Sign language to Speech conversion – Digital voice for Speech Impaired people

[1]Sumit Ramesh Mahajan, [2]Pratik Ashok Kante, [3]Parikshit Anil Pardeshi, [4]Pratik Raju Nikam, [5]Dr. G.S.Raghtate
[1,2,3,4]Student, [5]Asso. Professor
[1,2,3,4,5]Information Technology Department
[1,2,3,4,5]Datta Meghe College of Engineering, Airoli, Navi Mumbai

Abstract - As communication plays a very important role in human society, it's very difficult for people with speech issues (Mute people) to communicate with ordinary people. Sign language plays a vital role here as it uses the hand and body gestures for communication. But due to its learning difficulty and it not being very common, it is difficult for ordinary people to understand it. So, to overcome this issue, there is a need for a system which can translate the sign language into speech(voice). In this proposed system, the dataset of hand gesture images for A-to-Z letters is collected. Also, a custom gesture dataset for common day to day sentences is generated. Then the dataset is been through image preprocessing which made images ready for further feature extraction. Images are converted into grayscale images. As binary images contain minimum data, they are considered as an input for model. Convolutional Neural Network (CNN) model is used for the classification of the input. The output given by the model is converted to audio using python libraries. Also, a user-friendly UI is added in the system for people to easily interact with system.

Index Terms - Sign Language, Gesture Recognition, Convolutional Neural Network (CNN), Speech Impaired, Grayscale Image, ISL.

I. INTRODUCTION

Communication is very important requirement for surviving in society. Sign language is key technique of communication for speech impaired people. It makes use of hand gestures instead of voice to deliver the meaning. It has predefined gestures for various letters and sentences which have meaning associated with it. Most of the ordinary people cannot understand the sign language which makes it difficult for both speech impaired and ordinary people to communicate. This leads to the requirement of sign language interpreter who can act as a translator for people. However, such interpreters are limited. Also, there are different types of sign languages according to different nations. Examples of such languages are Indian sign language (ISL), American sign language (ASL), British sign language (BSL), etc. This results in development of a computer system which can automatically translate the sign language to respective audio. Various approaches are used before to tackle this problem in past times. Using the hand gloves which have sensors attached to it is one of the famous techniques. After the advancements in domains like computer vision and deep learning, using computer and camera-based system is the most effective technique.

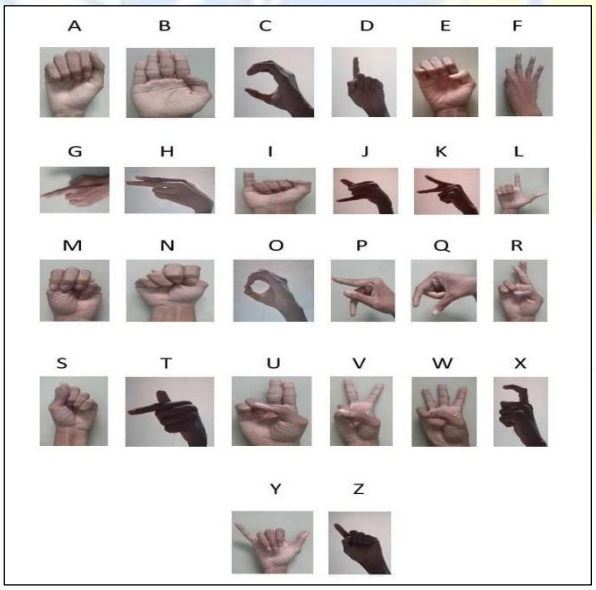

Fig.1 Hand gestures for ISL of alphabets from A - Z

The methodology proposed in this paper uses computer vision and deep learning technology to convert the hand gestures into audio. The system successfully classifies 26 alphabets of ISL (as shown in fig.1) and 14 predefined sentences with accuracy of 96%. Convolutional Neural Networks are used to classify the different gestures. After accurately classifying the gesture, the given output is converted into audio using pytts library. pyQT5 and tkinter is used to design the UI for application. The system also provides an option for custom sentences where user can create a new gesture and attach the meaning to it. That gesture then be used for communication. The further sections explain the related work and Implementation details of the project. Also, the generated results and future scope of project is also discussed in the further sections.

II. LITERATURE SURVEY

There has been different work in same field of sign language recognition field with different approach towards hand sign gesture recognition. A study of many different existing system based on sign language has been done to design a system that is able to recognize sign gesture and convert it into audio format.

In [1] the system was accurately able to track hand gesture of sign demonstrator for converting it into text form using different technique like object stabilization, face elimination, skin color extraction and then hand extraction. The system uses k-NN model to and HMM (hidden Markov model) chain for each hand gesture. This system work on real time Indian sign language recognition. The system is able to recognize one handed sign gestures of standard alphabet (A to Z). The output of the system is very efficient, consistent and able to translate sign to text.

In [2] which using CNN model to predict the hand sign gesture accurately and for taking both handed gesture in one frame. This system is able to recognize hand gesture and able to convert hand signs into audible format with accuracy 99.51%. The system recognizes only few alphabets (A to E) and only static images. The output of the system is able to translate sing language to audio format.

In [3] System use web camera to take hand gesture as input, for giving better or clear input to model, backgrounds are detected and eliminated using color extraction algorithm HSV (Hue, Saturation, Value) Segmentation is then perform to detect region of skin tone. The system using CNN to predict the correct hand sign gesture. The system able to recognize only A, B, C, D, K, N, O, T and Y. This system is only for Sign to text conversion.

In [4] This system uses one hand glove; The system focuses on Image based hand sign gesture recognition system by processing a scheme using data driven hand gesture recognition based upon skin color model approach and thresholding approach along with effective template matching using PCA. The system uses smart glove that translates a hand gesture using an FGPA. The translated letter will be performed on VGA monitor. This system is able to convert Hand sign gesture to audio format.

In [5] the system uses CNN for taking input from both hand and HOG (Histograms of oriented gradients) for feature extraction. In this method the input image is divided into small cells and then histogram of gradient is calculated for each cell. The use Google's text to speech for converting Sign hand gesture into audio format. This system is able track all the alphabets (A to Z).

In [6] which use computer vision-based hand gesture recognition using CNN (convolution neural network) for real time hand sign recognition and it works on American Sign Language (ASL). This system is able to track all alphabets and convert it into audible format, for converting signs into audio system use Python text to speech converter.

III. IMPLEMENTATION

From website of Kaggle, the required hand gesture image dataset is collected. The dataset contains gesture images for A-Z letters. Also, a custom dataset for 14 sentences is included. To make dataset ready for further classification training, the images went under preprocessing which include grayscale conversion which created binary images containing less data compared to RGB images. Few morphological operations and noise removal techniques were applied. For recognition, the images were feed to the model for training. Convolutional Neural Network (CNN) is used for training. After getting model ready for classification, the provided text output was converted into audio by using python text to speech libraries. The user-friendly UI is created and embedded with the model.

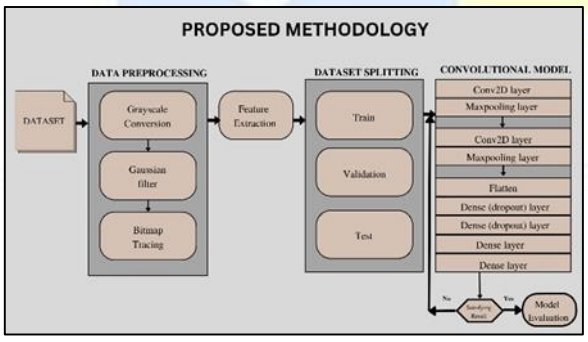

Fig.2 Complete process of pre-processing and model training

A. Dataset used

For A-Z alphabets, a dataset from Kaggle which includes 4000 images per alphabet is collected. So, total of 104000 images of alphabets are considered. Then, a custom dataset for 14 day-to-day sentences is generated using phone camera which included 4000 images for each sentence. Which is 56000 images in total. Few images were taken from mobile camera for sentences and then these images multiplied to generate a sufficiently big dataset. Final data which went under pre-processing contained 160000 images in total. For each sentence and alphabet, 3500 images were taken for training and rest 500 images were kept for testing. As the people do not talk only with alphabets, the sentences were chosen which are basic day to day sentences which are as follows. 1.Thank you so much 2. please help me 3. Hi, my name is xyz 4. I'm sorry 5. I'm not feeling well 6. I'm hungry/thirsty 7. Yes/no 8. No 9. I like it/love it 10. Nice to meet you 11. Goodbye/Good night 12. Fine, how are you? 13.I am confused 14.I am tired. All these sentences have one hand gesture associated with it.

B. Pre-processing

1) RGB to Binary conversion

The RGB images from dataset were converted into binary images. This is done because RGB images contain too much data which can extend the process of pre-processing and requires a lot of computational power. Binary images drastically reduce the data which is suitable for computation. Binary image means the object in the images is filled with solid white color and the background will be filled with solid black. Based on region of pixels, their numerical value in range of 0 to 1 is given to next process.

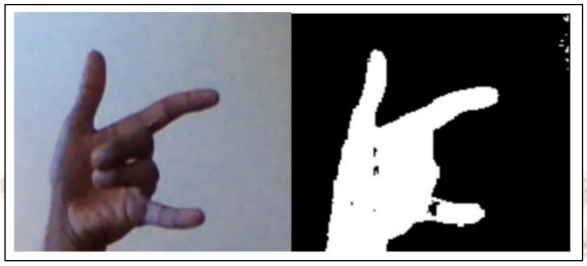

Fig 3. RGB image (to the left) to Binary image (to the right) conversion.

2) Data augmentation

The images used as dataset of sentences gestures were less in numbers. So, data augmentation[7] was used to expand the dataset. This provided us with flipped, rotated, and cropped set of images for same gestures. This resolves the issues of overfitting and data scarcity.

Fig.4 Augmented resultant images of sentence “Goodbye”

3) Noise reduction

Separation of background from the object is necessary in order to achieve high accuracy. So, noise reduction is done by convolving image with Gaussian smoothing filter. Further, the noise adjustment slider is added to GUI in order to adjust the noise in the area of interest (AOI) in the camera sample.

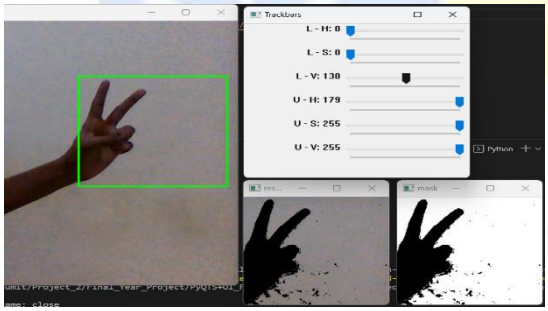

Fig 5. Noise adjustment settings when taking input from camera

C. Classification

The Convolutional Neural Networks[8] were used to classify the hand gestures. We have used TensorFlow and karas for implementing CNN model. There are total of 14 layers of CNN used in this project. Each 128x128 images were fed to the set of convolution and pooling layers. First, the 4 sets of convolution and max-pooling layers were used. These created the feature map and max-pooling operations filtered the key features. Further the flattening and dense-dropout layers were used for classification of gestures.

1) Convolutional layers

Convolutional layer is the most important layer of CNN. This layer filters the image with feature kernel to create a feature map. This provided us with a feature map.

2) Max-pooling layers

Max Pooling is an operation that calculates the maximum value for patches of a feature map, and uses it to create a down-sampled (pooled) feature map. This layer provided the most prominent features of images.

3) Flattening layer

The results provided by the max-pooling layer values in 2 dimensional arrays. So, the flattening layer was used to convert these 2-dimensional arrays into a single linear vector.

4) Dense layers

Dense layers are used as fully connected layers for classifications. Each neuron of this layer is connected to every neuron of its preceding layer.

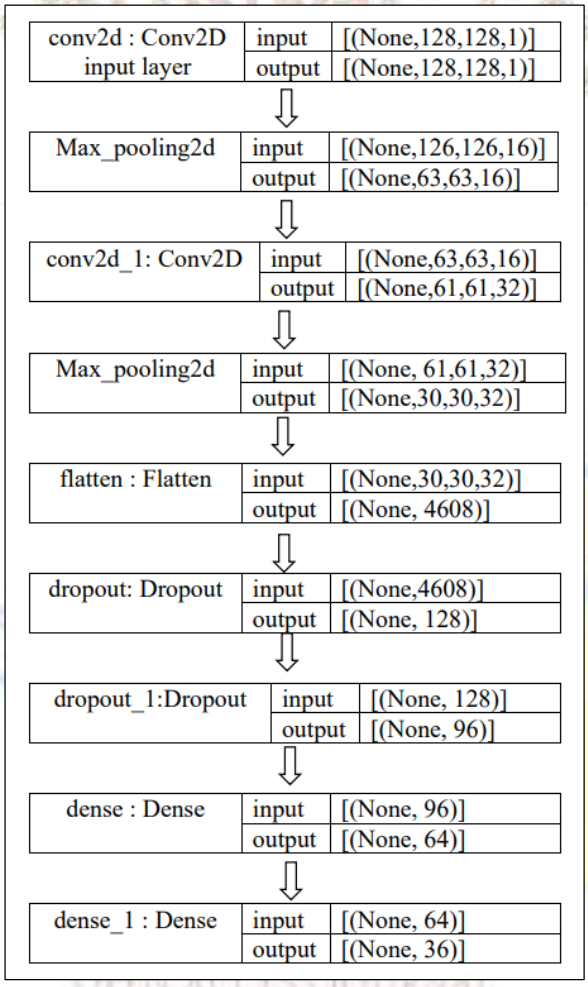

Fig 6. CNN Architecture and its layers used in this project

D. Text to speech conversion

As this CNN model provided text output as a result of classification, python text to speech libraries were used to convert this text outputs into audio outputs. Pyttx[9] library was used for this process.

E. Graphical User Interface (GUI integration)

In order to design the User Interface for our application, we have used Qt Designer which is a declarative framework for designing and building dynamic and custom UI. Qt Designer uses a number of widgets to design different aspects of an interface which consist of trackbar, buttons, label, text browser, etc. We have divided our UI functionality into 4 main parts which are gesture scanning, scanning complete sentence (sentence formation), Custom gesture creation and gesture viewer.

Fig 7. Home window of application

1) Custom Gesture Creator

Sometimes the gestures provided in this system are not sufficient for the sake of communication. So, the users need to add few custom gestures which have a unique sentence associated with it. This section offers the creation of custom and unique gestures that can be scanned and recognized to their associated values (letters or sentences). Scale-Invariant Feature Transform (SIFT) algorithm is used to achieve this functionality. The user has to add a gesture when camera is opened, after confirming the gesture, user adds sentence which gets associated with the gestures.

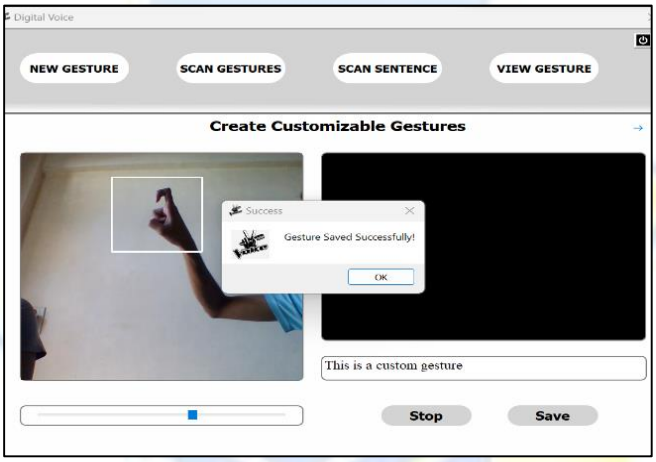

Fig 8. Creating custom gesture for sentence “This is a custom gesture”

2) Scan Gesture

The Scan Gesture offers scanning of custom created and consecrated gestures which will provide the desired text outputs.

3) Scan Sentence

This window will help us to scan the regularly used sentences that have been included in our created dataset and print them on the text window. Also, an ‘Audio’ button has been added in this window that will provide the vocal result of the text output generated. This interface covers the main function of the project which is to provide audio output from the dynamic live image provided in the camera frame.

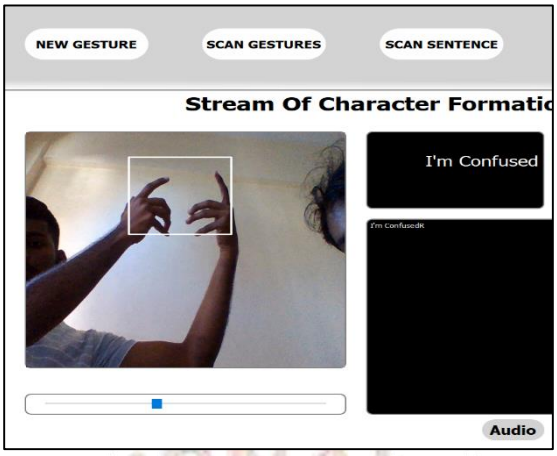

Fig 9. Scanning gesture for "I'm Confused"

4) Gesture Viewer

The Gesture Viewer depicts the series of the gestures provided to form a conversation using different hand signs of sentences and single alphabets. The next and previous buttons are used to hover through all the gestures performed.

IV. EXPERIMENTAL RESULTS

The CNN model was successfully able to recognize the hand gestures with the accuracy of 99.1%. After opening the application, user gets an introductory video which explains the 14 different hand gestures and sentences. Then user has 4 functionalities provided in main page. User can scan a gesture which will show the resultant text which is the meaning of that gesture. User can confirm that gesture by pressing the key c which also saves the text. User can create a complete sentence by following this same process. Then the saved text gets played as an audio after pressing the audio button provided in UI. User also has an option for creating a custom gesture where user has to scan and save the unique gesture and save the sentence for that gesture. SIFT algorithm uses this saved gesture to identify the same input and display the sentence associated with it. The same process then used to play audio speech.

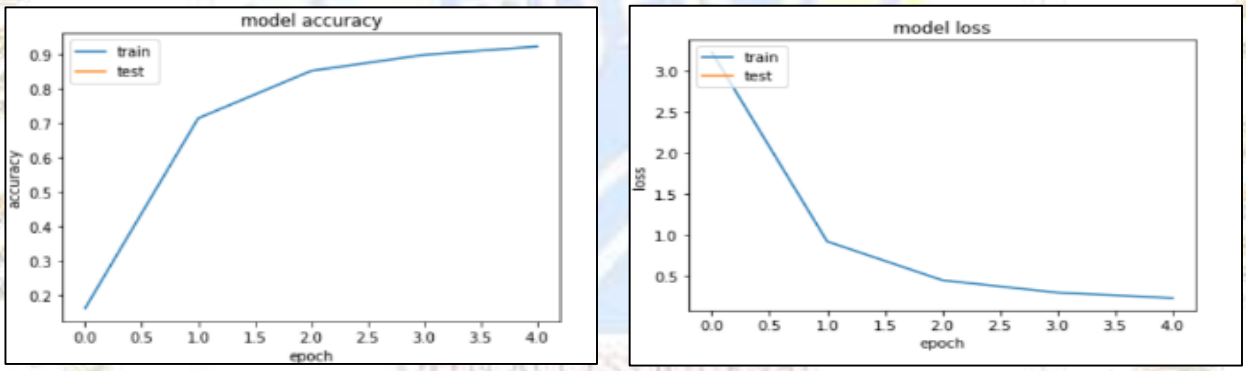

Fig 10. Model Accuracy result after five epochs

The model was trained by running 50 epochs which gave the accuracy of 99.1%. We have kept the Validation Steps up to 6500 in order to track the validation accuracy acquired i.e., 68.9%.

V. CONCLUSION

From the above results, it can be concluded that system is able to accurately classify the 26 alphabets and 14 sentences. Also, the custom gestures were getting recognized by the system with relatively low accuracy. This is because the SIFT algorithm compares the input images with only one saved image whereas CNN is trained on large dataset. All the recognitions were in real time and fast. The audio outputs were played accurately after playing the audio buttons. Currently the audio results were not in real time and requires an effort of clicking one button. The other paper discussed in this paper were helpful in improvising on many aspects of this project. This project can also be extended to other types of sign languages if required datasets of gestures are available.

VI. REFERENCES

[1] Dr. Dayananda P,Ankit Ojha, Ayush Pandey, "Sign Language to Text and Speech Translation in Real Time Using Convolutional Neural Network", ISSN: 2278- 0181,Volume 8, Issue 15, 2020.

[2] Kartik Shenoy, Tejas Dastane, Varun Rao, "Real-time Indian Sign Language Recognition", IEEE – 43488, 9th ICCCNT 2018 July 10-12, 2018.

[3] Mehreen Hurroo, Mohammad Elham Walizad, "Sign Language Recognition System using Convolutional Neural Network and Computer Vision" International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 9 Issue 12, December-2020.

[4] Muhammed Zahit Karacam, Umut Toksoy, "Sign Language Letter Translator",Diligent Contest, Europe, May 01, 2019.

[5] Omkar Vedak, Prasad Zavre, Abhijeet Todkar, Manoj Patil, "Sign Language Interpreter using Processing and Machine Learning", International Research Journal of Engineering and Technology (IRJET), Volume 06 Issue : 04, April 2019.

[6] Amrita Thakur, Pujan Budhathoki, Sarmila Upreti, Shirish Shrestha, Subarna Shakya, "Real Time Sign Language Recognition and Speech Generation", Journal of Innovative Image Processing (JIIP), Vol.02/No.02, 2020.

[7] An article on MachineLearningMastery for how to configure image data augmentation in keras. https://machinelearningmastery.com/how-to-configure-image-data-augmentation-when-training-deep-learning-neural-networks/

[8] An article on Medium by *Raiha Khan* on using CNN for Sign language recognition of hand gestures. https://medium.com/@raihakhan/applying-cnns-to-sign-language-recognition-of-hand-gestures-and-mouth-gestures-b02082911428

[9] Pyttx for converting text outputs into audio in https://hackthedeveloper.com/text-to-speech-pyttsx3-python/