The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Empirical Study of Speech Synthesis Markup Language and Its Implementation for Punjabi Language

Atul Kumar, Shyam S Agrawal

KIIT College of Engineering Gurugram India
atulkumar1508@gmail.com,ss_agrawal@hotmail.com

Abstract

This paper builds a prioritized list of requirements for speech synthesis markup which any proposed markup language should address. This study presents requirements and essential tags for specification development of Punjabi Language. A speech synthesizer works like written text into correct sounds to be spoken. To do this it uses an SSML document and one or more lexicons and dictionaries. We have presented how the different type of modules in TTS System helps to convert a text input of SSML document to spoken form in Punjabi Language. Since, Punjabi is the morphological rich Language, it is written in "Gurumukhi" Script and this is the official Language of Govt. of India. So, hence accordingly in this language Homograph problem will not occur. Tones in Punjabi pose big problems. The words written in similar ways, have different tones and there by changes their meanings for which the tags have been designed separately. In Punjabi orthographically the written symbols exactly corresponds to the specific words. Therefore in Punjabi, we do not any word which may be called Homograph.

Index Terms: Speech, SSML, Synthesizer, Punjabi

1. Introduction

The Speech Synthesis Markup Language (SSML) is a language that allows an author to change how text is spoken- for example, by marking some text as sentences and some as paragraphs, by telling the synthesizer when to change voices like low, high and medium, or even by telling the synthesizer exactly how to pronounce or spell a certain word [2]. This paper proposes some enhance SSML tags with respect to Indian Languages with Punjabi as Typical case study. In this paper we also introduce the concept of Homophones and Homograph problems will occur in Punjabi Language or not. If occur then how will handle it and this paper introduces the concept of prosodic nature in Speech Synthesis Markup Language for Punjabi Language. This paper builds a prioritized list of requirements and essential tags for speech synthesis markup which any proposed markup language should address.

This paper addresses requirements and essential tags for specification development. A speech synthesizer works like written text into sounds to be spoken. To do this it uses an SSML document and one or more dictionaries lexicons [4]. The Speech Synthesis Markup Language (SSML) is a language that allows an author to change how text is spoken- for example, by marking some text as sentences and some as paragraphs, by telling the synthesizer when to change voices, or even by telling the synthesizer exactly how to pronounce a certain word[7]. The lexicon documents used by a speech synthesizer, just like for speech recognizer, describe how words are to be pronounced First, it provided a single, standard XML- based language for describing pronunciations, both speech recognizers and speech synthesizers. Second, it requires support for IPA, the International Phonetic Alphabet. This Alphabet is a standard symbol set for representing pronunciations of all the language of the world. As we realize that the Speech is the very essential way for human beings to gather, deliver and share information and communicate with other people. It is reasonable that speech related techniques will soon become essential for human- machine interactions. So, speech synthesizer is a process in which written text converts into sounds to be spoken. Many Scientists have been working on the field of speech synthesis; it is becoming more important every day its value overcomes in remedying many development and educational milestones. The rest of paper is as follows, section 2 describes POS is usage for multiple pronunciations section 3 describes the SSML in Punjabi language. Section 4 describes the prosody analysis of Punjabi language. In this section we describe how we have chosen to design our SSML interpreter develops of other synthesis systems are free to interpret SSML key features as they understand and need not made their systems in the similar way. We have presented in this section simply to give rules on how SSML conversion may be carried out. In the last section some never interpreted features of SSML.

.

2. POS is usage for solving multiple pronunciations

In pronunciation specification different types of pronunciation of same words shown with different phoneme element. And “prefer “attribute can be used to give one pronunciation high priority among many pronunciation candidates but this does not solves the above problem for TTS. The attribute “prefer” can be used for defining the dialectal variation of the same orthographic information.

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0"
xmlns="http://www.w3.org/2005/01/pronunciation-
lexicon"
alphabet="ipa" xml:lang="Hi-in">
<lexeme>
<grapheme> ਇੱਕ ਸੌ ਤਿੰਨ </grapheme>
<phoneme prefer="true"> ɪ◌̆k səʊ tɪnn </phoneme>
<!-- IPA string is: " ɪ◌̆k səʊ tɪnn " -->
<phoneme> ɪ◌̆k, sɪphər, tɪnn </phoneme>
<!-- IPA string is: " ɪ◌̆k, sɪphər, tɪnn " -->
</lexeme>
</lexicon>
```

10.21437/SLTU.2018-22

3. SSML in Punjabi Language

For the speech synthesis Markup Language in Hindi language we follow the same steps which are followed by W3c. The process of transforming text into speech contains coarsely two phases: first the text goes through analysis and then the resulting information is used to generate the speech signal. In the block diagram in Figure 1, the former phase actually contains not only text analysis but also phonetic analysis in which the graphemes are converted into phonemes. And that is the different modules in the TTS system [4].

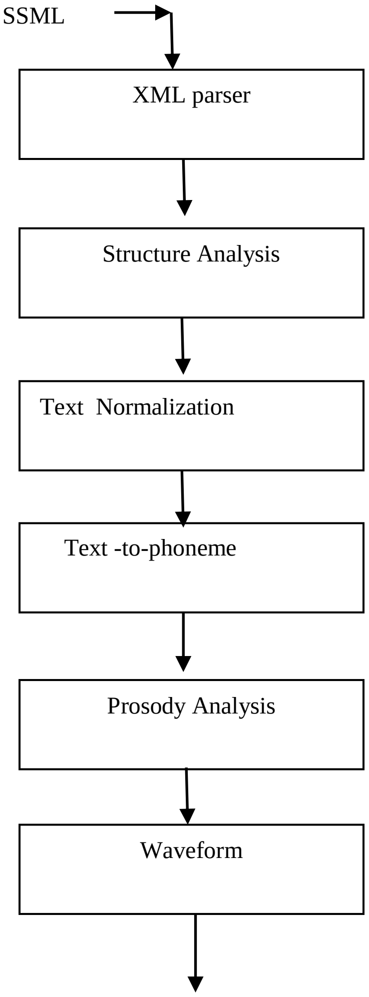

Figure-1: Block Diagram of Speech Synthesis

Example of SSML Document

```
<? Xml version= “1.0” encoding = “UTF-8”>
<speak version= “1.0”
Xmlns=          “http://www.w3.org/2005/01/pronunciation-lexicon”
Alphabet= “ipan” xml:lang= “bn”>
<? xml version="1.0" encoding="UTF-8"?>
<speak version="1.0"
 xmlns="http://www.w3.org/2001/10/synthesis"
xml:lang="HI-in">
<p> ਇਸ ਪ੍ਰਸ਼ਨ ਦਾ ਕੀ ਹੱਲ ਹੈ।</p>
<say-as interpret-as=" chaːlhiː "40 </say -as>
</speak>
```

4. Structure Analysis:

The function is to segment input document into paragraphs and sentences, and to detect the natural language of the written content for each character piece. The structure of a document influences the way in which a document should be read. For example, there are common speaking patterns associated with paragraphs and sentences. The structure of a document influences the way in which a document should be read. For example, there are common speaking patterns associated with paragraphs and sentences. The<p> and <s> elements defined in SSML explicitly indicate document structures that affect the speech output.

.

4.1 **Text Normalization:**

This module converts all other written languages to engine – specific language and all written forms of normalization is process in which written form(orthographic form) automatically converted into spoken form and after the end of this step the text to be spoken is converted into token.

For example:- :- 103 may be spoken as " ਇੱਕ ਸੌ ਤਿੰਨ ", " ਇੱਕ, ਸਿਫਰ, ਤਿੰਨ", "½." may be spoken as " ਅੱਧਾ ","40/-" may be spoken as " ਚਾਲ੍ਹੀ ਰੁਪਏ"

```
<lexeme><grapheme>103</grapheme>
<phoneme> ਇੱਕ ਸੌ ਤਿੰਨ </phoneme>
<phoneme> ਇੱਕ, ਸਿਫਰ, ਤਿੰਨ </phoneme>
<grapheme>"½" </grapheme>
<phoneme>" ਅੱਧਾ " </phoneme>
<grapheme> "40/-" </grapheme>
<phoneme>" ਚਾਲ੍ਹੀ ਰੁਪਏ "</phoneme>
</lexeme>
```

4.2. Text - to Phoneme

The function of this module is to tokenize a sentence into words according to a lexicon, and then to derive the pronunciation for each word. Once the synthesizer processor has determined the set of words to be spoken, it must derive pronunciations for each word. Word pronunciations may be conveniently described as sequences of phonemes. The SSML phoneme tag enables users to provide a phonetic pronunciation for the enclosed text, In this module, we process the following SSML element

Example

```
<?xml version="1.0" encoding="UTF-8"?>
<speak version="1.0"
 xmlns="http://www.w3.org/2001/10/synthesis"
xml:lang="HI-in">
<p> ਤੁਸੀਂ <say-as interpret-as=" ʧɑr"> 4 </say-as> ਕਿੱਥੇ ਜਾ
ਰਹੇ ਹੋ </p>
</speak>
```

So the pronunciation dictionary of this document is below.

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon                                    version="1.0"
xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
alphabet="ipa" xml:lang="Hi-in">
<lexeme>
<grapheme> ਤੁਸੀਂ </grapheme>
<phoneme> tʊsiː◌̇ </phoneme>
<!-- IPA string is:" tʊsiː◌̇ " -->

<grapheme> ਕਿੱਥੇ </grapheme>
<phoneme> kɪ◌̆thɛ </phoneme>
<!-- IPA string is:" kɪ◌̆thɛ” -->

<grapheme> ਜਾ </grapheme>
<phoneme> zaː</phoneme>
<!-- IPA string is:"zaː" -->

<grapheme> ਰਹੇ </grapheme>
<phoneme> rəhɛ </phoneme>
<!-- IPA string is:" rəhɛ " -->
<grapheme> ਹੋ </grapheme>
<phoneme> ho </phoneme>
<!-- IPA string is:"ho " -->

</lexeme>
</lexicon>
```

4.3 Prosody Analysis:

The function of this module is to generate prosodic structures (boundaries for prosodic word and phrase), and to predict target prosodic information (pitch, duration, emphasis) for a certain word. Prosodic structures have been playing important roles in speech communication. Prosodic word and prosodic phrase are the two most important grouping levels for producing synthetic speech with high intelligibility and naturalness. According to the SSML Specification Version 1.0, there are four tags which belong to the categories of “Prosody and Style”: emphasis, voice and break. But I have just focus on the prosody and emphasis tags for Punjabi Language.

4.4 **Prosody tag:**

Prosody is the set of features of speech output that includes the pitch, the timing, the pausing, the speaking rate, the emphasis on words and many other features. Producing human –like prosody is important for making speech sound natural and for correctly conveying the meaning of spoken language. By the use of prosody we see the variation of speech in speech synthesis.
These are prosody tags used in SSML

1. Pitch – set pitch relatively or by number of Hz, Legal values are: a number followed by “Hz”, a relative change or “x-low”, “medium”, “high”, “x-high”, or “default”.
2. Contour- set actual pitch contour,
3. Range – set pitch range, Legal values are: a number followed by “Hz”, a relative change or “x-low”, “low”, “medium”, “high”, “x-high”, or “default”
4. Rate – adjust speaking rate relative, Legal values are : a relative change or “x-slow”, “slow”, “medium”, “fast”, “x-fast”, or “ default”.
5. Duration – adjust duration
6. Volume>>- change the volume of the element content. Legal values are: number a relative change or “silent”, “x-soft”, “soft”, “medium”, “x-loud”, or “default”. The default value is 100.0.

   Example of prosody tags:

Example of prosody tags:

```
<prosody pitch='medium'> ਮੁਖੀ ਜੇ ਤੁਸੀਂ ਇਸ ਸਮੇਂ ਵਿਹਲੇ ਹੋ ਤਾਂ
ਕਲਾਸ ਵਿੱਚ ਚਲੇ ਜਾਉ < prosody>
<prosody volume='loud'> ਸਕੂਲ ਦੇ ਮੁੱਖ ਅਧਿਆਪਕ ਨੇ ਟੀਚਰ ਨੂੰ
ਬੁਲਾਇਆ ਅਤੇ ਕਿਹਾ, ਬੱਚੇ ਬਾਹਰ ਕਿਉਂ ਘੁੰਮ ਰਹੇ ਹਨ < prosody>
ਮੰਮੀ <prosody rate='x-slow'> mɛriː kɪtaːb </prosody> kɪ◌̆thɛ
həɪ
```

4.5. The Emphasis Tag

The emphasis tag indicates that the contained text will be spoken with emphasis. This tag comes with an optional attribute level, which specifies the strength of emphasis. The usage of the emphasis is tag and its attribute is illustrated below:

Example

ਮੰਮੀ –<emphasis level='strong'> ਸ਼ੋਰ ਨਾ ਮਚਾਉ </emphasis>. ਮਚਾਉ "
" ਮੈਨੂੰ ਕੁਝ ਸਮਾਂ ਦਿਉ, <emphasis level='moderate'> ਮੈਂ ਰੋਟੀ ਬਣਾ ਕੇ ਆ ਰਹੀ ਹਾਂ </emphasis> ਰਹੀ ਹਾਂ"

" Ti:chər nɛ kɪha: – sər, <emphasis level='none'> mɛri: kəla:s </emphasis> nəhi:◌̇ həɪ "

4.6 Waveform Production:

This is a final step in producing audio waveform output from the phonemes and prosodic information. This module uses the phonemes and the generated parameters from the prosodic information to produce the audio waveform. There are many approaches to this processing step so there may be considerable processor specific variation.

5. Performance Limitation to SSML for Punjabi Language

This paper has concentrated on elaborating the basic features of basic implementation of Punjabi SSML. It is never easy to categorize the functionally of a document processing system into logical and practical which does not fall cleanly into either of these various categories. This paper supplies a style file which gives information about this voice. These various specifications are not compulsory in the particular sense , that if the system cannot give a voice of this nature the default voice will be used.

6. Future Direction and Conclusion

In this paper, we have presented how the different type of module in TTS System helps to convert a text input of SSML document to spoken form in Punjabi Language. Since, Punjabi is the morphological rich Language, It is written in "Gurumukhi" Script and this is the official Language of Govt. of India. So, According to me In Punjabi Homograph problem will not occur. Because as we write in Punjabi same as we spoke (read). So that is why in Punjabi Language there is no such type of words are their which shows the nature of Homograph. Since, Prosody is an important aspect of speech that helps to maintain expressiveness and intelligibility in speech synthesis

6. References

[1] Silvia Quazza, Laura Donetti, Loreta Moisa, Pier Luigi Salza, "ACTOR:A Multilingual Unit-Selection Speech Synthesis System", Proc. Of 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Atholl, Scotland, 2001.

[2]Davide Bonardo and Paolo Baggia, "SSML 1.0: an XML-based language to improve TTS" White Paper Loquendo Vocal Technology and Services January 19, 2005.

[3]Ze, Heiga, Andrew Senior, and Mike Schuster. "Statistical parametric speech synthesis using deep neural networks." In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 7962-7966. IEEE, 2013

[4]Mohd Bilal Ganai and Jyoti Arora, "Text –to- Speech Conversinon" In International Journal of Innovative Studies in Sciences and Engineering Technology Volume: 2 Issue 1 January 2016.

[5] Shyamal Das Mandal, Somnath Chandra and Swaran Lata, "Use of Part of Speech (POS) and morphological information for resolving Multiple Pronunciations in Pronunciation Lexicon Specification (PLS) for Indian Languages- Bengali as a Case Study" In: W3C Conversational Workshop June 2010.

[6] Daniel C. Burnett and Zhi Wei Shuang, "Speech Synthesis Markup Language (SSML) Version 1.1" In: W3C Recommendation 7 September 2010

https://www.w3.org/TR/speech-synthesis11/

[7]Chris Ward, "An Introduction to Speech Synthesis Markup Language" In: Article of DZone/IoT Zone Apr 21 2017.
https://dzone.com/articles/an-introduction-to-speech-synthesis-markup-language

[8] Paul Taylor and Amy Isard, "SSML: A Speech synthesis markup language" In: Speech Communication Volume 21 Issues 1-2 Pages 123-133, 1997

[9].Wael Hamza, Raimo Bakis, Ellen Eide, Michael Picheny, and John Pitrelli, "The IBM Expressive Speech Synthesis System" In: In Proc. ICSLP 2004.