Prosody Cues For Classification of the Discourse Particle "hã" in Hindi

Sankalan Prasad [1], *Kalika Bali* [2]

[1] Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India
[2] Microsoft Research Labs India Pvt. Ltd. Sadashivnagar, Bangalore 560080, India
Sankalan.Prasad@iitkgp.ac.in, kalikab@microsoft.com

Abstract

In Hindi, affirmative particle "hã" carries out a variety of discourse functions. Preliminary investigation has shown that though it is difficult to disambiguate these different functions, there seems to be a distinct prosodic pattern associated with each of these. In this paper, we present a corpus study of spoken utterances of the Hindi word "hã". We identify these prosodic patterns and capture the specific pitch variations associated with each of the various functions. We also examine the use of prosodic cues in classification of the utterances into different functions using k-means clustering. While certain amount of speaker dependency, as well as lack of contextual and lexical information resulted in high classification entropy, however, the results were consistent with comparable studies in other languages.

Index Terms: discourse particle, prosody, pitch variations, k-means

1. Introduction

The affirmative particle "hã" in Hindi is used both as a discourse particle as well as the lexical equivalent of "yes". Further, "hã" can occur in isolation, in the middle of the sentence as well as to mark the beginning or ending of a discourse segment. Among the various functionalities of "hã", it is seen that "hã" is usually used to convey acknowledgement or agreement, or as a back channel to indicate that one interlocutor is still attending to another, or as a question to seek confirmation or clarification for a particular discourse segment [1].

There is evidence that prosodic features contain information that can be used to disambiguate different intentions and moods of the speakers. Price et al. [2] has already demonstrated the use of prosody in syntactic disambiguation. Hockey [3] found that utterances of "okay" in English differed in terms of pitch contours. Later, Gravano et al. [4] further studied the use of prosodic information for disambiguation of interpretation of "okay". Horne et al. [5] studied the use of prosody for disambiguation of different usages of "men" (lexical equivalent of "but" in Swedish). Other than [1] which describes the different possible discourse functions of "hã" in term of f_0 contours, no significant work has been done in Hindi on the interaction of prosody and discourse meaning .

The prosodic patterns of Hindi and its numerous dialects have not been well researched. In this paper we wish to investigate the connection between prosodic cues such as pitch, intensity and duration, and the interpretation of the discourse particle "hã". The intuition is that every intention will have a distinct prosodic pattern associated with it. We want to first identify these idiosyncratic features and then use it in a classification task. That is, we want to examine to what extent can the function of "hã" be extracted from an isolated utterance? Even with contextual information, the interpretation and identification of the discourse function of these particles may be ambiguous. However, it is not very clear about the role prosodic cues play in disambiguation or which prosodic cues are salient in differentiating the utterances of the same word. In particular, we want to investigate the importance of pitch and its associated prosody for the identification and classification of discourse functions.

In the past classification tasks based on prosody only without the help of lexical or contextual information have not been very encouraging. Lai [6] used similar features for a classification task of the word "really" in which she found that the results were just better than at random. Although, prosody alone has been shown to classify discourse functions with 60% [6-9] accuracy in previous studies, we know from literature that f_0 is an important cue. "hã" is used as "yes" in its regular linguistic function it is extremely important that a Spoken Dialogue System interprets it correctly to avoid confusion and interruptions. This is problematic because as it is ASR accuracies for a small word like "hã" are notoriously low. And if we understand better the role of pitch in classifying the different functions of "hã" then it may be employed more effectively in such cases. According to the 2001 census, Hindi and its dialects are spoken as the first language by 41% of people in India. Considering the high frequency of usage of "hã" in spoken Hindi (1322 utterances in about 170 min. of recording data), an efficient intention determination of it will go a long way in improving spoken language understanding of the language.

The rest of the paper has been organized as follows: Section 2 describes the corpus used for the study. Section 3 identifies the various types. Section 4 describes the features of the tokens. Section 5 describes the feature selection process and Section 6 describes the experiment performed elucidates the results.

2. Corpus

The corpus consisted of 35 spontaneous telephonic conversations in colloquial Hindi. We had a set of 70 unique speakers, 49 male and 21 female. Subjects were asked to call friends or relatives and talk on any topic of personal interest ranging from exams, upcoming marriages, welfare of families, weather, trips, visits home etc. Five dialects were represented in the speaker set, namely Bundeli, Awadhi, Kanauji, Chhattisgarhi and Haryanvi All recordings are channel separated with the caller inline and the receiver on

[1] This work was carried out during the author's internship at Microsoft Research labs India Pvt. Ltd.

the outline. All recordings were made at 8 KHz sampling rate, 8-bit Mu-law or A-law.

The "hã" tokens were extracted manually from these conversations. While extracting, care was taken that only the isolated utterance was retained and no contextual information could be inferred from the utterance. A total of 1322 "hã"s were extracted. Out of the 1322 utterances 133 questions, 550 agreement/acknowledgement/yes and 639 backchannels were annotated [section 3 and 4].

3. Types

The extracted "hã" tokens were identified to be of three types discussed below. Even though there are many other functionalities of "hã" these three were identified to be most frequent in usage. Also they were identified to have the most distinct prosodic patterns.

3.1. Question

An utterance wherein the speaker seeks clarification on the preceding discourse segment was marked as a question. A question signifies that the speaker is actively seeking more information from the other side. It is a strong cue for the other interlocutor that a response from his side should follow immediately. It causes a change in direction of discourse flow. An example of a question is included below.

Speaker 1: बारिश [“rain”]
Speaker 2: हाँ? [“*hã* (?)”]
Speaker 1:यहाँ बारिश शुरू हो गई [“(It) has started raining here”]

3.2. Backchannel

. There has been a lot of research into accurately defining backchannels however most of the work has been inconclusive [10-12]. Backchannels can be loosely defined as feedback. It is a cue for the other interlocutor to continue the discourse and a signal that the speaker is attentively listening. It is a passive way to seek information. However, it may be argued that backchannels are similar to agreements as in a way the speaker is agreeing with the ongoing discourse. Therefore, some ambiguity exists between agreements and backchannels. In general, if the other interlocutor can ignore the utterance completely and continue with whatever he/she was saying then it is considered as backchannel. Essentially a backchannel doesn’t cause a change in the flow of the discourse. An example of a backchannel is included below.

Speaker 1: हम तो गए थे दिल्ली [“I had gone to Delhi”]
Speaker 2: हाँ [“*hã”*]
Speaker 1: उधर ही गिर पड़े [“There (emph) (I) fell down”]
Speaker 2: हाँ [“*hã”*]
Speaker 1: और बाजू टूट गया [“and (my) arm broke”]

3.3. Agreement/Acknowledgement/Yes

An agreement or acknowledgement is an utterance where the speaker wants to convey that he is in agreement with whatever the other interlocutor had said. It essentially means that the speaker had contemplated on whatever the interlocutor had just said and is agreeing with it. It may also mean that the answer to the question asked by the interlocutor is “yes”. Such responses are usually in response to a question.

Notation	Interpretation
H	gentle rise
H*	Peak
H%	smooth rise
!H*	dip after a peak
L	gentle fall
L*	trough
L%	smooth dip
+	to indicate change in profile shape

Table 1: Notations and their interpretations

Label	Profile
Backchannel	H,L,H+L%,H+H%,L+H+L%, !H*,L+H%
Agreement	H+!H*, L,L*+H+L%,H+L%
Question	L*+H+!H*,L*+H,L*+H+L%, H*+!H*, L*+H*, L*+H*+H+L%

Table2: Profiles of individual labels

The other interlocutor in the mean time waits for the response of the speaker and the direction of discourse thereafter depends upon the response of the speaker. An example of an agreement has been included below.

Speaker 1: “हैलो, शर्मा जी हैं?”[“Hello# is Mr. Sharma there?”]
Speaker 2: “हाँ, आ रहे हैं” [“*hã* # (he) is coming”]

4. Features of Tokens

We examined the pitch contour, power and duration of each utterance separately and tried to analyze why different utterances of the same word get interpreted differently.

"hã" can be broken down to two phonemes /h/, / ã /. Of these /h/ is usually unvoiced. Duration of /ã/ can vary speaker to speaker and is usually nasalized. Sometimes nasalization is ignored and in places it can also be realized as a nasal stop. The labels used for annotating the pitch contours are given in table 1. Table 2 gives the types of pitch contours for each of the discourse type.

Questions are usually indicated by a rising profile. Speakers use questions both with nasalization and without nasalization. Utterances with a sharp peak either at the end (fig. 1.1), or near the middle (fig. 1.2) are invariably interpreted as questions. In fig. 1.2 the part after the peak is due to nasalization of the utterance. Questions without nasalization have a steep and steady rise in the pitch contour till the end; it might be followed by a slight dip at the end of the contour occasionally. Backchannels, which act as a gap filler, on the other hand were characterized by short durations, low power and relatively flat pitch contours (fig. 1.5). Agreements also have a flat profile (fig. 1.8) but they are marked by longer durations and more power.

As is also evident from Table 2, questions and agreements have very distinct profiles except in certain cases when the profile is L*+H+L% (fig. 1.3). Both backchannels and agreements (fig. 1.4 and 1.7) may have a rising profile in which case it might be confused with questions. Additionally both backchannels and agreements have a trailing pitch (fig. 1.6 and 1.9). The main point of distinction between the two is the duration and power. The ones with a longer duration and more power are more likely to be agreements. Backchannels have comparably shorter durations and lesser power.

However the trends are very speaker specific. Some speakers have a tendency to speak shorter "hã" with a lot more power for agreement. In such cases prosodic information is insufficient to successfully determine its type. Also most profiles of agreements overlap the ones of backchannels. However agreements had significantly longer duration than backchannels.

5. Feature Selection

As discussed in section 4, all utterances were discernible based on their pitch, duration and power. Pitch period is an important parameter in analysis and synthesis of speech signals. In our work we used an lpc based cepstrum approach towards pitch determination [13]. To remove pitch doubling errors peaks at approximately half the extracted frequency was searched for. If there was a peak above a specific threshold, determined empirically, then it was accepted as the fundamental frequency. Thereafter curve smoothening was used to remove sudden jumps in f_0. The pitch extraction was compared to PRAAT [14] and it showed comparable results.

While selecting the features, we tried to have a good coverage of features and feature extraction regions. Apart from the standard statistical features in prosody, attempt was made to capture features related to the entire profile of the pitch contour. The entire utterance was divided into 5 regions and mean and standard deviations were calculated for each of these. We also had parameters to capture the number of points with positive and zero slope, the sum of frequencies with a positive slope, whether the profile is rising or not, the number of points in the beginning with a positive slope, the number of points on the contour at the end with a negative slope. Apart from these, the duration of the utterance was also included as a feature because agreements tend to have longer utterances and backchannels relatively short ones. Energy and power features were also included in our feature list. A total of 42 features (Table 3) were identified which were expected to have an important role in efficient classification.

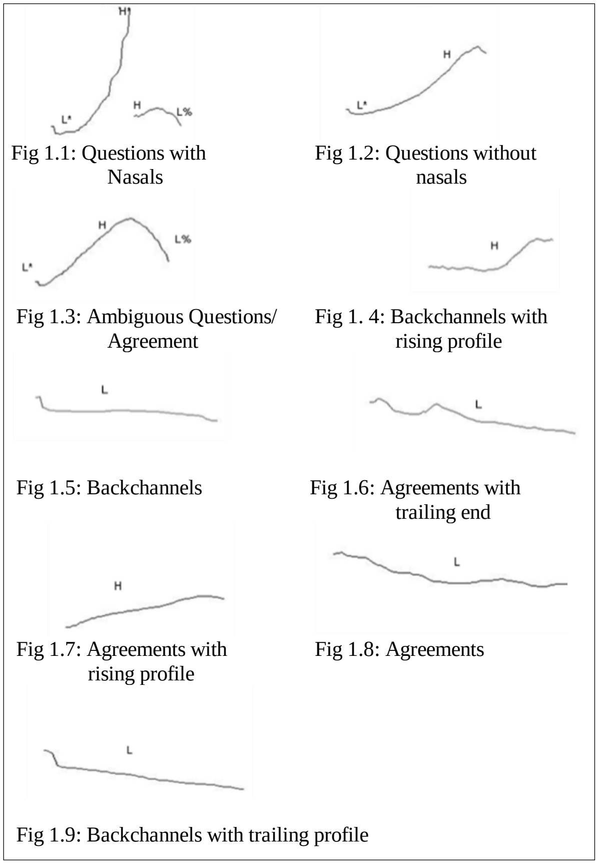

Fig 1.1: Questions with Nasals

Fig 1.2: Questions without nasals

Fig 1.3: Ambiguous Questions/ Agreement

Fig 1. 4: Backchannels with rising profile

Fig 1.5: Backchannels

Fig 1.6: Agreements with trailing end

Fig 1.7: Agreements with rising profile

Fig 1.8: Agreements

Fig 1.9: Backchannels with trailing profile

Figure 1: Distinct Prosodic Patterns of the various types of utterances

All the features were normalized with respect to each individual speaker and then pre-processed to zero mean and unit variance across all speakers. A random set of 300 samples was selected and hand annotated to indicate question agreement or backchannel. A regression tree was run on these samples and the features used in the decision making process were isolated. A total of 23 features were identified as useful for distinguishing the three classes with reasonable accuracy. They have been marked in italics in Table 3. Mean was considered important only in 2nd 3rd & 4th sections, standard deviation in 1st 2nd & 3rd sections, slope of best fit line in the 2nd & 4th section.

Duration Features
Duration: length of the utterance
Num_bound: length of the trailing part of the pitch
Num_accent: length of the rising part of the pitch

Power Features
Power: the power in the utterance.
utt_nrg_mean: the mean square energy
End_nrg: the energy in the last spoken frame
Start_nrg: the energy in the first spoken frame
Pen_nrg: the energy in the penultimate spoken frame

F0 Features
Start_f0: the starting value of the pitch
End_f0: the end value of the pitch
Max_f0: the maximum value of the pitch
Min_f0: the minimum value of the pitch
Range_f0: the variation in pitch
Mu: the mean pitch of the utterance
Mu_part: the entire utterance was divided into 5 equal spaced intervals and the mean pitch of each of the utterances were considered separately (they were normalized with respect to the mean of the speaker)
Sigma_part: the entire utterance was divided into 5 equal spaced intervals and the variance in pitch of each of the utterances were considered separately (they were normalized with respect to the variance of the speaker)
Median: the median of the utterance
Mode: the mode of the utterance
Inc_delta: the percentage of number of points in pitch profile with a positive slope
Eq_delta: the percentage of number of points in pitch profile with a zero slope
Peak_pos: the position of the maxima in speech
Peak_weight: a weight assigned to the position to denote how sharp the maxima is
Rising_sum: the sum of all the frequencies having a positive slope
Falling_sum: the sum of all the frequencies having a negative slope
Max_intensity_pos: the position of maximum intensity in the utterance
Max_intensity: the absolute value f the maximum intensity
Isrising: 1 if rising_sum> falling_sum else 0.
Slope: the entire utterance was divided into 5 equal spaced intervals, and the slope of the best fit line in each interval was considered separately.
Increasing_sum: the weighted sum of frequencies having a positive slope
Distance: the mean square deviation from the best fit line for the utterance

Table 3: Feature set

6. Experiment & Results

A classification experiment was designed to test the usefulness of this feature set in a. Each utterance was manually labeled as Question, Backchannel or Agreement. On the reduced feature set from section 5, singular value decomposition was applied. The first 10 values were selected thus retaining 79% of the total energy. All the data points were projected along the basis vectors. A standard k-Means algorithm was used to linearly classify the n-dimensional space into different classes using correlation distance. We clustered the data into 10 clusters, as the number of questions was the least and approximately one tenth the total number of utterances. Ten points with maximal distance were chosen and were set as the initial seed centers for the clusters. Based on the dominant type, each cluster was assigned that label. One cluster was found to have a dominant number of questions; five clusters were identified as agreements and four as backchannels. A detailed analysis of the cluster has been included in Table 4. For the calculation of the entropy an equal prior was used. The average entropy of classification was 0.4184. The classification accuracy has been studied in Table 5.

Cluster	Label	Entropy
Class 1	Agreement/Acknowledgement	0.4466
Class 2	Question	0.5473
Class 3	Back Channel	0.4029
Class 4	Agreement/Acknowledgement	0.3972
Class 5	Back Channel	0.3886
Class 6	Back Channel	0.3333
Class 7	Agreement/Acknowledgement	0.4286
Class 8	Agreement/Acknowledgement	0.5384
Class 9	Agreement/Acknowledgement	0.4302
Class 10	Back Channel	0.2467

Table 4: Cluster with their assigned label and entropy values

Type	Total Number	Positive id	Precision	Recall
Questions	133	67	0.453	0.504
Backchannel	639	371	0.654	0.581
Agreement	550	336	0.554	0.611

Table 5: Analysis of Classification task

7. Conclusions

"hã" is used both as a discourse particle and also as a lexical equivalent of "yes" in Hindi. Three types of functions of "hã" were considered; questions, backchannels and agreement/acknowledgement/yes. Their respective prosodic patterns were identified. It was observed, the utterances of the three functionality types varied on a total set of 42 features. However a regression study on the corpus showed that only 23 of them were relevant. A classification experiment was carried out based on these features. The average entropy of the classification was 0.4184. Which meant on an average about 41% of the utterances was misclassified. Identification of backchannels had the highest precision, i.e. among those classified as backchannels about 65% were indeed backchannels, while that of agreements had the highest recall value, i.e. of all the agreements 61% was positively identified as agreements.

The misclassification percentage though high was comparable to similar work in other languages [6-9]. The high misclassification percentage can be due to the vast variations in intonation patterns across the various dialects. Also prosody is not mono-functional. Especially in spontaneous conversation there could be a number of interacting factors that influence the prosodic patterns, like focus, stress, emotion, intention etc. We believe, however, that prosody provides a very strong cue for disambiguation and if combined with a language model may give much improved results. Since the prosody information like power, duration and intonation was insufficient for efficient classification, as a next step, we would like to combine prosody with lexical information. Also, taking into account the context vis-à-vis utterances from the other interlocutor adjoining the isolated utterance will give us stronger cues for disambiguation.

8. References

[1] Bali, Kalika, "F0 cues for the discourse functions of "*hã*" in Hindi", *Proceedings of the 10thAnnual Conference of the International Speech Communication Association*, Brighton, September 2009

[2] Price, Patti; Ostendorf, Mari; Shattuck-Hufnagel, Stefanie; Fong, Cynthia, "The Use of Prosody in Syntactic Disambiguation", *The Journal of the Acoustical Society of America*, Volume 90, Issue 6, December 1991, pages 2956-2970.

[3] Hockey, Beth Ann "Prosody and the role of okay and uh-huh in discourse", *Proceedings of the Eastern States Conference on Linguistics*, 1993 pages 128-136.

[4] Gravano, Agust´in; Benus, Stefan; Ch´avez, H´ector; Hirschberg, Julia; Wilcox, Lauren; "On the role of context and prosody in the interpretation of 'okay'" *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, June 2007, pages 800-807.

[5] Horne, Merle; Hansson, Petra; Bruce, Gösta; Frid, Johan; Filipsson, Marcus "Cue words and the topic structure of spoken discourse: The case of Swedish men 'but'" *Journal of Pragmatics* Volume 33, Issue 7, 2001, pages 1061-1081.

[6] Lai, Catherine "Prosodic cues for backchannels and short questions: really?", Speech Prosody, 2008, pages 413-416

[7] Shriberg, Elizabeth; Bates, Rebecca; Stolcke, Andreas; Taylor, Paul Jurafsky, Daniel; Ries, Klaus; Coccaro, Noah; Martin, Rachel; Meteer, Marie; Van EssDykema, Carol, "Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?" *LANGUAGE AND SPEECH* 41(34): Special Issue on Prosody and Conversation, 1998, pages 439-487.

[8] Shriberg, Elizabeth; Hillard, Dustin; Ostendorf, Mari, "Detection of Agreement vs. Disagreement in Meetings: Training with Unlabeled Data". *Proceedings of HLT-NAACL* 2003--short papers - Volume 2, pages 34-36.

[9] Seppänen, Tapio; Väyrynen, Eero; Toivanen, Juhani, "Prosody-based classification of emotions in spoken Finnish", *Proceedings of 8th European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2003 pages 717 - 720.

[10] Schegloff, E.A, "Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences". *In: D. Tannen ed., Analyzing discourse: Text and talk, Washington, DC: Georgetown University Press, 1982, pages: 71-93*

[11] Clancy, P.M.; Thompson, S.A.; Suzuki, R. and Tao, H, "The conversational use of reactive tokens in English, Japanese and Mandarin", *Journal of Pragmatics* 26, 1996 pages: 355-387.

[12] Ward, Nigel; Tsukahara, Wataru, "Prosodic features which cue back-channel responses in English and Japanese". *Journal of Pragmatics* Volume 32, Issue 8, July 2000, Pages 1177-1207.

[13] Ding, Hui; Qian, Bo; Li, Yanping; Tang, Zhenmin, "A Method Combining LPC-Based Cepstrum and Harmonic Product Spectrum for Pitch Detection", *Proceedings of the 2006 International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIH-MSP'06*, December, 2006, pages: 537-540

[14] Boersma, Paul; Weenink, David, "Praat: doing phonetics by computer" (Version 5.1.07) [Computer program]. Obtained in May, 2009, from http://www.praat.org/