Heba Khashogji

Assignment 8: De-black boxing of virtual assistant (Alexa)

De-black boxing of virtual assistant (Alexa)

Alexa is a well-known virtual assistant developed by amazon using AI in 2014 (Wikipedia, 2021). Alexa can play music, interact with our voices, make to-do lists, setting alarms, provide weather information, etc. (wikipedia, 2021). We can use Alexa as a home automation system controlling our smart devices (Wikipedia, 2021) (Amazon, 2021). Besides that, we can install extension functionality called skills, adding them to Alexa. Device manufactures can integrate Alexa voice capabilities in their products using the Alex voice service. In this way, any products built with this cloud-based service have access to a list of automatic speech recognition and natural language processing capabilities. Amazon uses the long-short term memory LSTM for generating voices (Amazon, 2021). In 2016, Amazon released Lex, making the speech recognition and natural processing language NLP available for developers to create their chat-bots (Barr, 2016). Less than a year later, Lex became generally available (Barr, AmazonLex–NowGenerallyAvailable, 2017). Now, web and mobile chat is available using Amazon connect (Hunt, 2019).

Any virtual assistant's main components include a light ring, volume ring to control voice level, microphone array used to detect, record and listen to our voices, power port to charge the device and audio output. Virtual assistance, after that, recognize voice and store conversation in the cloud.

De-black boxing of virtual assistant (United States Patent No. US2012/0016678 A1, 2012)

Level 0:

Here, the virtual assistant is just a black box whose input is a voice commands from the user while the output is the voice response. Fig.1 includes the black box of the virtual assistant (Alexa, for example).

Fig1. Black box of Virtual Assistant

Level 1:

For level-1 de-black boxing, we can see the following components:

1- ASR (Automatic Speech Recognition): returns Speech as Text.
2- NLU (Natural Language Understanding): Interpret text as a list of possible intents (Commands).
3- Dialog manager: Look at intent and determine if it can handle it. The specified rules define which speechlet to be processed.

4- Data store: Includes the voice in a text response.
5- Text to speech: Translates skill outputs into an audible voice.
6- The third-party skill: The third party writes and is responsible for skill actions and operations. Fig.2 shows the level-1 de-black boxing of a virtual assistant (Alexa).
7-

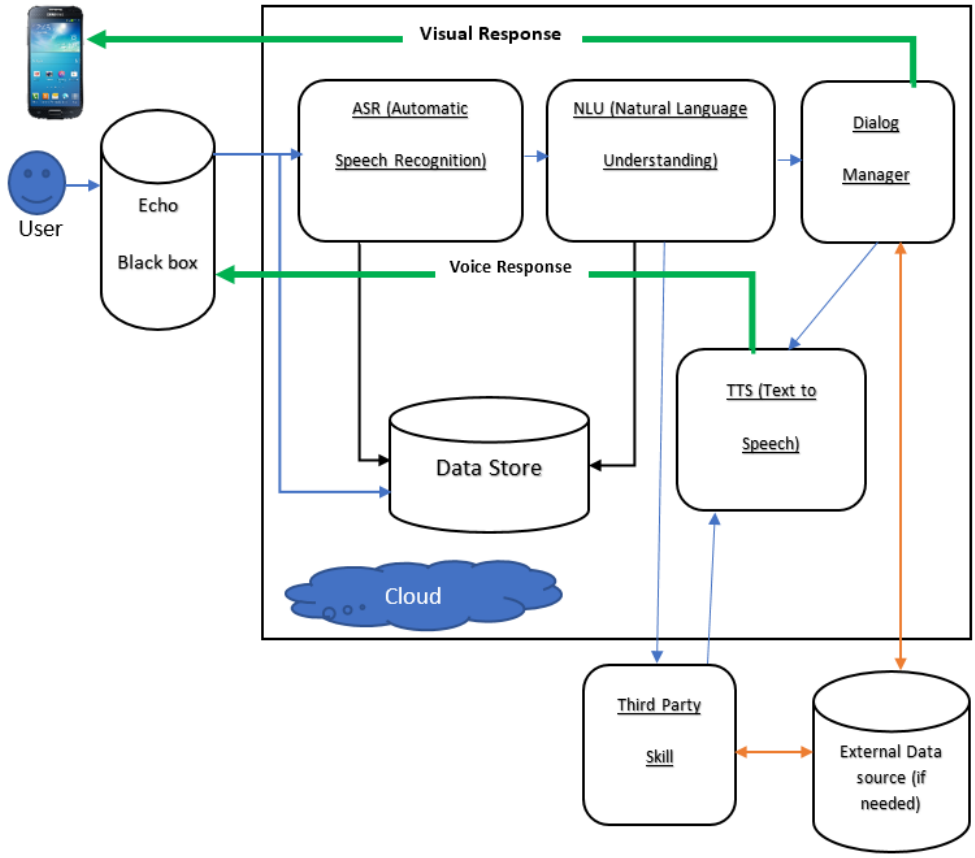

Fig2. Level-1 of De-black boxing of the Alex System

Level 2:
De-black box the ASR

The acoustic front-end takes care of converting the speech signal into corresponding features (speech parameters) via a process called feature extraction. The parameters of word/phone models are estimated from the acoustic vectors of training data. The decoder functions though the search of all possible word sequences to find the sequence of words that is most likely to generate. In a training phase, the operator will read all the vocabulary words and the word patterns are stored. Later, when for the recognition step, the word pattern is compared to the stored patterns and the word that gives the best match is selected. Fig3 illustrates the de-black box of ASR.

Heba Khashogji

Assignment 8: De-black boxing of virtual assistant (Alexa)

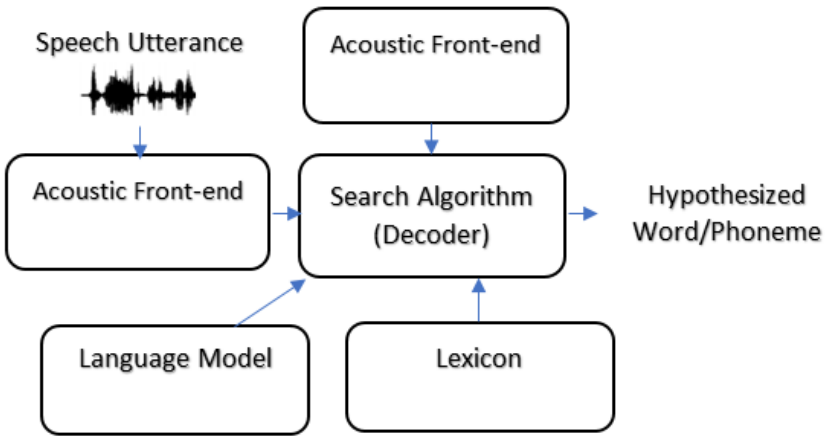

Fig3. Level2 of De-black box the ASR

De-black box of NLU (Natural Language Understanding)

Intent Classification (IC) and **Named Entity Recognition** (NER) use machine learning to recognize natural language variation. So, to identify and categorize key information (entities) in text, we need the **NER** of NLU. NER is a form of NLP, including two steps: detecting the named entity and the categorizing step. In step1, **NER** detects a word or thread of words that form a whole entity. Each word signifies a token: “The Great Lakes” is a thread of three tokens representing one entity. The second step requires the creation of entity categories like a person, organization, location, etc. **IC** labels the utterances of an NLP from a predetermined set of intents. **Domain Classification** is a text classification model that determines the target domain for a given query. It is trained using many labelled queries across all domains in an application. **Entity Resolution** is the last part of NLU that disambiguate records that correspond to real-world entities across and within datasets. So, to play "Creedence Clearwater Revival", the NER will be "CCR (ArtistName)", the Domain classifier is "music", the IC is "PlayMusicIntent", and the entity resolution will be " Creedence Clearwater Revival". Fig.4 includes the de-Blackbox of the NLU.

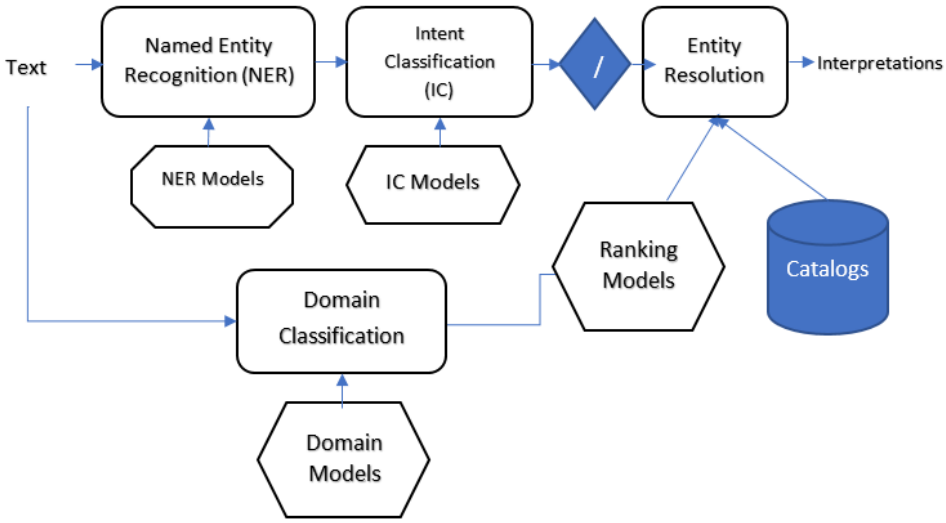

Fig4. Level2 of De-black box the NLU

Dialog Manager (DM)

DM selects what to report or say back to the user, whether to take any measure and decide to handle any conversation. DM includes two parts: dialog state tracking that estimates the user's goals tracking the dialog context as input, and dialog policy which generates the next system action. Dialog state tracking can be done using RNN and neural belief tracker (NBT), while the dialog policy can be done using reinforcement learning (RL). Fig.5 shows Level2 of De-black box the DM.

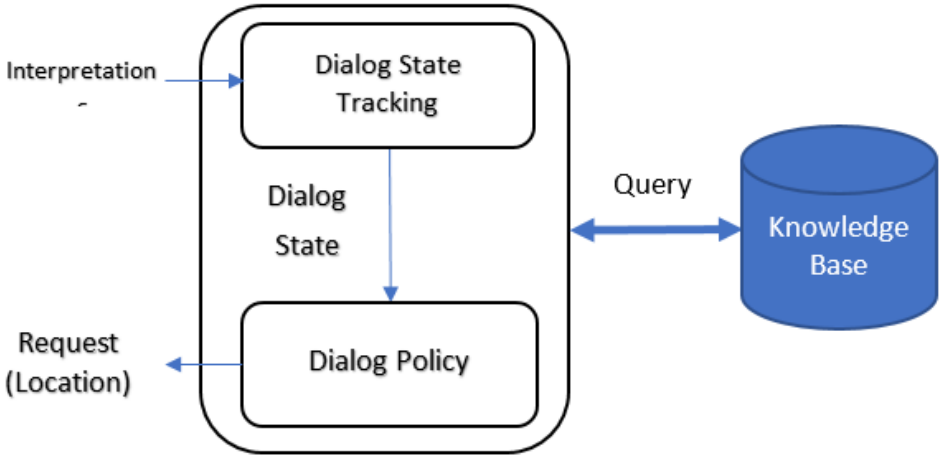

Fig5. Level2 of De-black box the DM

De-black box of Text-To-speech TTS

The last part of the Virtual assistant allows computers to read text aloud. **The linguistic front-end** is used to convert input text to a sequence of features such as phonemes and sentence type. **The prosody model** predicts pattern and melody to form the expressive qualities of natural speech. **The acoustic model** is used to transform linguistic and prosodic information into the frame-rate spectral feature. Those features are fed into the **neural vocoder** and used to train a lighter and smaller vocoder. Neural Vocoder generates 24 kHz speech waveform. It consists of a convolutional neural network expanding the input feature vectors from frame rate into sample rate and a recurrent neural network synthesizing audio samples auto-regressively at 24,000 samples per second. Fig6 includes the details of TTS.

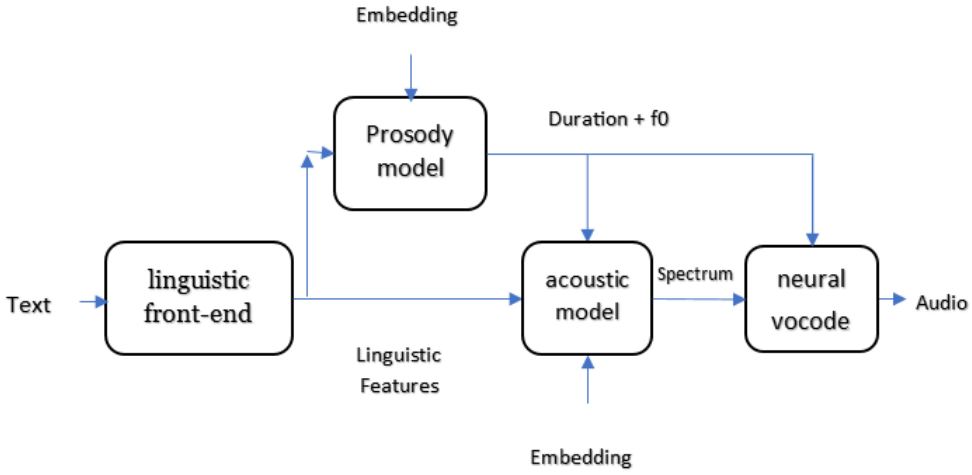

Fig6. Level2 of De-black box the TTS

Heba Khashogji

Assignment 8: De-black boxing of virtual assistant (Alexa)

Fig7 shows the De-black boxing of the Alex Echo system.

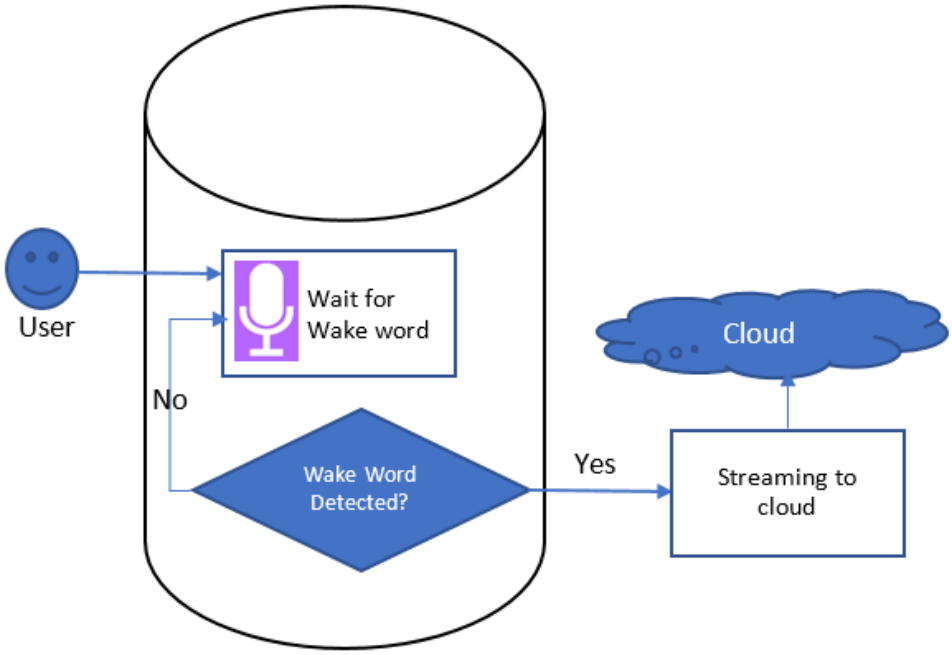

Fig7. Level2 of De-black boxing of the Alex System (De-black boxing of Echo system)

Heba Khashogji

Assignment 8: De-black boxing of virtual assistant (Alexa)

References:

Amazon .(2021 ,) .*Amazon Lex* . Retrieved from: https://aws.amazon.com/lex/

Gruber.et.al .(2012) .*United States. Patent no. US2012/0016678 A1 .*

Jeff Barr .(2016) .*amazon-lex-build-conversational-voice-text-interfaces* Retrieved from: AWSNewsBlog: https://aws.amazon.com/ar/blogs/aws/amazon-lex-build-conversational-voice-text-interfaces/

Jeff Barr .(2017) .*AmazonLex–NowGenerallyAvailable*. Retrieved from: AWSNewsBlog: https://aws.amazon.com/blogs/aws/amazon-lex-now-generally-available/

Randall Hunt .(2019) .*Amazon-Connect* .Retrieved from: AWS Contact Center: https://aws.amazon.com/ar/blogs/contact-center/reaching-more-customers-with-web-and-mobile-chat-on-amazon-connect/

Wikipedia .(2021) .*Amazon_Alexa* Retrieved from: Wikipedia: https://en.wikipedia.org/wiki/Amazon_Alexa

wikipedia .(2021) .*Virtual_assistant* . Retrieved from: wikipedia_Virtual_assistant: https://en.wikipedia.org/wiki/Virtual_assistant