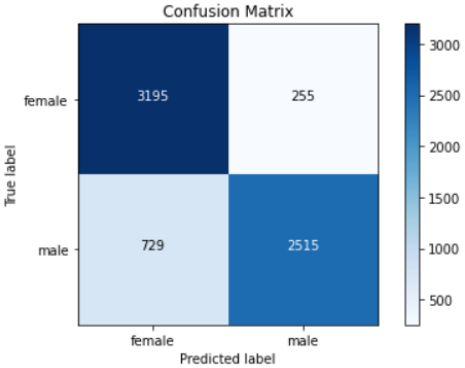



Fig. 14. Confusion Matrix for Gender Prediction

In Fig. 14. the Confusion Matrix for gender prediction is shown. The confusion matrix shows that the model was better at predicting gender with an accuracy of 91%. The classification model scored very well, but there is one thing that needs to be more understood. Creating a brief helper function to help to analyze the confusion matrix. This will give a better understanding of which categories the model predicted correctly, and which ones it struggled with. With this information, they can then adjust the model to improve overall accuracy. The confusion matrix is given in Fig. 15.

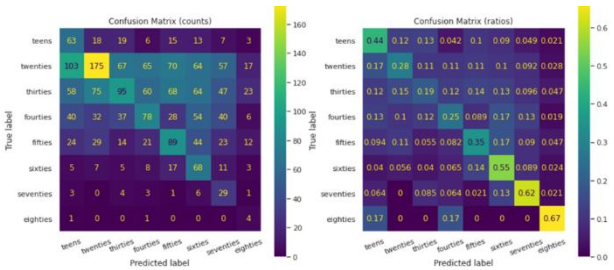



Fig. 15. Confusion Matrix for Age Prediction

The left confusion matrix shows that the model was better at detecting twenties samples than others (e.g., with an accuracy of 59%), it did better in classifying teens and sixties entries.

Training the model on a system with specifications outlined in Table II, set the hyperparameters as 100 epochs, 64 as batch size, and 0.01 as the learning rate. The training process took approximately 47 minutes.

TABLE II. THE HARDWARE SPECIFICATIONS OF THE SYSTEM USED TO TRAIN THE MODEL

| **System Specifications** | |
|---|---|
| Processor | 10th generation Intel Core i5-1035G1 quad-core processor |
| Graphics | NVIDIA GeForce MX250 (2 GB DDR5 dedicated) |
| Memory | 8 GB DDR4-2666 SDRAM |
| Storage | 512 GB PCIe NVMe M.2 Solid State Drive |

## VI. CONCLUSION AND FUTURE ENHANCEMENTS

Voice Recognition System has shown promising results in identifying the gender and age of a speaker. By using the sequential model, the model has shown good results in terms of gender prediction and by using the grid search model it was able to predict age. But the age prediction had some limitations like the model was unable to predict age accurately for voices with different accents. The Voice Based age prediction and with further improvements to the system, can be used in a variety of applications like fraud detection, surveillance, and customer service. Additionally, more accurate results can be obtained by using more advanced technologies like Artificial Neural Networks and Deep Learning algorithms. This could potentially lead to more accurate results and better applications. In the future, research can be conducted to make the system faster, more reliable, and more accurate. Furthermore, research can be conducted to make the system more robust to different accents and dialects.

## REFERENCES


[1] X. Qiu, Z. Du, and X. Sun, "Artificial Intelligence-Based Security Authentication: Applications in Wireless Multimedia Networks," IEEE Access, vol. 7, pp. 172004–172011, 2019, doi: https://doi.org/10.1109/access.2019.2956480.

[2] Ashok Kumar Konduru and M. Iqbal, "Multidimensional feature diversity based speech signal acquisition," International Journal of Speech Technology, vol. 23, no. 3, pp. 527–535, Sep. 2020, doi: https://doi.org/10.1007/s10772-020-09736-5.

[3] W. Endres, W. Bambach, and G. Flösser, "Voice Spectrograms as a Function of Age, Voice Disguise, and Voice Imitation," The Journal of the Acoustical Society of America, vol. 49, no. 6B, pp. 1842–1848, Jun. 1971, doi: https://doi.org/10.1121/1.1912589.

[4] S. E. Linville, "Source Characteristics of Aged Voice Assessed from Long-Term Average Spectra," Journal of Voice, vol. 16, no. 4, pp. 472–479, Dec. 2002, doi: https://doi.org/10.1016/s0892-1997(02)00122-4.

[5] Yair Even-Zohar and D. Roth, "A Sequential Model for Multi-Class Classification," Empirical Methods in Natural Language Processing, Jun. 2001.

[6] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

[7] L. C. Jain and L. R. Medsker, Recurrent Neural Networks. Informa, 1999. doi: 10.1201/9781420049176.

[8] Sainath, Tara N., et al. "Convolutional, long short-term memory, fully connected deep neural networks." 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). Ieee, 2015.

[9] Fairbanks, Grant, and Wilbert Pronovost. "An experimental study of the pitch characteristics of the voice during the expression of emotion." Communications Monographs 6.1 (1939): 87-104.

[10] A. Maedche et al., "AI-Based Digital Assistants," Business & Information Systems Engineering, vol. 61, no. 4, pp. 535–544, Jun. 2019, doi: https://doi.org/10.1007/s12599-019-00600-8.

[11] A. E. Rosenberg, "Automatic speaker verification: A review," Proceedings of the IEEE, vol. 64, no. 4, pp. 475–487, 1976, doi: https://doi.org/10.1109/proc.1976.10156.

[12] T. F. Cleveland, "Acoustic properties of voice timbre types and their influence on voice classification," The Journal of the Acoustical Society of America, vol. 61, no. 6, pp. 1622–1629, Jun. 1977, doi: https://doi.org/10.1121/1.381438.

[13] H. Fujisaki, "Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing," Springer eBooks, pp. 39–55, Jan. 1983, doi: https://doi.org/10.1007/978-1-4613-8202-7_3.

[14] E. Joliveau, J. Smith, and J. Wolfe, "Vocal tract resonances in singing: The soprano voice," The Journal of the Acoustical Society of America, vol. 116, no. 4, pp. 2434–2439, Oct. 2004, doi: https://doi.org/10.1121/1.1791717.

[15] H. Abdi and L. J. Williams, "Principal component analysis," Wiley Interdisciplinary Reviews: Computational Statistics, vol. 2, no. 4, pp. 433–459, Jun. 2010, doi: https://doi.org/10.1002/wics.101.

[16] P. Jafarian and M. Sanaye-Pasand, "A Traveling-Wave-Based Protection Technique Using Wavelet/PCA Analysis," IEEE Transactions on Power Delivery, vol. 25, no. 2, pp. 588–599, Apr. 2010, doi: https://doi.org/10.1109/tpwrd.2009.2037819.

[17] D. Kwasny and D. Hemmerling, "Gender and Age Estimation Methods Based on Speech Using Deep Neural Networks," Sensors, vol. 21, no. 14, p. 4785, Jul. 2021, doi: https://doi.org/10.3390/s21144785.

[18] Syed, Dipan Sadekeen, MAqib Alfaz, and Rifat Shahriyar, "One Source to Detect them All: Gender, Age, and Emotion Detection from Voice," Computer Software and Applications Conference, Jul. 2021, doi: https://doi.org/10.1109/compsac51774.2021.00055.

[19] Mavaddati, S. "Voice-based Age and Gender Recognition using Training Generative Sparse Model." International Journal of Engineering 31.9 (2018): 1529-1535.

[20] M. A. Uddin, R. K. Pathan, M. S. Hossain, and M. Biswas, "Gender and region detection from human voice using the three-layer feature extraction method with 1D CNN," Journal of Information and Telecommunication, vol. 6, no. 1, pp. 27–42, Oct. 2021, doi: https://doi.org/10.1080/24751839.2021.1983318.

[21] Umesh, A. S., and Ramesh Patole. "Automatic Recognition, Identifying Speaker Emotion and Speaker Age Classification using Voice Signal."

[22] M. Buyukyilmaz and A. O. Cibikdiken, "Voice Gender Recognition Using Deep Learning," Proceedings of 2016 International Conference on Modeling, Simulation and Optimization Technologies and Applications (MSOTA2016), 2016, doi: https://doi.org/10.2991/msota-16.2016.90.

[23] M. Araya-Salas and G. Smith-Vidaurre, "warbleR: anrpackage to streamline analysis of animal acoustic signals," Methods in Ecology and Evolution, vol. 8, no. 2, pp. 184–191, Sep. 2016, doi: https://doi.org/10.1111/2041-210x.12624.

[24] B. McFee et al., "librosa: Audio and Music Signal Analysis in Python," Proceedings of the 14th Python in Science Conference, 2015, doi: https://doi.org/10.25080/majora-7b98e3ed-003.

[25] Vergin, Rivarol, and Douglas O'Shaughnessy. "Pre-emphasis and speech recognition." Proceedings 1995 Canadian Conference on Electrical and Computer Engineering. Vol. 2. IEEE, 1995.

[26] Z. S. Bojkovic, B. M. Bakmaz, and M. R. Bakmaz, "Hamming Window to the Digital World," Proceedings of the IEEE, vol. 105, no. 6, pp. 1185–1190, Jun. 2017, doi: https://doi.org/10.1109/jproc.2017.2697118.

[27] Heckbert, Paul. "Fourier transforms and the fast Fourier transform (FFT) algorithm." Computer Graphics 2 (1995): 15-463.

[28] J. M. Gowdy and Zekeriya Tufekci, "Mel-scaled discrete wavelet coefficients for speech recognition," International Conference on Acoustics, Speech, and Signal Processing, Jun. 2000, doi: https://doi.org/10.1109/icassp.2000.861829.

[29] G. H. Wakefield, "Chromagram visualization of the singing voice.," MAVEBA, pp. 24–29, Jan. 1999.

[30] H. A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso, and M. Rosa-Zurera, "Age group classification and gender recognition from speech with temporal convolutional neural networks," Multimedia Tools and Applications, Jan. 2022, doi: https://doi.org/10.1007/s11042-021-11614-4.