CHARACTERISTICS OF INDIAN LANGUAGES

MADHAVI VARALWAR and NIXON PATEL
Bhrigus Inc. Hyderabad,India
{madhaviv@bhrigus.com, npatel@bhrigus.com}

A text to speech system often requires simple information such as language of the input text; voice-gender (male/female) to be used, pronunciation of a telephone number as isolated digits etc. A raw input text could be embedded with such information using XML like tags often referred to as Speech Synthesis Markup Language (SSML) which aims to produce a better content by a TTS in various contexts. In this positional paper, we discuss some of the possible SSML extensions keeping in the view of Indian language scripts and the corresponding TTS systems.

1. INTRODUCTION :

Bhrigus Inc. is actively involved in developing **TTS** and **ASR** for Indian languages, and is currently developing unit selection voice for Telugu. The goal is to build high quality voices and speech recognition for many of the Indian languages and interface them with computer-telephony applications. Some of these applications include verticals such as entertainment, health care, financial in the context of India. In this paper, we describe the nature of the Indian languages and describe and discuss our proposal where we feel the requirements of some more SSML elements to improve the rendering of Indian languages.

FEATURES OF INDIAN LANGUAGES AND SCRIPTS

Some of the features of Indian languages and the scripts used to express them are :

PHONEME SET :

Indian languages have a more sophisticated notion of a character unit or **akshara** that forms the fundamental linguistic unit. An akshara consists of 0, 1, 2, or 3 consonants and a vowel. Words are made up of one or more aksharas. Each akshara can be pronounced independently as the languages are completely phonetic. Aksharas with more than one consonants are called samyuktaksharas or combo-characters. The last of the consonants is the main one in a samyuktakshara.

All Indian languages have essentially the same alphabet derived from the Sanskrit alphabet. This common alphabet contains 33 consonants and 15 vowels in common practice. Additional 3-4 consonants and 2-3 vowels are used in specific languages or in the classical forms of others. This difference is not very significant in practice. Individual consonants and vowels form the basic letters of the alphabet.

DIFFERENT GRAPHEME'S :
The commonality in the alphabet does not extend the graphic forms used to express them in print. Each language uses different scripts consisting of dissimilar grapheme's for printing. Thus, printed matter in other scripts are inaccessible to readers of one script. There are 10-12 major scripts in India. The Devanagari script is the widest used one, being used to write Hindi (the most spoken language), Marathi, Konkani, and Nepali, the language of the neighboring Nepal. Different scripts use different philosophies for the individual grapheme's and their combinations. Some have a head-line or shirorekha that persists for a whole word. Others have non-touching grapheme's. The grapheme of one of the consonants is usually at the heart of the printed akshara. The vowel appears as a matra or vowel modifier. These can appear to the left,right, above or below it or in combinations. The supporting consonants of a samyuktashara also appear as modifier grapheme's to the left, right, above, or below of the main one. These modifiers could be truncated or scaled down forms of the basic consonant, but could also be completely different. They may touch each other or the main consonant in some cases or may be separate. These rules are not consistent even within a script and certainly not across scripts.

REDUPLICATION :
All languages employ reduplicated form in varying degrees and for different functions, extensive use of reduplication is a particular characteristic of Indian Languages. In this section we present an overview of the different kinds of re duplicative expressions found in the subcontinent . This sets the context for the echo formations, which will be introduced in the next section. The final section will review the literature on echo formations in other Dravidian languages. Data on echo expressions in Indian languages, will also be presented.

1.ONOMATOPOEIC :
Onomatopoeic expressions and expressive share the general property that neither of the two halves of the expression are independently meaningful, whereas the base of an echo expression forms a lexical item in its own right, and this is true of both elements of paired words. The semantics associated with reduplication are relevant too. Onomatopoeic expressions of Indian languages in general display a 'predilection for onomatopoeia'. Onomatopoeic expressions may or may not be reduplicated in structure, and the repetition may be exact, as in (*), or there may be segmental differences between the two halves, or a non-reduplicated case.

(*) kiiccukiiccu-n nu	_	'chirping'	(TAMIL)
taaraa zuvva,	_	'like a star'	(TELUGU)
sura-sura batti	_	'sparklers'	(HINDI)
kaakara-puvvu-vatti	_	'sparklers'	(TELUGU)
kora kora choochu	_	' frowning'	(TELUGU)
kasa kasa namulu	_	' chewing'	(TELUGU)
gora gora gunju	_	(probably only in our telangANA dists.)	

2.EXPRESSIVE :
This next category shares many similarities with the onomatopoeic expressions, but involves terms that are not sound symbolic in the strict sense. onomatopoeics and expressives are largely limited to positive utterances.
Some representative examples of expressive from Hindi,Telugu and from Bengali

Chi~ chi~	_	'dirty, filthy'	(Telugu)
chip chip	_	'sticky'	(Hindi)
thik thik	_	'sense of teeming with maggots'	(Bengali)
pil pil	_	'sense of being overcrowded'	(Bengali)
Rama rama	_	' expressing disgust'	(Telugu)
Kani kani	_	' wait / let us see what happens next '	(Telugu)

3.PAIRED WORDS :
These examples involve the juxtaposition of two lexical items and are variously referred to in the literature as 'synonymic compounds' , 'synonymic repetition' 'semantic reduplication' and 'redundant compounds' Both parts belong to the same semantic field, and may stand in several possible relations to one another.

tikku ticai	_	'point of direction direction' i.e. 'direction'	(Tamil)
kuti makkal	_	'subjects children' i.e. 'citizens'	(Tamil)
taay takappan	_	'mother father' i.e. 'parents'	(Tamil)
pilla jalla	_	' children'	(Telugu)

Parallel examples of each type can be given for Hindi

dhan daulat 'wealth wealth' i.e. 'wealth, riches'
saanii paanii 'cattle-cake water' i.e. 'cattle'
raat din or din raat 'night day' i.e. 'continually'

Paired synonyms frequently involve words taken from different registers, dialects or even languages.

4.ECHO:
The echo expressions, variously known as 'echo words', 'echo formations' form yet another category of reduplicated forms. In terms of their semantic structure, they fit between the onomatopoeias and expressives, in which neither part is independently meaningful, and the paired words and examples of complete and syntactic reduplication, where both parts can be assigned a meaning of their own . The base of an echo formation, which in the vast majority of cases appears first, it is always a lexical item in its own right, but the reduplicated part has no independent lexical meaning. A consistent cluster of meanings tend to characterize echo expressions cross linguistically and setting them apart from other sections of the continuum the echo expressions there is also considerable variation, which has implications for the question of lexical storage. Different languages favor different fixed segments, with ki(i)-/gi(i)- being

characteristic of the Dravidian languages. Echo expressions with idiosyncratic phonology, and possible Semantics, seem to be a common feature of the Indian subcontinent Such formations are highly productive: Examples of an echo expressions are

uppu cappu	_	‘taste’	(Telugu)
illu gillu	_	‘house’	(Telugu)
ekkada akkada	_	'here and there'	(Telugu)
kaappi kiippi	_	‘coffee and other beverages’	(Tamil)
paampu kiimpu	_	‘snakes and other reptiles/pests’	(Tamil)
puli kili	_	‘Tiger and such like animal’	(Tamil)
Aisa Waisa	_	' this way that way'	(Hindi)
Eidhar Udhar	_	' here or there'	(Hindi)

tendency amongst younger speakers of Telugu to produce sentences such as.
Pelli ayindi kaani gilli avaleedu
‘The marriage took place but not gilli (the consummation).’
Another Dravidian example, this time from Malayalam, although here the usage is partly metalinguistic.
ummaan ko uttaal ammaavan alle kil kimmaavan -‘As long as he feeds me, he’s called “Uncle”, but otherwise “Kincle”.’

This is certainly a marginal phenomenon, but represents a potential development from the echo expressions, with the two halves beginning to diverge.
Scornful or sarcastic connotations are widely reported for echo words, and, indeed, form one of the meanings associated with reduplication cross-linguistically. one further alternative is that the echo expression merely intensifies or emphasizes the meaning of the base word, the echo words are said to ‘encode the speaker's affective state,expands at some length upon the kinds of emotions that may be involved, directed either towards the addressee or to what is described in the base word, listing playfulness, hesitation, ridicule and emphatic negation.

Bhrigus Inc had done significant work regarding the characteristics of Indian Languages. In the previous **SSML** conference three new tags were proposed namely

Syllable Element <syllable>
For example if the Telugu word .naatoo. has to be spoken character by character, and the use of phoneme tag would split it as below.

```
<phoneme alphabet="itrans-3" ph="n aa t oo"> naatoo </phoneme>
```

However, for a native Telugu speaker there is no sound called .n. exists. For him/her the sound .n. always exists with a vowel which is a syllable (note characters in Indian languages are syllables). Hence it makes sense to have a syllable tag which would split the word .naatoo. as follows and which is more meaningful for the native speaker.

```
<syllable alphabet="i trans-3" syl="naa too"> naatoo </syallable>
```

2.Loan-Word Element <alien>

3 Dialect Element <dialect>

```
<?xml version="1.0"?>
<speak version="1.0" xml:lang="tel-in">
<voice gender="female">
<dialect name = .andhra.> yekkad.iki vel.laali </dialect>
<dialect name = .telengana. pro = .yaad.iki poovaale.> yekkad.iki
vel.laali </dialect>
</voice>
</speak>
```

PROPOSED EXTENSIONS IN THIS WORKSHOP ARE:

1.**Echo element <echo>**

Echo expressions are very common in Indian Languages as cited above.
Throughout the Indian languages, echo expressions are predominantly restricted to colloquial speech Since the colloquial variety has no standard written form, the occurrence of echo expressions in texts is therefore minimal. They are, however, in widespread use in the spoken language, at all levels of society .

```
<?xml version="1.0"?>
<speak version="1.0" xml:lang="tel-in">

        <emphasis>Welcome</emphasis> to the echo sample tag.
       <s> neku leda  <echo>illu gillu </echo></s>

</speak>
```

2.EXPRESSIVES :

```
<?xml version="1.0"?>
<speak version="1.0" xml:lang="hindi">

    <s> <say as> dhan dualat </say as> pa ne ke liye <say as> din raat </say as> kaam karna hai
</s>

</speak>
```

3.ONOMOTOPEIA:

```
<?xml version="1.0"?>
<speak version="1.0" xml:lang="tel-in">
          <p> pillalu andaru deepawali rojuna   <say-as> </say as>sura-sura bati
          <say-as> taara zuvva <say-as </say-as> kaalcharu</p>

 </speak>
```

CONCLUSIONS:

In this paper, we discussed the position of Bhrigus Inc. Hyd in building the TTS systems for Indian languages. We described the nature of Indian Langauages Keeping in view of the issues with the characteristics of these languages, we discussed the importance of **Paired words** and **Echo elements** and their behavior to improve the quality of TTS system in the context of Indian languages and other similar languages in the Asia-Pacific.