# Connecting the dots: An exploratory study of the workings of intelligent voice assistants

Abhishek Shah, Miloni Sangani, Aryan Singh, Siddhart Rao Deb

### Abstract

Communication between people and machines started back in the early 1960s. However, the era of smart voice assistants began ever since Apple launched the Siri voice assistant for iOS in 2010. Soon after its arrival, many other tech giants joined this bandwagon with similar voice assistants such s IBM Watson, Microsoft's Cortana, Amazon's Alexa, and Google's assistant to name a few.

Artificial intelligence technologies are commencing to be actively used in human life, this is facilitated by the appearance and wide dissemination of the Internet of Things (IoT). Autonomous devices are becoming smarter in their way to interact with both a human and themselves.

New capacities lead to creation of various systems for integration of smart things into Social Networks of the Internet of Things. One of the relevant trends in artificial intelligence is the technology of recognizing the natural language of a human. New insights in this topic can lead to new means of natural human-machine interaction, in which the machine would learn how to understand human’s language, adjusting and interacting in it. One of such tools is voice assistant, which can be integrated into many other intelligent systems.

In this paper, the principles of the functioning of voice assistants are described, their main shortcomings and limitations are given.

### Introduction

Voice is the future of brand interaction and customer experience. Voice control is the next evolution of human-machine interaction, thanks to advances in Artificial Intelligence, Natural Language Processing, Machine Learning, and Cloud computing. In recent years, Artificial Intelligence (AI) has shown substantial progress, and its potential is growing. The most common application of Artificial Intelligence is smart speakers also commonly known as Intelligent voice assistants.

Voice assistants are easy to use and thus millions of devices incorporate them in households, schools, and universities nowadays. They incorporate Artificial Intelligence by using Natural Language Processing (NLP), cloud computing and can communicate with the users in natural language. Today the evolution of artificial intelligence (AI) systems that are able to organize a natural human-machine interaction (through voice, communication, gestures, facial expressions, etc.) are gaining in popularity. It is no longer a human who learns to communicate with a machine, but a machine learns to communicate with a human, exploring his actions, habits, behavior and trying to become his personalized assistant.

Cloud platforms are now facilitating voice assistants in millions of homes. Voice assistants rely on a cloud-based architecture, since data has to be sent back and forth to centralized data centers.

The basic idea is that the user makes a request through the voice-activated device, and then, the voice request gets streamed through the cloud, and here voice gets converted into text. Then, the text request goes to the backend and after processing, the backend replies with a text response. Finally, the text response goes through the cloud and gets transformed into voice, which will be streamed back to the user.

The popularity of these devices is constantly rising since 2015. According to Canalys (2019), installed base will approach 225 million by 2020 and 320 million by 2022. Amazon Echo and Google Home devices are considered to reside in over 60% of US households by 2024.

The capabilities of voice assistants are continuously extending. The work on creating and improving such personalized assistants has been going on for a long time. These systems are constantly improving and enhancing, go beyond personal computers, and have already firmly established themselves in various mobile devices and gadgets.

Some key elements that distinguish voice assistants from ordinary programs are:

* NLP: The ability to understand and process human languages. It is important to fill the gap in communication between humans and machines.
* The ability to use stored information and data and use it to draw new conclusions.
* Machine learning: the ability to adapt to new things by identifying patterns.

The remainder of the paper is organized as follows. Section I, II presents a brief introduction to the architecture and construction of voice assistants. Section III provides an example of the work of a voice assistant. Section IV describes the shortcomings of existing voice assistants and our opinions on how to solve them.

### Methodology

### **1. Wake up word:**

### It is used to activate to trigger voice-enabled smart assistants so that it understands that we are speaking to it right now.

### Accurate wake word detection is important to protect user privacy and reduce processing overhead on the server as then it will constantly listen to the user and keep sending voice data to the server. Also, It always stays in a listening mode, so it does not take any voice commands until a wake-up word wakes it up.

# Technologies used

Conventional wake word detection methods make use of 2 stage models.

1. Deep NN Acoustic model
2. Hidden Markov Model

It may also have additional classifiers like SVM to increase accuracy. The Hidden Markov model output provides the endpoints of the wake word.

# Acoustic Model + Hidden Markov Model

Deep neural network (DNN) acoustic models have driven tremendous improvements in large vocabulary continuous speech recognition in recent years. Deep neural networks (DNNs) are now a central component of nearly all state-of-the-art speech recognition systems while HMM lies at the heart of most speech recognition systems currently in use.

This 2-stage AM+HMM KWS is de-facto the golden standard for keyword end pointing in the industry.

In this algorithm, the posteriors are produced by an Acoustic Model DNN for a set of Senones, based on an input audio stream, and an HMM is tuned to force-align a sequence of Senones expected in the keyword to those detected.

This is the classical approach used in Automatic speech recognition. The keyword endpoints are naturally produced in that algorithm as the times of the first and the last Senones in the HMM state sequence correspond to a keyword detection.

2. Information transfer:

Speech after wake up word is sent to automatic speech recognition software in the cloud, which takes audio and converts it to text.

Simultaneously, noise reduction is done as well using digital signal processing.

# Automatic Speech Recognition

Automatic speech recognition (ASR) is the 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, “Hey Alexa!”

With ASR, voice technology can detect spoken sounds and identify them as words. ASR is the cornerstone of the entire voice experience, allowing computers to finally understand us through our most natural form of communication which is "speech".

3. Sequencing and Analysis of order:

A decoder will determine what the most likely sequence of words is.

It then sends the user’s instruction to a cloud-based service where it processes the response and identifies the user’s intent.

The cloud service platform acts as the brain in the working of intelligent voice assistants as it performs all the complex operations such as automatic speech recognition and natural language understanding. It then makes the web service request to a third-party server if needed.

Natural Language Understanding:

Once the words are converted to text, it now tries to understand the meaning of the sentence. NLU converts the words into meaningful representations.

Sentiment Analysis:

Human communication isn’t just words and their explicit meanings. Instead, it’s nuanced and complex. You can tell based on the way a friend asks you a question whether they’re bored, angry, or curious.

The modern deep learning approach for sentiment analysis can be used for morphology, syntax, and logical semantics, of which the most effective one is Recursive Neural Networks(RNN).

As the name implies, the main assumption for Recursive Neural Net development is that recursion is a natural way of describing language. Recursion is useful in disambiguation, helpful for some tasks to refer to specific phrases, and works extremely well for tasks that use a grammatical tree structure.

Dialogue and Conversations:

4. Answering Questions:

A powerful deep learning architecture, known as a dynamic memory network(DMN) is used specifically for questions and answer problems. Given a training set of input sequences (knowledge base) and questions, it can form episodic memories, and use them to generate relevant answers.

The architecture has the following components:

The Semantic Memory Module(analogous to a knowledge base) consists of pre-trained GloVe vectors that are used to create sequences of word embeddings from input sentences. These vectors will act as inputs to the model.

The Input Module processes the input vectors associated with a question into a set of vectors termed facts. This module is implemented using a Gated Recurrent Unit.

The Gated Recurrent Unit enables the network to learn if the sentence currently under consideration is relevant or has nothing to do with the answer.

The Question Module processes the question word by word and outputs a vector using the same GRU as the input module, and the same weights.

Both facts and questions are encoded as embeddings. The Episodic Memory Module receives the fact and question vectors extracted from the input and are then encoded as embeddings. This uses a process inspired by the brain’s hippocampus, which can retrieve temporal states that are triggered by some response, like sights or sounds.

Finally, the **Answer Module** generates an appropriate response. By the final pass, the episodic memory should contain all the information required to answer the question. This module uses another GRU, trained with the cross-entropy error classification of the correct sequence, which can then be converted back to natural language.

And finally, the response is given back to the user.

### Conclusion:

In conclusion, the generation of an appropriate answer to a user’s input is a complex task that involve several stages, namely the correct detection of the user’s input, interpretation of user’s intent, and context coreference resolution.

These components involve use of state of the art techniques such as DNN acoustic model and Hidden Markov model for wake word detection, Automatic speech recognition for information transfer, Natural language understanding for sequencing and analysis of order and using deep learning for giving the user appropriate response.

Over the past few years, smart assistants have learned to carry over the context from one query to another and understand follow up questions. They are much more capable of understanding natural language much better. These improvements are the result of continued introduction and refinement of machine learning and deep learning techniques.

### References:

[1] Terzopoulos, George and Maya Satratzemi. “Voice Assistants and Smart Speakers in Everyday Life and in Education.” Informatics Educ. 19 (2020): 473-490.

[2]

[3]

[4]