Comparison and usefulness of ASR11 scheme over previous schemes for transliteration and label set purposes for Indian languages.

Tauseef Hussain and Samudravijaya K
Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005
Email: tauseef@nitc.ac.in, chief@tifr.res.in

Abstract

An Automatic Speech Recognition system for any language, using present day software tools, requires that a conversion scheme is present which will transliterate words from the given language to ASCII text. In the case of Indic languages which have an almost phonetic based alphabet, certain simple yet consistent rules could be devised to construct such a scheme. Such a scheme is described here and its advantages over other schemes are mentioned. In order to have a very visible one-to-one correspondence between lines of transliterated ASCII text and respective utterances in audio files it is suitable that the name of the audio file reflects the transliterated content of the file. The modification of the scheme to adequately suit the above purpose is also discussed.

1 Introduction

The Technology Development in Indian Languages [7] program of Department of Information Technology has sponsored a Automatic Speech Recognition (ASR) project in consortium mode [15]. The goal of the project is to develop a voice interface to an existing website http://www.agmarknet.nic.in/ that displays the prices of commodities traded in various mandis all over India. The consortium consists of educational and scientific institutions which would develop associated ASR systems for six Indian languages viz., Assamese, Bengali, Hindi, Marathi, Tamil, Telegu. Accordingly the consortium members devised a common label set (referred to as ASR10) for the above mentioned languages [1][2]. This paper describes this common label set ASR10 and its advantages over earlier label sets and transliteration schemes. A useful feature associated with the use of a modified version of ASR10 is also presented.

2 Representation of Linguistic Resources for an ASR system

A spoken sentence is represented as a sequence of acoustic-phonetic units such as phones. The following linguistic resources are required for training an ASR system :

(a) Audio Wav files containing utterances .
(b) Transcription (or transliterated text) of the above wav files.
(c) A list of basic acoustic units in terms of which spoken words of the language will be described/represented for speech recognition. In most cases, the acoustic units correspond to phones of the language.
(d) A pronunciation dictionary: Here each word in the transcription will be represented as a sequence of the labels of the acoustic units that compose the word.

A standard encoding scheme such as UTF-8 [13] is available to display the script for Indian languages on computers. However current software tools used in ASR systems have not been upgraded to process the format of transliterated text in a format other than ASCII.

Therefore a task in ASR system building would be to represent the transliterated sentences as well as the basic acoustic-phonetic units in ASCII text. A typical example of transliteration of a spoken Hindi sentence to its counterpart in ASCII is given next.

तिरुवनन्तपुरम् से जम्मु_तवि के लिये एसी फर्स्ट क्लास का किराया कितना होगा.

Tiruvanantapuram se jammu_tawi ke liye ac first class kaa kiraayaa kitna hogaa.

A typical example of phone wise separation (given a predefined label set of phones) of the first word of above transliterated sentence would be :
Tiruvanantapuram   t i r uu v a n a n t a p u r a m.

A pronunciation dictionary would contain entries such as above, one entry per line. Hence there is a need for a scheme/convention to help construct the transliterated text and the pronunciation dictionary. More often, when such a scheme serves both the purposes it is called as a sound based transliteration scheme. A label set such as ASR10 would be one which represents the individual phonetic sounds by labels. It must be emphasized here that in the stricter sense transliteration scheme is needed for ASCII representation of original script while label set is needed for ASCII representation of the phones or various sounds of language(s). In the pronunciation dictionary whilst one scheme may be used to represent the words in ASCII text, another may be used to represent (in ASCII) the phones for those words in the pronunciation dictionary. A desirable feature of a label set is that there are some simple yet consistent rules to match the acoustic phonetic properties of the phones of a language with the labels. For example the labels should reflect the place of articulation and manner of articulation in a consistent manner. The label set should also preferably be case independent and operating system independent. Existing schemes did not satisfy all the desired features of such a scheme. Hence the ASR consortium decided to devise such a scheme (ASR10) that satisfies all the above mentioned features.

3 Overview of schemes for transliteration and label set purposes for Indic scripts

Indic scripts have originated from *Brahmi* script [8]. Devanagari script, which is used to write Hindi and Marathi, is part of the Brahmic family part of scripts. The letter order of Devanagari, like nearly all Brahmi scripts, is based on phonetic principles which consider both the manner and place of articulation of the consonants and vowels they represent. The format of Devanagari for Sanskrit serves as the prototype for its application, with minor variations or additions, to other languages [9]. Hence there is an almost common phonetic based alphabet among Indian languages. Unlike in English, vowel consonant in Devanagari does not change its phonetic value to denote some other character in the alphabet. For instance the way 'c' is used to spell ‘cat’ in English is different from the way it is used to spell 'China' or 'chutzpah'. Hence there is an accepted notion of What You Speak Is What You Write (WYSIWYW) [8] for Indic scripts.

Even though at first it may seem that there is a one-to-one correspondence between the representation of written script of Indian languages and the representation of acoustic-phonetic sounds in them, it is not exact. At best there is a near one-to-one correspondence. Following examples are cited to support the equivocalness in terming Indic scripts to be absolutely phonetic.

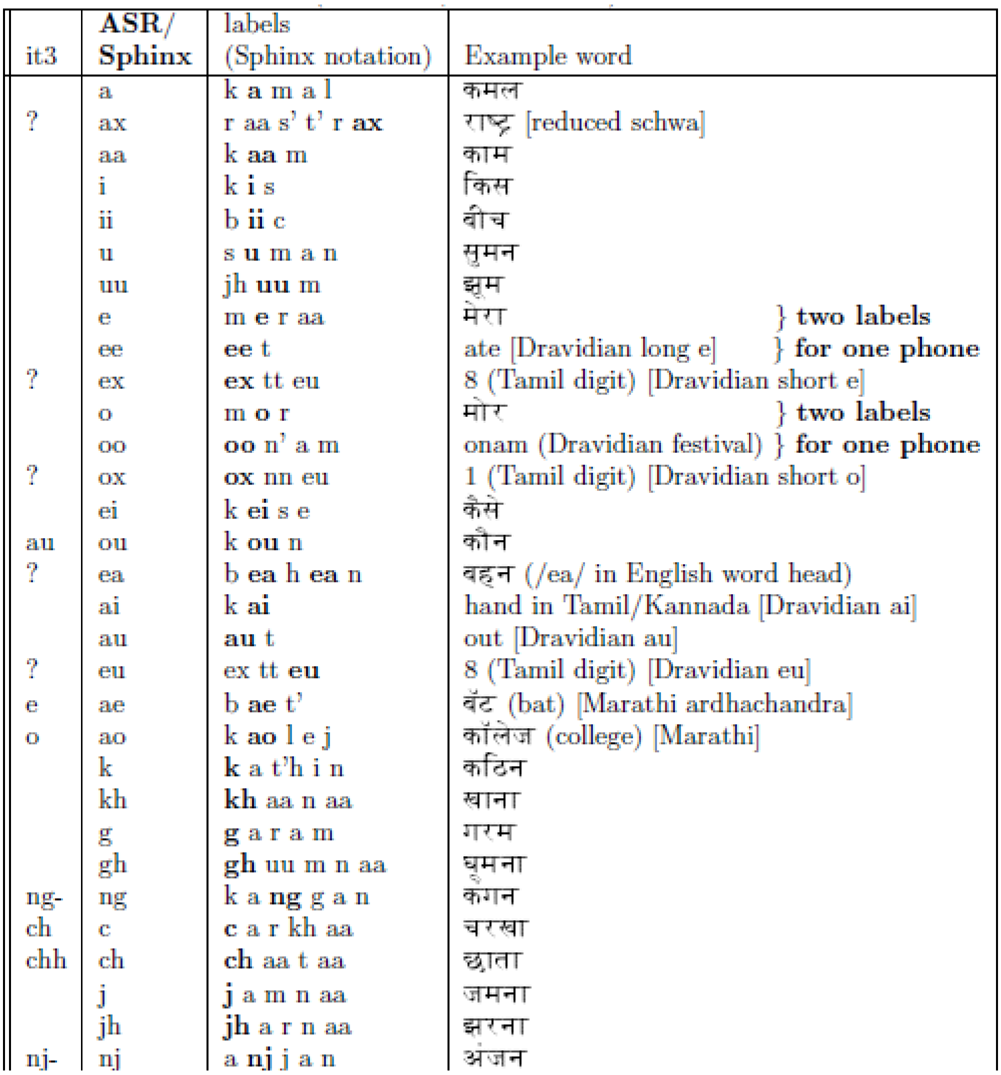
(1) Consider the way कमल is often written and pronounced. Strictly speaking there must occur a halanth at the end of last letter to indicate the way it is often pronounced.
(2) Consider the many variants of usage of the sound indicated by the letter letter र such as in कृषि , राष्ट्र or बर्तन.
(3) Use of different writing style in different languages to represent same phone sound e.g. Marathi बँक and Hindi बैंक which refer to same word and also pronounced in same way.
(4) In certain cases same matra sound is pronounced differently in different languages e.g. ऐ sound as uttered in कैसे in Hindi and that of मैना in Marathi.
(5) In certain cases depending on context, same word can sound and thus mean different. In Marathi जग could either have a palatal affricate or a dental affricate for the first consonant . In first case it would mean ‘the world’ and in second it would mean ‘to live’.

However, in spite of such exceptions, the commonality of phonetic sounds in various Indian languages led to attempts to create sound based ASCII transliteration scheme which could represent common, if not all, acoustic-phonetic sounds in Indian Languages.

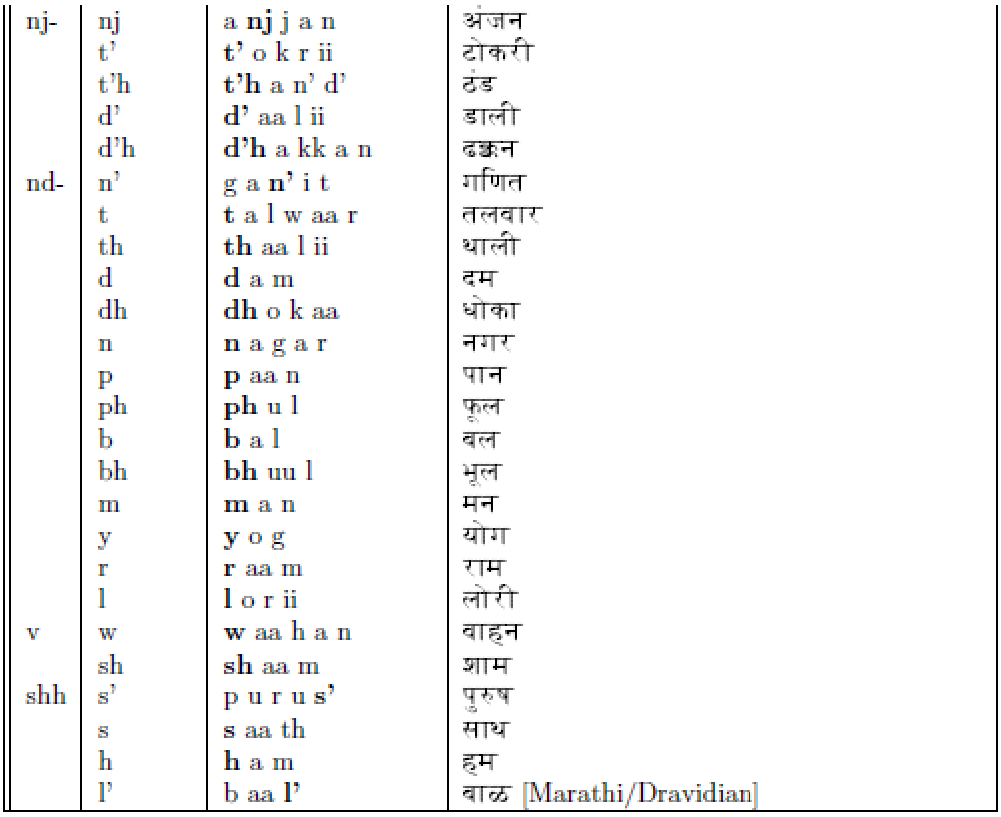
Several schemes/conventions have been proposed in the past to cater to transliteration and label set purposes for Indian languages. Some notable schemes are ITRANS [4], IT3 [5][6], INSROT [14]. Each of these schemes was devised to improvise upon an already existing scheme. For instance prior to IT3 there were many transliteration schemes such as ITRANS to key-in the Indian language scripts [5]. The focus of those schemes was mainly to represent the script of Indian languages. They paid less attention on the importance of user readability of the transliterated text. They used key combinations, for different Indian characters, which was difficult for user to remember. The ITRANS transliteration scheme was developed for the ITRANS software package, a pre-processor for Indic scripts. The user inputs in anglicized Roman letters found on QWERTY keyboards and the ITRANS preprocessor converts the Roman letters into Devanāgarī (or other Indic scripts). The ITRANS package is developed by Avinash Chopde which makes use of an approach to printing documents through LATEX [10]. ITRANS now has the support for several Indian languages but the transliteration scheme is not uniform for all Indian languages [10].

ITRANS used capital letters in its scheme e.g. retroflex sounds such as ट , ठ , ड , ढ were represented by Ta, Tha, Da and Dha respectively. TIFR 1998 [12] scheme also used capital letters for retroflex sounds. This was not suitable for searching contents on Internet and hence a case insensitive transliteration scheme INSROT (Indian Script Roman Transliteration) Table was proposed in [14] to facilitate searching Indic script related contents on web. A revised version of INSROT appeared in [8]. IT3, which is an acoustic sound based case-insensitive transliteration scheme, was developed by IISc Bangalore, India and Carnegie Mellon University [5]. Amongst the major features of IT3 was the phonetic nature of script which implied that the characters corresponded to the actual sound being spoken and hence IT3 could be used for all Indian languages [5][6]. Both INSROT and IT3 did not suffer from problems of case sensitivity. ASR10 was devised to further improvise upon these above mentioned schemes.

Figure 1 gives the label set of ASR10



it3	**ASR/ Sphinx**	labels (Sphinx notation)	Example word
	a	k **a** m a l	कमल
?	ax	r aa s' t' r **ax**	राष्ट्र [reduced schwa]
	aa	k **aa** m	काम
	i	k **i** s	किस
	ii	b **ii** c	बीच
	u	s **u** m a n	सुमन
	uu	jh **uu** m	झूम
	e	m **e** r aa	मेरा } **two labels**
	ee	**ee** t	ate [Dravidian long e] } **for one phone**
?	ex	**ex** tt eu	8 (Tamil digit) [Dravidian short e]
	o	m **o** r	मोर } **two labels**
	oo	**oo** n' a m	onam (Dravidian festival) } **for one phone**
?	ox	**ox** nn eu	1 (Tamil digit) [Dravidian short o]
	ei	k **ei** s e	कैसे
au	ou	k **ou** n	कौन
?	ea	b **ea** h **ea** n	बह्न (/ea/ in English word head)
	ai	k **ai**	hand in Tamil/Kannada [Dravidian ai]
	au	**au** t	out [Dravidian au]
?	eu	ex tt **eu**	8 (Tamil digit) [Dravidian eu]
e	ae	b **ae** t'	बॅट (bat) [Marathi ardhachandra]
o	ao	k **ao** l e j	कॉलेज (college) [Marathi]
	k	**k** a t'h i n	कठिन
	kh	**kh** aa n aa	खाना
	g	**g** a r a m	गरम
	gh	**gh** uu m n aa	घूमना
ng-	ng	k a **ng** g a n	कंगन
ch	c	**c** a r kh aa	चरखा
chh	ch	**ch** aa t aa	छाता
	j	**j** a m n aa	जमना
	jh	**jh** a r n aa	झरना
nj-	nj	a **nj** j a n	अंजन



nj-	nj	a **nj** j a n	अंजन
	t’	**t’** o k r ii	टोकरी
	t’h	**t’h** a n’ d’	ठंड
	d’	**d’** aa l ii	डाली
	d’h	**d’h** a kk a n	ढक्कन
nd-	n’	g a **n’** i t	गणित
	t	**t** a l w aa r	तलवार
	th	**th** aa l ii	थाली
	d	**d** a m	दम
	dh	**dh** o k aa	धोका
	n	**n** a g a r	नगर
	p	**p** aa n	पान
	ph	**ph** u l	फूल
	b	**b** a l	बल
	bh	**bh** uu l	भूल
	m	**m** a n	मन
	y	**y** o g	योग
	r	**r** aa m	राम
	l	**l** o r ii	लोरी
v	w	**w** aa h a n	वाहन
	sh	**sh** aa m	शाम
shh	s’	p u r u **s’**	पुरुष
	s	**s** aa th	साथ
	h	**h** a m	हम
	l’	b aa **l’**	बाळ [Marathi/Dravidian]

Figure 1. ASR10 label set

The first column in Figure 1 gives the corresponding IT3 symbols for some phones and ? against some phones meant that such labels are not available in IT3 for those phones. The second column gives the label in ASR10 for a phone. The third column gave a phone wise separation of words appearing in fourth column. Table 1 gives the representation of words/letters that shows certain phonemic features in languages governed by Devanagari script.

	Aspiration (without nukta) and Frication(with nukta)	Retroflexion(with /without nukta)	Nasalized vowel	Reduced vowel	English vowel	Sibilants
	ख फ ख़ फ़	ट ड़ ष	आँखें	राष्ट्र	बैंक	सारांश
ITRANS	kha pha Ka fa	Ta .Da Sha	aa.Nkhe.n	raashhTra	baiMka	saaraaMsha
INSROT	kha pha kh'a ph'a	t'a d'.a s'a	aam_'khem'	raas't'ra	baim'ka	saaraam'sha
IT3	kha pha kh~a ph~a	t'a d~a shha	aam:akhen'	raashht'ra	bain'ka	saaraan'sha
ASR10	kha pha khqa phqa	t'a d'qa s'a	aangkhen	raas't'rax	baengk	saaraansh

Table 1. Representation of words/letters to show certain phonemic features in Indic languages

Table 2 gives the representation of native Indic script influenced by typical acoustic -phonetic variations .

	Half form letter	Use of :	Non join	Variants of bindu usage	Geminated sounds	English sounds
	रत्न रतन	दुःख	पक्व पक् व	अंक अंत कंपन	अम्मा अण्णा	fool zoo
ITRANS	ratna ratan	duHkha	pakva pak.hva	aMka/a.nka aMta kaMpan	ammaa aNNaa	?
INSROT	ratna ratana	du:'kha	pakva pak\va	am'ka am'ta kam'pan	ammaa an'n'a	?
IT3	ratna ratana	duh'kha/duh:kha	pakva ?	an'ka an'ta kan'pana	ammaa and-nd-aa	?
ASR10	ratna ratan	duhkh	pakva ?	a ngk a n t k a m p a n	a mm a a nn' a	f oo l z uu

?= no such transliteration available

Table 2. Representation of native Indic script influenced by typical acoustic -phonetic variations

For the unaspirated plosives such as च ,ट ,त ,ज ,ड ,द both IT3 and INSROT assigned the labels cha, t'a, ta, ja, d'a, da respectively. IT3 used labels chh and shh to represent छ (a palatal aspirated affricate as in छाता) and ष (a retroflex fricative as in पुरुष) respectively. INSROT used labels involving the quote character such as c'ha for छ and also for all retroflex sounds such as t'a for ट . Following a simple rule of using h to denote aspiration ASR10 assigned ch for छ, gh for घ and so on. Following a simple rule of using ' to denote retroflexion ASR10 assigned s' for ष, n' for ण and so on. ASR10 accounted for additional sounds such as that of reduced vowel e.g. reduced schwa occuring at end of uttering राष्ट्र. Such a provision was absent in IT3 and INSROT. Despite its usefulness in transliterating purposes for Indic scripts, IT3 mapping table was not adequate to represent sounds (indicated by ? in Fig 1) from many different Indian languages and hence additional symbol set specific to a language such as that for Marathi had to be conceived for transliterating Marathi.

The label set ASR10 [2] as devised by the ASR consortium [1] would be among, but not uniquely, the first attempt to provide a set of phone-like labels for common Indian language sounds and hence for Indian Language Automatic Speech Recognition. Some of the provisions for phonetic representations, taking care of various sounds in the six languages of ASR consortium, as purportedly devised by ASR10 could be summarised as below :

a) Special use of certain suffixes:
    1. Aspiration : Suffix h is used to denote aspiration e.g. k (क) versus kh (ख).
    2. Retroflex consonants : Quote ' is appended to denote a retroflex constant e.g. t (त) versus t' (ट).
    3. Flap : Suffix q is used to denote a flap (nukta in Hindi) e.g. d' a g a r (डगर) versus lad'qkaa (लड़का).
    4. Nasalized vowel : Suffix n is used to denote nasalisation of a vowel e.g. kahaa (कहा) versus kahaan (कहाँ).
    5. Reduced vowel : Suffix x is used to denote reduction of a vowel e.g. राष्ट्र and also to denote two dravidian vowels ex (dravidian short e used in Tamil digit 8), ox (dravidian short o used in Tamil digit 1).
    6. Dental affricates : Suffix – is used to denote dental affricate of Marathi.

b) Provision was made for having two labels for one phone as illustrated by utterances of vowel sounds in Hindi मेरा and the dravidian long e as in ate.

c) Two labels for phone as illustrated by utterances of vowel sounds in Hindi मोर and Dravidian elongated o as in onam.

d) Label for the /ea/ sound as in English word head.

e) Label for the Dravidian au sound as given by /ou/ in Engish word out.

f) Label for Marathi archachandra as represented by बॅट .

g) Label for Marathi/Dravidian sound of ळ in बाळ .

h) Label for Malayalam retroflex r.

i) Assign different labels for bindu usage in different letters. When bindu was used with stop consonants belonging to the same place of articulation it was substituted with the symbol for the nasal consonant belonging to that place of articulation (Refer e.g. of ASR10 representation fo r bindu substitution in Table 2) .

4 Overview of tools and stages during transliteration process

Current existing methodologies for transliteration do not provide a tool to transliterate from Marathi script in UTF-8 (encoding of Unicode) to any of the known conventions (such as IT3) which can be written in ASCII. Hence with the already available tools and new scripts/programs a method was adopted to start with a commonly used transliteration convention, switch between conventions to arrive at the ASR10 scheme for transliteration. Figures next illustrate the stages leading upto a transliteration as per ASR10 scheme.

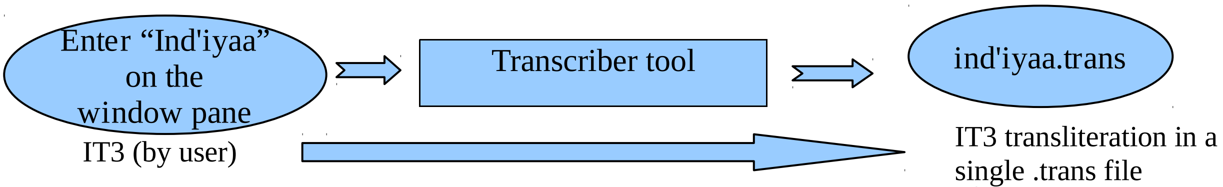
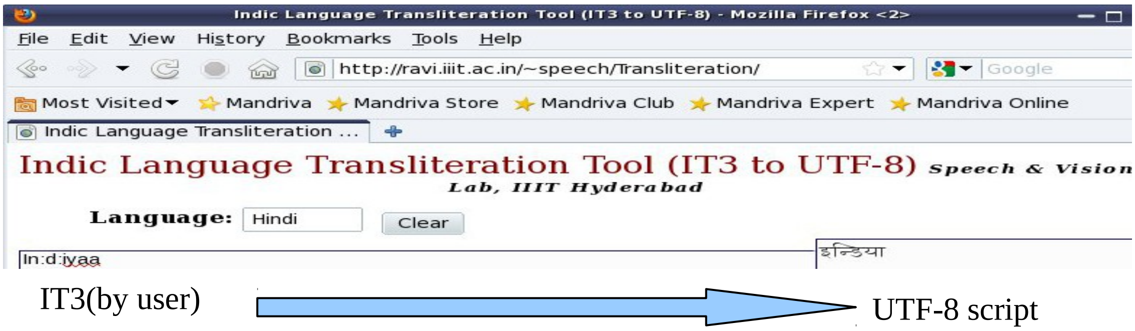

Enter "Ind'iyaa" on the window pane
Transcriber tool
ind'iyaa.trans
IT3 (by user)
IT3 transliteration in a single .trans file

Figure 2. Transcriber tool to get transliterated text in a single file



Indic Language Transliteration Tool (IT3 to UTF-8) - Mozilla Firefox <2>

http://ravi.iiit.ac.in/~speech/Transliteration/

Indic Language Transliteration Tool (IT3 to UTF-8) Speech & Vision Lab, IIIT Hyderabad

Language: Hindi

In:d:iyaa

इन्डिया

Figure 3. Indic Language Transliteration Tool by Speech and Vision Lab (IIT Hyderabad)

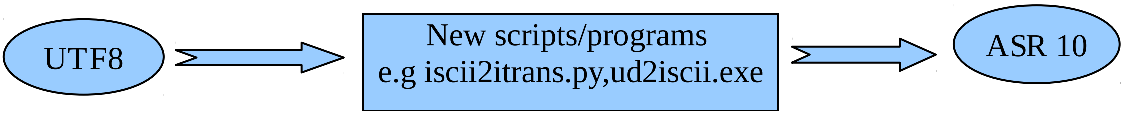


Figure 4. Conversion from UTF-8 to ASR10

As can be seen in Figure 2, first a transcriber tool takes the IT3 transliterated text in its pop up window as input by the user and generates a .trans file which contains the IT3 transliterated text. This tool also allows the user to listen to the audio wav whose content is to to be transliterated as per IT3 scheme. Hence a preliminary aural verification is also in place while transcribing in this manner. Whenever there is a need to get the original language script counterpart of the transliterated text, HTML based Indic Language Transliteration Tool [11] (developed by IIIT Hyderabad) (Figure 3) is used to get the conversion from IT3 wise transliterated text to the appropriate native script in UTF-8 format. This helps to ensure correctness of the IT3 wise transliterated text by visual inspection of its counterpart as both the IT3 format and UTF-8 format texts appear side by side in two windows of the html web page. This provision of HTML based transliteration tool helps in the tool being operating system and Unix scripts independent.

Finally with the use of Unix shell scripts and executable programs (Figure 4) the UTF-8 format text is converted to transliterated text as per ASR10 scheme. Descibed above, IT3, rather than ASR10 makes a good choice for first level of transliteration because of generally accepted standardisation of IT3 scheme as well as availablity of html based Indic Language Transliteration Tool to immediately verify that IT3 wise transliterated text reflects back the original native script in UTF 8.

5 ASR11 : An enhanced version of ASR10

ASR10 scheme was used for representing the transliterated text as well as the label set i.e the two columns of the dictionary were as per ASR10 scheme. A desired objective was to adopt a transliteration scheme whereby ASCII alphanumeric characters could be used so that each transliterated utterance was used in the naming of the corresponding audio wav file. The rationale for having the audio wav file same as that of its transliterated content would be to ease natural language processing by scripts and tools. To illustrate this consider an audio wav file which contains the utterance of the word saayekheda then the audio wav would be saayekheda.wav.

ASR10 used special characters such as quote('), colon(:) and dash(-) for representation of its label set and these characters have a special meaning in the context of Unix based environments e.g. some Unix based scripting languages treat these special characters in a different way as compared to other ASCII characters. For instance, quote or forward tics ' (situated next to Enter key on a QWERTY keyboard) is used in Unix shell/Perl scripts where character or command substitution is not required. Similar other special uses are there for colon and dash. This meant that names containing these special characters could not be used as names for the audio wav files. Another problem occured in Windows based OS where such audio wav files, with special characters in their names, could not be played.

To deal with the aforementioned problems ASR11 was devised to provide alternatives to usage of quote ('), colon (:), dash (-) so that transliterated text (in ASCII) could still be used for naming of corresponding audio wav files and also be suitable for Unix script based processing.

5.1 *Brief description*

The revised ASR11 [3] scheme consisted a revision of label set provided by ASR10 so as to use alphanumeric characters limited to a-z, A-Z, roman numerals 0-9 and underscore _ .
Figure below gives changes of ASR11 scheme over ASR10 scheme [2].
Following are the changes of ASR11 over ASR10 :

a) Quote ', which was used to indicate retroflex sounds and Dash – , which was used to indicate dental affricates , were replaced by x
b) Colon : was involved to indicate chandrabindu and bindu (dot above). Its use was substituted by q.
c) Use of q for nukta (dot below) was substituted by use of Q

Table 3 next illustrates the ASR11 scheme with examples.

Grapheme	Symbols converted		Example as per ASR10	Example as per ASR11
	ASR10	ASR11		
Retroflex plosives	{t,d}[']\[h]	{t,d}[x]\[h]	aat'ha(आठ)	aatxha
Retroflex nasal	n'	nx	gan'esha(गणेश)	ganxesha
Retroflex lateral	l'	lx	mul'aa(मुळा)	muulxaa
Retroflex liquid	s'	sx	purus'a (पुरुष)	purusxa
Retroflex fricative	r'	rx	Tamil retroflex r = vocalic /r/	
Dental Affricate	j- , c-	jx , cx	pyaaj- (प्याज)	pyaajx
Candrabindu	m:	mq	kahaam: (कहाँ)	kahaamq
Bindu	n:	q	an:ka (अंक)	aqka
Nukta	q	Q	chattiisagadxhaq(छत्तीसगढ़)	chattiisagadxhaQ

Table 3. ASR11 scheme

The pronunciation dictionary would now contain the text/words on left column in ASR11 format while right side phone wise separation would be in ASR10 format. An example of such an audio file would be jaamakhedxa.wav while its entry in pronunciation dictionary would look like below.
jaamakhedxa   j- aa m kh e d'

6 Conclusion and Future Work

In this paper the advantages of ASR10 scheme were described over earlier transliteration and label set schemes. A slightly modified version of ASR10, i.e. ASR11 that solved problems associated with naming audio wav files, was presented. Current ASR11 scheme does not provide for labels based on tonal sound changes e.g. reduction of pitch while uttering the word भारत in Punjabi. Future work would be focused in incorporating labels as per tonal sound changes by suitable labels.

7 Acknowledgements

The contribution from the following people associated with Marathi ASR project viz., Dr. Samudravijaya K, Dr. Nalini Bondale, Tejas Godambe, Nikita Acharekar of Tata Institute of Fundamental Research and Dr. Preeti Rao, Pranav Jawale of Indian Institute of Technology Bombay is both appreciated and acknowledged with regard to development of this work.

8 References

[1] “Phonetic Dictionary”, http://speech.tenet.res.in/wiki/uploads/3/39/LabelSetAsrConsortium_v0.1.pdf

[2] “Label Set ASR10 by ASR consortium”, http://speech.tifr.res.in/resources/data/labelSetASR100815.pdf

[3] “ASR11 transliteration scheme”, http://speech.tifr.res.in/asrProject/transcription/marathiTransliteration110601.pdf

[4] Avinash Chopde, “ITRANS - Indian Language Transliteration Package”
Version 5.3, http://www.aczoom.com/itrans/

[5] Prahallad Lavanya, Prahallad Kishore, Ganapathiraju Madhavi, “A simple approach for building transliteration

editors for Indian languages", Journal of Zhejiang University ISSN 1009-3095, Sep. 2005.

[6] Madhavi Ganapathiraju, Mini Balakrishnan, N.Balakrishnan, Raj Reddy, 2005. "Om: One Tool for Many (Indian) Languages". Proceedings from the International Conference on Universal Digital Library (ICUDL), Hangzhou, China. Journal of Zhejiang University SCIENCE, 6A(11):1348-1353.

[7] Technology Development for Indian Languages, http://tdil.mit.gov.in/

[8] "INSROT – Indian Script Roman Transliteration Table", Journal of Language Technology, ISSN No. 0972-6454, No.9, Apr. 2003, pp.138-142, http://tdil.mit.gov.in/insrotapril03.pdf

[9] "Wikipedia entry on Devanagari", http://en.wikipedia.org/wiki/Devanagari

[10] Systems Development Laboratory Indian Institute of Technology Madras, "Acharya Website", http://acharya.iitm.ac.in/multi_sys/transli/schemes.php

[11] "Indic Language Transliteration Tool", http://ravi.iiit.ac.in/~speech/Transliteration/

[12] Samudravijaya K, P.V.S.Raos and S.S.Agrawal, "Hindi Speech Database", Proc. Int. Conf. on Spoken Language processing(ICSLP00), Beijing, China, October 2000.

[13] "Wikipedia entry on UTF-8", http://en.wikipedia.org/wiki/UTF-8.

[14] "8 INSROT – Indian Script Roman Transliteration Table", Journal of Language Technology, ISSN No. 0972-6454, No.9, Jul. 2002, pp.31-32, http://tdil.mit.gov.in/tdil-jan-2002.zip

[15] "DIT Consortium Project Wiki", http://speech.tenet.res.in/wiki/index.php/Main_Page

[16] "Pronunciation Specification for LexiconDevelopment", http://www.w3cindia.in/presentations/PLS-workshop/milton.pdf