Quarterly Programmatic Report

Text-to-Speech Synthesizer for Indian Languages

May 2013
to
August 2013

Contents

Highlights

- Mahiti got in touch with organisations/individuals working for persons with blindness to test the voices for the languages developed in the previous quarter.
- We have done preliminary work on development of e-speak for the next set of four languages with Mahiti: Assamese, Manipuri, Oriya and Sindhi.
- We are also working with another team on development of e-Speak in Punjabi and Gujarati.

Text-to-Speech Synthesizer

Work with Mahiti

For the languages developed in the previous quarter, the Mahiti team was put in touch with organisations and individuals working for persons with blindness so that these voices could be tested along with the NVDA screen reading software. Following organisations were involved in this process:

- Saksham for Hindi.
- National Association for the Blind, Kolkata for Bengali.
- Young Power in Social Action and Visually Impaired Persons Society in Bangladesh for Bengali.
- Mr. Janardhan Naidu for Telegu.
- Chakshumathi for Malayalam.

Next Languages Identified

We began work on development of the beta version for four Indic languages in this quarter:

1. Assamese (inclusion and improvisation of pronunciation).
2. Manipuri (inclusion and improvisation of pronunciation).
3. Oriya (inclusion and improvisation of pronunciation).
4. Sindhi (inclusion and improvisation of pronunciation).

Assamese

The phonemes were inherited from Bengali with “Hindi” as the base. The phoneme source file for Assamese defined the vowel and consonant sounds, many of which had the same sound as Bengali excluding the additional consonants in Assamese. The phoneme sounds as well as the rule for inherent vowel needs to be defined. The chandrabindu in Assamese has a very important role unlike Bengali and Oriya. The rule defining pronunciation that affects the meaning because of the chandrabindu is very critical. Discussion and interaction regarding these rules were the principal concerns while working on the Assamese language.

Manipuri

The phonemes were inherited from Bengali. As of today, Manipuri is still written with Bengali alphabets but are subject to adopt Meitei script with an absolutely different pronunciation pattern. The rule and list in the dictionaries will be different from that of Bengali files. As per the remarks from Jonathan, one of the developers, "*Manipuri and Mizo are "tone languages" which may make them more difficult. E-speak has some features for tone languages (for example, the e-speak Vietnamese voice)*". This would call for phoneme sound modification and import from other non-English and non-Indic language phonemes. Present Manipuri script (written in Bengali) does not indicate tone. Manipuri dictionary is the 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

Phonemes in Oriya are inherited from Hindi. The pronunciation rules are a mixture of Hindi and Bengali. The rules for inherent vowels are similar to Bengali except in certain cases. Oriya being in the same group of Indo-Aryan languages, the phoneme sounds for Bengali is applicable in Oriya as well. Experts were consulted to advice on the near appropriate

pronunciation of individual vowels and consonants. The issue of non-native voice synthesization is also a major obstacle for a better Oriya output through e-speak.

Sindhi

The Sindhi language uses two different scripts. The Urdu script, which is widely used in Pakistan and parts of north India, and the Devanagari script used for Hindi. We adopted the Devanagari version for two basic reasons. First, the Devanagari script is widely used in India, and second, the existing Hindi dictionary with modifications in phoneme rules could be used for Sindhi in e-speak. The pronunciation rule for Sindhi differs from Hindi. It has a similarity with other Indian languages like Gujarati and Punjabi, since the region where Sindhi is widely spoken is close to these parts of India. Adopting Hindi dictionary for Sindhi has more advantages as it follows similar phoneme sounds for the vowels and consonants like the Hindi vowels and consonants. *The dictionaries for Assamese, Manipuri, Oriya and Sindhi were created in e-speak 1.47.11., by defining common phoneme strings among similar languages like Assamese and Bengali, Hindi and Oriya, and Sindhi*.

Initial work on the languages in the list and rules dictionaries was done on the 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 was also done. The rule and phoneme strings for conjunct, complex and typical rule are under research.

Implementation on e-speak Online Upgraded Versions

Assamese, Manipuri, Oriya, and Sindhi

The next version of e-speak after 1.47.11 is yet to be upgraded by the developer. The outputs 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 e-speak 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 e-speak.

Even though Bengali and Hindi belong to different group of languages (the Indo-Aryan group), and 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.

Work in Punjabi and Gujarati Languages

Mahesh Khosla has been hired to work on Punjabi and Gujarati languages. Punjabi language has already been integrated in e-speak and work on Gujarati was started in July 2013. Andhjan Kalyan Trust has been the interface organisation for testing the Gujarati language.