State of the Art Research for Bangla Text to Speech on Android Platform

Sheikh Abujar
Britannia University
Comilla, Bangladesh

M. S. I. Shahin
Jahangirnagar University
Savar, Dhaka, Bangladesh

Anisur Rahman
Jahangirnagar University
Savar, Dhaka, Bangladesh

Abdus Sattar
Britannia University
Comilla, Bangladesh

ABSTRACT
There are different kinds of TTS (Text to Speech) systems are already available for Personal computers and web applications. In the Platform of Smart Phone, few of TTS systems are available for Bangla Language. Nowadays android is a popular platform considering Smartphone. There are few Bangla TTS Systems are Available with different kind of Mechanisms and techniques, various kind of tools were used. Here we tried to introduce all mechanisms together and proving a summary above all existing system.

Keywords
TTS, Speech Synthesis, Bangla.

1. INTRODUCTION
There are more than 250 million people over 4 states of 2 countries in the world speaks Bengali. We are looking for a device which would be able to read any Bangla text aloud. So now there is no other device than mobile phone as a better option.

There are more than 14 million mobile users in Bangladesh and 30% of them are using smart phones. Use of smart phones are increasing day by day because of reliability, maximum features, capable of using faster internet and eligible for open source application based system. So these kind of features are making our communication very easier and maximum communication is happening over text messaging. So for making our life very easier there are many TTS engines are available for English and many other languages. For bangla there are few more TTS Systems are available in smart phones Platform.

Text and Speech both are very powerful communication infliction. If we can make it easier by converting from text to speech or vice versa than it would be a great achievement in communication life cycle, it will make communication easier than before. People would be able to speak their own words by texting only via Mobile Phone.

Speech is the most natural form of communication and interaction. Speech Synthesis is a major part of TTS engine and for Bangla it is done in many different ways by different authors. From all those we will get the basic idea of Speech Synthesis Techniques.

It is apparent that we are using prerecorded voices for TTS engines yet. Maximum system renders symbolic linguistic representation. So we will discuss about the existing system and possibilities of making the voice very much realistic. The concatenation of final token of speech should be patterned as like real communication.

Recorded voices are stored in Database. System differ in the size of the stored units. As for being the speeches or words recorded by human then the clarity may vary.

Maximum author tried to put most of the effort to code optimization and database compression. They've tried to found many new methods of Speech Synthesis also.

Android is a popular Smart phone operating system because of it allows open source applications to install and use, For this reason anyone can try for making better applications for using or business purpose. So it is very important to build a Bangla TTS for android.

The purpose of our research is to introduce with all of the best TTS Existing systems for Bangla in Android Platform, and ensuring the quality research outputs, findings and Placing possible future works .We discussed about the key points of individual authors and at the end we shown the comparison between all of those.

2. LITERATURE REVIEW
Edification and research for Bangla TTS Engine was improved very highly in last few years. For Android mobile there are many publications available. So here we will discuss about few of them.

Case Study 1:
After studying the paper Title (A benglai Speech Synthesizer on Android OS), authors' names (Sankar Mukherjee and Shyamal Kumar Das Mandal), we have found that they were trying to develop Bengali speech synthesizer on mobile device. They have used Epoch Synchronous Non Overlap Add (ESNOLA) based concatenative speech synthesis technique for Speech generation. They work hard for database compression because where as space was very limited, small diaphone database was being used in previous days which reduced the quality of synthesized Speech. But in other hand (Pucher, M. and Frohlich, 2005) introduced with large unit selection database, they used a Server for synthesized output speech. It was mandatory to transferred the wave form to a mobile device over a network. They tried a quality output in almost real-time on Mobile device.

Speech synthesis is the method of input text data to speech waveforms conversion. The Synthesis method ascertained by the vocabulary size. For utterances of the speech need to be modeled. There are many speech synthesis techniques such as rule-based, articulatory modeling and concatenative technique. But here they developed their synthesizer based on Epoch Synchronous Non Overlap Add (ESNOLA) concatenative speech synthesis method. ESNOLA provides moderate processing for proper matching between different segments during concatenation and it supports unlimited

vocabulary without decreasing the quality. So this could be proposed as a good technique of Speech Synthesis.

They have designed their full operational method as the given diagram. They divided the system in 4 parts including Input text and output speech state. In between they have planned two important states which is Text analysis module and Synthesizer Module. Where the major operations designed to be performed.

A perfect speech required many things such as intonation, prosody, phonological words. And specially handling exception is mandatory while converting text to speech. In this paper they have tried to work with all those parts have mentioned. In their system model they introduced a module named Text analysis module. Which have two sections named phonological analysis module and other one is Analysis of the text for prosody and intonation. They work with the exceptional words at the first Phonological rule part. They developed and implemented phonological rule analysis of the text for prosody and intonation as (Basu, J et al., 2009). They have also work with the exceptional dictionary due to requirement of language analysis. So total processing of text related part ends in phonological analysis module. And synthesizing will be done by the next module.

Synthesizer module works for generating a realistic and quality speech .after getting the finalized text from text analysis module they generate a token and then combine splices of pre-recorded Speech and generate the synthesized voice output using ESNOLA approach as in Shyamal Kr Das Mandal, et al. (2007). In ESNOLA approach, the synthesized output speech is generated by concatenating the basic signal segments from the signal dictionary at epoch positions.

They synthesized like

e.g "ভালো" = bh + bha + a + aL+o .

They had implemented their application in below System specification.

Table 1: System Specifications

Features	LG Optimus One P500
Operating System	Android OS v2.2
Processor	ARM 11
CPU speed	600 MHz
RAM	512 MB
Display	256K colors, TFT
input method	Touch-screen
Connectivity	USB

Memory management is a major issue in android platform otherwise it wouldn't be used broadly. In this paper they have mentioned that this context will live as long as this application is alive and does not depend on the activities life cycle. It is obtained by calling Activity.getApplication(). They kept the partneme database in external storage card. And the best part is after producing output the final speech file will be deleted.

For this TTS system there are total 596 sound files stored in the partneme database. Total size of the database is 1.0 Mb and application size is 2.26 Mb. The best part of this TTS system is it can read Bengali message from phones inbox and it also can generate speech by writing the Bengali word using English alphabet format.

Performance And Quality Evaluation is the major part of any Application. Here the total processing time is counting from the starting time (button is pressed to speak) to the first speech sound is pronounced. They had test the application in many ways and the output of all result is given below.

Table 2 speed time test

Utterance (words)	No. of syllables	Processing Speed [in sec.]
2	6	0.45
3	8	0.56
4	11	0.86
5	15	1.19

They have also judged their application by audience. To measure the output speech quality 5 subjects, 3 male (L1, L2, L3) and 2 female (L4, L5), are selected and their age ranging from 24 to 50. 10 original (as uttered by speaker) and modified (as uttered with android version) sentences are randomly presented for listening and their judgment in 5 point score (1=less natural – 5=most natural). The result is given below.

Table 3 result of listing test

		L1	L2	L3	L4	L5
Modified Sentences	Avg	3.82	1.76	2.62	2.73	3.5
	Stdev	0.73	1.15	0.82	0.81	0.5
Original Sentences	Avg	4.91	4.33	4.82	4.76	4.8
	Stdev	0.11	0.23	0.83	0.42	0.3

The total average score for the original sentences is 4.72 and the modified sentence is 2.88.

In their paper, they describe about implementation of a Bengali speech synthesizer on a mobile device. Their goal was to develop a text-to-speech (TTS) application that can produce real time Speech. They modified several components in ESNOLA to make it run on android device.

Case Study 2:

The objective of a TTS engine is to convert some language Text into its spoken equivalent by a series of modules. For a better TTS engine language modeling and Speech synthesis is major units. After Studying the paper Title(Text to speech for Bangla Language using Festival) authors names (Firoj Alam , Promila Kanti Nath and Dr. Mumit Khan) we found they have used the open-source third party tool Festival TTS engine. Festival provides a frame work for building speech synthesis systems for any TTS engine. The Festival system is written in C++ and uses the Edinburgh Speech Tools Library for low level architecture and has a Scheme (SIOD) based command interpreter for control. Festival Provides API documentation. In their TTS engine they have used two different kind of concatenative methods: unit selection and multisyn unit selection which supported in Festival.

In their research they have discussed about Text Analysis, Phonetic analysis Grapheme to phoneme Conversion, Prosodic Analysis, Speech Database or Waveform Synthesis, Speech Output and Analysis of output result.

The input text may come in nonstandard way, considering this problem they have used the text analysis part to convert all nonstandard words to standard words. Their grapheme-to-phoneme module produces strings of phonemic symbols

based on information in the written text. Final speech synthesis is accomplished by concatenative unit selection technique and multisyn unit selection technique.

In their proposed system the first step is text analysis. The job of a TTS engine is to convert the input text to equivalent Speech, for this reason the input text should convert to a standard format. There is always a chance that the input text may contain NSW (Non-Standard Word) type words. Here the author listed the NSW words as e.g. numbers (year, time, ordinal, and cardinal, floating point), abbreviations, acronyms, currency, dates, URLs. They have used Text normalization for formatting NSW to SW (Standard Word) and they disambiguate the ambiguous token using rule.

In their research they didn't work with Unicode directly because Festival doesn't support Unicode, So that they convert Unicode text to ASCII.

In text analysis part they Split the token based on white-space and punctuation. They consider white space as a separator and Punctuation can separate the raw tokens. Festival Ordered list of tokens, each with features of white-space, and punctuation. For tokenization White-space is the most commonly used .they have identified Bangla Language have more than 10 types of NSW, so each NSW can identify as separate token by token identifier rules. They used scheme regular expression in festival to identify the token. After identifying of all NSW they convert it to standard word by pronunciation lexicon or (letter to sound) LTS rule.

Pronunciation of a word sometimes doesn't match with the pronunciation form. They have solved this problem by using list of lexicon and LTS rule. They inserted 900 lexicons with its pronunciation in the lexicon dictionary.

The Steps of Phonetic Analysis within festival:

1. Building large amount of lexicon.
2. Building letter-to-sound rules.

They have used three techniques for concatenative synthesis: diphone, unit selection and multisyn-unit selection.They identified 45 phones excluding 31 diphthongs with their features based on articulatory analysis. To build diphone database they include diphthong as well. In their implementation they excluded the diphthongs. The duration they added is taken from Kiswahili TTS system but This is not exact duration for the phone set of Bangla language.

They have approximately recorded 500-900 utterance to cover most frequent words of language. The analogy of the system was tested in two ways: in terms of acceptability/naturalness and in terms of intelligibility. Synthesized speech was evaluated on three levels: sentence level, word level and phrase level. In case of sentences level the intelligibility rate being close to 85%. On phrase level it is 83.33% and word level it is 56.66%. In their second experiment, degree of naturalness of the synthesized speech was assessed, again on sentence 90%, phrase 85% and word level 65%. The results Obtained are shown in below Figure.

Case Study 3:

Their model consist of three part, 1st one is "LINGUISTIC MODULE" what generate a linguistic representation from text. 2nd one is "ACOUSTIC MODULE" which generates speech from the linguistic representation. And the 3rd and final one is "VISUAL MODULE" which driving a talking head based on the linguistic representation.

They created a relational lexical database from three source lexica: The Carnegie Mellon Pronouncing Dictionary, Moby Pronunciation II and COMLEX English pronouncing lexicon. There have almost entered 200,000 word, of which over 1500 are non-homophonous homographs. The interesting part of their project is they used animated image which will moved on the subject. In their Linguistic Module they token textual input and looks up word pronunciations and tags in the lexical database. Which words are not present in their lexical database they used a dynamic programming alignment algorithm that algorithm described for aligning sequences from the same alphabets. In Letter-to-sound neural network they defined features for a letter to be the union of the features of the phones that that letter might represent. When they get competitive results they thought that improved performance will come from simplifying the phonological representations found in the dictionary. By this they built a preliminary linguistic representation of the utterance. Then the linguistic representation submitted to a postlexical module where lexical pronunciations derived from the lexicon are converted to postlexical pronunciations typical of the speaker. They consider the distance to word, phrase, clause, and sentence boundaries was included.

After converting the linguistic representation they send it to the Acoustic Module, which has three stage 1.Duration Neural Network , 2.Phonetic Neural Network and 3.Waveform Synthesizer . The acoustic module established the timing of the speech signal by associating segment duration with each phone in the linguistic representation. An acoustic representation, consist of input parameters for the synthesis portion of a vocoder, is generated for each ten-millisecond frame of speech. Finally, the synthesis portion of the vocoder is used to generate speech from these acoustic descriptions. The most interesting part of their module is that they are providing the video for the speech, so it looks like natural. And that reason they collect the animated image from the nature. The video subsystem takes the output of the linguistic module and the output of the duration neural network and generates an animated figure by using an additional neural network.

Case Study 4:

Sanghamitra Mohanty has developed a very intelligent tool, which provides four Indian language Speech output at a time Hindi, Odiya, Bengali and Telegu. For all language she has considered a common system what she named Priyambada. She found Indian languages are phonetic in nature, and the progenitor phoneme mapping is linear. So the vowel and the consonant of the language are almost same except some of them. She took those in consider and apply algorithm for that. We found three stage on this TTS system. First one is Speech Corpora Creation. Here she identified speakers for four native languages, and get them in a laboratory environment using noise cancellation microphone. The sampling rate is 16 bit in single channel of 16000 Hz.By this way she collect the voice from the speakers. Secondly she creates a database for the Different Syllables from the text. She also stored individual polysyllables for different languages in a .wav file format. Finally she played the .wav files for the represented data. There she does not give the solution for the new word what is not in her present. With C++ language she developed a very interesting tool what plays very important role.

Case Study 5:

They actually focus to normalize the text. Most probably their work is same, their processes are tokenization, token classification, token sense disambiguation and word

representation. They found some ambiguous tokens in bangla language. Like, Bangla use many language(English, Arabic, Hindi etc) in their language. the most challenging part of token are the numbers, dates, year, time, multi-text genre etc. To solve this problem they found two ways. One is to token normal bangla language and another table is to handle the ambiguous words.

They levels three stage to token a word i) Tokenizer what will used to token the English and other South Asian scripts Bangla ii) Splitter is used for Punctuation and delimiter and iii) to token phone number, year, time and floating point is used Classifier. It also check the contextual rules, different form of delimiters was removed in this stage, for each type of token, regular expression were written in .jflex format all are checked in this stage.

To make the ambiguous token natural this part is used for. The ambiguous words like non-natural number cardinal, ordinal, acronym, and abbreviations will sound natural. For this the used some stages. Those are (i) Traverse from right to left. (ii). Map first two digits with lexicon to get the expanded form (i.e. 10 → ten). (iii) After the expanded form of the third digit insert the token "hundred". (iv). Get expanded form of each pair of digit after third digit from the lexicon. (v). Insert the token "thousand" after the expanded form fourth and fifth digit and "lakh" after expanded form of sixth and seventh digit. They will continue those stages. After each of second block they insert the token "koti" to make it natural

By this way they believe they can make perfection of 99% of the ambiguous words.

Table 4: Summary of all case studies

Topics	**Case study 1**	**Case study 2**	**Case study 3**	**Case study 4**	**Case study 5**
Tools	ESNOLA	FESTIVAL	NA	Priyambada	JFlex
Processing text type	ENGLISH	ASCII, UNICODE	ENGLISH	NOT DEFINED	ENGLISH
Input text type	BANGLA	ENGLISH	ENGLISH	ENGLISH	ENGLISH
Voice source	Pre recorded	Pre recorded	Pre recorded	Pre recorded	Pre recorded
Total Modules	2	3	NA	NA	NA
Audio format	Not define	Not define	Not define	.Wav	Not define
intonation	Yes	Yes	Yes	Yes	Yes
Utterance	Yes	Yes	Yes	Yes	Yes
Prosody	Yes	Yes	Yes	Yes	Yes
Phonological words	Yes	Yes	Not defined	Not defined	Yes
Exception Handling	Yes	Yes	No	No	Yes
Database length	596 files	Not defined	200,000	Not defined	Not defined
Database size	1.0 Mb	Not defined	Not defined	Not defined	Not defined
Speech quality evaluation	2.88 out of 5.00				
Intelligibility rate	No	85%	No	No	Yes
Word Processing speed	0.45 sec/ 2 word (no of syllable -6)	Not defined	Not defined	Not defined	Not defined
Accuracy	57.8%	85%	87%	Not define	99% for Ambiguous word

Table 4 explains, different parameters covered in all stated case studies. A simple overview was added in that table. .The tools they have used in their research and the depth of their research covered were stated in there. The details of their research were discussed in individual case studies. This table is just a gist of those case studies.

3. RESULT DISCUSSION

Several Text to speech development were held for English language. Bangla language is very much different than other languages. Linguistic rules are very different and uniformed. The level of utterance is more critical in Bangla language. In this paper, it was discussed that from many other research findings- few of them are suitable to apply for Bangla TTS or at least the initial understanding of how Bangla TTS could be developed, could be found. Detail materials were being used for different TTS development were mentioned individually. Different method of processing TTS were discussed and provide the gist for better understanding.

4. CONCLUSION AND FUTURE SCOPE

Text to Speech is one of the most important features of machine learning. Several Text to speech models are available for English language. Though a small research contribution was held for Bangla TTS. This research article states several case study of different TTS approaches and possible applications and discussing research outputs. Based on this literature reviews, a better Text to speech model could be proposed and developed.

Developing Bangla TTS requires large dataset or a corpus or text data as well as voice data. Without using supervised learning it would be difficult to train the TTS engine for better output. Bangla language utterance have several forms based on their use or text position. Bengali vowels have many different form of utterance. So, a better Text and voice dataset is must for Bangla TTS development.

5. REFERENCES

[1] Frances Alias, Xavier Servillano, Joan Claudi socoro and Xavier Gonzalvo "Towards High-Quality Next Generation Text-to-Speech Synthesis:A multi domain Approach by Automatic Domain Classification",IEEE Transactions on AUDIO,SPEECH AND LANGUAG PROCESSING, VOL16,NO,7 september 2008.

[2] Qing Guo, Jie Zhang, Nobuyuki Katae, Hao Yu , "High – Quality Prosody Generation in Mandrain Text-to-Speech system", FujiTSu Sci.Tech,J., vol.46, No.1,pp.40-46 ,2010.

[3] Gopalakrishna anumanchipalli, Rahul Chitturi, Sachin Joshi, Rohit Kumar, Satinder Pal Singh, R.n.v Sitaram, D.P.Kishore, "Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition System",

[4] A.Black, H.Zen and K.Tokuda "Statistical parametric speech synthesis", in proc.ICASSP, Honolulu, HI 2007, vol IV, PP 1229-1232.

[5] G.Bailly, N.Campbell and b.Mobius, "ISCA special session: Hot topics in speech synthesis", in proc.Eurospeech,Genea, Switzerland, 2003, pp 37-40.

[6] M.Ostendorf and I.Bulyko, "The impact of speech recognition on speech synthesis", in proc, IEEE Workshop Speech Synthesis, Santa Monica,2002,pp. 99-106.

[7] Text To Speech Synthesis - a knol by Jaibatrik Dutta .

[8] Silvio Ferreia,Celina Thillou, Bernaud Gosselin, "From Picture to Speech: an Innovative Application for Embedded Environment",

[9] M.Nageshwara Rao, Samuel Thomas, T.Nagarajan and Hema A.Muthy, "Text-to-Speech Syntheis using syllable line units"

[10] Jindrich Matousek, Josef Psutks, Jiri Krita, "Design of speech Corpus for Text-to-Speech Synthesis". Beckman M. and Elam G. "Guidelines for ToBI Labeling". Manuscript, version 3, 1997.

[11] Corrigan G., Massey N., and Karaali O. "Generating Segment Durations in a Text-to-Speech System: A Hybrid Rule-Based/Neural Network Approach". Proc. Eurospeech '97, Rhodes, September 1997.

[12] Gerson I., Karaali O., Corrigan G., and Massey N. "Neural Network Speech Synthesis". Speech Science and Technology (SST-96), Australia, 1996.

[13] Karaali O., Corrigan G., and Gerson I. "Speech Synthesis with Neural Networks". Invited paper, World Congress on Neural Networks (WCNN-96), San Diego, September 1996.

[14] Karaali O., Corrigan G., Gerson I., and Massey N. "Text-to- Speech Conversion with Neural Networks: A Recurrent TDNN Approach". Proc. Eurospeech '97, September 1997.

[15] Kiparsky P. "Lexical phonology and morphology". Linguistics in the morning calm, ed. by I.S. Yang. Seoul: Hanshin, 1982.

[16] Kruskal J. "An overview of sequence comparison". Time Warps, String Edits, and Macromolecules, edited by Joseph Kruskal and David Sankoff. Reading, MA: Addison- Wesley, 1983.

[17] Linguistic Data Consortium. COMLEX English pronouncing lexicon. Trustees of the University of Pennsylvania, version 0.2, 1995.

[18] Miller C., Karaali O., and Massey N. "Variation and Synthetic Speech". NWAVE 26, Quebec, October 1997.

[19] Nusbaum H., Francis A., and Luks T. "Comparative valuation of the quality of synthetic speech produced at Motorola". Research report, Spoken Language Research Laboratory, University of Chicago, 1995.

[20] O'Shaughnessy, D. "Modeling fundamental frequency, and its relationship to syntax, semantics, and phonetics". Ph.D. thesis, M.I.T., 1976.

[21] Sejnowski T. and Rosenberg C. "NETtalk: a parallel network that learns to pronounce English text". Complex Systems 1.145-168, 1987.

[22] Seneff S. and Zue V. "Transcription and alignment of the TIMIT database". M.I.T., 1988.

[23] Tuerk C. and Robinson T. "Speech Synthesis using Artificial Neural Networks Trained on Cepstral Coefficients". Proc. Eurospeech '93, Berlin, September 1993.

[24] Ward G. Moby Pronunciator II, 1996.

[25] Weide R. The Carnegie Mellon Pronouncing Dictionary. cmudict.0.4, 1995.