Urdu Text to Speech – Dictionary Based Synthesis – For Software Localization

Abdullah Hanif, Muhmmad Shaharyar Siddiqui, Syed Ali Hasan Shah, Wasi Khan

ABSTRACT

Text-To-Speech (TTS) synthesis is the conversion of text or a sequence of words to speech. Generally divided into two phases. Text processing, where the input text is transliterated into phonetic or some intermediate linguistic representation. The second one is the generation of speech waveforms. In this Text-To-Speech system, text processing is done using dictionary-based approach, this way each syllable is converted into appropriate phonetic transcription. Phone-based synthesis is used for concatenating speech in the latter process, speech synthesis, this is done using Microsoft speech SDK11 as platform's speech engine having capable of working with IPA and UPS, and producing efficient results, fast. The output of this Text-To-Speech system is, speech and SSML (Speech Synthesis Markup Language).

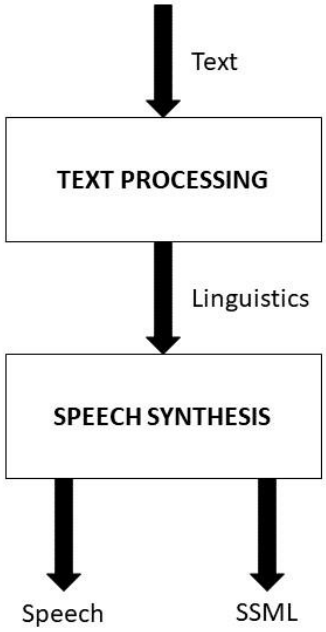

Figure 1. General block diagram of TTS

1. INTRODUCTION

The Text-To-Speech (TTS) synthesis converts an arbitrary input text or a sequence of words for any language into intelligible and natural sounding speech. The raw text given as input to Text-To-Speech system can be of any form. It may consist of numbers, time, dates, symbols and any miscellaneous characters. Therefore, before converting it into speech it must be converted to some form that can be spoken by the Text-To-Speech system. The Text-To-Speech system includes mainly two parts: text processing and speech synthesis. The general block diagram of Text-To-Speech system is shown in figure1.

Text analysis includes segmentation, text normalization, and parts of speech tagger. Phonetic conversion is to assign phonetic transcription to each word. There are two approaches in phonetic conversion. They are rule-based and dictionary-based approaches. Rule-based is applied for unknown words whereas dictionary based is used for known words. The most important qualities of a speech synthesizer are naturalness and intelligibility. Naturalness expresses, how the output sounds like human speech, whereas intelligibility is the easiness with which the output is understood. The technologies for generating synthetic speech waveforms are concatenative synthesis, formant synthesis and articulatory synthesis. Among these, concatenative synthesis is the primary

technology for speech synthesis. It is based on pre-recorded natural sound database. But it is limited to one speaker and usually require more memory capacity. This approach uses a real recorded speech as the synthesis units, such as: phoneme, syllable or word and concatenate the units together to produce speech. Text-To-Speech synthesis is a useful hardware and software tool in many application areas such as: vocal monitoring system for blind people, web browser, mobile phones, personal computer and so forth. Text-To-Speech system is currently developed for teaching aids, text reading, and talking books/toys. However, most Text-To-Speech systems only focus on a limited domain of applications. In this Text-To-Speech system dictionary based synthesis is used.

2. PROBLEM STATEMENT

Having such vast numbers of native speakers, Urdu is still struggling for its digital existence. Working and researching for Urdu Text-To-Speech was not an easy task to complete, one thing we realized very early that there is no community nor some basic completed work available for starters, the very first thought came in our mind is to develop such utility which can read Urdu text, as well as can generate and save data for further development.

3. METHODOLOGY

Text-to-speech system has two parts namely text processing and speech synthesis (digital signal processing). Our approach includes dictionary based synthesis. It takes Urdu text as input, text processing takes place, then text is parsed, and tokens are generated according to the words and diacritics. It then selects the appropriate UPS/IPA representation of the words from the dictionary. UPS/IPA is provided to the speech engine and output is generated.

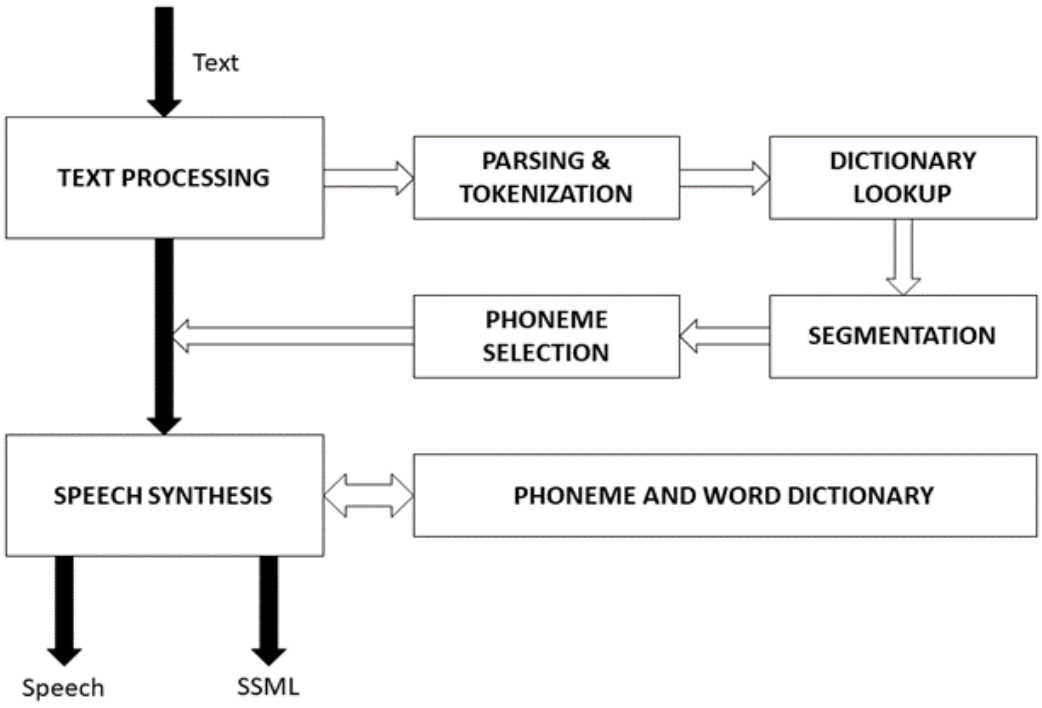

Figure 1. General block diagram of TTS

The output is produced of two types.

1. *Speech*
   Audio output, which is sent to the default output device and which can also be saved to a wav file.
2. *Speech Synthesis Markup Language*
   SSML which comprises of XML like constructs to define the output speech. This SSML can be passed to any other open-source SSML compatible speech engine to produce desired results

3.1. Text Processing

In Text-To-Speech system, first phase is text processing. It takes the raw input sequence and provide useful information to speech synthesizer to produce speech.

3.1.1. Parsing and Tokenization

The input sentence is segmented into tokens of words and delimiters. Some common delimiters are pre-defined and more can be added.

3.1.2. Segmentation

A dictionary lookup is performed against each token and the corresponding UPS/IPA representation of known words is selected, forming a new list of UPS/IPA representations of the words and delimiters.

3.2. Speech Synthesis

This phase of Text-To-Speech system, takes appropriate speech engine representations and convert them into speech

3.2.1. SSML Generation

An SSML string is generated, encoding the UPS/IPA in the standardized format and appropriate pauses for delimiters.

4. SIMULATION RESULTS

Sample 1:

با اصول شخص زندگی کے ہر شعبے میں کامیاب رہتا ہے۔
Points Reader 1: 7 (4-3)*
Points Reader 2: 4 (3-1)
Points Reader 3: 4 (2-2)
Points Reader 4: 7 (4-3)
Points Reader 5: 6 (3-3)

Sample 2:

اسلام ایک مکمل ضابطہ حیات ہے۔
Points Reader 1: 9 (5-4)
Points Reader 2: 8 (4-4)
Points Reader 3: 8 (5-3)
Points Reader 4: 7 (4-3)
Points Reader 5: 9 (5-4)

Sample 3:

ہر ہاتھ ملانے والا دوست نہیں ہوتا۔
Points Reader 1: 9 (5-4)
Points Reader 2: 8 (5-3)
Points Reader 3: 7 (4-3)
Points Reader 4: 8 (4-4)
Points Reader 5: 8 (5-3)

Sample 4:

وقت کا پابند ہونا کامیاب زندگی کی علامت ہے۔
Points Reader 1: 9 (5-4)
Points Reader 2: 8 (4-4)
Points Reader 3: 8 (4-4)
Points Reader 4: 8 (5-3)
Points Reader 5: 9 (5-4)

Sample 5:

استاد قوم کا محسن ہوتا ہے۔
Points Reader 1: 7 (4-3)
Points Reader 2: 6 (3-3)
Points Reader 3: 5 (3-2)

Points Reader 4: 7 (4-3)
Points Reader 5: 5 (3-2)

Sample 6:

محنت میں عظمت ہے
Points Reader 1: 9 (5-4)
Points Reader 2: 8 (4-4)
Points Reader 3: 8 (4-4)
Points Reader 4: 8 (5-3)
Points Reader 5: 9 (5-4)

5. DISCUSSION

In the simulation results section, as you can notice many listeners have rated the speech output as the output is understandable. Urdu Text-To-Speech provides now many ways to utilize this effort, either you can use it as domain specific Text-to-Speech system, where you can provide speech of certain words and Text-To-Speech will speak them fluently, just like native speaker, as far as limitations are concerned there are always tradeoff. Accuracy in speech output there are still no measures for Urdu Text-To-Speech's accuracy test. The best speech you can provide the best the output it'll speak, while entering records we've listened to many output which sounds closely to real word sound but in action when the word participated in certain sentence it made little noisy/glitch sound, so tweaking the input can produce some great speech output, Urdu Text-To-Speech also provide facility to generate IPA for the same sound which can be used in very wide variety, L10n dictionary output can be use also far wide as world wide web is seeking for some entry level Urdu text to speech which is just able to speak up the inputted text. Concisely there are many ways to use this engineered Text-To-Speech system to use.

6. CONCLUSION

In this paper, text to speech system is developed for words and sentence. It is necessary to remove delay in speech when speech waveforms are done concatenation. So, dictionary based synthesis can smoothly produce speech. But the output speech of words are discontinuities between transitions of words. For two or more syllable words, syllabification method is more appropriate. The sound quality is intelligible. Thus, dictionary based, if combined with domain specific synthesis are very easy and efficient to implement unlike other methods which involve many complex algorithms. But in unit selection synthesis, the implementation is not easy as these two methods. The output sentence of speech has little glitch.

7. ACKNOWLEDGMENTS

Thanks to Mr. Badar Sami, PhD(Candid) and Assistant Professor, Umaer Basha Institute of Information Technology for his help and thanks to the supervisor Dr. Muhammad Saeed, PhD(UoK) and Assistant Professor, Umaer Basha Institute of Information Technology for his guidance, support and encouragement.

8. REFERENCES AND LINKS

1) For understanding Urdu structure for speech synthesis (TTS + STT)
   a) http://www.ciil-lisindia.net/Urdu/urdu_struct.html
   b) http://aboutworldlanguages.com/urdu

2) Few working examples along with funded organizations for TTS
   a) http://182.180.102.251:8080/UrduTTS/
   b) http://www.cle.org.pk/software/langproc/UrduText2SpeechAPI.htm

3) Publications
   a) https://www.scribd.com/document/103044106/Urdu-Phonemic-Inventory
   b) http://www.cle.org.pk/Publication/Crulp_report/CR02_16E.pdf
   c) http://dl.acm.org/author_page.cfm?id=81458654458&CFID=746185594&CFTOKEN=70219667
   d) http://www.ijstr.org/final-print/july2015/Text-To-Speech-Conversion-Using-Different-Speech-Synthesis.pdf

4) Urdu words
   a) http://1000mostcommonwords.com/1000-most-common-urdu-words/
   b) Urdu 5000 Highest Frequency Words

5) Microsoft Speech API
   a) Speech Platform: https://msdn.microsoft.com/en-us/library/jj127861.aspx
   b) Speech Synthesis API: https://msdn.microsoft.com/en-us/library/hh362831(v=office.14).aspx
   c) UPS by Microsoft: https://msdn.microsoft.com/en-us/library/hh378346(v=office.14).aspx

6) Source Code
   a) https://github.com/sherry-exec/urdu-tts-lib.git
   b) https://github.com/sherry-exec/urdu-tts-frontend.git

*- The numbers defined in the parenthesis define the rating given by the listeners while listening to the output. The number entered before and after the hyphen (–) defines the ratings of Intelligibility and Naturalness of the speech out of 5 respectively, whereas the number outside the parenthesis defines the total sum of the speech quality out of 10.