Text-To-Speech Synthesis System for Punjabi Language

Parminder Singh[a1], Gurpreet Singh Lehal[b]
[a] *Asstt. Professor, Department of Information Technology, Guru Nanak Dev Engineering College, Ludhiana (Punjab) – INDIA*
[b] *Professor, Department of Computer Science, Punjabi University, Patiala (Punjab) – INDIA*

This paper discusses the approach used to develop a Text-To-Speech (TTS) synthesis system for the Punjabi text written in Gurmukhi script. Concatenative method has been used to develop this TTS system using syllables as the basic units of concatenation. After analyzing a carefully selected Punjabi corpus, we have selected nearly thirty three hundred syllables out of about ninety three hundred valid Punjabi syllables. The system is based on a Punjabi speech database that contains the starting and ending positions of syllable-sounds labeled carefully in a wave file of recorded words. The input text is first processed and then syllabified with an automatic syllabification algorithm that has been developed based on grammatical rules of Punjabi language. Then these syllables are searched in the database for corresponding syllable-sound positions in recorded wave file. The paper also discusses the criteria used for the selection of these syllables and the minimum number of words those cover these syllables for recoding.

Keywords: speech database, speech synthesis, Punjabi syllables, syllable sounds.

1. INTRODUCTION

Text-To-Speech (TTS) synthesis system has a vide range of applications in every day life. In order to make the computer systems more interactive and helpful to the users, especially physically and visibly impaired and illiterate masses, the TTS synthesis systems are in great demand for the Indian languages. Punjabi language is being spoken by about 104 million peoples in India, Pakistan and other countries with Punjabi migrants. The language is being written in Gurmukhi script in Indian Punjab, whereas in Shahmukhi script in Pakistani Punjab. In this paper we describe the approach used to develop TTS system for Punjabi text written in Gurmukhi script. Because Punjabi is a syllabic language, so output waveform is generated by concatenating the syllable sounds marked in recorded sound file. Syllable sounds in different contexts have been marked and stored in the speech database to get natural sounding synthesized speech.

2. PUNJABI PHONOLOGY

2.1 Punjabi Phonemes

Punjabi phonemes can be classified as: segmental phonemes and supra-segmental phonemes. Punjabi language, like other Indian languages includes segmental phonemes, but not supra-segmental phonemes in its alphabet.

Segmental phonemes in Punjabi include twenty vowels and thirty eight consonants. Out of twenty vowels ten (ਇ, ਈ, ਏ, ਐ, ਅ, ਆ, ਔ, ਉ, ਊ, ਓ) are non-nasalized and ten (ਿਂ, ੀਂ, ੇਂ, ੈਂ ਅਂ, ਆਂ, ਔਂ ੁਂ, ੂਂ, ੋਂ) are nasalized. And out of thirty eight consonants five (ਙ, ਞ, ਣ, ਨ, ਮ) are nasalized and the remaining consonants are non-nasalized. Punjabi vowels can be classified based on: opening of mouth, position of tongue tip and rounding of the tongue, whereas Punjabi consonants can be classified based upon: place of co-articulation and manner of articulation.

Supra-segmental phonemes include stress, nasality, juncture, tone and intonation [1]. In case of mono-syllablic words, stress is given on the whole word during the utterance of the sentence, whereas in case of poly-syllablic words, stress is given on long, middle syllables of the word, during utterance of word in a sentence [2]. In Punjabi stress is represented by / ੱ/, called *‘addak’*. Presence of *‘addak’* in words changes word meaning and so is phonemic. Stress is multi-dimensional supra-segmental phoneme in Punjabi. In Punjabi nasality is distinctive and its presence changes the meaning of the word. Unlike nasalized consonants, nasalized vowels are not segmental. The nasality depends upon vowels and so is supra-segmental. In Gurmukhi nasality is represented by / ੰ/ (called *‘tippi’*) and / ਂ/ (called *‘bindi’*). Juncture marks the breaks in the speech continuum. Juncture, as in other languages, is important in Punjabi, and it changes the meaning of the sentence. Among all Indo-Aryan languages, only Punjabi is a tonal language [3], in which presence of tone on one or more syllables of the word changes its accent and may change its meaning also.

[1]Corresponding Author: Parminder Singh, Asstt. Professor, Department of Information Technology, Guru Nanak Dev Engineering College, Gill Road, Ludhiana (Punjab) – 141006 (INDIA), Email: parminder2u@rediffmail.com

There are three types of tones in Punjabi – High tone, Level tone and Low tone. Stress and tone are used with each other and can not be separated. Syllable having tone is also stressed and one syllable has only one tone in Punjabi. Word level tone sequences produce intonation that is at sentence level [4].

2.2 Punjabi Syllables

Combination of phonemes gives rise to next higher unit called syllables which is one of the most important units of a language. A syllable must have a vowel called its *nucleus*, whereas presence of consonant is optional. In Punjabi seven types of syllables are recognized [1]. These syllable types are: V, VC, CV, VCC, CVC, CCVC and CVCC; where V and C represent vowel and consonant respectively. Out of these seven types, occurrence of last two syllable types having sound clusters, is very rare in Punjabi.

3. OFFLINE PROCESS

The offline process includes preparation of the Punjabi speech database. The following steps have been followed to prepare Punjabi speech database:

3.1 Selection of Punjabi Syllables

For the development of this TTS system, we have selected syllables as the basic unit of concatenation [5]. The reason for selecting syllables as basic speech units is that, being relatively longer than phonemes, these preserve within unit co-articulation [6, 7]. Out of total seven types of syllables in Punjabi, mentioned in sec. 2.2, we have selected first five types, which gave total 9317 valid syllables. These syllables have been selected after analyzing total possible syllables on carefully selected Punjabi corpus [8] having nearly 233009 unique and more than four million words. Out of these 9317 valid syllables, we finally selected 3312 syllables, which account for more than 99% of commutative percentage frequency of occurrence in the selected corpus.

3.2 Selection of Text for Recording

For labeling the syllable sounds, we have selected words having these syllables for recording. A simple algorithm has been developed based on the set covering problem for selecting the minimum number of words containing above selected syllables, which gave us 2013 words covering 3312 syllables. The algorithm is outlined as under:

1. Get the list of selected syllables.
2. Get the list of corpus words to be searched for the selected syllables.
3. For each word; set *SylsInWord* = number of syllables from the given list, present in that word.
4. Repeat until all syllables are searched for words.
   4.1 Set *CurrWord* = word having maximum value of *SylsInWord* & not already selected.
   4.2 Repeat step 4.3 for each syllable.
   4.3 if (syllable is present in the *CurrWord* & already not found) then
      (a) Output the syllable and *CurrWord* to the output file.
      (b) Set *SylsInWord* = *SylsInWord* -1; for the *CurrWord*.
      (c) Set this syllable as found syllable.
      (d) Set *CurrWord* as the selected word.
5. Exit.

In order to increase the naturalness of the synthetic speech, we have selected first 1000 most frequently used words from corpus to be stored in database as such. Also, for the first 100 most frequently used syllables, we have stored three syllable sounds for each, based upon its occurrence at initial, middle and last position in a word. This is because of the fact that syllable sound varies depending upon its position in the spoken word. This gave us total 3085 words for recording.

3.3 Recording

The selected words have been recorded by a native female speaker of Punjabi. The recording has been

done in a studio with following characteristics:

(a) Sampling Rate: 8000 Hz

(b) Bit Depth: 16 bit

(c) Channels: Stereo

Also, to avoid clipping and nonlinearity, the characteristics of microphone has been set for high performance. However, to reduce the storage size of the recorded sound file, channels has been changed to *mono*, which has reduced the file size by 90%, without much affecting the speech quality for normal use. The total recording time of 3085 words is about 2 hrs and 26 min, having size of 147 MB as wave file.

3.4 Labeling Syllable Sounds

After recording of the selected words, the next phase was to label the syllable sounds in the recorded sound file. This is one of the most important and time consuming task and needs to be done very carefully, because the naturalness of the synthetic speech produced by TTS system depends upon the quality of the syllable sounds and hence quality of speech database. For this purpose we have used sound editing software - *Sonic Foundry Sound Forge 5.0b* and the syllable sounds have been labeled manually one by one, after carefully listening and analyzing the word sounds. The starting and end positions of each syllable sound in the recorded sound file have been noted down.

3.5 Storing Syllable Sounds

The speech database is an important part of a concatenative TTS system, which must be optimized for high quality TTS system. The Database of our TTS system has four fields: syllable, starting position, end position and position of syllable in recorded word (initial, mid or end). The first field contains syllable itself, second and third fields contains starting and end positions of the syllable in the sound file, fourth field mentions the position of syllable in the word from which its sound is being extracted, as mentioned in the section 3.2. Only first 100 syllables have been stored for three positions, resulting in 300 entries in the database. Remaining syllables have been stored for any position in the corresponding recorded word.

4. ONLINE PROCESS

4.1 Text Pre-processing

The input text may have abbreviations, numeric values, special symbols etc., which must be processed before passing the text for the Text-To-Speech conversion. There may be the abbreviations in the input text, which are first searched and then replaced with expanded form, so that written abbreviations be spoken in full word form. Also the numeric values are first analyzed and then expanded to the form required for speaking out that numeric value. Similarly, this module of the developed Punjabi TTS system also searches and replaces the special symbols present in the input text, like $, %, +, -, /, * with the corresponding words. The entered Punjabi text is then segmented into words.

4.2 Syllabification

Since we are using syllables as the basic unit of concatenation, so the words of input text are to be segmented into syllables. There is no fixed procedure in Punjabi language to segment a word into syllables [9]. We have developed a rule based algorithm for automatic syllabification of the Punjabi words, by considering the different grammatical and phonotactic constraints of Punjabi. This algorithm analyses the positions of vowels and consonants in a word, and after taking into account the different grammatical rules, segment the word into syllables.

4.3 Concatenation

This module of the Punjabi TTS system is responsible for finding the boundary values of a syllable sound in the recorded sound file, from the database. The input text that is segmented into words is passed to this module. First search in the database is made to look if the word as such is present in the database (first 1000 most frequently used words have been stored in database as such, as mentioned in Section 3.2). If it is present, the starting and end positions of word sound in the recorded sound file, are returned from the

database, otherwise, syllabification module is called to segment the word into syllables. Now these syllables are searched in the database. First the search is made according to the position (starting, middle or end) of the syllable in the word. If search is successful for that particular position of syllable, starting and end positions of syllable sound in recorded sound file are returned from database. If there is no entry for that particular position of syllable in the database, then that syllable is searched for any position. If the syllable is found at any location, the starting and end positions of its sound in the recorded sound file, are returned from the database. Otherwise this syllable is segmented further into smaller speech units and these units are searched in the database.

5. ARCHITECTURAL DESIGN

The functioning of our TTS system is depicted in Fig. 1.

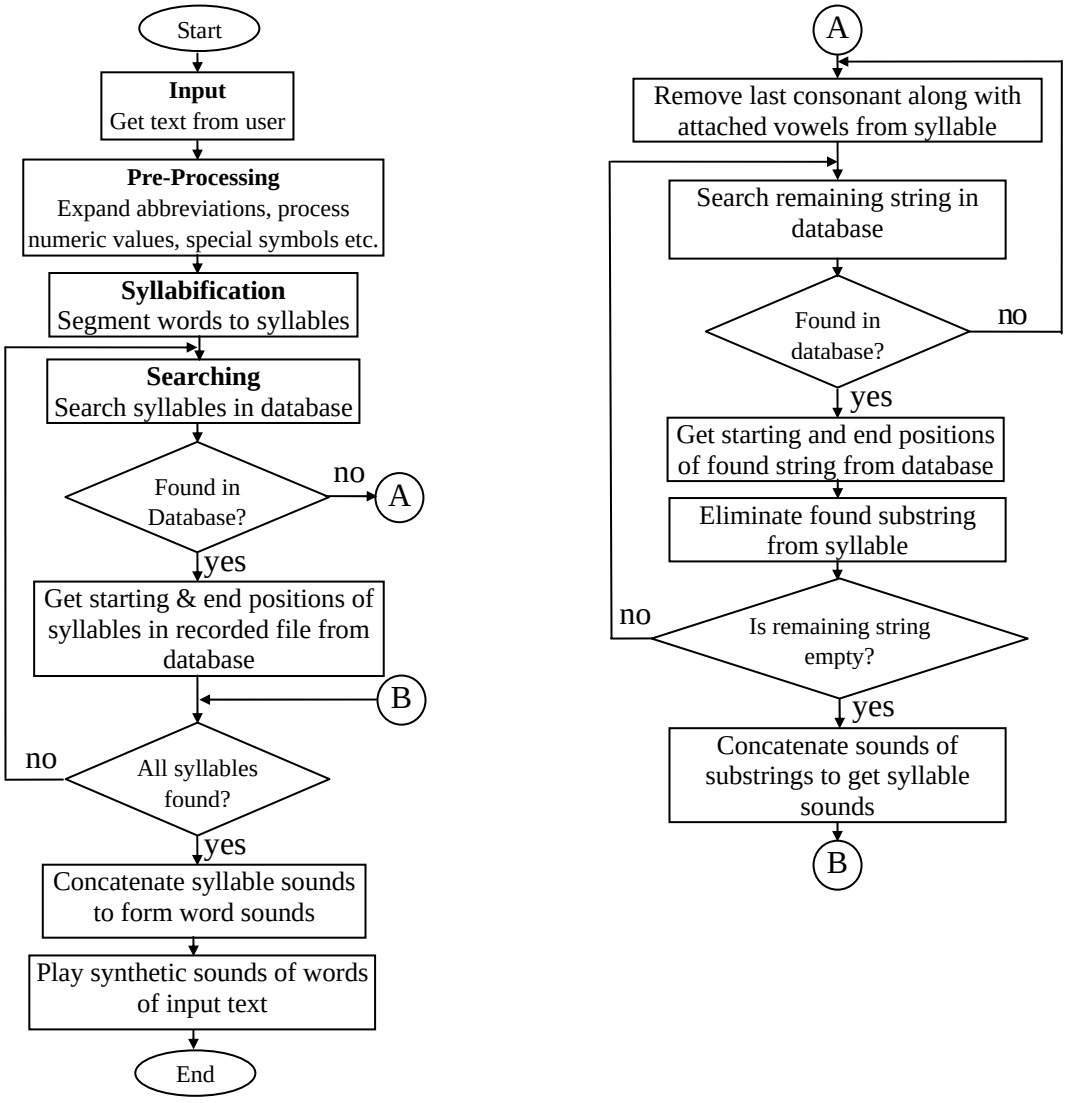

Fig. 1. Flow chart of the Punjabi Text To Speech Synthesis system

6. CONCLUDING REMARKS

By using the above discussed approach, a fairly good quality Punjabi Text-To-Speech synthesis system has been developed. Still, there are some improvements need to be done, especially to achieve smoothness at the concatenation of two syllables.

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.

REFERENCES

[1] Dr. Prem Singh, “Sidhantik Bhasha Vigeyan”, Madan Publications, Patiala, pp. 391.

[2] Harjit Singh Gill and Henry A. Gleason Jr. 1969, “A Reference Grammar to Punjabi”, Revised Edition, Dept. of Linguistics, Punjabi University, Patiala, pp. 25-40.

[3] Narinder K. Dulai and Omkar N. Koul, “Punjabi Phonetics Reader”, C.I.I. Languages, Mysore

[4] S. S. Joshi 1989, “Phonology of the Punjabi Verb”, Classical Publishing Company, New Delhi, pp. 49-88.

[5] Anupam Basu, Debasish Sen , Shiraj Sen and Soumen Chakraborty 2003, “An Indian Language Speech Synthesizer – Techniques and Applications”, Proc. of national systems conference, NSC 2003, Indian Institute of Technology, Kharagpur, India, pp. 17-22.

[6] S. P. Kishore and Alan W. Black 2003 “Unit Size in Unit Selection Speech Synthesis”, EUROSPEECH 2003 – Geneva, Italy.

[7] Donovan R., “Trainable Speech Synthesis”, Ph.D. Thesis, pp. 4-17, Cambridge University Engineering Department, England.

[8] Sunia Arora, Karunesh K. Arora and S. S. Agarwal 2004, *“Speech Corpora Design for the Indian Languages TTS”*, Proceedings of International Symposium on Speech Technology and Processing Systems, Vol. II, New Delhi, India, pp. 122-126.

[9] N. I. Tolstaya 1981, “The Punjabi Language – A descriptive Grammar”, Boston, USA, pp. 10-12.