www.ijsrnsc.org

Available online at www.ijsrnsc.org

IJSRNSC

Vol.8, Issue.2, April 2020
Research Paper

Int. J. Sc. Res. in Network Security and Communication

E-ISSN:2321-3256

A Voice Signal Interpreter using Machine Learning Techniques

Rosy Mishra[1*], Y. Sowjanya[2], Sushanta Meher[3], Mousumi Meher[4]

[1,2,3,4] Dept. of CSE, Vikash Institute of Technology, Biju Patnaik University Of Technology,Baragrh,Odisha, India

Received: 19/Mar/2020, Accepted: 15/Apr/2020, Published: 30/Apr/2020

Abstract-As we all know computer generated speech in immense now a days. In the current era of IT usage of computer generated speech is in great demand where as these speeches are still not human, and it's not easy on the ears. The production quality system like Google Assistant, Apple Siri, Amazon Alexa, or Bixby as different from what a human generated speech would sound like. Many advances have been made in speech synthesis using neural network for easy accessing. Prime goal of this paper is to create an interpreter (Machine translator), to translate text to speech using a neural network. In this paper the main objective is to combine two techniques to produce a new product i.e. a new version of the translator. An artificial communicator is being developed to communicate with the universal language that is English for communication in between System & humans. These two technologies are used to link with each other & competence thoroughly. The purpose of this Application software is to permit different picture to convert its text in a language which is displayed in vocalizations. And it also permits for weakening vision in worst case seen in fields in picture format to understand language. A skill is being generated in machine through software for the Suitableness of humans for easy accessing.

Keywords- Neural network, machine translator, Text to speech Translator, perceptron model

I. INTRODUCTION

Machine learning techniques are being used to generate, synthesize and recognize speech signal now a days hence it is important to understand how these generation are being done traditionally. A translator designed in such a way that is easily applicable in any systems. To generate this translator two specific methods are used for conversion from text-to-speech (TTS), parametric TTS and Concatenate TTS[2].

Extracting the morphological feature of the input image by using the Digital image processing technique. These feature such as phonemes, duration. As we have implemented this technique in Matlab-2018 a platform, the code is generated in such a way that exactly the text of the image is extracted by using regionprops () keyword in the code. The same way speech synthesis technique, i.e. Deep Neural Network is to convert text-to-speech which text is extracted the speech will be delivered at the output.

The image to speech which is inputted in the code, the features of image or text written on the image is extracted through regionprops(). It is a keyword used in the code, the code is designed using artificial intelligence techniques so that catching up for human will be easy. The text is extracted from the image, i.e. converted to gray image, and the gray image is converted to speech using DNN. Then the Vocoder features of the speech signal are extracted. These features can be spectra, spectrogram and fundamental frequency. In this feature, some inherited features of human speech are there[4]. These parameters, with extracted features are inputted into a mathematical mode called a vocoder, Vocoder does multiple complex transforms to these features to generate an audio waveform. This waveform is highly modular& workable. Done some approximations to the parameter to generate all kinds of speech.

The portable text to speech converter is designed to help the visually impaired, listen to an audio reading back of any scanned text. The person having a weakening vision problem they cannot recognize some of the words written. Like restaurant name, office works and many. They can easily hear those words in audio in our software that generated by these techniques[5]. Even deaf-and-dumb who cannot hear and speak, they can also easily understand their required information by just inputting the image only.

In our day-to-day lives, many problems arise in getting the required information. The solutions of certain problems are easily solved by these techniques. The person just has to input the image, where ever he/she cannot understand the text on the image is converted to speech and delivered in audio as well as text. So, that person can easily get the information. Even the deaf-and-dumb can access the data as per their requirement in the certain cases.

The remaining paper is organized as follows: The proposed work discussed in section 2. Is related to materials and methods the neural network model using perceptron algorithm and architectural descriptions using some methods and extracting models. In section 3. The simulation result and recognition rate are discussed by representing set of features. In section 4. We draw the conclusion with future work.

II. MATERIALS & METHODS

The portable text to speech machine translator is designed to assist the visually impaired, listen to an audio reading back of any scanned text.

It allows a much wider range of algorithms to be applied to the input data and can avoid problems such as the build-up of noise and signal distortion
during processing. Since images are defined over two dimensions (perhaps more) digital image processing may be modeled in the form of multidimensional systems.

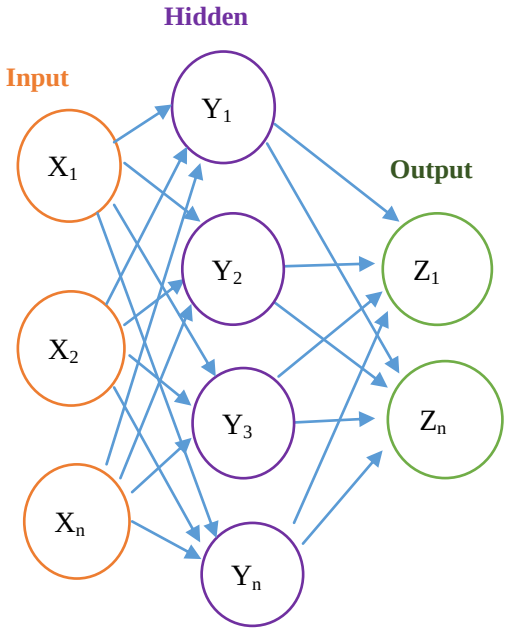

Fig. 1: Text to speech conversion Neural Network Model.

$$y = \phi(\sum_{i=1}^{n} WiXi + b) \text{---------- (1)}$$

Where **b** is the bias and ϕ is the non-linear activation function, wdenotes the vector of weights, **x** is the vector of inputs.
The major actions involved in this paper are:

A. Methodology

Digital image processing is the use of computer algorithms to perform image processing on digital images. Digital image processing allows the use of much more complex algorithms, and hence, can offer both more sophisticated performance of simple tasks, and the implementation of methods which would be impossible by analog means. Median filtering is a nonlinear operation often used in image processing to reduce "salt and pepper" noise. Median filtering is more effective than convolution when the goal is to simultaneously reduce noise and preserve edges.

```
If (a>=1) & (a<=15)
    Output= [Output '0'];

elseif (a>=16) & (a<=30)
    Output= [Output '1'];
Obj=System.Speech.Synthesis.SpeechSynthesizer;
Speak(obj,a);
a = RECOG(i);
```

*B. **Architectural** Flow of Proposed Model*

It comprises of two phases training and synthesis in the training period, text and speech features are extracted from the database and execute its alignment part, between them there is an acoustic features and text feature are used and then acoustic and duration models are trained using aligned data[6]. For training of an acoustic model input text features are repeated in each frame to the entire duration of a device to spread and then speech parameters are mapped to the corresponding frame-level. And at call-level the input text features are directly mapped to the corresponding device for training the duration model.

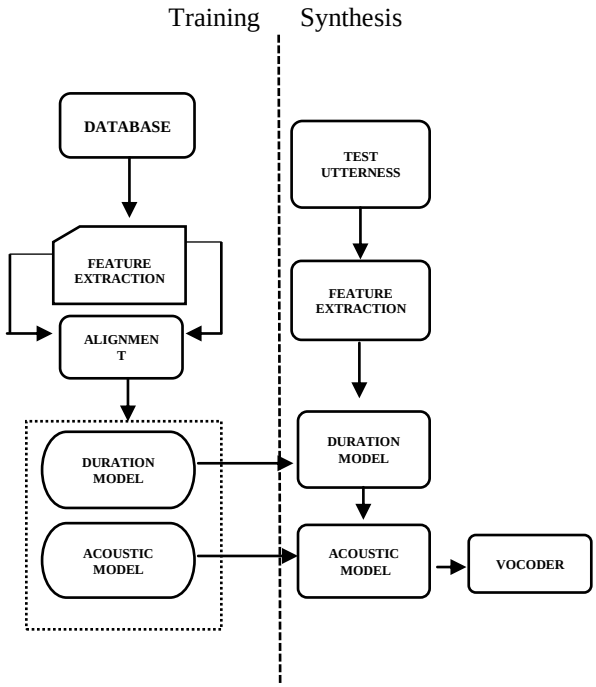

Fig.2: Proposed Model for Text To Speech conversion.

Another side of the model the synthesis stage, using the duration model first predicted durations for each call. Using the learning of the duration be synthesized, of each device. From the acoustic model the acoustic speech parameters are predicted. And from synthesizing speech signal the predicted speech parameters are passed through the vocoder.

A. Extracting the text from the input image

In Matlab 2018a platform this was developed. As we all know, the main motive of this software is to convert the text to speech &vice versa for easy accessing of humans.

At first the image is inputted into the program, the image is converted from RGB to Gray image. Using regionproprs(), and imfill() is used to extract the text on the image The text of the particular image is extracted .Text of that images are recognized and each letter of that text are extracted separately with the help of OCR system again with the regionprops().

B. Optical Character Recognition (OCR)

Optical Character Recognition of every single letter in the text goes under scanning, Binarization, segmentation, feature extraction and recognition. The recognized text will be pre-processed with the help of digital scanner i.e. Digitizedan.

C. Neural Network (text-to-speech conversion)

The single character of the text, i.e. text processing is extracted after that the linguistic features of the text are recognized i.e phenomes, duration etc. Vocoder features of the speech signal is extracted. These features can be spectra, spectrogram and fundamental frequency. In this feature, some inherited features of human speech are there. These parameters, with extracted features are inputted into a mathematical mode called a vocoder.

$$p(X) = \prod_{i=0}^{T-1} p(x_{i+1}|x,\ldots,xi) \text{----------(2)}$$

WaveNet is built using stacks of convolutional layers with residual, and skip connections in between. It takes data waveform as input, which then flows through these convolutionlayers and form outputs audio waveform sample.

Vocoder does multiple complex transforms to these features to generate an audio waveform. This waveform is highly modular & workable. Done some approximations to the parameter to generate all kinds of speech in audio.

III. SIMULATION AND RESULTS

The audio processing will be delivered finally along with the text displayed on the screen[3]. The portable text to speech converter or translator is designed to help the visually impaired people to listen to an audio read-back of any scanned text. The advantage of this system is it creates a scanner which scans the whole page containing the text. Then the people no need to give the exact image with text because the software is developed in such way that it can exclude the background by just extracting the text part that is scanned by optical character recognition system. After that extracted text will be converted to speech and will be produced in audio by the neural network i.e. Speech synthesizer.

A. Representation set of features:

A. Captured image input:

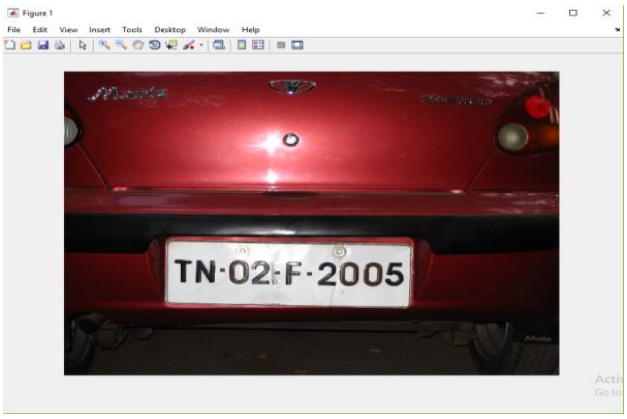

Fig 3: input image1

Fig 4: input image2

We have to take the RGB values for each pixel and make as output a single value reflecting the brightness of that pixel. One such approach is to take the average of the contribution from each channel: This is how the RGB image is converted to gray image.

I=(R+B+G)/3

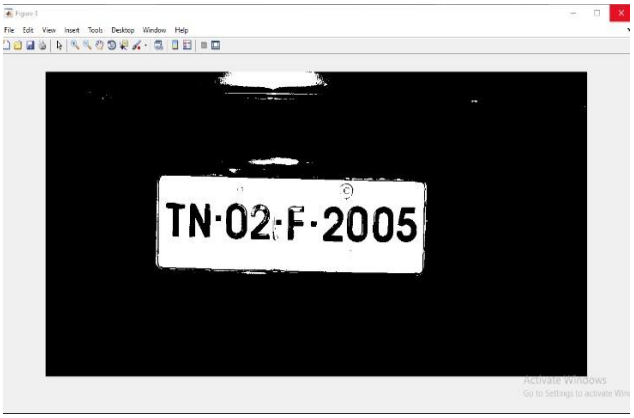

Fig 5: gray image.

Grayscale images, a kind of black-and-white or gray monochrome, are composed exclusively of shades of gray. The text is extracted from the image by ignoring the background by using region props(). The value of each pixel is a single snippet representing only a quantity of light Grayscale image is one in which, carries only intensity information.

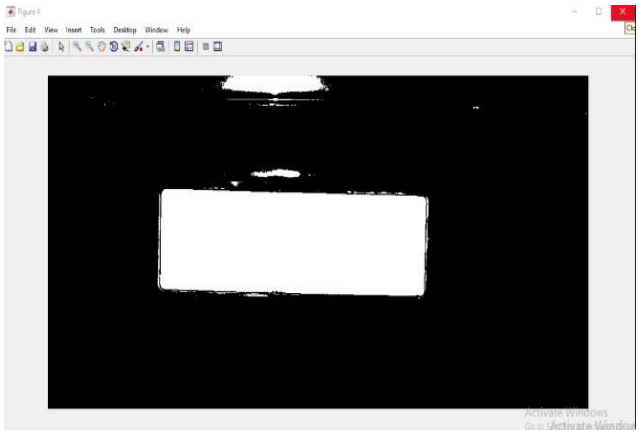

Fig 6: region props image.

The text of the image is extracted separately by just excluding the background using imfill().The fields of the structure array denote different measurements for each region, as specified by properties.The regionprops function returns the centroids in a structure array. s = regionprops(BW,'centroid'); Store the x- and y-coordinates of the centroids into a two-column matrix.Calculate centroids for connected components in the image using regionprops.

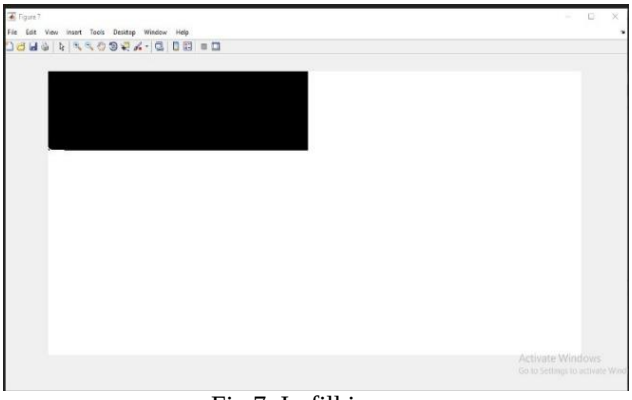

Fig 7: Imfill image.

imfill() fills the area defined by locations, where specifies the connectivity and fills holes in the input binary image.Each letter of the text is extracted by using OCR system and displayed separately in the page and that letteris processed in the software. In this a set of background pixels that cannot be reached by filling in the background from the edge of the image.

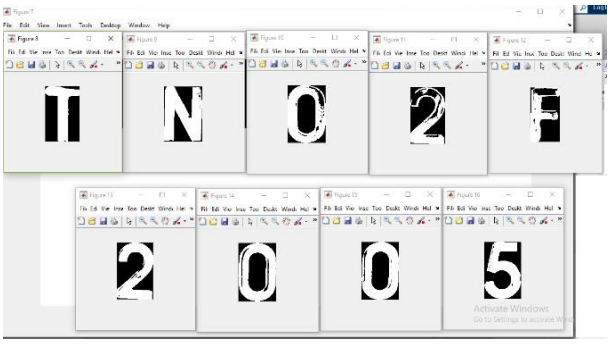

Fig 8: OCR system.

Optical Character Recognition is the machine recognition of printed text characters.An OCR (Optical Character Recognition) system is a computerized scanning system enabling you to scan text documents into an electronic computer file which you can after that edit using a word processor on your computer[9].

The audio is produced finally and even the text is displayed on the screen by using speech synthesis i.e. Deep neural network.

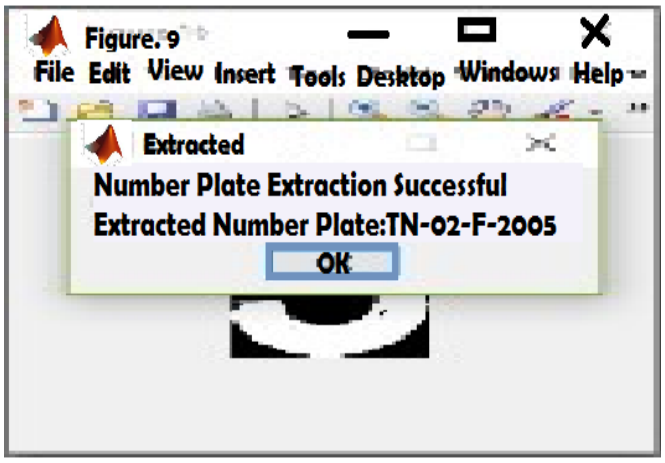

Fig 9: Audio processing.

Audio has been produced in the last final output .Audio signal processing is a subfield of signal processing that is concerned with the electronic manipulation of audio signals. Audio signals are electronic representations of sound waves—longitudinal waves which travel through air, consisting of compressions and rarefactions.

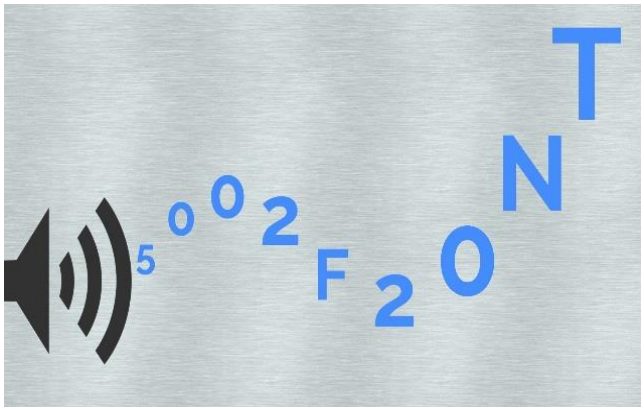

Fig 10: Speech Recognition, Image 1.

Speech recognition is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format. Basic speech recognition has a limited vocabulary of words and phrases, and it may only identify these if they are spoken very clearly.

B. Text input:

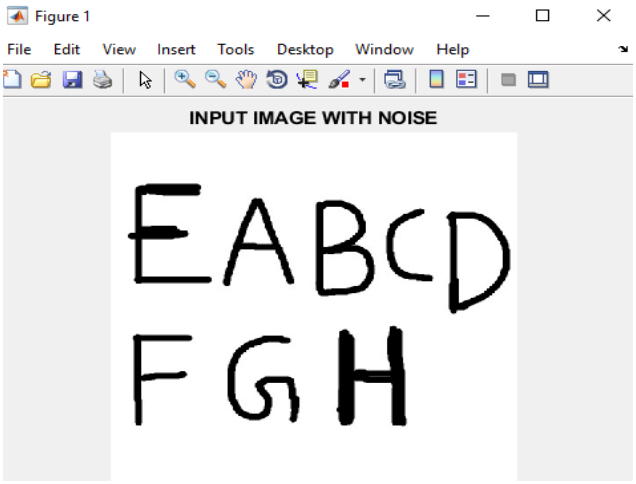

Fig 11: input image3.

Text Recognition is the ability of a computer to receive and interpret intelligible handwritten input from sources such as paper documents, photographs, touch-screens and other devices.The performance of modern text recognition systems implemented as neural networks. They can be trained and are able to read them and only make very few mistakes. Such tasks would be very difficult for most of us: look at Fig. 11extract text from pictures of documents, which you can use to increase accessibility or translate documents.

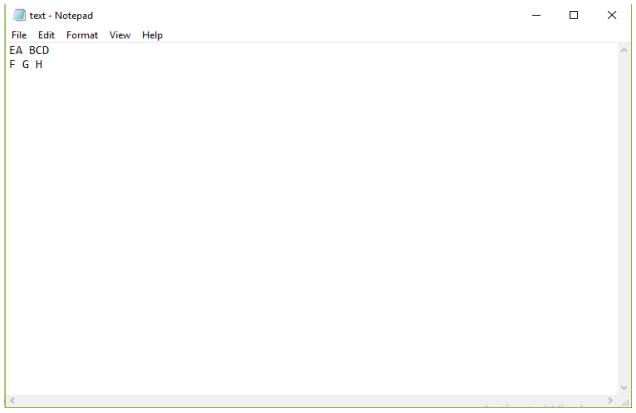

Fig 12: output image 2.

In the following text we'll look at two experiments to get a better understanding of what's happening inside such a neural network.These systems actually work at which parts in the image do these systems look at to identify the text. They exploit some smart patterns.For our first experiment, given an input image and the correct class i.e. direct observation which pixels in the input image elect for and which is against the correct text.

We can compute the influence of a single pixel on the result by comparing the score of the correct category in some scenarios.

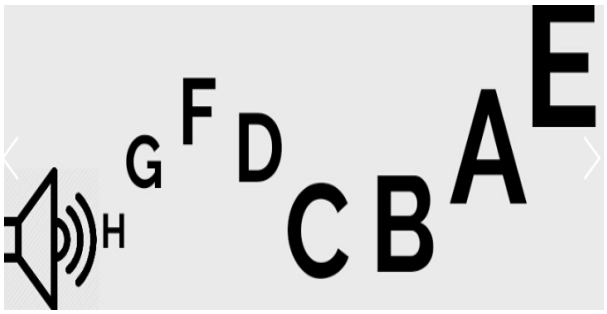

Fig 13: Speech Recognition Image 2.

Speech recognition is an interdisciplinary subfield of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers and also natural language processing, information retrieval[8]. Voice user interfaces refers to identifying the speaker, rather than what they are saying. Recognizing the speaker can simplify the task of translating speech in systems that have been trained on a specific person's voice or it can be used to authenticate or verify the identity of a speaker.

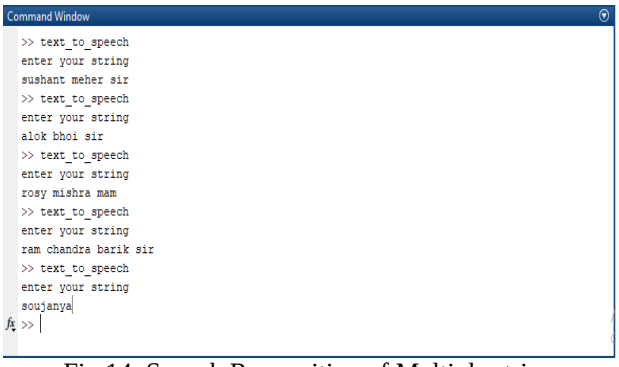

Fig 14: Speech Recognition of Multiple string

Method Append(), store the appended text into a hash set to prevent multiple text. But take note that doing it would prevent the same words, if it look what words were repeated. Its use distinct and except to get the words repeated in a collection.

Let's say its storing the appended text, use distinct make all records unique, and then Except it with the original to get the words removed from distinct and the remaining should be the repeated.

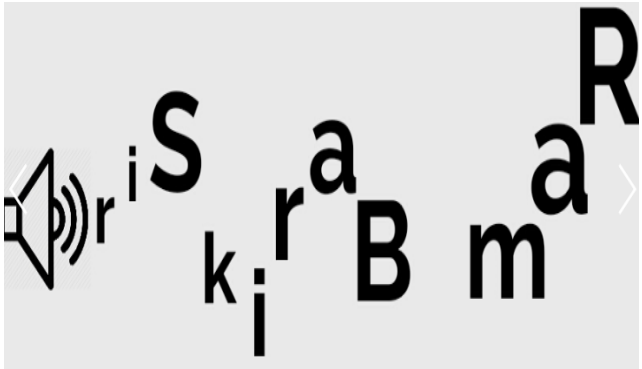

Fig 15: Speech Recognition Image.

The speech recognizer can throw an exception when using a speech recognition grammar that contains duplicate semantic elements with the same key name or multiple semantic elements that could repeatedly modify the value of the same semantic element. The methods provide another mechanism by which you can combine various types to create diversity and flexibility[7]. These methods correspond to the static methods, which are also defined on the class. The order of the parameters determines the order of the element.

B. Recognition table

The performance of the device is high enough and it achieves a readability of less tolerance with the average time processing less than one minute for various paper and font size, with good lighting, the average error rate from the image processing module is better[1]. The proposed system is working effectively for extracting features of the images.

Table 1: Speech Recognition Rate.

Total Images	Text Recognition Rate	Error (TRR)	Speech Recognition Rate	Error (SRR)
IMAGE-1	90%	10%	80%	20%
IMAGE-2	93%	7%	82%	18%
IMAGE-3	97%	3%	90%	10%

IV. CONCLUSION AND FUTURE WORK

The proposed systems are working effectively for extracting features of the images and generate voice. This Text-to-Speech technology for visual impaired people that can change the text image input into sound is implemented in different way[10]. This is a portable device and it does not require internet connection, and can be used independently by people with low vision or visual weakening. This device also has a user interface that allows people to interact easily.

By using digital image processing, and neural network analysis of image and voice recognition is more accurate as well as this method is efficient in terms of cost and time consuming compared to existing techniques. Day by day research work is increasing in this field and various speech processing techniques are implemented in order to get a more accurate result.

REFERENCES

[1] Wouter Gevaert, Georgi Tsenov, Valeri Mladenov, Senior Member, IEEE, "Neural Networks used for Speech Recognition*" in the Journal of Automatic Control, University of Belgrade*, vol. 20:1-7, 2010©.

[2] Malti Bansal, Shivam Sonkar "Text Image to Speech Conversion using Matlab and Microsoft SAPI" *International Journal of Electronics, Electrical and Computational System* IJEECS ISSN 2348-117X Volume 6, Issue 11 November 2017.

[3] Mohd Bilal Ganai1, Er Jyoti Arora2, "Implementation of Text to Speech Conversion Technique" *in International Journal of Innovative Research in Computer and Communication Engineering* Vol. 3, Issue 9, September 2015.

[4] Chaw Su Thu Thu1, Theingi Zin 2, "Implementation of Text to Speech Conversion" *in International Journal of Engineering Research & Technology (IJERT*) ISSN: 2278-0181Vol. 3 Issue 3, March – 2014.

[5]1N.Swetha,2K.Anuradha"Text-to-Speech Conversion" in *International Journal of Advanced Trends in Computer Science and Engineering,* Vol.2 , No.6, Pages : 269-278 (2013)Special Issue (ICETEM) 2013 ISSN 2278-3091.

[6] Jisha Gopinath1, Aravind S2, Pooja Chandran3, "Text to Speech Conversion System using OCR" in *International Journal of Emerging Technology and Advanced Engineering*, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 5, Issue 1, January 2015.

[7] Rubi Debnath, Vivek Hanumante, Disha Bhattacharjee, Deepti Tripathi, Sahadev Roy, "Multilingual Speech Translator using MATLAB" *in International Conference on Electrical, Electronics, Signals, Communication and Optimization* (EESCO) – 2015.

[8] Barik R.C., Mishra R. (2016) "Comparative Analogy on Classification and Clustering of Genomic Signal by a Novel Factor Analysis and F-Score Method". *In: Artificial Intelligence and Evolutionary Computations in Engineering Systems. Advances in Intelligent Systems and Computing*, Vol 394. Springer, New Delhi.

[9] R. C. Barik, R. Pati and H. S. Behera," Robust signal processing compression for clustering of speech waveform and image spectrum", *IEEE International Conference on Communication and Signal Processing*, April 2-4, 2015, India.

[10] Nisha Agrawal , Sanjukta Urma , Sonam Padhan , Ram Ch. Barik, "Indian Agro Based Pest Region Detection by clustering and Pseudo- Color Image Processing" in *International Journal of Engineering Research & Technology* (IJET) ISSN:2278-0181.

