Gender and Age Detection Using MFCC

1st Awsaf Mahmood Lisan
ECE Department, Batch:181
North South University
Dhaka, Bangladesh
awsaf.lisan@northsouth.edu

2rd Rubayet Kabir Tonmoy
ECE Department, Batch:181
North South University
Dhaka, Bangladesh
kabir.tonmoy@northsouth.edu

3nd Fahim Istiak
ECE Department, Batch:181
North South University
Dhaka, Bangladesh
fahim.istiak@northsouth.edu

4th Moshiur Rahman Faisal
ECE Department, Batch:181
North South University
Dhaka, Bangladesh
moshiur.faisal@northsouth.edu

***Abstract*—Gender and age estimation from speech has long been a popular study topic among computer scientists. Estimating gender and age from speech can be a very valuable technique in a variety of situations, including targeted marketing, national security, and surveys, among others. The majority of machine learning work on this subject is based on images of audio spectrogram or specialized handcrafted features like time domain frequency and pitch [4]; however, in this research, we use the MFCC feature extracted from raw audio data to estimate the speaker's gender and age.**

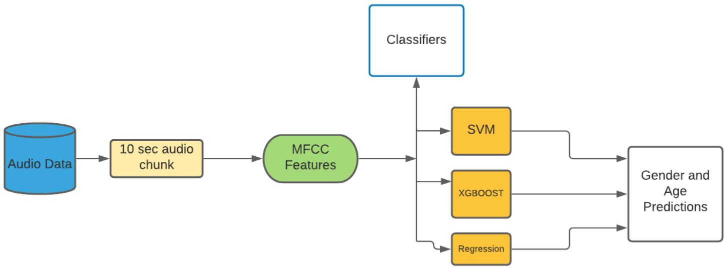
Introduction

Machine learning has become one of the most important methods of problem solving in computer science during the last few decades. Thanks to machine learning, many complicated real-world problems now have feasible answers. Other science domains, such as physics, chemistry, and bioinformatics, have benefited from machine learning as well.

Gender and voice estimation from audio samples is unquestionably a valuable tool that may be applied to a variety of research fields. Spectrogram images from audio samples or other user-crafted features such as frequency band, pitch, formants, and so on are nearly often used to tackle the gender and voice estimation task.

In our experiment, we want to extract MFCC (Mel Frequency Cepstral Coefficient) from raw audio input in wav format and convert it to a 1D vector so that we can predict from pre-existing classifier and regressor models using XGBoost, Support Vectors and so on. As a result, a problem that was previously solved with the help of deep neural networks or complex RGB images may now be solved using pre-existing machine learning models that can be employed by even the most inexperienced machine learning engineers.

Diagram of our process:



Literature Review

For our experiment, we opted for a architecture that we could follow along with existing machine learning models. Audio signal is usually accompanied by background noise, hence extracting useful elements from it is extremely challenging. It is also well knowledge that convolutional neural networks (CNNs) are effective at classifying highly non-linear input. In the field of image processing, they already surpass. As a result, engineers use it to carry out the duty of age and gender classification. Spectrograms are created because spectrogram images are a possible input that we might feed to CNN and automatically compute the appropriate features. [3]

Spectrograms help in analyzing the time-frequency information effectively. The frequency modulation can be observed in the case of spectrograms where it is not possible in the time domain. In general, the frequency domain gives information concerning single-frequency components. Time-frequency distribution (TFD) resolves the issue by providing both time and frequency information. Spectrograms give the information related to the moving sequence of the local spectra to any audio signal. [4]

On the contrary, Mel spectrum is computed by passing the Fourier transformed signal through a set of band-pass filters known as Mel-filter bank. A Mel is a unit of measure based on the human ears perceived frequency. It does not correspond linearly to the physical frequency of the tone, as the human auditory system apparently does not perceive pitch linearly. The Mel scale is approximately a linear frequency spacing below 1 kHz and a logarithmic spacing above 1 kHz [2].

$$f(mel) = 2595 \log_{10}\left(1 + \left(\frac{f}{700}\right)\right)$$

We determined that MFCC features extracted from raw audio data performs better and shows more accurate results compared to data received from spectrogram images. As a result, this experiment will make use of a set number of MFCC's extracted from raw wav formatted audio as our input data for the purpose of both age and gender prediction.

Our Model Setup

A. Dataset

We set out to collect our dataset for audio clips so that we can control variety and amount of data needed for our experiment. Our dataset contains 3617 males and 907 females. At a glance, the dataset is quite imbalanced; which is not a drastic issue for age prediction. However, this imbalance will result in a significant skew with regards to gender prediction. So, to tackle the issue at hand, we opt for evaluation metrics, like precision and recall, besides accuracy.

B. Prediction Models

As input, we make use of 13 MFCC features extracted from the audio files, which are exactly 10 seconds in duration. Audio files less than 10 seconds are attached with a padding. MFCCs from the male voices were labeled as 0 while MFCCs from the female voices were labeled as 1. For gender prediction, we utilized XGBoost and SVC(Support Vector Classifier) as our classifier.

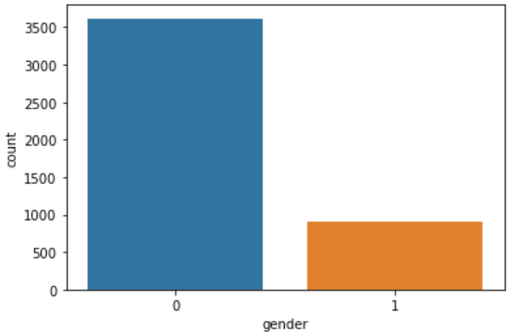
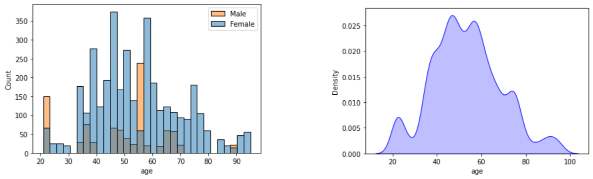


Fig. 1: Gender Distribution

XGBoost uses gradient boosting technique to ensemble decision trees. XGBoost stands for ”eXtreme Gradient Boosting”. Small, medium structured and tabular data uses XGBoost for classification [1]. So, XGBoost seemed perfectly adequate for our relatively small dataset. XGBOOST Decision Tree Image We chose SVR and XGBoost regressors to estimate the speaker's age using MFCC characteristics taken from the speaker's audio clip for age prediction similar to the classifier models. We chose the regression method because it works best with continuous data, and the age range in our sample reflects exactly that. [3] The best fit line for our data plots will be found using regression approach, and the age will be estimated using that best fit line. Image of data scatter plot

C. Data Pipeline

We needed to create a single data pipeline to feed both classifiers from our imbalanced DataFrame because we are using two distinct models to predict age and gender. For the age detection model we made use of the entire dataset for training the model.



(a) Age Distribution

(b) Age Density

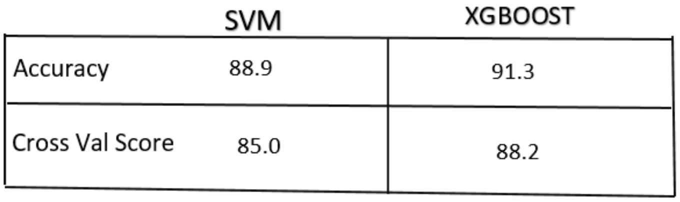
Fig. 2: Data for age predictor model

After training the age prediction model with the entire dataset, we subsequently fit the training data into the XGBoost and SVC classifier.
To simulate a standard use-case, we determine a function utilizing both trained models to predict age and gender from a single audio input, at the same time.

Results

The model with the SVM classifier had an accuracy of 88.9% in gender prediction, whereas the model with the XGBoost classifier had an accuracy of 91.3%.

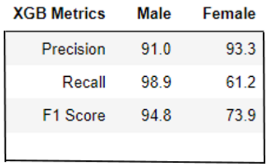
Gender Classification Results



	SVM	XGBOOST
Accuracy	88.9	91.3
Cross Val Score	85.0	88.2

In terms other of metrics for evaluation; the model with XGBoost had a **precision** of 91% for male and 93.3% for female, a **Recall** score of 98.9% for male and 61.2% for female and **F-1** score of 94.8% for male and 73.9% for female.

Metrics from model with XGBoost

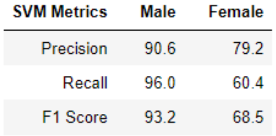


XGB Metrics	Male	Female
Precision	91.0	93.3
Recall	98.9	61.2
F1 Score	94.8	73.9

On the other hand, the model with the SVM classifier had a **precision** of 90.6% for male and 79.2% for female, a **Recall** score of 96.0% for male and 60.4% for female and **F-1** score of 93.2% for male and 68.5% for female.

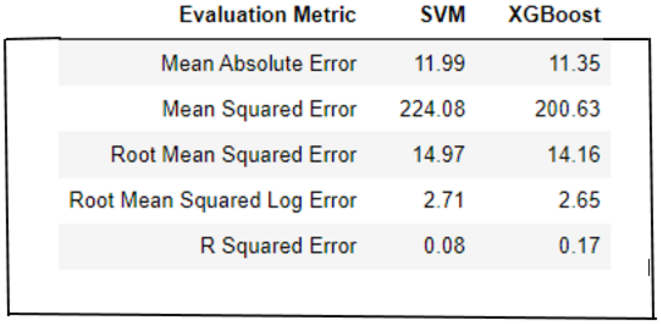
We employed assessment criteria such as mean absolute error, mean squared error, R-squared error, and others to evaluate our age prediction model.

Metrics from model with SVM



SVM Metrics	Male	Female
Precision	90.6	79.2
Recall	96.0	60.4
F1 Score	93.2	68.5

Age prediction model metrics



Evaluation Metric	SVM	XGBoost
Mean Absolute Error	11.99	11.35
Mean Squared Error	224.08	200.63
Root Mean Squared Error	14.97	14.16
Root Mean Squared Log Error	2.71	2.65
R Squared Error	0.08	0.17

Based on these metrics, we can say that our model forecasts age with an inaccuracy of 11-12 years from the speaker's actual age and model made with XGBoost Regressor performs slightly better compared to model made with Support Vector Regressor.

Conclusion

Because of its importance in different applications, recognizing gender and age from the human voice has been regarded as one of the most demanding jobs. We developed a machine learning model that can determine a person's gender and age based on their speech signals only. Our bespoke multi-language audio samples are used to test the gender and age categorization algorithms with a 10-second time limit on each sample statement.

Characteristics from 13 MFCCs were chosen for gender recognition and supplied into the SVM and XGBoost classifiers. The model, with two classes, has the best gender recognition accuracy of 91.3 percent (males and females). For age recognition, the XGBoost and SVR Regressor models were tested, with XGBoost achieving a better mean absolute error of 11.35 years.

Finally, we correctly predict age and gender from a human voice with the aforementioned accuracy and error metrics, combining the best Classifier and Regression models.

References

[1] Jie-Min Long, Zhang-Fa Yan, Yu-Lin Shen, Wei-Jun Liu, and Qing-Yang Wei. Detection of epilepsy using mfcc-based feature and xgboost. In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–4. IEEE, 2018.

[2] YV Srinivasa Murthy, Shashidhar G Koolagudi, and TK Jeshventh Raja. Singer identification for indian singers using convolutional neural networks. *International Journal of Speech Technology*, pages 1–16, 2021.

[3] Volodymyr Osadchyy, Ruslan V Skuratovskii, and Aled Williams. Analysis of the mel scale features using classification of big data and speech signals. *International Journal of Applied Mathematics, Computational Science and Systems Engineering*, 2, 2020.

[4] Dr. Radhika S.N.V. Jitendra1. Singer gender classification using feature-based and spectrograms with deep convolutional neural network. *International Journal of Advanced Computer Science and Applicationsy*, pages 1–10, 2021.