MKSSS's CUMMINS COLLEGE OF ENGINEERING, PUNE

ARTIFICIAL INTELLIGENCE IN ALEXA BY AMAZON

PRESENTED BY:

ESHA CHAUGULE-3610/C22018441668
GAYATRI DESHMANE-3618/C22019442603

GUIDED BY:

PROF. RADHIKA BHAGWAT

TABLE OF CONTENTS

ABSTRACT

Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving. The ideal characteristic of artificial intelligence is its ability to rationalize and take actions that have the best chance of achieving a specific goal.

Conversational AI signals a huge advancement in the way we interact with computers. Unlike menus, touchscreens, or mouse clicks, using our voices to have conversations is one of the most natural ways to use a computer; it requires no learning curve.

This new method of human-computer interaction makes powerful computer applications even easier to use and accessible to more people. For example, instead of having to make several swipes and clicks to play music, you can simply say, “Alexa, play the top songs in Seattle.”

Conversational AI systems are computers that people can interact with simply by having a conversation. With conversational AI, voice-enabled devices like Amazon Echo are enabling the sort of magical interactions we’ve dreamed of for decades. Through a voice user interface (VUI), voice services like Alexa can communicate with people in ways that feel effortless, solve problems, and get smarter over time.

INTRODUCTION

WHAT IS ALEXA?

Alexa is a smart speaker. It is a type of speaker and voice command device which has an integrated virtual assistant that offers interactive actions and hands-free activation with the help of one or several “wake words”. It is controlled by our voice using technologies such as ASR(automatic speech recognition), STT(speech to text) and NLU(Natural Language Understanding). Alexa is invading our homes, by making our homes smart, and assisting in even turning off the lights, or changing the TV channel.[1]

The wake word was chosen as ‘Alexa’ because of its sound. The letter ‘X’ is a hard consonant and creates a unique pitch, which is easier to detect even if there is a lot of noise going around the device. It continually learns and adapts to our speech patterns. Its dataset and accuracy increases the more we use it, and it adapts to our speech pattern, vocabulary and personal preferences. It has been designed to work in 8 languages, including Hindi. Its revenue was 24 million dollars when it was just started in 2014, and now it has increased to 12,000 million dollars in 2020. Alexa has made our life much easier and happier.[2]

But these are just the common things that everyone knows about Alexa. Now let's see what happens after the user makes a request.

WHAT HAPPENS AFTER THE REQUEST IS SENT?

- Amazon records the user's words. Interpreting sounds takes up a lot of computational power, the recording of the user's speech is sent to Amazon's servers to be analyzed more efficiently.
- Amazon's "Alexa Voice Service(AVS)" identifies skill and recognizes intent through ASR and NLU.
- Then it sends the customer intent to "Your Service" which is the Amazon Server, which processes the request and decides on what the response should be.
- The server responds to intent through text(Wikipedia information) or renders graphical components(image or voice response).

SIGNAL PROCESSING

Signal processing gives Alexa as many chances as possible to make sense of the audio by cleaning the signal. Signal processing is one of the most important challenges in far-field audio. The idea is to improve the target signal, which means being able to identify ambient noise like the TV and minimize them. To resolve these issues, **seven microphones** are used to identify roughly where the signal is coming from so the device can focus on it. **Acoustic Echo Cancellation** can subtract that signal so only the remaining important signal remains. Acoustic Echo Cancellation (AEC) is used to cancel acoustic feedback between a speaker and a microphone in loud-speaking audio systems, teleconferencing devices, hands free mobile devices, and voice-controlled systems.

The next task is "Wake Word Detection". It determines whether the user says one of the words the device is programmed to need to turn on, such as "Alexa". This is needed to minimize false positives and false negatives, which could lead to accidental purchases and angry customers. This is really complicated as it needs to **identify pronunciation differences**, and it needs to do so on the device, which has limited CPU power.

If the wake word is detected, the signal is then sent to the speech recognition software in the cloud, which takes the audio and converts it to text format. The output space here is huge as it looks at all the words in the **English language,** and the **cloud is the only technology capable of scaling sufficiently**. This is further complicated by the number of people who use the Echo for music — many artists use different spellings for their names than there are words.

To convert the audio into text, Alexa will analyze characteristics of the user's speech such as frequency and pitch to give you feature values. A decoder will determine what the most likely sequence of words is, given the input features and the model, which is split into two pieces. The first of these pieces is the prior, which gives the most likely sequence based on a huge amount of existing text, without looking at the features. The other is the acoustic model, which is trained with **deep learning** by looking at pairings of audio and transcripts. These are combined and dynamic coding is applied, which has to happen in real time.[5]

ALEXA SKILLS

In Alexa Skills, the word 'skills' does not have the literal English meaning of it. Alexa Skills are like smartphone apps. They are created by developers from third-party companies. Some skills work entirely on their own, while others are required to make Alexa interact with something else, such as an online service like Spotify, or a smart home product like a Roomba robotic vacuum cleaner.

- **Wake word**
  When users say 'Alexa' which wakes up the device. The wake word puts Alexa into the listening mode and ready to take instructions from users. The letter 'X' is a hard consonant and creates a unique pitch, which is easier to detect even if there is a lot of noise going around the device.
- **Invocation name**
  Invocation name is the keyword used to trigger a specific "skill". Users can combine the invocation name with an action, command or question. All the custom skills must have an invocation name to start it.

- **Launch Word** Action, command or question. E.g., ask, add, play.
- **Utterance**
  ‘Taurus’ is an utterance. Utterances are phrases the users will use when making a request to Alexa. Alexa identifies the user’s intent from the given utterance and responds accordingly. So basically, the Utterance decides what user wants Alexa to perform.[8]

WHAT HAPPENS INSIDE ‘ALEXA VOICE SERVICE’(TTS)?

Text-to-speech (TTS) is a type of assistive technology that reads digital **text** aloud. It's sometimes called “read aloud” technology. With a click of a button or the touch of a finger, TTS can take words on a computer or other digital device and convert them into audio.

The basic challenge of speech synthesis from text is to produce **natural and pleasant sound with correct pronunciation**. So the input to the speech-synthesis engine in such a case would be a **string of phonemes along with insertions of necessary accents and pauses**. Some transformations would be applied to this to obtain the acoustical transcript. These include models to generate the fundamental frequency and duration of each speech segment. The last step is synthesis of the speech waveform using the parameters generated in the earlier stage. Three types of speech synthesizers are used: articulatory, format and concatenative synthesizers.[11]

After a text response is sent by Amazon's server to AVS. It converts the text into its **normalized form.** Here, the dates, numbers, acronyms or abbreviations are converted into "normal word form". For example, 200$ is written as "two hundred dollars".

Phoneme is the smallest unit of the sound and Grapheme is a way that we write the sound. So now these **graphemes are converted into phonemes** after which these are converted into **waveforms** which take the pauses into consideration as well. To do this, Concatenative Synthesis is used. Concatenative synthesis is based on the concatenation (or stringing together) of segments of recorded speech. Generally, concatenative synthesis produces the most natural-sounding synthesized speech. Here, audio is sliced into tiny units and the machine tries to find the optimal sequence of pieces to maximize the naturalness of the audio given the sequence of words.

And finally, it gets converted into the **speech** accordingly.

TECHNOLOGIES USED BY ALEXA

❖ ASR(Automatic Speech Recognition)

Automatic speech recognition (ASR) is technology that converts spoken words into text. In short, it's the first step in enabling voice technologies like Amazon Alexa to respond when we ask, "**Alexa**, what's it like outside?" With **ASR**, voice technology can detect spoken sounds and recognize them as words.

The Automatic Speech Recognition (ASR) Evaluation tool allows you to batch test audio files to measure the ASR accuracy of the skills that you've developed. The main benefit of the tool is to use the feedback from your evaluation reports to improve the ASR performance for your skill. Use the report results to modify your sample utterances to improve your skill's interaction model and accuracy.

❖ TTS(Text To Speech)

Text To Speech allows your Alexa enabled device to say anything you want. Whatever you type, Alexa will speak. This is an invaluable tool for Alexa skill developers and enthusiasts alike.

Text-to-speech (TTS) is a type of assistive technology that reads digital **text** aloud. It's sometimes called "read aloud" technology. With a click of a button or the touch of a finger, TTS can take words on a computer or other digital device and convert them into audio.

❖ NLU(Natural Language Understanding)

With natural language understanding (NLU), computers can deduce what a speaker actually means, and not just the words they say. In short, it is what enables voice technology like Alexa to infer that you’re probably asking for a local weather forecast when you ask, “Alexa, what’s it like outside?”

Today’s voice-first technologies are built with NLU, which is artificial intelligence centered on recognizing patterns and meaning within human language. When a computer understands what you mean to say without you having to ask it in one specific way, using your voice starts to feel like having an actual conversation.

AGENT

An agent is anything that can perceive its environment through sensors and acts upon that environment through effectors. Artificial Intelligence is defined as a study of rational agents. A rational agent could be anything which makes decisions, as a person, firm, machine, or software. An **AI** system is composed of an agent and its environment. The agents act in their environment. The environment may contain other agents.

Agent work upon PEAS model:

P: Performance

E: Environment

A: Actions

S: Sensors

Alexa is an intelligent agent as AI assistants, like Alexa and Siri, are examples of intelligent agents as they use sensors to perceive a request made by the user and the automatically collect data from the internet without the user's help. They can be used to gather information about its perceived environment such as weather and time

PEAS MODEL

- ❑ Performance Measure: Performance measure is the unit to define the success of an agent. Performance varies with agents based on their different percepts. If the objective function to judge the performance of the agent. For example, in case of pick and place robot, no of correct parts in a bin can be the performance measure.
- ❑ Environment: Environment is the surrounding of an agent at every instance. It keeps changing with time. It's the real environment where the agent needs to deliberate actions.
- ❑ Actuators: These are the tools, equipment or organs using which agent performs actions in the environment. This works as output of the agent. The devices, hardware or software through which the agent performs any actions or processes any information to produce a result are the actuators of the agent.
- ❑ Sensors: These are tools, organs using which agent captures the state of the environment. This works as input to the agent. The devices through which the agent observes and perceives its environment are the sensors of the agent.

ENVIRONMENT

- **Fully Observable and partially observable**

A fully observable AI environment has access to all required information to complete the target task. Image recognition operates in fully observable domains. Partially observable environments such as the ones encountered in self-driving vehicle scenarios deal with partial information in order to solve AI problems.

- **Static and dynamic**

Static AI environments rely on data-knowledge sources that don’t change frequently over time. Speech analysis is a problem that operates on static AI environments. Contrasting with that model, dynamic AI environments such as the vision AI systems in drones deal with data sources that change quite frequently.

- **Discrete and continuous**

Discrete AI environments are those on which a finite [although arbitrarily large] set of possibilities can drive the final outcome of the task. Chess is also classified as a discrete AI problem. Continuous AI environments rely on unknown and rapidly changing data sources. Vision systems in drones or self-driving cars operate on continuous AI environments.

- **Deterministic and stochastic**

Deterministic AI environments are those on which the outcome can be determined based on a specific state. In other words, deterministic environments ignore uncertainty. Most real world AI environments are not deterministic. Instead, they can be classified as stochastic. Self-driving vehicles are a classic example of stochastic AI processes.

- **Episodic**

Subsequent episodes do not depend upon what actions occurred in previous episodes. Some tasks can be divided into different phases or episodes. Alexa works in an episodic environment.

ALEXA IS AVAILABLE IN 8 LANGUAGES

- English
- Hindi
- Portuguese
- Italian
- German
- Japanese
- French
- Spanish

APPLICATIONS

- Voice interaction
- Music playback
- Making to-do lists
- Setting alarms
- Real time information
- News
- Providing weather
- Traffic
- Sports

LIMITATIONS

- It's a cloud-based device so it can create problems if there is any trouble in the cloud.
- It does not have internal memory, the information is stored on the server.
- It facilitates some activities at home but also makes the user lazy.
- Alexa can record private conversations.
- Alexa cannot understand multiple accents.
- Perform multiple actions with one command.[10]

WHY SOME CONSUMERS HAVE NOT PURCHASED A SMART SPEAKER

- Privacy: Privacy is a concern, especially involving smart speakers. While waiting for a wake word, smart speakers are always listening. Smart speakers and other AI assistants, like those on a smartphone, save these recordings and allow the user to go into their account and delete them.
- Accuracy: Voice assistants don’t always understand what we are asking. Sometimes, it’s how we speak. Other times, it is simply because artificial intelligence hasn’t yet learned how to do something.
- Hackability and Security: Even though voice assistants communicate with their servers using encrypted connections, there is still a concern about hackability and security.
- Plan to buy
- Not interested
- Too expensive [10]

SMART SPEAKER MARKET SHARE

Google had experience with searches, languages, and data, so they had the ability to quickly jump into the smart speaker market, Mutchler explains. Apple’s HomePod is lagging behind, mostly because of features, the same factor that hinders Siri on iPhones. “She’s not open to developers. She’s very restrained,” Mutchler says.

Google Assistant spoke many languages from the very beginning, something Alexa had to learn how to do. Google Home was the first to go on sale outside the U.S., and Amazon Echo was the first to offer commerce. Both companies have always understood multiple voices and can differentiate between them. Siri has not.

When one company unveils a new feature, the others are usually not too far behind. In the beginning, making calls to a phone was not something smart speakers could do. Now, they can. Today, all options can play music, add items to calendars, perform searches, send messages, answer questions, control some smart devices, and much more. “It’s a constant feature game. Creating a voice assistant is more difficult than people think. They [competitors] make each other better,” Mutchler emphasizes.

Smart speaker companies do not make money on the sales of the hardware; the money is in how we will use smart speakers in the future. The key is getting people on board with a brand early, much like with smartphones. “Once you start with one assistant, you might not go back and try another one. It’s difficult to get people to switch,” Mutchler suggests.[10]

References:

1. https://en.wikipedia.org/wiki/Smart_speaker
2. https://seekingalpha.com/instablog/3009741-msu-eli-broad-student-research/4976301-amazon-s-echo-shows-promise-for-future-growth
3. https://bernardmarr.com/default.asp?contentID=1830
4. https://www.wired.com/story/amazon-alexa-2018-machine-learning/
5. https://channels.theinnovationenterprise.com/articles/how-amazon-alexa-works
6. https://www.amazon.science/blog/the-scalable-neural-architecture-behind-alexas-ability-to-select-skills
7. https://www.techrepublic.com/article/video-top-5-things-to-know-about-amazon-voice-services/
8. https://towardsdatascience.com/how-amazon-alexa-works-your-guide-to-natural-language-processing-ai-7506004709d3
9. https://developer.amazon.com/en-US/docs/alexa/ask-overviews/build-skills-with-the-alexa-skills-kit.html
10. https://www.smartsheet.com/voice-assistants-artificial-intelligence
11. https://www.pcquest.com/how-text-speech-works/