RESEARCH ARTICLE OPEN ACCESS

# Automatic Syllabification Rules for ASSAMESE Language

Laba Kr. Thakuria[1], Prof. P.H. Talukdar[2]
Department of Instrumentation & USIC, Gauhati University Guwahati, India
Department of Instrumentation & USIC, Gauhati University Guwahati, India

***Abstract***
For unit selection based text-to-speech system, syllabification acts as a backbone. Based on different structures of different languages syllabification rules are also varies. The purpose of this study is to examine and analyse the syllabification rules for Assamese language. Imparting education and training, preferably, in the local/regional language is urgently necessary in today's context in order to maintain social harmony and homogeneity. Language heterogeneity is a global problem in bringing all the benefits of Information Technology (IT) to our doorsteps. Syllabification rules are implemented into an algorithm which later can be integrated into a text-to-speech system. The analysis of these rules has been taken using 10000 phonetically rich words which reports to produce a comparable result of 99% accuracy as compared to manual syllabification.
***Keywords*:** Assamese language, diphthongs, phonemes, syllable, text-to-speech.

## I. Introduction

Syllable is a unit of sound which is larger than phoneme and smaller than a word **[**5**]**. Syllabification algorithms are mainly used in text-to-speech (TTS) systems in producing natural sounding speech, and in speech recognizers in detecting out-of-vocabulary words**[**4**]**. Syllable forms as a gap between a phonemes and words **[**1**]**. Various attempts have been made to define a syllable earlier. According to phonetics it is defined with respect to its articulation whereas in phonology it is simply termed as a sequence of phonemes **[**3**]**.When a word is broken down into its syllables the process is called syllabification. As we humans also syllabify a word, as far as possible before speaking if possible and phonemic segmentation if not, text-to-speech systems usually use syllable based approach as a basic unit **[**6**]**. A syllable based text-to-speech system performs always better than a phoneme level approach in terms of naturalness and easy in boundary analysis.In this study there is an honest endeavour to syllabify ASSAMESE language as well as implemented an algorithm which further automatically divides words into its constituent syllables.It is the aim of the proposed paper to study the phonetic variations of Assamese language to develop syllabification rules for Assamese language. the algorithm was tested using 5000 phonetically rich words. The degree of accuracy of the algorithm is 99%.

The research paper is organized as Assamese Language and its Phonological structure. It also describes the Family tree of the Assamese Language. Then it describes the Syllable structure and its algorithm with rules. This section also includes its working methodology.

## II. Assamese Language And Its Phonological Structure

Assamese is an Indo-Aryan language spoken by the Assamese people in general. The mixed Aryan culture and the mongoloid culture gave birth to a new culture. So, every community from this region always exhibits their indigenous culture with diversity. It is the link language for the people living in Assam and adjoining states of Arunachal Pradesh, Meghalaya, Nagaland etc. This language has come from Sanskrit as its offshoot, through different stages of development.

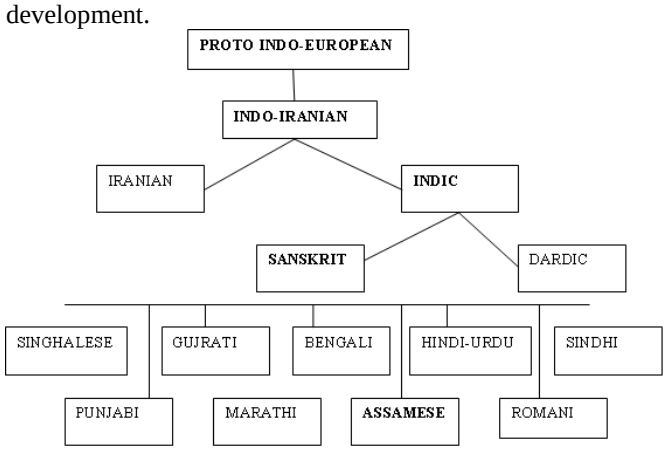

Fig-1: Proto Indo-European Family Tree

The ASSAMESE phonemic inventory consists of eight vowels, ten diphthongs, twenty-one consonants and two semi vowels **[**3**]**. The ASSAMESE vowels and consonants are shown in the tables below.

**Table 1:** Vowels in ASSAMESE.

| | Front | Central | Back |
|---|---|---|---|
| **Close** | i | | u û |
| **Half-close** | e ê | | o ô |
| **Half-open** | ɛ | | ɔ |
| **Open** | | A | |

The vowel sounds in Assamese Language occur in all the three positions, namely word initially, medially, and finally. Examples are shown below.

**Table 2:** Examples of Vowel Sounds

| Monophthongs | Initially | Medially | Finally |
|---|---|---|---|
| /i/ | /i/ 'he' | /aaitaa/ 'grandmother' | /xi/ 'he' |
| /e/ | /ek/ 'one ' | /deutaa/ 'father | /de/ 'give' |
| / ɛ / | /ɛ taa/ 'one piece' | /bɛ l/ 'kind of Fruit' | /kɛ nɛ / 'how' |
| /a/ | /aam/ 'mango' | /bal/ 'child' | /kalaa/ 'black' |
| /u/ | /ur/ 'fly' | /phul/'flower' | /saku/ 'eye' |
| /û/ | / ûkani/ 'leech' | /bûl/ 'color' | |
| /o/ | /osar/ 'near' | /gondha/ 'smell' | /lo/ 'iron' |
| / ɔ / | / ɔ kanmaan/ 'little' | /bɔ l/ 'strength' | /lɔ /e 'take' |

There are ten diphthongs in Assamese Language as follows:

**Table 3:** Diphthongs in Assamese Language

| Diphthongs | Initially | Medially | Finally |
|---|---|---|---|
| /ai/ | /aai/ 'mother' | /kaait/ 'thorn' | /paai/ 'to reach' |
| /ei/ | /eitu/ 'this ' | /xeitu/ 'that' | /dei/ 'okay' |
| / oi / | /oi/ 'interjection' | /ghoiniiyek/ 'wife' | /ekaanabboi/ 'ninety one' |
| / ɔ i/ | / ɔ inya / 'somebody else' | /p ɔ itaa / 'cockroch' | /h ɔ i / 'yes' |
| /ui/ | /ui/ 'white ant' | | /dui/ 'two' |
| /iu/ | | | |
| /ou/ | / ouxadh / 'medicine' | / aaloukik / 'spiritual' | /mou/ 'honey' |
| / au/ | / auxi/ 'no moon day' | /paautaa/ 'one who get' | /xaau/ 'curse' |
| /eu/ | | /deuri/'personal title' | /deu/ 'priest' |
| /ua/ | /uaar/ 'cover' | /guaal/ 'milkman' | |

There are twenty-three consonant sounds including two semi-vowels in Assamese Language.

**Table 4:** Consonants in ASSAMESE

| Nature of Articulation | Bi-labial | Alveolar | Palatal | Velar | Glottal |
|---|---|---|---|---|---|
| **Plosive** | p b | t d | | K g | |
| | ph bɦ | $t^{h}$ $\underset{ɦ}{d}$ | | $k^{h}$ $\underset{ɦ}{g}$ | |
| **Fricative** | | s z | | x | ɦ |
| **Nasal** | M | n | | ŋ | |
| **Lateral** | | l | | | |
| **Rolled** | | r l | | | |
| **Semi Vowel** | w | | j | | |

## III. Syllabification

Syllables are considered as the set of smallest speech sound in a language that distinguishes one word from another. The key objective of this study is to set the syllabification rules for ASSAMESE language using an algorithm which syllabify a word. In syllabification the location of some syllables in word plays a great role. Syllable shows different behaviour in a form of articulation features and pronunciation if it occurs in word initial, final or intermediary position**[**7**]**. In this research work the syllabification is actually achieved by looking at different patterns of syllables meaning that we can experiment with monosyllabic, disyllabic, trisyllabic and polysyllabic.

Thus, the purpose of this work is to formulate an algorithm for dividing Assamese words into syllables.

## IV. Methodology

The methodologies followed in the present study are:

a) Examine Assamese syllables structures from linguistic literature.**[**6**]**
b) Group discussions with linguists.
c) Dialect variations.
d) Study of previous works performed **[**2**]**.

## V. Subjects

Data was obtained from the Assamese Dictionary (**Hemkosha**) and also from the native speakers of the respective language and was syllabified by five different speakers between 20 and 25 years of age participated in the study. The entire test was done on two male speakers and three female speakers respectively. On the basis of these experiments almost all the syllable templates available in Assamese were found. About 5000

monosyllabic, disyllabic and polysyllabic words were experimented to find out the syllable structure in Assamese. For the words of Assamese having different structure see the Appendix A.

## VI. Procedure

The entire five subjects were asked to record the words of Assamese language, which were selected for contain almost all the syllable templates. The recording was done in a quiet room with a noise cancelling microphone using the recording facilities of a typical multimedia computer system. The boundaries of syllable were marked according to the pronunciation of the team. The voice was recorded and processed on Audacity and Cool Edit Pro. By observing formants through spectrogram, syllable structure and boundary were evaluated. The equipment used in recording was a microphone with a frequency response of 48000 Hz with a 16 bit sound card and high quality speakers.

## VII. Result

The recording was done on 5000 distinct words extracted from a Assamese corpus and then compared with manual syllabification of the same words to measure accuracy. Heterogeneous nature of texts obtained from the News Paper, Feature Articles Text books, Radio news, short stories etc**[**2**]**. A list of distinct words of 5000 most frequently occurring words chosen for testing the algorithm. The 5000 words results some **20,655 syllables** .The algorithm achieves an overall accuracy of **98.05%** when compared with the same words manually syllabified by an expert.

### A. Syl*lable Struc*ture

Syllable structure of ASSAMESE phonemes of vowels (V) and consonants(C) are as follows:-

1. V
2. VC
3. CV
4. CVC
5. CVV
6. CCVC
7. CCV
8. VCC

### B. *Syllabification rules*

An assumption is taken in this rule that diphthongs are considered as a single vowel unit. The syllabification rules developed after study are given below:-

i. The syllable boundary of a word having single vowel is at the end of the word. (V, CV)
ii. For a word having CVCC* structure, then syllable boundary is marked after first vowel. ie CV/C, CV/CVV, CV/CV,CV/CVC etc
iii. For a word having VCV* structure , then syllable boundary is marked after SECOND vowel. ie VCV/, VCV/C, VCV/CV, VCV/CVCetc
iv. For a word having CVV* structure , then syllable boundary is marked after SECOND vowel. ie CVV/,CVV/C,CVV/CV,CVV/CVC,CVV/V etc
v. For a word having VCC* structure , then syllable boundary is marked after second CONSONANT. ie VCC/VCV, VCC/V, VCC/VC etc
vi. For a word having CCV structure , then NO syllable boundary is marked. SYLLABIFICATION is only makred after completion of the word. ie CCV/
vii. For a word having CCVC* structure , then syllable boundary is marked after first vowel. ie CCV/C, CCV/CV, CCV/CVC,CCV/CVV etc
viii. For a word having CVCV structure , then syllable boundary is marked after first vowel. ie CV/CV, CV/CVV etc
ix. For a word having CVVCX structure , then CHECK X
   A) If x=vowel(V) ie CVVCV, boundary will found after second vowel
   B) If x=consonant(C) ie CVVCC , boundary will found between the two vowels.

### C. *Implementation*

All the above rules were implemented as a recursive function in a C++ environment which checks the presence of syllables in the sub word till the word boundary is reached. The words are processed from left to right.

## VIII. Algorithm

In this section, the Bodo syllabification rules identified in the section (4.2) are presented in the form of a formal *algorithm*. The function ***syllabify()*** accepts an array of phonemes generated, along with a variable called ***current_index*** which is used to determine the position of the given array currently being processed by the algorithm. Initially the ***current_index*** variable will be initialized to 0. The ***syllabify()*** functionis called recursively until all phonemes in the array are processed.

The function ***mark_syllable_boundary(position)*** will mark the syllable boundaries of an accepted array of phonemes. The other functions used within the ***syllabify ()*** function are described below.

i) ***total_vowels(phonemes):*** accepts an array of phonemes and returns the number of vowels contained in that array.

ii) ***is_vowel(phoneme):*** accepts a phoneme and returns true if the given phoneme is a vowel.
iii) **total_vowels_bet_cons ():** accepts a word with current position and returns the total number of vowels till the next consonant from current position. It also returns the length of the total vowels, length of the first vowel and length of the first consonant, after current position, as reference.
iv) ***total_cons_bet_vowels ()*** : accepts a word with Current position and returns the total number of Consonants till the next vowel from current position. It also returns the length of the first consonants, after current position, as reference.
v) ***get_syllables () :*** accepts a word with current position, syllable boundary and returns the syllable.

## IX. Conclusion

The above algorithm was tested on 5000 distinct Assamese Words, collected from Assamese corpus of dictionary (**Hemkosha**), short stories and news and compared it with manual syllabification we get the result of about 20,665 syllables with accuracy of 99%.This paper is presented an algorithm for dividing Assamese words into syllables. The algorithm is based on grammatical rules proposed in section (4.2) and does not require high computational cost. Syllabification is an important component of many speech and language processing system**[4]**, and it is expected that this algorithm is going to be a significant contribution to the researchers working on various aspects of the Assamese language.

## X. Acknowledgement

We would like to express my sincere gratitude and heartfelt thanks to my guide Prof. Pran Hari Talukdar, Professor, Department of Instrumentation and USIC, Gauhati University. we would not have been able to complete the research work and shape it in the form of the research paper without his consistent advice, and never ending enthusiasm, positivity, encouragement, support and understanding. we are very fortunate for having an opportunity to work with him from which I benefited enormously. we are also grateful to Gauhati University for giving me to access the data from GU_Galo_Adi corpus.

## References

[1]. ChandanSarma, U.Sharma, C.K.Nath, S.Kalita, P.H.Talukdar. **Selection of Units and Development of Speech Database for Natural Sounding Bodo TTS System**, *CISP Guwahati*, March 2012.
[2]. Jyotismita Talukdar, Chandan Sarma, Prof. P.H Talukdar. **Automatic Syllabification Rules for Bodo Language**, *International Journal Of Computational Engineering Research*, vol. 2 issue. 6, pp 110-114, October 2012.
[3]. Banikanta Kakati. **Assamese, its formation and development.**
[4]. Ruvan Weerasinghe, Asanka Wasala and Kumudu Gamage. **A Rule Based Syllabification Algorithm for Sinhala**
[5]. Parminder Singh, Gurpreet Singh Lehal. **Syllables Selection for the Development of Speech Database for Punjabi TTS System**, *IJCSI International Journal of Computer Science Issues*, vol. 7, Issue 6, November 2010.
[6]. Nungleppam Gopil Singh**,** Purnendu Acharjee, Prof. P. H. Talukdar**. An Improved Automatic Syllabification Rules for BODO Language**. *International Journal of Computing, Communications and Networking*, vol. 1, No.3, November-December 2012.
[7]. Jehangir zaman khan. **Syllabification rules in pashto**
[8]. Heriberto Cuay´ahuitl. **A Syllabification Algorithm for Spanish**
[9]. Upendra Nath Goswami. **An Introduction to Assamese**
[10]. Hemchandra Barua. **HEMKOSHA**
[11]. Couto I., Neto N., Tadaiesky, V. Klautau, A. Maia, R.2010. **An open source HMM-based text-to-speech system for Brazilian Portuguese**, *in Proc. 7th International Telecommunications Symposium Manaus*.
[12]. Kevin Hock, Julian Smart, Stefan Csomor. **Cross-Platform GUI Programming with Wxwidgets**, Pearson edition, 2006.
[13]. Juliette Blevins,**The syllable in phonological theory**,1995
[14]. George Kiraz and Bernd M¨obius. **Multilingual syllabification using weighted finite-state transducers, i**n Proc. *of the 3rd Workshop on Speech Synthesis*, 1998.
[15]. Robert Damper. **Learning about speech from data: Beyond NETtalk**, *in Data-Driven Techniques in Speech Synthesis*, pp 1–25, 2001

**Appendix A: Word List**

**Table 5** : Word list

| Assamese | Roman | Structure | Type |
|---|---|---|---|
| এই | ei | V | monosyllabic |
| এক | ek | Vc | monosyllabic |
| এ | e | V | monosyllabic |
| ঊধ্ | uudh | Vc | monosyllabic |
| ঊণ্ | uun | Vc | monosyllabic |
| ঊস্ | uux | Vc | monosyllabic |
| কৈছিল | koisil | cv/cvc | Disyllabic |
| কৈছে | koise | cv/cv | Disyllabic |
| গৈছিল | goisil | cv/cvc | Disyllabic |
| থৈছিল | thoisil | cv/cvc | Disyllabic |
| থৈছে | thoise | cv/cv | Disyllabic |
| দূৰৈৰ | duuroir | cv/cvc | Disyllabic |
| হৈছে | hoise | cv/cv | Disyllabic |
| এখৈৰা | ekhoiwaa | v/cv/cv | Trisyllabic |
| কৈছিলা | koisilaa | cv/cv/cv | Trisyllabic |
| কৈছিলে | koisile | cv/cv/cv | Trisyllabic |
| ঘৈণীয়েক | ghoiniiyek | cv/cv/cvc | Trisyllabic |
| তোলৈকে | toloike | cv/cv/cv | Trisyllabic |
| দৈৱযোগ | doiwajog | cv/cv/cvc | Trisyllabic |
| এপৰলৈকে | eparaloike | v/cv/cv/cv/cv | polysyllabic |
| তেতিয়ালৈকে | tetiyaaloike | cv/cv/cv/cv/cv | polysyllabic |