**Improving Text to Speech (TTS) conversion in Bangla Language**

Name: Hanif, Md.

ID: 17-33673-1

Section: G

**Motivation:**

Firstly, I want to build a program which will talk to me whenever I want. Sometimes I feel alone and I talk to google assistant software. Though the software is pretty good but it disappoints that it doesn’t ask questions back. There is no curiosity in the software. There are many programs available like google assistant, Alexa, Siri, Cortana. But existing software doesn’t make conversation like the conversation between me and my friend. They don’t ask any question back and they don’t speak in Bengali language. They just give response to its user. I want to build a software that will make a good conversation with human in Bangla language. The software will detect one’s emotion from voice and reply back. It will make a conversion like a conversation between two human friends. It will play games with humans and talk when playing game like two human friends play chess or any game and talk to each other. Existing software does not do these things. A person’s speech can motivate, cheer up, encourage his / her stressed or exhausted friend or calm down an angry person. I want to build this type of software which will motivate, cheer up or calm down a person. The software will act like an artificial friend to a person. In order for me to build a software like that, capable of making conversation like a human friend, I will have to give it plenty of intelligence. It has to be able to detect emotion from human voice, speak in Bangla language and speak accordingly. In case of speaking in Bengali language the software has to generate answers in text and convert the text into spoken output. I have found some website and software where Bangla text to speech conversion has been applied. But The outputs are not natural enough, not intelligent and not smooth enough that will make my software enough friendly to human. In order to apply Bangla language text to speech conversion in the software, the spoken output have to be more natural and smoother and the software has to be intelligent. Thus, the software will be friendlier to human. So, for me to take next step to build an artificial friend like Bangla speaking program, I will have to study a lot of Artificial Intelligence and improve Bangla language text to speech synthesis system.

**Literature Review:**

**Bangla language Introduction:** Bengali Language came from the Indo-European group of dialect’s Indo Aryan part. The language is the primary language of the country of Bangladesh and also the official state language of West Bengal of the Eastern Indian territory. It is additionally spoken in neighbor states of West Bengal such as Assam, Tripura, Bihar and Orissa. Bangla language had gone through a time of intense Sanskritization that began in the twelfth century and proceeded all through the medieval times, producing the tremendous hole between the writing and the accentuation. In this language, the quantity of characters is huge. It has a special type of character called compound or grouped character. At least two characters consolidate to frame a compound or grouped character. Thus, there is around 300 perceivable characters in this language. It has an inflectional nature. (Haldar, 2000)

**TTS conversion and benefits:** Text-To-Speech Synthesis is a method that provides a means of converting written text from a descriptive form to a spoken language that is easily understandable by the end user. In other words, A Text to speech conversion system is able to readout text characters strings to form original spoken sentences. On these days, numerous organizations search for methods for limiting their expenses while conveying improved client experience. Text to Speech (TTS) which is simply the essential innovation behind top notch administration applications answers this need by lessening costs with higher robotization and by giving customized client collaborations. With the assistance of high caliber and common sound it produces, Text to Speech applications empowers organizations to convey wise self-administration and to ensure upgraded client experience. A portion of the further advantages of content to-discourse innovation can be recorded as Normal and human-like voices add to improved client experience and steady brand image can be ensured by utilizing a solitary TTS voice over various contact focuses. With its capacity to vocalize dynamic substance it limits the remaining task at hand of call focus specialists. By expanding robotization it quickens client benefits and lessens operational expenses.

**Existing conversion methods and issues:** There are several speech synthesis systems available such as Domain specific synthesis, Unit Selection Synthesis, Diphone Synthesis. Several works are done in the past to build Bangla language Text to speech synthesis system. There are many processes exists for text to speech conversion in Bangla. Conventional default technique utilizes a lot of setting subordinate phonological standards said Ainsworth (1973).

Bandyopadhyay (2002) Proposes some important aspects to develop a Bangla language text to speech synthesizer system. Phoneme and part name (like Diphone) are utilized to create voice database and ESOLA system utilized for link. Be that as it may, quality may languish over absence of smoothness.

The works of Seddiqui, Azim, Rahman and Iqbal (2002) handles Bangla text to Phoneme from an alternative point of view. The work depends on the possibility of Bangla text to speech that is covered by the modifier phonemes and produces a single diphthong.

Eker (2002) has proposed a study which manipulates the Turkish language structure and attempted to execute the framework that accepts a book as its info. It expects that the text comprises of words and it forms word by word. At the point when a word is gotten from the text, it is passed to a unit that can process word as content and delivers the relating discourse. This part isolates the word into diphones; utilizing diphone database, it gets a discourse document comparing to diphone and its pitch esteem. At last it links the recently recorded speech sections utilizing PSOLA calculation and figures out how to create sound. As a future work, the scientist prescribed that the main thing ought to be done is to finished the diphone database and apply more trials on words. The created yield is worthy for little sentences, yet it requires a lot of time for long sentences. In this way, in request to make some genuine memories understanding framework, the framework ought to be quicker. Taking everything into account, this implies the strategy utilized right now a relevant one which with some exertion on finishing and setting up a better diphone database will bring about a framework that will create progressively reasonable yield for all Turkish words.

In 2003 Indian Natural Language Processing Lab Centre for Development of Advanced Computing (CDAC) used various speech unit to build up the speech combination. It basically utilizes syllable and phonemes. The speech corpus contains generally visit words and initials. They have been sectioned and named into various discourse units as required for advancement of a Hindi discourse synthesis framework. This work doesn't show regarding whether any sort of usage has been made or not. Moreover, it isn't known to have been tried or demonstrated in any related way.

Sen (2004) proposed pronunciation rules and text to speech synthesizer system for Bangla language. The article shows the elocution rule and text to speech synthesizer utilizing formant synthesis procedure.

Weerasinghe, Wasala, Welgama, and Gamage (2007) described a method of tokenization and text normalization part for Sinhala utilizing festival text to speech synthesis system. Gorman and Sproat (2016) introduced an insignificantly directed way to deal with text normalization for Khmer. Be that as it may, the work concentrated distinctly on number normalization and the framework was most certainly not made accessible openly.

Rashid, Hussain and Rahman (2010) proposed the process of text normalization and Diphone preparation for Bangla language. They presented a method for text normalization and Diphone creation based on concatenation-based framework.

Narendra, Rao, Ghosh, Vempada and Maity (2011) proposes the development of syllable based Bangla text to speech synthesis system. The article presents the presents the structure and advancement of unlimited TTS framework for Bangla language. Unlimited TTS framework can incorporate great nature of discourse in various areas.

Alam, Habib and Khan. (2011) proposed a TTS framework that makes the voice information for celebration, and moreover expands out the utilization of celebration to its installed conspire scripting interface to fuse Bangla language support. The authors’ TTS usage utilized two various types of concatenative strategies bolstered in Festival: unit selection and multi-syn unit determination. The specialists on their future work showed that various likely arrangements should be made to build up the total TTS synthesiser for Bangla language including the followings: report investigation, content investigation, phonetic examination, growing huge number elocution vocabulary, programmed dictionary passages as opposed to including physically, discover LTS or Grapheme to Phoneme (G2P) rule with the goal that it can handle obscure words, prosody investigation, and waveform blend by diphone procedure. In the end, the specialists saw that unit determination and multi-syn unit choice has a downside on account of the prerequisite of huge arrangement of speech corpus. The authors demonstrated that the framework had 99% precision for tokens in three classes: floating point numbers, tokens and time.

Mukherjee and Mandal (2012) proposed a Bengali speech synthesizer on the Android operating system. The article depicts a usage of a Bengali speech synthesizer for cell phones. Epoch Synchronous non overlap add (ESNOLA) is utilized for generating speech. The plot depends on connected speech synthesis strategy and part names are utilized as the littlest sign units for connections. But The intelligibility and smoothness are absent in the system.

**Summary:** In these works, researchers have covered most of work to develop a text to speech synthesizer system. Some researchers covered Text normalization, some covered Bangla Grapheme to Phoneme conversion, some covered developing Bangla Text to speech synthesizer using festival. But no authors could show the expected naturality, intelligibility of the systems and smoothness of the spoken output.

**References:**

Ainsworth, W. A. (1973). *A system for converting English text into speech*. Transactions on Audio and Electroacoustics. pp.288-290.

Alam, F., Habib, S.M. M., Khan, M. (2011*). Bangla Text to Speech using Festival*, Conference on Human Language Technology for Development. pp. 154-161.

Bandyopadhyay, A. (2002). *Some Important Aspects of Bengali Speech Synthesis System*. IEMCT Pune.

Eker, B. (2002). *Turkish text to speech system*, The Institute of Engineering and Science of Bilkent University.

Gorman, K. and Sproat, R. (2016). *Minimally supervised number normalization.* Transactions of the Association for Computational Linguistics. vol. 4. pp. 507–519.

Haldar, G. (2000). *Languages of India*. India: National Book Trust.

Mukherjee, S., Mandal, S. K. D. (2012). *A Bengali Speech Synthesizer on Android OS*, Proceedings of the 1st Workshop on Speech and Multimodal Interaction in Assistive Environments. pp. 43–46.

Narendra, N. P., Rao, K. S., Ghosh, K., Vempada, R. R., Maity, S. (2011). *Development of syllablebased text to speech synthesis system in Bengali*. Int J. Speech Technology 14. pp.167–181.

Natural Language Processing Lab Centre for Development of Advanced Computing. (2003) *Building speech corpora for unit selection based concatenative text to speech system for Indian Languages*. India: Vishwabharat.

Rashid, M. M., Hussain, M. A., Rahman, M. S. (2010). *Text Normalization and Diphone Preparation for Bangla Speech Synthesis*. Journal of Multimedia.

Seddiqui, Md. H., Azim, M. A., Rahman, M. S., Iqbal, M. Z. (2002). *Algorithmic Approach to Synthesize Voice from Bangla Text*. 5th ICCIT. pp. 233-236.

Sen, Aniruddha. (2004). *Bangla Pronunciation Rules and a Text-to-Speech System*. Symposium on Indian Morphology, Phonology & Language Engineering. pp. 39.

Weerasinghe, R., Wasala, A., Welgama, V. and Gamage, K. (2007). *Festival-si: A Sinhala Text-to-Speech System*. Text, Speech and Dialogue, 10th International Conference. pp. 472–479.