IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Abstracting clean Assamese text from a mixed Assamese and English corpus

Nabamita Deb[1], Shikhar Kr. Sarma[2]

Department of IT, Gauhati University[1, 2]

Abstract: Content to discourse handling has been performed on numerous dialects both Indian and western. In an exertion to get discourse from content in setting of Assamese dialect, a point by point exploratory examination has been done; the corpus considered for this reason comprised of two dialects: Assamese and English. The Assamese sentences have been separated from the blended corpus after which the framework was prepared utilizing a phoneme set, and a couple of solid sentences were acquired.

Keywords: Assamese and English, MARY TTS, SAMPA.

I. INTRODUCTION

The system generated speech is mostly artificial and its accuracy is also questionable mostly, and lacks sentiment, the quality of such synthesizers are mostly judged on its speech output and its closeness to natural speech. A system that converts text input to a speech output, is called a text to speech system (TTS), the input is mostly raw text in the respective language, and based on the phonetic transcriptions; a speech signal is rendered as output. Typically a text to speech system can be divided into two parts viz: a front end and a back end[1]. The front end is responsible for accepting text in its raw form; and does pre-processing on the text, which is then goes through a series of steps so that the text can be given to the backend part; where the speech output is produced. Major work of the front end is to clean up the raw text into tokens, which might be sentences ,clauses and finally words; in some systems and thereafter convert the same into phonemes. Phoneme and word are both the basic types of linguistic unit, and just like words each phoneme are distinct and differ from one another , and in contrast just as words carry meaning ,phonemes don't. Taking about verbal communication; as much as phoneme and words mater; prosody (prosodic component) is an integral part it, and has helps give emotion to speech. Each language has its own unique characteristic; the front end gives the information of prosody and the phonetic transcription and divides based on a markup language. The backend, called the synthesizer is mainly responsible for producing speech output; after taking the linguistic information from the front end.

Assamese is one of the 22 recognised languages in India. It is a major language in the state of Assam and a few neighbouring areas, still it is amongst the most under-resourced dialects which need discourse applications The point of this undertaking is to create an openly accessible Assamese content to discourse framework. A uninhibitedly accessible and open-source TTS framework for Assamese dialect can significantly help human computer communication: the potential outcomes are huge – such a framework can beat the education obstruction of the normal masses, enable the outwardly disabled populace, expand the potential outcomes of enhanced man-machine communication.

MARY TTS:
MARY (Modular Architecture for Research on Speech Synthesis)[2] TTS stage is a toolbox whose point is to give the instruments and nonexclusive reusable runtime framework modules so that it can support another dialect and make new voices. The toolbox has been effectively connected to the production of British English[3], Turkish, Telugu and Mandarin Chinese dialect segments and voices. These dialects are presently upheld by MARY TTS and additionally German and US English. The toolbox can be effectively utilized to make voices in the dialects officially upheld by MARY TTS.

The voice creation toolbox is chiefly expected to be utilized by examination bunches on discourse innovation all through the world, eminently the individuals who don't have their own previous innovation yet. The toolbox is produced in Java and incorporates natural Graphical User Interface (GUI) for the vast majority of the regular assignments in the production of an engineered voice.

SAMPA
SAMPA (Speech Assessment Methods Phonetic Alphabet)[5] is a machine-coherent phonetic letter set. It was initially created under the ESPRIT venture 1541, SAM (Speech Assessment Methods) in 1987-89 by a universal gathering of phoneticians, and was connected in the first example to the European Communities dialects Danish, Dutch, English, French, German, and Italian (by 1989); later to Norwegian and Swedish (by 1992); and in this way to Greek, Portuguese, and Spanish (1993).

Under the BABEL venture, it has now been stretched out to Bulgarian, Estonian, Hungarian, Polish, and Romanian (1996). Under the aegis of COCOSDA it is would have liked to stretch out it to cover numerous different dialects (and on a fundamental level all dialects). On the activity of the OrienTel venture, Arabic, Hebrew, and Turkish have been included. Other late increments: Cantonese, Croatian, Czech, Russian, Slovenian, Thai.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

The Assamese script (অসমীয়া লিপি Ôxômiya Lipi)[4] is a written work arrangement of the Assamese dialect. It used to be the script of decision in the Brahmaputra valley for Sanskrit and in addition different dialects, for example, Bodo (now Devanagari), Khasi (now Roman), Mising (now Roman) and so forth. The script has a sum of 11 vowel letters, used to speak to the eight fundamental vowel hints of Assamese, alongside various vowel diphthongs. These are utilized as a part of both Assamese and Bengali, the two fundamental dialects utilizing the script. A percentage of the vowel letters have diverse sounds relying upon the word, and various vowel qualifications saved in the written work framework are not affirmed in that capacity in present day communicated in Assamese or Bengali. For instance, the Assamese script has two images for the vowel sound [i] and two images for the vowel sound [u]. This repetition comes from the time when this script was utilized to compose Sanskrit, a dialect that had a short [i] and a long [iː], and a short [u] and a long [uː]. These letters are safeguarded in the Assamese script with their conventional names of hôrswô i (lit. 'short i') and dirghô i (lit. 'long i'), and so forth.

The list of the phones used for assamese in this research has been given below :

Letter	SAMPA	Name of letter	Vowel sign with [kɔ] (ক)	Name of vowel sign	IPA
	kO	ô	(none)	(none)	kɔ
র	ko	o	none) or ক'	(none)	ko
	ka	a	কা	akar	ka
	ki	hôrswô i	কি	hôrswôikar	ki
	ki	dirghô i	কী	dirghôikar	ki
	ku	hôrswô u	কু	hôrswôukar	ku
	ku	dirghô u	কূ	dirghôukar	ku
	kri	ri	কৃ	rikar	kri
	kE and ke	e	কে	ekar	e and ke
	koj	ôi	কৈ	ôikar	kɔj
	kU	û	কো	ûkar	kʊ
	kOw	ôu	কৌ	ôukar	kɔw

Table1: Sampa for vowels of assamese

Letter	Name of Letter	IPA	SAMPA
ক	Kô	k	k
খ	Khô	k^h	k_h
গ	Gô	ɡ	ɡ
ঘ	Ghô	$ɡ^ɦ$	$ɡ^ɦ$
ঙ	Ngô	ŋ	N
চ	prôthôm sô	s	s
ছ	ditiyô sô	sh	sh
জ	bôrgiyô zô	z	z
ঝ	Jhô	zh	zh
ঞ	Niô	j	j
ট	murdhônyô tô	t	t
ঠ	murdhônyô thô	t^h	t_h
ড	murdhônyô dô	d	d
ঢ	murdhônyô dhô	$d^ɦ$	$d^ɦ$

DOI 10.17148/IJARCCE.2015.412107

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

ণ	murdhônyô nô	n	n
ত	dôntyô tô	t	t
থ	dôntyô thô	t^h	t_h
দ	dôntyô dô	d	d
ধ	dôntyô dhô	$d^ɦ$	$d^ɦ$
ন	dôntyô nô	n	n
প	Pô	p	p
ফ	Phô	p^h	p_h
ব	Bô	b	b
ভ	Bhô	$b^ɦ$	$b^ɦ$
ম	Mô	m	m
য	ôntôsthô zô	z	z
ৰ	Rô	ɹ	r\
ল	Lô	l	l
ৱ	Wô	w	w
শ	talôibbô xô	x~s	x~s
ষ	murdhônyô xô	x~s	x~s
স	dôntyô xô	x~s	x~s
হ	Hô	h	h
ক্ষ	Khyô	k^hj	k_hj
ড়	dôre ṛô	ɾ	4
ঢ়	đhôre ṛô	ɾ	4
য়	ôntôsthô yô	j	j

Table 2: sampa for assamese consonants

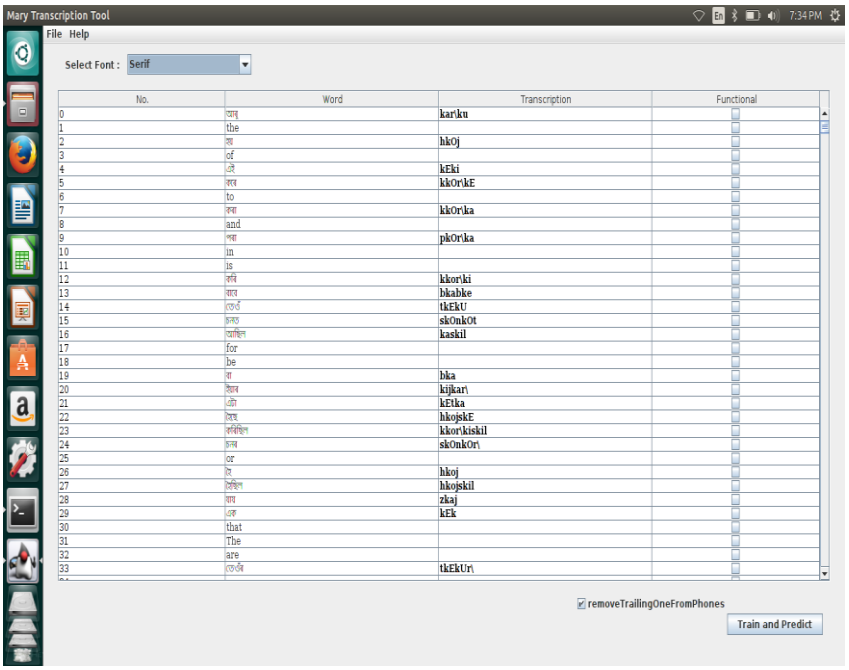

Fig1: Result of Transcription GUI in MARY TTS

DOI 10.17148/IJARCCE.2015.412107

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

II. RESULTS

The main aim of this project is to obtain reliable Assamese sentences from a mixed corpus. Initially we had to set up the environment by compiling the MARY TTS builder [6], and creating a wiki data directory for our language. Next the xml dump of Assamese language (84.9MB) was downloaded. Next we had cleant up this xml dump and found out the Assamese words and their frequency. Post this analysis we had manually given a the words found previously; their phoneme structure, next by using learing mechanism we had transcribed the rest of the words by using letter to sound rules in coherence with the underlying language. The rules were given by us using the SAMPA phoneme structure and the pronunciation of the Assamese script.Next With the files generated by the Transcription tool, we can now create a first instance of the NLP components in the TTS system for our language.Run feature maker with the minimal NLP components. The FeatureMaker program splits the clean text obtained in step 2 into sentences, classifying them as reliable or non-reliable (sentences with unknown words or strange symbols) and extracts context features from the reliable sentences. All this extracted data were kept in the database.

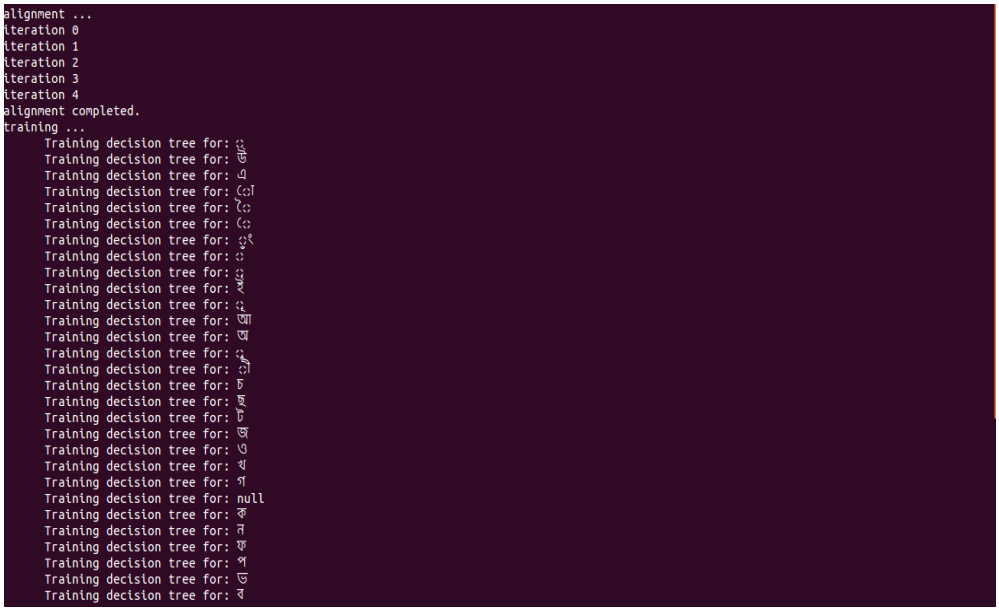

Fig2: Training decision tree for the assamese script

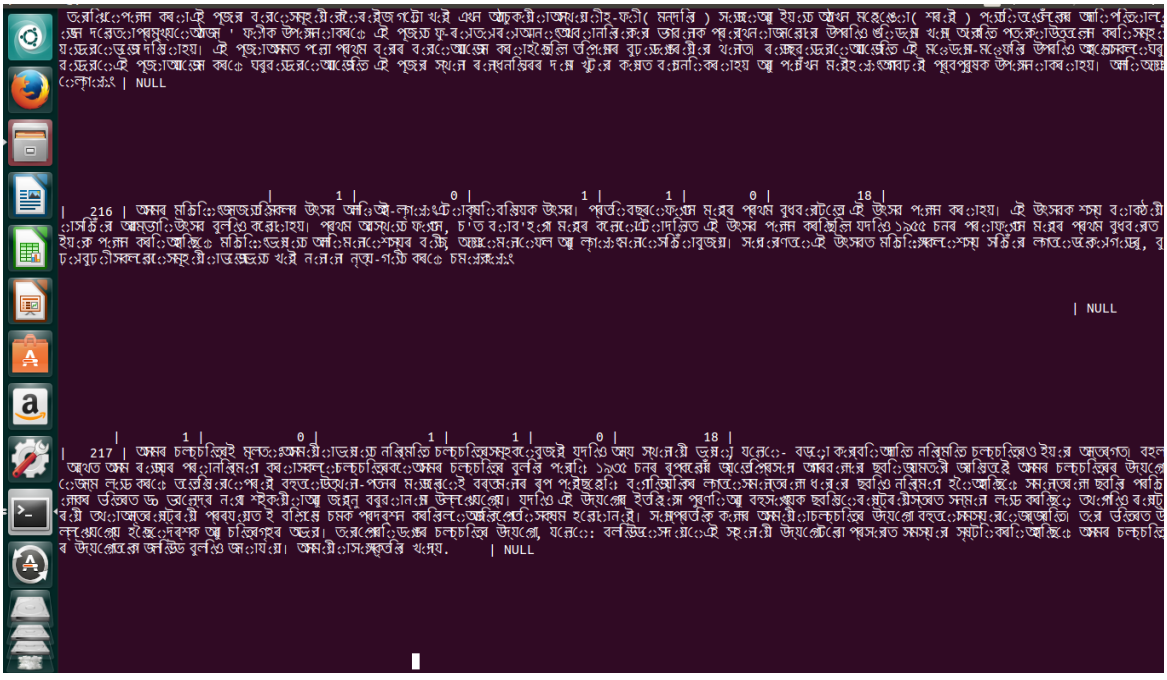

Fig3: A snapshot of the reliable text obtained from the mixed corpus.

DOI 10.17148/IJARCCE.2015.412107

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

III. CONCLUSION

We had obtained around 41 assamese sentences from the mixed courpus. In this experiment we had used SAMPA as a phoneme base for the assamese script , it is to be noted that the letters শ(talôibbô xô), ষ(murdhônyô xô) and the letter স (dôntyô xô) have been assigned similar sounding phoneme , in relatily the pronunciation of these letters differ. In future, the same will be tried with a better set of phonemes.

REFRENCES

[1] https://en.wikipedia.org/wiki/Speech_synthesis
[2] Multilingual Voice Creation Toolkit for the MARY TTS Platform, by Sathish Pammi, Marcela Charfuelan, Marc Schroder
[3] The MARY TTS entry in the Blizzard Challenge 2008, by Marc Schroder, Marcela Charfuelan, Sathish Pammi, Oytun T'urk.
[4] Bora, Mahendra (1981). The Evolution of Assamese Script. Jorhat, Assam: Assam Sahitya Sabha.
[5] https://github.com/marytts/marytts/wiki/New-Language-Support
[6] https://www.phon.ucl.ac.uk/home/sampa/

DOI 10.17148/IJARCCE.2015.412107