Comparing acoustic cues to phrase boundary in Hindi and Indian English

Mildred Pereira[1], *Preeti Rao*[1], *Hansjörg Mixdorff*[2]

[1]Department of Electrical Engineering, Indian Institute of Technology Bombay, India
[2]Berliner Hochschule für Technik, Berlin, Germany

mildredp@iitb.ac.in, prao@ee.iitb.ac.in, Hansjoerg.Mixdorff@bht-berlin.de

Abstract

Prosody is key to spoken language communication, integral to conveying meaning, emotion and emphasis beyond the literal speech. While several languages share the linguistic elements of phrase boundary and prominence, their acoustic realizations can vary considerably. With the goal to investigate whether languages can be discriminated based purely on prosody, we carry out a perception experiment with delexicalised speech of Hindi and Indian English model readers. With Hindi surpassing English, the overall language recognition accuracy was 63.5%, with highest individual listener accuracy of 79.0%. Acoustic cues at boundaries were examined across languages via word level features related to f0, intensity, duration and pause. Intensity features were crucial indicators of different boundary types in Hindi while duration based cues dominated automatic boundary classification in English.

Index Terms: prosody, delexicalisation, acoustic features

1. Introduction

Through spoken language, humans communicate messages, ideas and emotions. Prosody, of interest in this work, deals with the variations in pitch, intensity and duration which makes speech or read aloud text more comprehensible to the listener. For example, the chunking of words into higher level meaningful units such as phrases and sentences is achieved via prosodic boundaries, a feature of nearly every language, with the precise acoustic cues to their spoken realization differing across languages. For second-language learners, the carry over from L1 that tends to remain relatively less affected by L2 training is the speaker's prosody, thus contributing to the percept of non-native accent.

The properties of text play a vital role in phrasing and prominence. In the English language, phrase boundaries are more likely to follow content words than function words. Also, the nouns, adjectives, verbs, adverbs are made more prominent than articles, pronouns, conjunctions and determiners [1]. Indian English shares the prosodic elements of UK English while differing in the segmental aspects such as certain pronunciations. The acoustic cues like pitch, intensity and duration serve an essential purpose in indicating prosodic phrasing and emphasis. Longer durations of certain syllables or words, and/or variations in pitch signal boundaries between phrases. A rise in pitch often marks the end of a question. Hindi prosody is less researched. Linguistic structural differences in Hindi include the extensive use of case markers, verb final sentences and the presence of auxiliary verbs. A study of listener perception with Hindi narrative speech revealed that native speakers show considerable agreement in the placement of prosodic boundaries while the notion of prominence is weak (but for the case of trained expert transcribers) [2].

In this paper, we address the following research questions:

- Are the listeners able to distinguish between the utterances of Hindi and Indian English when only prosodic cues are available ?
- Which acoustic cues for boundary are distinctive in each of the two languages ?

In this work, we target a comparative study of language dependent acoustic cues to prosodic events. We start with verifying whether prosody alone (as represented in de-lexicalised speech) enables bilingual listeners to discriminate between Hindi and Indian English read aloud speech, each by a fluent reader of the language. The same utterances are labeled for perceived prosodic boundaries and utilised in the acoustic study for pitch, duration, pause and loudness based cues.

2. Dataset and Annotation

We used Hindi and English texts drawn from Grade 3 level narrative stories selected by language teachers[1]. Each of the 6 stories comprised of 2 paragraphs each between 55 to 75 words (4 or 5 sentences). Further, we generated the corresponding translation to the opposite language for each story in order to ensure a comparable overall number of words across the languages, with an automatic translation tool [3]. Translations of original English text to Hindi increased the number of words by 11.01% while Hindi to English reduced by 6.97%. This is consistent with previous studies[2][4]. This bidirectional translation ensured that the semantic content matched across the two language sets. While sentence-level matching is ensured, syntactic equivalence across Hindi and English is not possible due to inherent differences in word order (Hindi being an SOV language while English being SVO). Each language had 24 paras(12 stories x 2 paras) read by each of the 5(2M,3F) model readers. We define a model reader as an experienced language teacher teaching Hindi/English in English medium schools. They were instructed to go over the text precisely once before recording it on the WhatsApp voice recorder. This was done to ensure the readings appeared natural and unpracticed. The recordings were screened for background noise and re-recorded as needed. With each model reader reading 24 paragraphs, the total number of recorded utterances was 240. The total number of words were 15,339 (English : 7295, Hindi : 8044).

Three annotators who were fluent speakers of English and Hindi assigned one of three boundary types to each word in the spoken utterances. The boundary types were intermediate phrase boundaries (intP, prosodic boundaries within a sentence), Intonational Phrase boundaries (IP, sentence endings) and 'nb' for words not perceived as phrase final. This was done on LMEDS interface [5]. The text shown to the annotators for marking was without punctuations and capitalization to avoid

[1] For more dataset and other details : https://tinyurl.com/y2vt7w5b
[2] text expansion/contraction : https://tinyurl.com/dm2uzhdh

bias. The correlation between annotators for marking intP and IP boundaries for English was 0.81 and 0.96. For Hindi it was 0.81 and 0.95 respectively. Based on majority vote (2 of 3 annotators), English and Hindi had 745 and 754 IP boundaries, 1252 and 1078 intPs, and 5298 and 6212 'nb's, respectively. A comparison with text punctuation revealed that all readers realized IP boundaries (full stops); intP boundaries were commonly aligned with commas, conjunctions, meaningful phrase ends, or the penultimate word to highlight the final word.

Properties of text play crucial role in realizing boundaries. By analyzing PoS distributions at intP and IP boundaries, we can better understand how syntax and prosody interact, and whether different types of boundaries attract different kinds of words. For text analysis, part-of-speech (PoS) tags were generated for each word in all English and Hindi stories using the Stanza natural language processing toolkit[6] developed by Stanford. UPOS (Universal PoS tags) were used for both languages to maintain consistency. NOUNs formed the highest percentage of words in both languages. For English the PoS tags after which most intP boundaries were realised, were in the order of NOUN, VERB and ADV(adverb) while for Hindi it was ADP(adposition), NOUN and VERB. IP boundaries in English mainly occurred after NOUNs while in Hindi after auxiliary verbs(AUX).

A subset of the 240 utterances in this dataset was used to examine how listeners discriminate between Hindi and English when lexical information is obscured. The following section reviews relevant work on perceptual cues to language discrimination.

3. Language identification with delexicalised speech

Previous studies across languages have shown that listeners can identify a language even when speech is delexicalised(lexical content is made unintelligible but the prosodic cues like pitch, rhythm, intensity are retained). Strong correlations were seen between trained raters' score when listening to sound clips in filtered (500 Hz cutoff) and unfiltered condition[7]. Also, the raters could clearly distinguish between Mandarin and English prosody even when English prosody was produced by novice L2 learners whose first language was Mandarin.

In [8], from spontaneous Swedish speech utterances, 2 sec and one word stimuli before boundaries, filtered at 400 Hz were presented to Taiwanese and English speakers with no prior knowledge of Swedish. Results revealed that native as well as non-native speakers could predict the upcoming boundary strength in both filtered and unfiltered speech. Speech rate, f0 and HNR were the major cues to language identification in filtered speech. In [9], the authors conducted human language identification experiment with reduced segmental information in Japanese and American English. Results suggested that humans can identify languages with reduced segmental information. The results also showed variation due to the differing typologies of languages and subjects' linguistic knowledge.

We performed an experiment to check for potential differences between the perceived prosody of English and Hindi in adult readers' utterances by delexicalising the speech.

3.1. Experiment and Results

Sentence level utterance(stimuli) were clipped(from paragraph level utterances). Each stimulus was thus 8-15 s in duration. These were low pass filtered at 0-300 Hz with 100 Hz transition. 300 Hz cut-off was used since informal testing had shown that some Hindi male speech was not sufficiently delexicalized at 450 Hz. The total number of stimuli was 100 (50 per language, equal number of male and female stimuli). The experiment was setup in PsycoPy [10]. 20 bilingual listeners (10F and 10M, who used both English and Hindi in oral and written form) were engaged for the task of language identification. All the listeners had studied Hindi as a language in school where the primary medium of instruction was English. They were provided with headphones and volume was checked. At the start of the experiment listeners were presented with sample Hindi and English stimuli. They also received the following instructions: The audio stimuli presented would be either in English or Hindi. Each stimulus would be played only once, and listeners should label each sample as 'English,' 'Hindi,' or 'cannot make out' based on the language perceived. Any stimulus labeled as 'cannot make out' would be replayed once. They were further asked to write down any words they may have discerned. Any stimulus with any correctly identified words by any listener was discarded from analysis. General listener comments, solicited after the experiment, suggested that Hindi could be identified from the characteristic sentence ending and the overall speech flow. Also, intonation was a distinguishing factor. They were given a small honorarium for participation.

Figure 1 shows language-wise recognition accuracy across 20 listeners, with Hindi exhibiting a slightly higher median and English greater variability. The % 'cannot make out' is more for English as compared to Hindi. It could be that some acoustic or phonetic cues in Hindi are more resistant to suppression. Highest overall accuracy for individual listener is 79.0%. 536 out of 876 English stimuli were correctly identified for the language while for Hindi it was 470 out of 703. The overall accuracy for all listeners is 63.5%. Fluency and daily language usage inputs as taken from the listeners didn't have a direct correlation with the individual listener recognition accuracy.

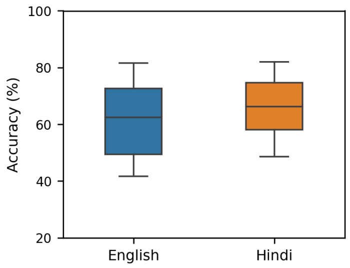


Figure 1: *Language recognition accuracy across 20 listeners*

3.1.1. Statistical Analysis

Hypothesis 1: Listeners' language identification accuracy exceeded chance

We conducted a one-sample t-test to assess whether listeners' accuracy exceeded chance (50%). Accuracy was significantly higher than chance, t(19) = 7.304, $p < 0.001$, with listeners achieving a mean accuracy of 63.5% (SD = 8.24%)

Hypothesis 2: The language identification accuracy is different for Hindi and English stimuli

We compared listeners' accuracy for Hindi and English stimuli to test whether one language was identified more accurately than the other. Accuracy was similar for Hindi (M

= 66.22, SD = 10.96) and English (M = 61.34, SD = 13.59). A paired-samples t-test indicated no significant difference between English and Hindi stimuli (p = 0.271)

4. Acoustic features

As noted in the last section, listeners were able to identify the language of delexicalised stimuli above chance overall. This suggests that the stimuli preserved prosodic information that could cue language identification. To further study these potentially language-dependent cues, we present an acoustic-prosodic analysis of the original speech, examining f0, intensity, duration and pause based features across the two languages.

For acoustic analysis, we take into account prosodic boundaries which indicate the end of a prosodic unit like phrase, clause or sentence. These were marked by 3 annotators for our dataset as explained in Section 2. Prosodic boundaries are known to be realised with pitch reset and pre-boundary final syllable lengthening in different languages [11]. Hence, compared to non-boundary words, boundary words with their pronounced acoustic cues could potentially aid in distinguishing languages. Phrasing in English and Hindi differ because their phonemes, stress, intonation, syllable structure are not the same. Existing literature highlights studies in English that examine how different acoustic features change at boundaries[12], but a similar study in Hindi is relatively unexplored. Additionally, a comparative study of acoustic feature changes at boundaries in Hindi and English is lacking. This work attempts to fill in the gap.

All the utterances in the dataset underwent forced alignment with the corresponding text prompts yielding word and phone durations(ctm files), which were further used along with a manually generated syllable lexicon to obtain the syllable durations. In rare cases, manual corrections for durations were needed. Syllable level acoustic feature extraction was done since they serve as the basic building blocks of speech production. f0 (acoustic correlate of pitch) and intensity, extracted every 10ms using Parselmouth[13], were aggregated at the syllable level using aggregates like mean, median, max. and min. For the i^{th}(last) syllable of any word, cross boundary feature ($F_c(i)$) was calculated as in Equation 1. *f(i)* and *f(i+1)* are the syllable level aggregates for i^{th} and $(i+1)^{\text{th}}$ syllable. For pre-boundary feature (Equation 2), the last(i^{th}) syllable of the word and the just previous $((i-1)^{\text{th}}$ syllable are considered.

$$F_c(i) = f(i+1) - f(i) \qquad (1)$$

$$F_p(i) = f(i) - f(i-1) \qquad (2)$$

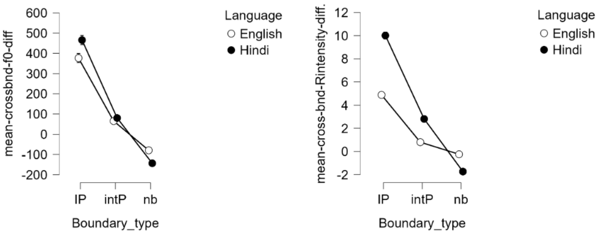
Duration was obtained from the forced alignment using available automatic speech recognition models for each language. The features used in this work are listed in Table 1

4.1. f0 features

The f0 in cents (with respect to the average f0 of the utterance) was taken into consideration to reduce the variability caused by the differences in individual speaker's pitch. Figure 2a shows that f0 differences at intP boundary for both languages are comparable but IP boundary ones are slightly apart; further the 'no boundary' ones don't show much difference(the dots represent the mean value while the bars show the std error). Manual Praat observations show that in English, f0 rises before intP boundary and falls before IP boundary thus confirming observations in [14]. Further, the syllable preceding an intP shows a rising f0 contour compared to the post boundary syllable, the post+1 may or maynot have high f0(it will have rising f0 contour if the word is made prominent). The post IP boundary words like (but, now, and, some, these) show rising f0 contours. Similar to English, f0 rises before intP boundary and falls before IP boundary for Hindi. f0 contours in Hindi shows a rising trend for after boundary words which fall in the conjuction part of speech(PoS) category. Other words after boundary which shows rising pitch contours were from NUM(Numerals) and DET(Determiners) category.

Table 1: *Feature set used in each feature category*

Feature category	**Features**
f0 (8)	cross boundary f0 difference(mean, median, min, max); pre-boundary f0 difference (mean, median, min, max)
Relative intensity (8)	cross boundary relative intensity difference(mean, median, min, max); pre-boundary relative intensity difference(mean, median, min, max)
Duration (3)	Duration difference(final-penultimate), duration ratio(final/penultimate), duration ratio (post boundary/final)
Pause(2)	Pause duration, Normalized pause duration



(a) *Cross boundary f0 difference*

(b) *Cross boundary relative intensity difference*

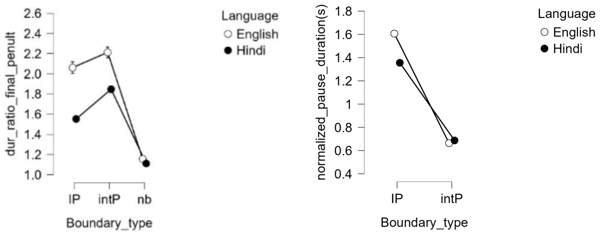
Figure 2: *Cross boundary mean feature difference at Intonational (IP), Intermediate (intP), and no boundaries (nb) for English and Hindi*

4.2. Relative intensity features

Calculated at every 10 ms frames, the absolute intensity is converted to relative by subtracting the average intensity(in dB) of the entire utterance from it. Further, these values are aggregated at syllable level. Praat observations of Hindi utterances reveal that intensity contour drops significantly at the last syllable before a boundary, more prominently for IP boundaries than intP boundaries. Comparing this with English utterances, the intensity drop at IP(sentence) end syllables is more pronounced for Hindi. This could be attributed to the fact that the IP end words in Hindi are usually the auxiliary verbs(AUX) and these are mostly spoken with reduced vocal effort. The default word order in Hindi is SOV(Subject Object Verb) while that for English is SVO and hence the verbs always appear at the sentence end positions. Figure 2b confirms the manual Praat observations since the IP boundary plots for both languages are quite apart and the error bars donot overlap.

4.3. Duration features

Duration is an important cue which signals boundary. Pre-boundary lengthening(increased duration of syllable prior to



(a) *Duration ratio of final and penultimate syllables*

(b) *Normalized pause duration*

Figure 3: *Duration and pause related features at different boundary types for English and Hindi*

boundary) is a proven phenomenon for English. In [15], the authors have analyzed syllable duration before 5 different types of boundaries in German read aloud prose text and inferred that syllable duration is maximum for intP end syllable as compared to IP one. Vandana Puri's PhD research [16] examined preboundary lengthening in Hindi by analyzing the durations of penultimate and final syllables at phrase and sentence boundaries. Using short sentences containing two disyllabic words with either final or penultimate stress, read by 30 speakers (150 utterances), she found that final syllables lengthen more than penultimate syllables at boundaries, and both syllable types are longer in sentence-final than in sentence-medial positions. In our dataset too, we investigate preboundary lengthening in the two languages both at intP and IP boundaries. The syllable immediately before a boundary was referred to as the final syllable, the one preceding final was the penultimate, and the one following final was the post-boundary syllable. Features evaluated were difference and ratio of speaker normalized durations of final and penultimate syllable; duration ratio of post boundary to final syllable. For Hindi language, the IP boundaries show a low duration ratio(Figure 3a), this could be because the last syllables are usually the auxiliary verbs which are usually spoken quickly and are not elongated. The duration ratio for no boundary words is very close to 1 suggesting almost equal syllable durations. Our results confirm the observations from [15] with intP boundaries having maximum syllable lengthening.

4.4. Pause features

Two pause features (pause duration and normalized pause duration) were measured with pauses below 100 ms ignored. The normalized pause duration feature was computed by dividing each pause duration by the mean pause duration across the utterance. The nb tokens were not included for pause features due to the threshold of 100 ms. The two languages showed comparable pause feature values at intP boundaries, but significantly higher values for English at IP boundaries (Figure 3b). This indicates that English readers discriminated between boundary types to a greater extent with pause.

4.5. Statistical analysis and classification experiment

MANOVAs (Multivariate ANOVAs) were carried out to evaluate whether boundary type and language influenced the acoustic realization of prosodic cues. For f0, intensity and duration-related features, boundary type showed a strong multivariate effect (Wilks' $\Lambda = 0.506$, $F = 321.02$, $p < .001$), denoting clear acoustic differentiation across 3 boundary categories. A smaller but significant effect of language (Wilks' $\Lambda = 0.936$, $F = 54.09$, $p < .001$) and a significant boundary type*language interaction (Wilks' $\Lambda = 0.909$, $F = 38.53$, $p < .001$) further suggest cross-linguistic differences and language-specific patterns of boundary marking. Pause features were omitted from this last analysis, being unavailable for nb words.

A Random Forest classifier was trained separately on English and Hindi datasets to predict boundary types within each language.Using f0, intensity and duration based features, a 10-fold cross validation showed strong classification performance. Hindi performed marginally better than English. Hindi achieved accuracy of 0.887, with 0.877 F1 and 0.938 AUC; for English it was 0.819, 0.803 and 0.900 respectively. Permutation-based feature importance from the Random Forest classifier showed that for English, duration features (duration diff(final-penult) = 0.111; duration ratio(final/penult) = 0.077) were the strongest predictors of boundary type, alongwith min. crossbnd. relative intensity difference = 0.078 followed by f0 related ones. For Hindi, 3 relative intensity related features were in the top 5 suggesting that the language prominently uses intensity cues for signaling different boundary types (Table 2).

Table 2: *Top 5 best ranked features for boundary classification task in English and Hindi(Random Forest classifier using Permutation Importance(PI) based feature ranking)*

English		**Hindi**	
Feature	**PI**	**Feature**	**PI**
duration diff.(final - penult)	0.111	max. crossbnd. relative intensity diff.	0.170
min. crossbnd relative intensity diff.	0.078	mean. crossbnd relative intensity diff.	0.098
duration ratio(final/penult)	0.077	min. cross bnd f0 diff	0.076
median crossbnd f0 diff	0.077	duration diff.(final - penult)	0.073
min crossbnd f0 diff.	0.066	median crossbnd relative intensity diff.	0.066

5. Discussion and conclusion

From the text analysis, the PoS categories in English preceding most intP boundaries were in order of frequency, NOUN, VERB, and ADV(Adverb). In Hindi, the corresponding categories were ADP(adposition), NOUN and VERB. For IP boundaries, English most commonly exhibited boundaries after NOUNs but for Hindi, after auxiliary verbs which typically qualify verb tense.

Language identification experiments with delexicalised speech showed that accuracy exceeded chance. Next, perceptually annotated boundaries were studied for acoustic feature cues to different types of boundaries. Intensity based features were crucial indicators of different boundary types in Hindi while duration based cues dominated boundary classification in English. We speculate that these language dependent boundary cues underlie, at least partially, the above-chance discrimination obtained in the delexicalised speech experiment. The observed language-specific cue weighting is also consistent with prosodic typology. Hindi, as a broadly syllable-timed language with verb-final syntax, places less functional load on pre-boundary lengthening and more on relative intensity reduction at sentence-final auxiliary verbs. Indian English maintains stress-based rhythmic organization, making duration cues more robust at phrase boundaries.

In future, we would like to extract spectral shape features to complement intensity features, while further extending the study to prosodic prominence in Hindi and Indian English.

6. References

[1] S. Ananthakrishnan and S. S. Narayanan, “Automatic prosodic event detection using acoustic, lexical, and syntactic evidence,” *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 216–228, 2007.

[2] P. Jyothi, J. Cole, M. Hasegawa-Johnson, and V. Puri, “An investigation of prosody in Hindi narrative speech,” in *Proceedings of Speech Prosody*, vol. 7, 2014, pp. 623–627.

[3] M. of Electronics and I. T. (MeitY), “Bhashini: National language translation mission (NLTM),” https://www.bhashini.gov.in, 2022, government of India.

[4] R. Udupa U and T. A. Faruquie, “An English-Hindi statistical machine translation system,” in *International Conference on Natural Language Processing*. Springer, 2004, pp. 254–262.

[5] T. Mahrt, “LMEDS: Language markup and experimental design software,” *URL https://github.com/timmahrt/LMEDS*, 2016.

[6] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, “Stanza: A Python natural language processing toolkit for many human languages,” *arXiv preprint arXiv:2003.07082*, 2020.

[7] A. Cunningham, “Investigating holistic measures of speech prosody through ratings of filtered speech,” in *Speech Prosody*, 2012.

[8] G. Kuo, “Processing prosodic boundaries in natural and filtered speech,” in *Proceedings of the 7th International Conference on Speech Prosody*, 2014, pp. 983–986.

[9] M. Komatsu, K. Mori, T. Arai, M. Aoyagi, and Y. Murahara, “Human language identification with reduced segmental information,” *Acoustical Science and Technology*, vol. 23, no. 3, pp. 143–153, 2002.

[10] J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv, “Psychopy2: Experiments in behavior made easy,” *Behavior research methods*, vol. 51, pp. 195–203, 2019.

[11] J. Kuang, M. P. Y. Chan, and N. Rhee, “The effects of syntactic and acoustic cues on the perception of prosodic boundaries,” in *Speech Prosody*, 2022.

[12] A. Rosenberg, “AutoBI-a tool for automatic toBI annotation.” in *Interspeech*. Makuhari, Chiba, Japan, 2010, pp. 146–149.

[13] Y. Jadoul, B. Thompson, and B. de Boer, “Introducing Parselmouth: A Python interface to Praat,” *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[14] J.-Y. Choi, M. Hasegawa-Johnson, and J. Cole, “Finding intonational boundaries using acoustic cues related to the voice source,” *The Journal of the Acoustical Society of America*, vol. 118, no. 4, pp. 2579–2587, 2005.

[15] G. Kentner, I. Franz, C. A. Knoop, and W. Menninghaus, “The final lengthening of pre-boundary syllables turns into final shortening as boundary strength levels increase,” *Journal of Phonetics*, vol. 97, p. 101225, 2023.

[16] V. Puri, “Intonation in Indian English and Hindi late and simultaneous bilinguals,” Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2013.