2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC)

Human Voice Analysis to Determine Age and Gender

Bhushan Kiran Munoli
Dept. of ECE
PES University
Bengaluru, India
bhushankiranmunoli@pesu.ped.edu

K Abheeshta Kumar Jain
Dept. of ECE
PES University
Bengaluru, India
kabheeshtakumarjain@pesu.pes.edu

Prem Kumar
Dept. of ECE
PES University
Bengaluru, India
premkumar@pesu.pes.edu

Aditya Ram P S
Dept. of ECE
PES University
Bengaluru, India
adityaramps@pesu.pes.edu

Ashwini
Dept. of ECE
PES University
Bengaluru, India
ashwinib@pes.edu

2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC) | 979-8-3503-9619-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICRTEC56977.2023.10111890

***Abstract*— Age estimation from a speaker's voice is a research area that has garnered significant interest in recent years. The most popular method to extract voice features presently involves using MFCC coefficients which are the statistical features acquired from a kind of Spectral depiction of the audio clip (a nonlinear "spectrum-of-a-spectrum"). In the Mel scale, the divisions of the frequency band closely mimic the human auditory systems response. Post extraction, the needed features are extracted and the relevant classification or regression models/operations are applied. Numerous deep learning algorithms have been used so far in the studies for age and gender estimation. In this paper, the attempt is the same. A new labeled dataset is built by collecting the audio samples of people belonging to a variety of age groups and two genders (male and female) from the web. To this dataset the MFCC feature extraction program was applied; after extraction, different regression models were used to compute the age of the speakers and classification models to identify their genders.**

Keywords*— Age, MFCC, XGBoost, Decision Tree, CatBoost, Stochastic Gradient descent, Regression, Classification.

I. INTRODUCTION

Estimation of age and gender from audio samples provides a myriad of opportunities for investigation in several areas: targeted advertisements from google assistants, IVRs(interactive voice response), in criminal forensics, security access cases, etc[1]. The research in this field has started to gain notoriety only in recent years. There are multiple components to human speech: accent, gender, age, physical and mental conditions, and such several other attributes, physical and behavioural, influence how human speaks. In this project, the intention is to only make use of the frequency of the audio signals to estimate the age and gender of the speaker. As explained in the abstract, Mel frequency is used for extracting the features from voice. After that machine learning models are employed to estimate the age and gender of the speaker. Regression models are used to get the exact age of the person rather than a class or a range of age. This paper is structured in the following way: firstly in the abstract, the entire paper is briefly summarized and the findings are mentioned. Following the abstract, the introduction section where the subject and structure of the project is articulated and clarified. Literature survey follows the introduction and explains the research history of the previous years on the central subject of the project. Methodology succeeds the survey and elucidates the procedure employed in the project, and finally there is results and conclusions section that elaborates on the outputs of the research.

II. RELATED WORKS

In this section, the existing literature related to this project is observed. Presently, several studies are being published on the subject of Human age estimation from the voice of a speaker. A unique approach to identify gender, age, and emotion from audio and speech was presented by Syed Rohit Zaman et al. [2]. The suggested technique transformed all of the dataset's audio files into 20 statistical traits that could be quantified numerically. Later, many machine and deep learning prediction models (CatBoost, Random Forest Gradient Boosting, XGBoost, AdaBoost, K - closest neighbours (KNN), Decision Tree, Artificial neural networks (ANN), Naive Bayes, and Support vector machine (SVM)) were developed using the transformed numeric dataset. With a test accuracy of 70.4%, they discovered that the Random Forest algorithm outperformed all other prediction models at predicting age. Damian Kwasny and Daria Hemmerling applied deep learning methods to estimate age and gender using speech signals [3]. For the classification tasks, they used embedded deep neural network designs such as the d-vector and x-vector architectures. Voxceleb 1, Common Voice, and DARPA-TIMIT are the three datasets used. Using the well-known TIMIT dataset, they produced brand-new, cutting-edge age estimation results with mean absolute errors (MAE) of 5.12 years for men and 5.29 years for women and RMSEs of 7.24 years for men and 8.12 years for women. 99.60% of the time, gender could be correctly identified. A study from 2011 suggested a novel method for determining a speaker's gender and age using a hybrid architecture made up of Weighted Supervised Non-Negative Matrix Factorization (WSNMF) and General Regression Neural Networks (GRNN) [4]. The overall accuracy across three different age ranges and two genders (Young, Middle, Senior, Male, and female) was 96%. In the context of local languages and Lingua-Franca, another study [5] made an attempt to speculate on the gender and age of Nigerian speakers. There were three different audio lessons offered: (1) All subjects' English-language utterances (2) All subjects' Native-language utterances (3) The dominant Native tribe's English utterances (4) The dominant Native tribe's Native utterances. Overall accuracies obtained over the four models: 83.33%, 84.85%, 82.32%, and 85.1%, respectively. Study [6] compared the two popular approaches, Mel-Frequency Cepstrum Coefficient (MFCC) and Perceptual Linear Prediction (PLP), to extracting age estimation features from

the data. After feature extraction and dimensionality reduction, the voice features were fed to a Multi-layer Perceptron (MLP), which post training yielded the estimation results. The dataset used in the study was the Mozilla Common Voice. The age estimated belonged to one of the eight label classes; the highest accuracy recorded was 94.34%. [7] The paper that was published at the IEEE 17th India Council International Conference (INDICON) used MLP (Multi-Layered Perceptron) for gender and age group predictions where the authors of the paper applied and compared two different approaches (Single model and Sequential Model) to group predictions of gender and age. Using the Mozilla Common Voice dataset, they recorded an overall accuracy of 89.58%. LSTMs were used by a study to aid in the estimation of the age of the speaker based on short-duration utterances [8]; it outperformed the state-of-the-art results in some areas. GMM (Gaussian Mixture Model) and MFCC (Mel Frequency Cepstral Coefficients) have also been used to identify the gender of the speaker [9].

As you can see from the papers referenced here so far, most studies have used and still use MFCCs for the extraction of features from the speech signals [10], and so inspired by the positive results of those studies, this project also includes the same feature extraction procedure. Based on the literature surveyed, there are eight major models that are used often in the detection and estimation of age and; hence those eight models are used for the estimation and detection purposes and are compared with their outcomes.

III. METHODOLOGY

The following approach is adopted to this project methodology (It is compactly and visually illustrated in the flow chart provided at the end of this section (Figure 1)):

A. Procuring the Dataset:

The dataset used in this project is self-built. The audio samples were collected from the web by extracting them from various online video footage. The samples comprise the voices of people, men and women, belonging to the age range of 5-75 years [11,12,13,14,15,16,17,18].

B. Frequency Spectral Analysis:

Depicting signals in the frequency domain as sinusoidal and analysing their features, such as amplitude, phase, frequency, etc. is called Frequency Spectral Analysis. MFCC coefficients, a number of statistical features obtained from plotting a signal in the Mel frequency scale are used. Mel Frequency is a particular pattern of arrangement of the audio signals in the frequency domain, where divisions of the bands of these signals echo the construction of the human auditory system. In the project, to extract the Mel coefficients a sklearn based program is built.

MFCC coefficient computing procedure:

1. Set the sampling frequency for the audio and divide it into an even number of equal sized frames. This process is called windowing.
2. Take the DFT of the sampled signal.
3. The Mel-spaced filter bank should be computed. Mel frequency scale maps the actual frequency to the frequency which humans can perceive. The formula for this transformation is mentioned below (1).

$$Mel(f) = 1127\ln\left(1+\frac{f}{700}\right) \quad (1)$$

4. Human beings are less sensitive to the changes in frequency at higher audio ranges. Log function also works in a similar fashion. Thus, log is applied to the Mel frequency signal for better approximations.
5. After these modifications the signal are converted back into a time domain signal through IDFT. The inverse of the log transform of the magnitude of the signal is called cepstrum.
6. The MFCC model takes the first twelve coefficients of the IDFT, and in addition it includes the energy of the signal sample also as a feature and the first and second order derivatives of the signal. Thus, a total of 39 coefficients or Mel frequency features are obtained, among them in this project, 21 Mel features are extracted and used.

C. Applying deep/machine learning models:

After extracting the features, the next task to perform is age estimation. As previously mentioned, eight different models are used for this task (four classification models, and four regression models):

1. *Decision Tree Regressor:* Decision tree is a non-parametric supervised learning technique that can be employed towards both classification and regression purposes. By learning simple decision rules extracted from the characteristics of the data, this model can forecast the value of a target variable. You can think of a tree as a piecewise constant approximation.[19]
2. *CatBoost:* CatBoost or categorical boosting is modelled from gradient-boosted decision trees. A number of decision trees are constructed one after another during the training. Every second tree is given less loss than the first one. The number of trees is decided from the initial settings. When it is turned on, the tree stops growing. [20]. From CatBoost, classifier and regressor models are used.
3. *XGBoost:* Similar to CatBoost, XGBoost is also modelled based on the gradient boosted decision trees and hence the name (Extreme Boosting) [21]. Some differences between the two to be noted: Types of tree splits (symmetric and unsymmetric), computation speed (CatBoost is faster owing to the symmetric split), and the type of boosting (light in the case of CatBoost and Extreme in XGBoost). Again, both the Classifier and Regressor models are used.
4. *Stochastic Gradient Descent*: By merging various binary classifiers in an "OVA" (one versus all) framework, SGDClassifier provides multi-class classification. A binary classifier is learned for each class that can distinguish it from all other classes. When testing is completed, the score of each classifier's confidence is calculated (i.e., the signed distances to the hyperplane) and the class with the highest confidence is selected.

   The class SGDRegressor implements a simple stochastic gradient descent learning procedure for fitting linear regression models, which supports a variety of loss functions and penalties.[22]
5. *Output Prediction:* This step is self-explanatory. After the application of machine learning algorithms, the

individual accuracies were evaluated. Those results will be discussed in the Results and Conclusions section.

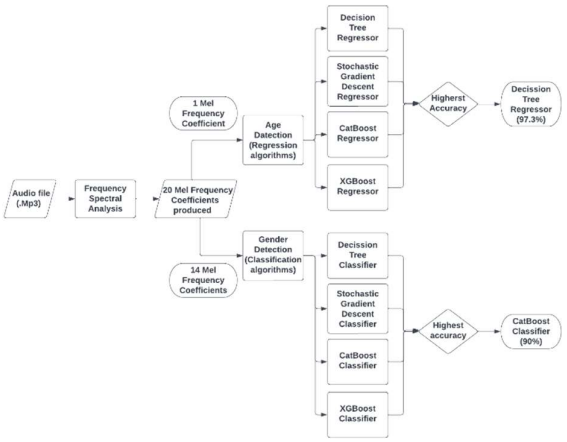

Figure 1. Visual illustration of steps of the project methodology

IV. RESULT AND CONCLUSIONS

Dataset: Self Prepared Dataset compiled by clipping the audio from YouTube videos. Age range of the voices varies from 5-75 years. Two genders: male and female. Total 494 samples (Test-train-split: 20:80).

The models used for predictions: Stochastic Gradient Classifier, Decision Tree Regressor, Logistic Regression, XGBoost classifier, CatBoost classifier, XGBoost Regressor, CatBoost Regressor, Stochastic Gradient Regressor.

Results: The percentage accuracy results of the models can be viewed from the following tables:

TABLE I. Results of Regression Models (Age Estimation)

Model	Accuracy
Decision Tree Regressor	97.3%
Stochastic Gradient Descent Regression	92.6%
CatBoost Regressor	65.7%
XGBoost Regressor	89%

TABLE II. Results of Classification Models (Gender Detection)

Model	Accuracy	Precision	F1-Score
CatBoost Classifier	90%	90%	90%
XGBoost Classifier	89%	88%	89%
Stochastic Gradient Descent Classifier	88%	88%	88%
Decision Tree Classifier	80%	80%	80%

Conclusions: Noting the results, it can be concluded that the Stochastic gradient regressor and Decision Tree Regressor perform the best at predicting the ages. For gender classification CatBoost, XGBoost, and SGD yield the closest and the most satisfactory results. In the future, the aim is to take the help of more factors such as the accent, vocabulary, etc. and improve the results even further.

REFERENCES

[1] Rohan Narayan Koli, “Classification of Speaker’s Age, Gender and Nationality using Transfer Learning ”, 2018

[2] 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), 2021, pp. 338-343, doi: 10.1109/COMPSAC51774.2021.00055.

[3] M. H. Bahari and H. Van Hamme, "Speaker age estimation and gender detection based on supervised Non-Negative Matrix Factorization," 2011 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS), 2011, pp. 1-6, doi: 10.1109/BIOMS.2011.6052385.

[4] Ogechukwu Iloanusi, Ugogbola Ejiogu, Ife-ebube Okoye, Ijeoma Ezika, Samuel Ezichi, Charles Osuagwu, “Voice Recognition and Gender Classification in the Context of Native Languages and Lingua Franca”, 6th Intl. Conference on Soft Computing & Machine Intelligence, 2019.

[5] A. Fidan, R. O. Bircan and S. Karamzadeh, "A New Approach For Age Estimation System Based on Speech Signals," 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2021, pp. 388-393, doi: 10.1109/ISMSIT52890.2021.9604611.

[6] S. Ravishankar, P. Kumar M.K., V. V. Patage, S. Tiwari and S. Goyal, "Prediction of Age from Speech Features Using a Multi-Layer Perceptron Model," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-6, doi: 10.1109/ICCCNT49239.2020.9225390.

[7] R. Zazo, P. Sankar Nidadavolu, N. Chen, J. Gonzalez-Rodriguez and N. Dehak, "Age Estimation in Short Speech Utterances Based on LSTM Recurrent Neural Networks," in IEEE Access, vol. 6, pp. 22524-22530, 2018, doi: 10.1109/ACCESS.2018.2816163.

[8] E. Yücesoy and V. V. Nabiyev, "Gender identification of a speaker using MFCC and GMM," 2013 8th International Conference on Electrical and Electronics Engineering (ELECO), 2013, pp. 626-629, doi: 10.1109/ELECO.2013.6713922.

[9] Mirhassani SM, Zourmand A, Ting HN. Age estimation based on children's voice: a fuzzy-based decision fusion strategy. ScientificWorldJournal. 2014;2014:534064. doi: 10.1155/2014/534064. Epub 2014 Jun 5. PMID: 25006595; PMCID: PMC4070543.

[10] Glamour. “70 Men Ages 5-75: What's Your Biggest Insecurity? | Glamour”, YouTube, Dec. 1, 2020 [Video file]. Available: https://www.youtube.com/watch?v=g1aUj2kLm Bw&list=PL1TPHg7HzcUo1-ewKRu1zCoaDRwPoGxCh&index=2. [Accessed: Oct. 25, 2022].

[11] Glamour. “70 Women Ages 5 to 75: What's the Bravest Thing You've Ever Done? | Glamour”, YouTube, May. 16, 2019 [Video file]. Available: https://www.youtube.com/watch?v=mDhXzzNNPqc&list=PL1TPHg 7HzcUpMHIwoOht6FFHELHN-Z66s&index=40. [Accessed: Oct. 25, 2022].

[12] Glamour. “70 Women Ages 5-75 Answer: What's a Good Life Mean to You? | Glamour”, YouTube, Nov. 5, 2018[Video file], Available: https://www.youtube.com/watch?v=ohXsM7cfSxI&list=PL1TPHg7H zcUpMHIwoOht6FFHELHN-Z66s&index=31[Accessed: Oct. 25, 2022].

[13] Glamour. “70 Women Ages 5-75: What Do You Find Offensive? | Glamour”, YouTube, Nov. 5, 2018[Video file]. Available: https://www.youtube.com/watch?v=ohXsM7cfSxI&list=PL1TPHg7H zcUpMHIwoOht6FFHELHN-Z66s&index=31[Accessed: Oct. 25, 2022].

[14] Glamour. “70 Men Ages 5-75: What Is The Most Difficult Thing You've Ever Done? | Glamour”, YouTube, May. 3, 2021[Video file]. Available: https://www.youtube.com/watch?v=b4jtYKlSBIM&list=PL1TPHg7HzcUo1-ewKRu1zCoaDRwPoGxCh&index=3[Accessed: Oct. 25, 2022].

[15] Glamour. “70 People Ages 5-75: Advice For Someone Younger Than You? | Glamour”, YouTube, March. 15, 2017[Video file].Available: https://www.youtube.com/watch?v=jyLFl-O3L0g[Accessed: Oct. 25, 2022].

[16] Glamour. “70 Men Ages 5-75: What Do You Feel Pressure About? | Glamour”, YouTube, April. 5, 2021[Video File], Available: https://www.youtube.com/watch?v=hAFHWNjjxUs&list=PL1TPHg7HzcUo1-ewKRu1zCoaDRwPoGxCh&index=4[Accessed: Oct. 25, 2022].

[17] ImagineVideoClips. “0 - 100 years in the USA”, YouTube, Sept. 28, 2017 [Video file]. Available: https://www.youtube.com/watch?v=86qBAryeFco[Accessed: Oct. 25, 2022].

[18] scikit-learn.org, “sklearn.tree.DecisionTreeRegressor”, 2011. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html. [Accessed: Nov. 14, 2022].

[19] catboost.ai, “load_model”, 2022. Available: https://catboost.ai/en/docs/concepts/python-reference_catboost_load_model. [Accessed: Nov. 14, 2022].

[20] xgboost.readthedocs.io, “Python API reference”, 2022. [Online]. Available:https://xgboost.readthedocs.io/en/stable/python/python_api.html. [Accessed: Nov. 14, 2022].

[21] scikit-learn.org, “Stochastic Gradient Descent”, 2011. [Online]. Available:https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html. [Accessed: Nov. 14, 2022].