



Konkani Integer Phonetic Transcription System

Swapnil Fadte¹, Edna Vaz Fernandes^{2,3}, Hanumant Redkar¹, Teja Kundaikar¹, Ramdas Karmali¹,
Jyoti D. Pawar¹

¹Discipline of Computer Science and Technology, Goa Business School, Goa University, India

²Department of Konkani, Govt. College of Arts Science and Commerce Quepem, Goa India

³University of Mumbai, India

swapnil.fadte@unigoa.ac.in, edna.vaz22@gmail.com, hanumantredkar@unigoa.ac.in,
teja.kundaikar@unigoa.ac.in, rnk@unigoa.ac.in, jdp@unigoa.ac.in

Abstract

This paper describes an ongoing work on the Phonetic Dictionary for Konkani language. In this work, we have build a resource that would phonetically transcribe Konkani Integers and generate their written form in the Devanagari script. The algorithm developed in this work takes an integer as an input and generates its written form in the Devanagari script, along with its phonetic transcription in the International Phonetic Alphabet (IPA). Our algorithm is a rule-based system in which phonetic transcriptions of numerals are created using rules from the available literature and for some cases we have proposed new forms for the numerals. The algorithm has been made robust enough to automatically give a written form of any Konkani numeral in the Devanagari script, along with its equivalent IPA transcription. This work is the first step towards providing an open-source phonetic dictionary for Konkani language. We have tried to keep the phonetic transcriptions as much as closer to their natural pronunciations. This is done for the purpose of capturing the general tendency of the language. So, for example, while the number '8' आठ [aʈ^h] is written with an aspirated retroflex consonant ठ [ʈ^h], the final consonant [ʈ^h] is heard without aspiration in the actual speech. This loss of aspiration at word final position generally happens across all the consonants of the language, in the Konkani varieties spoken in Goa.

Index Terms: Konkani, Konkani Speech data, phonetic dictionary, integers, Devanagari, integer dictionary

1. Introduction

India is a multilingual country having various languages and dialects. Konkani is the official language of the state of Goa (India). It belongs to the Indo-Aryan language family. The constitution of India - the document that lays down the framework which demarcates fundamental political code, structure, procedures, powers, and duties of government institutions and sets out fundamental rights, directive principles, and the duties of citizens, in its Eighth Schedule ("list") (Articles 344(1) and 351), recognizes 22 Regional languages as official languages [1]. Speakers of Scheduled Languages enjoy some advantages over the speakers of the non-scheduled languages. For example, the members of the parliament are allowed to speak and present their views in their language if it is one of the Scheduled languages. The Seventy-First Amendment to the Constitution on 20th August 1992, added Konkani to the list of Scheduled languages. The official script for Konkani is Devanagari. Numerals are usually written in the Devanagari script. However, they are also written as per the Arabic writing system for ease of communication.

2. Motivation

When we read any text written in a certain script, we encounter characters, numerals and punctuation marks. Anyone who understands a certain script will be able to read any language written in that script. However, when it comes to the reading (pronunciation) of numerals, one needs to follow certain rules. For example, if the integer 1234 is to be read as some value of currency, it needs to be read as "One Thousand Two Hundred Thirty Four" or "Twelve Hundred Thirty Four". Similar rules for integers will also be applicable to the pronunciation of Konkani integers. There are a few machine translation systems available for Konkani [2, 3], but they do not provide text transcriptions of Konkani Numerals. Also, testing of Konkani integers in the text representation on Google translator showed a good amount of errors. Following are some representative examples where we can see a combination of text with numerals. Such instances of data in a text corpus could pose a big challenge for any system that aims to transcribe data accurately.

1. "ता. १ फेब्रुवारी २०२२" [ta. ek februvāri don hədʒar bavi:s] ('dtd. 1st February 2022'). The above character and numeral combination refers to a specific date and a year. While '१ one' in १ फेब्रुवारी '1st February' might be easily transcribed as *ek (februvāri)* by any system, the numeral 2022 which is a year, has to be read and transcribed as दोन हजार बावीस [don hədʒar bavi:s] "two thousand twenty-two" and not as दोन शून्य दोन [don junj^o don don] "two zero two two". Thus, it is important that the system correctly identifies the context in which a numeral appears so that it can generate a proper transcription of the same.
2. "सकाळी ०८:०० ते १०:०० वरामेरेन" [səka:lī a:t^h tɛ d^ha vərəmərən] '(from) morning 8 to 10 a.m.' (Lit. morning 8 to 10 hours till). This phrase specifies a certain time of the day. The system needs to acknowledge this context of time and generate a string that reads the numerals as hours (and minutes in some other temporal context).
3. "गोंयचें क्षेत्रफळ ३७०१ चौखण किलोमिटर आसा." [gōjɕɛ̃ kʃɛtrə'fə̃l tɪm hədʒar satʃɛ ek tʃovk^həŋ kilomiṭər asa] ('The (total) area of Goa is 3701 sq. kms.'). The numeral in the above sentence specifies the area of the region of Goa. The accepted way of reading the numeral in this sentence is considering the entire string of numbers as one unit, i.e., as तीन हजार सातशे एक [tɪm hədʒar satʃɛ ek] (Lit. "Three thousand Seven Hundred One") and not as individual numbers- तीन सात शून्य एक [tɪm sat junj^o ek] 'Three Seven Zero One'.
4. "माशेलाचो पिन कोड ४०३ १०७." [maʃɛla:tʃo pin'kod tʃar junj^o tɪm ek junj^o sa:t] ('The pin code of Marcela is 403 107'). Postal Index Number (PIN or simply PIN Code) refers to the

six-digit number used by India Post in its postal code system. More commonly, the numbers indicating such a code are read by spelling out the numerals as discrete units. The Pin code in the above example needs to be read and transcribed as चार शून्य तीन एक शून्य सात [tʃar ʃunjə tɪn ek ʃunjə sa:t] 'Four Zero Three One Zero Seven'.

5. "ताचो फोन नंबर 9850 403 107" [tatʃo fon nambər nɔv a:tʰ pã:ts ʃunjə tʃa:r ʃunjə tin ek ʃunjə sa:t / tatʃo fon nambər nain eʃ faiv dʒiro fo:r dʒiro tʰri vən dʒiro sɛvən]('His phone number is 9850 403 107'). Phone numbers can be read differently by different speakers. However, reading the numbers as discrete units would be a good way to spell out the long number string.

From the above examples, it is clear that the numerals in any given sentence do not lend themselves to same type of output in the spoken and consequently in the written form. While the numerals in example 1 and example 3 are read as (a date and) a year and area respectively, the numerals in example 2 do not undergo much change in the way they are read/pronounced (and hence written) except for the need to write them in the Devanagari script. Similarly, while the numerals in example 4 are read as individual items, there could be other ways some speakers might want to read (pronounce) these. The same holds for the numerals mentioned in example 5 which allow different combinations for pronunciations. Given this background, an accurate transcription of the numerals that adheres to the speaking and writing rules of the language is of great relevance for developing a good transcription System.

3. Scope of the work

Through our work, we have made an effort to develop an automatic system for Konkani language that gives the phonetic as well as Devanagari transcription of a given integer. This is the first kind of work that aims to automatically transcribe Konkani numerals appearing in different contexts into the officially recognised Devanagari script along with the pronunciation of the numerals (given in IPA). As of now, our system only handles the transcriptions of numerals in a positional number system.

4. The proposed system

We are presenting here an automatic phonetic transcription system for Konkani Integers. The system takes an integer as an input and generates its representation in word (the written form) along with its phonetic transcription. Figure 1 diagrammatically presents the system developed by us.

5. Methodology

This section describes the implementation of the system and transcription rules that were used in the Algorithm design.

5.1. Implementation details

The implementation consists of two parts. In the first part, the input is processed and standardised. In the second part, the standardised input is transcribed.

5.1.1. Script conversion

This is the first and the simplest component of the system which takes the input either in Devanagari or Roman digits and converts

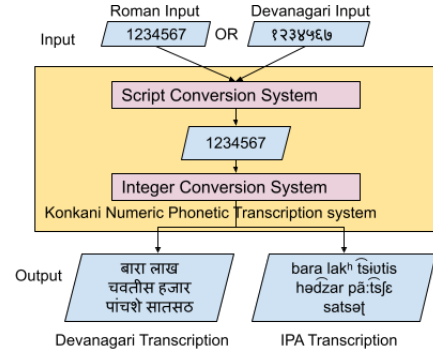


Figure 1: Integer Transcription System Diagram.

it to Roman numerals. This is achieved by converting the input value to a new string representation by one-to-one mapping of the digits.

5.1.2. Integer transcription

This is the main component of the system. It converts the integer into its spoken form (its pronunciation in Konkani) by applying conversion rules and using a dictionary database for the numbers and their corresponding pronunciation. These rules and mappings are presented in a tabular form and explained in detail in section 5.2 below. The algorithm for the transcription is created using these rules. It takes the integer and checks its length. Depending on the length of integer, it then calls for two sub algorithms: Left and Right transcription. Both Left and Right transcription works recursively and completes the transcription and results are merged and returned as final transcription of integers. Although it is rare to get an integer with a large number (length) in practice, ours algorithm can handle an integer of any length.

Data: *integer*

Result: *transcription_text*

$y \leftarrow ""$;

$X \leftarrow input$;

$N \leftarrow len(X)$;

if $N \geq 12$ **then**

$RIGHT \leftarrow assign\ last\ 11\ digits$;

$LEFT \leftarrow X/10^{11}$; /* removing last 11 digits from X */

$y \leftarrow left_transcription(LEFT) +$

$+position_mapping(12) +$

$right_transcription(RIGHT)$

else

$y \leftarrow right_transcription(X)$;

end

Algorithm 1: Integer transcription.

5.2. Identification of transcription rules

First, we have identified word transcription for integers. Devanagari transcription provided in Table 1 is from [4]. However, phonetic transcription for integers is not available. This is proposed and provided here for the first time. Rules for transcription for the integers one to hundred are directly mapped and are provided in Table 1. For example, the integer 63 will be mapped

Sr No	Devanagari	Roman	IPA	Devanagari Transcription	Sr No	Devanagari	Roman	IPA	Devanagari Transcription
1	१	1	[ek]	एक	51	५१	51	[ekavən]	एकावन
2	२	2	[dom]	दोन	52	५२	52	[bavən]	बावन
3	३	3	[ti:n]	तीन	53	५३	53	[treppən]	त्रेपन
4	४	4	[tʃa:r]	चार	54	५४	54	[tʃoupən]	चौपन
5	५	5	[pā:fs]	पाच	55	५५	55	[pāntsavən]	पचावन
6	६	6	[si:]	स	56	५६	56	[tʃ ^h appən]	छापन
7	७	7	[sa:t]	सात	57	५७	57	[səttavən]	सत्तावन
8	८	8	[a:t]	आठ	58	५८	58	[əttavən]	अठ्ठावन
9	९	9	[ŋəv]	णव	59	५९	59	[ekunʃat]	एकुणसाठ
10	१०	10	[d ^h a:]	धा	60	६०	60	[sa:t ^h]	साठ
11	११	11	[ikra]	इकरा	61	६१	61	[eksət]	एकसठ
12	१२	12	[bara]	बारा	62	६२	62	[bāsət]	बांसठ
13	१३	13	[tera]	तेरा	63	६३	63	[trəət]	त्रेसठ
14	१४	14	[tʃivda]	चवदा	64	६४	64	[tsiusət]	चवसठ
15	१५	15	[pəndra]	पंदरा	65	६५	65	[pəsət]	पांसठ
16	१६	16	[səla]	सोळा	66	६६	66	[sāsət]	सांसठ
17	१७	17	[sitra]	सतरा	67	६७	67	[satsət]	सातसठ
18	१८	18	[itra]	अठरा	68	६८	68	[atsət]	आठसठ
19	१९	19	[ekunʃis]	एकुणिस	69	६९	69	[ekunʃittir]	एकुणसत्तर
20	२०	20	[vis]	वीस	70	७०	70	[sitti]	सत्तर
21	२१	21	[ekvis]	एकवीस	71	७१	71	[ekjattar]	एक्यात्तर
22	२२	22	[bavis]	बावीस	72	७२	72	[bjattar]	ब्यात्तर
23	२३	23	[tevis]	तेवीस	73	७३	73	[trjattar]	त्र्यात्तर
24	२४	24	[tʃovis]	चोवीस	74	७४	74	[tsəvdjattar]	चवद्यात्तर
25	२५	25	[pəntʃvis]	पंचवीस	75	७५	75	[pəntʃattar]	पंच्यात्तर
26	२६	26	[sivvis]	सवीस	76	७६	76	[jattar]	शात्तर
27	२७	27	[səttavis]	सत्तावीस	77	७७	77	[səttjattar]	सत्त्यात्तर
28	२८	28	[əttavis]	अठ्ठावीस	78	७८	78	[ətt ^h jattar]	अठ्ठ्यात्तर
29	२९	29	[ekunʃis]	एकुणतीस	79	७९	79	[ekunʃəjʃi]	एकुणअंशशी
30	३०	30	[tis]	तीस	80	८०	80	[əjʃi]	अंशशी
31	३१	31	[ektis]	एकतीस	81	८१	81	[ekjəjʃi]	एक्यांशशी
32	३२	32	[bəttis]	बत्तीस	82	८२	82	[bjəjʃi]	ब्यांशशी
33	३३	33	[təttis]	तेत्तीस	83	८३	83	[trəjʃi]	त्र्यांशशी
34	३४	34	[tsiutis]	चवतीस	84	८४	84	[tsəvdjəjʃi]	चवद्यांशशी
35	३५	35	[pəstis]	पस्तीस	85	८५	85	[pəntʃəjʃi]	पंच्यांशशी
36	३६	36	[tʃ ^h əttis]	छत्तीस	86	८६	86	[ʃəjʃi]	श्यांशशी
37	३७	37	[səttis]	सात्तीस	87	८७	87	[səttjəjʃi]	सत्त्यांशशी
38	३८	38	[əttis]	आठ्ठीस	88	८८	88	[ətt ^h jəjʃi]	अठ्ठ्यांशशी
39	३९	39	[ekunʃʃalʃis]	एकुणचाळीस	89	८९	89	[ekunʃəvəəd]	एकुणणव्वद
40	४०	40	[tʃʃalʃis]	चाळीस	90	९०	90	[ŋəvəəd]	णव्वद
41	४१	41	[eketʃʃalʃis]	एकेचाळीस	91	९१	91	[ekjanŋəv]	एक्याणव
42	४२	42	[betʃʃalʃis]	बेचाळीस	92	९२	92	[bjanŋəv]	ब्याणव
43	४३	43	[tretʃʃalʃis]	त्रेचाळीस	93	९३	93	[trjanŋəv]	त्र्याणव
44	४४	44	[tsəvetʃʃalʃis]	चवेचाळीस	94	९४	94	[tsəvdjanŋəv]	चवद्याणव
45	४५	45	[pəntʃʃalʃis]	पंचेचाळीस	95	९५	95	[pəntʃjanŋəv]	पंच्याणव
46	४६	46	[ʃetʃʃalʃis]	शेचाळीस	96	९६	96	[ʃanŋəv]	शाणव
47	४७	47	[səttetʃʃalʃis]	सत्तेचाळीस	97	९७	97	[səttjanŋəv]	सत्त्याणव
48	४८	48	[əttetʃʃalʃis]	अठ्ठेचाळीस	98	९८	98	[ətt ^h janŋəv]	अठ्ठ्याणव
49	४९	49	[ekunʃpənnas]	एकुणपन्नास	99	९९	99	[ŋəvjanŋəv]	णव्याणव
50	५०	50	[pənnas]	पन्नास	100	१००	100	[ʃimbir]	शंबर

Table 1: Integer mapping rules for integers till 100.

to [trəsət, त्रॅसठ]. Table 2 provides the details of positional transcription for hundred, thousand, lakh and crore (ten million) and so on. Table 2 can be consider as a snapshot of the algorithm for various lengths. For example, the integer 1234567 gets positional transcription from Sr. no 4 in Table 2. The algorithm calculates the length and splits the integer into two sub-strings and calls two sub-tasks which recursively calculate the transcription for these two integer strings. There are other interesting cases like, the numeral multiples of 50 usually have multiple pronunciations. for e.g. integer 150 has two common transcriptions and 1500 has four transcriptions. Table 3 shows a few examples for the similar cases. Pronunciation rules for powers of ten till 10³⁰ are provided in Table 4. Technically, it can go to infinite with recursive logic for integer transcription.

6. Results and discussion

The output of the phonetic transcription system is shown in figure 2. The algorithm is coded using Python and can be accessed using this link¹. As regards the phonetic transcription of the integers, we have attempted an accurate transcription of the same. However, one needs to remember that these pronunciations are with regard to our observations about the Konkani spoken in Goa. E.g., as pointed out earlier, the integer written as आठ [a:t^h] is mostly pronounced as आट [a:t] i.e., without aspiration unlike its written counterpart. The same is true for other integers having aspirated consonants in word final positions or

¹https://github.com/SwapnilFadte/Konkani_integer_transcription.git

```

Data: integer
Result: transcription_text
y ← "";
X ← input;
N ← len(X);
if N ≤ 2 then
  if N == 0 then
    y ← "";
  else
    y ← integer_mapping(); /* Using table
    1 transcription rule */
  end
else
  if N ≤ 3 then
    if N == 100 then
      y ← integer_mapping()
    else
      y ← integer_mapping() +
      position_mappings(3) +
      right_transcription(RIGHT)
    end
  else
    if N ≤ 5 then
      y ← integer_mapping() +
      position_mapping(4) +
      right_transcription(RIGHT)
    else
      if N ≤ 7 then
        y ← integer_mapping() +
        position_mapping(6) +
        right_transcription(RIGHT)
      else
        if N ≤ 9 then
          y ← integer_mapping() +
          position_mapping(8) +
          right_transcription(RIGHT)
        else
          y ← integer_mapping() +
          position_mapping(10) +
          right_transcription(RIGHT)
        end
      end
    end
  end
end

```

Algorithm 2: right_transcription.

```

Data: integer
Result: transcription_text
y ← "";
X ← input;
N ← len(X);
if N ≥ 11 then
  RIGHT ← assign last 9 digits;
  LEFT ← X/109; /* removing last 9
  digits from X */
  y ← left_transcription(LEFT) +
  position_mapping(10) +
  right_transcription(RIGHT)
else
  y ← right_transcription(X)
end

```

Algorithm 3: left_transcription.

Sr No	Integer length	IPA	Devanagari Transcription
1	3	[ʃɛ]	शे
2	4	[hədʒar]	हजार
3	6	[lak ^h]	लाख
4	8	[koʃi]	कोटी
5	10	[əɾəb]	अरब
6	12	[k ^h əɾəb]	खरब

Table 2: Position mapping rules.

Sr No	Integer	IPA	Devanagari Transcription
1	150	[ekʃɛ pənnas]	एकशे पन्नास
2	150	[dɛdʃi]	देडशी
3	250	[dɔnʃɛ pənnas]	दोनशे पन्नास
4	250	[əɟɟɟʃɛ]	अडीजशे
5	350	[tɪnʃɛ pənnas]	तिनशे पन्नास
6	350	[sɑɟɛ tɪnʃɛ]	साडेतिनशे
7	1150	[ek hədʒar ekʃɛ pənnas]	एक हजार एकशे पन्नास
8	1150	[ek hədʒar dɛdʃi]	एक हजार देडशी
9	1150	[ikraʃɛ pənnas]	इकराशे पन्नास
10	1150	[sɑɟɛ'ikraʃɛ]	साडेइकराशे

Table 3: Multiple transcription rules for numerals above 100.

Sr No	Integer	IPA	Devanagari Transcription
1	10 ¹	[d ^h a:]	धा
2	10 ²	[ʃɛ/ʃɪmbɪr]	शे/शंबर
3	10 ³	[ɛ:k hədʒar]	एक हजार
4	10 ⁴	[d ^h a: hədʒar]	धा हजार
5	10 ⁵	[lak ^h]	एक लाख
6	10 ⁶	[d ^h a: lak ^h]	धा लाख
7	10 ⁷	[ɛ:k koʃi]	एक कोटी
8	10 ⁸	[d ^h a: koʃi]	धा कोटी
9	10 ⁹	[ɛ:k əɾəb/ʃɪmbɪr koʃi]	एक अरब/शंबर कोटी
10	10 ¹⁰	[d ^h a: əɾəb]	धा अरब
11	10 ¹¹	[ɛ:k k ^h əɾəb]	एक खरब
12	10 ¹²	[d ^h a: k ^h əɾəb/lak ^h koʃi]	धा खरब/एक लाख कोटी
13	10 ¹³	[ʃɪmbɪr k ^h əɾəb]	शंबर खरब
14	10 ¹⁴	[ɛ:k hədʒar k ^h əɾəb]	एक हजार खरब
15	10 ¹⁵	[d ^h a: hədʒar k ^h əɾəb]	धा हजार खरब
16	10 ²⁰	[ɛ:k əɾəb k ^h əɾəb]	एक अरब खरब
17	10 ²⁵	[ɛ:k lak ^h əɾəb k ^h əɾəb]	एक लाख अरब खरब
18	10 ³⁰	[d ^h a: əɾəb əɾəb k ^h əɾəb]	धा अरब अरब खरब

Table 4: Transcription rules for powers of ten.

word medial positions (as in the case of अठरा, अठ्ठावीस, etc.). Loss of aspiration especially at word-final position is common in Konkani varieties spoken in Goa.

```

Input          : 100563910
Output in Devanagari : धा कोटी पांच लाख त्रेंसठ हजार णवशे धा
Output in IPA     : dha: koʃi pã:ths lakh trēsəʈ
                  hædʒar ɳəʋʃɛ dha:

```

Figure 2: Integer Conversion.

7. Conclusion and future work

In this work, we have presented a system that transcribes Konkani integers into the officially recognised Devanagari script along with the IPA transcriptions of the numerals. The current system can technically handle integers of infinite length. As a future work it can be extended for fractions, dates, scientific numbers, phone numbers, etc.

8. References

- [1] *The Constitution of India*, Eighth schedule, Article(s): 344(1) and 351, Description: Official languages, 1950.
- [2] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016.
- [3] M. Junczys-Dowmunt, "Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation," in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 225–233. [Online]. Available: <https://aclanthology.org/W19-5321>
- [4] G. K. Academy, *Konkani Shuddhalekhanache Nem 5th Edition*. Secretary,Goa Konkani Akademi 243, Patto Colony, Panaji , Goa - 403 001: Goa Konkani Academy, 1972.