The Phonemic model from India for Bi-modal Applications

Ragunath K. Joshi
Visiting Design Specialist
Centre for Development of
Advanced Computing (formerly NCST)
Gulmohar Cross Road No. 9
Juhu, Mumbai 400 049,India.
Web: http://staff.ncst.ernet.in/rkjoshi
Email: rkjoshi@cdacmumbai.in

The Indian oral tradition

In ancient India various types of information as well as instructions were imparted through oral (vak) and listening (shravana) mode. The ancient sacred Vedas are reservoir of information, knowledge and wisdom of Indian past. This knowledge has been passed on from the teacher to the student through guru-shishya parampara using oral mode and its associate techniques such as recitation, repetition, memorization and oral reproductions, etc. Clear and correct pronunciations of syllabic clusters by the teacher as well as attentive and meaningful listening by the pupils were integral parts of this methodology. In addition to 4 vedas (Rigveda, Yajurveda, Atharvaveda and Samveda) the ancient text such as Brahmans, Aaranyaks and Upanishads dealt with theoretical, philosophical and practical aspects of human life and beyond. Enormous amount of information on various topics was thus preserved from generation to generation in the Indian sub-continent through oral tradition. Therefore the detailed study of phonetics became an important activity to support oral media. The Shikshas and Pratishakhyas attached to different Vedas and shakha/upshakhas, constitute parametric studies on phonetics and formations of linguistic elements. The Rigveda Pratishakhya (by Shaunak) dealt with the norms of phonetics as applied to Rigveda whereas Samvedic `Gaan' tradition has been dealt in the Samaveda Pratishakhya by Pushpa. Two branches (Shakha) of Yajurveda namely shukla (white) and Krishna (Black), have contributed significantly towards vedic phonetics through their own Pratishakhyas.

The Taittiriya Pratishakhya deals with the ancient phonetic system as applicable to the Taittiriya Samhita of the Black Yajurveda. A well-structured treatise on phonetics, this Pratishakhya deals with subject matter of phonetics into three groups: 1. Sudharana-vidhi consisting of enumeration and classification of sounds of alphabets vowels, diphthongs and consonants, 2. Samhitadhikara rules for the construction of euphonically combined text and 3. Uccarankalpa formation of articulate sounds and mode of their production. All rules laid down in this Pratishakhya were applicable to the recitation of, the Samhita, Pada, Kramapatha and also at times to the Jata patha. `Patha' being the most effectively uttered text for a given purpose. The elaborate explanations of various terms used in the process were provided by Sanskrit scholars, grammarians using Sutras in various shikshas.

A documentation of such oral information/text became necessary in due course and therefore a writing system based on phonetic principles was developed. A spoken word was transferred into a written word. The minimal sound unit called 'Varna' (phoneme) was identified as a root base for the writing system. 'Varnas' were

divided into two groups: Consonants and Vowels and into further classifications. This range was called Varna Samamnaya.. In the context, it is observed that 4 types of ARya Varnamalas (VarnaMatruKas) agree on the fundamental separate existence of vowel sounds and consonant sounds (Varna) as a basic linguistic matter. There is also an agreement on the classification and the quantity of the SparshaVarnas (as 5 Vargas x 5 = 25). However there are some differences, mainly on the types and counts of vowels, semivowels (Antashta Varnas) as well as fricatives. For example the Shukla Pratishakhya identifies 23 vowels and 42 consonants + others, whereas the Taitariya Pratishakya refers to 60 Varnas constituting 16 vowels and 44 others. Paninan Grammar identified 42 representative root Varnas from which 190 Speech sounds (Upadhwanis) could be created by adding various parameters to vowels (time, stress, nasalisation etc.) and consonants. Systematic combinations of elements of these groups resulted into a vast range of written syllables (Aksharas). Also, an elaborate markup system (Vaidik Tonal Signs) was developed to write phonetic nuances of 'Varnas'. (Udatta, Anudatta, Swarit, Kamp etc.) Thus the parlance and conversions between textual and oral modes were established through elaborate orthographic representation of phonetic syllabic structure. Such a well-defined model of written text with its correlation to spoken text exists in India and is found in many ancient manuscripts in Sanskrit language.

Tasks undertaken

Panini's classification of the Indian alphabet into pure consonants (C) and vowels (V) and the simple rendering rules as 1. In a sequence of vowels and consonants (pure), say VCCVVCVCCCVVV..., the vowel imediately following a consonant must always be rendered in its matra form (dependent). All other vowels must be rendered in their stand alone form (independent). 2. The sequence of consonants imediately preceeding a vowel form a conjunct (ligature). Constitutes the dynamic essence of Indian writing has served as the basis of the following implementation tasks which were undertaken:

1. In early 70s R. K. Joshi worked out a project called Deshanagari, a common script for all Indian languages using phonemic approach as a common link in all Indian languages. 7 bit Desha codes were designed by him for ISCII in 1982, at NCSDCT, TIFR.
2. Vividha, the multilingual text processor designed and developed at NCST was based on phonemic approach and was called pure consonant approach (refered as dead consonant in Unicode).
3. Vidura, an electronic publishing environment designed and developed at NCST (for IGNCA) included text processing applications supported by Vividha coding scheme and I/O devices. This environment was based on phonemic approach.
4. Many bilingual multilingual turnkey projects undertaken during 1985 to 1990 at NCST used phonemic approach for Indian language text processing very effectively.
5. The present IndiX2 environment (2005) on Linux platform for Vedic Sanskrit language text processing strongly supports the phonemic approach which has been successfully implemented using proposed Vedic codes.

Vedic Sanskrit Coding Scheme

A phonemic based, additive approach for text processing in computers for interchanging text and speech. Sanskrit Grammar has distinguished the terms Varna (phoneme) and Akshara (syllable) in the context of spoken languages and written languages. Since the oral tradition in India was of a higher order, the right pronunciation was given utmost importance. Various chinhas (signs) were

introduced in Vedic Sanskrit to reflect the speech nuances in written language, and to strike the equivalence in spoken and written expressions. The realization of such a system in the context of new technology seems to be imperative where writing is seen in the context of speech and speech in the context of writing. The scheme attempts to maintain the correspondence between spoken and written syllables firstly by giving each phoneme a distinct code and secondly by giving each chinha denoting nuances of speech, a distinct code.

Range of characters

As compared to modern historical derivatives from Sanskrit such as Hindi, Marathi, Nepali etc., the Vedic Sanskrit text demands adequate range of characters as well as exhaustive rendering rules to achieve the advanced typographic quality. The range includes Swara Varnas, Swara Bhedak Chinhas, Vyanjan Varnas, Vyanjan Bhedak Chinhas, various other chinhas to show nuances of spoken language like kaal (time duration), bala (stress), kamp (vibration) as well as specific Vedic chinhas such as Vedic Anuswaras, Visargas, Swaritas and Saamvedic intonation chinhas. The provision has been also made for further speech-related controls such as the tonality, peculiarity of voice, pauses, etc.

The Code charts

The coding scheme consists of two charts. The first chart is meant for text generation and processing in classical Sanskrit as well as for taking care of the spoken nuances from other Indian languages (marked with [R] in the chart). The second chart includes about 100 Vedic Sanskrit Chinhas needed for Rigveda, Yajurveda, Atharvaveda and includes Saamvedic intonation marks of various schools.

Vedic Code Chart 1

	080	081	082	083	084	085	086	087
0	०	[R]	ऐ	[R]	[R]	द्	[R]	LUPT
1	१	ड	[R]	क्	झ्	ध्	[R]	ARLU ◌
2	२	[R]	ओ	[R]	[R]	न्	ल्	◌ः
3	३	ढ	ओ	ख्	ञ्	[R]	[R]	◌ः
4	४	ण	[R]	[R]	ट्	[R]	व्	◌ः
5	५	[R]	आ	ग्	ठ्	प्	[R]	◌ः
6	६	ऊ	औ	[R]	ड्	फ्	श्	◌ः
7	७	[R]	◌ॅ	घ्	[R]	[R]	ष्	ऽ
8	८	ऋ	◌ॅ	ङ्	[R]	ब्	स्	।
9	९	ॠ	◌ॢ	च्	ढ्	[R]	ह्	SNDH □+□
A	अ	ऌ	◌ॅ	[R]	[R]	भ्	ळ्	ASND □ □
B	[R]	ऌ	ो	छ्	ण्	म्	[R]	RIKT
C	[R]	ऐ	◌ॅ	[R]	[R]	य्	[R]	ARRI □
D	[R]	ए	◌ं	ज्	त्	[R]	ळ्ह्	DRSK
E	आ	अं	◌ः	[R]	[R]	र्	[R]	VIST
F	[R]	[R]	[R]	[R]	थ्	[R]	[R]	[R]

Vedic Code Chart 2 (Revised December 2005)

	088	089	08A	08B	08C	08D	08E	08F
0	◌ॅ	◌ॢ	◌ॅ	◌ॅ	◌ॅ	◌ॅ	◌ॅ	[R]
1	◌ॅ	◌ॅ	◌ॅ	◌ॅ	◌ॅ	◌ॅ	◌ॅ	[R]
2	◌ॅ	[R]	◌ॅ	◌ॅ	◌ॅ	◌ॅ	◌ॅ	[R]
3	◌ॅ	◌ॅ	◌ॅ	◌ॅ	[R]	◌ॅ	◌ॅ	[R]
4	◌ॅ	◌ॅ	◌ॅ	◌ॅ	[R]	◌ॅ	[R]	[R]
5	◌ॅ	◌ॅ	◌ॅ	◌ॅ	[R]	◌ॅ	◌ॅ	[R]
6	◌ॅ	◌ॅ	◌ॅ	◌ॅ	[R]	◌ॅ	◌ॅ	[R]
7	◌ॅ	◌ॅ	◌ॅ	◌ॅ	◌ॅ	◌ॅ	◌ॅ	[R]
8	◌ॅ	◌ॅ	◌ॅ	◌ॅ	◌ॅ	◌ॅ	◌ॅ	[R]
9	◌ॅ	◌ॅ	◌ॅ	◌ॅ	◌ॅ	◌ॅ	◌ॅ	[R]
A	◌ॅ	[R]	◌ॅ	◌ॅ	◌ॅ	◌ॅ	◌ॅ	[R]
B	◌ॅ	◌ॅ	◌ॅ	◌ॅ	◌ॅ	◌ॅ	◌ॅ	[R]
C	◌ॅ	◌ॅ	[R]	[R]	◌ॅ	◌ॅ	◌ॅ	[R]
D	◌ॅ	◌ॅ	[R]	◌ॅ	◌ॅ	◌ॅ	◌ॅ	[R]
E	◌ॅ	◌ॅ	◌ॅ	[R]	◌ॅ	◌ॅ	◌ॅ	[R]
F	◌ॅ	◌ॅ	◌ॅ	[R]	◌ॅ	◌ॅ	◌ॅ	[R]

1.Vedic Sanskrit Anka: **10** (U+0800 to U+0809)

2.Vedic Sanskrit Swara Varna: **18** (U+080A to U+0826)

3.Vedic Sanskrit Swara Bhedak Chinha: **5** (U+0827 to U+082B) to U+08AB)

4.Vedic Sanskrit Swaradi Chinha: **3** (U+082C to U+082E)

1.Vedic Sanskrit Phonetic Break-up Signs: **9** (U+0880 to 0888)

2.Vedic Sanskrit Anuswara: **16** (U+0889 to U+0899)

3.Vedic Sanskrit Ardha Visarga: **7** (U+089B, U+08A6

4.Vedic Sanskrit Visarga: **10** (U+089C to U+08A5)

5.Vedic Sanskrit Vyanjan Varna: **35** (U+0831 to U+086A)

6.Vedic Sanskrit Vyanjan Bhedak Chinha: **4** (U+0872 to U+0875)

7.Vedic Sanskrit Other Signs: **3** (U+0876 to U+0878)

toU+08E3)

5.Vedic Sanskrit Taarata Chinha: **2** (U+08AE to U+08AF)

6.Vedic Sanskrit Swarita Chinha: **13** (U+08B0 to U+08BD)

7.Vedic Sanskrit Swarita Kampa Chinha: **3** (U+08C0 to U+08C2)

8.Vedic Sanskrit Samavedic Swarochchar Chinha: **29** (U+08C7

9.Vedic Sanskrit Special symbols: **11** (U+08E5 to U+08EF)

Indian Languages and Bi-modal Applications

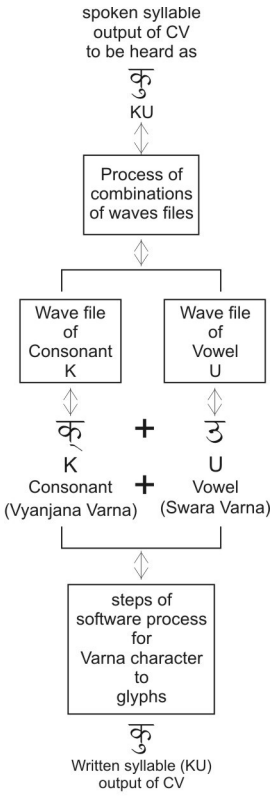
In the event of forthcoming sound/speech applications, there is an urgent need to look at text processing from the phonemic angle and not just from the orthographic angle. The phonemic scheme for Indian language will serve a good base for establishing parallel between the repertoire of linguistic spoken sounds and their counterpart in writing mode. The phonemic approach with its phonetic mode and pure consonant makes the processing task very easy, convenient and logical. Most important of all, it uses the classic traditional phonetic system of joining consonant and vowel into a syllable in a set pattern, unique to India. Thus, the software complexity of supporting Indian languages in different applications can be controlled by the use of a unified technique based on phonemic codes.

Additive approach

Unlike the full consonant approach as present in ISCII and Unicode which is of subtractive nature, this K model approach is purely additive, i.e., words through syllables are formed only by addition of phonemic units. As a result, this scheme can be more effective in text processing area like speech synthesis. In speech, sound units are added together and they are not subtracted. Any speech synthesis algorithm, which processes sound units linearly, may prove more effective with this phonemic scheme. Thus, the phonemic approach can well operate at bi-modal levels textual level would take care of text generation, text processing and linguistic operations. The oral mode would cater to the speech synthesis, recognition and advanced oral presentations.

The Position Statement

In realization of the further potential of exploratory applications it is proposed that:
- Varna (phoneme) should be treated as a basic smallest unit of encoding, this 'K' model being the minimal and atomic/root of Indian syllables (written/spoken).
- In present standards (Unicode 4.0, ISCII 91) the basic smallest unit for encoding is treated as ‘KA’ (consonant varna K + vowel varna A). This encoding could be treated as at higher level.
- The K model should be used to draw parallelence between International Phonetic Alphabet (IPA) and Indian Phonemic Alphabet.
- The authentic phonetic breakup of words in Indian language dictionaries should be initiated using ‘K’ model.
- To adopt `K' model for Indian Languages for bi-modal applications including TTS and STT.



Model for TTS and STT, based on 'Varna' encoding (Vedic Sanskrit)

References

- Dr. Kelkar A.R. 1989. *Transliteration of South Asian languages, a brief review and a proposal for a standard.* Centre of advanced study in linguistics, Deccan College, Pune.
- *Handbook of the International Phonetic Association.* 1999. Cambridge University Press.
- Naravane V.D. 1961. *Bharatiya Vyavahar Kosh.* Triveni Sangam.
- Peter Ladefoged, 2001, *Vowels and Consonants.* Blackwell publishers, UK.
- *Prabodh Primer,* Department of official languages, Ministry of home affairs, Govt. of India.
- R. K. Joshi, Prague 2003, A unified phonemic code based scheme for effective processing of Indian languages. 23rd Internationalization and Unicode.
- R. K. Joshi. October 2002, Vedic Code, a draft, Vishwabharat No. 7.
- R. K. Joshi, December 2005, InPA – Indian Phonemic Alphabet for Bi-modal Applications – ICON 2005, IIT Kanpur.
- R. Shama Sastri, K. Rangacharya, 1985 reprint, *The Taittiriya Pratishakhya,* Motilal Banarasidas, Delhi.
- Shrinath Shanbaug, Durgesh Rau, R. K. Joshi 2002, An intelligent multi-layered input scheme for phonetic scripts. ACM International conference proceeding series, Hawthorne, New York.
- *Unicode 4.1* - The Unicode Consortium 2004. The Unicode standard is the universal character-encoding scheme for written characters and text. It defines a consistent way of encoding multilingual text through software.
- Vinod Kumar, 2005, IndiX information leaflet, C-DAC Mumbai.
- W.S. Allen, 1961, *Phonetics in Ancient India,* London, Oxford University Press.

14th April 2006