ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

An Improved Grapheme to Phoneme rules for Assamese Language

Bitopi Sharma[1], PurnenduBikash Acharjee[2], Prof. P.H. Talukdar[3]

1. Deptt of Instrumentation & USIC, Gauhati University, India,
2. Asian Institute of Management and Technology, Guwahati, India,
3. Deptt of Instrumentation & USIC, Gauhati University, India,

Manuscript Info

Manuscript History:

Received: 14 October 2015
Final Accepted: 22 November 2015
Published Online: December 2015

Key words:

G2P, Schwa, TTS, syllable

****Corresponding Author***

Bitopi Sharma

Abstract

This paper deals with the study and development of the grapheme to phoneme rules for the Assamese language. Like every other Indian languages, Assamese Language also shows some regular mapping from graphemes-to phonemes. It has a systematic relationship between the written form of a word and its pronunciation. So, we preferred to write down letter-to-sound rules by hand.In the study we found some fundaments rules. Grapheme to phoneme conversion rules are the roots to the proper text to speech system.

INTRODUCTION

Grapheme to phoneme conversionis defined in the words of SittichaiJiampojamarn[1] as – "Given an input words *s* containing *n* graphemes, $s_1, s_2 .... s_n$, the task is to find the $t_1, t_2, ... t_m$phonemes sequence that corresponds to the input word *s"*.

A grapheme is the smallest semantically distinguishing structural writing unit in of any language and a phoneme is the smallest structural unit of a language. The terms grapheme and phoneme are analogous to each other, a phoneme is an abstraction where as a grapheme is the physical writing or appearance.A phoneme is considered as a basic distinctive unit of speech sound by which morphemes, word and sentences are represented. Combination of two graphemes, known as diagraph, may also generate a phoneme. A grapheme may or may not carry meaning by itself, and may or may not correspond to a single phoneme. Grapheme to phoneme (G2P) conversion is a process of generating pronunciation rules for a word. G2P rules differ for every language. This is a process of converting a sequence of letters into a sequence of phonemes. G2P rules are the base to develop a good and efficient TTS system.

Notation:

- Graphemes are often notated as <a>, <b> etc.
- Phonemes are often notated as /a/, /b/ etc.
- Phonetic transcriptions are often notated as [a], [b].

1. Assamese Language and its structure

Assam located in the foothills in Himalayan range, popularly known as the land of red rivers and blue hills. Assamese is an old language developing since ancient times having its roots from Sanskrit, Orriya and Magadhi Prakrit. Today's modern Assamese language has a lot of influence from local dialects.

Structure of Assamese Language

The Assamese phonemic inventory consists of eight oral vowel phonemes, three nasalized vowel phonemes, fifteen diphthongs and twenty-one consonant phonemes. The syllable is a single or a group of phonemes in a word articulate in single chest pulse. It may consist of one vowel and a consonant or more. It may be a word by itself or a part of the word. It may be meaningful or meaningless. A syllable ending in a vowel is called open and a syllable ended in a consonant is called closed.

Assamese words may be of two types monosyllabic that is having only one syllable and polysyllable that is having more than one syllable. The Assamese language have 11 vowels and 40 consonants.

In Assamese language phoneme chart contains less alphabets. The tables below list the Assamese phonetic chart[2]-

Assamese Phoneme	IPA	Assamese Letter	IPA
ক<k>	/k/	ক	/k/
খ<kh>	/kʰ/	খ,ক্ষ	/kʰ/
গ<g>	/g/	গ	/g/
ঘ<gh>	/gɦ/	ঘ	/gɦ/
ত<t>	/t̪/	ত,ট,ৎ	/t̪/
থ<th>	/t̪ʰ/	থ,ঠ	/t̪ʰ/
দ<d>	/d̪/	দ,ড	/d̪/
ধ<dh>	/d̪ɦ/	ধ,ঢ	/d̪ɦ/
প<p>	/p/	প	/p/
ফ<ph>	/pʰ/	ফ	/pʰ/
ব<b>	/b/	ব,ৱ	/b/,/ʋ /
ভ<bh>	/bɦ/	ভ	/bɦ/
ম<m>	/m/	ম	/m/
ন<n>	/n/	ন,ণ,ঞ	/n/,/n/,/ɲ/
ঙ<ng>	/ŋ /	ঙ,ং	/ŋ /
স<x>	/x/	চ,ছ,শ,স,ষ	/s//s/,/x//x//x/
য<j>	/dʒ/	জ,ঝ,য়	/dʒ/,/dʒɦ//dʒ/
ষ<x>	/x/	শ,স,ষ	/x/,/x/,/x/
হ<h>	/ɦ /	হ,ঃ	/ɦ/,
ৰ<r>	/r/	ৰ,ড়,ঋ	/r/,/r/
ল<l>	/l/	ল	/l/

Table1: Phoneme chart of Assamese Language

2. METHODOLOGY

Phonological Approach, hand written is taken as the main approach for the formulation of Grapheme to Phoneme Rule for Assamese Language[3].The formulation of the Grapheme rule the following steps have been followed.

a) Systematic analysis between the selected texts from thecorpus and the recorded speech.

b) Formulating the Hand written rule.

c) Verification of the rule by a linguistic expert.

3. SYLLABLE STRUCTURE OF ASSAMESE LANGUAGE

Assamese Language is official language of Assam, India. It is an easternmost Indo-Aryan language. Its development can be traced at least before 7th century AD from MagadhiPrakritand Sanskrit. The Assamese phoneme inventory consists of eight vowels and twenty one consonants, fifteen diphthongs are attested. In a word three syllables may appear in successioncomprising of five vowels [4].Phonemes segments of the vowels(V) and the consonants(C) describes syllables. The Assamese syllable structures may be divided into the following types based on the distribution of the segmental phonemes-

a. V
b. VV
c. VC
d. CV
e. V/CV
f. (CVC)*
g. CV/V
h. (CV)*/V
i. (CV)*/CVC

4. FUNDAMENTAL G2P RULES OF ASSAMESE LANGUAGE

In Assamese language based on the pronunciation the phonemes (letters) and the graphemes are classified [5][6] –

a. অ /ɔ̱/, আ /a̱ː/, ক /k/, খ /kʰ̱/গ /g/, ঘ /gɦ̱/, ঙ /ŋ̱/, হ /ɦ̱/ are the letters that are pronounced by vocal track vibration and are classified as glottalsounds popularly known as kanthabarna(কণ্ঠযুৰ্ণ)in assamese.
b. ই /i̱/, ঈ /i̱ː̱ /, এ /e/, ঐ /oj/, চ /s̱/, ছ /s̱/, জ /dʒ/, ঝ /dʒ̱ɦ̱/ ঞ /ɲ/, ষ /x/, শ /x/ are the sound pronounced by the articulation of the tip of the tongue towards the alveolar ridge and are classified as alveolar sound popularly known as talbarna (তালব্যবর্ণ)in assamese.
c. ট /t̪ /,ঠ /t̪ʰ /,ড /d̪/, ঢ /d̪ɦ/, ণ /n/, ৰ /r/,য /dʒ/ are the sounds pronounced by the articulation of the tip of the tongue toward the back of the teets and are classified as dental consonant sound popularly known as dantabarna (দন্ত্যুৰ্ণ) in assamese.
d. উ /u /, ঊ /̱ʊ̱ /, ও /o/, ঔ /oʊ̱/, প /p/, ফ /pʰ̱/, ব /b/, ভ /bɦ / ম /m/ are the sounds pronounced by the articulation of lips against each other and are classified as bilabial sound popularly khown as ushabarna (ওষ্ঠ্যবৰ্ণ)in assamese.
e. ঙ /ŋ̱/, ঞ /ɲ/,ন /n/, ম /m/, \$^0$ are the sounds pronounced by the articulation of the tongue between the teeth and are classified as interdental consonant sound popularly known as anunaxikbarna (অনুনাসিকবৰ্ণও) in assamese.

These are the general grapheme to phoneme rules for Assamese language. These rule are followed while pronunciation of a Bodo word. They are[3][5] —

a. In assamese language অ /a/ is pronounced in two ways –
   i. Tenseness way: Tenseness or laxness classification of articulation is based on the position of the tongue root. For example :
   অকলৈ /ɔ̱//k//ɔ//l//oj/
   অলপ/ɔ̱//l//ɔ̱//p//ɔ̱/
   ii. Lip rounding way: Lip roundedness refers to the amount of rounding in the lips during the articulation of the vowel. For Example:
   অগুণ /ɔ̱//g//ʊ̱//n//ɔ̱/
   অতীজত/ɔ̱//t̪//i̱//dʒ//t̪/

b. Some word ending with consonants do not pronounce the অ /ɔ/ sound. For example: ঘৰ্ /gɦ̱//r/
   আম্ /a̱ː̱ //m/
c. Some word ending with consonants do pronounce the অ/ɔ̱/ sound. For example: পাৰ /p//a̱ː//r//ɔ̱/
   বাট /b//a̱ː//t̪//ɔ̱/
d. The অ/ɔ̱/ sound is pronounced if a word ends with conjuncts. For example: দুষ্ট্ /d̪//ʊ̱//x//t̪//ɔ̱/
   বৈদ্য /b//e//d̪//dʒ//ɔ̱/
e. The অ/ɔ̱/ sound is pronounced if a word ends with ৰ /ʊ̱/. For example: জীৰ /dʒ//i/ /ʊ̱//ɔ̱/.
f. The অ/ɔ̱/ sound is pronounced if ◌ং(Anuswar) and ◌ঃ(bisarga) is followed by a consonant at the end.

For example: কংশ /k/ /ŋ/ /ɔ/, দুঃখ /d̪/hq /kʰ/

g. In assameseঈ /i/ and ঊ /ʊ/ is prounced as /i/ and /ʊ/.

h. In assamese language এ /e/ is pronounced in two ways –

iii. Tenseness way: Tenseness or laxness classification of articulation is based on the position of the tongue root. For example :
এক/e//k//ɔ/
এঙ্গাৰ /e//ŋ//aː/ /r//ɔ/

iv. Lip rounding way: Lip roundedness refers to the amount of rounding in the lips during the articulation of the vowel. For Example:
এই/e//i/
এতিয়া/e/t̪//i//j//aː/

i. In Assamese language চ /s / and ছ/s / is pronounced Sanskrit স /s̪/.

j. In Assamese language জ and ঝ is pronounced as /dʒ/

k. In assamese language ন and ণ is pronounced as /n/

l. If a word ends with য /dʒ/ or it is in the middle of the word then is pronounced as অ /ɔ/. Example: জয়

m. When অ /ɔ/ is not pronounced with ৱ/ʊ/ then it is sometimes pronounced as উ/u/.

Example: জীৱ is pronounced as জীউ.

n. Sometimes ৱ/ʊ/ is directly pronounced as ও/o/.

Example: পাৱ is pronounced as পাও.

o. In assamese language শ,ষ,স /x/ is pronounced as soft হ /ɦ/.

p. In Assamese language ক্ষ/kʰ / is pronounced as খ/kʰ /.

In Assamese language for every conjunct the pronunciation varies. The general grapheme to phoneme rules for conjuncts in Assamese language is –

a. If ঙ /ŋ/ precede গ /g/ then the গ /g/sound is not pronounced. Example:ভাঙ্গিলে is pronounced as ভাঙিলে.

b. If জ/dʒ/ precede ঞ/ɲ/ then গ্যঁ sound is pronounced. Example:জ্ঞান is pronounced as গ্যঁন

c. If ঞ/ɲ/ precede জ/dʒ/ then ন /n/ sound is pronounced. Example: আঞ্জা is pronounced as আনজা.

d. If ঞ/ɲ/ precede চ or ছ /s/ then ন /n/ sound is pronounced. Example: কাঞ্চন is pronounced as কানচন, লাঞ্ছনা is pronounced as লানছনা.

e. If য /dʒ/ succeed any consonant then হ /ɦ/ sound is pronounced.Example: অন্য is pronounced as অইন ধন্য is pronounced as ধইন.

f. If শ,ষ,স /x/ is conjunct with any consonant then the sound pronounced as in Sanskrit.

Example: বিশ্বাস, অষ্ঠমী

In Assamese language deep orthographic depth is observed that is no difference is noticed in pronunciation of two similar words but both are written with different graphemes having different meanings.

Example: 1. কুঁজ means hump of the back and কুজ means the planet march.

2. গিৰি means mountain and গিৰী means head of specific position such as nation, family and so on.

3. ঠাল means small branch of tree and থাল means platter.

In Assamese, word having same grapheme combination and same phoneme may derive different meaning depending on the use of the word.

Example 1: The word খৰ have same phonology and graphemes but derives different meaning depending on its use.

Sentence 1: ৰাম কামত খৰ

(Ram completes task quickly)

Sentence2: ৰামৰ খৰ হৈছে
(Ram is suffering from a ringworm skin disease)
In the first sentence the word means quick, fast, so forth. In the second sentence it means a kind of skin disease.

Example 2: The word বৰ have same phonology and graphemes but derives different meaning depending on its use.
Example 1: সিতাৰ বৰ ৰাম
(Ram is Sita's husband)
Example 2: সিতা বৰ ধুনীয়া
(Sita is very beautiful)
In the first sentence the word বৰ means groom. In the second sentence it acts as an adjective very.

Like every other Indian languages, Assamese Language also shows some regular mapping from graphemes-to phonemes. It has a systematic relationship between the written form of a word and its pronunciation. So, we preferred to write down letter-to-sound rules by hand

Nonsense Rule

A nonsense word is a syllable or group of syllables that can pronounced based on the phonetic rules of a language but which transmit no meaning to a reader or listener. Such words appear in a variety of literary contexts but generally only exist because of the sound of the nonsense word. In a speech synthesis process, detection of such cases is important for the naturalness of the synthesizer. The Basic nonsense rule for Assamese Language are-

i. No word in Assamese starts with / ঙ /ŋ /,ং /ŋ /, and ঃ. ৱ /ʋ /, য় /j/.
ii. No word in Assamese ends with হ /ɦ /.
iii. The diphthongs /ɔi/, /ui/, /ɯi/, /iu/ never occur at the starting of a word.

5. RULES FOR SCHWA DELETION OR RETENTION

The schwa is the vowel sound in many lightly pronounced unaccented syllables in words of more than one syllable. It is also known as neutral-vowel. Like every language Assamese has also schwa deletion or retention issue stated as follows-

1. The schwa of the first syllable is never deleted.
2. If the word ends with the consonants like ৱ/ʋ /, then the associated schwa is retained.
3. If য়/j/ comes after a syllable which consists higher vowels like ই/i/ or উ/u/ than the associated schwa is retained.
4. The schwa preceding a full vowel is retained to maintain lexical distinctions.
5. If the last syllable of the word contains a schwa and contexts 1 through 4 described above for the retention of the schwa do not occur, then the schwa is to be deleted.

6. CONCLUSION

In this paper I have tried to discuss about the basic Grapheme to Phoneme rules for the Assamese Language. During the formulation of the rules Assamese corpora of about 7000 phonetically rich words has been used. The words are selected from continuous texts of Assamese newspaper, story-book, articles etc. All these rules are tested in a HTS (A HMM based Speech Synthesis Process). All these rules were able to produce 90% correct pronunciation of the Assamese words. After the error analysis, it was found that the 10% error was due to the non-pure Assamese words, which came into Assamese Language as an influence of other languages

REFERENCES

SittichaiJiampojamarn, Grapheme-to-phoneme conversion and its application to transliteration.
Dr. GolukchandraGoswami,Axomiyabiyakaranarmoulik bisar,1987, reprint Seventh Edition August 2011 P66.

Satyanath bora, bahalbyaakaran, September 2012.
http://www.iitg.ernet.in/rcilts/pdf/assamese.pdf .P7.
SantanuKoushikBaruah, AdhunikAxomiyarosonaxamagrabyakaran, jotuwathamqs, phakarajojanaa, sithi-patra,daliladixambalit,second revised edition 2011, P 516.
Dr. Pran Hari Talukdar, Speech Production, Analysis and Coding Introduction to speech processing,P11-18.
Nabankur Pathak, Prof. P.H. Talukdar The Basic Grapheme to Phoneme (G2P) Rules for Bodo Language.