DOI 10.29042/2019-4931-4936

Helix Vol. 9 (3): 4931- 4936

Recent Trends in Text to Speech Synthesis of Indian Languages

*[1]Sarang L. Joshi, [2]Vinayak K. Bairagi
[1]Research Scholar, AISSMS IOIT, Pune
[2]Professor, AISSMS IOIT, Pune
*Email: jsarang70@gmail.com

Received: 30th November 2018, Accepted: 13th February 2019, Published: 30th June 2019

Abstract
A Text To Speech (TTS) synthesizer is a computer application capable of converting arbitrary input text into speech. This conversion broadly involves two steps, namely, text processing and speech synthesis. Text processing converts the entered text to a sequence of synthesis units, while speech synthesis is the generation of an acoustic wave form corresponding to each of these units. Naturalness and intelligibility are the most important qualities expected from a TTS system. In this paper we aim to provide an overview of various techniques for text to speech synthesis, discuss their characteristics, summarize and compares advantages and drawbacks. We have listed various Text-to-Speech synthesis frameworks developed and implemented at different Indian institutes.

Keywords
Concatenative, Prosody, Speech Synthesis, Syllable, TTS, Text to Speech

Introduction

Speech is one of the premier forms of everyday life communication among humans. Most of the information in digital world is accessible only to those few people who can read or understand a particular language. Computers communicating with the common man in the language he understands and convenient for him can make the digital content reach to the masses by facilitating the exchange of information across different people speaking different languages. For visually impaired, speech impaired, educationally under privileged and the rural communities, text to speech (TTS) conversion can reduce the trouble in human interaction with computers and make it as simple as ABC (*Kishore et al.,2003*; *Gaikwad et al., 2014).*

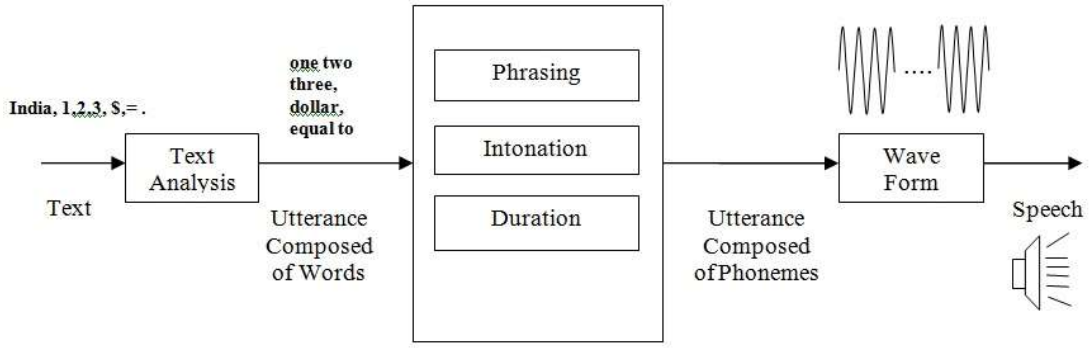

Figure 1: Overview of Text to Speech Synthesis

Text Preprocessing

The written text is a series of electronically coded alphabetic, alphanumeric or special characters (ASCII, ISCII, UNICODE) or a transliteration scheme of various fonts. Segmentation, a first step in text preprocessing, decomposes the long chain of characters into blocks easier to handle, e.g. sentences, which are further decomposed to words. The case of sentences is more complicated than words, as for example, the character "." may represent the end of a sentence, an abbreviation or in some cases decimal point *(Repe et al., 2010)*. Accurate segmentation is required for text to speech synthesis systems. Automatic segmentation using statistical approaches leads to poor boundary information when small training data is used whereas manual segmentation leads to human errors *(Pradhan et al., 2015)*.

Normalisation

The arbitrary text is an unrestricted set of characters which may consists of different numbers (e.g. 0,2.1,X, VI;), abbreviations such as Mr., Mrs., Prof., Dr., Ar., Er. and/or special characters such as #, &, etc. and/or currency($, €, £, ₹), upper and lowercase, punctuation(!, ?," "), date(01/06/1989), mathematical expressions (=, >, <,~, %), acronyms etc. Normalization means transforming the input text to a series of pronounceable words [2].e.g. 1968 if a number, is pronounced as "one thousand nine hundred sixty eight"; however if

 Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

represents a year then to be pronounced as “nineteen sixty eight”. 01/06/1989 to be pronounced as “first of June nineteen eighty nine”. The pronunciation of a certain word may depend on the context. In some cases e.g. currencies, the order of some character and value is changed. For example $20 is converted as twenty dollars while $200 million is to be converted as two hundred million dollars and not two hundred dollars million.

Phonetic Transcription

This is mapping of characters in the language alphabet to easily understandable symbolic representation, often known as grapheme to phoneme conversion, is based on pronunciation dictionary, letter to sound rules and lexical accents (as sound is context dependent). In most languages the written text does not correspond to its pronunciation. Hence, some kind of symbolic presentation is required for describing correct pronunciation.

Synthesis Techniques

Formant Synthesis

Formants are the resonance frequencies of the vocal tract that make sounds distinct. At least three formants are usually needed for producing intelligible speech while up to five formants result in high quality speech. Formant synthesis estimates these frequencies by modeling the vocal tract transfer function and generates speech. To artificially reconstruct the formant characteristics a set of resonators are excited using a voicing source or noise generator and the desired speech spectrum is achieved. Voicing or voicelessness is achieved by controlling the excitation source. The additional set of anti-resonators can simulate nasal tract effects, plosives and fricatives. The specifications of more than 20 such parameters are required for satisfactory restitution of the speech signal. However, majority of parameters are yet required to be manually optimized as automatic techniques used for specifying formant parameters are still not satisfactory. To generate the filter parameters formant synthesis relies on large set of rules written by linguists which is difficult even for simple words. MITalk, KlatTalk, and DECTalk are the applications of this technique (*Tabet et al. 2011)*.

Articulatory Synthesis

Produces high quality speech by direct modeling of the human articulator behaviour with the help of control parameters such as tongue tip position, tongue tip height, tongue position and tongue height, lip aperture, lip protrusion, velic aperture and excitation parameters such as glottal aperture, cord tension and lung pressure. Data acquisition for articulatory model is a difficult task as it is usually derived from X-ray analysis of natural speech. X-ray data is two-dimensional, while the real vocal tract is naturally three-dimensional. The tongue movements are too complicated which makes precise modeling almost impossible (*Tabet et al. 2011; panda et al. 2015)*.

Concatenative Synthesis

Speech is generated by connecting natural, prerecorded speech units such as phonemes, diphones, triphones, half syllables, syllables, words or sentences. The number and size of individual units decides the database size. Longer units, e.g. phrases or sentences, increase the naturalness, need less concatenation points but require more storage . Shorter units, e.g. phones, require less memory, can synthesize a wide range of words or sentences, however degrade the speech quality and also need more complex sample collecting and labelling techniques.

The speech unit should cause minimal concatenation distortion, minimal prosodic distortion and should be of general nature for unrestricted text to speech synthesis. Syllables are not influenced by neighbouring sound elements, are acoustically and perceptually more stable units than phones. The syllable unit performs better than diphone, phone and half phone *(Kishore et al.,2003)*.

Unit Selection Synthesis

Unit selection synthesis stores multiple instances of each unit with varying prosodies in the unit inventory; however results in large database (1-10 hours). With the help of a unit selection algorithm, the closest matching unit to the target prosody is selected thereby minimizing the need of prosodic modification. As compared to diphone based concatenative synthesis, unit selection synthesis technique provides more natural speech output as it requires less amount of digital signal processing to the recorded speech.

Hidden Markov Model (HMM) Synthesis

It involves training and synthesis phase. Mel frequency cepstral coefficients (MFCC) and their first and second derivatives are the most commonly used training features. Feature vector is formed by extracting features per frame. In the synthesis phase the feature vectors for a given phone sequence are estimated first and then transformed into audio signals by filter implementation.

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

TECHNIQUE	ADVANTAGES	LIMITATIONS
FORMANT SYNTHESIS	don't use a database of speech samples hence useful for limited memory and processing costs , e.g. embedded system,	artificial, robotic sound; difficulty in finding parameters from the input specification, created by the text analysis process
ARTICULATORY SYNTHESIS	Produces high quality speech	Most difficult to implement, Unavailability of sufficient data. Computationally more complex
CONCATENATIVE SYNTHESIS	Simple data driven approach(uses natural prerecorded speech samples)	The unit length affects the speech quality; limited to 1 speaker and 1 voice
UNIT SELECTION SYNTHESIS	Produces more natural speech requires much less modification of the speech units	high development time and cost for collecting and labelling the data, large memory resource requirements as it relies on a very large database
HIDDEN MARKOV MODEL SYNTHESIS	Only parameters of the models are stored and not the data itself, flexibility in changing voice characteristics, speaking styles and emotions, with little modification the same model can be applied to various languages	naturalness is still far from that of natural speech, suffer from buzziness. Conversion of prosodic features is difficult

Table 1: Comparison of Text to Speech Synthesis Techniques [5][7]

The formant synthesis was dominant for long time, however nowadays the concatenative synthesis is gaining the popularity. The articulatory synthesis has received less attention as compared with other techniques, quite rarely used in present systems as it is still very much complicated for high quality implementations however, with rapid development of analysis techniques and the computational resources it may arise as a potential method in near future.

Prosody

Prosody, an important aspect of speech, maintains expressiveness and intelligibility in speech. Continuous speech without the breath pauses or breath pauses at wrong places not only makes the speech sound unnatural but may also lead to misunderstand the meaning of the sentence e.g., the text "Anil says Sunil is a cheater" can be expressed in two different ways with two different meanings as "Anil says: Sunil is a cheater" or "Anil, says Sunil, is a cheater". In first case Anil is a cheater, while in the second case the cheater is Sunil.

Prosody is language dependent .Prosodic features pitch, duration and stress depend on gender, age, physical and emotional state and attitude of the speaker. Unfortunately, written text hardly contains any information about these features. The pitch contour depends on the meaning of the sentence e.g. the pitch slightly decreases toward the end of the sentence for normal speech and raise to the end of sentence when it is in a question form.

The most crucial and challenging task to achieve high quality speech in a TTS system is to derive as much relevant information (e.g. correct intonation, stress, duration) from the input text as possible. Predicting prosody for text to speech synthesizers is heavily dependent on the punctuation marks and the part of speech (POS) tags of the words in the text.

TTS Development for Indian Languages

For languages like English, user Interfaces for IT applications and services have become more and more prevalent. However, in a country like India, with relatively lower rates of literacy, the majority of the population is not comfortable using English, hence interfaces in local language needs to be developed to access IT applications, information and services on health, agriculture, travel, etc. through internet and/or telephones.

Among the 22 official languages of India. Nepali, Hindi and Marathi follow 'Devanagari' script; the others such as Tamil, Kannada and Telugu have their own scripts however they all share a common phonetic base. Indian language scripts are originated from the ancient Brahmi script. The basic units of the Indian language writing system called 'Aksharas'(Letters), orthographically represent a speech sound, are syllabic in nature. All Indian languages consist of 15-18 vowels and 35-38 consonants *(Pradhan et al., 2015)*. The most frequently occurring syllables hardly exceed 300 for each language.

Text to speech for 13 Indian languages have been developed under the consortium formed consisting of IITs and CDAC institutes across India under TDIL programme of MeitY, Govt. of India [8].

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

DEVELOPER	LANGUAGE	TECHNIQUE
IIT MANDI	RAJASTHANI	UNIT SELECTION
SSNCE CHENNAI	TAMIL	UNIT SELECTION
C-DAC SNLP LAB NOIDA	HINDI	CONCATENATIVE
C-DAC MUMBAI	MARATHI	UNIT SELECTION
C-DAC KOLKATA	BANGLA	ESNOLA CONCATENATIVE
ESpeak	ENGLISH,PUNJABI, MALAYALAM,HINDI	FORMANT SYNTHESIS
C-DAC THIRUVANANTHPURAM	MALAYALAM	ESNOLA CONCATENATIVE
IIT GUWAHATI	ASSAMESE ,MANIPURI, INDIAN ENGLISH	UNIT SELECTION
DA-IICT , GANDHINAGAR	GUJARATI	FESTIVAL & HTS FRAMEWORK
UTKAL UNIVERSITY,ORISSA	ODIYA, HINDI , BENGALI, TELUGU	CONCATENATION
IIT MUMBAI	MARATHI	CONCATENATIVE
B.A.M.U. AURANGABAD	PALI	UNIT SELECTION
SIMPUTER TRUST (DHVANI TTS)	HINDI,KANNADA, MARATHI, GUJARATI,TAMIL	-

Table 2: TTS Development for Indian Languages [9][10][11][12][13][14][15][16][17][18]

INSTITUTE	RESEARCH LAB	TTS FRAMEWORK
IIT MUMBAI	CENTRE FOR INDIAN LANGUAGE TECHNOLOGY	VANI (HINDI)
IIT MADRAS	SYSTEMS DEVELOPMENT LABORATORY	TELUGU
UTKAL UNIVERSITY, ORISSA	SPEECH TECH GROUP	ORIYA
HYDERABAD CENTRAL UNIVERSITY	LANGUAGE ENGINEERING RESEARCH CENTRE	VAANI (TELUGU)
IIIT HYDERABAD	SPEECH AND VISION LABORATORY, LANGUAGE TECHNOLOGIES RESEARCH CENTRE	TELUGU
IISC BANGALORE	MEDICAL INTELLIGENCE AND LANGUAGE ENGINEERING	THIRUKKURAL(TAMIL) VACHAKA(KANNADA)
CDAC BANGALORE	-	MATRUBHASHA

Table 3: TTS Research Activity in India [19][20][21][22]

Database

In speech synthesis the basic need is the speech corpora. Developing speech corpora is essential to enable the researchers study the acoustic and linguistic properties of speech and to develop speech synthesis models. To build a corpus a phonetically and prosodically rich text file (e.g. news bulletin, forum interviews, everyday conversations in an organization or in road traffic etc.,) should be selected which is then read and recorded by a native speaker. The recording may range from several minutes to hours *(Kiruthiga et al. , 2012)*.

Desirable characteristics of database:

- Sentences from diverse sources.
- Grammatically correct, meaningful and natural sentences.
- Simple, short, easy to read sentences.

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

The quality of the synthesized output depends heavily on the quality of the data collected. Hence, studio recording by professional male/female speaker must be done. Proper care and necessary measures should be taken during the multiple recording sessions to ensure that the same quality of speech is maintained.

Database Developed by	Language
Punjabi University, Patiala	Punjabi
Utkal University, Bhubaneswar	Hindi, Odiya, Bengali & Telugu
Islampur, Maharashtra	Konkani (Goan)
SJ College of Engineering, Mysore	Kannada
Linguistic Data Consortium for Indian Languages	Marathi & 21 other Indian Languages
IIIT , Hyderabad	Kannada, Hindi, Bengali, Malayalam, Marathi, Telugu, Tamil
Technology Development for Indian Languages	Punjabi, Hindi, Marathi, Manipuri, Bangla, Assamese

Table 4: Speech Database (Corpus) Development in India [24][25]

Applications

Proofreading your own writing, helping people with reading disability or low vision, creating answering machine messages, help to reduce eye strain from too much reading, listening an eBook or information during your commute, daily walk or run, study a second language, listen to a text read in different languages, amusing kids by letting your PC read stories to them, preparing for a big speech by hearing your words read aloud, public announcements at bus stands or railway stations, reading emails or web pages.

Research Challenges

1] Written text does not contain explicit emotions, pronouncing proper and foreign names is sometimes very aberrant. Correct prosody and pronunciation analysis from written text is required.

2] Speech quality is a multidimensional term and the evaluation technique must be carefully selected. To evaluate the quality of output speech, a subjective parameter Mean Opinion Score (MOS) is commonly used. There is a need to define new objective parameters for comparing TTS techniques. Objective assessment techniques will provide the most efficient type of evaluation.

3] Existing prosody models used in unit selection synthesis predict duration and intonation at phone level units. Suitable techniques need to be developed for prosody modelling at longer units like syllable and polysyllable.

4] Prosody models for many languages such as Japanese, French and English are based on set of rules used for predicting tones and break indices (ToBI). However, for Indian languages being rarely punctuated, incorporating the appropriate prosody for a given text in a TTS system is particularly a hard task due to the lack of information present in the text (punctuation such as commas denoting prosodic phrase breaks are usually absent.). Tools to POS tag Indian language text are still not completely effective. Techniques predicting appropriate phrase boundaries are required.

Conclusion

The linguistic structure of each language is different hence synthesis rules are different for all languages. Though the research work has extended to cover few more languages, still many languages are left to be covered. A syllable can be the best unit for Indian language speech synthesis as Indian languages are syllable centred i.e. pronunciation is mainly syllable based. Prosody plays a crucial role in making speech sound natural. As a field, study of the prosody of Indian languages is still in its infancy. There is dire need of working on common platforms / fonts / standards / procedures / software tools. More resources should be made available electronically with fewer restrictions.

References

1. Kishore, S. Prahallad, and Alan W. Black . 2003. Unit size in unit selection speech synthesis. Eighth European Conference on Speech Communication and Technology.
2. Gaikwad, Prakash B. and Dr VK Bairagi. 2014. Hand gesture recognition for dumb people using Indian sign language. International Journal of Advanced Research in computer Science and Software Engineering. 193-194.
3. Repe, Madhavi R., S. D. Shirbahadurkar, and Smita Desai. 2010. Natural Prosody Generation in TTS for Marathi Speech Signal. International Conference on Signal Acquisition and Processing, ICSAP'10, IEEE. 358-361.
4. Pradhan, Abhijit, Anusha Prakash, S. Aswin Shanmugam, G. R. Kasthuri, Raghava Krishnan, and Hema A. Murthy. 2015. Building speech synthesis systems for Indian languages", in Twenty First National Conference On Communications (NCC). IEEE 1-6.

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

5. Tabet, Youcef, and Mohamed Boughazi.2011. Speech synthesis techniques. A survey. 7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA). 67-70.
6. Panda, Soumya Priyadarsini, Ajit Kumar Nayak, and Srikanta Patnaik. 2015. Text-to-speech synthesis with an Indian language perspective. International Journal of Grid and Utility Computing 6, no. 3-4 : 170-178.
7. Zen, Heiga, Keiichi Tokuda, and Alan W. Black. 2009. Statistical parametric speech synthesis. Speech Communication 51, no. 11: 1039-1064.
8. TDIL, http://tdil-dc.in/index.php?option=com_vertical&parentid=85&lang=en, last accessed 2018/11/19
9. https://cdac.in/index.aspx?id=mc_st_speech_technology, last accessed 2018/11/19
10. Mandal, Shyamal Kr Das, and Asoke Kumar Datta. 2007. Epoch synchronous non-overlap-add (ESNOLA) method-based concatenative speech synthesis system for Bangla. SSW 351-355.
11. eSpeak, http://espeak.sourceforge.net/, last accessed 2018/11/19
12. Gopi, Arun, P. Devi Shobana, T. Sajini, and V. K. Bhadran. 2013. Implementation of malayalam text to speech using concatenative based TTS for android platform. International Conference on Control Communication and Computing (ICCC). 184-189.
13. Mahanta, Deepshikha, Bidisha Sharma, Priyankoo Sarmah, and SR Mahadeva Prasanna. 2016. Text to speech synthesis system in Indian English. Region 10 Conference (TENCON). 2614-2618.
14. Speech research lab DA-IICT, https://sites.google.com/site/speechlabdaiict / projects/tts-gujarati, last accessed 2018/11/19
15. Mohanty, Sanghamitra. 2011. Syllable based Indian language text to speech system. International Journal of Advances in Engineering & Technology 1. 2: 138-143.
16. RCILTS, http://www.cfilt.iitb.ac.in/westernmeetjune12/, last accessed 2018/11/19
17. Mache, Suhas, and C. Namrata Mahender. 2016. Development of Text-to-Speech Synthesizer for Pali Language. IOSR Journal of Computer Engineering (IOSR-JCE). 18, 3: 35-42.
18. Dhvani, http://dhvani.sourceforge.net/index.html, last accessed 2018/11/19
19. Jain, Harsh, Varun Kanade, and Kartik Desikan. 2004. Vani-An India Language Text to speech Synthesizer. IIT, Mumbai.
20. SVL projects, http://speech.iiit.ac.in/index.php/projects-list.html, last accessed 2018/11/19
21. Rama, GL Jayavardhana, A. G. Ramakrishnan, R. Muralishankar, and R. Prathibha. 2002. A complete text-to-speech synthesis system in Tamil. Proceedings of 2002 IEEE Workshop on Speech Synthesis, IEEE . 191-194.
22. Sarma, Sireesha, R. K. V. S. Raman, S. Sridevi, and Rekha Thomas,. Matrubhasha–An Integrated Speech Framework for Indian Languages.
23. Kiruthiga, S. and K. Krishnamoorthy. 2012. Design issues in developing speech corpus for Indian languages—A survey. International Conference on Computer Communication and Informatics (ICCCI), IEEE . 1-4.
24. TDIL resources,http://tdil-dc.in/index.php?option=com_download &task =fsearch &lang=en , last accessed 2018/11/19
25. http://speech.iiit.ac.in/index.php/research-svl/69.html , last accessed 2018/11/19

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495