Rule based Automated Pronunciation Generator

Ayesha Binte Mosaddeque, Naushad UzZaman, and Mumit Khan
Center for Research on Bangla Language Processing, BRAC University, Dhaka, Bangladesh
lunaticbd@yahoo.com, naushad@bracu.ac.bd, mumit@bracu.ac.bd

Abstract

This paper presents a rule based pronunciation generator for Bangla words. It takes a word and finds the pronunciations for the graphemes of the word. A grapheme is a unit in writing that cannot be analyzed into smaller components. Resolving the pronunciation of a polyphone grapheme (i.e. a grapheme that generates more than one phoneme) is the major hurdle that the Automated Pronunciation Generator (APG) encounters. Bangla is partially phonetic in nature, thus we can define rules to handle most of the cases. Besides, up till now we lack a balanced corpus which could be used for a statistical pronunciation generator. As a result, for the time being a rule-based approach towards implementing the APG for Bangla turns out to be efficient.

1. Introduction

Based on the number of native speakers, Bangla (also known as Bengali) is the fourth most widely spoken language in the world [1]. It is the official language in Bangladesh and one of the official languages in the Indian states of West Bengal and Tripura. In recent years Bangla websites and portals are becoming more and more common. As a result it has turn out to be essentially important for us to develop Bangla from a computational perspective. Furthermore, Bangla has as its sister languages Hindi, Assamese and Oriya among others, as they have all descended from Indo-Aryan with Sanskrit as one of the temporal dialects. Therefore, a suitable implementation of APG in Bangla would also help advancement of the knowledge in these other languages.

The Bangla script is not completely phonetic since not every word is pronounced according to its spelling (e.g., ব □□□ /bɔḍ̲ḍ̲^{h}o/, □□□□ /moḍ̲ḍ̲^{h}o/). Therefore, we need to use some pre-defined rules to handle the general cases and some case specific rules to handle exceptions. For example, the words 'অ □□□ /ɔnek/' and '□□□ /o̲ti/' both start with 'অ /ɔ/' but their pronunciations are [ɔnek] and [o̲ti] respectively. These changes with the pronunciation of 'অ /ɔ/' are supported by the phonetic rules:

অ + C + ে (এ কার) > অ

অ + ই / C + ি (ই কার) > ও, where C= Consonant

There are some rules that have been developed by observing general patterns, e.g., if the length of the word is three full graphemes (e.g. কলম /kɔlom/, খবর /k^{h}ɔbor/, □□□□□ /baḍ̲la/, □□□□ /kolmi/ etc.) then the inherent vowel of the medial grapheme (without any vocalic allograph) tends to be pronounced as [o], provided the final grapheme is devoid of vocalic allographs (e.g., কলম /kɔlom/, খবর /k^{h}ɔbor/). When the final grapheme has adjoining vocalic allographs, the inherent vowel of the medial grapheme (e.g. □□□□□ /baḍ̲la/, □□□□ /kolmi/) tends to be silent (i.e., silenced inherent vowels can be overtly marked by attaching the diacritic '□').

2. Previous work

A paper about the Grapheme to Phoneme mapping for Hindi language [2] provided the concept that, an APG for Bangla that maps graphemes to phonemes can be rule-based. No such work has yet been made available in case of Bangla.

Although Bangla does have pronunciation dictionaries, these are not equipped with automated generators and more importantly they are not even digitized. However, the pronunciation dictionary by Bangla Academy provided us with a significant number of the phonetic rules [3]. And the phonetic encoding part of the open source transliteration software 'pata' [4] provided a basis.

3. Methodology

In the web version of the APG, queries are taken in Bangla text and it generates the phonetic form of the given word using IPA[1] transcription. Furthermore, there is another version of the system which takes a corpus (a text file) as input and outputs another file containing the input words tagged with the corresponding pronunciations. This version can be used in a TTS[2] system for Bangla.

In terms of generating the pronunciation of Bangla graphemes a number of problems were encountered. Consonants (except for 'শ /ʃ/' and 'স /s/') that have vocalic allographs (with the exception of '□/e/') are considerably easy to map. However there are a number of issues: Firstly, the real challenge for a Bangla

[1] International Phonetic Alphabet
[2] Text To Speech

pronunciation generator is to distinguish the different vowel pronunciations. Not all vowels, however, are polyphonic. 'অ /ɔ/' and 'এ /e/' have polyphones ('অ /ɔ/'' can be pronounced as [o] or [ɔ], 'এ /e/' can be pronounced as [e] or [æ], depending on the context) and dealing with their polyphonic behavior is the key problem. Secondly, the consonants that do not have any vocalic allograph have the same trouble as the pronunciation of the inherent vowel may vary. Thirdly, the two consonants 'শ /ʃ/' and 'স /s/' also show polyphonic behavior. And finally, the 'consonantal allographs'(□য /ɟ/,□□ /r/,□□ /b/,□□ /m/), and the grapheme 'য় /j/' complicate the pronunciation system further.

Hypothetically, all the pronunciations are supposed to be resolved by the existing phonetic rules. But as a matter of fact they do not; some of them require heuristic assumptions. The phonetic rules that have been used are described in Table 1. Some of the rules are described below:

- If a word starts with 'অ /ɔ/' followed by 'ক্ষ' or 'জ্ঞ /ŋ/' then the 'অ /ɔ/' is pronounced as [o]. For example, 'অক্ষাংশ /okkʰaŋsho/', 'যজ্ঞ /joggo/'.
- If the first grapheme of a word is devoid of any vocalic allographs and a '□□ /r/' or '□□ /ɟ/'is attached to it then the implicit 'অ /ɔ/' (associated to the first grapheme) is pronounced as [o]. For example, '□□□□ /krom/', '□□□□ /oddo/'. There is one exception to this rule that is when there is a 'য় /j/' after the '□□ /r/' then the implicit 'অ /ɔ/' (associated to the first grapheme) is pronounced as [ɔ]. For example, '□□□□ /krɔj/'.
- If the first grapheme of a word is devoid of any vocalic allographs and the consonant following it is accompanied with a '□ /ri/' (□ □□□) then the implicit 'অ /ɔ/' (associated to the first grapheme) is pronounced as [o]. For example, '□□□□ /mosrin/'.
- If a word starts with 'অ /ɔ/' followed by 'ই /i/' or 'উ /u/' or their corresponding vocalic allographs (□ /i/,□ /u/) then the 'অ /ɔ/' is pronounced as [o]. For example, '□□□□□□ /obʰidʰan/'.
- If a word starts with '□/e/' associated with a consonant followed by a '□ /ŋ/' or 'ঙ /ŋ/' and if there is no 'ি /i/', 'ী /i/', 'ু /u/' or 'ূ /u/' then the 'ে /e' is pronounced as [æ]. But if there is a 'ি /i', 'ী /i/','ু /u/' or 'ূ /u/' then the 'ে /e/' is pronounced as [e]. For example, 'বেঙ /bæŋ/', but 'বেঙি /beŋi/'.
- If the consonantal allograph '্ব /b/' is associated with the first grapheme of a word then the '্ব /b/' is silent. For example, 'ধ্বনি /dʰoni/'.
- If the consonantal allograph '্ব /b/' is associated with a grapheme in the middle or at the end of a word then that grapheme's pronunciation is doubled. For example, 'বিশ্ব /bisʰsʰo/'.
- If the consonantal allograph '্ম /m/' is associated with a grapheme in the middle or at the end of a word then that grapheme's pronunciation is doubled and the last grapheme is pronounced in a slightly nosal tone. For example, 'রশ্মি /rosʰsʰi/'. But there are some graphemes ('গ /g/', 'ঙ /ŋ/', 'ট /t/', 'ণ /n/', 'ন /n/', 'ম /m/', 'ল /l/') when associated with '্ম /m/' keep the pronunciation of 'ম /m/' unchanged. For example, 'বাগ্মী /bagmi/', 'জন্ম /jɔnmo/'.
- The consonantal allograph '্য /ɟ/', when associated with the middle or end grapheme of a word, is not pronounced. For example, 'সন্ধ্যা /shonḏʰa/'.
- If there is 'য় /j/' at the end of the word and no vowel is associated with .and the previous letter contains a 'ি /i/' or 'ী /i/', the 'য় /j/' is pronounced as 'য়ো /jo/'. For example, 'জাতীয় /ɟatijo/'.

Table 1 contains the phonetic rules. Apart from these some heuristic rules are used in APG. Few such rules are shown in Table 2. They were formulated while implementing the system. Most of them serve the purpose of generating pronunciation for some specific word pattern.

Note: C = Consonant. C□C = Conjunct Consonant

Table 1: Pronunciation Rules

Letter/IPA/	Pronunciation	Rules	Example

অ/ ɔ /	অ/ ɔ /	অ+□□□+ঈ (□□□ conjuncted with a consonant)	ল□□□ণ > □□□ □□□
		Implicit অ +□ র (□ □□□) + য়	□□□□
		অ +C□C+ □য (□ □□□)	□□□□ > □□□□□
	ও / o /	অ at the begining	
		অ +ই / C+□ (□ □□□)	অভিধান > ওভিধান
		অ + উ / C + □ (□ □□□)	অনুকূল > ওনুকূল
		অ + ন	মন > মোন
		অ +C+□য (□ □□□)	অদ্য > ওদ্দো
		অ + □□□	লক্ষ > লোকেখা
		অ + জ্ঞ	যজ্ঞ > জোগ্গো
Letter/IPA/	Pronunciation	অ + C+ ৃ (ঋ কার) Rules	মসৃণ > মোসৃন ঊীধসঢ়ষব
		Implicit অ +□□ (□ □□□)	ক্রম > ক্রোম, দ্রব্য > দ্রোব্য
		অ + □□ (□□□)+□□ (□ □□□) the spelling has been changed to অ +□□ (□□□) but first 'অ' is pronuounced 'ও'	পর্যন্ত (prevoius spelling পর্য্যন্ত) > পোর্জোন্তো
অ/ ɔ /	ও / o /	অ at middle (implicit with consonents)	
		If a word has 3 or more letters, bfeore the 'অ' at middle there is 'অ','আ','এ' and 'ও', then the mid –'অ' is pronounced as 'ও'.	কলস > কলোস, পাগল > পাগোল, চেতন > চেতোন, শোষণ > শোশোন্ ,
		অ at end	
		Words that end with '-আন', '-ত', '-ইত' are pronounced as ও at the end	করান > করানো, খেলান > খেলানো, হত > হতো, চলিত > চলিতো
		if there is 'য়' after 'ই' or '□' then it is pronounced as 'ও'	স্মরণীয় > শঁরোনিয়ো
		Words that end with '-তর','-তম' are pronounced as 'ও' at the end	অধিকতর > ওধিকতরো, অধিকতম > ওধিকতমো
		if there are conjuncted consonants at the end of the word then it is pronounced as 'ও'.	শক্ত > শক্তো
		if there is 'হ' or 'ঢ়' at the end of the word then it is pronounced as 'ও'	কলহ > কলহো , গাঢ় > গাঢ়ো/ rho
		if there is 'য়' after 'অ' or 'আ' then it is pronounced as 'য়' followed by '□' (□□□□□)	জয় > জয়্
আ / a/	আ/ a/	আ or C + া	আম , রাগ
	অ্যা / æ/	□□□/ □□ + আ	□□□□□ > □□□□□□ , □□□□□ >□□□□□□
□ এ / e/		এ At the begining	

	এ / e /	□ /এ + ই /□ / ঈ / □ /উ / □ / ঊ / □ /এ / □ / ও / □ /য় / র / ল / শ / হ	একি > একি, দেখি > দেখি, খেয়ে, একুশ, খেয়া, মঠে◌া, তলে, বশে, কহে
		□ /এ + □ / ঙ / □□□ + □ / ই / □ / উ	□□□□ > □□□□, □□□□ □□ > □□□□ □□
		Monosyllabic pronouns	□□, এ , □□
	অ্যা / æ/	এ / □+ (□ / □□□ / ঙ)+ □ / not (□/ ই /□ / উ)	□□□□□□□□ > □□□□□□□□□□, □□□ > □□□□□□
ঙ □ / ŋ/	অঙ্ / Ong/	-	রঙ
ঞ / ɲ/	ন / n/	-	পঞ্চ > পন্ চ◌া
		□□ + জ	অঞ্জন > অন্জন্
		□ □ + ঞ	জ্ঞান > গ্যাঁন্
ব / ন as □□□ Letter/IPA/	Changed Pronunciation	At the begining C+ □□ is not pronounced Rules	□□□□□□□□□ > □□□□□□□ Example
		At middle or end C+ □□ is doubled	□□□□□ > □□□□ □□
		At the begining, middle or end C◌্C + □□ is not pronounced	□□□□□ব্ল > □□□□□
	Unchanged	Words that start with '□□□ 'and have 'ব' as '□□□', retain the pronunciation of 'ব'	উদ্বগে > উদ্ বগে
		ব +□□ or ম + □□ , retain the pronunciation of 'ব'	তব্বিত > তব্বিত, লম্ব > লম্বে◌া
ম /m as □□□	Changed	At the begining C+ □□ is not pronounced, but C is pronounced with a slight nasal tone	□□□□□□ > □□□□□□
		At middle or end C+ □□ is doubled and C is pronounced with a slight nasal tone	□□□□ > □□□□□□
		At the begining, middle or end C◌্C + □□ is not pronounced and the last C is pronounced with a slight nasal tone	□□□□□□□ > □□□□ □□□
	Unchaned	At middle or end গ /ঙ /ট /ণ /ন /ম /ল + □□, retain the pronunciation of 'ম'	□□□□□□ > □□□□ □□ , □□□□ > □□□ □□
য / ɟ as □□□	অ্যা / æ/	At the begining C + □□ or C + □□+□	ব্যাথা > ব্যাথা
	এ / e/	At the begining C + □□+ □	ব্যক্তি > বক্তে ি
	Unpronounced	At middle or end with conjuncts C◌্C+ □□	মর্ত্য > মর্ ত◌া
	Doubled	At middle or end C + □□ , C is doubled	ধন্য > ধ◌ান্ ন◌া

র / ৎ as □□□	র / r/	At the begning C + □□ , if C does not have any vowel associated with it then the it is pronounced as 'ও', and C is not doubled	□□□□□□ > □□□□□□□□
		At middle or end C + □□, C is doubled	□□□□□□□> □□ □□ □□□□□□
		At middle or end with conjuncts C + □□, C is not doubled	□□□□□□□ > □□□□ □□□□
ল / l as □□□	ল / l/	At the begning C + □□ , C is not doubled	□□□□□□ > □□□□□□□□
		At middle or end C + □□, C is doubled	□□□□□□□□□□ > □□□□□□□□□□□□□□ □
শ / ʃ/	স / s /	when conjuncted with ত / থ / ন / র	প্রশ্ন > প্রস্নো
	শ / ʃ/	when not conjuncted	□□□ > □□□□
স / s/	স / s/	when conjuncted with ত / থ / ন / র	স্নান > স্নান্
		For foreign words	সার্ভসি > সার্ভস্ি
	শ / ʃ/	when not conjuncted	সচতেন > শচতেন্
ষ / ʃ/	শ / ʃ/	Always pronounced as শ / ʃ	□□□□ > □□□□□
ঞ	হ্ / h	For exclamatory words	□□ > □□□
	ও / o	At the end C+ □ , C is pronounced as 'ও' if no vowel associated with it	□□□□ > □□□□
	doubled	At middle C+ □ , C is doubled	□□□□□□ > □□□□□□□□

Table 2: Heuristic Rules

Letter/IPA/	Rules	Example
ষ / ʃ /	Always pronounced as শ / Σ	□□□□ > □□□□□
জ/ ɟ /	If there is '□□'before 'জ' and no vowel is associated with 'জ', then 'জ' is pronounced as '□□'	□□□□□ > □□□□□□□
	If 'জ' is followed by '□□□' or '□□□', and no vowel is associated with 'জ', and then if there is any '□' or '□', then 'জ' is pronounced as '□□', else it is pronounced as 'জ'.	□□□□□ > □□□□□□ □□□□□ > □□□□□□□
য় / ল /	If there is 'য়' at the end of the word and no vowel is associated with 'য়' and the previous letter contains a ' □ ' or '□ ', the 'য়' is pronounced as '□□'.	□□□□□ > □□□□□□

4. Implementation

APG has been implemented in Java (jdk1.5.0_03). The web version of APG contains a Java applet that can be used with any web client that supports applets. The other version of APG is also implemented in Java. Both
the versions generate the pronunciation on the fly; to be precise no look up file has been associated. Figure 1 illustrates the user interface of the web version and Figure 2 illustrates the output format of the other version.

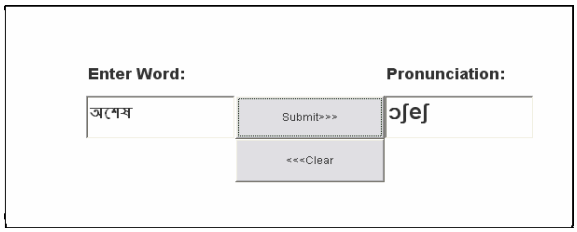

Figure 1 : The web interface of APG. The input word is 'অশেষ' and the correponding output is 'ɔʃeʃ'.

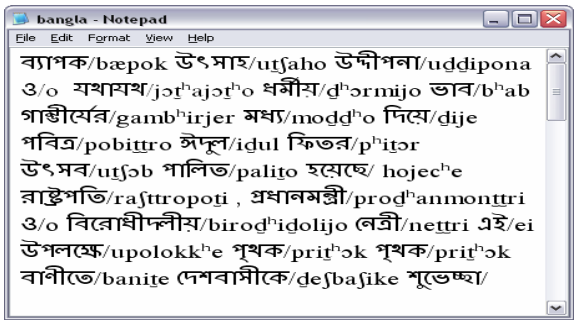

Figure 2 : the output file generated by the plug-in version of APG.

5. Result

The performance of the rule-based APG proposed by this paper is challenged by the partial phonetic nature of Bangla script. The accuracy rate of the proposed APG for Bangla was evaluated on two different corpora that were collected from a Bangla newspaper. The accuracy rates observed are shown in Table 3.

Table 3: Evaluation of accuracy

Number of words	Accuracy Rate (%)
736	97.01
8399	81.95

The reason of the high accuracy rate of the 736-word corpus is that, the patterns of the words of this corpus were used for generating the heuristic rules. The words in the other corpus were chosen randomly. The error analysis was done manually by matching the output with the Bangla Academy pronunciation dictionary.

6. Conclusion

The proposed APG for Bangla has been designed to generate the pronunciation of a given Bangla word in a rule based approach. The actual challenge in implementing the APG was to deal with the polyphone graphemes. Due to the lack of a balanced corpus, we had to select the rule-based approach for developing the APG. However, a possible future upgrade is implementing a hybrid approach comprising both a rule based and a statistical grapheme-to-phoneme converter. Also, including a look up file will increase the efficiency of the current version of APG immensely. This will allow the system to access a database for look up. That way, any given word will first be looked for in the database (where the correct pronunciation will be stored), if the word is there then the corresponding pronunciation goes to the output, or else, the pronunciation is deduced using the rules.

7. Acknowledgement

This work has been supported in part by the PAN Localization Project (www.panl10n.net) grant from the International Development Research Center, Ottawa, Canada, administrated through Center for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, Pakistan.

8. References

[1] The Summer Institute for Linguistics (SIL) Ethnologue Survey, 1999.

[2] M. Choudhury, "Rule-based Grapheme to Phoneme Mapping for Hindi Speech Synthesis", *Proceedings of the International Conference On*

Knowledge-Based Computer Systems, Vikas Publishing House, Navi Mumbai, India, 2002, pp. 343 – 353. Available online at: http://www.mla.iitkgp.ernet.in/papers/G2PHindi.pdf

[3] *Bangla Uchcharon Obhidhan*, Bangla Academy, Dhaka, Bangladesh.

[4] Transaliteration Software - Pata, developed by Naushad UzZaman, CRBLP, BRAC University. Available online at: http://www.naushadzaman.com/pata.html