© INNOVATIVE RESEARCH THOUGHTS
Refereed | Peer Reviewed | Indexed
ISSN : 2454 – 308X | Volume : 09 , Issue : 04 | July - September 2023
DOI : https://doi.org/ 10.36676/irt.2023-v9i4-013

Learning Model for Autistic and Dyslexic Children

Rachit Mehul Pathak
rachitmehul.pathak2020@vitstudent.ac.in
Ajay Varma Mudunuri
mudunuriajay.varma2020@vitstudent.ac.in
Varun Patrikar
varun.patrikar2020@vitstudent.ac.in

VIT University, Chennai Campus, Kelambakkam - Vandalur Rd,
Rajan Nagar, Chennai, Tamil Nadu 600127

Abstract

The purpose of the project is to create a learning software for autistic and dyslexic children. The goal of the project is to make the children choose a language and then using a custom speech -to-text classifier make the children repeat and pictographically learn various words. Teaching autistic and dyslexic is a very difficult task as they have a difficulty in learning and differentiating between different languages. We aim to create the model and categorize words from different languages, mainly English and Hindi. Currently, with models present, the problem in the models is that they are unable to detect the accent in which native Indian speakers speak. Thus, creating a model which can be used naturally over the whole country is a difficult task. The created model would be able to differentiate and understand a handful of words, and with a much powerful engine and dataset, itwould be able to act as a modern-day text to speech and a teaching tool for the autistic and dyslexic children.

Keywords: Speech-to-Text, Text-to-Speech, Web Crawler, VGG16, Inception, Wav Files, Similarity based Ranking, Metrics

Introduction

Nowadays, it has become very important to create a support system for children who are suffering with a form of disability such as autism and dyslexia. These children suffer from various problems such as avoidance
behaviors and hyper vigilance. They show

obsessive compulsive behavior along with a variety of phobias. All this leads to a decline in mental and physical health. These childrensuffer from social phobias due to their learning disabilities. They are often misjudged and undermined due to their learning skills.
In order to help autistic and dyslexic children in their language and academic skills, various models and applications have been created.

These models have been created with the soul aim of assisting autistic and dyslexic kids and increase their self confidence in life to cope up with day to problems. The problem with such models is that they fail to assist disabled kids in their learning ventures. As the complexity of words increases, these models decline in precision and accuracy.
Autistic and dyslexic kids need to be provided with both visual and auditory related assistance. It can be observed that some models lack visual assistance,while others lack audio assistance.
To assist autistic and dyslexic kids, this research has created a model that provides both audio and visual aid. The model is effectively able to rank words based on their pronunciations and lexicographical complexities. As the user progresses through with time, the complexity of

© INNOVATIVE RESEARCH THOUGHTS
Refereed | Peer Reviewed | Indexed
ISSN : 2454 – 308X | Volume : 09 , Issue : 04 | July - September 2023
DOI : https://doi.org/ 10.36676/irt.2023-v9i4-013

the words increases so that the skill level of the user increases with time. The model created is not only useful to kids suffering with disabilities, but to teachers who are trying to help them as well. It provides them with a learning routine, so that they can effectively guide their students on the correct learning path. Various datasets for both Hindi and English have been used to categorize words based on their difficulty. The accuracy of the model has also been repeatedly tested in order to choose the right words for the user. For Hindi words, a custom dataset was built that consists of various voice recordings from 2-3 sources.

Literature Review

[1] The application converts synthesizer is an application that converts text into spoken words using NLP and then using Digital signal processing. Various operations and processes were involved in text to speech analysis; however it was only done for the English language.

[2] Non-autoregressive text to speech (TTS) models such as FastSpeech can synthesize speech significantly faster than previous autoregressive models with comparable quality. The training of FastSpeech model relies on an autoregressive teacher model for duration prediction (to provide more information as input) and knowledge distillation (to simplify the datadistribution in output), which can ease the one-to-many mapping problem (i.e., multiple speech variations correspond to the same text) in TTS. The model is very quick compared to other models, however the duration extracted from the teacher model is not accurate enough, and the target spectrograms distilled from teacher model suffer from information loss due to data simplification, both of which limit the voice quality.

[3] The Web Speech API allows users to record audio from the microphone, which is then sent via a HTTP post request to the speech recognition web service. The model is very efficient with 96.63% and 82.78% accuracy for Indonesia and English language.

[4] The model uses deep learning-based machine learning models which shapes significant results in speech recognition and numerous vision related tasks. The research uses neural networks which give very good accuracy although, the accuracy of the model decreases as the number of words are increased.

[5] This project compares 3 major image processing algorithms: Single Shot Detection (SSD), Faster Region based Convolutional Neural Networks (Faster R-CNN), and You OnlyLook Once (YOLO) to find the fastest and most efficient of three.

[6] Considering how closely object detection relates to video analysis and visual comprehension, it has received a lot of study interest lately. The foundation of approaches is about shallow trainable architectures and handmade features. The advantage of this model is that it solves the issues with traditional architectures, more potent tools that can learn semantic, high-level, deeperfeatures are being offered as a result of deep learning's quick development. The disadvantages ofthis model is that in terms of network architecture, training methodology, optimization function, etc., these models behave differently. We have overcome this disadvantage by creating a model that behaves similarly for all important features.

[7] All visual media are interpreted by a computer as a collection of numerical values. They need image processing algorithms to inspect the contents of images as a result of this method. In orderto determine which method is the fastest and most effective among three, this research analyses Single Shot Detection (SSD), Faster Region based Convolutional Neural Networks (Faster R- CNN), and You Only Look Once (YOLO). The performance of these three

© INNOVATIVE RESEARCH THOUGHTS
Refereed | Peer Reviewed | Indexed
ISSN : 2454 – 308X | Volume : 09 , Issue : 04 | July - September 2023
DOI : https://doi.org/ 10.36676/irt.2023-v9i4-013

algorithms is compared in this comparative analysis using the Microsoft COCO (Common Object in Context) dataset, and their advantages and disadvantages are examined using metrics including accuracy,precision, and F1 score.

[8] A particular voice can be recognized using voice recognition technology. The foundation for speaker identification is voice signals. The advantage of this model is that it is useful in a wide range of applications, including voice mail, database access, phone banking, and phone purchasing. Where one can enter their voice for verification is one of the most effective voice recognition security applications. The fundamental method of interpersonal communication is speech. Speech sounds are translated into text through the process of speech recognition. Over the past few years, speech recognition technology has advanced significantly. However, there are other significant study hurdles, including variations in speaker and language, ambient sound, word size, and more.

[9] The technique of translating human sound impulses into words or commands is known as speech recognition. The foundation of speech recognition is speech. It is a crucial area for research in both pattern recognition and voice signal processing. Computer science, artificial intelligence, digital signal processing, pattern recognition, acoustics, linguistics, and cognitive science are justa few of the disciplines that are involved in the study of speech recognition. It is a broad, diversearea of inquiry. According to the requirements of the speaker's way of speaking, these areas can be divided into isolated words, connected words, and continuous speech recognition systems.
According to the degree of dependence on the speaker, these areas can be divided into speech recognition systems for the specific person and nonspecific person. According to the size of vocabulary, they can be divided into small vocabulary, medium vocabulary, large vocabulary, and infinite vocabulary speech recognition systems. The accuracy of speech detection modelsreduces as the vocabulary and complexity of the word becomes larger.

[10] Automatic speaker recognition systems use the machines to recognize an individual via a spoken sentence. Those systems recognize a specific individual or confirm an individual's claimed identity.
The most common type of voice biometrics is the Speaker Recognition. Its task focused on validation of a person's claimed identity, using features that have been obtained via their voices. In this paper, a brief overview of speech processing is given firstly, then some feature extraction and classifier techniques are described, also a comparative and analysis of some previous research are studied in depth, all this work leads to determine the best methods for speaker recognition. Various disadvantages regarding speech detection have been highlighted such as thedecline of accuracy for speech recognition models for words with complex pronunciations. We have dealt with this problem by creating our model systematically for English and Hindi words.

[11] Non-autoregressive text-to-speech (TTS) models such as FastSpeech can synthesize speech much faster than previous autoregressive models of comparable quality. FastSpeech model training relies on an autoregressive supervised model for long-term prediction (to provide more information than the input) and knowledge distillation (to simplify the data distribution of the output), with a one-to-many Solve the mapping problem (that is, multiple). TTS can facilitate language variations corresponding to the same text). However, FastSpeech has some drawbacks one of which is the teacher-student distillation pipeline is complex and time-consuming.

[12] The use of technology is very important nowadays. Most of everyday activities are

© INNOVATIVE RESEARCH THOUGHTS
Refereed | Peer Reviewed | Indexed
ISSN : 2454 – 308X | Volume : 09 , Issue : 04 | July - September 2023
DOI : https://doi.org/ 10.36676/irt.2023-v9i4-013

documentedand stored in digital form. One of those activity is recording. Activity of an organization is very important, therefore recording of meeting material is one of those important activities. Usually, the meeting material is documented by writing them into papers or typing them and stored them into computer. Sometimes, the meeting information is written or typed incorrectly so a speech- to-text application is required to solve this problem. In this study, the solution offered is to implement a web-based automation speed-to-text application which can record the voice of meeting participants then converted them into text automatically, so the results of the recording process of meeting materials are more effective and efficient by using voice recognition.

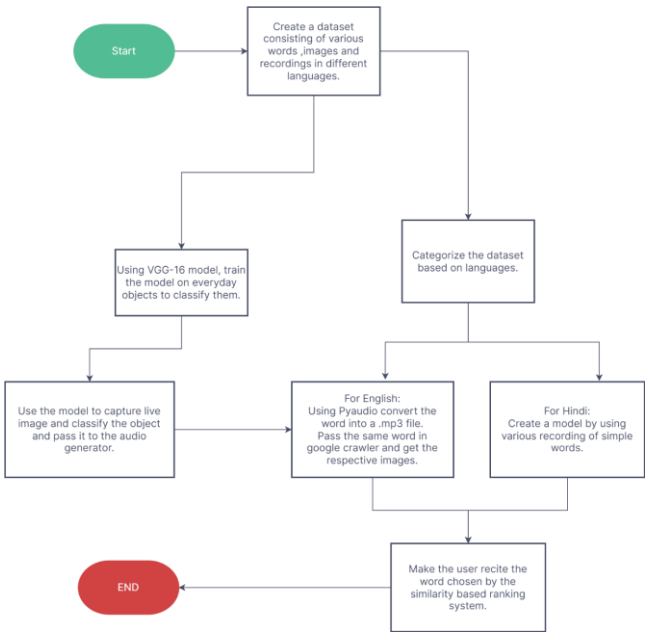
Methodology

In this project, a learning model is created for dyslexic and autistic children. The work has been divided into three segments. The first segment covers the making of the dataset using google crawler and feature ranking of the words. The customized dataset was run through amodel with features considering the length, complexity, and cosine similarity of the words. The model was able to rank the words based on their pronouncing and lexicographical complexity. This also stands as atool for the teachers to rank the words and segregate them into hard, medium, and easy to speak words. This segment uses spacy en_core_web_sm dataset and on top of it, categorizes the words with respect to their complexity.

On top if this, the project also includes a custom-built model for Hindi language, since the accent and pronunciation are different for people across the globe. The model built can classify speech into categories andwith a larger dataset, will act as a speech-to-text model. Various optimizers are used to make the prediction accuracy higher. The next part was consisted of making a software framework which was made using multiple threads for audio and visual predicate. The frameworks workflow is designed so that a student can choose the language at first, and then go about learning various words provided by the teacher. The audio and visual threads run simultaneously to give an active and interactive study session. The framework is designed to stop once the user says goodbye.

The last segment of the project consists of a visual model which using VGG16 model, and custom training, can detect everyday objects. The model is a type of convolutional neural network with 16-19 depth layers making it almost with 138 trainable parameters. It can classify the object, run the same framework and is able to make the word a part of itsdataset and teach the child what the object is and how to say it.

Fig 1: Methodology Chart



Results and Analysis

Hyperparameters and Parameters taken and adjusted around are:

- Train-to-test split ratio
- Learning rate was used as 0.0001.

© INNOVATIVE RESEARCH THOUGHTS
Refereed | Peer Reviewed | Indexed
ISSN : 2454 – 308X | Volume : 09 , Issue : 04 | July - September 2023
DOI : https://doi.org/ 10.36676/irt.2023-v9i4-013

- Dataset Information and Optimizing algorithms

The dataset considered for the training of everyday objects consisted of objects like hair comb, bottle, pens, and phone. The model was able to distinguish between the various objects. The parameters in the activation functions were Relu and Sigmoid. The model optimizer was tried and tested on Adam, RMSprop and SGD. The best result was observed with Adam Optimizer.

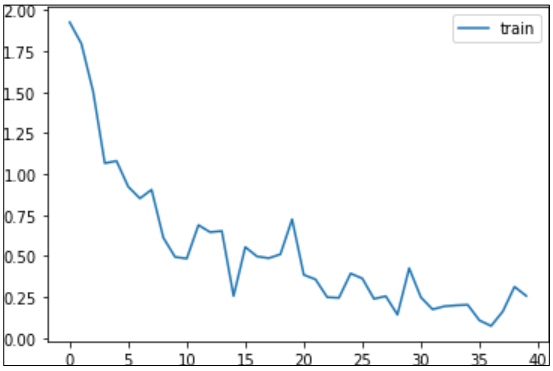
The image classifier was built on a pre-built VGG16 model. The trainable layers weretrained on the dataset containing 153 images from 4 classes. The model was built on a set of 30 epochs with 10 steps per epochs and the accuracy got was found to be 0.9000. Total params: 138,357,544

Trainable params: 138,357,544

Time taken for execution was around 10 minutes per model. Due to the dataset size, an EarlyStopping monitor and a ModelCheckpoint. Image augmentation techniques used were geometric transformation and random flipping.

The last segment of the project consists of a visual model which using VGG16 model, and custom training, can detect everyday objects. The model is a type of convolutional neural network with 16-19 depth layers making it almost with 138 trainable parameters. It can classify the object, run the same framework and is able to make the word a part of its dataset and teach the child what the object is and how to say it.

Fig 2: Adam Model Optimizer for VGG16 model



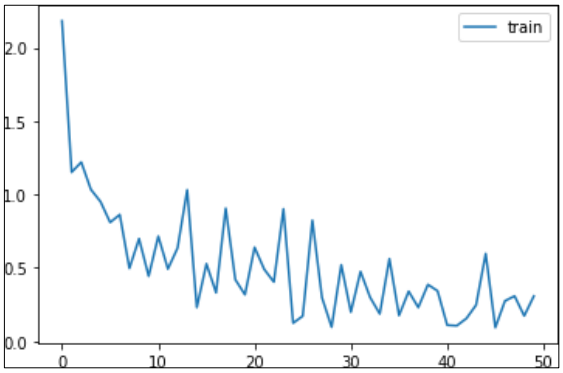


Fig 3: RMS Prop Optimizer for VGG16 model

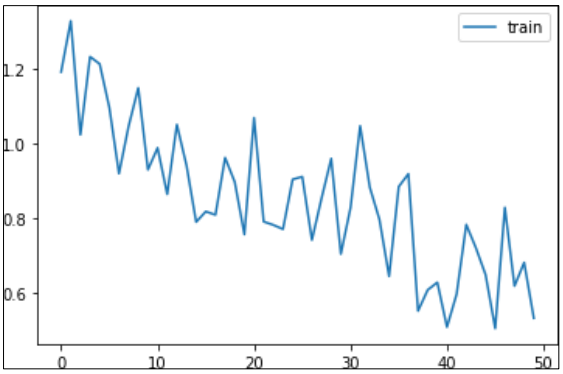


Fig 4: SGD Optimizer for VGG16 model

Table 1:Training time and Accuracy for different optimizers (VGG16 model)

Model	Optimizer	Accuracy	Training Time	Loss
VGG16	RMS Prop	0.9000	2:40 minutes	0.3068
VGG16	SGD	0.7500	2:37 minutes	0.5316
VGG16	Adam	0.9000	1:47 minutes	0.2572

I. Model with RMS PROP Optimizer

© INNOVATIVE RESEARCH THOUGHTS
Refereed | Peer Reviewed | Indexed
ISSN : 2454 – 308X | Volume : 09 , Issue : 04 | July - September 2023
DOI : https://doi.org/ 10.36676/irt.2023-v9i4-013

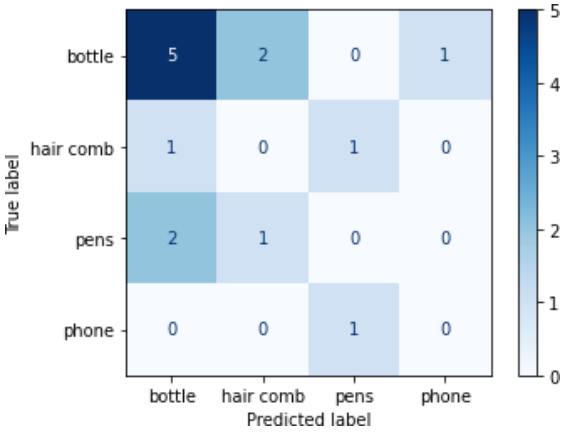


Fig 5: Confusion Matrix of Test Data for RMS PROP optimizer

	Precision	recall	F1 score	support
bottle	0.62	0.62	0.62	8
hair comb	0.00	0.00	0.00	2
pens	0.33	0.33	0.33	3
phone	0.00	0.00	0.00	1
accuracy			0.43	14
macro avg	0.24	0.24	0.24	14
weighted avg	0.43	0.43	0.43	14

Table 2: Test data classification report for RMS PROP optimizer

II. Model with Adam Optimizer

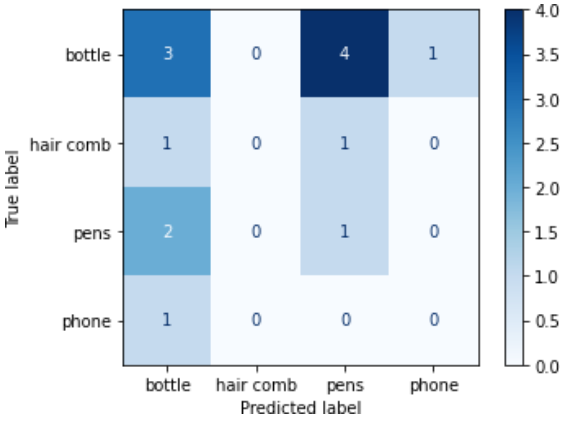


Fig 6:Confusion Matrix with Adam Optimizer

	Precision	recall	F1 score	support
bottle	0.43	0.38	0.40	8
hair comb	0.00	0.00	0.00	2
pens	0.17	0.33	0.22	3
phone	0.00	0.00	0.00	1
accuracy			0.29	14
macro avg	0.15	0.18	0.16	14
weighted avg	0.28	0.29	0.28	14

Table 3: Test data classification report for Adam optimizer

III. Model with SGD Optimizer

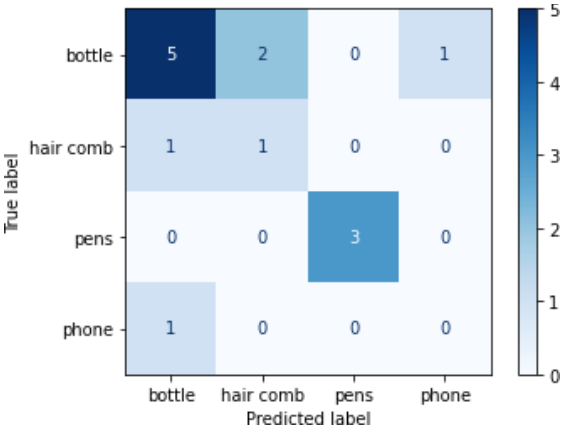


Fig 7:Confusion Matrix with SGD Optimizer

	Precision	recall	F1 score	support
bottle	0.33	0.25	0.29	8
hair comb	0.00	0.00	0.00	2
pens	0.20	0.33	0.25	3
phone	0.00	0.00	0.00	1
accuracy			0.21	14
macro avg	0.13	0.15	0.13	14

© INNOVATIVE RESEARCH THOUGHTS
Refereed | Peer Reviewed | Indexed
ISSN : 2454 – 308X | Volume : 09 , Issue : 04 | July - September 2023
DOI : https://doi.org/ 10.36676/irt.2023-v9i4-013

weighted avg	0.23	0.21	0.22	14

Table 4: Test data classification report(SGD optimizer)

For the model trained on the audio custom dataset, it consists of 5 labels namely hello, chai, namaste, tara and shabash. The confidence values of the words were represented as –

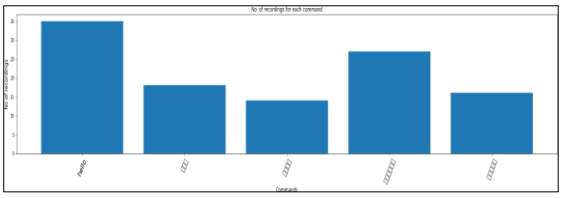


Fig 8

Plotting of a Raw wave of an audio data point. The wave is first sampled in sets of the minimum frequency and then the minimum frequency is used as the input shape for the model.

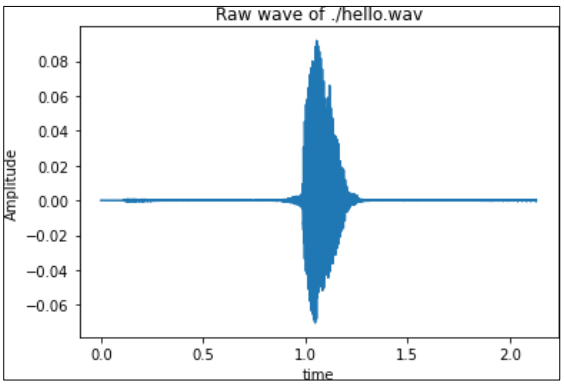


Fig 9: Raw wave for the word hello

The best accuracy and graph were formed using RMS Prop as an optimizer and thetrain test split ratio of 0.2. Adam gave similar result with a lesser loss function.
The model used was a base TensorFlow model with

Total params: 2,839,653
Trainable params: 2,839,653

Model	Optimizer	Accuracy	Training Time	Loss	Early Stop Epoch
Customized CNN	RMS Prop	0.3684	25.9 seconds	1.8011	Epoch 00024
Customized CNN	SGD	0.3014	41.7 seconds	1.5944	Epoch 00042
Customized CNN	Adam	0.3684	46.5 seconds	1.5046	Epoch 00037

Table 5:Training time and Accuracy for different optimizers (Customized CNN model)

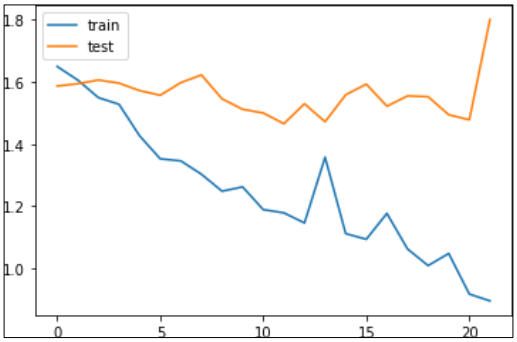


Fig 10: RMS Prop Optimizer for Customized CNN model

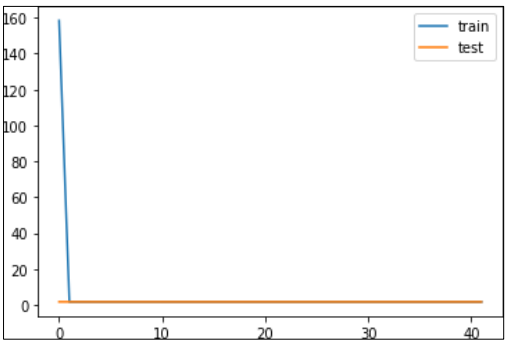
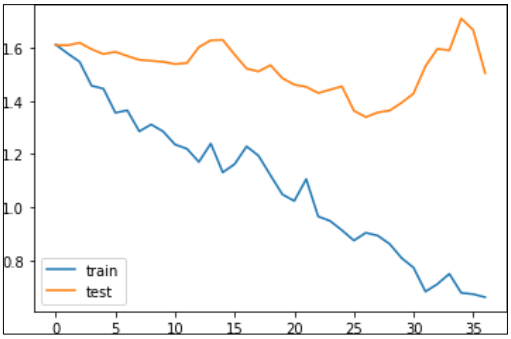


Fig 11: SGD Optimizer for Customized CNN model



© INNOVATIVE RESEARCH THOUGHTS
Refereed | Peer Reviewed | Indexed
ISSN : 2454 – 308X | Volume : 09 , Issue : 04 | July - September 2023
DOI : https://doi.org/ 10.36676/irt.2023-v9i4-013

Fig 12: Adam Optimizer for Customized CNN model

Conclusion

The purpose of the project was to create a learning software for autistic and dyslexic children. The goal of the project was to make the children choose a language and then using a custom speech -to-text classifier make the children repeat and pictographically learn various words. Teaching autistic and dyslexic is a very difficult task as they have a difficulty in learning and differentiating between different languages.

This research was able to create the model and categorize words from different languages, mainly, English and Hindi. The created model can differentiate and understand a handful of words, and with a much powerful engine and dataset, it will be able to act as a modern-day text to speech and a teaching tool for the autistic and dyslexic children. The custom-built model, was able to achieve the software workflow as planned.

A feature was included for the teachers to add and get the word ranked as per the complexity of the word. In addition, a visual everyday object detecting model was createdwhich would detect the objects present and teach the children how to spell and learn aboutthe objects.

The future scope of the project includes using a larger dataset of both audio andwords, to teach even more words and better the model created by increasing theaccuracy and hosting the scaled model on an open source to make it available to the world.

References

[1] Isewon, I. (n.d.). Design and implementation of text to speech conversion for visually impaired people.

[2] Kim, W., & Nam, H. (2021). End-to-end non-autoregressive fast text-to-speech. *Phonetics and Speech Sciences*, *13*(4), 47–53. https://doi.org/10.13064/ksss.2021.13.4.047

[3] Yasir, M., Nababan, M. N., Laia, Y., Purba, W., Robin, & Gea, A. (2019). Web-Based Automation Speech-to-Text Application using Audio Recording for Meeting Speech. Journal of Physics: Conference Series, 1230(1), 012081. https://doi.org/10.1088/1742-6596/1230/1/012081

[4] M.A.Anusuya, S.K.Katti, (2009). Speech Recognition by Machine: A Review. International Journal of Computer Science and Information Security, 6(1).

[5] Srivastava, S., Divekar, A.V., Anilkumar, C. *et al.* Comparative analysis of deep learning image detection algorithms. *J Big Data* **8**, 66 (2021). https://doi.org/10.1186/s40537-021-00434-w

[6] Zhong-Qiu Zhao, Member, Peng Zheng, Shou-tao Xu, Xindong Wu. (n.d.). Object Detection with Deep Learning: A Review.

[7] Srivast, S., Divekar, A. V., Anilkumar, C., Naik, I., Kulkarni, V., & V., P. (2020). *Comparative analysis of deep learning image detection algorithms*. Research Square Platform LLC. http://dx.doi.org/10.21203/rs.3.rs-132774/v1

[8] *Speech Recognition - An overview*. (n.d.). ScienceDirect Topics. Retrieved July 30, 2023, from https://www.sciencedirect.com/topics/engineering/speech-recognition

[9] Spalanzani, A. (2007). Evolutionary speech recognition. In *Robust Speech Recognition and Understanding*. I-Tech Education and Publishing. http://dx.doi.org/10.5772/4744

[10] Ali, Akbas & Hassan, Nidaa. (2019). Advantages and Disadvantages of Automatic Speaker Recognition Systems. 21-30.

ISSN : 2454 – 308X | Volume : 09 , Issue : 04 | July - September 2023
DOI : https://doi.org/ 10.36676/irt.2023-v9i4-013

[11] Yi Ren , Chenxu Hu, Xu Tan , Tao Qin, Sheng Zhao , Zhou Zhao , Tie-Yan Liu. (n.d.). FastSpeech 2: Fast and High Quality End-to-End Text to Speech.

[12] Yasir, M., Nababan, M. N., Laia, Y., Purba, W., Robin, & Gea, A. (2019). Web-Based Automation Speech-to-Text Application using Audio Recording for Meeting Speech. *Journal of Physics: Conference Series*, *1230*(1), 012081. https://doi.org/10.1088/1742-6596/1230/1/012081