Exploring Phonological Concepts in Hindi in the context of a Short-Vocabulary Speech Corpus

Aashish Yadavally AASHISH.YADAVALLY@UGA.EDU
Institute of Artificial Intelligence
University of Georgia

Shrinidhi Adke SHRINIDHI.ADKE@UGA.EDU
Institute of Artificial Intelligence
University of Georgia

Abstract

In this paper, we describe two aspects of the work we carried out as a part of this project: building a short-vocabulary Hindi speech corpus; studying the phonetic and phonological aspects of Hindi in the context of the data in this corpus. The basic thought behind building the speech corpus stems from it's use for the purpose of Automatic Speech Recognition. The larger goal is to extend it to facilitate the recognition of large vocabulary, continuously spoken fluently by any native speaker(1). We used a crawler (2), to automatically crawl through URLs of YouTube videos and retrieve them as audio files, which we further trimmed into individual sentences for transcribing them into Hindi text and their corresponding phonemic transcriptions. Secondly, we performed acoustic and phonemic analysis on the corpus, using it to study two different hypothesis, corresponding to the phonological aspects in Hindi.

1. Introduction

The task of creating a corpus consisting of phonetically rich sentences is very important, because it is not practical to use a large set of data for automatic speech recognition and synthesis (3). This can be achieved by performing analysis on a large corpus, and extracting sentences from it based on the uniform distribution of the phonemes in the extracted sentences. The construction of a phonetically rich corpus is very significant to the applications automatic speech recognition (ASR) and speech synthesis (TTS).

According to Ethnologue 2019 (4), Hindi is the most natively spoken language in the world after Mandarin, Spanish, and English. It is recognized as an official language of India in Devanagari script. There are a total of 33 consonants and 13 vowels used for speaking, as well as writing in Hindi (5). On the basis of place of origin, the consonants can be largely categorized into four categories, namely, Regular Consonants (25 in number), Semivowel Consonants (4 in number), Sibilant Consonants (3 in number), and Fricative Consonants (1 in number).

Acoustic study of the stop consonants is one of the most challenging tasks in Speech Recognition due to the dynamic variable context and speaker-dependent nature of stops.(6) We study the unaspirated stop consonants in the initial position in a CVC context with three mainly used vowels in hindi, /a:,i,u/. We have eight stop consonant classes of different place of articulations taking initial place of CVC syllables. The acoustic parameters like Voice Onset Time(VOT), Burst Duration(BD), Burst Frequency(BF) are measured and studied in a further section in more detail.

Group Name	Tongue Position	Group Members				
		VL*	VLA#	V¶	VA§	N†
क वर्ग	Velar	क	ख	ग	घ	ङ
च वर्ग	Palatal	च	छ	ज	झ	ञ
ट वर्ग	Retroflex	ट	ठ	ड	ढ	ण
त वर्ग	Dental	त	थ	द	ध	न
प वर्ग	Labial	प	फ	ब	भ	म

*VL = VOICE LESS
#VLA = VOICE LESS ASPIRATED
¶V = VOICED
§VA = VOICED ASPIRATED
†N = NASAL

Figure 1: Regular Consonants Categorized According to Tongue Position During Articulation (5)

Hindi Orthography	IPA Symbol	PLU Symbol	Example
अ	/ə/	A	अध्यापक 'Teacher'
आ	/a:/	AA	आम 'Mango'
इ	/i/	I	इत्र 'Perfume'
ई	/i:/	II	ईश्वर 'God'
उ	/u/	U	उत्तर 'Answer'
ऊ	/u:/	UU	ऊष्मा 'Heat'
ए	/e:/	E	एक 'One'
ऐ	/æ:/	EE	ऐनक 'Specs'
ओ	/o:/	O	ओस 'Dew'
औ	/ɔ:/	OO	औषधि 'Medicine'
अं	/aŋ/	MN	अंगूर 'Grape'
अः	/əh/	AH	अतः 'So'
ऋ	/r̥/	RI	ऋषि 'Sage'

Figure 2: Glance on Hindi Vowels with Example and Phonemic Presentation (5)

All Indian languages scripts are phonetic in nature and share a common phonetic base. Graphemes are the basic units of all language scripts, and are orthographic representations of speech sounds. The presence of retroflex consonants, fewer fricatives (compared to English/European languages), various types of taps or trills are some of the distinct properties of Indian languages (7). The mapping between the vowel graphemes with it's corresponding IPA symbols is represented in Figure 2, and that between the consonant graphemes with it's corresponding IPA symbols is represented in Figure 3.

2. Methodology for Constructing the Corpus

In this section, we describe the strategy followed towards building the short-vocabulary Hindi corpus. This dataset was built in the context of automatic speech recognition (ASR), and was further used as data to investigate a few phonological concepts in Section 4 and Section 5.

2.1 Crawler

Liao et al, in (8) and Lecouteux et al in (9) demonstrated that YouTube can be used to create a speech corpus, by using the YouTube API as a crawler, and the captions in YouTube videos as the text transcriptions for the audio files retrieved from their corresponding videos. The thought behind creating a dataset by crawling (10) through the videos in YouTube and retrieving them as audio files stemmed from this research.

As in the "\src" directory in (2), we wrote a Python wrapper script around "youtube-dl"(11), to retrieve the videos from YouTube as audio files. Each of these retrieved audio files were saved in the default "\audios" directory. Furthermore, the audio files were further trimmed into sentences using Audacity (12), and the trimmed audio files were saved in the "\corpus\data" sub-directory.

2.2 Files in the Corpus

The corpus was designed based on the files required for the acoustic modeling and language modeling, as in Kaldi (13), the speech-recognition toolkit. A "text.txt" file was created, containing the Hindi script transcriptions for each of the sentences, with their corresponding utterance ID's. Furthermore, a "speaker_to_gender.txt" file was created mapping each of the speakers in the audio files to their respective gender. These are the files generally required for the acoustic modeling in a corpus for Automatic Speech Recognition.

Furthermore, a "lexicon.txt" file and a "nonsilence_phones.txt" file were added, which are generally used for language modeling. The former contains information mapping all the words in the dataset with their corresponding phoneme transcriptions, while the latter includes all the individual phonemes used in the dataset. Also, there exists a "mapping.txt" file in the corpus which explains the one-on-one mapping relation between each of the graphemes with it's corresponding IPA symbols.

Hindi Orthography	IPA Symbol	PLU Symbol	Example
क	/k/	K	कर ‘Tax’
ख	/k^{h}/	KH	खेल ‘Sport’
ग	/g/	G	गाय ‘Cow’
घ	/$g^{ɦ}$/	GH	घर ‘House’
ङ	/ŋ/	WN	पङ्क ‘Mud’
च	/ʧ/	CH	चमत्कार ‘Miracle’
छ	/$ʧ^{h}$/	CHH	छाया ‘Shadow’
ज	/ʤ/	J	जल ‘Water’
झ	/$ʤ^{ɦ}$/	JH	झरना ‘Waterfall’
ञ	/ɲ/	ZN	कञ्चे ‘Marbles’
ट	/ʈ/	TT	टहलना ‘Amble’
ठ	/$ʈ^{h}$/	TTH	ठीक ‘Fine’
ड	/ɖ/	DD	डाक ‘Post’
ढ	/$ɖ^{ɦ}$/	DDH	ढोल ‘Drum’
ण	/ɳ/	AN	गणना ‘Calculation’
त	/t/	T	तरबूज ‘Watermelon’
थ	/t^{h}/	TH	थाना ‘Station’
द	/d/	D	दर्पण ‘Mirror’
ध	/$d^{ɦ}$/	DH	धातु ‘Metal’
न	/n/	N	नमक ‘Salt’
प	/p/	P	परोपकार ‘Charity’
फ	/p^{h}/	PH	फल ‘Fruit’
ब	/b/	B	बरसात ‘Rain’
भ	/$b^{ɦ}$/	BH	भक्ति ‘Devotion’
म	/m/	M	मदद ‘Help’
य	/j/	Y	यंत्र ‘Instrument’
र	/ɾ/	R	रक्षा ‘Defence’
ल	/l/	L	लगन ‘Diligence’
व	/ʋ/	V	वास्तु ‘Architectural’
स	/s/	S	समुदाय ‘Community’
ष	/ʂ/	SHH	धनुष ‘Bow’
श	/ʃ/	SH	शक्ति ‘Power’
ह	/ɦ/	H	हल ‘solution’

^{h}VOICELESS ASPIRATED
^{ɦ}VOICED ASPIRATED

Figure 3: Glance on Hindi Consonants with Example and Phonemic Presentation (5)

3. Corpus Analysis

In all, there are around 162 audio files that we considered for the corpus, each generated from four particular audio files retrieved from YouTube. There are two Male speakers, and two female speakers in the audio files used in the corpus. The Hindi script transcriptions of each of the audio files have been recorded in the ”text.txt” file. This gets the dictionary

to around 1070 words, for each of which we have written a phonemic transcription in the "lexicon.txt" file.

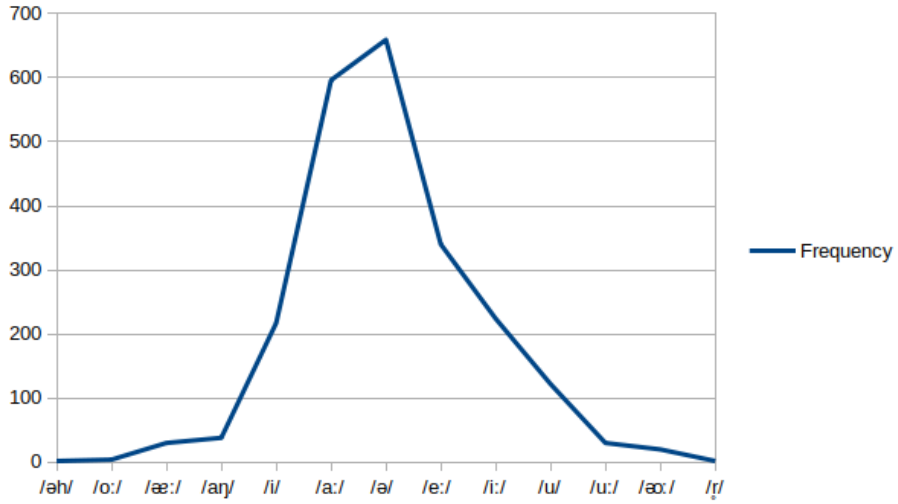

Figure 4: Distribution of Vowels in the Corpus

All the individual phonemes used in the corpus can be tracked in the "nonsilence_phones.txt" file. The distribution of the vowels according to their respective counts in the corpus is illustrated in Figure 4, while that of the consonants according to their respective counts in the corpus is illustrated in Figure 5.

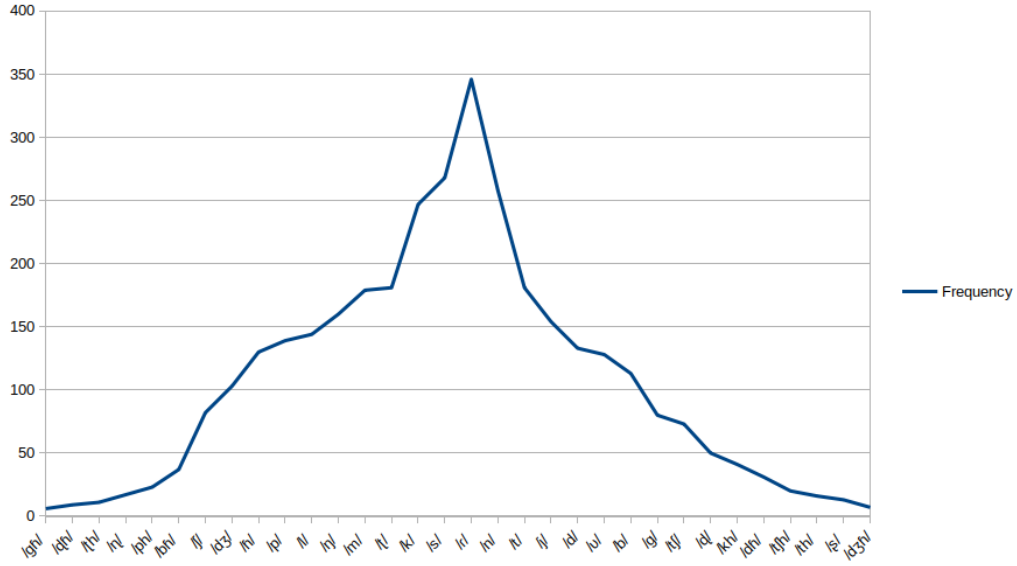

Figure 5: Distribution of Consonants in the Corpus

4. Phonological Study of /h/ in Hindi

There has often been a constant debate and controversy around the status of the phoneme /h/, with the linguists swinging between considering /h/ as a fricative or as an approximant. In (14), Catford states that it's unclear whether to call /h/ as a fricative or an approximant. However, he adds that if we define approximant as a type of articulatory stricture which, at normal conversational flow velocities generates turbulent flow when voiceless but not when voiced, as opposed to fricative, which has turbulent flow with both these phonation types, then these kinds of /h/'s are at times fricative.

In (15), Ladefoged argues that it would be improper to call /h/ as a voiceless glottal approximant, considering the place of articulation is not the glottis at all, and that, at best, it can be described as a voiceless approximant without any place of articulation. Infact, he adds that considering the place of articulation is not fixed, it should not be in the IPA chart, and should be a part of 'Other Symbols' as a voiceless approximant.

In the present section, we try to investigate along the lines of Pandey et al's work in (16) and (17), who address the variant realization of /h/ in Hindi. By selecting words from our corpus which align with the data used by them for the analysis, we performed a phonemic analysis along the lines of the authors for our data.

4.1 Data

In the corpus, there are 130 occurences of the phoneme /h/, of which, 41 are in the word-initial position, 79 are in the word-medial position and 10 are in the word-final position. Of these, we shall take into account the following words for analysis:

1. Word-initial position
   [həm] 'we'
   [həkiːkət] 'reality'
   [həmeːʃaː] 'forever'
   [huaː] 'happened'

2. Word-medial position

   - Without schwa fronting
     [tʃəhiːtaː] 'favourite'
     [nəhiː] 'no'
   - With schwa fronting
     [bəhən] 'sister'
     [ʃəhəriː] 'from the city'

3. Word-final position
   [kəh] 'say'
   [tərəh] 'as, like'
   [rəh] 'stay'
   [aːgrəh] 'insistence'

Schwa gets fronted before and after /h/ segment, but schwa fronting is not applicable before a consonant other than segment /h/. Also, if the segment /h/ is followed by a vowel other than schwa, schwa fronting is not applicable (20).

4.2 Average Formant Values for /h/ in the Data

In Figure 6, we have plotted the average formant values for /h/ in the words from the data containing /h/ in the word-initial position, i.e, [həm], [həkiːkət], [həmeːʃaː], [h uaː] respectively. Each of these words were spoken by male speakers.

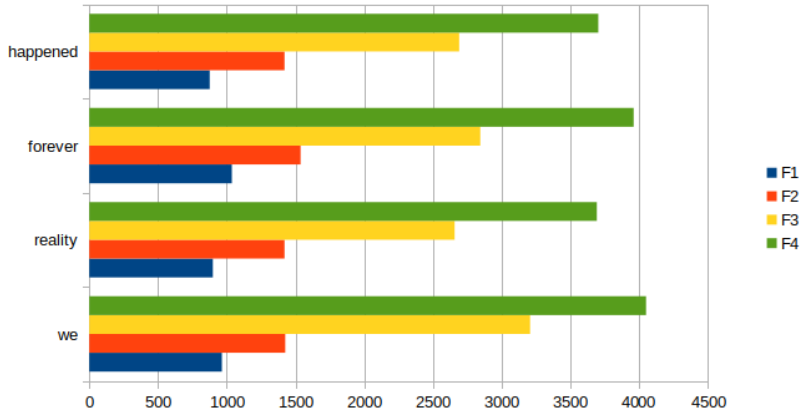

Figure 6: Formant values of /h/ in word-initial data

In Figure 7, we have plotted the average formant values for /h/ in the words from the data containing /h/ in the word-medial position, i.e, [tʃəh raː], [nəh iː], [bəh ən], [ʃəh əriː] respectively. This plot contains words both with and without schwa fronting. The first three words were spoken by a male speaker, while the last word was spoken by a female speaker. This is also evident from the plot, where, the formant values are significantly higher for the word [ʃəh əriː], as compared to the others.

In Figure 8, we have plotted the average formant values for /h/ in the words from the data containing /h/ in the word-final position, i.e, [kəh], [tərəh], [rəh], [aːgrəh] respectively.

4.3 Discussion and Summary

In Hindi, /h/ has two different variants, as a voiceless glottal fricative /h/, and as a voiced glottal approximant [ɦ]. In the word-initially position, /h/ always occurs as a fricative [h], while in the intervocalic positions, it occurs as an approximant [ɦ]. Whenever /h/ occurs in the intervocalic position, Ohala (18) noted that the approximant is inseperable from the following vowel, and that the approximant is merged into the vowel. The authors in (16) also justified the above statement through oscillograms and spectrograms.

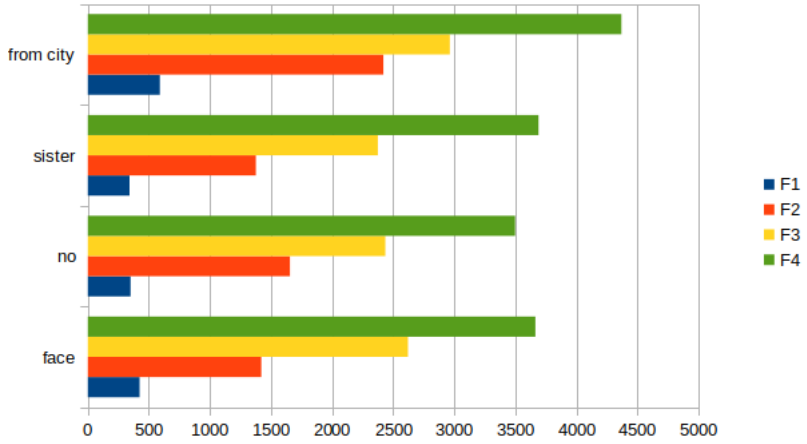

Figure 7: Formant values of /h/ in word-medial data

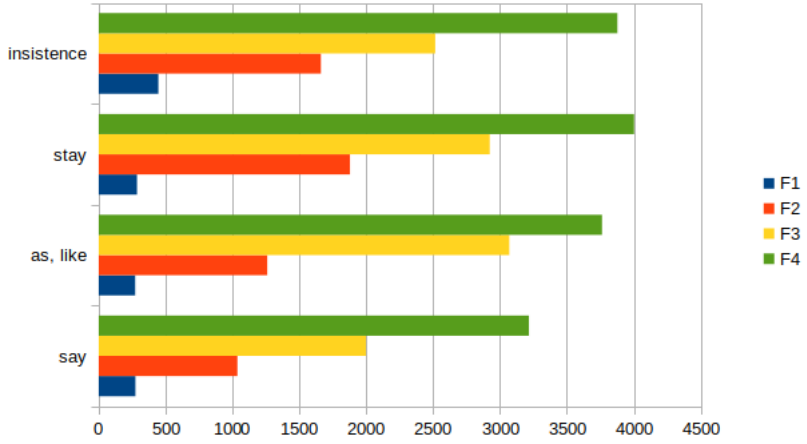

Figure 8: Formant values of /h/ in word-final data

As can also be observed from Figure 6, the average F1 values of /h/ in the word-initial position are in the rang of 950Hz, it is justifiable that fricativization of /h/ occurs. Thus, it can be concluded that in the data instances when /h/ occurs word-initially, it has the properties of a voiceless fricative.

In (16), the authors said that without schwa fronting, the average values of the first and second formants, i.e, F1 and F2 tend to be closer to either the preceding or the following vowel. The average F1 value and F2 value for schwa are 500Hz and 1500Hz respectively. The average F1 and F2 values for the words in our data with /h/ in word-medial position without schwa fronting was 474.3 Hz and 1410.2Hz respectively. Thus, these values justify the hypothesis that without schwa fronting, F1 and F2 values tend to be closer to the formant values of the schwa preceding them. The approximant character of ɦ is attested for the word-final position occurances too, as can be seen from the average formant values in

the formant graph in Figure 8.

Thus, it can be concluded that as /h/ is either preceded or followed by a vowel, it acquires the properties of an approximant, and acquires the voicing property. The position of the phoneme /h/ in the word is also an important property to decide whether it is an approximant or a fricative, as was seen with the /h/-schwa inseparability in the word-medial position (without fronting), thus acquiring the voicing property and the approximant nature. In the word-initial position, /h/ always showed properties of a voiceless fricative.

5. Acoustic Study of Unaspirated Stop Consonants

In Hindi, there are 16 stop consonants, while English has only six. The features used for English may not be useful for Hindi. This encourages to identify distinguishing features to classify the Hindi stop consonants uniquely. From our data, for CVC word structures, we selected Eight initial unaspirated consonants, both voiceless and voiced, /p, t, , k, b, d, , g/ and four final unaspirated voiceless consonants with three vowek sounds /a:, i, u/ in between them. From these 8 x 3 x 4 = 96 combinations, we selected 24 words in our dataset. We selected the words which are spoken by different speakers in similar contexts. Then the word is analyzed in Praat tool for different measurements.

5.1 Voice Onset Time (VOT)

Voice Onset Time is the time of the beginning of vocal cord vibration in CV sequences relative to the timing of the consonant release. In other words, it is the moment at which the vocal cords start to vibrate, measured in reference to the time of release of the plosive. The time difference between release burst of stop consonant and the start of periodic activity (i.e., start of vocal cord vibrations) gives the VOT. As it has been shown previously, VOT fails to distinguish between voiced unaspirated and aspirated stops, we further study the speech Burst.(6)

5.2 Burst Frequency and Duration

A speech burst has the form of an impulse and is produced by the release of the closure in the vocal tract. While measuring the duration of the burst, onset of the burst is marked by fixing the points where pattern shows an abrupt change in the overall spectrum after occlusion. The offset of the burst is noted. In unaspirated stops, the offset of the burst is noted as soon as regular glottal pulsing starts. While in aspirated stops, the burst is separated either by the high frequency noise or by a brief period of silence before the onset of aspiration noise.

The offset of the burst in unaspirated stops is found easily by observing the absence of acoustic energy in the spectrogram. Burst Frequency was measured from the spectra of each consonant. Spectra were obtained by taking the Fast Fourier Transform of the signal to determine the frequencies present. The burst frequency was chosen as the frequency corresponding to the highest amplitude present in the signal spectrum.

5.3 Discussion and Summary

We did the manual measurements of the selected CVC syllables and we noted the acoustic parameters for initial CV syllable, voiceless and voiced, summarized in Figure 9 and Figure 10. The tables shows the approximate average values computed along with their standard deviation. The VOT for voiced stops is negative and large while for voiceless it is positive and small. For unvoiced stop consonants, the average VOTs for /p, t, ʈ, k/ are 15.5ms, 14.7ms, 8.7 ms, and 39.33ms respectively. Thus the average VOT for different place of articulations is around 15 ms, with the exception of 39.33 for velar /k/. For voiced stop consonants, the average VOTs for /b, d, ɖ, g/ are -93.07ms, -135.04 ms, -107.8 ms and -97.4 ms, respectively which shows that VOT is a very important cue for distinction between voiced and unvoiced stop consonants.

From the study of BD and BF, it can be seen that, the frequency range varies with place of articulation. Labial stops has low frequency range with average of 1637Hz, whereas for dental stops it is 4316Hz, for Retroflex dental stops 3087Hz, and for velar stops average is 2978Hz. Hence it can be concluded that labial stops have lower BF. Finally, we could say that, our results closely matches with the experiment done on large scale in (6), with some small deviation in average values.

stop	Following Vowel		VOT (ms)	BD (ms)	BF (H)
/p/	/a:/	Mean	7.1	6.2	840
		S.D.	2.2	3.1	123
	/i/	Mean	13.1	9.2	3100
		S.D.	2.5	2.1	1500
	/u/	Mean	26.2	13.2	1100
		S.D.	5.6	3.6	600
/t/	/a:/	Mean	10.5	8.5	4500
		S.D.	3.1	3.2	1000
	/i/	Mean	13.4	12	4650
		S.D.	8.3	2.6	650
	/u/	Mean	20.2	12.1	3500
		S.D.	4.8	3.5	500
/ʈ/	/a:/	Mean	9.4	5.5	3500
		S.D.	2.5	3.5	200
	/i/	Mean	7.5	4.3	4900
		S.D.	1.3	7.1	250
	/u/	Mean	9.2	8.6	2100
		S.D.	2.1	3.1	350
/k/	/a:/	Mean	31	25	2500
		S.D.	9.1	6.4	900
	/i/	Mean	45	21.2	5600
		S.D.	6.3	9.8	1200
	/u/	Mean	42	30.1	2300
		S.D.	12.3	10.2	900

Figure 9: Average (mean) values with their standard deviations (S.D.) of various acoustic parameters measured for initial VOICELESS stop consonant from CVC syllables

stop	Following Vowel		VOT (ms)	BD (ms)	BF (H)
/b/	/a:/	Mean	-92.5	9	1202
		S.D.	10.1	4.5	200
	/i/	Mean	-101.3	11.1	2235
		S.D.	2.4	6.2	125
	/u/	Mean	-85.3	14.2	1350
		S.D.	21.6	3.6	230
/d/	/a:/	Mean	-134.6	10.2	4500
		S.D.	30.1	4.6	250
	/i/	Mean	-136.3	16.3	3950
		S.D.	36.3	7.2	430
	/u/	Mean	-134.2	15.2	4800
		S.D.	40.2	6.1	500
/ɖ/	/a:/	Mean	-105.3	13.1	3120
		S.D.	4.8	6.3	1000
	/i/	Mean	-108.5	12.1	3450
		S.D.	5.9	3.1	640
	/u/	Mean	-109.5	11	1450
		S.D.	5.3	3.2	530
/g/	/a:/	Mean	-103.5	10.3	2350
		S.D.	9.5	4.5	730
	/i/	Mean	-102.3	13.2	3420
		S.D.	30.1	6.5	600
	/u/	Mean	-86.3	30.1	1700
		S.D.	14.2	10.9	600

Figure 10: Average (mean) values with their standard deviations (S.D.) of various acoustic parameters measured for initial VOICED stop consonant from CVC syllables

In Figure 11, we tracked the voice onset time (VOT) and burst duration (BD) of the voiceless stop consonants, when followed by the vowels /a/, /i/, /u/ in the words in our data. It can be seen that the VOT durations is affected by the vowel that follows the

stop consonant, and the stop consonant followed by the vowel /u/ has the highest VOT. In Figure 12, we tracked the burst frequency (BF) of both the voiced and voiceless stop consonants, when followed by the vowels /a/, /i/, /u/ in the words in our data.

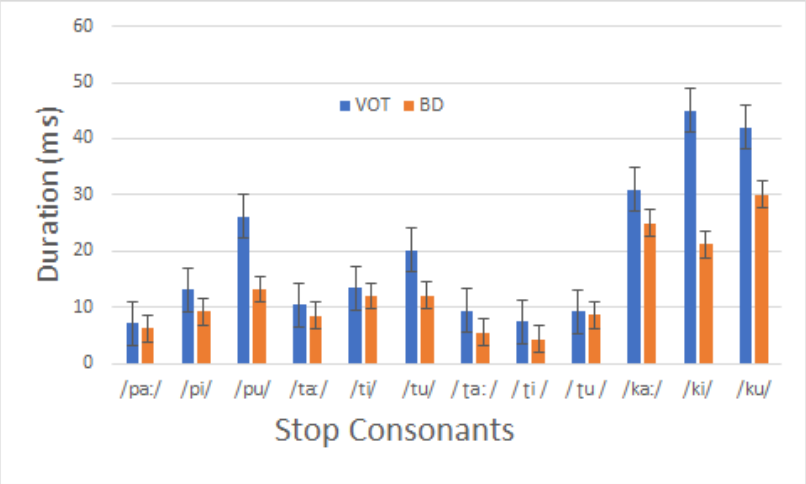

Figure 11: VOT and BD values for voiceless stop consonants when followed by vowels

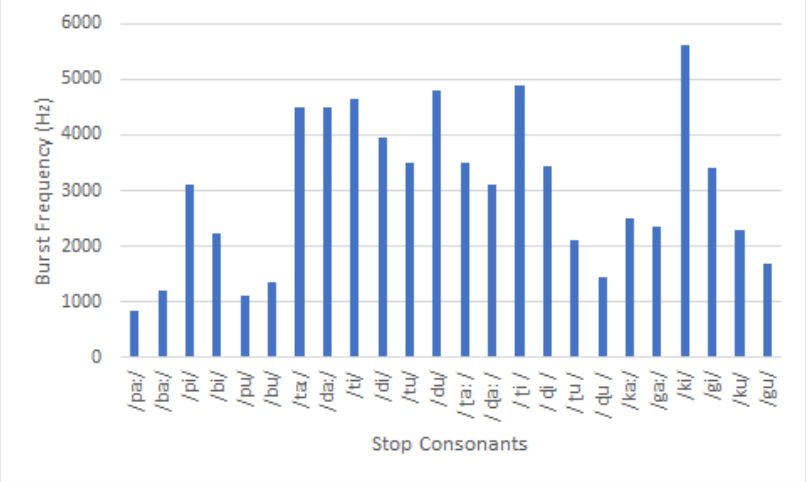

Figure 12: Burst frequencies for unaspirated stop consonants followed by vowels

Thus, from the acoustic study, it can be concluded that the vowel which follows a stop consonant in syllables in Hindi plays a very important role in a more accurate classification of these stop consonants. Even the acoustic attributes like Voice Onset Time (VOT), Burst Frequency (BF) and Burst Duration (BD) are affected by the vowel which follows the stop consonant in the initial position.

6. Acknowledgement

We are grateful to Dr. Keith Langston, who was particularly helpful throughout the course of the semester, also giving constant inputs at various stages of the project.

References

[1] Vishal Chourasia, Samudravijaya K, Manohar Chandwani, *"Phonetically Rich Hindi Sentence Corpus for Creation of Speech Database"*

[2] https://www.github.com/aashishyadavally/Hindi-Speech-Corpus

[3] Karunesh Arora, Sunita Arora, Kapil Verma, S S Agrawal, Automatic Extraction of Phonetically Rich Sentences from Large Text Corpus of Indian Languages INTERSPEECH2004 -ICSLP Jeju,Korea

[4] Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2019. "*Ethnologue: Languages of the World*". Twenty-second edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com

[5] Malviya, S., Mishra, R., Tiwary, U.S. (2016), "*Structural analysis of Hindi phonetics and a method for extraction of phonetically rich sentences from a very large Hindi text corpus*", 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), 188-193.

[6] Sharma, Ravi Prakash, Imran Khan and Osman S Farooq. *"Acoustic Study of Hindi Unaspirated Stop Consonants in Consonant-Vowel (CV) Context."* (2014).

[7] Shyam S Agarwal, "Development of Resources and Techniques for Processing of some Indian Languages", Invited Lecture, LDC, University of Pennsylvania, July 17, 2008

[8] Hank Liao, Erik McDermott, and Andrew Senior, "*Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription*". In 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pages 368373.

[9] Benjamin Lecouteux, Georges Linares, and Stanislas 'Oger, "*Integrating imperfect transcripts into speech recognition systems for building high-quality corpora*", Computer Speech Language, 26(2):6789.

[10] Lakomkin, Egor, et al. "*KT-Speech-Crawler: Automatic Dataset Construction for Speech Recognition from YouTube Videos.*" arXiv preprint arXiv:1903.00216 (2019).

[11] https://ytdl-org.github.io/youtube-dl/

[12] https://www.audacityteam.org/

[13] https://github.com/kaldi-asr/kaldi

[14] Catford, J. (1990). Glottal consonants another view. Journal of the International Phonetic Association, 20(2), 25-26. doi:10.1017/S0025100300004229

[15] Ladefoged, P. (1990). Some proposals concerning glottal consonants. Journal of the International Phonetic Association, 20(2), 24-25. doi:10.1017/S0025100300004217

[16] Pramod Pandey, Mahesh. M, Hemanga Dutta, "*/h/ in Hindi*"

[17] Pandey, Pramod, M. Mahesh, and Hemanga Dutta. "The Glottal Fricative and Schwa Deletion in Hindi: Implications for Speech Synthesis." (2017).

[18] Ohala, M.; Ohala, J.J.: Phonetic universals and Hindi segment duration; in Ohala, Proceedings, International Conference on Spoken Language Processing, Banff 1992, pp 831834, University of Alberta, Edmonton 1992.

[19] R. Prakash Dixit, Peter F. MacNeilage, "Glottal dynamics during Hindi bilabial plosives and the glottal fricative", The Journal of the Acoustical Society of America

[20] Shailendra Mohan, "*A note on Schwa Fronting in Hindi*", Indian Linguistics: 70:223-225