

The background features a large, light blue watermark of the BRAC University logo, which consists of a stylized open book with the text 'BRAC UNIVERSITY' overlaid in a serif font.

PHONETIC ENCODING FOR BANGLA AND ITS APPLICATION TO SPELLING CHECKER, TRANSLITERATION, CROSS LANGUAGE INFORMATION RETRIEVAL AND NAME SEARCHING

**Presentation of Undergraduate Thesis
Naushad UzZaman**

**Supervised by
Dr. Mumit Khan**

Phonetic Encoding

- Encode a word based on how it is pronounced.
- In English, **realise** and **realize** both are similar sounding words. So, their phonetic code should be same.
- In Bangla, ক্ষত and খত both are similar sounding words. So, their phonetic code should be same.

Why do we need Phonetic Encoding

- Bangla does not have very good spelling checker that can give word of same pronunciation in suggestions considering complex Bangla rules, this encoding helps us to develop that.
- Bangla have many transliteration applications, but all of those give a one to one mapping. It will convert each English letter or letters to fixed Bangla letter or letters. There are no transliterations available where if you write in English it will give dictionary word of same pronunciation. Using phonetic encoding we can develop that.
- Name searching is a very useful application in census, hospitals, educational institutes, offices, etc. There is no such name searching application in Bangla that gives names with almost same pronunciation in suggestion. This encoding helps to develop this application too.
- This encoding can work as an intermediate code in multi-lingual information retrieval, where a user issues a query in one language (such as English) to search a collection in a different language (such as Bangla). More specifically, writing the pronunciation of a word in English one can search words with same pronunciation in a Bangla document.

Why do we need Phonetic Encoding

- Handle complex cases in:
 - Spelling checker
 - Name searching
 - Transliteration
 - Multi-lingual information retrieval
 - And many other applications

Phonetic Encoding in English

- Soundex
- Metaphone
- Phonix
- Double Metaphone

Soundex

- Proposed by Odell and Russel
- Proposed at 1918
- Partitions the set of letters into seven disjoint sets, assuming that the letters in the same set have similar sound.

Soundex Table

<i>Code</i>	<i>Letters</i>
0 (not coded)	A, E, I, O, U, H, W, Y
1	B, F, P, V
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M, N
6	R

Example of Soundex

- Realize – 6004020 – 642
- Realise – 6004020 – 642

- Knight – 250203 – 2523
- Nite – 5030 – 53

Problems of Soundex

- Examples show that soundex can not handle the complex cases of English.
- Hence a better phonetic encoding was needed, which can handle the complex cases that considers the context of the letter before giving it a code.
- For example, in knight, k doesn't have any sound at the beginning and also gh is silent in this context. So, we need to consider all these for a better phonetic encoding.

Metaphone

- Proposed by Lawrence Philips in 1990.
- Unlike soundex, metaphone gives the code to letters or group of letters considering the context of the letter.

Metaphone Transformations

- B -> B unless at the end of a word after "m" as in "dumb"
- C -> X (sh) if -cia- or -ch-
- S if -ci-, -ce- or -cy-
- K otherwise, including -sch-
- D -> J if in -dge-, -dgy- or -dgi-
- T otherwise
- F -> F
- G -> **silent if in -gh-** and not at end or before a vowel
- in -gn- or -gned- (also see dge etc. above)
- J if before i or e or y if not double gg
- K otherwise
- H -> silent if after vowel and no vowel follows
- H otherwise
- J -> J
- K -> silent if after "c"
- K otherwise
- L -> L
- M -> M
- N -> N

- P -> F if before "h"
- P otherwise
- Q -> K
- R -> R
- S -> X (sh) if before "h" or in -sio- or -sia-
- S otherwise
- T -> X (sh) if -tia- or -tio-
- 0 (th) if before "h"
- silent if in -tch-
- T otherwise
- V -> F
- W -> silent if not followed by a vowel
- W if followed by a vowel
- X -> KS
- Y -> silent if not followed by a vowel
- Y if followed by a vowel
- Z -> S
-
- Initial Letter Exceptions
-
- Initial **kn-**, gn- pn, ac- or wr- -> **drop first letter**
- Initial x- -> change to "s"
- Initial wh- -> change to "w"

Example of Metaphone

- Knight – NT
- Nite – NT

- Basinger is pronounced in both way as “Basin-gger” or “Basin-ger”.
- Basinger - BSNJR
- Basin-gger - BSNKR
- Basin-ger - BSNJR

- In the previous example we saw that one word can be pronounced in two different ways. So, in the best case it may be coded to one pronunciation. Like in the example, we saw that it matched with Basin-jeer.
- This can be solved, if we can give two or multiple codes to the same word.
- So considering this issue in 2000, Philips again proposed another phonetic encoding, which he named as double metaphone.

Key concepts from English

- Soundex: groups the letters of same pronunciation.
- Metaphone & Phonix: also considers the context of a letter to encode it.
- Double metaphone: gives multiple codes to same word, if it is pronounced in more than two ways.

Existing Encoding in Bangla

- Hoque and Kaykobad's soundex type encoding, 2002
- Zaman and Khan's soundex type encoding, 2004

Hoque and Kaykobad's phonetic encoding Table

Name	Group Member
1	ক, খ, গ, ঘ, ঙ্গ
2	চ, ছ, জ, ঝ, য
3	ট, ঠ, ড, ঢ
4	ত, থ, দ, ধ, ত্
5	প, ফ, ব, ভ
6	ঙ, ঞ, ং
7	শ, স, ষ
8	র, ড়, ঢ়, ঝ
9	ন, ণ
α	ম
β	18 May 2005, BRAC University

Example of Hoque and Kaykobad's soundex type encoding, 2002

- For example, কর্ম will be converted to a 4 lengthen sound code as “ক8α0”, with zero padding.

Zaman and Khan's soundex encoding, 2004

Code	Group members
0	্, ো, ঁ
“a”	আ, া
“i”	ই, ঈ, ি, িী
“u”	উ, ঊ, ু, ু
“e”	এ, ে, ঐ, ঐে
“o”	অ, ও, ঔ, ৌ
“k”	ক, খ
“g”	গ, ঘ
“m”	ম, ঙ, ং
“c”	চ, ছ
“j”	য, জ, ঝ

Example of Zaman Khan's soundex

Input	Encoding	Suggestion
খুমাড়	kumar	কুমার
পাসান	pasan	পাষণ
দগধ	dgd	দগ্ধ (দগ ্ধ)

Limitation of existing encodings

- Bangla has many consonant clusters or juktakkhor with unusual pronunciations (i.e., ক্ষ, ক্ষা, etc.): let us consider ক্ষ. ক্ষ = ক+্+ষ; ক্ষত /kʰɔ̃t̪o/ is pronounced as খত /kʰɔ̃t̪o/, where ষ does not have any sound.
- Bangla has different uses of *Phalaa's*, such as BA, MA, YA, RA and LA phalaa.

Limitation of existing encodings

- Different pronunciation of letters or conjuncts in different contexts: consider again ক্ষ.
 - At the beginning of word
 - (ক্ষত → খত /kʰɔt̪o/);
 - In the middle or at the end of a word
 - (দক্ষ → দকখ /dɔkkʰio/).
- Multiple pronunciations of some letters in the same context, such as হ with ব:
 - আহ্বান → আওভান /aovan/.
 - আহ্বান → আহভান /aɦobɦian/

Double metaphone phonetic encoding

Sample Encoding Rules for ক্ষ

Soundex Encoding

"k"	ক	KA	"0995"
0 (zero)	্	Virama/Hasant	"0981"
"s"	ষ	SSA	"09B7"

Double Metaphone Encoding

ক্ষ	"0995""09CD""09B7"	"k"	@the beginning	ক্ষত
ক্ষ	"0995""09CD""09B7"	"kk"	@ middle/end	দক্ষ

Spelling checker

Spelling checker using phonetic encoding

Lexicon	Encoded
Word List	Word List
অকালপক্ক	“okalpkk”
স্বামী	“sami”
চাঁদ	“cad”
দগ্ধ	“dgd”

Encoded Test word	Misspelled Word
“sami”	ষামি

Search the encoded test word
in the encoded word list rather
than searching the test word in
the Dictionary word list

Name Searching

- Wide variety of uses like land registry, census, hospital, educational institute, industries, etc.
- Name searching uses the same technique as spelling checker.
- Any words can be names. So, all the rules for Bangla words are applicable for names.
- But we have to modify the encoding for some special behavior for names.

One example of modification for name searching

Similarly pronounced names	Encoding
মরতুজা, মুরতোজা, মরতোজা, মোরতুজা	“mrtj”

- The major variation of a name comes from interchanging the vowels.
- For that reason, vowels are *Not Coded* for names.

Another example of modification

Short cut	Elaborated form	Encoding
মোঃ	মোহাম্মদ	“mmmD”
ডঃ	ডক্টর	“DkTr”
ডাঃ	ডাক্তার	“DkTr”
এডঃ	এডভোকেট	“DbkT”

- মোঃ is the same as মোহাম্মদ /mohammɔd/.
- one-to-one transformations are used before encoding process.
- So, to encode মোঃ we will first transform it to মোহাম্মদ before the final encoding.

Transliteration

Types of Transliteration

- **Direct Mapping**

- One to one mapping
- shondha - শোনধা

- **Phonetic Mapping**

- Gives dictionary word with the same pronunciation
- shondha - সন্ধ্যা

Challenge

- We encode the Bangla words to English codes.
- Main challenge is to encode the English words, so that both similar sounding Bangla and English words get the same code.

Example সন্ধ্যা

- সন্ধ্যা is encoded to “shndha”
- User writes *shondha* in English.
- We need to encode *shondha* to “shndha”.

Transliteration using phonetic encoding

Lexicon	Encoded
Word List	Word List
অকালপক্ক	“okalpkk”
সন্ধ্যা	“shndha”
চাঁদ	“cad”
দগ্ধ	“dgd”

Encoded Like Bangla	English Word
“shndha”	shondha

Search the *encoded like Bangla word* in the *encoded word list*. We can get the *similar sounding word to replace*.

Example of transliteration

English word	Output in direct mapping	Encoding like Bangla	Output in phonetic mapping	Selected word
shondha	শোনধা	shndha	সন্ধ্যা া	সন্ধ্যা া
bepar	বেপার	bepar	বেপার / ব্যাপা	ব্যাপা র
ami	আমি	ami	আমি	আমি

Multi-lingual information retrieval

What does it handle

- User issues a query in one language to search a collection in different language.
- In simple word, it will take a word in English and will find the similar sounding word in a Bangla document.

Multi-lingual information retrieval using phonetic encoding

Lexicon	Encoded
Word List	Word List
অকালপক্ক	“okalpkk”
সন্ধ্যা	“shndha”
চাঁদ	“cad”
দগ্ধ	“dgd”

Encoded	English Word
Like Bangla	
“shndha”	shondha

Search the *encoded like Bangla word* in the *encoded word list*. We can get the *similar sounding word to replace*.

Summary and Conclusion

- Proposed a phonetic encoding
- Can be used in many useful applications.
- Future work: Digital pronunciation dictionary, Text to Speech