Quarterly Programmatic Report

Text-to-Speech Synthesizer for Indian Languages

March 2013
to
May 2013

Contents

1. Languages Identified

Eight Indic languages were identified for inclusion and improvisation on pronunciation in espeak text to speech synthesiser. Effort began with the version 1.47.08. The languages were:

- Bengali: Inclusion and improvisation of pronunciation
- Hindi: Improvisation of pronunciation by redefining phoneme strings
- Malayalam: Improvisation of pronunciation by redefining phoneme strings
- Telugu: Improvisation of pronunciation by redefining phoneme strings
- Assamese: Inclusion and improvisation of pronunciation. Effort began with the version 1.47.11
- Manipuri: Inclusion and improvisation of pronunciation. Effort began with the version 1.47.11
- Oriya: Inclusion and improvisation of pronunciation. Effort began with the version 1.47.11
- Sindhi: Inclusion and improvisation of pronunciation. Effort began with the version 1.47.11

Bengali dictionary was created with Hindi as a base dictionary. The phoneme rules were initially followed those of Hindi but later redefined and written by the developer, based on the research by us with the help of local and external language experts.

Hindi pronunciations, especially that of the numbers and conjunctives were found far from appropriate, even after considering the synthesising limitation of espeak. Work on modification of phoneme strings, defining the rules of pronunciation, modification of phoneme sounds were the principal task accomplished with Hindi. The input from experts and scholars were the basis of suggestions on modification sent to the developer during the period.

Malayalam pronunciation rules for numbers following hundreds and above were not correct. The basic rules of change of the words of such numbers were not present. The phoneme sounds for certain phonemes too were not appropriate or close to an acceptable level. Considering the variation of pronunciation based on various dialect of the language, a general method of defining specific phoneme strings those are intelligible was the primary objective of improvisation and modification of this language in espeak.

Telugu phonemes inherited from ph Hindi with the same phoneme for the inherent vowel. The phoneme rules and subsequent phoneme strings were modified based on the feedback from local and external experts, native speakers including former user of other TTS software developed for visually impaired people. The individual vowel and consonants pronunciation, barring a few, were intelligible; however pronunciation of combined words or sentences was the primary issue and task for the team to work upon.

Assamese was inherited from Bengali with 'hi' as base. The ‘ph’ source file Assamese defined the vowel and consonants sounds of Assamese, many of which had same sound as Bengali excluding the additional consonants in Assamese. The phoneme sounds for Assamese to be defined or modified as well as the rule for inherent vowel in Assamese. The chandrabindu in Assamese has a very important role unlike Bengali and Oriya. The rule defining pronunciation that affects the meaning because of chandrabindu is very critical. Discussion and interaction regarding these rules are the principal concern while working on Assamese.

Manipuri too was inherited from Bengali. As of today, Manipuri is still written with Bengali alphabets, but are subject to adopt Meitei script with an absolute different pronunciation pattern. The rule and list dictionaries will be different from that of Bengali files. As per the remarks from Jonathan "*Manipuri and Mizo are "tone languages" which may make them more difficult. espeak has some features for tone languages (for example the espeak Vietnamese voice)*". This would call for phoneme sound modification and import of other non English and non Indic languages phonemes. Present Manipuri script (written in Bengali) does not indicate tone. Manipuri dictionary files are same as that of Bengali to the extent of alphabets only. The inherent vowel or the phoneme sounds are absolutely different from that of Bengali. Experts are being consulted on this particular issue.

Oriya inherited from Hindi. The pronunciation rules are a mixture of that of Hindi and Bengali. The inherent vowel rule is similar to Bengali except certain cases. Oriya being in the same group of Bengali languages, indo Aryan - eastern branch, the phoneme sounds for Bengali are applicable for this language too. The challenges anticipated, were ambiguity of inherent vowel rule similar to Bengali rules. The team was consulted to advice on the near appropriate pronunciation of the individual vowels and consonants. The issue of non native voice synthesisation is also an important obstacle for a better Oriya output through espeak.

Sindhi uses two different scripts. The Urdu script that is widely used across Pakistan and parts of India and the Devanagari script (used for Hindi) that is in use in India. We adopted the Devanagari version of script for two basic reasons. Devanagari script is more in use in India, and secondly, the existing Hindi dictionary with modifications in phoneme rules could be used for Sindhi in espeak.

The pronunciation rule for Sindhi differs from Hindi and has a similarity with other Indian languages, Guajarati and Punjabi, since the region where Sindhi is most spoken are close to this parts of India. Inheriting Hindi dictionary for Sindhi has one more advantages as it follows the similar phoneme sounds for the vowels and consonants like the Hindi vowels and consonants.

The above dictionaries for Assamese, Manipuri, Oriya and Sindhi were created in espeak 1.47.11. with defining common phoneme strings among the similar languages like Assamese and Bengali, Hindi and Oriya and Sindhi.

2. Language specific adoption

Bengali, Hindi, Malayalam, Telugu

All the Indian languages mentioned above inherited phoneme files from Hindi and the rules for each of them were defined / redefined gradually based on the research and feedbacks. Even though Bengali and Hindi belong to a different group of languages, Indo Aryan group, whereas Malayalam and Telugu belong to Dravidian languages, all of them contain similar alphabets or more precisely similar vowels and consonants. There are additional vowels and consonants for which phoneme strings were defined and in certain cases phoneme sounds were modified to achieve a near correct pronunciation of those vowels and consonant.

Hindi and Bengali have similar phonemes but there are vowels those are common to both the languages have short, long, suppressed or aspirated version, thus arising a need for a specific rules to pronounce them in isolation or in combination with or without allograph based on their origin of articulation.

For example /e/ in Bengali and Hindi help to define similar pronunciation but it had to be modified as /e: / for the same pronunciation and use /e/ to define a suppressed pronunciation of the vowel.

Bengali had to import phoneme from non Indian language to define pronunciation of a rare pronunciation of the same vowel. The phoneme /&/ used to define a specific pronunciation of the same vowel that in general is represented by /e/ but changes its pronunciation similar to the Latin 'æ', when a word begins with this vowel.

Assamese, Manipuri, Oriya, Sindhi

Assamese has two additional alphabets which had a similar phoneme of the corresponding phonemes in Bengali. /r/ and /j/ the second one is not present in Bengali, in the sense it is used in Assamese. In such cases we have to follow Hindi rules to pronounce a sound like 'wo'. The specific pronunciation of 'sh' in Assamese is pronounced like /x/ similar to non Indic language phonemes.

The inherent vowel of Oriya /V/ is same as Bengali but unlike the rule defined in Bengali, the suppressed pronunciation of the inherent vowel in the words ending with it, had to be modified in Oriya. The rule had to be modified in reverse to some extent.

Manipuri list file configured based on the present Meithei terms which differs in pronunciation even if it is expressed through Bengali scripts. The phoneme /n^/ in Manipuri changes the meaning like chandrabindu in Assamese.

Initial works on the languages in the list and rules dictionaries done on local system. The redefining of numbers in the list files and a few basic pronunciation modifications based on the common phoneme string to express general words and individual alphabets and consonants could be done. The rule and phoneme strings for conjunct, complex and typical rule are under research.

3. Implementation on espeak online upgraded versions – last quarter

Bengali, Hindi, Malayalam, Telugu

Changed the sound of [t. #] for Hindi (in espeak 1.47.09).This affects most Indian languages. And also changed the sound of [t. #] and [d. #] for Telugu in espeak 1.47.10a. t. #] is ok, but not appropriate so far, for Hindi, but [d. #] is not. However the Telugu team confirmed that both the sounds somewhat improved and closer to the actual pronunciation

Other changes / modification uploaded:

1. Fixed pronunciation of "sw" at the beginning of a word.
2. Change the sound of the inherent vowel after "w".
3. Change the sound of the [t. #] phoneme ("tth") so that now has the sound of [t.] + [h]. This affects all Indian languages.
4. Changed the sound of the [c] phoneme.
5. Changed the sound of anusvara at the end of a word from [n] to [N].
6. Changed the pronunciation of voiceless double stop consonants (kk, t.t., tt, pp) to a single long stop [k:] [t.:] [t:] [p:]. This seems to match better the recorded pronunciation.
7. Changed the pronunciation of [t.t.#] to a single long [t.#:]
8. Bengali: Changed the sound of [tS#] phoneme to have more aspiration.

Assamese, Manipuri, Oriya, Sindhi

The next version of espeak after 1.47.11 is yet to be upgraded by the developer. The works and modification would be available under test voice once they are uploaded by the developer for the public version.

The main developer Jonathan has not upgraded espeak after the 1.47.11 version, which was uploaded during mid June. We are still awaiting communication from him regarding these new languages as well as queries regarding the other four Indic languages we worked earlier.

Considering Jonathan's unavailability, we are trying to find out ways to get help and assistance on technical aspect to include and improve Indic languages in espeak

Even though Bengali and Hindi belong to a different group of languages, Indo Aryan group, whereas Malayalam and Telugu belong to Dravidian languages, all of them contain similar alphabets or more precisely similar vowels and consonants. There are additional vowels and consonants for which phoneme strings were defined and in certain cases phoneme sounds were modified to achieve a near correct pronunciation of those vowels and consonant.