IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. I (May-Jun. 2016), PP 35-42
www.iosrjournals.org

Development of Text-to-Speech Synthesizer for Pali Language

Suhas Mache[1], C. Namrata Mahender[2]
[1](*Research Scholar, Department Of CS & IT, Dr.Babasaheb Ambedkar Marathwada University Aurangabad, India)*
[2](*Asst. Professor, Department Of CS & IT, Dr.Babasaheb Ambedkar Marathwada University Aurangabad, India)*

Abstract : *We introduced a new method for Text-To-Speech (TTS) synthesis for Pali language. We discuss the efforts in collecting speech database of Pali language and relevant design issues in development of TTS system. This system is based on unit selection concatenative speech synthesis using phonemes, syllables and words as an elementary unit for Pali speech synthesis. The speech units picked by the selection algorithm optimally. An important advantage of this approach leads to reduced prosody mismatch and spectral discontinuity that occurs syllable concatenation. The results obtained from the proposed system are far superior as compared to the traditional unit based Text-To-Speech system. The most important output of this system is the improved naturalness and intelligible synthesized speech.*

Keywords : *Text-To-Speech synthesis, Pali Speech Database, Unit selection, Speech Analysis*

I. Introduction

Generating a human like sound by machine is the dream of many scientists from last centuries. Synthesizing human speech is difficult due to the complexity of human speech. The production of human speech involves the lungs, vocal fold and vocal tract (oral cavity and nasal cavity) functioning collectively [1] [2]. Text-to-Speech (TTS) is the process of converting unknown text into sound. It is an artificial production of human speech. A computer system used for this purpose is called a speech synthesizer and that is implemented in both software and hardware [3]. The Text-to-Speech synthesis procedure consists of two main phases. The first one is text analysis, where the input text is transcribed into a phonetic or some other linguistic representation and the second one is the generation of speech waveforms [4].

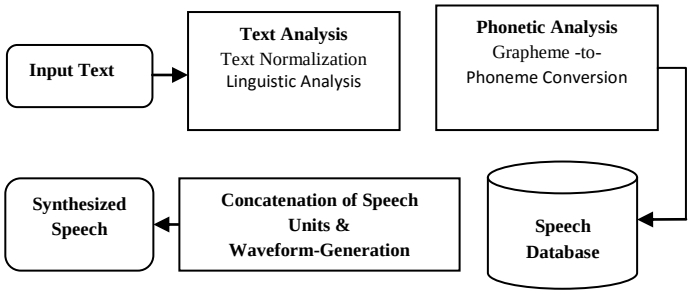

Fig.1 Block diagram of Text-to-Speech System

Techniques of speech synthesis [5].

a) Articulatory Synthesis
b) Formant Synthesis
c) Concatenative Synthesis

Unit selection based concatenative speech synthesis system has become popular now a day because of their highly natural sounding synthetic speech [6].

To develop a Text-to-Speech system for Pali language, we have chosen a unit selection based concatenative speech synthesis. The development process of Pali language TTS system consist of simplified phone set of Pali language. That includes vowels, consonant and syllables. The inputted text i.e. character string is then preprocessed and analyzed [4]. The speech signal is generated by concatenating prerecorded sound units. In this approach several fundamental periods of pre-recorded phonemes are simply concatenated. The phonemes are then connected to form words and sentences [7]. The detailed process is explained in section 3.

II. Previous work

In India 15 official languages are spoken in different forms across different places. From the point of view of TTS development, the most helpful aspect of Indian scripts is that they are basically phonetic in nature, and there is one-to-one correspondence between the written and spoken forms of most of the Indian languages. This makes the task of automatic phonetization simpler [8]. In our study we explore many speech synthesis methods, techniques, applications and products as possible in our investigation. The world's first mechanical speech synthesizer machine was developed by Gerbert [9] after that a continuous development was done. In present scenario sophisticated speech synthesizer is available in English language as compared in Indian languages where still a lot is to be done. The available work for Indian languages is discussed bellow.

Table No.1: Text to Speech systems in Indian languages.

Institute / Organization	**Covered Languages**	**Synthesis techniques**	**Database**	**Performance**
JNU Delhi	Sanskrit	Rule based	Database is designed in SQL server, phones and 2,00000 words	Unlimited TTS - more than average
C-DAC Mumbai	Marathi, Odia	Unit selection Concatenative Festival based speech synthesis	Units of phones, syllables and words	Unlimited TTS - more than average
IIIT Hyderabad	Bengali, Hindi, Kannada, Tamil, Malayalam, Marathi, Sanskrit	Letter to Sound rule (grapheme-to phoneme or Akshara-to-sound)	(1000 sentences of each language) ph – 47, 58, 51, 48, 57, 35, 51 sy – 866, 890, 851, 1191, 660, 930, 997 wd – 2285, 2145, 2125, 2077, 2097, 2182, 2310 ph – Phonemes, sy – Syllables, wd - Words	Unlimited TTS - more than average
TIFR Mumbai	Marathi, Hindi	Formant synthesis	Phones and words	Average
HCU Hyderabad	Telugu	Concatenative MBROLA based	Diphone	Average
CEERI, Delhi	Hindi Bengali (partly)	Formant Synthesis (Klatt – type)	Syllables & Phonemes (Parameter Data Base)	Excellent Unlimited TTS-Average
IIT, Chennai (Madras)	Hindi, Tamil	Concatenative diphone synthesis (1400 diphones)	Diphone Syllable (Mainly)	Unlimited TTS-Average
C-DAC Pune	Hindi, Indian English	Concatenative	Phonemes and recorded word	Limited TTS - more than average
C-DAC Noida	Hindi	Concatenative	Multi form units Diphones, Syllable, frequent words and phrases.	Domain specific, Excellent unlimited TTS Average
C-DAC Kolkata	Bangla	Concatenative	Phonemes and new set of signal units	Unlimited TTS Average
Utkal University, Bhubaneshwar	Oriya	Concatenative	Phones, Diphones & Triphones	Unlimited TTS Average
IISc Banglore (Dhvani)	Bengali, Gujrati, Hinhi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, Telugu.	Concatenative	Phomes (consonant, vowels & half consonant)	Unlimited TTS - more than average

TTS system in Indian Languages [9]-[23].

III. Methodology

3.1 Nature of the Pali Scripts

The name pali means lines or canonical text [24], basic units of writing system in Pali are characters which are an orthographic representation of speech sounds. Pali has been written in variety of scripts including Bramhi, Devanagari and other Indic scripts. Today Pali is studied mainly by those who wish to read original Buddhist scriptures. There are non-religious texts in Pali including historical and medical texts. A typical structure of Pali language script is in syllabic in nature the form of an Akshara are V, CV, CVC and CCCV

where C is consonant and V is vowel [25]. There is good correspondence between what is written and what is spoken.

3.2 Phone Set

An Akshara is an orthographic representation of a speech sound in Pali language. Pali language in Devnagari script is represented by 8 vowels and 32 consonants making a total of 40 alphabets as shown in fig. 1.

Vowels

अ आ इ ई उ ऊ ए ओ

Consonants

क	ख	ग	घ	ड.	च	छ
ज	झ	ञ	ट	ठ	ड	ढ
ण	त	थ	द	ध	न	प
फ	ब	भ	म	य	र	ल

व स ह ळ

Fig. No.2 Phone set for Pali language in Devnagari script [27].

3.3 Development of Speech Database

The purpose of building Pali speech synthesizer is to generate high quality synthetic speech. The database has been collected in order to investigate how well the Pali TTS system is used by humans in intelligible verbal communication [28]. In recent years speech synthesis technology has progressed remarkably because of large scale speech corpus. .

3.3.1 Text Corpus

It is important to have an optimal text corpus balanced in terms of phonetic coverage. The text corpus consist of phone set vowels and consonants, commonly used in daily words, body parts, days, months, names of worker role, animals, birds, relationships, colors, sessions, budha historical places, names of historical educational institutes, names of bhikhu-bhikhuni, budha grantha, short stories Tipitak and paragraphs and sentences from budha's ascent literature. The text corpus consists of total 676 sentences, 12359 total words and 3226 unique words. In these sentences we tried to cover the possible conversions related to our daily life.

3.3.2 Speech Recording

Data is collected from professional speakers, with very good voice quality. The quality of digital sound is determined by discrete parameters. The discrete parameters are the sample rate, bit capacity and number of channels [29]. One important conclusion that we can make at this point is that our digitization standards should be able to faithfully represent acoustic signals. The Linguistic Data Consortium for Indian Languages (LDC-IL) under Central Institute of Indian Languages has designed the standards for capturing the speech data according to application for which the speech data is collected and the devices that were used for recording the speech samples [30]. To record the speech samples, a process of speaker selection was carried out. A professional male speaker (Pali language expert) in the age group of 20-25 with very good voice quality was selected. The recording was done in noise free environment. The speech signals was recorded by using a standard headset microphone connected to the laptop. A 16 KHz sampling rate is often used for microphone speech. We have used PRAAT software tool to record the speech at 16 KHz sampling frequency and represented using 16 bits/sample [30].

3.3.3 Speech Processing and Labeling

The most important tasks in building speech database are the speech data segmentation and labeling. The quality of synthesized speech depends on the accuracy of the labeling process [31]. Manual labeling of speech data is still the most common form of labeling means to temporally define discrete names to them using symbols from an appropriately defined set corresponding to acoustic, physiological, phonetic or higher level linguistics terms [32]. In this work manual speech segmentation and labeling are very tedious and time consuming task and require much efforts. We record all the necessary speech data and then some processing are done to remove the noise of the recorded speech. Then normalize the speech by using audacity software tool. The processed speech was segmented using PRAAT tool and labeled the speech files as "k.wav".

3.3.4 Unit Selection and Speech Generation

One approach to the generation of natural sounding synthesized speech is to select and concatenate recorded speech units from a large speech database [33][34]. The task of unit selection procedure is to find the most appropriate recorded speech units in the corpus. Input data received from language processing modules in the speech synthesizer are sequence of phonemes to be pronounced, whereby prosodic parameters for pronunciation of each phoneme are provided. These parameters contain data on the fundamental frequency and duration of the phoneme pronunciation [35]. We propose that the units in a synthesis database can be considered as a state transition network. For a given textual input which is mapped into present database, the unit selection algorithm concatenates the corresponding wave files sequentially from left to right. The developed algorithms detects corresponding audio file present in the speech database and concatenates them.

The algorithm first checks weather the user has entered a word or a sentence by detecting the number of spaces present in the entered text. The inputted text is split into words, syllables and phonemes. A search is then made to find a best i.e. first priority of search for word into database. If exact word is found in database, then it fetches the corresponding wave file in the database. If not then split the word into syllables. Search syllables into database, if syllables are not found in database split into phonemes and select corresponding wave files stringing together (concatenating) all wave files and generate the speech.

Following figure shows searching of the optimal sequence of recorded speech units for target Pali word पालिभासा

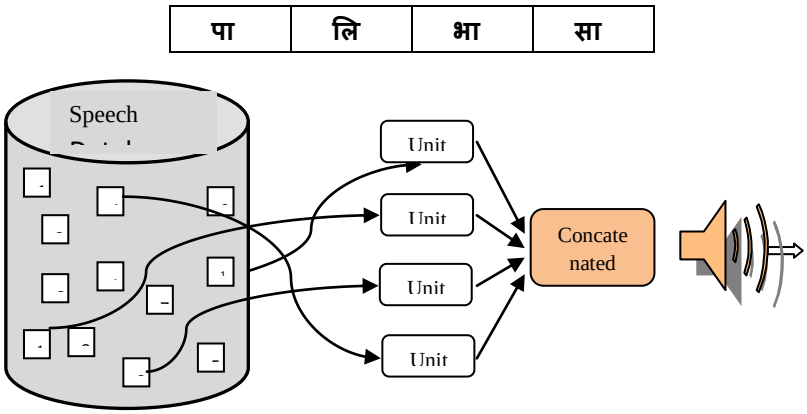

Fig.3 Searching of Speech Units

IV. Results and Discussion

The result of this work is evaluated to check the naturalness of the synthesized speech. The GUI shown in Fig. No. 4 where speech is synthesized by exact or complete words and words formed by concatenating syllables and phones

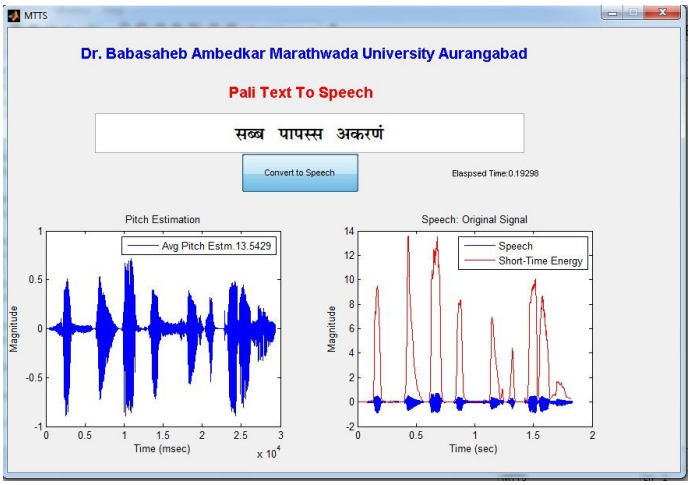

Fig. 4 Snapshot of Pali Text To Speech Synthesizer GUI

The speech waveform and spectrogram plot for the Pali phrase सब्ब पापस्स अकारण is shown in following figure.

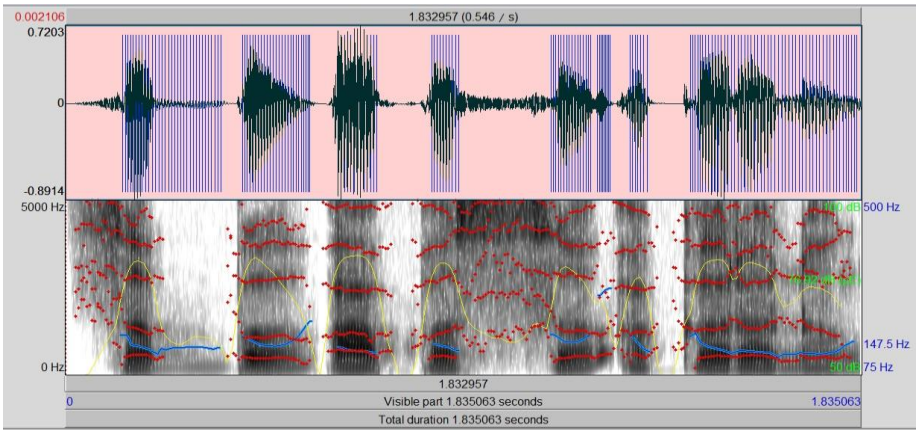

Fig. 5 Spectrogram and wave form of synthesized speech

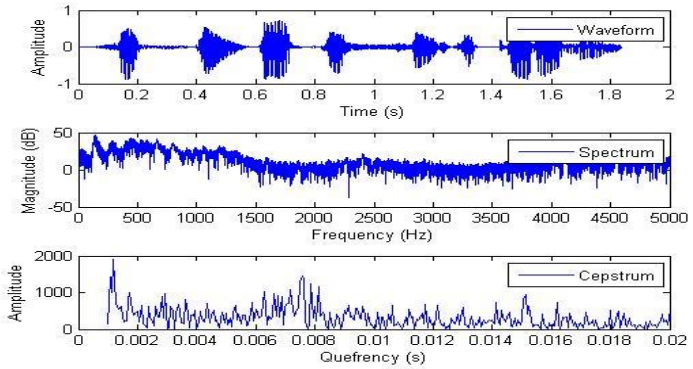

Fig.6 Waveform, Spectrum and Cepstrum

The spectrogram shows that the formant changes are not abrupt at the concatenation points. The average energy of synthetic speech was calculated by short time speech measurement. This measurement can distinguish between voiced and unvoiced speech segments, since unvoiced speech has significantly smaller short time energy. The energy of the frames is calculated using the relation

$$E_n = \sum_{m=inf}^{m=inf} [x(m)w(n-m)]^2 \qquad \ldots(1)$$

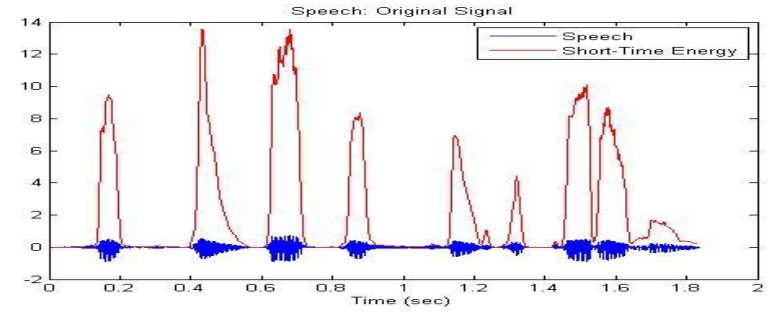

Fig. No.7 Short time energy of synthetic speech

Moreover, spectral changes are uniform across the syllable boundaries and hence reinforce the idea that the syllable like unit is indeed a good candidate for concatenative speech synthesis.

The results of Pali language vowels, consonants and words

Table No.2 Statistical results of vowels

Sr. No.	Pali Vowels	Pitch of Uttr_1	Pitch of Uttr_2	Pitch of Uttr_3	Pitch of Uttr_4	Pitch of Uttr_5	Mean	SD
1	अ	163.187	158.791	156.19	151.48	153.15	156.56	4.65
2	आ	148.307	158.872	151.669	150.75	149.5	151.82	4.14
3	इ	155.838	174.455	153.526	152.16	150.69	157.33	9.76
4	ई	148.4	162.934	150.967	151	155.28	153.72	5.71
5	उ	158.299	158.193	151.534	151.29	149.2	153.70	4.24
6	ऊ	153.286	163.536	161.836	153.9	152.81	157.07	5.17
7	ए	151.477	162.311	148.062	155.38	149.37	153.32	5.74
8	ओ	138.717	146.605	140.412	150.83	130.45	141.40	7.81

Table No.2 Statistical results of consonants

Sr. No.	Pali Consonants	Pitch of Uttr_1	Pitch of Uttr_2	Pitch of Uttr_3	Pitch of Uttr_4	Pitch of Uttr_5	Mean	SD
1	क	155.73	158.36	154.467	160.54	153.46	156.51	2.90
2	ख	152.2	153.48	146.706	164.05	151.75	153.64	6.37
3	ग	149.22	156.33	150.324	150.32	153.74	151.99	2.96
4	घ	143.61	146.45	143.093	157.79	148.07	147.80	5.94
5	च	157.67	159.19	147.87	149.66	155.49	153.97	4.97
6	छ	150.94	157.8	156.442	151.87	148.63	153.14	3.86
7	ज	151.52	154.22	153.343	153.5	155.91	153.70	1.59
8	झ	143.77	153.59	152.89	148.35	146.79	149.08	4.15
9	ट	146.22	162.16	154.159	154.35	157.81	154.94	5.86
10	ठ	135.69	143.47	137.912	139.2	138.02	138.86	2.87

Table No.3 Statistical results of words

Sr. No.	Word	Pitch of Uttr_1	Pitch of Uttr_2	Pitch of Uttr_3	Pitch of Uttr_4	Pitch of Uttr_5	Mean	SD	Variance
1	वानरीदो	150.368	149.133	147.829	147.126	148.142	148.52	1.26	2.52
2	मक्क्ट	159.577	157.515	156.735	153.542	150.918	155.66	3.42	6.85
3	मयुरो	150.153	147.311	149.171	143.63	150.067	148.07	2.73	5.46
4	गरुड	134.67	136.701	144.365	139.919	134.041	137.94	4.26	8.52
5	सुवो	157.743	157.122	155.488	158.264	148.427	155.41	4.04	8.08
6	सारिका	153.045	151.671	152.847	150.664	151.761	152	0.97	1.94
7	कोकिलो	162.627	150.858	157.704	156.924	149.426	155.51	5.39	10.78
8	किपिलो	145.234	165.986	167.629	161.007	155.162	148.50	9.1	18.21
9	काक	150.466	149.372	151.96	149.203	147.434	154.89	1.67	11.29
10	कुकुटो	165.191	159.453	161.212	161.175	158.799	161.17	2.49	4.98

The overall performance of the system is computed by calculating the percentage of correct phonemes (i.e. consonants and vowels), 1 to 100 digits, short words and connected words calculated on 100 random unknown words (i.e. words that are not in the database except connected words) of Pali language.

Table No.4 Test Results

Sr. No.	Type of Data	Accuracy (%)
1	Vowels	100 %
2	Consonants	100 %
3	Syllables	100 %
4	Digits (1 – 100)	100 %
5	Short words	71 %
6	Connected words	42 %

V. Conclusion

The developed Text-To-Speech system we observed that with initial experiments showing that word unit performs better than the phoneme and syllable units. In this experiment speech units picked by the selection algorithm optimally, to produce a natural sounding synthetic speech. This speech synthesizer is capable of generating natural sounding synthesized speech with no prosodic modeling.

It was observed that when the database of small speech units, the speech synthesizer is likely to produce a low quality speech. As the database of units increases, it increases the quality of the synthesizer. Creation of maximum coverage of units for concatenation synthesis gives greatest naturalness. This system can generate natural and intelligible synthesized speech for Pali language. Overall these test in table number 4 shows that the accuracy of the TTS system is 85 %. Here we conclude that the Text to Speech conversion accuracy is good.

Future scope- The current system is not meant to read words those are formed by combination of Half Symbol + Full Symbol (i.e. Jodakshar) e.g. बुध्द, राठ्ह, बलीवद्द. Thus in future development of this system such words will allow be included to make full-fledged system for Pali language.

References

[1]. Archana Balyan, S.S. Agrwal and Amita Dev, Speech Synthesis: Review, International Journal of Engineering Research and Technology, ISSN 2278-0181 Vol. (2), 2013, 57 – 75.

[2]. Jonathan Williams, Prosody in Text-to-Speech Synthesis Using Fuzzy Logic, M.S. diss., West Virginia University, USA, 2007.

[3]. D.D. Pande, M. Praveen Kumar, A Smart Device for People with Disabilities using ARM7, International Journal of Engineering Research and Technology, ISSN 2278-0181 Vol.(3), 2014, 614 – 618.

[4]. Sami Lemmetty, Review of Speech Synthesis Technology, M.S. diss., Helsinki University of Technology, Finland, 1999.

[5]. Mark Tatham and Katherine Morton, Developments in Speech Synthesis (John Wiley & Sons, Ltd. ISBN: 0-470-85538-X, 2005)

[6]. Jithendra Vepa and Simon King, Subjective Evaluation of Join Cost and Smoothing Methods for Unit Selection Speech Synthesis, IEEE Transactions on Audio, Speech and Language Processing, Vol. (14), 2006, 1763 – 1771.

[7]. Baiju Mahananda, C.M.S. Raju, Prahallad Kishore et. al., Building a Prototype Text-to-Speech for Sanskrit, Proc. 4th International Sanskrit Computational Linguistic Symposium, New Delhi, 2010, 39-47.

[8]. Krishna N Sridhar, Murthy Hema A, Gonsalves Timothy A., Text-to-speech in Indian languages, Proc. International Conference on Natural Language Processing, ICON-2002, Mumbai, 2002, 317-26.

[9]. Mattingly, I.G., Speech Synthesis for Phonetic and Phonological Models, T.A. Sebeok (Ed.) Current Trends in Linguistics, Linguistics and Adjacent Arts and Sciences, Vol.(12), 1974, 2451-2487.

[10]. Computational Linguistics R&D Special Centre for Sanskrit Studies JNU, Delhi, Online available http://sanskrit.jnu.ac.in/samvacaka/index.jsp (2016).

[11]. Girish N. Jha, Sudhir K. Mishra, R. Chandrashekhar et. al., Developing a Sanskrit Analysis System for Machine Translation, Proc. National Seminar on Translation Today: state and issues, Dept. of Linguistics, University of Kerala, Trivandrum, 2005, 23-25.

[12]. Development of TEXT to SPEECH SYSTEM in Indian Languages Phase-II, Online available http://cdac.in/index.aspx?id=mcst_speech_tchnology, 2016.

[13]. Kishore Prahallad, E. Naresh Kumar, Venkatesh Keri, S. Rajendran and Alan W Black, The IIIT-H Indic Speech Databases, Proc. of Interspeech Portland, Oregon, USA, 2012.

[14]. A. Sen and K. Samudravijaya, Indian Accent Text to Speech System for Web Browsing, International Journal Sadhana, Vol. (27), February 2002, 113-126.

[15]. Spoken Language Processing, Online available http:// http://speech.tifr.res.in 2016.

[16]. S.S. Agrwal, Rajesh Verma, Shailendra Nigam, Anuradha Sengar, Development of Rules for Unlimited Text To Specch Synthesis of Hindi, Proc. Conf. on ICA, Rome, 2001, 2-4.

[17]. Speech and Vision Laboratory IIT Madras, Online available http://www.cse.iitm.ac.in/speech, 2016.

[18]. Ramani B, S.L. Christina, G.A. Rachel, S. Solomi V, M.K. Nandwana, A. Prakash, A Common Attribute based Unified HTS framework for Speech Synthesis in Indian Languages, Proc. 8th ISCA Conf. on Speech Synthesis Workshop, Barcelona, Spain, 2013, 291-296.

[19]. Hindi Text-to-Speech (TTS) System C-DAC, Online available http://cdac.in/index.aspx?id=mlc_gist_speechtech, 2016.

[20]. ESNOLA based Bangla TTS C-DAC, Online available http://cdac.in/index.aspx?id=mc_st_TTS_Bangla, 2016.

[21]. Sanghamitra Mohanti, Syllable Based Indian Language Text To Speech System, International Journal of Advances in Engineering & Technology ISSN: 2231-1963, May 2011, 138-143.

[22]. Preranasri Mali, A Survey on Text To Speech Translation of Multi Language, International Journal of Research In Advanced Engineering Technologies ISSN: 2347-2812, 2014, 29-31.

[23]. Dhvani a Text to Speech System for Indian Languages, Online available http://dhvani.sourceforge.net, 2016.

[24]. Omniglot the Online Encyclopedia of writing Systems and Languages, Online available http://www.omniglot.com/writing/pali.htm, 2016.

[25]. Baiju Mahananda, C.M.S. Raju, Ramalinga Reddy Patil, Narayana Jha, et. al., Building a Prototype Text to Speech for Sanskrit, Proc. 4th International Conf. on Sanskrit Computational Linguistics Symposium, New Delhi, India, 2010, 39 – 47.

[26]. Branko Markovic, Slobodant Jovicic, Jovan Galic, Dorde Grozdic, Whispered Speech Database: Design, Processing and Application, Proc. 16th International Conf. on Text, Speech and Dialogue, Springer ISBN 978-3-642-40585-3-74, 2013, 591 – 598.

[27]. Susila Muljadhav, Pali Bhasha Parichya *पाली भाषा परिचय*, (Ramai Publication, 2015.)

[28]. Hiromichi Kawanami, Tsuyoshi Masuda, Tomoki Toda and Kiyohiro Shikano, Designing Speech Database with Prosodic Variety for Expressive TTS System, Proc. IREC, Japan, 2002, 2039-2042.

[29]. Primer on Audio PC various issues associated with PC based audio technology, Online available on http://www.totalrecorder.com/primerpc.htm, 2016.

[30]. Standards for Speech Data Capturing and Annotation, available online http://www.ldcil.org/download/samplingstandards.pdf, 2016.

[31]. Marian Boldea, Cosmin Munteanu, Labeling A Romanian Speech Database, Proc. Second International workshop on Speech and Computer, Napoca Romania 1997.

[32]. Barry, W.J. Fourcin, A.J., Levels of Labeling Computer Speech and Language, Proc. Abstract Computer Science, Australian National University, Australia 1992.
[33]. Hunt A.J.,Unit selection in a concatenative speech synthesis system using a large speech database, Proc. International Conference on ICASSP-96, Acoustic, Speech and Signal Processing IEEE Vol.(1), 1996, 373-376.
[34]. Andrew J. Hunt and Alan W. Black, Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database, IEEE Transactions on Audio, Speech and Language Processing, Vol. (3), 1996, 373-376.
[35]. Jerneja Gros and Mario Zganec, An Efficient Unit-selection Method for Concatenative Text-to-speech Synthesis Systems, International Journal of Computing and Information Technology, 2008, 69-78.