A New Approach for Generating Text Description of Images and Speech Synthesis

Mrunmayee Patil
ME Student, Dept. of Computer Engg,
Dr.D.Y.Patil, SOET,
Department of computer engineering
Lohegaon, Pune.

Ramesh Kagalkar
Research Scholar and Astt.Professor,
Dr.D.Y.Patil, SOET,
Department of computer engineering
Lohegaon,Pune.

ABSTRACT
An image can be defined as a matrix of square pixels arranged in rows and columns. Image processing is a leading technology, which enhances raw images received from gadgets such as camera or a mobile phone in normal day to day life for various applications. An image to text and speech conversion system can be useful for improving accessibility of images for visually impaired as well as physically challenging people understand the scenario from the images and also train the system as that of human brain. The techniques of image segmentation and edge detection play an important role in implementing proposed system. The system generates text descriptions for an input image given by the user. Object wise generation of sentences, preposition and conjunction mapping is a challenging task. We formulate the interaction between image segmentation and object recognition in the framework of Canny algorithm. The system goes through various phases such as pre-processing, feature extraction, object recognition, edge detection, image segmentation and Text To Speech (TTS) conversion. The proposed system database consists of huge set of sample images, which help to perform training of database. The accuracy of proposed system is achieved due to the proper recognition of objects and sentences are formed making use of the recognized objects. These sample images consists of several categories of images. The system

mainly consists of two main modules such as image to text and text to speech. An image to text module generates text descriptions in natural language based on understanding of image. A text to speech module generates speech synthesis in English from description of natural language.

Keywords
Image Processing; Image Segmentation; Speech Synthesis; Text to Speech Conversion; Edge Detection.

1. INTRODUCTION
The field of image processing is widely used in many areas of research. Common people can easily sense what all is going around them and can view all that is present in this universe. Blind people or visually impaired people find it difficult to interact with the world and they cannot exactly sense the things so they need some or the other human intervention. There are around 15 million blind people in India. In order to make provision for such people there is need of converting images into text and speech However, there is a need to develop an interface for such visually disabled people to communicate with the world. The proposed system makes a better provision for converting captured images as well as stored images to be converted into text and speech. In this conversion process there are various techniques used such as image pre-processing, image segmentation, edge detection and text to speech synthesis. Step by step execution of these techniques helps to achieve the final output. Input to the system is an image and final output is speech output. Image recognition and computer vision are major fields of research in today's world. Language, whether written, spoken or typed, makes much of human communication. This language describes the visual world either around us or in the form of images and video. Combining visual imagery and visually descriptive language is a challenge for computer vision. Proposed work mainly focuses on generating text descriptions for particular input from user and speech output for generated text. The output of our system which is an automatically generated description of images has many related applications such as improving accessibility of images for visually impaired. Proposed work focus is also on general research in the study of visually descriptive text and going deeper into the connection between images and language that has potential to suggest new directions for research in computer vision.

The previous work done has several drawbacks and did not prove efficient to achieve the proper text and speech output. Lack of firm platform is also a factor to be considered. There is need of proper object recognition from images and generating more accurate and correct text descriptions for images. Hence there is need to develop a new methodology for translation of image to text description and generating speech synthesis.

2. RELATED WORK
Benjamin Z. Yao, Xiong Yang, Liang Lin, Mun Wai Lee and Song-Chun Zhu [4] proposed an image parsing to text description that generates text for images and video content. Image parsing and text description are the two major tasks of his framework. It computes a graph of most probable interpretations of an input image. This parse graph includes a tree structured decomposition contents of scene, pictures or parts that cover all pixels of image. Over past decade many researchers form computer vision and Content Based Image Retrieval (CBIR) domain have been actively investigating possible ways of retrieving images and videos based on features such as color, shape and objects [5][6]. Paper [7] introduced by Yi-Ren Yeh, Chun-Hao Huang, and Yu-Chiang Frank Wang presents a novel domain adaptation approach for solving cross domain pattern recognition problem where data and features to be processed and recognized are collected for different domains.

S. Shahnawaz Ahmed, Shah Muhammed Abid Hussain and Md. Sayeed Salam [8] introduced a model of image to text conversion for electricity meter reading of units in kilo-watts by capturing its image and sending that image in the form of Multimedia Message Service (MMS) to the server. The server will process the received image using sequential steps:

1) read the image and convert it into a three dimensional array of pixels, 2) convert the image from color to black and white, 3) removal of shades caused due to nonuniform light, 4) turning black pixels into white ones and vice versa, 5) threshold the image to eliminate pixels which are neither black nor white, 6) removal of small components, 7) conversion to text.

In [9] Fan-Chieh Cheng, Shih-Chia Huang, and Shanq-Jang Ruan gave the technique of eliminating background model form video sequence to detect foreground and objects from any applications such as traffic security, human machine interaction, object recognition and so on. Accordingly, motion detection approaches can be broadly classified in three categories: temporal flow, optical flow and background subtraction.

CBIR(content based image retrieval) system enables digital images to be processed and used to extract the features vector on the basis of low level properties of the image. Color, texture and shape are to be considered. A solutions to this is given in [10].

Another approach in feature extraction is given in analyzing edges of image. Jain and vailaya [11] used edge direction method to build edge direction histogram firstly we have to find edges of image and then quantize them. It had limited performance.

Then Shandehzadeh [12] improved this method by considering the correlations between edges by using a weighted function.

3. SYSTEM OVERVIEW

In the proposed system our main focus is on identifying input image and converting it into relevant text and speech. So firstly we create a database of images. The input to the system can be images form database or captured image. These images when taken as input from the system, system checks for similar kind of images to map the features of objects. Once objects are detected then it identifies the scenario in the image and system gives text output. This generated text is then undergoes into text to speech synthesis and we get the speech output. This kind of approach may help the blind people to know the scenario around them. Following figure: 1 gives idea about system overview. [19,20,21]

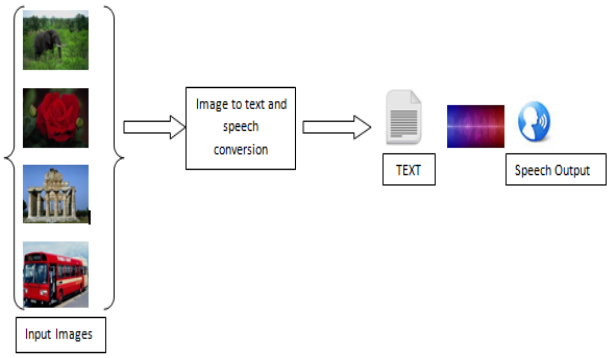

Fig 1: System Overview

4. PROPOSED SYSTEM ARCHITECTURE AND WORKING

In the proposed system our main goal is to identify objects in the image. When objects in the image are recognized it is easy to identify that image and convert it into respective test and speech. Techniques mainly used in the conversion process are:

- Pre-processing
- Gray scaling
- Edge detection
- Segmentation
- Feature extraction
- Speech synthesis

Based on these techniques of image processing we convert image into text as well as speech.

1) Pre-processing: Pre-processing of images involves mainly of removal of low-frequency background noise, removing reflections and masking portions of images. Pre-processing technique enhances data images in order to do further computational processing.[22]

2) Gray scaling: In gray scaling each pixel value of an image is represented using shades of gray. These kind of images are also known as black and white images. Each pixel intensity is expressed within the range of minimum and maximum where range is 0(black) and 1(white), any fractional value is in between.

Fig 2: Example of grayscaling. (A) Original image. (B) Gray scale image.

3) Edge Detection: Edge detection is the technique used to identify the fine edges in digital images. It identifies the points in the image at which the brightness of image changes very sharply. Point at which the image brightness changes are organized into set of curved line segments known as edges. Edge detection is mainly an important tool in the field of image processing for detection of features and feature extraction[23,24]

Approaches of edge detection-

Two main methods of edge detection are search based and zero crossing based. Search based method detects edges by computing the edge strength. The zero crossing based method searches for the zero crossing second-order derivative expression computed form images in order to find edges.

4) Segmentation: The aim of segmentation is converting an image into more meaningful and easy to analyze portions.

Segmentation does the job of partitioning an image into multiple segments which help to locate the objects and boundaries (curves, arcs, lines, etc.) in an image. With the help of image segmentation we can assign a label to each pixel which then same labels share the certain characteristics. We can characterize the pixels in a region with respect to the characteristics such as color, intensity or texture [25]

5) Feature extraction: In the fields of machine learning, pattern recognition and image processing, feature extraction plays an important role of building derived values which are known to be the features. These features are intended to be informative, non-redundant, facilitating the subsequent and generalized steps. Extracted features should contain relevant data from input data. This technique plays an important task in our proposed system.

Feature Extraction is the key concept in CBIR. A certain number of features for each image are extracted, describing its high level content information. Then, according to the similarity of these vectors, we can compare two specific images to each other. This class uses different techniques to extract features related to a single or the group of images[26].

6) Speech Synthesis: It is the process of artificially producing the human speech. Systems used for such purpose are called speech synthesizer which can be a software or hardware product. Concatenations of several pieces of recorded speech are stored in database and then synthesized speech is created.

A text -to-speech conversion system is used in our work to convert the generated text of images into speech output.

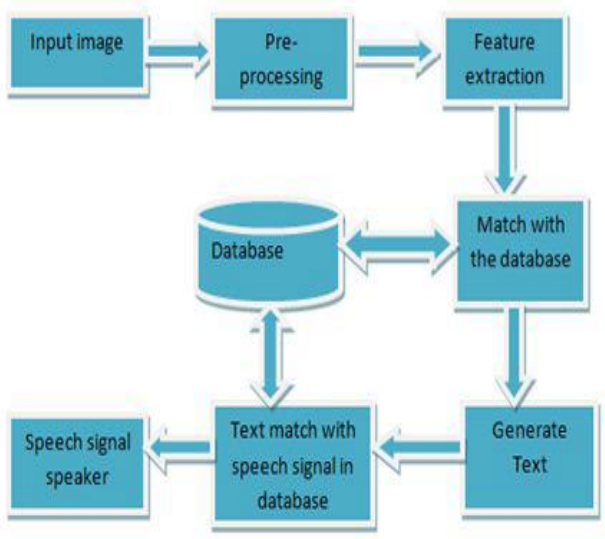

Fig 3: Architecture of proposed system

5. Flow of proposed system

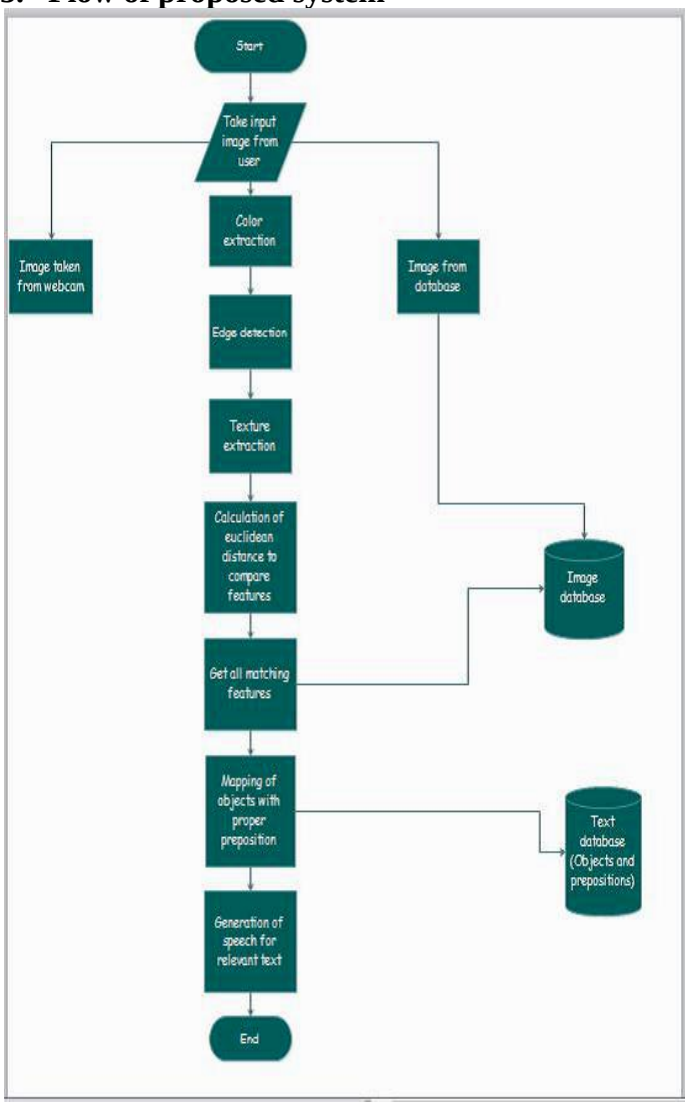

Fig 4: Flow of proposed system

In the above figure 4, the flow of our proposed system of image to text and speech conversion is given. Here, the input to the system is an image. Then pre-processing and pattern matching phases are carried out. In these phases the gray scaling of images is done. After segmentation and edge detection the next phase is object detection. In this phase, the important objects of the image are analyzed. Then the matching and comparing of images with database are carried out. The relative text for the identified objects is generated and then that text is converted into speech which is easy to be heard by the blind people.

6. ALGORITHMIC STEPS

The proposed system consists of following algorithmic steps:

Step 1. Color extraction using RGB values

- RGB-HSV extraction

Step 2. Edge detection

- Detect edges
- Convert obtained RGB image into binary values

Step 3. Texture extraction

- Color

- Shape
- Texture

Step 4. Calculate Euclidean distance to compare features with database images

Step 5. Get all the matching features

Step 6. Get objects relevant to features

Step 7. Map objects with relevant prepositions and conjunctions

Step 8. Sentence generation

Step 9. Convert generated text to speech by using the voice engine of system

The flowchart for above algorithm is given as:

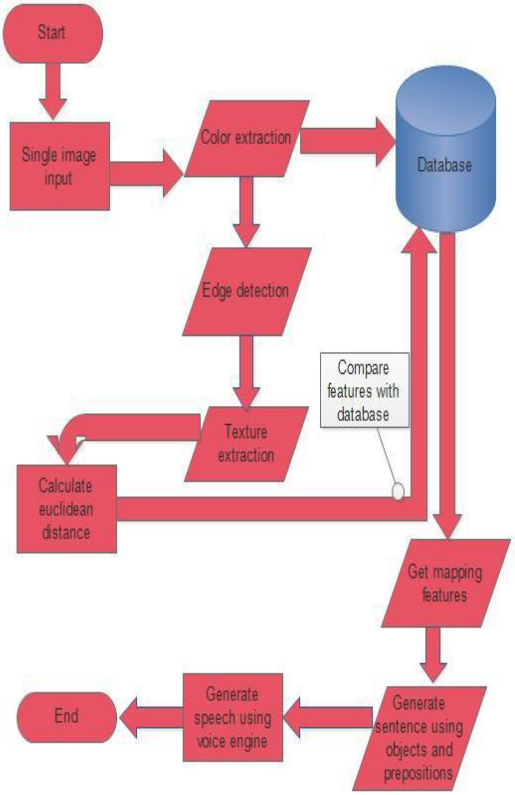

Fig 5 Flow chart for testing phase algorithm of proposed system.

7. ANALYSIS

For evaluation of this system, we use dataset taken from UCI repository which contains of total 14 categories of images. The image dataset contains of about 1000 images. From this set we make use of around 400 images for the purpose of training the proposed system. In the existing system, two forms of quantitative evaluation were performed, automatic evaluation using standard methods for evaluating generated sentences and human forced evaluations to directly compare the results between their method and several previous methods. In each case they also quantitatively evaluate and compare to two previous approaches for image description generation used on the same dataset [17]. The first comparison method is the bottom up HMM approach from Yang et al. [18], which detects objects and scenes, and then hallucinates plausible verbs for generation (using text statistics). The second comparison method is a retrieval-based approach from Farhadi et al. [16]. This method detects objects, scenes and actions and then retrieves descriptive sentences from similar images through the use of a meaning space.

Summarization of Experimental Results

The experimental results are outlined in the tables given below. There are total 14 categories of images, the results of 4 categories is given in the form of tables and their graphs.

Table 1: Comparison of time taken by proposed system for text and speech generation.

Id	Category	Text generation (In milliseconds)	Speech generation (In milliseconds)
1.	Sea Shore	11246	4382
2.	Historical monument	6137	3245
3.	Bus	5167	3211
4.	Elephant	5917	4174

Table 1 shows the time taken by the proposed system for generating text descriptions and speech translations. The categories of images from dataset considered for text and speech translations are sea shore, historical monument, bus and elephant images. The text generation time is comparatively more than the time taken for speech generation. The text generated by the proposed system consults the database tables of object and preposition for proper matching object names and prepositions. When the text is displayed on the text field of the screen, the speech is generated for the same text using the voice engine of the laptop.

Graphical Representation

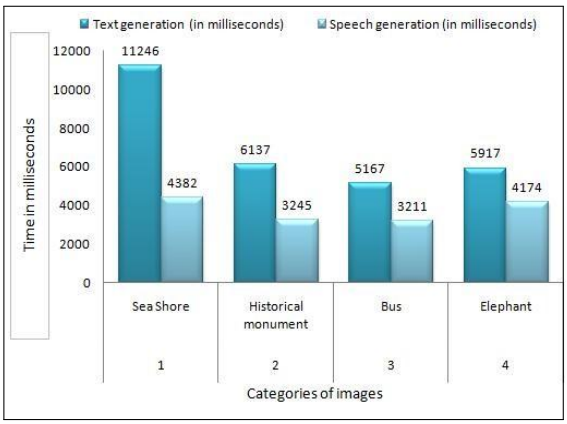

Fig 6:- Graph for text and speech generation

Figure 6 shows graph for the time taken by the proposed system for generating text descriptions and speech translations. The categories of images from dataset considered for text and speech translations are sea shore, historical monument, bus and elephant images. The text generation time is comparatively more than the time taken for speech generation. The text generated by the proposed system consults the database tables of object and preposition for proper matching object names and prepositions. When the text is displayed on the text field of the screen, the speech is generated for the same text using the voice engine of the laptop.

Thus from the analysis of normal person using the system it shows that the objects matching with the database is accurate

and clear. From the analysis it is clear that system is accurate and except some flaws which are acceptable. Thus proposed system can be efficient but it still depends on the ability of user which is using it.

8. CONCLUSION

The proposed image to text as well as speech conversion system provides the solution to the problems faced by blind people. In proposed system we have applied a simple and fast method which works suitably to recognize image and convert it into text as well as speech. It is low time-consumption approach, so that the real time recognition ratio is achieved easily. In the proposed system Canny edge detection algorithm is used which will recognize the input image by detecting the edges of objects in the image. It is capable of handling the different input images and translates them into text and speech. The proposed system is designed to translate The dataset contains the number of hand images that are taken from multiple user of different size which helps to recognize the correct output to any user using the system. The proposed system is trained on predefined dataset.

In future work we are looking for the dynamic system which will identify the dynamic actions or images taken dynamically. The dynamic image recognition system contains not only shapes but also the many other objects in image. So the system will continuously recognize the dynamic movements in videos/ images. Also in future work we are trying to have the advanced technology such as video conferencing, and try to make android application.

9. REFERENCES

[1] Mrunmayee Patil and Ramesh M. Kagalkar, "A Review On Conversion of Image To Text As Well As Speech Using Edge Detection and Image Segmentation",International Journal of Science and Research (IJSR 2014), ISSN (Online): 2319-7064 , Vol-3, Issue 10 Oct- 2014.

[2] Mrunmayee Patil and Ramesh M. Kagalkar," An Automatic Approach For Translating Simple Images Into Text Descriptions And Speech For Visually Impaired People ", Inter- national Journal of Computer Applications (IJCA), Vol- 118 , No.3, May 2015.

[3] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg and Tamara L. Berg, "Baby Talk: Understanding and Generating Simple Descriptions," IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 35, No. 12, December 2013.

[4] Benjamin Z. Yao, Xiong Yang, Liang Lin, Mun Wai Lee and Song-Chun Zhu, "I2T: Image Parsing to Text Description" ,IEEE transactions on image processing, 2008.

[5] Iasonas Kokkinos, Member, IEEE, and Petros Maragos, Fellow, IEEE "Synergy between Object Recognition and Image Segmentation Using the Expectation-Maximization Algorithm", IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 31, No. 8, August 2009.

[6] Fan-Chieh Cheng, Shih-Chia Huang, and Shanq-Jang Ruan, Member, IEEE "Illumination-Sensitive Background Modeling Approach for Accurate Moving Object Detection", IEEE Transactions On Broadcasting, Vol. 57, No. 4, December 2011.

[7] Dhiraj Joshi, James Z. Wang And Jia Li, The Pennsylvania State University, "The Story Picturing Engine—A System for Automatic Text Illustration", ACM Transactions on Multimedia Computing, Communications and Applications, Vol. 2, No. 1, February 2006.

[8] Munawar Hayat, Mohammed Bennamoun and Senjian An "Deep Reconstruction Models for Image Set Classification", IEEE Transactions on Pattern Analysis and Machine Intelligence.

[9] Mina Makar, Member, IEEE, Vijay Chandrasekhar, Member, IEEE, Sam S. Tsai, Member, IEEE, David Chen, Member, IEEE, and Bernd Girod, Fellow, IEEE, "Interframe Coding of Feature Descriptors for Mobile Augmented Reality", *IEEE Transactions On Image Processing, Vol. 23, No. 8, August 2014.*

[10] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," IEEE Trans. PAMI, vol. 22, no. 12, 2000.

[11] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 2, no. 1, pp. 1–19, Feb. 2006.

[12] A. Mian, M. Bennamoun, and R. Owens, "An efficient multimodal 2d-3d hybrid approach to automatic face recognition," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 29, no. 11, pp. 1927–1943, 2007.

[13] S. Feng, D. Xu, X. Yang, Attention-driven salient edge(s) and region(s) extraction with application to CBIR, Signal Processing 90, pp. 1–15, 2010.

[14] A. Vailaya, A. Jain, H.J Zhang, On Image Classification: City Images vs. Landscape, Proceeding of the IEEE workshop on Content-Based Access of Image and Video Libraries, pp. 3-8, 1998.

[15] J. Shanbehzadeh, F. Mahmoudi, A. Sarafzadeh, A.M. Eftekhari-Moghaddam, Image Retrieval Based on the Directional Edge Similarity, Proceeding of the SPIE: Multimedia Storage and Archiving Systems, Vol. IV, USA, pp. 267-71, 1999.

[16] A. Farhadi, M. Hejrati, A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D.A. Forsyth, "Every Picture Tells a Story: Generating Sentences for Images", Proc. European Conference On Computer Vision, 2010.

[17] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting Image Annotations Using Amazons Mechanical Turk",*Proc.* NAACL HLT Workshop Creating Speech and Language Data with Amazons Mechanical Turk, 2010.

[18] Y. Yang, C.L. Teo, H. Daume III, and Y. Aloimonos, "Corpus- Guided Sen- tence Generation of Natural Images", Proc. Conference on Empirical Methods in Natural Language Processing, 2011.

[19] Amitkumar Shinde & Ramesh Kagalkar,"Advanced Marathi Sign Language Recognition using Computer Vision , International Journal of Computer Applications

(IJCA), Volume 118 - No. 13,(ISSN No: 0975 8887), pp:1-7, April 2015.

[20] Amitkumar Shinde&Ramesh Kagalkar,"Sign Language Recognition for Deaf Sign User",International Journal For Research in Applied Science and Engineering Technology (IJRASET),Volume 2, Issue XII, December 2014, (ISSN No: 2321-9653), pp:67-69.

[21] Kaveri Kamble and Ramesh Kagalkar, "A Review: Translation of Text to Speech Conver- sion for Hindi Language " , International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064, Vol. 3 Issue 11, November 2014.

[22] Kaveri Kamble and Ramesh Kagalkar, "Audio Visual Speech Synthesis and Speech Recog- nition for Hindi Language", International Journal of Computer Science and Information Technologies(IJCSIT) ISSN (Online): 0975-9646, Vol. 6 Issue 2, April 2015.

[23] Kaveri Kamble and Ramesh Kagalkar, "A Novel Approach for Hindi Text Description to Speech and Expressive Speech Synthesis " , International Journal of Applied Information Systems (IJAIS) ISSN 2249-0868, Vol. 8 Issue 7 May 2015.

[24] Shivaji J. Chaudhari and Ramesh M. Kagalkar, "A Review of Automatic Speaker Age Classification, Recognition and Identifying Speaker Emotion Using Voice Signal", International

[25] Journal of Science and Research (IJSR 2014), ISSN(Online)2319-7064, Volume 3, Issue 11, November 2014.Shivaji J. Chaudhari and Ramesh M. Kagalkar, "Automatic Speaker Age Estimation and Gender Dependent Emotion Recognition ", International Journal of Computer Applications (IJCA) (0975 - 8887), Volume 117 No. 17, May 2015.

[26] Shivaji J. Chaudhari and Ramesh M Kagalkar, "A Methodology for Efficient Gender Dependent Speaker Age and Emotion Identification System ", International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE) ISSN 2319-5940, Volume 4, Issue 7, July 2015.