ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

Schwa Deletion: Investigating Improved Approach for Text-to-IPA System for Shiri Guru Granth Sahib

Sandeep Kaur[1], Dr. Amitoj Singh[2]

Pursuing M.E, CSE, Chitkara University, India [1]

Associate Director, CSE, Chitkara University , India [2]

Abstract: Punjabi (Omniglot) is an interesting language for more than one reasons. This is the only living Indo-Europen language which is a fully tonal language. Punjabi language is an abugida writing system, with each consonant having an inherent vowel, SCHWA sound. This sound is modifiable using vowel symbols attached to consonant bearing the vowel. Shri Guru Granth Sahib is a voluminous text of 1430 pages with 511,874 words, 1,720,345 characters, and 28,534 lines and contains hymns of 36 composers written in twenty-two languages in Gurmukhi script (Lal). In addition to text being in form of hymns and coming from so many composers belonging to different languages, what makes the language of Shri Guru Granth Sahib even more different from contemporary Punjabi. The task of developing an accurate Letter-to-Sound system is made difficult due to two further reasons:
1. Punjabi being the only tonal language
2. Historical and Cultural circumstance/period of writings in terms of historical and religious nature of text and use of words from multiple languages and non-native phonemes.
The handling of schwa deletion is of great concern for development of accurate/ near perfect system, the presented work intend to report the state-of-the-art in terms of schwa deletion for Indian languages, in general and for Gurmukhi Punjabi, in particular.

Keywords: Gurmukhi Punjabi, IPA, Schwa deletion, Sikhism

I. INTRODUCTION

Language of Shri Guru Granth Sahib(SGGS) is different from modern Punjabi due to certain reasons which makes it difficult for anyone to read.. Punjabi is an indo-Aryan language spoken by 130 million native speakers worldwide, making it 10th most widely spoken language in the world.[12][13].It has been classified as belonging to the CENTRAL group under INNER sub-branch of indo Aryan languages under NIA sub Classification Scheme. For Sikhs, the Punjabi (GURMUKHI) is the official language for all ceremonies and rituals.
Introduction to Gurmukhi Punjabi and SGGS is given in the following section:

1.1 GURMUKHI SCRIPT

Gurmukhi script is an abugida writing system, where each consonant has an inherent vowel (/ə/) modifiable using vowel symbols which can be attached to the relevant vowel-bearing consonant. Gurmukhi Punjabi has 35 native characters and another 6 characters to accommodate sounds from foreign languages. These characters represent 3 vowel characters, 2 fricatives, 25 consonants, 5 semi-vowels. In addition there are nine vowel symbols, two symbols for nasal sounds and one symbol which duplicates the sound of any consonant. The Punjabi phoneme inventory has twenty-five consonant phonemes, ten vowel phonemes, three tones (High, Mid, Low), and seven diphthongs. A number of non-native speech sounds are also found, however, these sounds occur in loan words only (mostly Persian and Arabic loans). Punjabi is the only tone language in the Indo-European language family, making it of considerable interest to both phonologists and historical linguists[UCLA,2010].
Gurmukhi script has quite a different structure and system as compared to all the other Indian Scripts. This has been attributed to two main facts:
1. The tonal system and some other phonetic features,
2. Different cultural and historical circumstances.[1]

1.2 SHRI GURU GRANTH SAHIB

Sri Guru Granth Sahib Ji is a voluminous text of 1430 pages(called angs),compiled and composed during the period of Sikh Gurus, from 1469 to 1708. It is written in the Gurmukhi Script. It has a total of 398,697 words with a total of 29,445 unique dictionary words. Many of these words have been used only once[1]. Guru Granth Sahib contains 5894 hyms(shabad) written in 60 melodies(raag) by 35 authors, including 6 sikh gurus, 15 bhagats, 3 Divines and 11 Poets, all from different social classes, religions and spiritual traditions.Although written in the Gurumukhi script, the text are in different languages including Braj, Punjabi, Khariboli (Hindi), Sanskrit, Arabic, Sindhi, Lehndi, Dakhni, Bengali, Marathi and Persian, given the generic title of Sant (Saint) Bhasha (language)(meaning the language of the saints).[1]

Copyright to IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

1.3 DIFFERENCE BETWEEN THE LANGUAGE OF GURU GRANTH SAHIB JI AND MODERN PUNJABI

The uniqueness of the language is due to the fact that the original scripture was written as a continuous text without any spaces between the words.

The differences between the two, Modern Gurmukhi and Gurmukhi in Gurbaani, are as follows:

1. The Gurmukhi script used in Shiri Guru Granth Sahib Ji uses the only first 35 characters of the 41 characters used in modern script.

2. There is no use of Adhdhak (/ੱw/) in Shiri Guru Grant Sahib Ji. In modern writing this symbol is used frequently for the purpose of doubling the sound.

3. No Punctuation marks were used; instead vowel symbols fulfilled the purpose.

4. The use of vowel sounds (i) and (U) at the end of the word, with some exceptions, does not form the syllable. These were used for grammatical reasons.

5. Use of characters (c, t, q, n, X, v) in the foot of other letter and symbols and half characters is not seen in the modern writing.

6. Bindi (N) : It has been written on the left hand side of vowel signs, which is not permissible in modern language.[1]

II. INTERNATIONAL PHONETICS ASSOCIATION (IPA) REPRESENTATION

IPA letters are used in the system to represent the sounds[1].The International Phonetic Alphabet (IPA) is an alphabetic system of phonetic notation based primarily on the Latin alphabet[6]. It was devised by the International Phonetic Association as a standardized representation of the sounds of spoken language. The IPA is designed to represent only those qualities of speech that are distinctive in spoken language: phonemes, intonation, and the separation of words and syllables[7].

As of 2008, there are 107 letters, 52 diacritics, and four prosodic marks in the IPA. The general principle of the IPA is to provide one letter for each distinctive sound (speech segment). This means that it does not use combinations of letters to represent single sounds The principal vowels for Punjabi are shown in Figure-1. The principal vowels are symmetrically distributed on a standard vowel chart: three front vowels, two central vowels, and three back vowels[9]. The three back vowels are rounded.

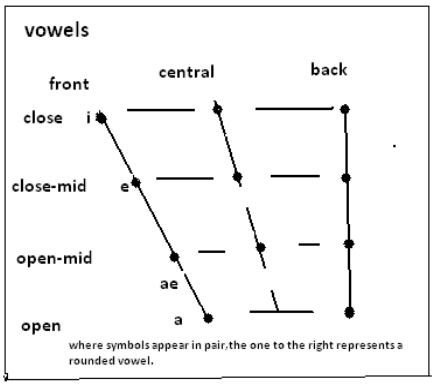

Figure1: Principal vowels for Punjabi

III.SCHWA

Schwa is a mid-central vowel that occurs in unstressed syllables. Phonetically, it is a very short neutral vowel sound, and like all vowels, its precise quality varies depending on its adjacent consonants. Each consonant in Punjabi (written in Gurmukhi script) is associated with one of the vowels,however schwa vowel is not explicitly represented in orthography. The orthographical representation of any language does not provide any implicit information about its pronunciation and is mostly ambiguous and indeterminate with respect to its exact pronunciation. The problem in many of the languages is mainly due to the existence of schwa vowel that is sometimes pronounced and sometimes not, depending upon certain morphological factors. In order to determine the proper pronunciation of words, it is necessary to identify which schwas are to be deleted and which are to be retained. *Schwa deletion* is a phonological phenomenon where schwa is absent in the pronunciation of a particular word, although ideally it should have been pronounced [4]. The process of schwa deletion is one of the complex and important issue for grapheme-to-phoneme conversion, which in turn is required for the development of a high quality text-to- speech (TTS) synthesizer. In order to produce natural and intelligible speech, the orthographic representation of input has to be augmented with additional morphological and phonological information in order to correctly specify the contexts in which schwa vowel is to be deleted or retained[5].

IV.LITERATURE REVIEW

In the simplest rule based system developed to start with Gurmukhi Panjabi being a phonologically transparent writing system, corresponding arrays of Gurmukhi alphabets and vowels symbols were generated and the only rule was to replace the Gurmukhi symbol by the Corresponding IPA symbol. This system, as expected, had several shortcomings as a LTS system for the Gurbaani. When results were compared to the words in the corpus, several repeated errors or shortcomings were noticed[1].Singh proposed various rules used in

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

development of letter-to-sound system. His work so far has been focused on the rules which can produce a system capable of handling the text in the corpus.

Parminder Singh and Gurpreet Singh Lehal[2] proposed a rule based schwa deletion algorithm for Punjabi TTS system. The decision for retention or deletion of schwa is very much obvious for native speaker, but for machine processing purpose this decision will be based on language specific rules.They have proposed a set of eleven rules which are based on grammar rules, inflectional rules and morphotactics for Punjabi. Sheilly Padda, Rupinderdeep Kaur and Nidhi have proposed some steps for converting Punjabi text to IPA.The conversion system is based on syllables. The text entered by the user is analyzed and the text is normalized. The normalized text is passed to the syllabification module in which the text is segmented into syllables [10]. The syllable is a combination of vowel and consonant pairs such as V, VC, CV, VCC, CVC, CCVC and CVCC.In Punjabi all combinations are represented except the last one.Next, the syllables generated are passed to tokenizer which breaks them into tokens. These tokens generated are mapped with the help of token mapper module [11]. The token mapper module maps the token with its corresponding IPA representation[3].Parminder Singh and Gurpreet Singh Lehal have also proposed a Text-to-speech synthesis system for Punjabi language. During the development of this TTS system, it was observed that for a concatenative speech synthesis system, the important features that must be taken care of are: selection of basic speech unit of concatenation, statistical analysis of selected speech units on corpus, corpus must be carefully selected and unbiased, labeling of the speech units. The last one is most important and the quality of the output speech depends, how carefully the speech units are labeled in the recorded sound file.[10]

V. PROPOSED WORK

The work so far has been done on the conversion of text-to-speech of Gurmukhi Punjabi. In this research the text is converted into IPA and then the IPA is converted to sound. Here in this research a modified method for converting text to IPA for Gurbani is proposed in which the schwa identified for deletion, is deleted first and then the text is converted into IPA. Various algorithms have been proposed for the deletion or retention of schwa in punjabi but this work has not been done for Gurmukhi Punjabi. After converting the text to IPA the accuracy of proposed approach is compared with the traditional approach for converting text-to-speech.

ACKNOWLEDGMENT

I would like to express my gratitude to **Dr. Amitoj Singh** and **Mr Gurpreet Singh** for their support, help and guidance during my research work. They have been a constant guiding force and source of illumination. Finally thank the almighty god with whose grace I'm always motivated and deeply engrossed with my thesis work during the entire duration from its conception to success.

REFERENCES

[1]Gurpreet Singh, "Letter-to-Sound rules for Gurmukhi Panjabi(Pa):First step towards Text-to-speech for Gurmukhi", Language Resources and Evaluation Conference: LRE-Rel: Language Resources and Evaluation for Religious Texts, May 2012, Istanbul, Turkey.

[2]Parminder Singh, Gurpreet Singh Lehal,"A rule based schwa deletion algorithm for Punjabi TTS system", Volume 139, 2011, pp 98-103.

[3]Sheilly Padda, Rupinderdeep Kaur, Nidhi,"Punjabi Phonetic:Punjabi Text to IPA Conversion" International Journal of Emerging Technology and Advanced Engineering,Volume 2,Issue 10,October 2012.

[4]Choudhury, M. Basu, A., Sarkar, S.: A Diachronic Approach for Schwa Deletion in Indo Aryan Languages. In: Workshop of the ACL Special Interest Group on Computational Phonology (SIGPHON), Association for Computations Linguistics, pp. 20--27, Barcelona ,2004.

[5]Narasimhan, B., Sproat, R., Kiraz, G.: Schwa-Deletion in Hindi Text-to-Speech Synthesis. Intl. J. Speech Tech., Volume 7, no. 4, 319--333,2004.

[6]International Phonetic Association (IPA), Handbook viewed on Oct 7,2014.

[7]Judy Thompson, "English Phonetic Alphabet (EPA), 2001.

[8]UCLA: Language MaterialsProject, 21 - Dec 2010. http://www.lmp.ucla.edu/Profile.aspx?LangID=95&menu=004.

[9]Anthony Atkielski, "Using Phonetic Transcription in Class",2005

[10]Parminder Singh and Gurpreet Singh Lehal, "Text-To-Speech Synthesis System for Punjabi Language".

[11]M. H. Mateus, E. d'Andrade, "The Phonology of Portuguese: The phonology of the world's languages", Oxford University Press Inc.,New York, 2000.

[12]Nationalencyklopedin "Världens 100 största språk 2010",The World's 100 Largest Languages in 2010

[13]"What Are The Top 10 Most Spoken Languages In The World?". Retrieved 2012-2013.

[14]Parminder Singh and Gurpreet Singh Lehal, "Text-To-Speech Synthesis System for Punjabi Language".

[15]Dr. Prem Singh, "Sidhantik Bhasha Vigeyan", Madan Publications, Patiala.

[16]Er. Sheilly Padda, Ms. Rupinderdeep Kaur, Er. Nidhi, " Architecture and Implementation of Punjabi Text to Speech System Using Transcriptions Concept".

[17]Parminder Singh and Gurpreet Singh Lehal, "Punjabi Text-To-Speech Synthesis System".