Pronunciation of Nouns in Text to Speech systems

Veera Raghavendra, Lavanya Prahallad
IIIT Hyderabad, India

Agenda

- Nature of Indian Language Scripts
- Convergence and Divergence
- Fonts and Transliteration Scheme
- SSML Extensions for Proper Nouns

Nature of Indian Language Scripts

- Indian language (IL) scripts originated from the ancient Brahmi script.
- Basic units of the writing system are Aksharas
- An Akshara is an orthographic representation of a speech sound
- Akshara is syllabic in nature
- A syllable is defined as C*VC*
- C is a consonant
- V is a vowel
- Examples: V, CV, CCV, CVC, CCCV
- amma:
  - Phone sequence: /a/ /m/ /m/ /aa/
  - Syllables: (/a/) (/m/ /m/ /aa/)
- Written from left-to-right
- Words are separated by space as in European languages
- Roman digits (0...9) are used as numerals.

Convergence and Divergence

- India is a multi-lingual nation with 21 recognized official languages and ~1652 dialects.
- These languages are: Assamese, Tamil, Malayalam, Gujarati, Telugu, Oriya, Urdu, Bengali, Sanskrit, Kashmiri, Sindhi, Punjabi, Konkani, Marathi, Manipuri, Kannadam, Bodo, Dogri, Maithili, Santhali and Nepali.
- Apart from Hindi and English
- While all of these languages share a common phonetic base, some of the languages such as Hindi, Marathi and Nepali also share a common script known as Devanagari.
- Languages such as Telugu,Kannada and Tamil have their own scripts.

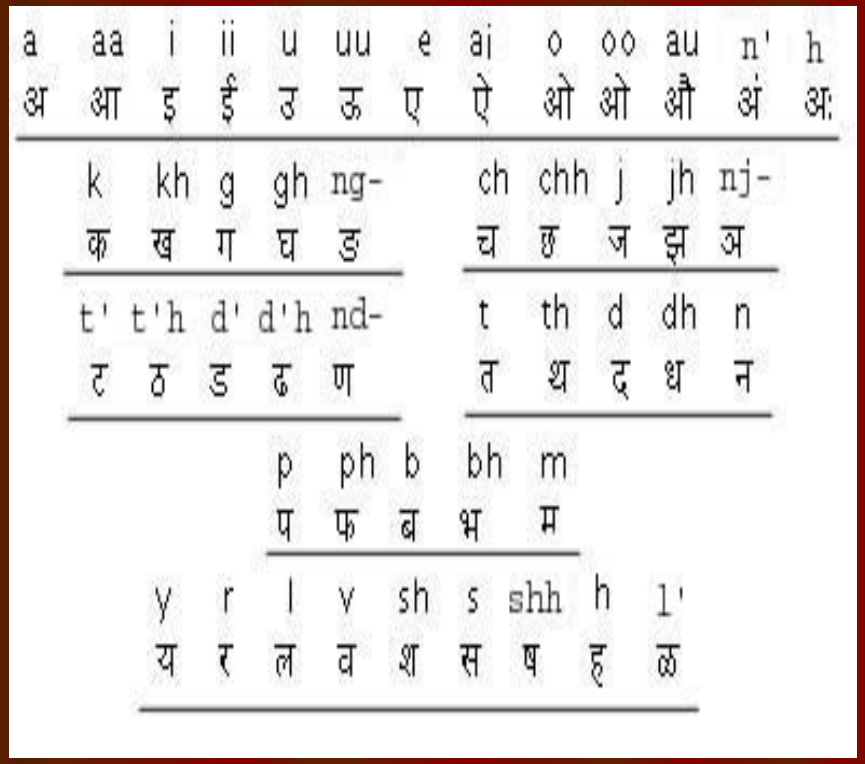
Fonts and Transliteration scheme

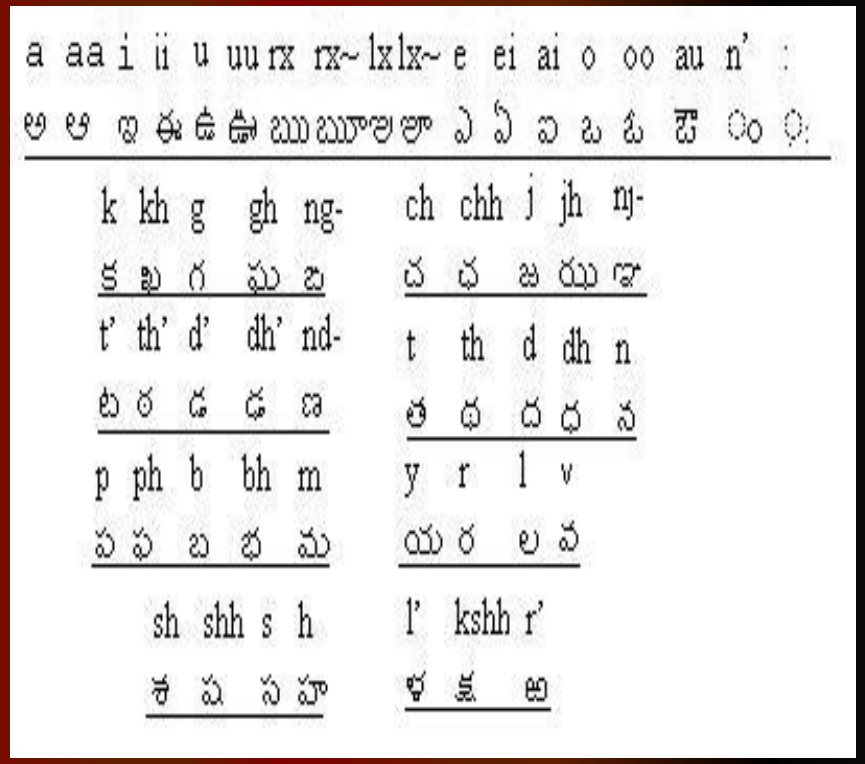
- True Type Fonts
  - Uses 1- 256 ASCII characters to represent characters
  - Character representation is different from one font to other [even in the same language]
  - Separate converter required for each font
  - Proprietary fonts
- Unicode
  - A universal character set
  - provides a unique number for each character in a language
  - Supports all platforms
  - Supports all the languages

- Transliteration (OM / IT3)
  - Developed by IISc Bangalore and Carnegie Mellon
  - Developed from the user readability aspects – Easier to read and type
  - It is case- insensitive.
  - Thus a single transliteration scheme is used for all the Indian languages, as they share the same set of sounds.
  - Each character (corresponding to a phone/ sound) is not more than three letters length.

Reference:
http://speech.iiit.ac.in/Transliteration/

http://www.cs.cmu.edu/~madhavi/Omtelugu





Particles

- Hindi and some other Indian languages have a practice of adding a particle 'ji' or 'saaheba‘ etc., after proper nouns.
- They are added when the speaker wants to give respect to the person he is referring to in his speech.

Examples:

- Huma maasat’arajii sei milnei gayei
  (We went to meet the teacher)
- Aaja pitaajii ghara para rahein’gei
  (Father will be at house today)

Example of Particle

- <?xml version="1.0"?>
  <speak version="1.0" xml:lang="hin-in" xml:type="IT3">
  <voice gender="female">
  Huma
  <particle type="ji">maastaar</particle>
  sei milnei gayei
  </voice>
  </speak>

Use of Loanword

- A **loanword** (or *loan word*) is a word directly taken into one language from another with little or no translation.
- Informal experiments suggested 33%of errors of TTS of IL occur while rendering loan words
- Such loan words could be automatically detected due to syllabic properties of the Indian languages

Example of loanword

- CANCER has to be pronounced as /K/ /AE/ /N/ /S/ /A/ /R/
- /AE/ phoneme does not exist in Indian language phone set
- <loan> kaansar </loan>
- loan (non-native) words could be rendered using different pronunciation dictionaries or letter-to-sound rules

Use of Mention

- What is mention
  - I mention – refers to first occurrence of a noun
  - II mention – refers to second occurrence of a noun
- More emphasize on the first occurrence of the proper noun in a sentence or paragraph
- Tag, <mention>, should be used to identify similar words in synthesizing the speech

Duration prediction using Mention Information

- Duration modeling using mention information of US English

	RMSE	Correlation
Without MENTION	0.876	0.4580
With MENTION	0.869	0.497

Example of Mention

- <?xml version="1.0"?>
  <speak version="1.0“>
  <voice gender="female">
  <mention occ=1>Gandhi</mention> was a major political and spiritual leader of the Indian Independence Movement. <mention occ=2>Gandhi</mention> was the pioneer of satyagraha
  </voice>
  </speak>

Conclusion

- Issues in Indian scripts are discussed
- Discussed the usage of <particle>, <loan> and <mention> extensions for SSML

Thanks...