Text to Speech Synthesis of Hindi Language using Polysyllable Units

Aparna N S
PG Scholar
Department of E&C, SJCE, Mysuru

Shreekanth T
Assistant Professor
Department of E&C, SJCE,Mysuru

ABSTRACT

A Text To Speech (TTS) synthesis is a computer based system that should be able to read any text aloud. Thus TTS technology is essential to those people who are visually impaired. It also plays a very important role in the field of Telecommunication, Industrial and educational applications. Thus TTS has been developed for foreign languages and is well established. As Indian language characters are complex in nature, it is not a straight forward approach to build the TTS system for Indian languages as compared to English. India is a country of multi languages among them Hindi is one of the 23 official language. Hence this paper discusses development of Hindi TTS system. Syllable units in Hindi language are better choice than any other units because each character in Hindi language is close to syllable which is in the form of CVC (C: consonant, V: vowel).Existing Hindi TTS can be done using phone and diphone. The disadvantage in the existing system is that it requires larger concatenation points and has low quality speech output. It is observed that the quality of the synthesized sentences can be improved by using polysyllable units. In the proposed system, the developed database consists of more than 25,000 bisyllable and 1,200 syllables considering 3 positions of syllable in a word i.e. start, middle and end. The obtained results were compared with monosyllable based TTS, it indicated that the naturalness and intelligible of speech output is high compared to monosyllable based TTS system.

Keywords

Hindi TTS, polysyllable unit, Concatenative speech synthesis, MOS (Mean opinion score)

1. INTRODUCTION

The objective of a text to speech system is to convert an arbitrary given text into a spoken waveform [1]. The general block diagram of TTS is shown in Figure 1, this generally involves two steps, (i) Text processing and (ii) Speech generation.

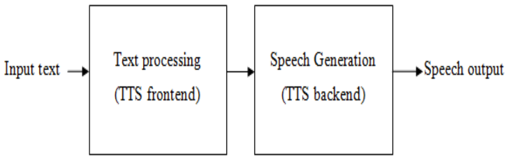


Figure 1: Block diagram of TTS

Text processing is used to convert the given text to a sequence of synthesis units. Speech generation is generation of an acoustic waveform corresponding to each of these units in the sequence. Synthesized speech can be developed by using different methods, mainly there are three methods are available.

1.1 Articulatory synthesis

Articulatory synthesis is a method in which the human vocal organs are modeled. The name of the method is inspired by the term ‘Articulators’ which implies speech organs like jaw, tongue, lips etc. This system contains physical models of both the human vocal tract and the physiology of the vocal cords. It transforms a vector of anatomic or physiologic parameters into a speech signal with predefined acoustic properties. The implementation of such a system is very difficult and therefore it is not widely in use yet.

1.2 Formant synthesis

Formant synthesis does not use human speech samples at runtime. Instead, the synthesized speech output is created using additive synthesis and an acoustic model. Parameters such as fundamental frequency, voicing, and noise levels are varied over time, to create a waveform of artificial speech. This method is sometimes called rules-based synthesis. The simplifications made in the modeling of the source signal and vocal tract inevitably lead to somewhat unnatural sounding result.

1.3 Concatenative synthesis

Concatenative synthesis is based on the concatenation of segments of recorded speech. Generally, it produces the most natural sounding synthesized speech. For a given text, these segments are joined based on some joining rules. This method is easy to produce intelligible and natural sounding synthetic speech. There are 3 vital subtypes of Concatenative synthesis, there are follows below**.**

1.3.1 Domain specific synthesis

Domain specific synthesis normally concatenates words or phrases of speech and can be used when the output of the synthesis system is limited to a small domain of utterances.

1.3.2 Diphone synthesis

Uses a minimal speech database containing all the Diphones (sound-to-sound transitions) occurring in a given language [9]. In diphone synthesis, only one example of each diphone is contained in the speech database [12].

1.3.3 Unit selection synthesis

This is the dominant synthesis technique in text to speech, uses large speech databases (more than one hour of recorded speech).The natural sounding speech is the main advantage this technique hence it is more popular Several research works have been carried out on Text to speech synthesis for Hindi language. Hindi synthesis is built by using different choices of unit size like syllable, diphone, phone and half phone. Phoneme is the smallest unit of speech. Database built using phoneme level requires a lot of memory storage and more time for speech synthesis, it is not preferred in Concatenative synthesis as it lacks co-articulation at concatenation points. To overcome this problem diphone was used as a basic unit. Diphone is the set of 2 phonemes. The issue with phonemes and diphones is that their usage will have more number of concatenation points for the given text. In case of generating polysyllables this has inherited the problem of concatenation junctions and time delay in concatenating larger units. To overcome these problems we should limit the number of concatenation points required for speech synthesis. This is achieved only by considering syllable as a basic unit of concatenation. Mainly Indian languages are syllabic in nature it is a combination of consonant (c) and vowel (v). This above survey motivated us to built Hindi TTS using polysyllable units, that contains cluster units of more than one type (monosyllable, bisyllable and trisyllable) and Concatinative technique is used to develop this system.[2,3,4].

2. GENERAL ARCHITECTURE OF TTS SYSTEM

General Architecture of text to speech system is shown in Figure 2.Text analysis is preprocessing parts which analyzes the input text and organizes into manageable list of words [5.6]. It contains of numbers, symbols, abbreviations are replaced by their corresponding whole words. Text detection is localizing the text area from any kind of printed documents. Phonetic analysis converts the orthographical symbols into phonological ones using a phonetic alphabet, it is also known as grapheme to phoneme conversion. Prosody is a concept that contains the rhythm of speech, stress patterns and intonation. Prosody plays a very important role in the understandability of speech and its features carry lots of information about the speaker, for example his or her emotional state, and even the social background. Speech synthesis finally generates the speech signal. The speech will be spoken according to the voice characteristics of a person, there are 3 types of speech synthesis techniques and these are briefly explained in section 1. In the proposed work Concatenative speech synthesis method is choosen for Hindi TTS system [7, 8, 11, 12].

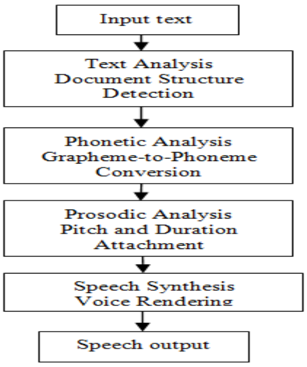


Figure 2: General Architecture of TTS

3. HINDI SCRIPT

Hindi text is written in the Devanagari script. The alphabets used in the Devanagari script are scientific and well organized. They are divided into 2 groups, (i) Vowels and (ii) Consonants. The letters which represent a simple vocal sound are called vowels (Svara), it have 2 forms, the independent form and dependent form. The independent form vowels are 'stand alone'. The dependent form vowels are always attached to consonant. The letters which can be sounded only with a vowel are called consonants (Vyanjan). In Hindi language there are 13 vowels and 33 consonants. Each and every language in India can be represented using English and known as transliteration. Transliteration of all local languages in English will help us to obtain an easy way to develop new applications for illiterates and visually impaired community. The main advantage of transliteration is that, it helps us to translate any length word or sentence of local languages to English. This guides us to obtain clear idea about how the breakup of a syllable can be done after transliteration using rules. Hindi Devanagari characters with English transliteration as shown in below Table 1.

Example1: Hindi word: कबूतर
Transliteration: ka/boo/ta/ra
Syllable breakup: cv/cv/cv/cv

Table 1. Hindi Devanagari characters with English transliteration

अ	आ	इ	ई	ऋ	उ	ऊ	ए	ऐ	ओ	औ	अं	अः
a	ā	i	ī	r̥	u	ū	e	ai	o	ou	aṃ	aḥ

क ka	ख kha	ग ga	घ gha	ङ ṅa					
च ca	छ cha	ज ja	झ jha	ञ ña					
ट ṭa	ठ ṭha	ड ḍa	ढ ḍha	ण ṇa					
त ta	थ tha	द da	ध dha	न na					
प pa	फ pha	ब ba	भ bha	म ma					
य ya	र ra	ल la	व va	श śa	ष ṣa	स sa	ह ha		
क्ष kṣa	त्र tra	ज्ञ jña/gya	क़ qa	ख़ kẖa	ग़ g̱a	ज़ za	फ़ fa	ड़ ṛa	ढ़ ṛha

From above Example 1 we can clearly say that Hindi language is having one to one correspondence with written and spoken form. Transliteration of local language to English will direct us to use syllable rules developed for speech database building for TTS system.

4. SOFTWARE IMPLEMENTATION

Developing text to speech is not a single step several Steps are involved in the building of TTS system which are as follows, the flow chart of proposed TTS system as shown in Figure 3. The first and the most important step involved in this is to build speech database by using PRAAT tool. Mainly 3 steps are involved in building speech database.

i. First records the bisyllable based Hindi words by using standard microphone and choose one person for recording these basic units, who has uniform characteristics of speaking.

ii. Moving on, Segment the recorded words by creating text grid.

iii. Finally label the segmented words and save it as WAV file.

ALGORITHM

Steps involved in the algorithm are as follows

1. Enter the text using windows mapped keyboard as the standard reference in input textbox.
2. Print the Hindi output in output textbox.
3. TTS read character by character for proper pronunciation and Check whether the entered text as a vowel or a virama or a visarga or consonant.
4. If it is a vowel, then append appropriate padding Unicode digit to previous consonant.
5. If no, check for anusvara and if it is then print append appropriate padding digit on the group.
6. If the read character belongs to consonants group then algorithm checks whether it is a independent or dependent vowel group. The read character is independent it directly map their corresponding UNICODE. If it is a dependent vowel sign then Unicode of that consonant is padded with corresponding two digit values.
7. If the word length has 4 letters then it picks the bisyllable units as first. If the word length has 3 letters then the algorithm appropriately chooses the first as bisyllable then phoneme.
8. Check for the blank space, if yes, delete the previous entered character and make changes to the generated code.
9. Check for space, if space is available then keep tracks of Unicode changes of the corresponding text.
10. Print the result and go to Step2.

For example if the entered word is “कटपुतलि” the modified UNICODE output for each letter is क-2325 ट-2335 पु -234604 त-2340 लि-235403 the presence of spaces between each Unicode helps us to differentiate the individual characters in the entered words.

TTS system first reads the word as character by character for the proper pronunciation. Then TTS system picks up the first matching word as the bisyllable unit followed by the monosyllable then phoneme unit from the database. This can be understood from the above given example “कटपुतलि” as “कट” (2535), “पुत” (460440), followed by the monosyllable unit “लि” (5403).

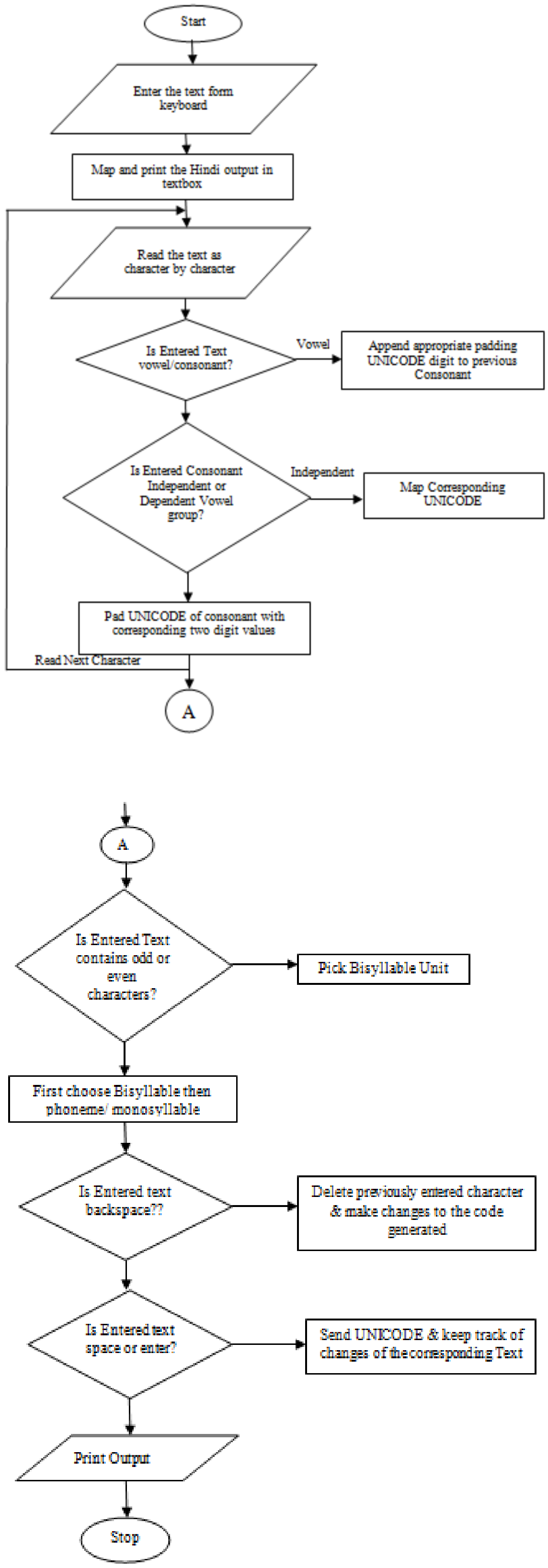

Start
Enter the text form keyboard
Map and print the Hindi output in textbox
Read the text as character by character
Is Entered Text vowel/consonant?
Vowel
Append appropriate padding UNICODE digit to previous Consonant
Is Entered Consonant Independent or Dependent Vowel group?
Independent
Map Corresponding UNICODE
Pad UNICODE of consonant with corresponding two digit values
Read Next Character
A
A
Is Entered Text contains odd or even characters?
Pick Bisyllable Unit
First choose Bisyllable then phoneme/ monosyllable
Is Entered text backspace??
Delete previously entered character & make changes to the code generated
Is Entered text space or enter?
Send UNICODE & keep track of changes of the corresponding Text
Print Output
Stop

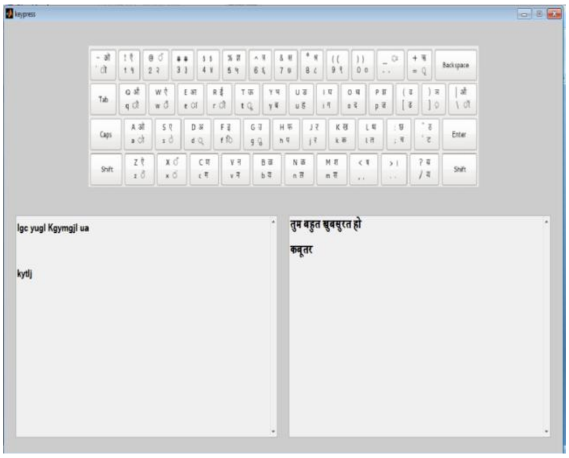
Figure 3: Flow chart of proposed TTS system

11. RESULT AND DISCUSSION

This section discusses about the results of the proposed work in terms of the quality of the synthesized speech.TTS is developed on 2 sets of unit i.e. Syllable (DB1) and polysyllable (DB2). To verify the naturalness of the synthesized speech obtained from these two databases Mean Opinion Score (MOS) is taken from various listeners. We have created an executable file Graphical User Interface (GUI) for the whole project using MATLAB. Once we start the GUI, the user is allowed to enter the data in the left text field based on the reference keypad at the top. The entered text is then converted into Hindi text along with the speech output, the screen shot of the GUI with speech output is shown in Figure 4.From Figure 4, we can clearly see the Hindi input entered by user on right end of text box and its corresponding English transliteration on the left side of text box. Once we stop entering the input, the program automatically creates a text file for Unicode. It converts the decimal value and writes the displayed values into the text files specified then it saves it in the path provided. The path can be provided any where depending on the path of executable file. Mean Opinion Score is the arithmetic mean of all the individual scores which gives the numerical indication of the perceived audio quality. A listeners test was conducted to check the naturalness of the concatenated speech. In order to compare the results, the algorithm is tested on both the databases DB1 and DB2.The MOS that is expressed as a single number in the range 1 to 5, where 1 is lowest perceived quality and 5 is the highest perceived quality. The MOS is generated by averaging the results of a set of standard, subjective tests where a number of listeners rate the perceived audio quality of test words read aloud by the female speakers over the communications medium being tested. A listener is required to give each word a rating using the rating scheme in Table 2. The perceptual score of the method MOS is calculated by taking the mean of the all scores of each word [31].

Table 2. Parameter for Mean Opinion Score

Mos	Quality	Distortion
5	Excellent	Imperceptible
4	Good	Slightly imperceptible
3	Fair	Slightly Annoying
2	Poor	Annoying
1	Bad	Very Annoying


lgc yugl Kgymgjl ua

kytlj

तुम बहुत खुबसुरत हो

कबूतर

In order to test the improvement of synthesized speech quality using syllable-like units a perceptual evaluation of 15 Hindi words that are provided in Table 3, are used for tests and 10 native Hindi speakers employed in the evaluation for both the database DB1 and DB2.MOS range and score for each word in both the database DB1 and DB2 are shown in Table 4 and 5.

Table 3. Test words

Word	Hindi
1	कबूतर
2	कथानक
3	विवेचन
4	कदन
5	परकीय
6	कमल
7	कटपुतलि
8	अगर
9	अचनक
10	समिति
11	कल्पन
12	सोपहार
13	कचहरि
14	बेगम
15	चमक

Table 4.MOS Range and score for each word in syllable based TTS (DB1)

W.no	Excellent 5	Good 4	Fair 3	Poor 2	Bad 1	No. of listeners	Obtained marks	Avg MOS
1	3	3	2	2	-	10	37	3.7
2	3	3	4	-	-	10	39	3.9
3	4	3	2	1	-	10	40	4.0
4	2	4	3	1	-	10	37	3.7
5	3	3	3	1	-	10	38	3.8
6	4	4	2	-	-	10	42	4.2
7	3	4	3	-	-	10	40	4.0
8	3	4	3	-	-	10	40	4.0
9	4	3	3	-	-	10	41	4.1
10	3	3	3	1	-	10	38	3.8
11	4	3	2	1	-	10	40	4.0
12	3	4	3	-	-	10	40	4.0
13	3	3	3	1	-	10	38	3.8
14	3	4	2	1	-	10	39	3.9
15	4	3	3	-	-	10	41	4.1
Sum	49	51	41	9		150	472	3.94

The concatenated speech output of the word "कबूतर "from the database DB1 is shown in Figure 5.

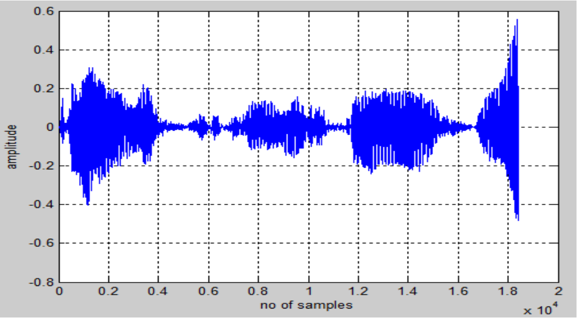


Figure 5: कबूतर concatenated speech sound from DB1

Table 5.MOS Range and score for each word in polysyllable based TTS (DB2)

W.no	Excellent 5	Good 4	Fair 3	Poor 2	Bad 1	No. of listeners	Obtained marks	Avg MOS
1	3	4	3	-	-	10	40	4.1
2	4	4	2	-	-	10	42	4.2
3	4	3	3	-	-	10	41	4.1
4	3	4	3	-	-	10	40	4.0
5	5	4	1	-	-	10	44	4.4
6	4	4	2	-	-	10	42	4.2
7	4	3	3	-	-	10	41	4.1
8	4	4	2	-	-	10	42	4.2
9	4	3	3	-	-	10	41	4.1
10	3	4	2	1	-	10	39	3.9
11	4	3	2	1	-	10	40	4.0
12	4	4	2	-	-	10	42	4.2
13	3	4	3	-	-	10	40	4.0
14	3	4	3	-	-	10	40	4.0
15	4	4	2	-	-	10	42	4.2
Sum	56	56	36	2		150	616	4.11

The concatenated speech output of the word "**कबूतर "**from the database DB1 is shown in Figure 6**.**

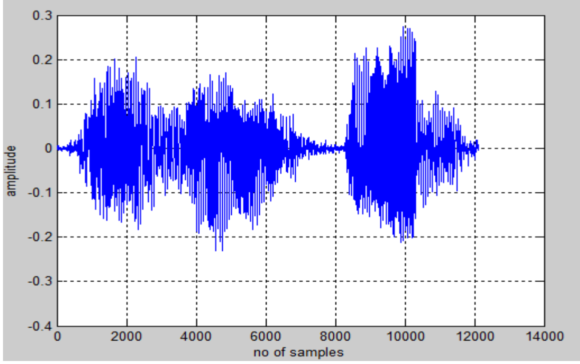


Figure 6: कबूतर concatenated speech sound from DB2

The average MOS values for both DB1 and DB2 are shown in Figure 7.

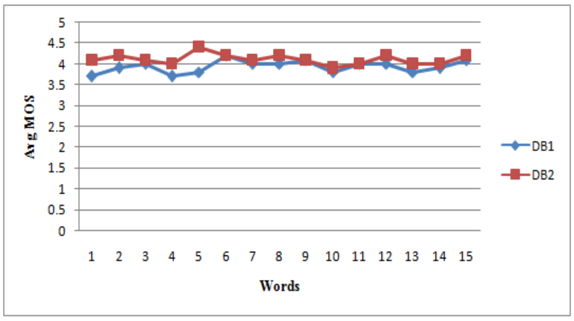


Figure 7: Avg MOS for both DB1 and DB2

From the Mean Opinion score it is observed that the quality of the proposed TTS system (DB2) is better than compared to the syllable based TTS system (DB1).

12. CONCLUSION

A Concatinative based polysyllable TTS has been developed for Hindi language. These results are compared with the monosyllable based TTS system. From the experimental results, we obtained the average MOS for DB1 as 3.94 and for DB2 as 4.11. Hence it is observed that the quality of speech will be improved if polysyllable units are used. The concatenation based speech synthesis system is developed using MATLAB programming, where it uses text processed decimal code as input. This concatenation algorithm is developed and implemented with the developed database and results were simulated.

6. ACKNOWLEDGEMENT

I would profound gratitude, immense pleasure and heartfelt thanks to our guide Mr. Shreekanth T, Assistant professor, Department of Electronics and Communication, SJCE, Mysuru for involvement, meticulous guidance, and constant encouragement throughout the project.

7. REFERENCES

[1] Sanghamitra Mohanty,"Syllable based indian languagetext to speech system ", International Journal of Advances in Engineering & Technology,Vol. 1, ISSN: 2231-1963,May 2011.

[2] D.J.Ravi, Sudarshan Patilkulkarni, "Evaluation of kannada text-to-speech [ktts] system", ISSN: 2277 128X, Volume 2, Issue 1, January 2012.

[3] Kiruthiga S and Krishnamoorthy K, "Annotating Speech Corpus for Prosody Modeling in Indian Language Text to Speech Systems", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January 2012.

[4] S P Kishore and Alan W Black ,"Unit Size in Unit Selection Speech Synthesis ",1317- 1320 EUROSPEECH – GENEVA, pg 1317-1320, 2003.

[5] Laba Kr Thakuria,Prof. P.H Talukadar "Text to speech synthesis of GALO and ADI languages using polysyllabic units", ,ISSN 2347-7393,2014.

[6] D.Sasirekha, E.Chandra "Text to speech: a simple tutorial", ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[7] Shruti Gupta,ParteekKumar,"Comparative study of text to speech system for Indian language" ,International journal of advances in and information technology ISSN 2277–9140 April 2012.

[8] Marian Macchi, Bellcore. “Issues in text-to-speech synthesis”, In Proc. IEEE International Joint Symposia on Intelligence and Systems,pp. 318-325,1998.

[9] Shreekanth.T , Udayashankara.V ,Arun Kumar.C “An Unit Selection based Hindi Text To Speech Synthesis System Using Syllable as a Basic Unit” IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 4, Issue 4, Ver. II Jul-Aug. 2014.

[10] Rajeswari K C and Uma Maheswari P. “Prosody Modeling Techniques for Text-to Speech Synthesis Systems”.International Journal of Computer Applications 39(16):8-11, February 2012.

[11] Rahul sawant, H.G Virani, Chetan desai ,”Personalized English speech synthesizer using Concatenative synthesis “,International journal of pure and applied research in engineering and technology .Research Article, IJPRET, 2013; Volume 1(8):260-267.

[12] B. Sudhakar R. Bensraj Development of Concatenative Syllable based Text to Speech Synthesis System for Tamil International Journal of Computer Applications (0975 – 8887 Volume 91 – No 5, April 2014.