BANSILAL RAMNATH AGARWAL CHARITABLE TRUST'S

VISHWAKARMA INSTITUTE OF TECHNOLOGY PUNE-411037

(An Autonomous Institute Affiliated to University of Pune)

SIGNAL AND IMAGE PROCESSING(SAIP)

Case Study on

TOPIC: - DEEP LEARNING APPROACHES IN SPEECH RECOGNITION

SUBMITTED BY:-

SAURABH JEJURKAR (73)

UNDER THE GUIDANCE OF:

PROF. DR. SHILPA SONDKAR

DEPARTMENT OF INSTRUMENTATION AND

CONTROL ENGINEERING 2020-2021

BANSILAL RAMNATH AGARWAL CHARITABLE TRUST`S

VISHWAKARMA INSTITUTE OF TECHNOLOGY

PUNE-411 037

(An Autonomous Institute Affiliated to University of Pune.)

CERTIFICATE

This is to certify that the SAIP home assignment titled **"Case study- DEEP LEARNING APPROACHES IN SPEECH RECOGNITION"** has been completed by **Saurabh Jejurkar** in the academic year 2020 – 2021, in partial fulfillment of S.Y. B. Tech. **in Instrumentation and Control Engineering.**

Prof. Dr. Shilpa Sondar

(HOD- IC and GUIDE)

Vishwakarma institute of technology ,Pune

Place: Pune **Date:** 23/05/2021

ACKNOWLEDGEMENT

I am pleased to have Vishwakarma Institute of Technology as our Institute.

I would like to thank our honorable Director **(DR.) RAJESH M. JALNEKAR**; HOD of S.Y. Instrumentation and Control Department and our guide **PROF. (DR.) SHILPA SONDKAR** their valuable guidance during the course of this home assignment work. The home assignment would have been uphill task without their continuous direction and unwavering support.

INDEX

CHAPTER 1: INTRODUCTION	06
CHAPTER 2: SIRI	08
CHAPTER 3: CORTANA	14
CHAPTER 4: ALEXA	16
CHAPTER 5: GOOGLE VOICE ASSISTANCE	21
CHAPTER 6: SAMSUNG BIXBY	25
CHAPTER 7: CONCLUSION	28
CHAPTER 8: REFERENCES	29

ABSTRACT

To call any technology that makes our lives easier by one name is almost impossible. There are a variety of terms that refer to agents that can perform tasks or services for an individual, and they are almost interchangeable — but not quite. They differ mainly based on how we interact with the technology, the app, or a combination of both. Here are some basic definitions, similarities, and differences: Intelligent Personal Assistant: This is software that can assist people with basic tasks, us Automated Personal Assistant: This term is synonymous with intelligent personal assistant. Smart Assistant: This term usually refers to the types of physical items that can provide various services by using smart speakers that listen for a wake word to become active and perform certain tasks. Amazon's Echo, Google's Home, Apple's HomePod, Samsungs Bixby are types of smart assistants. Virtual Digital Assistants: These are automated software applications or platforms that assist the user by understanding

Voice Assistant: The key here is voice. A voice assistant is a digital assistant that uses voice recognition, speech synthesis, and natural language processing (NLP) to provide a service through a particular application. Deeps learning change the voice recognition field .

CHAPTER 1

INTRODUCTION

Speech recognition is the computer interpreting the word spoken by human and converting them to machine understandable language. Deep learning which is subset of machine learning and basically known for neural network opens the gate for development in speech recognition. Deep learning make the speech recognition task easy for developer as well as for customers also. So companies like Apple, Amazon, Samsung, Microsoft and Google came up with there voice assistance which is based on deep learning based speech recognition. These voice assistant made huge change in human life. In home automation, education and daily work it helps a lot to human by saving time. So for this speech recognition there are following approaches

1. Simple pattern matching -where each spoken word is recognized in its entirety—the way you instantly recognize a tree or a table without consciously analyzing
2. Pattern and feature analysis -where each word is broken into bits and recognized from key features, such as the vowels it contains
3. Language modeling and statistical analysis -in which a knowledge of grammar and the probability of certain words or sounds following on from one another is used to speed up recognition and improve accuracy
4. Artificial neural networks -brain-like computer models that can reliably recognize patterns, such as word sounds, after exhaustive training.

HMM's in Speech Recognition

HMM stands for hidden Markov model. HMMs are useful for detecting patterns through time. This is exactly what we do with an acoustic model. HMMs can solve the challenge, we identified earlier, of time variability. For instance, in example of speech versus speech, the same word but spoken at different speeds. We could train an HMM with label time series sequences to create individual HMM models for each particular sound unit. The units could be phonemes, syllables, words, or even groups of words. Training and recognition are fairly straightforward if our training and test data are isolated units.

Language Models

We have HMM models that can convert those features into phonemes and address the sequencing problems for our full acoustic model. We haven't yet solved the

problems in language ambiguity though. With automatic speech recognition, the goal is to simply input any continuous audio speech and output the text equivalent. The system can't tell from the acoustic model which combinations of words are most reasonable.
That requires knowledge. We either need to provide that knowledge to the model or give it a mechanism to learn this contextual information on its own.

Deep Neural Network

There is problem with another approaches but deep neural network solve that problem. The algorithms like convolutional neural network, recurrent neural network and long term short term memory algorithm are plays important role in speech recognition. LSTM can solve back proportion problem which is happening in RNN algorithm. Alexa, Siri, Bixby, Cortana and google voice assistance uses these algorithm and they got the very good customer response.
In the deep neural network millions of millions of data is trained and why we are getting the good accuracy during speech recognition. But there was problem before for deep learning. Deep learning training data steps are take too much times in GPU so Google came up with TPU after that this problem got solve. TPU takes less time to run this algorithms.

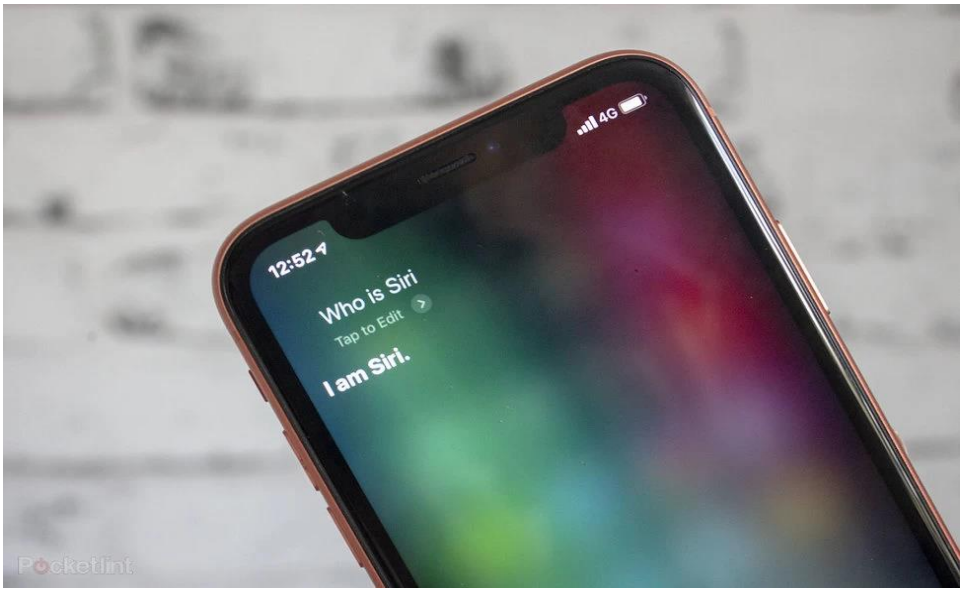
So how this deep neural network work in companies like Google, Microsoft, Amazon, Apple and Samsung understand by the this case study.

CHAPTER 2

CASE STUDY

1] SIRI:-

Siri is the voice assistant for Apple devices. It’s available on Apple's devices like iPhone, iPad, Mac, Apple Watch, Apple TV, and HomePod.



Uses of the Siri:

Siri can do following tasks:

- Make calls
- Send and read texts
- Send messages on third-party messaging apps like whatsapp
- Set alarms and timers
- Set reminders and check calendar
- Split a check or calculate a tip
- Play music (specific songs, artists, genres, playlists)

- Identify songs as well as providing song information like artist and release date
- Control HomeKit products
- Play TV shows and movies, answer questions about tv shows
- Do translations and conversions
- Solve math equations
- Offer up sports scores
- Check stocks
- Surface photos based upon person, location, object, and time
- Apple Maps navigation and directions
- Make reservations
- Open and interact with apps
- Find files (on Mac)
- Send money via Apple Pay
- Check movie times and ratings
- Search for nearby restaurants and businesses
- Activate Siri Shortcuts
- Search and create Notes
- Search Twitter and other apps
- Open the Camera and take photo
- Increase/decrease brightness
- Control settings
- Tell jokes, roll dice, flip a coin
- Play voicemails
- Check the weather

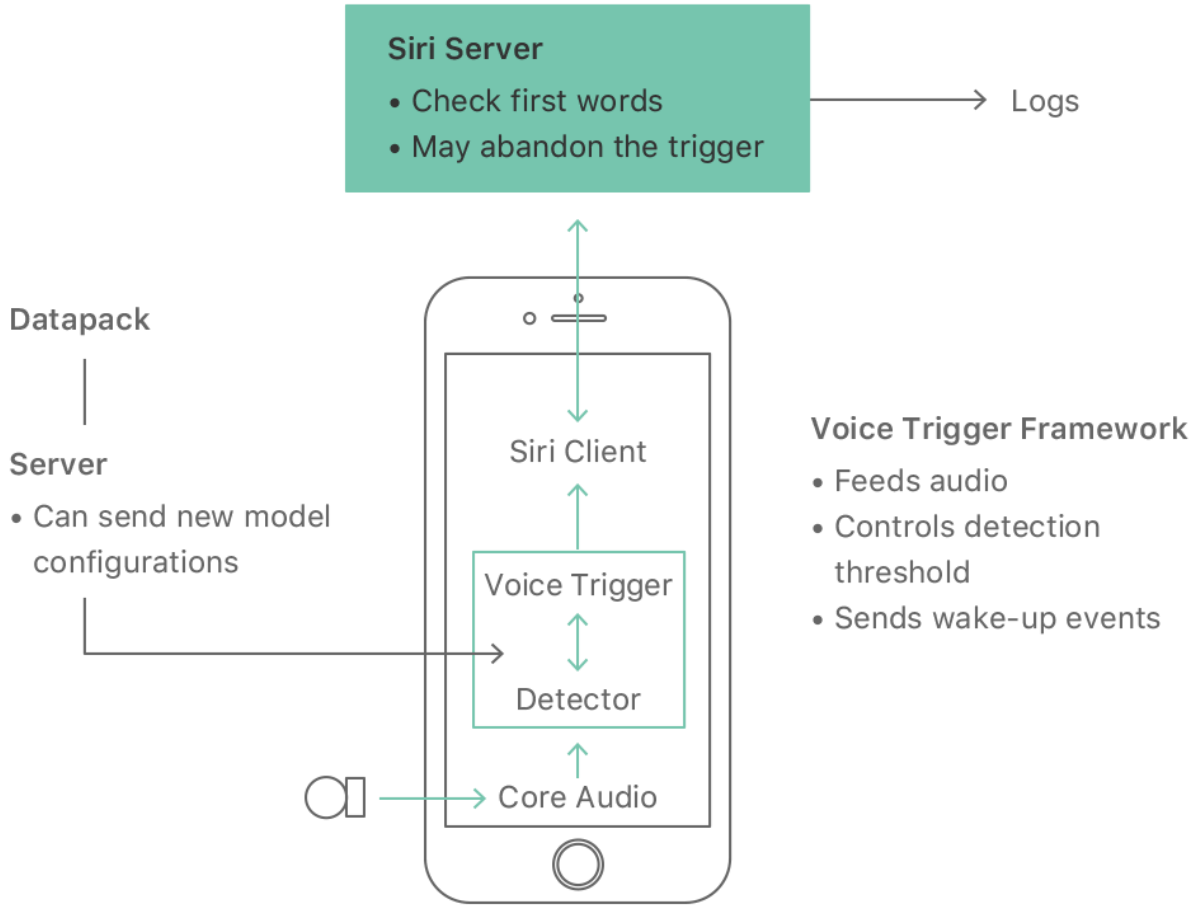
From the above tasks it shows that Siri is very useful in human life. It saves human power and do automatic work , helpful to stay updated in world and improves persons productivity.

Though Siri can do following task there are some disadvantages of siri

- Siri unable to work without active internet connection because it is depends upon the cloud computing
- For map more efficient for English language

- If person talks too fast or with strong accent Siri unable to understand customers query7
- Low quality audio, background noise and slow internet affect the efficiency of Siri
- Spelling variation affects the performance of Siri.

Siri work flow:

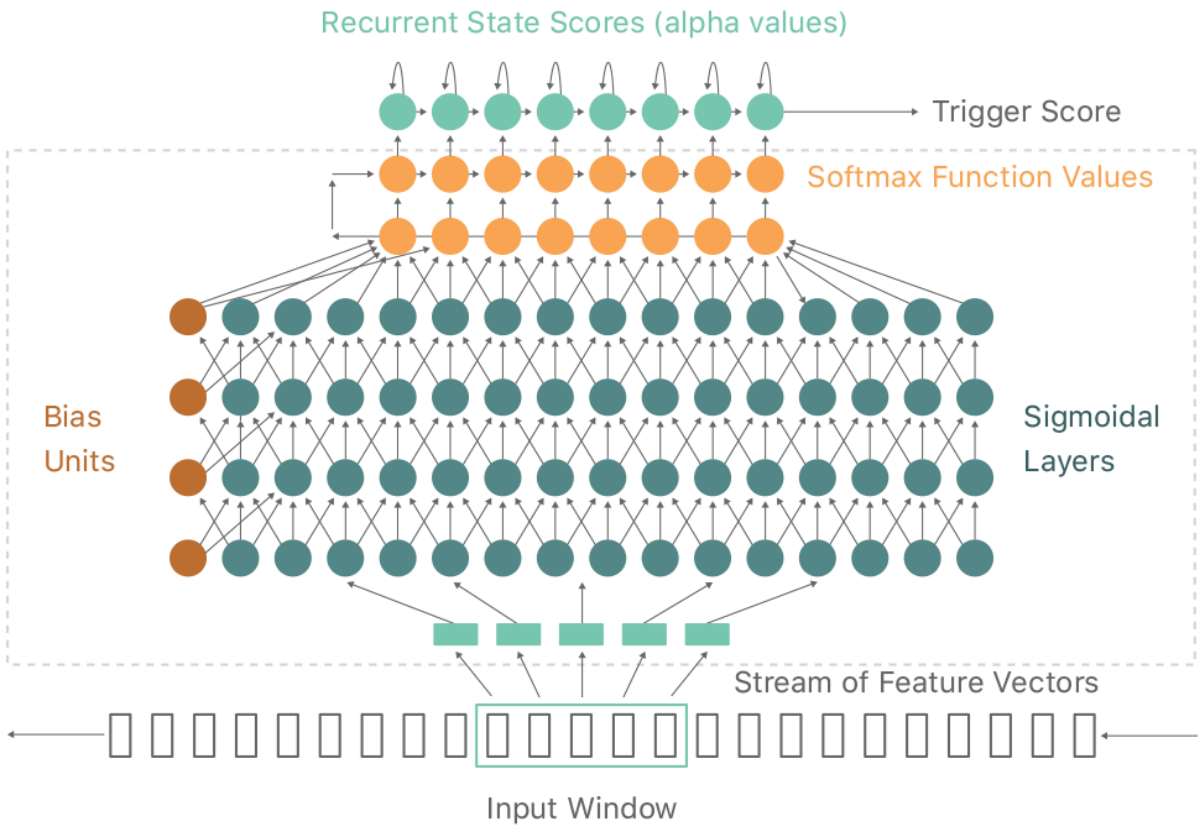


- To start the siri user need to say "Hey Siri".
- Then microphone in mobile/mac take the user sound input. Then that audio go to the Detector.
- Voice trigger framework feeds the audio, Controls detection thresholds and and sent wake up events.
- Siri takes some word to recognize the speech and search that word in acoustic model . Acoustic model is helpful in speech recognition. This model is in the cloud server.
- Then Siri do the task if that are in her acoustic model.

DEEP LEARNING MODEL IN SIRI

Siri detector uses the Deep Neural Network Algorithm for detection of speech. The microphone in an iPhone or Apple Watch turns persons voice into a stream of instantaneous waveform samples, at a rate of 16000 per second. A spectrum analysis stage converts the waveform sample stream to a sequence of frames, each describing the sound spectrum of approximately 0.01 sec. About twenty of these frames at a time (0.2 sec of audio) are fed to the acoustic model, a Deep Neural Network (DNN) which converts each of these acoustic patterns into a probability distribution over a set of speech sound classes: those used in the "Hey Siri" phrase, plus silence and other speech, for a total of about 20 sound classes.

The DNN consists mostly of matrix multiplications and logistic nonlinearities. Each "hidden" layer is an intermediate representation discovered by the DNN during its training to convert the filter bank inputs to sound classes. The final nonlinearity is essentially a Softmax function (a.k.a. a general logistic or normalized exponential), but since we want log probabilities the actual math is somewhat simpler.



THE ACOUSTIC MODEL

The DNN acoustic model is at the heart of the "Hey Siri" detector. How apple train it.Before there was a Hey Siri feature, a small proportion of users would say "Hey Siri" at the start of a request, having started by pressing the button. Apple used such "Hey Siri" utterances for the initial training set for the US English detector model. Apple also included general speech examples, as used for training the main speech recognizer. In both cases, Apple used automatic transcription on the training phrases. Siri team members checked a subset of the transcriptions for accuracy.

Apple created a language-specific phonetic specification of the "Hey Siri" phrase. In US English, Apple had two variants, with different first vowels in "Siri"—one as in "serious" and the other as in "Syria." Apple also tried to cope with a short break between the two words, especially as the phrase is often written with a comma: "Hey, Siri." Each phonetic symbol results in three speech sound classes (beginning, middle and end) each of which has its own output from the acoustic model.

Apple used a corpus of speech to train the DNN for which the main Siri recognizer provided a sound class label for each frame. There are thousands of sound classes used by the main recognizer, but only about twenty are needed to account for the target phrase (including an initial silence), and one large class class for everything else. The training process attempts to produce DNN outputs approaching 1 for frames that are labelled with the relevant states and phones, based only on the local sound pattern. The training process adjusts the weights using standard back-propagation and stochastic gradient descent. Apple have used a variety of neural network training software toolkits, including Theano, Tensorflow, and Kaldi.

This training process produces estimates of the probabilities of the phones and states given the local acoustic observations, but those estimates include the frequencies of the phones in the training set (the priors), which may be very uneven, and have little to do with the circumstances in which the detector will be used, so we compensate for the priors before the acoustic model outputs are used.

Training one model takes about a day, and there are usually a few models in training at any one time. They generally train three versions: a small model for the first pass on the motion coprocessor, a larger-size model for the second pass, and a medium-size model for Apple Watch.

"Hey Siri" works in all languages that Siri supports, but "Hey Siri" isn’t necessarily the phrase that starts Siri listening. For instance, French-speaking users need to say "Dis Siri" while Korean-speaking users say "Siri 야" (Sounds like "Siri Ya.") In Russian it is

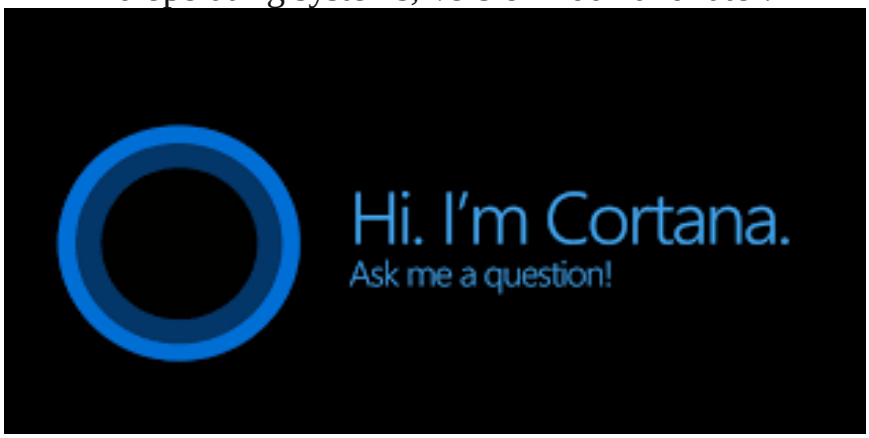
"привет Siri " (Sounds like "Privet Siri"), and in Thai "หวัดดี Siri". (Sounds like "Wadi Siri".)

So there can be another word instead of “Hey Siri” word. According to that acoustic model works.

CHAPTER 3

2]CORTANA

Cortana is voice assistance made by Microsoft company. Microsoft's Cortana is a cloud-based personal assistant that operates outside the realm of standard voice-enabled AI. Cortana doesn't just understand voice commands and carry out tasks but is integrated for use across Microsoft's 365 suite of products and all Windows 10 operating systems, version 2004 and later.



Uses of Cortana:

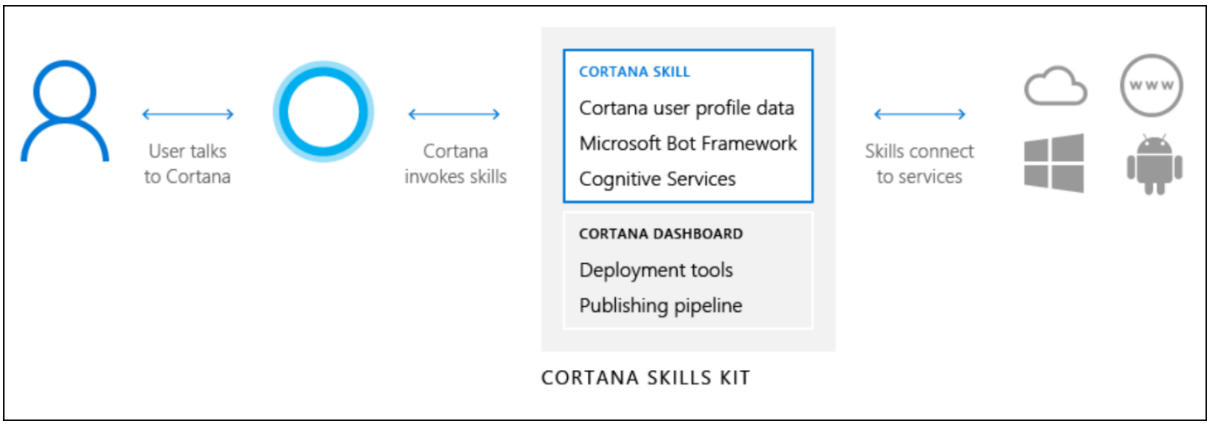
Cortana can do following task:

- Open the file from the pc
- Perform calculation
- Set reminders
- Set alarms
- Open apps
- Sending an email using Cortana
- Create calender event

There are some limitations of Cortana like:

- Lacks functionalities
- Can’t answer contextual answers
- Cannot send emails and text messages on WhatsApp, Instagram
- Can’t open folders through voice commands
- Can’ tell system information like internet speed, disk usage, memory consumption
- Can’t open all desktop applications

Flow Chart:



User give the command to Cortana like “Hey Cortana, open this file”. Then Cortana invokes that skill. Then send that skill to Microsoft Azure cloud. In cloud there is acuastic model who determined the the command and if that command found in its skill then it work according to that command.

Deep learning in Cortana

Cortana uses the deep learning algorithm like Convolutional Neural Network i.e. CNN and Long Short Term Memory deep learning algorithm i.e. LSTM for automatic speech recognition.

CHAPTER 4

3]ALEXA

Alexa is virtual voice assistance developed by Amazon. It is AI based application. It is available as mobile application as well as in hardware as shown in figure given below.

Uses of the Alexa:

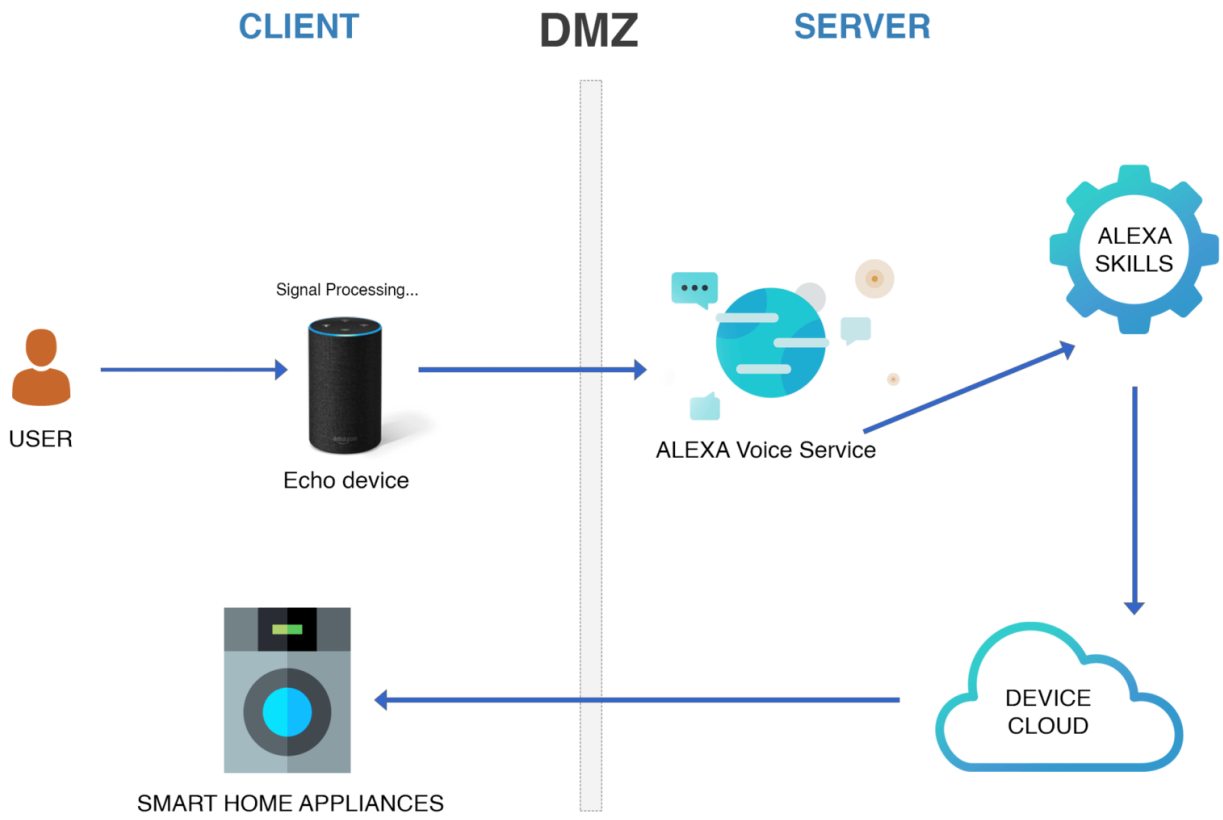
Alexa can be used for following purpose:

- For the home automation in home like to turn on/off light, TV, Fan and AC
- Sport news
- Deliver news
- For information search
- Allow Amazon prime members to buy product from Amazon
- Amazon prime members can listen Amazon prime music.
- She can tell about weather
- Command can be given in different language like Spanish, French and Hindi(in developing stage)
- Helps to check credit card balance
- User able to use as personal trainer
- Play games like tic-tac toe
- Scheduled the meeting
- It can control thermostat
- Drive the car

Limitations of the Alexa:

- It is cloud based so requires internet connection.
- Slow internet may delay the responses.
- Cannot be integrated with desktops and laptops

Work Flow:



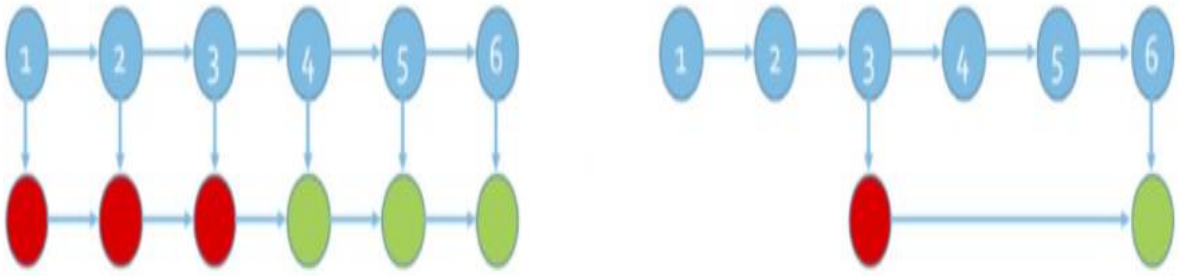
In Alexa there are 7 microphones to get the input from the user . It helps in noise cancelation inn Alexa. “Hey Alexa” is a wake word for Alexa. The wake helps in activation of echo device then echo device listens users instructions. Echo speaker(or Amazon Echo) is basically speaker device for user. By this device user able to speak with Alexa. These devices are available in many models and activate by the specific wake word. These devices are manufactured with preconfigured wake-words.
Innovation name is a keyword which requires to prompt particular Alexa skills. All custom skills must require an invocation name to get start the interaction.
An utterance is what user wants Alexa to execute.
“Hey, Alexa can you start action movie Terminator 3”
In the above example, “Terminator 3” is utterance. Utterances are nothing but the phrases what users use while giving instruction to Alexa. The response from Alexa is

decided and based upon the identified utterance requested by the user. Alexa Voice Service can be considered as the brain of Alexa. This is a suite of services say APIs and tools. These services are configured around Alexa (kinda AI assistant). This service holds the responsibilities of understanding human natural language by taking voice commands from users via echo device. As, AI has machine learning underneath, which further has capabilities like NLP – NLU. This resolves complex voice commands with advanced of computational power and deep learning algorithms.
The services in Alexa Voice Service, are nothing but Alexa skills. Depending upon a voice command, a most appropriate service gets invoked and cater users with the most meaningful response for user's request. Alexa skills development is a niche area which requires developers to implement commanding solutions. These skills are key to success while responding to users with expected results. This is the component, which makes a decision by looking at invocation name and utterance in a voiced sentence, which in turn, concludes user's input, processes it and respond expectedly. The utterance is the phrases which encapsulate the user's desired result. Device cloud receives inputs from Alexa Voice Service (a response by Alexa Skills based upon user's input received). Then, it sends response command signals to an appropriate device connected online with a device cloud to accomplish the action, as instructed by the user. For example, this could be starting of an Air Conditioner or playing a movie on TV.

Deep learning in Alexa

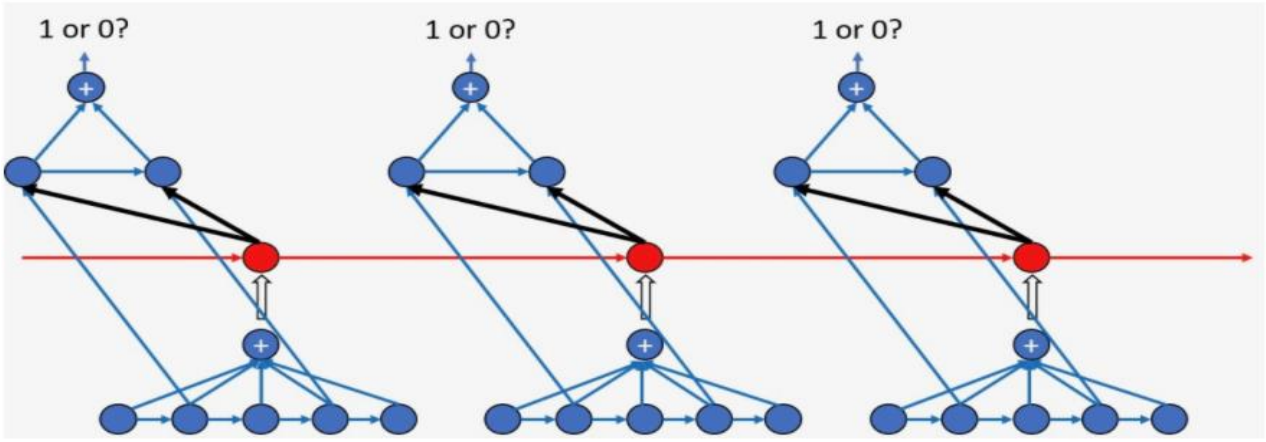
Alexa have Uses Transfer and Adapt Approach (TANDA). And this TANDA approach is build on Google's Transformer . It is trained on large question answering dataset. Alexa used semi-supervised learning to train a system developed from an external dataset to do acoustic-event detection. Semi-supervised learning uses small sets of annotated training data to leverage larger sets of unannotated data. In particular, they use tri-training, in which three different models are trained to perform the same task, but on slightly different data sets. Pooling their outputs corrects a common problem in semi-supervised training, in which a model's errors end up being amplified.
Alexas media detection system is based on the observation that the audio characteristics Alexa would most like to identify are those common to all instances of media sound, regardless of content. Alexas network design is an attempt to abstract away from the properties of particular training examples. Like many machine learning models in the field of spoken-language understanding, Alexa uses recurrent neural networks (RNNs). An RNN processes sequenced inputs in order, and each output factors in the inputs and outputs that preceded it. They use a convolutional neural network (CNN) as feature extractor, and stack RNN layers on top of it. But each RNN

layer has only a fraction as many nodes as the one beneath it. That is, only every third or fourth output from the first RNN provides an input to the second, and only every third or fourth output of the second RNN provides an input to the third.



A standard stack of recurrent neural networks (left) and the "pyramidal" stack Alexa use instead

Because the networks are recurrent, each output we pass contains information about the outputs we skip. But this "pyramidal" stacking encourages the model to ignore short-term variations in the input signal. For every five-second snippet of audio processed by their system, the pyramidal RNNs produce a single output vector, representing the probabilities that the snippet belongs to any of several different sound categories. But Alexa system includes still another RNN, which tracks relationships between five-second snippets. Amazon experimented with two different ways of integrating that higher-level RNN with the pyramidal RNNs. In the first, the output vector from the pyramidal RNN simply passes to the higher-level RNN, which makes the final determination about whether media sound is present. In the other, however, the higher-level RNN lies *between* the middle and top layers of the pyramidal RNN. It receives its input from the middle layer, and its output, along with that of the middle layer, passes to the top layer of the pyramidal RNN.



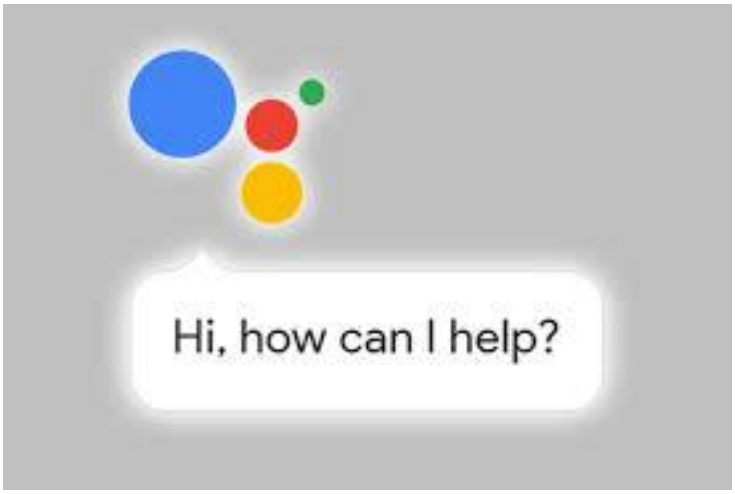
In the second of two contextual models, a high-level RNN (red circles) receives inputs from one layer of a pyramidal RNN (groups of five blue circles), and its output passes to the next layer (groups of two blue circles).

This is Alexas best-performing model. When compared to a model that used the pyramidal RNNs but no higher-level RNN, it offered a 24% reduction in equal error rate, which is the error rate that results when the system parameters are set so that the false-positive rate equals the false-negative rate.

CHAPTER 5

4]GOOGLE VOICE ASSISTANCE

Google voice assistance is AI technology based voice assistance created by Google.



Uses of Google Voice Assistance:-

Google Assistance can do following tasks:

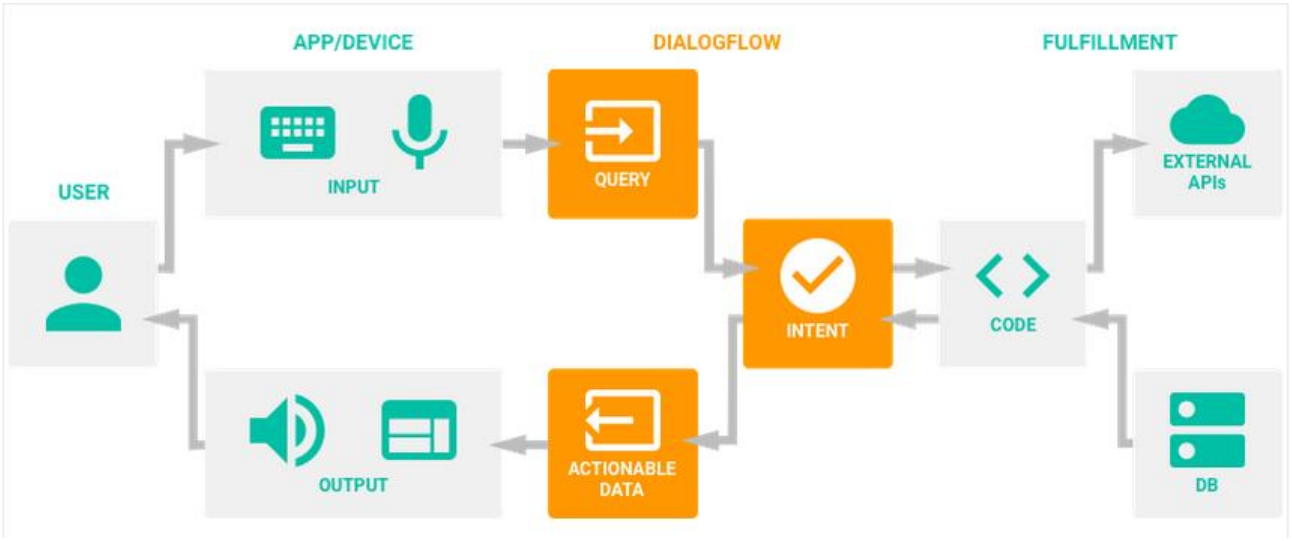
- Make phone calls
- Weather news
- Send SMS
- Search information on Google
- Control users devices
- Real-time spoken translation
- Control music
- Access information from users calender and other personal information
- Run timers and reminders
- Play content on Chromecast or other compatible devices
- Open the apps on the phone
- Read the notifications
- Make the appointment

- Play games

There are some limitations of Google Voice Assistance these are:-

- It is cloud based so it requires internet connection.
- Slow internet may delay the responses.
- Cannot be integrated with desktop and laptops.
- Alexa has edge over google assistant in terms of home automation.
- Alexa feels nicer than talking to Google.

Work flow:

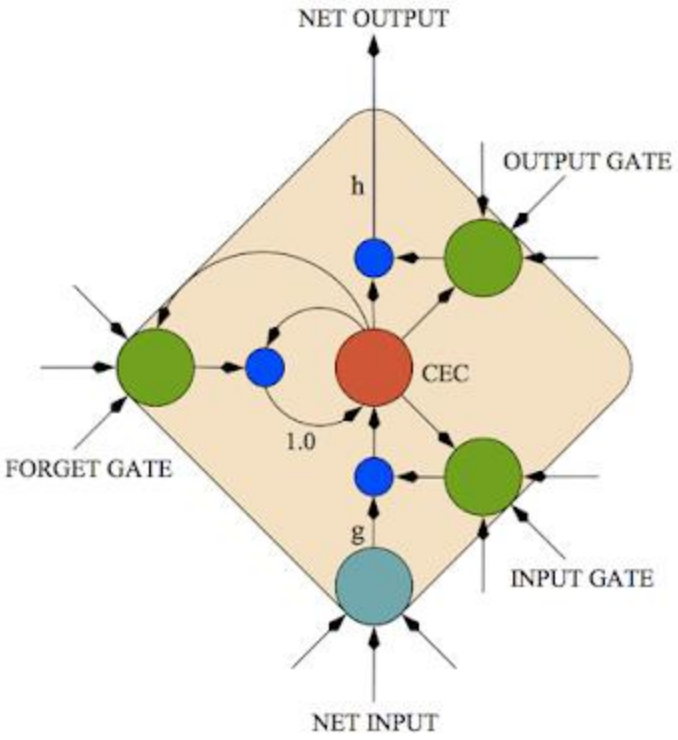


Deep Learning in Google Voice Assistance

In 2015 Google announced improvements to Google Voice transcription using Long Short Term Memory Recurrent Neural Network (LSTM RNNs)—yet another place neural networks are improving useful services. Since it launched in 2009, Google Voice transcription had used Gaussian Mixture Model (GMM) acoustic models, the state of the art in speech recognition for 30+ years. Sophisticated techniques like adapting the models to the speaker's voice augmented this relatively simple modeling method.

Then around 2012, Deep Neural Networks (DNNs) revolutionized the field of speech recognition. These multi-layer networks distinguish sounds better than GMMs by using "discriminative training," differentiating phonetic units instead of modeling each one independently.

But things really improved rapidly with Recurrent Neural Networks (RNNs), and especially LSTM RNNs, first launched in Android's speech recognizer in May 2012. Compared to DNNs, LSTM RNNs have additional recurrent connections and memory cells that allow them to "remember" the data they've seen so far—much as you interpret the words you hear based on previous words in a sentence.
By then, Google's old voicemail system, still using GMMs, was far behind the new state of the art. So we decided to rebuild it from scratch, taking advantage of the successes demonstrated by LSTM RNNs. But there were some challenges



An LSTM memory cell, showing the gating mechanisms that allow it to store and communicate information.

There's more to speech recognition than recognizing individual sounds in the audio: sequences of sounds need to match existing words, and sequences of words should make sense in the language. This is called "language modeling." Language models are typically trained over very large corpora of text, often orders of magnitude larger than the acoustic data. It's easy to find lots of text, but not so easy to find sources that match naturally spoken sentences. Shakespeare's plays in 17th-century English won't help on voicemails.
Google decided to retrain both the acoustic and language models, and to do so using existing voicemails. Google already had a small set of voicemails users had donated for research purposes and that Google could transcribe for training and testing, but

we needed much more data to retrain the language models. So Google asked their users to donate their voicemails in bulk, with the assurance that the messages wouldn't be looked at or listened to by anyone—only to be used by computers running machine learning algorithms. But how does one train models from data that's never been human-validated or hand-transcribed?

Google couldn't just use our old transcriptions, because they were already tainted with recognition errors— garbage in, garbage out. Instead, Google developed a delicate iterative pipeline to retrain the models. Using improved acoustic models, Google could recognize existing voicemails offline to get newer, better transcriptions the language models could be retrained on, and with better language models Google could recognize again the same data, and repeat the process. Step by step, the recognition error rate dropped, finally settling at roughly half what it was with the original system.

Sometimes the recognizer would skip entire audio segments; it felt as if it was falling asleep and waking up a few seconds later. It turned out that the acoustic model would occasionally get into a "bad state" where it would think the user was not speaking anymore and what it heard was just noise, so it stopped outputting words. When we retrained on that same data, we'd think all those spoken sounds should indeed be ignored, reinforcing that the model should do it even more. It took careful tuning to get the recognizer out of that state of mind.

It was also tough to get punctuation right. The old system relied on hand-crafted rules or "grammars," which, by design, can't easily take textual context into account. For example, in an early test our algorithms transcribed the audio "I got the message you left me" as "I got the message. You left me." To try and tackle this, we again tapped into neural networks, teaching an LSTM to insert punctuation at the right spots. It's still not perfect, but we're continually working on ways to improve our accuracy.

In speech recognition as in many other complex services, neural networks are rapidly replacing previous technologies.

CHAPTER 6

\

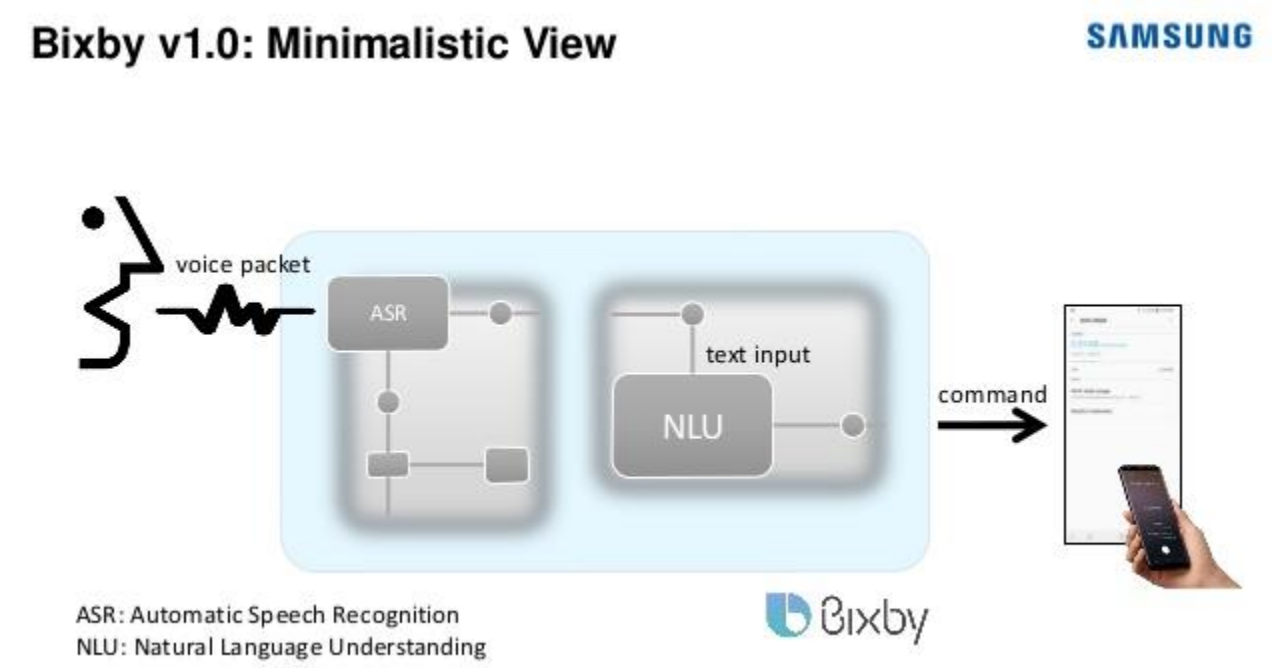
5]SAMSUNG BIXBY

Bixby is a virtual assistant developed by Samsung Electronics. It was developed in 2017.

Uses of Bixby:
Bixby can do following task:

- Weather news
- Book an appointment
- Bixby reminders
- Restaurant recommendation
- Internet search
- Calling
- Notification control
- Open the app

Work Flow



Deep Learning in Samsung Bixby

Bixby and other voice assistants use Automatic Speech Recognition (ASR) to transcribe spoken language to text. While this technology has come a long way since its inception, it still isn't perfect, and it's important to be able to train and retrain ASR models many times to achieve the best possible accuracy. Samsung, a close partner of Google Cloud, used CLOUD TPUS, Google Cloud's purpose-built machine learning processors, to train their ASR models faster and ultimately improve Bixby's accuracy.

For several years, the Deep Neural Network-Hidden Markov Model (DNN-HMM) hybrid ASR system architecture served as the standard approach for many speech-to-text services, including the previous generation of Bixby. DNN-HMM systems incorporated acoustic, pronunciation, and language models into the speech recognition pipeline. However, the training process was complex, which made it difficult to optimize overall accuracy.

To overcome the limitations of the DNN-HMM hybrid ASR system, the Bixby team decided to revamp their engine with a cutting-edge end-to-end deep learning

approach. By leveraging a single deep neural network model based on the Transformer architecture, the new engine would not only have a simplified training process, but it could also have access to a vast pool of training data.
However, this change in the system also meant new challenges. New experiments and tuning were needed to reach an accuracy comparable to the previous system. Additionally, fast training iterations were critical to keeping the live service up to date while continuously expanding into new languages. In order to meet those requirements while absorbing the ever-increasing amount of training data, the Bixby team decided to explore CLOUD TPUS.
Back in 2013, Google realized that our existing CPU and GPU infrastructure could not keep up with our growing computational needs for AI, so Google decided to build a new chip specifically for the purpose. The result was the Tensor Processing Unit (TPU), which has been deployed in Google data centers since 2015. Since then, Google have developed multiple new generations of TPU chips as well as large-scale supercomputing systems called TPU Pods that connect thousands of TPU chips together to form the worlds fasted training super computer.
The Bixby team's migration from GPUs to TPUs seemed to be happening smoothly, but when the time came to test the TPU-trained model back on GPUs, the Bixby team encountered a strange issue. While the model had been fine on TPUs, it started repeating itself once it ran inference on GPUs. For instance, "Hi Bixby, how's the weather today?" was being transcribed as "Hi Bixby, how's the weather today? Hi Bixby, how's the weather today?"Technical issues like this in machine learning projects are often very difficult to troubleshoot, as it's not always clear whether the problem is in the code, the platform, the infrastructure, or the data. The Bixby team opened a support case with Google Cloud. And after a combined effort with Customer Engineering, TPU Product team, and Google Brain, they found the root cause.While sequential models like Bixby's ASR engine deal with input data of variable length, the XLA compilation for TPU currently has a requirement that all tensor sizes must be known at graph construction time. Due to this difference, the model trained on TPUs that expected a long padding after each input audio sequence, could not reliably predict the end of a sentence when the padding was shortened or removed on GPUs. Hence, being unable to complete the sentence, it was repeating the same phrases over and over again.
Once the root cause was identified, it was quickly addressed. The Bixby team decided to simulate the TPU environment during GPU inference by introducing additional padding at the end of each input audio. Without making any modifications to the training code, this enabled their new model to predict the end of a sentence correctly on GPUs—and the repetition issue was successfully resolved.

CHAPTER 7

CONCLUSION

Neural network is best method which having good accuracy as compare to other method such as simple pattern matching, pattern and feature analysis and Langauge modeling. Companies like Amazon, Apple, Samsung, Google and Microsoft use this Neural network in their acoustic model. These companies voice assistance used the deep learning algorithm such as Convolutional Neural Network(CNN), RNN (Recurrent Neural Network) and Long Short Term Memory(LSTM) algorithm. Training the data these companies mainly use TPU(Tensor processing Unit) as hardware because these companies train millions of data to build their model for speech recognition. Natural language processing also plays important role in this speech recognition.

These deep learning based voice assistance helps human a lot in their daily life by saving their times and by boosting their productivity in work.

Although there are multiple advantages of this voice assistance but there is some problem regarding privacy protection, Noise in room, Slow internet and amplitude of voice.

CHAPTER 8

REFERENCES

1. https://www.cnet.com/home/smart-home/appliance-science-alexa-how-does-alexa-work-the-science-of-amazons-echo/
2. https://developers.google.com/assistant/conversational/overview
3. https://cloud.google.com/blog/products/ai-machine-learning/samsung-bixby-training-gets-speed-boost-with-cloud-tpu
4. https://www.analyticsvidhya.com/blog/2021/01/introduction-to-automatic-speech-recognition-and-natural-language-processing/
5. https://ai.googleblog.com/2015/08/the-neural-networks-behind-google-voice.html
6. https://www.explainthatstuff.com/voicerecognition.html
7. https://www.amazon.science/blog/two-new-papers-discuss-how-alexa-recognizes-sounds
8. https://www.maketecheasier.com/things-to-do-with-cortana/
9. Research paper -Google Search by Voice by Johan Schalkwyk
10. Research paper- DOMAIN AND SPEAKER ADAPTATION FOR CORTANA SPEECH RECOGNITION by Yong Zhao