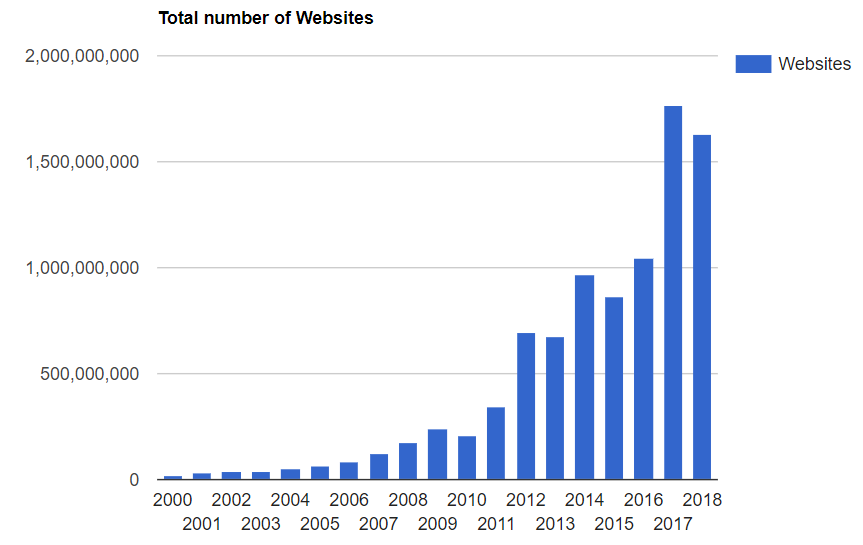
Introduction

Motivation

The internet has been, since it publicly launched in the 1990s, one of the biggest innovations and influencers towards building the technological world that we live in today. The World Wide Web has expanded over the course of the last 30 years to reach over 5 billion pages that are indexed through search engines, with over 1.6 billion web sites being active as of today.



But more importantly, over 50% of the world’s population is using the internet in 2019, either on a computer or mobile devices. Some of the most popular activities on the internet are researching, browsing social media or online shopping, as statistics show the more popular domains being *google.com, facebook.com, wikipedia.org, amazon.com.*

The high demand of internet usage makes it necessary to continue improving the browsing experience and create new technologies that can make it more efficient to accomplish certain tasks that are difficult to achieve otherwise. Using applications through voice commands has proven to be a necessary tool for the common user as most smartphones, TVs, speakers and car systems nowadays have such technologies implemented. Some of these include Google Assistant (enabled on 1 billion devices), Amazon’s Alexa (100 million devices), Apple’s Siri (over 500 million devices). It is also estimated that the usage of voice assistants will triple by 2023 as the home devices industry will keep improving their features and become more attractive to the public.

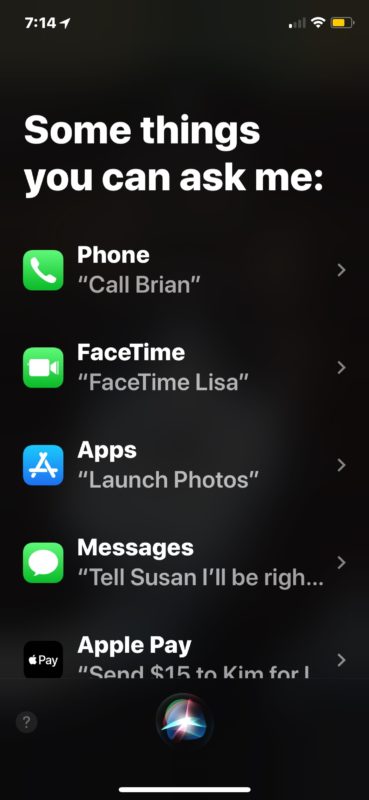
Regarding web accessing on computers, for over 3 years Google Chrome has been the most used browser because of its simplicity, speed, security and large catalog of user developed extensions. Just as mobile phones are getting more and more popular with using a voice assistant, the computer could also find it to be more convenient to be able to perform commands without having to use a mouse or keyboard. My application tackles the problem by dividing it in two parts, taking the vocal input from the user and initiating commands towards the browser based on that input.

Similar applications

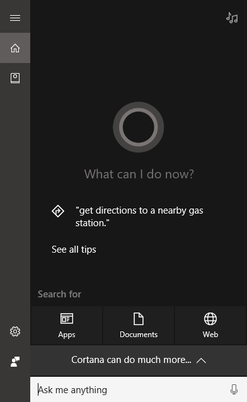
**Google Assistant** has launched in 2016 and it lets you interact with your Android device using voice commands. The way it works is starting up after the user says “Ok Google”, and wait for the user to specify their command. The assistant is able to find information on your device, such as contacts, messages, calendar notes or open applications, but also find information online such as the weather, bookings, points of interest, or random searching.



**Siri** is Apple’s voice assistant which launched in 2011, currently being compatible with every built-in application on Apple devices (interacting with them similarly to Google Assistant) and starts up with the command “Hey Siri”. It can only run on iOS devices while Google Assistant can run both on iOS and Android.



**Cortana** is a voice assistant launched by Microsoft in 2014 for Windows 8.1 and released also for Windows 10 and few other platforms. It is integrated in Microsoft’s Edge browser and uses Bing as a search engine, performing similar tasks to the other assistants, but on a computer.



Short description

While all these assistants are able to interact with your device to retrieve information and perform actions, on the browser level, their functionality is focused towards searching the web, finding topics on Wikipedia, telling the weather or finding certain locations or news, which only lead the user to the information that they need.

My application is focused towards handling the recurrent actions that a user performs on a specific web page, such a scrolling, clicking on buttons, selecting text, filling input fields or downloading content. To interact with the program the user needs to hold down the *RCtrl* key in order to input their command and then release it so the voice input can be recognized and transformed into text. Then, the text is processed by the application which will determine based on the similarity of it with the supported commands, what did the user want to happen, and then the function which launches the interaction with the browser on that specific command will be called and executed. If the command cannot be identified by the speech recognizer it will not be forwarded to the text analyzer and, similarly, if the text produced by the user’s input doesn’t resemble any of the supported commands, the user will be informed that they should try to repeat their request. Each request from the user is handled at a separate time and they will not be able to insert a new command until the previous one has finished its execution.

The complexity of the application depends on the function that the user requests at a certain time, many being executed in constant time such as pressing some button combinations on the keyboard or performing a javascript call, but others such as clicking on a button or scrolling the page require a linear time depending on the number of buttons that need to be searched on that page or the height that needs to be scrolled.

The purpose of my project is to create a program that can interact with the browser in a different way, that has not been popularized as of today and that, in my opinion, has the potential to be a more appealing way of using internet services to a certain group of users.

Chapters

1. Speech recognition and Browser functionality
2. Technologies used
3. Personal contributions
4. Application architecture
5. ???
6. Conclusions

Chapter 1.

Speech recognition and Browser functionality

**Speech recognition** refers to the ability of a program to transform an audio input representing a speech, into a text representing the words of what was said. It is a popular technology used in many devices nowadays, such as mobile phones, household devices, car systems or robotics.

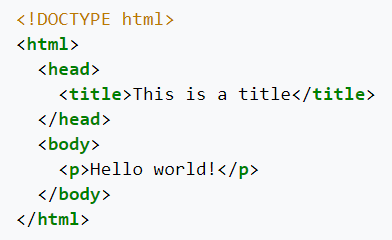
Some of the models that try to solve the problem of recognizing speech are:

* *Hidden Markov models* – “doubly stochastic process with an underlying stochastic process that is not observable (it is hidden), but can only be observed through another set of stochastic processes that produce the sequence of observed symbols” [1]
* *Dynamic time warping* – algorithm that analyzes words based on the speed of the speaker
* *Neural networks* – network trained to map vocal inputs to text outputs in order to be able to determine the output for new inputs

These technologies are used in implementing the speech recognizers available today, such as Google API, Microsoft API And CMU Sphinx.

Another important aspect in developing the application is to understand the **browser’s functionality**. The World Wide Web is a collection of web sites linked through URIs, that host resources which can be accessed through HTTP requests. In order to view the information listen on a web page, a user has to type an URL in the address bar of a browser and submit it, which will send a request to the host of the web site, returning the resources necessary to build the page in the browser.

Out of all the resources of a web site, the HTML file is the most important in determining the structure and the elements belonging to that specific page. The basic structure of an HTML file looks like this:



Finding the elements of a web page can be realized by searching through different tags or values inside the file, such as *<p>* or *<h1>* to determine paragraphs, *<button>* to find buttons, *<a>* for hyperlinks, or *<input>* for input fields. Then, the program would have to read or modify the information of that field.

Chapter 2.

Technologies used

**Speech Recognition Module**

Python’s *SpeechRecognition* module is a library that supports multiple speech recognition APIs, and has the ability to listen and record a microphone in the background, convert audio input to text and save audio data to files. Some of the APIs supported are: CMU Sphinx, Google Speech API, Wit.ai, Microsoft API, Houndify or IBM.

To use the **Microphone**, the user has first to install the additional *PyAudio* module, which is used to record audio input. On instantiation, the constructor of the *Microphone* can be given the following optional parameters:

**Microphone(device\_index: Union[int, None] = None, sample\_rate: int = 16000, chunk\_size: int = 1024) -> Microphone**

1. *device\_index*: an integer between 0 and the number of audio devices on the system, representing the index of the device that will be used to obtain audio input from; if unspecified, the default microphone will be selected
2. *sample\_rate*: the rate of samples per second at which audio chunks are recorded; a higher value can result in a better audio quality, but slower recognition
3. *chunk\_size*: the number of chunks in which the audio input will be separated

On instantiation, the *Microphone* object will check whether the user has assigned any optional parameters, and if so, whether they are the correct type (throws exception otherwise). Then it tries to import the *PyAudio* module and instantiate an object of it to use for the recording, raising *AttributeError* if the library hasn’t been installed. If no error occurs, all the components needed for the *Microphone* object will be saved as class variables.

The **Recognizer** is a class used for setting up different functionalities of speech recognition, starting and stopping the recording process and selecting the API that will perform the transcription. It has no additional parameters and on initialization, it will set some class variables to default values:

* *recognizer\_instance.energy\_threshold = 300 # type: float*

A variable that represents the perceived loudness of the sound, the value 300 being considered the threshold between silence and speech. But since each microphone has a different sensitivity that requires a different threshold, this value will be considered only as the initial (minimum) value and will need to be dynamically adjusted based on the hardware used and the loudness of the environment. This initial value can be modified to a lower value in the case that the microphone is not able to detect audio input, or to a higher value if the background noise makes it hard to understand and transcribe what the speaker is saying.

* *recognizer\_instance.dynamic\_energy\_threshold = True # type: bool*

In most cases, the energy threshold needs to be automatically adjusted based on the used microphone device and the background noise in the room. This value should be changed only if the ambient of the room remains constant and the initial energy threshold is set to the appropriate value.

* *recognizer\_instance.dynamic\_energy\_adjustment\_damping = 0.15 # type: float*

Approximates the speed at which the energy threshold is dynamically adjusted in one second (if the *dynamic\_energy\_threshold* property is enabled). A lower value means a faster adjustment, but it could miss certain parts of the audio input, while a higher value will decrease the rate of modification.

* *recognizer\_instance.dynamic\_energy\_adjustment\_ratio = 1.5 # type: float*

Represents the minimum ratio at which the speaker’s voice is louder than the background noise (if the *dynamic\_energy\_threshold* property is enabled). A smaller value would result in a harder differentiation between the two, while a larger value requires a quieter ambient noise.

* *recognizer\_instance.pause\_threshold = 0.8 # type: float*

The minimum number of seconds that represents silence between phrases. Lowering the value could generate a negative output for slower speakers, while increasing it could damage it for faster ones.

* *recognizer\_instance.operation\_timeout = None # type: Union[float, None]*

The number of seconds after which an internal operation (such as an API request) will timeout. The *None* default value means there is no timeout and the program will wait for the action to be performed.

In order to initiate the recording of the microphone, the following function will be used:

**recognizer\_instance.listen\_in\_background(source:AudioSource,callback: Callable[[Recognizer, AudioData], Any]) -> Callable[bool, None]**

Given an *AudioSource* instance as input (in this case, the *Microphone* object), it will transform it into an *AudioData* instance and send it as a parameter to the given callback function. It returns a function that, when called, will stop the background listening. The process will execute on a separate thread, so that the main thread will be able to perform other actions before it chooses to stop it.

When called, the function will check whether the given input is of the correct type (*AudioSource*), then it will initialize and start the thread that performs the listening. The thread is also marked as a daemon thread, which means that it will terminate automatically if the main thread completes (as long as there’s no other non-daemon thread present), and it can also be shut down at any time during its execution. The recording thread will continuously call a separate *listen* function that will handle the conversion of *AudioSource* to *AudioData* for each phrase, and, after it, the callback function which will do the recognizing. The thread will check every second if the stop function has been called.

The stop function is the function returned by the *listen\_in\_background* method, that, when called, will stop the recording thread from sending sound data to the converting function and thus finishing the thread. It has one optional parameter, *wait\_for\_stop*, that is set to a default *True* value, which will make the function wait for the thread to finish before returning. If set to *False*, the method will return right away, while the thread will still be on its way to end.

While *AudioSource* is an abstract class, implemented by the *Microphone* class, an *AudioData* object holds in the *frame\_data* variable a sequence of bytes that represent audio samples. This raw data can be then converted into different audio formats, if needed. The object also contains a *sample\_width*, representing the width of each sample in bytes, and a *sample\_rate* for the data.

**AudioData(frame\_data: bytes, sample\_rate: int, sample\_width: int) -> AudioData**

The *listen* function that performs the conversion from the *Microphone* to *AudioData* has the following structure:

**recognizer\_instance.listen(source: AudioSource, timeout: Union[float, None] = None, phrase\_time\_limit: Union[float, None] = None, snowboy\_configuration: Union[Tuple[str, Iterable[str]], None] = None) -> AudioData**

The recording takes place when the energy detected is above the *recognizer\_instance.energy\_threshold* value, and ends after few seconds of silence, given by the value of *recognizer\_instance.pause\_threshold*. It can stop also when there’s no more audio input sent to the method (when the stop function of *listen\_in\_background* is called). It has the following parameters:

1. *source*: the input *AudioSource* object
2. *timeout*: the number of seconds the function waits for an input phrase before exiting; if *None*, it will wait indefinitely
3. *phrase\_time\_limit*: the maximum number of seconds allowed for recording one phrase, before stopping and processing what was recorded in that timeframe; if *None*, a phrase will have no time limit
4. *snowboy\_configuration*: a configuration for a third party recognizer that can train a neural network based on the given inputs; if *None*, it will not be used

The way it works is verifying, first, that the source is a valid *AudioSource* object, then recording the audio input until there is a silence of *timeout* length (if not *None*), or until a phrase is long enough. If *dynamic\_energy\_threshold* is enabled, the *energy\_threshold* will be updated based on the current threshold, the *dynamic\_energy\_adjustment\_damping* and *dynamic\_energy\_adjustment\_ratio*:



The audio input of the microphone will be turned into a sequence of bytes that, concatenated, will represent the *frame\_data* element of an *AudioData* object. The transformation is done using *PyAudio*’s *read* function, that takes as a parameter the number of chunks to be read. After finishing, the function will return the newly created *AudioData* object.

**Google Speech API**

“Google has improved its speech recognition by using a new technology in many applications with the Google App such as Goog411, Voice Search on mobile, Voice Actions, Voice Input (spoken input to keypad), Android Developer APIs, Voice Search on desktop, YouTube transcription and Translate, Navigate, TTS. After Google, has used the new technology that is the deep learning neural networks, Google achieved an 8 percent error rate in 2015 that is reduction of more than 23 percent from year 2013. According to Pichai, senior vice president of Android, Chrome, and Apps at Google, “We have the best investments in machine learning over the past many years. Indeed, Google has acquired several deep learning companies over the years, including DeepMind, DNNresearch, and Jetpac”.”[2]

The way Google Speech API is integrated in the *SpeechRecognition* module is through the method *recognize\_google* of the *Recognizer* object.

**recognizer\_instance.recognize\_google(audio\_data: AudioData, key: Union[str, None] = None, language: str = "en-US", , pfilter: Union[0, 1], show\_all: bool = False) -> Union[str, Dict[str, Any]]**

The function would be called after obtaining an *AudioData* object from the *listen* function and has the purpose to obtain the transcription of audio data (represented as a sequence of bytes that build a specific format file) by sending a request to the API with the given input. It will return an error if the speech cannot be recognized or if the API cannot be accessed. As parameters, it takes:

1. *audio\_data*: the *AudioData* object that will be transcribed
2. *key*: a key required to access the API; if *None*, will use a default key
3. *language*: the language of the spoken text
4. *pfilter*: profanity filter
5. *show\_all*: if *False*, will return the response as a string, else, as a json dictionary

On implementation, the function will check that the given *audio\_data* is of the correct type, and then will convert the bit string of the *frame\_data* element in a byte representation of a FLAC file, using a third party executable. Then, it will send the request to the API and determine the result with the highest confidence rate from within the response dictionary:

****

Possible outcome:

{    
   **"result"**:[    
      {    
         **"alternative"**:[    
            {    
               **"transcript"**:"hello world",  
               **"confidence"**:0.80891722  
            },  
            {    
               **"transcript"**:"hello-world"  
            },  
            {    
               **"transcript"**:"hello word"  
            },  
            {    
               **"transcript"**:"hello work"  
            },  
            {    
               **"transcript"**:"hallo world"  
            }  
         ],  
         **"final"**:true  
      }  
   ],  
   **"result\_index"**:0  
}

**Levenshtein Module**

**Selenium’s Webdriver**

**Selenium’s Action Chains**

**Pynput Module’s Keyboard Listener & Controller**

**Javascript Commands**

**// + daca mai adaug ceva**

Sources:

// de unde am preluat niste informatii statistice si legate de aplicatii similare pentru introducere

<https://www.internetlivestats.com/total-number-of-websites/>

<https://www.internetworldstats.com/stats.htm>

<https://www.worldwidewebsize.com/>

<https://www.juniperresearch.com/press/press-releases/digital-voice-assistants-in-use-to-triple>

<https://www.pocket-lint.com/apps/news/google/137722-what-is-google-assistant-how-does-it-work-and-which-devices-offer-it>

<https://www.pocket-lint.com/apps/news/apple/112346-what-is-siri-apple-s-personal-voice-assistant-explained>

<https://en.wikipedia.org/wiki/Cortana>

// informatii din capitolul 1

<https://en.wikipedia.org/wiki/Speech_recognition>

<https://en.wikipedia.org/wiki/World_Wide_Web>

<https://en.wikipedia.org/wiki/HTML>

// informatii din capitolul 2

<https://pypi.org/project/SpeechRecognition/>

<https://github.com/Uberi/speech_recognition/blob/master/reference/library-reference.rst>

References:

// de pe scholar

[1] Rabiner, Lawrence R., and Biing-Hwang Juang. "An introduction to hidden Markov models." ieee assp magazine 3.1 (1986): 4-16.

[2] Këpuska, Veton, and Gamal Bohouta. "Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx)." Int. J. Eng. Res. Appl 7.03 (2017): 20-24.