IJETST- Vol.||04||Issue||08||Pages 5631-5637||August||ISSN 2348-9480 2017

Open access Journal International Journal of Emerging Trends in Science and Technology

IC Value: 76.89 (Index Copernicus) Impact Factor: 4.219 DOI: https://dx.doi.org/10.18535/ijetst/v4i8.30

An approach to prepare lexicons of Assamese text for unit selection concatenation TTS

Authos

Parismita Sarma

Department of Information Technology, Gauhati University, Guwahati, Assam, India

Abstract

Text To Speech synthesizer plays a very important role in modern day Information Technology. Communication technology can be enhanced by speech technology to a greater extent. There are a number of steps in natural language processing phase of a Text to speech synthesizer. In this paper the approach to make a TTS for Indian language is shown by taking Assamese as an example. First explanation is how to get lexicons of the language then an idea of using DONLabel tool is shown. To build a speech synthesizer for any language, researcher should acquire adequate knowledge about the phonetic and acoustic characteristic of that language. In this research paper important features of Indian languages, the proposed block diagram for synthesis process, preparation of lexicons from the written text for Assamese language is explained with snapshots.

***Keywords-TTS*, synthesizer, Assamese language, natural language processing, TTS, unit selection synthesis**

1. Introduction

A naturally speaking text to speech synthesizer can be made if sufficient amount of speech as well as text resource is available for the language. Till today a number of TTS are developed by the researchers including a number of Indian languages. Assamese is a North East Indian language and millions of people speak that language. Assamese language has its root in Indo-European family. Another root lies in Indo-Iranian sub category[1]. In spite of different dialects for upper Assam and lower Assam, written script is same for both of them. Dialect variation exists among different districts of lower Assam like Kamrup, Goalpara, Borpeta.

1.1 Deliberation for Building Voices for Indian Language

It is observed that for most of the Indian languages unit selection based synthesis technique is preferred[2]. It is because of some of the unique characteristics of the Indian languages. The method is concatenative which is done over some of the speech units like phone, syllable or words at most. According to this methodology already recorded speech units are picked up and have to join them at execution time. Success of this method depends upon accessibility of speech units in the database and their consistency. It is also called as unit selection concatenative speech synthesis. We are trying with Festvox speech tool and this tool works properly with phone as well as syllable as fundamental units.

Most of the Indian languages have some common property which are proved to be useful for unit selection concatenate synthesis process.

1.2 Some Unique Features of Indian Languages

From the study we are aware that nearly every Indian languages are derived from Brahmi script. Aksharas are fundamental units of these written scripts[3]. This akshara is graphical equivalent of speech units. Aksharas is generally in V, CV, C*V format[3]. Important thing is that akshara is syllabic in nature. For this reason in case of unit selection concatenative synthesis method syllable or syllable like units are taken as units to concatenate. Later it is extended to cope up with diphone and phone.

There are total 22 recognized languages in India and all use almost same set of phones and they have almost same sound[4]. All include almost fifty phones, fifteen vowels and thirty five consonants. Among theme languages like Hindi, Marathi uses Debanagiri script. The scripts used by Assamese, Bengali, Manipuri, Udiya and Tamil, Telegu etc south Indian languages different from each other.

From sources of phonetics study on Indian languages we can say that every language is unique from syllable distribution point of view. They are also different for the prosodic characteristic like prominence, tone, accent and duration.

1.3 Types of Specific Domain Related TTS

When designing a TTS for a language, we have to be very specific about its working domain. Here domain means whether it is related with limited or specific application or an unrestricted field.

To build a TTS for limited domain only few numbers of vocabulary is needed as this type of synthesizer work for a particular purpose, for example ticket reservation or station announcement system. Unrestricted domain work for any text of the language. That is why its vocabulary size should be too vast and whole procedure of synthesis becomes complex compared to limited domain.

In this approach we are considering Festival as the platform and Festvox as the toolset to make voices.

2 Literature Review

Concatenative speech synthesis is already done for a number of Indian languages. On festival framework diphone based TTS is already built for a number of Indian languages. For diphone based synthesizer frequently used or all the diphone of a language are selected and recorded by a sophisticated recording tool. This set is then stored in a database. These units can be annotated to give more natural speech. More signal processing is used in this methodology and it outputs a robotic speech[5]. These pitfalls are removed by using syllable as basic units for concatenation. Syllables were already used to build TTS for south Indian languages. When syllable is taken as the basic unit, position of the syllable in the word plays an important role. Two syllabi when concatenated must be compatible with respect to their energy and other acoustic parameters[6].

For most of the Southern, Eastern, Northern and West Indian language TTS are already build up using unit selection concatenation procedure. Speech corpora for Hindi, Punjabi and Marathi languages were developed by C-DAC Noida. There are around 1000 sentences which are phonetically quite rich. For the other Indian languages also C-DAC Noida is working. EMILLE-CIIL is a speech corpus prepared by BNC and Asia – Radio programmers' from human natural conversation[7]. For three East Indian languages C-DAC Kolkata prepared an annotated text and speech database of 8.5 GH size in the recent year 2015 and 2016.

2.2 Some Concerns To Build Voices For Indian Languages

In Festival definition of phoneset of the concerned language is not available, the implementer has to create this phoneset. While creating it, the phoneme features should also be put on it. Next the tokenizer file is cleaned by removing special symbol connected with the phone set. Grapheme to phoneme conversion rules are created to convert text to speech form. Intonation model is used for incorporating prosodic features to individual units.

2.3 Components of TTS Framework

a. **Speech Engine**: Among many speech engine, eSpeak is commonly used user friendly engine used for speech synthesis purpose. It can respond to many Indian languages. It is a fast responder and intelligible but require naturalness. This paper is an initial part of a Assamese TTS research work which is using Festival grown up with centre for speech technology research, a open source synthesis platform. A set of APIs are used for interaction and module execution.

b. **Screen Reader**: Screen reader interprets whatever is written on the screen. This tool has to convey anything on screen to the speech engine. For windows system JAWS (Job Access With Speech) is mostly accepted screen reader.

c. **Typing Instrument**: Typing instrument is the tool used to map general keyboard arrangement to different symbols of Indian languages. Mostly used typing tool is SCIM (Smart Common Input Method).

3 Problem Definition

Fig. 1 shows the block diagram of methodology how to design an unrestricted speech synthesizer for Assamese language using unit selection technique. In this paper the researcher gives a description about how to get formatted lexicon from the Assamese text. As already discussed that most of the Indian languages including Assamese are syllabic in nature. In case of unit selection concatenated synthesizer festival as a tool gives feasible output. This method of speech synthesizer need a huge speech corpus which is generally prepared manually. According to this methods the individual speech units should be inculcated with proper prosody like time duration between adjacent units, accented pitch and related supra segmental features. Making these rich phonetic speech units are not easy, festival provides a number of speech tools with the help of which appropriate prosody can be beaten to the units. The complete Assamese TTS should render natural output for unrestricted domain text.

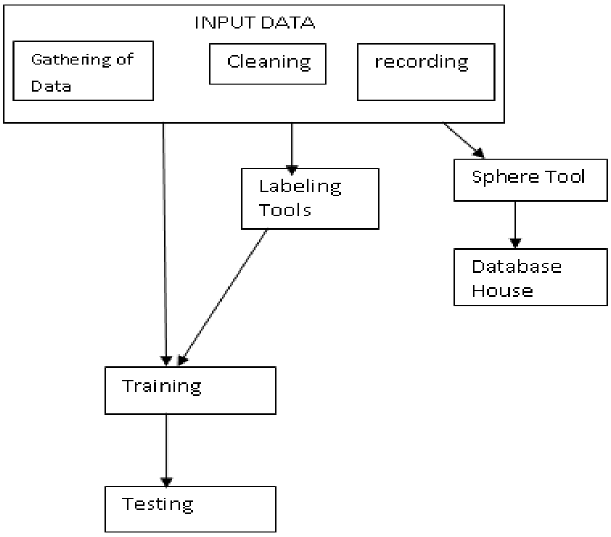


Fig. 1 Proposed problem block diagram

The above mentioned phases shown in Fig. 1 can be described as follows. From different sources like story books, news sites and speech by distinguished persons speech data are collected. In the data cleaning phase specific sentences which cover maximum number of syllabi are collected for synthesis purpose. Noise free environment and sophisticated microphone is used to record those sentences which must be uttered with necessary tone and accent. Sphere file is a repository where all the selected speech files are stored in a standard format called sphere. This type of file has a header where information about that speech file is achieved. Labelling of the speech file is necessary to detect the correct boundary of the syllabi[8]. So labelling of the files are done by using a labelling tool and then they are corrected again manually. For most of the Indian languages' unit selection synthesis Training phase is completed by transcription. Testing phase is completed by visually challenged persons.

4 Workflow Of Unit Selection Concatenated TTS

There are two phases which include all the works done by a TTS. Front End performs the text pre processing. This phase normalizes the text, expands abbreviation, conversion of numbers to date, time etc appropriate format are also done. The methodology is known as tokenization. There are other works like phonetic transcription, prosodic phrasing where marking is done on the basis of sentence and other smaller parts. In this paper we are discussing the way how phonemes of Assamese language are build from written text. This phonetic transcription phase has to pass over many tough works. It is because this phase has to prepare sound for every letter. Next phase deals with prosody generation. Text pre processing phase ends up with all the above mentioned works. The fig 2 shows the work flow of a general unit selection concatenation Text To speech synthesizer.

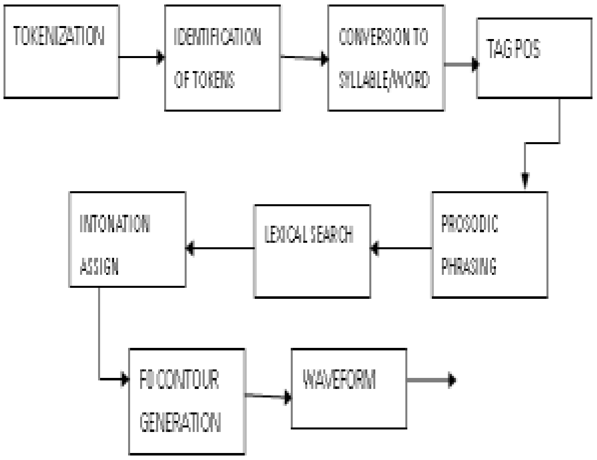


Fig. 2 Block diagram showing workflow to build TTS

4.1 Pre processing

As already mentioned, pre processing involves collection of text from different reliable sources then normalize them. Some conversion process help to acquire the required normalized text. Among all the conversion process, some of them are more important. They are conversion of number to its proper format like rupees, year, phone number etc. Acronym should be taken care of and converted to proper format. For example Mr. should be Mister, 5.6 should be five point six. Techniques are used for conversion word into syllable, diphone or phone.

Tokens are individual entity like word or parts of speech. Word segmentation means procedure of getting words of a sentence. Delimiters and all other special symbols are removed from the sentence. These words are then segregated to syllabi or phoneme or diphone. Our work is based on concatenation of syllable, hence a syllabification algorithm is used to syllabify the words.

The output of text pre processing phase for an Assamese sentence is shown below.

Input: তুমি কি কৰিছা

Output: SIL তু SIL মি SIL কি
SIL ক SIL ৰি SIL ছা
SIL (syllabification algorithm is applied)

SIL means silence, which acts like a delimiter to distinguish two words. The duration of silence must be measurable to distinguish a syllable from word and sentence. Fig. 3 shows the snapshot a number of phonetically rich Assamese text in Unicode format uttered by a female speaker. The serial number _f means it is a female spoken wav file. Fig. 4 shows the snapshot of the associated recorded wav file of the given text file.

1_f [illegible]
2_f [illegible]
3_f [illegible]
4_f [illegible]
5_f [illegible]
6_f [illegible]
7_f [illegible]
8_f [illegible]
9_f [illegible]
10_f [illegible]
11_f [illegible]
12_f [illegible]
13_f [illegible]
14_f [illegible]
15_f [illegible]
16_f [illegible]
17_f [illegible]
18_f [illegible]
19_f [illegible]
20_f [illegible]
21_f [illegible]
22_f [illegible]
23_f [illegible]
24_f [illegible]
25_f [illegible]
26_f [illegible]
27_f [illegible]
28_f [illegible]
29_f [illegible]

Fig. 3 Snapshot of selected text in Unicode format

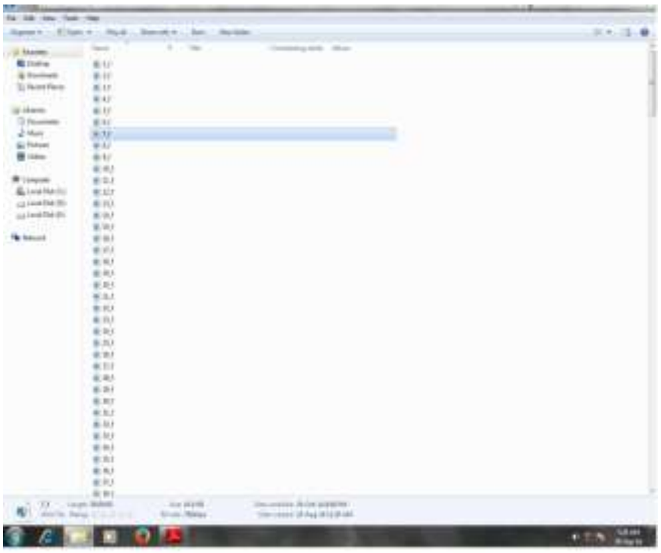
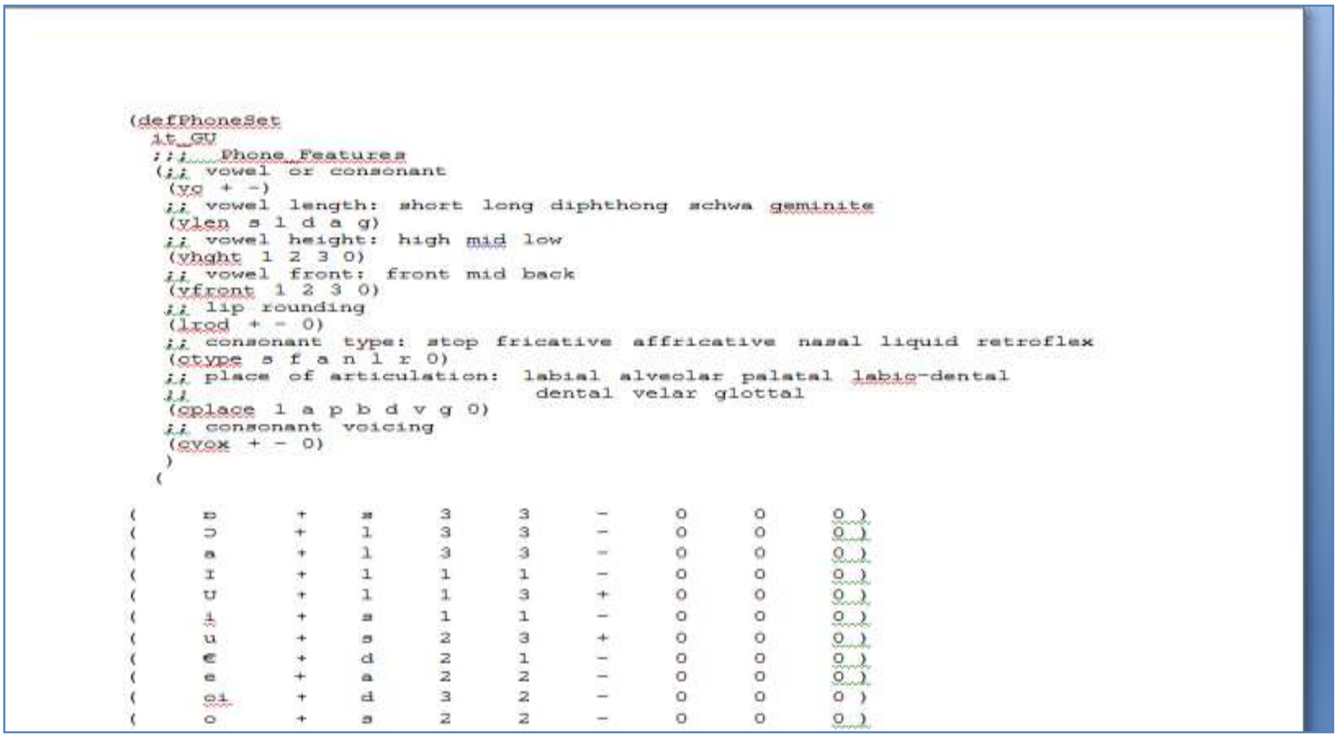


Fig. 4 Snapshot of wav file of associated text



```
(defPhoneSet
  it_GU
  ;;;   Phone_Features
  (;; vowel or consonant
   (vc + -)
   ;; vowel length: short long diphthong schwa geminite
   (vlen s l d a g)
   ;; vowel height: high mid low
   (vhght 1 2 3 0)
   ;; vowel front: front mid back
   (vfront 1 2 3 0)
   ;; lip rounding
   (lrod + - 0)
   ;; consonant type: stop fricative affricative nasal liquid retroflex
   (ctype s f a n l r 0)
   ;; place of articulation: labial alveolar palatal labio-dental
   ;;                          dental velar glottal
   (cplace l a p b d v g 0)
   ;; consonant voicing
   (cvox + - 0)
   )
  (

(     ɒ     +     s     3     3     -     0     0     0 )
(     ɔ     +     l     3     3     -     0     0     0 )
(     a     +     l     3     3     -     0     0     0 )
(     ɪ     +     l     1     1     -     0     0     0 )
(     ʊ     +     l     1     3     +     0     0     0 )
(     i     +     s     1     1     -     0     0     0 )
(     u     +     s     2     3     +     0     0     0 )
(     ɛ     +     d     2     1     -     0     0     0 )
(     e     +     a     2     2     -     0     0     0 )
(     oi    +     d     3     2     -     0     0     0 )
(     o     +     s     2     2     -     0     0     0 )
```

Fig. 5 Definition of a few Assamese vowels

4.2 Phonemic analysis

This work deals with the way a word or syllable is pronounced. Basically a lexicon is used for letter to sound mapping. The preparation of lexicon is easy if manually done. This conversion from letter to sound or grapheme to phoneme is carried out in a number of ways. The methodology are Rule based procedure, Data driven procedure and Statistical method. This approach is based on Rule based approach. A set of well defined rules are prepared for mapping purpose. It is a dictionary based way of matching letters to pronunciation. A huge Lexicon is prepared by hand. Proper pronunciation can be searched with the help of this dictionary. A newly appearing text is first searched in the dictionary, if matches then next field parts of speech is tried for matching. If both the entity matches then the pronunciation is returned. If no text matching occurs then already prepared G2P rules are discussed.

Pronunciation creation process demands the phone set of the language. The phone set of the language is needed in consequent phases of whole synthesis process. Every language has its unique phone set. They are characterized by a set of feature values of vowel, consonant, semi vowel etc. Features include vowel height and length, articulation place for a consonant pronunciation. The phone set for this work is described as shown below.

```
(definePhone
        NAME
        ATTRIBUTES
         PHONE_DEF  )
```

NAME is the symbol, ATTRIBUTES are the definition of the phonemic features. They are as discussed below.

Vowel and Consonants are denoted by V and C. They are meant by + and - . Vowel length is another property which influences pronunciation. This property name is named as vlen. This holds values like short, long, diphthong, schwa, geminate. Another parameter vowel height is denoted by vght. There are three types of vowel height found, they are high, mid and low. It is the tongue height at the time of pronunciation. Another parameter is where it is pronounced at front, mid or back. So denoted by the parameter vfront. Lip rounding is another parameter, denoted by lrod. Two types of values are possible, whether round or not. Like the vowels there are few characteristics of consonants also. Ctype parameter means the type of consonant whether it is stop, fricative, affricate, nasal, liquid or retroflex. According to place of articulation consonants are divided into labiel, alveolar, palatal, dental, laboi-dental, glottal and velar. For this parameter cplace is used. Consonants may be voicing or un voicing. Parameter used for this work is cvox. The Fig, 5 shows the snapshot of a few of the Assamese vowel with attribute values prepared for the work.

4.3 Structure of Lexicon

As already mentioned that lexicon is a place where pronunciation are searched for. Lexicon is constituent of three parts,such as Head Word, POS (Parts of Speech), list of pronunciation. Utterance of a word segregated into phoneme and syllable are stored in the list. For example an Assamese word “খেলিব” is shown below.

(“খেলিব” v (((hke) 0) ((l i) 1) ((b ɒ) 0)))

The above definition says that the word is verb, so it is denoted by v, by using syllabification algorithm or by some tool every word is syllabified. Every syllable is segregated to constitute phonemes. Next 0 or 1 is suffixed with them. 0 denotes no syllable stress is there and 1 means syllable stress is associated with. Different stress level impaired on individual syllable is important because it can determine the prosody of the words and tune as well as whole meaning of the utterance. It is seen that a word may have different meaning when parts of speech differs. For example Assamese word টান has two meanings. If it is used as a verb, it means to pull. If adjective then it means hard. That is why stress level on each syllable is different.

Depending upon the size of the database different techniques are used to tag the speech units. For a small sized database manual labelling can be preferred. But this approach results erroneous syllabification. For an unrestricted huge database, semi automatic tool DONLabel is used. DONLabel uses Group Delay algorithm to detect syllable boundary[8]. Due to almost matching phones as well as morphological structure of most of the Indian languages DONLabel tool can be applied to them. Text to be synthesized are sent to this tool in UTF 8 format. Along with this the tagged wav file is also feed to the tool as input. It produces associated labels of the wav files. As already seen that unique ids are assigned for individual text and its associated wav file, now DONLabel makes it easier to map text and wav files. For building the database another software called Pronunciation Rule Parser is attached with DONLab. Pronunciation Rule Parser has its own syllabification and phonification algorithm. Grapheme to Phoneme rules are also included with Pronunciation Rule Parser which helps to address sounds for the words in the text.

5. Conclusions

An efficient TTS can help all people from a farmer to a doctor. According to phonetic characteristics of most of the Indian languages, unit selection concatenate synthesis methodology is most suitable for them. The paper is a part of the project work, where an Assamese Text to Speech synthesizer is tried for unrestricted domain. Here a systematic manner is suggested to build an efficient TTS. This methodology shows a way how to get lexicons and the lab files by using DONLabel tool. In the later stage working on the framework will be done. In that phase ultimate speech will be produced by using festvox speech tool which is adjunct with festival open source software. Output of DONLabel tool will be input to festival software.

Acknowledgement

I am thankful to my guide Pro. S.K. Sarma sir and faculty of IT department of Gauhati University for giving me opportunity to carry on the research work in the laboratory.

References

1. Kakati Banikanta. “Assamese its formation and development” 5thed..Guwahati, India, LBS publication, 2007
2. Samuel Thomas, M. Nageshwara Rao, Hema A. Murthy, C. S. Ramalingam, Natural Sounding TTS based on Syllable-like Units., Proceedings of 14th European Signal Processing Conference, Florence, Italy, Sep 2006.
3. B. Yegnanarayana, S. Rajendran, V.R. Ramchandran and A.S. Madhukumar “*Significance Of Knowledge Sources For A Text-To-Speech System For Indian Languages”* Sadhana, pp 147-169, 1994.
4. “South asian language families,” http://http://en.wikipedia.org/wiki/File: South Asian Language Families.jpg.
5. A.W. Black and K.A. Lenzo, .Building synthetic voices., 2003, http://festvox.org/bsv/
6. N Sridhar Krishna, Partha Pratim Talukdar, Kalika Bali, and A G Ramakrishnan, “*Duration modeling for hindi text to speech*

synthesis system," in International conference on spoken language processing, South Korea, 2004.

7. Ed. S S Agrawal et al, Proc Intl. Symposium on Speech Technology and Processing Synthesis and O-COCOSDA-2004, vol II, Tata McGraw Hill, Nov 17-19,2004, New Delhi
8. Ashwin Bellur, Badri Narayanan, Raghav Krishnan, and Hema A Murthy, "*Prosody modeling for syllable-based concatenative speech synthesis for hindi and tamil*," in NCC, 2011, pp. 216–220.