A Study on Detection of Intonation Events of Assamese Speech Required for Tilt Model

Parismita Sarma
Research Scholar
Department of Information
Technology Gauahti University,
Guwahati,

Sikhar Kumar Sarma
Professor
Cotton College State University
Guwahati, Assam, India

ABSTRACT

This paper has done a study and experimental analysis on different intonation events of Assamese speech. Assamese is a North East Indian language and spoken by lacks of people in India. The researchers need intonation model to identify language specific intonation events, which are necessary for synthesis process of that particular language. The paper shows outcomes of some experiments done with different speech software and observes the intonation behavior like fundamental frequency, duration and boundary tone on segment of Assamese utterances. This paper is on Tilt intonation model and phonology of Assamese language which is necessary for speech synthesis. Some unique characteristics of Assamese utterances and syllable level behavior are also described in this paper.

Keywords

Intonation, speech synthesis, fundamental frequency, duration

1. INTRODUCTION

In the process of text to speech synthesis of a language we have to prepare prosody. Phonetic module preparation is a phase in speech synthesis and prosody have to generate in this phase. Prosodic modeling involves generation of F0 contour, accent prediction, prosodic phrasing etc[1]. There are two tasks associated with intonation prediction. They are accent prediction and F0 contour realization. Starting from initial intonation, many improvement taking place regarding supervised and unsupervised model of intonation. Quality of synthesis speech is growing day by day and at the same time, demand of more sophisticated intonation model is also increasing. Synthesis of Assamese language is based on unit selection concatenate synthesis method. Regarding this synthesis process more concern is given on the fundamental frequency (F0) of units and duration values of them. These parameters are rationally necessary to get natural sounding speech[2]. Tilt model helps to identify these phonemic characteristics. Without rhythm or tone any kind of speech becomes monotonic, just speaking like a robot. To get natural sounding speech, intonation of rhythm of individual units should be taken care of. Phonetic module is a phase of speech synthesis procedure, it entails phonemic significance of script[3]. Phonetic module also includes prosodic modeling. The module deals with F0 prediction and contour generation, prosodic phrase and accent. Tilt is a tool used for event detection in speech synthesis process. Tilt converts different intonation events to some parametric sequential events with help of some conversional methods. The model consists of a collection of functions that form a library. Fundamental functions of tilt are analysis of text, synthesis and rectification or modification of the values received from raw speech wave.

2. ASSAMESE LANGUAGE

Assam is situated in the north east part of India. Assamese is largely spoken language in that region. Origin of Assamese language lies mostly in Indo-European family. Dr. B. K. Kakoti mentioned in his PhD. thesis that there are a huge number of words exported from Indo Chinese family. This language is also associated with Indo-Aryan family[4]. Eight numbers of vowel phonemes, twenty one consonant phonemes and many diphthongs are there in the language[5]. Assamese has a number of characteristics of its own. It has no any retroflex pronunciation, but this sound is very much found in southern part of India. Velar nasal /ŋ/ is frequently found in many words of Assamese language. /w/ phoneme is extensively used in Assamese language. Most important feature of Assamese is that use of velar fricative /x/. This phoneme is totally absent in any other Indian languages.

3. SOME INTONATIONAL CHARACTERISTICS

3.1 Tone

Tone can be defined as assets of a syllable. In other words it is a type of pitch movement. Ancient Indian Vedic languages were tonal languages[6]. Tone and its specific structure is able to express meaning of a sentence. Sentence, phrases even individual word can carry different meaning depending upon tone involved in them. Tone is prominent on syllable of a word. Application of different types of tone to China Tibetan languages is unique characteristics of those languages. In a sentence or word some syllabi become more prominent due to application of tone and are capable of expressing different meaning. There are four types of tones found in different context of a sentence. Vocal cord vibration has crucial impact upon intonation of a language. When vocal cord vibrates and it grows up to maximum that type of tone is called as rising tone or acute. When residual vowel goes on decreasing and decreasing then that type of tone is called as falling tone. Sometimes vowel sounds rise up suddenly and immediately goes down, it is called as circumflex tone[7]. On the other hand neutral tone, which does not rise or fall is called as level tone. Intonation model has to predict speech parameters from the written text. Written text does not have any information regarding stressed syllable or F0 contour. Important thing is that prediction about accented syllable and type of accent should be correct to design a good intonation model. Next prediction is about F0 contour generation. If accent or tone is known, generation of F0 contour is not difficult. For example the Assamese sentence “ৰাতুল আজি আহিব ” (Ratul will come today) can be uttered in three different moods. In normal mood sentence gives an information that Ratul will come. The F0 contour for normal mood is shown in fig. 1. If the same sentence is converted for question asking, F0 contour will be as shown in fig. 2. It is seen that the normal sentence is changing its meaning to “Is it Ratul who will come today itself ? ”. In surprise or exclamatory mood the same sentence will express emotion as “ Ratul will come today ! very surprise!!”.

3.2 Stress

All phones or syllabi of an utterance are not equally stressed. It is the relative air pressure when we pronounce a text. Highly stressed units are more louder. Continuant sounds can bear stress. Stop sounds have very little stress bearing capacity. Vowel phonemes are more capable for heavy stress. Some languages have relatively more stressed units, phonemes, syllabi and words. Every syllable of a word will have its own specific stress, this phenomenon is called as word stress. This stress for some language always occurs at fixed location. Displacement of stress from its original position confers different meaning to the whole sentence. English is that type of language. Stress level in English is clearly detectable. In case of Assamese sentences the whole sentence express prominent stress level. This is the reason English people hear Assamese conversation as continuous series of utterances. For Assamese language, upper Assam dialogue which has same pronunciation mapping with written script, and has accented stress on middle phoneme compared to first and last phoneme[8]. The relative stress putting on individual words is called as sentence stress. Some words may be more prominent for a particular meaning and has more accented syllable.

4. DIFFERENT INTONATION MODEL

Intonation model was discussed from the time of speech synthesis started to function, many intonation models are designed by different researchers at different times. Still today we can think about designing a new model for detection of exact intonation event. It is a challenging area in speech research. One of them is tone sequence model. Tone sequence model is also called as linear model. According to this model, text from left to right is changed to a sequence of values. British school is one example of linear model which is derived from auditory analysis. Another is Pierrehumbert 1980 model, based on acoustic analysis of speech signal[9]. Phonetic base another model is Tilt-Taylor 1998. We are discussing in our paper about this model. Hierarchical models are also available, designed for intonation detection. A hierarchical model works with help of a group of partial modules generated separately for sentence, syllable, word and phone. At final stage all these modules have to combine together. Fujisaki 1983 is one of these type of model.

5. EXPERIMENTAL RESULTS AND ANALYSIS

In this paper researchers have used the Assamese sentence “**ৰাতুল আজি আহিব** ” (Ratul will come today) to discuss different intonation events. We have selected this sentence as it is a phonetically and grammatically rich sentence and able to express different moods by changing stress level on constituent syllabi. Recording of the sentence was done in a noise free multimedia system with a female radio announcer. A reliable syllabification algorithm was used to syllabify every word of the sentence. Syllable is a unit of utterance which play most important rule in concatenate speech synthesis process. It may be single phoneme or collection of phonemes which can be pronounced in one breath or can be pronounced with a resonance. Every syllable must have one sonority, is called as peak of the sound and other phonemes around it. The phoneme with sonority is called syllable nucleus. Table1 shows F0 or fundamental frequency of individual syllable in three different moods.

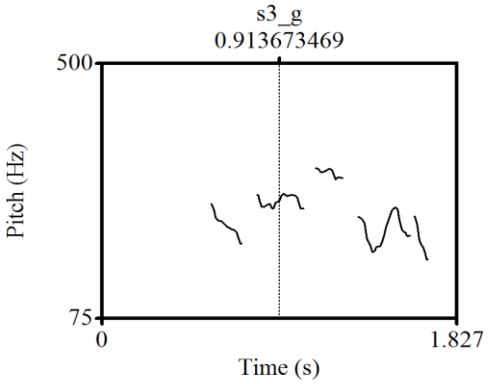

Fig1: Normal mood F0 contour of “ৰাতুল আজি আহিব”

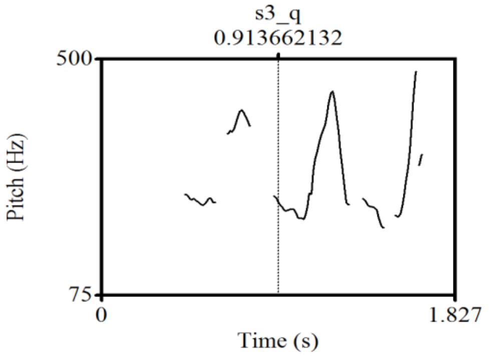

Fig2: Questioning F0 contour of “ৰাতুল আজি আহিব?”

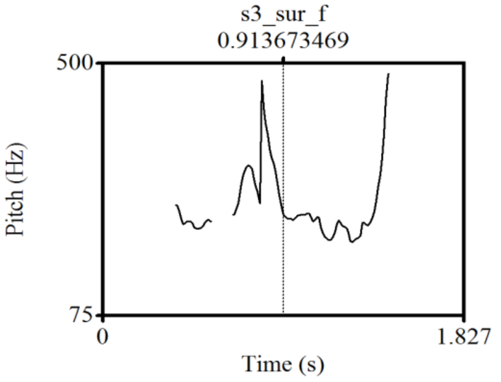

Fig3: Exclamatory F0 contour of “ৰাতুল আজি আহিব !”

Table1: F0 (Fundamental frequency) of individual syllable

Syllable	Normal	Questioning	Exclamatory
ৰা	231.08	238.37	227.36
তুল	268.58	398.15	355.44
আ	280.68	245.99	330.32
জি	230	257.41	243.65
আ	234.84	314.66	215.18
হি	223.45	244.15	234.2
ব	205.05	387.67	333.28

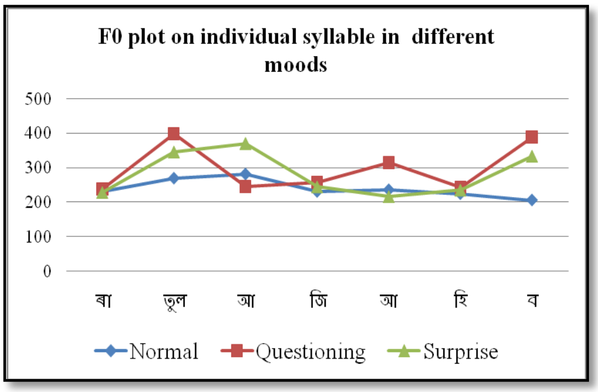

Fig 4: Plot of F0 in normal, questioning and exclamatory mood

Accented syllable of the sentence can be identified from graph in Fig 4. It is a case study for intonation of Assamese speech basically done for Tilt model. Fig1 shows the F0 contour of the sentence in normal mode. As in the sentence there are mixed types of phonemes (some phoneme does not have F0 contour), we get a discontinued F0 contour graph. Fig 2 shows the F0 contour of the same sentence in questioning mood. The contour has a rising end at end. Fig3 is F0 contour of surprise mood of the sentence. This contour also has some accented words and syllabi to express rising and falling level of exclaim nation. On the other hand syllable level F0 contour are shown in the figures 5, 6,7 and 8. Significance of these F0 contour can be studied from their pitch values as shown in the table 1. Fig. 5 is F0 contour of syllable /tul/(তুল) of name Ratul in question asking mood and fig. 6 is F0 contour for the same syllable in surprise mood. This syllable is accented for giving more stress on the particular name. Fig. 7 is F0 contour of syllable /a/ (আ) of the word (আজি) in question asking mood and fig. 8 is F0 contour of the same syllable in surprise mood. From table1 we can understand that /tul/(তুল) syllable of word “Ratul” is accented in question asking and surprise mood. They have high values compared to other syllable’s F0 values. It is obvious from fig. 5 and fig. 6. Like this another syllable /a/ (আ) of word “আজি” is also accented and it is noticeable from fig. 7 and fig. 8. When level of F0 value is changed in a syllable, meaning of the whole sentence also changes accordingly. Accurate detection of F0 value in a syllable is a confronting task. Fig 4 shows that at end of both question asking mood and surprise mood the last syllable /b/ (ব) is rising up or accented. It happens because in both of these tones people rise up their voice to express appropriate emotions.

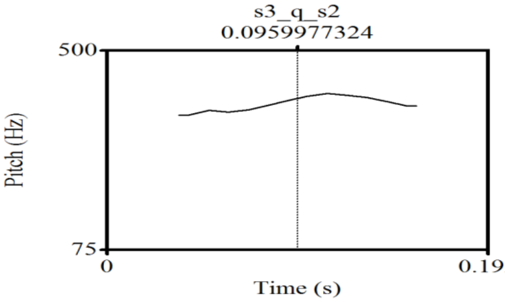

Fig 5: F0 contour of syllable /tul/(তুল) in word ৰাতুল (Ratul) in questing mood

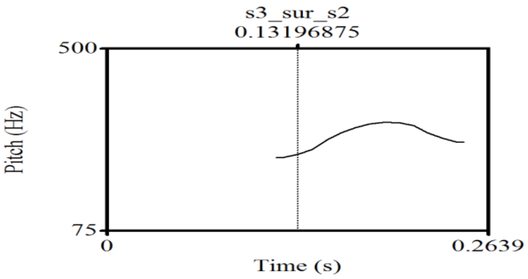

Fig 6: F0 contour of syllable /tul/(তুল) in word ৰাতুল in exclamatory/surprise mood

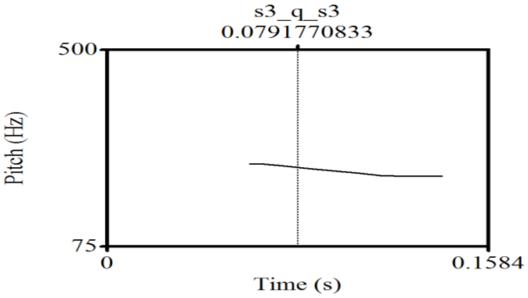

Fig 7: F0 contour of syllable /a/(আ) in word আজি in Questioning mood

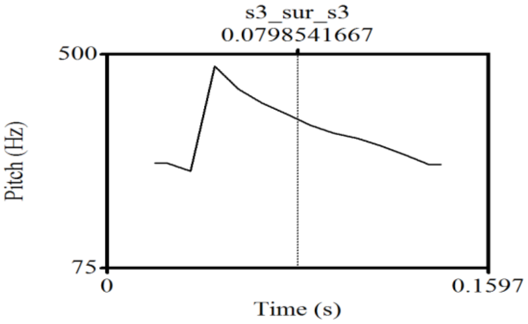

Fig 8: F0 contour of syllable /a/(আ) in word আজি in exclamatory/surprise mood

From the F0 contour graphs in all the above figures one thing becomes fairly understandable that vowel phonemes has great tendency for rising F0 values. Study in these experiments shows that prominent differences are observed mainly in the vowels and plosive consonants (here /t/ and /b/). According to Assamese phonology whole primary stress is seen in the vowels. Only vowels can act as syllable nucleus, not the consonants. Some remarkable differences between Assamese vowels and consonants are explained by Assamese language experts. These statements are useful for building any intonation model. According to them vowels can be pronounced without taking help of other phones. They may be voiced or unvoiced. Consonants (except the hissing sounds) cannot be pronounced without help of vowels. Vowels can be pronounced to a long extant at a time but consonant cannot (except the hiss sound). Vowels are audible from distance, consonants are none. Any intonation model should take care of these linguistic features for natural speech output.

6. EXTRACTION OF F0

We know that F0 plays the most important role in intonation modeling. But it is not an easy task to detect F0 in an utterance correctly. This is the core or fundamental tone in speech. The range of F0 are different for male and female. For female this value is comparatively more than male. The shape of F0 is different for different piece of utterances. F0 value reveals different amplitude, duration and position in sentences, words and syllable accordingly. In the experiments it is noticed F0 contours with different shapes. These shapes are not smooth lines or curves. Sometimes they are disconnected in between. It happens due to presence of different types of phones in a syllable or word. Extraction of F0 values from raw wave are not always recommended. Accurate F0 values can be detected if it can be measured at the articulator part glottis, where sound is produced. A number of corresponding electrical movement can be detected at glottis and from those parameters F0 value can be measured almost correctly. Unvoiced utterance do not have F0 values, to make up these discontinuity smoothing is required in F0 contour of words and sentences. Generally interpolated values are used for them. Some windows are made up with fixed duration and some values are assigned to them. Thus an approximate value of F0 can be assumed for the whole utterance.

7. STRUCTURE OF TILT MODEL

Tilt works in the concept of Intonation Event. This model is concerned about the events pitch accents, which is signified by ‘a’ and boundary tones denoted by ‘b’. Event ‘a’ is amount of F0 value of prominent syllable of words. Event b is related with raised F0 at boundary or edge of intonation. This b gives information about whether the intonation is a continuation, a question or emotion. Combining these two, another event ‘ab’ is declared. When a and b are occurring together, one of their less significant event can be ignored. It means any one of pitch accent and boundary tone can be ignored. This event occurs due to closeness of individual events a and b. According to tilt model only three parameters are enough to express intonation at syllable level. They are amplitude, duration and tilt value which is evaluated from the other two. Tilt is a supervised model. Annotated data are required for this supervised method. Hidden Markov Model(HMM) is used internally to detect the events.

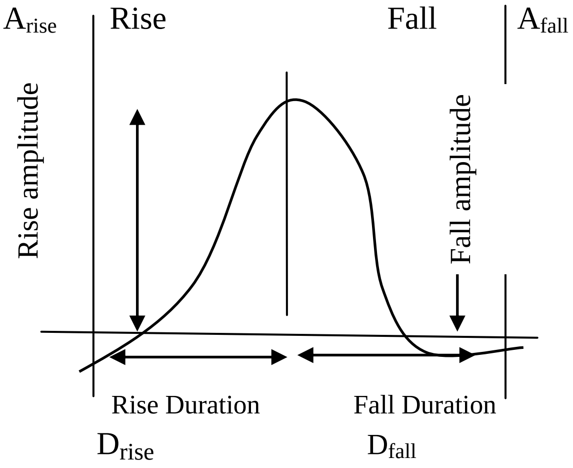

Fig 9: Mathematical explanation of Tilt model

As shown in Fig.9 for an accented syllable all the above mentioned parameters are measured according to mathematical expression. Amplitude is total sum of rise and fall amplitude of the accented syllable. Rise duration is duration of the accented syllable when it rise up. Fall duration is the time period when the syllable falls down to be flat. Together they gives total duration of the accented portion. On the other hand tilt parameter is a number, which conveys the event’s shape. Tilt is calculated as shown below.

$$Tilt = \frac{|A_{rise} - A_{fall}|}{2(A_{rise} + A_{fall})} + \frac{|D_{rise} - D_{fall}|}{2(D_{rise} + D_{fall})}$$

The above mentioned parameters are same as shown in fig. 9. In the tilt model corpus have to label methodically such that parameters can be evaluated properly for appropriate prosody design. Accent and boundary should be detectable and measurable. General hand labelling faces some difficulties. Sometimes manual perception gives the impression as there is a accent, but really it may not be there. This type of confusion gives rise to a new labelling called ’minor’ accent. When there is a rise there will certainly have a fall. Tilt is aware of this fact. But in hand labelling both rising and falling boundary are identified and these events are denoted by **rb** and **fb** [10]. Provision is kept to remove **fb** later. Tilt works well for Assamese synthesizer. One of the result of application of the model on an Assamese speech unit is shown in this paper. Model is processed on already stored data and build up of parameters are done from speech wave form. From the F0 contour tilt parameters can be directly derived. Duration and amplitude are two important tilt parameters. If F0 contour is determined then duration and amplitude is quite obvious to get. Tilt suppose to response for some basic events. Every event is isolated and they execute individually. In the definition of F0 contour. For most of the Indian languages including Assamese, more stress is given to prominent syllable of the word. Assamese is a language with pragmatic tune in the sentences and different tunes can give different meaning for the same sentence. As tilt function on every syllable, F0 value of every syllable can be described with four values. F0 values are transformed to another domain that is to log domain to segregate the components to constituent parts. Another component of the model called phrase is build like CART tree(Classification and Regression Trees). K number of accents are considered to evaluate the exact contour of the syllable. Tilt is build on bottom up approach. Starting from syllabic level to sentence level F0 contour is generated. Building of F0 contour is main aim of any intonation model. It helps in building natural sounding speech. The following segment is a result of tilt model application on an Assamese utterance. It is cited here to make functioning and result of tilt model application more understandable for the reader.

```
(Output)(IntEvent)

#0.60000  24  sil  ;  tilt_start_f0
0.0000 ;

0.91000  24  b  ;  tilt_start_f0
130.5240 ;

1.05000  24  a  ;  tilt_start_f0
103.7130 ; tilt_amplitude 23.2280 ;
tilt_duration  0.2200  ;  tilt_tilt
1.0000  ; tilt_peak_pos  0.0000  ;
syllink "4" ;

1.13000 24 c ; tilt_start_f0 130.90
;

1.37000  24  fb  ;  tilt_start_f0
133.4880 ; tilt_amplitude 31.2580 ;
tilt_duration 0.2500 ; tilt_tilt -
```

```
0.4220 ; tilt_peak_pos 0.0000 ;
syllink "4" ;

1.45000 24 a ; tilt_start_f0
104.0200 ;

1.64000 24 c ; tilt_start_f0
105.2820 ; tilt_amplitude 17.2260 ;
tilt_duration 0.2300 ; tilt_tilt
1.0000 ; tilt_peak_pos 0.0000 ;
syllink "6.11" ;
```

8. CONCLUSION

In this paper we have discussed some important intonation characteristics of speech units like syllable, words and sentences for Assamese language. Discussion is according to tilt model which was proposed during 2000 and we have done some exercise on Assamese language. As intonation characteristics are most important features for prosody design and if they can be understood well then it can be compared with other Indian languages and speech to speech synthesis will be easier. From this study we can comprehend similarity and dissimilarity of intonation characteristics of utterances of a language with others. Appropriate understanding will encourage to design more intonation model using new technology. Speech synthesis itself is a challenging assignment and designing a correct intonation model is not a easy task. Important thing to mind is that, tilt model has no significance from linguistic point of view. The model works strictly under mathematical rule but it can not detect linguistic consequences. This is the main drawback of tilt model. Detection of fundamental frequency and duration of individual syllable is also difficult. Sometimes it may leads to wrong assumptions. So in future we are hopeful of removing the drawbacks of tilt model and will try to make a more efficient and accurate revised tilt model. Tilt is a good model as it uses less input parameters for assumption of its intonation parameters.

9. ACKNOWLEDGEMENTS

Writer’s special thanks goes to Department of Information Technology, Gauhati university for offering helping hands to carry out different experiments from time to time. Secondly she want to thank her Guide Prof S.K. Sarma Sir for giving valuable suggestion to make the paper ready for submission.

10. REFERENCES

[1]. Thomas S. 2007 Natural Sounding Text-To-Speech Synthesis Based On Syllable-Like Units.

[2]. Anumanchipalli G.K., Oliveira L. C., Black A. W. 2011 “A Statistical Phrase/Accent Model for Intonation Modelling”

[3]. Taylor. P. 2000 “Analysis And Synthesis of Intonation Using the Tilt Model”

[4]. Kakati B. 2007 “Assamese its formation and development” 5th ed.. Guwahati, India.

[5]. Goswami G.C. 1982 “Structure of Assamese” FIRST EDITION, Department of Publication, GAUHATI UNIVERSITY.

[6]. Goswami G.C. 1982 “Structure of Assamese” FIRST EDITION, Department of Publication, GAUHATI UNIVERSITY

[7]. K . Sangramsing , M. Monica, G. Jayesh. “Hidden Markov Model based Speech Synthesis: A Review International Journal of Computer Applications” (0975 – 8887) Volume 130 –No.3, November 2014

[8]. Kakati B. 2007 “Assamese its formation and development” 5th ed.. Guwahati, India.

[9]. Oliver. D. 2002, “Modelling Polish Intonation for Speech Synthesis”

[10]. Anumanchipalli G.K., Oliveira L. C., Black A. W. 2011 “A Statistical Phrase/Accent Model for Intonation Modelling”