e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:07/July-2022 Impact Factor- 6.752 www.irjmets.com

RAY OF LIGHT, AN INTEGRATED SYSTEM FOR SPECIALLY ABLED

Prof. H.D Nandeesh*1, Suchith J Prakash*2, Skanda Tejaswi D*3, Nishanth S*4, Prajwal J*5

*1Professor, Department Of Computer Science And Engineering, SJCE, Mysuru, Karnataka, India.

*2,3,4,5Student, Department Of Computer Science And Engineering, SJCE, Mysuru, Karnataka, India.

ABSTRACT

Ray of light is a IOT based assistant system for specially abled people like deaf, dumb and blind. This solution is to address the problems of these people in a single device itself, so the goal is to create a device solution that is simple to use, efficient, rapid, and accurate. The study offers a Raspberry Pi and Google API-based assistance for the blind, deaf, and dumb. For communication in this project, we have three different scenarios: dumb, blind, and deaf. A dumb person who can't speak utilises text-to-speech technology. A deaf individual who cannot hear uses speech-to-text. Those who are blind and cannot see utilize an image-to-speech stack. Between these components, there is a Web Interface for smooth operation between these modules.

Keywords: IOT, Raspberry Pi, Google API, Speech Recognition, Tacotron, Pytesseract, Web Interface.

I. INTRODUCTION

A total of 1.3 billion people worldwide suffer from some form of vision impairment, of which 188.5 million have mild symptoms, 217 million have moderate to severe symptoms, 36 million are blind, and the majority are over the age of 50. The biggest concentration of blind individuals is thought to reside in India. There are 9.1 billion mute and deaf people worldwide. WHO estimates that 466 million individuals, or about 5% of the world's population, have hearing loss that is incapacitating. Although for some reason, our way of life has not given much thought to those who are differently abled. Despite not having access to scientific developments, they nonetheless encounter numerous issues on a daily basis.

The vital element of human existence is interaction. Unfortunately, there is still a gap. Individuals engage through Braille and sign language, but they find these methods cumbersome. They are effectively pressured to master these customary forms of communication or need outside assistance, like another person. This essay primarily focuses on bridging that gap by attempting to give them a sense of liberty and the ability to move among other everyday people.

The Tacotron model is used to translate text into emotion-based speech since a dumb individual is unable to talk. Blind people cannot see, so Pytesseract, Tacotron, and Google API models are used to convert images to speech. Deaf people cannot hear, so speech recognition is used to convert speech to text. Located between these parts is a web interface.

II. METHODOLOGY

As we examine the methodologies used in this project, let's first have a look at the essential hardware and software requirements that we need.

Requirements For Systems : Since our product has three main features, in each mode user interface has its know importance as we are dealing with the data that is provided by the end users. In deaf mode the resultant text that is the audio that is converted is displayed in Monitor. In dumb mode user needs the keyboard to give the input text. In Blind mode, the audio output is heard in high quality audio speaker. So all these interactions is done using the web application as well. So the browser that runs HTML,CSS as well as Javascript is a must.

The hardware requirements are Raspberry Pi, Camera Module, Monitor, Microphone, and Speaker. Operating system, platform requirements, Python compiler (3.0.0 and above), and vscode are the software prerequisites (Text editor). For the web interface, Google API, Tkinter, and Python libraries, we use the Django (python) backend programming language. In the Python libraries, we use the Data analysis and visualization (pandas, numpy, matplotlib, seaborn) and AI ML parts (scikit-learn, tensorflow, keras).

The entire procedure is broken down into three modules: dumb, blind, and deaf. Let's take a closer look at each module.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:07/July-2022 Impact Factor- 6.752 www.irjmets.com

Dumb Module: The user's choice to activate dumb mode does so. Here, we transcribe text into speech to facilitate communication for those with vocal problems. The model we use for speech synthesis is tacotron.The accompanying spectrogram is produced from input by the character sequence-to-spectrogram end-to-end generative text-to-speech model known as Tacotron. Tacotron is built on a seq2seq paradigm with attention. The model consists of an encoder, an attention-based decoder, and a post-processing net.The model creates spectrogram frames from input characters, which are subsequently converted into waveforms. We are utilizing this pre-trained model for improved accuracy because it has been trained and all the other parameters have been set.The model outputs audio after receiving text and an emotion type as input. The text that is used as input is delivered to Firebase, where the tacotron model is used to process it and turn it into audio. The audio file is then loaded into the Raspberry Pi, which plays the audio through the connected external speaker. As a result, speech is the output.

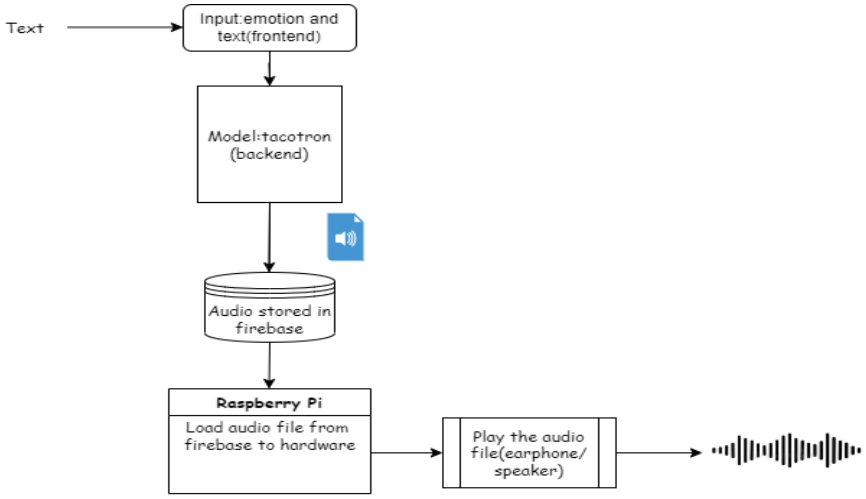

Figure 1: Flowchart of Dumb module.

Blind Module : Change the text-filled visuals we regularly encounter into voices so that blind people may interpret the information included in them. This system is divided into two components. OCR (optical character recognition) is used to first convert images to text, and then the tacotron model is used to turn text to audio. In this mode, we use the Raspberry Pi's attached camera to take the photographs. The firebase receives the image and begins processing it there. The OCR model, which has been trained to recognise the text in images, is then sent the image. Once the OCR model has produced its output, it is given to the tacotron model, which creates voice from the text. The Raspberry Pi receives the output audio file and plays the speech through the speaker.

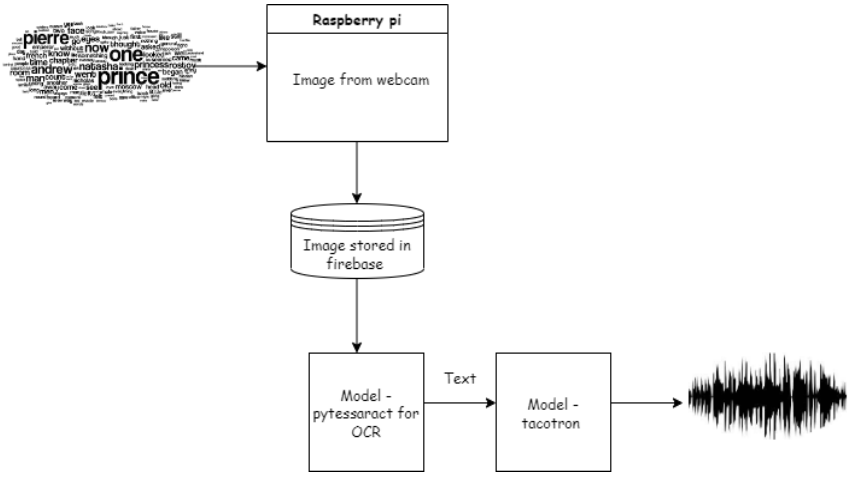

Figure 2: Flowchart of Blind Module.

Deaf Module : Speech to text conversion occurs here. It is a high priority because it is an integrated system and all three modes must function properly when used concurrently. The option has been set to deaf mode. The product's USB microphone, which is connected to the Raspberry Pi, records sounds or words spoken to the

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:07/July-2022 Impact Factor- 6.752 www.irjmets.com

user, who in this case may be deaf, and stores them as an mp3 file.We will utilize a subset of the Speech Instructions dataset, which contains brief (one-second or fewer) audio clips of commands like "down," "go," "left," "no," "right," "stop," "up," and "yes," to create and train a simple automatic speech recognition model for various terms.

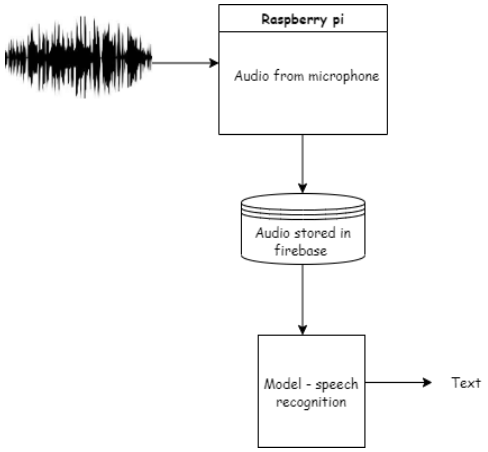

Figure 3: Flowchart of Deaf Module.

III. MODELING AND ANALYSIS

For a better understanding, let's look at each of the three popular models one at a time. As we saw in our previous discussion, these models require thorough examination.

1. Tacotron Model with Emotion Synthesis (For Dumb Module) :

1.1 Prerequisites : However, since we are employing pre-trained models in this case, a dataset is not necessary. Let's look at how pre-trained models work in our scenario before moving on to the methods involved in this model.Models that had been trained and saved with their weights. After that, use the bottom levels of the network once more; this is transfer learning.

- Tacotron model: Tacotron is a full-featured algorithmic text-to-speech model that builds speech fully from scratch using characters. It scales nicely to the use of enormous amounts of audio data along with transcripts because there is no necessity for phoneme-level alignment. Before the text-to-speech (TTS) interface, an acoustical model and a sound generation module were employed.
- Emotion Synthesis: No one ever communicates emotionlessly. The categorical system, which distinguishes between distinct emotion categories like anger, fear, or boredom, or the usage of emotional dimensions like arousal, valence, or dominance, is typically used to model emotional expression. The categorical paradigm benefits from being well established in regular human communication and being intuitively simple to understand. Aspects of emotional speech synthesis thus play a crucial role, ranging from use cases and emotion models to advanced technologies.

1.2 Model Architecture :

- Text is given to the Tacotron model.
- Characters serve as the tacotron model's input. To obtain encoded features, character embedding (512 x 1 vectors) is done, after which CNN (3 Convolution Layers consist of 5D Filters with each layer followed by Batch Normalization + reLU) and LSTM (512 units) are applied. IFR (Internal Feature Representation) is obtained after passing this via the Attention Unit (Generates 128-D Context Vectors) .This wraps up the encoder part.
- Next comes the Decoder part,Attention Unit (Generates 128-D Context Vectors) output along with Pre-Net (Two FC Layers of 256 neurons + reLU) output is given to LSTM Layers (2 Uni-Directional Layers with 1024 neurons). This output is given to Linear Transform to get Predicted Spectrogram Frame.Next That is given to Post-Net (5 Convolution Layers + Batch Normalization + tanh) to get Enhanced Prediction
- Next, that output is given to Modified WaveNet (30 Dilated Convolution Layers along with PixelCNN++) to get a Speech Waveform.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:07/July-2022 Impact Factor- 6.752 www.irjmets.com

- This finishes the Tacotron part.
- Next the other input is emotion is given to a pre-trained model(The models are trained using emotion samples from the emov database.) to get Emotion Frequency for that emotion.
- Speech Waveform and Emotion Frequency are the results of the models that were just mentioned. To create the final speech, which is an Audio.wav file, Speech Waveform and Emotion Frequency are combined.
- A flowchart of the complete procedure is shown in the picture below.

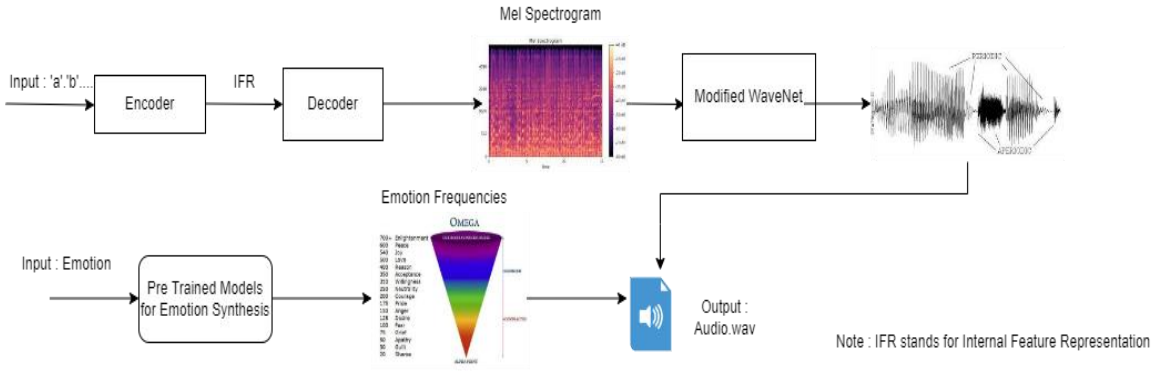

Figure 4 : Flowchart of Tacotron Model with Emotion Synthesis.

2. Optical Character Recognition (For Blind Module) :

2.1 Prerequisites : In our project, we're developing a character-based OCR model. We'll use two datasets for that. Many deep learning frameworks, including TensorFlow, Pytorch, and Keras, currently include the Standard MNIST dataset. We can identify the digits 0 through 9 using the MNIST dataset. Each image consists of a single digit of grayscale photos measuring 28 by 28. Since the capital letters A–Z are missing from MNIST, we are utilizing a different dataset that rescales the capital letters from NIST Special Database 19 to be 28 × 28 grayscale pixels to match the format of our MNIST data.

2.2 Model Architecture :

- The MNIST dataset and A-Z capital letters from the NIST Special Database are essential to our model.
- Those datasets are loaded and combined. Merge two datasets for the model's input. To make model fitting easier, convert labels from integer to vector. See the dataset's count of character weights as well as the class weights for each label.Merge two datasets for the model's input. To make model fitting easier, convert labels from integer to vector. See the dataset's count of character weights as well as the class weights for each label.
- By enhancing the training data input using ImageDataGenerator, we may enhance the performance of our ResNet classifier. We manipulate the images' horizontal and vertical translations, tilts, and other scaling rotations and size changes. We are building the ResNet architecture, the core of our project.
- The 1x1 Convolution Layer constitutes the initial building block of the ResNet module. The 3x3 Convolution Layer is the 2nd part of the ResNet module. Another set of 1x1 Convolution Layers makes up the third block of the ResNet module. Apply a Convolution Layer to the shortcut if we want to reduce the size of the spatial area. The last Convolution Layers and the shortcut added together likewise return to that state.
- After setting the input, run BatchNormalization. Check to see if the CIFAR dataset is being used. Use only one Convolution Layer. Verify that the Tiny ImageNet dataset is being used. Apply Convolution Layer, BatchNormalization, Activation Function (reLU), and MaxPooling to cut down on spatial size.
- Further, BatchNormalization,Activation Function (reLU), and AveragePooling are applied. Softmax classifier at last for multi-class output.
- Create the model,train the model,get text from images,that text is given to Tacotron Model with Emotion Synthesis to get Audio.wav as final output.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:07/July-2022 Impact Factor- 6.752 www.irjmets.com

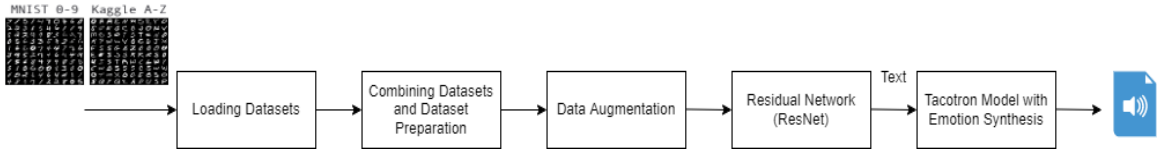

Figure 5: Flowchart of Optical Character Recognition.

3. Speech Recognition (For Deaf Mode) :

3.1 Prerequisites : Over 1.5 lakh WAV (Waveform) audio files of humans speaking 35 distinct words make up the original dataset. They are simply verbal instructions with a single word.There may already be too many screens all around us. Every day, it seems, fresh iterations of typical items are "re-invented" with wifi built-in and vibrant touchscreens. Voice interfaces provide a viable alternative to our screen addiction.

3.2 Model Architecture :

- Load the dataset and see the sampling rate and change the sampling rate.
- Pre-processing consists of two steps, mainly,Resampling and Removing less-than-one-second commands.
- Since this is a multi-classification problem, convert the output labels to integer encoding and then transform the integer encoding to a one-hot vector.
- 4 Conv1D with Activation Function (reLU) and Dropout of 0.3.Two Dense Layers with Dropout of 0.3.At last Activation Function (softmax) to get multi class output.
- Since there are multiple classifications involved, categorical cross-entropy should be used to define the loss function.
- The callbacks for early stopping and model checkpoints are used to save the best model after each epoch and to stop training the neural network at the appropriate moment.
- Now let us train the model with a 32 number batch size and assess its effectiveness on the holdout set.
- Create the model, run the model to get text output.

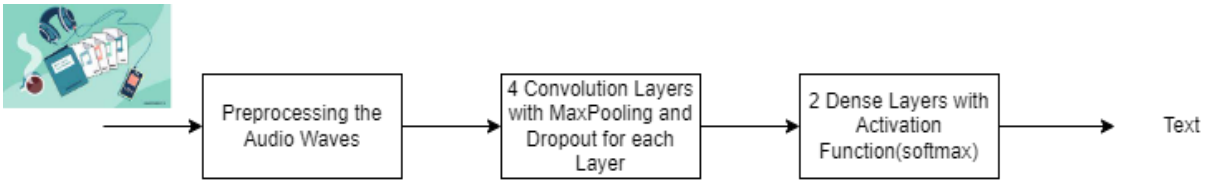

Figure 6: Flowchart of Speech Recognition.

IV. RESULTS AND DISCUSSION

The accuracy of all the discussed models are given in Table 1.But there are a few things about each model we've examined so far that need to be discussed.

- In the Tacotron Model with Emotion Synthesis For the first time, we witnessed excellent, comprehensible communication of an emotion! The emotion of anger was being produced quite effectively! However, the previous issues persisted with the other emotions. For all other emotions, the spectrograms that were created remained empty.The two hardest emotions to learn were amusement and sleepiness. This was due to the existence of non-verbal cues, which are missing from the transcripts, such as giggling and yawning. Regarding these feelings, the preprint made a similar statement. When we plotted the mel-spectrograms of several ground-truth samples, we found that the perceptual distinction between succeeding temporal frames for Disgust was lower than for Anger on the temporal axis. This, in our opinion, is what prevented the model from producing Disgust correctly. This is only a hunch, though.
- Although the effectiveness of the Optical Character Recognition model was higher, the prediction output for complicated photos containing text was not as good because the text in those images was too faded or dark. Therefore, in those circumstances, we employed pytesseract.
- The speech recognition model was only trained to recognise vocal commands, but we also needed it to recognise complete sentences.Then we carried out this task using the gTTS Google API. High-end computers can cut down on the processing time, which is typically between three and five seconds.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:07/July-2022 Impact Factor- 6.752 www.irjmets.com

Table 1. Model Performance

SN.	Model Type	Accuracy
1	Tacotron Model with Emotion Synthesis	95%
2	Optical Character Recognition	94%
3	Speech Recognition	85%

V. CONCLUSION

A field prototype for aiding the blind, deaf, and hard of hearing has been created as a result of this research. This initiative uses resources and space effectively while simultaneously aiming to empower and help people with impairments. Since every component of the device is affordable and useful, it is a less expensive solution. The most recent and well-liked innovations have made this item portable, adaptable, and convenient.

What we intend to create is an integrated help system with improved features and effective algorithms. Using our camera module, we can expand our project to include sign language support. People with diverse abilities can now use these improvements to meet their unique needs. The concept can be expanded by making the tool wearable and more portable to make it simpler for users to operate.

ACKNOWLEDGEMENTS

First and foremost, we would like to express our gratitude to Prof. H.D. Nandeesh, our lecturer, mentor, and project advisor, without whose help and direction at each stage, this endeavor would not have been successful. We would also like to express our gratitude to Dr. Srinath S, our Head of Department, for continually encouraging us to investigate new fields. Finally, we also want to express our gratitude to our family and friends for their support and assistance in helping us achieve our objective.

VI. REFERENCES

[1] Rastogi, Rohit, Shashank Mittal, and Sajan Agarwal. "A novel approach for communication among Blind, Deaf and Dumb people." 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2015.2015, pp. 605-610.

[2] Kumar, K. Naveen, P. Surendranath, and K. Shekar. "Assistive Device for Blind Deaf and Dumb People using Raspberry-pi." Imperial Journal of Interdisciplinary Research (IJIR) 3.6 (2017).

[3] Suvarna Nandyal, Shireen Kausar,Raspberry Pi Based Assistive Communication System for Deaf, Dumb and Blind:(2019).

[4] Deepak Sharma, KenilVora, Shivam Shukla, “HAND ASSISTIVE DEVICE FOR DEAF AND DUMB PEOPLE.”, International Journal of Advances Research.