Machine Learning for Speech-to-Text: A Custom-CNN Model Approach

Ifa Fatima
Department of Computer Science and Engineering
Aliah University
Newtown, Kolkata
fatimaifa786@gmail.com

Khondekar Lutful Hassan
Department of Computer Science and Engineering
Aliah University
Newtown, Kolkata
klh.cse@gmail.com

Moinuddin
Department of Computer Science and Engineering
Aliah University
Newtown, Kolkata
cse.moinuddin@gmail.com

***Abstract*— One of the most challenging topics for researchers to study is speech to text conversion. This engineering technology is experiencing rapid growth. Disabled persons can profit from its various potential applications and advantages, as it has a wide range of applications. Language problems prevent many individuals from communicating. With the intention of lowering this barrier, a machine-learning model was created in order to build systems that, in certain situations, can be a big assistance in enabling individuals to communicate information by using voice input to operate a computer. An easy-to-use and efficient voice recognition algorithm is described in this paper's research work. After converting the audio signal to the appropriate text, summarised text is produced. This paper presents research on a simple and efficient voice recognition technique. Summarised text is produced by converting the auditory signal to the appropriate text. The usage of machine learning in a variety of applications these days presents a challenge to model improvement. A custom-CNN model is developed to increase the precision of recognition. ReLU's quick computation makes it an excellent choice for CNN training. To confirm the effectiveness of the suggested approach, a great deal of experimentation is conducted. With an accuracy rate of 75.60%, the experimental results show how well the present CNN design performs.**

Keywords: Machine Learning, Speech, Audio to Text, Convolutional Neural Network (CNN)

I. INTRODUCTION

Human communication relies heavily on speech. Though there are several ways to express our thoughts and feelings, speech is considered the primary vehicle of communication. Speech recognition is the technique of having a machine recognise distinct people's speech based on specific words or phrases. Variations in pronunciation are obvious in each person's speech[11]. The original form of speech is a signal, which is processed so that all of the information included in the signal is transformed into text format. Even though speech is the most common mode of communication, there are various concerns with speech recognition such as fluency, pronunciation, broken words, stuttering, and so on. All of these must be addressed when processing a speech. Text summarisation is a key idea in the world of documentation [1]. Long texts are difficult to read and understand since they need a lot of time. This issue is resolved by text summarisation, which offers a condensed, semantic summary of the text.

II. LITERATURE REVIEW

A. A technique called "Memory-Based Phoneme-to-Grapheme Conversion" has been proposed by Bart et al. [10] to address Out-of-Vocabulary items (OOVs) in speech recognition. A method is described to develop the legibility of the textual output in a large vocabulary continuous speech recognition system (SRS) when out-of-vocabulary words strike. Final word-level accuracy for the P2G (phoneme-to-grapheme) conversion system on the dataset with phoneme recognition errors is 46%. However, for out-of-vocabulary (OOV) words, the accuracy is significantly lower, at just 7% at the word level, which improves slightly to 8% after post-processing with a spelling corrector. This technique uses machine learning concepts.

B. Lee et al. [5] have proposed “Using tone information in continuous speech recognition”. Automatic speech recognition and voice/ unvoiced boundary segmentation methods were used to recognize the Cantonese dialects and speech patterns. A different type of hidden Markov model was used here and that is context dependent tone modelling. By using Utterance wide normalization with context independent models, they achieved an accuracy of 54% and with context dependent models they achieved a 60% accuracy.

C. Shaheena et al. [12] focuses on “Bangla Speech-to-Text Conversion using SAPI” which supports only a limited number of languages, including English. It needs for expanding STT conversion to other languages, specifically Bangla, which is significant due to its cultural heritage and recognition by UNESCO as an important language. It suggests that using English as an intermediary language in SAPI could potentially facilitate Bangla STT conversion. This paper discusses how generating a grammar file with English character sets for Bangla words could improve recognition accuracy, and it reports a 74.81% recognition rate from an experimental study, indicating the potential of the approach while acknowledging areas for improvement.

III. PROPOSED METHOD

Our methodology for Speech to text conversion using machine learning involves converting spoken language into text using algorithms and models.

A. Data Collection:

In this initial phase of our project, we utilized a comprehensive audio dataset sourced from Kaggle to develop our speech recognition system. Kaggle, known for its extensive collection of datasets and data science resources, provided an ideal platform to obtain the diverse audio data necessary for robust model training. The dataset included a variety of audio recordings featuring different speakers, accents, dialects, and environmental conditions, which is crucial for ensuring the model's strong generalization across various real-world situations.

To prepare the dataset for effective training, several pre-processing steps are applied. This dataset includes the files contained 14 commands, out of them 10 commands (‘yes’, ’no’, ’up’, ’down’, ’left’, ’right’, ’on’, ’off’, ’stop’, ’go’) have been used in this Speech – to – Text Conversion’ projects Each command/order is a second .wav audio file, spoken in English language spoken by variety of different speakers. By leveraging the rich and diverse audio data from Kaggle[13], a sophisticated speech recognition system is developed which is capable of performing well in varied applications.

B. Working for Speech to Text:

- Converting speech to text, sometimes referred to as automatic speech recognition (ASR), converts spoken language into written text. Here's how it processed:

- Input Speech: The process begins with the capture of input speech, typically through a microphone. This audio can be from various sources like a person speaking, a recorded conversation, or any audio containing spoken words.

- Preprocessing: The raw audio data may undergo preprocessing steps, which can include noise reduction, filtering, and normalization to enhance the quality of the audio signal.

- Models: Acoustic models are used to represent the relationship between speech features and distinct sound units in language. Machine learning techniques, such as DNNs or CNNs, are often employed to build these models. They learn to recognize patterns in the audio that correspond to specific phonemes.

- Training dataset: Combine the acoustic and language models in a training dataset process. This involves optimizing the parameters of both models simultaneously to improve their collaboration and enhance the overall outcome of the Speech-to-Text system.

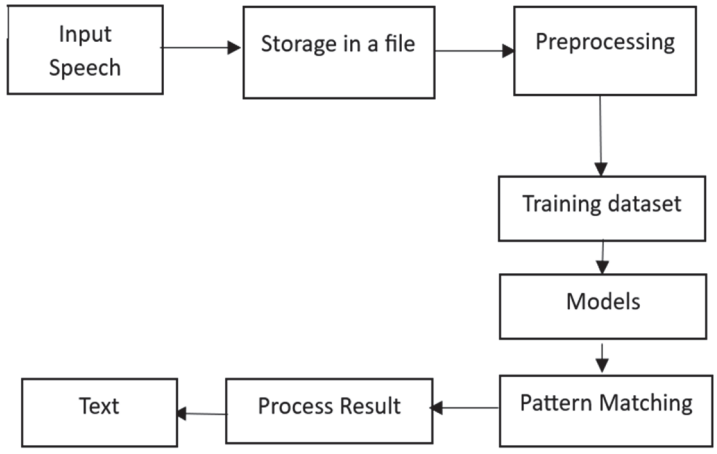

Fig.1. Working of Speech to Text

C. Speech Processing using Machine Learning:

Speech recognition for human-machine interaction depends on the human brain, much like machine learning does. Many tasks make advantage of the machine learning methodology via the feature learning capabilities. The data modelling capability outcomes obtained are superior to the performance of traditional learning methods. As a result, the speech signal recognition relies on a machine-learning algorithm to combine the voice aspects and qualities [8]. As a result of voice, the speech signal is transformed into an important component of speech enhancement.

Machine learning includes of supervised and unsupervised learning, with supervised learning being utilised for speech recognition purposes. Machine Learning (ML) software may measure spoken words using a set of numbers to represent the signals. There are numerous acoustic modelling machine-learning algorithms [7], including DNN and CNN.

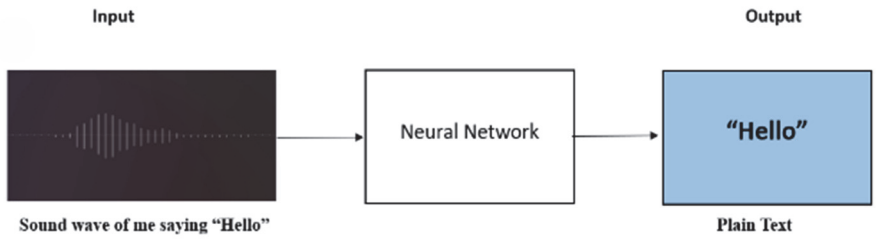

Fig.2. Neural Network based Sound Wave Detection.

D. Structure of Conv1D Model:

- *Input Layer:* The model starts with an Conv1D layer that has 128 units. The input-shape parameter directs the shape is (8000, 1) in this case. It indicates that the input data consists of a sequence of 8000 timesteps, with each per time step having 1 feature.
- *1st Layer of Conv1D Layer:* This layer has 8 filter of each kernel size is 13. The activation function is ReLU. No padding is added. This layer aids in the extraction of higher-level representations and the capture of temporal dependencies from the input sequence.
- *MaxPooling1D Layer:* A MaxPooling1D layer is added after the Conv1D layers with a pooling size of 3.
- *Dropout Layer:* A dropout layer is added after the MaxPooling1D layers with a dropout rate of 0.3
- *2nd Layer of Conv1D Layer:* This layer has 32 filters where each kernel size is 9. The activation function is ReLU. No padding is added.
- *Pooling Layer:* Another pooling layer is added after the Conv1D layers with a pooling size of 3.
- *Dropout Layer:* Another dropout layer with a dropout rate of 0.3 is included after the maxpooling1D layer for further regularization.
- *Flatten Layer:* This layer flattens the input, converting the 3D tensor into 1D tensor. It is used to prepare for the next fully connected (Dense) layer.
- *Dense Layer:* Repaired linear activation function (ReLU) and 256 units dense layer are added.
- *Dropout Layer:* Another dropout layer with a dropout rate of 0.3 is included to prevent overfitting.
- *Dense Layer:* A dense layer with 256 units and a rectified linear activation function (ReLU) is added.
- *Output Layer:* The last layer is a dense layer containing n units, corresponding to the number of classes to be predicted. The activation function applied is SoftMax, which transforms the model's output into a probability distribution across the classes, enabling the model to make predictions.

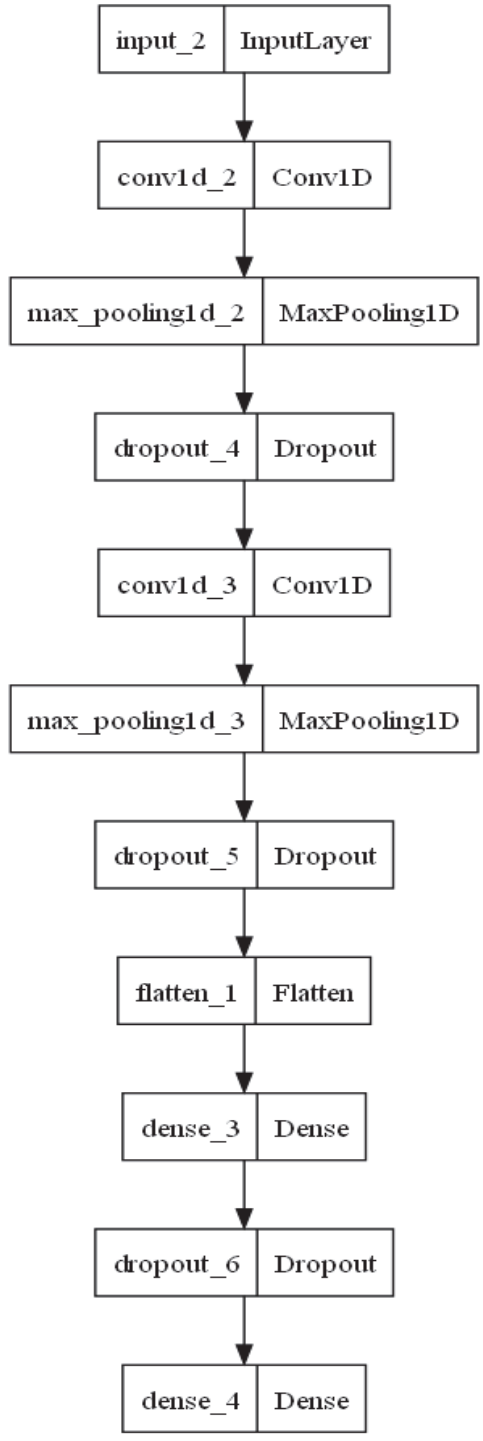

Fig.3. Structure of Conv1D Layer Model

ISBN : 978-81-978522-2-0

IV. RESULT AND ANALYSIS

A. *Speech Dataset*

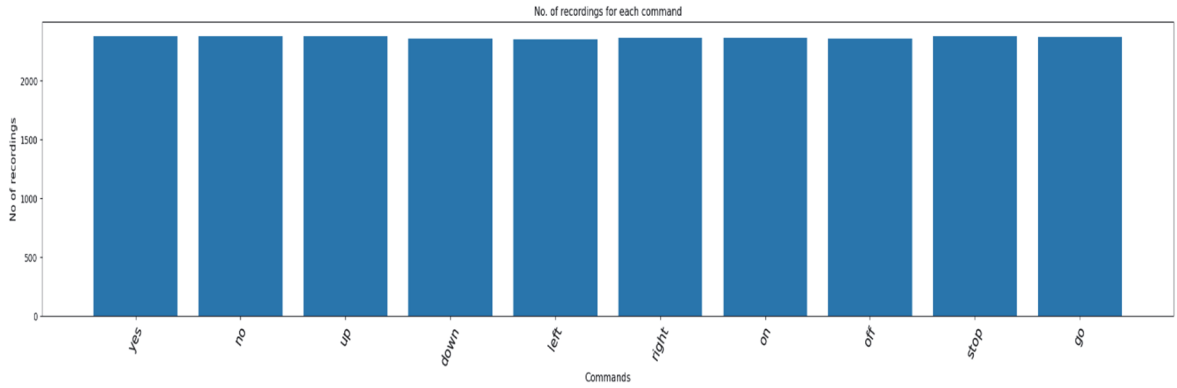

Fig.4. Plotting bar graph of Speech Dataset

B. *Evaluation:*

The model's precision, recall, F1, and support scores are shown by the classification report visualizer. We may see metrics like accuracy, which provides us with an overall measure of the accuracy of our model. In addition, we may observe precision, recall, and F1 Score, which provide information about how well our model is able to recognize classes.

	precision	recall	f1-score	support
yes	0.78	0.76	0.77	430
no	0.63	0.72	0.67	420
up	0.72	0.77	0.74	433
down	0.74	0.72	0.73	420
left	0.77	0.68	0.72	429
right	0.87	0.76	0.81	421
on	0.94	0.80	0.86	431
off	0.75	0.72	0.73	435
stop	0.62	0.77	0.69	412
go	0.79	0.81	0.80	432
accuracy			0.75	4263
macro avg	0.76	0.75	0.75	4263
weighted avg	0.76	0.75	0.75	4263

Fig.5. Classification report using Decision Tree

Here confusion matrix is used to assess how well a classification model performs. The machine learning model's predicted values and the actual target values are compared in the matrix.

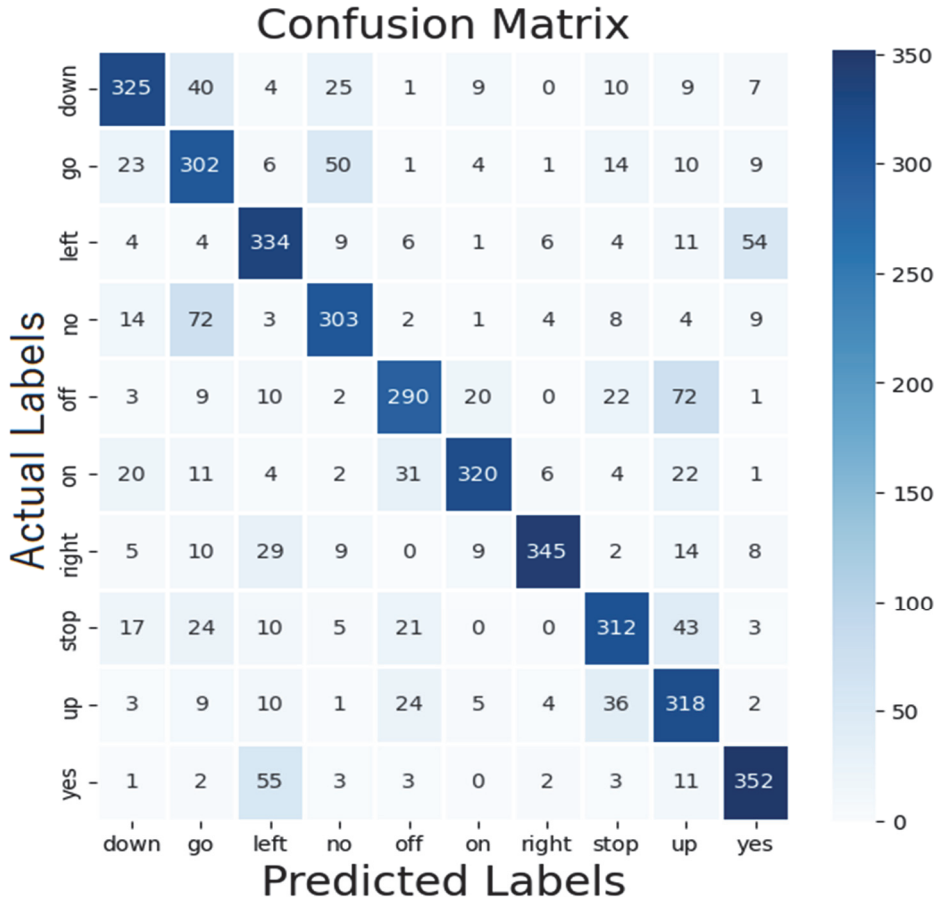

Fig.6. Decision Tree based Confusion Matrix

C. Training and Testing:

In results and analysis section, experimenting with various optimizers might lead to better results during hyperparameter tuning. Eighty percent of the data will be utilised for training, and twenty percent will be used for validation.

The training loss reduces steadily after 10 epochs while the validation loss staggers nearby. Finally, we have gotten better Accuracy result of 75.60%. The words are predicting fine. CNN help to convert or transcript the word perfectly[3]. But little some of words are not predicting in a noisy environment. We will try to improve the accuracy using a Conv1D model.

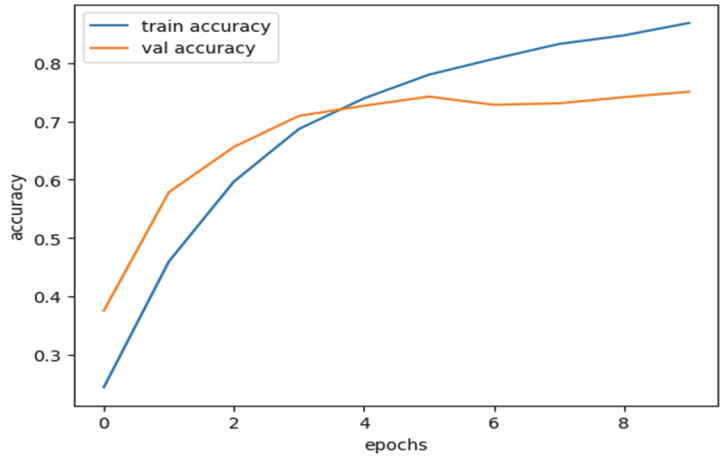

Fig.7. Accuracy v/s Epochs

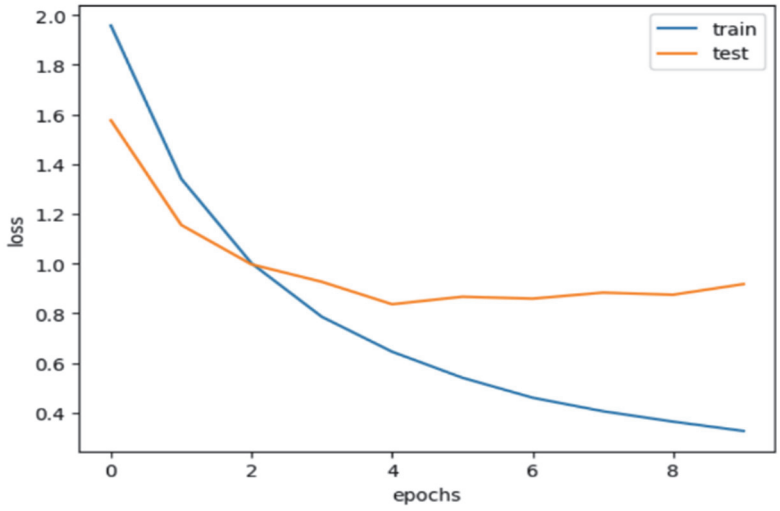

Fig.8. Loss v/s Epochs

D. Output:

In result and analysis section, the output of converting audio into written text using machine learning involves capturing spoken words through an audio input device like a microphone and transforming these sounds into text format. The process begins with preprocessing the audio signal to reduce noise and enhance clarity. Machine learning models, particularly those involving deep learning techniques such as neural networks, then analyze the audio features to identify phonemes, words, and sentences. The result is a written text that reflects the spoken content.

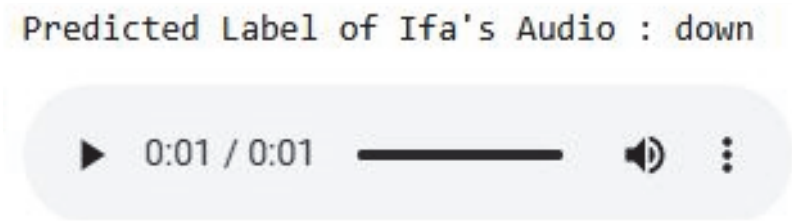

Fig.9. Result of Speech-to-Text Conversion

We created a small or mini dataset using my own voice recordings to evaluate the proposed Speech-to-Text conversion model. After testing, the model successfully transcribed the spoken words perfectly, demonstrating its effectiveness in converting speech into text.

V. CONCLUSION

We discussed the topics in this study paper that the developed speech recognition system, trained on a meticulously pre-processed dataset, achieved an accuracy of 75.60%. The use of CNNs was effective in converting spoken words into text, though there were challenges in noisy environments. The findings imply that although the model works well, it may still be improved, and future research will explore using a Conv1D model to further enhance accuracy, particularly in noisy conditions. The research highlights the importance of preprocessing, model selection, and the potential for optimization in achieving better speech-to-text conversion results.

ACKNOWLEDGEMEN

I, Ifa Fatima would like to thanks Prof. Khondelkar Lutful Hassan and Moinuddin, Aliah University, for their insightful discussions and suggestions, which greatly enriched the quality of this work during the preparation of this technical paper.

REFERENCES

[1] Omar Adil Mahdi, Mazin Abed Mohammed, and Ahmed Jasim Mohamed "Implementing a novel approach convert an audio compression to text coding via hybrid technique"-IJCSI, 2012.

[2] Chung Ming Chien, Mingjiamei Zhang, Ju Chieh Chou, and Karen Livescu "Few-shot spoken language understanding via joint speech-text models"-University of Chicago, 2023.

[3] Santosh K Gaikwad, Bharti W Gawali, Pravin Yannawar "A review on speech recognition technique"-IJCA10(3), 16-24, 2010.

[4] Osama A Hamid, Abdel R Mohamed, Hui jiang, Li Deng, Gerald Penn, Dong Yu "Convolutional Neural Network for Speech Recognition" –IEEE 22(10), 1533-1545, 2014.

[5] Tan Lee, Wai Lau, Y.W. Wong, P.C. Ching "Using tone information in continuous speech recognition"- ACM (2002).

[6] Suman K. Saksamudre, P. Shrishrimal, R. Deshmukh "A Review on Different Approaches for Speech Recognition System" -22 April 2015-IJCA.

[7] Sakshi Dua, Sethuraman Sambath Kumar, Yasser Albagory, Rajakumar Ramalingam, Ankur Dumka Rajesh Singh, Mamoon Rashid, Anita Gehlot, Sultan S. Alshamrani and Ahmed Saeed Alghamdi- "Developing a speech recognition system for recognizing tonal speech signals using a CNN"-2022.

[8] A. Kumar and R K Agarwal "Discriminatively trained continuous Hindi speech recognition using integrated acoustic features and recurrent neural network language modelling"- Journal of Intelligent Systems. Available: https://doi.org/10.1515/jisys-2018-0417.

[9] Shahana Bano, Pavuluri Jithendra, Gorsa Lakshmi Niharika, Yalavarthi Sikh "Speech to Text Translation enabling Multilingualism" - 2020 IEEE International Conference for Innovation in Technology.

[10] Bart Decadt, Jacques Duchateau, Walter Daelemans and Patrick Wambacq "Memory-Based Phoneme-to-Grapheme Conversion"- CLIN 2002.

[11] Vaishali A. Kherdekar, Dr. Sachin A. Naik "Convolution Neural Network Model for Recognition of Speech for Words used in Mathematical Expression" - Vol.12 No.6 (2021), 4034-4042 Turkish Journal of Computer and Mathematics Education (TURCOMAT).

[12] Shaheena Sultana, M. A. H. Akhand, Prodip Kumer Das, M. M. Hafizur Rahman "Bangla Speech-to-Text Conversion using SAPI" - International Conference on Computer and Communication Engineering (ICCCE 2012).

[13] Kaggle (http://download.tensorflow.org/data/speech_commands_v0.01.tar.gz.)