Prosodic Features of Marathi News Reading Style

Sanket Barhate, Shruti Kshirsagar, Niramay Sanghvi, Kamini Sabu, Preeti Rao,
Department of Electrical Engineering
Indian Institute of Technology, Bombay
Mumbai, India
prao@ee.iitb.ac.in

Nandini Bondale
School of Technology and Computer Science
Tata Institute of Fundamental Research
Mumbai, India
drnandini.bondale@gmail.com

Abstract- **Text-to-speech synthesizers present an attractive alternative to reading in hands-free communication scenarios. Speech intelligibility and naturalness are key to the user acceptability of synthesized speech. The accurate modeling of prosody plays an important role in both dimensions. While prosody is language dependent, it is also strongly dependent on the speaking style. In this work, we study the important prosodic features of news reading style in Marathi using publicly available radio broadcasts. Prominence and boundaries are among the important linguistic cues conveyed via a news reader's prosody. Using perception testing, we obtain boundaries and prominent words in broadcast recordings of two female news readers. We measure acoustic parameters known to serve as cues to prominence such as the fundamental frequency, duration and intensity. We also make observations on timing and pitch phenomena at inter- and intra-sentence breaks. Our results indicate that prominence depends strongly on achieved F0 span in the word and to a smaller extent on duration increase. Breaks are signaled by pauses and pre-boundary lengthening of the final syllable. We observe that, unlike English, sentence ending in Marathi is not always accompanied by a pitch fall in the final syllable. The implications of these observations on prosody generation are discussed.**

Keywords— TTS, prosody, speaking style, Marathi

I. INTRODUCTION

Text-to-Speech (TTS) synthesis forms an important component of human-machine communication where the emphasis is on achieving intelligible and natural sounding speech in a chosen voice. State-of-the-art TTS systems operate by concatenating stored waveform segments corresponding to the phone sequence specified by the text while applying spectral similarity and continuity constraints in prosodic parameters such as fundamental frequency and intensity. While this achieves a natural sounding voice, it tends to be of neutral prosody. The important linguistic functions of phrase and sentence boundaries as well as word prominence must typically be achieved by post-processing. While text processing predicts the boundaries and syllable-level prominences, these functions must be acoustically realized using language-dependent models for modification of F0, duration and intensity. Further, given that different speaking styles have differing salient characteristics represented mainly in prosody [1, 2, 3], it is important that the models match the prosody peculiar to the chosen style in natural speech.

Professional news readers use variations in speaking rate and speech properties for communicating their message effectively and efficiently. Previous studies [1] indicate that such speech tends to be faster, but also accompanied by longer pauses. Usually, topic shifts are accompanied by larger pauses than topic elaborations, and pauses are further reduced at punctuation marks. Besides, topic shifts tend to have their final rhyme to be lengthened the most and the speech rate reduced to the slowest [2]. Additionally it is reported that F0 maxima of segment following boundaries higher in hierarchy are correspondingly greater [3]. Further, causally related sentences had shorter pause and faster articulation rate than non-causally related sentences.

In several languages of the world, one or more of the prosodic parameters, namely, pitch, duration and intensity variation are used to draw a listener's attention to prominent words [14]. For example, studies of Dutch reading style note that prominent words are spoken with high intensity and are found to be longer, and with high median F0 and larger F0 range [4]. The degree of prominence depends on the uttered word type. Most function words are never perceived as prominent. Specific content words like nouns, adjectives and adverbs are always perceived as prominent to some degree. Verbs form a middle class which may or may not be perceived as prominent [5].

Marathi, a language spoken predominantly in the Indian state of Maharashtra with its population of over 100 million, is a relatively poorly studied language as far as the prosody is concerned. However there exist a few studies on the prosody of Hindi [6, 7, 8, 9, 10]. Hindi and Marathi share numerous similarities with regard to the written word as well as pronunciation since they are both derived from Sanskrit, like several other Indo-Aryan languages. In the present work, we present an analysis of the prosody of Marathi news reading style variation to analyze how prosody is achieved in news reading for Marathi language. Our work focuses on durational prosody and prominence. Section II describes the speech material and perception-based annotation. Section III presents the acoustic measurements carried out. Conclusions are discussed in section IV.

II. SPEECH MATERIAL AND PERCEPTION TESTS

Perception tests were carried out to determine boundaries and prominent words, as perceived by listeners, during news reading corresponding to available Marathi news broadcasts.

A. Data Set

Two radio news broadcasts from the archives of All India Radio [15, 16] by two different female news readers were chosen for the present study. They are from Marathi news broadcast on 21 January at 19:00 hrs and 22 January at 13:45 hrs. The first broadcast (Br1) contains 14 paragraphs over 8min 56sec. The second broadcast (Br2) contains 10 paragraphs over 4min 34sec.

For perceived boundary labeling, the first two paragraphs of Br1 were used. This 81 sec recording contains 10 sentences, 182 words and 469 syllables. For acoustic measurements, the entire Br1 and Br2 were considered using the text transcript information of sentence boundaries. For prominence perception labeling, the first three paragraphs of Br2 were considered in addition to the two paragraphs from Br1 comprising a total of 20 sentences with 358 words and 930 syllables in all.

For the acoustic analyses, the corresponding transcripts were manually aligned with the audio at syllable and word level with the help of listening coupled with PRAAT waveform and spectrogram views [14].

B. Perception test

Six native Marathi listeners were involved in the perception experiment. They were given the printed text transcript of the broadcast. The text contained punctuation in the form of full-stops and commas. The listeners were asked to mark perceived boundaries with vertical line and to underline the prominent/emphasized words. Listeners were allowed to make multiple passes over audio while marking transcripts.

1) Boundary Perception

The two paragraph listening dataset had 10 full stops and 7 commas in the text. Boundaries were categorized into topic boundaries (corresponding to topic shift), inter-sentence (indicated by full-stops, but excluding topic boundaries), and intra-sentence (occurring within sentence). With this type of classification, the text had 2 topic boundaries, 8 inter-sentence boundaries and 10 intra-sentence boundaries.

Table 1 reports the results of the perception test where we see that a total of 26 distinct boundaries are perceived by at least one listener. We discarded the last 6 boundaries of Table 1 (less than 50% agreement). The remaining 20 perceived boundaries then included the text-marked 8 inter-sentence and 2 topic boundaries, and 10 intra-sentence boundaries. Thus we found that all the topic and inter-sentence boundaries were correctly perceived by all listeners.

TABLE 1. RESULTS OF BOUNDARY PERCEPTION EXPERIMENT

No. of Listeners Agreed	No of perceived boundaries
6	13
4	5
3	2
2	1
1	5

As for the intra-sentence boundaries, out of 7 commas only 5 were consistently perceived across listeners, while the remaining perceived intra sentence boundaries were not accompanied by any textual punctuation marks.

In summary, full stops are always perceived as boundaries, but this is not the case with commas in the text transcript. Additionally, there are occurrences of perceived boundary without corresponding text punctuation marks.

2) Prominence Perception

Similar to previous work [11], we considered the degree of prominence on a scale corresponding to the number of listener agreements on that particular word. Table 2 shows number of words and their degrees of prominence as perceived by listeners.

TABLE 2. RESULTS OF PROMINENCE PERCEPTION EXPERIMENT

No. of Listener Agreements	Assigned degree of prominence	No of words
0	0	242
1 or 2	1	77
3 or 4	2	26
5 or 6	3	13

It can be seen that 32% of the total of 358 words are judged prominent by at least one listener. Considering the words with degree of prominence 2 and 3 in above table, prominence is perceived largely for adjectives (44%) followed by proper nouns (28%) followed by numbers (13%).

III. ACOUSTIC MEASUREMENT

To make acoustic measurements relating to boundaries, we syllabicated whole two paragraphs along with silence region marking.

A. Boundaries

The prosodic cues known to be linked to boundaries are pauses, pre-boundary lengthening and pitch contour slope on the final syllable [12]. To investigate these cues for Marathi news reading style, we carried out the measurements presented below.

1) Pauses

We were more concerned with duration cues relating to boundaries. It was found that out of 20 perceived boundaries, 16 were realized by pauses, where any silence region with

duration more than 20 ms is deemed as a 'pause'. In all, 80% of the boundaries are realized by pause. For all boundary types, mean and standard deviation (SD) of duration is given in Table 3. For mean and SD of topic boundaries, we considered whole transcript of 14 paragraphs as there were just 2 topic boundaries in the data set considered otherwise.

TABLE 3. MEAN AND SD FOR DURATION OF DIFFERENT BOUNDARY TYPES

Boundary Type	No of Observations	Mean Duration	Standard Deviation
Topic	14	1.39737	0.4492
Inter-Sentence	8	0.505	0.055
Intra-Sentence	6	0.42337	0.08

It can be seen that as we go higher in hierarchy i.e. from intra-sentence boundary to topic boundary, the mean duration of pause realized tends to increase, which is consistent with observations on English [2]. Moreover silence to speech ratio for first paragraph was found to be 8.3% while for second it turned out to be 10.4% which is consistent with figures for English news reading [1].

2) Pre-boundary Lengthening

To verify the hypothesis relating to lengthening of ultimate syllable relative to the penultimate syllable, we require instances of the word to occur both within sentence and at the end of sentence. In Marathi, verbs tend to occur at sentence boundaries as Marathi has Subject-Object-Verb structure. Intra-sentence occurrence of verbs is very low. We considered first broadcast with 14 paragraph transcript for this task. In this transcript we found that the verb /*ahe*/, meaning 'is', meets our requirement. There were 5 intra-sentence occurrences and 14 end-of-sentence occurrences for the verb /*ahe*/. We first syllabicated all the occurrences of /*ahe*/. The boundary of the end vowel was marked when formants died out as depicted in the spectrogram in Fig. 1.

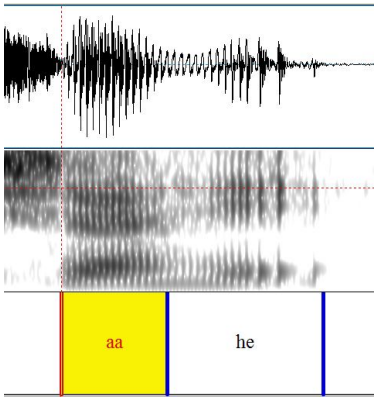

Fig. 1. PRAAT view for syllable duration of /*a*/(aa) and /*he*/

Fig. 2 shows the boxplot for duration of word /*ahe*/ for all the positions in sentences along with the syllable duration for /*a*/ *and* /*he*/ for all their occurrences.

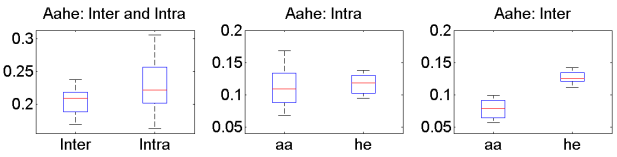

Fig. 2. Duration boxplots for /*ahe*/ at word-level (left) and syllable level (middle and right) for inter- and intra-sentence locations.

We investigated word lengthening first, and found as shown in Fig 2 that there is no significant difference between intra- and inter sentences occurrences of the word. However, measurements at the syllable level clearly capture the pre-boundary effects in terms of relative durations of the ultimate and penultimate syllables. Fig. 2 indicates that the separation between syllable durations of /*a*/ and /*he*/ is much higher when it occurs at sentence boundary than when it occurs within sentence. As a caution, at this point, we don't make any claim regarding lengthening of feet (Portion between two lexically stressed vowels) as mentioned in [12] where vowel nucleus before boundary gets lengthened.

From the above observations we conclude that in Marathi news reading style, pre-boundary lengthening is realized by changing relative durations of penultimate and ultimate syllable. This differs from English where it is realized by lengthening of vowel nucleus just before the boundary [12].

3) Pitch contour slope

In general, declarative sentences across languages are accompanied by falling pitch contours. In the Marathi news broadcast, we observed that some sentences showed their final syllables marked with rising pitch. To examine the above phenomena more closely, we used the second speaker's transcript which contained more sentences ending with pitch rise. This transcript had 32 sentences out of which 7 were accompanied by pitch rise at the end of the sentence, amounting to 22 % of the total inter-sentence boundaries. A closer examination of the associated syntax and semantics indicated that the news reader used pitch rises to convey the relatedness of the subsequent sentence with the previous one. The two sentences are related either during topic elaboration or continuation. We observed total 11 instances where two consecutive sentences were semantically related of which 64% exhibited pitch rise. Pitch rise was observed to be associated with certain key words, which observation needs to be further verified with more data. Semantically independent sentences were always accompanied by pitch fall.

B. Prominence

To study acoustic cues to prominence we measured–the following acoustical features: (1) F0 max per word in semitones with respect to paragraph mean, (2) maximum intensity per word normalized with paragraph mean intensity, and (3) log duration Z score per word (computed by adding the syllable-level Z scores) in seconds [13] using PRAAT scripts. We compare the acoustical measures with the

prominence judgement of the native Marathi listeners via the box plots in Fig. 3. We see that duration is relatively unaffected whereas pitch and intensity increase with increasing degree of prominence.

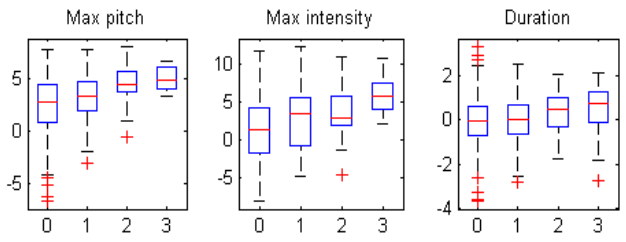

Fig 3: Boxplot for Acoustic parameters and perceived prominence of native Marathi listeners

Table 4 gives the Spearman's correlation coefficients between the degree of prominence and the measurements.

TABLE 4. SPEARMAN CORRELATION FOR DIFFERENT PROMINENCE CUES

Parameters	Spearman coefficient
F0 max	0.70
Intensity max	0.70
duration	0.56

The F0 maximum per word and intensity maximum per word show the highest correlation with perceived prominence. However In case of the normalized duration per syllable the relation towards prominence is not that strong as judged by native listeners. The same result was observed even when we analyzed the recordings of both speakers individually.

IV. DISCUSSION AND CONCLUSION

We studied the acoustic correlates of perceived boundaries and word prominence in Marathi news broadcast audio recordings. Perception tests showed that all sentence boundaries in the text transcripts were clearly perceived as boundaries by Marathi native listeners. The perceived boundaries were realised by pauses and pre-boundary lengthening of penultimate syllable with topic boundaries showing the longest pauses. While sentence endings are expected to have decreasing F0 contours, it was found that this is not universal as far as Marathi news reading style goes. Our broadcast audio dataset indicates that inter-sentence dependencies may actually lead to pitch rise at the boundary. The degree of perceived prominence of a word in Marathi news broadcast audio was found to correlate well with acoustic measurements of word level parameters, namely, maximum F0 and maximum intensity. Duration does not seem to play a significant role. This is contrary to previous findings on elicited utterances in Marathi by casual native speakers where word durations were clearly increased in focus and decreased post-focally [13]. We speculate that this difference arises because news readers tend to be time conscious and prefer to rely more on cues that do not entail time lengthening as far as possible. F0 and intensity models for prominence can be derived from these findings and applied to new text transcripts of Marathi news where text processing methods are used to determine prominent words from semantics and syntax. Similarly, acoustic properties of text-predicted boundaries can be suitably prescribed based on the results of our study. Finally TD-PSOLA methods as in [17] can be used to modify the speech prosody of the synthetic utterances, while taking care not to alter spectral quality as far as possible by choosing units that are acoustically close to the desired prosody in the concatenative synthesis, in order to achieve high quality natural sounding read news from text.

REFERENCES

[1] E. Strangert, "Prosody in public speech: analyses of news announcement and political interview," Proc. of Interspeech, 2005.

[2] C. L. Smith, "Topic transition and durational prosody in reading aloud," Speech Communication, 2003.

[3] L. N. J. T. Hanny den Ouden, "Prosodic realisation of global and local structure and rhetorical relations in read aloud news report," Speech Communication, 2008.

[4] B. Streefkerk, L.Poles and L. Bosch, "Acoustical Features as Predictors for Prominence in Read Aloud Dutch Sentences used in ANN's", In Eurospeech, 1999

[5] B. Streefkerk, L.Poles and L. Bosch, "Towards Finding Optimal Features of Perceived Prominence", In Proc. 14th ICPh.99

[6] Genzel, S. and Kügler, F.,"The prosodic expression of contrast in Hindi", In: Proc. 5th International Conference of Speech Prosody, Chicago, USA, 2010.

[7] Puri, V., "Intonation in Indian English and Hindi late and simultaneous bilinguals", Ph.D. thesis, University of Illinois at Urbana-Champaign, 2013.

[8] Féry, C.,"Indian Languages as Intonational 'Phrase Languages', In: *Problematizing language studies*, 2010, pp. 288–312.

[9] Harnsberger, J. D., "Towards an intonational phonology of Hindi", In Proc. Fifth Conference on Laboratory Phonology, Northwestern University, 1996.

[10] Patil, U., Kentner, G., Gollrad, A., Kügler, F., Féry, C., and Vasishth, S., "Focus, word order and intonation in Hindi," Journal of South Asian Linguistics, vol. 1, no. 1, pp. 53–67, 2008.

[11] Streefkerk, B. M., Pols, L. C. W. and Ten Bosch, L. F. M. 1998. Automatic detection of Prominence (as defined by listeners' judgments) in read aloud Dutch sentences. *ICSLP-98*, Sydney, Vol. 3, 683-687.

[12] Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). "Segmental durations in the vicinity of prosodic phrase boundaries," *The Journal of the Acoustical Society of America*, *91*(3), 1707-1717.

[13] P. Rao, Mixdorff, Deshpande, N. Sanghvi, S. Kshirsagar. "A Quantitative Study of Focus Shift in Marathi", Proc. of Speech Prosody 2016, Boston, 2016.

[14] Boersma, P., "PRAAT, a system for doing phonetics by computer", Glot International 5, pp. 341-345, 2001

[15] PRASARBHARTI:http://www.newsonair.com/Regional-Audio-Bulletins-News-schedule.asp

[16] PRASARBHARTI:–http://newsonair.com/Regional-Bulletins-Script-Schedule.asp

[17] Moulines, E., F. Emerard, D. Larreur, J. L. Milon, L. Le Faucheur, F. Marty, F. Charpentier, and C. Sorin. "A real-time French text-to-speech system generating high-quality synthetic speech." In Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on, pp. 309-312. IEEE, 1990.