Personal Information Extraction From Bangla Speech Signal Using MFCC and GMM

by

Maisha Munawara Hridy
14101037
Md. Hasib Hasan
14101033
Mahfuz Al Emon
14101007

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
BRAC University
August 2019

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at BRAC University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

____

Maisha Munawara Hridy
14101037

____

Md. Hasib Hasan
14101033

____

Mahfuz Al Emon
14101007

Approval

The thesis/project titled "Personal Information From Bangla Speech Signal Using MFCC and GMM" submitted by

1. Maisha Munawara Hridy (14101037)
2. Md. Hasib Hasan (14101033)
3. Mahfuz Al Emon (14101007)

Of Summer, 2019 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on August 28, 2019.

Examining Committee:

Supervisor:
(Member)

Jia Uddin, PhD
Associate Professor and Undergraduate Coordinator
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Mahbubul Alam Majumdar, PhD
Professor and Chairperson
Department of Computer Science and Engineering
BRAC University

Abstract

Our system extracts personal information from bangla speech. Dataset that was used consists real-life voice inputs from different age and gender groups. A set of Bengali speech samples from YouTube were used as input dataset. This system is based on basic machine learning algorithms. Mel frequency cepstral coefficient was used to train and construct this system. While calculating gender and age detection part, we will be using GMM to calculate the final scores on the samples having the MFCCs of the extracted speech samples. GMM model basically congregates some subsets among the whole set based on probability. Along with the gender determination process, age detection process will also be simulated using fundamental frequency of speech. Python is the programming language used to write the coding. Our system was successful in giving 88% accuracy for gender recognition and 75% accuracy for age detection.

Keywords: Machine Learning; Mel Frequency Cepstral Coefficient; Gaussian Mixture Model; Natural Language Processing; Python; Bangla

Dedication

We would like to dedicate this thesis to our loving parents, friends and everyone that helped us with this paper.

Acknowledgement

We would like to thank Dr. Jia Uddin for agreeing to supervise us with our thesis. His patience and confidence in us have been a source of encouragement and this thesis would not have been possible without the great support and influence of him. We believe his dedication to this paper deserves to be reciprocated with great gratitude. A special thanks to the Thesis committee for taking the time to review and evaluate our thesis as part of our undergraduate program.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

ASR Automatic Speech Recognition

CART Classification and Regression Trees

GMM Gaussian Mixture Model

GRNN Generalized Regression Neural Network

MFCC Mel Frequency Cepstral Coefficient

NMF Nonnegative Matrix Factorization

PLP Perceptual Linear Predictive

RANSAC Random Sample Consensus

ReLU Rectifier Linear Unit

SDC Shifted Delta Cepstral

SVM Support Vector Machine

UBM Universal Background Model

WSNMF Weighted Semi-NMF

Chapter 1

Introduction

In the present era, technology has won over manual human work in different complex situations. In this case, machine learning has become a very important part for improving human-machine interaction. Voice and speech recognition are two fundamental parts of biometric data. To extract personal information from text, many AI based processes have been improvised to some extent which can be overpassed in terms of cost and user friendliness of this machine learning based system. In this project, we will be trying to identify age and gender of the speakers from bangle language using voice processing. Recently, biometric features of a human are employed in various identification, interactive and security applications. This work focuses on age and gender recognitions by using user's utterances. Such recognition systems can be applied to interactive phones, interactive advertising devices, and digital signatures and so on. In everyday life many important purposes can be served by this system such as, gender specialized advertisements in different media. By identifying user's voice, it can predict his/her age group and gender and avoid showing irrelevant advertisements [14]. Also, while investigating criminal cases, this system can help in huge range. Using suspect's voice, it can shortlist from the preset suspects list in the server which will make the investigation process quicker and narrowing it down to get the result. Beside these official uses, it can be used in recreation purposes too. If this system can be used in online media players, it can detect the age and gender of the listener and can suggest songs more suitable and liked by them. Apparently, that will increase the customer satisfaction and acceptance of the software and thus it can make more profit. Moreover, we can use this system in statistical and economical purposes too. By keeping real life data, surveys can be held in particular sector for particular reason which will remove the hazard of surveying manually and will save a huge time. In order to avoid providing sensitive and risky personal information, institutions may opt to use speech recognition to authenticate identities of their clients. This has helped to curb fraud and phone crimes by use of voice biometrics in certain institutions like banks and other government institutions. To talk about social sectors, some social experiments or surveys can be directed by the Government where age or gender group is recorded from accessing particular voice and speech datasets controlled by the authority. The main benefit of this system is cost-effective as it is made by basic human-machine algorithm. So, our system can be used in huge range. Considering cost efficiency, it can be introduced vastly in commercial grounds. For being user friendly of the system, it will reach in different sectors of regular and high official users on different important purposes. Like other biometric data a lot of information can be extracted from voice. However, our core objective here is not to just generate useless data but to primarily focus on the concept of machine learning. For machine learning voice analyzing is just a variable. The key thing here is to highlight that providing different types of datasets (in this case voice samples) we can teach a computer program to expand itself and calculate almost anything based on the given data. Moreover, the program will improve its results accuracy by learning itself while we keep feeding data. So, gradually it will give more accurate and realistic outputs after some sessions of training.

1.1 Thesis Contribution

In most of the cases, English is vastly used language to be researched in Natural Language Processing field. We know that, Bangla language has approximately 260-300 million speakers. However, it is rare to find language processing projects using Bangla language. Whereas, Persian, Arabic, French, Hebrew and many other languages are not let behind to be researched on. On the other hand, in spite of being one of the popular languages, Bangla is left behind in this matter while other technological parts are going up in the curve. So, in this paper we used machine learning to detect age and gender using Bangla language database. We applied MFCC and GMM method together on our dataset to get output more accurately and in simpler manner from other existing methodologies. We trained our system by ourselves using online and offline voice data collected from different kind of sources and persons. This data feeding made our module more efficient. Our proposed method is more robust than using other Artificial Neural Network, Hidden Markov and so on. Moreover, our proposed method works faster and gives more accurate result with even a dataset that contains a lower number of speakers.

1.2 Thesis Orientation

In chapter 2, some approaches which have done by different authors related to this research field has been discussed. While running our primary background study, we have come across some published papers based on Natural Language Processing on different languages using different kinds of methods as well. We have discussed about some projects which used methods such- CART, SVM, SDC, WSNMF, GRNN, HMM based etc. Moreover, the accuracy level gained by each module were also compared in along with mentioning datasets that had been used in each experiment. In chapter 3, the detailed procedure and workflow of our proposed model have been illustrated. Firstly, we mentioned about some Python dependencies which have been some main libraries while running the algorithm. In other sub section, the detailed description of our own dataset that has been used in our proposed method has been brought up. Along with that, how the dataset will be manipulated through MFCC and GMM method to get the required output, has been demonstrated. To make the total procedure more illustrated and to show the clear picture of the process in front of readers, we have added some graphical and diagram-based work flow which can give a gist of the huge elaborated work. Lastly, how the main methods- MFCC and GMM methods, Python speech recognition engine have been demonstrated for making this module , has been given briefly with suitable figures and that too part by part sequentially which resulted a good workflow presentation. Finally, chapter 4 gives the results after successfully running and testing the module. For showing the comparative results, we have added two tables containing Gender prediction output and other one containing Age prediction output. Followed by that, the percentage calculation of required result has been shown in details. It showed approximately 88% accuracy in gender recognition and 75% accuracy in age recognition. Chapter 5, being the concluding chapter explains the future works and research scopes of this project. Overall, the general purpose, benefits, outcomes, goals and contributions of our proposed model have been highlighted to conclude it with satisfactory discussion and presentation.

Chapter 2

Literature Review

For extracting Age and gender from voice input, multiple approaches are observed. Proposes the use of CARTs to determine age, age group. Results showed that it worked best than its previous technique giving 72% accuracy for age group. In that techniques, 2048 elicitations of a word “rasa” produced semi-spontaneously in isolation by 428 adults. The male and female speakers aged from 17 years old to 84 years old. Each speaker recorded 3 to 14 elicitations of words. The words were normalized for study. It was quite successful as the gender determined successfully by 51.04% and age determined by 50% [5]. However, another approach was using Mel Frequency Cepstral Coefficient (MFCC). Results gave satisfactory accuracy for speech recognition. Shifted Delta Cepstral (SDC) is an alternative of MFCC. SDC is more robust under noisy Data. But if the recording does not contain much noise then MFCC gives better results than SDC. Here ELSDSR dataset is used. 22 speakers in which 12 male and 10 females was used to make ELSDSR. Their age was from 24 to 63. There was no priority control distribution by nationality and age. The group had relatively small variation in profession. High-quality microphone was used for recording the dataset. Deflection boards were kept at angles facing each other, and were set up in front of the table and speakers. Age and gender recognition are done based on pitch information then MFCC then a mean matrix was formed with the two different scores [15]. This fused model gave an accuracy of 64.20%. Dataset’s limitation was the main obstacle here. In another project, a supervised machine learning algorithm called SVM was used for the final recognition where accuracy is 64.20% and out of the 4 classes, 3 classes have higher recognition values. Finally, for classification of age and gender Gaussian Mixture Method is more robust than some other classifiers like Artificial Neural Network, Hidden Markov Model [11]. 6 age groups of various Persian speaker people were used. For implementation and extraction Perceptual Linear Predictive (PLP) and Mel-Frequency Cepstral Coefficients (MFCC) were used and for classification SVM was used. Dataset used was FARSDAT dataset; a Persian Speech Database. Farsdat is made with the recordings of 300 Iranian speakers. Another notation used a Gender database, acoustic and prosodic level information fusion [13]. Again, in another article proposes a new approach for speaker gender detection and age estimation. It is based on a hybrid architecture of WSNMF and GRNN the accuracy of the gender detection method was 96% and the corpora was consisting of 555 speakers [10]. A separate paper used super vectors of GMM means and variances and combining these features to determine age. There were two corpora used, gender corpus and Hebrew Corpus; succeeded in classifying the child and senior speakers while the young and adult groups had poor classification results [9]. Moving on to the next paper where age detection was done using conversational telephone speech using senone posterior based i-vectors [22]. More works have been done to determine age and gender detection using several methods such as HMM-based feature extraction method and tried to approach acoustic feature for alcohol detection of the speaker [8], [12]. Another study uses decision level fusion and ensemble-based techniques [6], linear discriminant techniques [7], combining regression and classification method and also min-max modular support vector machine [2]. In another paper used VAD method to make GFCC front-end extraction more robust. They built their own dataset consisting of 37 Arabic speakers. Where 21 are male and another 16 are female. The recording has been done by each speaker. Their entire identification system is implemented in MATLAB. To make the identification phase slower they used 72 parameters which make their memory more utilized. Their average accuracy rate combining with MFCC and GFCC are 65.97% and 66.83% [23]. Another article based on MEC 2017 corpus. They used two baseline system one is random forest algorithm, another is DNN with single task for emotion classification. They trained their DNN with SGD optimizer. Comparing their baseline system, they showed that DNN out performance the random forest by 4.1%. At the very end by using best baseline system (single task DNN) they improve their system by 7.8% absolute (29.5% relative improvement) [27]. In the next paper showed gender classification by using low level features to 256 sample windows. Their male sample has been resized to looking forward the mismatching sample sizes over two classes. Their linear kernel SMO achieves 67.3% classification accuracy consistent [4]. Another article was based on TREC 2003 video retrieval data set. The ratio of their proposed models' sample (female and male speaker is 3:4) and fully evaluate on the bootstrapping performance. All of their classification accuracy results are calculated by running the trained classifier over their data collection [3]. The next article explains their proposed model is based on RANSAC classifier. They used 200 male and 200 female voice sample and implement the dataset on MATLAB. They showed RANSAC gives better result compared to neural-network [16].

Chapter 3

Proposed Model

3.1 Workflow

For gender recognition, the system is divided into two phases: Training phase and Testing Phase. Figure 3.1 and 3.2 demonstrate both the phases

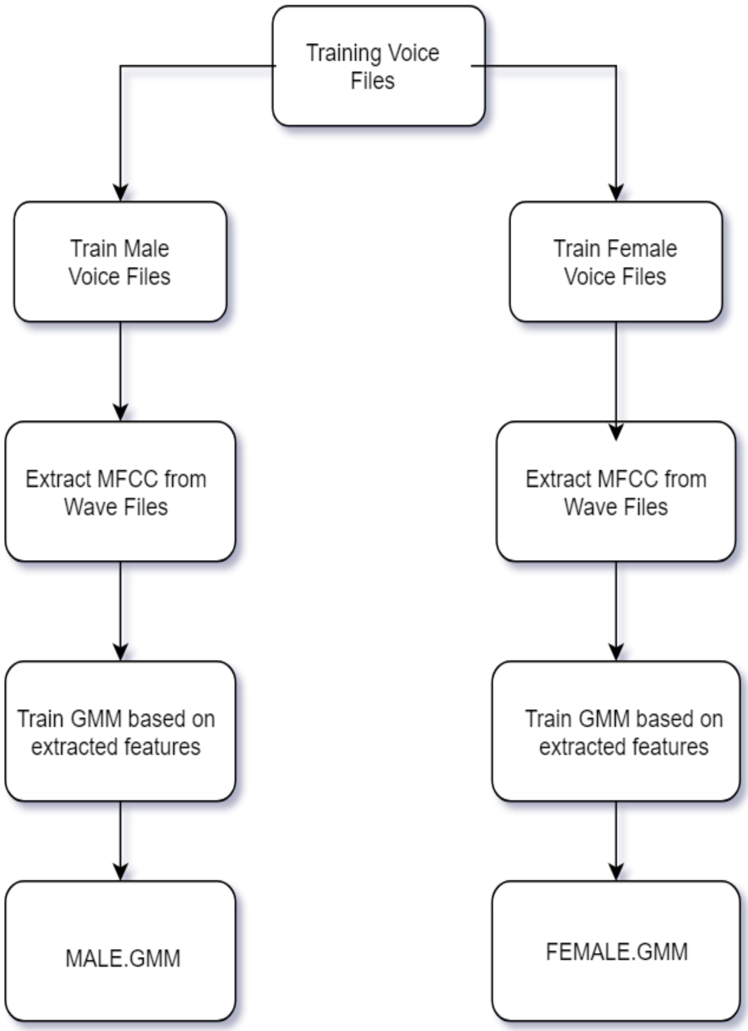

Training Voice Files
Train Male Voice Files
Train Female Voice Files
Extract MFCC from Wave Files
Extract MFCC from Wave Files
Train GMM based on extracted features
Train GMM based on extracted features
MALE.GMM
FEMALE.GMM

Figure 3.1: Training Phase of Gender Determination

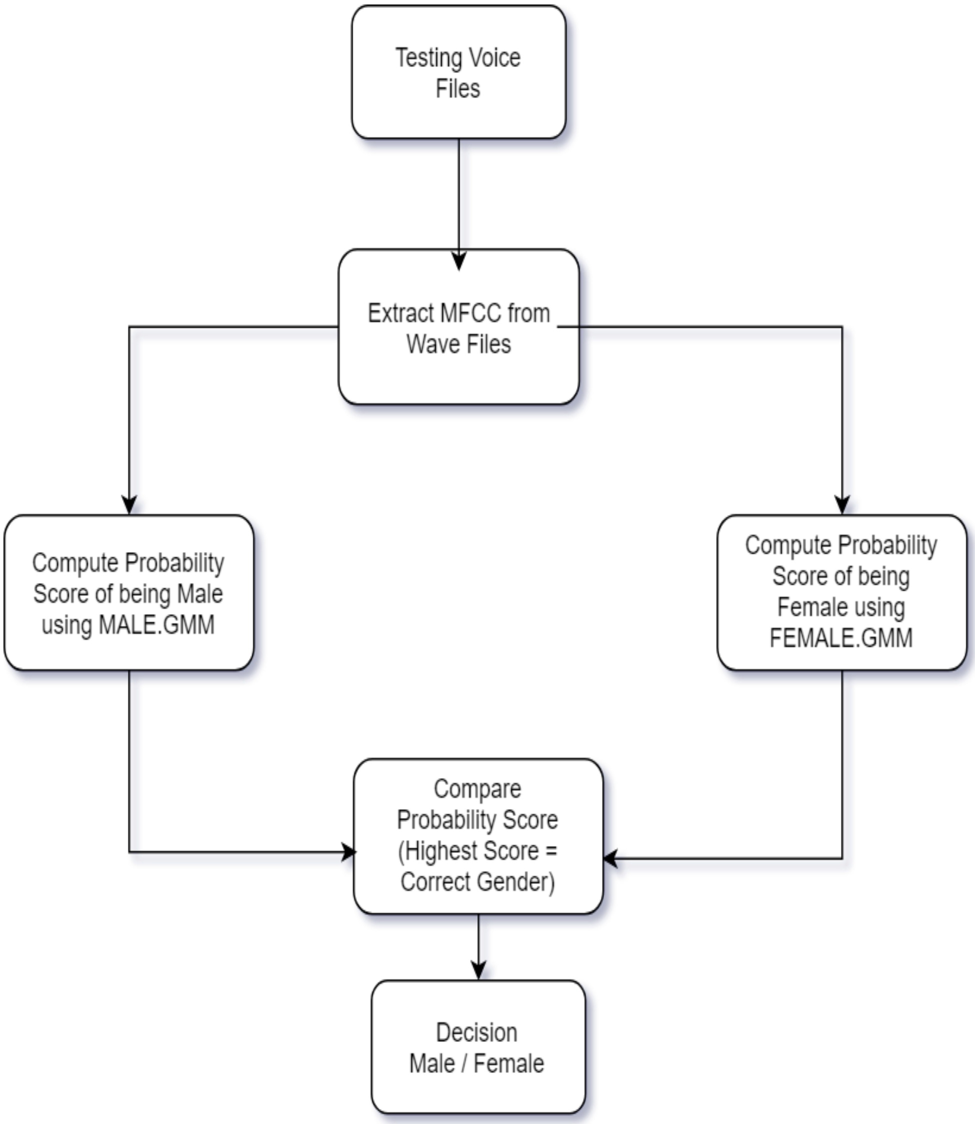

Testing Voice Files
Extract MFCC from Wave Files
Compute Probability Score of being Male using MALE.GMM
Compute Probability Score of being Female using FEMALE.GMM
Compare Probability Score (Highest Score = Correct Gender)
Decision Male / Female

Figure 3.2: Testing Phase of Gender Determination

Next, Figure 3.3 will illustrate the workflow of age determination of our system. After the gender recognition process, the age determination process will start.

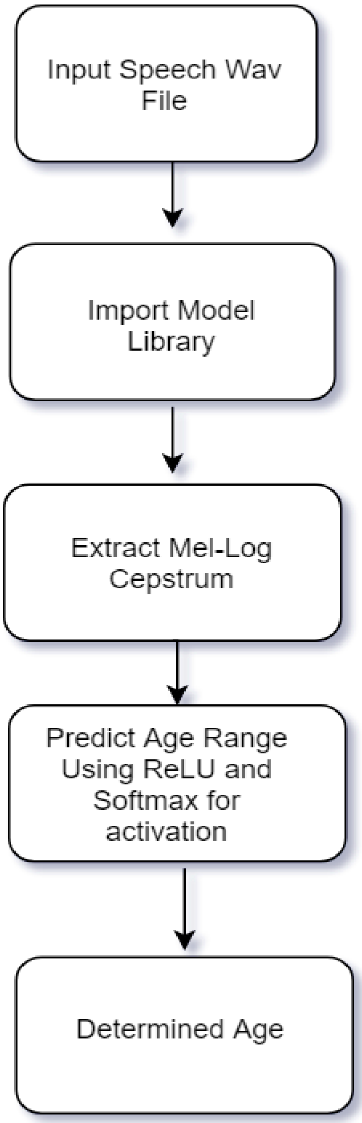


Figure 3.3: Workflow of Age Detection

3.2 Overview

The basic theories and concepts used in here are explained to get a better prospect of each stage of the whole program.
Firstly, the whole program is structured with Python. Is is a general purpose yet highly sophisticated and interpreted programming language. Modules and libraries are the core of any programming languages. The lines of codes are simply to utilize these resources to process the input data analyze it and show the result. Here the following python dependencies are used.

NumPy is the neccesary library for fundamental computation and mathematical operations. It is vastly used for creating arrays. Then Scipy is another library that use Numpy and use it as data structure and scientific programing. Moreover Sklearn is also known as sci-kit learn. It is the most important machine learning library for python. It uses NumPy, SciPy for its' operation. Furthermore, Python Speech Feature is the essential library for extracting MFCC.Lastly the Age Range

Classifier is library is used to determine age. The h5 format is the data file stored in Hierarchical Data Format which contains multidimensional array of age data.

3.3 Dataset

For our research we came across a number of Bangla Language dataset. Maximum of those were not available for free and some of them were author permission only. Bengali Speech Data – ASR has 1000 speaker but download was not permitted [26]. Finally, we managed a corpus called "Shruti Bangla Speech Corpora " which is made by Society for Natural Language Technology Research [28]. In the dataset male speakers are 75% and female speakers are 25%. That is, 34 speaker age varies in between 20 to 40 years were used. From them, 26 was male speakers and 8 was female speakers. They were chosen from West Bengal, India. All the speaker lived there during their childhood. Each audio file is in WAV format and there are 7383 sentences in total. The topic that are spoken in the dataset are sports, genaral news, geographical news, politics etc. Sentences used in the dataset was designed at IITKGP. Anandabazar magazine, multiple major news domain articles. These were collected to cover Bengali language phonetic variations . Moreover , generally regularly used sentences were collected and recorded. There were two recording session and phonetic compact sentences was used so that most commonly spoken words can be covered in the dataset.

3.4 Framing

voice/speech is highly non-stationary signal. For analyzing the speech signal we need to convert it to stationary signal.The speech signal is to be divided into short frames. The duration has to be around 20 to 30 milliseconds.Figure 3.4 and 3.5 illuminates the non-stationary and stationary signal accordingly[20], [21].

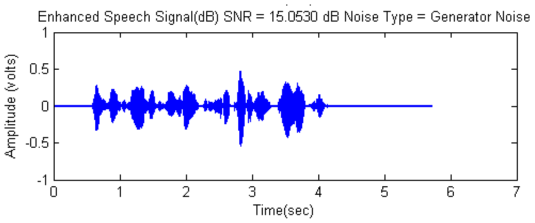


Figure 3.4: Non-Stationary Signal

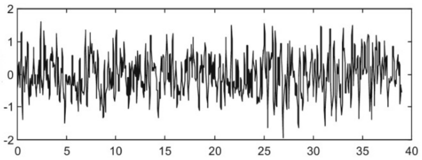


Figure 3.5: Stationary Signal

3.5 Windowing

Extracting raw frames from a speech can sometimes result non integer number of periods in the extracted waveform. That may lead to an incorrect frequency representation. Multiplying speech frame with window function can solve the problem. Figure 3.6 shows windowing a signal [29].

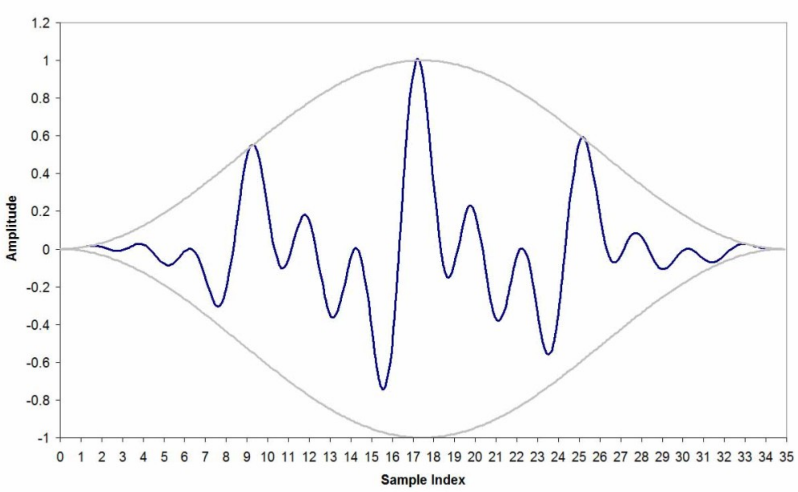


Figure 3.6: Windowing Signal

3.6 Overlapping

Samples are lost towards the start and end of the frame due to windowing, and this will lead to a wrong frequency representation. Samples lost from the end of the nth frame and the beginning of the (n+1)th frame are included in the frame formed by the overlapping.

3.7 MFCC Feature Extraction

Mel-frequency cepstral coefficients (MFCCs) have frequently been used in speech recognition. MFCC is the variation of the human ear's critical bandwidths with frequency. The MFCC technique makes use of two types of filter, namely, linearly spaced filters and logarithmically spaced filters [17]. MFCC calculation can be divided into following steps.

1. The time domain speech signal is to be converted into spectral domain signal using Fourier transform.
2. Take the logarithm of the value so the source and filter are additive.
3. Now we convert it to Mel spectrum.
4. We take the discrete cosine transform (DCT).

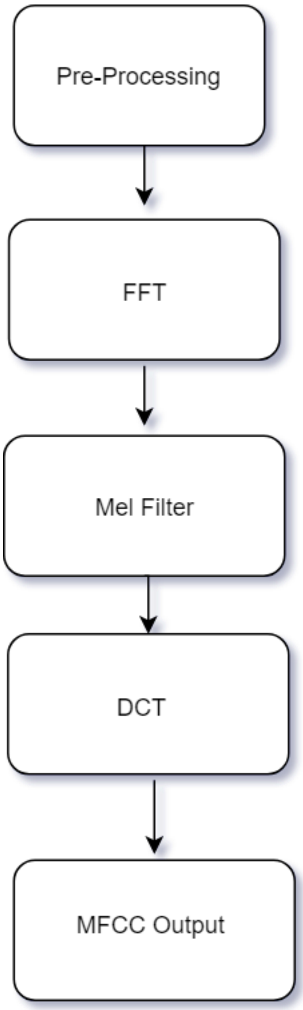
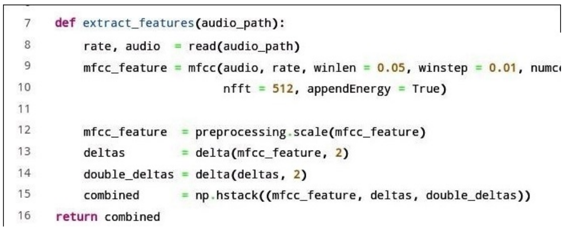


Figure 3.7: MFCC Feature Extraction Process

Figure 3.7 depicts the process of calculation of MFCC. In general, the spectrums obtained from the voice samples are measured with triangular overlapping signals. And these MFCCs are the amplitudes of the resulting signals of the dataset which are going to be fed in the Gaussian Mixture Model trainer. In Figure 3.8 the commands are shown which was used in python to direct the above process in which MFCC is used.



```
def extract_features(audio_path):
    rate, audio  = read(audio_path)
    mfcc_feature = mfcc(audio, rate, winlen = 0.05, winstep = 0.01, numc
                        nfft = 512, appendEnergy = True)

    mfcc_feature  = preprocessing.scale(mfcc_feature)
    deltas        = delta(mfcc_feature, 2)
    double_deltas = delta(deltas, 2)
    combined      = np.hstack((mfcc_feature, deltas, double_deltas))
return combined
```

Figure 3.8: MFCC Python Code

3.8 Model Training Using GMM

Gaussian Mixture Model or GMM is a function used to distribute probabilities of systems where measurements are continuous. For instance, in biometric features which in this case is features extracted from voice/audios. GMM works using EM algorithm. EM algorithm means Expectation-Maximization algorithm for Gaussian mixtures [18]. The algorithm is an iterative algorithm that starts from some initial estimate of a random value and then proceeds to iteratively update that value until convergence is detected. Each iteration consists of an E-step and an M-step The probability of data points here is calculated by GMM with the following equation-

$$p(x) = \Sigma_z p(x|z)p(z) = \Sigma_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \tag{3.1}$$

The Equation 3.1 is used to analyze Gaussian density of multiple samples to generate the most probable result [24]. And mixture of these probable samples output the most likely result which is the actual Gaussian Mixture Model.

3.9 Testing Phase

Testing phase is where the actual program runs. As shown in the overview flowchart, this process depends on the fruits of the training phase. The better the training process was, the more accurate the results are. For instance, more voice samples in the dataset as well as optimized audio files means the program can determine the result recognizing more voice pitch variations. The training wheel program also makes it faster to calculate the result, otherwise the testing program has to go through all the datasets and take hours to output the result. But since the training part already utilized those hours before to generate ready to go GMM files, the testing program only has to analyze a short amount of data.

3.10 Gender Determination

The gender identification is done over three steps: 1) Extract Mel-Frequency from the whole dataset just like training phase does for the train dataset. 2) Now, the program computes the probability of speaker being a male based on the male.gmm files, likewise it also computes the probability score of being a female voice based on the female.gmm files. Needless to say these GMM sampled were the results of the training program. So now the program has two values. 3) Lastly, the program compared the two scores of probability with each other, and displays the result by prioritizing the higher probability score. At this part we are done with the gender determining part and a command line embedded at this part of the script automatically executes the age determining python script.
The result calculation is as follows.

Given a speech K and speaker S, the gender recognition test can be restated into a basic hypothesis test between HmHm and HfHf, where:
- HmHm : K is a MALE

- HfHf : K is a FEMALE

p(K—Hm)
p(K—Hf)
If result is ¿=1 then Hm accept
If result is ¡1 then Hf accept

where p(K—Hi) is the probability density function for the hypothesis Hi evaluated for the observed speech segment K. [1]

3.11 Age Detection

The system execute command to run age determination.py Script. For determining the age unlike gender recognition, there is no separate training session. The training will be done real time when the program will run. First the wav file will be taken as input. Rectifier Linear Unit will predict the age range and SoftMax will be used to determine age probability. Python speech engine is used to compare the data with built in data which are from the dataset TIDIGITS and TIMIT. And Finally, the program will show the determined age . It is much simpler as it only focuses on the frequency of the speaker's voice. For example, if the speaker's voice frequency is similar to people of 20-25, it will calculate the average to be 23 years old. We calculate the age with numerical percentage and find the success rate of age detection.

Chapter 4

Experimental Results and Analysis

4.1 Gender Determination

The result of gender determination can be summarized in Table 1 which is illustrating in confusion matrix From Table-4.1 we can measure the following characteristic.

Table 4.1: Gender Determination Matrix

Gender	Male	Female
Male	23	3
Female	1	7

Precision for male recognition = 23/(23+3)=0.885
Precision for female recognition = 7/(1+7)=0.875
Accuracy = 0.88

Table 4.2: Gender Determination Result

Gender	Success Rate
Male	88.5%
Female	87.5%
Total	88%

For only male gender, 23 out of 26 were determined giving 88.5% success rate. For female gender 7 out of 8 runs were determined giving 87.5% success rate. Overall success rate based on these is 88% approximately.

4.2 Age Detection

Table 4.3: Age Detection Result

Age Group	Original Age	Determined Age	Success Rate
20-30	27	21	77.7%
30-40	7	5	71.4%

According to the result of the above table the overall success rate of age detection is approximately 75%. The age group of the dataset used was not equally divided and maximum of speakers were between 20-30 years. Among the 34 speakers, 30 of them was Graduated. So, the age group most of them are near 30 years. Moreover, the result of age detection our system shows +-3 years age. Thus, the speakers whose age are in around 30 years are not accurately determined always. We wish to work with another dataset in which we can get more wide range of age group speaker so our result can get better.

Chapter 5

Conclusions and Future Works

5.1 Conclusions

In this paper we propose a framework that can extract gender and age information from Bengali voice and speech using Machine Learning. After discussing all above, we see that, two features- Age and Gender prediction can be fairly done through our system. Though this system can be upgraded to some elaborated extends such as- personality, local area of speaker prediction and so on. We used MFCC and GMM to identify gender and used ReLU, SoftMax and Pythons' built in library to detect age group. So, we can say that this is a progressive system that too using Machine Learning process which makes it more acceptable. As we know, in today's world, biometric data are mostly used features for getting any experimental outputs in different important sectors. For this, user end prefers more user-friendly and cost-effective system to run the whole experimental process. In this case, this system can take a great place in the market for mass level use. Moreover, it will create scope in innovating and improvising more and getting into the depth of the "Machine Learning" research field. Lastly, there have been many systems already been developing on English and some other languages, but in that comparison, research on Bengali language processing is not that popular like those. Consequently, working on this framework is a bit challenging because of the limited research and resources. So, successful run of this system, can create a new milestone in this region for the native language speakers. As most of the countries having Bengali as native language is developing countries, this process can be a new milestone in their technological as well as commercial sector. Overall, this system can be very useful and user-friendly while leading any biometric data analysis or experiment to get satisfactory results.

5.2 Future Works

We are currently working on only Age and Gender of the speaker. In future, we plan to add some other features in the system like detecting local bangla language. Since there are many kinds of local languages within many different districts in Bangladesh. There are also many tribal languages in the hills area of this country. So, we are planning to improve the system which will detect those different local bangla languages and the origins of the languages by following the voice of the speaker. We would like to make our own dataset using Bangladeshi people from major districts. However, we can also train our system with other dierent datasets to detect bangla speech recognition or command. For instance, we can train our system to recognize criminal's voice from database. On the other hand, our future system model should be based on neural network which is online based [19], [25]. It can utilize the google voice API engine, consisting of a large number of data updated which will result in more accuracy. Finally, a larger dataset is required to get more accuracy. But it is rare to find a larger bangla language dataset with more speakers. We are yet to have a complete bangla language dataset consists of larger number of male and female speakers. We aim to make our own bangla language dataset which will contain not only different age limit speakers but also speakers of different local language of our country to make the dataset more versatile and robust which will help further experiments and programs. Working with bangla language processing is a challenging work in a way because of low resources and guidelines on the internet. So, we had to work with less reliable voice sample that we found. There is no luxury of choosing between dataset and comparing those as we only found one bangla dataset to work with. We aim to share our dataset for free for educational and experimental uses.

Bibliography

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models”, *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[2] H.-C. Lian and B.-L. Lu, “Age estimation using a min-max modular support vector machine”, in *Twelfth International Conference on Neural Information Processing*, 2005, pp. 83–88.

[3] G. Tzanetakis, “Audio-based gender identification using bootstrapping”, in *PACRIM. 2005 IEEE Pacific Rim Conference on Communications, Computers and signal Processing, 2005.*, IEEE, 2005, pp. 432–433.

[4] P. Dutta and A. Haubold, “Audio-based classification of speaker characteristics”, in *2009 IEEE International Conference on Multimedia and Expo*, IEEE, 2009, pp. 422–425.

[5] S. Schötz, “Automatic estimation of speaker age using cart”, *Working Papers in Linguistics*, vol. 51, pp. 155–168, 2009.

[6] J. Bekios-Calfa, J. M. Buenaposada, and L. Baumela, “Revisiting linear discriminant techniques in gender recognition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 858–864, 2010.

[7] C. van Heerden, E. Barnard, M. Davel, C. van der Walt, E. van Dyk, M. Feld, and C. Müller, “Combining regression and classification methods for improving automatic speaker age recognition”, in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2010, pp. 5174–5177.

[8] F. Lingenfelser, J. Wagner, T. Vogt, J. Kim, and E. André, “Age and gender classification from speech using decision level fusion and ensemble based techniques”, in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[9] R. Porat, D. Lange, and Y. Zigel, “Age recognition based on speech signals using weights supervector”, in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[10] M. H. Bahari and H. Van Hamme, “Speaker age estimation and gender detection based on supervised non-negative matrix factorization”, in *2011 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS)*, IEEE, 2011, pp. 1–6.

[11] D. Mahmoodi, H. Marvi, M. Taghizadeh, A. Soleimani, F. Razzazi, and M. Mahmoodi, “Age estimation based on speech features and support vector machine”, in *2011 3rd Computer Science and Electronic Engineering Conference (CEEC)*, IEEE, 2011, pp. 60–64.

[12] R. Gajšek, F. Mihelič, and S. Dobrišek, "Speaker state recognition using an hmm-based feature extraction method", *Computer Speech & Language*, vol. 27, no. 1, pp. 135–150, 2013.

[13] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion", *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, 2013.

[14] S. E. Shepstone, Z.-H. Tan, and S. H. Jensen, "Audio-based age and gender identification to enhance the recommendation of tv content", *IEEE Transactions on Consumer Electronics*, vol. 59, no. 3, pp. 721–729, 2013.

[15] H. Erokyar, "Age and gender recognition for speech applications based on support vector machines", 2014.

[16] A. Ghosal and S. Dutta, "Automatic male-female voice discrimination", in *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, IEEE, 2014, pp. 731–735.

[17] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to dnn-based children's and adults' speech recognition", in *2014 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2014, pp. 135–140.

[18] R. Sridharan, "Gaussian mixture models and the em algorithm", *Avilable in: http://people. csail. mit. edu/rameshvs/content/gmm-em. pdf*, 2014.

[19] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks", in *Proceedings of the iEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 34–42.

[20] J. M. Giron-Sierra, *Digital Signal Processing with Matlab Examples, Volume 1: Signals and Data, Filtering, Non-stationary Signals, Modulation.* Springer, 2016.

[21] T. Omotayo, L. Afolabi, F. Ehiagwina, and O. Adewunmi, "Development of spectral subtraction algorithm for enhancement of noisy speech signal of electricity generator", *World Wide Journal of Multidisciplinary Research and Development*, vol. 2, pp. 7–14, Aug. 2016.

[22] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos, "Speaker age estimation on conversational telephone speech using senone posterior based i-vectors", in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, pp. 5040–5044.

[23] E. B. Tazi, "A robust speaker identification system based on the combination of gfcc and mfcc methods", in *2016 5th International Conference on Multimedia Computing and Systems (ICMCS)*, IEEE, 2016, pp. 54–58.

[24] M. Słoński, "Gaussian mixture model for time series-based structural damage detection", *Computer Assisted Methods in Engineering and Science*, vol. 19, no. 4, pp. 331–338, 2017.

[25] Z.-Q. Wang and I. Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks", in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 5150–5154.

[26] *Bengali speech data asr*, Jun. 2018. [Online]. Available: http://tdil-dc.in/index.php?option=com_download&task=showresourceDetails&toolid=2000&lang=en.

[27] F. Tao, G. Liu, and Q. Zhao, "An ensemble framework of voice-based emotion recognition system for films and tv programs", in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 6209–6213.

[28] B. Das, K. Das, P. Mitra, and A. Basu, *Shruti bengali bangla asr speech corpus.* [Online]. Available: http://cse.iitkgp.ac.in/~pabitra/shruti_corpus.html.

[29] Mark, *The fourier transform part viii – windowing.* [Online]. Available: http://www.themobilestudio.net/the-fourier-transform-part-8.

Appendix A

Shruti Corpora Details
Speakers 34
Male 26
Female 8
Sentences 7383
Phoneme 49
Total Words 22012
Duration 21.64 hours