Indic extensions
Accent marks and
Concrete text

Raghunath 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.cdacmumbai.in/rkjoshi
Email: rkjoshi@cdacmumbai.in

Internationalizing W3C's Speech Synthesis Markup Language
Workshop III
13-14 January 2007, Hyderabad, India

The following 3 topics have been briefly covered in this paper.

1) Requirements for extensions to SSML to improve the synthesis of languages (Indic).
2) Representation of accent marks to convey tones and other linguistic features
3) CSS to ACSS and ACSS to CSS

Part 1
Requirements for extensions to SSML to improve the synthesis of languages (Indic).

The following 4 extensions are being suggested

1a) **<Say-as-if>**
It is generally observed that many Indian Languages, and their script orthography are authentically represented in the spoken counterpart. However, it is not totally true. The co-relationship between written and spoken word cannot be established to the full extent for many reasons. Some of the reasons are:

- The glyph range is not adequate
  i.e. Tamil K has to stand for kh g gh

- The glyph range is adequate but 'traditionally' the spoken sound does not relate to the 'supposed to be the glyph' but to yet another glyph. i.e. Bengali 'A' is written as 'A' but pronounced as 'Awa'

- The glyph range is not updated to accommodate `borrowed' sounds from another language i.e. Traditional Hindi (Devanagari) has had no provision for 'short e'.
  The Manak Hindi (extended Devanagari) has made provision for 'short e'.

The above behaviours can be taken care by requisite rules provided in the synthesis processor. Yet there are still some 'oddities' which may need the tag of <say-as-if>
i.e. Marathi - Samshaya meaning doubt is spelt as Samshaya and pronounced as Savshaya
Bengali – Shojjo meaning tolerate is spelt as Sha Ha YaPhala and pronounced as Shojjo

1b) **<say-to-self>**
This linguistic feature is philosophical in nature. The mode of 'thinking aloud' is represented by this extension. While in the thinking aloud stage the linguistic expression, the oral behaviour may not be grammatically 'exact'. Such speech (internalized yet externalized) may be added with positive (ya…, Aa…, Aaa…, ss… etc) or negative tones (n…nn…nnna…) with irregular breaks (pauses) in between. This behaviour can be observed in the contrast to the 'regular conversation between the two'. The tone of (as if) the inner mind expression may vary intermittently and some requisite specifications need to be worked out.

1c) **<say-bil>**

Many years ago, a linguistic survey was conducted regarding the 'bi-lingual switching over codes phenomenon' as observed in some pockets of south India. In this behaviour the speaker changed his language of expression all of a sudden and that too frequently. He/She switched over linguistic codes very easily, casually, as if it came naturally. Such bi-lingual spoken mode could be marked differently.
than the lang tag (?) - perhaps <say bil>. The written mode could use either the script change or the same script written in 'bidi' mode.
i.e. TAMILHSILGNE.

1d **<phps>**

Proper Name Phoneme String

There are variety of proper names in various Indian Languages which need to be pronounced correctly in the 'speech mode'. The usage of various scripts for the proper name have added orthographic variations of the same. To help to pronounce them correctly, it is suggested that the phoneme string of proper names could be entered as part of 'the body' of the text under this tag.
i.e. Rudradharma should be spelt with phonetic string as R U D R A Dh A R M AA.
Such phonemic strings will help to identify basic consonant/vowel sound categories and their codes. This will further help in the proper concatenation techniques.

Part2

Representation of accent marks to convey tones and other linguistic features

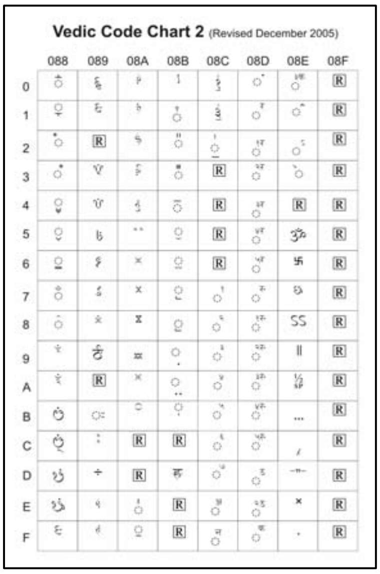
Sanskrit Grammar has distinguished the terms Varna (phoneme) and Akshara (syllable) in the context of spoken languages and written languages. Since the oral tradition in India was of a higher order, the right pronunciation was given utmost importance. Various chinhas (signs) were introduced in Vedic Sanskrit to reflect he 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 context of speech and speech in context of writing.

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 sign 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, pecularity of voice, pauses, etc. Vedic Sanskrit includes multi-tier usages of diacritic marks of complex compositions, above, below and at sides of the characters.

The chart includes about 100 Vedic Sanskrit Chinhas needed for Rugveda, Yajurveda, Atharveda and includes Saamvedic intonation marks of various schools. These diacritic marks can also be used further for indicating intonations needed for spoken words from various dialects of Indian languages.

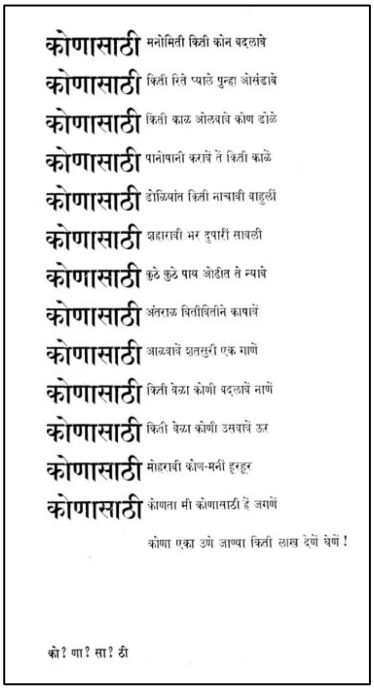
Vedic Sanskrit tonal/accent marks (chart on the next page)

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 to U+08AB)
4.Vedic Sanskrit Visarga: **10** (U+089C to U+08A5)
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 toU+08E3)
9.Vedic Sanskrit Special symbols: **11** (U+08E5 to U+08EF)



Part3
CSS to ACSS and ACSS to CSS

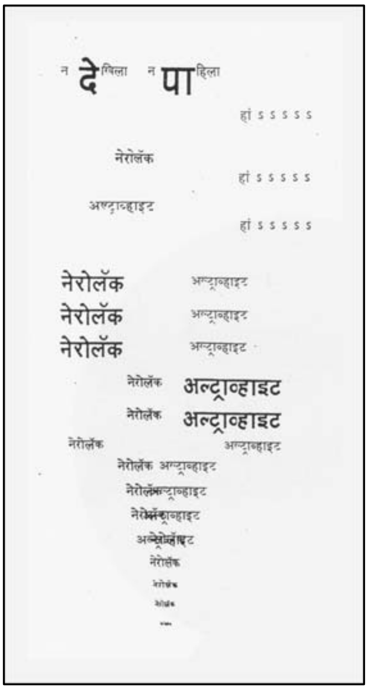
The different subtle shades of utterances is an integral part of a spoken mode and the aim of ACSS. It should be the equally important concern of the written text in orthographed mode. Well formatted, well calligraphed and typographed written text can be produced under CSS. There has to be a programmatic and powerful link between CSS to ACSS and ACSS to CSS. This will be a real contribution of computer technology and graphics. Some of my experiments in concrete text are presented below in this context.



कोणासाठी मनोमिती किती कोन बदलावे
कोणासाठी किती रिते प्याले पुन्हा ओसंडावे
कोणासाठी किती काळ ओलवावे कोण डोळे
कोणासाठी पानोपानी करावें तें किती काळें
कोणासाठी डोळियांत किती नाचावी बाहुलीं
कोणासाठी शहारावी भर दुपारीं सावली
कोणासाठी कुठे कुठे पाय ओढीत ते न्यावे
कोणासाठी अंतराळ वितीवितीने कापावें
कोणासाठी आळवावें शतसुरी एक गाणें
कोणासाठी किती वेळा कोणी बदलावें नाणें
कोणासाठी किती वेळा कोणी उसवावें ऊर
कोणासाठी मोहरावी कोण-मनीं हुरहूर
कोणासाठी कोणता मी कोणासाठी हें जगणें

कोणा एका उणे जाण्या किती लाख देणें घेणें !

को ? णा ? सा ? ठी



न दे^{खिला} न पा^{हिला}

हां ऽ ऽ ऽ ऽ ऽ

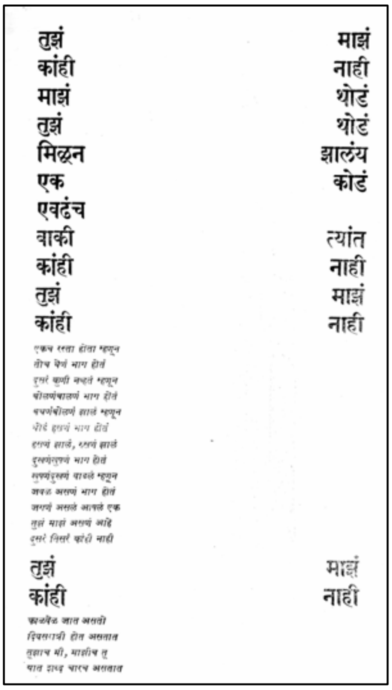
नेरोलॅक

हां ऽ ऽ ऽ ऽ ऽ

अल्ट्राव्हाइट

हां ऽ ऽ ऽ ऽ ऽ

नेरोलॅक अल्ट्राव्हाइट
नेरोलॅक अल्ट्राव्हाइट
नेरोलॅक अल्ट्राव्हाइट
नेरोलॅक अल्ट्राव्हाइट
नेरोलॅक अल्ट्राव्हाइट
नेरोलॅक अल्ट्राव्हाइट
नेरोलॅक अल्ट्राव्हाइट
नेरोलॅकअल्ट्राव्हाइट
नेरोलॅकअल्ट्राव्हाइट
अल्ट्रोलॅकइट
नेरोलॅक
नेरोलॅक
नेरोलॅक



तुझं माझं
कांही नाही
माझं थोडं
तुझं थोडं
मिळून झालंय
एक कोडं
एवढंच
बाकी त्यांत
कांही नाही
तुझं माझं
कांही नाही

एकच रस्ता होता म्हणून
तोच येणं भाग होतं
दुसरं कुणी नव्हतं म्हणून
बोलणंचालणं भाग होतं
बघणंबोलणं झालं म्हणून
थोडं हसणं भाग होतं
हसणं झालं, रुसणं झालं
रुसणंफुगणं भाग होतं
सुखणंदुखणं वाढले म्हणून
जवळ असणं भाग होतं
जगणं असलं आपलं एक
तुझं माझं असणं आहे
दुसरं तिसरं कांही नाही

तुझं माझं
कांही नाही

फक्त वेळ जात असतो
दिवसरात्री होत असतात
तुझाच मी, माझीच तू
यात शब्द चारच असतात

References

- *Satyakatha* magazine published by Mauj Publications, Mumbai.
- W.S. Allen, 1961, *Phonetics in Ancient India*, London, Oxford University Press.
- 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.
- R. K. Joshi, Prague 2003, A unified phonemic code based scheme for effective processing of Indian languages. 23rd Internationalization and Unicode.

Acknowledgements

CDAC Mumbai
Jui Mhatre, Supriya Kharkar, Gracy Abreo,
Ganti Santosh, Sobi Pillai.

15th December 2006