JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

SPEECH RECOGNITION USING TINY ML

[1]Rutik Namdeo Rathod,
[2]Sopan Nagorao Waghmare,
[3]Pooja Mahadev Survase,
[4]Aishwarya Jagatsing Gaherwar,
[5]Prof. Swapnali R. Bhujbal
[1](Student (B.E. Computer) P K Technical Campus (Chakan, Pune)),
[2](Student (B.E. Computer) P K Technical Campus (Chakan, Pune)),
[3](Student (B.E. Computer) P K Technical Campus (Chakan, Pune)),
[4](Student (B.E. Computer) P K Technical Campus (Chakan, Pune)),
[5]Professor (M.E Computer) P K Technical Campus (Chakan, Pune))

Abstract: Object We are going to build an embedded application that uses an 18 KB model, trained on a dataset of speech commands, to classify spoken audio. The model is trained to recognize the words "yes" and "no,' and is also capable of distinguishing between unknown words and silence or background noise. Our application will listen to its surroundings with a microphone and indicate when it has detected a word by lighting an LED or displaying data on a screen, depend it on the capabilities of the device. Understanding this code will give you the ability to control any electronics project with voice commands.

Keywords **– ESP EYE, SPECTOGRAM, EASE, LOW POWER CONSUMPTION, SCALABILITY, ETC.**

I. INTRODUCTION

The model we use in this chapter is trained to recognize the words "yes and "no, and is also capable of distinguishing between unknown words and silence or background noise. The model was trained on a dataset called the Speech Commands dataset. This consists of 65,000 one-second-long utterances of 30 short words, crowdsourced online.

Well-researched Although the dataset contains 30 different words, the model was trained to distinguish between only four categories: the words "yes" and no," "unknown" words (meaning the other 28 words in the dataset), and silence. The model takes in one second's worth of data at a time., It outputs four probability scores, one for each of these four classes, predicting how likely it is that the data represented one of them. However, the model doesn't take in raw audio sample data. Instead, it works with spectrograms, which are two-dimensional arrays that are made up of slices of frequency information, each taken from a different time window.

II. METHODOLOGY

Main loop : like the "hello world" example, our application runs in a continuous loop. All of the subsequent processes are contained within it, and they execute continually, as fast as the microcontroller can run them, which is multiple times per second. Audio provider : The audio provider captures raw audio data from the microphone. Because the methods for capturing audio vary from device to device, this component can be overridden and customized. Feature provider : the audio provider converts raw audio data into the spectrogram format that our model requires. It does so on a rolling basis as part of the main loop, providing the interpreter with a sequence of overlapping one-second window. TF Lite interpreter : the interpreter runs the TensorFlow Lite model, transforming the input spectrogram into a set of probabilities. Model: the model is include as a data array and run by the interpreter. The array is located in tiny_conv_micro_features_model_data.

Components

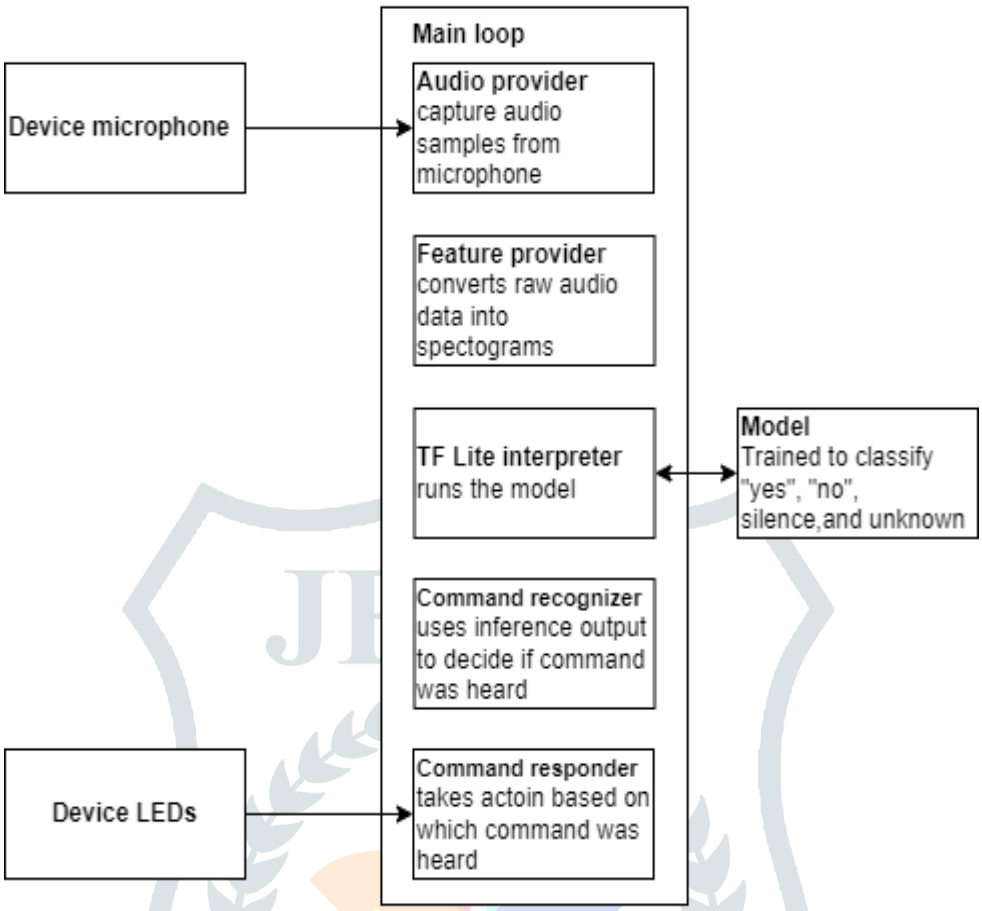


Fig: System architecture

III. COMPONENTS

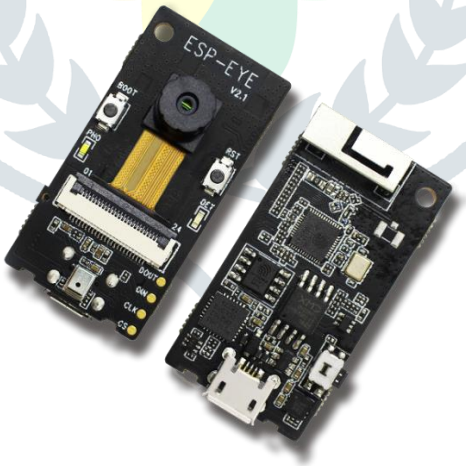


Fig: ESP EYE

ESP-EYE (ESP32) is a compact development board based on Espressif's ESP32 chip microphone ESP-EYE also offers plenty of storage, with 8 MB PSRAM and 4 MB SPI flash - and it's fully supported by Edge Impulse. You'll be able to sample raw data, build models, and deploy trained machine learning models directly from the studio. It's available for around 22 USD from Mouser and a wide range of distributors.

Fig: MICROPHONE DEVICE

Speech recognition starts by taking the sound energy produced by the person speaking and converting it into electrical energy with the help of a microphone. It then converts this electrical energy from analog to digital, and finally to text. It breaks the audio data down into sounds, and it analyzes the sounds using algorithms to find the most probable word that fits that audio. Speech recognition is a machine's ability to listen to spoken words and identify them and convert the spoken words into text, make a query or give a reply.

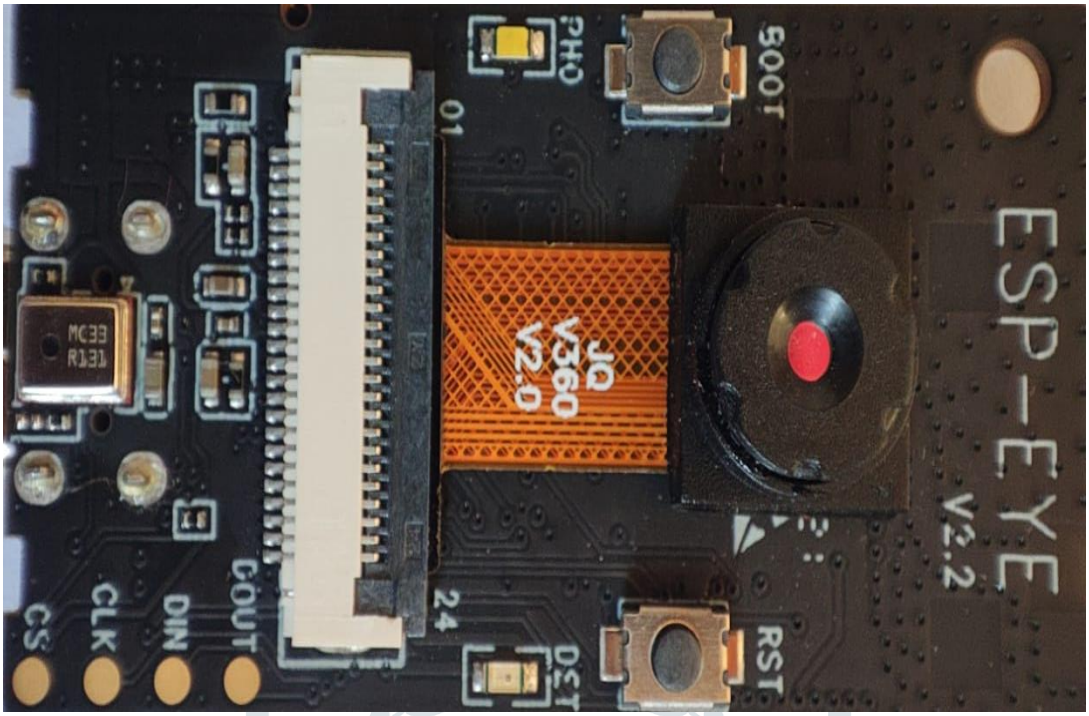
IV. LITERATURE SURVEY

[1] Speech recognition using Linear Predictive Coding (LPC) and Artificial Neural Network (ANN) for controlling movement of mobile robot. Input signals were sampled directly from the microphone and then the extraction was done by LPC and ANN. Ms.Vimala.C and Dr.V.Radha (2012) proposed speaker independent isolated speech recognition system for Tamil language. Feature extraction, acoustic model, pronunciation dictionary and language model were implemented using HMM which produced 88% of accuracy in 2500 words. Cini Kurian and Kannan Balakrishnan (2012) found development and evaluation of different acoustic models for Malayalam continuous speech recognition. In this paper HMM is used to compare and evaluate the Context Dependent (CD), Context Independent (CI) models and Context Dependent tied (CD tied) models from this CI model 21%. The database consists of 21 speakers including 10 males and 11 females. Suma Swamy et al. (2013) introduced an efficient speech recognition system which was experimented with Mel Frequency Cepstrum Coefficients (MFCC), Vector Quantization (VQ), HMM which recognize the speech by 98% accuracy. The database consists of five words spoken by 4 speakers at ten times. Annu Choudhary et al. (2013) proposed an automatic speech recognition system for isolated and connected words of Hindi language by using Hidden Markov Model Toolkit (HTK). Hindi words are used for dataset extracted by MFCC and the recognition system achieved 95% accuracy in isolated words and 90% in connected words. Preeti Saini et al. (2013) proposed Hindi automatic speech recognition using HTK. Isolated words are used to recognize the speech with 10 states in HMM topology which produced 96.61%.

[2] Noticed that End-to-End Automatic Speech Recognition models have shown superior performance over the traditional hybrid ASR models. But training an end-to-end ASR requires a lot of data which is not only expensive but may also raise dependency on production data. It consists of a multi-context text-to speech engine to generate synthetic speech, and an RNN model for speech recognition. The model for the text to-speech engine is trained and evaluated independently from the speech recognition model. The training data are sampled from a combination of real speech recordings and TTS based synthetic speech audio. The ratio between real recordings and synthetic audio seen during RNN training is optimized with sampling weights. This method well mixes the real and synthetic data in each batch so that the ASR model sees both data.

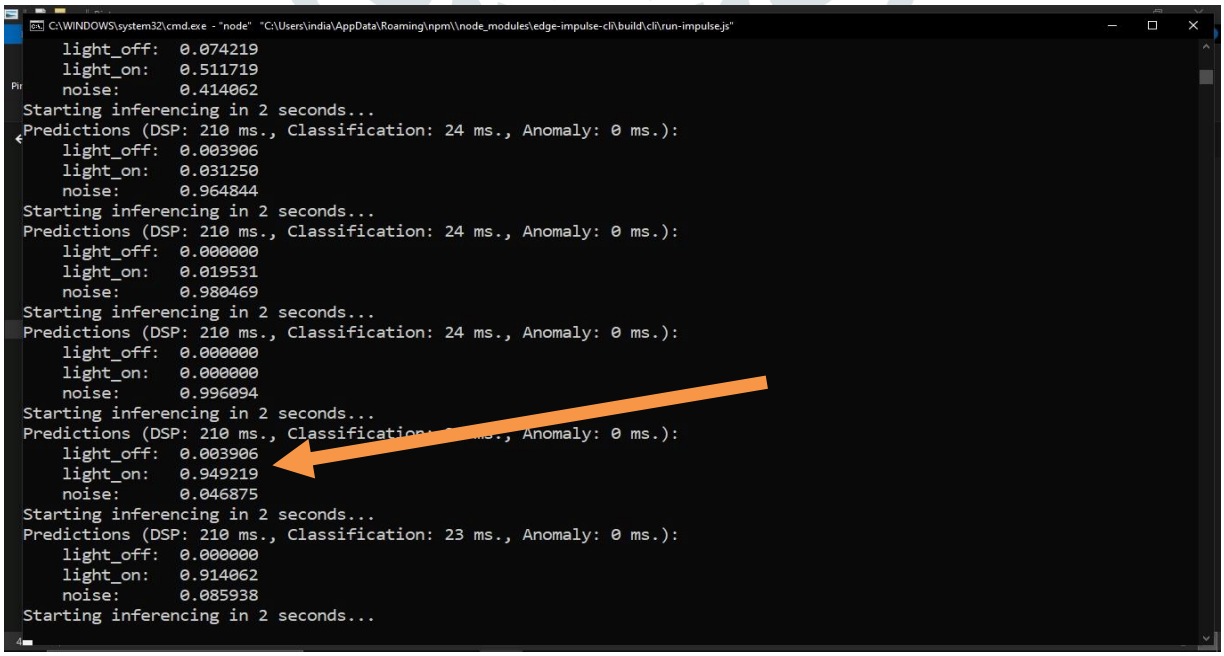
[3] Speech recognition is becoming a more useful technology in computer applications. Many interactive speech-aware applications exist in the field. In order to use this kind of easy way of communication technique into the computer field, speech recognition technique has to be evolved. The computer has to be programmed to accept the voice input and then process it to provide the required output, using various speech recognition software. Speech recognition is the process of converting speech signal to a sequence of words using appropriate algorithm. This provides an alternative and efficient way for the people who are not well educated or not having sufficient computer knowledge to access the systems and where typing becomes difficult. This speech recognition technique also reduces the manpower to accept and process the commands. In our research work, we have to implement this speech recognition technique in customer care center, where many queries have to be processed every day.

V. MODEL

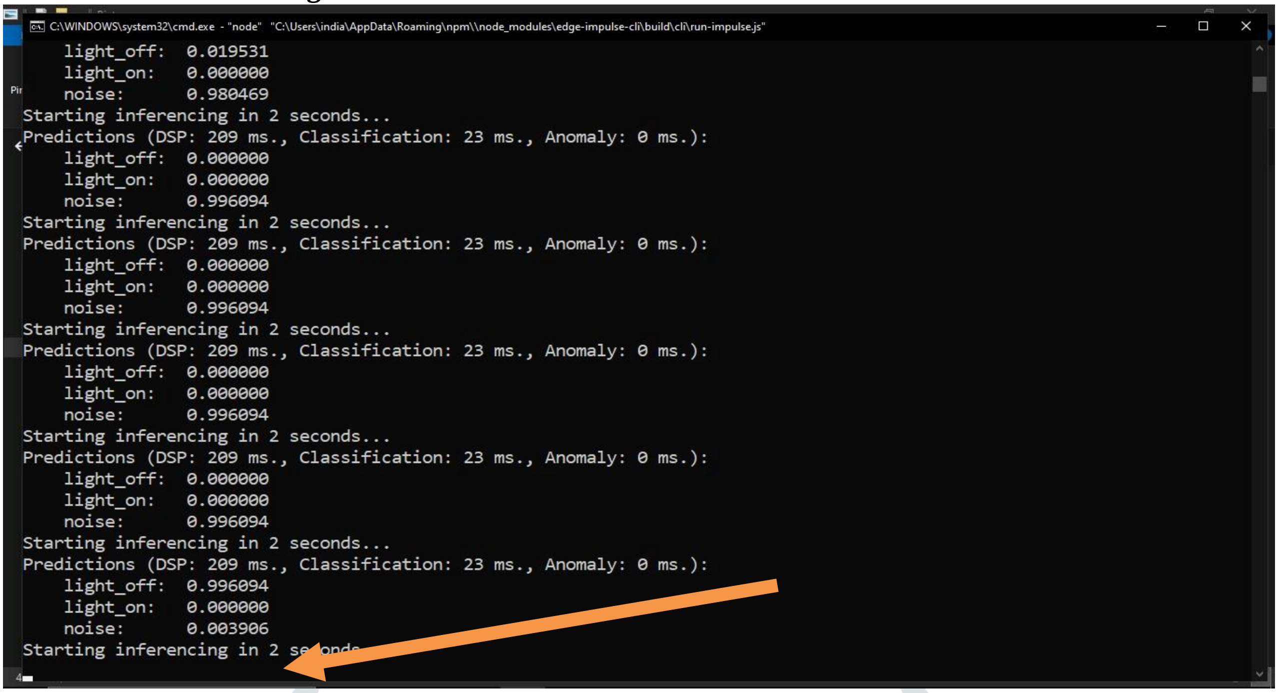


VI. OUTPUT

A). When Voice Detected For Light On



B). When Voice Detected For Light Off



VII. CONCLUSION

We have build and embedded applications that uses 18kb model, trained on a dataset of speech commands to classify spoken audio. The model is trained to recognize the world "yes" and "no" and is also capable of distinguishing between unknown words and silence or background noise. Our application will listen to its surroundings with a microphone and indicate when it has dedicated a word by lighting an LED or displaying data on a screen, depending on the capability of the device.

VIII. REFERENCES

- S. Davis and P. Mermelstein, "Comparison of parametric representationsfor monosyllabic word recognition in continuously spoken sentences,"IEEE Trans. Acoust. Speech Signal Process., vol. 28, no. 4, pp. 357–366,Aug. 1980, doi: 10.1109/TASSP.1980.1163420.
- De Wachter, M., et al., Template-based continuous speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 2007. 15(4): p. 1377-1390
- Resch, B., Automatic Speech Recognition with HTK. Signal Processing and Speech Communication Laboratory. Inffeldgase. Austria. Disponible en Internet: http://www. igi. tugraz. at/lehre/CI, 2003.
- Kekre, H., A.A. Athawale, and G. Sharma. Speech recognition using vector quantization. in Proceedings of the International Conference & Workshop on Emerging Trends in Technology. 2011..
- Moore, R.K., Twenty things we still don't know about speech, in Proc. CRIM/FORWISS Workshop on Progress and Prospects of speech Research an Technology. 1994.
- Tebelskis, J., Speech recognition using neural networks. 1995, Siemens AG.
- Pan, Y., P. Shen, and L. Shen, Speech emotion recognition using support vector machine. International Journal of Smart Home, 2012. 6(2): p. 101-108.
- Anusuya, M. and S.K. Katti, Speech recognition by machine, a review. arXiv preprint arXiv:1001.2267, 2010.
- Alhawiti, K.M., Advances in Artificial Intelligence Using Speech Recognition. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 2015. 9(6): p. 1397-1400.
- Furui, S., Theory and Applications. 1 ed. Speech Technology, ed. K.J. Fang Chen. US: Springer. XXVII, 331.
- Gulzar, T., et al., A systematic analysis of automatic speech recognition: an overview. Int. J. Curr. Eng. Technol, 2014. 4(3): p. 1664-1675.
- Sambur, M.R. and L.R. Rabiner, Statistical decision approach to the recognition of connected digits. The Journal of the Acoustical Society of America, 1976. 60(S1): p. S12-S12.
- Rabiner, L. and J. Wilpon, A simplified, robust training procedure for speaker trained, isolated word recognition systems. The Journal of the Acoustical Society of America, 1980. 68(5): p. 1271-1276.
- Lee, K.-F. and H.-W. Hon. Large-vocabulary speaker independent continuous speech recognition using HMM. in Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on. 1988. IEEE.
- Kita, K., T. Kawabaa, and T. Hana zawa. HMM continuous speech recognition using stochastic language models. in Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on. 1990. IEEE.
- Suzuki, H., et al. Speech recognition using voice characteristic- dependent acoustic models. in Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP`03). 2003 IEEE International Conference on. 2003. IEEE.