P. Rao, N. Sanghvi, H. Mixdorff, and K. Sabu. "Acoustic correlate of focus in Marathi: Production and perception," *Journal of Phonetics*, 65C (2017), pp. 110-125.

Acoustic correlates of focus in Marathi: production and perception

Preeti Rao*[1], Niramay Sanghvi[1], Hansjoerg Mixdorff[2], Kamini Sabu[1]

*Corresponding author: prao@ee.iitb.ac.in

[1]Department of Electrical Engineering, Indian Institute of Technology Bombay, India

[2]Beuth University Berlin, Germany

Abstract

Focus or prominence is an important linguistic function of prosody. The acoustic realisation of prominence in an utterance, in most languages, involves one or more acoustic dimensions while affecting one or more words in the utterance. It is of interest to identify the acoustic correlates as well as their possible interaction in the production and perception of focus. In this article, we consider the acoustics of focus in Marathi. Previous studies on Hindi, the more researched member of the Indo-Aryan family, have reported that the well-known rising *F0* pattern on non-final content words in an utterance becomes hyper-articulated when the word is in focus. The associated *F0* excursion, duration and intensity increase and are accompanied by post-focal compression of pitch range. A preliminary goal of the present study was to verify whether Marathi exhibits similar behaviour. We used Subject-Object-Verb (SOV) structured utterances with elicited focus on each word by 12 native Marathi speakers. We observed that each narrow focus location is accompanied by a distinct set of local and global acoustic correlates in *F0*, duration and intensity which closely parallel previous observations on Hindi. *F0* cues were also examined via the accent command amplitudes of the Fujisaki model. *F0* range, duration and intensity were found to vary significantly with focus condition prompting a study to examine their relative importance in the perceptual judgement of focus. Perception testing with synthetically manipulated utterances revealed that duration cues are interpreted in a categorical manner, relatively uninfluenced by the pitch cues. Only when duration is ambiguous, does the on-focus *F0* cue appear to play a role. An explanation for this may lie in the normal *F0*-rise characteristic of the content words in Marathi, making pitch a less dependable functional cue for focus.

Keywords

Focus, acoustic correlates, Fujisaki modeling, Marathi prosody

1.0 Introduction

Focus refers to the marking of new information in an utterance over given information by emphasizing it in some way. Focus may be of different widths with “broad focus” indicating all new information,

and “narrow focus” indicating that one word or phrase at a specific location in an utterance is emphasized over all others. Focus is typically achieved via the manipulation of prosody although in some languages it may be realised by changing word order (Swerts et al., 2002). The phonetic marking of focus itself is based on a common set of acoustic attributes although the precise cues, and the relative importance of each, are language dependent (Koreman et al., 2009). Although prominence is typically associated with a word, the acoustic cues may be described at the word level or sub-word levels such as syllable or even vowel nucleus (Rosenberg & Hirschberg, 2009). Experimental studies as early as 1958 on English demonstrated that manipulating duration, fundamental frequency (*F0*) and intensity influenced listeners’ judgement of stress on a syllable (Fry, 1958). Although segmental quality can also be a factor in stress or prominence perception due to possible hyper-articulation (Cho et al., 2011), nearly all languages use one or more of the following acoustic dimensions of speech: *F0*, duration and intensity. Systematic variations of these properties at the word or syllable level are utilised both in the production and perception of focus. Accentuation of a segment is often accompanied by the deaccentuation of other segments leading to further enhancement of the prominence effect. In stress accent languages, it is the lexically stressed syllables that undergo modification under changing focus (Bolinger, 1986). The acoustic correlates of focus in a particular language and their dependence on the location and type of focus are of interest both from a scientific viewpoint as well as for speech technology applications such as automatic speech understanding, computer-aided pronunciation training and text-to-speech (TTS) synthesis. In the latter, accurate models for prosody can serve to generate more intelligible and natural speech.

There have been numerous studies about the phonetic realisation of focus in several languages using elicited utterances by native speakers. Breen et al (2010) studied the prosodic realisation of information structure in American English, as a representative West-Germanic language. Based on utterances elicited in response to wh- questions, the acoustic properties of narrow focus were measured in different sentential locations. It was observed that focused material is produced with longer duration, higher *F0* and greater intensity than non-focused material. The strongest indicators for discriminating the different focus conditions were word duration, word maximum intensity, word mean and maximum *F0*. In an earlier study specifically on English declarative intonation, Xu and Xu (2005) reported that the pitch range of the focused word is expanded, the pitch range of the post-focus items, if any, is compressed and lowered, and the pitch range of the pre-focus items, if any, remains neutral.

Acoustic properties of prominence marking in German and Norwegian were compared by Koreman et al. (2009) to find that the former used *F0* and intensity cues while Norwegian relied on duration cues. In a more recent study on German, perceptual prominence of a word was observed to be strongly correlated with *F0* transition, syllable duration, maximum intensity and mean harmonics-to-noise ratio (Mixdorff, 2015). Unlike English, German and Greek, Georgian is found to use duration to signal focus, with focus showing very little influence on intonation (Skopeteas & Fery, 2010). Turkish uses duration and intensity to cue focus (Ipek, 2011). It was reported that while on-focus pitch variation was absent in Turkish, narrow focus in sentence initial position elicited post-focal pitch compression. A perception experiment with the same recorded utterances across different focus locations showed that, in fact, initial focus was recognized most accurately by listeners. In a tone language like Mandarin, on-focus increase in *F0*, duration and intensity have been observed across dialects with post-focal compression of intensity and pitch ranges playing a prominent role with Beijing Mandarin speakers (Chen et al., 2009). Korean also shows both narrow focus and post-focal effects that involve all the acoustic parameters (Lee, 2010).

Most studies investigating the acoustic correlates of focus have been based on production, i.e. by using measurements on elicited speech by native speakers. More definitive conclusions can be expected from

the systematic testing of focus perception by native listeners provided suitable stimuli are available. The relationship between variations of stimulus parameters and perceived prominence can reveal the importance of the different acoustic cues provided they can be teased apart, which is usually possible only with synthesized stimuli. There have been a few studies of this type on intonation based features that have attempted to relate *F0* variation to perceived prominence (Terken, 1989; Gussenhoven & Rietveld, 1988; Gussenhoven et al., 1997; Ladd & Morton, 1997; Hermes, 1997; Terken & Hermes, 2000). The role of other prosodic parameters—mainly intensity and segmental durations— has been studied much less. Vainio and Jarvikivi (2006) in the course of studying the dependence of prominence perception on tonal features in Finnish, found that intensity has an effect and that speakers could exploit a perceptual trade-off between pitch height and intensity increase. That languages differ in the degree to which they exploit duration, F0 and intensity in production but also to some extent in perception was recognised in the course of a recent large study on information structure across languages (Andreeva et al., 2015).

Among Indo-Aryan languages, Hindi has been studied relatively extensively for its prosody, as also specifically for the prosody of prominence. Hindi is a head-final (SOV) language with relatively free word order. While particles (equivalent to the English words "only", "also" or "even") can serve to mark focus, this is not necessarily the case (Patil et al., 2008). Several studies on Hindi intonation have noted that while declarative sentences follow the universal pitch down-step pattern, the individual (non-final) content words each show an *F0* rise from left to right across the word (Moore, 1965; Pandey, 2015; Rajendran & Yegnanarayana, 1996; Harnsberger, 1996). Harnsberger (1996) assigned an LH intonation pattern to all but phrase-final words in his study with 3 native speakers, and observed that syllable weight had no role to play in this. Similar observations by several others have led to the view that each content word constitutes a separate phrase (Fery, 2010; Patil et al., 2008). Several other Indian languages share this characteristic (Hayes & Lahiri, 1991; Khan, 2007; Khan, 2016; Fery & Fanselow, 2008) including Marathi (Rao & Srichand, 1996). The property has also been exploited for word segmentation of continuous speech in Hindi (Rajendran &Yegnanarayana, 1996; Rao & Srichand, 1996). The F0 rise characteristic may mean that the more significant *F0* changes would necessarily occur towards the end of the word which raises the question of the "lexically stressed" syllable within a word. Hindi and Marathi are considered syllable-timed languages and lexical stress per se is not a well understood aspect. However there have been some commentaries that assign lexical stress based on syllable weight in Hindi (Dryud, 2001) and Marathi (Dhongde & Wali, 2009) although the acoustic correlates of stress are unknown or even considered non-existent as in some other South Asian languages (Khan, 2016). Ohala (1986) suggests that stress is weaker than in English but only one syllable can be made prominent in a word.

Given the word-level *F0* rise characteristic, focus, in Hindi, has been shown to bring in additional prosodic effects, namely (i) exaggeration of the *F0* rise on the focus word by way of greater excursion (Harnsberger 1996, 1999; Moore 1965; Dyrud, 2001, Patil et al., 2008). (ii) compression of the post-focal pitch range although rising pitch accents are still realised on the post-focal content words (Patil et al., 2008). Pre-focal elements do not undergo modification. Genzel and Kugler (2010) found that contrastive focus in Hindi is linked with an increase in duration of the stressed syllable as well as increased *F0* rise across the word that came as much from lowering the L tone as raising the H tone in the hammock shape. Puri (2013) found the main acoustic correlates of focus in Hindi by bilingual speakers (of Indian English) to include increased duration, as well as an increase in *F0* excursion, both on the focused element, and post-focal reduction in duration, intensity and *F0* excursion. No intensity increase was observed on the focused word. Post-focal compression of pitch range was also noted previously (Harnsberger, 1996; Patil et al., 2008, Fery, 2010).

In summary, studies of focus in Hindi are in agreement about the existence of post-focal compression of *F0* range, duration and intensity. Further, there is a general agreement that both duration increase and *F0* rise are observed on the focused word. A few other Indian languages have been studied for the acoustics of focus via elicited utterances by native speakers. Khan (2007) observed that focus on a word in Bengali leads to a further *F0* rise on the final syllable and is accompanied by post-focal compression in pitch range so that post-focal content words do not exhibit the normal word-level *F0* rise character. Fery (2013) discusses Konkani as a typical Indian language where focus is right-aligned in a prosodic phrase. There have been studies on the prosody of Indian English which discuss the acoustic cues of focus. While Moon (2002) reported for Indian English that pitch range does not differ between focused and unfocused content words, Maxwell (2010) reports that Bengali as well as Indian English by Bengali speakers shows an increase in pitch excursion on the focused word accompanied by the lengthening of the accented syllable. In our previously reported work on Marathi SOV utterances (Rao et al., 2016), it was noted that focus on a word led to a significant increase in the *F0* rise on the final syllable except in in the case of subject focus. Duration lengthening of the stressed syllable of the word was noted in all narrow focus locations where the stressed syllable as defined as the quantitatively heaviest syllable in the word. Post-focal compression of pitch range, duration and intensity were also observed.

In this work, we consider the production and perception of narrow focus in Marathi in SOV utterances. Marathi is the state language of Maharashtra in India with over 70 million native speakers. We elaborate upon the previously reported work (Rao et al., 2016) on elicited utterances by 12 native speakers of Marathi in contrastive and non-contrastive focus conditions, while extending the acoustic analyses further to include utterances with focus on a function word. We also study the variation in the phonetic realization of narrow focus across words with different stress patterns. Acoustic measurements of raw *F0*, intensity and duration are used to derive syllable- and word-level features which form the basis for detecting the significant effects in terms of both on-focus word and across-utterance phenomena. Fujisaki modeling of the *F0* contour (Fujisaki, 1988) is exploited to compactly describe the intonation pattern over the utterance in terms of physically meaningful constituents leading to a clearer understanding of the pitch cues. The quantitative modeling of the pitch contour also helps us to design and implement a perceptual study based on manipulated stimuli to derive the relative importance of the different acoustic dimensions in the judgement of focus by native Marathi listeners.

In the next section, we present the speech materials used in this study. This is followed by a description of acoustic measurements and analyses. The methodology and results of the perception study are presented next. Finally we discuss our findings and present a picture of the salient aspects of focus realisation in Marathi SOV utterances with reference to the literature on other languages.

2.0 Speech materials

Like many other Indian languages, Marathi is an SOV language with flexible word order that permits any constituent to appear in sentence-initial position and become the topic (Junghare, 1988). Thus, it is possible that speakers choose between syntactic and prosodic means to varying extents to signal focus although there are no known studies on this aspect for Marathi. The authors' experience with news reading style analyses (Barhate et al., 2016) indicates that purely prosodic means of conveying prominence are indeed commonly encountered, especially in professional speaking.

Since focus can also be signaled by change in syntactic word order and/or by a particle, it was important to consider whether eliciting prominence by constraining the target sentence came easily and naturally to native Marathi speakers. The selected target sentence, as seen in Table 1, has four critical words: subject (*Amol*), object (*aai*), function word (*barobar*) and verb (*bolat*), and thus allows the study of conditions across both focus type and location. We found that Marathi speakers and listeners were

perfectly comfortable with the constrained text and took effortlessly to the prosodic means of focus expression. The words are multisyllabic including the function word. Lexical accent is not distinctive in Marathi, but Dhongde and Wali (2008) provide a tentative set of stress rules based on the syllable weight in multi-syllabic words. Using the rules provided by them, we hypothesize that *mol* in *Amol*, *ro* in *barobar*, and *bo* in *bolat* are the lexically stressed syllables, and hence potential candidates for receiving emphasis due to focus (henceforth referred to as stressed syllables). The position of the stressed syllable is different in each of the three multi-syllable words allowing us to also examine this aspect in the context of focus realisation. Further, to observe dependence on stressed syllable position for focus word in a fixed sentential position, we designed an additional prompt with a different word in subject focus (*Aanand*) in which the initial syllable *Aa* is the stressed syllable, in contrast to the previous subject word *Amol*.

A total of 20 native speakers of standard Marathi were recorded. All were young adults studying or working in Mumbai with no exposure to any study of phonetics. Eventually, we used the data of 12 speakers (6 male and 6 female) for our analyses based on a perceptual verification procedure described later. Eight pre-recorded questions by a different native Marathi speaker were used as prompts to elicit appropriately focused responses as shown in Table 1 for narrowly focused non-contrastive and contrastive statements, and one for the broadly focused. In the case of the prompt for contrastive focus elicitation, the speaker used stress on the alternate word, e.g. on *Rohit* in the subject focus prompt, thus providing the contrastive context. While both types of focus were used for the subject, object and verb, the function word was restricted to the contrastive focus context given that this is the only available mode to elicit focus on this word.

Two instances of each of the (non-function word) target forms were elicited providing 18 utterances per speaker. The target utterance for the contrastive form includes the prefix *Nahi* (translates to “no”) which the speakers were instructed to articulate silently. As an introduction to the task, examples of questions and responses by a native Marathi speaker corresponding to a different SOV sentence were played out to the subjects without further explanation. We recorded our subjects with a high quality microphone in a quiet room at 16 kHz sampling rate.

Table 1: ***Target and prompt texts in Romanized script with the English translation. Each utterance (i.e. response to each question on the right) was recorded twice by each of the 12 speakers, except the function word focus (marked *) which was recorded at a later time as 4 repetitions by 5 of the 12 speakers.***

Target	Prompt
Amol aai barobar bolat hota *(Amol/mother/with /talking/was)* IPA: əmol ɑi bərobər bolət hotɑ Translation: Amol was talking with mother	*Tumhi kay mahnalat?* What did you say? *Kon aai barobar bolat hota?* Who was talking with mother? *Amol konabarobar bolat hota?* With whom was Amol talking? *Amol aai barobar kay karat hota?* What was Amol doing with mother?
Nahi…Amol aai barobar bolat hota Translation: No… Amol was talking with mother	*Rohit aai barobar bolat hota ka?* Was Rohit talking with mother? *Amol bhavabarobar bolat hota ka?* Was Amol talking with brother? *Amol aai baddal bolat hota ka?** Was Amol talking about mother? *Amol aai barobar khelat hota ka?* Was Amol playing with mother?
Aanand aai barobar bolat hota IPA: ɑnənd ɑi bərobər bolət hotɑ Translation: Aanand was talking with mother	*Tumhi kay mahnalat?* What did you say? *Kon aai barobar bolat hota?* Who was talking with mother?

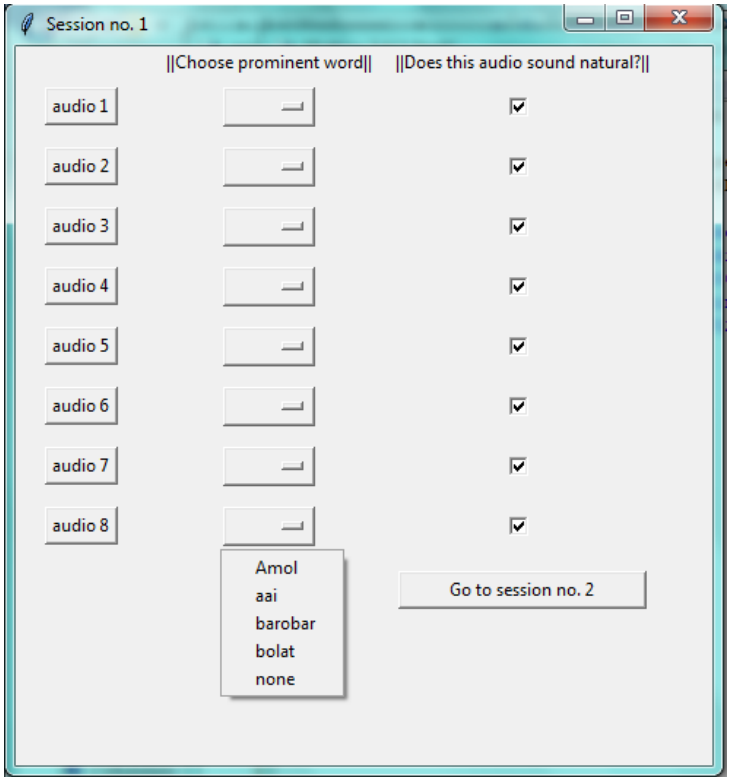
The data collection was followed by perceptual verification. Two listeners (native Marathi speakers who did not participate in the recording) were presented each recorded statement over headphones and asked to identify the width and, if narrow, the location of focus to verify that it matched the intended width and location in a 5-way forced choice test as presented in the GUI of Figure 1. For the subsequent acoustic analyses, we retained only those speakers who had all 18 utterances pass the verification with both listeners. That is, we excluded those speakers who failed to convey the intended focus in one or

more utterances. The perceptual verification, before acoustic analyses, was an important step especially given that our subjects were naive speakers without any phonetic training. Retaining those speakers who conveyed the intended meaning to both listeners in all conditions made for a perceptually validated dataset of utterances for our investigation of the acoustic correlates of focus.

At this point, we had a dataset of 216 utterances (12 speakers x 9 focus conditions x 2 instances). The 10th focus condition (focus on the function word) was recorded as an afterthought at a later time in this study and we could get only 5 of our 12 speakers back to the laboratory then to provide us a dataset of 20 utterances (5 speakers x 1 focus condition x 4 instances) all of which also passed the perceptual verification by the same two listeners. The listeners agreed that all the recordings sounded natural.

We also conducted a separate 2-way forced choice test for focus type discrimination between contrastive and non-contrastive forms with 12 speakers x 7 focus conditions x 1 instance. The contrast-distinction remained around chance. Thus we conclude that all the focus conditions were reliably discriminated across the 12 speakers while in the case of focus type, the contrastive form was not discriminated from non-contrastive. The prosodic similarity of the two forms indicates that the linguistic distinction between the contrastive and non-contrastive forms is restricted to the lexical difference provided by the negation word *Nahi*. It must be remarked that at least a few other studies have found that contrastive and non-contrastive forms are not prosodically distinguished (t'Hart et al., 1990; Koreman et al., 2009; Mixdorff et al., 2015).

Figure 1. The *Graphical User Interface for perception testing*



3.0 Acoustic parameters and analyses

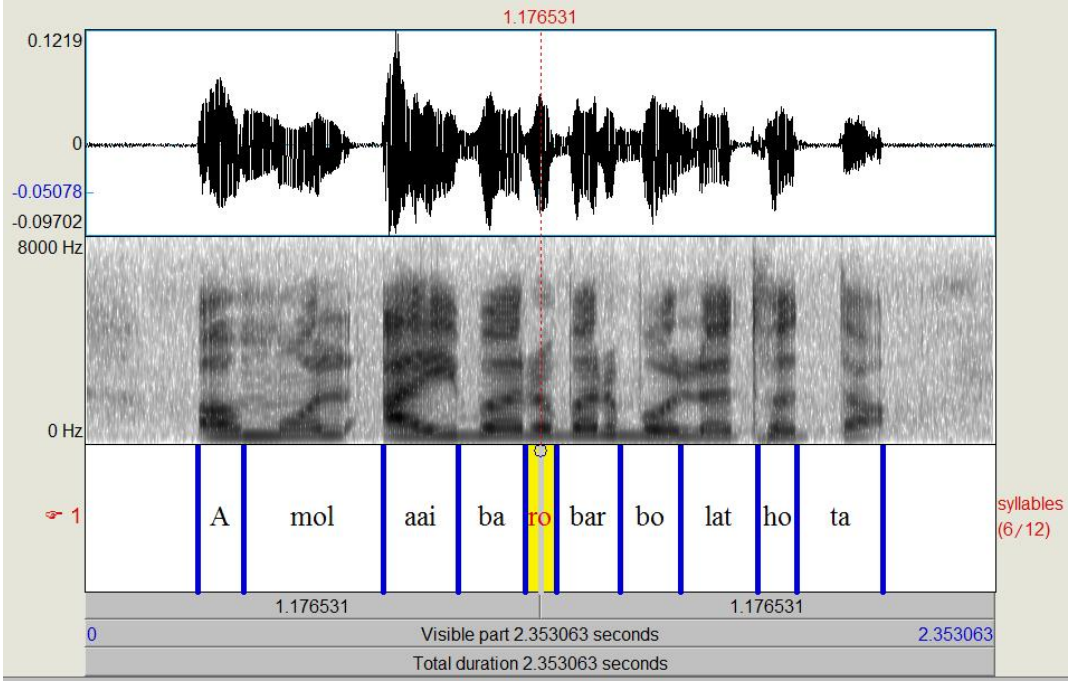
The phonetic realisation of prosodic cues within an utterance can be described in terms of measurements involving the acoustic attributes of duration, intensity and *F0* within the utterance. Narrow focus on a

given word influences the acoustics of the word itself apart from more global modifications that involve the surrounding words. Quantitative measurements of relative change are typically carried out with reference to the corresponding neutral or broad focus utterance. Syllable level segmentation of the utterance facilitates the acoustic measurements at syllable and word level. Further, it is important to normalise the measurements to account for speaker dependencies and speaking rate effects.

Each utterance was manually aligned at syllable level with the help of the spectrogram and waveform views in the Praat interface (Boersma, 2001) keeping in mind the phones constituting the syllables (V, CV or CVC corresponding to our critical words: *A-mol*, *aai*, *ba-ro-bar*, *bo-lat*). We explicitly segmented speech pauses of duration over 100 ms. Silences of shorter durations were merged with the preceding syllable. Figure 2 shows a screenshot of the PRAAT interface with a segmented utterance.

Intensity and *F0* contours sampled uniformly at 10 ms intervals across the utterance were estimated using PRAAT scripts using appropriate parameter range settings and thresholds. The occasional pitch octave errors and spurious values that occurred at phone boundaries were manually corrected by adjusting pitch analysis parameters in local neighbourhoods, and finally, signal resynthesis was performed to verify perceptual similarity with the original utterance.

Figure 2: ***Screenshot of Praat views of waveform and spectrogram of an utterance with the syllable boundaries marked***



3.1 Acoustic measurements and normalized features

The phonetic realisation of prosodic cues is described in terms of the relative changes in the acoustic attributes of duration, intensity and *F0* at the syllable or word levels. Consistent with previous work on focus in other languages, as reviewed in Sec. 1, specific acoustic measurements related to duration, intensity and *F0* were carried out at syllable and word level for each of the utterances. Every measurement is normalised with respect to utterance level observations so as to eliminate speaker and speaking-rate dependencies. This enables the grouping of the syllable (word) measurements for each syllable- or word-condition across speakers and instances, and allows us to observe the underlying

variations in the distributions, for each syllable (word), with the different focus conditions (broad, narrow, pre-focus and post-focus). We describe next the specific acoustic measurements that were carried out.

Duration: Speech segment durations are known to follow a lognormal distribution and hence are more reliably characterized on a log-transformed, rather than linear, measurement scale (Rosen, 2005). The mean and standard deviation of the log-transformed syllable durations is computed for each utterance from the sequence of syllables comprising the utterance. These quantities are used to normalise each of the individual syllable's log-transformed duration to get a Z-score (by subtracting from it the mean and dividing by the standard deviation). The normalised score for each word is obtained by summing the normalised scores of its constituent syllables.

Intensity: Intensity measurements, are computed in dB via short-time analysis at 10 ms intervals on the waveform. We recorded the mean and maximum intensity of each syllable, and the same measurements for each word. The intensity measurements are normalised by subtraction of the average intensity computed across the utterance.

F0: We measured *F0* in semitones (with respect to utterance mean *F0*) in order to obtain the word-level maximum, minimum and *F0* span. The word-level minimum and maximum constitute *F0* min and *F0* max respectively with the difference providing the *F0* span. Given that we're measuring the rise in F0 across the word segment, we consider the earliest occurring local minimum as the *F0* min with the word-level maximum following it as the *F0* max.

Next, the instances of a given syllable (word) across speakers and utterances are each categorised as belonging to one of the following four conditions: (i) broad focus, (ii) pre-focus, (iii) on-focus and (iv) post-focus. For example, an instance of the word *aai* is categorised as post-focus when it occurs within a subject-focus utterance; an instance of the syllable *bo* is on-focus when narrow focus occurs on the verb *bolat*. Such grouping provides us a variable number of instances per category for each of the segments of interest, i.e. the first 4 words and 8 syllables, of the text prompt. Based on the previously noted perceived similarity between the non-contrastive and contrastive forms we pooled the data across the two forms to get four utterances for each syllable-condition per speaker. Only the broad focus condition for each syllable was represented by two utterances per speaker.

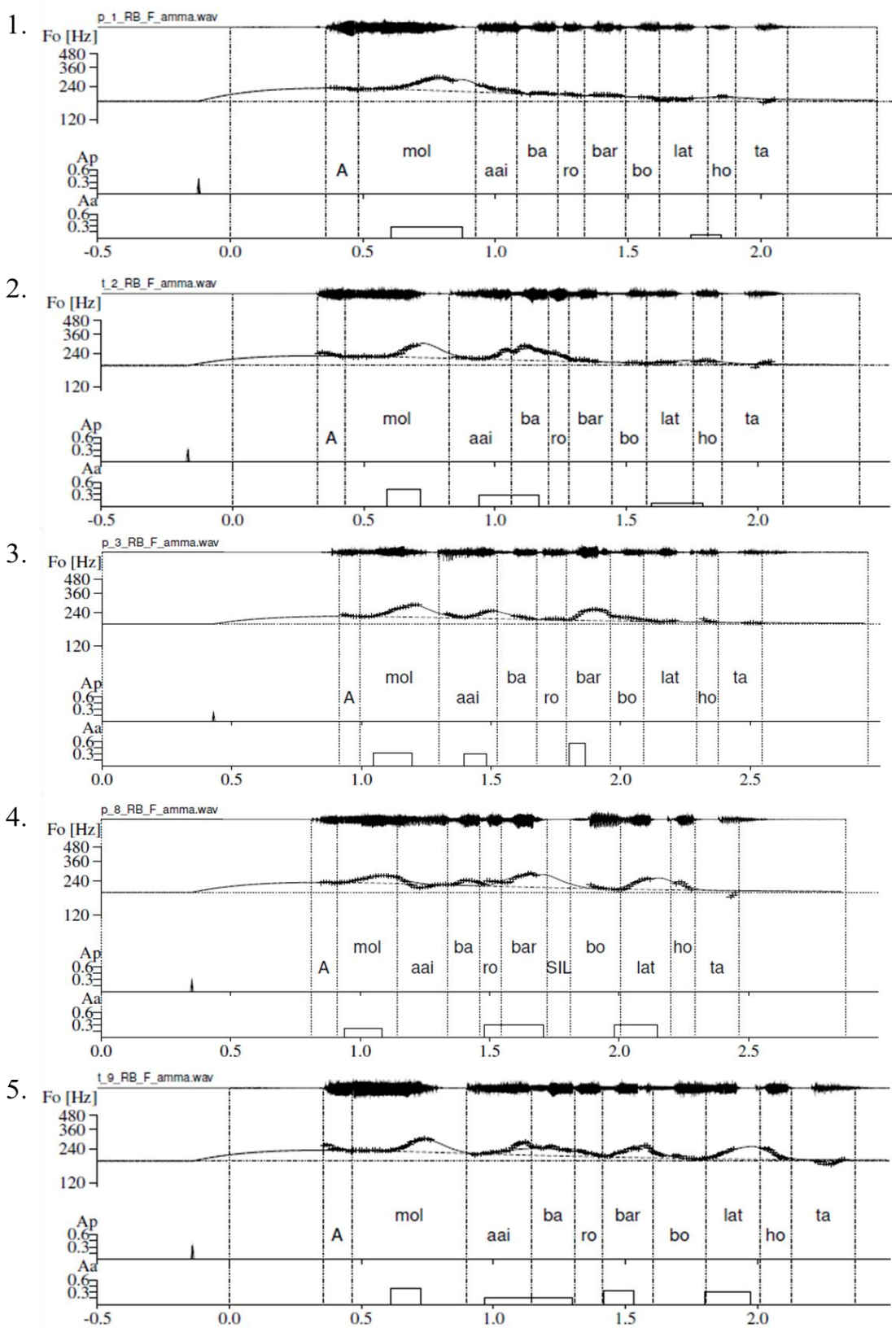
3.2 Fujisaki Modeling of the F0 Contour

The Fujisaki model provides a linguistically motivated representation of the *F0* contour in terms of the component phrases and accented syllables of the underlying utterance (Fujisaki, 1988). The Fujisaki model parameter estimation process approximates the *F0*-contour by the superposition of three components: a constant base frequency *Fb*, exponentially decaying phrase components (*Ap*) which are the responses to the phrase commands and accent components (*Aa*) which are the smoothed responses to the accent commands. Mean square error minimization in log(*F0*) space is used on an interpolated and smoothened extracted *F0* contour to estimate the locations and amplitudes of the phrase and accent commands and *Fb* (Mixdorff, 2000). The obtained contour, represented by a small set of parameters, is a smooth and continuous approximation to the measured contour. Fujisaki modeling of *F0* contours has been applied in speech analysis as well as prosody perception studies motivated partly by its ability to support *F0* modification via the easy manipulation of the linguistically meaningful model parameters. For example, it has been shown that perceived prominence of an accented syllable correlates strongly with the detected accent command amplitude (Mixdorff and Widera, 2001; Mixdorff et al., 2015). This makes the accent command amplitude a useful alternate characterization for *F0* variations related to focus.

In Figure 3, panel (1), we see an example of *F0* contour decomposition for a non-contrastive utterance with focus on *Amol* uttered by male speaker AK. The Fujisaki model parameters underlying the natural *F0* contour (indicated by +++ signs) were estimated by applying the method presented by Mixdorff (2000). The *F0* contour is modelled using one phrase component triggered by an impulse-wise phrase command with magnitude *Ap* of 0.36 at t=-0.11 s and one box-shaped accent command with *Aa*=0.51 and onset time *T1* at t=0.54 s and offset time *T2* at t=0.75 s. A second, but much weaker, accent command is seen at t=1.8 s. The modelled contour is indicated by the solid line which, as we see, approximates the natural contour closely. *Fb* of 109 Hz is indicated by the horizontal dotted line. The amplitude of the accent command, *Aa*, on a syllable can be interpreted as the strength of the pitch excursion on that syllable. The accent command amplitude, *Aa*, is measured for each syllable of the utterance by noting the value of *Aa* in the command that overlaps with that syllable. If there is no command on a syllable, its *Aa*=0.0. A command overlapping 2 syllables is assigned to both (with its actual value). We thus note that changes in sentence intonation with focus condition can be described compactly by accent command locations and amplitudes. The amplitudes can be associated with specific syllables in the utterance dependent on the location of the corresponding command.

The use of Fujisaki model parameters to represent the *F0* contour further facilitates a principled approach to the modification of the contour for perception testing (Vainio et al., 2003). That is, the quantitative modeling of the contour overcomes the difficulty inherent in the direct scaling of pitch ranges where, irrespective of whether the *F0* values are scaled or shifted, the different parts of the contour are affected in different ways (Ladd and Morton, 1997).

Figure 3: ***Speech waveform, F0 contour (extracted: ++; model: solid line) and Fujisaki model phrase and accent commands for the sentence 'Amol aai barobar bolat hota' in the following focus conditions: 1. subject, 2. object, 3. function word, 4. verb and 5. broad focus. Plots created with software by Mixdorff (2009).***



3.3 Analyses

For each syllable-condition, we compute a distribution from the acoustic measurements of each type across all speakers and utterances. Box-plot distributions of normalised log duration, mean intensity and max *F0* are presented in Figures 4, 5 and 6 respectively for each of the word- and syllable-conditions. With the dependent variable being the acoustic measurement of interest, Table 2 presents the results of one-way ANOVA showing the significance of the differences between the distributions of measured values for each word in neutral focus and in a given focus condition of interest (i.e. narrow focus or post-focal). Finally, we study the behaviour of Fujisaki model accent command location and amplitude with changing focus condition for each syllable.

Duration: We observe in Figure 4 that focus type clearly influences word duration for all focus locations in the following manner. Word duration is extended in narrow focus with respect to the neutral focus. Pre-focus durations (of *Amol* and *aai*) are not discriminated from neutral focus while the word-final verb (*bolat*) shows a clear duration reduction, relative to neutral focus, when it is post-focal. For each word, the syllable durations are unequally affected as would be expected from the syllabic stress pattern of the word. From the rules of Dhongde and Wali (2009), we note that the stressed syllables are *–mol*, *-ro-* and *bo-*, and that, indeed, the most prominent changes occur on these syllables. Thus it is the stressed syllable (i.e. final in *Amol*, and penultimate in *barobar* and *bolat*) that undergoes elongation in narrow focus on the word and correspondingly reduction when the word is post-focus. We note from Table 2 that the differences from neutral focus are significant for all on-focus locations. In the case of post-focal duration, only the verb location shows significant difference. A separate observation was that of pre-focal pauses. We detected 47 inter-word pauses in all across the 12 speakers' data. Of these 37 were pre-focal (26 before *aai*, and 11 preceding *bolat*). The remaining 10 were observed after the word *Amol* in subject or broad focus utterances.

In Table 3, we show the 1-way ANOVA for log duration for the *Aanand* utterances. The stressed syllable in *Aanand* is the initial syllable. When 1-way ANOVA measurements are carried out separately for each of the syllables of *Aanand*, we observe that it is indeed the initial syllable that undergoes more significant modification in narrow focus on the word, in contrast with the *Amol* utterance where the final syllable was most affected.

Intensity: We observe in Figure 5 that narrow focus on a word does affect the mean intensity of the word in all locations relative to broad focus. However the only consistent trend in post-focal intensity observed across speakers was the post-focal decrease on the verb (*bolat*), as is borne out by ANOVA results for mean intensity in Table 2. In the disyllabic and trisyllabic words, it was observed that all syllables are similarly affected, as seen in Figure 5, except possibly for an exaggerated effect on the final syllable in *barobar*.

F0: Figure 6 shows the distributions of *F0* max for the different words and focus conditions. One-way ANOVA results on both *F0* measures (max and span) are presented in Table 2. We note that the maximum *F0* and the *F0* span are significantly higher for narrow focus at object (*aai*) and verb (*bolat*) locations with reference to neutral focus. The *F0* cues are not so clear for the subject focus (*Amol*) or function word focus (*barobar*). Post-focus on the verb (*bolat*) and on the function word (*barobar*) influences both the *F0* measures, with significant decrease in value compared with neutral focus. Post-focus on the object (*aai*), on the other hand, is marked only with a decrease in *F0* span.

Figure 4: ***Normalized log duration distributions across speakers and utterances at word-level (top) and syllable-level (bottom). 1: in broad focus, 2: on narrow focus, 3:pre-focal, 4:post-focal***

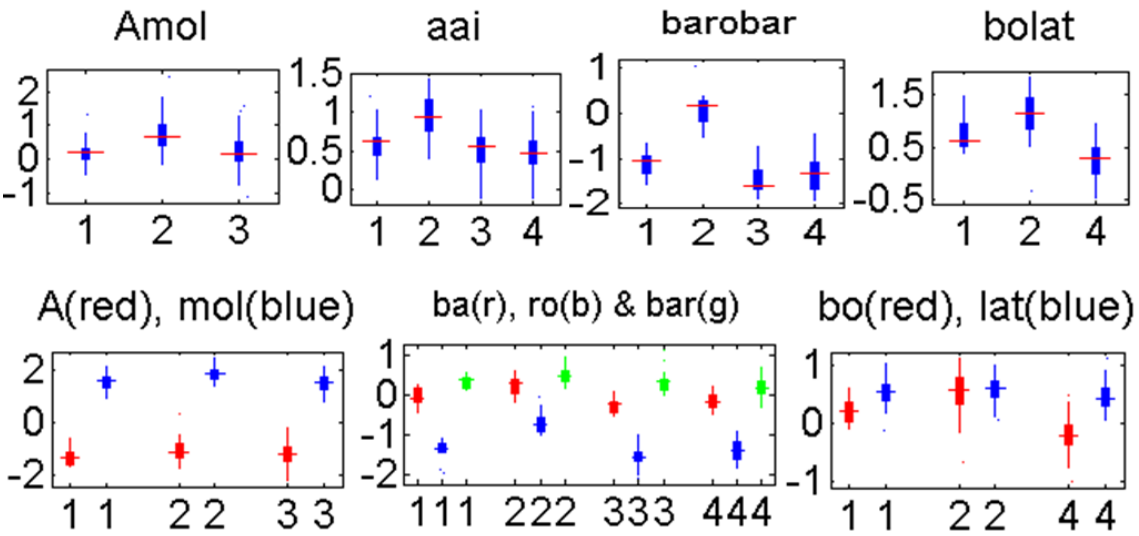


Figure 5: ***Normalized mean intensity distributions across speakers and utterances at word-level (top) and syllable-level (bottom). 1: in broad focus, 2: on narrow focus, 3: pre-focal, 4: post-focal***

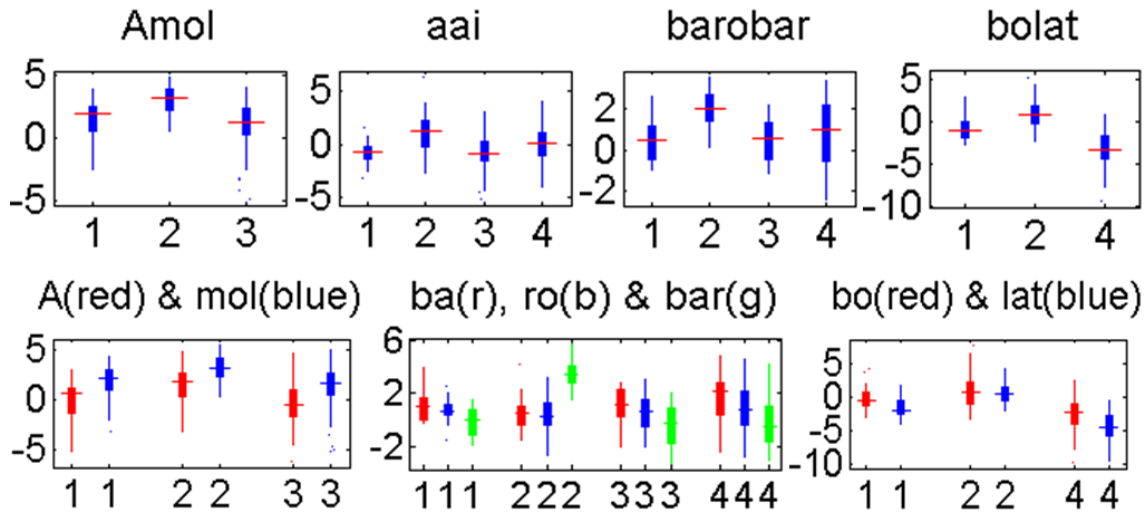


Figure 6: ***Max F0 distributions (semitones with respect to utterance mean F0) across speakers and utterances at word-level (1:broad focus, 2:on narrow focus, 3:pre-focal, 4:post-focal)***

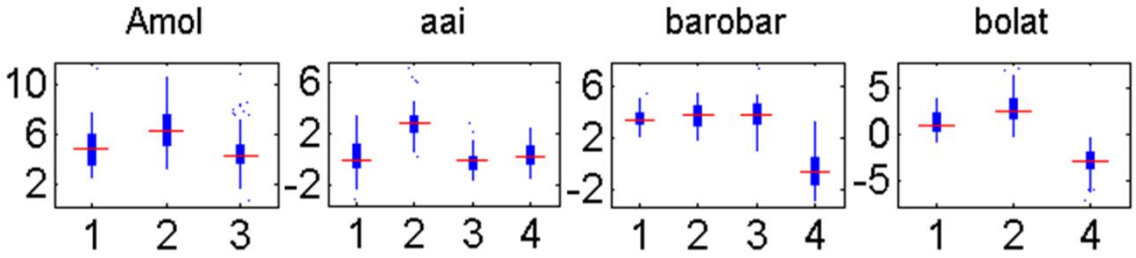


Table 2: ***F values (p values in parentheses) for different focus words, calculated with respect to broad focus; p-values less than 0.001 set to zero***

Parameter	**On focus**				**Post focus**		
	Amol	***aai***	***barobar***	***bolat***	***aai***	***barobar***	***bolat***
Logdur	20.3 **(0)**	28.2 **(0)**	88.82 **(0)**	17.4 **(0)**	3.2 **(0.07)**	3.43 **(0.07)**	43.6 **(0)**
I (max)	11 **(0.001)**	17.3 **(0)**	40.95 **(0)**	9.8 **(0.002)**	0.6 **(0.43)**	0.5 **(0.48)**	25.2 **(0)**
I (mean)	22.1 **(0)**	18.3 **(0)**	14.35 **(0)**	16.6 **(0)**	3.8 **(0.05)**	0.28 **(0.6)**	33.2 **(0)**
F0 (Max)	10 **(0.002)**	62.1 **(0)**	0.23 **(0.63)**	17.5 **(0)**	0.4 **(0.5)**	55.84 **(0)**	206.0 **(0)**
F0 (Span)	0.43 **(0.5)**	50.5 **(0)**	2.4 **(0.13)**	16.1 **(0)**	21.0 **(0)**	169.5 **(0)**	201.6 **(0)**

Table 3: ***F and p values for duration with respect to broad focus duration for 2 distinct subject focus words (Amol, Aanand) and their syllables; p-values less than 0.001 set to zero.***

"*Aanand*" utterances			**"*Amol*" utterances**		
Syll/word	**F**	**p**	**Syll/word**	**F**	**p**
Aanand	14.03	0	*Amol*	20.28	0
Aa	9.6	0.003	*A*	6.78	0.011
nand	4.48	0.04	*mol*	18.63	0

Table 4: ***Mean accent command amplitudes for the syllables of the target sentence, with the values for the focused words set in bold type. Row headings indicate focus condition***

Syllable	*Amol*	*aai*	*barobar*	*bolat*	Broad
A	0.00	0.00	0.00	0.00	0.00
mol	**0.46**	0.45	0.37	0.36	0.45
aai	0.12	**0.37**	0.32	0.12	0.18
ba	0.07	0.33	0.03	0.16	0.18
ro	0.07	0.26	0.20	0.26	0.28
bar	0.08	0.18	**0.41**	0.48	0.50
bo	0.05	0.19	0.14	0.22	0.29
lat	0.03	0.02	0.06	**0.49**	0.38
ho	0.02	0.03	0.00	0.04	0.14
ta	0.00	0.00	0.00	0.00	0.00

Figure 3 presents results of the Fujisaki analysis for speaker AK's target sentence for the following conditions: (1) subject (2) object (3) function word (4) verb (5) broad. The narrow focus conditions are non-contrastive except for the function word focus. Each one of the five panels displays, from the top to the bottom: the speech waveform, the *F0* contour (extracted and modeled versions), and the phrase and accent commands obtained from the model. The syllable segmentation is indicated by the dotted vertical lines. Marathi syllable texts are provided in a Romanized transcription.

As can be seen, *F0* contours differ clearly across all five conditions. Except for post-focal words we indeed observe a well-defined hammock-like shape or F0 rise on each word as reported previously for several Indian languages as reviewed in Section 1. In the Fujisaki model representation of the *F0* contour, this shape results from accent commands which are aligned with the right edge of the constituent words. In the case of *aai* narrowly focused, the accent command actually extends into *barobar*. This continuation is even more striking in the broad focus case: A relatively weak accent command on *aai* is further boosted by a second command associated with the last syllable of *barobar* (Rao et al., 2016). This suggests that *aai barobar* is treated as a prosodic unit. AK, like a few other subjects, further emphasizes narrow focus on *bolat*, by introducing a short pause before the focused constituent. Further, as may be expected from the language characteristic hammock shaping at word level, the significant *F0* increases in the disyllabic words were observed to take effect in the final syllables (*-lat*, *-mol*) as supported by the alignment of the accent command corresponding to each of these words.

If we compare the four narrowly focused conditions, the main difference is actually seen on the accent commands associated with the post-focal items. With *bolat* in focus the associated accent command is boosted, whereas the command becomes deleted or very weak when narrow focus is placed on any of the preceding words. In contrast, the accent command aligned with *Amol* is never suppressed. This suggests that narrow focus only has a marginal effect on pre-focal items, but suppresses pitch gestures on the post-focal ones.

Table 4 displays means of accent command amplitude *Aa* for each syllable in the five focus conditions. The non-existence of an associated accent command is taken into account with an *Aa* of zero. As can be seen, *Aa* is universally low for post-focal items (set in grey). With narrow focus on a word, the final syllable has the prominent accent. A Kruskal-Wallis test for independent samples shows that *Aa* assigned with the critical words is significantly different for *aai* and *bolat* depending on whether the focus condition is narrow or broad focus ($p < 0.001$) whereas the significance is lower for *Amol* and *barobar* ($p < 0.007$).

3.4 Discussion and implications for perception testing

Based on the analyses just conducted as well as the visual inspection of the variation of acoustic attributes across utterances, a summary of the salient phonetic correlates of focus for each location in the utterance (S, O, F and V) is presented in Table 5. Although our statistical analyses for post-focal effects is reported for the pooled data, visual observations across speaker datasets indicated that the verb was affected similarly when it was post focus regardless of which previous word was in narrow focus.

We note that focus in a given sentence location typically influences, both, local changes in acoustic parameters associated with the word itself and acoustic changes in the segment of the utterance following it. Narrow focus in all locations is associated with an increase in duration (of the stressed syllable) and mean intensity of the focus word. As for pitch cues, these are present in narrow focus in the object and verb locations only, in the form of increases in *F0* maximum and *F0* span as well as in the accent command amplitude. In object focus, we observe a lot of variability in realization in *F0* correlates across speakers and instances but, overall, the accent command is completely aligned with *aai* unlike in all other focus conditions where *aai* and *barobar* form a unit with the stronger *F0* excursion (accent) on the final syllable of *barobar* (a function word). Further, narrow focus in all non-final locations is marked by the following post-focal cues: (i) *F0* span (and accent command amplitude) compression in all post-focal words. (ii) compression in duration, intensity and *F0* span of the verb.

Table 5. ***Summary of local and global phonetic correlates of focus for broad focus and different narrow focus locations in the utterance "Amol aai barobar bolat hota"***

Focus word	**On-focus cues**	**Post-focal cues**
Broad	Reference (every word has an *F0* accent on the final syllable)	-
Subject (*Amol*)	Duration: increase in the final syllable *mol* Intensity: mean intensity increase on word *F0*: no change in *F0* span or in *Aa*; only an increase in *F0* max on word	Duration: decrease in the syllable *bo* Intensity: mean intensity decrease on *bolat* *F0*: span compression (*aai, barobar, bolat*)
Object (*aai*)	Duration: increase in syllable *aai* Intensity: mean intensity increase in *aai* *F0*: *Aa* increase aligned with *aai*, increase in *F0* max and span	Duration: decrease in the syllable *bo* Intensity: mean intensity decrease in *bolat* *F0*: span compression (*barobar, bolat*)
Function word (*barobar*)	Duration: increase in syllables (*ba, ro*) Intensity: mean intensity increase in word *F0*: accents on both *aai* and *barobar*, no change in *F0* max or span.	Duration: decrease in the syllable *bo* Intensity: mean intensity decrease in *bolat* *F0*: span compression (*barobar, bolat*)
Verb (*bolat*)	Duration: increase in syllable (*bo*) Intensity: mean intensity increase in word *F0*: *Aa* increase, changes in *F0* max and span (*lat*)	-

Given our observations on the acoustic correlates for the different focus locations, we see that each narrow focus location could be marked with possibly multiple perceptual cues. These comprise of on-focus and post-focal cues, each of which is realised by variations in one or more of the *F0*, duration and intensity features of the corresponding words. Although it is of interest to examine the relative importance of the different perceptual cues in each of the different focus conditions, we see that it is a challenging task in general given the presence of the multiple concurrent cues across the utterance. We note however that verb focus is characterised by strictly local cues due to the absence of post-focal segments. This makes it a suitable test case to examine the relative importance of the different acoustic correlates in the production and perception of narrow focus.

While the significant acoustic correlates of narrow focus on the verb are evident in the comparison of distributions as in Table 2, further summarised in Table 5, speaker-specific observations of the acoustic changes on the verb can be more instructive. We consider the following two acoustic features: the duration of the first syllable *bo* as a fraction of the corresponding sentence duration (to compensate for speaking rate), and the amplitude of the Fujisaki model accent command aligned with the final syllable *lat*. We employ scatter plots as graphical means to examine the relationship between the two acoustic features in the narrow and broad focus conditions separately. Figure 7 shows the two scatter plots of

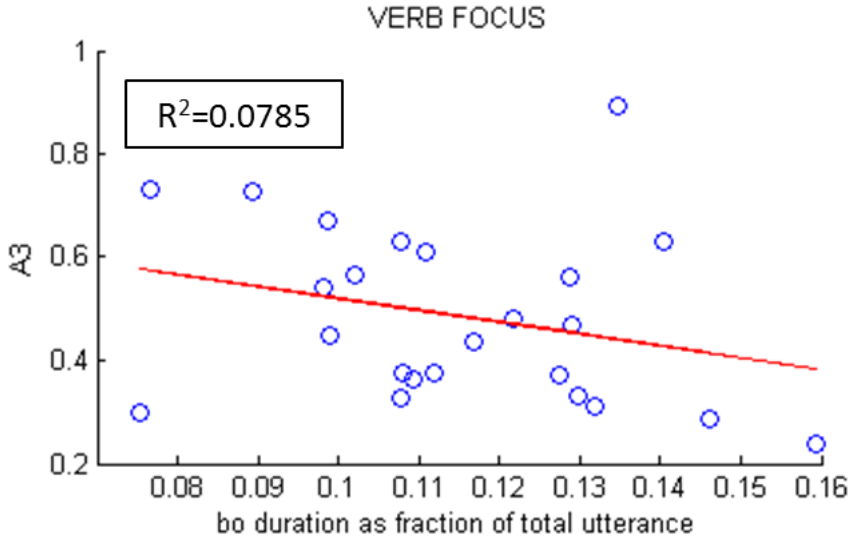
the individual measurements of the duration and the accent command amplitude as they co-occur on the verb (*bolat*) in each speaker-instance as obtained from across the set of utterances (2 instances each of non-contrastive narrow focus and of broad focus) across the 12 speakers. Similarly, Figure 8 plots the individual measurements of duration of syllable *bo* and mean intensity of the word. We observe that the overall spread of the data points is higher in the narrow focus condition compared to broad focus. We also separately noted that the 3 extreme points in the verb focus scatter plots (i.e. points with duration value less than 0.095 s in Figures 7(a) and 8(a)) were marked by pre-focal pauses, unlike all other points.

The scatter plots provide a visual representation of how the acoustic features covary. Next, we apply linear regression to study the quality of a linear fit between the two acoustic parameters on the data in each scatter plot. The goodness-of-fit parameter, R^2, also known as the coefficient of determination that varies from 0 to 1, indicates a superior linear fit to the verb focus data rather than to the corresponding broad focus data in the Figures 7 and 8. We observe from Figure 7 that the duration and accent command amplitude are negatively correlated in the verb focus condition but only weakly related in the broad focus condition. That is, in verb focus, speakers tend to strike a trade-off between duration increase and the accent command amplitude increase with some exercising more of duration change and others more of pitch change. Similarly, we observe in Figure 8 that duration and mean intensity are negatively correlated in verb focus and relatively uncorrelated in broad focus.

The above observations on elicited data suggest that there is an interaction between the distinct acoustic properties in the production of verb focus, and this phenomenon could influence the perception of verb focus. Perception testing could further facilitate more quantitative observations in terms of the relative importance of the different cues via the use of systematically varied synthetic stimuli, where a measure of the perceived prominence of each stimulus is recorded via listening tests with native speakers. This is similar to the classic approach of Abramson and Lisker (1985) who tested for the relative power of *F0* shift versus voice onset timing in the perception of stop consonant voicing. In the next section, we present our perception testing method and outcomes.

Figure 7: ***Scatter plots of accent command amplitude on the syllable 'lat' (A3) vs relative duration of the syllable 'bo' (in 24 non-contrastive utterances across 12 speakers) with linear regression fit and goodness-of-fit coefficient R^2 for (a) verb focus and (b) broad focus***

a.



b.

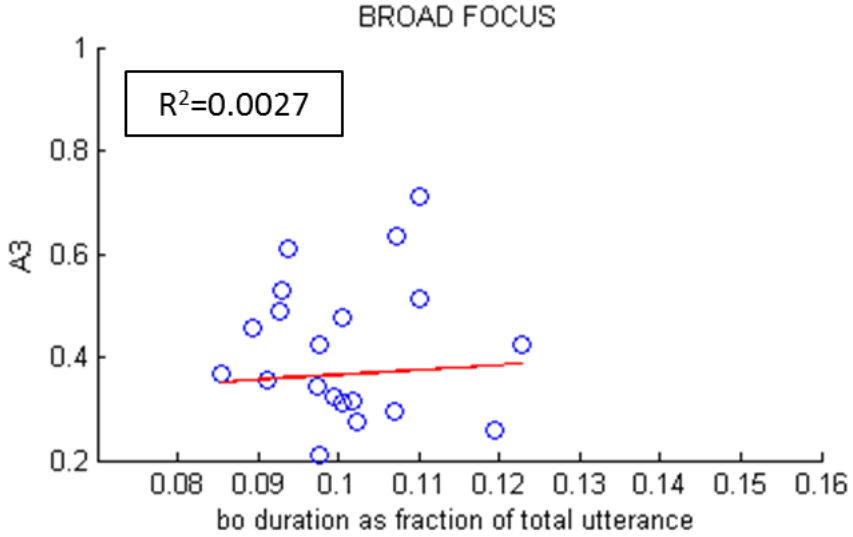
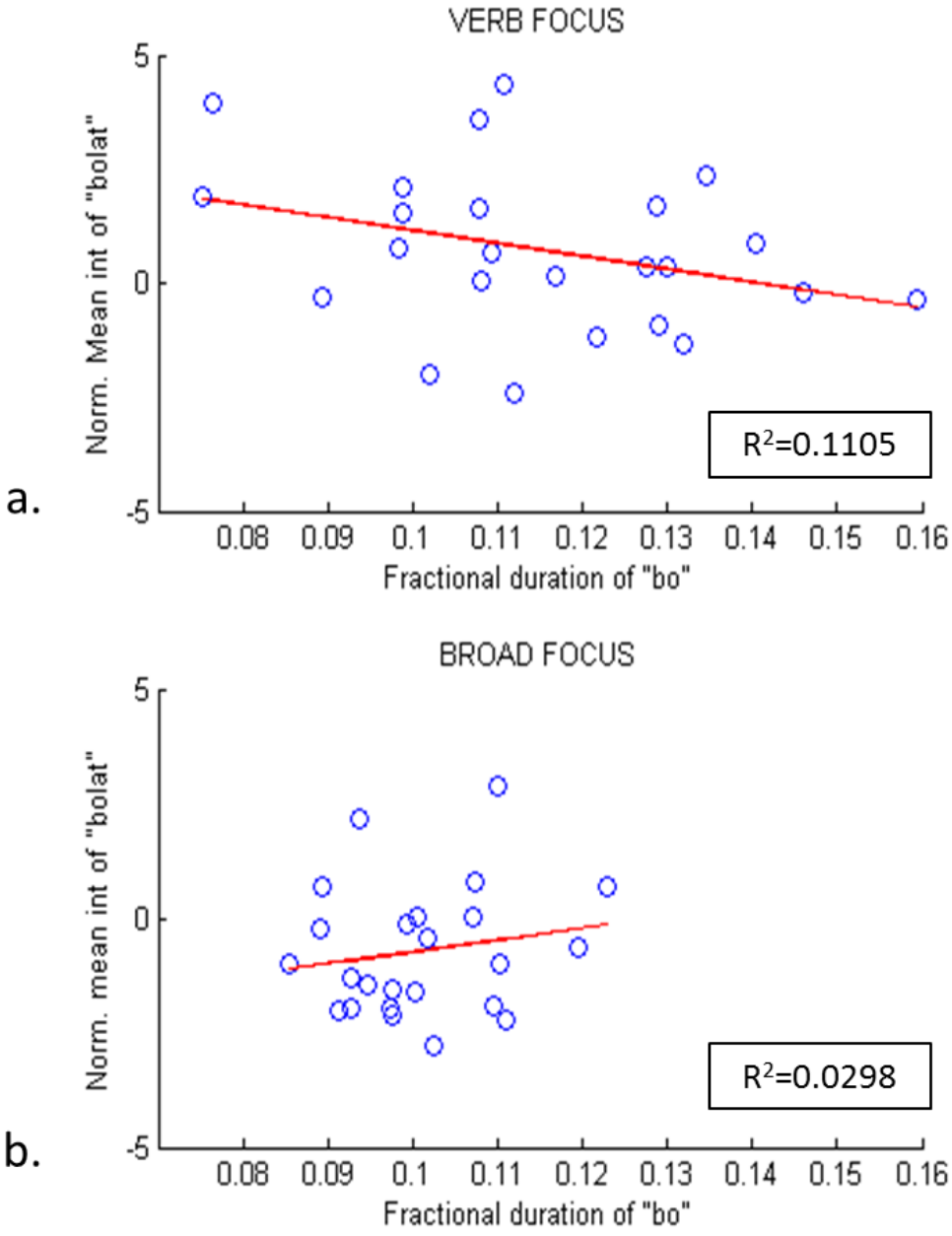


Figure 8: ***Scatter plots of normalized mean intensity of word 'bolat' vs relative duration of the syllable 'bo' (in 24 non-contrastive utterances across 12 speakers) with linear regression fit and goodness-of-fit coefficient R^2 for (a) verb focus and (b) broad focus***



4.0 Perception testing of verb focus

The production data on verb focus showed that speakers use a variety of combinations of on-focus duration increase (including that manifested via the occurrence of a pause), word intensity increase and *F0* span increase. These are represented by the relative duration of the initial syllable of the word, the mean intensity of the word and the Fujisaki model accent command on the final syllable of the word. The range of variation of each attribute in the production data seen in the Figures 7 and 8, helps us

select combinations of the parameters for the generation of realistic stimuli simulating verb focus. We pick a set of points that sample the observed space of duration, accent command amplitude and mean intensity. Given the more common mention of *F0* and duration features in the context of focus in previous studies on Indian languages, we sample these feature ranges more closely compared to that of intensity. We choose 4 approximately equally spaced values of syllable relative duration and 3 separate values of accent command amplitude to obtain 12 combinations. In order to keep the total number of stimuli under test moderate, we use only 2 different mean intensity levels to eventually obtain 24 synthetic stimuli. The selected two mean intensity levels are the average of the set of the natural broad focus utterances and the average of the natural verb focus utterances in Figures 8 (a) and (b) respectively. In this section, we describe the synthesis of the stimuli followed by the perception testing procedure and discussion of the outcome.

4.1 Synthesis of stimuli

From amongst the data set of natural speech, a prototypical broad focus utterance by a male speaker is chosen as the base instance upon which parameter modifications can be carried out. At the outset, the natural speech stimulus is replaced by its *F0*-stylised version as follows. A close-copy stylization of the natural *F0* contour is obtained via Fujisaki modelling in the FujiParaEditor (Mixdorff, 2009), and the utterance is resynthesized by applying the needed F0 contour changes to the waveform using the TD-PSOLA function in PRAAT. This step serves to eliminate micro variations in F0 that are perceptually irrelevant and obtains for us a base version on which the acoustic parameter modifications can be more robustly realized. Figure 9 shows the stages in the *F0* contour stylization of the broad focus utterance.

We then manipulate the F0 and duration attributes of the stylized broad focus utterance, keeping intensity unchanged, to obtain each of the 12 different stimuli. Next, the corresponding 12 modified intensity stimuli are obtained by applying the needed intensity scaling to each of these. The F0 and duration modifications are realized by changes in the relative duration of stressed syllable of the focus word, and in the accent command amplitude, *Aa*, of the focus word. Durational modifications are carried out via the duration tier and TD-PSOLA time scaling in PRAAT, using it to change the syllable durations of the prototype stylized broad focus utterance. After durational modifications are performed, the *F0* contour of the duration-modified utterance is manipulated. Fujisaki modeling of the *F0* contour followed by resynthesis with the desired accent command amplitude is carried out in the FujiParaEditor (Mixdorff, 2009).

The different steps involved in the duration-F0 manipulation are illustrated in Figure 10 where the targeted modification is an increase in the duration of the syllable *bo*, with an increase in the accent command amplitude associated with the word *bolat*. In changing the duration of *bo*, we take into account the fractional value of the same with respect to the utterance duration. For the manipulated utterance, we supply the duration scaling parameter such that the *bo* segment has the same fractional value relative to the overall duration of the modified utterance. By scaling up the intensity of the verb segment in each of the 12 stimuli, we obtain the additional set of 12 intensity-modified stimuli for the perception test. The amount of scaling up is equal to the average increase in mean intensity in the verb focus condition as compared to the broad focus condition.

Finally, three dummy stimuli are added to the perception test. These correspond to a subject, object or function-word focus utterance chosen from the same speaker's dataset and further replaced by close-copy *F0* stylised versions. The dummy stimuli serve as decoys to confirm that the listener is paying attention to the task of marking the focus condition.

Figure 9: ***Different stages in F0 stylization of the utterance (from top to bottom: the original utterance, the extracted Fujisaki accent command parameters, and the F0 contour resynthesized from the model parameters)***

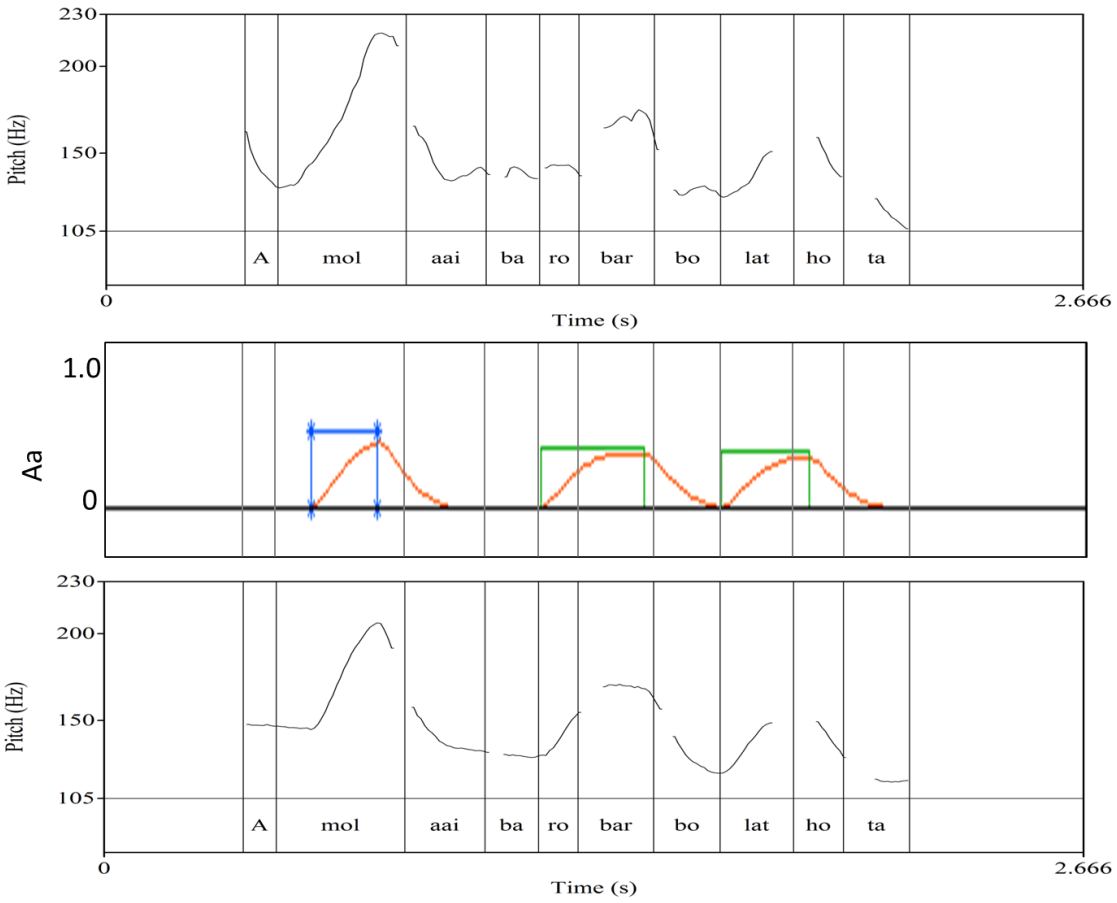
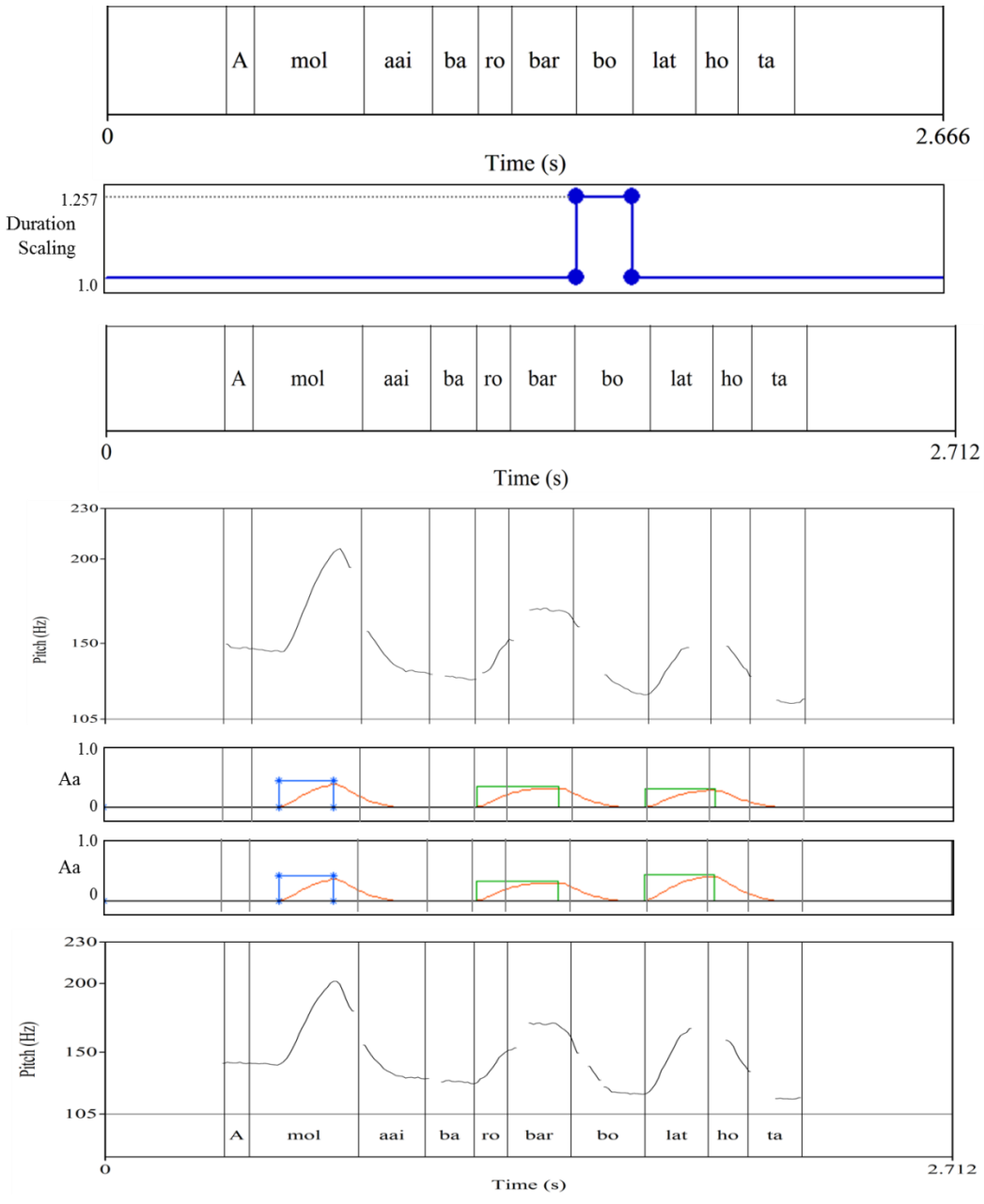


Figure 10: ***Different stages in the combined manipulation of duration and F0 contour of the utterance (from top to bottom: the original syllable segmentation, the desired duration scaling contour, the modified syllable segmentation, the extracted F0 contour of the duration modified utterance, the corresponding Fujisaki model accent commands, the manipulated accent commands, and the resulting final manipulated F0 contour)***



4.2 Perception test

The listeners for the perception experiments are 10 native Marathi speakers living in Mumbai who have not participated in our study so far as either speakers or listeners. To introduce the listeners to the concept of focus and prominence, we played out examples of questions and answers corresponding to a different SOV sentence compared to the one used in the experiment. For the perception experiment itself, the listeners were given a 5-way forced choice task in which they indicated the focus location (i.e. *Amol, aai, barobar, bolat*, or none in case of no perceivable prominence) via the GUI shown in Figure 1. The GUI allows for them to indicate the prominent word via a dropdown menu and the perception of naturalness, i.e. the box can be unchecked to indicate an unnatural stimulus. The stimuli are played out via headphones whenever they click on the corresponding audio buttons in the GUI with repeated listening permitted. The 24 manipulated stimuli, described in the previous section, are presented to the listeners in 3 randomized sessions, with the three dummy stimuli added in each session. It took a listener about 15 minutes on the average to complete a single session with the 27 stimuli. It was observed that the dummy stimuli were correctly identified by all the listeners in all trials.

Figure 11 presents the results of the perception test with the 12 manipulated broad focus stimuli without intensity modification. Each test stimulus, characterised by fractional duration (relative to utterance length) of the syllable *bo* and *A3*, the accent command amplitude of the syllable *lat*, is indicated by a circle at the corresponding point on the graph. The degree of prominence for each stimulus, plotted on the y-axis, is estimated by the number of times it is correctly recognised as narrow focus on the verb, across all the 3 trials for all listeners (a total of 30 observations per stimulus). The points corresponding to a fixed accent command amplitude (*A3*) value are connected by a curve. Similarly, Figure 12 presents the degree of prominence for each of the 12 intensity-modulated stimuli.

In Figure 11, we observe a relatively steep increase in perceived prominence with duration increase at mid values of duration for any given value of *A3*. The curves are all relatively flat at the low and high ends of the duration scale. Further, the three distinct curves for the different *A3* converge at the extremes but are more separated at the mid values of duration. This indicates that when duration is ambiguous (corresponding to the mid-range values), the perceived focus varies appreciably with *F0* modifications. However, at the low and high values of duration, we see that listeners make a clear choice between verb focus and non-verb focus conditions irrespective of the strength of the F0 gesture. It was observed that broad focus was by far the most chosen category whenever verb focus was not picked. Given that broad and verb focus share the same pre-focus prosodic representation, the manipulation of on-focus local prosodic parameters is expected to restrict the perceived choices to these two conditions. The remaining conditions, constituting less than a fifth of the total trials, were predominantly subject focus.

Thus we see that while perceived prominence of the verb depends on, both, its duration and accent command strength, focus is close to being categorically interpreted based on duration. Figure 12 presents perceived prominence with changing duration and accent command amplitude for the intensity-modulated stimuli. We see that that all the perceived prominence values have shifted upwards to some extent with the intensity modification, relative to the same duration-pitch values of Figure 11. The three curves also largely retain their characteristic of steep change in perceived prominence with changes in duration, relatively weakly influenced by the accent command amplitude value. We would also like to note that although the listeners were instructed to indicate unnatural sounding stimuli by unchecking the box on the GUI as in Figure 2, there were no such markings recorded indicating either that all stimuli were perceived as natural by all listeners or that this particular instruction went unheeded by some.

Figure 11: ***Fractional duration of 'bo' versus degree of verb prominence judgement for different values of A3 across 3 trials with 10 listeners; no intensity manipulations used***

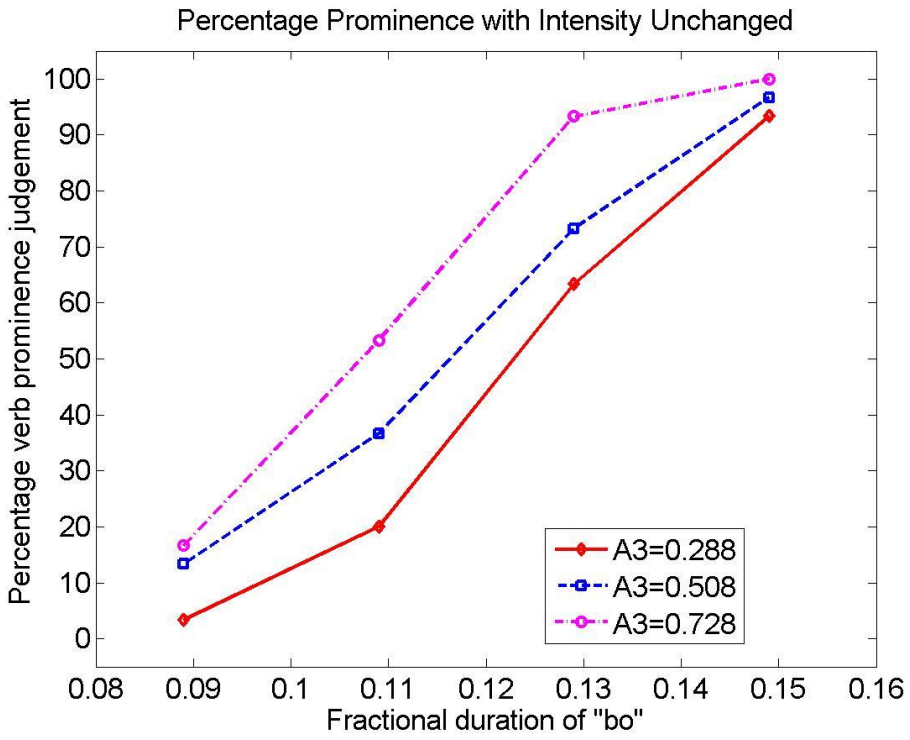
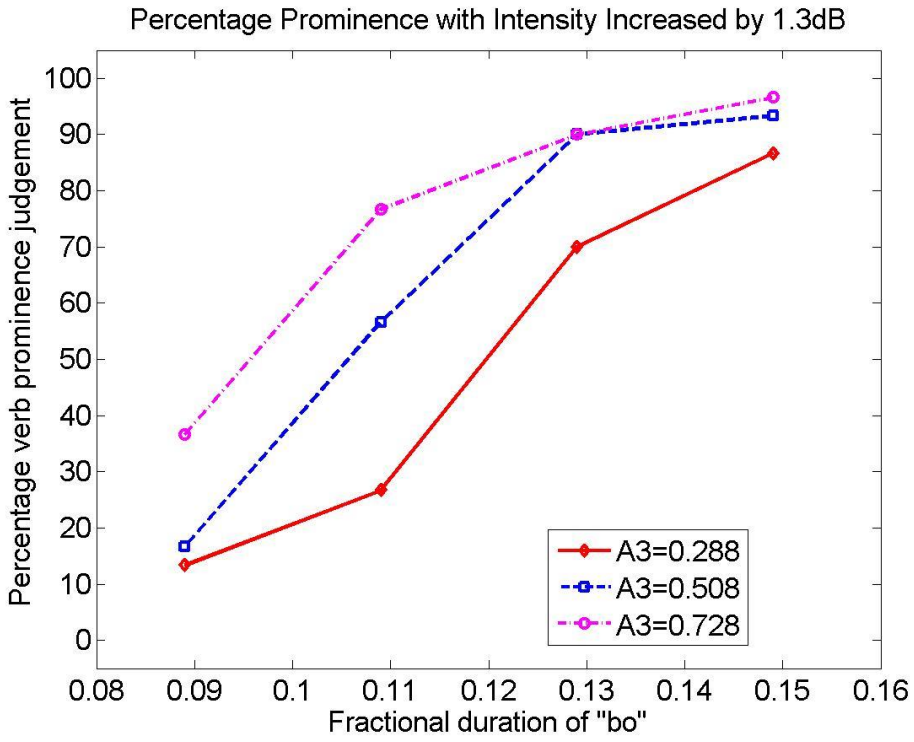


Figure 12: ***Fractional duration of 'bo' versus degree of verb prominence judgement for different values of A3 across 3 trials with 10 listeners; normalized mean intensity of 'bolat' increased by 1.3 dB***



5. Discussion and Conclusions

The phonetic correlates of focus in Marathi were investigated in the context of SOV utterances using production data together with perception testing with its scope limited to verb focus. Marathi speakers can use particles, word order or prosody to communicate focus conditions. It was observed that the purely prosodic means of expressing focus studied in this work come naturally to native speakers. No prosodic differences are observed between contrastive and non-contrastive forms; neither is the difference perceived by listeners when the lexical difference corresponding to the initial word "no" is eliminated. Considering that we had a limited dataset from 12 speakers and one reference sentence, the present work can be considered preliminary but providing a clear basis to build upon further.

From production data, we conclude that narrow focus on a word typically influences both local and post-focal prosodic features. The local cues consistent across sentential position are increases in word duration and mean intensity with respect to the values in the corresponding broad focus segment. Marathi is not known to exhibit lexical stress. However based on a previously proposed definition of the stressed syllable in terms of syllable weight, we observe that the duration change occurs primarily in the stressed syllable of the word. Further, there is typically a reduction in *F0* span, and correspondingly a decrease in the Fujisaki model accent command, of all post-focal words.

On-focus *F0* cues appear in the object and verb locations only. Here the final syllable *F0* rise that occurs in all content words in broad focus becomes hyper-articulated when the word is in narrow focus. Thus, the on-focus cues of *F0* accent and the duration elongation actually occur in different syllables of the word if the stressed syllable is not the final syllable. Subject focus is not discriminated from broad focus in the *F0* span, probably owing to the topic marker effect in the latter. The function word forms a prosodic word with the object in the broad focus utterance. With narrow focus on the object however, the most consistent cue is the *F0* accent alignment, shifting from the final syllable of the prosodic word

to appear on the object word. In this case, the remaining *F0* magnitude cues may or may not be realized. Narrow focus on the verb, which is in sentence-final position, is indicated purely by local prosodic correlates. The verb *bolat* undergoes duration elongation (of the syllable *bo*), and *F0* span or accent command amplitude increase on the final syllable *lat*. Individual speakers display differences in the relative strengths of the syllable duration and F0 accent cues possibly from personal habit. Further, we did not measure the precise alignments of F0 changes with respect to the accented syllables (Gussenhoven, 2004); this can be studied in future via the obtained Fujisaki modeled accent commands.

The above conclusions, related to Marathi on-focus and post-focal cues at the word-level, match with what is known about the phonetic correlates of narrow focus in English using elicited speech (Breen et al., 2010; Xu & Xu, 2005). Further, they are largely in line with previous studies on production of focus in Hindi, as reviewed in Section 1. Like Hindi, Marathi shows an *F0* rise on the final syllable in every content word in broad focus utterances. In the present work, we observed that this behaviour transferred to the final syllable of the prosodic word formed by the function word attaching itself to the preceding object. Only when the object itself is in focus does an additional pitch accent right aligned with the object appear. It was also clearly demonstrated that in general the duration variation and *F0* variation affect different syllables of the word under narrow focus depending on syllable weight and position.

The perceptual experiment proves that native Marathi speakers are indeed capable of perceiving focus on the verb for Marathi SOV sentences, and the contribution of each acoustic feature is observable via systematic manipulation of the synthesized stimuli. There have been relatively few past studies using perception tests to study the acoustic correlates of perceived prominence. Most of these have been restricted in scope to the role of intonation, and more specifically, to the local *F0* characteristics of the target word. One of these is that of Ladd and Morton (1997) who considered the effect of changing pitch range on focus perception via manipulated stimuli to find that the normal versus emphatic distinction is categorically interpreted as pitch range is gradually increased. Vainio and Jarvikivi (2006), in the course of examining the interaction between word order and acoustic cues, observed a trade-off between intensity and pitch height in the perception of focus. In the present work, we have attempted to understand the interaction of the distinct acoustic attributes (duration, *F0* and intensity) in the perception of narrow focus. Perception testing was carried out with manipulated stimuli in the space between broad focus and verb focus, which differ only in terms of the acoustic properties of the verb segment. We conclude from the perception test that prominence is categorically interpreted with gradual changes in segment relative duration. We infer that duration is the dominant cue based on the observation that values of the segment relative duration are interpreted in a manner that is little affected by the accompanying F0 cues or intensity cues. Only when duration cues are ambiguous, does the pitch accent influence the categorization. An explanation for this may lie in the normal pitch-rise characteristic of the content words in Marathi (similar to the "repeated rising contours" of South Asian languages in the words of Khan (2016)), making pitch a less dependable functional cue for focus. The salience of duration in the prosody of focus may also be linked to the syllable-timed nature of the language.

Perception studies involving other sentential positions would be topics for future work. Such work would lead to the better understanding of the importance of post-focal cues versus on-focus cues. Further, given that native speakers of Marathi and other Indo-Aryan languages can also make use of particles to signal focus in Marathi, the interaction between focus particles and prosodic cues is of interest apart from the interaction with word ordering itself. While the present work considered *F0,* duration and intensity attributes, future research can involve the effects of prosodic focus on phonetic details at the segmental level (Cho, 2011) as well as on voice quality features such as spectral balance that have been shown to vary with change in focus condition in a number of languages (Rosenberg et

al., 2012). Finally, for any practical applications of such work, a larger study of Marathi prosody including phonetic cues to phrase boundaries apart from prominence is warranted for the different speaking styles (Barhate et al., 2016).

References

Abramson, A.S. and Lisker, L., (1985). Relative Power of Cues: F0 Shift Versus Voice Timing. *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, pp.25-33.

Andreeva, B., Barry, W. and Koreman, J., (2015) Cross-language and individual differences in the production and perception of syllabic prominence. Rhythm-typology revisited. Report of the DFG Project BA 737/10.

Barhate, S., Kshirsagar, S., Sanghvi, N., Sabu, K., Rao, P. and Bondale, N., (2016) Prosodic Features of Marathi News Reading Style. *In: Proceedings of IEEE TENCON 2016*, Singapore.

Bolinger, D. (1986). Intonation and its Parts: Melody in Spoken English. *Stanford University Press, Palo Alto.*

Boersma, Paul (2001). Praat, a system for doing phonetics by computer. *Glot International* **5:9/10**, 341-345.

Breen, M., Fedorenko, E., Wagner, M. and Gibson, E., (2010). Acoustic Correlates of Information Structure. *Language and Cognitive Processes,* 25(7), pp.1044-1098

Chen, Szu-wei, Wang, B. and Xu, Y. (2009). Closely Related Languages, Different Ways of Realizing Focus. *In: Proceedings of Interspeech*, Brighton.

Cho, T., Lee, Y., & Kim, S. (2011). Communicatively driven versus prosodically driven hyper-articulation in Korean. Journal of Phonetics, 39(3), 344-361.

Dyrud, L. O. (2001). Hindi-Urdu: Stress Accent or Non-Stress Accent?. Ph.D. thesis, University of North Dakota.

Dhongde, R. V. and Wali, K. (2013). Marathi. Vol. 13, *In: John Benjamins Publishing*.

Fery, C. and Gisbert Fanselow (2008). A Sentence-based Typology of Intonation Systems. *Ms., Universit¨at Potsdam*.

Féry, C. (2010). Indian Languages as Intonational Phrase Languages. *In: Problematizing Language Studies*, 2010, pp. 288–312.

Féry, C. (2013). Focus as Prosodic Alignment. *Natural Language & Linguistic Theory,* 31(3), pp.683-734.

Fry, Dennis B. (1958). Experiments in the Perception of Stress. *Language and speech,* 1(2), pp.126.

Fuchs, R. (2015). You're Not from Around Here, Are You?. *In Prosody and Language in Contact*, Springer Berlin Heidelberg, pp. 123-148.

Fujisaki, H. (1988). A Note on the Physiological and Physical Basis for the Phrase and Accent Components in the Voice Fundamental Frequency Contour. *In: (Fujimura, Osamu ed.) Vocal Physiology: Voice Production, Mechanisms and Functions*, New York (Vocal Fold Physiology; 2), pp. 347-355.

Genzel, S. and Kügler, F. (2010), The Prosodic Expression of Contrast in Hindi, *In Proceedings of 5th International Conference of Speech Prosody*, Chicago, USA, May 2010, pp. 1-4.

Gussenhoven, C., & Rietveld, T. (1988). Fundamental Frequency Declination in Dutch: Testing Three Hypotheses. *Journal of Phonetics*, 16, pp.355–369.

Gussenhoven, C., Repp, B. H., Rietveld, A., Rump, H. H., & Terken, J. (1997). The Perceptual Prominence of Fundamental Frequency Peaks. *Journal of the Acoustical Society of America*, 102(5), pp.3009–3022.

Gussenhoven, C. (2004). The Phonology of Tone and Intonation. Cambridge: Cambridge University Press.

Harnsberger, J. D. (1996). Towards an Intonational Phonology of Hindi, *In Proceedings of 5th Conference on Laboratory Phonology*, Northwestern University.

Harnsberger, J. D. (1999). The Role of Metrical Structure in Hindi Intonation. *In South Asian Language Analysis Roundtable*, University of Illinois, Urbana-Champaign.

Hayes, B. and Lahiri, A. (1991). Bengali Intonational Phonology. *Natural Language and Linguistic Theory*, 9(1), pp.47–96.

Ipek, C. (2011). Phonetic Realization of Focus with No On-focus Pitch Range Expansion in Turkish. *In Proceedings of the 17th International Congress of Phonetic Sciences,* Hong Kong: Department of Chinese, Translation and Linguistics, City University of Hong Kong, pp. 140-143.

Junghare, I. Y. (1988). Topic prominence and zero NP-anaphora in Marathi and Hindi. Languages and cultures. Berlin: Mouton De Gruyter, 309-328.

Khan, S. D. (2007). *Phrasing and Focus in Bengali*. Poster presented at the *International Congress of Phonetic Sciences Satellite Workshop on Intonational Phonology: Understudied or Fieldwork Languages*, Saarbrücken, Germany, August 2007.

Khan, S.D. (2016). The intonation of South Asian languages: towards a comparative analysis. Proceedings of Formal Approaches to South Asian Languages 6, 23-36.

Koreman, J., Dommelen W., Sikveland, R., Andreeva, B., Barry, W., Vainio, M., Aulanko, R. and Aaltonen, O. (2009). Cross-language Differences in the Production of Phrasal Prominence in Norwegian and German. *In Proceedings of the 10th Conference of Nordic Prosody*, pp. 139-150.

Ladd, D. R., and Morton, R. (1997). The Perception of Intonational Emphasis: Continuous or Categorical?. *Journal of Phonetics*, *25*(3), pp.313-342.

Lee, Y. C. and Xu, Y. (2010). Phonetic realization of contrastive focus in Korean", *In Proceedings of Speech Prosody*, Illinois.

Maxwell, O. (2010). Marking of Focus in Indian English of l1 Bengali Speakers. *In Proceedings of Speech science and Technology*, pp.58-61.

Mixdorff H. (2000). "A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters", *In Proceedings of ICASSP*, Istanbul, Turkey, 3, pp.1281-1284.

Mixdorff, H., Fujisaki, H. (1995): Production and Perception of Statement, Question and Non-Terminal Intonation in German. *In Proceedings of the ICPhS*, Stockholm, Schweden, 2, pp.410-413.

Mixdorff, H., Cossio-Mercado, C., Hönemann, A., Gurlekian, J., Evin, D. and Torres, H. (2015). Acoustic Correlates of Perceived Syllable Prominence in German. *In Proceedings of 16th Annual Conference of the International Speech Communication Association*, Dresden.

Mixdorff, H. and Widera, C. (2001). Perceived Prominence in terms of a Linguistically Motivated Quantitative Intonation Model. *In Interspeech*, pp. 403-406.

Mixdorff, H. (2009). FujiParaEditor, http://public.beuth-hochschule.de/~mixdorff/thesis/fujisaki.html, last retrieved in May, 2016.

Moon, R. (2002). A Comparison of the Acoustic Correlates of Focus in Indian English and American English. PhD dissertation, University of Florida.

Moore, R. R. (1965). A Study of Hindi Intonation. *(Unpublished dissertation)*, University of Michigan.

Ohala, Manjari (1986). A Search for the Phonetic Correlates of Hindi Stress. *South Asian languages: structure, convergence, and diglossia*, pp.81-92.

Pandey, P. (2015). Indian English Prosody. *In: Communicating with Asia*, Cambridge University Press, Cambridge.

Patil, U., Kentner, G., Gollrad, A., Kügler, F., Féry, C., & Vasishth, S. (2008). Focus, Word Order and Intonation in Hindi. *Journal of South Asian Linguistics*, 1(1).

Puri, V. (2013). Intonation in Indian English and Hindi Late and Simultaneous Bilinguals. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Rajendran, S. and Yegnanarayana, B. (1996). Word Boundary Hypothesization for Continuous Speech in Hindi based on F0 Patterns. *Speech Communication*, 18(1), pp. 21-46.

Rao, P., Mixdorff, H., Deshpande, I., Sanghvi, N., & Kshirsagar, S. (2016). A Quantitative Study of Focus Shift in Marathi. *In Proceedings of Speech Prosody*.

Rao, G. and Srichand, J. (1996). Word Boundary Detection using Pitch Variations. *In Proceedings of International Conference on Spoken Language Processing*, Philadelphia.

Rosen, K.N. (2005). Analysis of Speech Segment Duration with the Lognormal Distribution: A Basis for Unification and Comparison. *Journal of Phonetics*, 33, pp.411–426.

Rosenberg, A., & Hirschberg, J. (2009). Detecting Pitch Accents at the Word, Syllable and Vowel Level. In *Proceedings of human language technologies: The Annual Conference of the North American chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pp. 81-84.

Rosenberg, A., Cooper, E., Levitan, R. and Hirschberg, J., (2012). Cross-language prominence detection. *In Proceedings of Speech Prosody 2012.*

Skopeteas, S., & Féry, C. (2010). Effect of Narrow Focus on Tonal Realization in Georgian. *In Proceedings of speech prosody*, Chicago.

Swerts, M., Krahmer, E., & Avesani, C. (2002). Prosodic Marking of Information Status in Dutch and Italian: A Comparative Analysis. *Journal of Phonetics*, *30*(4), pp.629-654.

Terken, J. (1989). Reaction to C. Gussenhoven and A. Rietveld: Fundamental Frequency Declination in Dutch: Testing Three Hypotheses. *Journal of Phonetics*, 17, pp.357–364.

Terken, J., and Hermes, D. (2000). The Perception of Prosodic Prominence. *In (M. Horne Ed.), Prosody: Theory and Experiment*, pp. 89–127.

't Hart, J. Collier, R. and Cohen, A. (1990). A perceptual study of intonation. Cambridge University Press, Cambridge.

Vainio, M., Mixdorff, H., Järvikivi, J. and Werner, S. (2003). The Production and Perception of Focus in Finnish. *In Proceedings of ICPhS.*

Vainio, M., & Järvikivi, J. (2006). Tonal Features, Intensity, and Word Order in the Perception of Prominence. *Journal of Phonetics*, *34*(3), pp.319-342.

Xu, Y. and Xu, C.X. (2005). Phonetic Realization of Focus in English Declarative Intonation. *Journal of Phonetics*, *33*(2), pp.159-197.