International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-05, Issue-12, Mar 2020

Speech Translation Device Using Raspberry Pi

[1]Amrita Dwivedi, [2]Anushka Shukla, [3]Chanchal Sahgal, [4]Kajal Dubey, [5]Prof. Prashant Raghuvanshi, [6]Dr. Saurabh Jain

[1,2,3]Electronics and Electrical Engineering, [4]Computer Science and Engineering, Medi-Caps University, Indore, INDIA, [1]*amritadwivedi9816@gmail.com*, [2]*akhsuna03@gmail.com*, [3]*chanchalsahgal17@gmail.com*, [4]*kajal91097dubey@gmail.com*

***Abstract*: Electronic translators are small portable devices that translate words from one language to the other. They are mainly devised for tourists and entrepreneurs. Using various different techniques, it can scan words that you speak and then it translates them into English or the language and the person infant can listen to the speech in their language which is the output of the device. Most of the devices carry out text to speech conversion or speech to text conversion but some can also do speech to speech conversion. It is genius translation device that can scan the words you speak, translate it in fractions of seconds and say them right back at you in another language with or without the use of internet. Aside from the shortcomings that involves text translation, the device also has to deal with some other problems occurring in speech-to-speech translation such as inarticulate translation of the spoken language and the linguistic errors of the spoken language and the correction of these errors. In this particular case, we would be translating English sentences to Marathi (or any other regional language). The device works on the following operating principle: The device has a pre-stored database that consists of vast characters of the language which the system uses and decide the translation of a given sentence or a word or a whole paragraph. Thus, larger the database of the words or sentences and their key translation, better will be the accuracy of the translation system.**

***Keywords:* Decoder, Encoder, Python, Raspberry Pi, RNN, Speech Translation.**

I. INTRODUCTION

Speech translation is a process in which the vernacular spoken phrases are instantly translated and spoken aloud in another language. This differs from phrase translation in which the system only translates a fixed and finite set of phrases that have been manually entered into the system. Speech translation helps speakers who use various languages to communicate with each other. It is therefore of immense importance for humans due to diversity in languages and dialects and also for the progress of global business. So, when you speak something in one language, and you can press a button and it will read it out and pronounce it for you so you know how to say it in a different language. A compact and easy to use translator device allowing you to travel, communication, with which social networking is no longer a language barrier. In this particular case, we would be translating English sentences to Marathi (or any other regional language). The principle of translating in STD is simple: a system decides an appropriate translation of an input sentence by analyzing the pre-translated sentences in the database. Therefore, the larger the database of pre-translated sentences, greater will be the accuracy of the STD system. "Overcoming the Language Barrier with Speech Translation Technology" by Satoshi, Nakamura in Science & Technology Trends - Quarterly Review No.31 April 2009.

A speech translation system would typically integrate the following three software technologies: automatic speech recognition (ASR), machine translation (MT) and voice synthesis (VS).

A person who speaks certain language say language X talks into the microphone of the device and the speech recognition module then identifies it. It then compares this data input with the input in the library which consists of large amounts of data. The input is then converted into a string of words, using dictionary and grammar of language X. The machine translation module then translates this string. Early systems replaced every word with a corresponding word in another language say Y. Current systems do not use the word to word translation approach as it takes a lot of time and instead, they take try to analyze the meaning of the sentences altogether and then translates them at once. This translated text then goes to the speech synthesis module where the waveforms matching the text are selected the database and speech synthesis connects them and gives the output.

"Speaker Independent Connected Speech Recognition- Fifth Generation Computer Corporation". Fifthgen.com.

 DOI : 10.35291/2454-9150.2020.0201

Archived from the original on 11 November 2013. Retrieved 15 June 2013.

Speech recognition is a subfield of computational linguistics that develops the technology and methodologies which authorizes the recognition and the translation of spoken language into text with the help of computers. It is also called as the Automatic Speech Recognition (ASR).Some speech recognition devices need require a lot of training which comes under the machine learning area and these devices are "system dependent" whereas others have a library of several keywords and the their meanings that help in the speech translation and are "speaker independent"

"British English definition of voice recognition". Macmillan Publishers Limited. Archived from the original on 16 September 2011. Retrieved 21 February 2012.

The term voice recognition refers to generally discerning the speaker and instead of what they are speaking. Distinguishing the speaker can ease the task of translating the speech in systems that are generally speaker dependent as they are trained with a specific person's voice and can be used to recognize the speaker for security purposes.

II. RECURRING NEURAL NETWORK

RNNs are a special sort of neural network with loops that allow information to persist throughout different steps during a network. The loop makes the neural network return and check what happened altogether of the previous words before deciding what the present word actually means. A RNN are often thought of as copy-pasting an equivalent network over and once again, with each new copy-paste adding a touch more information than the previous one. The applications for RNN's are vastly different from traditional NNs because they do not have an output and input set as a concrete value, instead, they take sequences because the input or output. RNN can be used for Natural Language Processing, image/Video Captioning, translation and much more.

In Recurrent Neural Network (RNN) the output from previous step is fed back as the input to the present step. Generally, in the neural networks, all the inputs and outputs are not dependent on each other, but when there is prediction of the next word is involved, we need the previous words and therefore a requirement to collect the previous words is needed. Thus, RNN came into existence, which solved this issue with the assistance of a Hidden Layer. The most and most vital feature of RNN is Hidden state, which remembers some information a few sequence.RNN have a "memory" which remembers all information about what has been calculated. It uses an equivalent parameter for every input because it performs an equivalent task on all the inputs or hidden layers to supply the output. Therefore, all three layers are generally joined together such that the weights of the hidden layers is in one recurrent layer. Advantages of Recurrent Neural Network: An RNN remembers each information through time. It's useful in statistic prediction only due to the feature to recollect previous inputs also. This is often called Long Short-Term Memory.

Training an RNN may be a very difficult task.

The different RNN models that it can follow are:

- Fixed to Sequence: The RNN takes an input of fixed size and outputs a sequence.
- Sequence to Fixed: The RNN takes in an input sequence and outputs a hard and fast size. Sentiment analysis where a given sentence is assessed as expressing positive or negative sentiment 3.
- Sequence to Sequence -The RNN takes in an input sequence and outputs a sequence.

Drawbacks of current devices include:

1. The translator devices which already exists do not convert Indian regional languages.

2. Most of the translators which convert Indian languages, are in the form of apps supporting IOS or android, and they do not exist as an individual single purpose device. Thus, the whole system cost increases. The devices which already exists, use internet for database. They do not have their individual data base. Thus, with the failure of internet, the device stops working. Even if there is offline conversion, they only convert from speech to text but not speech to speech.

So we are trying to design a device with the sole purpose of having a device which converts English to any other Indian regional languages. This device is only for the use of language translation, thus would definitely cost much lower than phones and will work with same efficiency without the internet as well.

III. EXPERIMENTAL APPARATUS

A schematic illustration of the experimental apparatus is given in figure -1. It consists of three main sections:

1) Microphone and speaker

2) Button

3) Raspberry Pi 3 model B+

4) Technology

Microphone and speaker: A microphone is an energy conversion device that converts a sound signal into an electrical signal and a speaker is a transfusing device which converts an electrical signal into an acoustic signal which is produced by humans.

 DOI : 10.35291/2454-9150.2020.0201

Button: A switch is a lever that can be pressed to turn a device on or off. When the switch is turned ON, a current start flowing in the device and it starts functioning.

Raspberry Pi 3 Model B+: Raspberry Pi is a low-cost, basic computer. The Raspberry Pi is contained on a single circuit board and features ports for HDMI, USB 2.

Figure 1: Schematic illustration of translation device

Analog audio and SD Card: The computer run entirely on open-source software and is able to mix and match software according to the work we wish to do. In our model Raspberry pi 3 model B+ is used.

Technology: RNN

In this project, we used the recurrent neural network (RNN) that functions as part of a machine translation process. RNN is a class of artificial neural networks where nodes are connected to form a directed graph as a temporal sequence, and thus exhibits temporal dynamic behavior. They are called recurrent because the network's hidden layers contextual information performs zut from previous time steps becomes input at the current time step. This recurrence works in a form of memory. RNN can use this memory to process different length sequences of inputs. This process works as follows: it accepts English speech as input and returns the Marathi translated speech.

IV. LIBRARIES USED

A. *Pandas*: Python programming language has a software library- pandas for analysis and data manipulation . It offers data structures and operations for manipulating time series and numerical tables and data structures. It is a free software . in machine learning pandas is mainly used in form of data frames. Pandas allow importing data of various file formats such as excel, csv etc. Operations such as concatenation, group by , merge, join, melt, as well as data cleaning features such as replacing ,filling or imputing null values for data manipulations are offered by pandas.

B. Nu*mPy* :With a large collection of high-level mathematical functions to operate on , multi-dimensional arrays and matrices , Python programming language has a NumPy library .

C.*String*: Python string module contains a number of functions to process strings used in the program., as string methods most string operations are made, and many functions in the string module call the corresponding string method using simply wrapper functions.

D. *Re*: A special sequence of characters are used in python that helps you match or find other strings or sets of strings these sequences are called regular expression.

F. *Sklearn*: In python, Scikit-learn is a free machine learning library. Sklearn supports Python numerical and scientific libraries like NumPy and SciPy. Various algorithms like support vector machine, k-neighbors and random forests are sklearn features.

G. *Keras*: In python an open-source neural-network library is keras. For operating on TensorFlow, Microsoft Cognitive Toolkit, R, Theano, or PlaidML, we use keras. Keras is being user-friendly, modular, and extensible

V. DATASET

We used this dataset : mar-eng.zip https://www.manythings.org/anki/ .For our project and increased the dataset with few hundreds of data to improve the efficiency .We can see that this is quite small vocabulary for the dataset. As this allows us to train the model in a reasonable time.

Before starting to build the models, we need to prepare some data. For both English and Marathi, we compute the vocabulary. We also compute the length of maximum sequence for both languages to convert a given token into integer and vice versa we create two python dictionaries for each language, and then to load the data in batches we write a python generator function by making 90-10 train and test split. After this model required for training is defined. It takes slightly more than 2 hours to train.

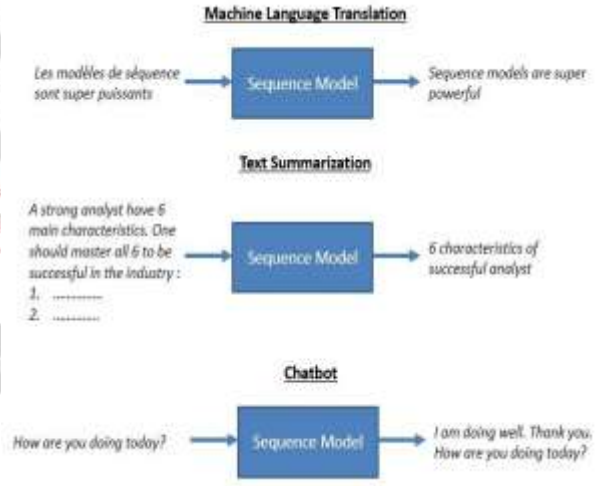

Figure2- Machine Language translation using sequence models.

(source: https://www.analyticsvidhya.com/blog/2018/04/sequence-modelling-an-introduction-with-practical-use-cases/)

FEATURES

- Works offline.
- Support for Indian Regional Languages.
- Easy to use.
- Removes language barrier.
- It can help a person to learn a new language in a better and easier way.
- It could work as a handy oral dictionary.
- Best handy and low-cost device for travelers.

 DOI : 10.35291/2454-9150.2020.0201

VI. PROPOSED MODEL

Recurrent neural network (RNN) is a category of artificial neural networks in which connections between locations form a directed graph in a temporal order.

This allows it to display dynamic temporary performance.

Extracted from the feed forward neural networks, RNNs can use their internal state (memory) to process a sequence of dynamic changes.

Recurrent Neural Networks (or LSTM / GRU directly) have been found to be very effective in solving sequence related problems given the large amount of data.

They have real-time applications in speech recognition, Natural Language Processing (NLP) problems, time series prediction, etc.

Sequence to Sequence (usually abbreviated to seq2seq models) is a special category for the reconstruction of Neural Network structures commonly used (but not limited to) to solve complex language-related problems such as Machine Interpretation, Question Answer, Chat-bots, Text Summary, etc.

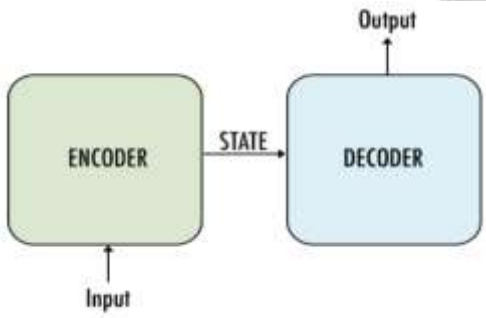

Figure3 - Encoder decoder flowchart
(*source: https://towardsdatascience.com/sequence-to-sequence-model-introduction-and-concepts-44d9b41cd42d)*

A. *Encoder-Decoder Architecture*

The most common architecture used to create Seq2Seq models is the Encoder Decoder architecture.

This is the one we will use for this device. Below is the highest-level view of the building.

B. *Encoder*

We start by embedding our name embedding into one hot word processor. Our vector (embedded in words) is multiplied by the matrix W (hx) of a particular weight.

Our already calculated hidden state (which is the maximum traffic flow of the RNN node) is multiplied by the differential weight of the W (hh) matrix. The effects of these 2 repetitions are added simultaneously and in the absence of a word type such as Relu / tanh. This will be now our hidden state which is h. This process is repeated for the length of our filing sentence. Obviously in the first name of input x0 there is no hidden state in the past so we just set this h0 to be for all zeros.

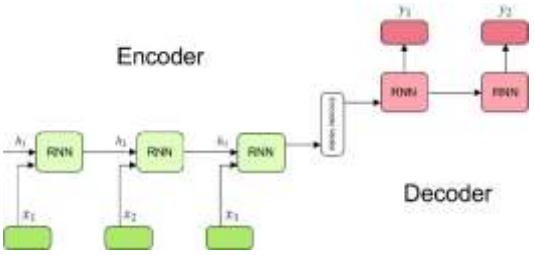

Figure 4-Encoder Decoder RNN

(*source: https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346)*

By looking at the given diagram above, we see that there are three components of the LSTM model:

a) x_i = sequence of input at the i^{th} time stamp.

b) h_i = hidden unit at the i^{th} time stamp.

c) y_i = output unit at the i^{th} time stamp.

We will read the input sequence (English sentence) word by word and observe the internal regions of the LSTM network made after the last step hk, ck (assuming the sentence has the Figure: Encoder decoder flowchart

word 'k'). These integers (say hk and ck) are called input sequences, as they are included in (summarize) all entries in the vector form. Since we will start outputting once we have read all the sequences, the Encoder effect at each time is discarded. Additionally, we should also understand what kind of vectors are Xi, hi and Yi.

C. *Decoder*

The Encoder Long Short-Term memory has same role to play in the training as well as the simulation phase, whereas, a very different role is being played in the training as well as simulation phase by the Decoder Long Short-Term memory. The initial states h0 of the decoder are set to the final states of the encoder is the most important thing. So, this means by intuition that the sequence of output generated by the decoder directly depends in some way or the other by information encoded by the encoder. Obviously, the Marathi sentence which is being translated must in some way depend on the given sentence in English.

In the first step we provide the token START so that the decoder can start generating the next token (first name of the Marathi sentence). We make the decoder learn to predict the _END token as soon as we get the last word of the Marathi sentence. This will be used as a stand-in during the abuse process, basically it will mean the end of the translated sentence and we will stop the inference loop.

Finally, the loss is calculated from the predicted results for each step and the back-propagation errors are propagated in time to refine the network parameters. Training the network for a long time with a large enough number of data results in good guesses (versions) as we will see later.

Ultimately, the encoder summarizes the input sequence to state vectors (sometimes also referred to as hypothetical signals), which are provided in a code that starts outputting the output sequence to the vectors. The decoder is just a language model set in the first regions.

 DOI : 10.35291/2454-9150.2020.0201

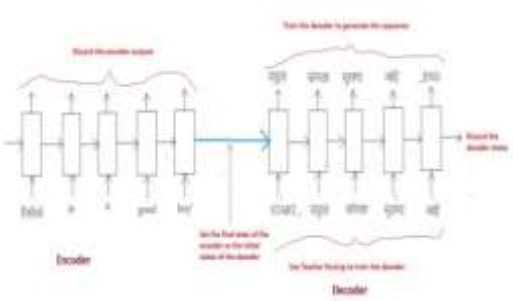

Figure5-Final state

(source: https://towardsdatascience.com/word-level-english-to-marathi-neural-machine-translation-using-seq2seq-encoder-decoder-lstm-model-1a913f2dc4a7)

VII. CONCLUSION

This project takes a speech input then converts it into text. The already present text corpus provided as dataset with a vocabulary size of few hundred words helps to make the machine learn how to translate sentences from English from Marathi .The translated text in Marathi then converts into speech form for output .Even though the results are not the best, but we can note that the words predicted are semantically quite close to the actual meaning of the input sentences. In order to get more accurate results a larger vocabulary of words would be needed.

Input: what is your name?

Actual: □□□□ □□□ □□□ □□□

Predicted: □□□ □□□ □□□□ □□□

Input: where do you live?

Actual: □□ □□□□ □□□□□□

Predicted: □□□□ □□ □□□□□□

Input: i am a doctor

Actual: □□ □□□□□□ □□□

Predicted: □□ □□□□□□

Input: my name is amrita

Actual: □□□□ □□□ □□□□□ □□□

Predicted: □□□□ □□□ □□□ □□□□□

VIII. FUTURE SCOPE

Currently, speech translation technology is available as product that instantly translates free form multi-lingual conversations. These systems instantly translate continuous speech. Challenges in accomplishing this include overcoming speaker-dependent variations in style of speaking or pronunciation are issues that have to be dealt with in order to provide high quality translation for all users. Moreover, speech recognition systems must be able to remedy external factors such as acoustic noise or speech by other speakers in real-world use of speech translation systems. For the reason that the user does not understand the target language when speech translation is used, a method "must be provided for the user to check whether the translation is correct, by such means as translating it again back into the user's language" .In order to achieve the goal of erasing the language barrier worldwide, multiple languages have to be supported. As the collection of corpora is extremely expensive, collecting data from the Web would be an alternative to conventional methods.

REFERENCES

[1] Satoshi Nakamura, "Overcoming the Language Barrier with Speech Translation Technology" Science & Technology Trends - Quarterly Review No.31 April 2009.

[2] "Speaker Independent Connected Speech Recognition- Fifth Generation Computer Corporation". Fifthgen.com. Archived from the original on 11 November 2013. Retrieved 15 June 2013.

[3] "British English definition of voice recognition". Macmillan Publishers Limited. Archived from the original on 16 September 2011. Retrieved 21 February 2012.

[4] "Voice recognition, definition of". Web Finance, Inc. Archived from the original on 3 December 2011. Retrieved 21 February 2012.

[5] "The Mailbag LG #114". Linuxgazette.net. Archived from the original on 19 February 2013. Retrieved 15 June 2013.

[6] Reynolds Douglas, Rose Richard, "Robust text-independent speaker identification using Gaussian mixture speaker models" (PDF). IEEE Transactions on Speech and Audio Processing. 3 (1): 72–83. doi:10.1109/89.365379. ISSN1063-6676. OCLC 26108901. Archived (PDF) from the original on 8 March 2014. Retrieved 21 February 2014.

[7] "Speaker Identification (WhisperID)". Microsoft Research. Microsoft archived from the original on 25 February 2014. Retrieved 21 February 2014.

[8] Albas Thomas Fritz. "Systems and Methods for Automatically Estimating a Translation Time." US Patent 0185235, 19 July 2012.

[9] Rubin P. Baer, T. Mermelstein, P. "An articulatory synthesizer for perceptual research". Journal of the Acoustical Society of America,1981.

[10] MTIL2017: Machine Translation Using Recurrent Neural Network on Statistical Machine Translation.

[11] Sequence to Sequence Learning with Neural Networks, Ilya Sutskever Google ilyasu@google.com, Oriol Vinyals Google vinyals@google.com, Quoc V. Le Google qvl@google.com

[12]https://medium.com/@dev.elect.iitd/neural-machine-translation-using-word-level-seq2seq-model-47538cba8cd7

[13] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[14] G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing - Special Issue on Deep Learning for Speech and Language Processing, 2012

[15] https://www.guru99.com/seq2seq-model.html , seq2seq (Sequence to Sequence) Model for Deep Learning with PyTorch

[16] https://www.analyticsvidhya.com/blog/2018/04/sequence-modelling-an-introduction-with-practical-use-cases/

 DOI : 10.35291/2454-9150.2020.0201