Impact Factor (SJIF): 5.301

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 5, Issue 5, May-2018

Designing and spontaneous recognition of acquired speech signals

[1]Akanksha Singh , [2]S. Kolangiammal, [3]K. Chaitanya , [4]P.G. Vinay

[134]B.Tech-Electronics and Communication Engineering, SRM University , Chennai, India.

[2]Assistant Professor , Sr. G , Electronics and Communication Engineering , SRM university , Chennai , India.

[1]akanksha96@ieee.org , [2]kolangiammal.s@ktr.srmuniv.ac.in, [3]chaitu2012@gmail.com

***Abstract*— The aim of this paper is to present how speech signals can be recognized on the basis of image processing and feature analysis. It involves the processes like detection, extraction and classification. The main focus is to provide the users with user-friendly vocal with the computer and to allow certain handicapped people (such as blind, dumb, poor vision, visual dyslexia) to use the computer or to read any type of document.**

The three major processes involved are Image capturing (the text to be read is captured and processed), Text recognition(text is filtered from image) and finally Speech output (audio of the word or sentence is formed).

Keywords—recognition; image processing; segmentation; extraction; gesture; acquisition(key words)

I. INTRODUCTION

Use of digital technology to represent artistic work in a creative manner is called Digital Art. All the photo edits, paintings that we make using different softwares in our computers are all a part of Digital Art. Not only these drawings but also the sculptures, music/sound art, net art, digital art and virtual reality have also been recognized as artistic practices. These digital visual arts have progressed in the same way as electronically produced musics have been.

The concept of this paper lies in designing a system to produce an audio signal by recognizing the texts from the images and also through hand held gestures. By means of recognition and conversion to speech signals, the physically challenged people such as the blind can go through all the written or printed text documents or books easily just by listening to them.

To achieve that ideology it has been worked in an environment which not only is user friendly but also is accurate enough to recognize texts from different images and also recognize the gestures based on different delay timings. By introducing different environment, such as python, it also increases the efficiency of producing the speech signals and also coding in a more understandable way.

Also not much of a hardware is required since all one has to do is click the picture. Web cameras can be externally attached to the desktops and laptops but these days they are in-built in the laptops which reduce the needed hardware leaving the raspberry pi module to be the only hardware.

These programs run better if the OS is highly supportive and so we are providing Raspbian OS interfaced through the Raspberry module. The possibility of every package running in this OS is very high here.

The texts need not be black and white, it can be RGB too since the algorithm used allows RGB colors and converts them into grayscale. Text is identified and is segregated from the background and is given binary numbers. This text recognition process takes place in segments, also based on threshold value the noise signals are also removed.

The accuracy over gestures recognition is based on the delay timing.

The recognition of the characters is based on the input library that we have stored for the software to relate and match the characters and identify which letter or word it is. The regular characters can be recognized by saving letter from few known fonts. But for the irregular fonts (handwritten) a software called the tesseract is used which draws boundary around the letters and based on its feature factors, it identifies which letter it is.

II. LITERATURE REVIEW

Some research papers were referred which had different methods as well as different concepts. It gave a way to bring out alternative environment to work on. Moreover, our previously contributed works were the ones where similar concepts were dealt.

Rustam Shadiev el [1] had aimed to explore the effectiveness of applying STR(speech to text recognition) in English lectures, thereby, providing a better learning process. The non-native english speaking students who would often find it difficult to understand few words while listening could get themin texts. By this method the students could clarify the words or concepts and have also showed outstanding results outnumbering the native English speaking students.

Yu-Liang Hsu et al. [2] suggested the use of inertial pen through which we could write on a board and respective text could be recognized. This digital pen used DTW (Dynamic Time Warping) algorithm for both handwriting and gesture recognition. By holding the pen, the users could write the numerals or lower case letters or even make the gestures in their preferred style. These signals were wirelessly transmitted and recognized online. Its procedures involve signal acquisition, signal pre-processing, motion detection, template selection and recognition. It has taken this technology to next level by eradicating the usage of traditional pen and board.

Hyung II et al. [3] suggested that the camera based text recognition is going very popular these days and is in great demand. He used state estimation (a method of state selection) in his paper to carry forward the work which focussed on the scene text detection problem. A text line detection algorithm was used for the camera captured documents which contributed in major development in this field.

Koki Yoshida et al. [4] focussed on the recent trends in growing craze on E- books. This concept became popular because these E- books have features that the traditional books don't have. One can share what they are reading and one can highlight the text as well. But even though its growing trend, there are still many people who are fond of reading paper books. So this paper basically deals with the proposal on information adding system using a projector camera system and has proved to be really effective.

These texts could be recognized in different languages too, is what one paper concluded. Hence this led to the idea of working on popular, national and regional languages as a part its advancements.

III. PROPOSED CONCEPT

The interest arose as a part of advancement in the previous work which also had gesture recognition ideologies. By means of eye blink sensor and hand gestures, the user can control the mouse pointer and also open/close any folder without even touching the screens or the mouse.

Extending this idea of using the gestures helped finding the answer to the question on how user friendly and helpful it would be if one writes through the gestures and text is obtained in a word document.

This thought brought light on the various applications of such a device and also satisfying the physically challenged group of the society where we live in. Also, it is not limited to certain characters or one or two languages. This a whole set of dictionary loaded with different fonts, styles, textures etc to be recognised. Moreover, this can obtain speeches in many languages which proves the fact that this project can also be used for native and non-native language speakers using the chart shown in Fig. 1.

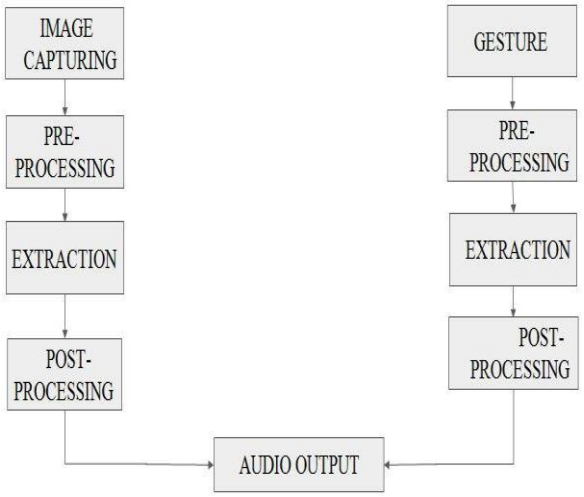

Fig. 1 Flow chart

The system consists of Raspberry Pi system module, as shown in Fig. 2, running on a Raspbian OS which is a type of linux environment. It is used because it is comparatively more flexible and user-friendly at the same time. It is interfaced with a camera which helps in capturing the gestures and the images. The camera can be attached externally or the inbuilt webcams can be used effectively for this purpose.

According to Fig. 1, a monitor is needed to display the output. Although, one can even connect laptops for increasing the portability of the system. SD card is used to store the the required data such as compiling codes, images captured, output files etc. Any coloured (specific) object can be used for gesture recognition. The particular colour can be fed in the codes so that the program recognises it while interpreting the codes and giving the outputs. The software used in in here is Python. Again, this is used because it is flexible and can be modified as per the need easily. The inbuilt libraries make it an effective and efficient way of writing the codes. The OpenCV library (Computer Vision) is used for such user interfaces to make the task a lot easier.

Efforts have been made so that the coloured objects can be recognized instead of the traditional ways of dealing with the grayscale images only and geometric algorithms are used to carry forward the project which give accurate data based on the input data sets.

For image to text conversion, firstly image is converted into gray scale image and then black and white image. After that it is converted into text. Microsoft Win 32 SAPI library has been used to build speech enabled applications, which retrieve the voice and audio output information available for computer.

The developed software has set all policies corresponding to each and every alphabet, its pronunciation methodology and also the way it is used in grammar and dictionary. In general, a particular signal has a specific power requirements which are shown in different graphs. These graphs are understood by the software developed and recognized as distinct words.

This approach can be used in part as well (if not whole). If a user wants only image to text conversion, then it can be possible and if the user wants only text to speech conversion, then it is possible too.

IV. SYSTEM ARCHITECTURE

A. Specifications

A **Raspberry pi** module is used in order to interface the OS and external hardwares such as web camera, audio jacks etc. All the packages with respect to the program are installed and stored using the SD card. The module has a CPU of 1.2GHz frequency with 4-ARM cortex A-53. A Broadcom videocore IV GPU with 1GB RAM is being used. A networking slot allowing ethernet connection with 2.4GHz frequency and following 802.11 IEEE

protocol is present. Also the module has a 40 pin GPIO header, HDMI slots, 3.5mm audio-video jack slot and 4-USB 2.0 slots to interface the camera and the display.

The **web camera** is interfaced with the raspberry pi module in order to capture images or capture the gestures. For laptops, these are inbuilt but for desktop

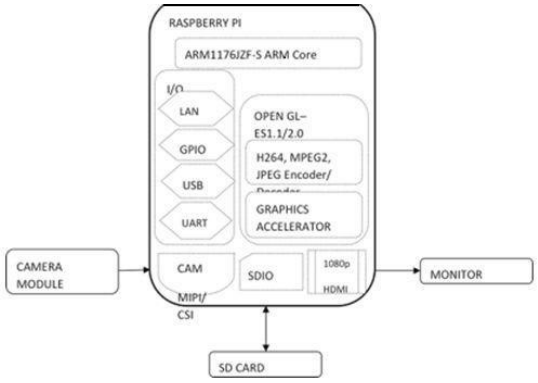

Fig. 2 Raspberry pi module

it has to be externally interfaced. Basic specifications of such a camera is expected to have a 60° field of view, a focal length minimum of 4.0 and a minimum of 1.2 MP camera to capture better images which help reading the texts better and eradicate the issues such as sensitivity to light and stability. The resolution of camera would lie around 1280*960.

The main output of this paper is the audio signal, in order to listen to that we must interface an earphone or a speaker to the AV jack provided in the module.

The other types of hardware used is the **SD card** which is used to store all the packages and ethernet cable which helps connecting to the raspbian OS through desktop monitoring.The hardware setup is shown in Fig. 3 which shows the block diagram of the working of the system.

Coming to the software specifications, the application has been created in **Python** environment. Since it is less complex, its one of the reasons why its is widely used. Also the library **OpenCV** (Open Computer Vision) is a library that helps code or command human interactive programs in a less complex and much easily understandable way.

The **Raspbian OS,** in which the python programs are being worked on, is connected through the module by means of desktop monitoring. It is based on debian linux. This OS includes tools for browsing, python programming and GUI desktop. It was built using the X- window system software with point-click interface.
The next most important software used is the **Tesseract**. This software consists the feature factors of every letter i.e both regular and irregular characters. This is an advantage as it allows a wide range of fonts and styles for the same letter itself.

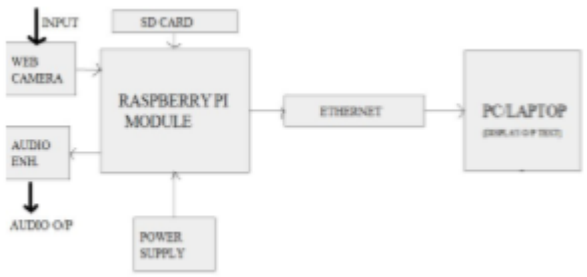

Fig. 3 Block diagram of the system module

B. Concept and Algorithm

There are many popular techniques for image segmentation, namely, Thresholding method, edge detection based, region based techniques, clustering based techniques, watershed based techniques, partial differential equation based and artificial neural network based techniques etc.

Since RGB letters and words will be dealt in this phase, the technique which will be used is the **Thresholding** method where the user can obtain a binary image from a grayscale image as it is the simplest method if image segmentation. The image intensity is represented by $\mathbf{I}_{\mathbf{I,J}}$. Each pixel in an image is replaced by a black pixel if I is less than $\mathbf{T}_{I,J}$ (a fixed constant) that is, $\mathbf{I < T}$ or by a white pixel $\mathbf{I,J}$ otherwise.

Furthermore, the thresholding methods are categorised into many parts such as histogram shape, clustering, entropy, object attribute, spatial as well as local. The **Clustering method** has been chosen where the clustering of gray-level samples are done in two parts as background and foreground (object) or alternatively are modelled as a mixture of two Gaussians.

Under this method, the **Estimation** Algorithm is used (Expectation-Maximization) to overcome the main difficulty in the **Gaussian Mixture Model.** The problem is that one doesn't know that which the data points come from which latent component. If this data is collected, it would become very easy to separate and mark Gaussian distribution for each set of points.

To incorporate information about the covariance structure of the data, one can go for the concept of the **K-means.** One can follow the steps given below by referring to Fig. 4 given below:

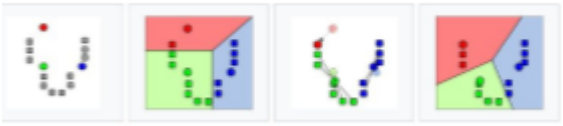

Fig. 4 K- means concept in steps

1) K initial "means" (in this case k=3) are generated randomly within the domain of data.
2) K clusters are created by associating every observation with the nearest mean. **Voronoi** diagram is used to represent the partitions in the above figure.
3) The centroid of each of the new clusters become the new mean.
4) Steps 2 and 3 are repeated until the convergence is reached.

V. WORKING

A. Methodology

The system model includes the following steps of operation:

1) Pre-processing
2) Segmentation
3) Feature extraction
4) Classification
5) Recognition
6) Post processing

Pre-processing is the step where the image is captured and the foreground and background pixels are differentiated. Segmentation algorithm is used in this case to get a clearer picture of the text in three steps

excluding noise using suitable threshold values. The features are extracted if in case gestures are being used, if not then the text is extracted as foreground and background is neglected. Classification is done as to what the portrayed letter or word actually is. Post processing is the process where the output is found as per the need, i.e. audio signal in this case.

The methodology includes the following processes to be followed in detail:

Image processing: It is processing of the images using mathematical operations by using any form of signal processing for which the input is an image, a series of images, or a video, such as a photograph or video frame; Computer vision, on the other hand, is often considered high-level image processing out of which a machine intends to decipher the physical contents of an image or a sequence of images.

Digital image processing is the use of computer algorithms to perform image processing on digital images.

Object recognition:- A typical image recognition system consists of pre-processing, segmentation, feature extraction, classification and recognition, and post processing stages. In general, image recognition is classified into two types as offline and online image recognition methods. In the off-line recognition, the writing is usually captured optically by a scanner and the completed writing is available as an image.

Image acquisition :- In Image acquisition, the recognition system acquires a scanned image as an input image. This image is acquired through a scanner, digital camera or any other suitable digital input device.

Pre- processing:- The pre-processing is a series of operations performed on the scanned input image. It essentially enhances the image rendering it suitable for segmentation.

Image used in image processing:- Image recognition enhances the processing of scanned images by allowing you to automatically recognize and extract text content from different data fields. For example, when the user scans a form and uses a document imaging software to process it, CR allows the user to transfer information directly from the document to an electronic database.

Image recognition:- Since commercial IR engines achieve high recognition performance when processing black and white images at high resolution. To extend the recognition capability of the IR for image and video text, the main research efforts focus on text segmentation and enhancement.

Segmentation:- In the segmentation stage, an image of sequence of images is decomposed into sub-images of individual image. The labelling provides information about number of images in the image. Each individual image is uniformly resized into pixels for classification and recognition stage. Image segmentation methods are performed on the extracted image regions to remove the background surrounding text images. However, these methods are unable to filter out background regions with similar grayscale values to the images.

B. Data sets

The data sets are the inputs that are being given to test the system and get the desired output.

Fig. 5 Input data sets taken

Here are some of the input samples as shown in Fig. 5 which is used to get the audio output after text recognition. One can hold the picture with the text written on it and camera will capture the image, recognise the text and give the output.

C. Output

The mode is chosen as 1 or 2 depending upon which feature is to be used. If 1 is used, then it is for the image to audio conversion as shown in Fig. 6. If mode 2 is used, then it is for gesture to audio conversion in Fig. 7.

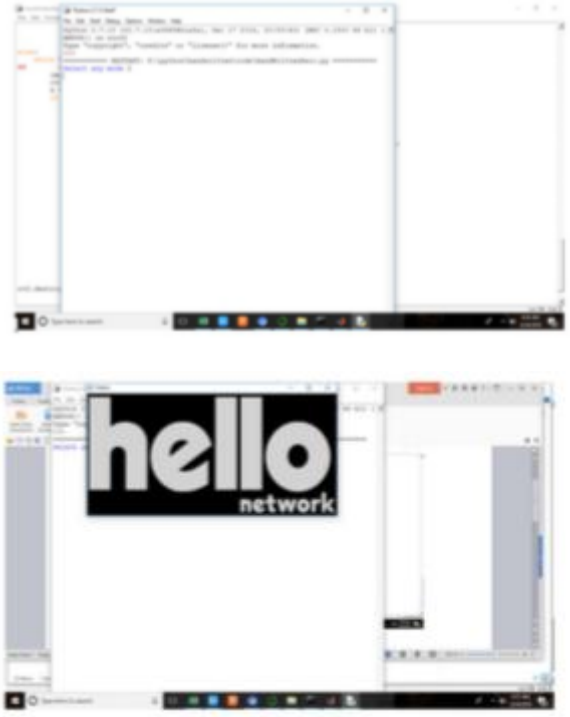

Fig. 6 output when module 1 is selected for image recognition and the text is extracted.

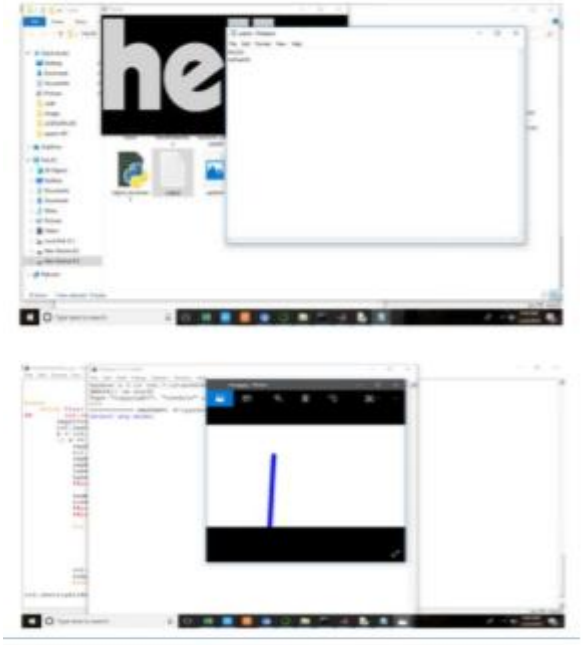

Fig. 7 output when module 2 is chosen for gesture recognition and text extraction.

VI. CONCLUSION

The project represents recognition of speech signal based on the extraction of features of every character. The recognition system involves detection of the signals, extraction and classification of the signals.

The mentioned library allows selecting voice and audio device one would like to use. People with poor vision or visual dyslexia or totally blindness can use this approach for reading the documents and books easily. People with speech loss or totally dumb person can utilize this approach to turn typed words into vocalization. Certain hi-tech applications are also using this methodology where the men can not physically go.

Since printed document images are more of historical and poor in quality, recognition results in a number of errors. Also ,there is a scope of developing a Multilingual system so that we can read more than one language documents. The facility to stop, pause and continue reading can be provided, i.e., the user should be able to pause the synthesizer at any time and then continue reading using just a mouse-click.
The simulated results have shown that the filter based feature extraction gives much better accuracy with lesser algorithmic complexity than other speech expression recognition approaches. It also has decreased the usage of heavy hardware.

REFERENCES

[1] Rustam Shadiev,Yueh-Min Huang, Wu-Yuin Hwang,Narzikul Shadiev, Department of Engineering Science,"National Cheng Kung University, Taiwan.Year: 2015.2015 IEEE 15th International Conference on *Advanced Learning Technologies*).

[2] Yu-Liang Hsu, Cheng-Ling Chu, Yi-Ju Tsai, and Jeen-Shing Wang, Member, IEEE.Year: 2015.IEEE SENSORS JOURNAL, VOL. 15, NO. 1, JANUARY 2015.

[3] Hyung Il Koo.Year: 2016.IEEE Transactions on *Image Processing*(Volume: 25, Issue: 11, Nov. 2016.

[4] Koki Yoshida; Hirotake Yamazoe; Joo-Ho Lee.Year: 2017.*Ubiquitous Robots and Ambient Intelligence* (URAI), 2017 14th International Conference.

[5] Hilal Kandemir,Büşra Cantürk, Bigisayar Mühendisliği Bölümü, Turgut Özal Üniversitesi, Ankara, Türkiye,Muhammet Baştan,*Signal Processing and Communication Application Conference* (SIU), 2016 24th.Year: 2016.