Duration Modeling in Hindi

Somnath Roy
Jawaharlal Nehru University
Centre for Linguistics
Jnu, New Delhi-110067

Nishant Sinha
IMS Engineering College
CSE Department
Ghaziabad UP

ABSTRACT
Duration is one of the important cues for finding the prosodic variation in human speech and other cues are pitch(F_0), amplitude(Intensity) and pauses. Two important concept of duration modeling is in play nowadays, 1- segmental duration modeling, 2- syllable duration modeling. Since speech is a complex continuous signal, hence finding the boundary of segments and syllables is a manual intensive work and prone to error. As per our observation, it is relatively more error prone to find the segment boundary than the syllable boundary. This paper presents the effect of duration both at the level of syllable and segment and its distinctive role in finding the prosodic variation at the sentence level. The findings could be adopted as a model for implementing the prosodic variation in text to speech synthesis(TTS) and automatic speech recognition(ASR). The approach for finding the the model is purely based on the statistical inference derived from the duration values with respect to other cues extracted from the recorded speech data..

General Terms
Prosody, duration, syllable, segment, speech synthesis, automatic speech recognition

Keywords
Duration Modeling, Prosody Hindi, Hindi speech synthesis

1. INTRODUCTION
Duration of an event in general is defined as the interval between the start and end of that event. Duration is the most important cue which makes the sentence sound rhythmic when spoken. It also determines the speed of the sentence being spoken[1]. However duration also convey different types of important information like structure of the sentence, which part of the discourse is most important and loaded with more information[1]. Only phonetic correlation is not sufficient for this study, i.e. phonological, syntactic and semantics [11] also plays crucial role in finding the pattern of loaded information in discourse analysis. Duration model is more or less language dependent phenomena. The approaches to segmental duration modeling can be divided into two categories: rule-based and corpus- based. Klaat[2] proposed a rule-based duration model is based on sequential rule, which is implemented in the MITalk system [5]. In this system, starting from some intrinsic rule, the duration of a segment is modified by rules that are applied sequentially. However, rule-based models[3] try to generalize considering the rule formation process and in this process some exceptions generate which certainly could not be handled using this process without getting exceedingly complicated. When large speech corpora and the computational means for analyzing these corpora became available, new data-driven approaches or machine learning techniques are preferred because in these cases accurate duration finding is important for natural and intelligible speech. Machine learning techniques to predict the duration: Bayesian Network Model, Neural network model, Classification and Regression Model, and Linear Statistical Model. Linguistic analysis about the language like where in the sentence in general the duration is elongated and where it is compressed is also contribute to the analysis.
The rest of the paper is organized as follows: Section 2 is about a brief description about the Hindi language and the linguistic phenomena helpful for duration modeling. Section 3 describes the proposed prosodic modeling[6][12] for TTS using a schematic diagram on the basis of linguistic analysis. Section 4 describes the speech database creation for the analysis of duration modeling in our analysis. Section 5 describes the analysis of cues and statistical inference for the analysis of the modeling. Section 6 we draw conclusion based upon the analysis and result.

2. HINDI LANGUAGE AND LINGUISTIC ANALYSIS
The date of origin of Hindi language is quite debatable and remained a topic of disagreement among the scholars in past. Most of the scholars perfer the 11^{th} century but the earliest date suggested is 9^{th} century. Modern Hindi is supposed to have developed from Old Indo-Aryan(OIA) or Vedic Sanskrit through the growing phases of Classical Sanskrit like the Pali-Prakrits and Apabhransha. The route of origin of the language is not yet fully agreed upon by scholars but the general consesus is that it is a direct descendant of Sauraseni Prakrit via Apabhransha[13]. It is spoken by more than 41% of the total population of Inida (source: 2001 Census of India, www.censusindia.gov.in). The language has 30 consonants, 10 oral vowels with their nasal conuterparts. Two important distinctive features of its phonology uses retroflexion and aspiration in its consonants inventory. Stress is not distinctive in the language and tied to the syllable weight[14]. Most derivational and inflectional morphology in Hindi is affixal; forms of nouns shows agreement with number, gender, and case. The prenominal adjectives agree with the head noun in number, gender, and case[14]. The orthography of Hindi is based on Devnagri Script, which is a derivative of Brahmi script. The later went under continious evolution in its orthographic symbols without affecting the askshara- based character[16]. The akshara is a grapheme consisting of an optional Onset which may be simple or complex and an obligatory nucleus.
Word stress phenomena in Hindi reveals that the stress falls on the alternate syllables if it contains even number of syllables. In general it falls on the syllable having high moraic weight. Other important thing in Hindi is schwa-deletion and schwa-retention issues which also plays an important role in make the language rhythmic . Schwa deletion[10] and retention are described below with the

help of examples.

Table 1. Example of Word stress pattern

Word	IPA	Gloss	D/R	Stress
कमल	'kəməl	Lotus	D	1st syll.
कमला	'kəmla:	a name	D	1st syll.
चमत्कार	tʃə'mət'ka:r	miracle	R	2nd & 3rd
आश्चर्य	a:ʃ'tʃərjə	surprise	R	2nd syll.
संस्थान	sən'stʰa:n	institution	D	2nd syll.

As we see from the above examples, in some cases schwa deletion rules applies and in some cases it retains.

3. PROSODIC MODELING FOR HINDI TTS

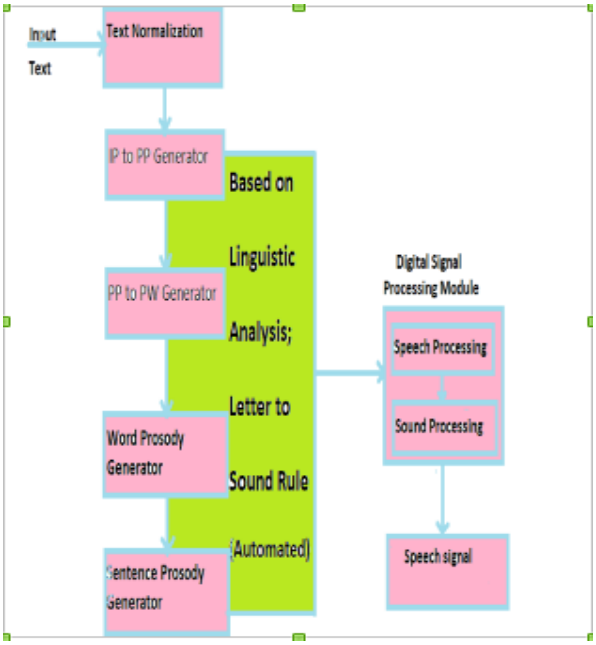


Fig 1. Schematic diagram for prosodic modeling in TTS. (This image is reproduced from [15])

The proposed model is based on the following prosodic hierarchy, shown below in top-down fashion[15]:

Intonational Phrase (IP)

Phonological Phrase (PP)

Phonological Word (PW)

Foot (∑)

Syllable (σ)

Mora(x)

According to the prosodic hierarchy described above, syllables are made up of moras. Syllables combine to make foot and foot combine to make phonological word and so forth. In this process of speech synthesis[7][8][9], we have used top down approach for segmental processes (From IP to PW) and bottom-up approach for prosody generation. A brief description of each module is given below:

3.1 Text Normalization

The very first step of speech synthesis is that the input text should be transformed into string of phonemes. This module converts non-textual tokens into textual form. It has basically two components; Number Converter and Acronym Converter.

3.2 IP to PP Generator

This module will take the input in the form of sentences i.e. sequence of words (textual form), and then it will mark the IP chunks in the input sentence based on some of the mono-syllabic function words which are quite prevalent in Hindi like 'ki:', 'hi', 'per' and 'se:' .This is important because the prosodic features in Hindi are very dependent on the phrase boundary structure. Next from chunked IP sentences, chunking of PP will be done. Chunking labels will be decided later.

3.3 Phonological Word Generrator

This is a module which will take Phonological Phrase as the input, taking white spaces between words as separator for words. But detecting compound words in Hindi is not a simple task. Either one can use some linguistic convention to show that these words are compound words(e.g. using hyphen) or one can attempt to make an algorithm for correct phonetic form determination of compound words for treating those as simple words. But the latter part seems far more difficult because for this algorithm concrete morphological word formation rules are required.

3.4 Word Prosody Generator

This is the main module of prosody generation. This module itself will be categorized into several sub-modules which are: IPA Mapping, Final Schwa Deletion, Syllabification, Syllabic Labelling, Standardized Phonetic Form and Labeled foot structure for Stress.

3.5 Sentence Prosody Generator

The study of prominence and intonation is based on acoustic cues (discrete acoustic values). Total effort is to correlate the statistical model of prominence and intonation with the phonological understanding of prominence and intonation. Based on that analysis, label all the words, PPs and IPs, in order to show the prominence shifting and the intonation pattern realized in the sentences.

4. SPEECH DATABASE

The present study on duration modeling is based on the speech database of size 20 minutes, which consists of around 200 sentences. Speech recorded at the frequency of 16 Khz which is the standard speech frequency being used in the TTS and ASR applications. Also the mono channel were used and the number of bits per channel was 16. The data were recorded in the style of semi-spontaneous speech corpus. One most important and distinctive approach is, the recording was not done in sound proof recording studio, but it has been recorded in a common living room. The whole idea was to take background noise in the analysis so that the model could be useful for ASR system. The recorded data is manually annotated and segmented at

segment and syllable level using praat (a software for speech analysis).

5. ANALYSIS AND STATISTICAL INFERENCE

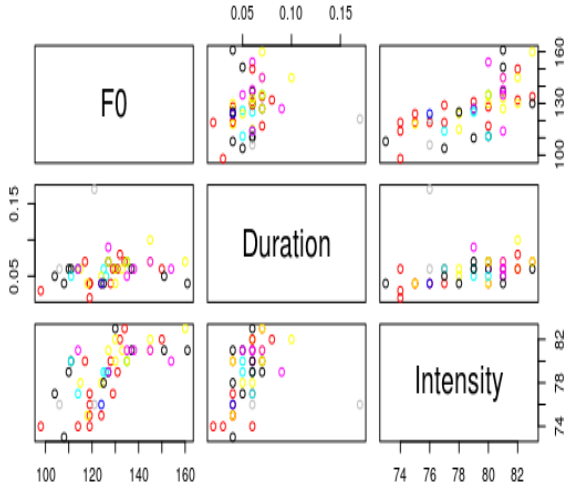


Fig. 2 Analysis of vowel segment /schwa/ using k-mean clustering technique

All sentences are the Hindi declaratives, which generally shows a falling pitch pattern. Since the speech recording is done in semi-spontaneous style and the length of sentences were larger than the usual sentence to cover the contextual influences, hence it is first chunked into smaller portions for manual labeling process. For manual labeling only two text-grids are formed called Vsegment grid and Syll grid. Using manual labeling process, all chunked speech segments are listened carefully and the boundary of vowels and syllables are marked. Using praat scripting, a script is written to automatically extract the values of F_0 , Intensity, and Duration from the text-grid. In the continuous speech, it is nearly impossible to find the exact boundary for either a segment or a syllable. Hence to remove the contextual effect or the adjacent effect (co-articulatory effect in case of segment boundary detection) 12.5% from the beginning and 12.5% from the end is removed to raise the confidence upto 75%.

From the extracted quantitative values of F_0 , Intesity and Duration, a K-mean clustering technique is applied to find the correlation among these objective cues of prosody. We found in the analysis that intensity is a degenerated prosodic factor because same prosodic information can conveyed by varying intensity values by keeping F_0 and duration unchanged in Hindi speech. Hence it should be considered as a secondary prosodic factor. Other two cues F_0 and duration are of utmost importance which induces prosodic effect at supra-segmental levels. Fig.2 shows that there is a strong correlation between F_0 and duration, which can be captured mathematically using large speech database.

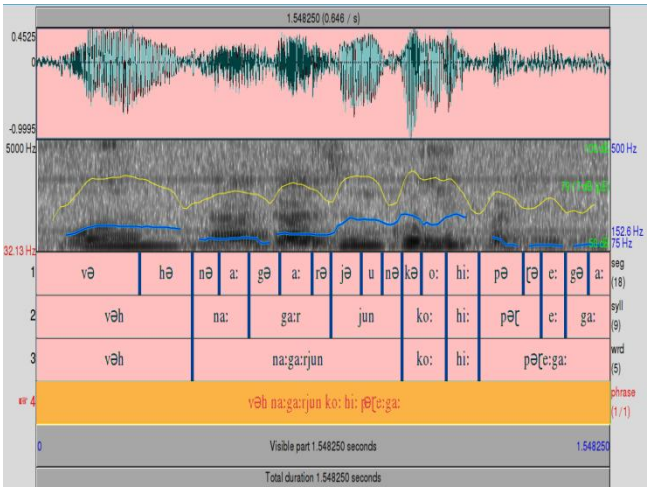
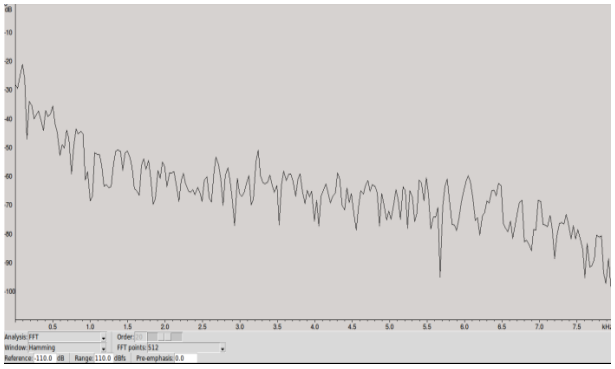


Fig. 3. Praat image of: vah nagarjun ko hi padhega(He will read nagarjuna only).



Amplitude specturm of the waveform shown in image 2 after applying FFT at 512 point scale and then Hamming window(using wavesurfer speech analysis software).

Table. 2. Extracted mean pitch, intensity, and duration of segments of the above wave file shown in image 1

Segment	Seg.pitch (Hz)	Seg.Intensity (dB)	Seg.Duration (sec)
və	165.281	81.131	0.279
hə	166.800	78.789	0.142
nə	123.485	73.796	0.072
a:	125.80	76.602	0.081
gə	119.889	74.978	0.084
a:	141.991	79.810	0.087
rə	138.590	75.633	0.060
jə	184.058	81.870	0.072

u	195.512	82.430	0.057
nə	187.445	73.940	0.050
kə	210.715	84.374	0.047
o:	184.978	82.399	0.067
hə	212.728	81.320	0.042
i:	207.898	72.812	0.046
pə	111.108	74.167	0.118
rə	99.676	69.897	0.040
hə	98.758	72.513	0.066
gə	100.182	67.883	0.062
a:	101.578	70.544	0.059

Table 3. Extracted Mean Pitch, Intensity and Duration for syllables

Syllable	Syll.Pitch (Hz)	Syll.Intensity (dB)	Syll. Duration (sec)
vəh	165.820	81.131	0.305
na:	124.876	78.789	0.152
ga:r	134.645	75.499	0.230
jun	81.126	77.501	0.181
ko:	196.277	190.509	0.120
hi:	211.407	83.378	0.088
per	107.485	78.438	0.167
e:	98.758	72.513	0.066
ga:	100.878	100.878	0.120

Prosodic cues related to segments and Syllables are extracted in Table 2 and Table 3 respectively. Duration values of consonants(shown in Table 2) does not show any pattern during the analysis hence for semental study we chose to find the pattern in vowels. And as shown above in Fig. 2, using K-mean clustering technique it is found that there is a strong correlation between the cue which gives intonational effect(i.e F_0) and duration. But the values in Table 3 clearly shows that the syllable duration is capable of distinguishing the prominent syllables. Prominent syllables have been highlighted. According to [1] the functional formation of prominent syllable finding can be done by the formula shown below.

Syllable prominence = max{ (x+y)*z} –---------(1)

Where x is ratio of pitch in a word in all possible permutation. Similarly y is the ratio of intensity. And z is the value of duration for the syllable in numerator, * denotes the multiplication operation. Our analysis validates this model for syllable duration modeling because the proposed equation (equation 1) gives higher priority to the duration values, because other values are just added but the duration value multiplies with the added ones and is the deciding factor.

6. CONCLUSION

Hindi is a language where lexical stress plays an emphatic role in prosodic changes hence segmental effect at the lexical level need to computed or analyzed. The data extracted for segments shows that pattern can be derived if it is studied in context of another segments in the syllable, word or phrase. Because the same segment in different left-right context shows different durational values as shown above in Table 2 for the vowel segment "a:". From the analysis of vowel segments it shows that the intonational contour at the lexical level changes as soon as change in duration occurs but vice-versa is not true. Hence, it can be said that duration is the most important cues while finding the lexical stress in Hindi. The durational values for syllables clearly shows that prominent syllables have high duration value. The syllable nucleus model can be derived in many ways from the data extracted as shown in Table 2 and is discussed in details in [1]. Further study related to duration model need to be done in a way to find a concrete mathematical formulation suitable to be used in ASR and TTS.

7. REFERENCES

[1] Roy, Somnath (2014). Conference Proceedings of Computational Science and Computational Intelligence, Las Vegas, USA, 10 March-13 March 2014, IEEE. .

[2] Dennis H Klaat, (1979). Synthesis by rule of segmental durations in English sentences, In B. Lindblom and S. Ohman, Editors, Frontiers of Speech Communication Research, pp 287-300, American Press, New York.

[3] Anderson, M., Pierrehumbert, J. & Liberman, M.Y. (1984). Synthesis by rule of English intonation patterns. *IEEE Congress on Acoustics, Speech, and Signal Processing*: pp 77-80.

[4] Ashwin Bellur, K Badri Narayan, Raghava Krishnan K, Hema A Murthy. Prosody Modeling for Syllable-Based Concatenative Speech Synthesis of Hindi and Tamil. *IEEE* 2011.

[5] Jonathan Allen, M. S. Hunnicut, D. H. Klatt (1987). From Text to Speech: The MIT Talk System. Cambridge University Press,Cambridge.

[6] Bagshaw Christopher Paul. (1994). Automatic prosodic analysis for computer aided pronunciation teaching, Ph.d thesis.

[7] Black & Kominek. (2009). Optimizing Segment Label Boundaries for Statistical Speech Synthesis. *IEEE*: pp. 3785-

[8] Black W Alan, Hunt J Andrew. (1996). Unit Selection Synthesis in a Concatenative Speech Synthesis Using a Large Speech Database. IEEE: pp.373-376.

[9] Roy Somnath (2014), A Technical Guide to Concatenative Speech Synthesis for Hindi using Festival. International Journal of Computer Applications, Vol. 86, pp-30-34.

[10] Pandey Pramod, Roy Somnath, D. Kumar , M. Mahesh. Inconsistencies in the Pronunciation of Hindi for a Pronunciation Lexicon, unpublished.

[11] Chomsky & Halle (1968). The sound pattern of English. New York: Harper & Row Publishers.

[12] Cutler, A., Dahan D.& Donselaar. (1997),Prosody in the comprehension of spoken language: A literature review, Language and Speech,141-201.

[13] Singh Rajendra, Agnihotri R. K (1997). Hindi Morphology: A Word based Description. Motilal Banarsidas Publishers, New Delhi.

[14] Kachru Yamuna (1987). Hindi . John Benjamin Publishing Company. London. Vol-12. ISBN: 1382-3485.

[15] Roy Somnath(2013). Statistical Approach to Prosodic Modeling in Speech Synthesis, Phd. Synopsis. Jawaharlal Nehru University, New Delhi. Unpublished.

[16] Pandey Pramod (2007). Orthography-Phonology Interface in Devnagri for Hindi. Written Language & Processing. ISSN:1387-6732. 227-236, 1989.