Chapter 3

A Quest for Dialect-Distinctive Features

3.1 Introduction

Assamese, is the principal language of the state of Assam in North East India and one of the official languages of the Republic of India. Assamese is regarded as the lingua-franca of the whole of North East India. It is the easternmost member of the Indo-European family and is spoken by most natives of the state of Assam. The linguistic affiliation of the Assamese language is indicated in Figure 3.4. As reported by RCILTS IITG, over 15.3 million people speak Assamese as the first language and including those who speak it as a second language, a total of 20 million people speak Assamese primarily in the North-Eastern state of Assam and also in some parts of the neighboring states of West Bengal, Meghalaya, Arunachal Pradesh and other North East Indian states[1]. Majority of its speakers reside in the state of Assam in the Brahmaputra valley districts, a few in the neighboring states of Arunachal Pradesh, Nagaland and Meghalaya[2]. Outside India, Assamese is spoken by several thousand people in the countries of Bhutan and Bangladesh [3]. The distribution of Assamese speakers in and around the state of Assam can be seen in Figure 3.2. The Assamese language grew out of Sanskrit, however, the original inhabitants of Assam, like the Bodos and the Kacharis, have substantially influenced its vocabulary, phonology and grammar. Assamese and the cognate

[1] http://www.iitg.ac.in/rcilts/assamese.html

[2] http://www.ciil-lisindia.net/Assamese/Assa_demo.html

[3] https://en.wikipedia.org/wiki/Assamese_people

languages, Maithili, Bengali and Oriya, developed from Magadhi Prakrit which is the eastern branch of the Apabhramsa that followed Prakrit[4].

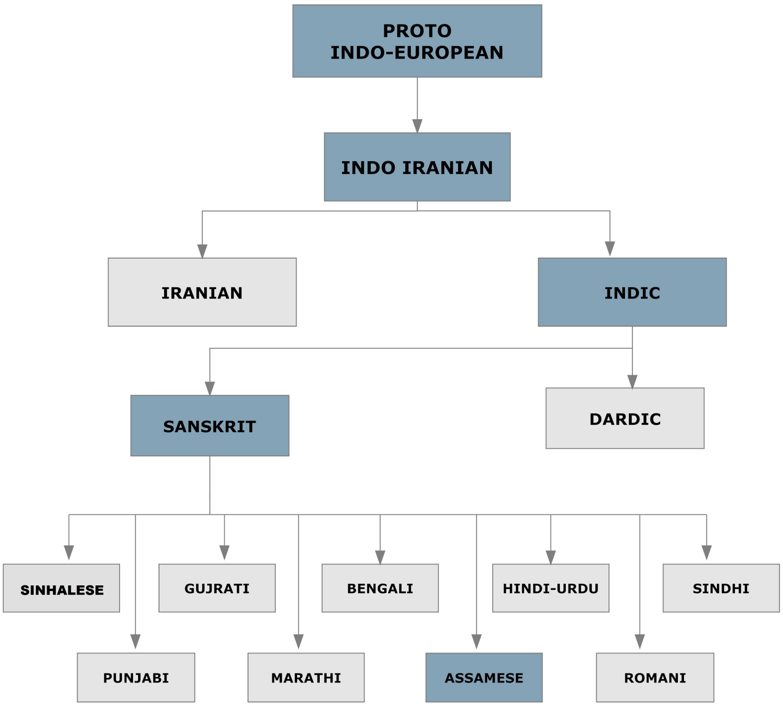


Figure 3.1. Linguistic Affiliation of the Assamese language. [5]

3.1.1 Assamese and its dialects

During the middle of the 19th century, Bengali was made the official language of Assam, but towards the end of the century Assamese was reinstated alongside Bengali. British colonizers decreed Eastern Assamese to be the standard Assamese dialect. Presently, however, Central Assamese is accepted as the principal dialect. Most literary activity takes place in this dialect, and it is often called the 'lik^{h}itɔ-b^{h}axa', though regional dialects are often used in novels and other creative works. A number of regional dialects are recognized. These dialects vary from each other primarily in terms of phonology and morphology, inspite of sharing a high degree of mutual intelligibility. Dr. Banikanta Kakati [67], an eminent linguist, has divided the Assamese dialects into two major groups, Eastern Assamese and Western Assamese. Goswami and Tamuli [45] divided Assamese into three distinct varieties, the Eastern, the Western and the Central variety. Goswami also claims that the Western variety can be further divided into the three sub-varieties belonging to (i)

[4] http://en.wikipedia.org/wiki/Assamese_language

[5] `http://www.iitg.ac.in/rcilts/assamese.html`

Barpeta, (ii) Nalbari and (iii) Palashbari-Chaigaon. However, recent studies have shown that there are four major dialect groups or dialectal varieties, listed from east to west[6]. The distribution of speakers of the four major dialectal groups, in Assam can also be seen in Figure 3.2.

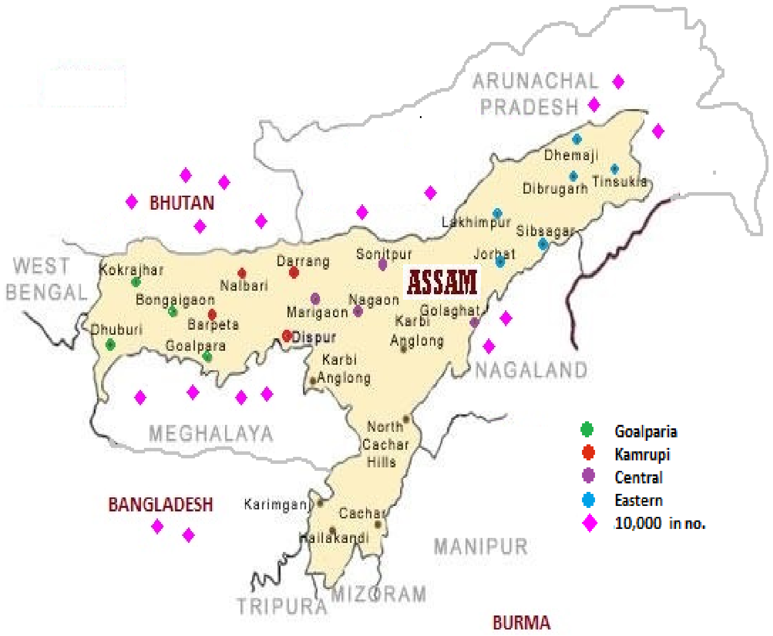


Figure 3.2. Distribution of Assamese speakers in and around the state of Assam

1. The Eastern variety spoken in and other districts around the district of Sibsagar.

2. The Central variety spoken in present Nagaon district and in the adjoining areas.

3. The Kamrupi variety spoken in the undivided districts of Kamrup, Nalbari, Barpeta and Darrang.

4. The Goalparia variety spoken in Goalpara, Dhubri, Kokrajhar and Bongaigaon districts.

3.1.2 The Assamese script

The Assamese script presently has a total of 41 consonants and 11 vowels representing the eight main vowel sounds of Assamese as shown in Figure 3.3. The script used

[6] http://www.iitg.ac.in/rcilts/assamese.html

by the Assamese language is a variant of the Eastern Nagari script which traces its descent from the Gupta script. The names of the consonant letters in Assamese are typically just the consonant's main pronunciation plus the inherent vowel /ɔ/. In addition to this the language has a number of 'juktakhars' or consonant clusters. Assamese spelling is not always phonetically based. Current Assamese spelling practices are based on Sanskrit spelling, as introduced in Hemkosh[7], the second Assamese dictionary written in the middle of the nineteenth century by Hemchandra Baruah [8]. The Hemkosh is in fact considered as the standard reference of the Assamese language. The Assamese phoneme inventory is unique in the Indic group of languages in its lack of a dental-retroflex distinction among the coronal stops. Historically, the dental and retroflex series merged into alveolar stops. Assamese is also unusual among Eastern Indo-Aryan languages for the presence of the voiceless velar fricative /x/ .

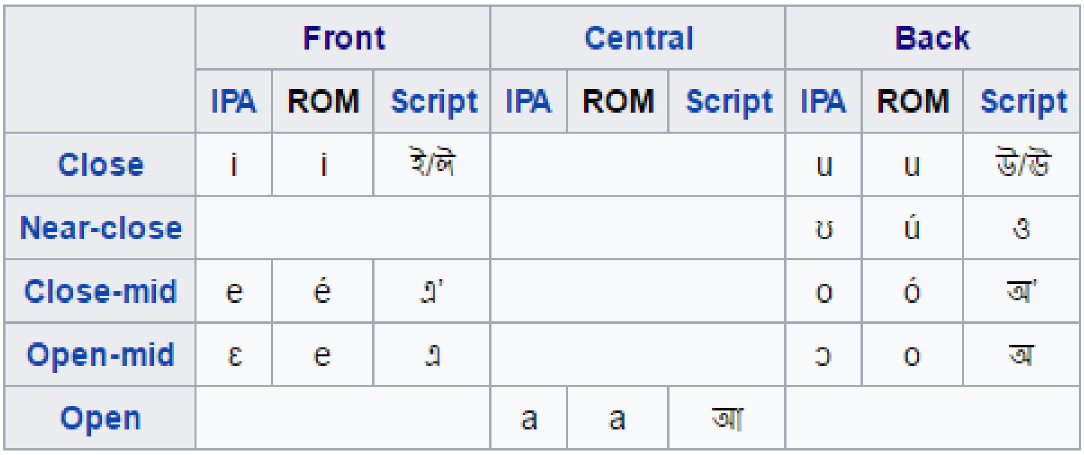
3.1.3 Linguistic variations in Assamese dialects

The most significant variations in a language occur at the level of the lexicon (vocabulary), phonology (pronunciation), grammar (morphology and syntax)and usage [121]. The first three levels of variations are more common. Lexical differences i.e., differences in vocabulary, are one aspect of dialect diversity which people notice readily. The dialects of Assamese too exhibit a number of differences at the lexical level. A word may be entirely different in the different varieties or may undergo a change in its structure or form. For example, the English name 'guava' in Central Assamese is /mod^{h}uɹiam/ and in Kamrupi it is /xop^{h}ɹam/, 'beautiful' in English is /d^{h}unia/ in Central Assamese and /t^{h}ouga/ in Kamrupi, 'papaya' in English is /ɔmita/ in Central Assamese and /moitp^{h}ɛl/ in Kamrupi. If a word in the standard or Central variety of Assamese, has two /a/ sounds side-by-side, the first /a/ turns into an /ɔ/ or /ɛ/. In the Kamrupi variety, this is tolerated. For example 'star' in English is /taɹa/ in Kamrupi and /tɔɹa/ the standard variety, 'drain' is /nala/ in Kamrupi and /nɔla/ in the standard variety. In words having two syllables, the second /ɔ/ becomes an /a/. For example 'hot' is /gɔɹam/ in Kamrupi and /gɔɹɔm/ in the standard variety of Assamese. A prominent feature of the Kamrupi dialect group is the use of initial stress, in contrast to the use of

[7] https://en.wikipedia.org/wiki/Hemkosh

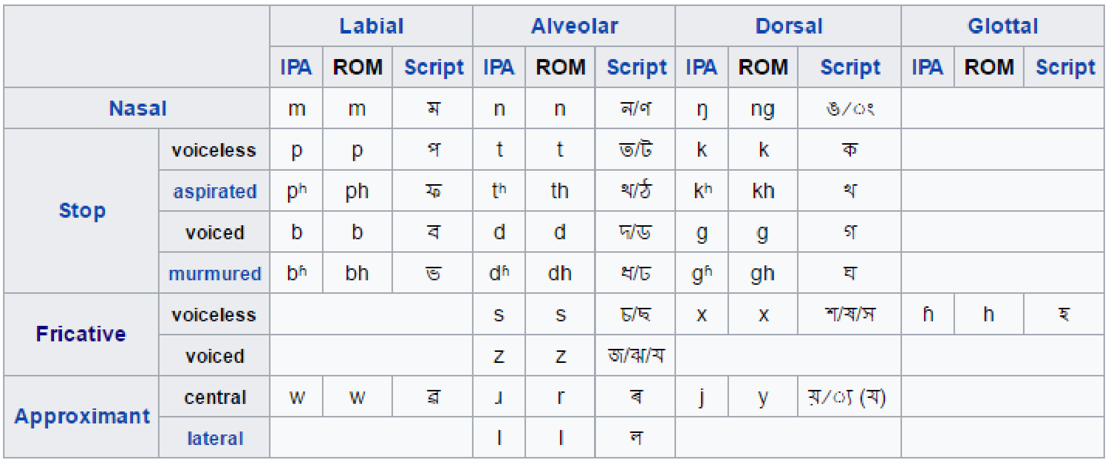
[8] https://en.wikipedia.org/wiki/Hemchandra_Barua

[9] http://en.wikipedia.org/wiki/Assamese_language



	Front			Central			Back		
	IPA	ROM	Script	IPA	ROM	Script	IPA	ROM	Script
Close	i	i	ই/ি				u	u	উ/ু
Near-close							ʊ	ú	ও
Close-mid	e	é	এ'				o	ó	অ'
Open-mid	ɛ	e	এ				ɔ	o	অ
Open				a	a	আ			

(a)



		Labial			Alveolar			Dorsal			Glottal		
		IPA	ROM	Script	IPA	ROM	Script	IPA	ROM	Script	IPA	ROM	Script
Nasal		m	m	ম	n	n	ন/ণ	ŋ	ng	ঙ/ং			
Stop	voiceless	p	p	প	t	t	ত/ট	k	k	ক			
	aspirated	p^h	ph	ফ	t^h	th	থ/ঠ	k^h	kh	খ			
	voiced	b	b	ব	d	d	দ/ড	g	g	গ			
	murmured	$b^ɦ$	bh	ভ	$d^ɦ$	dh	ধ/ঢ	$g^ɦ$	gh	ঘ			
Fricative	voiceless				s	s	চ/ছ	x	x	শ/ষ/স	ɦ	h	হ
	voiced				z	z	জ/ঝ/য						
Approximant	central	w	w	ৱ	ɹ	r	ৰ	j	y	য়/্য (য)			
	lateral				l	l	ল						

(b)

Figure 3.3. (a) Vowels in Assamese Phoneme Inventory (b) Consonants in Assamese Phoneme Inventory [9]

penultimate stress in the eastern dialects[10]. This results in shortening the word, /komoɹa/ in Central Assamese becomes /kumɹa/ in the Kamrupi dialect, /nazanu/ in Central Assamese becomes /naznu/ in the Kamrupi dialect. Medial vowels are therefore rarely pronounced or usually slurred over.

The dialects (both eastern and western) of Goalpara straddling the Assamese-Bengali language boundaries display phonetic features from both languages. The phonemes of eastern Goalpara dialect approach those of Assamese while those of the western dialects approach those of Bengali. The distinctive velar fricative /x/

[10] http://en.wikipedia.org/wiki/Kamrupi_dialect

present in Assamese is present in the eastern dialect, but absent in the western dialect[11]. The dental and cerebral (retroflex) phonemes present in Bengali are found in the western dialect, but they approach the alveolar sound in the eastern dialect in consonance with Assamese. The aspirated /c^{h}/ is present both in Bengali and the western dialect, but is not found in the eastern Goalparia dialect and in standard Assamese. Epenthesis or the insertion of an extra sound in a word, is a distinguishing feature in western Assamese dialects of Kamrup and Goalpara. For example, /kazia/ in Central Assamese becomes /kaiza/. Epenthetic vowels are the rule in Kamrupi dialects, and even diphthongs and triphthongs appear in initial syllables such /haula/ in Kamrupi is /haluwa/ and /kewla/ in Kamrupi is /kewalia/ in the standard variety. Diphthongs do not occur in the final syllables. It is also observed that pronunciation changes with time. This is observed in the frequent use of high vowels in the Eastern variety in place of medial vowels. Use of medial vowels is a feature of the Kamrupi variety. /kapuɹ/, /mul/, /tamul/ and /k^{h}alu/ in Eastern Assamese as against /kapoɹ/ (cloth), /mol/ (worth), /tamol/ (betel-nut) and /k^{h}alo/ (I have eaten) in Kamrupi.

With respect to morphology also, the western dialects exhibit differences from the standard variety. The plural suffixes of Kamrupi are also very different from that of the standard variety of Assamese (Kamrupi: -gila, -gilak; Standard: -boɹ, -bilak). The instrumental sense -di in Kamrupi is increasingly accepted in the standard variety now. For example /hatedi/ in Kamrupi is /hateɹe/ in the standard variety, but /hatedi/ is also acceptable. Kamrupi has a large variety of adverbial formations such as - ita, - et^{h}en, - enke and - kahai, which are also different in Eastern and Central Assamese. The Kamrupi variety uses - lak and - ilak for third personal affix while East Assamese uses - le and - ile for the same. For example ‘he ate’ in English is /xi k^{h}alak/ in Kamrupi and /xi k^{h}ale/ in the standard variety. The suffix -tu is used for both genders in Kamrupi, for example, /loɹatu/ (the boy), /apitu/ (the girl) and /solitu/ (meaning both the boy/ the girl). Another noticeable difference is the way verb forms are used with the personal pronoun /tɔi/. In the standard variety, the verb form with /tɔi/ ends with /ɔ/ (example:/tɔi k^{h}aisɔ/) and that with /tumi/ ends with /a/ (/tumi k^{h}aisa/). In the western variety though, both /ɔ/ and /a/ are equally used with /tɔi/. For example, /*tɔi* kiɔ xui *asɔ*?/ in the standard variety is /*tɔi* kja g^{h}umi *asa*?/ in the western variety.

[11] http://en.wikipedia.org/wiki/Goalpariya_dialect

3.1.4 A preliminary analysis of two varieties of Assamese

The main goal of our research work is to incorporate dialectal features into speech synthesised by a TTS built for the standard variety of Assamese so that the synthesised speech, sounds natural with respect to the dialect considered. The two varieties of Assamese that we have considered for our study are the AIR variety and the Nalbaria variety. The AIR variety is a form of the central group of Assamese generally spoken by the readers of Assamese news of All India Radio, this we consider as the standard variety. In fact speech in the AIR variety is generated from a TTS built for the AIR variety but with text in the Nalbaria variety. This results in Nalbaria speech lacking naturalness, i.e., as if spoken by a speaker of the standard variety, i.e., AIR. Therefore this is considered to be the AIR variety. The Nalbaria variety is a form spoken by the people in around the district of Nalbari in Assam. It belongs to the Kamrupi variety. Though a number of dialects of Assamese exist, we have chosen Nalbaria for our study specially because it greatly differs from the standard variety at all the above mentioned levels of variations in a language. For our study we have concentrated only on the phonological features of a pair of Assamese varieties. As mentioned in Section 2.5, differences between dialect pairs of different languages may vary. The first step therefore is to analyse speech in the two varieties of Assamese that we have considered for our study and find out features distinctive to each variety. As a preliminary step a perceptual study is carried out where 4-5 speakers fluent in both the varieties, AIR and Nalbaria, are asked to listen to speech in both the varieties and point out perceptual differences. In general, the subjects have been able to point out some basic differences which are listed below:

1. Nalbaria speakers generally tend to speak faster than AIR speakers. This would mean that the number of segments uttered per unit time would be much higher in Nalbaria. **Feature to analyse:** Segmental duration.

2. The duration of the secondary vowel in most diphthongs in Nalbaria, is much smaller than the primary vowel, therefore making the perception of diphthong difficult. **Feature to analyse:** Dynamic diphthong trajectories or amount of diphthongisation.

3. Some vowel sounds are perceived differently in the two varieties. For example, the vowel /o/ is pronounced differently in the two varieties. **Feature to analyse:** Vowel formant space.

4. Nalbaria speakers tend to pronounce most sounds, such as phonemes like /t/, /t^{h}/, etc with ease. **Feature to analyse:** Voice onset time.

5. Some sounds are produced by constricting the airflow in the glottis, for example the /h/ in /aihbana/. **Feature to analyse:** Spectral Tilt (Open Quotient).

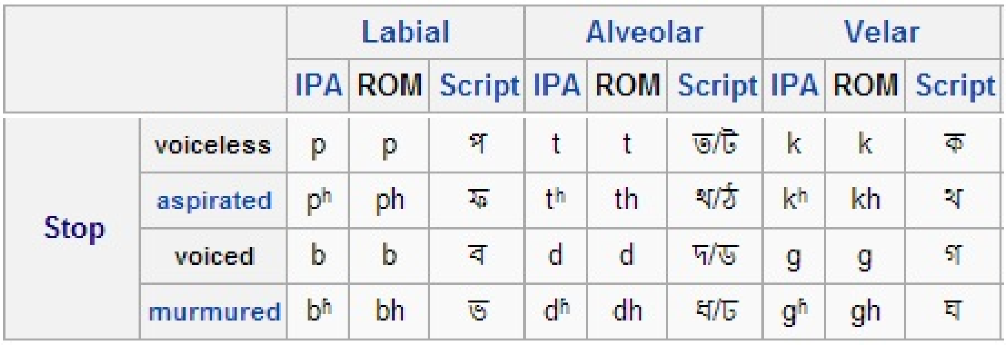
In this chapter, we carry out a detailed analysis of the phonological features, spectral and prosodic, of AIR and Nalbaria. The differences in the two varieties as pointed out by the preliminary perceptual analysis are in fact cues to what features are to be taken up for analysis. In addition to these features, features such as Cepstral coefficients and Pitch Contours are also analysed. Based upon the analysis some features are selected for transformation or manipulation in order to make the synthesised dialectal speech more natural. The analysis is carried out on speech data collected from 2-5 male speakers. We could not use speech data from female speakers as the female speakers of Nalbaria we contacted were too shy to speak in the dialect. Some female speakers who agreed for the recordings were too conscious of the way they spoke. This led to unnatural recorded speech which had to be discarded. A point to note is that the following representations of vowels using IT3 codes[12], will be used in the rest of the thesis in addition to IPA symbols: /ɔ/ as /ax/, /a/ as /aa/, /i/ as /i/, /u/ as /u/, /e/ and /ɛ/ as /e/, /ʊi/ as /oi/, /ʊ/ as /o/ and /ʊu/ as /ou/.

The rest of the chapter is organised as follows. Section 3.2 introduces the feature of Voice Onset Time(VOT) and carries out an analysis of the VOT of stop consonants in the two Assamese varieties under study. Section 3.3 analyses the vowel space in the two varieties. It also carries out an analysis of the duration of vowels in the concerned varieties. Section 3.4 presents an analysis of the diphthongs to find out the extent of diphthongisation in the two varieties while another feature, spectral tilt, is analysed in Section 3.5. Section 3.6 presents an analysis of the cepstral coefficients and pitch contours and finally Section 3.7 summarizes and concludes the analysis.

[12] `http://www.iitg.ernet.in/cseweb/tts/tts/Assamese/transliteration/IndicLanguageTransliteration`

3.2 Analysis of VOT

The speech signal contains various analytical features and one such feature is the VOT or voice onset time which has proved to be a very important feature for classifying stops into different phonetic categories with respect to voicing. The stops or stop consonants are those consonantal sounds which are produced by blocking the vocal tract. The occlusion may be made with the tongue blade, tongue body or glottis. The consonants /p/, /t/, /k/, /b/, /d/, /g/, /ph/, /th/, /kh/, /bh/, /dh/, and /gh/ are considered as stops. The stop consonants are classified as shown in Figure 3.4. Voice Onset Time or Phonation Onset, commonly known as VOT is generally defined as the difference in time (in milliseconds) between the instant of stop consonant closure release and the start of vocal cord vibration. Negative values of VOT mean that the vocal cords begin vibrating before time of end of vocal tract closure while positive VOT indicates that the vocal cords start vibrating after the beginning of vocal tract closure. VOT may also have a zero or near zero value when the vocal cords start vibrating at the time of closure release. VOT plays an important role in perceptual discrimination of phonemes of the same place of articulation, for example labial stops (/p/,/b/), alveolar stops (/t/ and /d/) and velar stops (/k/ and /g/) [11]. VOT has come to be regarded as one of the most important methods for examining the timing of voicing in stops (especially in word-initial position) and has been applied in the studies of many languages [24]. VOT has been studied by researchers from various fields like language typology, phonetics, second language impact on the first language, speaker identification and also dialect and accent detection and identification [84], [30], [5]. VOT is also used as a parameter for speech synthesis [122].



		Labial			Alveolar			Velar		
		IPA	ROM	Script	IPA	ROM	Script	IPA	ROM	Script
Stop	voiceless	p	p	প	t	t	ত/ট	k	k	ক
	aspirated	p^h	ph	ফ	t^h	th	থ/ঠ	k^h	kh	খ
	voiced	b	b	ব	d	d	দ/ড	g	g	গ
	murmured	b^h	bh	ভ	d^h	dh	ধ/ঢ	g^h	gh	ঘ

Figure 3.4. Classification of Assamese Stop Consonants

Most existing studies concentrate on the English language. No known attempt has been made to examine the VOT patterns in Assamese and its dialectal variants so far. The stops in Assamese may be classified into three groups according to the place of articulation. They are labials, alveolars and velars. Historically the dental and retroflex stops have both merged into alveolar stops. Also, for each group there are two different types based on the manner of voiced/unvoiced distinction, i.e., aspirated and murmured. We therefore make a comparative study of the VOT patterns of the standard variety of Assamese (AIR) and one of its dialectal variants, Nalbaria (NAL). This would provide a better understanding of the phonological differences that exist among the different dialectal variants of the language which may prove to be useful for dialect recognition, translation and synthesis. For our initial study we concentrate on the stops in word initial positions.

3.2.1 Literature Review

Lisker and Abramson [87] was the first to describe VOT in their well known cross-language study of voicing in initial stops and they defined VOT as 'the time interval between the burst that marks release of the stop closure and the onset of quasi-periodicity that reflects laryngeal vibration'. They investigated 11 languages and classified them into three groups based on the number of stop categories each language contained. They also suggest that each stop category falls into one of three ranges, 125 to 75 ms (lead), 0 to +25 ms (short lag), and +60 to +100 ms (long lag), respectively. Following Lisker and Abramson's categorization, both the AIR and the NAL variety of Assamese fall into the four-category group of languages. A four-category language is one in which there are four stops at one place of articulation [114]. For example four labials (/p/, /ph/, /b/, /bh/), four alveolars (/t/, /th/, /d/, /dh/), four dorsal stops (/k/, /kh/, /g/, /gh/). However the stops in Assamese can be distributed in the above mentioned ranges with slight modifications of the lower and upper end values.

Measurements of VOT before the release of stop closure are stated as negative numbers and called 'voicing lead', while after the release are stated as positive numbers and called 'voicing lag' [88]. If release and voicing are simultaneous, VOT is considered to be zero.

Depending on VOT, Lisker and Abramson divided languages into two groups: group A languages which have long VOT, over 50 milliseconds, for a voiceless stop but short VOT for voiced: and group B languages which have short VOT, less than 30 milliseconds, for voiceless, but negative VOT for voiced stops.

VOT values for stops are found to vary in relation to the place of articulation [27]. In most languages it is confirmed that VOT values get longer when the place of articulation moves from an anterior to a posterior position. However exceptions do exist. Lisker and Abramson in their works further demonstrated that, for both unaspirated and aspirated stops, mean VOT values are longer in velar stops than in alveolar and bilabial stops.

Lisker and Abramson also reported that the influence of vowels on the VOT of stop consonants is not significant, however later research prove that the quality of the vowel do effect the VOT of stops. Moreover researchers also claim that in rapid speech, the speaking rate also might influence the stop VOTs [93].

Different tones have different pitch levels, which are determined by the vibrating frequency of the vocal cord. As speculated by many researchers, VOT durations are in fact affected by tone as different tones have different fundamental frequencies and pitch levels. Therefore VOT values may vary when they occur in different lexical tones [105].

3.2.2 Experimental Framework

In order to carry out the analysis, the first step is to build the required corpus, followed by the measurement of VOT, analysis of VOT data, and finally drawing conclusions based on experimental observations. The following subsections elaborate on these points.

3.2.2.1 Building the Speech Corpus

A corpus is developed having two parts, one part for the standard variety of Assamese i.e., AIR, and the other part for the dialectal variety i.e., Nalbaria. A list of words having the voiced and voiceless plosives in word initial position is prepared. Since the quality of the vowel following the stop influences the VOT of the stop, for each stop consonant, five words are selected where the stop is followed by one of the five vowel sounds /aa/, /e/, /i/ ,/o/ and /u/. For e.g., for the plosive /p/ we have five words starting with ‘paa’, ‘pe’, ‘pi’, ‘po’ and ‘pu’ such as ‘paani’, ‘pepaa’, ‘pithaa’, ‘pohaxr’ and ‘puthi’. Likewise the list of words is prepared for the rest of the stops. Therefore the list consists of a total of sixty words. Four speakers, all male, are chosen, two speaking the standard variety and two speaking the dialectal variety. Each word is spoken by the speaker five times. The corpus is thus prepared by recording the word list in the voices of the four speakers and consists of twelve hundred word samples. The recording for the NAL variety was carried out at the Dr.Bhupen Hazarika Regional Government Film and Television Institute, Guwahati, in a sound proof recording room at a sampling rate of 44.1 kHz and bit resolution of 16. The recording for the AIR variety was carried out in a sound proof room at Tezpur University, Dept. of Computer Science and Engineering with a Sony recorder at the same sampling rate and resolution.

Sl No.	Stop Consonant	Word(IPA)	Word(IT3)
1	/p/	/pani/-water	/paani/
2	/p^{h}/	/p^{h}atek/-jail	/phaatek/
3	/b/	/bamun/-priest	/baamun/
4	/b^{h}/	/b^{h}agɔɹ/-tiredness	/bhaagaxr/
5	/t/	/taɹik^{h}/-date	/taarikh/
6	/t^{h}/	/t^{h}ali/-plate	/thaali/
7	/d/	/dapun/-mirror	/daapun/
8	/d^{h}/	/d^{h}aɹɔna/-idea	/dhaaraxna/
9	/k/	/kali/-yesterday	/kaali/
10	/k^{h}/	/k^{h}am/-envelope	/khaam/
11	/g/	/gahɔɹi/-pig	/gaahaxri/
12	/g^{h}/	/g^{h}am/-sweat	/ghaam/

Table 3.1 List of stop consonants of Assamese

3.2.2.2 Measuring VOT

The voice onset time of a plosive is defined as the duration between the release of a plosive and the beginning of vocal cord vibration as shown in Figure 3.5. VOT can be positive, negative or 0.

1. If the onset of voicing follows the release, measure the interval in milliseconds between the release of the plosive until the onset of voicing. This is positive VOT or voicing lag.

2. If the onset of voicing coincides with the release, this is 0 VOT. The measurement would be 0 milliseconds.

3. If the onset of vocal cord vibration precedes the plosive release, then measure the voicing duration from the onset of voicing (or the onset of closure if there

is voicing throughout), again in milliseconds. This is negative VOT or voicing lead.

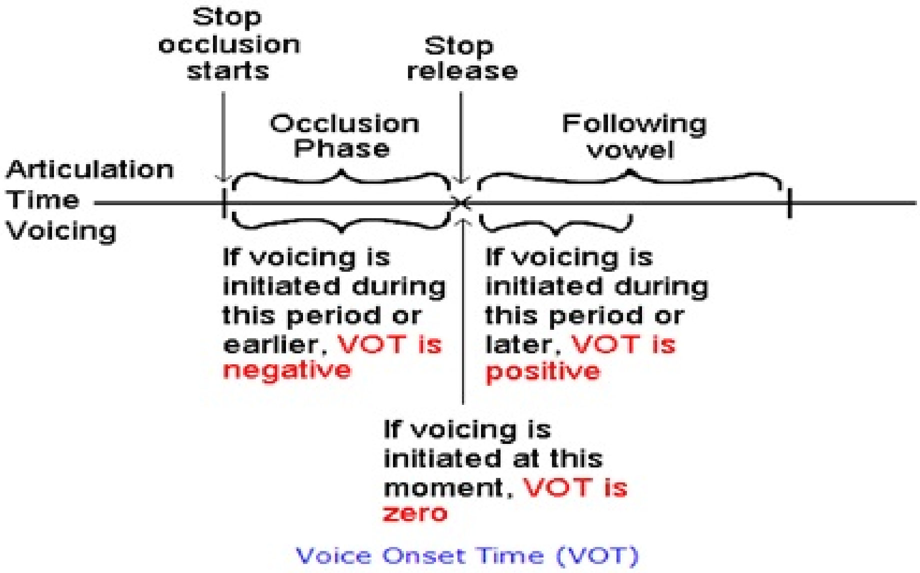
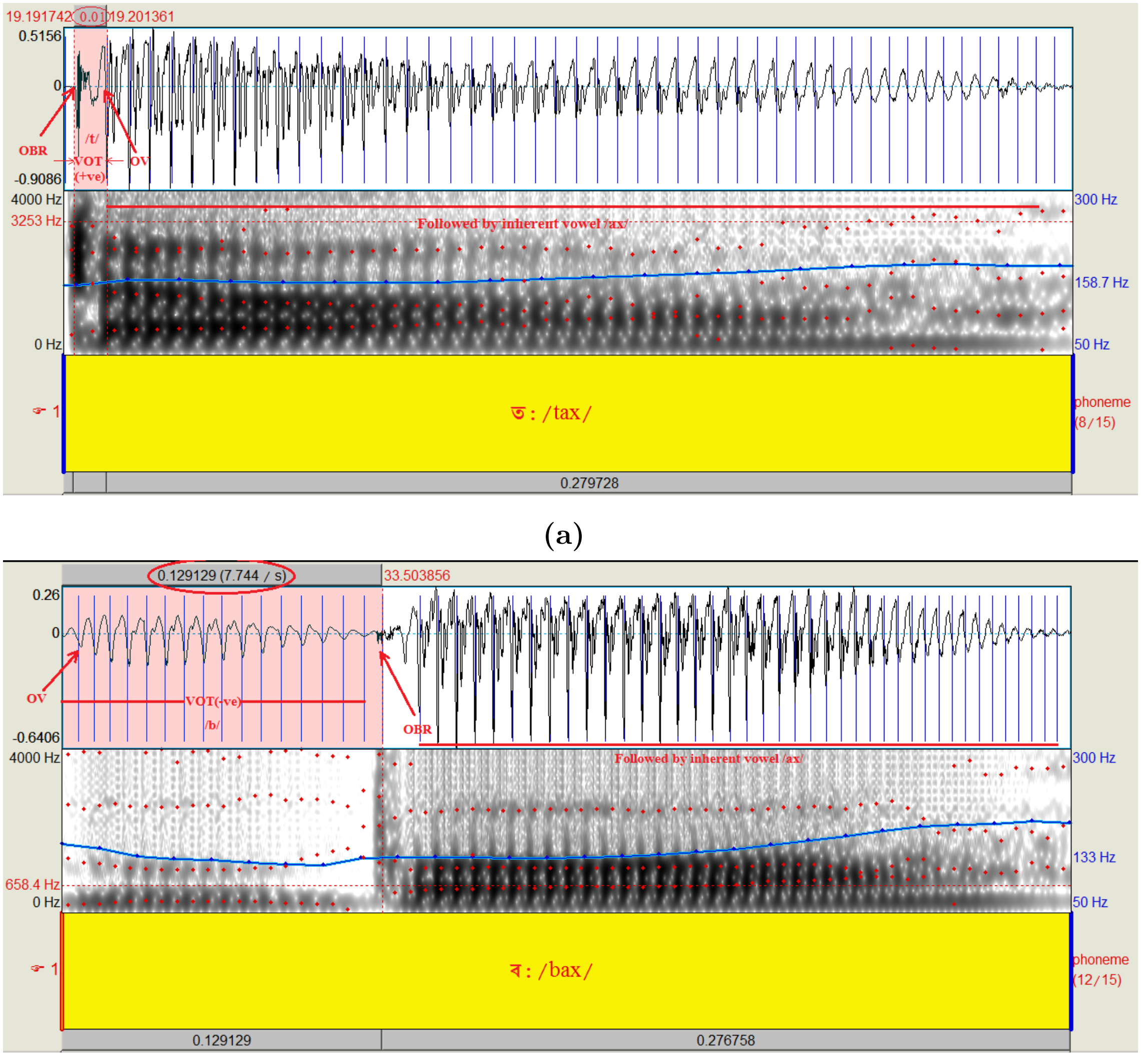

Stop occlusion starts
Stop release
Occlusion Phase
Following vowel
Articulation
Time
Voicing
If voicing is initiated during this period or earlier, VOT is negative
If voicing is initiated during this period or later, VOT is positive
If voicing is initiated at this moment, VOT is zero
Voice Onset Time (VOT)

Figure 3.5. Voice Onset Time (source: [158])

3.2.2.3 Methodology

The experiment carried out for our research depends mainly on extracting VOT values of the voiced and unvoiced stops of AIR Assamese and Nalbaria Assamese. Our analysis is carried out with the help of PRAAT speech analysis software. PRAAT is used to generate the waveform and spectrogram for each word utterance containing the plosive in word-initial position. On each waveform two points in time are located: the onset of burst release (OBR) marked by the onset of low amplitude, aperiodic noise and the onset of voicing (OV) marked by the onset of high amplitude periodic energy. Onset of voicing, i.e., starting of vocal cord vibration, can be observed by noticing low frequency periodicity in the wide band spectrogram. VOT is calculated as the latency between OBR and OV.

We use signal energy and vocal cord vibration information to locate the beginning of stop release, closure and voicing. Audio monitoring of the signal is also made to check the sound of each segment. The closure release is marked at the beginning of abrupt increase in the energy level and can be easily identified in the signal waveform in PRAAT. VOT values are measured to a precision that permitted rounding to the nearest 0.5ms. The starting mark is set at the sharp increase of signal energy which signaled the release of /b/, /d/, /g/, /p/, /t/, /k/. The



(a)

(b)

Figure 3.6. Measurement of VOT (a) vot(/t/)=10ms (b) vot(/b/)=-129ms

end mark is set at the first upward going zero-crossing which signaled voicing onset. If a positive VOT or voicing lead is produced, the end mark is set at the first burst of /b/, /d/, /g/ [99]. Furthermore, the VOT values are measured by one and cross checked by another investigator for more reliability. Figure 3.6 illustrates the measurement of VOT for the stops /t/ and /b/ in AIR. In this case the latency between OBR and OV, i.e., VOT=10ms for /t/ and -129ms for /b/.

3.2.3 Results and Observations

The VOT range and mean of the NAL variety and the AIR variety are presented in Table 3.2, also in Figure 3.7, and compared with different snapshots of VOT measures as shown in Figures 3.8 - 3.10.

Phones	Range(N)	Mean(N)	Range(AIR)	Mean(AIR)
/p/	14:42	24	11:14	12
/b/	-98:-21	-59	-120:-56	-88
/t/	17:28	21	10:24	12
/d/	22:42/-56:-42	26/-46	-110:-70	-89
/k/	22:47	34	21:39	30
/g/	17:45/-58:-14	30/-35	-130:-58	-89
/p^h/	43:73	58	100:150	126
/b^h/	28:72	57	76:130	112
/t^h/	58:79	65	67:120	95
/d^h/	38:62	50	90:120	105
/k^h/	56:125	79	80:150	127
/g^h/	62:91	75	100:160	118

Table 3.2 Range and mean of VOTs for NAL—Nalbaria variety and AIR—All India Radio variety

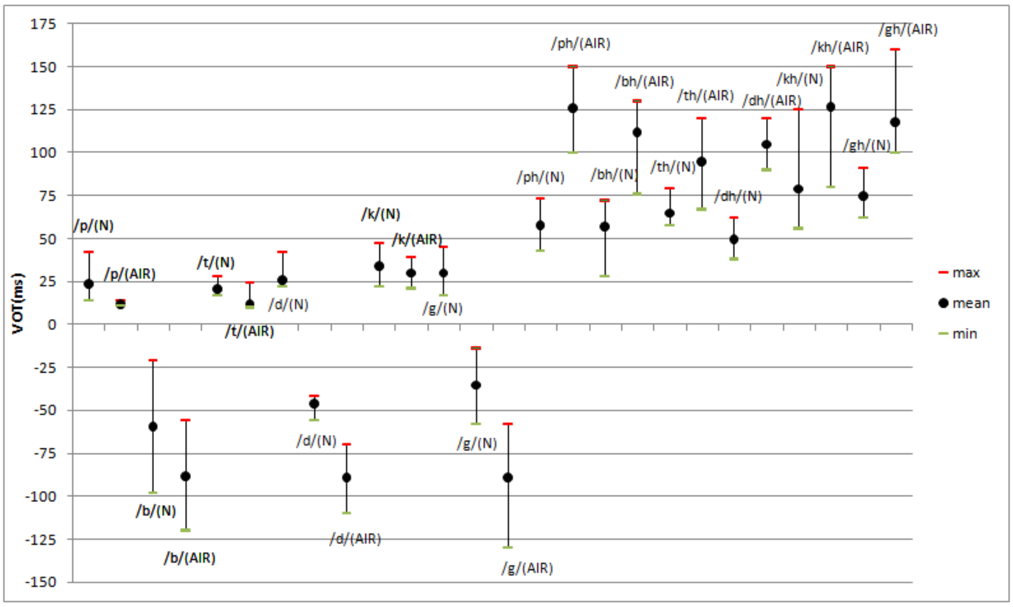
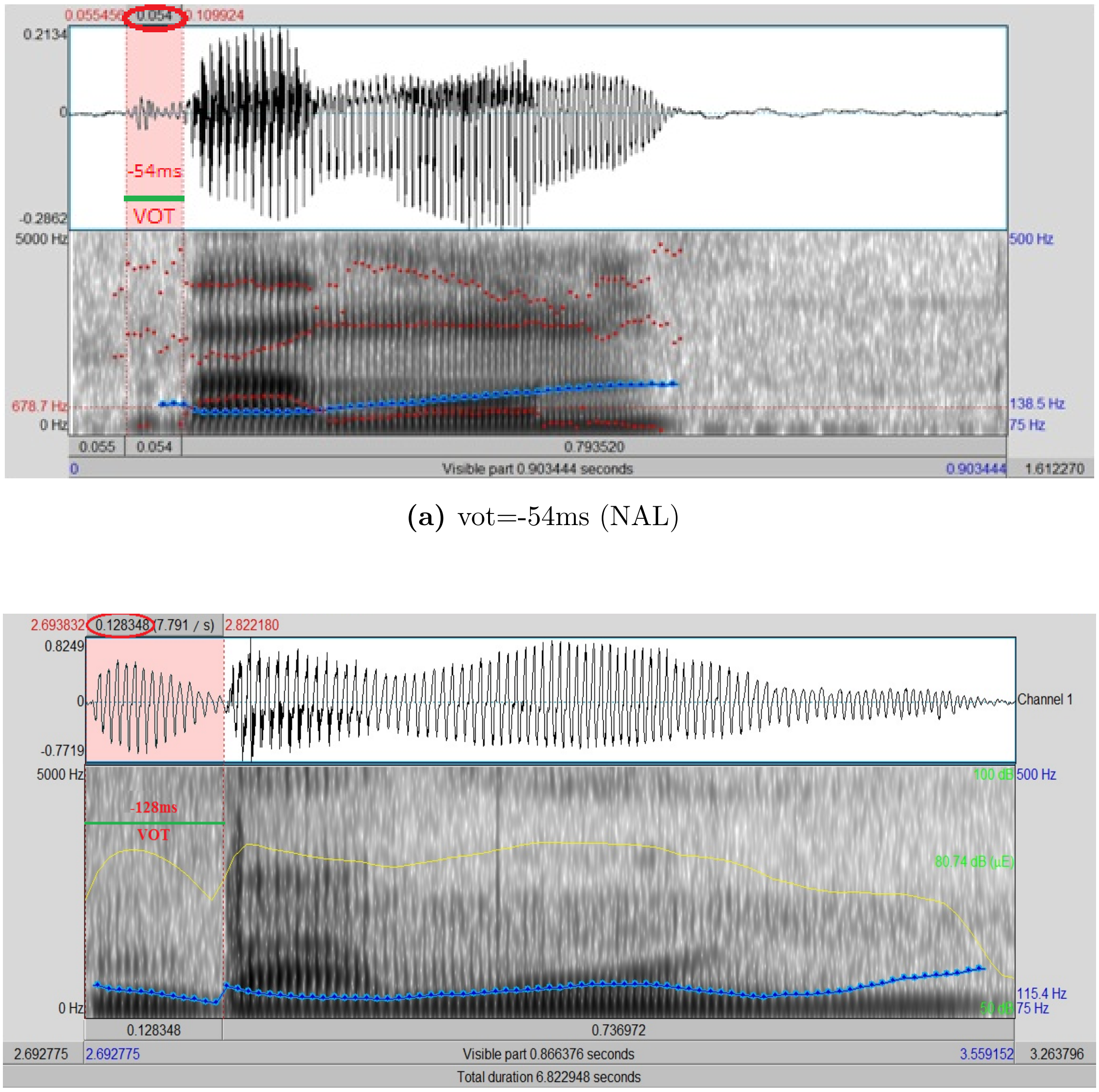


Figure 3.7. Voice Onset Time (min, max, mean) in NAL and AIR

After analysing the results presented in Table 3.2 we find that the VOT values for the voiceless stops /p/, /t/ and /k/ in the Nalbaria variety extend from +14 ms to +47 ms while those for the AIR variety range from +10 ms to +39 ms. The mean VOT values for /p/ and /t/ are almost same, which does not conform to the general agreement that further back the place of articulation, the longer the VOT.

The VOT values for the voiced stops /b/, /d/ and /g/ extend from -98ms to -14 ms /17ms to 45ms in the Nalbaria variety and extend from -130ms to -56ms



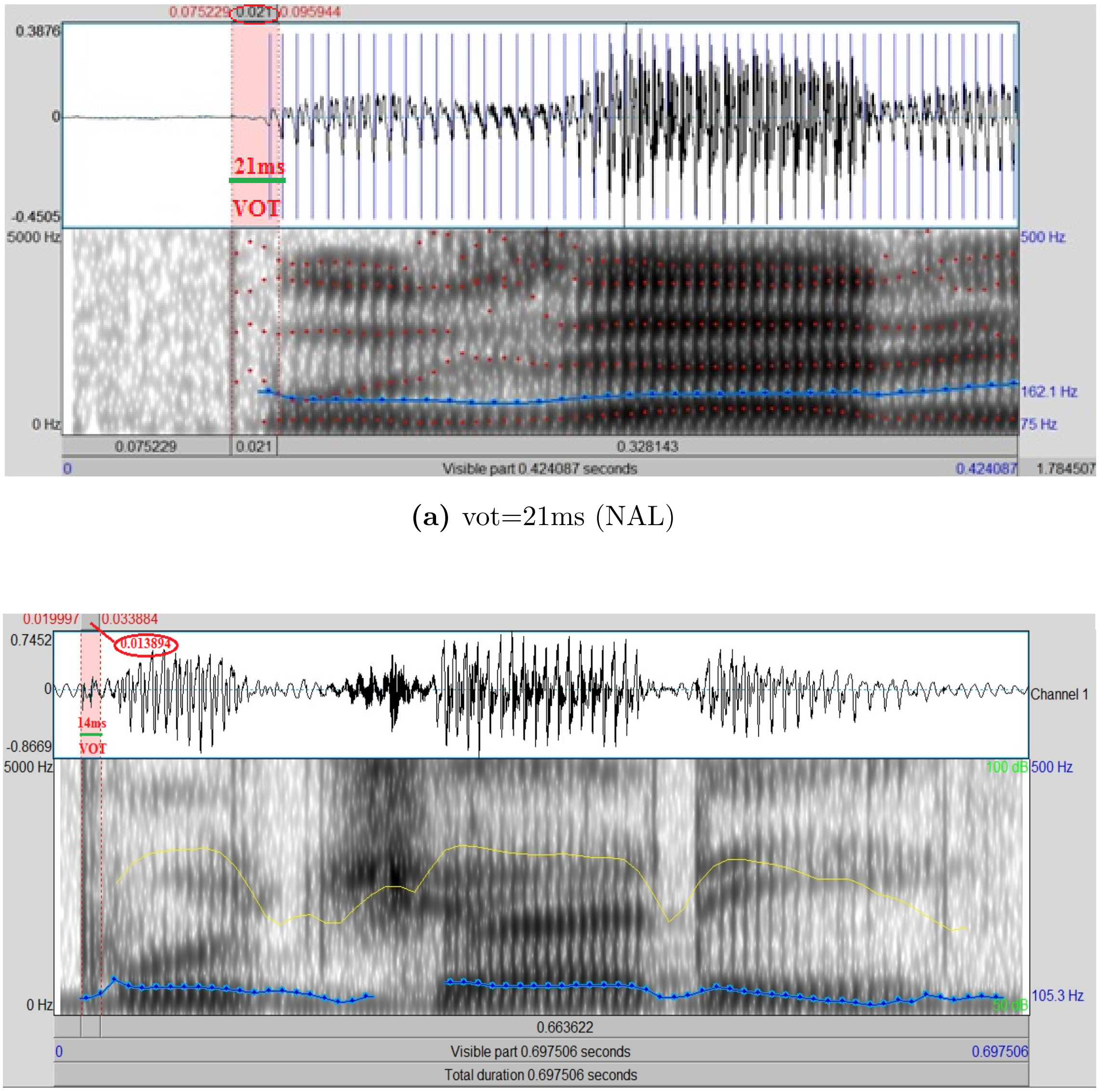
(a) vot=-54ms (NAL)

(b) vot=-128ms (AIR)

Figure 3.8. Waveform & Spectrogram for the word 'baamun'

in the AIR variety. While the VOT values for both voiceless aspirated and voiced murmured in both the varieties of Assamese are higher up in the VOT continuum and extend from +45 ms to +160 ms, the stops in the AIR variety are much more aspirated (almost twice) than its Nalbaria counterparts.

It should be noted that we give two sets of values for /d/ and /g/ in case of the Nalbaria variety because otherwise it would have meant lumping both positive and negative values for VOT as members of a single population which would have been misleading. This can be seen in Figure 3.11 which shows the mean VOTs for the stops in the two varieties. It can be seen that /d/ and /g/ in Nalbaria have two means, one positive and one negative, marked in red and blue bars. This

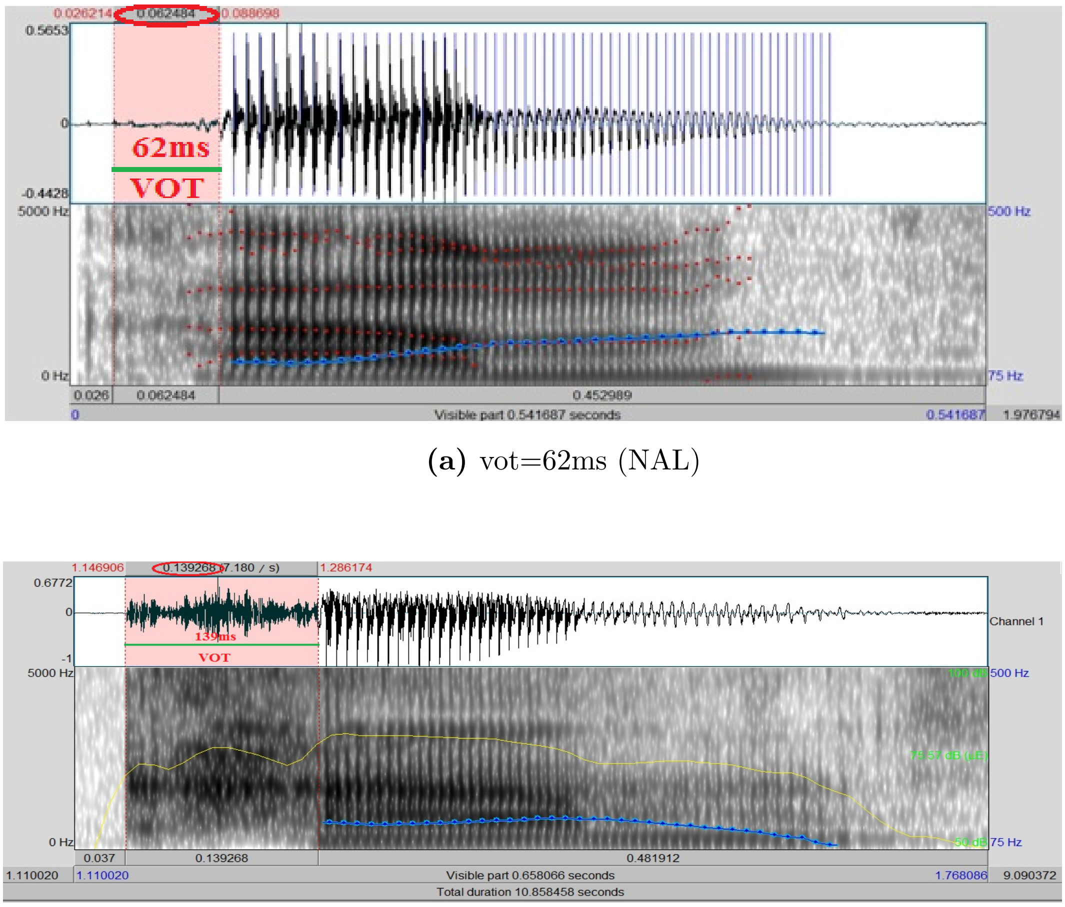


(a) vot=21ms (NAL)

(b) vot=14ms (AIR)

Figure 3.9. Waveform and Spectrogram for the word 'pujaari'

is similar to the case observed by Lisker and Abramson for the English language. This implies that for voiced stops in Nalbaria, which are alveolar and velar, the onset of vocal cord vibration may follow the plosive release which is not usually the case. One explanation may be that the speakers of the AIR variety of Assamese, follow the common rule of producing the voiced stops in a manner that the vocal cord vibration is preceded by the stop release. While the speakers of the dialectal variant may not adhere to the common rule. Another explanation may be the effect of speaking rate or tempo. Studies show that the speaking rate has an influence on VOT. The speakers of the Nalbaria variety generally tend to speak fast which may result in the positive values of the voiced stops.



(a) vot=62ms (NAL)

(b) vot=139ms (AIR)

Figure 3.10. Waveform and Spectrogram for the word 'ghaam'

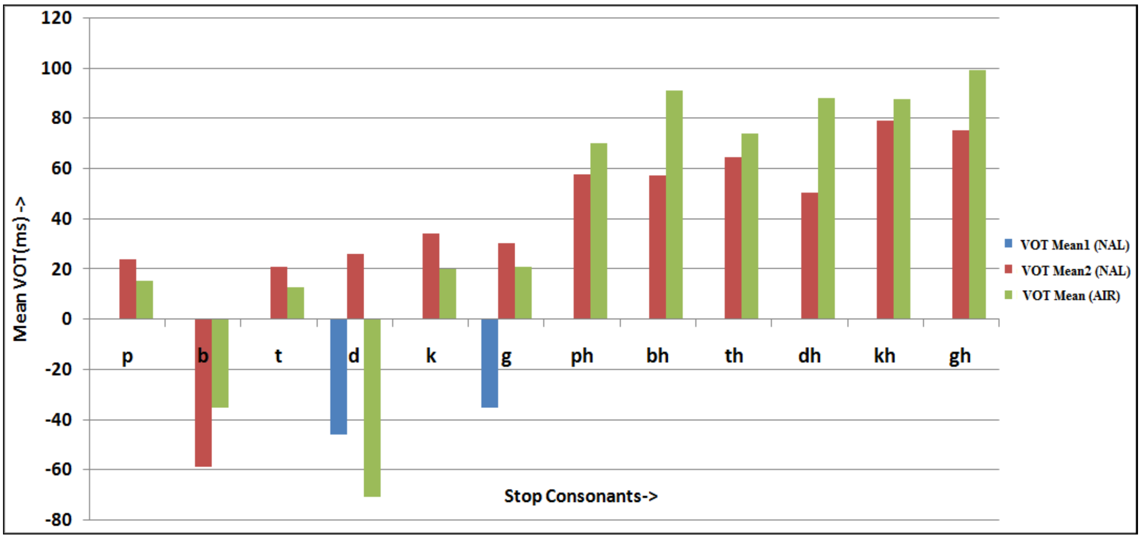


Figure 3.11. Mean VOTs of Nalbaria variety and AIR variety

3.3 Analysis of Vowel Space and Duration

Vowels can be described by setting up an imaginary "Vowel Space (VS)" and then defining each vowel by their position in that space[13]. A cross-section of the human head looking to the left is imagined and the vowels are defined according to the position of the highest point of the tongue while producing each vowel. Human vowel sounds in any language can be partly defined in relation to this vowel space. We therefore have high (= close) or low (= open) and front or back vowels. We can also define vowels as close-mid, open-mid, or centralized. The VS tells us where in the mouth the different vowels are produced with respect to the position of the tongue and what are the maximal values or corners of the vowel space.

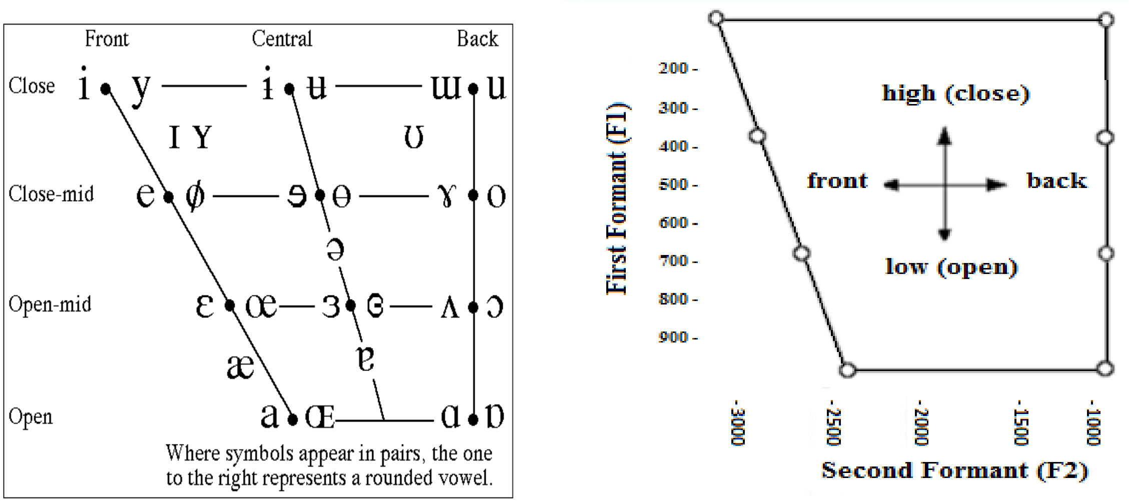
More commonly the VS is represented by the formant frequencies. Formants or formant frequencies, are concentrations of acoustic energy around particular frequencies of the speech spectrum. There can be any number of formants but the first and second formants have the most significant contribution to the quality of vowels. Points on an F1-F2 plane which in fact is the VS, are often used to represent a speaker's pronunciation [140]. The first formant F1 corresponds to the vowel openness (vowel height). On the IPA vowel chart as shown in Figure 3.12a, F1 varies along the *y-axis*. F1 increases in frequency as the vowel becomes more open and decreases to its minimum as the vowel sound closes. The second formant F2 corresponds to vowel frontness. It follows along the *x-axis*. F2 increases in frequency the farther forward that a vowel is and decreases to its minimum as a vowel moves to the back, i.e., back vowels have low F2 frequencies and front vowels have high F2 frequencies. However, in open vowels, a high F1 forces F2 to be high. So an alternative measure of frontness is F2 - F1. In short, the perceived vowel sound will depend on how the vowel is articulated and therefore on its position on the F2-F1 plane, i.e., the vowel space. Figure 3.12b shows the VS in terms of the formant frequencies F1 and F2. Thus the vowel /u/ will be close/high + back (low F1, low F2), vowel /a/ will be open/low + back (high F1, low F2), the vowel /e/ will be mid + front (high F1, medium F2) and the vowel /i/ will be close/high + front (low F1, high F2).

Vowels more specifically deal with the vocal sounds. In the spectrum of vocal sounds, reinforcement of several frequency zones is perceived as vowels. This makes

[13] `https://notendur.hi.is/peturk/KENNSLA/02/TOP/VowelSpace.html`

[14] `http://www.phonetics.ucla.edu/course/chapter1/vowels.html`

[15] `https://notendur.hi.is/peturk/KENNSLA/02/TOP/VowelSpace.html`



(a) The IPA Vowel Chart.[14]

(b) F2-F1 representation of Vowel Space.[15]

Figure 3.12. Vowel Chart and Vowel Space

formant analysis of vowels important for speech synthesis and recognition. The following subsections focus on analysing the formant structure of vowels as well as vowel duration in the two varieties of the Assamese language, for use in recognition and synthesis of dialects. IPA vowel charts of both the varieties are plotted for comparison.

3.3.1 Literature Review

In the synthesis of speech, vowels play a very important role because vowels are known to carry more articulatory information. Since the main goal of this thesis is to synthesise dialectal speech by incorporating dialect-distinctive features, we decided to carry out an acoustic analysis of vowels and diphthongs in the two dialects of the Assamese language, i.e., AIR and NAL, in terms of formants and segmental duration since these features are likely to be candidates of distinctive features. Clopper [28] provides evidence on variation in vowel duration across American dialects in the course of a survey of characteristics of six major regions. An acoustic investigation is carried out by Jacewicz et al. [62] on the duration of five American English vowels across three major dialect areas. There are systematic differences across all vowels studied, with the longest durations in the South and the shortest in the Inland north, with the Midlands in an intermediate but distinct position. Sinha et al. [131] investigated the influence of dialect on phoneme duration in six dialects of Hindi, i.e., Awadhi, Bagheli, Bhojpuri, Bundeli, Haryanvi and

Khariboli. Results show that phoneme duration is a dialect-distinctive feature with respect to the dialects of Hindi and can be used for their identification. Holt et al. [55] analysed the duration of vowels in African American English (AAE) and White American English (WAE), to get a better understanding of temporal variation in the two dialects. Their findings indicate that both extensive vowel lengthening before voiced stops and smaller temporal contrast between tense and lax vowels are distinctive features of AAE, signaling a differential use of vowel duration in AAE as compared to WAE. Another study [40] aimed at characterising the nature of the dynamic spectral change in vowels in three distinct regional varieties of American English spoken in the Western North Carolina, in Central Ohio, and in Southern Wisconsin, reported variation in formant dynamics as a function of phonetic factors such as vowel emphasis and consonantal context. Ewald et al. [35] investigated regional variation in Swedish vowels using DCT coefficients. Their results showed that in the three varieties of Swedish, i.e., Central, Estonian and Finland Swedish, Central Swedish vowels exhibited a higher degree of formant movement than the vowels in the other two varieties.

3.3.2 Experimental Framework

3.3.2.1 Building of Speech Corpus

A list of words having the vowels, ax(IPA:ɔ), aa(IPA:a), i(IPA:i), u(IPA:u), E(IPA:ɛ), e(IPA:e), o(IPA:u) in word initial, medial and end positions is prepared. The listed words are recorded from four speakers (two speaking the AIR variety & two speaking the NAL variety) at a sampling rate of 44.1kHz and 16 bit resolution in a noise free environment using a Sony recorder. The speakers are asked to read out each word, twice, at a normal pace. If any mispronunciation occurred, that particular word is discarded. Sample vowels in the two varieties along with their carrier words are shown in Table 3.3

3.3.2.2 Measuring formants and vowel duration

The recorded acoustic material is transferred to a standard PC and PRAAT speech analysis software [15] is used to generate the waveform and spectrogram for each word utterance containing the vowels in word-initial, median and end position.

Vowel duration is considered to extend from the beginning of the periodic

Table 3.3 Vowels in AIR and NAL along with their carrier words

Vowel	Carrier Word(AIR)	Carrier Word(NAL)
ax	axmitaa	baxjaar
aa	aam	baamun
i	itaa	bilehi
u	aathuwaa	ghumti
e	etiaa	kheti
o	ojaa	godhlaa

marking to the end of the periodicity. Therefore, on each waveform two points in time are located: the onset of vowel/diphthong marked by the onset of high amplitude, periodic energy and the offset of vowel marked by a change in the amplitude of the waveform and in the energy of the formants. The pitch contour is also used as an indicator of voicing, since without voice there is no pitch. Vowels are also indicated by darker regions in the spectrogram due to the presence of resonant frequencies. In other words, both audio playback and visual cues from the spectrogram are used to mark the vowel boundaries. The highlighted portion is played back through headphones, to confirm that the vowel/diphthong under study has been marked correctly. The markers are shifted if correction is required. The segments are then annotated with appropriate labels.

A vowel is supposed to exhibit a steady state formant pattern and therefore, the average of the formants at 25%, 50% and 75% of vowel duration, are considered. Therefore frequency of the first two formants at 25%, 50% and 75% of vowel duration are measured using formant trackers in PRAAT [16], with the maximum frequency set at 5500Hz, the dynamic range set to 30 dB and the window length set to 0.025 seconds. If mistracking of formants occurred it is hand corrected. The duration of vowel segments are also measured. The average F1, F2 values together with vowel duration, are then imported into an Excel sheet.

3.3.2.3 Methodology

The first step is formant normalisation in order to normalise the vowel formants from the various speakers. Formant normalisation is carried out to minimize inter-speaker variation due to physiological or anatomical differences, and at the same time preserve inter-speaker variations due to dialect or social differences. For our

[16] http://courses.washington.edu/l453/Formants.pdf

work we have used the Lobanov [38] measure for normalising the formants. The Lobanov measure expresses values relative to the hypothetical centre of a speakers vowel space. We therefore subtract a speakers mean formant frequency from a formant value and then divide by the standard deviation for that formant. Average values of F2 versus F1 for each of the vowels for each of the speakers are plotted and shown in Figure 3.13. Normalised formant values are also shown in Figure 3.14. Further vowel duration for both the varieties of Assamese are also separately presented in a bar diagram as shown in Figure 3.15.

3.3.3 Results and Observations

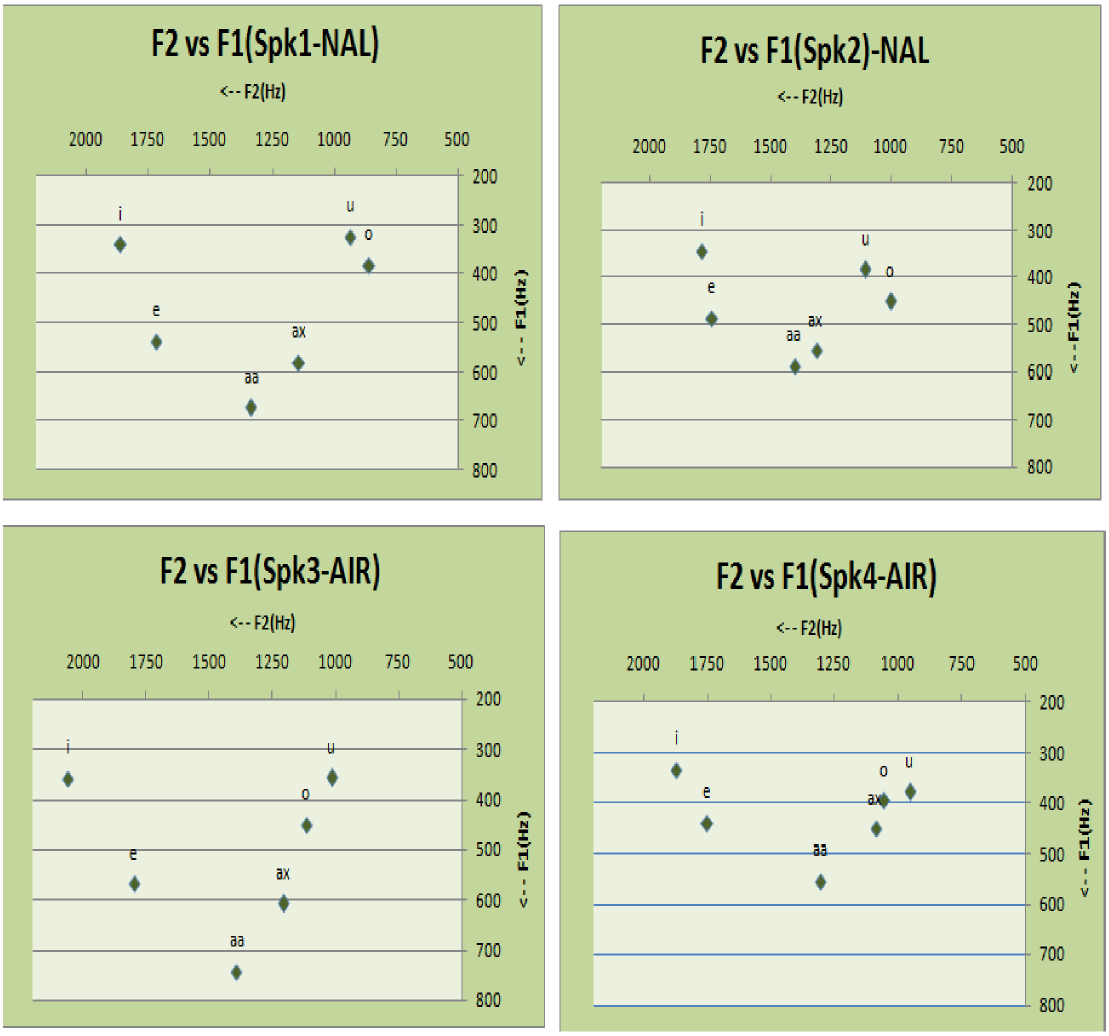
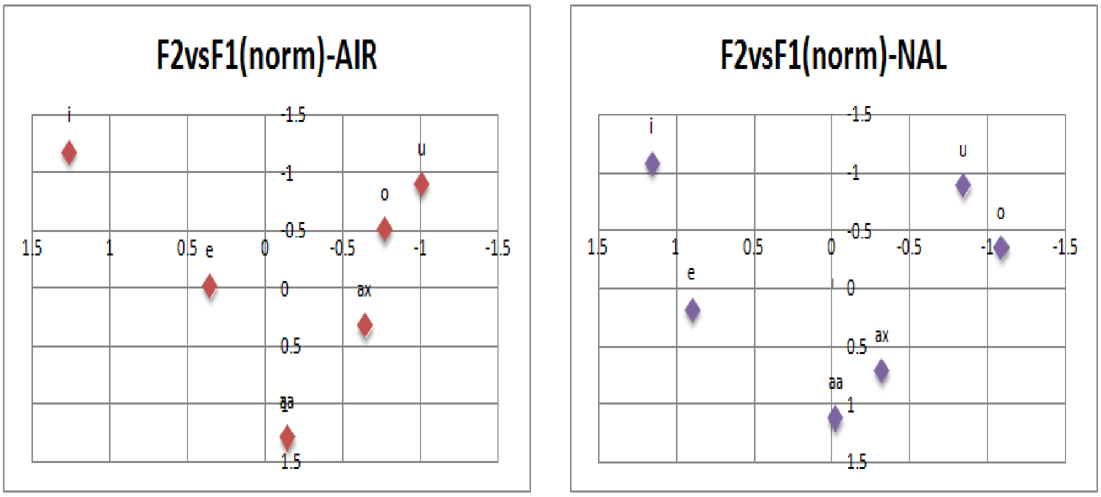


Figure 3.13. F2 vs F1 plots for NAL & AIR

After comparing the formant plots of the vowels uttered by the four speakers in Figure 3.13 as well as the normalised formant plots for AIR and Nalbaria in



(a)

(b)

Figure 3.14. Normalised vowel spaces for (a) AIR and (b) NAL

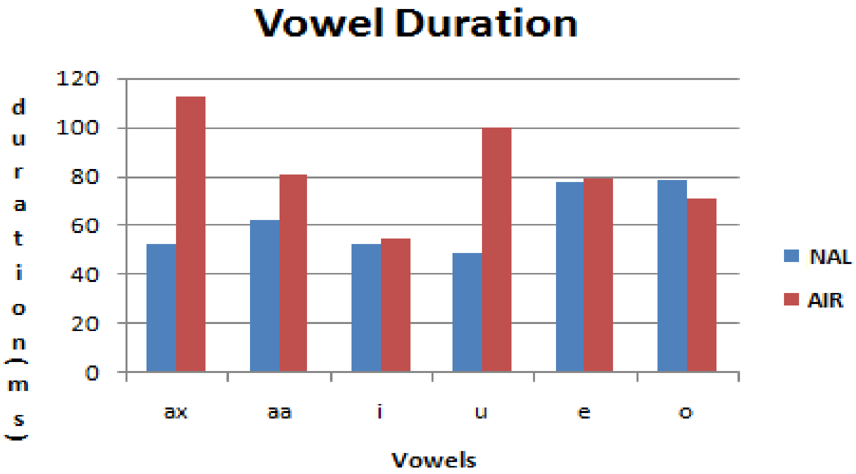


Figure 3.15. Average Duration of vowels in AIR and NAL

Figure 3.14, the following observations are made:

Observation 1: In the Nalbaria variety, the /ax/ is more close to the /aa/, i.e., the backness of /ax/ is less than that of the AIR variety.

Observation 2: In the Nalbaria variety, the /u/ is more central and /o/ is more back, while in the AIR variety, /u/ is more back and /o/ is more central. Or we can also say that in AIR /u/ is more commonly used while in NAL /ʊ/ is more commonly used.

Observation 3: The plot showing average vowel duration in the two varieties shows that the vowels in the Nalbaria variety are considerably shorter than their counterparts in the AIR variety. This may be attributed to the fact that Nalbaria speakers tend to speak faster than most AIR speakers.

3.4 Analysis of Diphthong formant trajectories

A diphthong is a complex vowel made of two components, the beginning vowel referred to as the *nucleus* and the end vowel referred to as the *offglide*. Formant movement in diphthongs starts about 30% to 40% of the way through the vowel and continues till the end. The 20% point (20% of the way through the diphthong) in a diphthong represents the nucleus whereas the offglide is usually represented by the 80% point. The amount of offglide movement in a diphthong can be measured by subtracting the first and second formant measures at the 20% point, i.e., the nucleus, from the same measures at the 80% point, i.e., the offglide[17]. A distance measure can be the absolute Euclidean distance between the nucleus value and the offglide value. The bigger the difference (distance), the easier it will be to hear the offglide. For example, the diphthong /aai/ has a big difference between the 80% and 20% points, while the diphthong /ei/ has only a small difference. As a result, it is easier to perceive the /aai/ diphthong. A diphthong however, has three critical points at which meaningful information can be derived from the formants [125]. They are the ‘Onglide’ which represents only the first vowel, the transition phase where there is a shift from the first vowel to the second, and the ‘Offglide’ which represents the second vowel of the diphthong. Therefore for diphthongs, formant measurements are generally taken at 20%, 40%, 60% and 80% to cover the onglide, offglide as well as the transition phase.

3.4.1 Literature Review

A number of studies have been carried out to analyse the acoustic nature of diphthongs and also to bring forth their cross-dialectal variation. Sarwar et al. [125] observed that the acoustics of diphthongs are dependent to a great extent on the speed of utterance, intonation, tone and rhythm. Studies include various measurements like distance between onglide and offglide, formant trajectory length and spectral rate of change of a diphthong. In another study, Mayr and Davies [92] revealed interesting cross-dialectal differences in spectral dynamics across their tra-

[17] https://depts.washington.edu/phonlab/resources/measuring-formants.pdf

jectories. Findings of this study suggest a number of cross-dialectal differences in the phonetic realisation of Welsh monophthongs. Furthermore, the results for spectral rate of change revealed interesting inter-dialectal and intra-dialectal differences in diphthong dynamics. Spectral change not only affected the extent of formant movement, but also the relative distribution of peaks and troughs across diphthong trajectories thereby revealing significant dialect-specific differences. Keerio et al. [71] analysed the diphthongs and glides of the Sindhi language and it is observed that the first two formants, F1 and F2, are sufficient to differentiate between vowels and phonemes whereas F2 plays a significant role to differentiate between the vowel-diphthong category and the glides. Another important point regarding diphthongised vowels, reported by Magen [89] in her study, is that for diphthongised vowels the perceptually dominant portion of the diphthong is variable.

3.4.2 Experimental Framework

3.4.2.1 Building of Speech Corpus

A list of words having frequently used diphthongs (such as /axi/, /aai/, /ei/, /oi/, /ui/, /ou/, /aau/, /eu/, /aao/, /eaa/, /iaa/, /uaa/, /ie/, /ue/, etc) in word initial, medial and end positions is prepared. The listed words are recorded from four speakers (two speaking the AIR variety & two speaking the NAL variety) at a sampling rate of 44.1kHz and 16 bit resolution in a noise free environment using a Sony recorder. Some frequently used diphthongs in the two varieties of Assamese along with examples of their carrier words are shown in Table 3.4

3.4.2.2 Formant measurement in Diphthongs

A PRAAT script has been written to extract the first two formants F1 and F2 at 20%, 40%, 60% and 80% of diphthong duration. The F1, F2 formants at the nucleus (20-25%) and offglide (75-80%) of the diphthong are also recorded in the excel sheet. The Euclidean distance between the nucleus and the offglide in each of the diphthongs is calculated using Equation 3.1 and recorded in the excel sheet.

$$d = \sqrt{(F1n - F1o)^2 + (F2n - F2o)^2} \tag{3.1}$$

Where ‘d’ is the Euclidean Distance, ‘F1n’, ‘F1o’ are the first formants at the nucleus and offglide of the diphthong respectively and ‘F2n’ and ‘F2o’ are the

Table 3.4 Frequently used diphthongs in AIR and NAL along with their carrier words

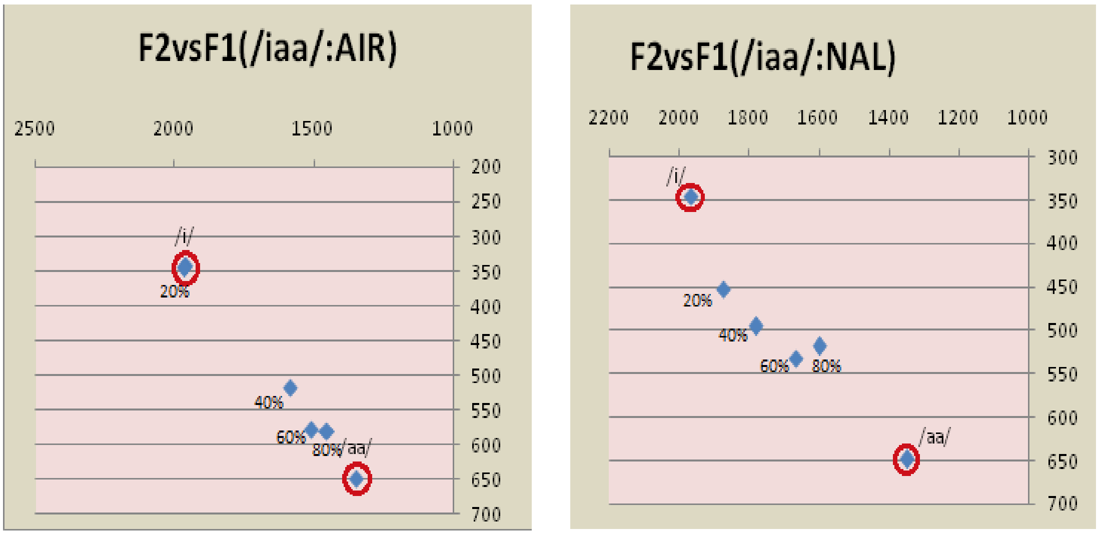
Diphthong	Carrier Word(AIR)	Carrier Word(N)
axi	maxi	naxhaxi
aai	naai	xaxdaai
ei	eijoni	geislu
ou	bou	thougaa
oi	hoise	boihaan
iaa	etiaa	giliaas
aao	raaonaa	gheraao
ui	xuinyax	puihbax
iu	xiu	naalgiliu
eaa	eaa	jeinbeaa
uaa	jonuaar	bhuaa

second formants at the nucleus and offglide respectively. For example, for the diphthong /eu/ in the word 'keusaa' in NAL, F1n=308, F1o=313, F2n=2107 and F2o=1684. Therefore, distance d=$\sqrt{(308-313)^2+(2107-1684)^2}$=422.

3.4.2.3 Methodology

Similar to vowel formants, the formants at 20%, 40%, 60% and 80% of diphthong duration are normalised. Among the frequently used diphthongs which have been taken up for analysis, the diphthong /iaa/ is taken as an example for representative purpose only, and the dynamic F2-F1 plot, i.e., plot of F2 versus F1 at 20%, 40%, 60% and 80% of vowel duration, of /iaa/ in the two varieties are plotted for comparison as shown in Figure 3.16. The average F1-F2 (AIR) values of the component vowels assuming AIR to be the standard variety are also plotted on the same graph for comparison.

The average values of distance between 'offglide' and 'onglide' for the commonly used diphthongs in the two varieties are also plotted in a bar graph in Figure 3.20 for comparison. Furthermore the range of distance values of diphthongs in the two varieties are also presented in Table 3.5 .



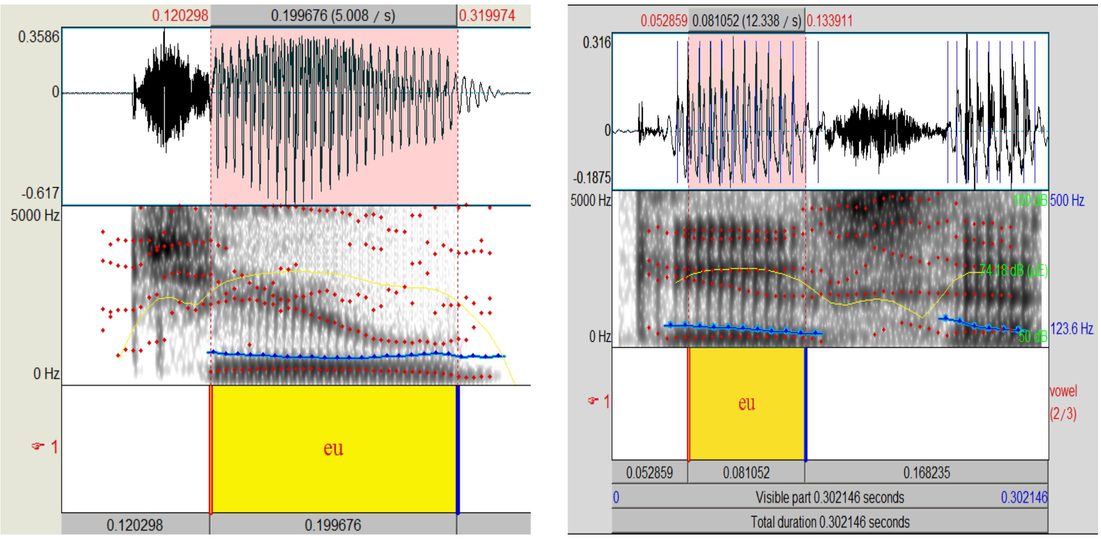
(a) **(b)**

Figure 3.16. Dynamic F2-F1 plot of a diphthong in (a) AIR & (b) NAL

3.4.3 Results and Observations

Comparing the dynamic F1F2 trajectory for the diphthong /iaa/ in AIR and NAL, the following observation is made:

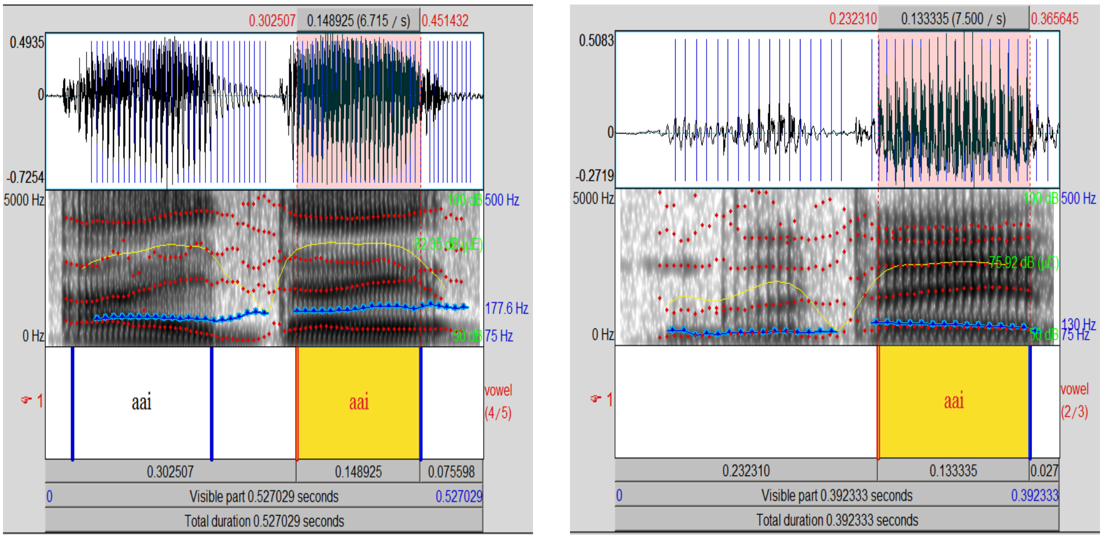
Observation 1: The dynamic F1F2 plot of the diphthong /iaa/ in AIR almost reaches the target vowel /aa/ while the dynamic F1F2 plot of the diphthong /iaa/ in Nalbaria lies somewhere between the vowel sounds /i/ and /aa/ .



(a) offglide=1172 **(b)** offglide=422

Figure 3.17. Diphthong /eu/ in (a) AIR and (b) NAL

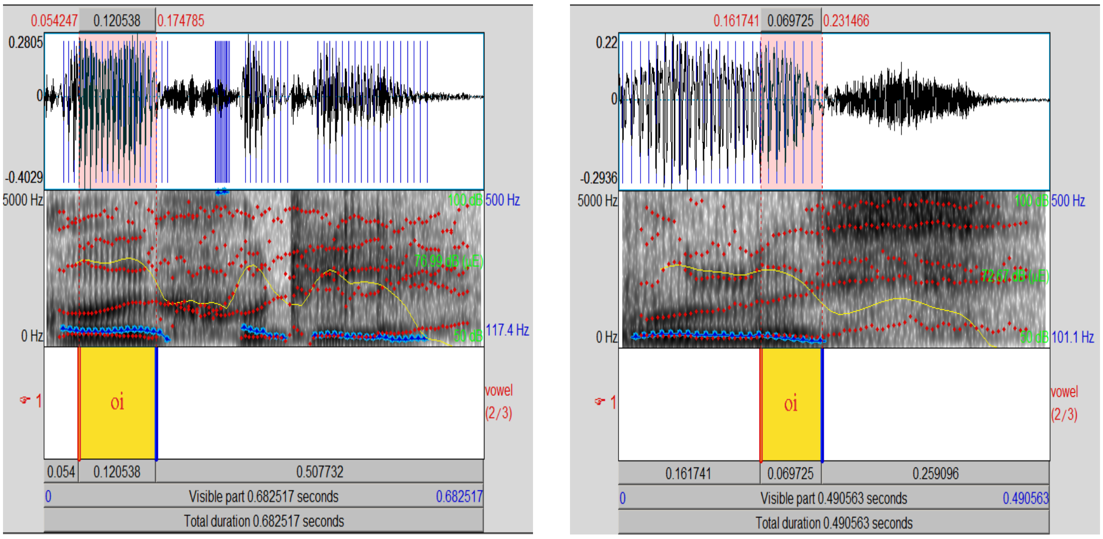
In Figure 3.17a, the diphthong /eu/ in AIR has a distance of 1172 whereas



(a) offglide=331

(b) offglide=182

Figure 3.18. Diphthong /aai/ in (a) AIR and (b) NAL



(a) offglide=323

(b) offglide=157

Figure 3.19. Diphthong /oi/ in (a) AIR and (b) NAL

in Nalbaria as seen in Figure 3.17b, has a distance of 396 between nucleus and offglide. Likewise, in Figure 3.18a, the diphthong /aai/ in AIR has a distance of 331 whereas in Nalbaria as seen in Figure 3.18b, has a distance of 182 between nucleus and offglide. Figure 3.19a shows the diphthong /oi/ in AIR with a distance of 323 whereas /oi/ in Nalbaria as seen in Figure 3.19b, has a distance of 157 between nucleus and offglide.

Observation 2: The distance between nucleus and offglide of some frequently used diphthongs in the two varieties of Assamese are calculated and plotted in

Figure 3.20 and it can be observed that in almost all cases the distance is much larger in the AIR diphthongs making them more prominent.

Table 3.5 Range of 'd' values of diphthongs for speakers in AIR and NAL. ('d': Euclidean distance between nucleus and offglide of a diphthong)

Spk1(N)	Spk2(N)	Spk3(AIR)	Spk4(AIR)
28<d<600	75<d<500	200<d<1350	500<d<1100

Observation 3: Table 3.5 shows that the distance between the nucleus and offglide for diphthongs in the AIR variety is usually much higher than that in the Nalbaria variety.

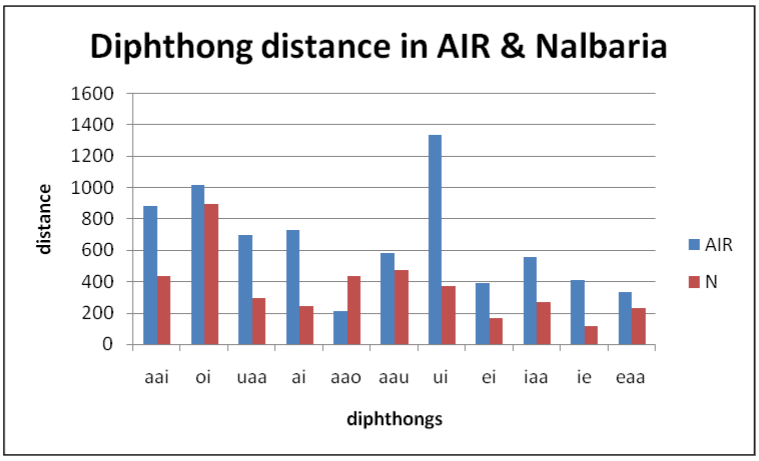


Figure 3.20. Distance between nucleus and offglide of some commonly used diphthongs in AIR and NAL

3.5 Analysis of Spectral Tilt

The term Spectral Tilt is used to describe the overall slope of the power spectral density (PSD). For speech, it is one of the many features responsible for the prosodic feature of accent. A speaker may modify the tilt, by raising the slope of the spectrum of a vowel, in order to put stress on a particular syllable. In phonetic research, spectral tilt is commonly used as a measure of creaky phonation. One of the major acoustic parameters that reliably differentiates the three basic phonation types, i.e., breathy, creaky and modal, in many languages is spectral tilt, i.e., the degree to which intensity drops off as frequency increases. Spectral tilt can be quantified by comparing the amplitude of the fundamental, i.e., H1, to that of higher frequency harmonics, e.g., the second harmonic (H2), the harmonic closest to the first formant (A1), or the harmonic closest to the second formant (A2). It is characteristically most steeply positive for creaky vowels and most steeply negative for breathy vowels.

3.5.1 Literature Review

Pharyngealisation is a secondary articulation that involves a retraction of the body and root of the tongue towards the pharyngeal wall. In Arabic, pharyngealisation is associated with dental/alveolar consonants, though it extends to other places of articulation also. Tamimi [4], in his works, evaluated the role of spectral tilt in describing the acoustic characteristics of pharyngealisation in Jordanian and Moroccan Arabic. Results show high classification rates implying that spectral tilt can be used to distinguish between pharyngealised vs non-pharyngealised vowels, advocating its importance as an acoustic cue for pharyngealisation in the Arabic dialects. Gordon and Ladefoged [44] in their cross-linguistic study on phonation types describe various features for differentiating phonation types in various languages; one such feature is the spectral tilt. In the creaky vowel, the amplitude of the second harmonic is slightly greater than that of the fundamental. At the other extreme, in the breathy vowel, the amplitude of the second harmonic is considerably less than that of the fundamental. Spectral tilt has been used by several researchers in works such as [34] and [68] for the detection or evaluation of prominence or emphasis of words in a sentence.

3.5.2 Experimental Framework

3.5.2.1 Building the Speech Corpus

In this analysis of the feature Spectral Tilt, we not only compare the tilt values in AIR and Nalbaria speech but also compare the tilt values with those in TTS generated Nalbaria. This is done for a better understanding of the differences between AIR, Nalbaria and TTS generated speech. Speech data is collected from 3-5 speakers in both AIR and Nalbaria. Care is taken to use similar text prompts for recording speech in the two varieties. Speech data is also generated from two Text-to-Speech systems built for the standard variety of Assamese. For our study we have generated speech from TU-TTS(The TTS that we have developed for the standard variety of Assamese) and IITG-HTS(TTS developed for the standard variety of Assamese at IIT-Guwahati). In this case, the prompts that are used for recording Nalbaria speech are transcribed and these transcriptions are provided as input to the respective TTSs.

3.5.2.2 Measuring Spectral Tilt

The first step towards measuring spectral tilt is to take the spectral slice of each vowel in the sample data. Spectral slices are the results of a fast fourier transform (FFT) done on a very small portion of the speech signal. It provides very specific information about the frequencies present in the speech segment and their relative amplitudes. A spectral slice as shown in Figure 3.21 shows amplitude on the Y-axis and frequency (from 0 upto the Nyquist frequency) on the X-axis and can be easily displayed in PRAAT and the required amplitudes can be measured manually. Harmonic amplitudes H1, H2, A1, A2 of a vowel segment can be seen in Figure 3.21.

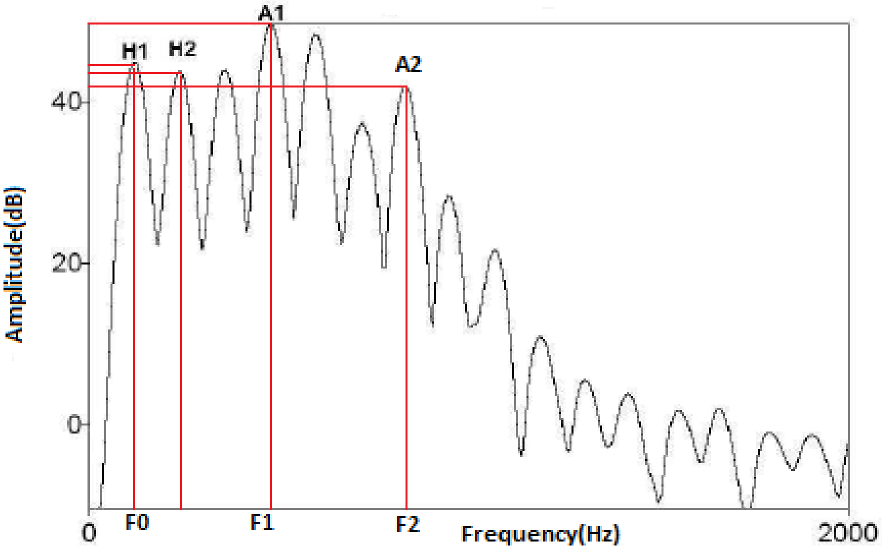


Figure 3.21. H1, H2, A1, A2 in a vowel segment

3.5.2.3 Methodology

The speech files in the speech corpus, both recorded from AIR and NAL speakers, and TTS generated, are annotated in PRAAT using textgrids. A PRAAT script is written to extract the tilt values of H1-H2, H1-A1 and H1-A2 for each segment in the vowel tier in the textgrids, where H1 and H2 are the amplitudes of the first two harmonics and A1 and A2 are the amplitudes of the first two formants F1 and F2. The script also saves the measured values in a text file from where the data is imported into an Excel file. The Excel file contains tilt values of H1-H2, H1-A1 and H1-A2 for vowels in the three varieties, i.e., AIR, Nalbaria and synthesised Nalbaria (from standard TTS).

3.5.3 Results and Observations

Boxplots used to plot the measured tilt values of H1-H2, H1-A1 and H1-A2 for the three varieties, i.e., AIR, Nalbaria and synthesised Nalbaria (from standard TTS) are presented in Figures 3.22a, 3.22b and 3.22c.

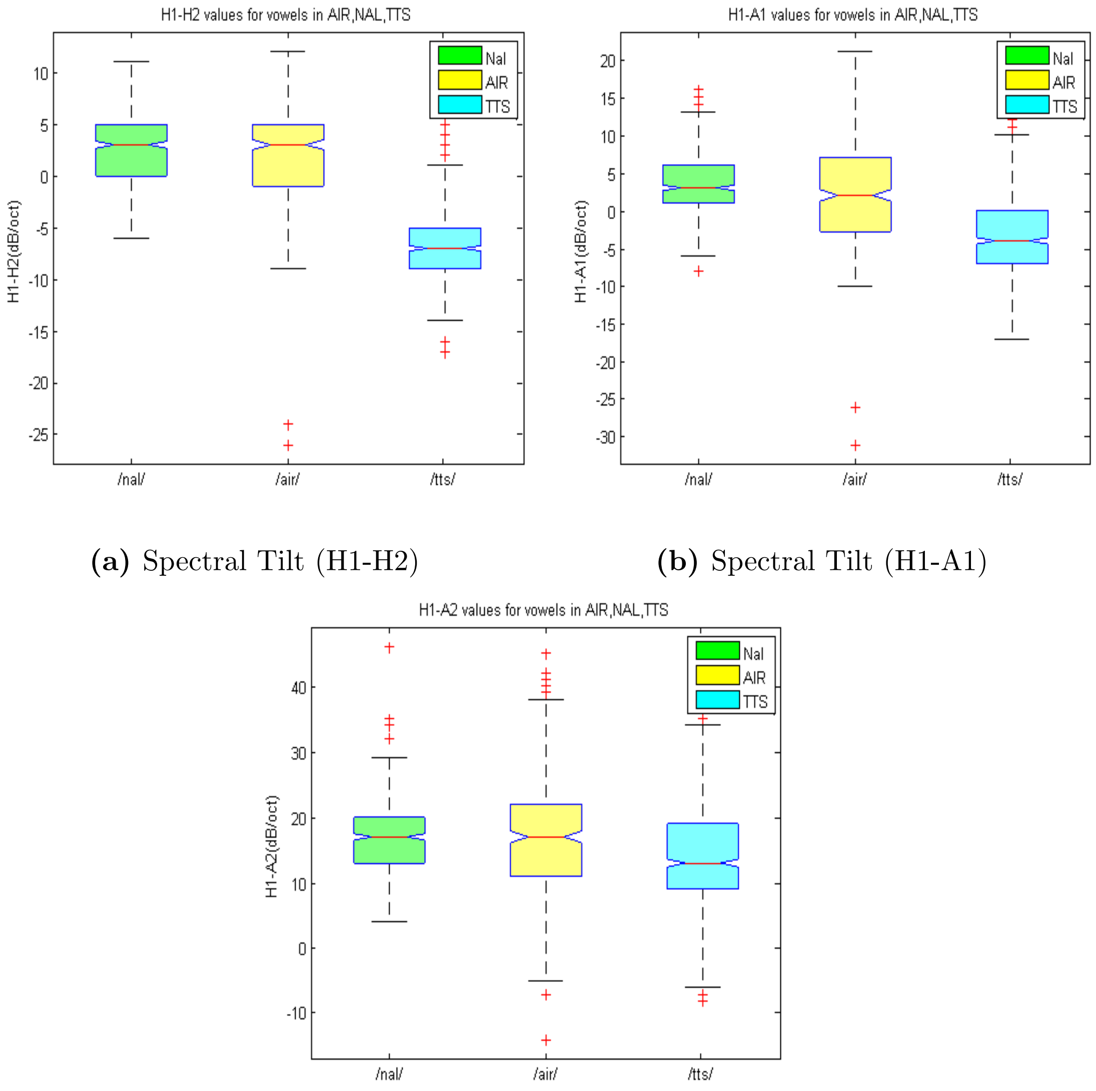
Observation 1: Box plots representing spectral tilt data in the three varieties, show that negative values of H1-H2 and H1-A1 in AIR vowels extend beyond the 1st quartile indicating more instances of breathy vowels in AIR than in NAL.

Observation 2: In case of TTS generated vowels, values of H1-H2 and H1-A1 are most steeply negative indicating breathy vowels.

3.6 Analysis of Spectral (MCEP) and Prosodic Features (F0)

Dialect specific information is known to be present at different levels of the speech signal. At the segmental level such information can be observed in the form of unique sequences of the shapes of the vocal tract for producing the sound units. The shape of the vocal tract is characterised by the spectral envelope which can be represented by Mel frequency cepstral coefficients (MFCCs). MFCCs have been ascertained as the state-of-the-art for feature extraction especially since it is based on actual human auditory system and has a perceptual frequency scale called Mel-frequency scale. In speech processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound[18]. Such a representation is based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. MFCCs as well as MCEPs are coefficients that collectively make up a Mel Frequency Cepstrum(MFC). They are derived from a type of cepstral representation of the speech signal. In our work we have used Mel Cepstral

[18] https://en.wikipedia.org/wiki/Mel-frequency_cepstrum



(a) Spectral Tilt (H1-H2)

(b) Spectral Tilt (H1-A1)

(c) Spectral Tilt (H1-A2)

Figure 3.22. Spectral Tilt values in (a) AIR, (b) NAL and (c) TTS speech data

Coefficients or MCEPs for analysing the spectrum because like MFCCs they are also used to represent the speech spectrum, and at the same time they can be directly used with the Mel Log Spectral Approximation (MLSA) filter for synthesising speech using the Speech Processing Toolkit (SPTK). At the suprasegmental level, the dialect specific knowledge is embedded in the duration patterns of the syllable sequences and the dynamics of the pitch (F0) and energy contours. In linguistics, speech synthesis, and music, the pitch contour of a sound is a function or curve that tracks the perceived pitch of the sound over time[19]. The pitch contour is often considered to be an important signal of linguistic stress and therefore we

[19] https://en.wikipedia.org/wiki/Pitch_contour

have considered this as a candidate for our analysis.

3.6.1 Literature Review

Mel cepstral coefficients, specially MFCCs, and pitch (F0), are very popular features in the field of dialect identification. Rao [117] showed that using prosodic features such as pitch together with MFCCs, increases the performance of a dialect recognition system. Various other works [56], [90] and [57], are seen using these features for identification/recognition of dialects indicating that these features are distinctive in many dialects.

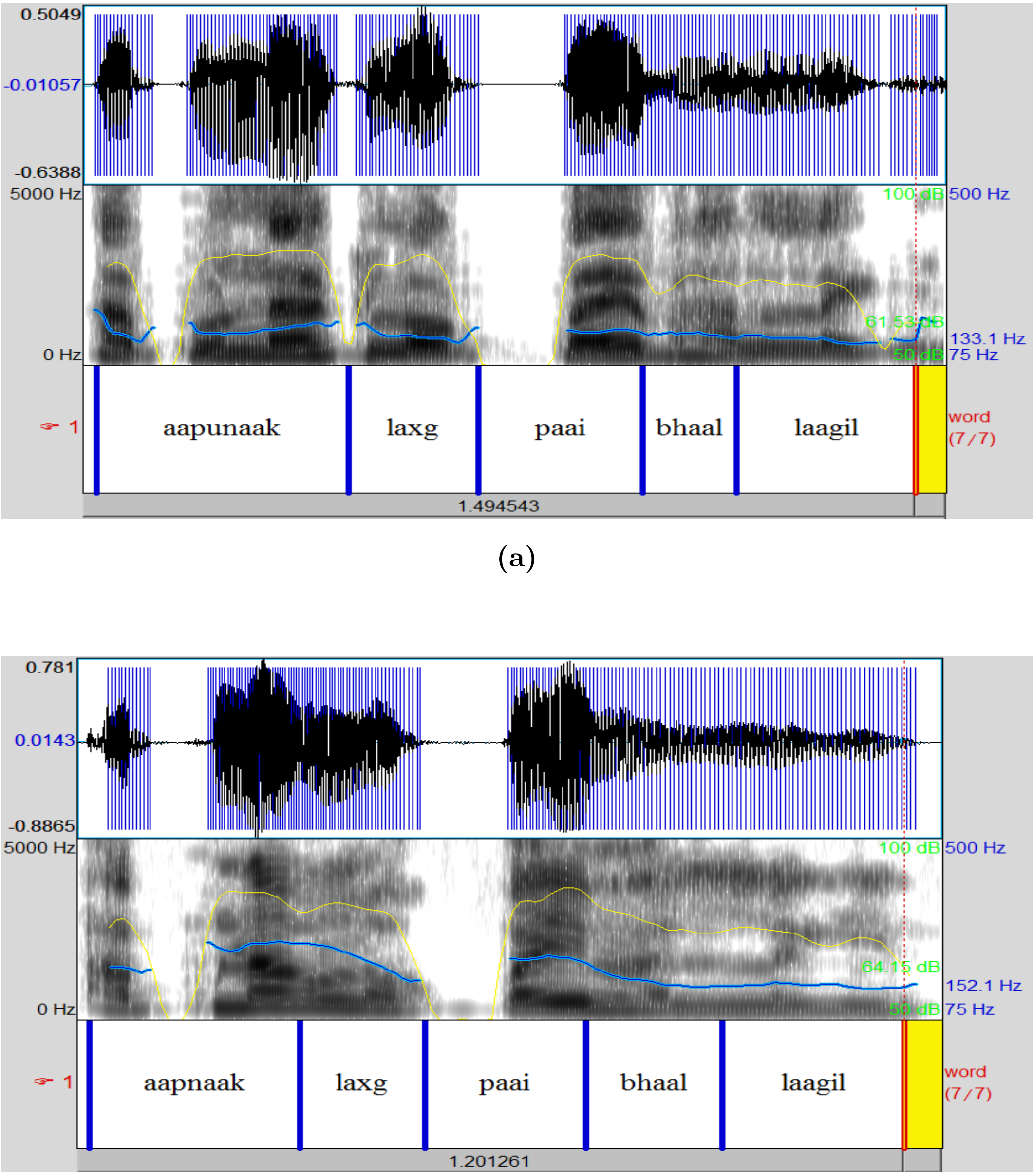
3.6.2 Experimental Framework

3.6.2.1 Building of Speech Corpus

A set of ten sentences are recorded from three speakers in Nalbaria and from three speakers in AIR. The data is cleaned to remove silence and noise, and kept in two separate folders, one for AIR and another for Nalbaria. This corpus, C_M containing 10 sentences in AIR and 10 in Nalbaria from three speakers in AIR and three in NAL, i.e., a total of sixty short sentences, are used for analysing the cepstral coefficients of the two dialectal varieties of Assamese under consideration. In order to compare and analyse the pitch contours in the two varieties, a set of fifteen sentences, five in each sentence type (declarative, interrogative and exclamatory), are recorded from three speakers in AIR and three in Nalbaria. The speakers are asked to speak in the three different styles as naturally as possible. The corpus C_P therefore contains a total of 90 short sentences. Sample pitch contours in the two varieties in the three different styles, are presented in Figures 3.23a, 3.23b and 3.24a, 3.24b.

3.6.2.2 Methodology

The SPTK toolkit is used to extract 21^{st} order mel cepstral coefficients (MCEPs) from the wav files in C_M using the 'mcep' function. A MATLAB script is written to read the MCEP files, concatenate them and store in two separate matrices of dimension $[m \times n]$ where 'm' is the number of samples and 'n' is the order of coefficients, one for AIR and one for NAL. The MCEPs for the two varieties are

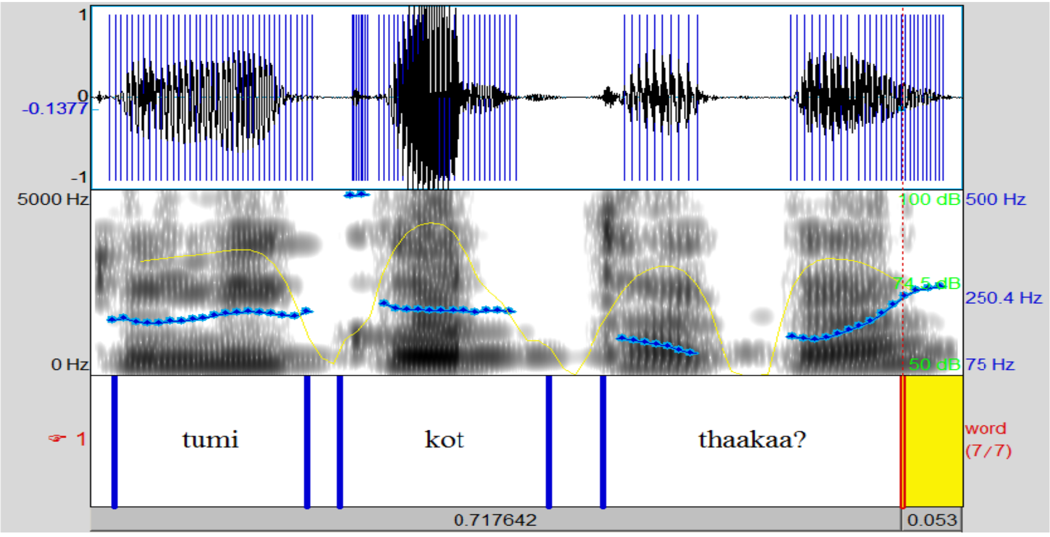


(a)

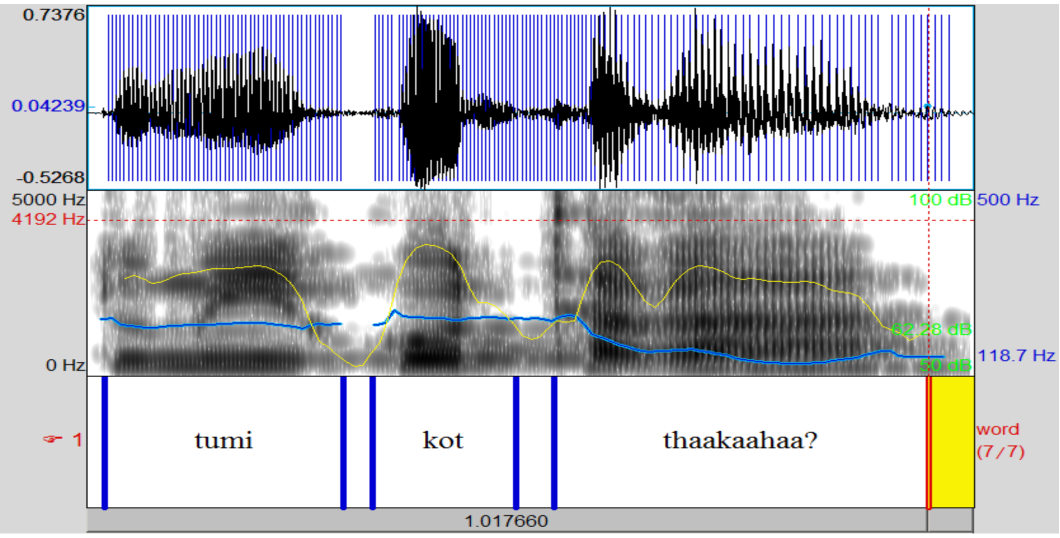
(b)

Figure 3.23. Pitch contours in (a) AIR and (b) NAL (Declarative)

plotted for comparison in Figure 3.26. The MCEPs are normalised using the min-max normalisation method to have them in the range of 0 and 1. Figure 3.27 shows normalised MCEPs in the two varieties. Gaussian Mixture Models (GMMs) are then built to model separately the MCEPs of AIR and NAL with the number of mixtures randomly set to 8. The distance between these two GMMs is measured using the popular Kullback-Leibler Divergence (KLD) [51]. KLD is a measure which is based on relative entropy, and is often used as a measure of difference between two probability distributions. The KLD values between the corresponding mixtures are also plotted for comparison in Figure 3.28.



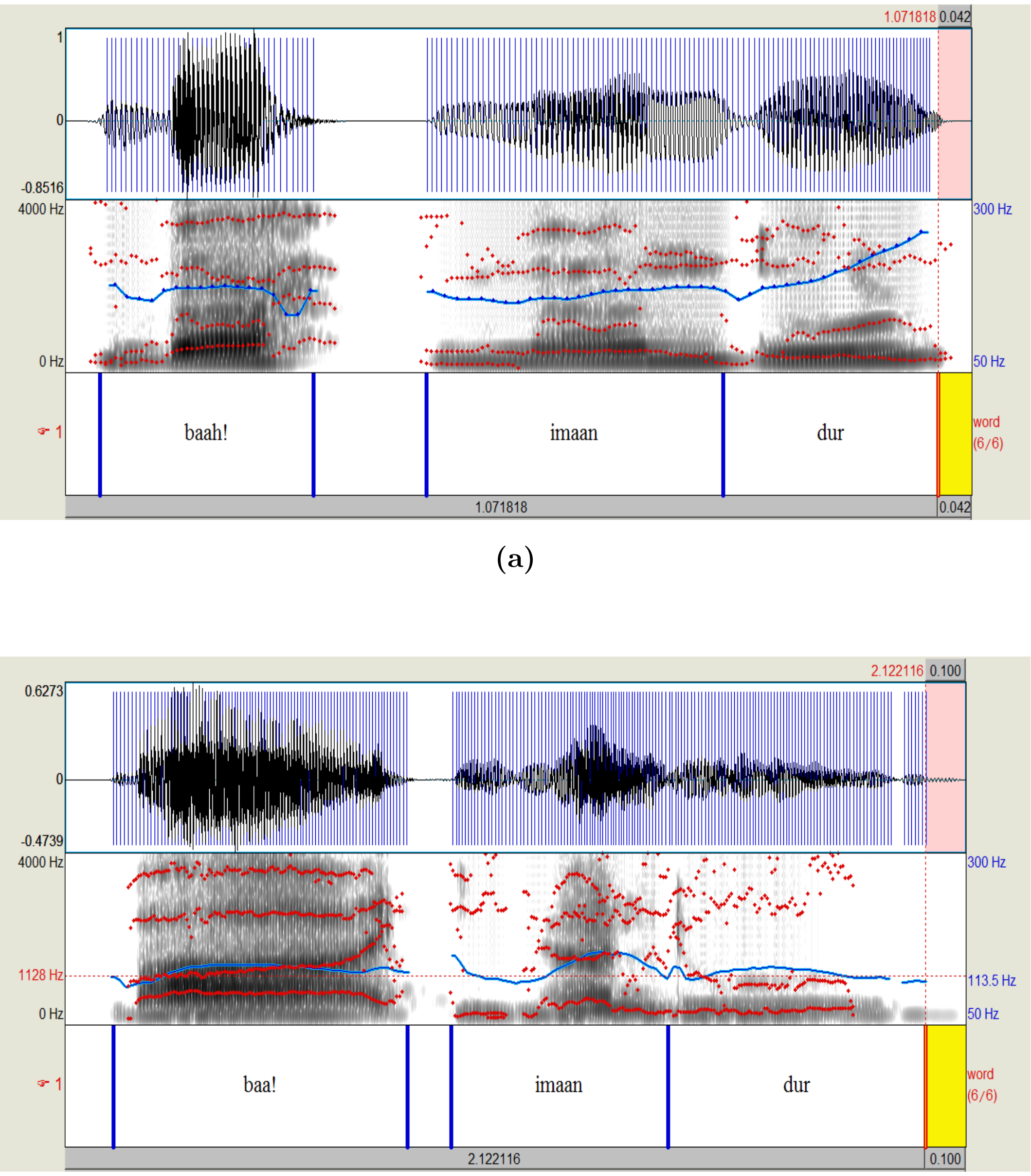
(a)



(b)

Figure 3.24. Pitch contours in (a) AIR and (b) NAL (Interrogative)

Pitch contours from the wav files in C_P are extracted in Praat, and a Praat script is also written to measure minimum, maximum, mean and standard deviation values of F0 in each of the sentences for the different speakers. Bar diagrams presented in Figure 3.29, 3.30, 3.31 and 3.32 are used to compare the minimum, maximum, mean and standard deviation values of F0 for the sentences in the three different styles considered, i.e., declarative, interrogative and exclamatory, in AIR and NAL.



(a)

(b)

Figure 3.25. Pitch contours in (a) AIR and (b) NAL (Exclamatory)

3.6.3 Results and Observations

The analysis on MCEP and F0 values in AIR and NAL speech data leads to the following observations:

Observation 1: A limitation of the cepstral coefficients is the lack of physical interpretation. The 0^{th} order coefficients, c0, is the power over all frequency bands, and 1^{st} order coefficients, c1, is the balance between low and high frequency components within the signal frame. The other cepstral coefficients have no clear interpretation other than that they contain the finer detail of the spectrum to dis-

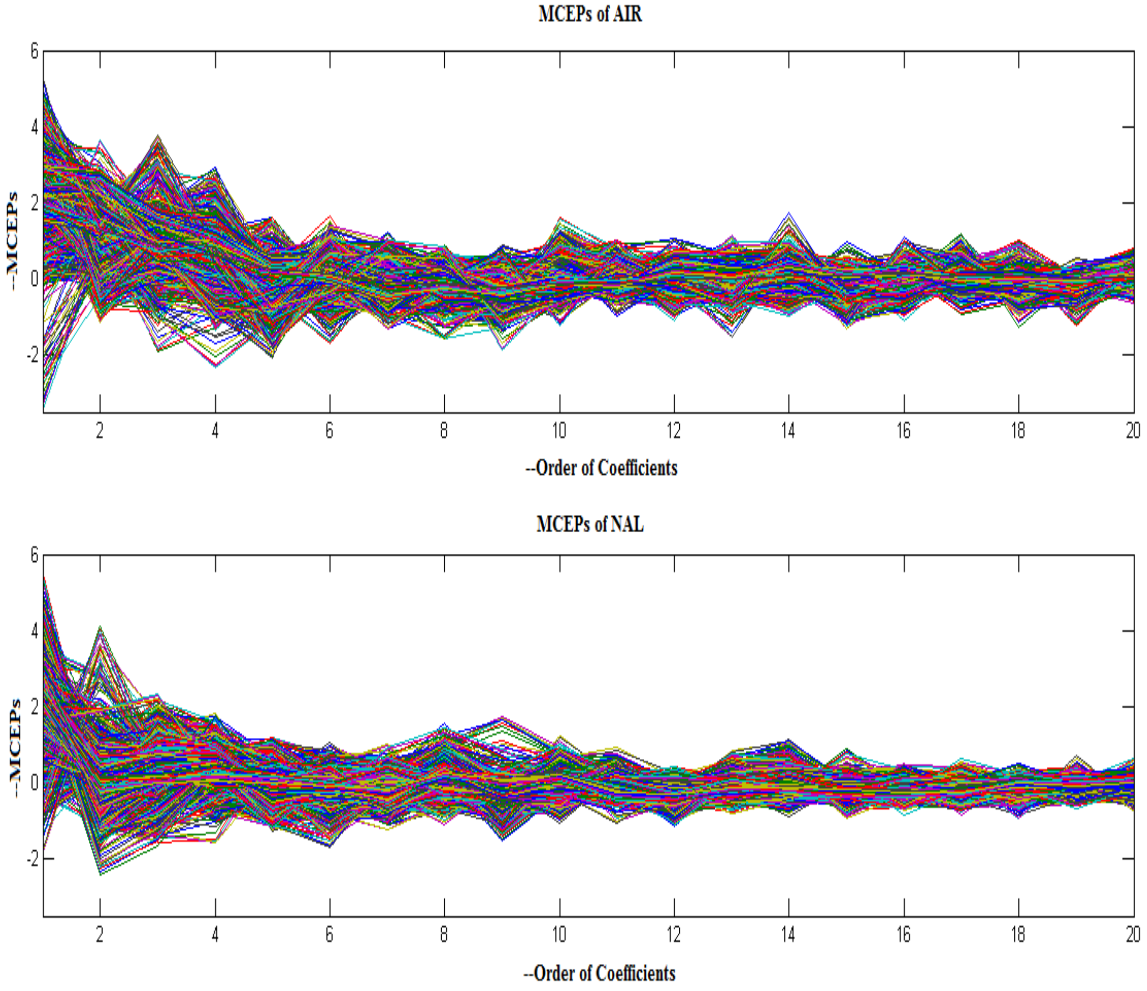


Figure 3.26. Mel Cepstral Coefficients in AIR and Nalbaria speech data

criminate the sounds. However it can be seen from the normalised MCEP plots that the 1^{st} order MCEP coefficients in AIR lie in the range of 0.4 to 1 while in NAL they lie in the range of 0 to 1. For higher order coefficients too, the variation in values is higher in NAL.

Observation 2: The approximate KL divergence and also the $[8 \times 8]$ matrix containing the exact KL divergence between the corresponding components of the two GMMs, as indicated in Figure 3.28, further establish the fact that the GMMs modeling the MCEPs in AIR and Nalbaria, are also different. From the figure it can also be seen that one component i.e., the first mixture components of the two GMMs modeling AIR and Nalbaria speech data has the highest KLD value indicating greater distance. This indicates that MCEPs may also be used as a distinctive feature for the two varieties under study, i.e., AIR and Nalbaria. However, further investigations and experiments are required to find whether MCEPs contain dialect

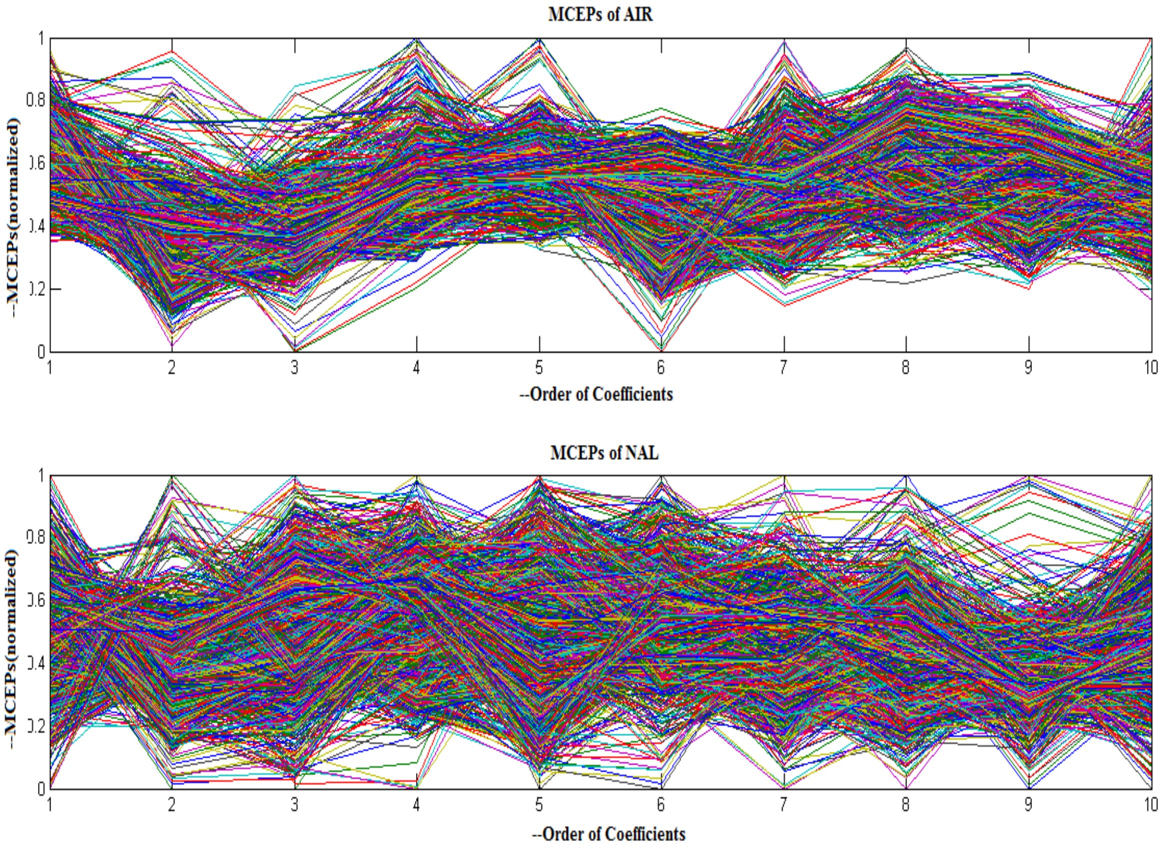


Figure 3.27. Normalised Mel Cepstral Coefficients in AIR and Nalbaria speech data

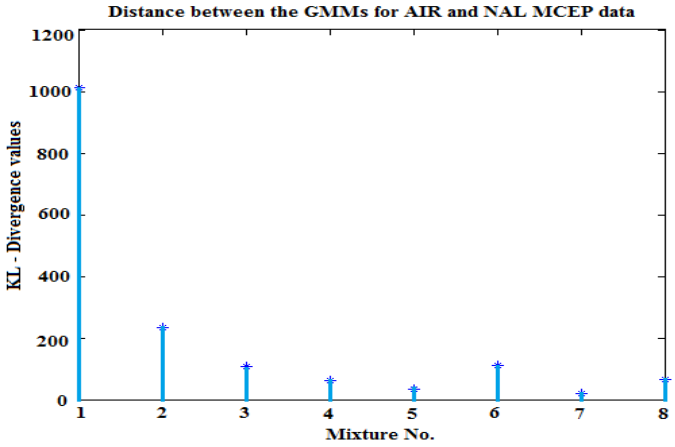


Figure 3.28. KLD distance between mixtures in AIR and Nalbaria GMMs

specific information or not.

Observation 3: For the set of interrogative sentences, there is no significant

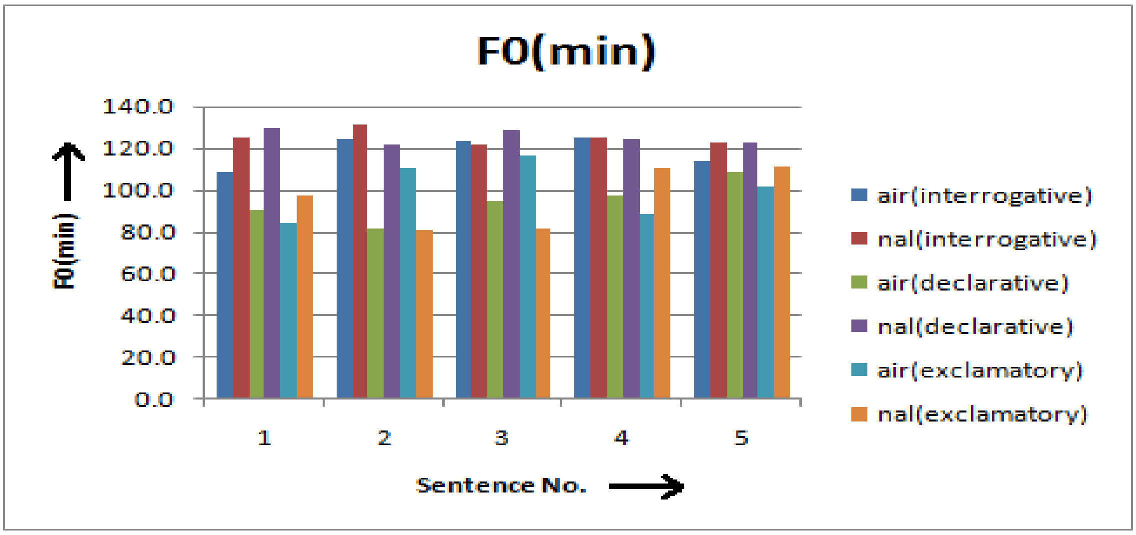


Figure 3.29. F0(min) for utterances (Declarative, Interrogative, Exclamatory) in AIR and NAL

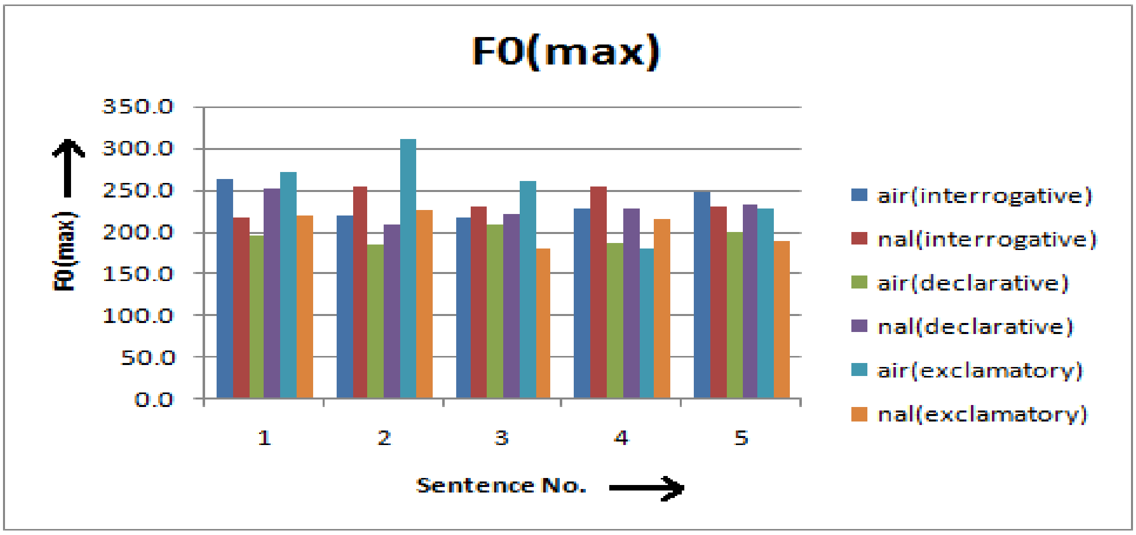


Figure 3.30. F0(max) for utterances (Declarative, Interrogative, Exclamatory) in AIR and NAL

difference in the average values of F0 in AIR and NAL. For the set of declarative sentences, average F0 values in NAL are higher than that in AIR. Exclamatory sentences in AIR have higher F0 values compared to those in NAL.

Observation 4: Another noticeable difference is that F0 values deviate much more from their means in the set of exclamatory sentences in the two varieties than in the declarative and interrogative sentences.

Observation 5: From the example pitch contours of an interrogative sentence in AIR and Nalbaria, it is observed that the contour is rising at the end of the

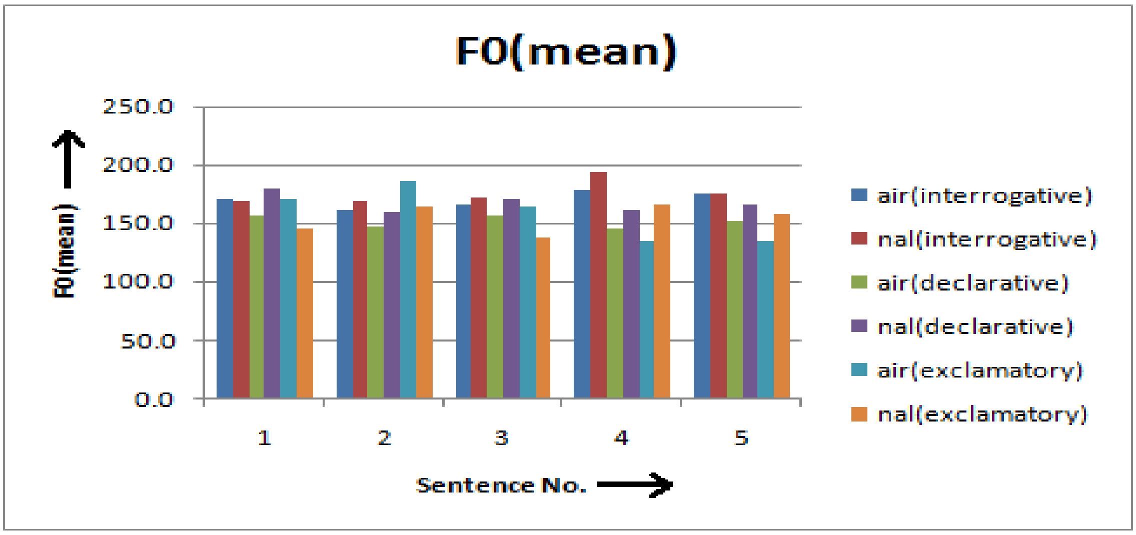


Figure 3.31. F0(mean) for utterances (Declarative, Interrogative, Exclamatory) in AIR and NAL

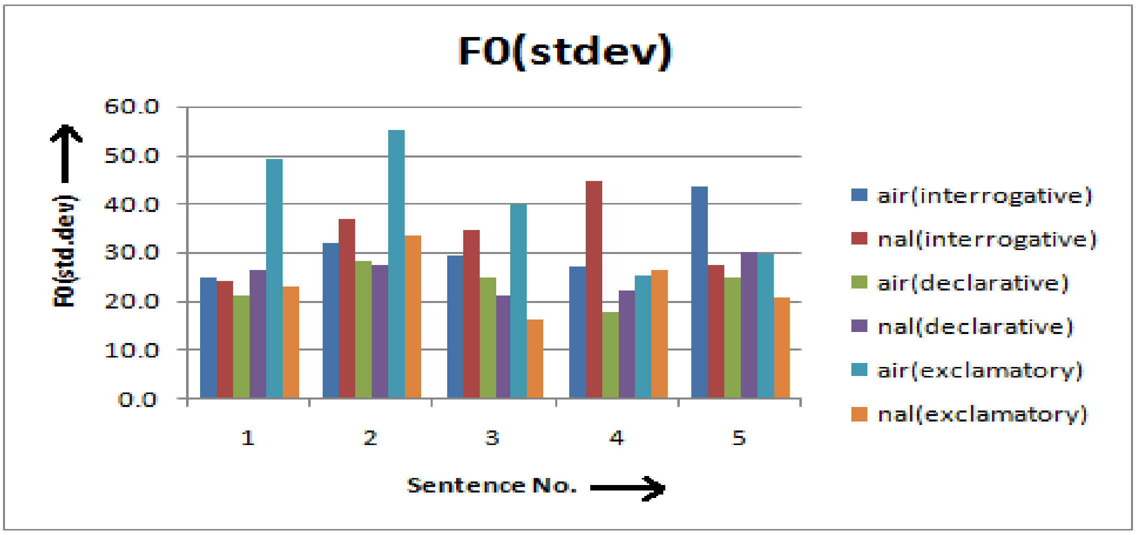


Figure 3.32. F0(std.dev.) for utterances (Declarative, Interrogative, Exclamatory) in AIR and NAL

sentence in AIR while it is almost flat in Nalbaria. This may be because of the extra segment 'haa' which is usually appended to the end of most interrogative sentences in Nalbaria. At the same time it is seen that the pitch contour in the declarative sentence in AIR is much flatter, indicating a flatter tone.

3.7 Discussion

A total of seven features, namely ‘VOT’, ‘Vowel Space’, ‘Vowel Duration’, ‘Diphthong Formant Trajectories’, ‘Spectral Tilt’, ‘Mel Cepstral Coefficients’ and ‘Pitch Contours’, have been analysed in this study. It has already been mentioned that the choice of features for analysis is not random but based on a preliminary perceptual query conducted on a few speakers familiar to both the varieties of Assamese under study. All of these features, in varied extents, have shown some amount of distinctness relative to a particular variety. For more conclusive results the experiments need to be carried out in a more elaborate manner with more number of speakers and the speech corpora also needs to be expanded to cover a wide variety of text. However the results obtained seem to comply with the differences pointed out in the preliminary analysis of speech data from two varieties of Assamese in Section 3.1.4. A point to note is that for some of the experiments, speech data in the two varieties include speech from a single speaker fluent in the two varieties. So we cannot always say that the features may be speaker dependent. From the synthesis point of view, we only need a qualitative comparison of features not quantitative. More experiments can obviously provide a more reliable comparison but for our work a qualitative comparison would be equally sufficient.

Cepstral coefficients are a popular representation of the speech signal and is known to characterise the vocal tract which acts as the filter while producing speech. Pitch is equally important with respect to the prosody of speech. Our review of literature shows that these two features can be easily converted from one variety to another using the method of Voice Conversion. Therefore the first approach that we have used as elaborated in the next chapter, i.e., Chapter 4, is to use the method of VC to convert from one variety of Assamese to another.

It is also observed that out of the seven features that have been explored, the formant space, i.e., F2 versus F1, is significantly different in the two varieties, AIR and Nalbaria. Position of vowels in the formant space is an indication of how the vowel sounds are produced and therefore affects the perception of the vowels and diphthongs which comprise two or more vowels. Therefore our second approach towards incorporation of dialect-distinctive features in synthesised speech, is to transform the vowel formant space (VFS) from one variety to another, in a bid to make the vowel/diphthong sounds in one variety sound closer to their counterparts in the other variety. This is elaborated in Chapter 5.