International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue III, March 2018- Available at www.ijraset.com

A Survey :On Text To Speech Synthesis

Ramita P. Karpe[1], Nagaraj Vernekar[2]
[1, 2] *Department of Computer Science and Engineering, Goa College of Engineering*

Abstract: ***Text-to-speech (TTS) synthesizer converts an input text into speech waveforms. It has been an effective tool for many visually challenged people for reading through hearing feedback. The ability to convert text to voice reduces the dependency, frustration and sense of helplessness of such people. India is called as the land of unity and diversity, with 22 official languages being spoken throughout the country. TTS systems are mostly available in English; however, it has been observed that people feel more comfortable in hearing their own native language. There has been a significant improvement in the research related to the development of the TTS system in Indian Languages in the past few years. An attempt is made to address the important results reported so far and it has also tried to highlight the beneficial directions of their search till date. Moreover, the paper also contains a comprehensive bibliography of many selected papers appeared in reputed journals and conference proceedings as an aid for the researchers working in the field of Text to Speech Synthesizer.***
Keywords: TTS; Syllable; phoneme; vowel; consonant

I. INTRODUCTION

Communication has prevailed in various forms since humans first appeared on the earth. All species in this universe have communication system. Speech is one of the most important ways of communicating with each other and especially, it is the main medium of human communication. Earliest form of communication includes cave painting, usage of drums, storytelling to give information for the next generations. Natural Language processing (NLP) is an interesting area of research, which takes natural language text or speech and explores useful things. The dependence of human on computer interaction on written text and images makes the use of computers impossible for visually and physically impaired and also the illiterate masses. These obstacles can be overcome by using automatic speech generation from sentences of natural language. Unfortunately, as we can see that in the present era of human computer interaction, the educationally backward and the rural communities of India are being deprived of technologies that spread the growing interconnected web of computers and communications. A good solution for this problem would be communications. A good solution for this problem would be computers talking to the common man in the language he is comfortable to communicate in. Speech is a natural means of communication among human beings and gives a very good platform for man-machine interaction. Further it is also desirable that human-machine interface permits one's native language of Communication. India is a multilingual country having 22 official languages. A text to speech conversion is very important for human computer interactions which allow environmental barriers to be removed for people with a wide range of disabilities.

A. Challenges in Text to Speech Systems

Speech synthesis has been developed steadily over the recent decades and it has been integrated into several new applications. Developing speech synthesis system is a complicated process and, it includes the following important challenges:

1) Development of TTS systems requires knowledge about human speech production and about languages being developed.
2) The actual implementation of a fully functional system requires good software skills.
3) Most TTS systems do not generate semantic representations of their input text; as a result, various heuristic techniques are used to guess the proper way to disambiguate homographs [1], like examining neighboring words and using statistics about frequency of occurrence.
4) The most important qualities of a speech synthesis system are naturalness and intelligibility [2]. Naturalness describes how closely the output sounds are related to each other.
5) Intelligibility is the ease with which the output is understood. The ideal speech synthesizer should be both natural and intelligible [16].

B. Nature of Indian Scripts

Most of the Indian scripts have originated from ancient Brahmi script through various transformations. The basic units of the writing system are referred to as Aksharas. The properties of Aksharas are as follows: (1) An Akshara is an orthographic representation of a speech sound in an Indian language;(2) Aksharas are syllabic in nature;(3) The typical forms of Akshara are V,CV,CCV and CCCV, thus have a generalized form of C*V. The shape of an Akshara depends on its composition of consonants and the vowel, and

sequence of the consonants. In defining the shape of an Akshara, one of the consonant symbols acts as pivotal symbol (referred to as semi full form). Depending on the context, an Akshara can have a complex shape with other consonant and vowel symbols being placed on top, below, before, after or sometimes surrounding the pivotal symbol (referred to as half-form)[3].

II. SPEECH SYNTHESIS METHODS

There exist several different methods to synthesize speech. Each method falls into one of the following categories: articulatory synthesis, formant synthesis, and concatenative synthesis. A succinct description of these categories is presented in the following subsections.

A. Articulatory Speech Synthesis

Articulatory synthesis consists of computational biomechanical models for the production of speech. These models simulate the effect of natural speech production organs (articulators) such as tongue, lips, glottis and moving vocal tract. The simulation of these articulators is based on time- dependent, 3-dimentional differential equations that compute synthetic speech output. The execution of such models requires extremely high computational resources on one hand and on the other hand the output of such models is not ranked near to natural-sounding fluent speech [4, 5, 6].

B. Formant Speech Synthesis

Formant synthesis consists of a set of rules that controls a highly simplified source-filter model. The model is based on the assumption that it is possible to treat the source (glottal) as completely independent from the filter [7]. The operation of the filter is based on the bandwidth and formant frequencies which are used as the control parameters. The values of the format frequencies are related to the specific resonance of the vocal tract. Though the production of speech with formant synthesis methods has advantages of requiring moderate computational resources and producing highly intelligible speech; however, it does not produce completely natural sounding speech.

C. Concatenative Speech Synthesis

Concatenative synthesis uses recorded snippets of smallest (elementary) speech units called phonemes. Phonemes consist of both phones (a vowel or a constant) and phone-to-phone transitions called diphones. The phonemes are recorded using high quality apparatus in a sound-proof environment to develop an inventory called voice database. Each phoneme in the voice database is stored either as waveform or it is encoded with a suitable speech encoding algorithm. When the text is applied to concatenative synthesizer, it first performs character to phoneme mapping and then assembles the corresponding phonemes to generate the speech. As the speech generated with this method uses recorded snippets of actual sounds it has highest potential for sounding natural [8].

III.SYLLABLE BASED TEXT TO SPEECH

Current TTS systems can be classified into 2 categories, viz., (1) rule-based synthesis and (2) concatenative synthesis. A rule-based synthesizer is a knowledge-based system based on the explicit description of the control parameters over time. It consists of a set of rules derived from an explicit knowledge of speech production mechanism, at least at the acoustic-level. Rule- based synthesis approach includes, articulatory synthesis and formant synthesis [10].

In concatenative speech synthesis approach, there is no necessity to determine the speech production rules. Hence, concatenative synthesis is easier than rule-based synthesis. Concatenative synthesis generates speech by combination of natural, pre-recorded speech sound units (e.g., words, syllables, half-syllables, phonemes, half- phone, diphones or triphones) [9]. Unit selection-based speech synthesis technique (USS), which is a kind of concatenative synthesis, where numerous instances of each sound unit is stored with varying speech prosody. The unit that best matches the target prosody is selected and joined. The following sections discuss the various methods reported so far for Syllable based Text to Speech for Indian languages.

A. The Development of Syllable Based Text to Speech System for Tamil language[11]

The methodology described by M .Karthikadevi, et al. [12], the text to speech system is divided into 2 phases namely Text analysis and Phonetic analysis phase. NLP component is responsible for handling text analysis.it requires 2 databases namely text database and a phonetic database. The input given to the system is checked with the speech corpus, whether the speech exists. If the word does not exist, the word is segmented into syllables using syllabification rules. Followed by letter to sound conversion using pronunciation dictionary. The phonetic analysis maps each phoneme is with corresponding wav files in the speech database. The Unit Selection Algorithm describes the optimal sequence of the units and concatenating these units to get the synthesized speech. This algorithm leads to concatenate the units with the help of unit cost, which yields low concatenation points. Unit Selection cost

function should ensure that the selected optimal unit sequence should closely match with the target unit specification and with other adjacent units in the sequence [13].

B. Algorithms for Speech Segmentation at Syllable-Level for Text-to-Speech Synthesis System in Gujarati[15]

The proposed system by Hemant A. Patil, et al. [15] makes use of Minimum Phase Group Delay-based Segmentation and Gaussian Filter-Based Speech Segmentation at syllable level. Voice building in Gujarati TTS is carried out with fallback mechanism, i.e., in order to synthesize everything (i.e., all possible text); at least one instance of each syllable must be present in the speech corpus. As the number of syllables is also not fixed, it is very difficult to cover all the syllables of a language. TTS systems were evaluated for their speech intelligibility. It has been observed that the intelligibility is better in case of Gaussian-based segmentation approach.

C. Algorithms for Speech Segmentation at Syllable-Level for Text-to-Speech Synthesis System in Gujarati[15]

Hemant A Patil, et al. [15] proposed an algorithm based on Festival framework [16] .few modifications were done to the festival framework in order to incorporate all the 13 languages. As Indian languages are syllable-timed, a syllable- based framework was developed. As quality of speech synthesis is of paramount interest, unit-selection synthesizers are built. For the speech corpus a large amount of text from the Internet was collected. This included news data, blogs and short stories in Indian languages. To ensure coverage of all domains, and given that Indian languages are low resource languages, efforts were made to generate text to cover domains not available on the web. After the collection of data the speech corpus was recorded. DON Label tool was used for labeling the entire speech corpus. As syllable encompasses coarticulation, the prosody modification required was significantly less.

D. Text to Speech Synthesis System in Indian English[17]

Deepshikha Mahanta, et al. [17]. An effort has been made to modify the existing English grapheme to phoneme dictionary by implementing specific rules for one particular variety of Indian English, namely Assamese English. The proposed method of dictionary modification are applied at the front end of the Indian English TTS, developed using unit selection synthesis and statistical parametric speech synthesis frameworks for Assamese English, phonemes as basic units are considered for concatenation both in the USS and SPSS based approaches. The data is segmented in phone level with Hidden Markov Model (HMM) based forced Viterbi alignment [18]. In SPSS, the spectral parameters representing vocal-tract and excitation information are extracted from the given speech utterances. Spectral parameters typically include Mel generalized cepstral coefficients (MGCs) [19] and their dynamic features or Line Spectral Pairs (LSPs) and their dynamic features. The dynamic features are the first and second or- der derivatives of speech parameters, which are included to accommodate dependency of features of one frame over its nearby frames [20].

IV. HMM BASED TEXT TO SPEECH

HMM-based speech synthesis technique is a statistical parametric synthesis technique. Unlike USS, in training phase, the parameters are extracted from the speech data to form context-dependent hidden Markov models (HMMs) which is obtained from context-independent HMMs.

The major advantages of HMM-based speech synthesis system are (i) free from sonic glitches and (ii) footprint size is significantly low such that these systems can be implemented in hand-held devices too. Presently, HMM-based speech synthesis systems exist for languages like Japanese, Mandarin, Korean, English, German, Portuguese, Swedish, Finnish, Slovenian, etc.

A. Development and Evaluation of Unit Selection and HMM Based Speech Synthesis Systems for Tamil[21]

Ramani Boothalingam, et al. [21] designed and evaluated a system based on unit step ad hmm model. It consists of training phase and synthesis phase. During the training phase, context-dependent models are trained for different phonemes by extracting the duration, spectral, and excitation parameters. During synthesis, for the given text, based on the context-dependent label files generated, the corresponding models are concatenated. The spectral and excitation parameters are obtained from the concatenated models and a speech waveform is synthesized from these. Tree-based clustering is then carried out to generate context dependent models, for which a question-set, specific to the language, plays a vital role. The question-set may contain any number of questions as required by the language, to describe the contextual features of the phonemes. Greater the number of appropriate questions, more specific is the clustering. A subjective evaluation of the developed speech synthesis systems is carried out using the conventional mean opinion score (MOS), which is a five-point grading scale.it is concluded that HMM-based voice built using the same amount of speech data outperforms FestVox based voice. The reason is due to the fact that there are no sonic-glitches present in the speech synthesized.

V. TD-PSOLA BASED TEXT TO SPEECH

TD-PSOLA [23] is an effective technique that can be used for duration and pitch scale modifications. It helps in scaling the duration and pitch period without losing any source or system information. The modification is carried out at signal level and hence there is no significant distortion in the synthesized speech. To modify the pitch contour in the neutral speech and synthesize using TD-PSOLA, the GCIs are obtained using DYPSA and the keywords in the sentence are located by the process of keyword identification.

A. LP and TD-PSOLA Based Incorporation of Happiness in Neutral Speech Using Time Domain[22]

Sreenidhi S, et al. [22] it focuses on incorporating happiness into neutral speech using signal processing algorithms. In this regard, neutral and happy speech are analysed and it is found that happiness can be perceived in certain emotive words in a sentence. Thus, in order to introduce happiness into neutral speech, these emotive keywords are identified and the above mentioned time-domain parameters are modified. Linear Prediction based synthesis of happy speech is initially performed. To improve the quality of the synthesized speech, TD-PSOLA is then used. An attempt has been made to incorporate happy emotion in neutral speech. Certain emotive words in the sentence such as “fantastic”, “great”, “wonderful”,etc., express happiness. These emotive words are identified in the neutral speech and the time domain parameters of these words are modified to incorporate happy emotion. It is concluded that TD-PSOLA based synthesized speech is rated higher than the LP approach, indicating that the emotional content is perceived more clearly in the former method.

VI. COMPARISON OF TECHNIQUES

In this section, we propose a comparative study of major speech synthesis techniques. In this study, we focus on the three performance factors that consist of effectiveness, flexibility and simplicity. In the case of effectiveness, we emphasize in the naturalness of the speech include almost like human voice and without noise. Flexibility, we consider the difficulty of editing when applied to applications. In addition, it can adjust the parameters. Simplicity is a state of being a simple technique an easy to understand thus we represent on the complexity of an implementation. To better understand, the comparative study is depicted in Table I. Furthermore, the advantages and disadvantages of these techniques are also discussed. As shown in Table I, Unit Selection generates a better naturalness than the conventional techniques. In contrast, Unit Selection has some limit to allow the voice modification and reduces the range of its applications. TD-PSOLA gives a good quality voice when used on both speech signals and very fast in computation. Finally, HMM has most flexibility to change its voice characteristics than the other techniques and also smooth speech sounds. Unit Selection gives a high quality speech, intelligible as natural human speech, and without noise environment. However, Unit Selection has difficulty in adapting the parameters and some limit to allowing the voice modification. Meanwhile, HMM is the most flexible and simplest. HMM is also able to change its voice characteristics, speaking styles, and emotions.

TABLE I PERFORMANCE COMPARISON TABLE

Techniques	Performance		
	Effectiveness	Flexibility	Simplicity
Unit Selection	Speech quality has better naturalness than the conventional techniques.	-	Difficult to produce voice quality variations
HMM	Synthesize highly intelligible and smooth sounds	Changes its voice characteristics and emotions	Relative ease with which HMM based systems adapted to speakers not present in dataset
TD-PSOLA	Gives good result when used on both speech signal		Ease of Implementation

VII. CONCLUSION

In This paper, we have presented a survey of several speech synthesis techniques. The focus nowadays is on the unit selection synthesis. Such synthesis methods allow for more natural-sounding modifications of the signal. The main limitation of the unit selection synthesis combined with HNM is high processing cost and fewer variations are allowable on the recorded data. Hidden

Markov Model Synthesis is statistical methods that allow more variations on the recorded data and this method will become dominant.

REFERENCES

[1] Klatt, D. H.,.The Klattalk text—to—speech conversion system, in Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82, pp 1589 – 1592.
[2] Choudhury, M., Rule Based Grapheme to Phoneme Mapping for Hindi Speech Synthesis, in 90th Indian Science Congress of the International Speech Communication Association (ISCA)2003, Bangalore, India.
[3] Prahallad L.,Prahallad K.,and GanapathirajuM.,"A simple approach for building transliteratio neditors for indian languages," Journalof Zhejiang UniversityScience, vol. 6A,no.11,pp.1354–1361,2005
[4] Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., and Syrdal, A. (2000), "The AT&T Next-Gen TTS System", IEEE Proceedings. R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
[5] Black, A.W. and Lenzo, K. A., "Multilingual text-to- speech synthesis", EUROSPEECH '99, Budapest, Hungary.
[6] Sondhi, M. M., and Schroeter, J, "Speech Production Models and Their Digital Implementations", The Digital Signal Processing Handbook,V.K. Madisetti, D. B. Williams (Eds.), CRC Press, Boca Raton,Florida , 1997
[7] Jilka, M., Syrdal, A. K., Conkie, A. D. and Kapilow, D. A. (2003), "Effects on TTS quality of methods of realizing natural prosodic variations",ICPhS2003,Barcelona,Spain.
[8] [Klatt, D.H. (1987), "Review of Text-to-Speech Conversion for English", Journal of the Acoustical Society of America. 82 (3): 793- 857.
[9] Y. Tabet and M. Boughazi, "Speech Synthesis Techniques. A Survey," in 7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA), pp. 67-70, 2011
[10] D. Jurafsky and J.H. Martin, Speech and Language Processing.: Prentice Hal,2000.
[11] M.Karthikadevi, Dr.K.G.Srinivasagan, The Development of Syllable Based Text to Speech System for Tamil language,International Conference on Recent Trends in Information Technology , 2014.
[12] Benoy Kumar Thakur, Bhusan Chettri, Krishna Bikram Shah, "Current Trends, Frameworks and Techniques Used in Speech Synthesis-A Survey," Int. J. of Soft Computing and Engineering, ISSN: 2231-2307, vol. 2, No.2, pp. 442-446, May 2012.
[13] Narendra N.P, K.Sreenivasa Rao, "Optimal weight tuning method for unit selection cost functions in syllable based text-to-speech synthesis," Elsevier, Applied soft computing, vol. 13, February 2013.
[14] Hemant A. Patil, Tanvina Patel, Swati Talesara, Nirmesh Shah, Hardik Sailor, Bhavik Vachhani, Janki Akhani, Bhargav Kanakiya, Yashesh Gaur and Vibha Prajapati, Algorithms for Speech Segmentation at Syllable-Level for Text-to-Speech Synthesis System in Gujarati, Oriental COCOSDA held jointly with Conference on Asian Spoken Language Research and Evaluation(O-COCOSDA/CASLRE), International Conference , 2013
[15] Hemant A Patil1, Tanvina B Patel1, Nirmesh J Shah1, Hardik B Sailor1, Raghava Krishnan2, G R Kasthuri2 T Nagarajan3, Lilly Christina3, Naresh Kumar4, Veera Raghavendra4, S P Kishore4, S R M Prasanna5, Nagaraj Adiga5, Sanasam Ranbir Singh5, Konjengbam Anand5, Pranaw Kumar6, Bira Chandra Singh6, S L Binil Kumar7, T G Bhadran7, T Sajini7, Arup Saha8, Tulika Basu8 K Sreenivasa Rao9, N P Narendra9, Anil Kumar Sao10, Rakesh Kumar10, Pranhari Talukdar11, Purnendu Acharyaa11, Somnath Chandra12, Swaran Lata12, Hema A Murthy 2, A Syllable-Based Framework for Unit Selection Synthesis in 13 Indian Languages, Oriental COCOSDA held jointly with Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), International Conference , 2013
[16] A.W. Black, P. Taylor, and R. Caley, "The Festival speech synthesis system," http://festvox.org/festival/, 1998.
[17] Deepshikha Mahanta, Bidisha Sharma, Priyankoo Sarmah, S R Mahadeva Prasanna, Text to Speech Synthesis System in Indian English, IEEE Region 10 Conference (TENCON) — Proceedingsof the InternationalConference , 2016
[18] M. Nageshwara Rao, Samuel Thomas, T. Nagarajan, and Hema A. Murthy, "Text-to-speech synthesis using syllable-like units," in National Conference on Communication, 2005, pp. 227–280.
[19] S. P. Kishore and A. W. Black, "Unit size in unit selection speech synthesis," in Eurospeech, 2003, pp. 1317–1320.
[20] M Sreekanth and A G. Ramakrishnan, "Festival based maiden TTS sys- tem for Tamil language," in 3rd Language and Technology Conference, Poznan, Poland, October 2007, pp. 187–191.
[21] Ramani Boothalingam , V Sherlin Solomi, Anushiya Rachel Gladston, S Lilly Christina, P Vijayalakshmi 1, Nagarajan Thangavelu, Hema A Murthy, Development and Evaluation of Unit Selection and HMM-Based Speech Synthesis Systems for Tamil, National Conference on Communications (NCC), 2013.
[22] Sreenidhi S, Anushiya Rachel G, Vijayalakshmi P, Nagarajan T, LP and TD-PSOLA Based Incorporation of Happiness in Neutral Speech Using Time Domain, Power and Computing Technologies [ICCPCT],2014
[23] E. Moulines and F. Charpentier, "Pitch-synchronous waveform pro- cessing techniques for text-to-speech synthesis using diphones," Speech communication, vol. 9, pp. 453–467, 1991.