SIT 788

Assignment 9.1: Advanced intelligent system

Problem Introduction

This task discusses an advanced intelligent system – Amazon Alexa and provides an overview of the different AI modules and services used. Amazon's Alexa[1] is a cloud based, AI powered voice-based service which offers natural voice experiences to customers and can act as an Intelligent Personal Assistant (IPA). Alexa can be used with devices which have built-in device integrations as well as it can be used to control IOT enabled devices. This gives users the capabilities to do routine tasks while just using their natural voice query, which not only provides a better customer experience when using such device but also democratises cutting edge technologies for business use cases.

The following sections discuss:

- Functionalities offered by Alexa
- Use cases for Alexa
- Alexa design overview

Alexa Functionalities

Apart from out of the box experience from Amazon devices such as Echo speaker, Alexa can be integrated with 3rd Party devices – both these devices constitute **Alexa Built-In devices** [2]. Users can interact with these devices through a natural voice query, this query is processed, and Alexa infers the action(s) to perform based on the user query. To activate these devices a **wake-up word** (the user calls out Alexa!) is used which precedes any query. A second set of devices – **Alexa Connected devices** [3] are the IOT enabled devices which can be controlled through an Alexa Built-In device. The user query is processed via Alexa Built-in device and delegated to the service (referred as *skills* in Alexa documentation) responsible for controlling Alexa Connected device. This flow is displayed in Fig 1.

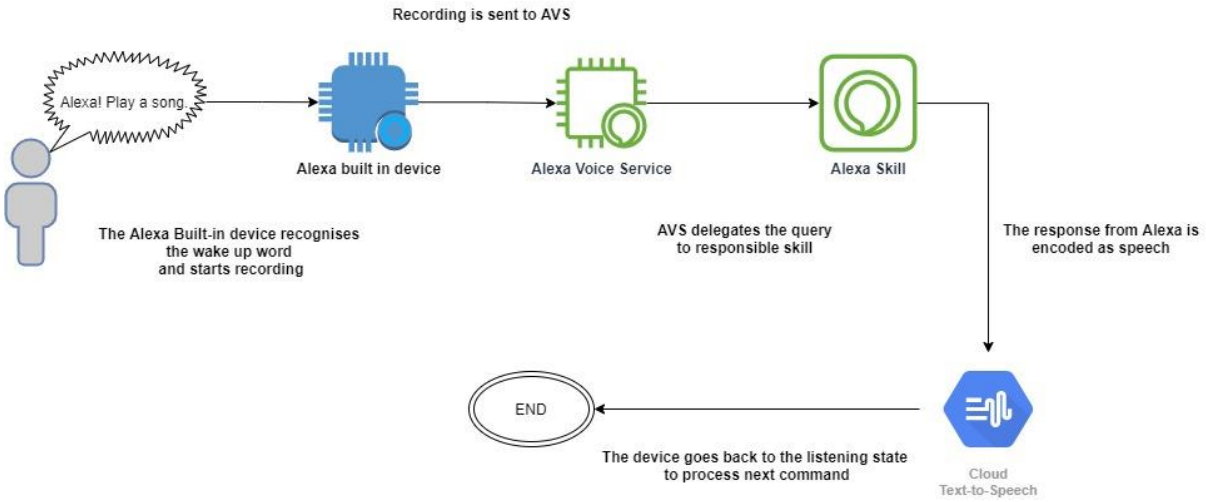


Fig 1: The life cycle of a sample Alexa query.

Alexa use cases

This section explores the use cases based on two device types discussed in previous section.

Alexa Built-In device: This scenario involves using an Alexa Built-in device which consists of a speaker and a mic which is used for talking to the customers. The Alexa Built-In devices are powered by **Alexa Voice Service (AVS).** AVS handles the complete workload for providing an IPA experience. This is provided as SDK and APIs which are free to use when building a built-in device. Some example scenarios where such devices are used are:

- Smart Speakers
- Smart Headphones
- Personal Computers/Phones
- Smart TVs and Screens

Alexa Connected device: This scenario involves using an Alexa connected device which is controlled via an Alexa Built-in device. The voice processing part is same as the built-in device where AVS is used to process natural voice queries. The additional step in these scenarios is mapping to a predefined skill or a custom skill which is a reference to a routine to perform a given task e.g., "*Turn on the bedroom light"* will invoke a routine responsible for turning on the entity named as bedroom light. Skills can also be mapped to user customized actions/routines. Some example scenarios are:

- Lights, Switches, and bulbs
- Cameras
- Door locks
- Other Smart home devices

Alexa design overview

As both the Alexa Built-In device and Connected device use the same AI processing service – **AVS**, this section explores the system design of this service. The underlying assumption while exploring the system design is that the user query is in English, whereas in real-life scenarios multiple languages are supported which can be handled by a sperate speech translation service (not discussed here). Another assumption in case of Alexa connected devices is that the skills/routines required to operate the IOT device is available, i.e., the design of all such routines are not discussed here as they can be varied according to the use case.

The devices are connected to the internet all the time, where some modules are located on the device itself and some modules are on the cloud. The core modules required for designing AVS are discussed below.

Keyword identifier: This module is responsible for recording the user query when it hears the wake-up word (ALEXA!) and sends the recorded (encoded) audio to the cloud for further processing. This module is located on the device itself and hence is a lightweight module.

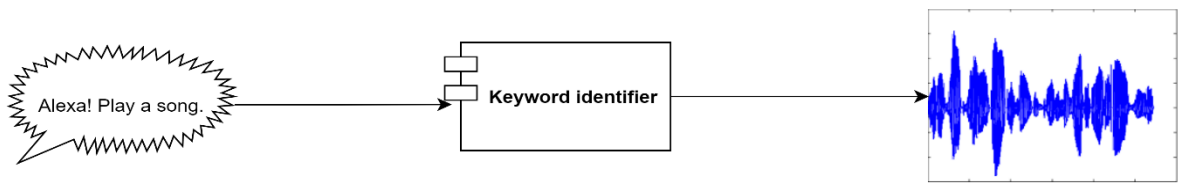


Fig 2: Keyword identifier block

Automatic Speech Recognizer: The ASR is located on the cloud and is responsible for converting a raw encoded audio to a natural language output text. The ASR module is a deep learning architecture which consists of a 1-d CNN and fully connected layer with SoftMax to output natural language representation of the spoken command.

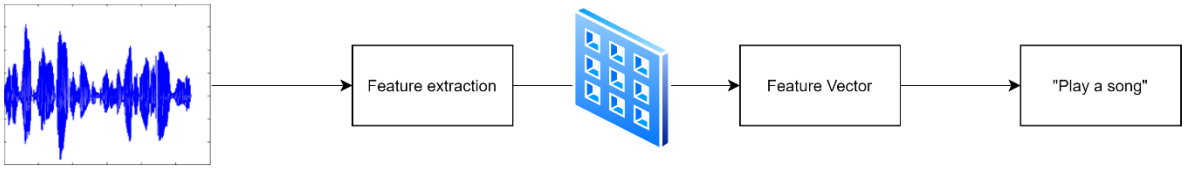


Fig 3: Encoded signal transformation to output natural language text.

Natural Language Understanding Service: This service is also located on the cloud and is responsible for structuring the natural language input into the format expected by a conversation manager service. The service maps the words to their intents and then to their respective domains. E.g., *Play a song* will be mapped to *domain* music which will be further mapped to *intent* of playing a song of a particular artist (in this case, a random song from the liked tracks). For Alexa, *play a song* plays a random song from the liked tracks of the user.

Conversation Manager Service: This service is also hosted on the cloud; this service is responsible for maintain the logical flow of the conversation between the user and the device (IPA). This service is also responsible for parsing the output from NLU service and delegating it to the responsible skills/backend services to perform the requested action (intent). This service interacts with the user through a set of questions until a tangible *intent* is found. A tangible intent is found when an action can be performed Alexa device can return to the starting (listening) state. Thus, this service has two outputs, either a *clarification* or a tangible *intent.*

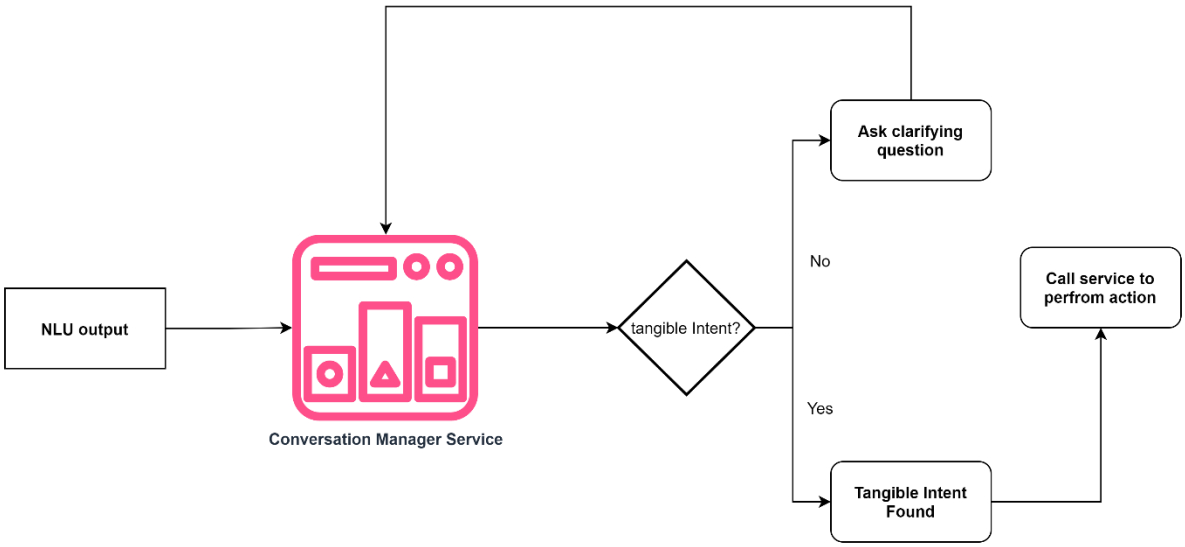
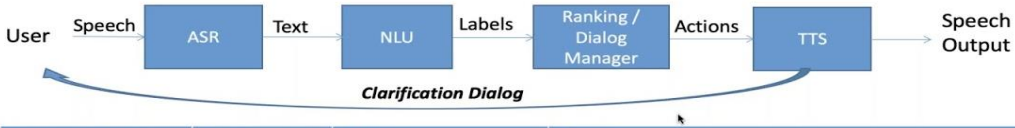


Fig 4: Shows the logical flow maintained in Alexa – user interactions.

Text to Speech Service: The text to speech service includes a **spectrogram predictor** and a **neural vocoder.** The spectrogram predictor produces a spectrogram of the text from conversation manager service. The neural vocoder takes in the output from spectrogram predictor and converts it into an audio message which is played back to the user. This service gives the speech ability to the IPA and is used to convey messages from the service to the user.

Finally, these systems come together to form an end-to-end system known as *Alexa* which uses AVS to achieve IPA capabilities. The complete interaction can be captured in fig 5 where a sample query "Play Two Steps Behind by Def Leppard" is processed.



Component	Input	Output	Example
Automatic Speech Recognition (ASR)	Speech	Text (1-best or lattice of alternatives)	"Play Two Steps Behind by Def Leppard"
Natural Language Understanding (NLU)	Text	Slots and Intent Type	**Intent**: PlayMusicIntent **Slots**: Artist Name=Def Leppard Song Title: Two Steps Behind
Ranking / Dialog Manager (DM)	Labels & Context	Dialog Actions	Ask the application to play the song or clarify
Text-to-Speech (TTS)	Text	Speech	*"Which artist?"* or *"Playing Two Steps Behind by Def Leppard"*

3

Fig 5: The overall flow of the user – Alexa conversation [4]

References

1. https://developer.amazon.com/en-US/alexa
2. https://developer.amazon.com/en-US/alexa/devices/alexa-built-in
3. https://developer.amazon.com/en-US/alexa/devices/connected-devices
4. https://www.youtube.com/watch?v=U1yT_4xcglY&ab_channel=AmazonWebServices