Voice based Age, Accent and Gender Recognition

Deepa Angadi[1], Manoj K R[2], Nagendra N S[3], Nithin Kumar B[4]
[1]*Assistant Professor, Department of Computer Science and Engineering, MVJ College of Engineering, Bengaluru.*
[2,3,4]*UG Student, Department of Computer Science Engineering, MVJ College of Engineering, Bengaluru*

Abstract **- Gender, Age and Accent recognition that is solely achieved by human speech which is an interesting subject in the field of Automatic Speech Recognition systems. In this project, we have introduced a way of classifying a human's speech into different classes of Gender, Age and Accent. The classification used in this project is mainly based on Machine Learning Models. The Models are systematically trained on the audio data that is obtained from Mozilla Common Voice. The dataset only contained the audio files with the respective labels associated with it. The noisy data from the dataset have been filtered out before it is fed into the Machine Learning models. In our project, we are using Mel Frequency Cepstral Coefficients (MFCC). We are extending Accent classification into American, European and Indian. But, in our project "Voice based Age, Accent and Gender Recognition" we have extended our work to the classification of speech into Gender, Age and Accent.**

Index Terms **- Random Forest, K Nearest Neighbors, XGBoost, MFCC features.**

I.INTRODUCTION

Identification of human traits is an important aspect in voice processing that can be employed in different signal processing fields. Humans are very good at recognizing people, besides visual cues, they also heavily rely on auditory cues. Pretty quickly we can assess from speech alone a person's gender, age and accent. Classifying speaker characteristics is an important task in Dialog Systems, Speech Synthesis, Forensics, Language Learning, Assessment Systems, and Speaker Recognition Systems. By adapting the Speech Translation system to the speaker's gender, age and accent, the overall recognition accuracy increases dramatically. In moments such as threatening calls from an unknown person asking for the bank details pretending to be one of the bank officers, electronic banking, criminal cases. Foreign accent identification in English was the subject of the Interspeech Computational paralinguistic challenge. Speaker identification technology can help rapidly identify suspects' voices and isolate conversations of interest in a wide range of law enforcement cases. "Voice based Gender, Age and Accent recognition" can be used in the law enforcement community to identify the voices of unknown individuals into categories of age, accent and gender.

For a robust and low latency predicting system, the input signal is transferred into a feature domain to simplify the signal analysis while maintaining the features. The proposed classification algorithm includes training steps to provide the appropriate trained atoms/features for each data class. The input signals stored in the form of the wav file are then processed and the MFCC (Mel-Frequency Cepstral Coefficients) are employed to train the basis of the Machine learning models. The speech database chosen to accomplish the task is Mozilla Common Voice. Mozilla Common Voice is a crowdsourcing project started by Mozilla to create a free database for speech recognition software. The project is supported by volunteers who record sample sentences with a microphone and review recordings of other users.

The speech audio files will be 'compiled' into acoustic models for use with open-source speech recognition engines. The Mozilla Common Voice dataset that we have used contains around 3,80,368 audio samples with each sampled at 48KHz with 16-bit PCM encoding. The dataset is split into 3 sets for training, cross-validation and testing. In the training phase, the Machine Learning model is trained on the extracted features. The audio samples in the training dataset are from various countries with various accents as specified above. The MFCC features are extracted from each audio sample in the dataset. It is the representation of the short-term power spectrum of a sound. Acoustic features are the characteristics of a speech sound that helps in distinctive feature analysis.

Some of the acoustic features being Kurtosis, Standard deviation, IQR, peak frequency etc.

II. RELATED WORK

A. The use of long-term features for GMM- and i-vector-based speaker diarization systems

This section discusses the various aspects that contribute to the performance and key reasons that impede the process of partitioning an input audio stream into homogeneous segments according to the speaker identity in diarization systems. Numerous features have been used for the Gaussian mixture modeling and i-vector-based diarization systems. It is observed that the usage of the i-vector based and cosine-distance clustering with the signal parameterization which involves static cepstral coefficients, delta and prosodic features have their accuracy improved cutting down the error rate of the diarization. In the mentioned reference paper, it is noticed that the optimal outcome is about 24% relative diarization error rate enhancement when collated with the baseline system which is based on Gaussian mixture modeling and short-term static cepstral coefficients.

B. Deep Learning Model based Mandarin Accent Identification for Accent Robust ASR

The strategies used in this paper for the Mandarin Accent Identification are i-vectors, DNN and bLSTM accent classifiers. i-vectors v are obtained from the reconciled Gaussian Mixture Models (GMMs) with mean supervector and GMM universal background models (UBM) with mean supervector. Analogous to the GMMs that are trained and the features based on the Linear Discriminant Analysis (LDA) were gained from windows of voice inputs. DNN and bLSTM Accent Classifiers used to train deep learning accent identification includes 45 Mel-frequency Cepstral Coefficients (MFCCs) and 7 fundamental frequency variation (FFV) features obtained at a rate of 10 milliseconds and with a window size of 25 milliseconds. The structure contains two bidirectional Long Short Term Memory layers of size 512, each having 256 Long Short Term Memory units for the forward and backward directions, in combination with the softmax output layer.

C. Gender Recognition by Voice Using an Improved Self-Labeled Algorithm

In this reference paper that is solely based on the self labeled algorithm, the main objective of gender recognition by speech is performed upon the usage of a new ensemble semi-supervised self-labeled algorithm. The steps in finding the necessary labeled data for training classifiers to make the gender classification result in high performance is often expensive in terms of time and cpu resources taken for the task as it requires human efforts, while in contradiction it is said to be easy in finding unlabeled data in general. To overcome the problems of inadequate labeled data, semi-supervised learning (SSL) algorithms for gender recognition give the appropriate techniques to exploit new useful knowledge patterns in the unlabeled dataset resulting in better and more reliable classifiers in the process of gender recognition.

In the above mentioned work for gender recognition, it is seen that a new ensemble-based self-labeled algorithm, called iCST-Voting has been used. The algorithm combines the individual predictions of three of the most known and consistent self-labeled methods i.e, Co-training, Self-training, and Tri-training making good use of the ensemble as base learner for the algorithm.

III. PROJECT ARCHITECTURE

The project architecture gives you the visual representation of the model structure built for the classification of Age, Accent and Gender based on the input voice speech.

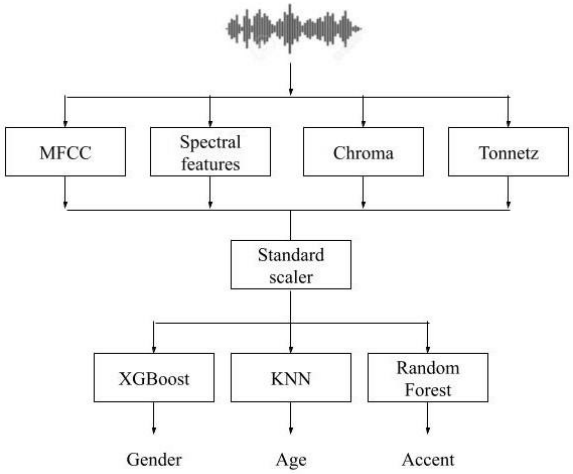

Fig 1. Proposed system

The above architectural diagram shows the object interaction and the flow of the project. The components shown in the project architecture contains, user interacting with the system, the MFCC features, the models used for the Age/Accent/Gender classification. As shown in the above diagram and in Section III describing the models, we have used XGBoost for the Gender classification, KNN model for the Age classification and Random Forest for the Accent classification.

The dataset is divided into 80:10:10 where 80% of data points are used for the training purpose of the model, 10% as the test data points and final 10% for the cross validation of classification models. During the preprocessing of the dataset, the outliers are removed by normalizing the dataset and all the NaN values are removed before it is used to train the classifiers. The dataset is scaled using sklearn's Standard Scaler which helps to get standardized distribution, with a zero mean and standard deviation of one (unit variance). It standardizes features by subtracting the mean value from the feature and then dividing the result by feature standard deviation

IV. PROPOSED SYSTEM

A. Gender Classifier using XGBoost model

In this paper, for the gender classification we have used XGBoost for identifying whether the input speech is Male/Female. XGBoost is a decision tree based on the ensemble technique which provides a regularizing gradient boosting framework provides scalable, portable and distributed boosting libraries. Gender classification is best done using XGBoost because of the model optimization reasons which involves tree pruning using depth-first approach for the dataset which contains numerous MFCC features that are fed into the model. It is very much efficient when building a model around the dataset which contains missing data and uses regularization to avoid overfitting.

The dataset for the Gender classification is taken from the Common Voice Corpus accumulated from a number of public domain sources where the text is read by the users. The dataset contains 55,000 Male and 18,000 Female speech samples which is obtained from over 300 hours of speech recordings.

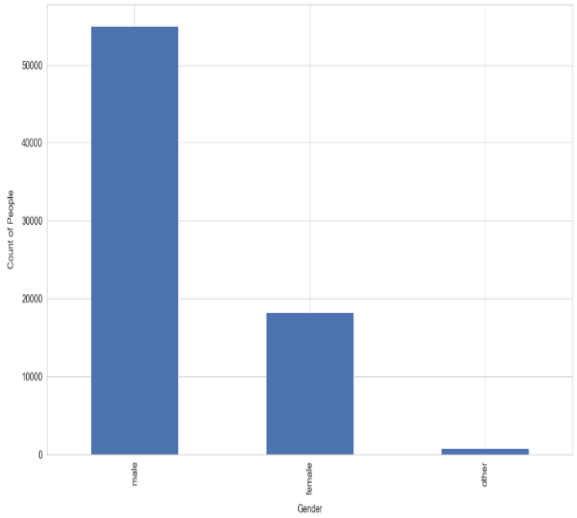

Fig 2. Gender distribution in the dataset

The dataset is filtered and the XGBoost classifier is only trained on the Male and Female data points. The extracted features from the audio samples of Male/Female are MFCC, Mel spectrum Frequency, Chroma and Contrast. Therefore, the total rows of the aforementioned audio features is 73,000(includes Male and Female) audios. The dataset is scaled using sklearn's StandardScaler.

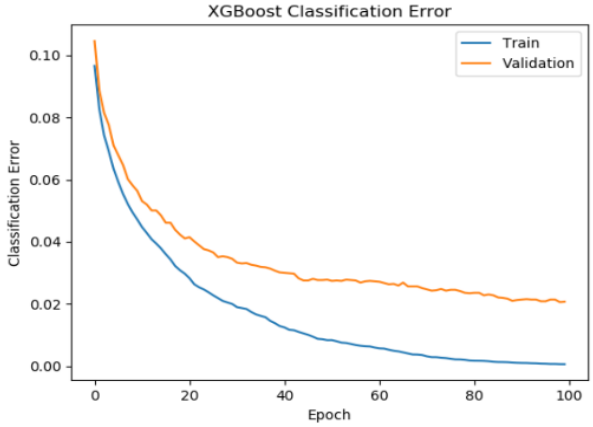

Fig 3. XGBClassifier error rate

The XGBoost classifier is then assessed on the test dataset with the default parameters of the model. From the above fig 2, we can conclude that the model doesn't overfit for the preprocessed dataset. The error rate decreases gradually for both training and validation set as the number of iterations or epochs increases. The metrics that are used to evaluate the model are Accuracy, Balanced Accuracy, F1 score and ROC AUC.

Metric	Train set	Validation set	Test set
Accuracy (%)	99.9	97.9	97.1
Balanced Accuracy (%)	99.9	97.0	95.8
F1 Score (%)	99.9	98.6	98.0
ROC AUC (%)	99.9	97.0	95.8

Fig 4. Performance metric of Gender classification

B. Age classification using K-Nearest Neighbors

In our paper, for age classification we have used K-Nearest Neighbors to classify the input speech audio into Youth/Adult/Senior. The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. KNN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically.

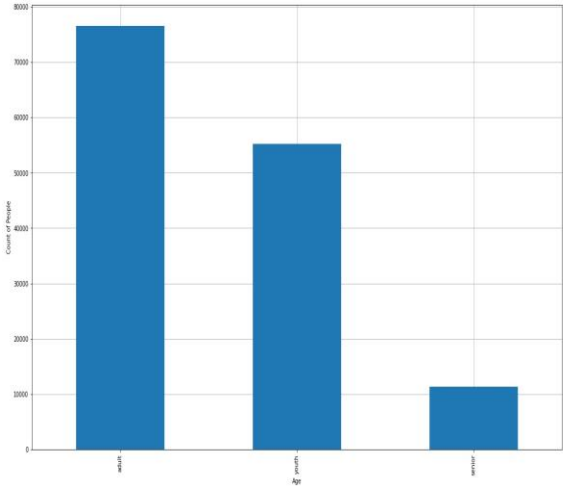

Fig 5. Age distribution in the dataset

The dataset used for Age classification is the same as the dataset that is used for the Gender classification which is taken from the Common Voice Corpus. It is obtained from recording where the text is read by the human. The dataset contains 76,567 Adult, 55,226 Youth and 11,377 Senior audio samples after the preprocessing of the dataset. The features used for the Age classification are 20 MFCC(Mel frequency cepstral coefficients), Spectral centroid, Spectral bandwidth, Spectral rolloff along with gender.

The parameters of the KNN model used are as follows, the number of neighbors that is the value of 'N' in the K-Nearest Neighbors is set to 5 and the value of 'p', the power parameter for minkowski metric is assigned 1 which is equivalent to the Manhattan distance(l1). The weight function used in the prediction is based on the 'distance', where the weight points are the inverse of their distance i.e., closer neighbors of a query point will have a greater influence than neighbors which are further away.

Metric	Train set	Validation set	Test set
Accuracy (%)	99.9	93.6	93.3
F1 Score (%)	100.0	94.0	93.0
ROC AUC (%)	99.9	98.3	98.4

Fig 6. Performance metric of Age classification

B. Accent classification using Random Forest model

In our paper, for accent classification we have used Random Forest to classify the input speech audio into American/European/Indian. The Random Forest algorithm is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

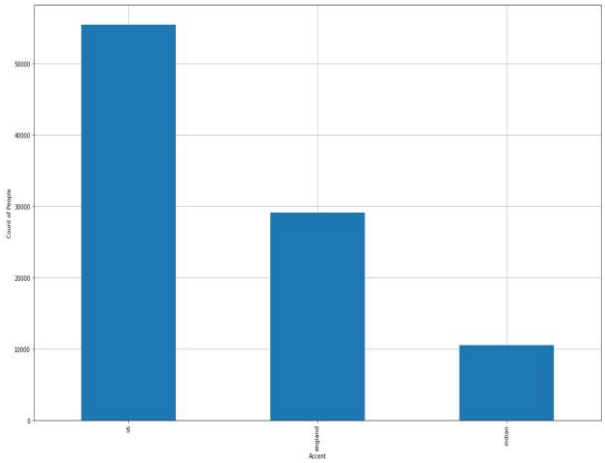

Fig 7. Accent distribution in the dataset

The dataset used for Accent classification is taken from the Common Voice Corpus. The dataset contains 55,501 American, 29,113 European and 10,521 Indian audio samples after the pre-processing of the dataset. During pre-processing, the dataset was balanced by undersampling majority classes before it is used to train the Random forest classifier. The features used for the Accent classification are 20 MFCC(Mel frequency cepstral coefficients), Spectral centroid, Spectral bandwidth, Spectral rolloff. The parameters of the Random Forest model used are default parameters.

Metric	Train set	Validation set	Test set
Accuracy (%)	97.0	97.0	96.0
F1 Score (%)	97.0	97.0	97.0
ROC AUC (%)	99.8	99.8	99.8

Fig 8. Performance metric of Accent classification

V. CONCLUSION

This work has proposed the use of acoustics features (MFCC, Spectral centroid, Spectral bandwidth, Spectral rolloff, Mel spectrum Frequency, Chroma and Contrast) to predict Age, Accent and Gender of human beings using their voice input. The work has also analyzed the use of above-mentioned speech features with different Machine Learning models to provide the prediction with less latency.

The experimental results show that training the Machine Learning models like XGBoost, KNN and Random Forest model with the mentioned features give a similar accuracy rate as compared to the GMM-UBM model and Deep learning models. The performance of the XGBoost for Gender classification is near perfect for the training set indicating no signs of overfitting and the error rate is very less on the test set because of the excellent training process. The KNN model for the Age classification gives accurate prediction because of the distinction and separation in the clusters resulting from the preprocessing of the dataset. Random Forest for Accent classification is impressive for the training set indicating no signs of overfitting and the error rate is very less on the test set because of the excellent training process.

REFERENCES

[1] Abraham Woubie Zewoudie, Jordi Luque, Javier Hernando. The use of long-term features for GMM- and i-vector-based speaker diarization systems. EURASIP Journal on Audio, Speech, and Music Processing, 2018.

[2] Felix Weninger, Yang Sun, Junho Park, Daniel Willett, Puming Zhan. Deep Learning based Mandarin accent identification for accent robust ASR. INTERSPEECH, 2019.

[3] T K Harshitha Devang, Sushma M Tilave, Sowmya M.S., Tejaswini. Accent Identification. International Journal of Engineering Research & Technology, Vol. 6, No. 13, 2018.

[4] Ioannis E. Livieris, Emmanuel Pintelas, Panagiotis Pintelas. Gender Recognition by Voice Using an Improved Self-Labeled Algorithm. Machine Learning and Knowledge Extraction, pp. 492-53, 2019

[5] S. B. Magre, P. V. Janse, G. Kollios, R. R. Deshmukh. A Review on Feature Extraction and Noise Reduction Technique. International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, No. 2, pp. 352-356, 2014

[6] Maryam Najafian, Martin Russell. Automatic accent identification as an analytical tool for accent robust automatic speech recognition. Speech Communication 122 – Elsevier, pp. 44-56, 2020.

[7] S. Mavaddati. Voice-based Age and Gender Recognition based on learning Generative Sparse Models. International Journal of Engineering, Vol. 31, No. 9, pp. 1529-1535, 2018.

[8] Noor Salwani Ibrahima, Dzati Athiar Ramlia. I-vector Extraction for Speaker Recognition Based on Dimensionality Reduction. Procedia Computer Science 126, pp. 1534–1540, 2018.

[9] Osman Büyük and Levent Arslan. An Investigation of Multi-Language Age Classification from Voice. 12th International Joint Conference on Biomedical Engineering and Technologies: BIOSIGNALS, Vol. 4, pp. 85-92, 2019

[10] Saeid Safavi, Martin J Russell, Peter Jancovic. Automatic Speaker, Age-group and Gender Identification from Children's Speech. Computer Speech & Language 50 – Elsevier, pp. 141- 156, 2018.