IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 6, 2020, accepted February 6, 2020, date of publication February 18, 2020, date of current version February 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2974799

Acoustic Analysis of the Speakers' Variability for Regional Accent-Affected Pronunciation in Bangladeshi Bangla: A Study on Sylheti Accent

SHAFKAT KIBRIA, M. SHAHIDUR RAHMAN, (Member, IEEE), M. REZA SELIM, (Member, IEEE), AND M. ZAFAR IQBAL
Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet 3114, Bangladesh

Corresponding author: Shafkat Kibria (shafkat.kibria-cse@sust.edu)

This work was supported by the Higher Education Quality Enhancement Project (AIF Window 4, CP 3888) for "The Development of Multi-Platform Speech and Language Processing Software for Bangla."

ABSTRACT Accented pronunciation variability is one of the key elements that deteriorate the accuracy of the automatic speech recognition (ASR). This article reports the results of the acoustic analysis of the two groups of speakers' variability caused by regional accent in Bangladeshi Bangla. The analysis considers the seven monophthongal and four diphthongal vowels of Bangla to investigate the acoustic characteristics of two groups of single-accent speakers and their correlation on the articulation of the Standard Colloquial Bangladeshi Bangla (SCBB). An accent is the speaker's regional signature and shaped by his/her community and educational background. This study examines both male and female speakers from the Sylhet region, which has one of the extremely deviant dialects in Bangla, and comparatively less deviant speakers from different districts of North-West and Middle Part of Bangladesh. Accent-related acoustic features such as pitch slope, formant frequencies, and vowel duration have been considered to examine the prominent characteristics of the accents and to classify the accents from these features. Both gender groups are distinctly analyzed. It has been found that there are significant deviations in formant frequencies and various steepness of the rise/fall in pitch slope within accents of both gender groups. In this study, it has been observed that accent related changes in speech affect the ASR performance. This has emphasized the need for accent-specific acoustic models to handle the speakers from highly deviant dialects as well as considering the accent-affected speakers' variability in the corpora development for robust ASR system in Bangladeshi Bangla.

INDEX TERMS Accent checker, accent analysis, accent classification, accent database, acoustic analysis, Bangladeshi Bangla, formant frequencies, intonation, pitch, pitch slope, speaker variability, Standard Colloquial Bangladeshi Bangla (SCBB), Sylheti Accent.

I. INTRODUCTION

In Bengali or Bangla (বাংলা/baŋla/) language, there are many different accents among native speakers [2]. Geographically, one can divide them in two major regions: people of Bangladesh and people of West Bengal (a part of India) [3]. Some Bangla native speakers also live in other countries of the world. So, Bangla can be broadly classified into two main accent groups: Bangladeshi Standard Bangla and Kolkata (capital of West Bengal) Standard Bangla. There are standard accents in every language; in English there are Received Pronunciation (RP) [43], which is a Standard British English accent, and General American (GenAm) [44], which is a Standard American English accent, etc. For Bangladeshi Bangla, Standard Colloquial Bangladeshi Bangla (SCBB - প্রমিত বাংলা) is Standard Bangladeshi Bangla accent of the educated people of Bangladesh. It is the affiliation of the standard diversity of the spoken language in Dhaka and other cities of Bangladesh. SCBB also varies in phonetic and some other linguistic context from the Kolkata Standard [1].

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

There is also some significant difference within each of these two regions. According to P. Sloka Ray *et al.* 1966, there are some highly deviant regional dialects in Chittagong, Chittagong Hill Tracts, Sylhet, Rangpur, Mymensingh, etc. in Bangladesh [2]. The *deviant dialect* refers to the dialect that departs from accepted standard dialect of a specific language. In this article, we have reported the acoustic analysis of the accent-related features of eleven (11) vowels, seven of them are monophthongal vowels and four of them are diphthongal vowels. The analysis involved examining the accent-affected pronunciations effect of inter-speaker variability on the acoustic features for the two groups of single-accent speakers in Bangladeshi Bangla language. The corpus is formed from two groups of single-accent speakers in Bangladesh; one group from the Sylhet region that has extremely deviant dialect and the other group from different districts of North-West and middle part of Bangladesh, which have less deviant dialects. Furthermore, we have examined the interrelationship of the chosen vowels' acoustic features within each of these accents on the pronunciation of the SCBB. Four machine learning (ML) classification methods have been tried to classify the speakers' accent groups based on the accent-affected acoustic features. At the end of the article, we have reported the observation of the performance of two automatic speech recognition (ASR) systems on the accent groups.

The term *dialect* refers to the difference in pronunciation, vocabulary and grammar among varieties of the same language that form a particular speech pattern whereas the term *accent* refers to the distinct pattern of pronunciation [4], [5]. "Accent" of a language reflects the people of a geographical region and/or a socio-economic class to which they belong [4]. It also reflects the speakers' educational background [4]. Researchers had accomplished several kinds of accent-related acoustic analysis for various languages across the world [6]–[11]. There are no research findings on the Bangladeshi Standard Bangla, except for our own on the accent-affected acoustic features analysis of four (4) monophthongal vowels [12]. The accent analysis researches in other languages, had reported to have different accent-affected acoustic features that help us to know the regional accent effect on speech for a particular language community. These reported acoustic features are the first three formants frequencies, phone duration, intensity and pitch slope of vowel sounds [6]–[11]. The formant frequencies F1 (first formant), F2 (second formant), and F3 (third formant) are resting on the disposition of the vocal tract for utterance of different types of vowels. Research on the speakers of six different regions of America has shown that formants F1 and F2 are effected while vowel category differed significantly by the regional dialect [11]. Other researches had shown that the phone durations varied for different vowels across different regional accents for the same language [9], [10]. Formants F1, F2, and F3 significantly differed in some of the vowels for two well-known regional accents in British English [10]. Similar researches had shown that there is also significant effect of regional accents on the pitch slope among the same language [10], [13]. Therefore, the vowels' acoustic characteristics are essential to do accents analysis on a particular language community [9]–[11], [13]. Vowels have significantly more feature details for accent analysis, however this study restricts on the acoustic features of the Bangla vowel phonemes for investigating regional accent effect.

Formants represent the resonant frequencies of the vocal tract during the articulation. So, one can analyze the formants frequencies over time to investigate the effect of accents in vowel acoustics. Previous research has analyzed the formants frequencies for a specific vowel by analyzing the average frequencies over time [10]; whereas, in this study, the linear regression has been used to generate the formants contour which has given better generalized representation (see details in Section IV-A1 and IV-A2). Previous research on accent classification shows that acoustic features like: (i) Formants frequencies – F1, F2, and F3 (ii) Phone duration (iii) Intensity (iv) Mel-Frequency Cepstral Coefficients (MFCCs) and (v) the prosodic features such as pitch contour, rise/fall in the pitch slope are shown to differ significantly within regional accents [6], [7], [10], [14]. In this study, we have considered acoustic characteristics such as Formants frequencies (F1, F2, F3), Phone duration and rise/fall in the pitch slope for the regional accents classification using the four (4) ML methods (see details in Section II-B and IV-D). The compared and analyzed results of classifications methods also been presented.

Deep learning techniques are making a deep impact towards huge advancement in ASR system with a large vocabulary recognition [15]. However, the quality of ASR depends on the quality of the speech corpus. On the contrary, Bangladeshi Bangla has inadequate annotated speech corpora for large vocabulary continuous speech recognition (LVCSR) system. The quality of the corpus depends on the hours collected of speech as well as on the speaker variability [7]. This study has shown the necessity of investigating the regional accents variation in Bangladeshi Bangla to categorize the speaker variability and to build a quality speech corpus for the robust LVCSR system.

A. VOWELS IN BANGLA

Previously, several researchers investigated Bangla vowels based on articulatory phonetics and acoustic attributes. According to Suniti Kumar Chatterji (1921), Manzur Morshed A. K. (2001) and Alam *et al.* (2007) have reported that there are 14 (fourteen) monophthongal vowels in Bangla. These are: ই/i/, এ/e/, অ্যা/æ/, আ/a/, অ/ɔ/, ও/o/, উ/u/, ইঁ/ĩ/, এঁ/ẽ/, অ্যাঁ/æ̃/, আঁ /ã/, অঁ/ɔ̃/, ওঁ/õ/, উঁ/ũ/ [16], [17], [21]. Whereas, Abdul Hai (1967) and Daniul Huq (2002) have reported following 8 (eight) monophthongal vowels: ই/i/, এ/e/, অ্যা/æ/, আ/a/, অ/ɔ/, ও/o/, উ/u/, and ঔ/ou/. They have ignored nasality of these phonemes because of their less frequent use in Bangladeshi Bangla [18], [19]. On the contrary of their claim most research studies have reported ঔ/ou/ as a diphthongal phoneme. C. A. Ferguson and M. Chowdhury (1960) and S.

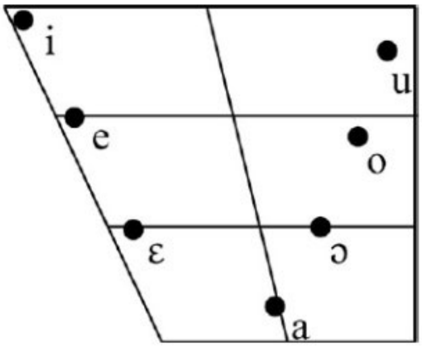


FIGURE 1. Vowels are plotted based on F1 and F2 Frequencies averaged across six speakers - Sameer Ud Dowla in 2010 [1].

Dowla Khan (2010), have reduced the monophthongal vowel phonemes number to 7 (seven): three front unrounded, three back rounded and a low neutral vowel (see Figure 1). Their corresponding nasal-vowels are much less frequent than the oral vowels in practice. The vowel phonemes are: ই/i/, এ/e/, অ্যা/æ/ or অ্যা/ɛ/, আ/a/, অ/ɔ/, ও/o/, and উ/u/ [1], [20].

Previous researches have not agreed on the actual numbers of the diphthongal vowels in Bangla. According to Abdul Hai (1967), there are about 31 diphthongal vowel phonemes [18]; on the other hand, Suniti Kumar Chatterji (1921) has shown only 25 and Manzur Morshed A. K. (2001) has reported only 29 [16], [17]. Alam *et al.* (2007) has reported the union of all the distinct findings and studied total of 38 diphthongs [21]], whereas, S. Dowla Khan (2010) has listed only 16 of them [1].

In our study of accented speech data annotation, we have considered 11 (eleven) vowels; 7 (seven) monophthongs: ই/i/, এ/e/, অ্যা/æ/ or অ্যা/ɛ/, আ/a/, অ/ɔ/, ও/o/, and উ/u/ and 4 (four) diphthongs: আও/ao̯/, এই/ej/, অয়/ɔe̯/, ওই/oj/. অ্যা vowel is denoted by æ or ɛ in IPA, however we followed the Bangladeshi standard Bangla IPA [1] and used ɛ throughout this paper. In this study, the acoustic and prosodic features of these vowels have been explored and investigated. The script for recording the accented speech corpus contains SCBB sentences (see detail in Section II).

B. WHY SYLHETI ACCENT

Sloka Ray *et al.* 1966, has reported that Sylheti (dialect of Sylhet region) is one of extremely deviant dialects in Bangladeshi Bangla [2]. Sylheti is also recognized by some of the linguists as a language in its own right [22]. It has extreme diversity mostly on pronunciation and vocabulary and few on grammar among the Bangla language. It also has alternative script called "Sylheti Nagri" and used more Arabic and Persian words compared to Sanskrit as this was mostly used by the Muslim writers of the Sylhet region [23]. The "Sylheti Nagri" script has 5 (five) vowels and 27 (twenty-seven) consonants. Its vowel phonemes are: (i) i, (ii) e, (iii) a, (iv) o, and (v) u [23]. The absence of "æ/ɛ" and "ɔ" vowels in "Sylheti Nagri" script shows an evidence of significant difference in pronunciations of SCBB in the Sylheti accent.

The outline of this article is as follows. In Section II, we are going to discuss the accent Database preparation and describe the experimental setup for accent analysis. The acoustic feature extraction from the accented speech is described in Section III. The most important parts – the experimental results presentation, accent classification and ASR performance on accents are illustrated in Section IV. A discussion on results and how accent analysis is used for robust Bangla speech recognition and a conclusion are presented in Section V.

II. ACCENT DATABASE AND EXPERIMENTAL SETUP

Based on the findings of Sloka Ray *et al.* 1966 [2], we have chosen a group of speakers from a highly deviant dialect and another group of speakers, from some moderately deviant dialects with neutral accent of Bangladeshi Bangla. The first group of speakers are from Sylhet region and the second group of speakers are from North-West and middle part of Bangladesh. Our study hypothesizes that the people from highly deviant dialect (i.e., Sylheti) in Bangladeshi Bangla have a more accented effect on pronunciation of SCBB sentences than the people who have a neutral accent. The people, who have moderate deviant dialects and spent some notable time of their life in Dhaka and suburb of Dhaka, usually have the neutral accented SCBB. Based on the hypothesis, we have prepared our Accent Database (see detail in Section II-A) and done the experimental setup for acoustic analysis and accent classification of the accented speech (see detail in Section II-B).

A. ACCENT DATABASE

The 4 (four) male and 3 (three) female subjects have been chosen from Sylhet region; and 5 (five) male and 3 (three) female subjects from different districts of North-West and middle part of Bangladesh. The speakers have been chosen based on their distinguishable accent. For the Sylheti (SYL) accented group, one could easily find the effect of the regional accent in their speech. In the neutral (NEU) accented group, one could easily recognize the neutral accent of SCBB in their speech data. The subjects, who have been chosen from the Sylhet region, lived all their life in the Sylhet region. The speakers from the neutral accent were raised in their home districts, lived and educated in Dhaka or suburb of Dhaka. The speakers of both groups are undergraduate students of Shahjalal University of Science and Technology (SUST) community. From our previous study on Sylheti and neutral accent, it was found that the steepness of the rise and fall in /E/ vowel's pitch contour differed significantly among the two accent groups [12]. This finding helped us evaluate accent groups at the preliminary stage of the speakers' selection.

1) AUDIO RECORDING

The data corpus has been recorded using "Audacity" (a freeware for digital audio processing and recording) with a USB Audio/MIDI Interface "M-Track 2 × 2" and a Dynamic Microphone in a studio acoustics environment. A recording

TABLE 1. The detail information about the accented corpus.

Items	Values	Male	Female
Speaker Distribution			
Age	18-23 yr.	9	6
Education	Undergraduate	9	6
Accent Category	Neutral	5	3
	Sylheti	4	3
Text Corpus Details			
Total Sentences (Nos.)	9		
Total Words (Nos.)	101		
Total Unique Words (Nos.)	69		
Total Unique Vowel Phonemes (Nos.)	12		
Total Unique Monophthongs (Nos.)	7		
Total Unique Diphthongs (Nos.)	5		
Speech Corpus Details			
No. of Utterances (Nos.)	135	81	54
Accent-wise Annotated Vowels (Nos.)	Neutral	460	276
	Sylheti	368	276
Total Annotated Vowels (Nos.)	1380	828	552
Average Duration of the Utterance (sec.)	4.11	4.34	3.77
Total Duration (sec.)	555.25	351.51	203.74

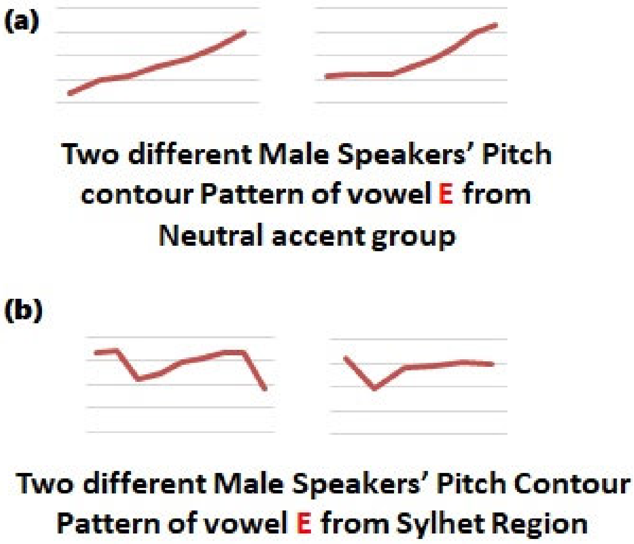


FIGURE 2. The examples of অ্যা/ɛ/E/ vowel's pitch contour patterns for male speakers from two accent groups.

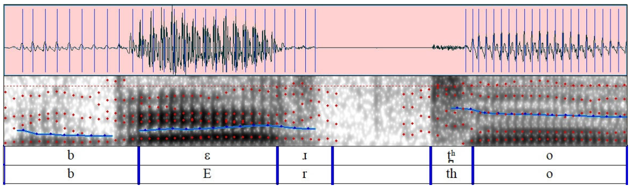


FIGURE 3. Annotation example of using *Praat* for the word "ব্যর্থ" (/bɛɹt̪ʰo/, /bErtho/) of a male speaker from North-west part of Bangladesh.

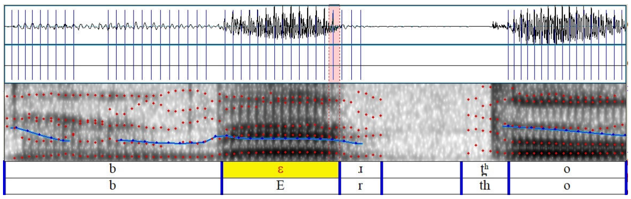


FIGURE 4. Annotation example of using *Praat* for the word "ব্যর্থ" (/bɛɹt̪ʰo/, /bErtho/) of a male speaker from Sylhet region.

script has been chosen with SCBB sentences. A story of nine sentences, which have been used in several research papers to study phonetics of standard Bangladeshi Bangla language [1], [19], has been chosen for recoding. All recorded speech has been sampled at 16 kHz. To maintain the recording and the voice quality, the recorded speech was double-checked manually, once during the recording and then during the annotation. The script has been recorded in a single session with each utterance of a single sentence. The recordings have been perceptually evaluated by a number of native-Bangla speakers, who are accustomed of both the Sylheti and neutral Bangladeshi Bangla accent. Because, research on listener's accent perception has shown that the listener's own accent fluency and a posteriori knowledge of specific accent affects his or her accent perception [37]. Furthermore, we have also evaluated the steepness of the rise and fall in vowels' pitch contours after collecting the speech [10], [12], [13] (see example pitch contours are in the Figure 2). The detail of the recorded corpus is given in Table 1.

2) SEGMENTATION AND ANNOTATION

The recorded data has been segmented and annotated in the following ways:

(a) Phoneme-wise
(b) Syllable-wise
(c) Word-wise
(d) Sentence-wise

Two types of phonetic transcription systems are used for annotation, *Bangladeshi Standard IPA transcription* [1] and *B-ToBI (Bengali Tones and Break Indices System) transcription* [24]. Both types of transcriptions have been used for phoneme and syllable level annotation. "*Praat*" (*Version 6.0.19*) [31], a well-known speech analysis and processing software, has been used for segmentation and annotation. The vowel phonemes carry more temporal detail in the acoustic features than the consonants [7]. The accent mainly affects the vowel-related acoustic and prosodic features [7]. It is required to extract and analyze the vowels' features to evaluate the effect of regional-accents variation on the standard accent [6]–[11]. So, the vowel phonemes have been manually segmented using the "*Praat*" [31]. Figures 3 and 4 show the examples of phoneme level annotation for the Bangla word " ব্যর্থ" (/bɛɹt̪ʰo/, /bErtho/) using "*Praat*" for two distinct male speakers from both accents. The chosen paragraph has covered all the monophthongs and 5 (five) diphthongs of Bangladeshi Bangla. In the script, the monophthongs were more frequent than the diphthongs. The script contained 101 words and out of these words 53 were chosen for the segmentation, labeling and extracting vowel acoustic features. Tokenization has been done in level by level. Initially we have tokenized the 9 recorded sentences, then the chosen

53 words were tokenized. After that, these 53 words were tokenized in 108 syllables and 92 vowels. The occurrences of monophthongs were from 4 to 18 times while the occurrences of diphthongs were from 1 to 3 times. Only the diphthongs, which occurred more than twice, i.e., আও/ao̯/, এই/ej/, অয়/ɔe̯/, and ওই/oj/ were considered here for accent analysis. The analyzed vowels with the corresponding words are listed in Table 2.

B. EXPERIMENTAL SETUP

For statistical analysis the mean of F1, F2 and F3, the pitch slope, and the duration of the vowels have been arranged word-wise, speaker-wise, accent-wise and then the vowel-wise. The analysis was done using the statistical toolbox of *MS Excel 2016*. For both accents we have calculated the mean, the standard deviation and the variance of each vowel's acoustic-features (i.e. formant frequencies and phone durations). We also calculated the vowel-wise p-value of the one-tailed t-test and the two-tailed t-test across the two accents. Likewise, the accented vowel-wise mean of the pitch slope also has been calculated.

Furthermore, the temporal details of the extracted vowels' features have been saved as CSV ("comma-separated values") files for ML-based analysis. *Python 2.7* based machine learning platform the "*GraphLab Create v2.1*" was used to apply ML methods i.e., Linear Classification, Support Vector Machine (SVM), Decision Tree (DT) and Nearest Neighbor Classifier (NNC) for accents classification [25]. Similarly, the *GraphLab Create* - based Linear Regression method has been used to generate the formants contour from the temporal detail of the formants.

III. ACCENTED ACOUSTIC FEATURE EXTRACTION

Researchers have found a significant effect of accent on vowels' formant frequencies F1, F2, and F3, the pitch slope and phone duration [6]–[11]. In this study, these vowels' features have been extracted using the "*Praat*" [31]. By applying the Gaussian-like window and computing the LPC (Linear Predictive Coding) coefficients with Burg's algorithm [26], the *Praat* [31] performs a frame-wise short-term spectral analysis for tracking the formants. We used two types of *Praat* settings for Male and Female speakers, as shown in Table 3.

To differentiate the silence, voiced and unvoiced frames the voicing and silence threshold of *Praat* were set. For each vowel utterance, the mean value of the formants' frequencies has been measured from the several repeated frames for comparing the correlation between the accents. Depending on the vowel duration, 6-12 repeated frames were considered avoiding carefully the adjacent consonant's effect on that vowel utterance.

IV. ACCENTED FEATURES ANALYSIS

The differentiating features of the vowel sounds can be associated with the differences in their first three formant frequencies [28]. Articulation manners of vowels are uniquely different for every accent; so the vowel duration also differs among the accents. Prosodic feature, i.e. pitch contour, reflects the regional accent background. These above mentioned acoustic and prosodic parameters have been utilized in

TABLE 2. List of words that are considered in the accent database for analysis.

Bangla Words	IPA Transcriptions	B-ToBI Transcriptions	Extracted Vowels in Bangla/IPA/B-ToBI
একদিন	ɛk d̪in	Ek din	অ্যা/ɛ/E; ই/i/i
উত্তর	ut̪ːoɹ	ut.tor	উ/u/u; ও/o/o
হাওয়া[a]	hao̯a	haWa	আও/ao̯/aW; আ/a/a
এবং	eboŋ	eboG	ও/o/o
সূর্য[a]	ʃuɹʥo	Surzo	উ/u/u; ও/o/o
তর্ক	t̪ɔɹko	tOrko	অ/ɔ/O; ও/o/o
করছিল	koɹʨʰilo	korchilo	ও/o/o; ই/i/i
তাদের[a]	t̪ad̪eɹ	tader	আ/a/a; এ/e/e
মধ্যে	mod̪ʱːe	modh.dhe	ও/o/o; এ/e/e
কে	ke	ke	এ/e/e
বেশি	beʃi	beSi	এ/e/e; ই/i/i
শক্তিমান	ʃokt̪iman	Soktiman	ও/o/o; ই/i/i; আ/a/a
সেই	ʃej	Sey	এই/ej/ey
মুহূর্তে	muhuɹt̪e	muhurte	উ/u/u; এ/e/e
ভারী	bʱaɹi	bhari	আ/a/a; ই/i/i
চাদর	ʨad̪oɹ	cador	আ/a/a; ও/o/o
পরা	pɔɹa	pOra	অ/ɔ/O; আ/a/a
একজন	ɛkʥon	Ekjon	অ্যা/ɛ/E; ও/o/o
পথিক	pot̪ʰik	pothik	ও/o/o; ই/i/i
দিকে	d̪ike	dike	ই/i/i; এ/e/e
হেঁটে	hete	heTe	এ/e/e
আসেন	aʃen	aSen	আ/a/a; এ/e/e
আর	aɹ	ar	আ/a/a
রাজি	ɹaʥi	razi	আ/a/a; ই/i/i
হয়	hɔe̯	hOY	অয়/ɔe̯/OY
তাকেই	t̪akej	takey	আ/a/a; এই/ej/ey
হিসেবে	hisebe	hisebe	ই/i/i; এ/e/e
করা	kɔɹa	kOra	অ/ɔ/O
হবে	hɔbe	hObe	অ/ɔ/O
এরপর	eɹ pɔɹ	er pOr	এ/e/e; অ/ɔ/O
সব	ʃɔb	SOb	অ/ɔ/O
বইতে	bojt̪e	boyte	ওই/oj/oy; এ/e/e
শুরু	ʃuɹu	Suru	উ/u/u
করে	kɔɹe	kOre	অ/ɔ/O; এ/e/e
কিন্তু	aʃen	aSen	ই/i/i; উ/u/u
যতই	ʥɔt̪oj	zOtoy	অ/ɔ/O; ওই/oj/oy
বয়	bɔe̯	bOY	অয়/ɔe̯/OY
ব্যর্থ	bɛɹt̪ʰo	bErtho	অ্যা/ɛ/E
হয়ে	hoe	hoe	ও/o/o; এ/e/e
বন্ধ	bɔnd̪ʱo	bOndho	অ/ɔ/O; ও/o/o
গরম	gɔɹom	gOrom	অ/ɔ/O; ও/o/o
তাপ	t̪ap	tap	আ/a/a
ছড়ায়	ʨʰɔɹae̯	chORaY	অ/ɔ/O
সঙ্গে	ʃɔŋge	SOGge	অ/ɔ/O; এ/e/e
খুলে	kʰule	khule	উ/u/u
ফেলে	fɛle	fEle	অ্যা/ɛ/E
অবশেষে	ɔboʃeʃe	OboSeSe	অ/ɔ/O; ও/o/o; এ/e/e
নিতে	nit̪e	nite	ই/i/i
দু'জনের	d̪uʥoneɹ	dujoner	উ/u/u
সূর্যই	ʃuɹʥoj	Surzoy	ওই/oj/oy

[a] Here is total 50 words and these marked words are considered twice from different sentences; so from each speaker's speech data - total 53 words are considered for analysis

TABLE 3. ***Praat*** **[31] setting in this study.**

Praat Parameters	Value for Male Speaker	Value for Female Speaker	Common value setting for both gender
Maximum formant frequency	5000 Hz	5500 Hz	-
Pre-emphasis	-	-	50 Hz (So that the emphasis factor =0.96)
Size of the Gaussian window	-	-	25 ms
Pitch (F0) range	75 Hz to 300 Hz	100 Hz to 500 Hz	-
Pitch (F0) Analysis Method			Autocorrelation method
Silence threshold	-	-	0.03
Voicing threshold	-	-	0.45

several research for accent classification [6], [7], [10], [14]. In this study, we have utilized these extracted features of the accented speech for both accent analysis and discrimination.

For the accent analysis, we have used the calculated mean, standard deviation, variance and the one-*tailed* and two-*tailed* *t*-test of all the extracted formants' frequencies for the chosen eleven vowels. The *t*-tests have been utilized to measure the statistical significance on the difference between two accents for F1, F2, F3 and phone duration. To make our formants analysis trustworthy, we have also crosschecked the calculated formant frequencies (F1-F3) from *Praat* with another publicly available formant tracker: that of DPPT (Differential-Phase Peak Tracking) algorithm [40]. For each extracted vowels, the pitch slope of the pitch contour has been calculated from the maximum change of rise and fall in minimum elapsed time. Then the means of the pitch slopes were calculated to analyze the steepness of rise and fall in vowel pitch contour across these two accents. To do the accent discrimination, some feature engineering have been applied on these extracted acoustic and prosodic features.

A. FORMANT FREQUENCIES ANALYSIS

Previous researchers, such as Zheng *et al.* [10], Adank *et al.* [9], Clopper *et al.* [11] and Ladefoged [27], have shown that F1, F2, and F3 formant frequencies have played huge role of holding the most noteworthy information of the vowels for the accent analysis and detection. As Sylheti dialect is one of the most deviant dialect in Bangladesh [2], we have investigated Sylheti accent effect on the pronunciation of SCBB dialogs by the speakers from Sylhet region. We have analyzed the F1, F2, and F3 frequencies to explore the correlation between SYL and NEU accented Bangladeshi Bangla speech.

Every speaker has their own configuration of vocal tracts into various shapes to articulate various types of sound in various accents. During the utterance, the resonant frequencies of the vocal tract can be modulated by the articulators such as movement of the palate, various parts of the tongue, the lips, the cheeks and the teeth. The manner of the articulation governs the vocal formants' frequencies. The first three formant frequencies (F1, F2 and F3) are important to understand the sound [10], [28].

The *first formant frequency (F1)* is associated with the jaw opening; the formant frequency rises as the jaw opens wider. So, this formant is interrelated with the vowel height, which is the distance between the tongue and roof of the mouth. The higher F1 frequency represent the open vowel that is lower the vowel height. Vice versa, the lower F1 represents the close vowel and higher vowel height [10], [28].

The *second formant frequency (F2)* is correlated roughly to the shape of the body of the tongue and the tongue advancement. This formant is mostly governed by the frontness or backness of tongue. Higher F2 formant represent the front vowels that is, the tongue body is in the front of the mouth and oral cavity is short. The back vowels have lower F2, because the position of the tongue body is in the back of the mouth, the month is elongated and pharynx is lowered [10], [28].

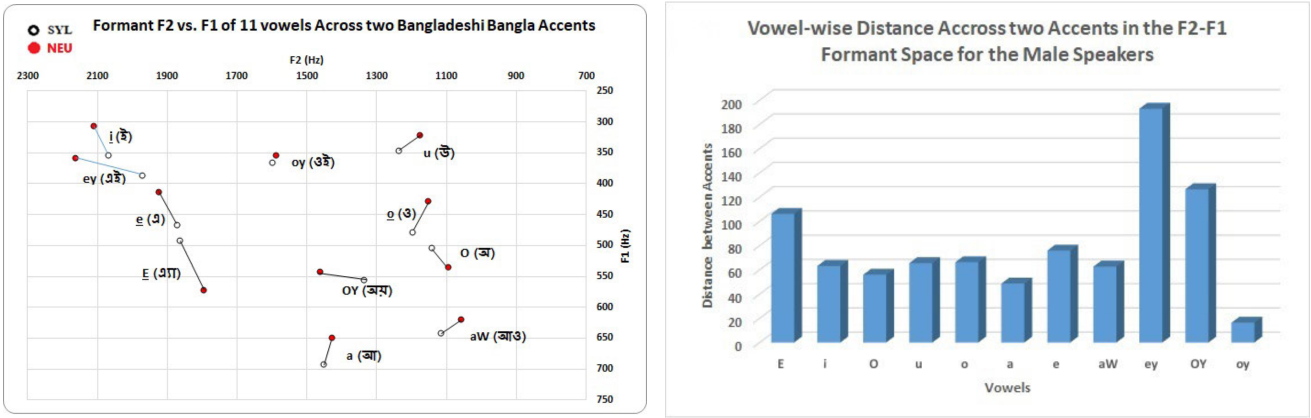
The *third formant frequency (F3)* differs with shape of the lip-rounding and also rest on the position of the vowel production. Higher F3 frequency relate with the rounded shape of the lip. Furthermore, both F2 and F3 are also associated with the lip-rounding and position of the vowel construction [10], [28].

Vowel formant frequencies are significantly different for male and female speakers depending on vowel, language and formant number [29]. So, we have presented our result of the formants analysis for male and female speakers in two different subsections (see Section IV-A1 and Section IV-A2). In Section IV-A3, We have illustrated the crosschecking *Praat* vs. DPPT formant frequencies (F1-F3) to make the formants analysis more reliable.

1) MALE SPEAKERS' FORMANTS ANALYSIS

The Figure 5a shows the eleven vowels distribution in F2-F1 formant space for the male speakers across NEU and SYL accents. Figure 5b shows the bar chart comparison of distance from the NEU accented and SYL accented vowels in F2-F1 formant space. From Figure 5a, it can be seen that the NEU accented /E/ sound is well seperated from the SYL accented /E/ sound. This is consisted with the fact that /E/ phoneme does not exist in Sylheti dialect, and the speakers tend to substitute the /e/ sound in its place. The Figure 5b shows that /E/ vowel has noteable distance between these accent groups. From Table 4, it can be seen that the vowel /E/ has higher F1 value for the NEU accent. The observed *p*-value (<0.0001) from the 1-*tailed* and 2-*tailed* *t*-test also indicates the significant difference in /E/ vowel for the F1 frequency. The SYL accented /e/ sound also differs from the NEU accented /e/ sounds. The Figure 5a shows that the SYL accented /e/ sound and /E/ sound are almost similar. Furthermore, in Table 4, *p*-value (<0.0001) from the 1-*tailed* and 2-*tailed* *t*-test suggest that both of these sounds significantly differed from the NEU accented sounds and /e/ vowel has higher F1 value in the SYL accent.

According to scatter plot of vowels in Figure 5a and bar chart of Figure 5b, it can also be seen that the NEU accented



(a) Scatter plot of male speakers' vowels distribution

(b) Bar chart of the distance of individual vowel phonemes among accents in the F2-F1 formants space

FIGURE 5. Formants F2 vs. F1 for the chosen 11 vowels of Bangladeshi Bangla (5a) scatter plot of male speakers' vowels distribution in the F2-F1 formant space across SYL and NEU accent (5b) comparison of the distance of individual vowel phonemes in the F2 vs. F1 formant space for the accented speech of male speakers.

TABLE 4. Mean, Standard deviation of F1 formant frequency and *p*-value of the 1-*tailed* and 2-*tailed t*-test of eleven vowels across two accents – SYL and NEU for male speakers.

Vowels	F1 frequency (Hz)					*t*-test	
	Mean			STD		*p*-value	
B-ToBI	SYL	NEU	SYL/NEU	SYL	NEU	1-*tailed*	2-*tailed*
E	491.02	571.02	0.86	39.86	61.45	<0.0001	<0.0001
i	353.50	306.56	1.15	54.21	37.92	<0.0001	<0.0001
O	502.86	533.88	0.94	63.72	72.99	0.0079	0.0157
u	345.28	321.25	1.07	48.68	42.32	0.0079	0.0159
o	477.55	427.17	1.12	68.12	72.39	<0.0001	<0.0001
a	691.85	649.13	1.07	84.35	96.11	0.0072	0.0145
e	465.40	412.16	1.13	69.17	57.06	<0.0001	<0.0001
aW	641.27	619.48	1.04	63.26	58.39	0.2322	0.4645
ey	385.56	357.05	1.08	23.89	29.17	0.0184	0.0368
OY	554.58	543.09	1.02	46.19	57.54	0.3333	0.6666
oy	364.98	352.81	1.03	25.94	30.32	0.1414	0.2828

TABLE 5. Mean, Standard deviation of F2 formant frequency and *p*-value of the 1-*tailed* and 2-*tailed t*-test of eleven vowels across two accents – SYL and NEU for male speakers.

Vowels	F2 frequency (Hz)					*t*-test	
	Mean			STD		*p*-value	
B-ToBI	SYL	NEU	SYL/NEU	SYL	NEU	1-*tailed*	2-*tailed*
E	1865.79	1796.13	1.04	218.02	145.96	0.1416	0.2833
i	2069.66	2111.83	0.98	173.92	202.47	0.1340	0.2680
O	1142.79	1096.24	1.04	154.21	145.69	0.0497	0.0994
u	1237.92	1177.02	1.05	226.29	249.07	0.1142	0.2284
o	1198.47	1155.47	1.04	157.12	191.15	0.0769	0.1538
a	1453.39	1430.20	1.02	136.92	176.66	0.2220	0.4439
e	1872.59	1926.56	0.97	164.77	206.58	0.0331	0.0662
aW	1117.76	1059.04	1.06	85.96	83.86	0.0831	0.1661
ey	1972.90	2163.68	0.91	304.92	177.87	0.0729	0.1457
OY	1338.59	1464.54	0.91	74.48	134.20	0.0120	0.0239
oy	1599.61	1588.69	1.01	220.60	263.93	0.4548	0.9096

diphthongs /ey/ and /OY/ sounds have distinguishable distance in the F2 axis from the SYL accented diphthongs. However, *p*-values from the 1-*tailed* and 2-*tailed t*-test, from Tables 4 – 6, suggest that the difference in /ey/ and /OY/ vowels for the F1, F2 and F3 frequencies are not statistically significant. In this study, we have less samples for the diphthongs (Figure 6 shows the accent-wise no. of samples of the phonemes considered in the acoustic features analysis). The *p*-values from two *t*-tests suggest that there is not sufficient evidence to conclude about the differences in the F2 frequency for these diphthongal vowels.

The scatter plot (see Figure 5a) of the vowel distribution for the male speakers also shows that SYL accented /o/ and /O/ vowels are positioned closely in the F2-F1 formant space.

TABLE 6. Mean, Standard deviation of F3 formant frequency and *p*-value of the 1-*tailed* and 2-*tailed t*-test of eleven vowels across two accents – SYL and NEU for male speakers.

Vowels	F3 frequency (Hz)					*t*-test	
	Mean			STD		*p*-value	
B-ToBI	SYL	NEU	SYL/NEU	SYL	NEU	1-*tailed*	2-*tailed*
E	2477.42	2447.86	1.01	310.90	155.30	0.3659	0.7319
i	2689.88	2664.33	1.01	220.60	204.67	0.2780	0.5559
O	2386.07	2432.40	0.98	282.11	244.22	0.1756	0.3512
u	2490.18	2475.53	1.01	203.45	194.81	0.3651	0.7301
o	2460.03	2529.67	0.97	268.09	259.30	0.0653	0.1306
a	2367.34	2523.97	0.94	252.69	229.42	<0.001	≈0.001
e	2539.82	2540.93	0.99	206.43	157.55	0.4851	0.9702
aW	2433.53	2614.64	0.93	305.48	275.32	0.1061	0.2123
ey	2666.39	2679.96	0.99	277.94	152.14	0.4518	0.9036
OY	2481.81	2444.70	1.02	269.62	147.98	0.3669	0.7338
oy	2528.39	2476.68	1.02	171.20	178.66	0.2315	0.4630

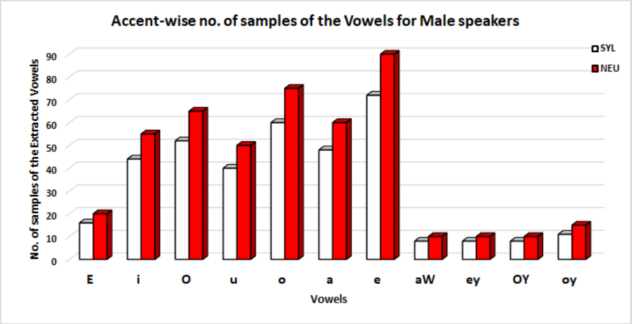


FIGURE 6. Accent-wise no. of samples of the vowel phonemes considered in accent analysis from the accent-database for male speakers.

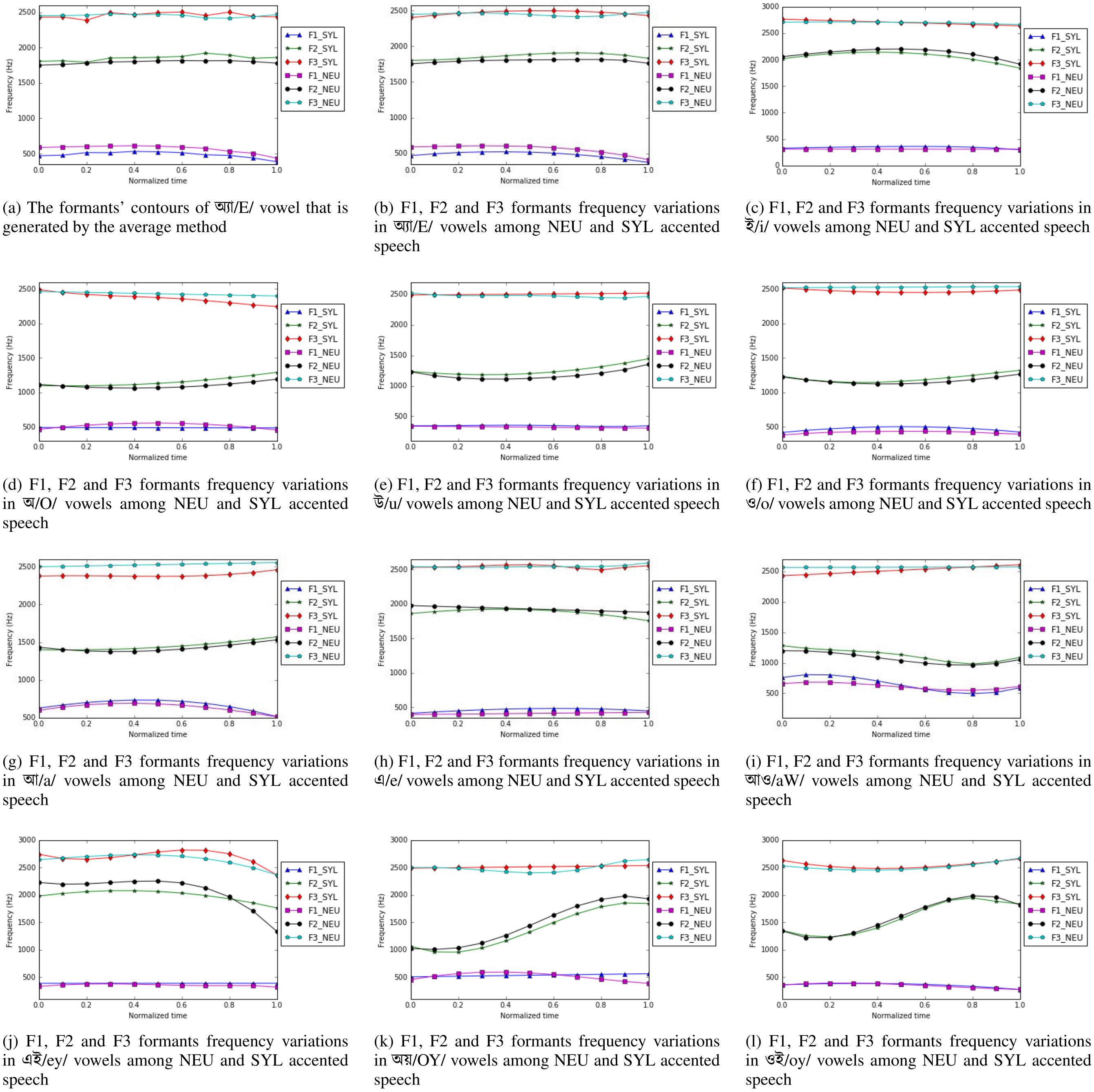
This indicates the fact that SYL accent cannot differentiate these sounds properly. It is not surprising as we know from the Section I-B, that "Sylheti Nagri" has the /o/ vowel but the /O/ vowel is missing. Besides, from the Figures 5a, 5b and Tables 4 – 6, it can be seen that the SYL accented /o/ and /O/ sounds differ from the NEU accented /o/ and /O/ sounds. Also, the *p*-value (<0.0001) from the 1-*tailed* and 2-*tailed t*-test approves that NEU accented /o/ vowel differs significantly in the F1 from SYL accented one.

For the other three vowels /i/, /u/, and /a/, F1 has higher values in accent SYL. The *p*-values (<0.0001) from two types of *t*-test also suggest that there is significant difference in F1 for the /i/ sound (see Table 4). On the other hand, F2 has higher values for /u/ and /a/ in accent SYL while /i/ has higher F2 value in accent NEU. However, it turns out that these difference are not statistically significant (see Table 5). Furthermore, for the /i/ and /u/ vowels have higher F3 in accent SYL and /a/ vowel has higher F3 in accent NEU. The *p*-values of 1-*tailed t*-test indicate that F3 value difference for /a/ vowel is statistically significant (*p*-value is <0.001), but 2-*tailed t*-test give *p*-value≈0.001, which is suggest that there is no sufficient evidence to conclude about the difference in the F3 frequency for the /a/ sound among the accents (see Table 6).

Figure 7a and 7b compares the average method and the linear regression method generated formants contour of /E/ vowel. It can be seen that linear regression method has given better generalized representation of formants contour than the averaged one.

Figure 7 shows the formants frequencies variations in eleven vowels among the accented speech for male speakers. There is not much variation of F1 along the time dimension for the sounds /i/, /u/, /ey/ and /oy/ among these accents. It is seen that during the time of the articulation for the sound /E/ (see Figure 7b), the tongue was raised in accent SYL and lowered in accent NEU (F1 was higher). On the other hand, /o/ sound has the opposite trend (see Figure 7f); the tongue was lowered in accent SYL (F1 was higher) and raised in accent NEU. For the /e/ and /a/ vowel (see Figure 7g and 7h), the tongue has approximately the same position at the beginning of the articulation, but was lowered in the middle then raised again and had a similar position in the end for the SYL accented speech. Whereas, /O/ sound has opposite trend for F1 (see Figure 7d), the F1 was higher in the middle (tongue was lowered) then same again for both accent in the end in accent NEU. For /aW/ sound (see Figure 7i), for NEU accented speech, the F1 was higher (tongue was lowered) from the beginning to normalized time 0.6 then same again for both accents up to the end. Furthermore, for /OY/ sound (see Figure 7k), from normalized time 0.1 to 0.6, the tongue was lowered after that it was started to raise in accent NEU with respect to accent SYL.

From the Figure 7, it can be seen that there is not much variation of F2 along the time dimension up to normalized time 0.4 for the sounds /E/, /O/, /u/, /a/ and /o/ for both accents. After that the tongue was advanced in accent SYL than it was in accent NEU during articulation for these sounds. Whereas from normalized time 0.4 to 1.0, the F2 was higher for /i/ sound in accent NEU (see Figure 7c). Similarly, from normalized time 0.2 to 1.0, the F2 was higher for /OY/ sound in accent NEU (see Figure 7k). Here, F2 is increased indicate that the tongue is further advanced at its maximum point in the mouth in accent NEU. Furthermore, the tongue has a similar position for both accents for /oy/ sound (see Figure 7l). Whereas, for /ey/ sound, F2 was higher at the beginning of the articulation in accent NEU. Then it was



(a) The formants' contours of অ্যা/E/ vowel that is generated by the average method

(b) F1, F2 and F3 formants frequency variations in অ্যা/E/ vowels among NEU and SYL accented speech

(c) F1, F2 and F3 formants frequency variations in ই/i/ vowels among NEU and SYL accented speech

(d) F1, F2 and F3 formants frequency variations in অ/O/ vowels among NEU and SYL accented speech

(e) F1, F2 and F3 formants frequency variations in উ/u/ vowels among NEU and SYL accented speech

(f) F1, F2 and F3 formants frequency variations in ও/o/ vowels among NEU and SYL accented speech

(g) F1, F2 and F3 formants frequency variations in আ/a/ vowels among NEU and SYL accented speech

(h) F1, F2 and F3 formants frequency variations in এ/e/ vowels among NEU and SYL accented speech

(i) F1, F2 and F3 formants frequency variations in আও/aW/ vowels among NEU and SYL accented speech

(j) F1, F2 and F3 formants frequency variations in এই/ey/ vowels among NEU and SYL accented speech

(k) F1, F2 and F3 formants frequency variations in অয়/OY/ vowels among NEU and SYL accented speech

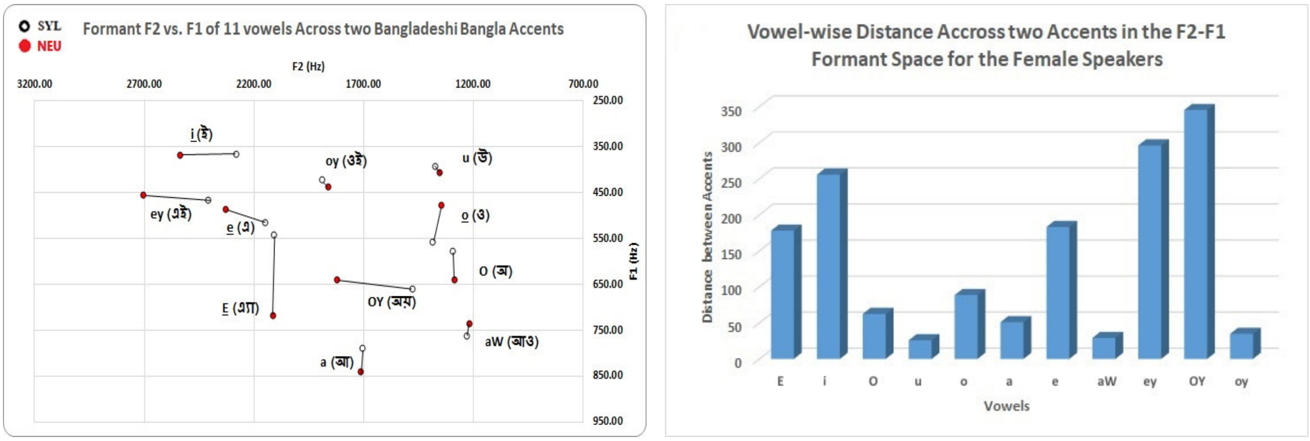
(l) F1, F2 and F3 formants frequency variations in ওই/oy/ vowels among NEU and SYL accented speech

FIGURE 7. F1, F2 and F3 formants frequency variations in eleven vowels among NEU and SYL accented speech for male speakers.

decreased from normalized time 0.7 than it was in accent SYL (see Figure 7j). On the other hand, for /aW/ sound, F2 was higher at the beginning of the articulation in accent SYL. Then it was same for both accents from normalized time 0.8 (see Figure 7i). For /e/ sound (see Figure 7h), F2 was lower at the beginning of the articulation. Then it was same for both accent from normalized time 0.3 to 0.7. From normalized time 0.7 to 1.0, it was lower again in accent SYL. Here, F2 is lower means that the tongue is less advanced.

2) FEMALE SPEAKERS' FORMANTS ANALYSIS

In the Figures 8a, 8b and the Tables 7 – 9, the formants analysis of the accented speech from the female speakers have been presented extensively. The Figure 8a shows the eleven vowels distribution in F2-F1 formant space in a scatter plot across two accents of Bangladeshi Bangla. However, the Figure 8b shows the horizontal bar graph comparison of distance among SYL versus NEU accented vowels in F2-F1 formant space. From the Figures 8a, 8b and the Tables 7 – 9, it can be seen that there are almost similar trend of accent



(a) Scatter plot of female speakers' vowels distribution

(b) Bar chart of the distance of individual vowel phonemes among accents in the F2-F1 formants space

FIGURE 8. Formants F2 vs. F1 for the chosen 11 vowels of Bangladeshi Bangla (8a) scatter plot of female speakers' vowels distribution in the F2-F1 formant space across SYL and NEU accent (8b) comparison of the distance of individual vowel phonemes in the F2 vs. F1 formant space for the accented speech of female speakers.

TABLE 7. Mean, Standard deviation of F1 formant frequency and *p*-value of the 1-*tailed* and 2-*tailed* *t*-test of eleven vowels across two accents – SYL and NEU for female speakers.

Vowels	F1 frequency (Hz)					*t*-test	
	Mean			STD		*p*-value	
B-ToBI	SYL	NEU	SYL/NEU	SYL	NEU	1-*tailed*	2-*tailed*
E	539.56	717.48	0.75	49.55	130.68	<0.001	<0.001
i	365.10	365.84	0.99	65.04	55.95	0.4802	0.9604
O	576.08	638.20	0.90	81.76	130.79	0.0072	0.0144
u	390.57	405.21	0.96	63.81	87.92	0.2317	0.4635
o	556.15	477.04	1.17	105.92	113.59	<0.001	<0.001
a	788.40	838.80	0.94	119.52	134.56	0.0487	0.0974
e	513.73	483.91	1.06	84.05	107.99	0.0562	0.1125
aW	760.46	733.68	1.04	101.19	88.35	0.3180	0.6360
ey	465.54	453.78	1.03	81.17	62.78	0.3925	0.7850
OY	659.47	637.62	1.03	134.29	105.67	0.3805	0.7609
oy	420.02	435.85	0.96	45.55	84.32	0.3145	0.6290

TABLE 8. Mean, Standard deviation of F2 formant frequency and *p*-value of the 1-*tailed* and 2-*tailed* *t*-test of eleven vowels across two accents – SYL and NEU for female speakers.

Vowels	F2 frequency (Hz)					*t*-test	
	Mean			STD		*p*-value	
B-ToBI	SYL	NEU	SYL/NEU	SYL	NEU	1-*tailed*	2-*tailed*
E	2112.60	2116.47	≈1.00	275.92	186.79	0.4842	0.9684
i	2283.68	2539.27	0.90	365.32	204.10	<0.001	<0.001
O	1294.62	1288.85	≈1.00	196.03	204.15	0.4495	0.8990
u	1376.47	1355.26	1.02	326.93	318.99	0.4000	0.8001
o	1386.18	1346.30	1.03	263.05	259.70	0.2355	0.4711
a	1706.41	1714.00	≈1.00	170.55	236.15	0.4382	0.8763
e	2149.79	2330.35	0.92	256.07	253.92	<0.001	<0.001
aW	1231.76	1220.48	1.01	148.12	98.78	0.4401	0.8801
ey	2409.85	2705.78	0.89	184.20	94.26	0.0045	0.0090
OY	1478.88	1823.47	0.81	111.20	233.71	0.0067	0.0134
oy	1892.20	1861.16	1.02	295.52	364.69	0.4227	0.8453

effect in the female speakers' vowels distribution in the formants space for the NEU and SYL accented speech like as the Male speakers' accent analysis results in the previous section (see Section IV-A1).

From the vowels distribution in the F2-F1 formants space (in Figure 8a), it can be seen that the NEU accented /E/ and /e/ sounds are well segregated from the SYL accented of these sounds. /E/ vowel has changed in F1 and /e/ has primarily changed in the F2 for NEU versus SYL accent. The difference of means among the accents in the F1 space for /E/ vowel is also statistically significant (*p*-values of 1-*tailed* and 2-*tailed* *t*-test are <0.001 – see Table 7). A similar trend has been

TABLE 9. Mean, Standard deviation of F3 formant frequency and *p*-value of the 1-*tailed* and 2-*tailed t*-test of eleven vowels across two accents – SYL and NEU for female speakers.

Vowels	F3 frequency (Hz)					*t*-test	
	Mean			STD		*p*-value	
B-ToBI	SYL	NEU	SYL/NEU	SYL	NEU	1-*tailed*	2-*tailed*
E	2816.02	2785.53	1.01	203.96	166.83	0.3463	0.6925
i	2940.56	3095.07	0.95	299.80	198.94	0.0084	0.0167
O	2765.56	2776.80	0.98	349.82	234.30	0.4341	0.8681
u	2746.90	2909.08	0.94	300.04	274.47	0.0165	0.0330
o	2956.66	2941.72	1.01	309.90	222.64	0.3967	0.7935
a	2830.72	2772.17	1.02	218.47	203.64	0.1218	0.2435
e	2887.47	2935.75	0.98	276.72	154.95	0.1333	0.2665
aW	2837.28	3064.59	0.93	120.07	209.92	0.0252	0.0505
ey	2935.66	3055.00	0.96	219.88	209.68	0.1794	0.3587
OY	2912.67	2828.56	1.03	250.42	118.67	0.2405	0.4809
oy	2892.66	2925.73	0.99	287.15	161.47	0.3841	0.7682

seen in Male speakers' accent analysis. The Figure 8b shows that /E/ vowel has significant distance among accent groups. It also proves the fact that no /E/ phoneme exists in Sylheti dialect. For the SYL accent, both /E/ and /e/ sounds have closer position in the F2-F1 formant space, which indicates that these speakers tend to substitute the /E/ sound with the /e/ sound. Furthermore, the p-values of two types of *t*-tests (*p*-values of 1-*tailed* and 2-*tailed t*-test are <0.001 – see Table 8) indicate that the difference of means among the accents in the F2 space for /e/ vowel is statistically significant. The bar chart in the Figure 8b shows the distance among accents for the /e/ sound.

From Figure 8a and 8b, it can be also understood that the SYL accented /o/ and /O/ vowels are well separated from the NEU accented one. From literature review in Section I-B, it can be known that between these vowels, Sylheti accent has only /o/ vowel. In Figure 8a, these vowels have closer position in the F2-F1 formant space for the SYL accent. This accepts the fact that articulation manner of these two vowels are similar in SYL accent. Whereas, from the Figures 8a, 8b and Table 7, it can be seen that the NEU accented /o/ and /O/ sounds differ from the SYL accented of these sounds. The *p*-value (<0.001) from the 1-*tailed* and 2-*tailed t*-test confirms that the means difference between NEU and SYL accented /o/ vowel is statistically significant in the F1 (see Table 7). Moreover, SYL accented /o/ sound has higher F1.

The Figure 8a and 8b suggest that the NEU accented diphthongs /ey/ and /OY/ sounds have higher values in the F2 axis from the SYL accented of these diphthongs. They have also notable distance in the F2-F1 space. But, from Tables 7 – 9, *p*-values from the 1-*tailed* and 2-*tailed t*-tests, it can be seen that the means difference in /ey/ and /OY/ vowels for the F1, F2 and F3 frequencies are not statistically significant. Moreover, in this study, these diphthongs have less samples (Figure 9 shows the no. of samples of the phonemes considered in the accent analysis from accented corpus). The *p*-values, from the Tables 7 – 9, indicate that there is not sufficient evidence to conclude about the means differences in the F2 formant space for these diphthongs. Other two diphthongs are closely positioned in the F2-F1 formant space (see Figure 8a). There is also no significant evidence of means difference in the F1, F2 and F3 among the accents for these diphthongs from the Tables 7 – 9.

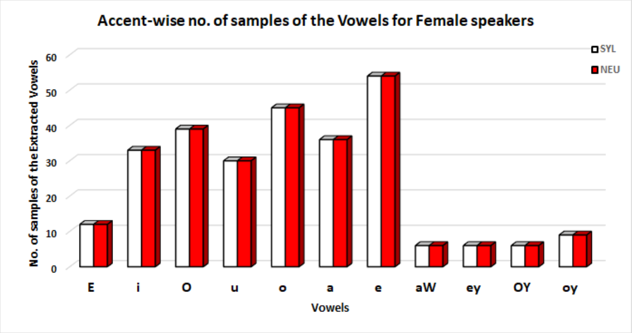
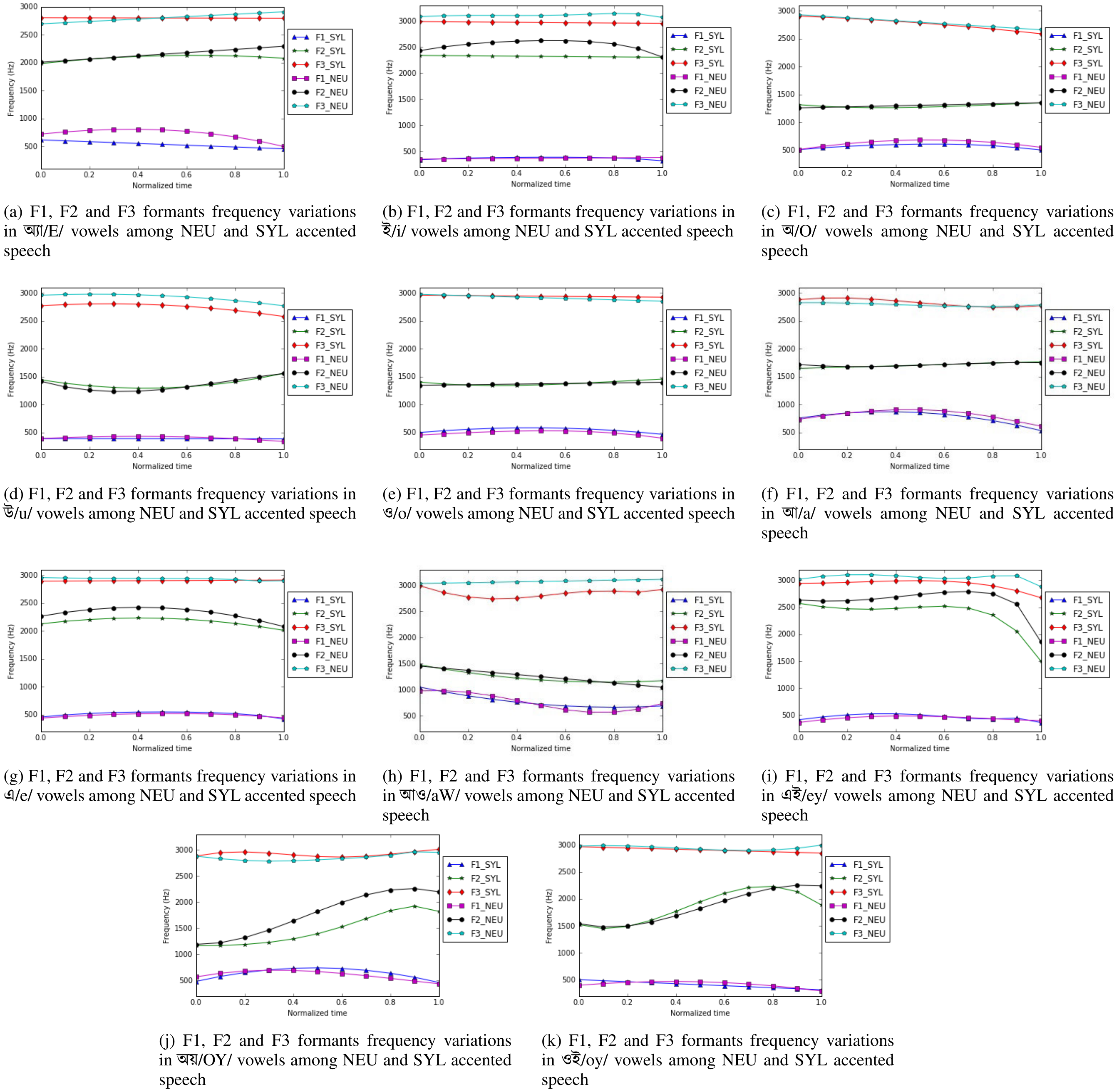


FIGURE 9. Accent-wise no. of samples of the vowel phonemes considered in accent analysis from the accent-database for female speakers.

For the rest of the three vowels /i/, /u/, and /a/, there are no significant difference in the means of F1 among the accents. For the /i/ and /u/ vowels have closer value in F1 among the accents. The /a/ vowel has higher F1 in accent NEU, which is not statistically significant (see Table 7). However, Figure 8a shows that /i/ sound has higher value in the F2 axis in accent NEU. The *p*-values (<0.001) from two types of *t*-test also suggest that there is significant difference in F2 for this sound (see Table 8). Other two sounds, /u/ and /a/ have closer values in F2 across the accents. On the other hand, F3 has higher values for /i/ and /u/ in accent NEU and /a/ has higher F3 value in accent SYL; but these difference are not statistically significant (see Table 9).

Figure 10 shows the formants frequencies' contour variations in eleven vowels among the accented speech for female speakers. There is not much variation of F1 along the time dimension for the sounds /i/, /u/, and /e/ among these accents. Whereas, for the sound /E/ (see Figure 10a), the tongue was raised in accent SYL and lowered in accent NEU (F1 was higher) almost all the time of the articulation. Then the tongue is raised in the end of the articulation in accent NEU. On the other hand, /o/ sound has the opposite trend



(a) F1, F2 and F3 formants frequency variations in অ্যা/E/ vowels among NEU and SYL accented speech

(b) F1, F2 and F3 formants frequency variations in ই/i/ vowels among NEU and SYL accented speech

(c) F1, F2 and F3 formants frequency variations in অ/O/ vowels among NEU and SYL accented speech

(d) F1, F2 and F3 formants frequency variations in উ/u/ vowels among NEU and SYL accented speech

(e) F1, F2 and F3 formants frequency variations in ও/o/ vowels among NEU and SYL accented speech

(f) F1, F2 and F3 formants frequency variations in আ/a/ vowels among NEU and SYL accented speech

(g) F1, F2 and F3 formants frequency variations in এ/e/ vowels among NEU and SYL accented speech

(h) F1, F2 and F3 formants frequency variations in আও/aW/ vowels among NEU and SYL accented speech

(i) F1, F2 and F3 formants frequency variations in এই/ey/ vowels among NEU and SYL accented speech

(j) F1, F2 and F3 formants frequency variations in অয়/OY/ vowels among NEU and SYL accented speech

(k) F1, F2 and F3 formants frequency variations in ওই/oy/ vowels among NEU and SYL accented speech

FIGURE 10. F1, F2 and F3 formants frequency variations in eleven vowels among NEU and SYL accented speech for female speakers.

(see Figure 10e); the tongue was lowered in accent SYL (F1 was higher) and raised in accent NEU. For the /a/ vowel (see Figure 10f and 10g), the tongue has approximately the same position (same value of F1) for both accents at the beginning of the articulation. But, the tongue was raised from the middle and up to the end of the articulation for the SYL accented speech. Whereas, /O/ sound has opposite trend for F1 (see Figure 10c), the F1 was higher from normalized time 0.1 to 0.6 (tongue was lowered) then it was getting lower in the end in accent NEU. But, from beginning to end of articulation SYL accent had lower F1 in /O/ sound. For /aW/ sound (see Figure 10h), for NEU accented speech, the F1 was higher (tongue was lowered) from normalized time 0.1 to 0.6 then it was getting lower up to normalized time 0.9. Furthermore, for /OY/ sound (see Figure 10j), from normalized time 0.1 to 0.6, the tongue was lowered after that it was started to raise in accent NEU with respect to accent SYL.

From the Figure 10, it can be seen that there is not much variation of F2 along the time dimension for the sounds /O/, /a/ and /o/ for both accents. For /E/ sound, the tongue

TABLE 10. ***Praat*** **vs. DPPT Mean of F1 and Standard deviation,** ***p*****-value of the 1-*****tailed*** **and 2-*****tailed t*****-test on DPPT's F1 of four vowels across two accents – SYL and NEU for four male speakers.**

Vowels	F1 frequency (Hz)							*t*-test		Holds the significant difference as observed in *Praat* analysis (see Table 4)
B-ToBI	Mean (*Praat* vs. DPPT)					STD (DPPT)		*p*-value (DPPT)		
	SYL		NEU		SYL /NEU	SYL	NEU	1-*tailed*	2-*tailed*	
	Praat	DPPT	*Praat*	DPPT						
E	476.81	481.08	577.96	585.57	0.82	56.29	40.51	<0.001	<0.001	Yes
i	367.28	363.19	301.29	315.23	1.15	38.25	46.31	<0.001	<0.001	Yes
o	493.78	516.04	433.20	450.90	1.14	60.89	83.44	<0.001	<0.001	Yes
e	485.46	496.54	395.58	414.57	1.19	55.62	60.48	<0.00001	<0.00001	Yes

was advanced in accent NEU than it was in accent SYL during articulation. Whereas, from beginning to end of the articulation, the F2 was higher for /i/ sound in accent NEU (see Figure 10b). For /OY/ sound, from the beginning of the articulation F2 was same for both accents. Then the F2 was getting higher in accent NEU (see Figure 10j). Here, F2 is increased indicate that the tongue is further advanced at its maximum point in the mouth in accent NEU. Furthermore, the tongue has a similar position up to normalized time 0.3 for both accent then F2 was getting higher up to normalized time 0.8 and staring to decreasing up to the end for /oy/ sound in accent SYL (see Figure 10k). Whereas F2 was higher at the beginning of the articulation then it was started to decrease in accent SYL than it was in accent NEU for /ey/ sound (see Figure 10i). On the other hand, F2 was same from the beginning of the articulation then it was getting lower from normalized time 0.8 in accent NEU for /aW/ sound (see Figure 10h). Here, F2 is lower means that the tongue is less advanced. For /e/ sound (see Figure 10g), F2 was higher during the whole articulation in accent NEU. Moreover, for /u/ vowel (see Figure 10d), F2 was lower from the beginning of the articulation in accent NEU then started to increase and has the same pattern from normalized time 0.5 up to end for both accents.

3) EXTRACTED FORMANT FREQUENCIES VERIFICATION

There are several formant trackers available for formant frequencies extraction. Some of the publicly available formant trackers are *Praat* [31], Wavesurfer [41], Winsnoori [42], DPPT [39], [40], [45] etc. These formant trackers are popular and reliable to speech-related clinicians, phoneticians, speech scientists, linguists etc. Most of these formant trackers use LPC-based formant estimation algorithms. *Praat*, Wavesurfer and Winsnoori are example of LPC-based formant tracker [31], [41], [42]. On the other hand, DPPT algorithm use differential phase spectrum processing for formant tracking [40], [45]. So, to validate the *Praat* formant frequencies, we have used the implemented DPPT algorithm form COVAREP (Cooperative Voice Analysis Repository) [39]. COVAREP is a publicly available repository for speech technologies [39]. DPPT is an efficient algorithm for formant tracking. The main advantage of DPPT, it can track high order formants effectively [40], [45]. The reason behind that the differential phase spectra has the spectral tilt-free property [40], [45]. It has been reported that, after comparing the DPPT with the *Praat* formant tracker in synthetic speech, "*the Praat's robustness on analysis of synthetic speech is lowest except for the F1 track*" [40]. They have also compared DPPT with the formant tracker of *Praat* and Winsnoori for the four real speech examples. They found that DPPT was best among the three methods for three of the four examples [40], [45]. It gave worst results among the three formant trackers for the fourth example [40], [45]. Whereas, the formant tracker of *Praat* was more consistent among the three algorithms for all of those four real speech examples [40], [45]. Bozkurt [45] has shown another test result after comparing the DPPT with the *Praat* and Wavesurfer formant trackers for the five male and five female real speech examples. These speech examples have contained Japanese, French, English and Danish sentences [45]. The research has reported that the results of the three formant trackers have provided equivalent and high quality formant tracks on this test set [45].

We have validated *Praat* vs. DPPT formant frequencies (F1-F3) for randomly selected 4 male speakers (2 NEU and 2 SYL accented) and 4 female speakers (2 NEU and 2 SYL accented). For these 8 speakers' accented speech, we have extracted the F1-F3 using the DPPT and compared our previously extracted F1-F3 from the *Praat*. We have only compared formant frequencies, where we have found statistical significant difference in vowels among the accents (see Section IV-A1 and IV-A2). It means that we have compared the result of /E/, /i/, /o/, /e/ vowels. With the *Praat* formant frequencies, we have found the statistical significant difference in F1 for all of these four vowels for male speakers among the accents (see Section IV-A1). Whereas, for female speakers, we have found /E/ and /o/ vowels have significant difference in F1, on the other hand, /e/ and /i/ vowels have significant difference in F2 (see Section IV-A2).

The cross-validation results between *Praa*t vs. DPPT have been presented in Table 10 for male speakers and in Table 11 and 12 for female speakers among the accents. Form Table 10 – 11, it can be seen that the means of F1 of *Praat* vs. DPPT have the closer values for both male and female within the same accent group. Furthermore, in Table 10, the means of F1 in DPPT show that /E/, /i/, /o/ and /e/ vowels have notable distance between the accent groups

TABLE 11. ***Praat* vs. DPPT Mean of F1 and Standard deviation, *p*-value of the 1-*tailed* and 2-*tailed* *t*-test on DPPT's F1 of two vowels (/E/ and /o/) across two accents – SYL and NEU for four female speakers.**

Vowels *B-ToBI*	F1 frequency (Hz)							*t*-test		Holds the significant difference as observed in *Praat* analysis (see Table 7)
	Mean (*Praat* vs. DPPT)					STD (DPPT)		*p*-value (DPPT)		
	SYL		NEU		SYL /NEU	SYL	NEU	1-*tailed*	2-*tailed*	
	Praat	DPPT	*Praat*	DPPT						
E	539.09	514.86	746.84	702.80	0.73	43.02	113.26	<0.001	<0.001	Yes
o	597.53	578.12	496.12	489.59	1.18	88.88	114.96	<0.001	<0.001	Yes

TABLE 12. ***Praat* vs. DPPT Mean of F2 and Standard deviation, *p*-value of the 1-*tailed* and 2-*tailed* *t*-test on DPPT's F2 of two vowels (/i/ and /e/) across two accents – SYL and NEU for four female speakers.**

Vowels *B-ToBI*	F2 frequency (Hz)							*t*-test		Holds the significant difference as observed in *Praat* analysis (see Table 8)
	Mean (*Praat* vs. DPPT)					STD (DPPT)		*p*-value (DPPT)		
	SYL		NEU		SYL /NEU	SYL	NEU	1-*tailed*	2-*tailed*	
	Praat	DPPT	*Praat*	DPPT						
i	2380.11	1493.65	2569.64	1892.17	0.79	300.14	453.27	<0.001	<0.001	Yes
e	2239.39	1679.74	2386.69	2008.57	0.84	297.30	357.47	<0.001	<0.001	Yes

for male speakers. The *p*-values (<0.001) of two types of *t*-test confirm that the difference of means between NEU and SYL accented of these vowels are statistically significant in DPPT's F1. This result is also valid for the *Praat*'s F1 for male speakers. From Table 11, it can be seen that the means of DPPT's F1 for female speakers have significant distance between these accent groups for /E/, and /o/ vowels. The *p*-values (<0.001) of two types of *t*-test approve that the difference of means between NEU and SYL accented of these vowels are statistically significant in DPPT's F1 for female speakers. This result is also valid for the *Praat*'s F1 for female speakers. Whereas, the means of F2 of *Praat* vs. DPPT for female speakers have the notable differences for /i/ and /e/ vowels within the accent groups (see Table 12). Also, DPPT's F2 have significant distance between these accent groups for those vowels (see Table 12). The *p*-values (<0.001) of two types of *t*-test confirm that the difference of means between NEU and SYL accented of these vowels are statistically significant in DPPT's F2 for female speakers. The means in *Praat*'s F2 for female speakers also support the statistical significant differences among these accents for /i/ and /e/ vowels. Table 4 – 9 show the *Praat* extracted formant frequencies analysis and from the comparison, it can be summarized that the formant frequencies achieved from the DPPT has held the similar statistical significant differences that have found with *Praat* extracted formants on these four vowels among the accents.

B. VOWEL PITCH SLOPE ANALYSIS

The prosodic feature, pitch or fundamental frequency (F0) contour is shaped by the several known factors such speaker's regional accent, language background, educational background, socio-economic class, anatomy, and emotional state [4], [10], [13]. During articulation, every language and accent have distinct patterns of intonation of speech that associate with the steepness of the rise and fall in the vowel pitch contour. The previous researches [6], [7], [10], [13] have shown that intonation play an important role to differentiate and investigate the foreign and regional accent influence; because the foreign or regional accented speech has different pitch slope from native or standard accented speech. The pitch slope can be computed by dividing the maximum change in the pitch contour in a minimum time elapsed for the target vowel [13]. The steepness of the rise and fall in vowel pitch contour is represented by the pitch slope. Figure 2 is the example of অ্যা/ε/E/ vowel's pitch contour patterns for male speakers from two accent groups.

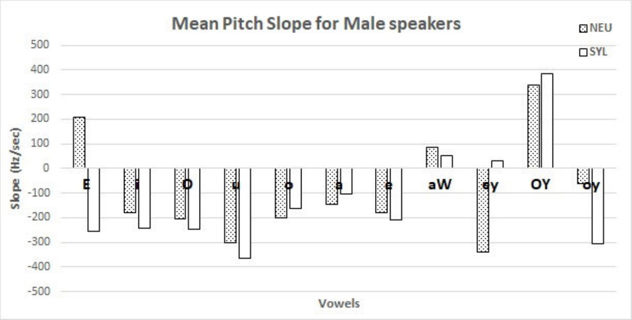


FIGURE 11. **The mean pitch slope of eleven vowel phonemes across two accents for male speakers.**

Figures 11 and 12 show the result of the mean pitch slope analysis of the eleven vowels across SYL versus NEU accents for male and female speakers. For male speakers (see Figure 11), it can be seen that /E/ sound has a negative pitch slope in accent SYL and positive in accent NEU. On the other hand, /ey/ sound has the opposite trend – positive pitch slope in accent SYL and negative in accent NEU. The other vowels' pitch slopes have a similar trend for both accent groups of the male speakers. But the SYL accent group has steeper fall

TABLE 13. Mean, Standard deviation of duration and *p*-value of the 1-*tailed* and 2-*tailed* *t*-test of eleven vowels across two accents – SYL and NEU for male speakers.

Vowels	Vowel Duration (ms)						*t*-test	
B-ToBI	Mean				STD		*p*-value	
	SYL	NEU	SYL-NEU	SYL/NEU	SYL	NEU	1-*tailed*	2-*tailed*
E	81.25	88.03	-6.78	0.92	16.64	17.75	0.1235	0.2471
i	75.01	71.90	3.11	1.04	23.85	29.54	0.2818	0.5635
O	81.83	88.51	-6.68	0.92	26.05	25.49	0.0836	0.1671
u	78.81	76.00	2.82	1.04	28.00	25.84	0.3125	0.6249
o	85.85	82.76	3.09	1.04	25.83	26.33	0.2476	0.4952
a	99.11	97.62	1.49	1.02	28.93	30.05	0.3972	0.7943
e	93.56	87.69	5.87	1.07	37.37	35.91	0.1570	0.3139
aW	154.59	136.95	17.64	1.13	21.29	29.87	0.0818	0.1636
ey	123.25	122.02	1.23	1.01	38.79	20.89	0.4685	0.9369
OY	197.69	201.28	-3.59	0.98	44.47	69.43	0.4480	0.8961
oy	177.48	152.02	25.46	1.17	96.63	31.36	0.2084	0.4169

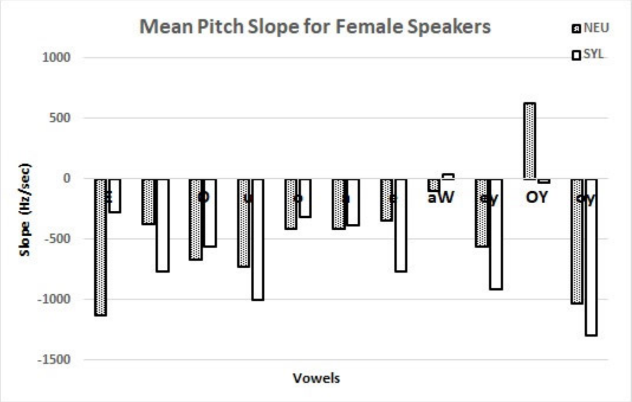


FIGURE 12. The mean pitch slope of eleven vowel phonemes across two accents for female speakers.

for /i/, /O/, /u/, /e/ and /oy/ vowels. Moreover, NEU accent has steeper fall for /o/ and /a/ vowels. For the rest of the two vowels, /aW/ has steeper rise in accent NEU and /OY/ has steeper rise in accent SYL.

For the female speakers, (see Figure 12), it can be seen that /aW/ sound has a positive pitch slope in accent SYL and negative in accent NEU. On the contrary, /OY/ sound has the opposite trend – positive pitch slope and steeper rise in accent NEU and negative in accent SYL. The other vowels' pitch slopes have a similar trend among the accents for the female speakers. From the rest, most of the vowels – /i/, /u/, /e/, /ey/ and /oy/, have steeper fall in accent SYL. Moreover, accent NEU has steeper fall for /E/, /O/ and /o/ vowels. For the /a/ vowel, both accents have almost similar steeper fall in the pitch slope.

C. VOWEL DURATION ANALYSIS

Vowels duration rest on several factors – these are the manner of articulation, stress, speaking style, rhythm, the endpoints of word and syllable, the pause location in utterance, and vowels articulation before a voiced consonant or before the voiceless consonants. For each vowel, every accent has a unique set of the manner of articulation. During the articulation, the shape of the vocal tract, which can be modified by the articulators, cause the variation in the phone duration [10], [28].

Tables 13 and 14 show the mean, standard deviation and *p*-values of *t*-tests of the duration of eleven vowels across SYL versus NEU accents for male and female speakers. For male speakers (see Table 13), /E/, /O/, and /OY/ sounds have been shortened by the range from 3 ms to ≈7 ms for accent SYL. The rest of the eight vowels - /i/, /u/, /e/, /a/, /o/, /ey/, /aW/ and /oy/ have been lengthened by the range from 1 ms to ≈25 ms for accent SYL. On an average, accent NEU has shorter vowel duration. The average durations over all eleven vowels are 113 ms and 110 ms for accent SYL and NEU, respectively. For the SYL accent, /E/ and /O/ sounds have been shortened by a similar margin with a length of ≈7 ms. On the other hand, the SYL accented vowels – /i/, /u/, /e/, /a/, /o/ and /ey/ have been lengthened by a smaller margin with the range of 1.2 to 5.9 ms. Furthermore, /aW/ and /oy/ vowels have been lengthened by a bigger margin with the range of 17.6 to 25.5 ms for the SYL from the NEU accent. The rest of one vowel, /OY/ has been shortened by a smaller margin with a length of ≈3.6 ms for the SYL accent.

On the contrary, for female speakers (see Table 14), /u/ and /oy/ sounds have been lengthened by the near about 2 ms for accent SYL. Furthermore, /E/, /O/, /i/, /e/, /a/, /o/, /ey/, /aW/ and /OY/ sounds have been shortened by the range from 1 ms to ≈21 ms for accent SYL. On an average, accent NEU has longer vowel duration. The average durations over all eleven vowels are 101 ms and 108 ms for accent SYL and NEU, respectively. For the SYL female accent, /E/ and /O/ sounds have been shortened by a similar margin with a length of greater than 14 ms. On the other hand, the SYL accented

TABLE 14. Mean, Standard deviation of duration and *p*-value of the 1-*tailed* and 2-*tailed* *t*-test of eleven vowels across two accents – SYL and NEU for female speakers.

Vowels	Vowel Duration (ms)						*t*-test	
B-ToBI	Mean				STD		*p*-value	
	SYL	NEU	SYL-NEU	SYL/NEU	SYL	NEU	1-*tailed*	2-*tailed*
E	75.06	89.11	-14.05	0.84	24.84	31.23	0.1181	0.2363
i	66.07	67.94	-1.87	0.97	22.33	20.70	0.3624	0.7248
O	74.85	90.72	-15.88	0.82	23.71	26.10	0.0031	0.0063
u	75.69	73.39	2.29	1.03	24.17	21.25	0.3490	0.6979
o	71.87	76.41	-4.53	0.94	16.70	18.09	0.1100	0.2200
a	88.77	95.43	-6.66	0.93	24.40	22.66	0.1170	0.2340
e	82.33	83.31	-0.97	0.99	32.41	27.73	0.4335	0.8670
aW	155.23	175.90	-20.66	0.88	38.33	30.76	0.1642	0.3284
ey	112.21	122.82	-10.61	0.91	30.75	29.34	0.2774	0.5547
OY	173.12	179.32	-6.20	0.97	100.84	22.54	0.4442	0.8884
oy	133.24	131.22	2.02	1.02	32.37	27.87	0.4445	0.8891

vowels – /i/, /e/, /a/, /o/ and /OY/ have been shortened by a smaller margin with the range of 1.0 to 6.7 ms. Furthermore, /aW/ and /ey/ vowels have been shortened by a bigger margin with the range of 10.6 to 20.7 ms for the SYL from the NEU accent. The rest of the two vowels, /u/ and /oy/ have been lengthened by a smaller margin with a length of ≈2 ms for the SYL accent.

D. ACCENT CLASSIFICATION

In this section, the problem of accent discrimination for Bangladeshi Bangla is considered. Here, we have investigated the accent discrimination among the neutral accented speech and the regional accented speech from a highly deviant dialect in Bangladeshi Bangla. From the data analysis and discussion of the previous sections, it can be seen that the acoustic features, i.e., formant frequencies and phone duration, and the prosodic feature, i.e., pitch slope, have been varied in various degrees in the different accented speech. So, the contributions of these acoustic and prosodic features in the accent classification were investigated using the four machine learning algorithms.

For the accent classification experiment, we have only considered the male speakers' data samples. Since 92 nos. of data points of eleven vowels' features were extracted from 9 speakers, there was a total of 828 nos. of data for accented vowels from the male speakers. From the total data samples, 55% of them are NEU accented vowels. The training data (train) set contains ≈69% of the data (i.e., 571 data points); moreover, the cross-validation data (cv) set contains ≈15% of the data (i.e, 125 data points) and the test data (test) set contains ≈16% of the data (i.e, 132 data points). The three sets of features (see Table 15) have been examined using four different ML methods, i.e., Linear Classification, SVM, DT, and NNC from the *GraphLab-Create* toolkit. The several settings of the hyper-parameters have been examined among these ML methods and considered only those settings that have better accuracy on train, cv, and test set and better F1 scores on cv set.

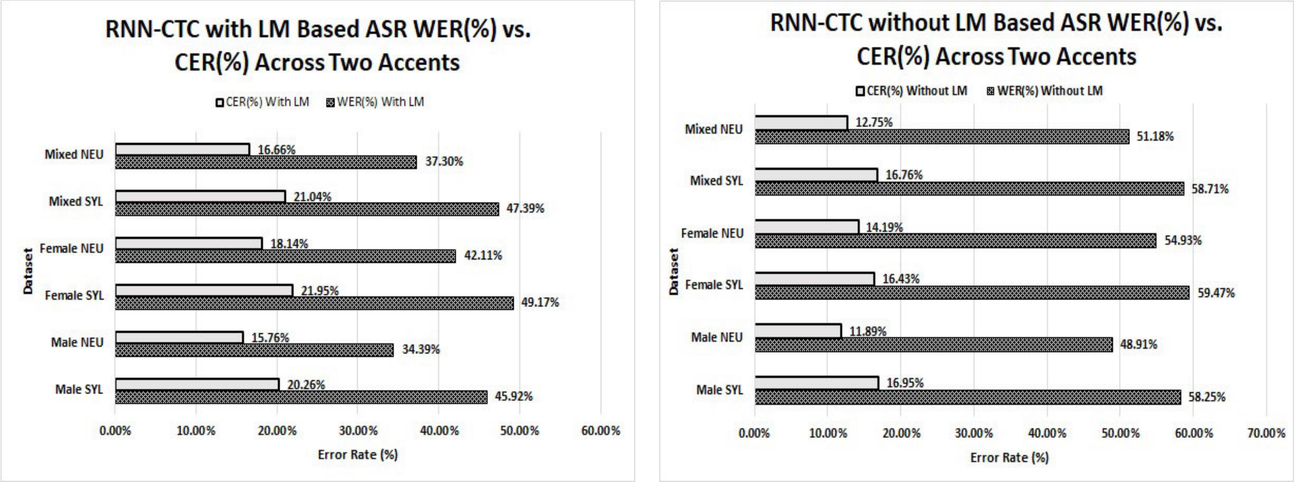
TABLE 15. Features sets for the accent classification.

Feature Sets	No. of features	Features list
Feature Set I	5 Nos.	F1, F2 , F3, Pitch slope, Vowel duration
Feature Set II	16 Nos.	F1, F2 , F3, Pitch slope,Vowel duration, Rise and Fall of the Pitch Slope of the eleven vowels (Rise=+1 and Fall=-1)
Feature Set III	27 Nos.	F1, F2 , F3, Pitch slope, Vowel duration, Rise and Fall of the Pitch Slope of the eleven vowels (Rise=+1 and Fall=-1), Distance of the accented vowel from the NEU accented centroid of the corresponding vowel in the F2-F1 formant space (for 11 vowels)

The Table 16 shows that the linear or logistic classifier has better classification and accents detection with the Feature set-II and has a decent F1 score of 0.68 on test set. Although the SVM has better F1 score of 0.67 on test set for the Feature Set-I & II but it has a better classification and balance accents detection with the Feature Set-III with F1 score 0.63 on test data. Besides, the Nearest Neighbor Classifier has balance accents detection with all of the feature sets, however it has better classification performance on test data with the Feature Set-I. On the contrary, the DT method has both better classification and balance accents detection on test set with all of the feature sets. Furthermore, DT has best classification and accents detection performance on all of the data sets with the Feature Set-II and has the F1 score 0.72 on the test data. From the Tables 15 and 16, it can be also seen that Feature Set-I contain the principal features and all the ML methods have boosted their maximum accuracy of classification and accents detection based on these features. Though other additional features in Set II & III help these ML methods to tune their accuracy for better performance. Moreover, features Set-II contains an additional vector of 11 vowels (1 X 11) that

TABLE 16. Accents classification results.

ML Method	Features set	Hyperparameters	Accuracy on data sets			F1 Score (test data)	Detection	
			Train	CV	Test		NEU	SYL
Logistic Classifier	Feature Set I	With default setting of *GraphLab Create*	57.44%	56.40%	52.27%	0.64	68.75%	26.92%
	Feature Set II		59.37%	55.60%	59.09%	0.68	72.50%	38.46%
	Feature Set III		61.82%	58.40%	53.03%	0.60	57.50%	46.15%
SVM	Feature Set I	$max\_iterations = 200$	54.47%	55.80%	54.55%	0.67	77.50%	19.23%
	Feature Set II		54.82%	58.40%	54.55%	0.67	77.50%	19.23%
	Feature Set III		61.65%	61.60%	53.79%	0.63	65.00%	36.54%
NNC	Feature Set I	With default setting of *GraphLab Create*	64.62%	60.00%	53.79%	0.59	60.00%	48.07%
	Feature Set II		64.97%	57.60%	52.27%	0.60	56.25%	48.07%
	Feature Set III		67.08%	52.80%	49.24%	0.54	52.50%	48.07%
DT	Feature Set I	depth=3	69.00%	59.20%	59.85%	0.69	75.00%	36.54%
	Feature Set II	depth=9	88.62%	60.80%	65.91%	0.72	72.50%	55.77%
	Feature Set III		71.80%	51.20%	61.36%	0.69	71.25%	46.15%



(a) Performance of RNN-CTC based ASR with LM

(b) Performance of RNN-CTC based ASR without LM

FIGURE 13. RNN-CTC based ASR performance on accented speech corpus. (13a) WER(%) and CER(%) of ASR with LM. (13b) WER(%) and CER(%) of ASR without LM.

represent the pitch slope feature (rise = +1 or fall = −1) for a particular vowel, which is corresponding to the data point, and others are set to zero. From the data analysis of Section IV-B, it can be known that the changes (rise or fall) for the pitch slope for the accented vowels have differed among the accents.. From Table 16, it can be seen that with the features Set-II, all these ML methods have balance accents detection with better F1 scores on test data. On the other hand, in the feature Set-III, we have added another additional feature information about the accented vowel distance from the NEU accented centroid of the corresponding vowel in the F2-F1 formant space. By using the feature Set-III, we have achieved better accuracies on the train set for most of the ML methods, but the performances have decreased on the cv and test set.

E. ASR PERFORMANCE ON ACCENTS

We have developed an ASR trained with the "Open SLR – Large Bengali ASR training data" [30] using the starter code of *Deep Speech 2 (DS2)* [36], which is provided by *Baidu Research*. DS2 is an End-to-End deep learning system. The model architecture of the DS2 is based on Recurrent Neural Network and usually trained with Connectionist Temporal Classification loss function (known as RNN-CTC). We have used improved MFCC method [38] for speech feature extraction and trained the RNN-CTC model. Then tested the performance with our accented speech corpus. The Open SLR data set contains ≈196k utterances (i.e. ≈250 hours) of speech. This End-to-End ASR system is released as a web app named "Sukothon" (সুকথন v0.1 Beta) [34] by the Department of CSE, SUST under the HEQEP-CP3888[1] project. Figures 13a and 13b show the performance of RNN-CTC based ASR system with and without language model (LM) on our accented speech corpus. The performance of the Google ASR system (with the Language option - "বাংলা (বাংলাদেশ)", which implies Bangladeshi Bangla language) has also tested with our accented speech corpus.

[1] Higher Education Quality Enhancement Project (HEQEP) (AIF Window 4, CP 3888) for "The Development of Multi-Platform Speech and Language Processing Software for Bangla"

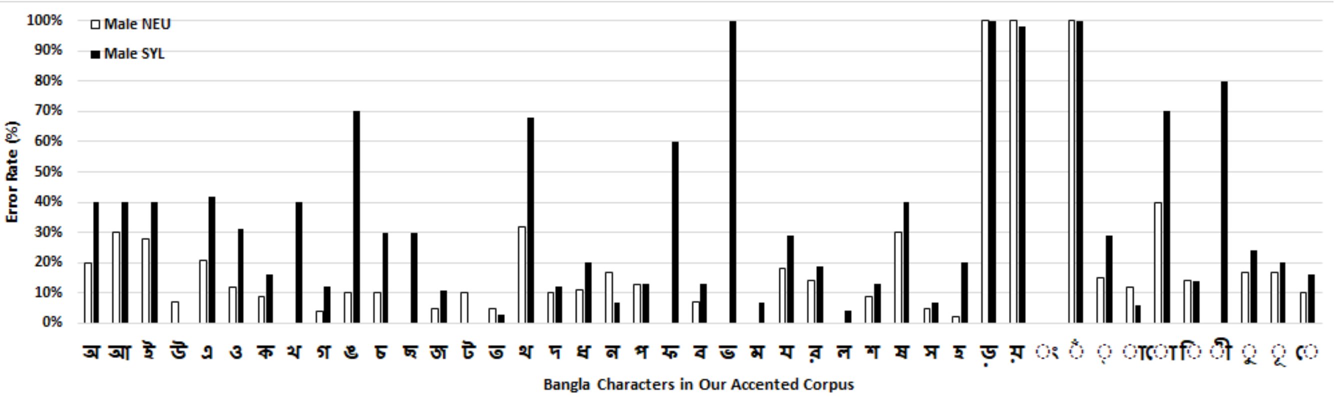


FIGURE 14. The comparison of character-wise recognition error rate (%) of our ASR system without the LM on our accented corpus across two accents for the male speakers.

To evaluate our accented speech corpus with our RNN-CTC based ASR and the Google ASR, we have divided the corpus into 6 (six) datasets –

(a) **Male SYL:** speech data of the SYL accented male speakers
(b) **Male NEU:** speech data of the NEU accented male speakers
(c) **Female SYL:** speech data of the SYL accented female speakers
(d) **Female NEU:** speech data of the NEU accented female speakers
(e) **Mixed SYL:** speech data of the SYL accented speakers from both genders
(f) **Mixed NEU:** speech data of the NEU accented speakers from both genders

Figures 13a and 13b show the performance of our RNN-CTC based ASR system with and without the LM, respectively, on the above mentioned six datasets. From the Figures 13a and 13b, it can be seen that WER (Word Error Rate) and CER (Character Error Rate) of our ASR system on the above mentioned three SYL accented datasets (i.e., *Male SYL, Female SYL,* and *Mixed SYL*) is higher than the NEU accented three other datasets (i.e., *Male NEU, Female NEU,* and *Mixed NEU*). Besides, our ASR system with LM has lower WERs and higher CERs than our ASR system without LM on the respective datasets. The performance of our ASR system with LM (see Figure 13a) on the six accented datasets show that the WERs on the *Male SYL* and *Mixed SYL* datasets are ≈10% higher than the WERs on the *Male NEU* and *Mixed NEU* datasets, respectively. Furthermore, the WER on the *Female SYL* dataset is ≈7% higher than the WER on the *Female NEU* dataset (see Figure 13a). From the Figure 13a, it can be also seen that the CERs on the *Male SYL* and *Mixed SYL* datasets are 4.5% and 4.38% higher than the CERs on the *Male NEU* and *Mixed NEU* datasets, respectively. Furthermore, the CER on the *Female SYL* dataset is 3.81% higher than the CER on the *Female NEU* dataset (see Figure 13a). On the other hand, from the Figure 13b, it can be seen that the WERs of our ASR system without LM on the Male SYL, Female SYL, and Mixed SYL datasets are 9.34%, 4.54%, and 7.53% higher than the WERs on the Male NEU, Female NEU and Mixed NEU datasets, respectively. Moreover, the CERs on the Male SYL, Female SYL, and Mixed SYL datasets are 5.06%, 2.24%, and 4.01% higher than the CERs on the Male NEU, Female NEU, and Mixed NEU datasets, respectively (see Figure 13b). The performance of the Google ASR has been evaluated on the four accented datasets i.e., Male SYL, Female SYL, Male NEU, and Female NEU (see Figure 16). The WERs of Google ASR system on the Male SYL, and Female SYL datasets are 6.15% and 0.12% higher than the WERs on the Male NEU, and Female NEU datasets, respectively (see Figure 16).

We have also examined the character-wise error rate (%) of our ASR system without the LM on the four accented datasets, i.e., *Male SYL*, *Female SYL*, *Male NEU*, and *Female NEU* (see Figures 14 and 15). Each of these four accented datasets contains parallel text (the aligned text with the speech data), which has 43 unique Bangla characters. We have only investigated the output of our ASR without the LM because it can give us the characters' recognition performance of the RNN-CTC part of the ASR system. Through this examination, we can identify the deficit of the corpus that we need to improve for better RNN training. Figures 14 and 15 show the character-wise error rate (%) of our ASR system without the LM for these 43 characters. From Figure 15, it can be seen that for the male speakers, খ/kh/, ছ/ch/, ফ/f/, ভ/bh/, ম/m/, ল/l/, and ◌ী/i/ characters have no error in recognition the NEU accented speech; on the contrary, খ/kh/, ছ/ch/, ফ/f/, ভ/bh/, and ◌ী/i/ characters have higher recognition error rate (RER), whereas ম/m/ and ল/l/ have lower RER in accent SYL. Furthermore, ◌ং/G/ has no recognition error for both accents. The উ/u/ and ট/T/ have no error in recognition of the SYL accented speech, but they have lower RER in accent NEU. On the other hand, ক/k/, গ/g/, জ/j/, দ/d/, ব/b/, শ/sh/, স/s/, and ◌ে/e/ have lower RER with the range from 5% to 10% in accent NEU, but they have a little bit higher RER in

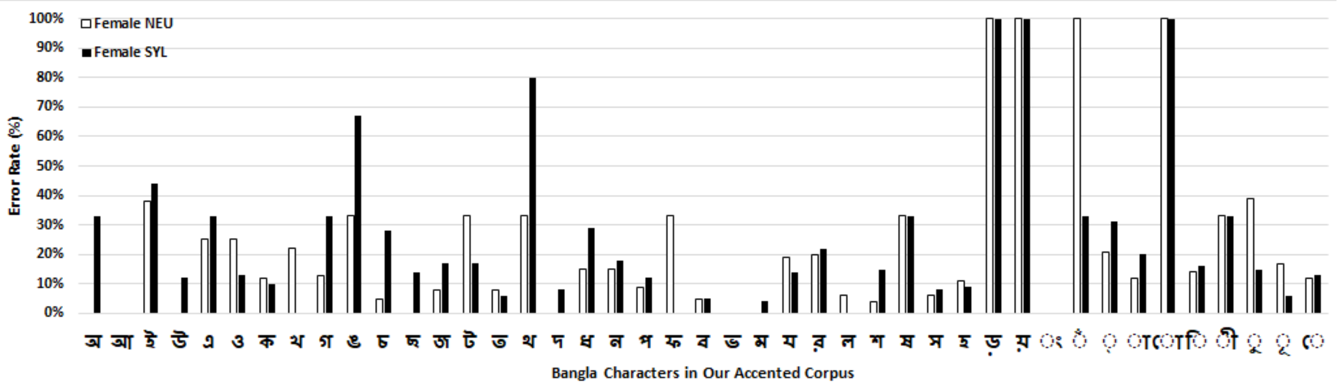


FIGURE 15. The comparison of character-wise recognition error rate (%) of our ASR system without the LM on our accented corpus across two accents for the female speakers.

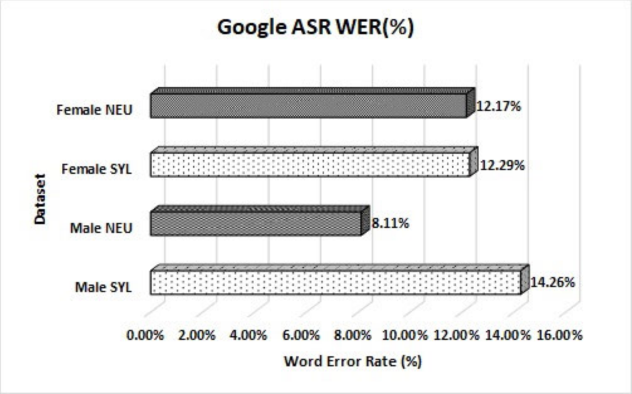


FIGURE 16. Google ASR performance on accented speech corpus.

accent SYL. Though, the characters - অ/O/, এ/e/, ও/o/, ঙ/G/, চ/c/, য/z/, হ/h/, and ্ (use for Geminating or Clustering two consonants) have lower RER with the range from 2% to 18% in accent NEU, but they have higher RER in accent SYL. The ত/t/, ন/n/, and া/a/ have lower RER with the range from 3% to 7% in SYL accented speech, but they have a little bit higher RER in accent NEU. Moreover, for both accents, ধ/dh/, প/p/, র/r/, ি/i/, ু/u/, and ূ/U/ have closer RER; whereas ড়/R/, য়/Y/ and ঁ (use for nasality) characters could not correctly recognize through our ASR. Finally, the rest of the characters have higher RER in both accents; still, SYL accented speech has higher RER.

From Figure 15, it can be seen that for the female speakers, অ/O/, উ/u/, ছ/ch/, দ/d/, and ম/m/ characters have no error in recognition the NEU accented speech; on the contrary, অ/O/ has higher RER and উ/u/, ছ/ch/, দ/d/, and ম/m/ have lower RER in accent SYL. Furthermore, আ/a/, ভ/bh/ and ং/G/ has no recognition error for both accents. The খ/kh/, ফ/f/ and ল/l/ have no error in recognition of the SYL accented speech, but খ/kh/ and ফ/f/ have higher RER and ল/l/ has lower RER in accent NEU. On the other hand, ন/n/, প/p/, শ/sh/, স/s/, া/a/, ি/i/, and ে/e/ have lower RER with the range from 4% to 15% in accent NEU, but they have a little bit higher RER in accent SYL. Though the characters এ/e/, গ/g/, চ/c/, জ/j/, ধ/dh/, and (use for Geminating or Clustering two consonants) have lower RER with the range from 5% to 25% in accent NEU, they have higher RER in accent SYL. The ক/k/, ত/t/, য/z/, and হ/h/ have lower RER with the range from 6% to 14% in SYL accented speech, but they have a little bit higher RER in accent NEU. Moreover, ব/b/, ষ/Sh/, and ী/i/ have similar RER in both accents, ব/b/ has lower and the rest two have higher RER; whereas ড়/R/, য়/Y/ and ো could not properly recognize in both accented speeches. Some the characters, i.e., ও/o/, ট/T/, ু/u/, and ূ/U/ have lower RER with the range from 6% to 17% in accent SYL, but they have higher RER in accent NEU. Furthermore, ঁ (use for nasality) could not recognize in NEU accented speeches and have higher RER in accent SYL. Lastly, the rest of the characters have higher RER in both accents; still, SYL accented speech has higher RER.

From Figures 13 and 16, it can be concluded that both ASR systems perform poor in the Sylheti accent for male speakers. Though the performance of our ASR on accented speech for both gender almost similar trend in accent SYL and better recognition accuracy on neutral accented speech (see Figures 13a and 13b); however, Figure 16 shows that the WER of Google ASR on female speakers from both accents has almost close, but still, SYL accented female speech has a little bit higher WER. On the contrary, Google ASR's performance deteriorates with approx. 6% higher WER on SYL accented speech for male speakers. Overall all the ASR system, which are used in this study, has performed better speech recognition for male speech. Still, from Figure 13, it can be seen that the speech recognition, which has been done by our ASR with the LM (see Figure 13a), has near about 10% higher WER and ≈4.5% higher CER in accent SYL for male speakers. Whereas our ASR without LM (see Figure 13b), there are ≈9% higher WER and ≈5% higher CER in accent SYL for male speakers. Though the overall performance on our accent database for both accents shows that our ASR with the LM system has better WER compares to without the LM system, but still, our ASR without the LM has better CER. On the other hand, our ASR with LM (see Figure 13a) has

≈7% higher WER and ≈4% higher CER in SYL accented speech for female speakers. Similarly, our ASR with LM system (see Figure 13b) has approx. 4.5% higher WER and approx. 2% higher CER in accent SYL for female speakers. From Figures 14 and 15, it can be concluded that most of the characters have higher RER for SYL accented speech. These trends imply that we need a different ASR system for SYL accented people or the people from a highly deviant dialect in Bangladeshi Bangla. The RNN-CTC performance on SYL accent (see Figures 14 and 15) also suggest that we should consider the variabilities of the speakers, which is caused by highly deviant regional dialect, to build a quality speech corpus for the robust LVCSR System in Bangladeshi Bangla.

V. CONCLUSION

In this study, the correlation between the two accents of Bangladeshi Bangla language is examined. The seven monophthongal and four diphthongal vowels of Bangla have been analyzed using the accent-related acoustic, i.e. formant frequencies and vowel durations, and prosodic features, i.e. pitch and pitch slope. The problem of accent classification for Bangladeshi Bangla is also studied. The Neutral accent of Bangla and the deviant Sylheti accent were chosen for this study. The results from the formant frequencies analysis show that the অ্যা/E/ এ/e/, ও/o/ and ই/i/ vowels formant frequencies have a significant difference between these two accents for both genders. The mean pitch slopes and the mean vowel durations of these vowels also differ between these two accents. The results show that NEU accented /E/ sound is well separated from the SYL accented /E/, which is consistent with the fact that /E/ phoneme does not exists in Sylheti dialect. Sylheti dialect has /e/ sound on their vowel phoneme inventory and so the SYL accented speakers tend to substitute the /e/ sound in its place. The paper has also reported that /E/ sound has a significant difference in F1 formant for both genders while /e/ sound has a significant difference in F1 formant for male speakers and in F2 formant for female speakers. The results also show that NEU accented /o/ sound is well segregated from the SYL accented one, and /O/ and /o/ sounds are placed closely in F2-F1 space. This observation is consistent with the fact that SYL accented speakers cannot distinguish these sounds properly. Sylheti dialect has /o/ sound on their vowel inventory, so the SYL accented speakers tend to substitute the /O/ sound with the /o/ sound. The paper has also showed that /o/ sound has a significant difference in F1 formant for both genders. Besides these findings, a new approach has been used to analyze the vocal tract shape in the accented speech more precisely. Instead of average method, linear regression has been used to generate the average contour of the formant frequencies of F1, F2, and F3 for each vowel. Linear regression has given better generalized representation of formants contour than the averaged one.

From the pitch slope analysis, it can be seen that there are lots of difference in the vowels' pitch slope between these two accents. The vowel duration analysis shown that, on the average, compared to accent SYL, accent NEU has shorter vowel duration for male speakers and longer vowel duration for female speakers. Classification results show that the acoustic and prosodic features play a significant role in accent classification. The ASR systems performance suggest the necessity of accent based ASR system for robust speech recognition for Bangladeshi Bangla. Though, we have investigated accent based speakers' variability among NEU and SYL accent on a small accent database. From the F2-F1 space, we have found significant differences in four identical vowels for both genders among the accents. Furthermore, other investigation results have also clarified that SYL accented speech has noteworthy deviant features than that of NEU. So, it can be said, this small dataset helps us make a good assumption that these results are also valid for other people with these dialects. Similarly, after having examined the correlation between the two accents of Bangladeshi Bangla language, it can be concluded that the people from highly deviant dialect (i.e., Sylheti) have a more accented effect on pronunciation of SCBB sentences than the people who have a neutral accent. Therefore, the hypothesis of our study is also proved to be correct (see the hypothesis in Section II). To investigate the regional accent based speakers' variability, many types of research already been performed through many languages (i.e. British English, American English, French etc.) These researches have helped them to build accent based robust speech recognizers and robust LVCSR System. This study coined the requirements of investigating the speech of the people from a deviant dialect in Bangladesh so that the speakers' variability in Bangladeshi Bangla can be considered before developing the speech corpus for a robust LVCSR System. This paper has also reported necessities of accent based Bangla ASR system for the people from an extremely deviant dialect in Bangladesh.

APPENDIX

Sentences used in this study–

Bangla transcription:

1) একদিন উত্তর হাওয়া এবং সূর্য তর্ক করছিল তাদের মধ্যে কে বেশি শক্তিমান।
2) সেই মুহূর্তে ভারী চাদর পরা একজন পথিক তাদের দিকে হেঁটে আসেন।
3) হাওয়া আর সূর্য রাজি হয় তাদের মধ্যে যে পথিকের গায়ের চাদর খোলাতে পারবে তাকেই বেশি শক্তিমান হিসেবে ধার্য করা হবে।
4) এরপর উত্তর হাওয়া তার সব শক্তি দিয়ে বইতে শুরু করে, কিন্তু সে যতই জোরে বয় পথিক তার চাদর চেপে ধরে রাখে।
5) ব্যর্থ হয়ে হাওয়া তার চেষ্টা বন্ধ করে।
6) এর পর সূর্যের পালা।
7) সূর্য তার গরম তাপ ছড়ায়।
8) পথিক সঙ্গে সঙ্গে তার গায়ের চাদর খুলে ফেলে।
9) অবশেষে উত্তর হাওয়া মেনে নিতে বাধ্য হয় যে তাদের দু'জনের মধ্যে সূর্যই বেশি শক্তিমান।

IPA transcription of above sentences:

1) ɛk d̪in ut̪ːoɹ hao̯a eboŋ ʃuɹdʑo t̪ɔɹko koɹtɕʰilo t̪ad̪eɹ mod̪ʱːe ke beʃi ʃokt̪iman.
2) ʃej muhuɹt̪e bʱaɹi tɕad̪oɹ pɔɹa ɛkdʑon pot̪ʰik t̪ad̪eɹ d̪ike hete aʃen.

3) hao̯a aɹ ʃuɹʥo ɹaʥi hɔe̪ t̪ad̪eɹ mod̪ʱːe ʥe pot̪ʰikeɹ gaeɹ ʨad̪oɹ kʰolat̪e paɹbe, t̪akej beʃi ʃokt̪iman hisebe d̪ʱaɹʥo kɔɹa hɔbe.
4) eɹ pɔɹ ut̪ːoɹ hao̯a t̪aɹ ʃɔb ʃokt̪i d̪ie bojt̪e ʃuɹu kɔɹe, kint̪u ʃe ʥɔt̪oj ʥoɹe bɔe̯ pot̪ʰik t̪aɹ ʨad̪oɹ ʨepe d̪ʱoɹe ɹakʰe.
5) bɛɹt̪ʰo hoe hao̯a t̪aɹ ʨeʃta bɔnd̪ʱo kɔɹe.
6) eɹ pɔɹ ʃuɹʥeɹ pala.
7) ʃuɹʥo t̪aɹ gɔɹom t̪ap ʨʰɔɹae̯.
8) pot̪ʰik ʃɔŋge ʃɔŋge t̪aɹ gaeɹ ʨad̪oɹ kʰule fɛle.
9) ɔboʃeʃe ut̪ːoɹ hauoa mene nit̪e bad̪ʱːo hɔe̯ ʥe t̪ad̪eɹ d̪uʥoneɹ mod̪ʱːe ʃuɹʥoj beʃi ʃokt̪iman.

BToBi transcription of above sentences:

1) Ek din ut.tor haWa eboG Surzo tOrko korchilo tader modh.dhe ke beSi Soktiman.
2) Sey muhurte bhari cador pOra Ekjon pothik tader dike heTe aSen.
3) haWa ar Surzo razi hOY tader modh.dhe ze pothiker gaer cador kholate parbe, takey beSi Soktiman hisebe dharzo kOra hObe.
4) er pOr ut.tor haWa tar SOb Sokti die boyte Suru kOre, kintu Se zOtoy jore bOY pothik tar cador cepe dhore rakhe.
5) bErtho hoe haWa tar ceSta bOndho kOre.
6) er pOr Surzer pala.
7) Surzo tar gOrom tap chORaY.
8) pothik SOGge SOGge tar gaer cador khule fEle.
9) OboSeSe ut.tor haWa mene nite badh.dho hOY ze tader dujoner modh.dhe Surzoy beSi Soktiman.

REFERENCES

[1] S. D. Khan, "Illustrations of the IPA Bengali (Bangaldeshi Standard)," *J. Int. Phonetic Assoc.*, vol. 40, no. 2, p. 221–225, 2010.
[2] P. S. Ray, M. Abdul Hai, and L. Ray, *Bengali Language Handbook*. Washington, DC, USA: Center for Applied Linguistics, 1966, pp. 1–3.
[3] *Summary by Language Size, Ethnologue*. Accessed: May 5, 2019. [Online]. Available: https://www.ethnologue.com/statistics/size
[4] J. C. Wells, *Accents of English*. Cambridge, U.K.: Cambridge Univ. Press, 1982.
[5] P. Foulkes and G. J. Docherty, *Urban Voices-Overview*, P. Foulkes and G. J. Docherty, Eds. London, U.K.: Arnold, 1999, p. 1–24.
[6] Q. Yan, S. Vaseghi, D. Rentzos, C.-H. Ho, and E. Turajlic, "Analysis of acoustic correlates of British, Australian and American accents," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2003, pp. 345–350.
[7] L. M. Arslan and J. H. L. Hansen, "Language accent classification in american english," *Speech Commun.*, vol. 18, no. 4, pp. 353–367, Jun. 1996.
[8] L. Loots and T. Niesler, "Automatic conversion between pronunciations of different English accents," *Speech Commun.*, vol. 53, no. 1, pp. 75–84, Jan. 2011.
[9] P. Adank, R. van Hout, and H. V. D. Velde, "An acoustic description of the vowels of northern and southern standard Dutch II: Regional varieties," *J. Acoust. Soc. Amer.*, vol. 121, no. 2, pp. 1130–1141, Feb. 2007.
[10] D. C. Zheng, D. Dyke, F. Berryman, and C. Morgan, "A new approach to acoustic analysis of two British regional accents Birmingham Liverpool accents," *Int. J. Speech Technol.*, vol. 15, p. 77–85, Jun. 2012.
[11] C. G. Clopper, D. B. Pisoni, and K. de Jong, "Acoustic characteristics of the vowel systems of six regional varieties of American English," *J. Acoust. Soc. Amer.*, vol. 118, no. 3, pp. 1661–1676, Sep. 2005.
[12] S. Kibria, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "Acoustic analysis of accent-specific pronunciation effect on Bangladeshi Bangla: A study on Sylheti accent," in *Proc. Int. Conf. Bangla Speech Lang. Process. (ICBSLP)*, Sep. 2018, pp. 1–4.
[13] C. Grover, D. G. Jamieson, and M. B. Dobrovolsky, "Intonation in English, French and German: Perception and production," *Lang. Speech*, vol. 30, no. 3, pp. 277–295, Aug. 2016.
[14] S. Ghorshi, S. Vaseghi, and Q. Yan, "Cross-entropic comparison of formants of British, Australian and American English accents," *Speech Commun.*, vol. 50, no. 7, pp. 564–579, Jul. 2008, doi: 10.1016/j.specom.2008.03.013.
[15] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012, doi: 10.1109/MSP.2012.2205597.
[16] S. K. Chatterji, "Bengali phonetics," *Bull. School Oriental African Stud.*, vol. 2, no. 1, pp. 1–25, Dec. 2009.
[17] A. K. M. Morshed, *Adhunik Bhasatatto (Modern Linguistics)*, 3rd ed. Dhaka, Bangladesh: Mowla Brothers, 2001, pp. 219–237.
[18] A. Hai, *Dhvani Vijnan O Bangla Dhvani-Tattwa*, 10th ed. Dhaka, Bangladesh: Mullick & Brothers, 2007, pp. 12–35.
[19] D. Huq, *Bhasha Bigganer Katha (Facts about Linguistics)*. Dhaka, Bangladesh: Mowla & Brothers, 2002, pp. 81–93.
[20] C. A. Ferguson and M. Chowdhury, "The phonemes of bengali," *Language*, vol. 36, no. 1, pp. 22–59, Jan./Mar. 1960.
[21] F. Alam, S. M. Murtoza Habib, and M. Khan, "Acoustic analysis of Bangla vowel inventory," Centre Res. Bangla Lang. Process., BRAC Univ., Dhaka, Bangladesh, Tech. Rep. 643, 2007. [Online]. Available: http://hdl.handle.net/10361/643
[22] S. M. Rasinger, *Bengali-English in East London: A Study in Urban Multilingualism*. Bern, Switzerland: European Academic Publishers, 2007, pp. 26–27.
[23] S. Nagri. *Banglapedia: National Encyclopedia of Bangladesh*. Accessed: May 23, 2019. [Online]. Available: http://en.banglapedia.org/index.php?title=Sylheti_Nagri
[24] S. D. Khan, "The intonational phonology of Bangladeshi standard Bengali," in *Prosodic Typology II: The Phonology of Intonation and Phrasing*, S.-A. Jun, Ed. Oxford Scholarship Online, Apr. 2014, doi: 10.1093/acprof:oso/9780199567300.001.0001.
[25] *Turi Machine Learning Platform User Guide*. Accessed: May 31, 2019. [Online]. Available: https://turi.com/learn/userguide/
[26] D. J. Childers, *Modern Spectrum Analysis*. New York, NY, USA: IEEE Press, 1978, pp. 252–255.
[27] P. Ladefoged, *Elements of Acoustic Phonetics*, 2nd ed. Chicago, IL, USA: The Univ. of Chicago Press, 1996.
[28] C. R. Nave. (2016). HyperPhysics. Department of Physics and Astronomy. Georgia State Universit. Atlanta, GA, USA. Accessed: Jun. 11, 2019. [Online]. Available: http://hyperphysics.phy-astr.gsu.edu/hbase/hframe.html
[29] E. Pépiot, "Voice, speech and gender: Male-female acoustic differences and cross-language variation in English and French speakers," in *Proc. XVèmes Rencontres Jeunes Chercheurs de l'ED*, Paris, France, Jun. 2012, pp. 1–13.
[30] *Open SLR Large Bengali ASR Training Data*. Accessed: Jun. 22, 2019. [Online]. Available: http://www.openslr.org/53/
[31] P. Boersma and D. Weenink. (2016). Praat: Doing phonetics by computer [computer program]. Praat—Version 6.0.19. Accessed: Aug. 15, 2016. [Online]. Available: http://www.praat.org/
[32] Audacity. *Audacity: Free, Open Source, Cross-Platform Audio Software, Version 2.1.0*. Accessed: Aug. 1, 2016. [Online]. Available: https://www.audacityteam.org/
[33] *M-Track 2 × 2—USB Audio/MIDI Interface, a Product of M-AUDIO*. [Online]. Available: https://m-audio.com/m-tracks/2x2
[34] Sukhthan. *v0.1 Beta Bangla Speech to Text*. Accessed: Aug. 27, 2019. [Online]. Available: https://stt.sustbanglaresearch.org
[35] *Cloud Speech-to-Text: Language—Bengali (Bangladesh)—A Google Cloud's AI & Machine Learning Product*. Accessed: Aug. 27, 2019. [Online]. Available: https://cloud.google.com/speech-to-text/
[36] D. Amodei *et al.*, "Deep speech 2: End-to-End speech recognition in english and mandarin," 2015, *arXiv:1512.02595*. [Online]. Available: http://arxiv.org/abs/1512.02595
[37] A. Ikeno and J. H. L. Hansen, "The effect of listener accent background on accent perception and comprehension," *EURASIP J. Audio, Speech, Music Process.*, vol. 2007, Dec. 2007, Art. no. 076030.
[38] S. Sigtia and S. Dixon, "Improved music feature learning with deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 109–113.

[39] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP—A collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Florence, Italy, May 2014. [Online]. Available: https://covarep.github.io/covarep/

[40] B. Bozkurt, B. Doval, C. D'alessandro, and T. Dutoit, "Improved differential phase spectrum processing for formant tracking," in *Proc. ICSLP*, Jeju Island, South Korea, Oct. 2004, pp. 1–4. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2004/i04_2421.html

[41] K. Sjãlander and J. Beskow. (2006). *WaveSurfer [Computer Program]*. [Online]. Available: http://www.speech.kth.se/wavesurfer/

[42] Y. Laprie, (2009). *WinSnoori*. [Online]. Available: https://members.loria.fr/ YLaprie/WinSnoori/index.html

[43] D. Jones, *English Pronouncing Dictionary*, 3rd ed. London, U.K.: J. M. Dent & Sons, 1926.

[44] W. R. Van Riper, "General American: An ambiguity," in *Dialect and Language Variation*. New York, NY, USA: Academic, 1986, pp. 123–135, doi: 10.1016/B978-0-12-051130-3.50013-6.

[45] B. Bozkurt, "Zeros of the z-transform (ZZT) representation and chirp group delay processing for the analysis of source and filter characteristics of speech signals," Ph.D. dissertation, TCTS Lab, Faculte Polytechnique Mons, Univ. Mons, City Univ., Mons, Belgium, 2005. Accessed: Dec. 31, 2019. [Online]. Available: https://theses.eurasip.org/media/theses/documents/bozkurt-baris-zeros-of-the-z-transform-zzt-representation-and-chirp-group-delay-processing-for-the-analysis-of-source-and-filter-characteristics-of-speech-signals.pdf

SHAFKAT KIBRIA was born in Sylhet, Bangladesh, in 1978. He received the B.Sc. degree in computer science from East West University, Dhaka, in 2001, and the M.Sc. degree in computer science from Umeå University, Umeå, Sweden, in 2005. He is currently pursuing the Ph.D. degree in computer science with the Shahjalal University of Science and Technology (SUST), Sylhet, Bangladesh.

From September 2006 to October 2007, he was a Lecturer with the Sylhet International University, Sylhet, Bangladesh. From 2007 to 2008, he worked as Ph.D. Research Fellow at an EU Project (DustBot) at the AASS Research Center, Örebro, Sweden. In 2011, he joined Manarat International University (MIU), Dhaka, as a Lecturer. Since 2016, he has been an Assistant Professor with the Department of CSE, MIU, Dhaka, where he is currently in study leave. Since 2016, he has been working as Ph.D. Research Fellow in HEQEP project - "Development of MultiPlatform Speech and Language Processing Software for Bangla" (CP3888 – https://sustbanglaresearch.org/) at the Department of CSE, SUST. His research interests include speech recognition, accent analysis, Bangla speech to text, Bangla natural language processing, human-robot interaction (HRI), machine learning systems (behavior-based robotics), mobile robotics (path planning, outdoor robotics), human–computer interaction (HCI), financial computation and analysis for decision making, and embedded system-based smart and ubiquitous environment.

M. SHAHIDUR RAHMAN (Member, IEEE) was born in Jamalpur, Bangladesh, in 1975. He received the B.Sc. and M.Sc. degrees in electronics and computer science from the Shahjalal University of Science and Technology, Sylhet, Bangladesh, in 1995 and 1997, respectively, and the Ph.D. degree in mathematical information systems from Saitama University, Saitama, Japan, in 2006.

In 1997, he joined the Shahjalal University of Science and Technology as a Junior Faculty, where he is currently serving as a Professor. From 2009 to 2011, he was a JSPS Postdoctoral Research Fellow at Saitama University. His current research interests include speech analysis, speech synthesis, speech recognition, enhancement of bone-conducted speech, and digital signal processing.

M. REZA SELIM (Member, IEEE) was born in Jamalpur, Bangladesh, in 1976. He received the B.Sc. and M.Sc. degrees in electronics and computer science from the Shahjalal University of Science and Technology, Sylhet, Bangladesh, in 1995 and 1996, respectively, and the Ph.D. degree in information and computer sciences from Saitama University, Saitama, Japan, in 2008.

In 1997, he joined the Shahjalal University of Science and Technology as a Junior Faculty, where he is currently serving as a Professor. His current research interests include P2P networks, natural language processing, and machine translation.

M. ZAFAR IQBAL was born in Sylhet, Bangladesh, in 1952. He received the B.Sc. degree in physics from the University of Dhaka, Bangladesh, in 1973, the M.Sc. degree in theoretical physics from the University of Dhaka, in 1974, and the Ph.D. degree in experimental physics from the University of Washington, Seattle, WA, USA, in 1982.

He worked as a Postdoctoral Researcher at the California Institute of Technology (Caltech), from 1983 to 1988 (mainly on Norman Bridge Laboratory of Physics). He then joined Bell Communications Research (Bellcore), a separate corporation from the Bell Labs (now Telcordia Technologies), as a Research Scientist, and left the institute, in 1994. He is currently a Professor of computer science and engineering with the Shahjalal University of Science and Technology (SUST). He is also a Bangladeshi science fiction author, physicist, academic, and activist. His current research interests include Bangla OCR, Bangla speech to text, Bangla natural language processing, WDM networks, optical communication, GPON, non-linear optics, and robotics.

Dr. Iqbal serves as the Vice President of the Bangladesh Mathematical Olympiad Committee. He played a leading role in founding the Bangladesh Mathematical Olympiad and popularized mathematics among Bangladeshi youths at local and international level. In 2011, he won the Rotary SEED Award for his contribution in the field of education.