University of Essex
Department of Mathematical Sciences

MA981: DISSERTATION

Automatic Age Recognition From Speech Using Machine Learning

Rifat Monzur
2200367

Supervisor: Dr Vasileios Giagos

March 25, 2024
Colchester

Abstract

As people age, various physical changes become quite evident. One of such change is voice. Only hearing voice for few seconds, we can generally recognise if the person is young, adult or old. We can also detect gender by hearing someone. There is scientific evidence that our voice changes as we age and also the effect of gender in our voice. There are many systems build to automatically recognise human age and some even done better job than human. We have also tried three different approaches on common voice database where age is divided into 8 classes. Firstly, we tried six traditional machine learning algorithms namely K nearest neighbor, Support vector machine, Random forest, Gradient boost, Decision tree and Gaussian naive bias with MFCCs as feature extractor. Among those, K nearest neighbor and support vector machine performed best with 0.89 and 0.88 f1 macro score respectively. We also experiment those algorithms with different train data size and providing gender as feature. We have seen positive influence of gender as feature and performance also got better with bigger size of database. Our second approach was to combine the learning of traditional models to make ensemble models. Results of ensemble learning was mixed to positive. It always performed better than support vector machine and many cases it performed better than k nearest neighbor. However, there are no specific combination of models that we found that always performed better for different data size. Overall, with ensemble learning we were able to achieve 0.90 f1 score for whole dataset. Our third approach was transfer learning based approach with LLMs, wav2vec2 as feature extractor and Hubert. Due to time constraint and absence of powerful GPU, we ultimately failed to materialize this approach. In future, it would be great extension of this research to explore the potential of LLMs in automatic age recognition.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Some say 'Age is just a number'. But when it comes to voice and human physical changes, it is not just a number. With age our various physical aspect of our body change so does our voice. As we grow old, our vocal cord become thinner. On the other hand our larynx's cartilage more rigid and less flexible. As a result, we see change in our voice pattern. We also see different effect on voice for different gender. Men experience increase in pitch as they age. On the other hand, women see decrease in pitch. Older people voice become frail or bit shaky due to changes in vocal muscles of laryngeal. Due to age, changes in other body parts like curvature of the spine or decrease in lungs capacity also have it's impact on voice [11].

1.1 Can human recognise age from voice?

From various research and observation, we can also observe that human can easily distinguish between younger and older adults. Within few seconds of listening voice, one person can easily determine if it is from young or old person. They tested on various types of speech and in all types of speech, they were at least 80 percent correct. For straight forward speech they almost 100 percent correct [26].

In another research, they made more detailed approach to guess the age by participant. They recorded the audio in more restricted and consistent environment. They divided the age group from 20-29, 30-39, 40-49, 50-59, 60-69, 70-79 and 80-89. Mean age of the subject is generally in the middle to keep the environment fair. For example, for 20-29 group, the mean

age is 24.4, for 30-39 group, the mean age is 34.9 and so on. They made 3 point system. 1 point for young, 2 points for not young or old and 3 points for old. So, the mean score for every age group will be between 1.0 to 3.0 [30].

Age Group	20-29	30-39	40-49	50-59	60-69	70-79	80-89
Mean Point	1.2	1.5	1.9	2.1	2.3	2.7	2.9

Table 1.1: Chronologic Age in Decades [30].

From the 1.1 table, we can see gradual increase over chronological age group. It is always strictly increasing over to next age group. For, 20-29 it is 1.2 and for 80-89 it is 2.9. So, we can observe that listener can distinguish age from voice.

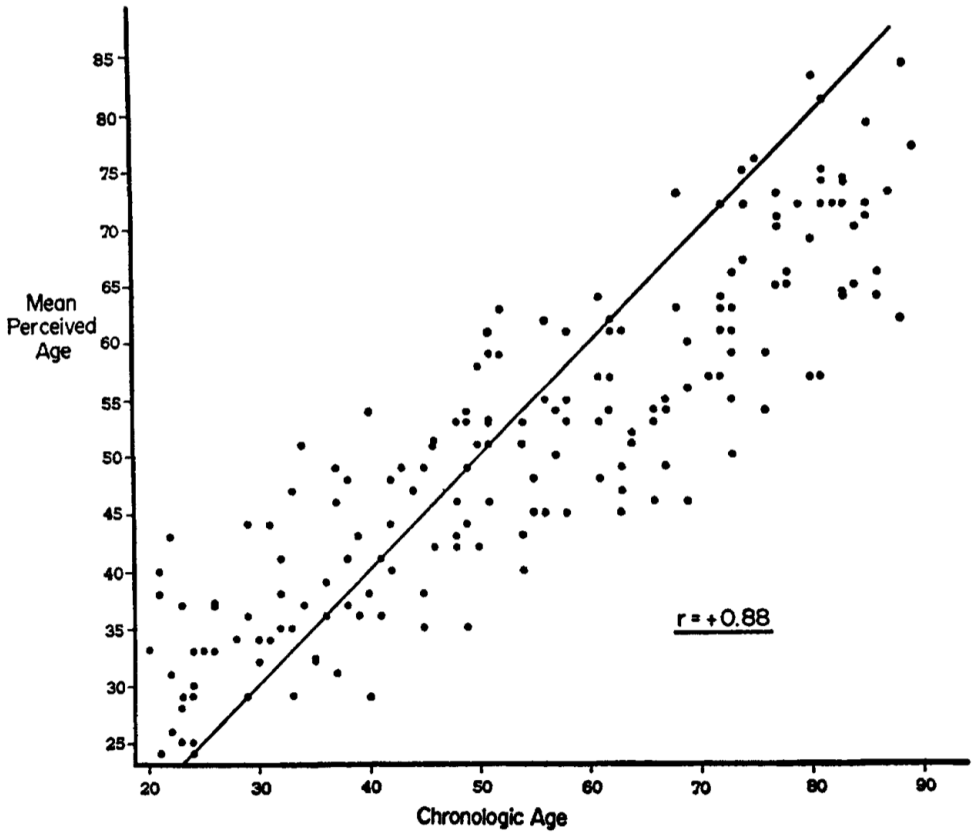

Figure 1.1: Scatter plot of the mean perceived age versus chronologic age of 175 speakers. Diagonal line is +1.0 correlation [30].

We can see similar pattern from the above scatter plot. Listener guess the age quite closer to actual age. Analyzing this illustration further reveals that, in comparison to the linear

function with a correlation of +1.0, the observers displayed a tendency to inaccurately assess the ages of younger speakers as older and older speakers as younger. This phenomenon aligns with a predictable regression effect [30].

1.2 Why voice change with age?

As it becomes clear that voice change with age, let's figure out why it happens. But, before that we need to understand few terms like vocal pitch, vocal cord, thyroid cartilage and larynx and what they do.

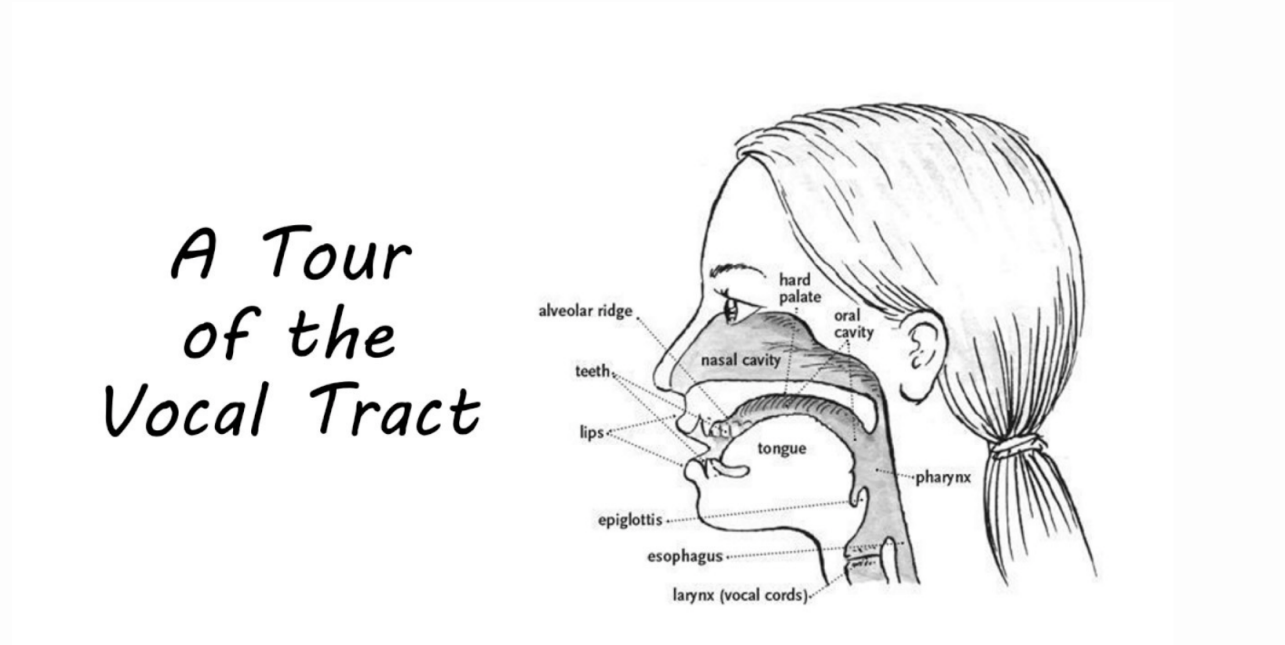

Figure 1.2: Human vocal tract [10].

Vocal pitch is one of the fundamental aspect of human voice. It means the lowness or highness of the voice. When air flow through vocal cords from lungs, the amount of vibrations defines if it is high pitch or low pitch. If the vibration is low then it is lower pitch. On the other hand, if the vibration is high, it is high pitch. Vocal pitch is depend on various factors like size and tension of vocal cords, lungs capacity, vocal tract and indirectly through human age.

Vocal cords is pair of muscle tissues located in larynx. It plays crucial role in producing human voice. Vocal cords vibrate rapidly when air pass through it from lungs. As a result sound waves are created and it is determined how vocal pitch will be. Then, sound waves

are shaped by tongue, lips and mouth to create speech. Individual can adjust tension and length of vocal cords to produce meaningful sounds.

Larynx also know as voice box is located in respiratory system and essential for human voice generation. Larynx contains vocal cords and is important for producing sound, breathing and airway protection. Larynx control vocal cord's length and tension to produce various sound wave. As a result, it also control voice pitch and volume. The larynx is an intricate structure that involves the coordination of muscles, cartilage, and other tissues. Changes in the larynx's size, shape, and function can result in variations in voice quality, pitch, and other vocal characteristics.

Thyroid cartilage is crucial part human voice box or larynx. In male, it's also called Adam's apple due to it's prominence. Hyoid bone is located just above thyroid cartilage and is horse shoe shaped. It is unique bone in human body which doesn't articulate with any other bone. Rather it is help in by complex structure of muscles, ligaments, and tendons [28].

An individual's vocal pitch primarily relies on the degree of contraction in the vocal cord muscles when they come into contact with the airflow from the lungs. In the case of a child, their voice is characterized by a high pitch due to the small size of the larynx (voice box) and the short, thin, and taut vocal cords. As an individual progresses through puberty, the larynx undergoes growth, and the vocal cords elongate and become thicker, resulting in a lower and deeper voice [11].

Larynx goes through a lot of over ther years. Larynx started to develop at third month of fetus. At birth, larynx is located high in the neck and gradually descent through out our life for both sexes. As larnyx descent so the vocal pitch become lower. Hyoid bone started to ossify around age of 2. At birth thyroid cartilage is about 110° and 120° respectively male and female. Thyroid cartilage remain in the same spot until puberty. That's why male and female voice sounds similar before puberty. In vocal fold, membranous and cartilaginous portion remain equal at infancy. However, by adulthood membranous becomes 60% of the total vocal folds. Whole vocal cord length increase as well. At infancy, it 6 to 8mm in length. But at increases to 12 to 17mm and 17 to 23mm for female and male respectively. We can see similar pattern for all the other parts of larynx [28].

At birth, child make noise of 500 Hz. As child grow, frequency of speech drop gradually. Around age of 8, it drops to 275 Hz. Until puberty, both sexes frequency remains similar. After puberty male vocal cord grow around 60%. On the other hand, female vocal cord grow

around 34%. During puberty angle of thyroid cartilage decrease to 90° while female angle remain same at 120°. After puberty female voice frequency drops at 220 to 225Hz and male drops at 120 to 130Hz [28].

Menopause also have effect on female voice. According to study 29% female experience change in voice around 50 years old compare to male 38%. Contraceptive pills also affect female hormones. As a result of that it affects female voice as well. [15]

As now it is clear that voice with age and why it happens. Human could easily recognise person's age group as we discussed in 1.1. Now, we want to know if it is possible for automatically detect in which age group that voice belongs using machine learning.

CHAPTER 2

Literature Review

Firstly, we are going to explore available researches which have worked on similar tasks with similar or different dataset. Then, we will also explore researches which are related to speech fields and have potential for better output in future at age classification from voice.

2.1 Similar Research

In 2007, age recognition was performed over 7 age classes [24]. Among the 7 classes, children below 14 were considered same class regardless of their gender as it is hard to distinguish difference between them. Other age groups were young (14-19 years), adult (19-64 years) and old (above 64 years) with both sexes have different groups. They used German SpeechDat II corpus , which contains telephone recordings of 4000 native speakers who read out specific set of words and digits. They tried four different approaches and among them Parallel Phoneme Recognizers (PPR) using Continuous Densities Hidden Markov Models (CDHMMs) performed the best with Mel Frequency Cepstral Coef- ficients (MFCCs) feature extraction. With 55% accuracy it performed similar to human efficiency.

In 2008, the above research was extended with adding another database [13]. There were two sets of data set. One is German SpeechDat II corpus [24], which contains telephone recordings of 4000 native speakers who read out specific set of words and digits. Data set was age class balanced. Another data set was VoiceClass corpus where speaker talk about their favourite fish dish. However, this data set was age class imbalanced and did not have

specific set of words or digits. Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) were utilized to predict age class. SVM performed better on age balanced data set than both GMM and Human. However, GMM approach performed slightly better on age imbalanced data set.

Another research was done on Kurdish language with similar type of 7 age class system with only difference is old and adult divided by 56 year instead of 64 year [19]. They got the best result using K-nearest neighbors classifier with 75% accuracy with MFCCs feature extraction technique. However, with help of wavelet based denoising algorithm, they got improved accuracy of 81%.

In 2021, a research [33] conducted to detect age, gender and emotion from audio. They used Frequency Spectrum Analysis (FSA) [20] to extract 20 statistical features which are mean frequency , standard deviation of frequency, median frequency, first quartile, third quartile, interquartile range, skewness, kurtosis, spectral entropy, spectral flatness or tonality coefficient, mode frequency, spectral centroid, average fundamental frequency, minimum fundamental frequency, maximum fundamental frequency, average dominant frequency, minimum dominant frequency, maximum dominant frequency, range of dominant frequency, and modulation index. They used 10 machine learning technique which are Random Forest, CatBoost, Gradient Boosting, K-nearest neighbors, XGBoost, AdaBoost, Decision Tree, Artificial neural networks (ANN), Naive Bayes, and Support vector machine. They also used 10fold cross validation. Their dataset have 9 age groups. Among all the technique Random forest and Catboost performed the best with accuracy and f1 macro score.

2.2 Related research with potential

Recently, self supervised learning is becoming more and more popular in Natural Language Processing (NLP) and computer vision fields. One of the main reason, self supervised learning is becoming popular is that it takes very little labeled data to get good accuracy. In many cases, good amount of labeled data is hard to get. As labelling data could be both costly and time consuming. In some cases, it could be even impossible to generate the labelled data necessary to produce good result. That is where self supervised learning coming in handy.

Another reason for self supervised learning to become popular is that you do not have to select feature manually. For example, MFCCs is very popular feature selector what we can

understand from 2.2 research. While MFCCs and other low-level descriptors function rather well, detractors note that these features are frequently very straightforward and can lose a lot of crucial information from the original signal. To solve this issue, we need feed the model raw audio data so it can find the best feature itself. Transfer learning exactly does this job. It finds the best feature from raw data without any human bias[14].

We can see a lots of use of transfer learning in NLP and computer vision. In 2019, Bert revolutionized NLP field and got state of art results in many nlp tasks[32]. We can say same about imagenet for computer vision[22]. What is more interesting is that all the current leading models which are performing state of art in both fields are pretrained transfer learning models.

In speech field, we started to see progress and influence of transfer learning. Wav2Vec[29] and Hubert[21] are example of transfer learning models for speech. Wav2Vec was pre trained in 960 hours of audio in unsupevised way. It was trained to solve automatic speech recognition (ASR) problem. It was evaluated in word error rate(WER). It performed better than any other system available at the time with 2.43%. In 2020, Facebook AI released wav2vec 2.0, which performed even better than wav2vec[12]. Next hidden unit Bert was released which make further improvent over state of art wav2vec 2.0[21].

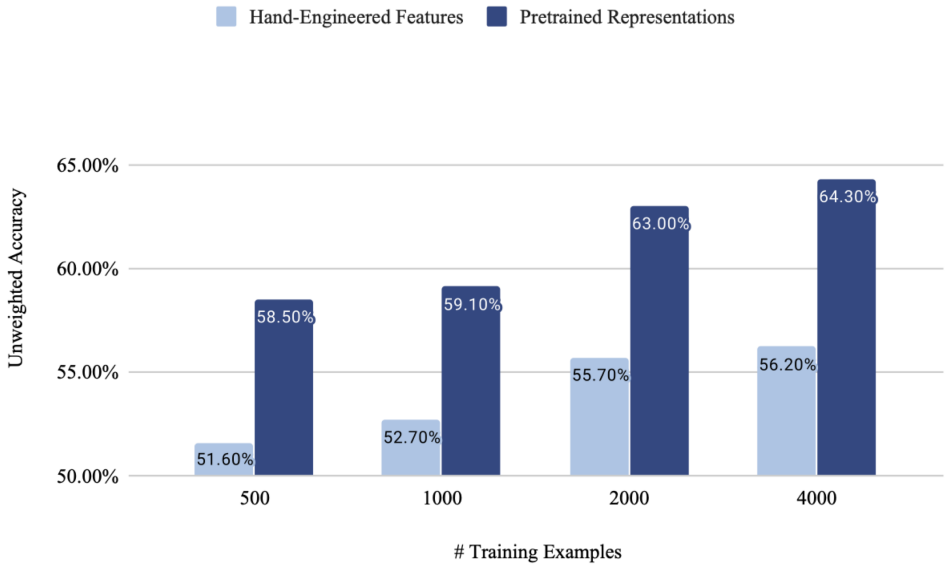

Figure 2.1: Evolution of performances for our best model in function of the amount of training data [14].

From figure 2.1, we can see transfer learning outperformed every other system on ASR task. Recently, we are seeing transferwise learning technique is also being used in speech classification problems. We can see in speech emotion recognition, transfer learning technique outperformed all other state of art models[14]. From figure 2.1, we can see pretrained representation performed better for small to large training samples. The difference between accuracy is quite large. For 500 training samples, it is 7% and for 4000 training samples, it is almost 8%. So, we can observe improvement in accuracy in both small and very small samples.

To conclude, we have seen various traditional machine learning techniques with good accuracy for age recognition from voice. We have also seen various pretrained models like hubert and wav2vec2, achieved state of art performance both in ASR and emotion recognition. So, there is potential for those models to perform well on age recognition from voice as well.

CHAPTER 3

Dataset

3.1 General Information

Common voice is a crowd sourced dataset of speech data[7]. Anyone can submit their speech and contribute to the corpus. It was primarily created for automatic speech recognition (ASR) but it is encourage to use for other purposes as well.

We are working with small portion of common voice dataset with only english speech[4]. Main corpus size is more than 100gb. However, the small portion we are going to work with is about 13gb. The corpus got license of creative common public domain.

3.2 Validation System

To validate user's voice input, a voting system is maintained. According to the voting dataset is divided into three parts: valid, invalid and others [4].

Valid: If an audio clips receive at least 2 votes from listeners and majority of them accept that audio matches with the transcription then it is valid clip

Invalid: If an audio clips receive at least 2 votes from listeners and majority of them accept that audio does not match with the transcription then it is invalid clip

Others: If audio clip receive less than 2 votes or no majority decision reached by the listeners or in another word valid and invalid votes are equal then that clip belongs to others.

3.3 Organization and Conventions

Valid, invalid and others are again divided into three parts each which are test, dev and train. All those three parts of data, contains a csv file each which contains the following columns [4].

filename: relative path of the audio file

text: proposed transcription of the speech

up_votes: number of listener who said audio matches the transcription

down_votes: number of listener who said audio does not match transcription

age: age group of the speaker if reported. There 9 age groups.

teens: below 19 years

twenties: from 19 year to 29 year

thirties: from 30 year to 39 year

fourties: from 40 year to 49 year

fifties: from 50 year to 59 year

sixties: from 60 year to 69 year

seventies: from 70 year to 79 year

eighties: from 80 year to 89 year

nineties: above 89 year

gender: gender of the speaker if reported. There are 3 gender groups: male, female and others.

accent: accent of the speaker if reported.

us: 'United States English'

australia: 'Australian English'

england: 'England English'

canada: 'Canadian English'

philippines: 'Filipino'

hongkong: 'Hong Kong English'

indian: 'India and South Asia (India, Pakistan, Sri Lanka)'

ireland: 'Irish English'

malaysia: 'Malaysian English'

newzealand: 'New Zealand English'

scotland: 'Scottish English'

singapore: 'Singaporean English'

southatlandtic: 'South Atlantic (Falkland Islands, Saint Helena)'

african: 'Southern African (South Africa, Zimbabwe, Namibia)'

wales: 'Welsh English'

bermuda: 'West Indies and Bermuda (Bahamas, Bermuda, Jamaica, Trinidad)'

CHAPTER 4

Experimental Data Analysis

Though there are three sets of data, we are planning to work only on valid dataset. As valid dataset is validated by listeners and also working with whole 13gb audio data would not be possible considering time constraint and processing power we have. Let's focus only on valid dataset for now. Valid dataset have three subsets called train, test and dev. Dev dataset is for development and experimentation [4]. We need to make sure if dev dataset is similar to train dataset. We need to have similar data ratio and same categorical values. We also need make sure that, train dataset has all the same categorical values as the test dataset.

4.1 Train Dataset

4.1.1 General Information

There are 195776 rows of data in valid train dataset. There are 8 columns which are filename, text, up_votes, down_votes, age, gender, duration and accent. We will not focus on up_votes and down_votes as they are used for data validation purpose. They are not suitable for our task. Our main focus will be on age, gender, duration and accent as they are more related to our task.

Column	Available	Missing
filename	195776	0
text	195776	0
up_votes	195776	0
down_votes	195776	0
age	73768	122008
gender	74059	121717
accent	64711	131065
duration	0	195776

Table 4.1: Train dataset statistics of missing and available values

4.1.2 Missing Values

From 4.1, we can see filename, text, up_votes and down_votes have all the values available. It should not come as surprise as it is mentioned 3.3 that those values are mandatory value. However, we can see missing values in age, gender, accent and duration as they are optional value as mentioned in 3.3. We can see observe that data is maintained it's convention. One thing that could come as little surprising is that the duration column as there is not a single value available. Let's take a deeper drive into the missing values of age, accent and gender.

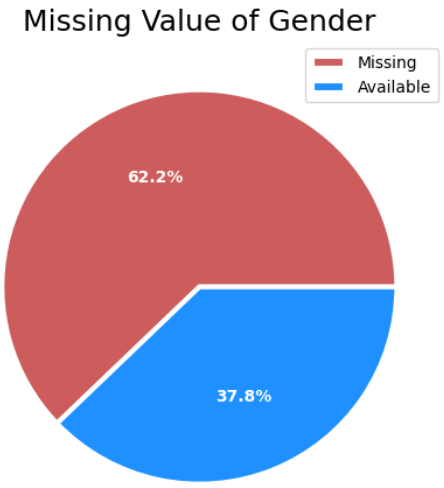

Figure 4.1: Missing value percentage of Gender

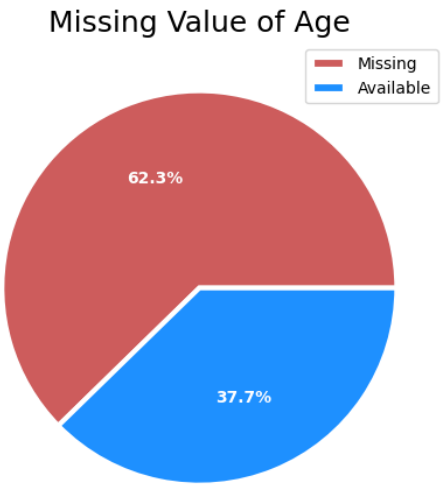

Figure 4.2: Missing value percentage of Age

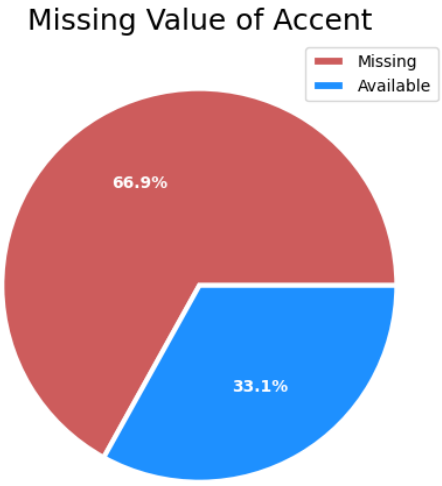

Figure 4.3: Missing value percentage of Accent

From table 4.1, we can see 122008 value missing and 73768 available in age column. If we see in the pie chart 4.2, we can see 62% value missing and 38% value available in age column. From table 4.1, we can see 121717 value missing and 74059 available in gender column. If we see in the pie chart 4.1, we can see 62% value missing and 38% value available in gender column. From Table 4.1, we can see 131065 value missing and 64711 available in accent column. If we see in the pie chart 4.3, we can see 62% value missing and 38% value available in age accent.

4.1.3 Categorical Values

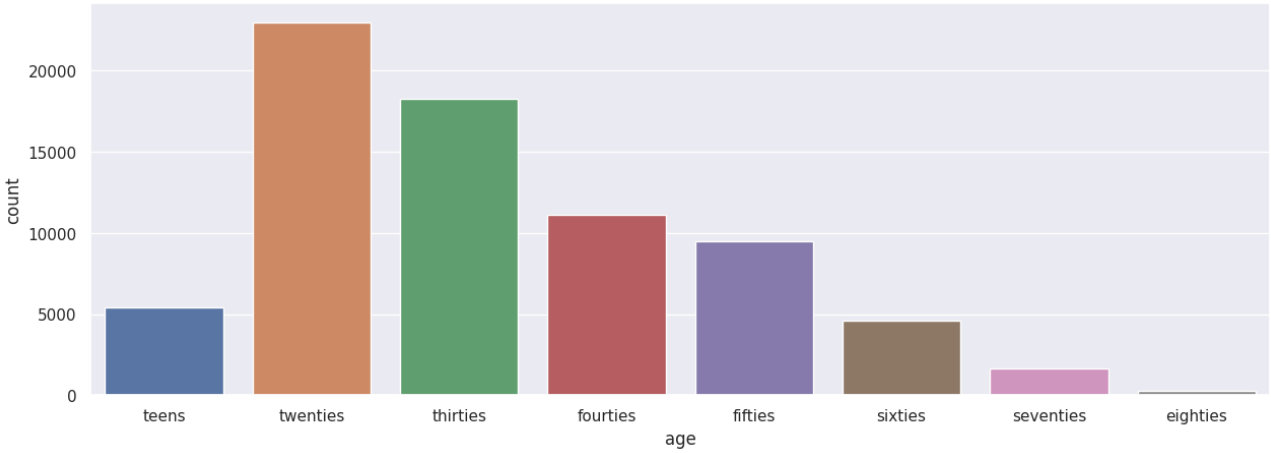

Figure 4.4: Bar chart of age values.

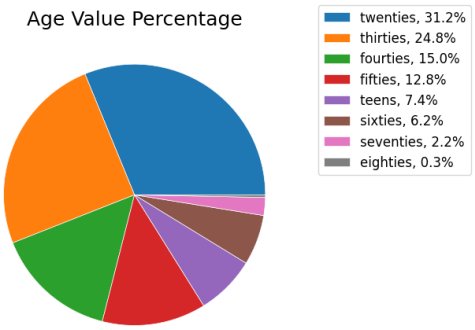

Figure 4.5: Age value pie chart

From 4.4 bar chart, we can observe age data categories are imbalanced. Some categories got few folds more values than other categories. If we think it through it is not that surprising. Considering data is crowd sourced and it can be inputted by everyone, it was bound to represent demographic of internet users. As we can see from the bar chart, older categories has less speech. Speech percentage is decreasing with age with the exception teens. Lack of teen's voice clip can be do with their lack of desire to contribute or unaware of benefits of contribution to crown sourced community.

Twenties and thirties have the most voice clips. From 4.5, we can see 31% and 25% clips are from twenties and thirties. They have the majority of the voice with 56%. Seventies and eighties have very few voice clips as there are very few internet users in that age group. Eighties only have 0.3% speech. We have to keep an eye how model perform with so much

imbalanced in different categories. For example, twenties with 31% has 100 times more data than eighties with 0.3%.

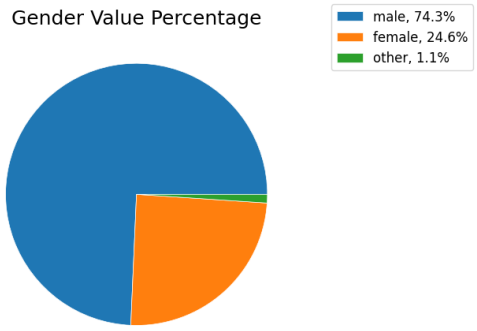

Figure 4.6: Gender value pie chart

Gender has three categorical values, male, female and others. From 4.6 pie chart, we can see male and female ratio is disproportionate. There are almost three times more male clips than female clips with male have 74% and female have 25%. Other gender is only 1.1%.

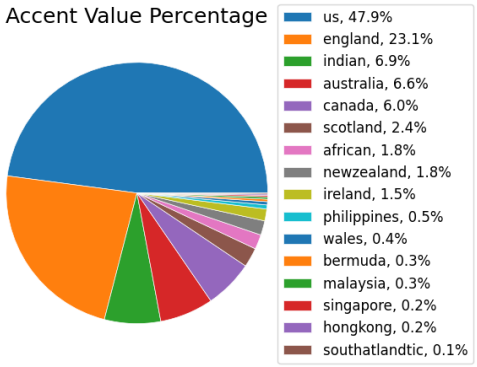

Figure 4.7: Accent value pie chart.

From 4.7 pie chart, we can observe that there are 16 types of english accents in dataset. However, us and england accent account for more than 70% of the values with american accent count for almost 48%. Other notable accents with more than 5% are indian, australia and canada. Among the 16 accents, there are about 7 accents with less than 0.5% percent values which are philiphines, wales, bermuda, malaysia, singpore, hongkong and south atlantic.

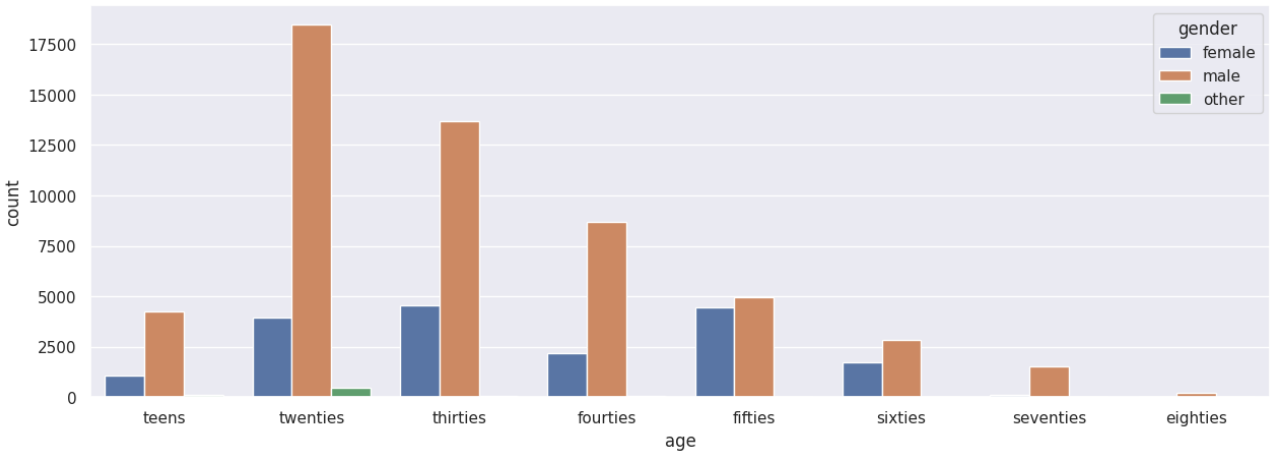

Figure 4.8: Age gender combination bar chart.

From above figure 4.8, we observe that teens, twenties, thirties and fourties age group have 3 or 4 times male population than female population. It is similar to the whole dataset ratio. However, fifties and sixties have better female and male ratio. Fifties age group has almost same male and female ratio. On the other hand, seventies and eighties have almost no female population. We could notice other gender population only on twenties age group.

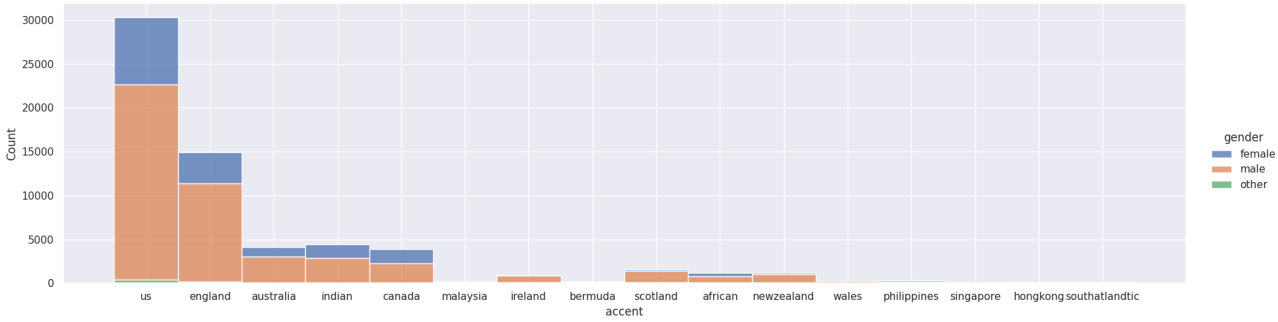

Figure 4.9: Age accent combination bar chart.

From above figure 4.9, we can see majority of the accents in the dataset are us, england, australia, indian and canada. All of those have about 3 to 4 times more male than female which is logical considering the whole dataset male female ratio. However, we can see significant portion of the other gender only speaking us accent.

4.2 Dev Dataset

4.2.1 General Information

There are 4076 rows of data in valid dev dataset. There are 8 columns which are filename, text, up_votes, down_votes, age, gender, duration and accent. We will not focus on up_votes and down_votes as they are used for data validation purpose. They are suitable with our need. Our main focus will be on age, gender, duration and accent as they are more related to our task. Also we will try to find out if dev dataset properly represent as sample of train dataset as we are initially going to experiment with dev dataset.

4.2.2 Missing Values

Column	Available	Missing
filename	4076	0
text	4076	0
up_votes	4076	0
down_votes	4076	0
age	1528	2548
gender	1540	2536
accent	1350	2726
duration	0	4076

Table 4.2: Dev dataset statistics of missing and available values

From 4.2, we can see filename, text, up_votes and down_votes have all the values available. It should not come as surprise as it is mentioned 3.3 that those values are mandatory value. However, we can see missing values in age, gender, accent and duration as they are optional value as mentioned in 3.3. We can see observe that data is maintained it's convention. One thing that could come as little surprising is that there is not a single value available in duration column. Let's take a deeper drive into the missing values of age, accent and gender.

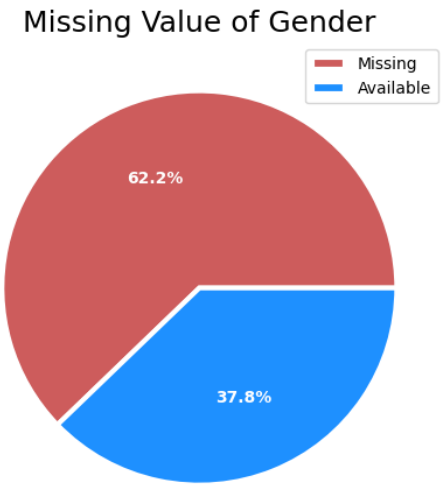

Figure 4.10: Missing value percentage of Gender

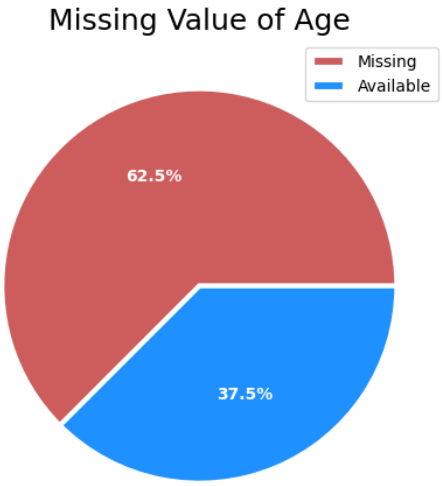

Figure 4.11: Missing value percentage of Age

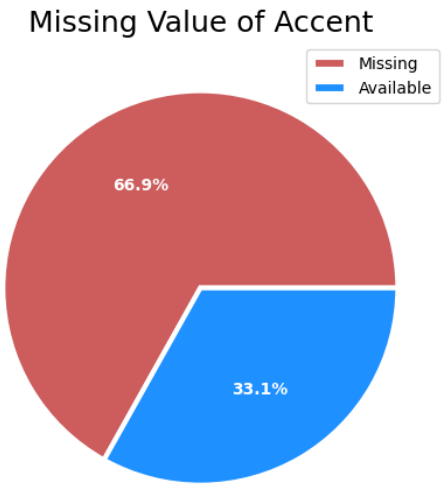

Figure 4.12: Missing value percentage of Accent

From table 4.2, we can see 2548 value missing and 1528 available in age column. If we see in the pie chart 4.11, we can see 62% value missing and 38% value available in age column. From table 4.2, we can see 2536 value missing and 1540 available in gender column. If we see in the pie chart 4.10, we can see 62% value missing and 38% value available in gender column. From Table 4.2, we can see 2726 value missing and 1350 available in accent column. If we see in the pie chart 4.12, we can see 67% value missing and 33% value available in age accent.

4.2.3 Categorical Values

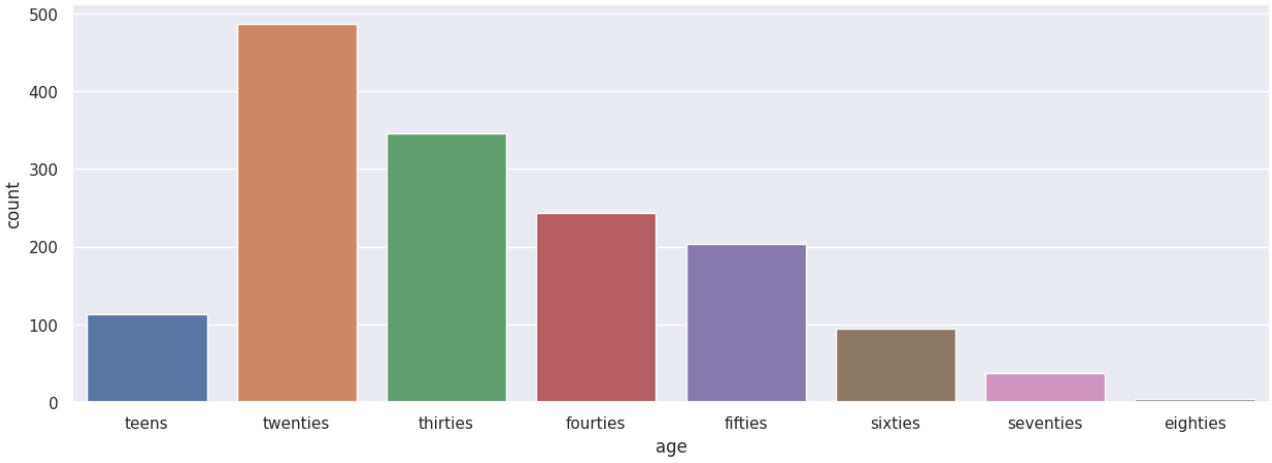

Figure 4.13: Bar chart of age values.

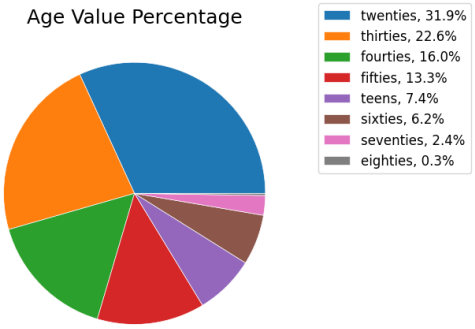

Figure 4.14: Age value pie chart

From 4.13 bar chart, we can observe age data categories are imbalanced. Some categories got few folds more values than other categories. If you think it through that is not that surprising. Considering data is crowd sourced and it can be inputted by everyone, it was bound to represent demographic of internet users. As we can see from the bar chart, older categories has less speech. Speech percentage is decreasing with age with the exception teens. Lack of teen's voice clip could be do with their lack of desire to contribute or unaware of benefits of contribution to crown sourced community.

Twenties and thirties have the most voice clips. From 4.14, we can see 32% and 23% clips are from twenties and thirties. They have the majority of the voice with 55%. Seventies and eighties have very few voice clips as there are very few internet users in that age group. Eighties only have 0.3% speech. We have to keep an eye how model perform with so much imbalanced in different categories. For example, twenties with 32% has 100 times more data than eighties with 0.3%.

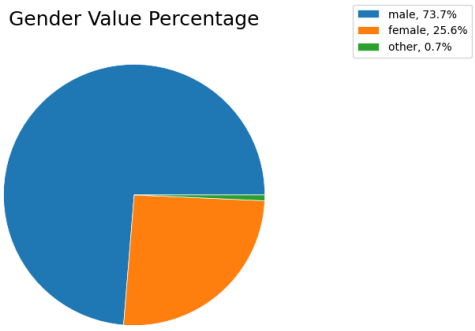

Figure 4.15: Gender value pie chart

Gender has three categorical values, male, female and others. From 4.15 pie chart, we can

see male and female ratio is disproportionate. There are almost three times more male clips than female clips with male have 73% and female have 26%. Others is only 0.7%.

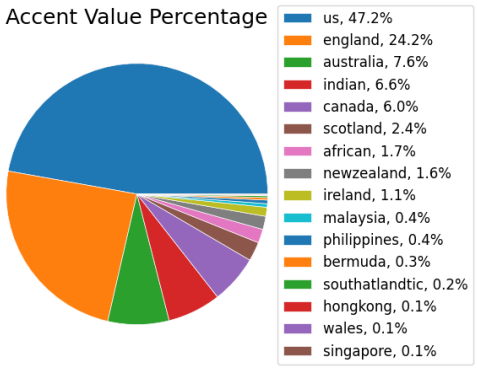

Figure 4.16: Accent value pie chart.

From 4.16 pie chart, we can observe that there are 16 types of english accent in dataset. However, us and england accent account for more than 71% of the values with american accent count for almost 47%. Other notable accents with more than 5% are indian, australia and canada. Among the 16 accents, there are about 7 accents with less than 0.5% percent values which are philiphines, wales, bermuda, malaysia, singpore, hongkong and south atlantic.

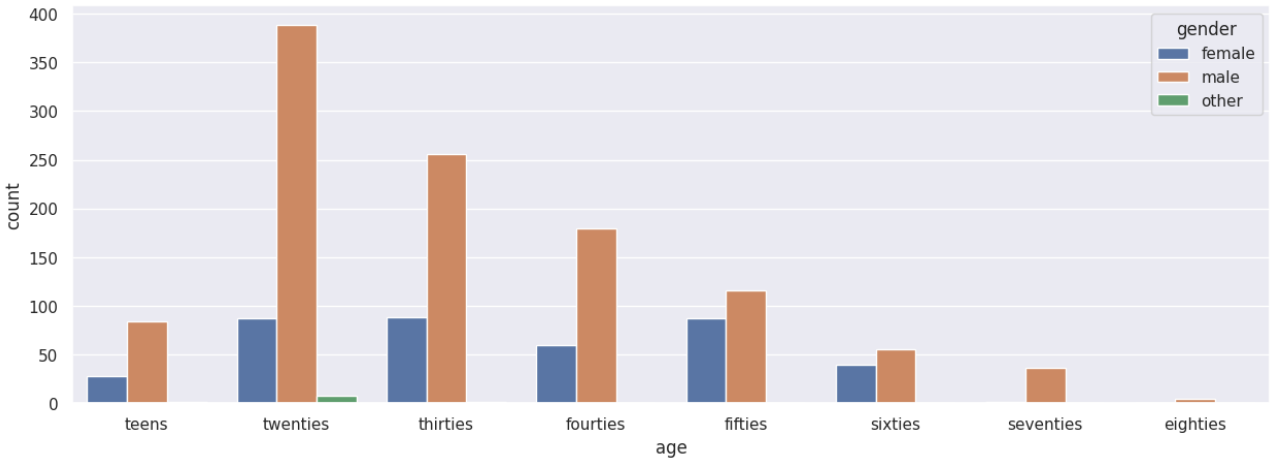

Figure 4.17: Age gender combination bar chart.

From above figure 4.17, we observe that teens, twenties, thirties and fourties age group have 3 or 4 times male population than female population. It is similar to the whole dataset ratio. However, fifties and sixties have better female and male ratio. With fifties age group has almost equal male and female ratio. On the other hand, seventies and eighties have

almost no female population. We could notice other gender population only on twenties age group.

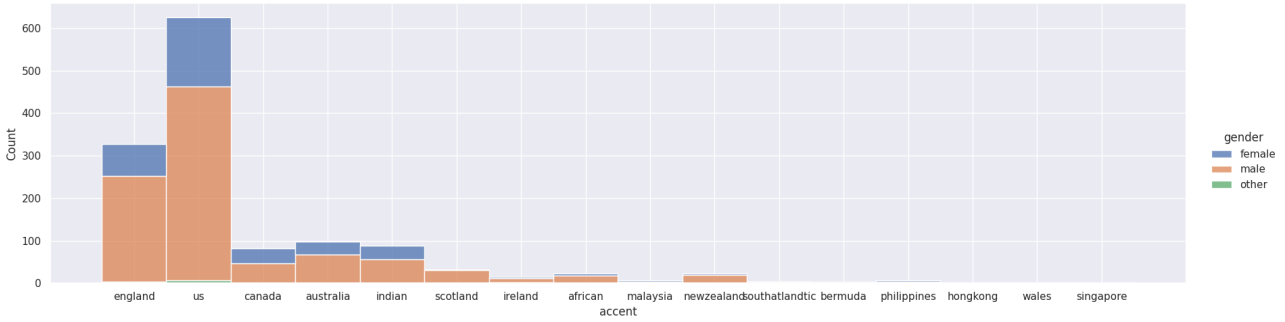

Figure 4.18: Age accent combination bar chart.

From above figure 4.18, we can see majority of the accents in the dataset are us, england, australia, indian and canada. All of those have about 3 to 4 times more male than female which is logical considering the dataset. However, we can see significant portion of the other gender only speaking in us accent.

From above observations, we can conclude that dev dataset is similar to train dataset. As our plan to start with small database to experiment and research, it would be good sample set of train dataset.

4.3 Test Dataset

4.3.1 General Information

There are 3995 rows of data in valid test dataset. There are 8 columns which are filename, text, up_votes, down_votes, age, gender, duration and accent. We will not focus on up_votes and down_votes as they are used for data validation purpose. They are suitable with our need. Our main focus will be on age, gender, duration and accent as they are more related to our task. We want to make sure that test dataset is similar in pattern and all the categorical values are same in both train and test or at least available in train dataset.

4.3.2 Missing Values

From 4.3, we can see filename, text, up_votes and down_votes have all the values available. It should not come surprise as it is mentioned 3.3 that those values are mandatory value.

Column	Available	Missing
filename	3995	0
text	3995	0
up_votes	3995	0
down_votes	3995	0
age	1542	2453
gender	1541	2454
accent	1338	2657
duration	0	3995

Table 4.3: Test dataset statistics of missing and available values

However, we can see missing values in age, gender, accent and duration as they are optional value as mentioned in 3.3. We can observe that data is maintained it's convention. One thing that could come as surprising is that the duration column as there is not a single value available. Let's take a deeper drive into the missing values of age, accent and gender.

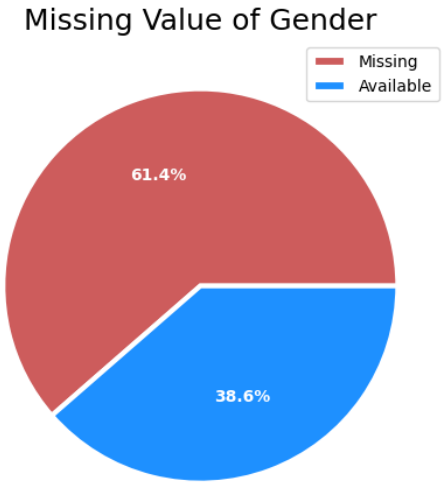

Figure 4.19: Missing value percentage of Gender

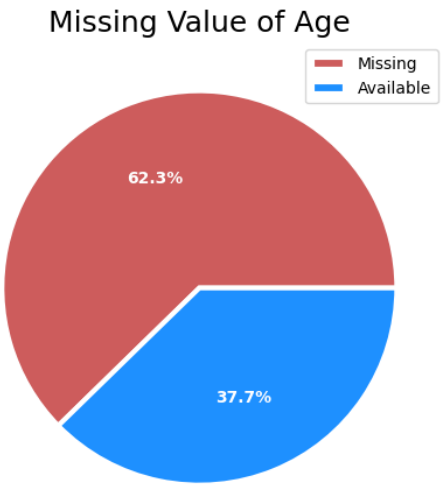

Figure 4.20: Missing value percentage of Age

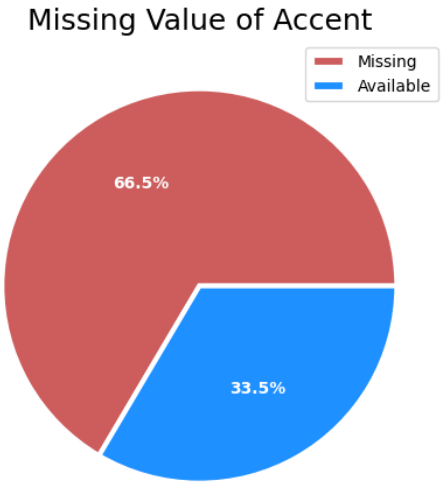

Figure 4.21: Missing value percentage of Accent

From table 4.3, we can see 2453 value missing and 1542 available in age column. If we see in the pie chart 4.20, we can see 62% value missing and 38% value available in age column. From table 4.3, we can see 2454 value missing and 1541 available in gender column. If we see in the pie chart 4.19, we can see 61% value missing and 39% value available in gender column. From Table 4.3, we can see 2657 value missing and 1338 available in accent column. If we see in the pie chart 4.21, we can see 67% value missing and 33% value available in age accent.

4.3.3 Categorical Values

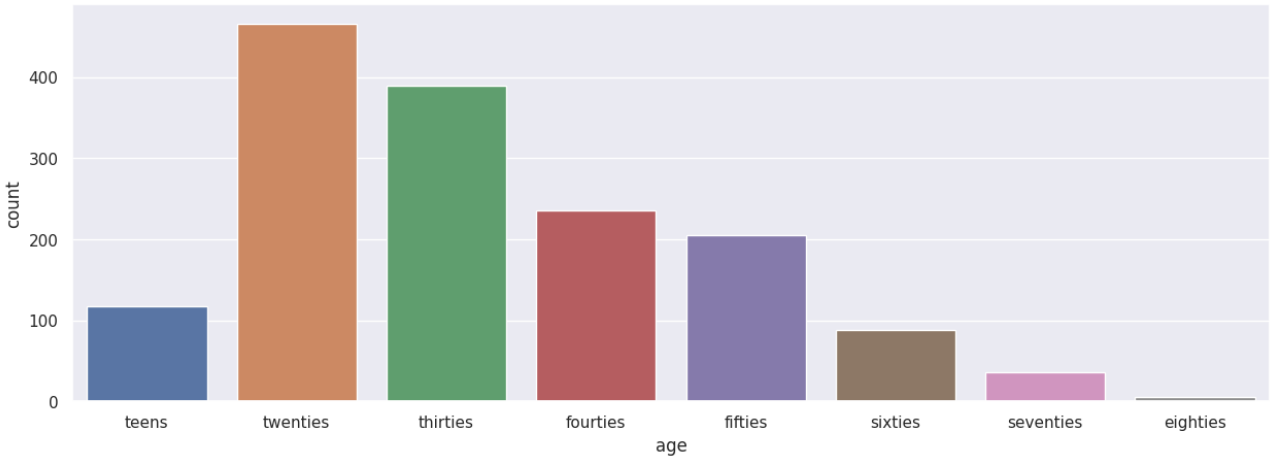

Figure 4.22: Bar chart of age values.

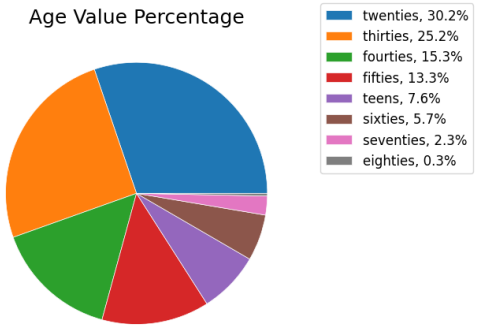

Figure 4.23: Age value pie chart

From 4.22 bar chart, we can observe age data categories are imbalanced. Some categories got few folds more values than other categories. If you think it through that is not that surprising. Considering data is crowd sourced and it can be inputted by everyone, it was bound to represent demographic of internet users. As we can see from the bar chart, older categories has less speech. Speech percentage is decreasing with age with the exception teens. Lack of teen's voice clip could be do with their lack of desire to contribute or unaware of benefits of contribution to crown sourced community.

Twenties and thirties have the most voice clips. From 4.23, we can see 30% and 25% clips are from twenties and thirties. They have the majority of the voice clips with 55%. Seventies and eighties have very few voice clips as there are very few internet users in that age group.

Eighties only have 0.3% speech. We have to keep an eye how model perform with so much imbalanced in different categories. For example, twenties with 30% has 100 times more data than eighties with 0.3%.

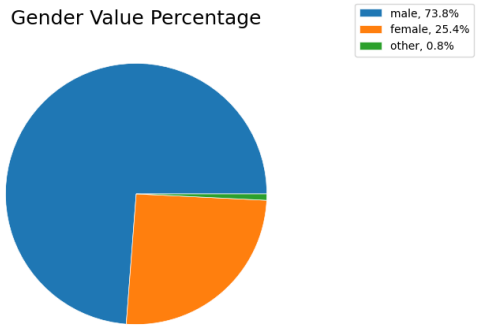

Figure 4.24: Gender value pie chart

Gender has three categorical values, male, female and others. From 4.24 pie chart, we can see male and female ratio is disproportionate. There are almost three times more male clips than female clips with male have 74% and female have 25%. Other gender is only 0.8%.

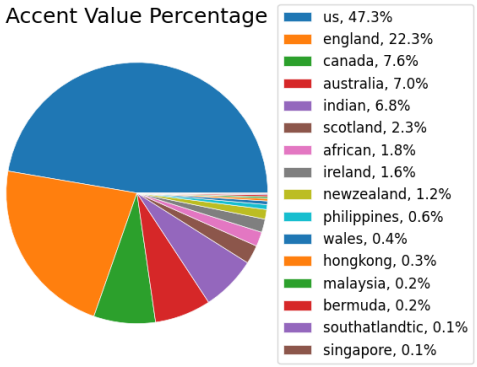

Figure 4.25: Accent value pie chart.

From 4.25 pie chart, we can observe that there are 16 types of english accent in dataset. However, us and england accent account for more than 69% of the values with american accent count for almost 47%. Other notable accents with more than 5% are indian, australia and canada. Among the 16 accents, there are about 7 accents with less than 0.5% percent values which are philiphines, wales, bermuda, malaysia, singpore, hongkong and south atlantic.

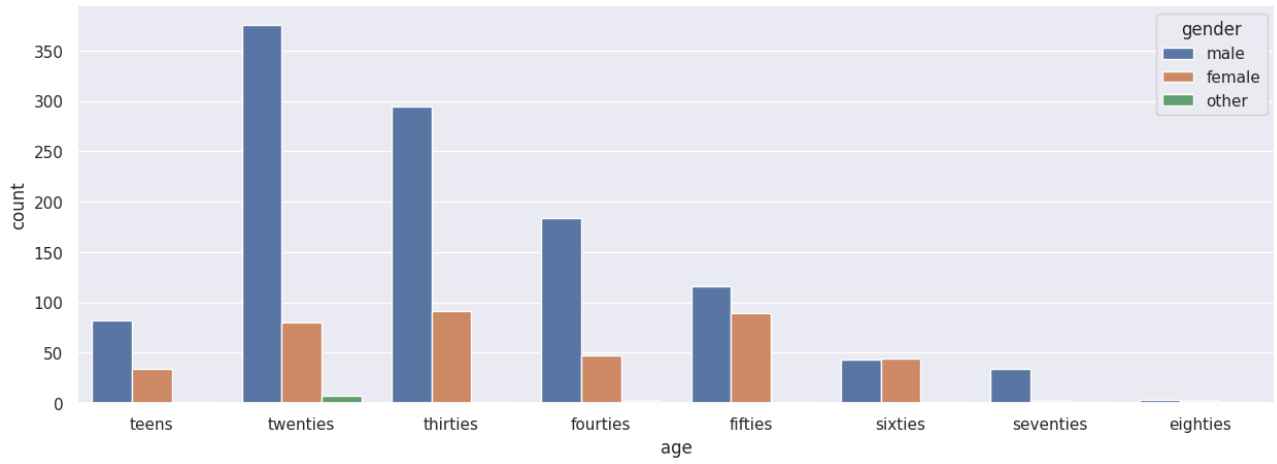

Figure 4.26: Age gender combination bar chart.

From above figure 4.26, we observe that teens, twenties, thirties and fourties age group have 3 or 4 times male population than female population. It is similar to the whole dataset ratio. However, fifties and sixties have quite a better female and male ratio. Sixties age group has almost same male and female ratio. On the other hand, seventies and eighties have almost no female population. We could notice other population only on twenties age group.

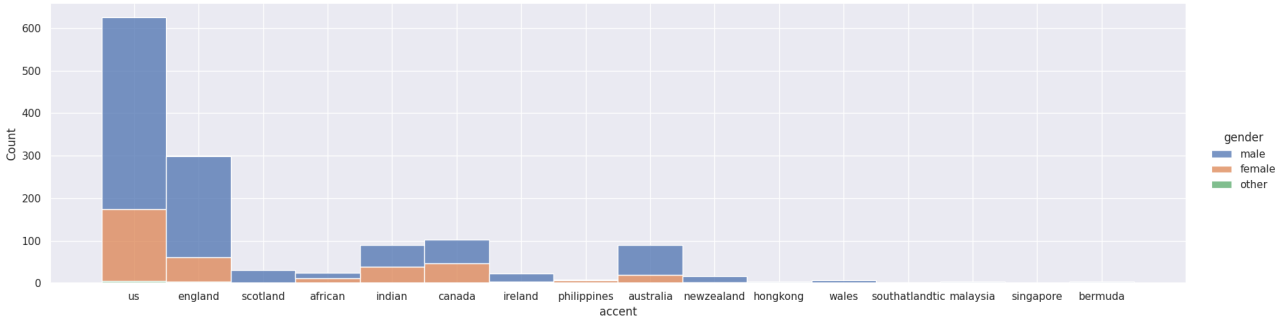

Figure 4.27: Age accent combination bar chart.

From above figure 4.27, we can see majority of the accents in the dataset are from us, england, australia, indian and canada. All of those have about 3 to 4 times more male than female which is logical considering the whole dataset female and male ratio. However, we can see significant portion of the of the other gender only speaking us accent.

Overall, test dataset has similar ratio and categorical values like train dataset. There is no extra categorical values found in test dataset. We should be good to test with test dataset after using train dataset for training.

CHAPTER 5

Data Preprocessing

5.1 Data Organization

After experimenting with dev dataset, we have found out that processing audio data could be time consuming. So, we needed to take some measures to save time and work efficiently. One of those measures was to use both Google Colab [2] and Kaggle notebook [5] simultaneously. Kaggle already had common voice database as standard kaggle database. As a result, using common voice with kaggle was quite straight forward. However, using common voice with Google colab is completely different story. I tried different procedure to use common voice database with google colab. But, most of the procedures did not work out. The main problems were size of database (13gb), format(zip) and one folder got more than 100k audio files. Only valid train dataset contains 195776 audio files 4.1.1. Though it was successfully uploaded, but it got timeout to fetch file from 195776 files. So, we ended up dividing the audio dataset into 196 different folders. Each folder have 1000 audio files except the last one which have 776 files. All the audio files were in chronological order. It started with 0,1,2 and so on until 195776. We parse the number from name and divide by 1000 with floor value to get the folder name. We used the following equation to get folder name.
$folder\_name = \lfloor \frac{number}{1000} \rfloor$ where number is the order of the audio file and folder_name is the subfolder, it belongs to.

5.2 Data Cleaning

Among the seven columns 3.3 in dataset, up_votes and down_votes columns are not related with our task. So, we are going to drop those columns. We are also going to drop both text and accent, as they are generally not going to be present with real time audio. However, we are going to keep gender column to see it's influence. We have discussed earlier about the influence of gender over voice [28]. We want to experiment if it is going help to recognize age or not. So, we are going to filter any row that have either age or gender missing. So, after those filters, we are left with 73768 rows in train dataset and 1541 rows in test dataset.

CHAPTER 6

Feature Engineering

Firstly, we have to figure out how to extract feature from audio clips. From 2.1, we found out that MFCCs is the most popular feature extraction technique from audio. It was first introduced by Davis and Mermelstein in the 1980's[17]. It is widely used for ASR and various speech classification tasks. We have already seen in 2.1, MFCCs is used to extract features in various similar researches about age recogntion from voice. We are also going to use MFCCs to extract features.

6.1 Mel Frequency Cepstral Coefficents (MFCCs)

To understand MFCCs, we have to understand how human make sound and how phoneme being produced. The sound that generated by humans are filtered by voice tract like tongue, teeth etc into shape of that voice tract itself as we discussed in 1.2. From envelope of short time power spectrum, we can see the shape of voice tract itself being created. MFCCs accurately represent this envelope from sound.

6.1.1 Steps

Let's take a look into high level steps on how to extract feature from audio using MFCCs [6].

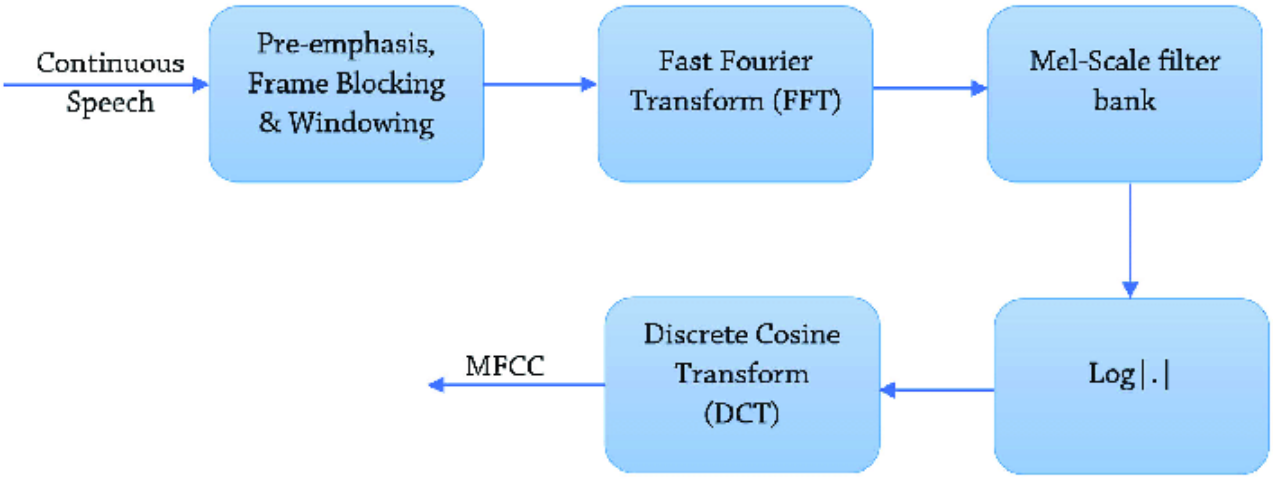

Figure 6.1: MFCCs feature extraction steps taken from [16]

1. Frame the signal of 20-40 ms into frames
2. Calculate peridogram estimate of power spectrum for every frame
3. Apply mel filterbank on various frequency regions and sum the energy in each filter
4. Generate logarithm of all the filterbank enegies
5. Generate Discrete Cosine Transform (DCT) of all log filterbank energies

It is an oversimplified version of MFCCs. Take a closer look at 6.1 to understand more clearly. We will discuss in more details at later sections about implementation and why it works [6].

6.1.2 Why MFCCs works?

Audio signal is constantly changing. To keep matter simple, we take 20 ms to 40 ms frame where it does not change that much. Anything smaller than that would be unreliable as there would not be enough sample. Anything larger than that, time frame could result in too much signal change [6].

Calculation of power spectrum is influenced by cochlea which is an organ in our ear. It vibrates in different spots depending on the sound we hear. From those vibrations we differentiate different frequencies. Periodogram estimate finds different frequencies presence in a frame similar to cochlea [6].

Periodogram estimate is still not good enough as it contains too much information. We need to filter out more to work more like cochlea. One of the behaviour of cochlea is that, it

can not differentiate between two closely spaced frequency. This behaviour is more observed with higher frequency. To replicate this behaviour, we use Mel filterbank. The first Mel filter is narrow and it give us how much energy exist near 0 Hz. However, as frequency increase so the filter size increases. As a result it becomes less concerned with frequency variation. Mel scale exactly does this job of making how wide the filter should be, for different frequencies [6].

After we get all the filterbank energies, we will take logarithm(log) value of them. This is also to do with human hearing system. Human hearing of sound volume is not linear to the energy of specific sound. To double the perceived volume of a sound for a human, double the energy of the sound is not enough. We have to create a sound with 8 times more energy to make human perceive double the volume. So, log values of filterbank energies help to represent similar to what human hears [6].

Lastly, we perform DCT on log values of filterbank energies. We do DCT because filterbank are overlapping and filterbank energies are heavily correlated. Applying DCT, decorrelates the filterbank energies [6].

6.1.3 Mel Scale

Mel scale help to perceive sound frequency like human ear. Human ear does much better job of detecting small change at lower frequency than higher frequency. Mel scale help to replicate similar behavior like human ear [6].
Equation for converting frequency to Mel scale,
$M(f) = 1125ln(1 + \frac{f}{700})$
Equation for converting Mel scale to frequency,
$M^{-1}(m) = 770(exp(\frac{m}{1125}) - 1)$

6.1.4 Implementation

Let's assume we start with 16000Hz audio signal.

1. Firstly, frame the signal into 25ms frames. For 16000Hz, frame length will be $16000 * 0.025 = 400$ samples. If we take frame step size 10ms then frame step size in samples would be $16000 * 0.01 = 160$. So, first frame would start at 0 with 400 samples and next

frame with 400 sample will start at 160, then 320 and so on. We can observe that frames will overlap [6].

Let's define some notation,

$s(n)$: time domain signal

$s_i(n)$ where n ranges over 1-400 and i ranges over number of frames

$S_i(k)$ is Discrete Fourier Transform(DFT) of i^{th} the frame

$P_i(k)$ is power spectrum of i^{th} frame

2. We will perform DFT of i^{th} frame using following formula,
   $S_i(k) = \sum_{n=1}^{N} s_i(n)h(n)e^{-j2\pi kn/N} \qquad 1 \leq k \leq K$
   where $h(n)$ is N sample analysis window and K is the length of DFT.
   Periodogram estimate of the power spectrum equation,
   $P_i(k) = \frac{1}{N}|S_i(k)|^2$
   We generally perform FFT with 512 point and keep the first 257 values [6].

3. Now, we are applying 26 triangular filter to the periodogram power spectral estimate. Filterbanks will be formed with 26 vectors of length 257 which we got from previous step. We multiply filterbanks with periodogram power spectral estimates and then add them up. We will end up with 26 numbers which is an indication of how much energy was in each filterbank [6]. Look closely into the figure 6.2 to understand more clearly.

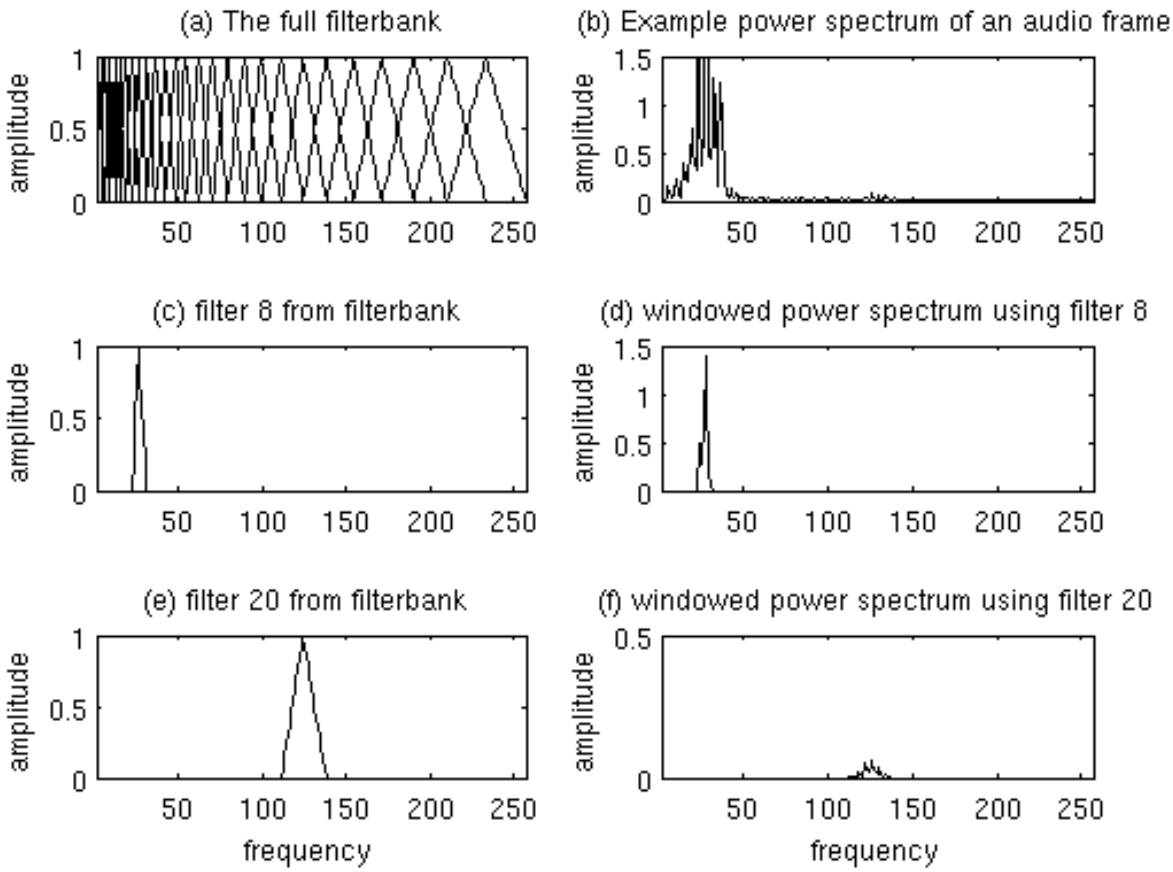

Figure 6.2: Mel filterbank and power spectrum estimates [6]

4. Convert 26 filterbank engeries into logarithm values

5. Take DCT of 26 values frrom previous steps to convert them into 26 cepstral coefficient. Those 26 values are called MFCCs [6].

6.2 Other features

Other than MFCCs, we are also using 3 other features named spectral centroid, spectral bandwidth and spectral rolloff. We will also experiment with Gender. We are going to discuss more about those features below and why they are important.

Gender: As research suggested in 1.2, gender has great influence over voice. We will try our models both with and without gender feature.

Spectral Centroid: From research, we found out that spectral centroid speaker centroid

feature could come handy during audio classification problem [23]. Spectral centroid is the center of mass of spectrum [9].

Spectral Bandwidth: Spectal bandwidth have worked as an influential feature in speech recognition [25]. It is the width, at half the maximum intensity, of the band of light leaving the monochromator [8].

Spectral Rolloff: Spectral Rolloff gives an idea where certain amount of energy (generally 85%) is contained at a spectogram. It is used in varous speech classification problem like emotion detection [27] and depression detection [31].

6.3 Feature Selection

Feature selection is important because it helps reduce complexity. As a result it reduce computational time and overfitting of model. It will also reduce training time. In our case since independent variables are numerical and dependent variable is classifier, anova is a good choice for feature selection.

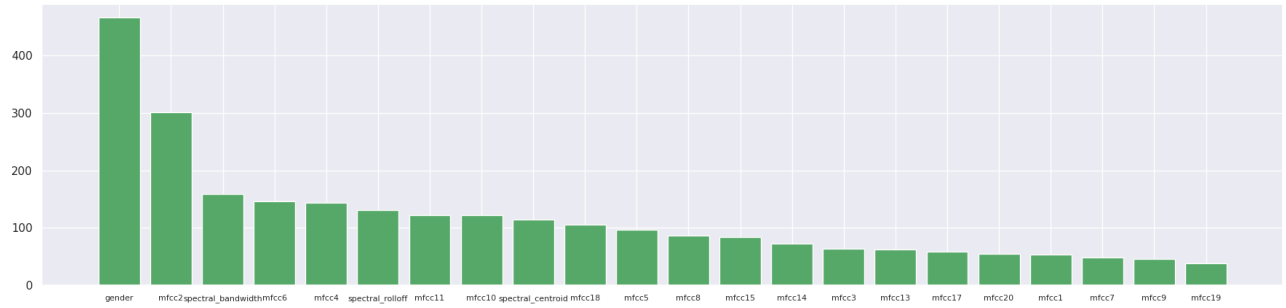

Figure 6.3: Feature selection using anova f score

Anova test gives us F values than we sort them to find most important features. We used anova test to select 22 most important features. From fig 6.3, we can see gender is the most important feature and followed by mfcc2. Spectral centroid, spectral bandwidth and spectral rolloff all those other features from 6.2, also playing an important role. They are all inside top 10 important features.

CHAPTER 7

Methodology

7.1 Traditional Machine Learning

7.1.1 Models

We have used 6 traditional machine learning models to recognise age from voice.

1. K nearest neighbor (KNN)
2. Support Vector Machine Classifier (SVC)
3. Random Forest (RF)
4. Gradient Boost (GB)
5. Decision Tree (DT)
6. Gaussian Naive Bayes (NB)

7.1.2 Dataset Size

We first experiment on dev dataset and found out that audio dataset processing time could be time consuming. So, we first experiment on small portion of database to find which parameter is performing better and if the algorithm is working or not. We also wanted to know to how much dataset is needed for best result and which algorithm perform better

on small data. We chose each portion randomly to avoid any bias in dataset. Check out following table 7.1 to get clear idea about the data sizes.

Data %	Rows
5%	3652
10%	7324
25%	18327
50%	36533
75%	54986
100%	73465

Table 7.1: Train dataset sizes

7.1.3 Hyper parameter Tuning

We used grid search with 3 fold cross validation to optimize model parameters. Since grid search is time consuming, we started grid search with only 5 % data. After that, we narrow our scope for parameters little bit for 10%. After getting 10% result, we narrow our parameters scope even more for 25% and this goes on until 100% data. We did this because if we had grid searched on 100% dataset, it would be taken days if not weeks to complete.

7.1.4 With or Without Gender

As we have seen influence of Gender over voice during research [28] and we also found out during feature selection 6.3, that gender is the most important feature . So, we wanted to see the result of adding gender as a feature. However, there could be scenario in real world, where gender might not be available in dataset. For those cases, we also wanted to check out how our models perform without gender.

7.2 Ensemble Learning (EL)

Ensemble learning is combination of individual models to get better result by accumulate their learnings. There are four categories of EL which are bagging, stacking, boosting and voting. We are going to use voting for our EL models [18].

7.2.1 Weighted majority voting ensemble(WMVE)

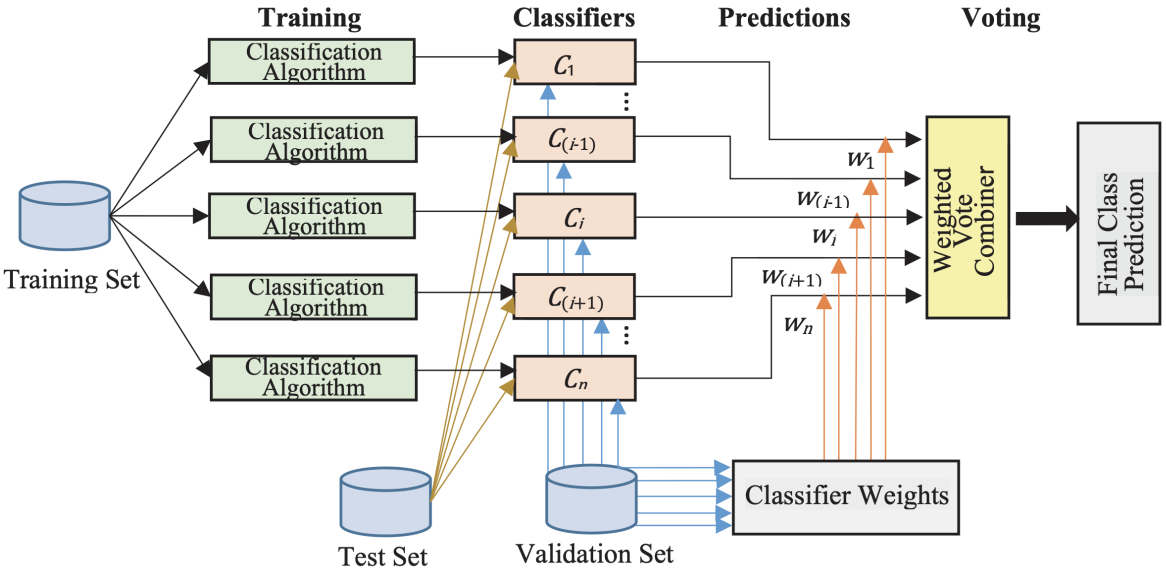

Figure 7.1: Weighted majority voting ensemble (WMVE) [18]

We will make use of classifiers in 7.1, to make our ensemble learning models. We trained 6 classifier algorithms. All those classifiers are not going to perform equally. So, giving them equal voting rights will not be fair. So, we will give different weight to each classifier.

The weight of the classifier will depend on the accuracy of the classifier. For our case, we will use f1 macro score. F1 macro score will be the weight of that classifier. Every predicted class will start 0 probability. We will add weight to predicted class probability if the model predicted that specific class. Class with max probability in the end, will be our EL model predicted class.

For example, we have three classifiers C1,C2,C3 with f1 macro score of 0.8, 0.7 and 0.2 respectively. Let's assume we have three prediction classes which are class1, class2 and class3. So, C1, C2 and C3 predicts class1, class2 and class2 respectively. So, the probability of class1 will be classifier C1's f1 score which is 0.8. The probability of class2 will be sum of classifier C2's f1 score and classifier C3's f1 score which will be $0.7 + 0.2 = 0.9$. No classifier predicted class3. So, the probability of class3 would be 0. So, class 2 has the maximum probability. EL model would predict class 2.

7.2.2 EL models classifiers selection

We created four EL models named E1, E2, E3 and E4. The reason behind creating four EL models is to create more simple model and remove model with less accuracy that could affect overall accuracy.

1. Sort all the six classifiers by F1 score
2. E1 will have all the six classifiers
3. E2 will have top five classifiers with best F1 scores
4. E3 will have top four classifiers with best F1 scores
5. E4 will have top three classifiers with best F1 scores

7.3 Hubert

As we discussed earlier in 2.2, hubert is a pretrained LLM model. We will use hugging face framework to use hubert [3]. We will also use wav2vec2 to extract feature automatically rather than manually extract feature. Firstly we have to convert our database to custom hugging face database. We will have to use gpu of google colab as without gpu it take ages to run LLM models.

7.4 Evaluation

7.4.1 Accuracy vs F1 score

The choice between accuracy and f1 score depend on the class distribution. For a long time accuracy was the most popular evaluation for classifier tasks. It works great for balanced dataset. But, for many cases like our's, where dataset is imbalanced. In imbalanced dataset, accuracy as an evaluation matrix could be deceptive [1].

For example, let's assume, there are 100 data records with two class types namely A and B. If 90 of them have class type A and 10 of them have B. If model always predict A, it will have 90% accuracy. 90% looks great on paper. However, our model is always wrong for B

class. The model just predict A. So, it is bad representation of model's performance. This is where f1 score comes to play. F1 score takes into account of class distribution[1].

$$f1score = 2 * \frac{precision * recall}{precision + recall}$$

7.4.2 Macro F1 score vs Micro F1 score

There are mainly two types of F1 score namely micro and macro. Micro F1 give equal importance to each observation. On the other hand, Macro F1 give equal importance to every class. So, macro is better for imbalanced dataset which is our case.

$$macrof1score = \frac{\sum f1score}{numberofclasses}$$

CHAPTER 8

Results

8.1 Traditional Machine Learning

We used six traditional machine learning algorithms as described in 7.1. We train the models once with gender feature and once without gender feature to figure out the influence of gender in age detection.

8.1.1 With Gender

Data %	KNN	SVC	RF	GB	DT	NB
5%	**0.602**	0.518	0.419	0.377	0.243	0.218
10%	**0.723**	0.670	0.516	0.268	0.287	0.214
25%	**0.795**	0.767	0.625	0.370	0.344	0.228
50%	**0.845**	0.820	0.714	0.305	0.362	0.230
75%	**0.876**	0.859	0.709	0.521	0.347	0.230
100%	**0.891**	0.882	0.732	0.426	0.332	0.223

Table 8.1: Test Dataset F1 macro score for Different Size train Dataset of K-nearest neighboor(KNN), Support Vector Classifier(SVC), Random Forest(RF), Gradient Boosting Classifier(GB), Decision Tree Classifier(DT) and Gaussian Naive Bayes(NB)

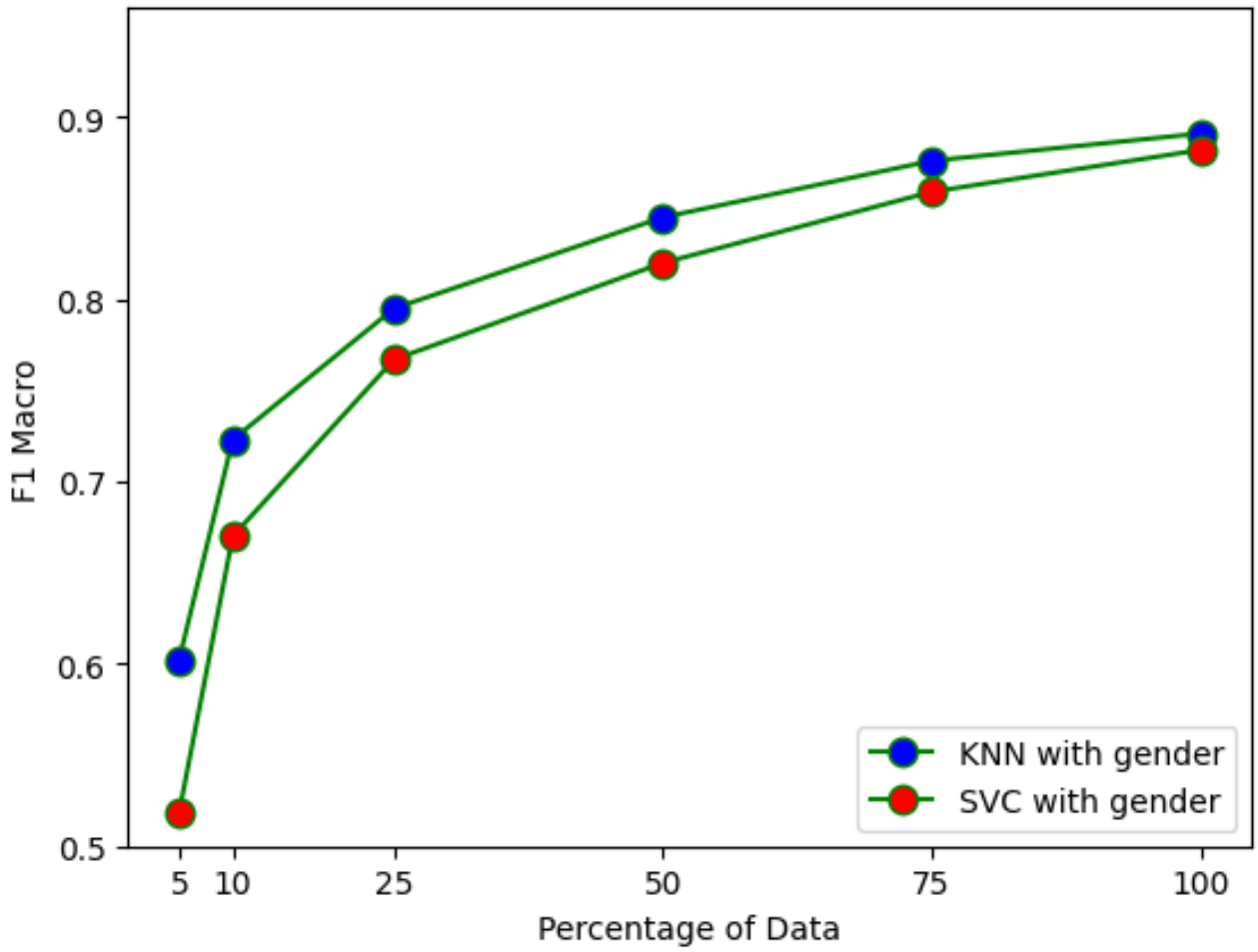

Figure 8.1: KNN vs SVC with gender for different data sizes

KNN always performed better than any other algorithm for all datasize. SVC was the second best performer. From figure 8.1, we can see the gap between SVC and KNN closing with data size. For 5% and 10% data, SVC's f1 score was quite far from KNN. However, with the increasing size of dataset, SVC performed quite close to KNN. For 100% train dataset, difference was merely 0.01. We could also see the influence of data size. From table 8.1 we can observe that, in most scenarios, model performed better with bigger data size.

8.1.2 Without Gender

Data %	KNN	SVC	RF	GB	DT	NB
5%	**0.595**	0.561	0.394	0.337	0.205	0.195
10%	**0.716**	0.663	0.456	0.412	0.254	0.193
25%	**0.797**	0.754	0.602	0.378	0.285	0.211
50%	**0.868**	0.790	0.684	0.348	0.314	0.213
75%	**0.885**	0.865	0.711	0.545	0.356	0.211
100%	**0.887**	0.881	0.697	0.467	0.305	0.209

Table 8.2: Test Dataset F1 macro score for Different Size train Dataset of K-nearest neighboor(KNN), Support Vector Classifier(SVC), Random Forest(RF), Gradient Boosting Classifier(GB), Decision Tree Classifier(DT) and Gaussian Naive Bayes(NB)

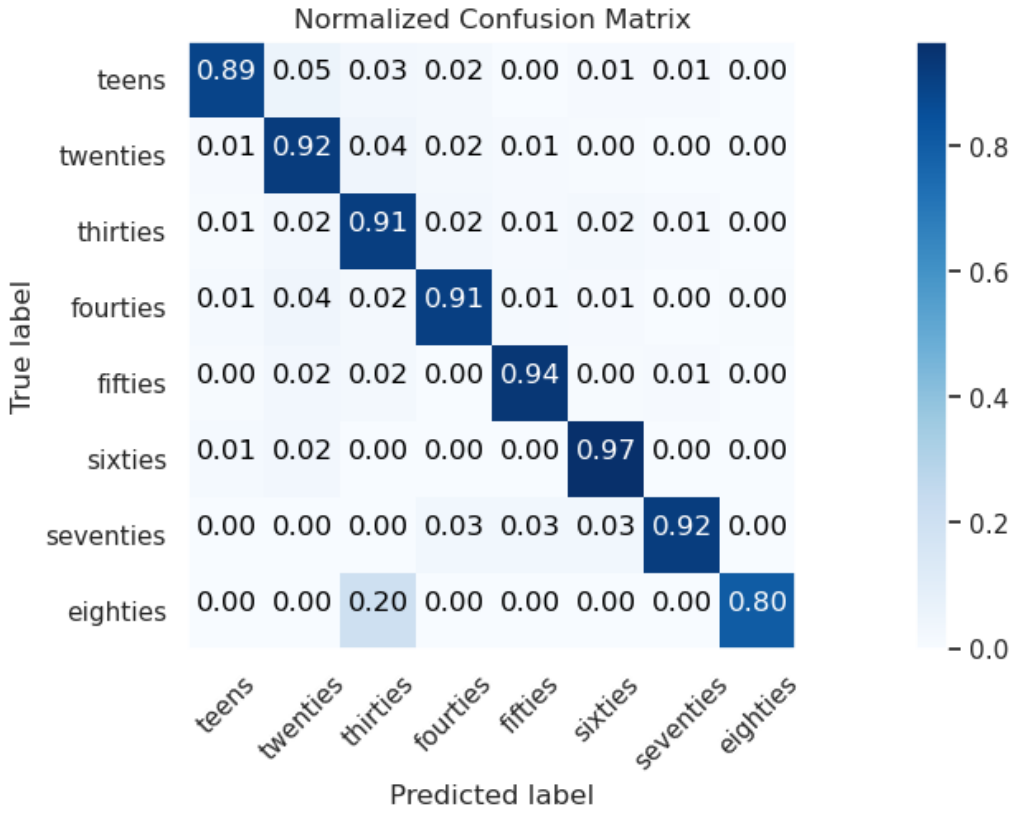

Figure 8.2: KNN normalised confusion matrix with 100% data

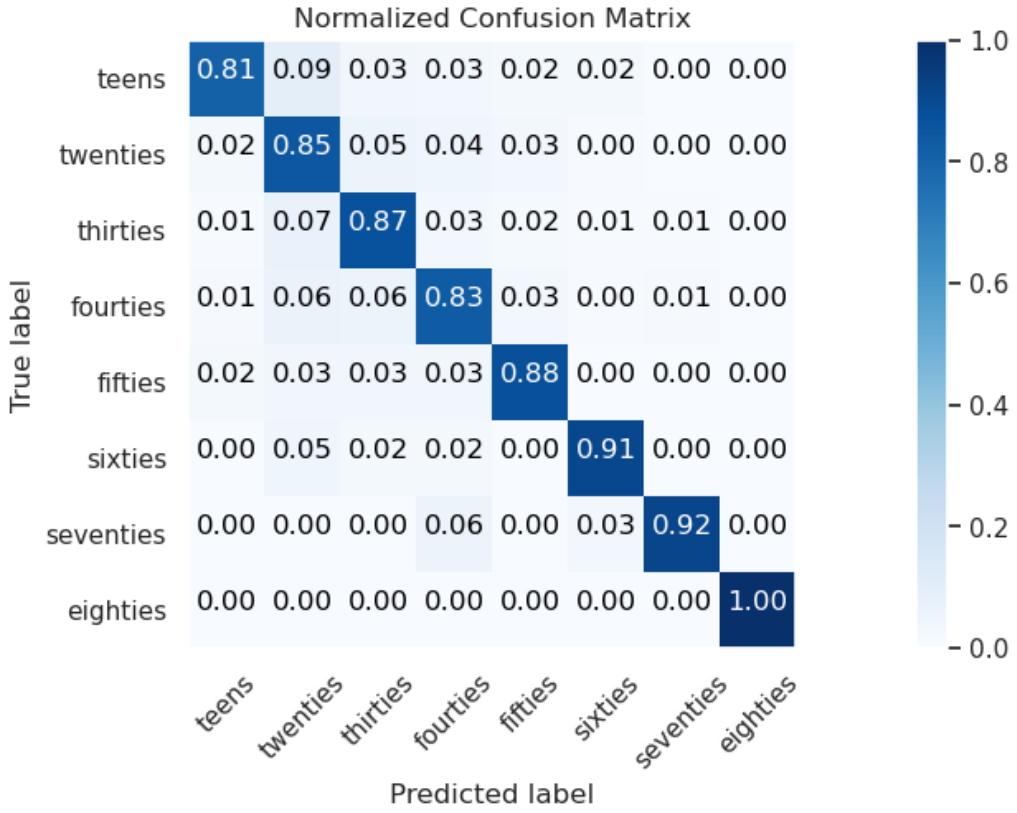

Figure 8.3: SVC normalised confusion matrix with 100% data

KNN always performed better than any other algorithm for all datasize. SVC was the second best performer. From confusion matrix 8.2 and 8.3, we can KNN performed better than SVC in all classes except for seventies and eighties. For seventies, performance was equal. For eighties, SVC's accuracy was 1.0 compared to KNN's accuracy 0.8.

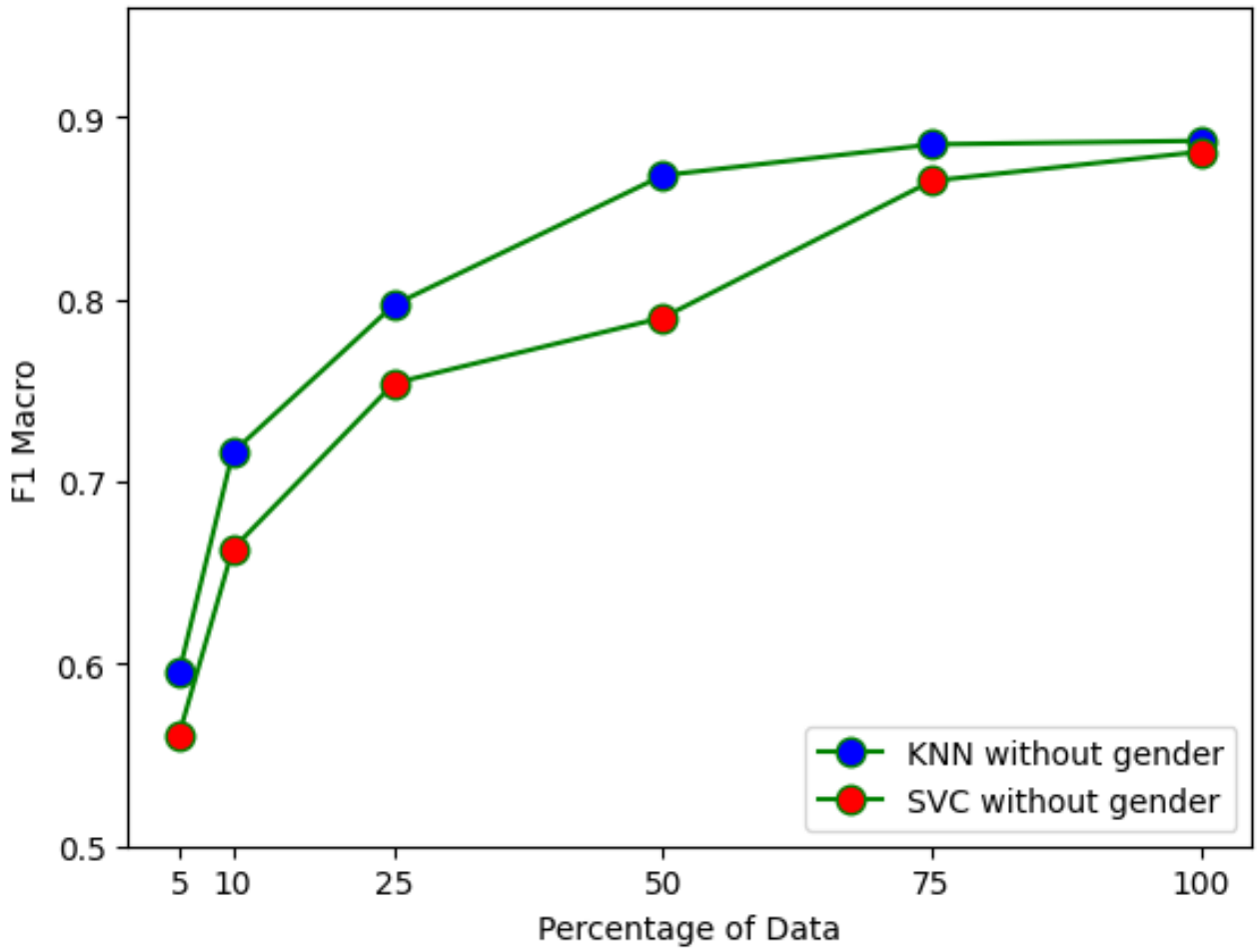

Figure 8.4: KNN vs SVC without gender for different data sizes

From 8.4, we can see the gap between SVC and KNN closing with data size except for 50%. For 5% and 10% data, SVC's f1 score was quite far from KNN. However, with the increasing size of dataset, SVC performed quite close to KNN with the exception of 50%. For 50%, the accuracy gap was quite big. But after that it decreased the gap for 75% and 100%. For 100% train dataset, difference was just 0.006. We could also see the influence of data size. From table 8.2 we can observe that, in most scenarios, model performed better with bigger data size.

8.1.3 With vs Without Gender

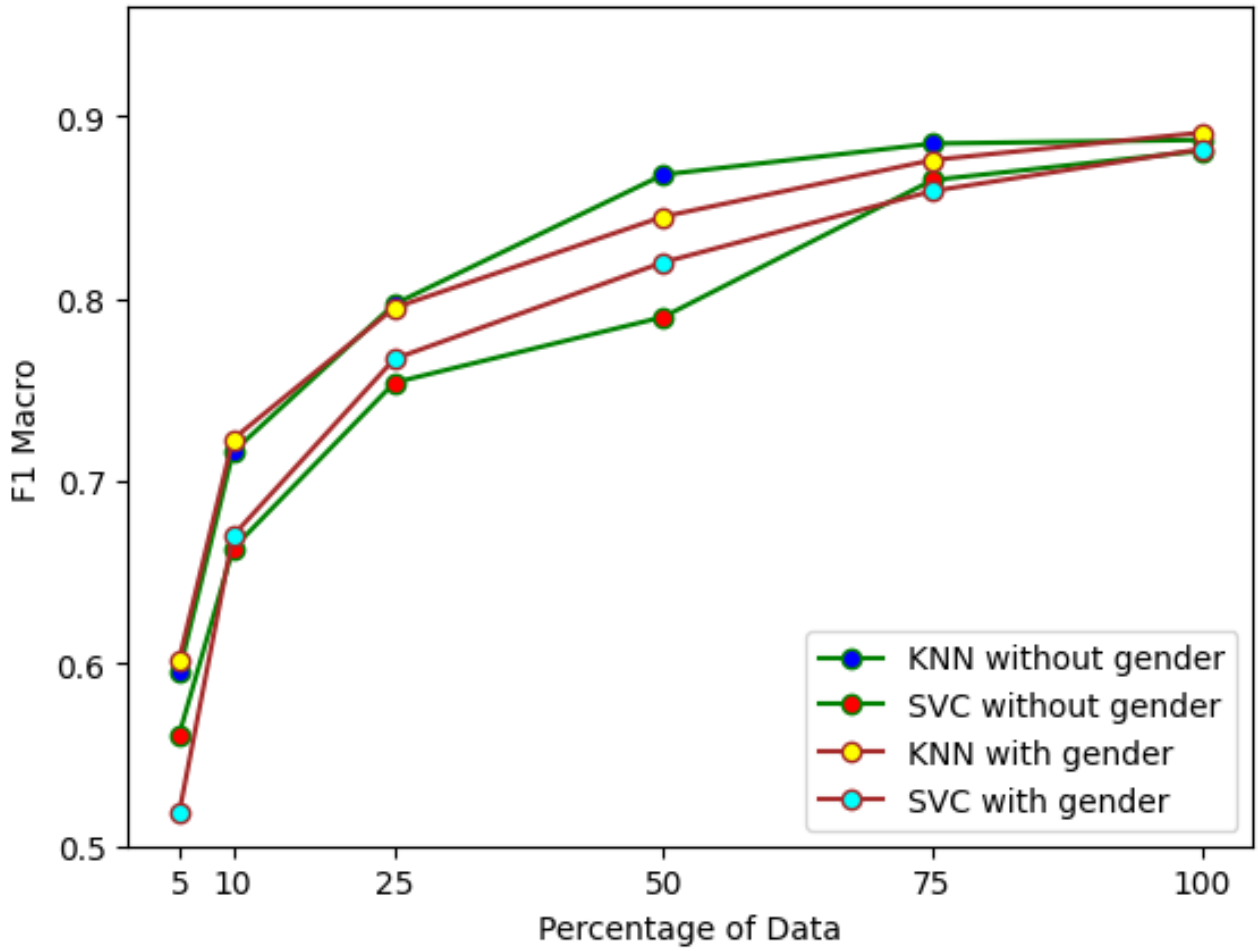

Figure 8.5: KNN vs SVC with and without gender for different data sizes

From table 8.1 and 8.2, we can clearly see influence of gender from f1 score for 100% data size. Except for GB, every other model performed better with gender for 100% data. However, for our best performing KNN and SVC difference was very little for 100%. We can see the biggest difference for DT and RF.

From figure 8.5, we see KNN with or without gender in both cases performed similar. KNN without gender even performed better for 50% data. For KNN, effect of gender is very minimal. On the other hand, SVC with gender always performed better or similar to SVC without gender. For SVC, gender influence is quite positive.

8.2 Ensemble Learning (EL)

Data %	KNN	SVC	E1	E2	E3	E4
5%	0.595	0.561	0.612	**0.616**	0.606	0.594
10%	0.716	0.663	**0.721**	0.713	0.710	0.711
25%	**0.797**	0.754	0.793	0.791	0.791	0.792
50%	**0.868**	0.790	0.822	0.840	0.841	0.861
75%	0.885	0.865	0.873	0.882	0.879	**0.896**
100%	0.887	0.881	0.894	0.894	**0.897**	0.888

Table 8.3: Test Dataset Statistics of Different Size train Dataset for K-nearest neighboor(KNN), Support Vector Classifier(SVC), E1(KNN,SVC,RF,GB,DT,NB), E2(KNN,SVC,RF,GB,DT), E3(KNN,SVC,RF,GB) and E4(KNN,SVC,RF)

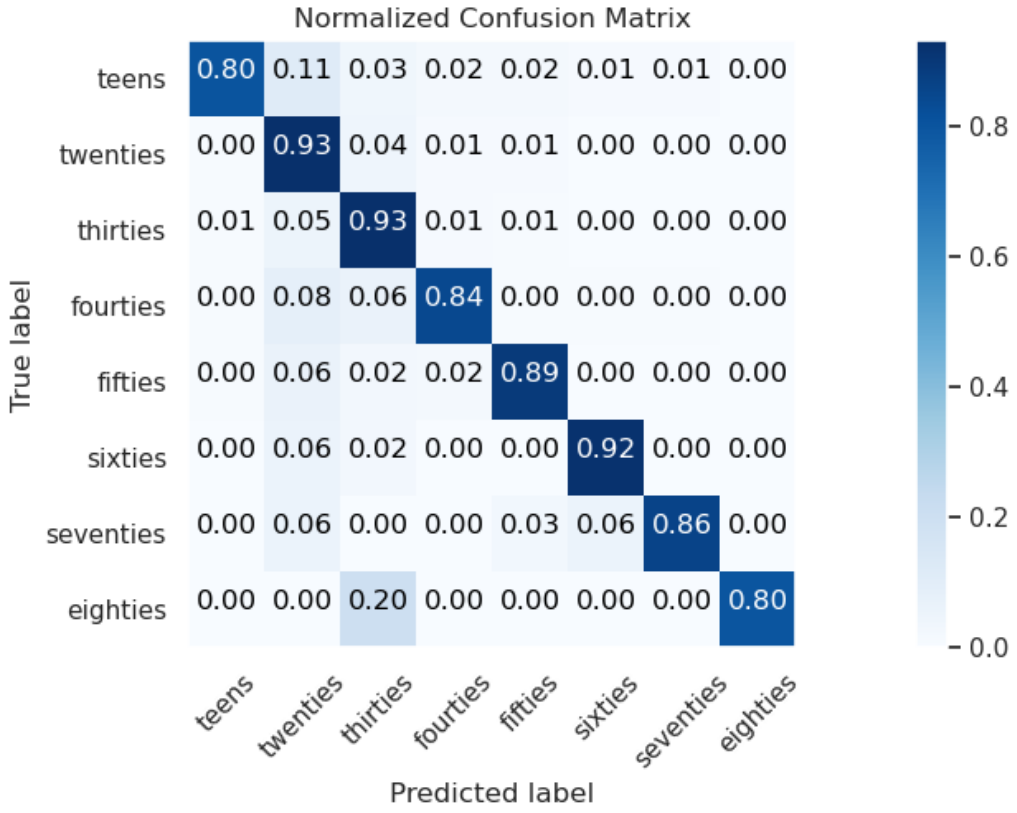

Figure 8.6: E3 normalised confusion matrix with 100% data

From 8.1.2, we get best performing models list by f1 score which is KNN, SVC, RF, GF, DT and NB respectively. Using 7.2.2 procedure, we got four different EL models. So, four EL models are E1(KNN,SVC,RF,GB,DT,NB), E2(KNN,SVC,RF,GB,DT), E3(KNN,SVC,RF,GB) and E4(KNN,SVC,RF). We trained on data set without gender. Overall, E3 gave the best f1 macro score with 0.897 for 100% train data.

We only compare between ensemble learning models and our best performing KNN and SVC models at table 8.3. EL models was not always the best performing model for all data sizes. SVC never outperformed any EL models. Out of six datasize, KNN outperformed EL models for 25% and 50% dataset training. But, EL models outperformed KNN on other four occasions. However, four different EL models were best performing on four occasions. So, we could not determined which EL model would perform best. One good sign for EL models is that they all performed better than KNN for 100% data size.

From 8.2 and 8.6 confusion matrix, we can see E3 performed better on twenties and thirties age group. For eighties, they performed same. For other cases, KNN performed better. SVC performed better on eighties class than both of those.

Overall, EL models results are satisfactory. It was always very close or better than KNN. However, we need more experiments to dig deeper to find it's potential.

8.3 Hubert

We have failed to run hubert due to gpu. Though we were able to create custom database for hugging face and extract features using wav2vec2, but could not run hubert properly. Even with minimum settings, gpu is overflowing in google colab. We tried to decrease batch size and epoch, still no avail. We could run for very small 3% of data for couple of times but gpu availability is very limited. In this situation, it was impossible to experiment with Hubert considering our limited time frame.

CHAPTER 9

Conclusions

We have tried to use three types ML approaches to recognize age from voice and also tried to find influence of gender and training data size in recognizing age.

Firstly, we used traditional ML models like KNN, SVC, RF, GB, DT and NB. Among those, KNN and SVC performed quite well with 0.89 and 0.88 f1 macro score respectively. Using those models, we also found out influence of gender for age detection. There is significant impact of gender in detecting age. So, if possible we should provide model with gender data. However, for KNN and SVC, we have found out influence of gender mitigate with datasize. In other word, if we have good amount of data, performance of KNN and SVC will be similar regardless of gender present or not. We also found that data size has a positive impact on f1 score. F1 score proportionally increased with data size.

Our second approach for age recognition, is ensemble learning. Result of EL models are mixed to positive. Most of the cases it performed similar to our best performing traditional model KNN and always performed better than our second best performing model SVC. However, performance of EL models were not consistantly better than KNN. We also could not find any EL model combination which always performed better than other EL models. However, when trained with whole data set, all EL models performed better than KNN.

Our last approach was recognising age with pretrained model Hubert using Huggingface. The advantage of those models is that we do not have manually extract feature ourselves. Traditionally, in many fields pretrained models are performing state of art in many tasks. However, we could not conduct this experiment due to lack of gpu processing power. Even

with as small as 3% training data, gpu of google colab could not take the load. In future, it would be great extension of this research, if we could recognise age using LLMs and test it's potential.

APPENDIX A

Data and Code Availability

Data: https://www.kaggle.com/datasets/mozillaorg/common-voice

Modified Data for Google Colab: https://drive.google.com/drive/folders/1M3RFg5BEI-E
usp=share_link

Code Link: https://drive.google.com/drive/folders/1ghI-tlaYp0-Mrk6WEPCnAM6ZtF
usp=share_link

APPENDIX B

Resources utilized

Audio processing and MFCCs feature extractor library: https://librosa.org

Machine learning library: https://scikit-learn.org/stable/

LLM library: https://huggingface.co

Bibliography

[1] Accuracy vs f1 score. https://www.v7labs.com/blog/f1-score-guide#:~:text=The%20macro%2Daveraged%20F1%20score%20is%20useful%20only%20when%20the,be%20a%20misleading%20performance%20metric. Accessed: 2023-07-30.

[2] Google colab. https://colab.research.google.com/?utm_source=scs-index. Accessed: 2023-07-30.

[3] Hugging face. https://huggingface.co. Accessed: 2023-07-30.

[4] Kaggle: Common voice. https://www.kaggle.com/datasets/mozillaorg/common-voice?datasetId=5793&sortBy=voteCount&sort=votes&select=README.txt. Accessed: 2023-07-30.

[5] Kaggle notebook. https://www.kaggle.com/docs/notebooks. Accessed: 2023-07-30.

[6] Mel frequency cepstral coefficient (mfcc) tutorial. http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/#computing-the-mel-filterbank. Accessed: 2023-07-30.

[7] Mozilla: Common voice. https://commonvoice.mozilla.org/en. Accessed: 2023-07-30.

[8] Spectral bandwidth. https://knowledge.cphnano.com/en/pages/what-is-a-spectral-bandwidth-1#. Accessed: 2023-07-30.

[9] Spectral centroid. https://www.sciencedirect.com/topics/engineering/spectral-centroid. Accessed: 2023-07-30.

[10] Why does your voice change as you age? https://jonathanbgn.com/speech/2020/10/31/emotion-recognition-transfer-learning-wav2vec.html. Accessed: 2023-07-30.

[11] Why does your voice change as you age? https://www.britannica.com/story/why-does-your-voice-change-as-you-age. Accessed: 2023-07-30.

[12] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[13] Tobias Bocklet, Andreas Maier, Josef G Bauer, Felix Burkhardt, and Elmar Noth. Age and gender recognition for telephone applications based on gmm supervectors and support vector machines. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1605–1608. IEEE, 2008.

[14] Jonathan Boigne, Biman Liyanage, and Ted Östrem. Recognizing more emotions with less data using self-supervised transfer learning. *arXiv preprint arXiv:2011.05585*, 2020.

[15] Monique J Boulet and Björn J Oddens. Female voice changes around and after the menopauseâan initial investigation. *Maturitas*, 23(1):15–21, 1996.

[16] Aankit Das, Samarpan Guha, Pawan Kumar Singh, Ali Ahmadian, Norazak Senu, and Ram Sarkar. A hybrid meta-heuristic feature selection method for identification of indian spoken languages from audio signals. *IEEE Access*, 8:181432–181449, 2020.

[17] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[18] Alican Dogan and Derya Birant. A weighted majority voting ensemble approach for classification. In *2019 4th International Conference on Computer Science and Engineering (UBMK)*, pages 1–6. IEEE, 2019.

[19] Fatima K Faek. Objective gender and age recognition from speech sentences. *ARO-The Scientific Journal of Koya University*, 3(2):24–29, 2015.

[20] Sean A Fulop. *Speech spectrum analysis*. Springer Science & Business Media, 2011.

[21] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[22] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

[23] Jia Min Karen Kua, Tharmarajah Thiruvaran, Mohaddeseh Nosratighods, Eliathamby Ambikairajah, and Julien Epps. Investigation of spectral centroid magnitude and frequency for speaker recognition. In *Odyssey*, page 7, 2010.

[24] Florian Metze, Jitendra Ajmera, Roman Englert, Udo Bub, Felix Burkhardt, Joachim Stegmann, Christian Muller, Richard Huber, Bernt Andrassy, Josef G Bauer, et al. Comparison of four approaches to age and gender recognition for telephone applications. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–1089. IEEE, 2007.

[25] Stefan Mlot, Emily Buss, and Joseph W Hall III. Spectral integration and bandwidth effects on speech recognition in school-aged children and adults. *Ear and Hearing*, 31(1):56, 2010.

[26] Paul H Ptacek and Eric K Sander. Age recognition from voice. *Journal of speech and hearing Research*, 9(2):273–277, 1966.

[27] P Sandhya, V Spoorthy, Shashidhar G Koolagudi, and NV Sobhana. Spectral features for emotional speaker recognition. In *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, pages 1–6. IEEE, 2020.

[28] Robert Thayer Sataloff, Karen M Kost, and Sue Ellen Linville. The effects of age on the voice. *Clinical Assessment of Voice, 2nd ed.; Sataloff, RT, Ed*, pages 221–240, 2017.

[29] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

[30] Thomas Shipp and Harry Hollien. Perception of the aging male voice. *Journal of Speech and Hearing Research*, 12(4):703–710, 1969.

[31] Melissa N Stolar, Margaret Lech, Shannon J Stolar, and Nicholas B Allen. Detection of adolescent depression from speech using optimised spectral roll-off parameters. *Biomedical Journal*, 2:10, 2018.

[32] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

[33] Syed Rohit Zaman, Dipan Sadekeen, M Aqib Alfaz, and Rifat Shahriyar. One source to detect them all: gender, age, and emotion detection from voice. In *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 338–343. IEEE, 2021.