Determination Of Acoustic Parameters Of Marathi Prosody

Submitted in partial fulfilment for the award of the degree

of

MASTER OF ENGINEERING

in

ELECTRONICS & TELECOMMUNICATION ENGINEERING

(Academic Year: 2015-16)

by

Shruti Rajendra Kshirsagar

University Registration Number: Thakur/209

Under the Guidance of

Dr. Bijith Marakarkandy

Associate Professor, TCET

University of Mumbai

Department of Electronics and Telecommunications (P.G.)

THAKUR TRUSTS

Zagdu Singh Charitable Trust's (Regd.)

THAKUR COLLEGE OF ENGINEERING & TECHNOLOGY

(Approved by AICTE, Govt. of Maharashtra & Affiliated to University of Mumbai)

(Accredited by National Board of Accreditation, New Delhi)* • ISO 9001 : 2008 Certified

* Accredited Programmes : • Computer Engineering • Electronics & Telecommunication Engineering • Information Technology (w.e.f.: 16-09-2011 for 3 years)

A - Block, Thakur Educational Campus, Shyamnarayan Thakur Marg, Thakur Village, Kandivali (E), Mumbai - 400 101.

Tel: 022-6730 8000/ 8106 / 8107 Telefax : 022-2846 1890 E-mail : tcet@thakureducation.org Website: www.tcetmumbai.in • www.thakureducation.org

Determination Of Acoustic Parameters Of Marathi Prosody

Submitted in partial fulfilment for the award of the degree

of

MASTER OF ENGINEERING
in
ELECTRONICS & TELECOMMUNICATION ENGINEERING
(Academic Year: 2015-16)

by

Shruti Rajendra Kshirsagar
University Registration Number: Thakur/209

Under the Guidance of

Dr. Bijith Marakarkandy
Associate Professor, TCET

University of Mumbai

Department of Electronics and Telecommunications (P.G.)

THAKUR TRUSTS

Zagdu Singh Charitable Trust's (Regd.)

THAKUR COLLEGE OF ENGINEERING & TECHNOLOGY

(Approved by AICTE, Govt. of Maharashtra & Affiliated to University of Mumbai)

(Accredited by National Board of Accreditation, New Delhi)* • ISO 9001 : 2008 Certified

* Accredited Programmes : • Computer Engineering • Electronics & Telecommunication Engineering • Information Technology (w.e.f.: 16-09-2011 for 3 years)

Shyamnarayan Thakur Marg, Thakur Village, Kandivali (E), Mumbai - 400 101.

Tel: 022-6730 8000 / 8106 / 8107 Telefax : 022-2846 1890 E - mail: tcet@thakureducation.org

Website : www.tcetmumbai.in • www.thakureducation.org

Spine/Side view of the Report –

M.E. (EXTC) SHRUTI RAJENDRA KSHIRSAGAR 2015-16

DISSERTATION APPROVAL CERTIFICATE

This is to certify that the dissertation entitled "**Determination Of Acoustic Parameters Of Marathi Prosody**" for M.E. in Electronics and Telecommunication Engineering submitted to University of Mumbai by "**Shruti Rajendra Kshirsagar**", a bonafide student of Thakur college of Engineering and Technology, Kandivali, Mumbai has been approved for the award of Master of Engineering Degree in Electronics and Telecommunication Engineering.

Signature:

Name- Dr. Bijith Markarkandy.
Associate Professor

Examiners:

1. Signature: …………………………
2. Name: …………………………….

1. Signature: …………………………
2. Name: …………………………….

Date:

Place:

CERTIFICATE

This is to certify that "**Shruti Rajendra Kshirsagar**" has satisfactorily carried out the dissertation work entitled "**Determination Of Acoustic Parameters Of Marathi Prodosy**" for the degree of Master of Engineering in Electronics and Telecommunication Engineering under University of Mumbai.

Dr. Bijith Markarkandy	Dr. B. K. Mishra
Project Guide	Principal
Associate Professor	HOD(PG)
Thakur College of Engineering & Technology	Thakur College of Engineering&Technology
Date	Date

Principal
Thakur College of Engineering and Technology
Kandivali, Mumbai

DECLARATION

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included. I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the institute and can also evoke penal action from the sources which have thus not been properly cited or from where proper permission has not been taken when needed.

(Signature)

Name: Shruti Rajendra Kshirsagar.

University Registration number: Thakur/209

Date:

Place:

SYNOPSIS OF PROJECT WORK

Name of Seminar:	Determination Of Acoustic Parameters Of Marathi Prosody.
Student's Name:	Shruti Rajendra Kshirsgar.
Class:	M.E.(Electronics and Telecommunication)
College:	Thakur College of Engineering and Technology
Semester:	IV
University Registration Number:	Thakur/ 209
Date of Registration:	04/08/2014
Name of Guide:	Dr. Bijith Markarkandy & Prof. Preeti Rao.

Semester	University Seat Number	Result (CGPA)
I	3210	8.36
II	2976	8.55

PUBLICATIONS ASSOCIATED WITH THIS THESIS

International Conference:

[1] P. Rao, H. Mixdorff, I. Deshpande, N. Sanghvi and S. Kshirsagar"Quantitative Study of focus shift in Marathi", Speech Prosody, 2016,Boston,U.S.A. (Paper accepted)

ABSTRACT

The purpose of this work is to study the effect of focus shift on the prosodic features for Marathi, which is a major Indian language.This study will help in building a prosody model which can be use in synthesis and speech recognition purpose.In this study, we aim to find the prosodic cues of focus in Marathi. This is, to the best of our knowledge, amongst the first works on the topic. Therefore we will follow the methods that have been employed in studying other languages.The eventual goal is to establish the features that are the most significant indicators of focus, in terms of acoustic features as well as their perception given the peculiarities of the language including the local pitch (F0) variation across a word and the possibly syllabic-weight based stress rules.

In our analysis we consider different focus locations and different focus widths for two short sentences.. F0 will be studied via the accent commands of the Fujisaki model. Contexts for narrowly focused items are either contrastive or non-contrastive. We will try to find out the difference for contrastive and non contrastive focus analysis. In news reading we consider those word which was spoken with emphasis by native news reader and which was perceived as prominence by native listeners and checked out the relation of these words with acoustic parameters

Our results show that narrow focus is marked by longer duration of the focused word, and a larger accent command in the focused word. Post-focal effects are observed for duration, intensity and *F0*. No differences were found between contrastive and non-contrastive focus. In news reading study, Spearman's correlation is having higher value for pitch and duration, but low value for intensity.Listenersrealibly perceived prominence by perception.topic marker effect seen on focus word.

The conclusion of this study is that prominence in Marathi speech is mainly a function of pitch rather than duration and intensity. Duration is also important factor because speakers tend to extend the duration when they want to give new information (which we called as focus or prominence). But loudness or intensity has the least significant role when it comes to signaling focus. We have yet to examine the relative perceptual contributions of the acoustic correlates of focus. To that end we intend to carry out perception experiments using resynthesized speech. In future research we will also concentrate on additional feature such as measurements concerning pitch and intensity, and on classification with a classifier such as an artificial neural network and a feed forward neural network.

Acknowledgement

I would like to take the opportunity to express my heartfelt gratitude to the people whose help and co-ordination has made this project a success. I sincerely express my deep sense of gratitude to my guide **Dr. Bijith Markarkandy** (Associate Professor), **Prof. Preeti Rao** and **Prof. Hansberg Mixdorff** for their valuable guidance, encouragement and timely need given to me throughout the course of this work.

I would like to thanks my co-author **Mr. Niramay Sanghavi** and **Mr. Ishan Deshpande** for their valuble work part and suggestions in this thesis.Alldaplab member of IIT, Bombay for their valuble suggestions and support time to time

I would like to thank my Honourable Principal **Dr. B. K**. **Mishra** for the conductive environment in the institution. I would like to express my heartfelt gratitude to all the teachers and staff members of Electronics and Telecommunication Engineering department of TCET for their full support.

My deep sense of gratitude to **Ms. Sujata Kulkarni** (ME Coordinator) *&* **Dr. Vinitkumar Dongre** (HOD,EXTC) for their constant support and guidance throughout the academic year.

I am grateful to the library staff of TCET for the numerous books, magazines made available for handy reference and use of internet facility.

Lastly, I am also indebted to all those who have indirectly contributed in making this project successful.

(Shruti Rajendra Kshirsagar)

Contents

List of Figures

List of Tables

Abbreviations and Symbols

SOV	Subject Object Verb
AIR	All India Radio
V	Vowel
CV	Consonant Vowel
CVC	Consonant Vowel Consonant
F0	Pitch
ANOVA	Analysis of variance
AMMA	Amol Marathi
AAMA	Aanand Marathi

Chapter 1. Introduction

1.1 Preview

Prosody

Prosody is the study of the tune and rhythm of speech and how these features contribute to meaning. Prosody is the study of those aspects of speech that typically apply to a level above that of the individual phoneme and very often to sequences of words (in prosodic phrases). Features above the level of the phoneme (or "segment") are referred to as suprasegmental. A phonetic study of prosody is a study of the suprasegmental features of speech.

At the phonetic level, prosody is characterized by:-

- Vocal pitch (fundamental frequency)
- Loudness (acoustic intensity)
- Rhythm (phoneme and syllable duration)

Phonetic studies of prosody often concentrate on measuring these characteristics. Prosody has been studied from numerous perspectives by people belonging to differing linguistic schools. There has been great diversity of approaches to prosody. Different approaches examine prosody from the perspectives of grammar, discourse, pragmatics, phonetics, phonology and conveying of emotions.

Speech contains various levels of information that can be described as:-

- Linguistic - direct expression of meaning.
- Paralinguistic - may indicate attitude or membership of a speech community.
- Non-linguistic - may indicate something about a speaker's vocal physiology, state of health or emotional state

Paralinguistic aspects of speech are those aspects that are not strictly linguistic, but which contribute to the meaning of an utterance. Paralinguistic features may help to indicate a speaker's attitude, although this may overlap with emotional aspects of speech. Prosody is an important paralinguistic aspect of speech.

Our analysis of prosody is aimed at providing its correlation with focus and prominence within a sentence. Focus refers to that part of the sentence which conveys more information on the important part of a sentence. Focus has relation to the meaning conveyed in a sentence (semantics). Change in the focus association of focus sensitive particles like only leads to distinctly different interpretations of the same sentence (sentence with the same word order). Focus information tries to give prominence to new information (new words) while old words (or given information) are not accented. The accented word(s) forms the focus domain. However, not all of the words in a focus domain need be accented. Focus has important implications for suprasegmental phonology (intonation, stress and duration), as specific intonational tunes are encoded in the grammar to encode focus. In most languages, speakers can use pitch accents on particular syllables to provide focus information in a particular utterance. Studies on focus and its ramifications on syntax and prosody have pointed out different types of focus phenomena. In human-human speech communication, focus can be realized and comprehended in an effortless manner. However, it is a scientific curiosity and also signals processing challenge to mimic the same on a digital machine. For this to happen, first we need to understand how it is manifested in the speech signal. This is because, from signal point of view, focus is such subtle information; there may not be very significant changes from the focus to non-focus part of the speech signals. Therefore a careful analysis and interpretation is required for the information present at various levels.

In order to elicit the data for our analyses on the production of focus and its relation to prosody, we elicit presentational/information focus by playing out various questions to native Marathi speakers and having them record their answers. The questions are intended to produce focus on various different elements of the sentence as well as the sentence as a whole.

Speech has special voice features that are inherent to communication. When we talk with someone we usually want to convey two things: some objective information or the WHAT and some sort of attitude or how we feel about it. Both of this aim is accomplished by speaking with special variation of voice pitch and rhythm, called prosodic features. We tell WHAT by choosing word and putting them in phrase that are spoken melodically according to certain rule. These rules for what content of our message are code rule of our language that distinguish among word and signal the grammar of sentence. These melodies are pattern of rhythm, tones and timing of the syllables. These patterns are a linguistic code; they are specific to one's native language and usually different from those of other languages. It is these set of melodic rules that contribute to the production and perception of focus, i.e. the WHAT we wish to convey.

Acoustic cues

Important acoustic cues are as follows:

1. Pitch: Pitch, in speech, the relative highness or lowness of a tone as perceived by the ear, which depends on the number of vibrations per second produced by the vocal cords. Pitch is the main acoustic correlate of tone and intonation. In addition, words which are in focus tend to have a higher pitch.

2. Intensity: Intensity of particular sound is a relative measure of the energy contained in a particular waveform signal. While vowels tend to have higher intensity, consonants have a lower Intensity. By placing prominence on a word, speakers tend to produce it with greater loudness and intensity.

3. Duration- Duration of a syllable or word is the length of time for which the particular sound exists.

These Acoustic cues can be directly extracted from PRAAT software.

Marathi, a language spoken predominantly in the Indian state of Maharashtra with its population of over 100 million, is a relatively poorly studied language as far as the prosody is concerned. However there exist a few studies on the prosody of Hindi [6, 7, 8, 9, 10]. Hindi and Marathi share numerous similarities with regard to the written word as well as pronunciation since they are both derived from Sanskrit, like several other Indo-Aryan languages. However Hindi is known to be influenced by Persian while Marathi has Dravidian influences in its phonetic inventory. Further, unlike in Hindi,phonological length opposition of vowels is not seen in present-day Marathi [11].

For various speech technology applications it is necessary to know which acoustical features play a role in the perception of prominence. Prominence and its realization in the speech signal can be useful in speech synthesis and speech recognition, especially in applications where ambiguous sentences occurred. In natural speech the relationship between prominence and certain acoustical correlates, such as F_0, duration and intensity, is complex. Much is known about the acoustical correlate F_0 and its close relation to pitch accents, much less is known about other acoustical correlates such as intensity and duration. Also less is known about the variability within and between speakers to emphasize words in fluent speech. In this study we present, next to F_0, some acoustical measurements on duration and intensity and its relation to prominence.

Despite the fact that prominence can be useful as an interface between acoustics and linguistics, prominence is not a very well defined term in literature. However, a common definition of prominence is that it refers to those words or syllables that are perceived as standing out from their environment. Or to put it in another way: prominence refers to a greater perceived strength of words in a sentence [21,22]. Therefore, in this study prominence was defined through judgments of naive listeners, who were instructed to mark all those words they perceived to be spoken with emphasis.

1.2 Problem statement

Prosody modeling is one a challenging task in designing a text to speech convertor. In general, prosody mainly consists of duration, pitch and intensity of the spoken unit. Besides, the break between units is one of its important elements as well. Therefore, one utterance’s prosody can be regarded as the elaborate composition of these four perceivable characteristics and the variations in prosody stem from a lot of factors in different dimensions which can be observed in the real speech corpus such as the syllable’s position in the sentence, lexical tone,the speaker’s emotions and so on. Furthermore the complex interactions between factors lead to another difficulty in designing the prosody model. As a result, in addition to inferring the reliable factors influencing the prosody, modeling the interactions between factors intelligently is also a challenge in this work.

Marathi is a very poorly studied language, but studies are available for other languages such as English and Dutch, which will help us in building a prosody model for Marathi. However, understanding of prosody is a challenging task as it varies with different people.

1.3 Motivation

Marathi, a language spoken predominantly in the Indian state of Maharashtra with a speaker population of over 100 million, is a relatively poorly studied language as far as the prosody is concerned. However there exist a few studies on the prosody of Hindi [refs 6, 7, 8,&9]. Hindi and Marathi share numerous similarities with regard to the written word as well as pronunciation since they are both derived from Sanskrit like several other Indo-Aryan languages. However Hindi is known to be influenced by Persian while Marathi has known Dravidian influences in its phonetic inventory. Further, unlike in Hindi, phonological length opposition of vowels is not seen in present day Marathi. There are some prosody models on English and other languages, but no study found on Marathi. So this study will help to build a model of Marathi prosody which potentially has use in speech recognition and speech synthesis.

1.4 Organization of report

Two types of data to be analyzed are as follow:

A. Elicited SOV sentences with different focus position

B. News reading recordings from AIR by 1 Marathi speaker.

Chapter 2 gives brief overview regarding previous literature on the topics.

Chapter 3 explains about the production and perception of elicited data set study which contains an overview about the data set elicited for Marathi speech. Recording, verification and annotation for these sentences, feature extraction and methods of analysis are briefly explained, followed by a detailed overview of the results and observations and the perception test experiment.

Chapter 4 describes about prominence detection in news reading which contains a brief explanation about the speech material used; procedure and design of the experiment; preprocessing and acoustical measurements; the usage of spearman's correlation coefficient in analysis and the results of listening experiments to identify prominence. We then describe the acoustical measurements obtained well as their relation to perceived prominence.

Chapter 5 describes the summary and discussion for elicited data set and detection of prominence in read speech used for news broadcasts.

Chapter 6 states the conclusion and ideas for future work.

Chapter 2. Literature Survey

2.1. Literature survey for elicited data set

Campbell et al [1] examines spectral correlates of stress and accent in a corpus of sentences with varying focus, produced by four speakers of American English. Analyses of the spectrum at vowel centre shows a clear effect on the spectral tilt - i.e. more energy at higher frequencies relative to energy nearer to the fundamental - when the vowel is in a nuclear-accented syllable. These results lend support to the hypothesis that syllables marking focal prominence are phonated in a more emphatic way than other syllables. That is, accented syllables may be louder and not simply intonationally prominent, but this effect does not distinguish an independent lexically specified level of 'primary stress' between the intonational prominence of accent and the basic rhythmic contrast between strong (full) and weak (reduced) syllables. We need to observe the same thing as in German where the ratings of syllable prominence are strongly correlated with the amplitude Aa of underlying accent commands, syllable duration, maximum intensity and mean harmonics-to-noise ratio is this applicable to Marathi language or not. This would be an interesting topic for further research

Breen et al [2] studied the effect of focus width (narrow, broad), focus location (subject, verb, object), and focus type (contrastive, non contrastive) on the acoustic features of utterances in English. A large set of parameters were observed, followed by a step-wise discriminant analysis to determine the strength of each of these in categorizing the utterances. They observed that the strongest indicators for classifying focus location and width were duration+silence, intensity, mean F0, and maximum F0. The selected parameters were indeed successful for classification into the different categories. For focus type, a clear distinction was seen only in the experiment where the speakers were aware of the ambiguity of contrastive and non-contrastive focus, i.e. when they were exposed to all the utterances of each type. The maximum contribution was from the intensity. It was found that for focus location and focus width, the selected parameters were indeed successful for classification into the different categories. For focus type, this was the case only in the experiment where the speakers were aware of the ambiguity of contrastive and non-contrastive focus, i.e. when they were exposed to all the utterances of each type. We can check which is the strongest indicator for classifying focus location and width. Contrastive and non contrastive are having any difference or not.

Lee et al [3] studied the full scale of focus effect in Korean in pre-focus, on-focus and post-focus words. Results show that focused words have significantly increased F0, duration and intensity, post-focus words have significantly reduced duration and intensity, but pre-focus words lack systematic changes in any of these parameters. Thus Korean seems to resemble languages like English and

Mandarin that exhibit post-focus compression (PFC), but differ from Taiwanese and Cantonese where PFC has been found to be absent. We have to observed that how pitch, intensity and duration changes according to focus location in Marathi and is these post focal compression and pre focal compression is observed or not. We can check in Marathi language whether post focus compression is achieved or not.

Mixdorff et al [4] explored the relationship between perceived syllable prominence and the acoustic properties of a speech utterance. Their acoustic analysis confirms earlier results in that focus and sentence mode modify the fundamental frequency contour, syllabic durations and intensity. Ratings of syllable prominence are strongly correlated with the amplitude Aa of underlying accent commands, syllable duration, maximum intensity and mean harmonics-to-noise ratio.

Swert et al [5] reports on a comparative analysis of accentuation strategies within Italian and Dutch noun phrases. They compare accent patterns in different languages from an acoustic, perceptual and functional point of view. They observed a strong language dependency in a comparison of Dutch and Italian. We can observe language dependency of Hindi and Marathi as they both have derived from Sanskrit

Genzel et al [6] observed that contrastive focus affects the pitch rise associated with a prosodic word in Hindi declaratives in two ways. First, duration is affected. The word is lengthened under contrastive focus and second, the scaling of the pitch contour is affected. We need to check this duration and pitch effect in Marathi language unlike in Hindi.

Puri [7] found the main acoustic correlates of focus in Hindi by bilingual speakers (of Indian English) to include increased duration as well as an F0 excursion on the focused element and post-focal reduction in duration, amplitude and F0 excursion. We can observe post focus reduction in Marathi language as Puri did for Hindi.

Fery [8] studied main Indian languages—Hindi, Bengali, Tamil and Malayalam—and observed that these languages show common intonational properties. These languages are termed as phrase languages as they don't have pitch accents. The pattern exemplified by declarative sentences in Hindi, Bengali, Tamil and Malayalam consist in a rising contour, decomposed in a low tone associated with the beginning of the prosodic phrase(p-phrase), and a high tone associated with the end of the same p-phrase. Sentence final p-phrases have a falling contour with an early high tone and a late low tone. We can verify whether this prosodic phrase is observed in Marathi language or not. This would be an interesting topic for further research

Hansberger et al [9] studied about the tone pattern of Hindi which was similar as Fery [8]. Hanrsberger used Yes/ No types questions to elicit contrastive focus. They found difference result for contrastive and non-contrastive questions from a separate study that use Wh-type to elicit non-contrastive focus. Whether there is difference between contrastive and non contrastive question in Marathi or other languages is an interesting topic for further research.

Patil et al [10] studied the intonational realization of focus and its interaction with different word orders. They used SOV and OSV word order and wide focus data for experiment. They observed that SOV sentences show greater post focal pitch range compression than OSV. OSV word order has higher F0 peaks as well as a larger F0 range than SOV order. Duration is longer in object focus for OSV sentences. Subject focus in SOV word order has a higher F0 max, a greater pitch range and longer duration as compared to wide focus. Changes in pitch and duration for different word order such as SOV and OSV for Marathi and other languages would be an interesting topic for further research.

Yardi et al [11] observed that, unlike in Hindi, phonological length opposition of vowels is not seen in present day Marathi.

Rao et al [12] proposes a method for detecting word boundaries. This method is based on the behavior of the pitch frequency across the sentences. The fundamental frequency (F0) is found to rise in a word and fall to the next word. The presence of this fall is proposed as a means of detecting word boundaries. Four major Indian languages are used and the results show that nearly 85% of the word boundaries were correctly detected. The same method used for German language shows that nearly 65% of the word boundaries were correctly detected.

Rajendran et al [13] uses the properties of F_0contour such as declination tendency, resetting and fall-rise patterns in Hindi to identify syllabic units, by using the energy contour, pitch and the first order LP coefficient. Each syllabic unit is assigned an accent value L (Low), H or h (High) by (i) comparing the F_0 value at the midpoint of each syllabic nucleus with that of the previous syllabic unit and (ii) comparing the F_0 values at two different points within each syllabic unit in a sequence having an accent value L. Word boundaries are placed between the adjacent syllabic units (i) H and L, (ii) h and L, (iii) L and L, (iv) L and h and (v) H and h. An evaluation conducted on a corpus of 50 sentences in Hindi read aloud by five native speakers in an ordinary office environment showed that about 74 percent of the word boundaries and about 28 percent of the function words were correctly identified. The results of the word boundary hypothesization can be used to improve the performance of the acoustic-phonetic, lexical and syntactic modules in a speech-to-text conversion system. This conclusion about F0 resetting was similar to Rao et al [12]. We can observe this behaviour of pitch (F0) for Marathi language.

Pandey et al [14] observed an F0 behavior called a “hammock shape” in Hindi also observed it in Indian English. This observation is important, since it may mean that the more significant F0 changes would occur towards the end of the word. We can observe is this hammock shape is observed in Marathi language or not unlike Indian English.

Dyrud et al [15] investigated of phonological predictions about stress in general, and specifically about stress in Hindi-Urdu.

Dhonge et al [16] studied about Marathi linguistic characteristics, morphology and lexical stress.

2.2. Literature survey for News reading study

Barbertje et al [17] investigated to see which acoustical features have a linear relation with perceived prominence (by judgment of native listeners) in Dutch sentence that were read aloud. They used F0 range per syllable in semitone, duration per syllable in seconds and loudness of vowel in dB as acoustical features. To determine the relation with the acoustical features, spearman’s correlation coefficient was used. It turn out that not only F0 range per syllable has a high correlation with prominence but also the loudness per vowel. In case of syllable duration the relation with prominence is not that strong. These results were then verified by using an artificial neural network classifier. Which acoustic parameter is prominent in defining prominence in Marathi can be found out unlike in Dutch study for other languages. Other classifier can be used for making it more accurate.

Barbertje et al [18] also presented acoustical as well as lexical features for classification purposes of perceived prominence in Dutch sentences. Via a perception experiment with 10 naive listeners, they derived prominence labels at the word level for 500 sentences. Parts of these sentences were used for lexical/syntactical analyses. It turns out that most of the function words are never perceived as prominent, and that specific content words namely adverbs, nouns and adjectives are almost always perceived with some degree of prominence, whereas verbs form a middle class. They use F0 range per syllable, both ‘raw’ and corrected for the declination line, to distinguish between the most prominent and non-prominent content words, although intensity and duration features can be used as additional features to improve the classification. As an initial result we can conclude that F0 range is a very good feature to distinguish between prominent and non-prominent content words. Other factor can be define for definition of prominence and other classifier can be used for verify the conclusion drawn from this factor.

Portele et al [19] demonstrated a prominence-based approach which turned out to be a useful interface between acoustics and linguistics.

Mayer [20] suggests using prominence of words to disambiguate sentences in which the pronominal reference is unclear. In this kind of ambiguous sentence, the notion of pitch accents is not enough for disambiguation. This underlines that prominence and its realization in the speech signal can be useful in speech synthesis and speech recognition, especially in applications where ambiguous sentences occur. We can see this effect in Marathi language.

Ladd [21], Terken [22] tried to define prominence in their study. Despite the fact that prominence can be useful as an interface between acoustics and linguistics, prominence is not a very well defined term in literature. However, a common definition of prominence is that it refers to those words or syllables that are perceived as standing out from their environment. Or to put it in another way: prominence refers to a greater perceived strength of words in a sentence.

T hart et al [23] observed the pitch moment for speech synthesis.

Kompre [24] and Wightman et al [25] performed a first attempt to solve the problemsin recognizing accented words in speech recognition.

Streetfkerk et al [26] studied prominence by two perception experiments with different instructions and different presentations were used to locate prominence in 81 read aloud sentences. Results show that, depending on the instruction and presentation (mark prominent words or prominent syllables), the subjects can listen more analytically or more globally. The results indicate that a word perception experiment is a good method to detect prominence at sentence level, coming closer to sentence accent than in a syllable perception experiment. Another pilot perception experiment was run to investigate prominence marking in monotonous sentences. This pilot experiment shows that listeners are indeed able to mark prominent words even in sentences with monotone pitch. Furthermore various acoustical measurements were done both on the most prominent as well as on the non-prominent words. With the help of an artificial neural network, prominence is classified on the base of acoustical information only. Prominence detection is very important in building a prosody model so we can check out this which is important factor in defining a prominence in Marathi language.

2.3 Problem defination

- We can observe whether the hammock shape is present in Marathi language as define by Pandey [14].
- To build a prosody model we need to study the effect of focus shift on the prosodic features for Marathi language.
- Whether Post focus compression is achieved in Marathi language.
- Whether there is difference between contrastive and non contrastive focus.

2.4 Reseach problem

- Do Marathi speakers use acoustic cues for prominence (or only lexical cues)?If so which cues are used and how?
- Do Marathi listeners reliably perceive prominence via acoustic cues?
- News reading style- Do listener reliably perceive prominence?
- If so, what are the acoustic cues that signal prominence

Chapter 3. Methdology

Methodology

This study was done in two phase. First phase is the elicited data set study, second phase is the news reading. In first phase we consider elicited data set from native Marathi speakers. In second phase we consider native Marathi listeners to mark prominence word by news reader. Annotate this audio by native Marathi person and did some statistical calculation on this data. All this Methodologies are explained briefly in this chapter.

3.1 Phase-1: Elicited Data Set study

Methodology:

1. Recording for data set- We first took recoding from native Marathi speakers in Python GUI. Seven pre-recorded questions by a different native Marathi speaker were used as prompts to elicit appropriately focused responses as shown in tables above for narrowly focused non-contrastive and contrastive statements, and one for the broadly focused.

2. Perception test- The data collection was followed by perceptual verification.We retained only those speakers for analyses who had all 14 utterances pass the verification with native Marathi listeners.

3. Acoustic Measurement- We normalized pitch, intensity and duration by PRAAT script and represent this data in boxplot.statistical calculation such as Anova used to show difference in data to prove the result and support the result of ox plot.fuzisaki model used to get result for pitch.

4. Draw conclusion- Conclusion was drawn from this results and observation.

5. Synthesis and perception-This result was crosscheck by perception test by native Marathi listeners. Perception test on natural stimuli is carried out in order to check whether the native listener reliably perceive focus. Native Marathi speaker speak the sentences with narrow and broad focus which contains answers with contrastive and non contrastive focus [1]

[1] **This phase-1 in this chapter is the result of joint work with my co-author P. Rao, H. Mixdorff, I. Deshpande, N. Sanghvi and S. Kshirsagar"Quantitative Study of focus shift in Marathi", Speech Prosody, 2016, Boston,U.S.A.**

3.1.1 Elicited Data Set

The statement for the first target sentence, along with its various written interpretations, is given in table 3-1:

Table 3-1 Amol target sentences with its various interpretations

	Original sentence in Devanagari
Target Sentence	अमोल आईबरोबर बोलत होता.
Non-Contrastive Subject Focus (Q)	कोण आईबरोबर बोलत होता?
Non-Contrastive Subject Focus (A)	<u>अमोल</u> आईबरोबर बोलत होता.
Non-Contrastive-Object Focus (Q)	अमोल कोणाबरोबर बोलत होता?
Non-Contrastive-Object Focus (A)	अमोल <u>आईबरोबर</u> बोलत होता.
Non-Contrastive-Verb Focus (Q)	अमोल आईबरोबर काय करत होता?
Non-Contrastive-Verb Focus (A)	अमोल आईबरोबर <u>बोलत</u> होता.
Contrastive Subject Focus (Q)	रोहित आईबरोबर बोलत होता का?
Contrastive Subject Focus (A)	<u>अमोल</u> आईबरोबर बोलत होता.
Contrastive object Focus (Q)	अमोल <u>भावा</u>बरोबर बोलत होता का?
Contrastive object Focus (A)	अमोल <u>आईबरोबर</u> बोलत होता.
Contrastive verb Focus (Q)	अमोल आईबरोबर <u>खेळत</u> होता का?
Contrastive verb Focus (A)	अमोल आईबरोबर <u>बोलत</u> होता.
Broad Focus (Q)	तुम्ही <u>काय</u> म्हणालात?
Broad Focus (A)	अमोल आईबरोबर बोलत होता.

2

The statement for the first target sentence, along with its various written interpretations, is given in table 3-2:

Table 3-2 Aanand target sentences with its various interpretations

	Original sentence in Devanagari
Target Sentence	आनंद आईबरोबर बोलत होता.
Non-Contrastive Subject Focus	कोण आईबरोबर बोलत होता?
Non-Contrastive Subject Focus	आनंद आईबरोबर बोलत होता.
Non-Contrastive-Object Focus	आनंद कोणाबरोबर बोलत होता?
Non-Contrastive Subject Focus	आनंद आईबरोबर बोलत होता.
Non-Contrastive-Verb Focus	आनंद आईबरोबर काय करत होता?
Non-Contrastive-Verb Focus	आनंद आईबरोबर बोलत होता.

3.1.2 Recording

Data from a total of 20 native speakers of standard Marathi were collected. All were young adults studying or working in Mumbai. Eventually, we used the data of 12 speakers (6 male and 6 female) for our analyses based on a verification procedure described later. Seven pre-recorded questions by a different native Marathi speaker were used as prompts to elicit appropriately focused responses as shown in tables above for narrowly focused non-contrastive and contrastive statements, and one for the broadly focused. Two instances were elicited of each of the target forms providing 14 utterances per speaker. The target utterance for the contrastive form includes the prefix Nahi, 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 test speakers without further explanation. We recorded our speakers with a high quality microphone in a quiet room at 16 kHz sampling rate.

3.1.3 Verification

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 location and width of focus to verify that it matched the intended location/width. We retained only those speakers for analyses who had all 14 utterances pass the verification with both listeners. This led to a dataset of 168 utterances (12 speakers x 7 focus conditions x 2 instances).

3.1.4 Annotation of Recordings

Before beginning the annotation, all utterances were manually trimmed to include ~300 ms of silence at the start and end. After that, a PRAAT script was run to create default textgrids for each utterance, with a fixed alignment. Manual realignment was then carried out, with the help of the spectrogram and waveform views in PRAAT, keeping in mind the phones constituting the syllable (V, CV, CVC corresponding to our critical words: A-mol, aai, bo-lat). Silences with duration greater than 100 ms were explicitly segmented, while silences lesser than 100 ms were merged with the preceding syllable.

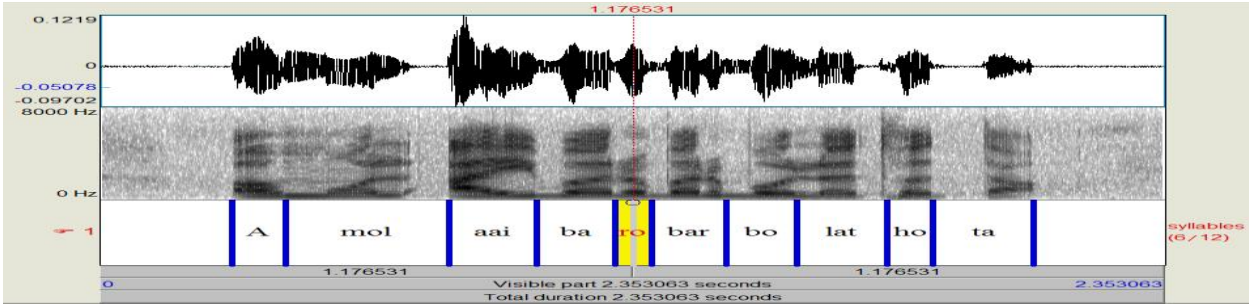

Figure 3-1 Screenshot of PRAAT with the syllable boundaries shown along with the waveform and spectrogram views.

In figure 3-1 first tier represent the waveform of audio data, second tier represent the spectrogram view of audio data which will help in annotation purpose, third tier represent the syllable level annotation of audio data.

3.1.5 Feature Selection

First, a PRAAT script was run in order to create text files for each utterance, which contained the normalized values of each of the measurements described later.

3.1.5.a Duration

We consider Z score normalization of log durations, by using the following formula

$$\text{Log Z} = \frac{x-\mu}{\sigma} \quad \textbf{(3.1)}$$

The log duration of each syllable(to the base 10) is taken into consideration. Doing so, we get 10 data points(i.e. normalized log durations) per utterance, for each of the ten syllables in that utterance. No silence greater than 100 ms is taken into consideration for calculating the durations of each their syllable as well as their log duration normalized values.

For normalized log duration at the word level, the values are obtained by adding the z-score normalized log durations of the syllables in the word. Taking these arrays into account, the box plots are obtained and a one-way ANOVA is performed.

3.1.5.b (Pitch) F0

We consider F0 maximum, F0 minimum, F0 span and F0 mean. For F0, we first removed any spurious pitch variations, by using the PRAAT Manipulation window. After that, we extracted the F0 contours, sampled at 10 ms and performed our analyses on them. We measured F0 by normalizing it in semitones with respect to utterance mean F0 for the word-level maximum, minimum and F0 span, as well as syllable-level mean F0.

$$\text{Minimum pitch in semitones}=39.863*\log_{10}\frac{minimumvalueoft\quad hatsyllable}{meanvalueofutterances} \quad \textbf{(3.2)}$$

Our F0 measurements involve scanning each word forward from the first encountered local minimum to the first local maximum. These constitute F0 min and F0 max respectively, with the difference in semitones providing the F0 span.

.After the pitch objects were generated, we ran a PRAAT script to generate pitch measures.

3.1.5.c Intensity

We consider maximum and mean intensity. Intensity measurements are also calculated using the intensity contours extracted from PRAAT. All measurements were normalized by subtracting the utterance mean intensity.

3.1.6. Method of Analysis

3.1.6.a Box plot

The box plot (whisker diagram) is a standardized way of displaying the distribution of data based on the following: minimum, first quartile, median, third quartile, and maximum.

The box plot is a quick way of examining one or more sets of data graphically.

They take up less space and are therefore particularly useful for comparing distributions between several groups or sets of data.

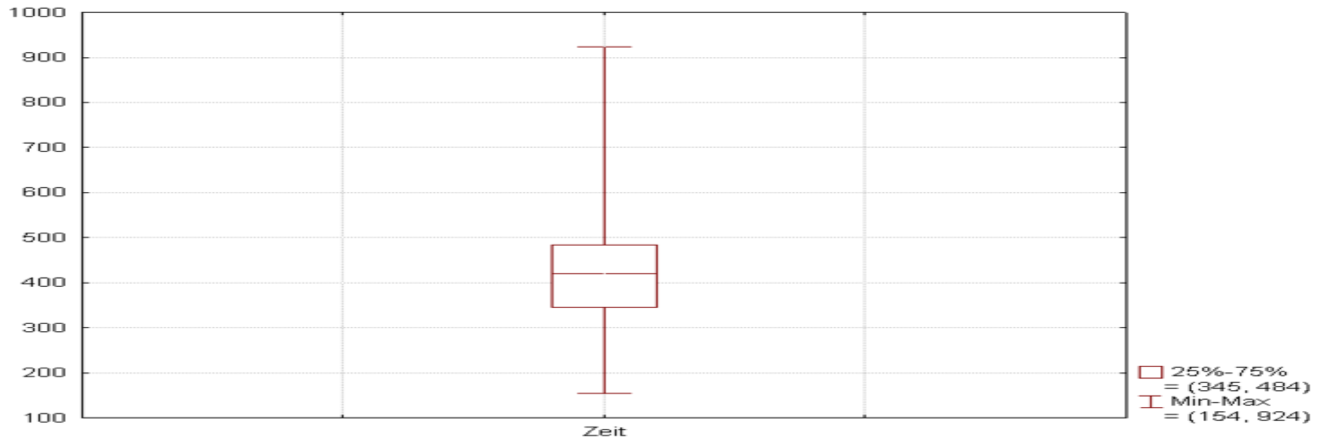

Figure 3-2 Box plot example.

In figure 3-2 two lines corresponding 154 and 924 represent minimum and maximum value of data.In box plot first quartile (25% of data) is at 345 and third quartile (75% of data) is at 475. Median of data is at 404. This boxplot will help in disguish the different data sets.

3.1.6.b Fujisaki Modeling-

The Fujisaki model approximates natural F0 contours by superimposing three components: A constant base frequency Fb (indicated by the dotted horizontal line), exponentially decaying phrase components which are the responses to the phrase commands and accent components which are the smoothed responses to the accent commands. We prefer this model over the TOBI labeling, since synthesis is not possible from TOBI markings, and no quantitative estimates can be made from them.

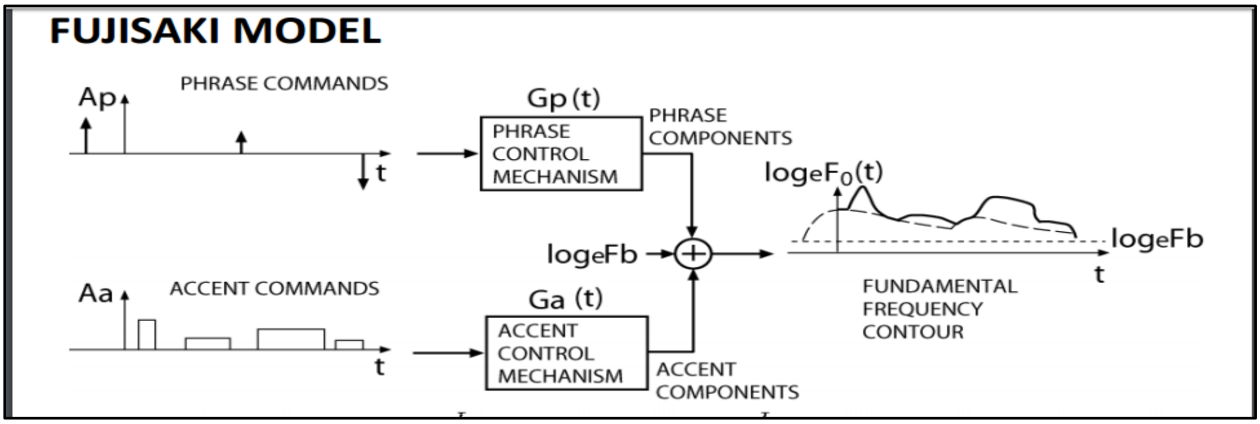

Figure 3-3 Fujisaki model detail explanation

$$\ln F0(t) = \ln Fb + \sum ApiGp(t - T0i) + \sum Aaj[Ga(t - T1j) - Ga(t - T2j)] \quad \textbf{(3.3)}$$

Fb: baseline value of fundamental frequency
I: number of phase command
J: number of accent command
Api: magnitude of the I th phase command
Aaj: amplitude of the jth accent command
T0i: timing of the I th phrase command
T1j: onset of the j th accent command
T2j: end of the j th accent command
.

In figure 3-3 phrase commands are passing to phrase control mechanism to get phrase command. Accent commands are passing to Accent control mechanism to get accent command. Accent command are associate with focus and phrase command are associate with starting phrase sentence.This both (phrase command and accent command) will help to get resynthesized contour.

3.1.6.c Anova-

Analysis of variance (ANOVA) is a collection of statistical models used to analyze the differences among group means and their associated procedures (such as "variation" among and between groups)

3.1.7 Perception test on elicited data set.

Perception tests will also be useful in to answer the following questions:

1. Do Marathi listener reliably perceive prominence via acoustic cues
2. Is there any difference between contrastive and non contrastive focus condition in terms of listener perception.

Perception test on natural stimuli is carried out in order to check whether the native listener reliably perceive focus. Native Marathi speaker speak the sentences with narrow and broad focus which contains answers with contrastive and non contrastive focus. The listener was informed of the acoustic cues that they had to keep in mind while producing the statement in different focus conditions. These acoustic cues were determined based on the previous production experiment. Only those cues that were found to be significant focus indicators were taken into consideration:

1. Rising pitch, intensity and duration on focus word.
2. Post focal compression
3. Pre focal silence.

Listeners who did participate in creating the previous recording dataset are not allowed to participate for the perception test. General overview of the experiment and some instructions were explained to each listener and then their performance was noted down.

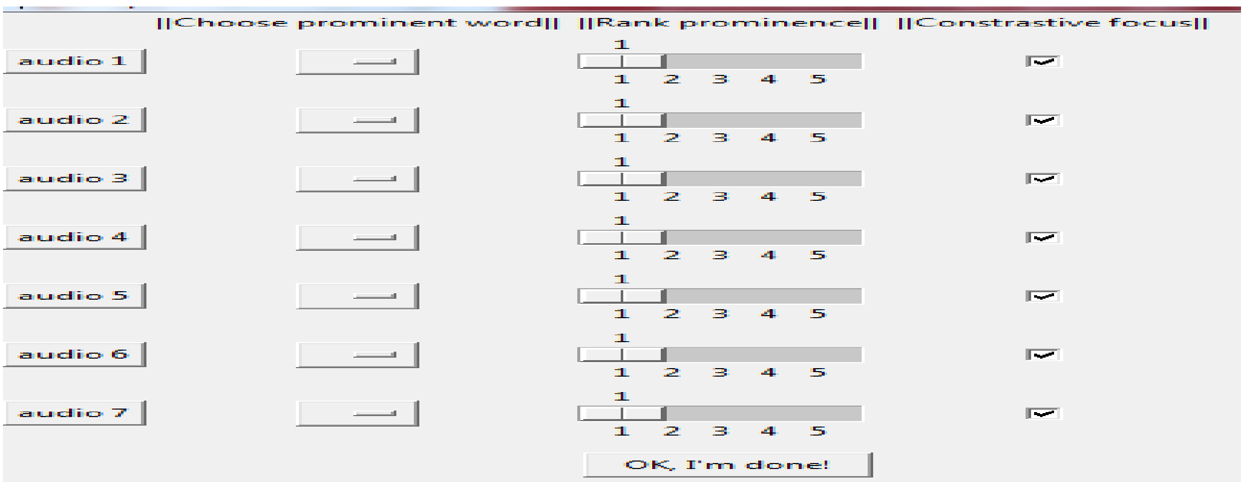

Figure 3-4 Perception test python interface.

In figure 3-4 represent the interface in python for perception test.In column 1, having audio 1, audio 2. This audios represent the recording data of native speaker which contain different focus location for recognition purpose of native listener. Second coloumn (choose prominent word) contain option for different focus location such as focus on Amol, aai, bolat. Third coloumn (rank prominence) represent the degree of prominence. 1 represent very low prominent, 2 represent vlow prominent, 3 represent moderate prominent, 4 represent high prominent, 5 represent very high prominent. Fourth coloumn recognise whether that particular audio is contrastive or non-contrastive.

Perception test result

- 4/4 listener's reliably perceived all sentences correctly in the perception test.
- Subject focus, object focus, verb focus and broad focus were perceived correctly.
- All speakers fail to perceive difference between contrastive and non-contrastive focus.

3.2 Phase-2:News reading study

3.2.1 Introduction

Acoustical features playan important role in the perception of word or syllable prominence, which is what we found out in the previous chapter. But which one is most important in defining a prominent word is still a question. Knowing the relevant features for perceived prominence can be very useful in speech technology applications such as speech synthesis. Knowing more about the realization of different degree of prominence can be helpful to generate more natural speech. From our literature survey, we can conclude F0 seems to be a very important feature to detect the most prominent words in Dutch language. We will try to observe this for Marathi language. An advantage of F0 is that it is more independent for intrinsic properties of speech sounds than duration, but the extraction of F0 features from the pitch movements is still a problem. Therefore, we concentrate on measurement within the syllable.

In this experiment we are trying to present some acoustical measurements on duration, Intensity and F0 and its relation to prominence.

In this chapter we first describe speech material used, the procedure and design of the experiment, and the result of listening experiment to define prominence. We then describe the acoustical measurements obtained well as their relation to perceived prominence.

3.2.2 Speech Material

The speech material was taken from PrasarBharati, a statutory autonomous body established under the PrasarBharati Act that came into existence on 23/11/1997. It is the Public Service Broadcaster of the country. The objectives of public service broadcasting are achieved in terms of PrasarBharati Act through All India Radio and Doordarshan, which earlier were working as media units under the Ministry of I& B and since the above said date became constituents of PrasarBharati.

It contains read speech of several Indian female speakers. We considered both transcripts of broadcast news and audio related to the transcripts.

3.2.3 Procedure for experiment

Native Marathi listeners participated in the experiment.

In our approach we use six native listeners with the aim to get them to mark the prominent words in the transcript, as per their perception.

Each listener has to mark those words that are spoken with emphasis. This instruction is used as an operational definition of prominence.

We find out the agreement between listeners and numbered it in transcript. For example, if two out of the six listeners agreed on a particular word as being prominent, we marked that word as '2'. As a first step the words with agreement of 4,5 and 6 are defined as the prominent words. Another possibility is, to treat the agreements by all speaker as very prominent and by 0 speaker mean non prominent.

3.2.4 Procedure and Design

10 phonetically rich sentences spoken by Indian female native Marathi news reader are presented to 6 native Marathi listeners. The recordings are played out through closed headphones. They listen to the audio of two subjects by the same speaker and mark on transcript. Multiple passes were allowed. Space does not permit us to discuss the within and between listener differences.

3.2.5 Result from listening experiment. (Agreement by listener

Each listener judged audio several times and were allowed to crosscheck. Two paragraphs from transcript were marked by the listener. These paragraphs contained 10 sentences, 180 words, 469 syllables.

Out of 180, 69 words were marked as prominent by at least one of the 6 listeners. This is 38% of the total number of words. On average, there are 11 words per sentence. Hence, we can conclude that there are about 3 or 4 prominent words per sentence. Furthermore, it must be mentioned that about more than half of the words (62%) were never judged as prominent. In these two paragraphs, 13 words are monosyllabic. Speech to silence ratio in terms of duration for the two paragraphs is 9.36.

3.2.6 Pre-processing and acoustical measurement

Before the acoustical features can be measured, the syllable level annotation was done with the help of PRAAT.As a first step we decided to measure the following acoustical features.

1. F0 mean per syllable in semitones.

2. Duration per syllable in seconds.

3. Intensity per syllable.

A Praat script was used to get value from Praat and to normalise it, then result were stored in csv file in Excel and the following row data plots were obtained to get a general overview about data. Box plots were then obtained by using Matlab and finally spearman's correlation coefficient was calculated to see which features have a linear relation with prominent words (as judged by listeners)

3.2.7 Prominence and Acoustical Features

In order to see the relation between the prominence judgments of the listeners and acoustical measurements, three graphs (scatter plot on listeners agree verses normalised data)are presented below.

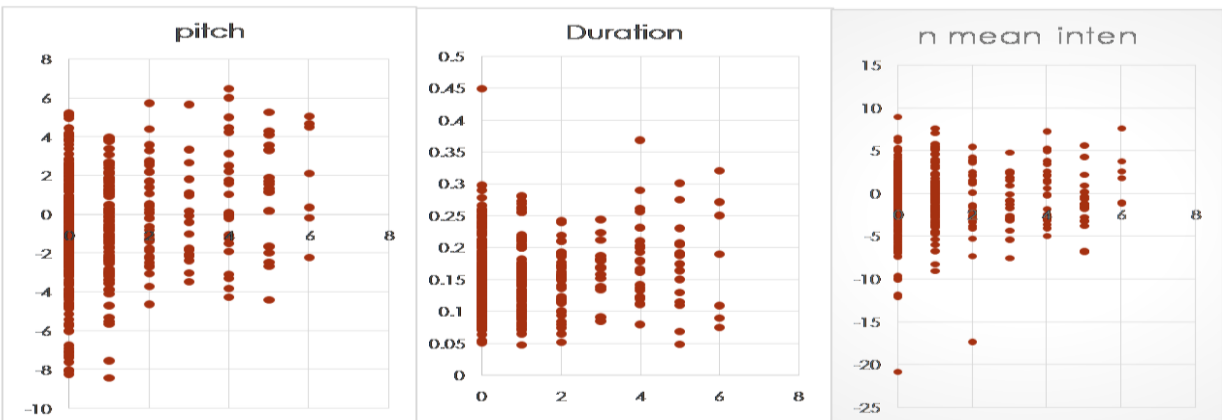

Figure 3-5 Scatterplot for row data.

In figure 3-5 represnt the scatterplot on listeners agree verses normalised data. This graph is giving visualisation of raw data behaviour which support the result of spearman's correlation coefficient.from this we can say pitch plot shows linear relation with number of agreement.

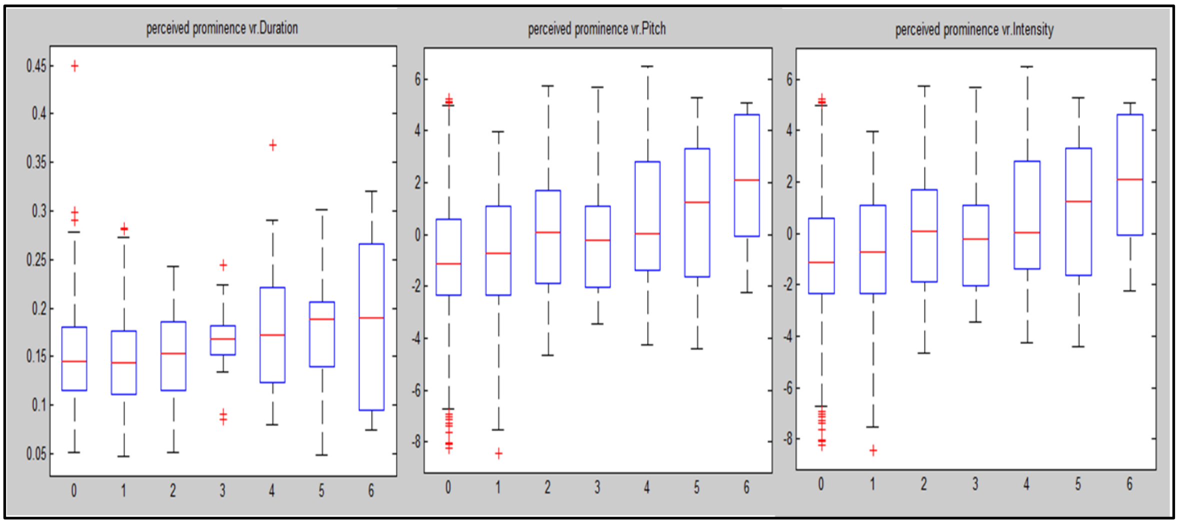

Figure 3-6 Box-plot for row data.

In figure 3-6 represent three graphs show the number of listeners who agree verses the normalized pitch, intensity and durations. The perceived prominence versus the F0 mean and duration per syllable show a higher correlation than perceived prominence versus the intensity of syllables.

To test if there is a linear relation between various acoustical measurements and perceived prominence, spearman's correlation coefficient were calculatated.

3.2.8 Spearman's correlation coefficient

The Spearman's Rank Correlation Coefficient is used to discover the strength of a link between two sets of data.

To test if there is a linear relation between various acoustical measurements and perceived prominence, Spearman's coefficients were calculated.

Spearman's Rank correlation coefficient is a technique which can be used to summaries the strength and direction (negative or positive) of a relationship between two variables. The result will always be between 1 and minus 1.

Interpretation of Spearman rank correlations

Positive and negative Spearman rank correlations

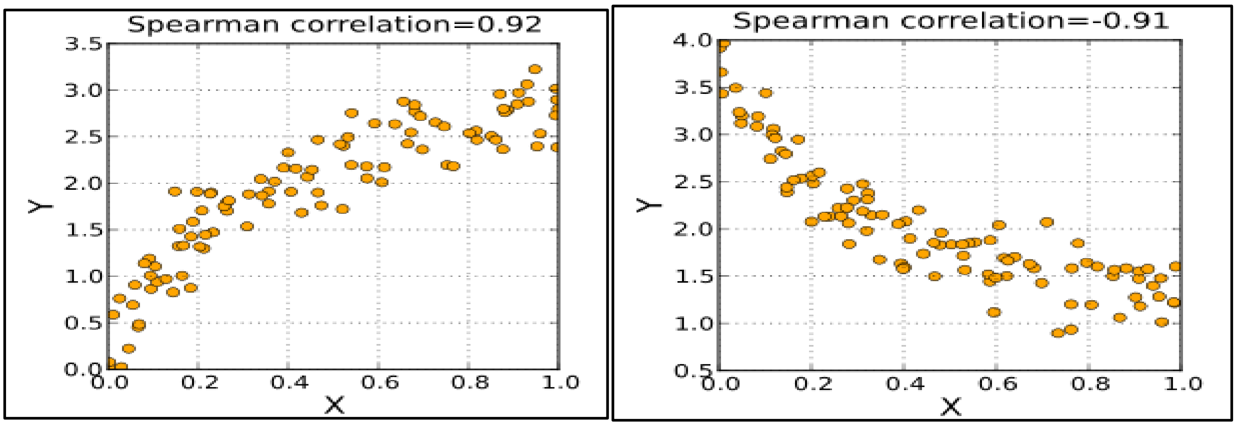

Figure 3-7 Interpretation of Spearman rank correlations.

In figure 3-7 the sign of the Spearman correlation indicates the direction of association between X (the independent variable) and Y (the dependent variable). If Y tends to increase when X increases, the Spearman correlation coefficient is positive. If Y tends to decrease when X increases, the Spearman correlation coefficient is negative. A Spearman correlation of zero indicates that there is no tendency for Y to either increase or decrease when X increases. The Spearman correlation increases in magnitude as X and Y become closer to being perfect monotone functions of each other. When X and Y are perfectly monotonically related, the Spearman correlation coefficient becomes 1. A perfect monotone increasing relationship implies that for any two pairs of data values X_i, Y_i and X_j, Y_j, that $X_i - X_j$ and $Y_i - Y_j$ always have the same sign. A perfect monotone decreasing relationship implies that these differences always have opposite signs.

Chapter 4. Result and discussion

4.1. Results for elicited data set study

The observations for duration, F0 and intensity are reported below, with neutral focus serving as a baseline for comparison in all cases. We observed that while most focus conditions were reliably discriminated across the 12 speakers, this was not true of the contrast-distinction which remained around chance. The listeners also agreed that the recordings sounded natural. Preliminary inspection of individual speakers' data did not reveal any differences in the acoustics between the non-contrastive and contrastive forms. We therefore pooled the data for each of the three focus location forms to get four utterances each per speaker. Only the broad focus was represented by two utterances per speaker.

4.1.1 Duration

We carried out 1-way ANOVA and obtained box plots for the z-score normalized log duration values, using the appropriate MATLAB script. The 1-way ANOVA results for log duration, at word level, are given in table 4-1. If p value is less than 0.05 than there is significant difference between two group.In this table all focus condition such as on focus, pre focus and post focus compare with neutral focus.

Table 4-1 Anova result for log duration at word level

	On focus	Pre focus	Post focus
Amol	F=20.28, p=2.611e-5	F=28.23, p=1.2147e-6	N/A
Aai	F=28.23, p=1.2147e-6	F=1.31, p=0.2559	F=3.21, p=0.0777
Bolat	F=17.47, p=8.323e-5	N/A	F=43.65, p=1.19e-9

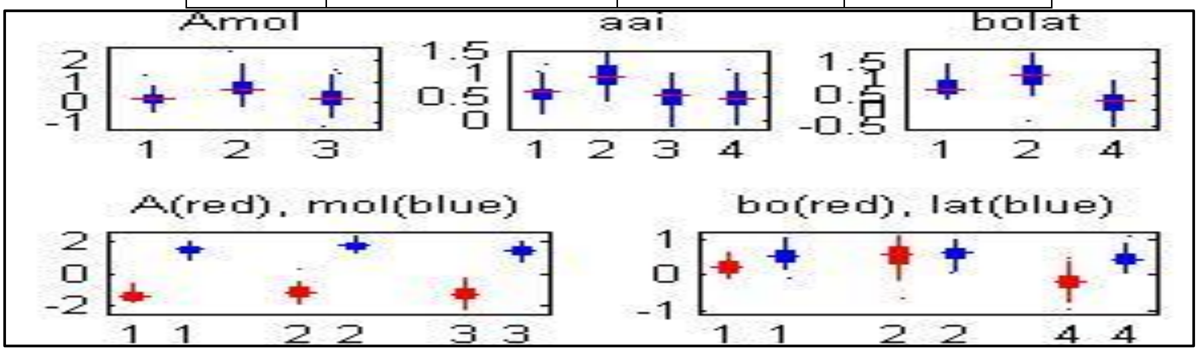

Figure 4-1 Boxplot for log duration at word level.

In figure 4-1 we observed that focus type is clearly cued by word duration for all focus locations. Word duration is extended in focus while it is reduced post-focally, all with respect to the neutral focus. Pre-focus durations (of Amol and aai) are not discriminated from neutral focus. Fig. 3.4 also captures the interesting observation that it is the stressed syllable (i.e. final in Amol, and penultimate in bolat) that undergoes elongation in focus and reduction when in post-focus. An additional observation was that of pre-focal pauses. We detected 47 inter-word pauses across the 12 speakers' data. Of these 37 were pre-focal (26 before aai, and 11 preceding bolat). The remaining ones appeared after the word Amol in narrow or broad focus.

4.1.2. Aanand Utterances

• The objective was to find if lexical stress manifests itself in "Aanand."While the concept of lexical stress is clearly defined for languages such as English, it is poorly studied for languages such as Marathi.

• According to the rules given by Dhonge, it can be hypothesized that the "Aa" in "Aanand" would be the tentative bearer of lexical stress, if "Aanand" were to come in focus. To investigate the presence of lexical stress in "Aanand", the Z scores of the log durational measurements of "Aa" and "nand" were considered. These Z scores were calculated in a manner similar to that for "amma" (Amol Marathi) utterances.

• Later, for all speakers combined, we obtained box plots and performed 1-way ANOVAs for the scores of "Aa", "nand", and "Aanand", for the neutral focus vs. on-focus condition, making observations on the difference in F and p values for each. In order to make sure that the first syllable in the topic marker doesn't receive stress by default, we also performed 1-way ANOVAs on "A", "mol" and "Amol", in a manner similar to that above.

The two results were then compared and conclusions drawn .the separations viewed in the box plots were quite good for "Aa", as would be expected, while they weren't as good for "nand" In 7 out of 11 speakers, "Aa" showed a noticeable separation in On-Focus condition, as compared to Neutral Focus and Pre-Focus conditions.

The 1-way ANOVA results for "Aa", "nand" and "Aanand" were gathered for on-focus vs. neutral focus conditions

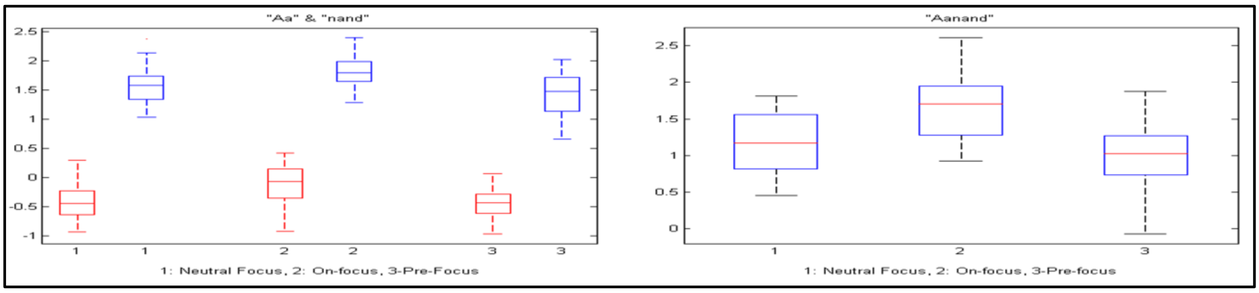

Figure 4-2 Box plot for Aanand and Amol syllable.

In figure 4-2 show a better separation for the syllable “Aa”, as indicated by it’s higher F and lower p values. “nand”, on the other hand, barely falls within the p<0.05 threshold.Post focus compression is achieved is clearly shown in above figure.

“Aanand” has higher F and lower p values than both its constituent syllables.

The 1-way ANOVA results are shown below table, for Subject focus vs. Neutral Focus are given below in table 4-2. If p value is less than 0.05 than there is significant difference between two group.In this table all focus condition such as on focus, pre focus and post focus compare with neutral focus.

Table 4-2 Anova result for Amol and Aanand at syllable and word level.

Syllable & word	F	P	Syllable & word	F	P
A	6.78	0.0112	Aa	9.6	0.0035
Mol	18.63	5.1e-5	Nand	4.48	0.0403
Amol	20.28	2.6e-5	Aanand	14.03	0.0005

4.1.3 Pitch (F0)

<u>Mean pitch</u>

The one way ANOVA results for mean pitch are reported, followed by the relevant box plots, for various syllables are given in 4-3. If p value is less than 0.05 than there is significant difference between two group.In this table all focus condition such as on focus, pre focus and post focus compare with neutral focus.

Table 4-3 Anova result for Mean pitch at syllable level.

	On focus	Pre focus	Post focus
A	F=13.23, p=0.0005	F=1.62, p=0.2059	N/A
Mol	F=27.79, p=1.427e-6	F=0.86, p=0.3563	N/A
Aai	F=6.87, p=0.0107	F=0.26, p=0.6086	F=24.95, p=4.155e-6
Bo	F=2.54, p=0.1157	N/A	F=12.66, p=0.0005
Lat	F=14, p=0.004	N/A	F=91.52, p=2.493e-16

The box plot of mean pitch is as follows:

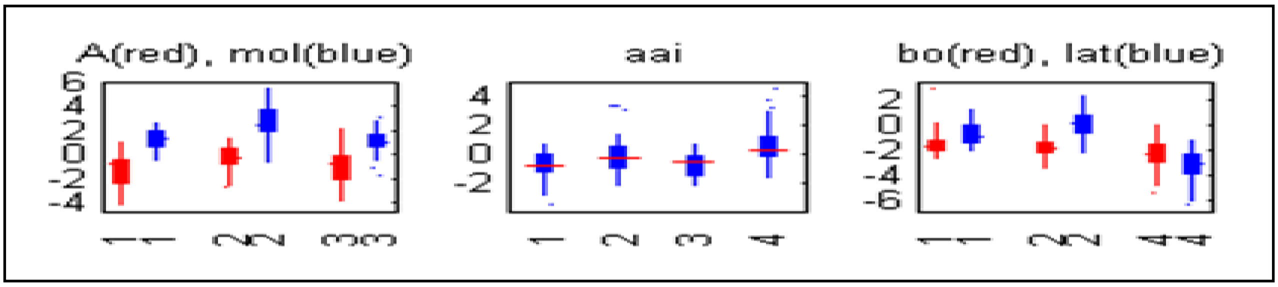

Figure 4-3 Box plot of mean pitch.

In figure 4-3 a peculiar effect is noted, wherein the mean pitch for “aai” in post focus condition is noticeably higher as compared to the neutral and on-focus conditions. This is due to the fact that the high pitch of “Amol” in subject focus carries forward to the pitch in “aai”, leading to an overall increase in the mean pitch.

Maximum pitch

The one way ANOVA results for maximum pitch are reported, followed by the relevant box plots, for various syllables are given below in table 4-4. If p value is less than 0.05 than there is significant difference between two group.In this table all focus condition such as on focus, pre focus and post focus compare with neutral focus.

Table 4-4 Anova result for Maximum pitch at word level.

	On focus	Pre focus	Post focus
Amol	F=10.11, p=0.0022	F=1.36, p=0.2454	
Aai	F=62.05, p=3.074e-11	F=1.15, =0.2868	F=0.35, p=0.5557
Bolat	F=17.54, p=8.0663e—5		F=205.69, p=1.2721e-27

Boxplot for maximum pitch are as follows

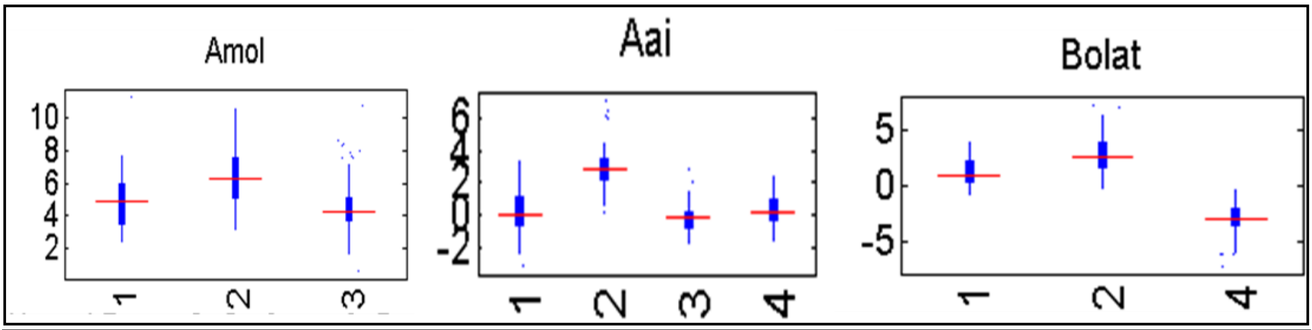

Figure 4-4 Boxplot of maximum pitch.

.

In figure 4-4 we observed that focus type is clearly cued by word level maximum pitch for all focus locations. Word level maximum pitch is extended in focus while it is reduced post-focally, all with respect to the neutral focus. Pre-focus maximum pitch (of Amol and aai) are not discriminated from neutral focus.

Pitch Range

The one way ANOVA results for pitch range are reported, followed by the relevant box plots, for various syllables are given in table 4-5. If p value is less than 0.05 than there is significant difference between two group.In this table all focus condition such as on focus, pre focus and post focus compare with neutral focus.

Table 4-5 Anova result for pitch range at word level.

	On focus	Pre focus	Post focus
Amol	F=0.43, p=0.516	F=2.58, p=0.1108	N/A
Aai	F=50.51, p=7.955e-10	F=3.24, p=0.0762	F=21.01, p=1.941e-5
Bolat	F=16.01, p=0.0002	N/A	F=201.63, p=2.693e-27

The boxplot for pitch range are as follow:

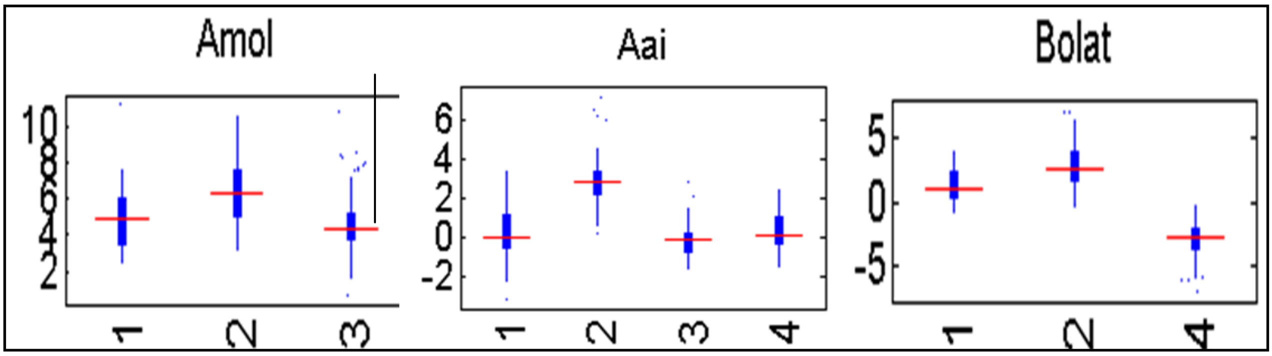

Figure 4-5 Boxplot for pitch range.

In figure 4-5 shows that on focus is clearly observed and post focus compression is achieved. we observed that focus type is clearly cued by word level pitch level for all focus locations. Word level pitch range is extended in focus while it is reduced post-focally, all with respect to the neutral focus. Pre-focus pitch range (of Amol and aai) are not discriminated from neutral focus.

4.1.4 Intensity

The one way ANOVA results for maximum intensity are reported, followed by the relevant box plots, for various words are given below in 4-6.

Table 4-6 Anova result for Maximum intensity at word level.

	On focus	Pre focus	Post focus
Amol	F=10.92, p=0.0015	F=1.94, p=0.1663	NA
Aai	F=17.32, p=8.84e-5	F=0.15. p=0.6964	F=0.61, p=0.4363
Bolat	F=9.75, p=0.0026	.NA	F=25.24, P=1.821e-6

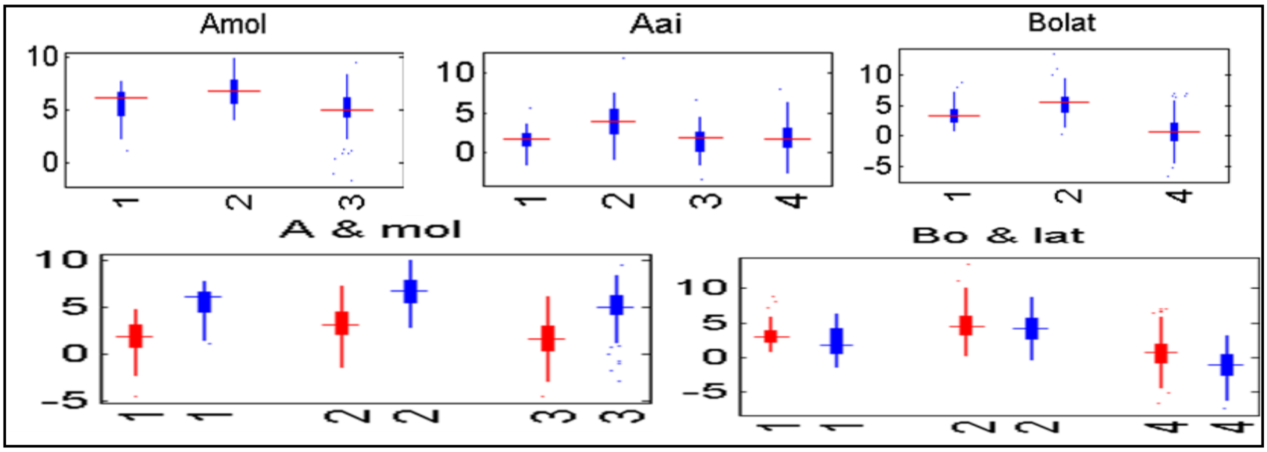

Figure 4-6 Boxplot for maximum intensity.

In figure 4-6 represent word level boxplot for maximum intensity which is clearly showing higher value on on focus post focus compression and second tier represent syllable level boxplot for maximum intensity. we observed that focus type is clearly cued by word level maximum intensity for all focus locations. Word level maximum intensity is extended in focus while it is reduced post-focally, all with respect to the neutral focus. Pre-focus maximum intensity (of Amol and aai) are not discriminated from neutral focus. Fig. 3.9 also captures the interesting observation that it is the stressed syllable (i.e. final in Amol, and penultimate in bolat) that undergoes elongation in focus and reduction when in post-focus.

The one way ANOVA results for mean intensity are reported, followed by the relevant box plots, for various words are given below in 4-7. If p value is less than 0.05 than there is significant difference between two group.In this table all focus condition such as on focus, pre focus and post focus compare with neutral focus.

Table 4-7 Anova result for Mean intensity at word level.

	On focus	Pre focus	Post focus
Amol	F=22.13, p=1.2423e-5	F=123, p=0.2696	
Aai	F=18.38, p=5.67e-5	F=0.08 .p=0.7729	F=3.8, p=0.0553
Bolat	F=16.67, p=0.0001	.	F=33.28, P=6.5451e-8

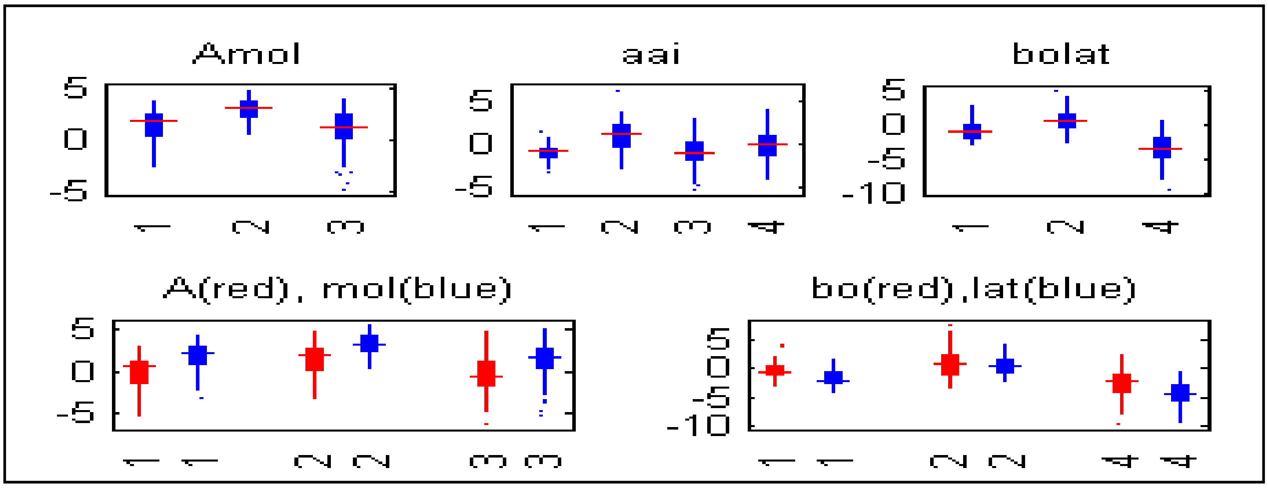

Figure 4-7 Boxplot for mean intensity.

In figure 4-7 represent word level boxplot for mean intensity which is clearly showing higher value on on focus post focus compression and second tier represent syllable level boxplot for mean intensity. we observed that focus type is clearly cued by word duration for all focus locations. Word level mean intensity is extended in focus while it is reduced post-focally, all with respect to the neutral focus. Pre-focus mean intensity (of Amol and aai) are not discriminated from neutral focus. Fig. 4-7 also captures the interesting observation that it is the stressed syllable (i.e. final in Amol, and penultimate in bolat) that undergoes elongation in focus and reduction when in post-focus.

4.1.5. Fujisaki modeling

Fig. 4-8 presents results of the Fujisaki analysis for speaker AK's target sentence for the following conditions: (1) Non-contrastive focus on Amol, (2) aai, (3) bolat, and (4) broad focus. Each of the four panels displays, from the top to the bottom: the speech waveform, the F0 contour, and the underlying phrase and accent commands. 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 for all four conditions. Except for post-focal words we indeed observe a well-defined hammock-like shape as reported by [3, 4]. 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. This indicates that aaibarobar is treated as a prosodic unit. AK, like many other subjects, further emphasizes narrow focus on aai and bolat, respectively, by introducing a short pause before the focused constituents. Further, as may be expected from the language characteristic hammock shaping at word level, the significant pitch increases in the disyllabic words were observed to take effect in the final syllables (-lat, -mol) as measured by syllable mean pitch.

If we compare the three 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 when narrow focus is placed on Amol. 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 F0 gestures on the post-focal ones.

The following table 4-8 displays means of Aa for all syllables in the four focus cases. The non-existence of an accent command associated is taken into account with an Aa of zero. As can be seen, Aa is low for post-focal items (set in grey). Independent samples Kruskal-Wallis test shows that Aa assigned with the critical words is significantly different for aai, bolat depending on the focus condition ($p < 0.001$) whereas the significance is lower for Amol ($p < 0.007$).

Table 4-8 Mean of Aa for all syllables in the four focus cases.

syll.	Amol	Aai	Bolat	Broad
A	0.00	0.00	0.00	0.00
Mol	**0.46**	0.45	0.36	0.45
Aai	0.12	**0.37**	0.12	0.18
Ba	0.07	0.33	0.16	0.18
Ro	0.07	0.26	0.26	0.28
bar	0.08	0.18	0.48	0.50
Bo	0.05	0.19	0.22	0.29
Lat	0.03	0.02	**0.49**	0.38
Ho	0.02	0.03	0.04	0.14
ta	0.00	0.00	0.00	0.00

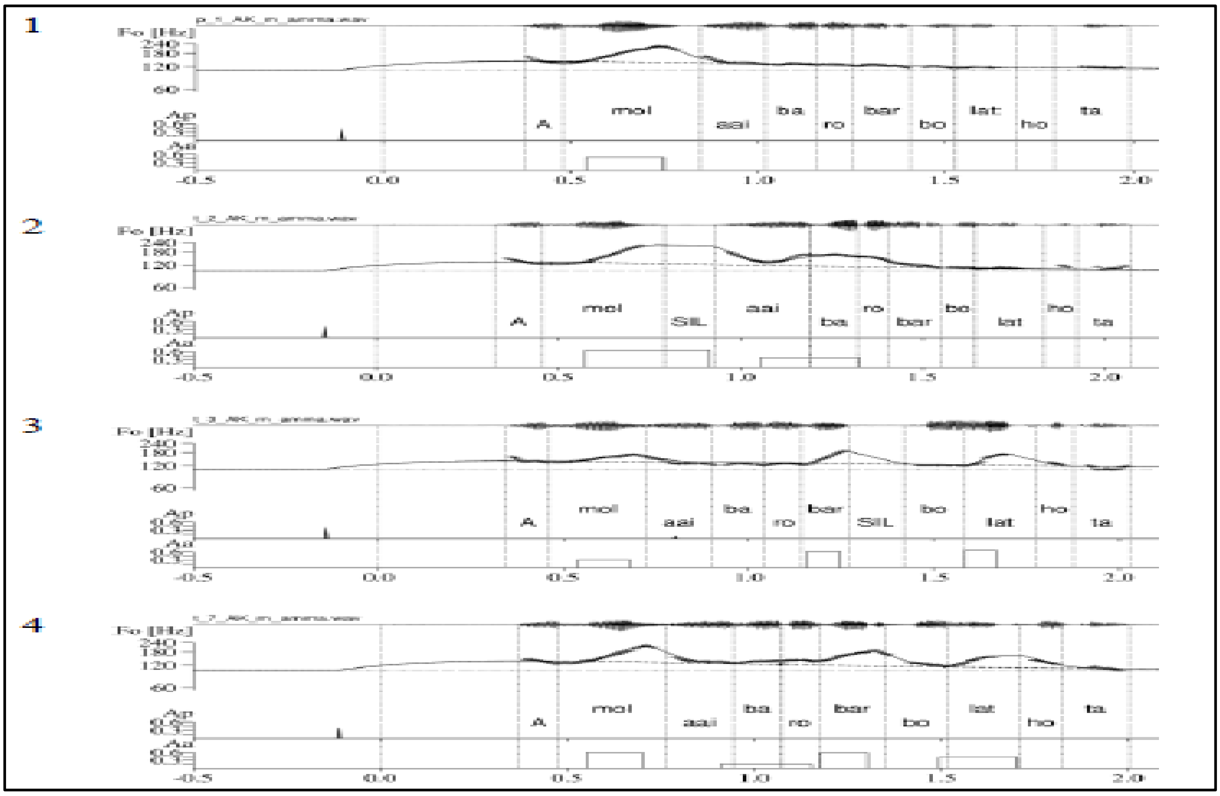

Figure 4-8 Fujisaki modelled pitch contour for various focus conditions.

Figures 4-8 show the Fujisaki modeled pitch contour for various focus conditions:

(1-Subject Focus, 2-Object Focus, 3-Verb Focus, 4- Broad Focus). First impulse represent phrase command and box in sentence represent focus location called as accent command

4.2. Result of spearman correlation coefficient for news reading.

The resulting correlations are presented in following table 4-9. The significant correlation shows that there is positive linear relation between prominence and duration, F0 range and intensity per syllable.

Table 4-9 Spearman's correlation coefficient result.

Acoustic parameter	Spearman's correlation
Pitch	0.53
Intensity	0.4462
Duration	0.4468

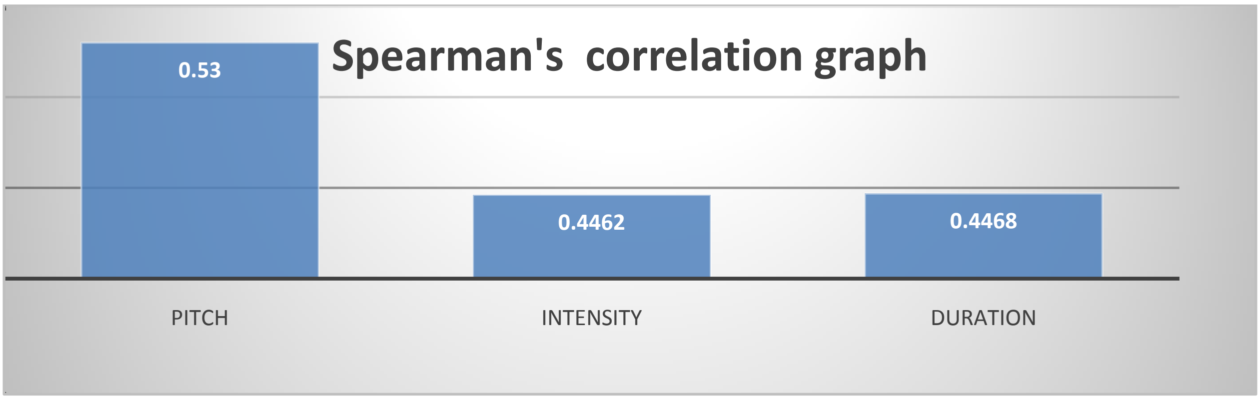

Figure 4-9 Result of spearman's correlation coefficient.

In figure 4-9 shows the result of spearman's correlation coefficient.we can observe that pitch is having high value than duration and intensity so we can say that pitch is having higher linear correlation with agreement of listeners than duration and intensity as spearman's correlation gives the relation of linearity.

4.3 Summary

A clear elongation in duration is seen. More specifically, the stressed syllables “mol” and “bo” in “Amol” and “bolat” respectively undergo elongation. The majority of silences occur before the focus word.

When “aai” is in focus, the pitch rise associated with “aai” has a carryover effect on the word “barobar”. Maximum F0 and the F0 span are significantly higher for narrow focus at object and verb locations with reference to neutral focus. The F0 cues are not so clear for the S focus (Amol). Post-focus on the V (bolat) is cued by both F0 measures, with significant decrease in value compared with neutral focus. Post-focus on the O (aai), on the other hand, is cued only with a decrease in pitch span.

Focus shift does affect the mean intensity of the word relative to broad focus. However the only consistent trend observed across speakers was the post-focal decrease in intensity on the verb. In the disyllabic words, it was observed that both syllables are similarly affected.

Excursion of pitch, high intensity and duration observed for focus words.

There was greater separation for the “Aa” syllable as compared to the “nand” syllable in terms of duration, while the opposite was observed for “A” vs. “mol”. The z-scores for the word “Aanand” were best separated, as they had the highest F and lowest p values. However, a clear separation for “Aa” was not observed for all speakers, hence proving that not all speakers clearly exert stress on a particular syllable in the word, at least not in terms of duration. Furthermore, the flexibility in the pronunciation of “Aanand” leads to some issues. Hence, we can say that while lexical stress does exist in Marathi, the application of it is still somewhat speaker-dependent.

Perception testing shows that all sentences reliably perceived by native Marathi listeners and there is no significant difference between contrastive and Non-contrastive focus.

For news reading there is linear relation between mean pitch and perceived prominence by native Marathi listeners. Duration is also has a linear relation with perceived prominence but Intensity is having very low linear relation with perceived prominence by native Marathi listeners.

4.4 Discussion

Observations from production experiments have been presented as part of a larger study on prosody and information structure in Marathi. Duration, intensity and F0 have been measured at word and syllable level for the target sentence in SOV form. The F0 measurements also involve accent parameters obtained from Fujisaki model fitting of the natural contours. Statistical analyses indicate that all the word-level acoustic parameters (duration, intensity mean, F0 max and span) cue on-focus in object and verb. For the subject, the same holds true, except that F0 cues are not so discriminative for on-focus with respect to broad focus probably due to topic marker effect. In agreement with this, Fujisaki model accent command amplitude shows a clear increase with narrow focus in the object and verb locations but not in subject focus. Pre-focus is not distinguished from broad focus in any of the attributes. Post-focal compression is most clearly observed in duration, intensity and F0 measurements in the verb location, again the post focus compression is also observed by box-plots related to the parameters of the verb bolat.

In the two disyllabic words (Amol, bolat), it is the stressed syllables ("A" and "bo", respectively) that are affected for duration. However, it is the word-final syllables that are affected for F0 increase as seen from the obtained alignment of the accent command. Thus, the regular hammock shape characteristic of content words is emphasized further with F0 increase on the final syllable when the word is in focus.

An interesting observation is the case of aai which forms a unit with the following function word barobar so much so that the strongest F0 excursion appears on the final syllable bar in broad and pre-focus conditions. Here the aai segment shows a variety of F0 realizations, most commonly a slight lowering of F0. When aai is in narrow focus however, there is a well-defined hammock on the focus word with barobar showing a non-increasing F0 including a fall over one or more syllables. This behavior is also captured by the changing alignment of the accent command with respect to the syllable aai.

These observations are not very different from what has been seen for other languages, including Hindi [6, 7], where focus has been seen to be marked by increased F0 span and elongation of the focused word. We find that post-focal reduction of duration and F0 span are very conspicuous too, especially on the verb (bolat). Again, it is interesting that it is not our hypothesized stressed syllable in bolat where the significant changes of F0 occur. The change occurs in the second syllable even when the first syllable is the stressed syllable according to the syllable weight based rule. We thus observe the trend of the stressed syllable to be lengthened or shortened when in or out of narrow focus respectively whereas intensity, if utilized by the speaker, is affected across all syllables of the word.

Finally, in our acoustic analysis we identified features that are modified under focus shift.

Perception tests which were based on natural stimuli spoken by native Marathi speaker states that listeners were able to perceived focus word and focus condition reliably. Listeners fail to identify the difference between contrastive and non-contrastive sentence.

For news reading study, Spearman’s correlation coefficient has a higher value for pitch than duration, and a low value for intensity which states that pitch is the prominent factor for defining prominence or focus in news reading speakers which is similar as Dutch language study.[17] The only difference is that in Dutch language, duration has a low linear relation with perceived prominence but in Marathi, intensity has a low linear relation with perceived prominence.

Chapter 5. Conclusion & Future Scope

5.1 Conclusions from the elicited SOV dataset:

We have studied the effect of focus shift on prosodic features for Marathi, which is a major Indian language. We consider different focus locations and different focus widths. We report observations of fundamental frequency, intensity, and syllabic durations of constituent words of the utterance.Our results indicate the following.

- Narrow focus is marked by longer duration of the focused word, and a larger accent command in the focused word except for the accent command of "mol" in subject focus (possibly due to the topic marker effect).
- Pitch exertion (hammock shape) is always observed on focus word.
- Post-focal effects are observed for duration, intensity and F0.
- No differences were found between contrastive and non-contrastive focus.
- Listeners reliably perceived prominence via acoustic cues.

5.2 Conclusions from prominence detection in news reading style:

Prominence, as defined by the judgments of native listeners, can be predicted from the acoustics as follows.

- Both, the mean pitch per syllable and duration per syllable, display a high correlation with prominence.
- In case of intensity, the relation with prominence is not that strong.

5.3 Overall conclusion

By considering both studies we can finally conclude that prominence in Marathi speech is mainly a function of pitch rather than duration and intensity. Duration is also important factor because speakers tend to extend the duration when they want to give new information (which we called as focus or prominence). But loudness or intensity has the least significant role when it comes to signaling focus.

5.4 Future work:

We have yet to examine the relative perceptual contributions of the acoustic correlates of focus. To that end we intend to carry out perception experiments using resynthesized speech. Thanks to the Fujisaki modeling, F0 contour modifications are straightforwardly realized. Applying PSOLA based resynthesis will provide direct control of durations and intensity. Hence we will be able to examine more quantitative relationships between features and perceived changes in focus.

For news reading, F0 mean per syllable measure in semitones is a useful feature to classify prominent and non-prominent content words. The distinction between the lexical stressed syllables of the prominent and non-prominent content words is promising for further analysis. In future research we will also concentrate on additional feature such as measurements concerning pitch and intensity, and on classification with a classifier such as an artificial neural network and a feed forward neural network.

References

REFERENCES

[1]. Campbell, N. and Beckman, M., “Stress, prominence, and spectral tilt”, In: Intonation: Theory, models and applications, 1997.

[2]. Breen, M., Fedorenko, E., Wagner, M. and Gibson, E., “Acoustic Correlates of Information Structure”, Language and Cognitive Processes 25.7, pp. 1044-1098, 2010.

[3]. Lee, Y. C. and Xu, Y., “Phonetic realization of contrastive focus in Korean”, In: Proc. Speech Prosody 2010, Illinois, 2010.

[4].Mixdorff, H., Cossio-Mercado, C., Hönemann, A., Gurlekian, J., Evin, D. and Torres, H., “Acoustic Correlates of Perceived Syllable Prominence in German”, In: Proc. Sixteenth Annual Conference of the International Speech Communication Association, Dresden, 2015.

[5] Swerts, M., Krahmer, E. and Avesani, C., "Prosodic marking of information status in Dutch and Italian: A comparative analysis", Journal of Phonetics, vol. 30, no. 4, pp. 629-654, 2002.

[6].Genzel, S. and Kügler, F., “The prosodic expression of contrast in Hindi”, In: Proc. 5th International Conference of Speech Prosody, Chicago, USA, pp. 1-4, May 2010.

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

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

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

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

[11] Yardi, V., “Teaching English pure vowels to Marathi learners: some suggestions”, ELT Journal. 4, pp. 303-307, 1978.

[12] Rao, G. and Srichand, J., “Word Boundary Detection using Pitch Variations”, In: Proc. International Conference on Spoken Language Processing, Philadelphia, 1996.

[13] Rajendran, S. and Yegnanarayana, B., “Word boundary hypothesization for continuous speech in Hindi based on F0 patterns”, Speech Communication, 18(1), pp. 21-46, 1996.

[14] Pandey, P., “Indian English Prosody”, In: Leitner, G., Hashim, A., and Wolf, H. (Eds.), Communicating with Asia, Cambridge University Press, Cambridge, 2015.

[15] Dyrud, L. O., “Hindi-Urdu: stress accent or non-stress accent?”, Ph.D. thesis, University of North Dakota, 2001.

[16] Dhongde, R. V. and Wali, K., “Marathi”, Vol. 13, In: John Benjamins Publishing, 2009.

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

[18] Barbertje M. streefkerk, “Towards Finding Optimal Features of Perceived Prominence”

[19].Portele. T. and Heuft B., “Towards a prominence based synthesis system”. Speech Communication, 21:61-71.1997

[20] Mayer, J., Intonation und Bedeutung, Ph.D. Thesis, Phonetik AIMS, Stuttgart, Vol. 3, 1997.

[21] Ladd, D. J., Intonational Phonology, Cambridge University Press, 1996.

[22] Terken, J., "Fundamental frequency and perceived prominence of accented syllables", Journal of the Acoustical Society of America 89: 1768-1776, 1991.

[23]‘T Hart J., Collier R. and Cohen A. 1990. A perceptual study of intonation. Cambridge, University Press.

[24] Kompe, R. 1997. Prosody in Speech Understanding Systems, Lecture Notes for Artificial Intelligence. Springer-Verlag, Berlin.

[25].Wightman, C. W. and Ostendorf, M. 1994. Automatic labeling of prosodic patterns. IEEE Transactions on Speech and Audio Processing, 2, 469-481.

[26].Streetfkerk, B. M. and Pols, L. C. W. 1998. Prominence in read aloud Dutch sentences as marked by naive listeners. Tagungsband KONVENS-98, Frankfurt a.M., 201-205.