FORMANT ANALYSIS OF BANGLA VOWEL FOR AUTOMATIC SPEECH RECOGNITION

Tonmoy Ghosh[1], Subir Saha[2] and A. H. M. Iftekharul Ferdous[3]

[1,3]Department of Electrical and Electronic Engineering,
Pabna University of Science and Technology, Pabna, Bangladesh
[2]Department of Computer Science and Engineering,
Pabna University of Science and Technology, Pabna, Bangladesh

ABSTRACT

To provide new technological benefits to the mass people, nowadays, regional and local language recognition draws attention to the researchers. Similarly to other languages, Bangla speech recognition scheme is demandable. A formant is considered as the resonance frequency of vocal tract. Formant frequencies play an important role for the purpose of automatic speech recognition, due to its noise robust characteristics. In this paper, Bangla vowels are investigated to acquire formant frequencies and its corresponding bandwidth from continuous Bangla sentences, which are considered as potential parameters for wide voice applications. For the purpose of formant analysis, cepstrum based formant estimation and Linear Predictive Coding (LPC) techniques are used. In order to acquire formant characteristics, enrich continuous sentences and widely available Bangla language corpus namely "SHRUTI" is considered. Intensive experimentation is carried out to determine formant characteristics (frequency and bandwidth) of Bangla vowels for both male and female speakers. Finally, vowel recognition accuracy of Bangla language is reported considering first three formants..

KEYWORDS

Automatic speech recognition, Bangla vowel, Formant, Linear Predictive Coding.

1. INTRODUCTION

Automatic speech recognition (ACR) is considered as a challenging task in the field of speech and sound engineering. Accurate recognition technique demands highly separable features of the phoneme, but due to acoustic similarity, it becomes very difficult. A formant is one of the acoustic features that are proven to become a well representative of vowels. Both Bangla and English languages are from Indo-European family, having many similarities as well as differences in the phonemic system. Bangla language is different from English due to it pronunciation and alphabet arrangement in a word. Apart from English, formant characteristic of Bangla vowels needs to be investigated.

Bangla is the seventh most spoken language in the world, around 300 million people speaks in Bengali [1]. Most of the Bengali native people prefer to use their mother language in recent technologies such as mobile phone, computer application etc. As a result Bangla speech

DOI : 10.5121/sipij.2016.7501

recognition, speech to text and text to speech conversion are highly demandable. However, very few attempts are taken to address Bangla speech recognition. Now a day, to provide technological benefit to the mass people, Bangla speech recognition technique draws attention to the researchers. Feature extraction for the purpose of Bangla speech recognition is presented in [2], where word base Bangla speech recognition is reported using mel-frequency cepstral coefficients (MFCCs). The major drawback of this approach is in real world application, speech is generated using continuous sentences. In [3], similarly Bangla word is recognized by MFCCs, and LPC but the simulation is performed in very few words, which does not ensure its reliability. Bangla vowel characterization is presented in [4], where Bangla vowel synthesis is used. In [5], dynamic time warping (DTW) and K-nearest neighbour (KNN) is considered for Bangla speech analysis for the purpose of recognition. Auto Correlation Function (ACF) based formant frequency estimation is presented in [6]. All the research works are done by the limited database, mainly recorded in their own environment and discrete words are used as an input voice. As a result, each method has its own limitation of the self-made dataset, but nowadays standard continuous sentences database (also called corpus) has been developed, considering both male and female speaker, which is publicly available. Formant characteristics determination in standard Bangla database is highly demandable.

In this paper, Bangla vowels are investigated to determine formant frequency and bandwidth. At first vowel signal is extracted from continuous sentences, which are collected from standard Bangla database. To enhance vowel formant characteristic, pre-processing steps are performed. Two widely used formant estimation technique: 1) cepstrum based, and 2) LPC based methods are used to calculate formant frequency and its corresponding bandwidth.

2. ANALYSIS PROCEDURE

The human voice is produced from the vocal cord which is referred to be the primary sound source. Vocal cord is a vibrating valve, generates audible pulses from the lung airflow or air pressure, and composes the laryngeal sound source. In general, the acoustic resonance of the human vocal tract is called a formant. However, in signal processing point of view, formant is a range of frequency in the sound spectrum, where there is an absolute or relative maximum occurred [7]. Thus, formant can be defined as either a resonance of vocal tract or the spectrum maximum that the resonance produces and it can be measured from amplitude peak in the sound frequency spectrum. Another approach to determining the formant frequencies and its bandwidth is by vocal tract modelling. There are 14 vowels and 28 consonants in Bangla language [8]. Bangla vowels are classified with it originating position in vocal tract and it is expected that vowels in same class show similar characteristics and, on the other hand, different classes show distinguishable natures. 14 Bangla vowels are listed in Table 1 with classification.

Table 1. Vowels in Bangla Language.

	Front	Central	Back
Close	ঈ ($\tilde{\imath}$) ই (i)		ঊ ($\tilde{u}$) উ (u)
Close-mid	ঐ ($\tilde{e}$) এ (e)		ঔ ($\tilde{o}$) ও (o)
Open-mid	এঁ ($\tilde{æ}$) এ্যা ($æ$)		অঁ ($\tilde{ɔ}$) অ ($ɔ$)
Open		আঁ ($\tilde{a}$) আ (a)	

2.1. Bangla Language Corpus

One of the essential element to doing research in Bangla language is database or corpus of Bangla language. There is very limited number of Bangla language corpus available. A standard Bangla language data plays an important role in Bangla speech recognition, Bangla text to speech and Bangla speech to text operations. Mainly there is two established corpus namely: 1) SHRUTI Bengali Continuous ASR Speech Corpus, created by Indian Institute of Technology, Kharagpur (IITKGP), [9] and 2) The EMILLE Corpus, created by Lancaster University, UK [10]. "EMILLE" corpus is developed for a minority language, as a result, there are very limited number of words and sentences available, and most of the sentences are not marked by phoneme, thus this corpus is not considered in this paper. "SHRUTI" is a dedicated Bangla language corpus, containing a total 7383 unique sentences. Sentences are spoken by 34 speakers, collected from different parts of the Indian state of West Bengal, where 75% speakers are male and rest of them are female. This corpus is well established and available in online [11]. But one of the drawbacks of this corpus is that phoneme information of uttering sentences is not available. In this paper, for the purpose of formant analysis of Bangla vowels "SHRUTI" corpus is considered, due to its wide variation and availability. Both Male and Female speakers and different sentences are considered.

2.2. Pre-processing of speech signal

In order to acquire desired vowel speech signal from the continuous sentences, it is essential to do some pre-requisite steps as vowel tracking, resample, windowing, band limiting and pre-emphasis. At first from a given sentence, vowels are tracked manually for the purpose of formant estimation. After taking the vowel portion from a given sample, it is resampled to 8 KHz signal, due to human voice maximum frequency is considered as 4 KHz, thus according to the Nyquist–Shannon sampling theorem, 8 KHz sampling rate is enough to represent human Bangla voice. Windowing operation is performed to remove ripples of that given speech signal. For the purpose of windowing, Hamming windowing is considered. Hamming windowing represents as

$$w(n) = ham(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) & for\ 0 \le n \le N-1 \\ 0 & otherwise \end{cases} \quad (1)$$

The human voice signal is a band limiting signal with the highest frequency of 3600 Hz, to limit given vowel signal, a low pass filter is designed with a cutoff frequency of 3600 Hz. From the nature of formant, it can be referred, that first formant shows the highest level of energy and level of energy is gradually decreased from second to third and so on. As a result, it is very difficult to trace the third or higher formant frequencies. To overcome this problem a preemphasis filter is implemented to that band limited signal. This pre-emphasis filter not only increases higher formant frequencies energy label, but also mitigates DC component of a given speech signal. The pre-processing steps are shown in Fig. 1.

2.3. Formant Estimation Methods

2.3.1 Cepstrum Based formant Estimation

Peak picking methods of formant estimation is an ancient but very effective method. According to the definition of formants, the absolute or relative maximum in sound spectrum is consider as formant frequency. In cepstrum based analysis, vocal tract shape is considered as a filter and its spectral peaks are considered as resonance and are commonly referred as formants. Here speech signal is represented as.

$$S(t) = g(t) \otimes h(t) \tag{2}$$

where the speech signal is a convolution product of excitation g(t) and vocal tract filter h(t). The cepstrum is computed by power spectrum, which is calculated using Fourier transformation, followed by an inverse Fourier transformation of the logarithm of that power spectrum. formants are estimated with the help of the range of formant frequencies and spectrum peaks.

$$c(n) = FFT^{-1}(\log(FFT(s(n)))) \tag{3}$$

In the cepstrum equation, excitation (g(n)) and vocal tract filter (h(n)) are superimposed, and can be separated by filtering. As a result, low order terms of cepstrum contain vocal tract information. After calculating cepstrum, extract amplitudes corresponding to the vocal tract resonances, and at resonance point local maxima is found. Finally from the local maxima of cepstrum spectral formant frequencies are estimated.

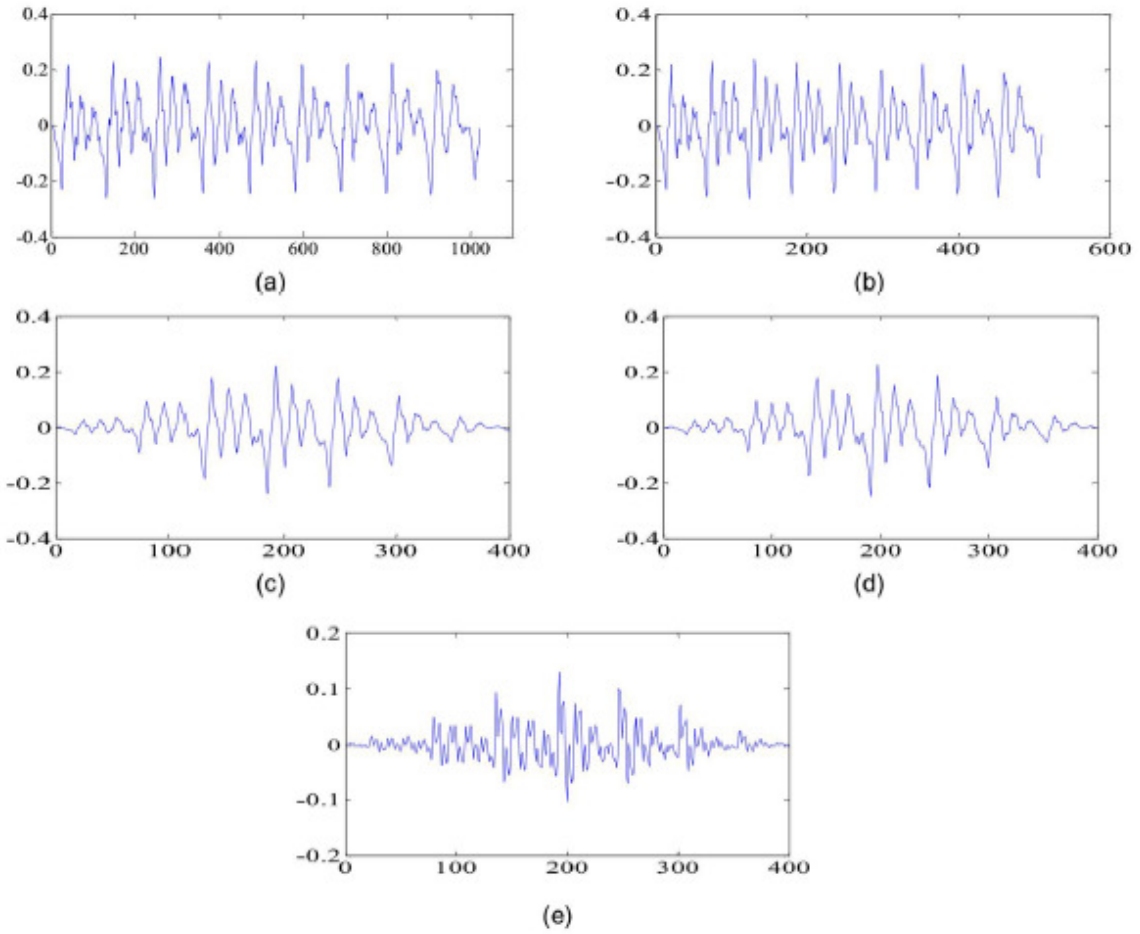

Figure 1. Illustration of pre-processing steps to acquire formant frequency of 'ই' vowel of "ইত্যাদি" word: (a) given speech signal; (b) resampled signal; (c) signal after hamming windowing; (d) signal after band limiting; (e) pre-emphasis signal

2.3.2 Linear Predictive Coding (LPC) Technique

From the previous discussion, it can be referred that, vocal tract can be modeled as a linear filter with resonances. And formant frequencies can be obtained from the resonance frequencies of the vocal tract. Graphically, the peaks of the vocal tract response correspond roughly to its formant frequencies. Therefore, a voice speech signal and vocal cords the system can be assumed as causal, stable, linear time-invariant and stationary autoregressive (AR) system. In short, vocal tract is all-pole linear system and its each pair of conjugate pole correspond to the formant frequency. So a voice speech signals given by

$$H(z) = \frac{G}{1 + \sum_{k=1}^{P} a_k z^{-k}} = \frac{G}{\prod_{k=1}^{P} (1 - p_k z^{-1})} \tag{4}$$

where G is the gain, P is the system order, $\{a_k\}$ are the system parameters, and $\{p_k\}$ are the system poles. The complex conjugate pole pairs of vocal tract system represents Formants. For a complex pole $p_k = r_k \exp(j\omega_k)$ with magnitude r_k and discrete angular frequency ω_k, the formant frequency (F_k) and bandwidth (B_k) can be computed as [12]

$$F_k = \frac{F_s}{2\pi}\omega_k; \qquad B_k = \frac{F_s}{2\pi}\ln(r_k); \tag{5}$$

The first absolute or relative maximum in sound frequency spectrum is defined as the first formant and called F1, the second one is defined as the second formant and called F2, and similarly third formant as F3, and so on. It is observed that generally, the first formant exhibits the highest energy, and the second formant exhibits lower energy than the first formant and so on. The first three formants show comparatively higher energy than the other formants, which can be detected with high precision. In most of the cases, it is quite enough to represent a voice sound or vowel phoneme by first three formants [13]. LPC power spectrum is shown in Fig 2. of 'ই' vowel, here peaks are considered as formant frequency.

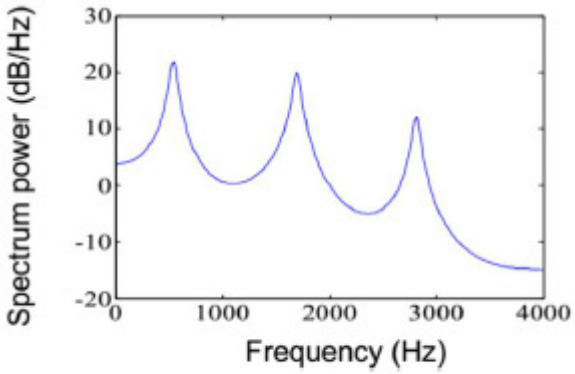

Figure 2. LPC power spectrum of 'ই' vowel

3. SIMULATION RESULT

In order to evaluate formant frequencies and its bandwidth, experimentation has been conducted in a subset of 'SHRUTI Bengali Continuous ASR Speech Corpus', containing 5 male and 5 female speakers, and 5 uttering sentences of each speaker. Spectrogram of sample vowels is represented in Fig. 3. In Fig. 3. (a) spectrogram of 'অ' vowel is shown which is collected from "অনেক" (aneka) word spoken by 'deb' speaker. Similarly, spectrum of 'আ', 'ই', 'উ', 'এ', and 'ও' vowel from different words, collected from continuous sentences, spoken by different speakers are demonstrated in Fig. 3 (b) to (f). In the figure, red or dark red represents resonance frequencies or formant frequencies whereas blue color represent lower spectrum power.

Different Bangla vowels exhibit different formant frequencies with variation of bandwidth is shown in spectrogram plot, which is considered one of the important characteristic for Bangla speech recognition technique.

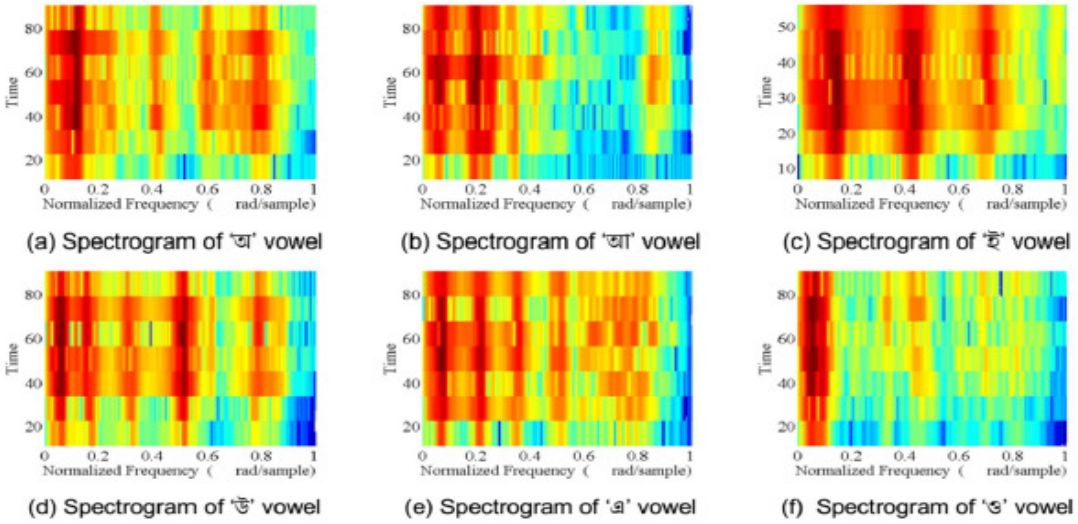

Figure 3. Illustration of spectrogram of six Bangla vowel

In this paper, formant frequency is computed by two different methods, which are described in subsection 2.3. At first, cepstrum based peak picking method is applied followed by LPC based formant estimation method. After that average formant value is calculated for a given vowel for 25 (5 speakers, 5 utterances each) samples of male and 25 samples of a female speaker. Formant bandwidth is computed using LPC method for all the sample vowels and taking average to represent final result. Order in LPC method is considered as 12, that ensure getting at least 3 pairs of poles. Using equation (5), formant frequencies and bandwidth are calculated from the estimated pair of poles.

Table 2. Average formant frequency and bandwidth for male.

Bangla vowel	F1		F2		F3	
	frequency	bandwidth	frequency	bandwidth	frequency	bandwidth
অ	607	100	1160	105	2388	164
আ	816	172	1330	591	2456	209
ই	480	153	1760	87	2760	1064
উ	367	215	1358	460	3056	705
এ	408	55	1875	485	2826	178
ও	445	97	1245	175	2766	155

Table 3. Average formant frequency and bandwidth for female

Bangla vowel	F1		F2		F3	
	frequency	bandwidth	frequency	bandwidth	frequency	bandwidth
অ	657	115	1080	165	2512	657
আ	867	155	1540	456	2644	867
ই	544	163	1830	108	2807	544
উ	388	230	1410	480	3130	388
এ	510	97	1920	360	2840	510
ও	470	121	1280	215	2815	470

Table 4: Bangla vowel recognition accuracy using first three formats

Male		Female	
LPC	Cepstrum	LPC	Cepstrum
75%	77%	71%	72%

For the purpose of analysis, only major six Bangla vowels are reported namely: 'অ', 'আ', 'ই', 'উ', 'এ', and 'ও'. Average formant frequency and bandwidth of six major Bangla vowels of male speakers are presented in Table 2. And for female speakers, formant frequencies and their corresponding bandwidth are reported in Table 3. From the analysis table, it is observed that formant frequency of female the speaker is higher than the male speaker, due to high fundamental frequency (pitch) of a female speaker. it is also found that Bangla vowels show similar formant characteristics with English vowels with a slight difference [14]. The reason behind that difference is Bangla language pronunciation, word formation, and tone of Bangla language. Bangla vowel recognition accuracy using first three formants is reported in Table 4. Here, the linear discriminant analysis (LDA) classifier is used and it is found that cepstrum based formant estimation method provides a better result than LPC.

4. CONCLUSION

Formants are very important parameter to represent a vowel signal. It is widely used in practical applications like voice recognition, vowel synthesis, and automatic speech generator. The major contribution of this work is to find formant frequencies of Bangla vowels using continuous sentences collected from the well-established database, whether most of the previous literature used their own recordings and discrete words. In order to estimate the formant frequency of Bangla vowels, two well establish methods are taken into consideration such as LPC and ceptrum based formant estimation. Finally, Bangla vowels are detected using those formant frequencies. In practical application, people generally utter sentences rather than discrete word thus our work is more related to practical applications. Another advantage of formant frequency is highly noise robust; as a result in noisy environment voice based application is highly depended on formant frequency.

REFERENCES

[1] (2015) The ethnologue website. [online]. Available: http://www. ethnologue.com/statistics/size

[2] N. J. Lisa, Q. N. Eity, G. Muhammad, M. H. Huda, and C. M. Rahman, "Performance evaluation of Bangla word recognition using different acoustic features," International Journal of Computer Science and Network Security, vol.10, pp.96-100, 2010.

[3] M. A. Ali, M. Hossain and M. N. Bhuiyan, "Automatic speech recognition technique for Bangla words", International Journal of Advanced Science and Technology, vol. 50, pp. 51-60, 2013.

[4] S. A. Hossain, M. L. Rahman, and F. Ahmed, "Bangla vowel characterization based on analysis by synthesis," International Journal of Computer, Electrical, Automation, Control and Information Engineering, vol. 1, pp. 354-357, 2007.

[5] A. Sayem, "Speech analysis for alphabets in Bangla language: automatic speech recognition," International Journal of Engineering Research, vol. 3, pp. 88-93, 2014.

[6] S. A. Fattah, T. Ghosh, A. K. Das, R. Goswami, A. S. M. M. Jameel, and C. Shahnaz, "An approach for formant based speech recognition in noise," in Proc. IEEE region 10 conference, TENCON, pp. 1-4, 2012.

[7] S. H. Lee, T. Y. Hsiao and G. S. Lee, "Audio–vocal responses of vocal fundamental frequency and formant during sustained vowel vocaliza-tions in different noises," Hearing Research, vol. 324, pp. 1-6, June 2015.

[8] D. Huq, Bhasha bigganer katha (Facts about linguistics), 1st Ed., Mowla Brothers, 2002.

[9] B. Das, S. Mandal, and P. Mitra, "Bengali speech corpus for continuous automatic speech recognition system," in Proc. Conf. Speech Database and Assessments, Oriental COCOSDA, pp. 51-55, 2011.

[10] (2015) The Emille corpus website. [online]. Available: http://www.lancaster.ac.uk/fass/projects/corpus/emille/

[11] S. Mandal, B. Das, P. Mitra, and A. Basu, "Developing Bengali speech corpus for phone recognizer using optimum text selection technique," in Proc. Conf. Asian Language Processing, IALP, pp. 268-271, 2011.

[12] B. Yegnanarayana and R. N. J. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," IEEE Trans. Speech Audio Processing, vol. 6, no. 4, pp. 313–327, 1998.

[13] X. D. Huang, A. Acero and H. W. Hon, Spoken Language Processing, 1st Ed., Englewood Cliffs, Prentice-Hall NJ, 2001.

[14] D. O. Shaughnessy, Speech Communications: Human and Machine, 2nd ed., IEEE Press, NY, 2000.

AUTHORS

Tonmoy Ghosh is now working in Pabna University of Science and Technology (PUST) as a Lecturer in the department of Electrical and Electronic Engineering. Previously he also worked as an engineer in the telecom industries of Bangladesh for two and half years. He achieved his B.Sc. and M.Sc. Engineering degree in Electrical and Electronic Engineering from Bangladesh University of Engineering and Technology (BUET). His research interests are Signal Processing, Image Processing, Boimedical Engineering, and Video Processing.

Subir Saha is now working in Pabna University of Science and Technology as a Lecturer of Computer Science and Engineering. Previously he also served for Green University in Bangladesh as Lecturer of CSE for 2 years. He achieved his B.Sc. degree in Computer Science and Engineering from Bangladesh University of Engineering and Technology in 2013. Now he is doing his MSc in the same department. His research interests are System Network, Embedded Systems, Bioinformatics, E-learning.

A.H.M.Iftekharul Ferdous is now working in Pabna University of Science and Technology as a Lecturer of Electrical and Electronic Engineering department. Previously he also served for UITS University in Bangladesh as Lecturer of EEE for 1.5 years. He achieved his B.Sc. degree in Electrical and Electronic Engineering from Islamic University of Technology in 2010. Now he is doing his MSc in the same department in RUET. His research interests are System Network, Renewable energy, Smart grid.