Punjabi Text to Speech using Phoneme Concatenation

Himmy, Dr. Dharamveer Sharma

***Abstract*— Speech is the natural and primary medium of communication among humans. This paper discusses the development of synthetic speech from Punjabi text given in Gurumukhi script using concatenation of phonemes. Phoneme is the basic unit of concatenation method for speech synthesis. We proposed Text to Speech Synthesis system in which input text is first converted to list of phonemes which are further divided into vowels and consonants. The aim of our research work is to develop the system which produces synthetic speech having qualities of intelligibility and naturalness. In this work, we use phoneme as a unit of concatenation. After analyzing selected Punjabi Corpus from many words, we have selected nearly 830 valid graphemes which are converted to phonemes. The system is based on database of phoneme units. The input text is tokenized to get list of phonemes, then these phonemes are searched from the database for corresponding phoneme in recorded wave file. In order to get natural speech, we use concatenation approach which depends on recording of phoneme units. The database consists of phoneme sound units labeled carefully recorded in wave file.**

***Index Terms*— Text to Speech, Phonemes, speech synthesis, Concatenation synthesis, speech database**

I. INTRODUCTION

NLP- Natural Language Processing

Natural language processing is a field of computer science which concerned with the interactions between computers and human (natural) languages.NLP involve many challenges as natural language understanding, enabling computers to derive meaning from human or natural language and others involve natural language generation. Natural Language processing is required when you want to make an intelligent system which works according to your instructions, when you want to hear decision from a dialogue based system, etc. TTS synthesis makes use of NLP techniques since text data is first input into the system and thus it must be processed in the first place.

Text to Speech Synthesis

The Text to Speech generation is based on natural language processing which deals with interaction between computers and natural language. In recent years, the use of computers in speech synthesis has become an important area in field of artificial Intelligence and natural language processing.

Speech is the primary medium of communication among humans. TTS system includes mainly two parts: natural language processing and digital signal processing.[1] Text to Speech system has wide range of applications in everyday life. In order to make the computer systems more interactive and helpful to the users, especially physically and visibly impaired the TTS synthesis systems are in great demand for the Indian languages. Punjabi language is spoken in India by millions of people which is written in Gurumukhi script in Indian Punjab. The text to speech conversion follows some steps: text preprocessing, text analysis, text phoneziation, prosody generation and speech synthesis algorithms.[2] The system of Text to Speech is developed using concatenation of phoneme sounds. Phoneme sounds are stored in database by doing proper labeling of recorded wave file.

***Manuscript received August, 2017*.**

***Himmy**, Department of Computer Science, Punjabi University, Patiala., Patiala, India, +91-750-802-1026*
***Dr. Dharamveer Sharma**, Associate Professor, Department of Computer Science, Punjabi University, Patiala., India, +91-981-561-5070.*

II. PUNJABI PHONEMES

Punjabi is a language spoken by inhabitants of historical region of Punjab. Punjabi language is subdivided into Gurumukhi (in India) and Shahmukhi (in areas of Pakistan). In Gurumukhi script, the Punjabi language has thirty eight consonants, ten non-nasal vowels and 10 nasal vowels.
A Punjabi phoneme can be defined as minimum sound unit of language. Phoneme represents a class of sounds, which speaker accepts as a single unit. Punjabi phonemes can be classified as: Segmental phonemes and Supra-segmental phonemes.[3]

Segmental Phoneme

These represent the units which have their independent existence. Vowels and consonants are the segmental phonemes. Vowels are the sounds produced by humans when the breath flows out through the mouth without being blocked by teeth, tongue and lips. In Punjabi language, there are total 20 vowels. Out of twenty vowels, ten are non-nasalized and other ten are nasalized vowels.

ਅ ਆ ਇ ਈ ਉ ਊ ਏ ਐ ਓ ਔ — Non-nasalized vowels
ਅੰ ਏਂ ਇੰ ਈਂ ਅਂ ਆਂ ਔਂ ਉਂ ਓਂ ਊਂ — Nasalized vowels

The speech units other than vowels are called consonants There are 38 consonants which are

ਸ ਹ ਕ ਖ ਗ ਘ ਙ ਚ ਛ ਜ ਝ ਞ ਟ
ਠ ਡ ਢ ਣ ਤ ਥ ਦ ਧ ਨ ਪ ਫ ਬ ਭ — Consonants
ਮ ਯ ਰ ਲ ਵ ੜ ਸ਼ ਖ਼ ਗ਼ ਜ਼ ਫ਼ ਲ਼

Out of which five (ਞ, ਙ, ਣ, ਨ, ਮ) are nasalized and the remaining consonants are non-nasalized.

Supra-segmental Phonemes

The phonemes which are used over more than one segment and so these cannot be segmented. Their presence changes

the meaning of graphemes. Supra-segmental phonemes include stress, nasality, tone. The stressed unit is spoken louder than other units to differentiate from others. In Punjabi stress is represented by / ˘ / called 'addak'. Punjabi sounds have nasalized vowels and nasalized consonants. The presence of nasality also changes the meaning of words. In Gurumukhi, nasality is represented by tippi (ੰ) and bindi (ਂ). Bindi and tippi are used with vowels to nasalize the vowel sounds and therefore come along with one of the matras. Both serve the same purpose but bindi is used with kanna(ਕਾਂ), bihari(ਕੀਂ), lanv(ਕੇਂ), dulavan(ਕੈਂ), horha(ਕੋਂ) and kanaorha(ਕੌਂ) whereas tippi is used exclusively with mukta(ਕੰ), sihari(ਕਿੰ), aaunkarh(ਕੁੰ) and dulaenkarh(ਕੂੰ). The languages, where word meaning are dependent on pitch level are known as tone languages. Among all Indo-Aryan languages, only Punjabi is a tonal language because it forms the three tones from the series of consonants- rising tone, falling tone, a mild tone.

III. Techniques of Text to Speech Generation

The methods which are used in conversion of text to speech are usually classified into three groups [4]:

• Articulatory synthesis: This attempts to model the human speech production system directly. This method typically involves models of the human articulators and vocal cords.

• Formant synthesis: It does not any database of speech samples. It uses the wide range of fixed frequency peaks called as formants.

• Concatenation synthesis: This uses different length prerecorded samples derived from natural speech. Concatenation synthesis is based on the concatenation of segments of recorded speech. Generally, concatenation synthesis produces the most natural-sounding synthesized speech.

IV. TTS Using Phoneme Concatenation

Concatenation approach concatenates pre-recorded speech units into the word sequences according to the pronunciation dictionary or set of rules. In this method, memory requirements are usually very high, especially when long concatenation units are used, such as syllables or words. So, to reduce memory requirements, phoneme units are used for concatenation as phonemes consume less memory space. This method follows some steps- first analyze the input text and convert to phonetic transcription. Then modify phonemes duration based on context. Finally, synthesized speech using phoneme units concatenation.

The three main subtypes of concatenation synthesis [5]:

• Unit selection synthesis: It uses large database of recorded speech. During database creation, each recorded sound is segmented into some or all of following- individual phones, diphones, half phones, syllables, morphemes, words, phrases and sentences. Unit-selection for text-to-speech is the common approach for near-natural speech synthesis systems.

• Diphone synthesis: It uses a minimal speech database containing all diphones occurring in a language. The number of diphones depends on phonotactics of language.

• Domain specific synthesis: It concatenates prerecorded words and phrases to create complete utterances. It is used in applications where the output from the system is limited to particular domain.

V. Generation of Punjabi speech system

There are some steps which have to be followed to generate speech from text:

1. *Pre-Processing of text:*
   The user can enter the text which is to be converted in speech. The entered text is divided into words which results into list of words.
2. *Segment words into graphemes*:
   The list of words is now segmented into graphemes. A grapheme is written representation of a phoneme. The user has to make all graphemes that represent the phonemes. As Speech generation system using concatenation of phonemes whereas phoneme is basic unit of concatenation. The list of phonemes is stored in form of linear array. There are always same number of graphemes and phonemes. Phoneme is sound representation of grapheme. Phonemes are building blocks of spoken words.
3. *Database Development:*
   Speech synthesis using concatenation method require database of recorded speech units. Database development includes following steps:
   - 3.1 Selection of Phoneme: Phoneme is selected as basic unit of concatenation. Phonemes consist of vowels and consonants. Selection of phonemes used Punjabi corpus, having a large number of unique words. By analyzing these unique words, it gives nearly 830 phonemes which contains both vowels and consonants (nasalized and non-nasalized).
   - 3.2 Recording of phonemes: The selected phonemes are then recorded by native male and female speaker of Punjabi. The naturalness of speech generated depends upon recording of each phoneme. So, recording has been done by professional or in the studio by using proper bit rate, pitch, sampling rate etc. As recording is important part to produce more natural sound. The recorded file must be saved in wave file format. The recording can be done by using any sound recorder. Use any audio editor to trim or edit the recording or to remove the empty spaces in recorded file.
   - 3.3 Labeling of Phoneme: The next phase is labeling of recorded phoneme. The recorded phoneme sound is labeled carefully because naturalness of synthetic sound is dependent on how phonemes are labeled. So, labeling is most important and time consuming part of database development of phonemes. The phoneme sounds have been labeled manually after carefully analyzing and listening the recorded sound. The recorded file must be in wave file format with proper labeling. It is the most important step in generating this system.
4. *Concatenation of phoneme sounds:*
   The next and important part of speech generation is, after getting list of phonemes from input text, particular phoneme is searched from database of recorded phonemes for concatenation. If search is successful, then it returns that particular recorded wave file.

VI. Architecture Design

The architectural design of TTS system is explained by the flowchart which describes each step of development of text to speech synthesizer. The architectural design of Punjabi TTS system is:

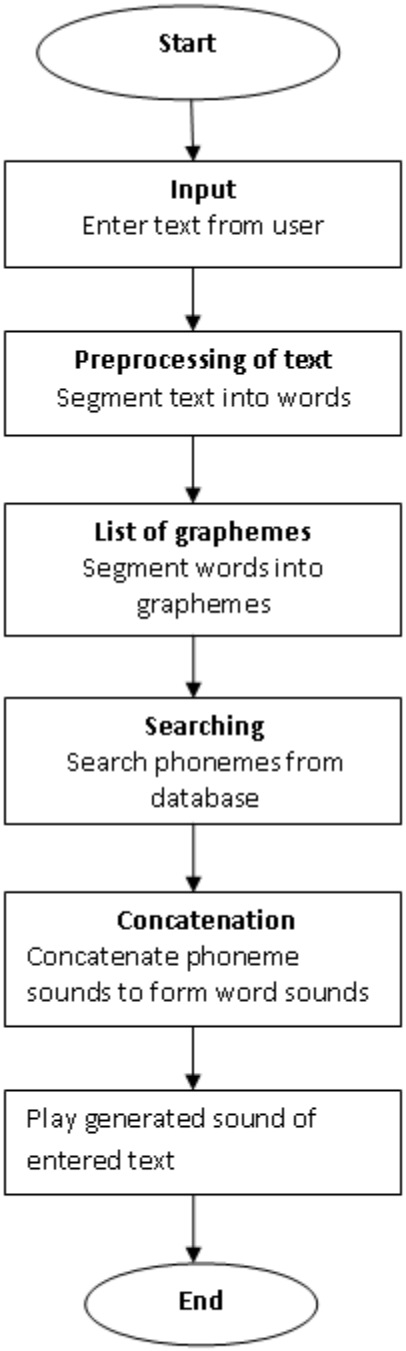

Fig1. Flow Chart of Punjabi Text to Speech Synthesis using Phoneme Concatenation

VII. Implementation

We developed MFC application to implement speech synthesis using concatenation method. MFC application is created in Visual Studio ultimate edition. MFC executables generally fall into five types: standard Windows applications, dialog boxes, forms-based applications, Explorer-style applications, and Web browser–style applications. We used dialog box to implement it. An MFC application is an executable application for Windows that is based on the Microsoft Foundation Class (MFC) Library. The easiest way to create an MFC application is to use the MFC Application Wizard. MFC projects are not supported in Visual Studio Express editions.

1. *Pre-processing of text*: The user can enter input text into text box. A dialog box opens to enter text which has following layout in fig.

2. *Segment entered text*: Next step is to segment the entered text into words and then into graphemes. Grapheme is the smallest unit of a writing system of any given language. So each word is divided into its grapheme units. For example:

The user entered the text " ਇਹ ਪੰਜਾਬੀ ਦਾ ਵਾਕ ਹੈ " ,

then this text is segmented into the following words :

ਇਹ , ਪੰਜਾਬੀ, ਦਾ, ਵਾਕ, ਹੈ ,

then each word is segmented to graphemes, which are:

ਇਹ = ਇ , ਹ

ਪੰਜਾਬੀ = ਪੰ , ਜਾ , ਬੀ

ਦਾ

ਵਾਕ = ਵਾ , ਕ

ਹੈ

ਦਾ , ਹਾ remains same in list of graphemes.

3. *Database Development:* The concatenation synthesis of speech uses prerecorded sound units which are stored in database. So the most important requirement to implement concatenation method is database development. The database development includes phoneme sounds i.e. recording each grapheme to form phoneme as phoneme is smallest speech unit of any language.

- Selection of graphemes: To develop database, first we have to found unique grapheme units of Punjabi language by forming Punjabi corpus of large number of unique words or text in Unicode format to find graphemes of Punjabi. By analyzing it, we found near to 830 graphemes which we have to convert to phoneme for developing database.
- Recording of phonemes: Each grapheme is recorded by native male or female Punjabi speaker or it can be done in studio at particular pitch, bit rate and other prosody characteristics. The software used to enhance or modify the recorded speech is Audacity which helps to remove noise in recorded speech. We recorded each phoneme having following characteristics:
  Project rate: 44100 Hz,
  Channels: mono, 32-bit float
- Labeling of phoneme:
  The phonemes are labeled carefully as to do concatenation of phonemes, these must be labeled carefully and correctly. Each phoneme is labeled with its grapheme symbol. For example , if sound of " ਦੀ " is recorded , then it must be labeled with the name itself i.e. " ਦੀ.wav " as we are saving each

recorded sound in wave file format having .wav extension.

4. *Concatenation of phonemes:* The phoneme units are concatenated to synthesize speech. It is done by using concatenation algorithm on phoneme units, code is written in C++ programming. As for the example, "ਇਹ ਪੰਜਾਬੀ ਦਾ ਵਾਕ ਹੈ" , it finds phonemes from database. Then it concatenates all phoneme sounds to produce speech by concatenation of phoneme units.

VIII. Conclusion and Future Scope

Text to speech system by using the method of concatenation of phonemes, a good quality and natural sound has been produced but it all depends on how each phoneme sound is recorded. To get more natural sound, database must be developed carefully. The aim of developing this system is to produce more natural sound like humans, so it is required to record and label the sounds carefully. During development of TTS system using concatenation approach, some features must be taken care of which are as follows:

- Selection of Punjabi corpus of words
- Selection of graphemes
- Quality of speech depends on recording and labeling of phonemes

We developed Punjabi text to speech system, but still, there are some improvements need to be done, especially to achieve smoothness of sound produced by concatenating phonemes.

REFERENCES

[1] Hay Mar Htun, Theingi Zin, Hla Myo Tun, "Text To Speech Conversion Using Different Speech Synthesis", INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH ISSN 2277-8616, vol. 4, Issue 07, pp. 104-108, July 2015.

[2] Sheilly Padda, Nidhi, Rupinderdeep Kaur, "Architecture and Implementation of Punjabi Text to Speech System Using Transcriptions Concept", International Journal of Engineering Research and Development ISSN: 2278-067X, vol. 1, Issue 5, pp. 08-11, June 2012.

[3] Simmi Luthra, Parminder Singh, "Punjabi Speech Generation System based on Phonemes", International Journal of Computer Applications (0975 – 8887) vol. 49, Issue No.13, pp. 40-44, July 2012

[4] http://research.spa.aalto.fi/publications/theses/lemmetty_mst/chap5.html accessed on 11.01.2017

[5] Pardeep Gera and R. K. Sharma, "Speech synthesis for Punjabi language using syllable-like units", National Conference on Recent Trends in Information Systems (ReTIS-06), Jadavpur University, Calcutta to be held on 14-15 July,2006.

Himmy is a student of M.Tech (CSE) in Department of Computer Science, Punjabi University, Patiala carrying out her research work under the guidance of Dr. Dharamveer Sharma. Her main research interest is Text to Speech for Punjabi language.

Dr. Dharamveer Sharma, Ph. D.(Computer Science), MCA. Presently serving as Associate Professor in Department of Computer Science, Punjabi University, Patiala. His key research areas are Optical character recognition, natural language processing, general computing. He has published more than 100 research papers in journals and conferences of repute including IEEE, ACM, Springer etc. He has developed various softwares and websites, prominent among those are his own website, Unicode Based Punjabi Language Utilities, Result Processor(OMR), Offline Oncampus Admission Counselling system, Encyclopaedia of Sikhism and many more. He has senior membership of Universal Association of Computer and Electronics Engineers (UACEE) and Life membership of Computer Society of India and UIPRAI.