Development of TTS for Marathi Speech Signal Based on Prosody and Concatenation Approach

[1]Surendra P. Ramteke, [2]Gunjal Oza, [3]Nilima P. Patil

[1,2]Department of E&TC Engineering, [3]Department of Computer Engineering
SSBT's College of Engineering & Technology Bambhori, Jalgaon (M.S.), INDIA

Abstract

Text to Speech conversion for natural language sentences can overcome obstacles of human computer interaction on written text and images for visually handicapped, physically impaired, educationally under privileged and the rural communities. Today, the quality of synthesized speech is not equivalent to the quality of real speech. Most research on speech synthesis focuses on improving the quality of the speech produced by Text-to-Speech (TTS) systems to meet intelligibility and naturalness. A general review of different speech processing techniques to improve speech quality is necessary before using a particular method. This paper presents an implementation of TTS for Marathi Speech Signals using Concatenation and Prosody approach. Unit selection method is used along with MFCC and HMM for Feature extraction and training of Database. Happy and Sad emotions are considered for prosody implementation.

Keywords: Speech synthesis, concatenation, Phoneme, Prosody, consonants & Vowels, Unit selection.

1. Introduction

Recently we find a rush in the development of spoken language systems that facilitate the human computer interaction. The objective of TTS is to provide textual information to people via voice messages.

A high degree of naturalness and intelligibility in the generated response is a significant factor of such a human-computer speech interface. Also it makes the use of computers possible for visually and physically impaired and illiterate masses, the educationally under privileged and the rural communities of India. It is also enviable that human-machine interface permits one's native language of communication.

Natural language processing (NLP) is a discipline that aims to build computer systems that will be able to analyze, understand and generate human speech. Sub Areas of research for NLP are Speech Synthesis and Speech Recognition. Speech synthesis is the process of converting the text into spoken language and Speech recognition is the process of converting spoken language to written text or some similar form.

In speech synthesis area, the use of corpus based concatenation techniques has been gaining popularity [1-6], due to its ability to achieve a high degree of naturalness and great quality natural speech. This paper reports an endeavour in designing a process for generating speech for Marathi language, by means of concatenating words, syllables, consonants and vowels. Marathi is the language spoken by the native people of Maharashtra.

In any language to cover expressive or emotional speech is one of these new relevant issues. The term "prosody" covers a wide range of features of speech, including phrasing, pitch, loudness, tempo and rhythm. According to review prosody has a great influence on the intelligibility and naturalness of speech perception.

Now a day emphasize is on the data-driven (corpus-based) approach of extracting prosodic features and focus on the design of a database. In this paper experiments on Marathi language show that algorithm can obtain appropriate emotional prosody features, and emotions like sadness and happiness can be synthesized.

1.1. Text To Speech Synthesis

A Text-to-Speech (TTS) Synthesizer is a computer based system that should be adept to read any text distinctly, whether it was enter in the computer by an operator or scanned and submitted to an Optical Character Recognition (OCR) system. The objective of a text to speech system is to convert a given text into a spoken waveform using correct algorithms. Main components of text to speech system are: Text processing and Speech generation. It was observed that various distinct techniques were there to convert a given text into speech and that each such system has its own advantages and disadvantages. However, the key challenge in all cases is the quality of the sound produced and its naturalness.

Any TTS is divided into two stages. Text is essentially a string of characters can be input into the computer either through a keyboard or through an optical character recognition (OCR) system. When the text is input through the keyboard, the characters are encoded using Unicode or ISCII (Indian Standard Code for Information Interchange) format. Optical character recognition (OCR) will be highly accurate for typed text but not for handwritten text.

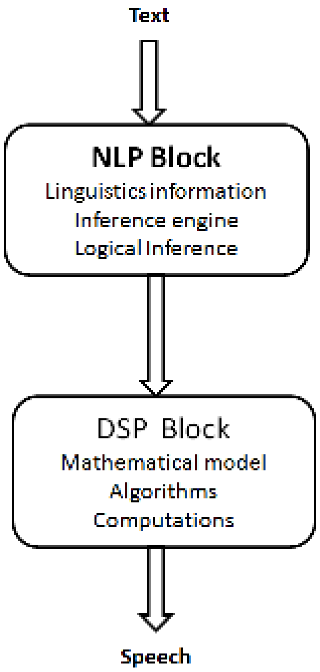


Figure 1: Basic block diagram of TTS

In the first step, whatever text is required to be spoken is passed to a Natural Language Processor (NLP). The first stage takes text input, processes it and converts it into precise phonetic string to be spoken. This conversion of input text into linguistic representation, usually called text-to-phonetic or grapheme-to-phoneme conversion.

The second stage takes this phonetic representation of speech and generates digital signal using a synthesis technique. Each phoneme maps to a signal unit of the audio database. So in this step the concatenation unit reads the phonemes one by one from the output phoneme string. For each such phoneme read, the unit fetches the corresponding sound unit that the phoneme maps to from the audio database and join it to the end of the voice output string.

The paper is organized as follows: Section II confers literature review. A concatenation approach and description of the database is provided in section III Then, the prosodic approach for emotion identification is described. Next section gives brief of algorithm used for both approaches. Discussion of the results is provided in section VI. Finally, conclusions are depicted.

2. Literature Review

Till now TTS synthesis for many languages is done like English, Maindrain, Arabic [1, 2], Chinese [4], Bangia, Hindi, Sindhi, Telgu, Marathi etc. There are three main approaches to speech synthesis: formant synthesis, articulatory synthesis, and concatenative synthesis. Out of which first two are rule-based synthesis techniques and concatenation uses data base approach [8].

In formant synthesis, the vocal tract transfer function can be duly modelled by simulating formant frequencies and formant amplitudes [9-13].

Direct modelling of the human articulator behaviour generates articulatory synthesised speech, predominantly it is the most satisfying method to produce high quality speech but at same time most difficult methods to implement [14, 15]. In general, the results of articulatory synthesis are not as good as the results of formant synthesis or the results of concatenative synthesis.

To trounce limitations of Articulatory and Formant synthesis, concatenative synthesis follows a data driven approach. By connecting natural, pre-recorded speech units, speech is generated in concatenative synthesis [16].

In Symbol based concatenation approach for Text to Speech System for Hindi using vowel classification technique algorithm developed by author [17] involves analysis of a sentence in terms of words and then symbols involving combination of pure consonants and vowel technique.

In sindhi TTS system[18] input Sindhi text is converted into sounding speech using concatenative synthesis method, in which sounds of phonemes, diphones and syllables are concatenated using novel algorithm. Also the approach used to develop a Text-To-Speech (TTS) synthesis system for the Punjabi text written in Gurumukhi script was discussed [20].

Work done on Marathi language [19] presents the concatenative TTS system and discusses the issues relevant to the development of a Marathi speech synthesizer using different choice of units like words and phonemes as a database.

Emotional speech is one of the key techniques towards a natural and realistic conversation between human and machines [27].Work on TTS for Marathi Language is done but not with natural prosody effect, therefore natural prosody generation in TTS for Marathi Speech Signal using concatenation synthesis technique was implemented [25]. A TTS is built using both neural networks and fuzzy logic [26] giving more natural sounding speech.

3. Concatenative Synthesis

In concatenation method, the synthetic speech is made by joining a speaker's natural short term waveform segments which have accumulated in a large speech corpus in advanced. In a concatenative synthesis approach, the DSP module obtains the sound files from an acoustic inventory corresponding to the string of phonemes or words and concatenates them. Finally, it modulates the sound according to the intonation and prosodic information if present.

In other words, the synthetic speech, generated by this method, is recycling of a speaker's natural voices that preserve naturalness and individuality. The most preferable waveform segments are searched from the corpus.

Although this methodology is simple compared with other conventional TTS methods such as LPC synthesis, larger amount of computation and wider memory space are needed. In general, the degree of quality of synthetic speech is proportional to the corpus size .Thus the high-quality synthetic speech can be realized by using concatenative speech synthesis with extra large speech corpus.

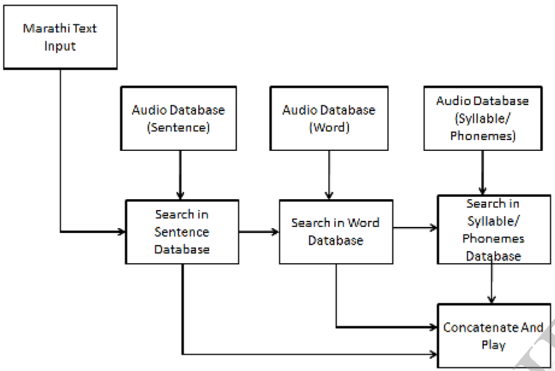

Marathi Text Input
Audio Database (Sentence)
Audio Database (Word)
Audio Database (Syllable/ Phonemes)
Search in Sentence Database
Search in Word Database
Search in Syllable/ Phonemes Database
Concatenate And Play

Figure 2: Block diagram for Concatenation Synthesis

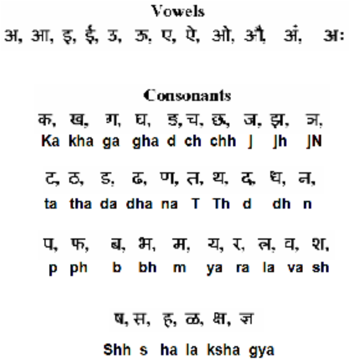
The text input is either standard words or numbers or symbols. If the input text is a number then it is handled by a digit processor. Whenever string of input text is given to the computer it starts searching word into word database. If word is not found in the database then word is split into syllables and syllables are searched in the syllable database. If the syllable also does not exist in the database then syllable split into phonemes and search into phoneme database. Finally concatenating phonemes audio is played as shown in figure 2

3.1. Nature of Marathi Language Script

Marathi script is originated from the ancient Brahmi script. The basic units of the writing system are referred to as Aksharas. An Akshara is an orthographic representation of a speech sound with are syllabic in nature, thus have a generalized form of C*V. Marathi has a complex system of signs to indicate consonant clusters or 'jodAkshare'.

3.2. Marathi Phonemes

Marathi script consists of 16 vowels and 36 consonants making a total of 52 alphabets as shown in figure 3. Out of the 36 consonants, first 25 are divided into 5 groups, each containing 5 letters. This classification is based on their pronunciation.


Vowels
अ, आ, इ, ई, उ, ऊ, ए, ऐ, ओ, औ, अं, अः

Consonants
क, ख, ग, घ, ङ, च, छ, ज, झ, ञ,
Ka kha ga gha d ch chh J Jh JN

ट, ठ, ड, ढ, ण, त, थ, द, ध, न,
ta tha da dha na T Th d dh n

प, फ, ब, भ, म, य, र, ल, व, श,
p ph b bh m ya ra la va sh

ष, स, ह, ळ, क्ष, ज्ञ
Shh s ha la ksha gya

Figure 3: Marathi Phonology

3.3. Marathi Syllables

Combination of phonemes gives rise to next higher unit called syllables which is one of the most important units of a language. A syllable must have a vowel called its nucleus, whereas presence of consonant is optional.

3.4. Format of input text

The script of Marathi language is stored in digital Computers in ISCII, UNICODE and in transliteration scheme of various fonts. Synthesis engine can separate input given in any of these formats. Marathi language has a common phonetic base.

3.5. Database creation and searching

TTS Database contains a simple Marathi words, consonants and vowels. Recorded data is stored in files named ka.wav, kha.wav upto gya.wav for consonants and a.wav upto aha.wav for vowels. Likely words.wav files are stored in word database.

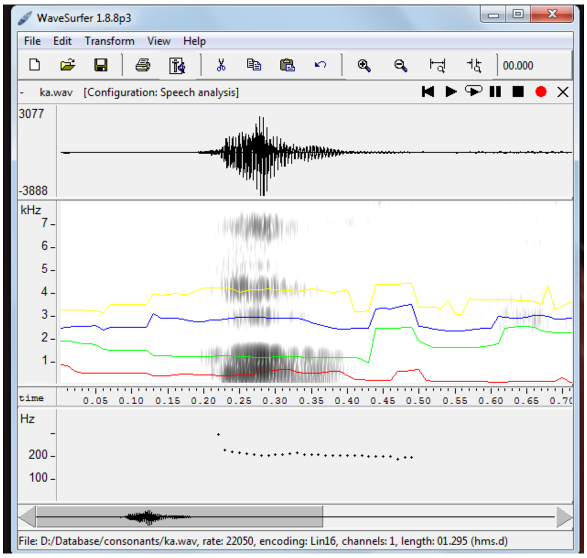
Basically there are two databases, audio database which stores the audio files and text database which stores the text files corresponding to audio files in the audio database. When the given word does not exist in database then it is synthesized from syllables where breaking of word is performed (CV structure). An Akshara (word) in Marathi language scripts is close to a syllable and can be of the following form: C, V, CV, CCV, VC and CVC where C is a Consonant and V is a vowel.

Here are some of Rules for words without joint alphabets-

CV structure	CV Break
CVCVCVC	CV+CV+CVC
CVCV	CV+CV
VCV	V+CV
CVCVCVCV	CV+CV+CV+CV

CV structures are split according to rules draw from experimentally based on the structure of

Marathi language. Normally there is one vowel in each syllable.



File: D:/Database/consonants/ka.wav, rate: 22050, encoding: Lin16, channels: 1, length: 01.295 (hms.d)

Figure 4: Recorded speech signal of Consonant 'K'

4. Text To Speech Conversion Using Prosodic Approach

Prosody related to the melody and rhythm of speech. Prosody conveys syntactic, semantic, as well as emotional information. Emotional speech is one of the key techniques towards a natural and realistic conversation between human and machines.Prosodic aspects are often divided into features such as in stress and intonation. Stress is a shorter-term variation that highlights a specific syllable or a semantically important word. Intonation is a longer-term variation that is linked to the grammatical structure. For instance, it applies differently to questions and declarations.

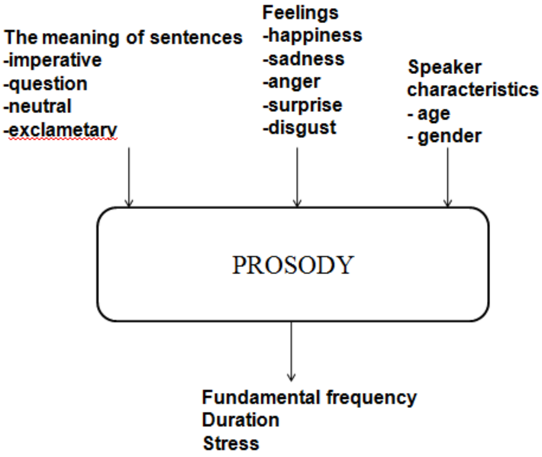


Figure 5: Prosody Characteristics

Prosodic features are roughly comparable to supra segmental features. The pitch, loudness, and quantity are among the most notable ones. They correspond to physical properties, respectively the fundamental frequency, the intensity (or amplitude), and the duration. Prosodic features extend over a sentence, a phrase, and a word syllable. Unfortunately, written text barely contains information of these features and some of them change dynamically during speech. Figure 5 shows few prosodic dependencies. However, with some specific control characters this information may be given to a Speech synthesizer.

4.1. MFCC

In this work, we selected the MFCC as representative features of the segmental information, as it is considered to be the best available approximation of human ear, which have demonstrated equitable performance in similar emotion identification tasks. The MFCC features have been calculated using Matlab, using 25 ms window length every 10ms, with a Hamming windowing and a pre-emphasis factor of 0.97. Figure 6 shows block diagram of MFCC

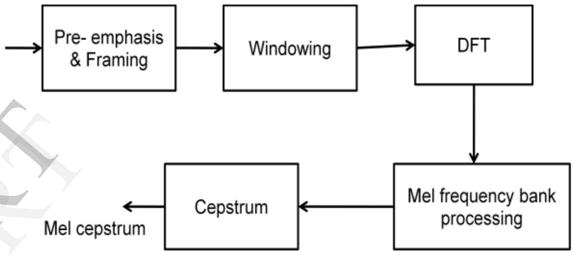


Figure 6: Basic block diagram of MFCC

4.1.1. Pre–emphasis

In this step signal is pass through a filter which emphasizes higher frequencies so that energy of signal at higher frequency get increases.

4.1.2. Framing

The process of segmenting the speech samples into a small frame with the length within the range of 20 to 40 msec. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M (M<N). Typical values used are M = 100 and N= 256.

4.1.3. Windowing

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines.

If the window is defined as W (n), $0 \leq n \leq N\text{-}1$ where

N = number of samples in each frame

Y[n] = Output signal

X (n) = input signal

W (n) = Hamming window,

The Hamming window equation is given as:

$$w(n) = \begin{cases} 0.54 - .46\cos\left(\frac{2\pi n}{N-1}\right) & 0 \le n \le N-1 \\ 0 & \text{otherwise} \end{cases}$$

4.1.4. Discrete Fourier Transform (DFT)

Input is Windowed signal x[n]…x[m] gives Output for each of N discrete frequency bands. A complex number X[k] representing magnitude and phase of that frequency component in the original signal Algorithm for computing DFT uses Fast Fourier Transform (FFT) with complexity N*log (N) where in general, N=512 or 1024

$$\mathrm{X[k]} = \sum_{n=0}^{N-1} \mathrm{X[n]} e^{-j2\frac{\pi}{N}kn}$$

4.1.5. Mel Filter Bank Processing

Mel-scale-Human hearing is not equally sensitive to all frequency bands as less sensitive at higher frequencies; roughly > 1000 Hz. Mel-scale is approximately linear below 1 kHz and logarithmic above 1 kHz and defines as

$$\mathrm{Mel(f)} = 2595 \times \log_{10}(1 + \frac{f}{700})$$

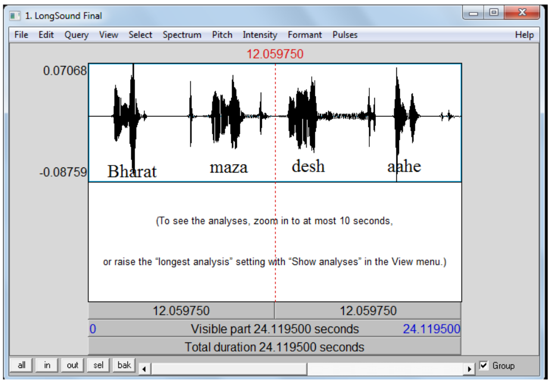
5. Algorithm

Concatenation synthesis with unit selection process is easy. For a given textual input which is mapped into present database, synthesis algorithm concatenates the corresponding wave files sequentially from left to right. TTS synthesis process along with prosody consideration is elucidating as follow.

Step 1 Provide input text
Step 2 Check for prosody in input text
Step 3 If there is no prosody, go to step 5
Step 4 If there is prosody, apply prosody rules and go to step 5
Step 5 Search word into word database
Step 6 If word is there select corresponding wave file and go to step 13
Step 7 If word is not there split the word into syllable
Step 8 Search syllable into syllable database.
Step 9 If Syllable matches with database, select corresponding wave file and go to step 13
Step 10 If syllable is not there in database split into phonemes
Step 11 Search corresponding phonemes into consonants and vowels database
Step 12 Select corresponding wave files and go to next step
Step 13 Store wave file
Step 14 Stringing together all wave files and play.

6. Results

Database contains all the texts (units) and the prosody. When the text will be given as input it will be compared with the database entered and according to the prosody entered the precise file will be played.



(To see the analyses, zoom in to at most 10 seconds,

or raise the "longest analysis" setting with "Show analyses" in the View menu.)

Visible part 24.119500 seconds

Total duration 24.119500 seconds

Figure 7: Concatenation Output for input text- Bharat maza desh aahe

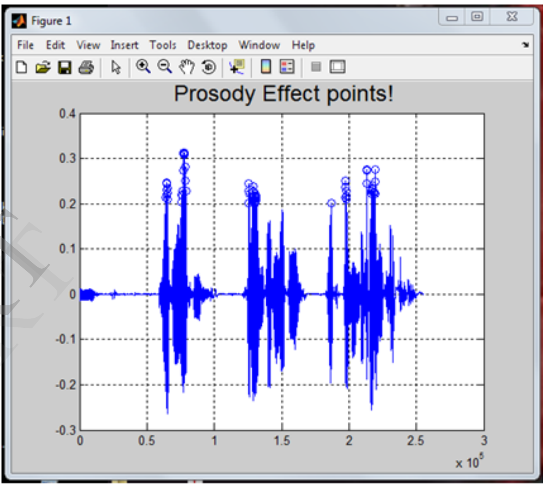


Figure 8: Prosody points for happy text

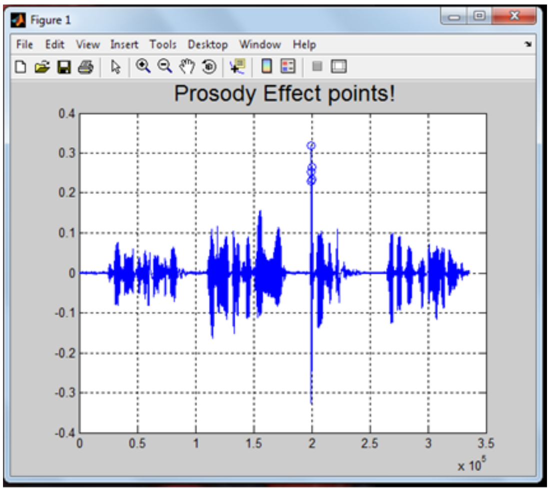


Figure 9: Prosody points for sad text

Figure 7 shows output for Marathi input text: Bharat maza desh aahe. Figure 8 shows output of prosody effect points for happy text input. All the prosody points are lies above 0.2 to 0.3 values of pitch. This means for given input text output pitch and intonation of speech signal is maintained between these points. Figure 9 shows output of prosody effect points for Sad text input. Output speech signal lies below 0.2 range of pitch. This means for given input text output pitch and intonation of speech signal is maintained between these points.

7. Conclusion

Different methodologies for synthesis, viz. Concatenation, unit selection, articulatory and formant, are overviewed. A comprehensive review of current synthesis methods & algorithms for text to speech quality improvement has been discussed. Most research on speech synthesis focuses on improving the quality of the speech produced by Text-to-Speech (TTS) systems to meet integibility and naturalality. In this paper concatenation method is used because of its simplicity and again easy to implement. It was observed that when the database of units is small, the synthesizer is likely to produce a low quality speech. As the database of units increases, it increases the quality of the synthesizer. Current Text-to-Speech systems based on concatenative method can synthesize different emotions (happy and Sad) with limited range of results due to lack of large amount of emotional speech data.

It was observed that the word unit performs better than the phoneme units, and seems to be a better representation for Indian languages such as Marathi. Creation of maximum coverage of units for concatenation synthesis can give greatest naturalness. Areas need to focus more are text analysis and prosody. Currently, unit selection is the best sounding speech synthesis method.

8. References

[1] Rachid Hamdi,Hamed D.Alsharari and Mouldi Bedda,"Arabic verbal sentences synthesis by hybridizing neural networks and genetic algorithms", The 1^{st} International Symposium on Computers and Arabic Language & Exhibition 2007 © KACST & SCS

[2] Zemirli, Z.; Khabet, S; Mosteghanem, M., (2007), “An effective model of streesing in an Arabic Text to Speech System”, IEEE AICCSA, pp. 700-707.

[3] Dakkak, O.; Ghneim, N. Abou Zliekha, M.; moubayed, S., (2005), “Emotion Inclusion in an Arabic Text-to-Speech”, 13th European Signal Processing Conference.

[4] Tien Ying Fung and Helen M. Meng, “Concatenating Syllables for Response Generation in Spoken Language Applications”, IEEE, Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings, Page(s): II933 - II936 vol.2

[5] R. Barra, J.M. Montero, J. Macías-Guarasa, L.F. D’Haro, R. San-Segundo, R. Cordoba, “PROSODIC AND SEGMENTAL RUBRICS IN EMOTION IDENTIFICATION",IEEE,Acoustics, Speech and Signal Processing,ICASSP 2006 Proceedings. Volume: 1,Page(s): I

[6] Sandeep Chaware, Srikantha Rao,"RULE-BASED PHONETIC MATCHING APPROACH FOR HINDI AND MARATHI”, Computer Science & Engineering: An International Journal (CSEIJ), Vol.1, No.3, August 2011.

[7] Anupam Basu, Debasish Sen, Shiraj Sen and Soumen Chakraborty,"An Indian Language Speech Synthesizer – Techniques and Applications “INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR 721302, DECEMBER 17-19, 2003.

[8] Youcef TABET and Mohamed BOUGHAZI,"SPEECH SYNTHESIS TECHNIQUES. A SURVEY", 2011 7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA).

[9] T .Styger, & E. Keller. “Formant synthesis,” In E. Keller (ed.), Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges (pp. 109-128). Chichester: John Wiley.1994N.

[10] J. Allen, S. Hunnicutt, R. Carlson and B. Granstrom, “MITalk-79: The 1979 MIT text-to-speech system,” Speech communications papers presented at the 97th meeting of the acoustical society of america, Cambridge, USA, pp. 507- 507,1979 .

[11] J. Allen, S. Hunnicutt and D.H. Klatt, “From Text-to speech: The MITalk System,” Cambridge University Press, Cambridge, 1987.

[12] D.H. Klatt, “The klattalk text-to-speech conversion system,” Proceeding on the international conference on acoustic, speech and signal processing, Paris, pp. 1589-1592, 1982.

[13] D.H. Klatt, “DecTalk user's manual,” Digital Equipment Corporation Report, 1990.

[14] B. Kroger , “Minimal Rules for Articulatory Speech Synthesis,” Proceedings of EUSIPCO92, pp.331-334,1992.

[15] D.H. Klatt “Review of text-to-speech conversion for English,” Journal of the Acoustical Society of America, vol. 82(3), 1987.

[16] T. Dutoit, “High-Quality Text-to-Speech Synthesis: an Overview,” Journal of Electrical & Electronics Engineering, Australia: Special Issue on Speech Recognition and Synthesis, vol. 17, pp. 25-37, 1999.

[17] Chaudhury, P.; Rao, M.; Vinod Kumar, K., "Symbol based concatenation approach for Text to Speech System for Hindi using vowel classification technique", Nature & Biologically Inspired Computing (NaBIC), Page(s): 1082 - 1087, 2009

[18] Mahar, J.A.; Memon, G.Q.; Shah, S.H.A.,"WordNet Based Sindhi Text to Speech Synthesis System”, Second International Conference on Computer Research and Development, 2010 Page(s): 20 - 24,2010.

[19] Shirbahadurkar, S.D.; Bormane, D.S., "Marathi Language Speech Synthesizer Using Concatenative Synthesis Strategy (Spoken in Maharashtra, India)", Second International Conference on Machine Vision, 2009. ICMV '09, Page(s): 181 - 185

[20] C.Singh,"Text-To-Speech Synthesis System for Punjabi Language", ICISIL 2011,CCIS 139,pp.302-303, © Springer-verlag berlin Heidelberg -2011

[21] Chang-Heon Lee, Sung-Kyo Jung, and Hong-Goo Kang,"Applying a Speaker-Dependent Speech Compression Technique to Concatenative TTS Synthesizers",IEEE TRANSACTIONS ON AUDIO,

SPEECH, AND LANGUAGE PROCESSING, VOL. 15, NO. 2,pp.632-640, FEBRUARY 2007

[22] Muhammad Masud Rashid, Md. Akter Hussain, M. Shahidur Rahman,"Diphone Preparation for Bangia Text to Speech Synthesis",12th International Conference on Computer and Information Technology (ICCIT 2009),21-23 December, 2009, Dhaka, Bangladesh

[23] Monojit Choudhury,"Rule Based Grapheme to Phoneme Mapping for Hindi Speech Synthesis" Department of Computer Science and Engineering Indian Institute of Technology, Kharagpur

[24] B. Gerazov,G. Shutinoski,G. Arsov,"A Novel Quasi-Diphone Inventory Approach to Text-To-Speech Synthesis",The 14th IEEE Mediterranean Electrotechnical Conference (MELECON 2008),Page(s): 799 - 804,2008

[25] Mrs. Madhavi R. Repe ,Mr. S.D. Shirbahadurkar,Mrs.Smita Desai,"Prosody Model for Marathi Language TTS Synthesis with Unit Search and Selection Speech Database", International Conference on Recent Trends in Information, Telecommunication and Computing (ITC),Page(s): 362 - 364 ,2010

[26] Thesis by Jonathan Brent Williams on "Prosody in Text-to-Speech Synthesis Using Fuzzy Logic",2005

[27] Ling CEN, Paul CHAN, Minghui DONG, Haizhou LI,"Generating Emotional Speech from Neutral Speech",7th International Chinese Spoken Language Processing (ISCSLP),Page(s): 383 - 386,2010

[28] Dan-ning Jiang; Wei Zhang; Li-qin Shen; Lian-hong Cai,"PROSODY ANALYSIS AND MODELING FOR EMOTIONAL SPEECH SYNTHESIS",IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05),Volume: 1, Page(s): 281 - 284,2005