RESEARCH ARTICLE OPEN ACCESS

Kannada Phonemes to Speech Dictionary: Statistical Approach

Mallamma V. Reddy *, Hanumanthappa M **
* Department of Computer Science, Rani Channamma University, Vidyasangam, Belagavi-591156, India
** Department of computer science, Bangalore University, Jnanabharathi Campus, Bangalore-560056, India

ABSTRACT
The input or output of a natural Language processing system can be either written text or speech. To process written text we need to analyze: lexical, syntactic, semantic knowledge about the language, discourse information, real world knowledge to process spoken language, we need to analyze everything required to process written text, along with the challenges of speech recognition and speech synthesis. This paper describes how articulatory phonetics of Kannada is used to generate the phoneme to speech dictionary for Kannada; the statistical computational approach is used to map the elements which are taken from input query or documents. The articulatory phonetics is the place of articulation of a consonant. It is the point of contact where an obstruction occurs in the vocal tract between an articulatory gesture, an active articulator, typically some part of the tongue, and a passive location, typically some part of the roof of the mouth. Along with the manner of articulation and the phonation, this gives the consonant its distinctive sound. The results are presented for the same.
Keywords: Natural Language Processing, Phonetics, Unicode

I. INTRODUCTION

Linguistics is the scientific study of language. It speaks three aspects of the language they are language form, language meaning, and language in context. Linguistics analyzes human language as a system for relating sounds and its meaning. Phonetics studies acoustic and articulatory properties of the production and perception of speech [1] sounds. The common users interface for phonetic to speech as shown in Fig.1.

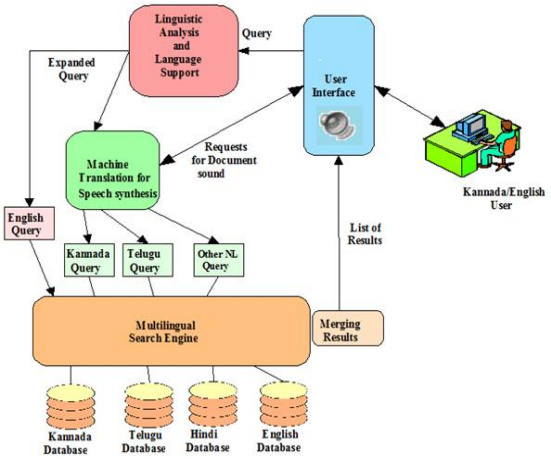

Fig.1: User interface for phonetics to speech

Kannada or Canarese is a south Indian language widely spoken in the state of Karnataka in India. Kannada [2] is originated from the Dravidian Language others are Telugu, Tamil, and Malayalam. Whose native speakers are called Kannadigas, number roughly 38 million, making it the 27th most spoken language in the world. Kannada as a language has undergone modifications since BCs. Based on the modifications it can be classified into four types: Purva Halegannada (from the beginning till 10th Century), Halegannada (from 10th Century to 12th Century), Nadugannada (from 12th Century to 15th Century), Hosagannada (from 15th Century). Hosagannada or Kannada language uses forty nine phonemic letters, classified into three groups they are:

1 *Swaragalu / vowels:* There are thirteen vowels, are the independently existing letters are ಅ ಆ ಇ ಈ ಉ ಊ ಋ ಎ ಏ ಐ ಒ ಓ ಔ, there are two types of Swaras (Vowels) depending on the time used to pronounce. They are Hrasva Swara: A freely existing independent vowel which can be pronounced in a single matra time (matra kala) also called as a matra are ಅ ಇ ಉ ಋ ಎ ಐ ಒ ಔ and Deerga Swara: A freely existing independent vowel which can be pronounced in two matras are ಆ ಈ ಊ ಏ ಓ.
2 *Vyanjanagalu/ Consonants:* there are thirty four consonants, are dependent on vowels to take a independent form of the Consonant. These can be divided into two types Vargeeya are ಕ್ ಖ್ ಗ್ ಘ್ ಙ್, ಚ್ ಛ್ ಜ್ ಝ್ ಞ್, ತ್ ಥ್ ದ್ ಧ್ ನ್, ಟ್ ಠ್ ಡ್ ಢ್ ಣ್, ಪ್ ಫ್ ಬ್ ಭ್ ಮ್ and Avargeeya are ಯ್ ರ್ ಲ್ ವ್ ಶ್ ಷ್ ಸ್ ಹ್ ಳ್, and

3 *Yogavaahakagalu*: there are two yogavaahakagalu are Anuswaras ಅಂ and Visarga ಅಃ .

Basic Language Rule in Kannada is when a dependent consonant combines with an independent vowel; an Akshara is formed as shown below:
Vyanjana + Vowel (matra) ---> Letter (Akshara)
Example: ಕ್ + ಅ ---> ಕ

Based on this rule we can combine all the Consonants (Vyanjanas) with the existing Vowels (matra) to form Kagunitha for Kannada alphabet.

II. TERMINOLOGIES RELATED TO SOUND

Sounds of all languages fall under two categories: Consonants and Vowels. Consonants are produced with some form of restriction or closing in the vocal tract that hinders the air flow from the lungs. Consonants are classified according to where in the vocal tract the airflow has been restricted. This is also known as the place of articulation. Vowel is a sound in spoken language, pronounced [3] with an open vocal tract so that there is no build-up of air pressure at any point above the glottis. The following sound system terminologies are useful in natural language processing they are:

- *Phoneme* is a unit of sound in a language that cannot be analyzed into smaller linear units and it is the smallest contrastive part of a word which may cause a change of meaning.
- *Phonological awareness* is the ability of a listener to recognize the sound structure of words.
- *Phonemic awareness* is the ability of a listener to hear, identify and manipulate phonemes.
- *Phonemic transcription* is a type of phonetic transcription that uses fewer phonetic symbols – only one for each phoneme.
- *Phonetic spelling* is a way to confirm the spelling of a word by pronouncing each letter as a word
- *Phonetic transcription* is the visual representation of speech sounds
- *Phonetics* is a science that studies the sounds of human speech sounds especially with regard to the physical aspects of their production. In the case of oral languages there are three basic areas of study are Acoustic phonetics is the study of the physical transmission of speech sounds from the speaker to the listener, articulatory phonetics is the study of the production of speech sounds by the articulatory and vocal tract by the speaker, and Auditory phonetic is the study of the reception and perception of speech sounds by the listener.
- *Phonology* is a branch of phonetics that studies systems of phonemes and pronunciation especially as they occur in a particular language.
- *Stress* is the relative emphasis that may be given to certain sounds or syllables in a word, or to certain words in a phrase or sentence.
- *Phonics* is the branch of linguistics concerned with spoken sounds; phonetics [4] the correlations between sound and symbol in an alphabetic writing system; the phonic method of teaching reading.
- *phonotactics* the branch of linguistics concerned with the rules governing the possible phoneme, sequences in a language or languages; these rules as they occur in a particular language.

III. METHODOLOGY

This section describes the new algorithm which is developed for morphological [5] generator for generating sound system for given input query. The main advantage for this algorithm is simple and accurate. Input/output Examples for Morphological Analyzer for inflections of a noun stem GARDEN and its corresponding meanings as shown in Table.1

Algorithm
Step 1: Get the word to be analyzed.
Step 2: Check whether the entered word is found in the Root Dictionary.
Step 3: If the word is found in the dictionary, present the sound of the word stop;
Else
Step 4: Separate any suffix from the right hand side
Step 5: If any suffix is present in the word, then check the availability of the suffix in the dictionary.
Then
Step 6: Remove the suffix present,
Then re-initialize the word without identified suffix, Go to Step 2.
Step 7: Repeat this process until the Dictionary finds the root/stem word.
Step 8: Store the root/stem word in a variable and then get the corresponding Kannada phonetic to speech/sound from the bilingual dictionary
Step 9: Check what all grammatical features does the word have given and then generate the corresponding features for the Kannada word
Step 10: Exit.

Kannada phonetic to Speech dictionary, we have proposed two way of producing the sound of the given input query. First entering the English word and transliterate it, for transliteration there many transliteration tools are available to type Kannada characters using a standard keyboard.

These include Baraha (based on ITRANS) and Quillpad (predictive transliterator). Nudi, the government of Karnataka's standard for Kannada input, is a phonetic layout loosely based on transliteration. Unicode for Kannada keyboard as shown in Fig. 2 and second directly enter the Kannada input query with the help of Virtual keyboard, we have developed for Kannada as shown in Fig. 3

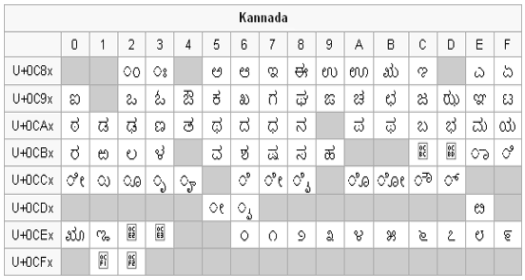

Fig.2: Unicode of Kannada

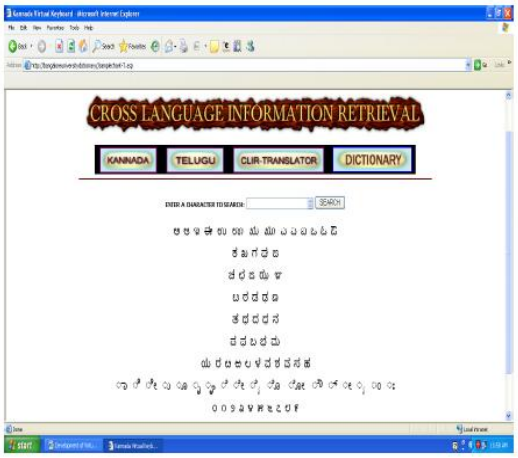

Fig.3: Kannada Virtual Keyboard

Table.1: Noun stem and its meanings

Inflected Nouns	Meaning in English
ಉದ್ಯಾನ – ವು	Garden
ಉದ್ಯಾನ – ವನ್ನು	The garden
ಉದ್ಯಾನ – ದಿಂದ	From the garden
ಉದ್ಯಾನ – ಕ್ಕೆ	To the garden
ಉದ್ಯಾನ – ದಸೆಯಿಂದ	Because of garden
ಉದ್ಯಾನ – ದ	Of the garden
ಉದ್ಯಾನ – ದಲ್ಲಿ	In the garden
ಉದ್ಯಾನ – ಗಳು	Gardens
ಉದ್ಯಾನ – ಗಳನ್ನು	The gardens
ಉದ್ಯಾನ – ಗಳಿಂದ	From the gardens
ಉದ್ಯಾನ – ಗಳಿಗೆ	To the gardens
ಉದ್ಯಾನ – ಗಳದಸೆಯಿಂದ	Because of gardens
ಉದ್ಯಾನ – ಗಳ	Of the gardens
ಉದ್ಯಾನ – ಗಳಲ್ಲಿ	In the gardens

IV. EXPERIMENTAL SETUP

Character encoding is a way of storing characters in a computer as bits. Character set detection works reliably in detecting UTF-8 for Kannada [6][7]. Due to the unreliability of heuristic detection, it is better to properly label datasets with the correct encoding. For example, HTML documents can declare their encoding in a Meta element, thus:

<Meta http-equiv="Content-Type" content="text/html; char-set =UTF-8" />.

Alternatively, when documents are conveyed through HTTP, the same metadata can be conveyed out-of-band using the Content-type header. Finally, if a Unicode encoding is used, text files can be explicitly labeled with an initial byte order mark.

Usually, the first step in building the Pronunciation Dictionary is to create a sorted list of the words contained in your Grammar, one per line, with pronunciations using statistical computational approach as:

- Character Unigram: A unique single letter (‘a’, ’b’, …’z’ for English, ಅ ,ಆ for kannada).
- Character bigram: A unique two-letter long sequence (“aa”, “ru” ಆ ,ಮು respectively).
- Character trigram: A unique three-letter long sequence (“kha”, “pha”, ಖು ,ಫ respectively).
- Character n-gram: A unique n-character long sequence of letters.
- N-gram frequency: How frequently an n-gram appears in (some sample) text.

We created machine-readable Kannada-Kannada phonetic to speech dictionary for query translation. The following steps are followed for dictionary generation the first step is speech recognition Engine: Language Model or Grammar, Acoustic sound second step is to build Phonetically Balanced Dictionary third step is recording the data: while recording the data we have to make sure that, the environment should be quite, adjust the Microphone, adjust the recording levels and configuring audacity preferences as shown in Fig.4 the fourth step is words level mapping using statistical approach and the final step is display its phonetic sound.

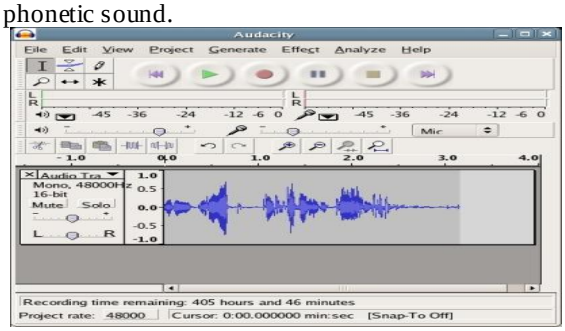

Fig.4: audacity of recording sound

V. RESULTS

Natural Language Processing Tool for phonetic to speech dictionary is built by using the ASP.NET as front end, and, for a database, the

Kannada text files are encrypted by using the UTF-8 Encoding system [8]. The statistical computational approach is used to map the elements which are taken from documents or input query to generate the sound. The sample results for Vowels, Consonants, bisectional consonant Ka ಠ and numbers are as shown in below Fig 5, Fig 6, Fig 7 and Fig.8 respectively.

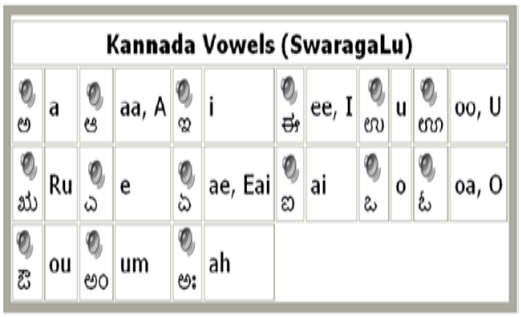

Fig. 5: Phonemes to speech for Kannada Vowels

Fig. 6: speech for Kannada numbers

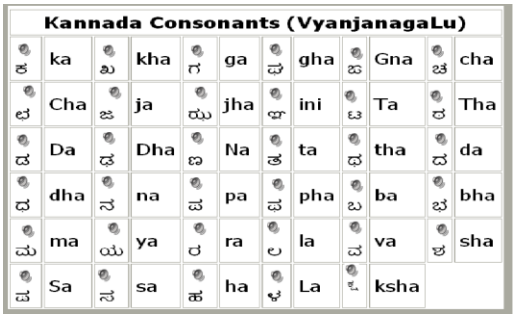

Fig. 7: Phonetics to speech for Kannada consonants

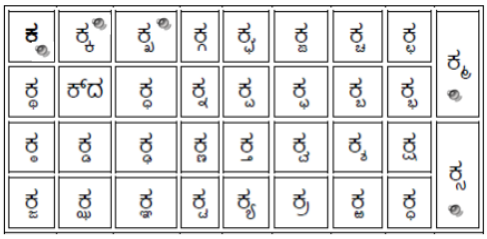

Fig. 8: Phonetics to speech for Kannada bisectional consonant ka ಠ

VI. CONCLUSION

This paper presents the Kannada phoneme to speech dictionary. The dictionary entries are based on phoneme stems of specified phonetics. Kannada phoneme and Kannada sound is the source and the target, respectively, for input query. The dictionary is useful to learn natural languages. A statistical computational approach is used to the generate Kannada speech dictionary. The proposed method can be easily extended to other language pairs that have different sound systems.

REFERENCES

[1]. Jakobson, Roman, Gunnar Fant, and Morris Halle. "Preliminaries to Speech Analysis: The Distinctive Features and their Correlates", MIT Press. 1976
[2]. The Karnataka Official Language Act, "Official website of department of Parliamentary Affairs and Legislation", Government of Karnataka. Retrieved 2007-06-29
[3]. Kingston, John. "The Phonetics-Phonology Interface", in the Cambridge Handbook of Phonology (ed. Paul DeLacy), Cambridge University Press. 2007
[4]. http://www.dailywritingtips.com/the-difference-between-phonics-and-phonetics/
[5]. Dr. Ramakanth Kumar P,et.al "Kannada Morphological Analyser and Generator Using Trie" published in IJCSNS International Journal of Computer Science and Network Security,VOL.11 No.1, January 2011.
[6]. http://en.wikipedia.org/wiki/Kannada
[7]. http://www.omniglot.com/writing/kannada.htm
[8]. http://www.ssec.wisc.edu/~tomw/java/unicode.html#x0C80