Proceedings of the SMART–2019, IEEE Conference ID: 46866
8th International Conference on System Modeling & Advancement in Research Trends, 22nd–23rd November, 2019
College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, India

A Machine Learning Approach to Automating Bengali Voice Based Gender Classification

S.M. Saiful Islam Badhon[1], Md. Habibur Rahaman[2], and Farea Rehnuma Rupon[3]
[1]*Dept. of CSE, Daffodil International University, Dhaka, Bangladesh*
[2]*Dept. of CSE, Daffodil International University, Dhaka, Bangladesh*
[3]*Dept. of CSE, Daffodil International University, Dhaka, Bangladesh*
E-mail: [1]1saiful15-7878@diu.edu.bd, [2]habibur15-7761@diu.edu.bd, [3]farea15-7707@diu.edu.bd

***Abstract*—Does thicker vocal folds produce sounds with longer wavelength? And can they produce higher pitches to human ears? We address these types of questions and try to identify the difference between male and female voices. By using Machine learning algorithm it's possible to identify the gender from voices. And for that we extract voice signal's MFCCs features by calculating Discrete Fourier Transform, Mel-spaced filter-bank and log filter-bank energies. Identify gender from natural voice can be one of the most important part of voice recognition. In normal voice to text conversion it's not important to detect the voices gender. But when we use this voice recognition for real life applications, it will be densely needed to identify the voices gender. Gender identifying from voice is a field of Natural language processing which is a branch of artificial intelligence. We followed a simple working sequence for getting the ultimate result. The sequence is, Input-audio-file, Pre-works, Feature Extraction, Creating CSV file with features, Train the model and finally test with test data. For feature extraction we used Mel-frequency cepstral coefficient (MFCC). And for mapping and selection we used Logistic Regression, Random Forest and Gradient Boosting. After all this work we get 99.13% accuracy on the dataset that containing 1652 data of more than 250 speakers and tested them with 400 male and 400 female voices.**

Keywords*: *Gender identification, Feature extraction, Voice to gender, Bangla voice gender, MFCCs

I. Introduction

Voice recognition is one of the most mooted topics of NLP. Speech is the medium of communication and human interaction. Speech is created by biological mechanism using several body parts. Human brain can automatically identify the gender difference hearing a person's voice, but computer cannot. A person's gender is essential for the interactions of social community and computers. Recently technology requires automatic gender classification which is nowadays playing a vital role in many ways. Most of these voice detection systems detect voice by reading word sequencing. This research makes a voice recognition application based on wave frequency of a person voice. This application can automatically detect the gender of a human using Bangla language. There are many efficient uses of gender detection through voice. Some of them are described below:

In crime detection, it will be helpful. People commit different types of crimes through phone calls or voice messages. Some- times Criminals hide their identity deliberately. The national security force can surveillance the criminals through this system. Thus, these types of crimes can be solved by categorizing genders through voice [1]. Demographic Investigation can be another use of gender's identifier. A nations demographic or census information can be automatically identified through human voice. Demographic statistical information such as gender, disability status, education status etc. can be collected by this type of application [2]. For Commercial Betterment we can use it. Gender detection is nowadays useful for guiding digital marketing and also smart shopping which creates initiation of new smart websites, online marketing and digital advertising etc. Thus, knowing the number of male and female customers would help building more effective commercial transaction [2]. In a mobile healthcare system or an online healthcare system this application can play a vital role. It would be easier for the healthcare professional to prescribe for the patient more accurately. There are also some vocal folds pathologist which are biased to a specific gender Such as vocal folds cyst which is found only in female patients [3].

II. Related Work

From the beginning of research science, several projects have been invented to recognize gender from a speech. So, detecting gender is not a new work in this decade. Some works have been done to recognize gender from a human voice. A new system Using Bootstrapping on audio classification is introduced in where the system identify gender from speech [4]. It shows more than 90% performance for k Nearest Neighbors, Neural Network, Naive Bayes, Logistic Regression, Decision Trees (C4.5) and Support Vector Machine (SVM) Classifiers.

They tried to present a mixture of Piece wise GMM and neural networks [5]. Which is Content based multimedia indexing segments and every segments duration being 1 second. It outputted 90% accurate result for every language and channel. Another system found which also provide 90% accurate result and they also classified the voice by using multimedia indexing of voices channel [6]. A support vector machine is applied on discriminative weight training in to identify gender [7]. This support vector machine (SVM) consist an optimal weighted Mel frequency cepstral co-efficient (MFCC) which based on MCE (Minimum Classification Error) and generates a gender decision rule. It introduced another method in which provides almost 100% accuracy [8]. A system introduced in where Gaussian Mixture Model is used for two stage classifiers for high accuracy and low complexity [9]. It shows more than 95% accuracy rate. In 1992 Konig and Morgan worked with Linear Prediction coding Coefficients. They extracted 12 LPC using a Multi-layer Perceptron's classifier and energy features in every 500 milliseconds. Based on DARPA resource management database it shows 84% accuracy where the database contains a clean speech of around 160 speakers in English (US). HMM- Hidden Markov Models used to identify gender from a speech where the engine is trained with one Hidden Markov Model speech to recognize each gender. This model is used to decode a signal from test speech. Parries and Carey in 1996 combined pitch and Hidden Markov Model to identify gender from a speech which shows more than 97% accuracy. The experimented on some sentences of 5 seconds from the database of OGI. In 1997, using GMM, Slomkaand Sridharan combined a general audio classifier and pitch-based approach. After remove the silence on OGI and based on 7 seconds speech the system reported 94% accuracy. Using MFCC and GMM as a classifier Tzanetakis and Cook (2002) applied to identify gender in a multimedia indexing context and it shows 74% accuracy. It is seeming that most of the gender detection research works with foreign language. In Bangla language there is not much research have done to recognize gender from a speech. In the there is a system which introduced us to detect gender from Bengali speech which extract their features using Fast Fourier Transform [10]. It also provides a low accuracy around 80%.

III. Methodology

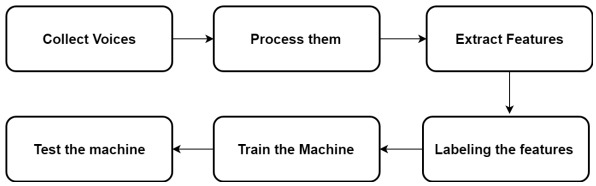

Fig. 1: Workflow of the Work

Bangla videos and through mobile recording system. All the speakers are native speakers of Bangla and they are the citizen of Bangladesh. The students of Daffodil International University helped a lot to collect these data. The average age of those speakers is 20-50. All the data collected from different location of Bangladesh. As speech are collected from call recording, YouTube and many other ways that's why speech is gathered from different location. Another thing is all those speeches is in standard Bangla. Speeches are recorded through a mobile recording application. For this, we didn't use any special room or any kind of special accessories which the data will be smooth or free from noise. Speeches are terminated by the software named Filmora. Additionally, some online platform which provides the option to terminate a large voice such as audiotrimmer.com and mp3cut.net. Most of the voice is in 128kbps and we didn't use any filter over the speech. All the speeches are in mostly 4-7 seconds. And we collected exactly 1652 (female voice=821 and male voice=831) voices from more than 250 people. In fig. 2 we compare the ratio between male and female voices number.

It's possible to classify lots of things from voice using artificial intelligence. And when we start work with natural voices which belongs to natural language processing (NLP) we found that we need to find out the features of voices. And for detecting male voices and female voices it's important to find out those features which can detect the differences between male and female voices. So, we planned for reaching our goal which is given at fig. 1.

First of all, we collected the voices for creating the data set.

A. Dataset

Speeches are collected from different sources. Some of them collected via a google form, audio call recording, YouTube

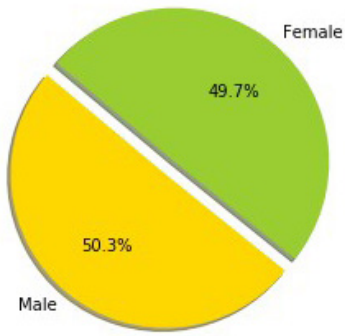

Fig. 2: Ratio of Male and Female Voice

Among 6 steps of fig. 1 features extraction part was most crucial and time consuming. It took almost 90% of our whole work time in data processing and features extraction. With 26 features of a voice we tried to identify male or female voice. Some of them are described below.

B. Feature Extraction

1). Zero Crossing Rate

This is the feature which records the changes of sign in a voice signal according to time. As we know in male voice is broader than female voice. We can find out this in this feature.

In fig 3 and 4 we took male and female voice of same Bengali sentence, though male and female both said the same sentence, there are clear differences found in their voice.

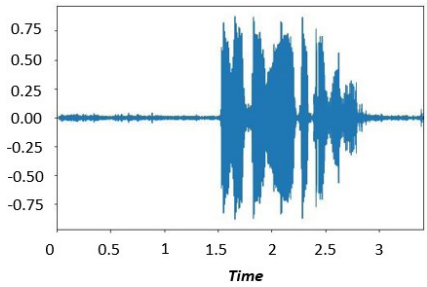

Fig. 3: Zero crossing rate of male voice

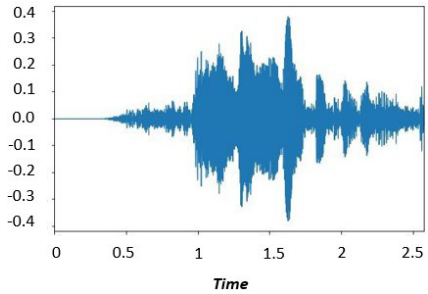

Fig. 4: Zero Crossing Rate of Female Voice

2). Spectral Centroid

This feature finds out the center of mass for a sound. More clearly it is finding out the loudness of a voice. So, depending on the loudness we tried to specify the voices, below figures (fig 5 and 6) will help us to specify the voices with the help of spectral centroid.

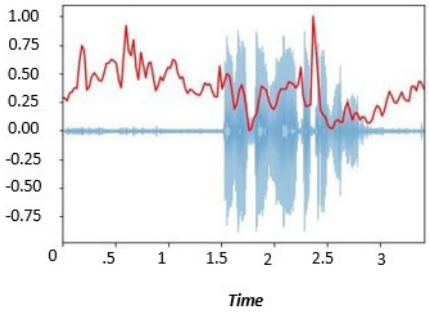

Fig. 5: Spectral Centroid of Male Voice

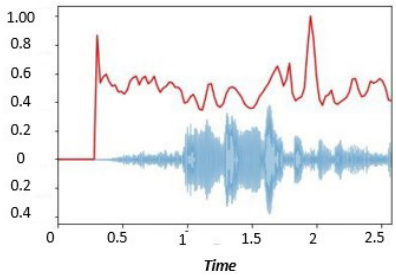

Fig. 6: Spectral Centroid of Female Voice

3). Chroma Feature

Chroma feature predominantly focus on tonal part of audio signal. It helps to recognize cords or finding harmonic similarities between audio signals [15]. Normally male voices are lower than female because of longer and thicker vocal folds which produce sounds with longer wavelengths which our ear identify as lower pitches. And all these varieties of vocal folds happen because of testosterone which is a male sex hormone [16]. So, our dataset shows also some verities of Chroma features between male and female voices which is shown in fig 7.

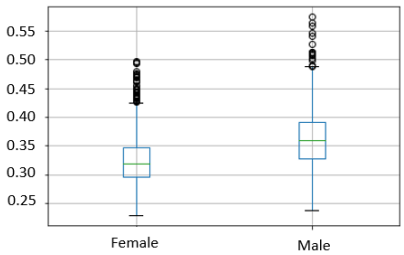

Fig. 7: Boxplot Presentation of Chroma Feature

4). Spectral Bandwidth

The difference between higher and lower point of a continuous frequencies is called Bandwidth. It is measured by Hertz. It may also refer to passband bandwidth or baseband bandwidth. The difference between upper and

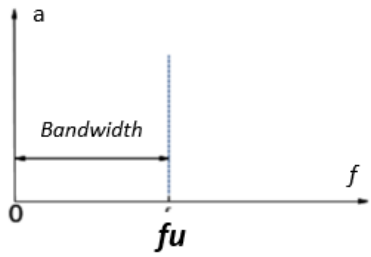

Fig. 8: Baseband Bandwidth

lower cutoff frequencies are called Passband Bandwidth, above figure (fig. 8) make a clear understanding of this. Such as a communication channel, a signal spectrum etc. On the other hand, the bandwidth which is equal to its upper cutoff frequency is called Baseband Bandwidth. It is applied on low pass filter or baseband signal. The boxplot presentation of spectral bandwidth according to our data given at fig. 9.

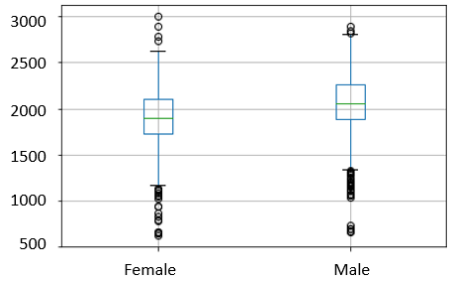

Fig. 9: Boxplot Presentation of Spectral Bandwidth

5). Rolloff

Explicitly roll-off alludes to the activity of a kind of channel, one intended to roll-off frequencies above or beneath a specific point. It is known as a roll-off in light of the fact that the procedure is continuous. Hi-pass and low pass channels both roll-off frequencies outside of their range, however, they don't promptly wipe out all frequencies outside their range. The sound is tenderly (or not all that delicately)" roll-off" with frequencies further above or beneath the cutoff recurrence ending up increasingly lessened. Roll off steepness is commonly expressed in dB per Octave, with higher numbers demonstrating a more extreme channel. 24 dB/Octave is more extreme than 12 dB/Octave and fig. 10 is illustrating boxplot of roll off.

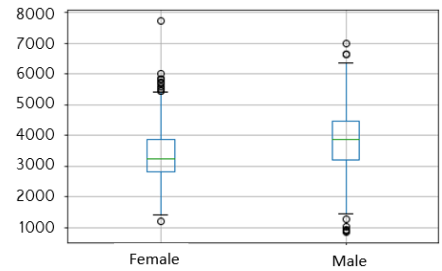

Fig. 10: Boxplot Presentation of Roll off

6). MFCCS Features

This is one of the most popular and effective method of feature extraction. The full form of MFCCs is Mel frequency cepstral coefficients. The sounds of human are filtered by the shape of the vocal tract including tongue, teeth etc. and the shape define what is the sound [11]. So, if it's possible to determine the accurate shape, it will be easy to work with human voices. For finding MFCCs we need to follow some steps those are given at fig 11 [11].

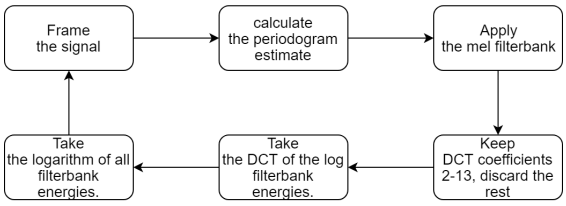

Fig. 11: Steps of MFCCs

For detecting pitch in linear manner so that system can understand those pitches of sound we need to use mel-scale[12]. The formula of converting normal frequency to mel-scale is given below:

$$M(f) = 1125 \times ln(1 + f/700) \quad (1)$$

$$M^{-1}(m) = 700 \times (exp(m/1125 - 1)) \quad (2)$$

Equation:1 is the formula of converting frequency to mel scale and equation:2 is the formula of converting mel scale to the frequency. After all of these a simple description of implementing steps are given below:

- First of all, its important to cut those voice signals into small frames and 20ms to 40ms is good but 25ms is the standard size for framing. So, if we have 32khz signal, we will get 0.025*32000=800 samples. Next steps will apply on every single frame that we got.
- This step will focus on Discrete Fourier Transform of the frames. For that we need to follow an equation.

$$Si(k) = \sum_{N}^{n=1} \mathrm{Si(n)h(n)}e^{-i2\pi kn/N} \text{ here, } 1 \le \mathrm{k} \le \mathrm{K} \quad (3)$$

Here, by calculating DFT we find Si(k), i denotes the number of frames, Pi(k) is the power spectrum of i number frame. Si(n) is the time domain frame by frame, K is the length of a frame. And we can extract the power spectrum of Si(n) as below:

$$Pi(k) = \frac{1}{N} \times |Si(K)| \quad (4)$$

- Computing Mel-spaced filter-bank is the main concern of this step. Mel-filter-bank is basically set of 20-40 filters. And the filters are those which we apply to the periodogram power spectral in previous step.
- We need to find out the log of every energy from previous step which provide us the log filter-bank energies.
- Finally, by transforming those filter-bank energies into discrete cosine we will get cepstral coefficients of those energies.

And finally, we extract 20 features of MFCC that shows as below (fig 12 and 13).

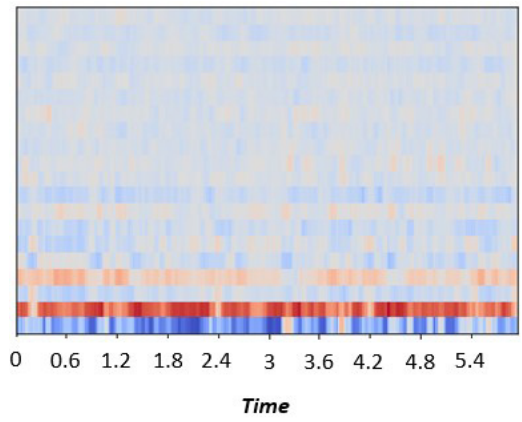

Fig. 12: MFCCs for male

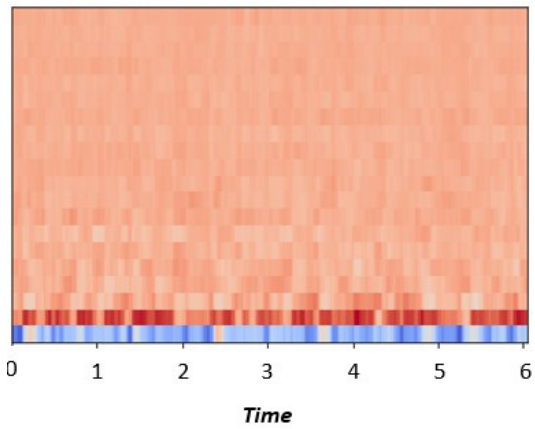

Fig. 13: MFCCs for female

So finally, we collected 26 features of a voice and those are: chroma feature, Root Mean Square Error, spectral centroid, spectral bandwidth, roll off, zero crossing rate and 20 features from MFCCS. And the heat map of fig 14 shows the correlations between all the features that we worked with. For clear understanding of fig .14 use the link: https://github.com/SaifulBadhon/heatmap/blob/master/heatmap.png

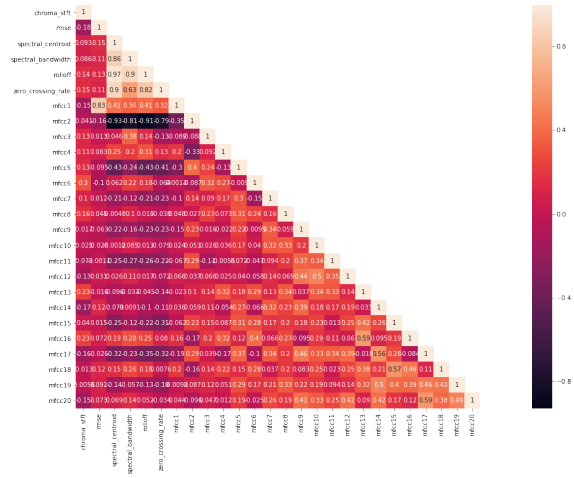

Fig. 14: Heat-map

IV. Experiments and Results

A. Experiment Setup

we tested our system in two ways. One is splitting the data in 8:2 train test data and another one is we input random voices which is not trained before and we matched the outcome with actual outcome. We tried some machine learning algorithm for this prediction Logistic regression, Random forest and Gradient Boosting showed better result among all of them. And gradient boosting was the best. We took help of confusion matrix for finding accuracy of randomly tested voices.

Confusion matrices of different models are plotted below (fig 15,16 and 17)

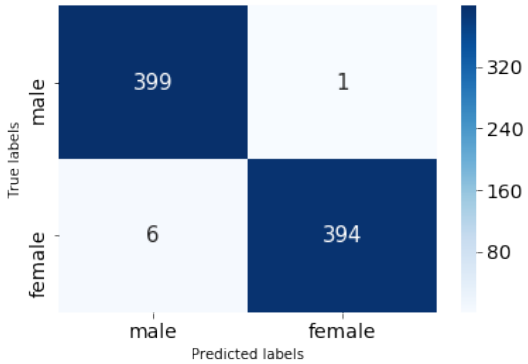

Fig. 15: Heat Map of Gradient Boosting Confusion Matrix

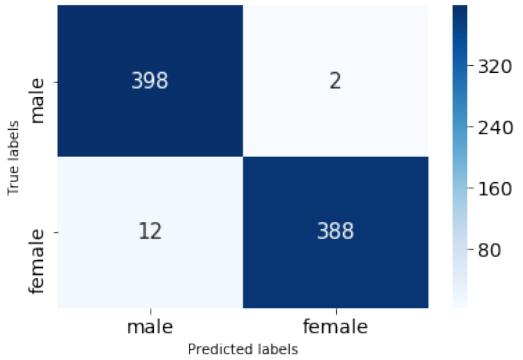

Fig. 16: Heat Map of Random Forest Confusion Matrix

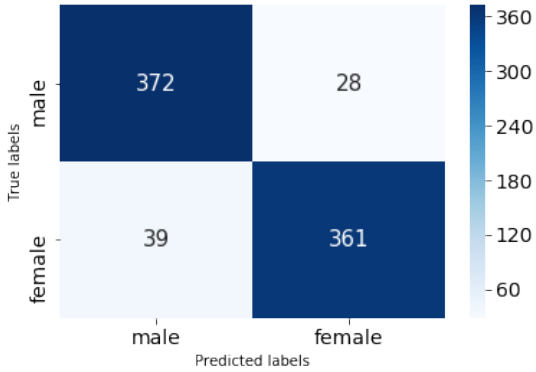

Fig. 17: Heat Map of Logistic Regression Confusion Matrix

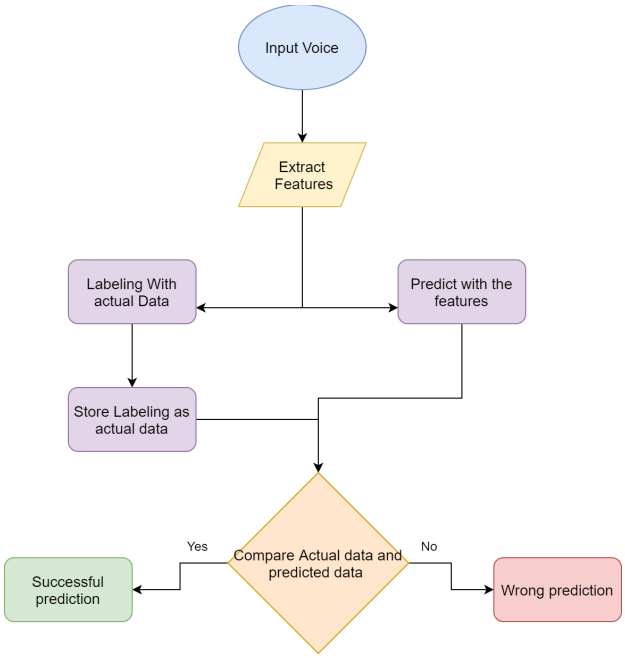

Fig. 18: Workflow of Detecting Result

After completing the training of machine, we need to test it. And for getting result we followed a workflow which is given at fig 18.

B. Result

After training the model we get highest accuracy in Gradient Boosting algorithm and that is 99.13%. For that we test with 400 male and 400 female voices for different algorithm those are given below:

TABLE 1: PERFORMANCE WITH GRADIENT BOOSTING ALGORITHM

	Precision	Recall	F1-Score	Support	Accuracy
Male	0.99	1.00	0.99	400	0.9913
Female	1.00	0.98	0.99	400	0.9913
Micro Avg	0.99	0.99	0.99	800	0.9913
Macro Avg	0.99	0.99	0.99	800	0.9913
Weighted Avg	0.99	0.99	0.99	800	0.9913

So, we can say that gradient boosting will be the best choice for us cause with this algorithm we get the best possible

TABLE 2: PERFORMANCE WITH RANDOM FOREST ALGORITHM

	Precision	Recall	f1-score	Support	Accuracy
Male	0.97	0.99	0.98	400	0.9825
Female	0.99	0.97	0.98	400	0.9825
Micro Avg	0.98	0.98	0.98	800	0.9825
Macro Avg	0.98	0.98	0.98	800	0.9825
Weighted Avg	0.98	0.98	0.98	800	0.9825

TABLE 3: PERFORMANCE WITH LOGISTIC REGRESSION ALGORITHM

	Precision	Recall	F1-Score	Support	Accuracy
Male	0.91	0.93	0.92	400	0.9162
Female	0.93	0.90	0.92	400	0.9162
Micro Avg	0.92	0.92	0.92	800	0.9162
Macro Avg	0.92	0.92	0.92	800	0.9162
Weighted Avg	0.92	0.92	0.92	800	0.9162

TABLE 4: PERFORMANCE IN DIFFERENT MODELS

Model	Accuracy	Error Rate	Precision	Recall	F1-score
Gradient Boosting	99.13%	.88	99	99	99
Random Forest	98.25%	.74	98	98	98
Logistic Regression	91.62%	.27	92	92	92

accuracy which is 99.13% and lowest Error Rate which is 0.88%.

V. CONCLUSION AND FUTURE WORK

This work tried to detect human gender from their voices in Bengali language. In near future there will be heavy use of voice-based application. By 2020 there will be 50% voice internet searching [13]. 100 million smartphone users will use voice assistant in 2020 [14]. Even in Bangladesh and native speaker of Bengali start using voice-based applications. For this upcoming future of voice recognition systems, it will be mandatory to detect voices gender. There are some research papers on gender detection in Bengali language with impressive accuracy but lake of verity on voices that's mean lake of speakers. The more speaker we have the more verity we have in voices. Here this paper tried to work with more verity of voices. We had more than 250 speakers and exactly 1652 voices. And the accuracy was 99.13%.

This paper didn't work with third gender in future we want to work with third gender and, we want to improve our data sets verity. We are focusing on verity of data set not in number. Cause in gender detection we need different type of speaker so that we get more and more verity of voices. Lots of voices of same speaker will not be helpful for gender detection.

REFERENCES

[1] P. Gupta, S. Goel, A. Purwar, "A Stacked Technique for Gender Recognition Through Voice", 2018 Eleventh International Conference on Contempo- rary Computing (IC3),2-4 Aug. 2018

[2] F. Lin, Y. Wu, Y. Zhuang, X. Long, 'Human Gen- der Classification: A Review', 2015. [Online]. Avilable: https://www.researchgate.net/publication/280105452 [Accessed: 29-Aug- 2019]

[3] M. Alhussein, Z.Ali, M.Imran and W.Abdul, 'Automatic Gender Detection Based on Characteristics of Vocal Folds for Mobile Healthcare System'. Avilable:https://www.hindawi.com/journals/misy/2016/7805217/ [Accessed: 29- Aug- 2019]

[4] G. Tzanetakis, Audio-based gender identification using bootstrapping, in Communications, Computers and signal Processing, 2005. PACRIM. 2005 IEEE Pacific Rim Conference on. IEEE, 2005, pp. 432433.

[5] H. Harb and L. Chen, Voice-based gender identification in multimedia applications, Journal of intelligent information systems, vol. 24, no. 2-3, pp. 179198, 2005.

[6] A general audio classifier based on human perception motivated model, Multimedia Tools and Applications, vol. 34, no. 3, pp. 375395, 2007.

[7] S.-I. Kang and J.-H. Chang, Discriminative weight training-based opti- mally weighted mfcc for gender identification, IEICE Electronics Express, vol. 6, no. 19, pp. 13741379, 2009.

[8] L. Kye-Hwan, K. Sang-Ick, K. Deok-Hwan, and J.-H. Chang, A support vector machine-based gender identification using speech signal, IEICE transactions on communications, vol. 91, no. 10, pp. 33263329, 2008.

[9] Y. Hu, D. Wu, and A. Nucci, Pitch-based gender identification with two- stage classification, Security and Communication Networks, vol. 5, no. 2, pp. 211225, 2012.

[10] M. S. Ali, M. S. Islam, and M. A. Hossain, Gender recognition system using speech signal,International Journal of Computer Science, Engineering and Information Technology, 2012.

[11] J. Lyons, 'Mel Frequency Cepstral Co- efficient (MFCC) tutorial',2013. [Online]. Available:http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/ [Accessed: 23- Aug- 2019].

[12] 'The mel frequency scale and coefficients', 2013. [Online]. Available:http://kom.aau.dk/group/04gr742/pdf/MFCC worksheet.pdf [Accessed: 27- Aug- 2019].

[13] R. Sentance, 'The future of voice search: 2020 and beyond' , 2018. [On- line]. Avilable:https://econsultancy.com/the-future-of-voice-search-2020- and-beyond/ [Accessed: 29- Aug- 2019].

[14] C.Ciligot, '7 Key Predictions For the Future of Voice Assistants and AI' 2019. [Online]. Avilable:https://clearbridgemobile.com/7-key-predictions-for-the-future-of-voice-assistants-and-ai/ [Accessed: 29- Aug- 2019].

[15] Kattel, M. & Nepal, Araju & Shah, Ayush & Shrestha, Dev, 'Chroma Feature Extraction', 2019[online]. Avilable: https://www.researchgate.net/publication

[16] 330796993 Chroma Feature Extraction[Accessed:21-Sep-2019]. H. Reith 'Why are male and female voices distinctive?', 2016 [online]. Avilable: https://www.quora.com/Why-are-male-and-female-voices- distinctive [Accessed: 21- Sep- 2019].

