CHAPTER I

THE STUDY AND ITS PROBLEMS

A. BACKGROUND OF THE STUDY

The artificial intelligence, voice recognition, natural language processing and learning machines space has been growing rapidly. The idea of having a personal assistant which connects us to the web and the ever growing Internet of Things is becoming ever more commonplace.

An intelligent personal assistant (or simply IPA) is a software agent that can perform tasks or services for an individual. These tasks or services are based on user input, location awareness, and the ability to access information from a variety of online sources (such as weather or traffic conditions, news, stock prices, user schedules, retail prices, etc.). Examples of such an agent are Apple's Siri, Google's Google Now, Amazon Alexa, Microsoft's Cortana, Braina (application developed by Brainasoft for Microsoft Windows), Samsung's S Voice, LG G3's Voice Mate, BlackBerry's Assistant, SILVIA, HTC's Hidi, IBM's Watson (computer), Facebook's M (app) and One Voice Technologies (IVAN).

According to venture capitalist Chi-Hua Chien of Kleiner Perkins Caufield & Byers, examples of tasks that may be

performed by a smart personal agent-type of Intelligent Automated Assistant include schedule management (e.g., sending an alert to a dinner date that a user is running late due to traffic conditions, update schedules for both parties, and change the restaurant reservation time) and personal health management (e.g., monitoring caloric intake, heart rate and exercise regimen, then making recommendations for healthy choices).

Intelligent personal assistant technology is enabled by the combination of mobile devices, application programming interfaces (APIs), and the proliferation of mobile apps. However, intelligent automated assistants are designed to perform specific, one-time tasks specified by user voice instructions, while smart personal agents perform ongoing tasks (e.g., schedule management) autonomously.

Simply put, artificial intelligence is a sub-field of computer science. Its goal is to enable the development of computers that are able to do things normally done by people -- in particular, things associated with people acting intelligently.

Stanford researcher John McCarthy coined the term in 1956 during what is now called The Dartmouth Conference, where the core mission of the AI field was defined. However, the five decades since the inception of AI have brought only very slow progress, and early optimism concerning the attainment of human-level intelligence has given way to an appreciation of the profound difficulty of the problem. Artificial Intelligence are designed to perceive human mind or cognitive functions. It is designed to think as something stated. It response as if you are communicating with someone. Today, artificial intelligence is widely used in medical diagnosis, robotics, games and etc. It is commonly used because Artificial Intelligence is very useful in their respective fields. Mainstream thinking in psychology regards human intelligence not as a single ability or cognitive process but rather as an array of separate components. Research in AI has focused chiefly on the following components of intelligence: learning, reasoning, problem-solving, perception, and language-understanding. But analysing use data and providing recommendations like that is relatively easy, and it doesn’t quite give the feel of a “personal assistant.” The real task – and the real difficulty – for IPAs is recording, understanding, and effectively responding to human speech.

While we take them for granted, our hearing and language systems are remarkable feats of computation.

While understanding language is relatively easy for most four-year-olds, decoding human speech is a remarkably complex task. In fact, decoding a verbal request requires more than 100 times as much processing power as responding to a textual search request. This is because while our ears have been honed by millions of years of evolution to pick up and decode human speech, IPAs only have built-in microphones, and these microphones do little to delineate human speech from surrounding noise.

Artificial Intelligence plays an important role in today’s technology. Researchers and Developers believe that the appropriateness of Artificial Intelligence used to a certain level may offer a chance for man to communicate without any restrictions and make computing an innovative and futuristic success.

B. ALGORITHMS

This study is focused on Intelligent Personal Assistant Platform Development. Before discussing the actual algorithms to be used in the development phase, this part of the section will briefly discuss a typical IPA Algorithm and how it works. This concept is important to fully understand the relevance of the to-be-discussed core algorithms and how will it all be combined to build a development platform.

An IPA Algorithm:

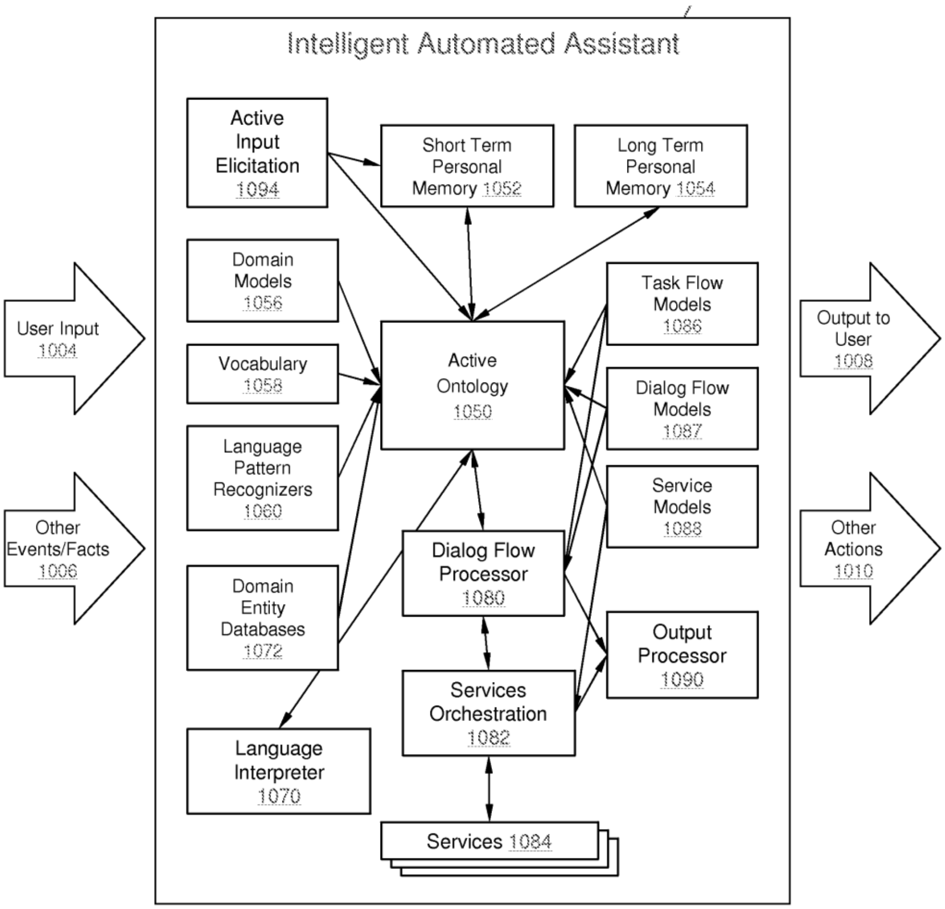

Intelligent Automated Assistant
Active Input Elicitation 1094
Short Term Personal Memory 1052
Long Term Personal Memory 1054
Domain Models 1056
Vocabulary 1058
Language Pattern Recognizers 1060
Domain Entity Databases 1072
Language Interpreter 1070
Active Ontology 1050
Dialog Flow Processor 1080
Services Orchestration 1082
Services 1084
Task Flow Models 1086
Dialog Flow Models 1087
Service Models 1088
Output Processor 1090
User Input 1004
Other Events/Facts 1006
Output to User 1008
Other Actions 1010

Figure 1.0 IPA Algorithm

On the above diagram we can see how many components work separately inside an IPA Core. Language Processing, Vocabularies, Dialog Management etc. act as separate modules and the tedious tasks of incorporating and making them work altogether is the burden of the developer.

After thorough analyzation, we conceptualized an algorithm that will simplify the above stated process. This is done by grouping the three main components (Speech Recognition, Language Understanding, Speech Synthesis) and combining them to develop a core which will act as the central processing module and message bus. External and Add-On modules and services will be then incorporated by accessing the Platform's API.

Here is the illustrated projected platform that will be developed using our algorithm:

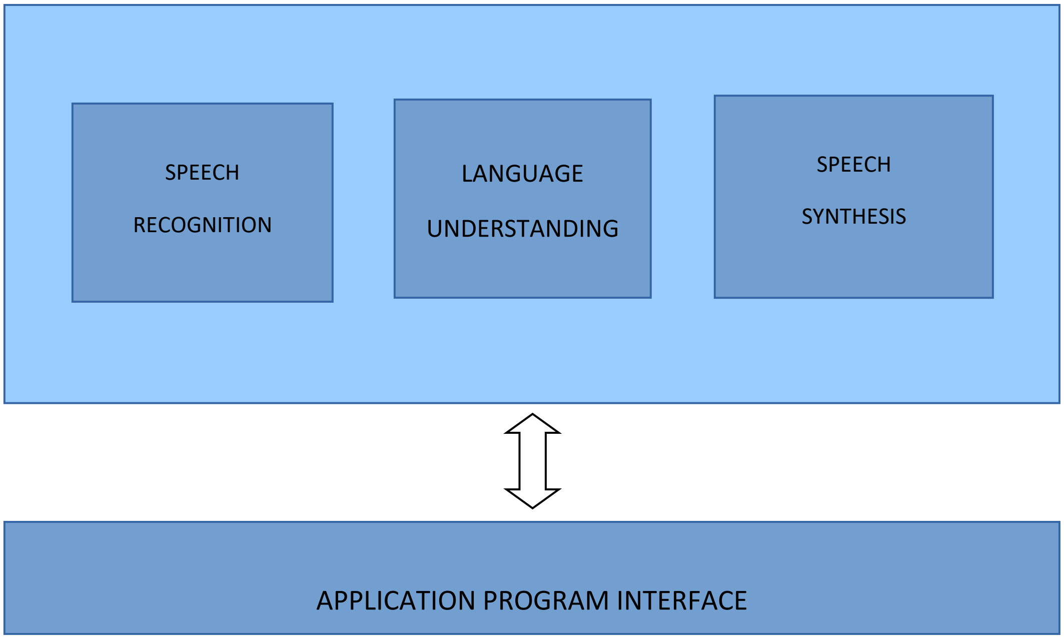


Figure 2.0 Projected Core Platform Architecture

Here are the existing algorithms that will be the core of this study:

1.) Sphinx Continuous Speaker-Independent Speech Recognition

Sphinx is a continuous-speech, speaker-independent recognition algorithm making use of hidden Markov acoustic models and anagram statistical language model.

Sphinx Speech Recognition Algorithm:

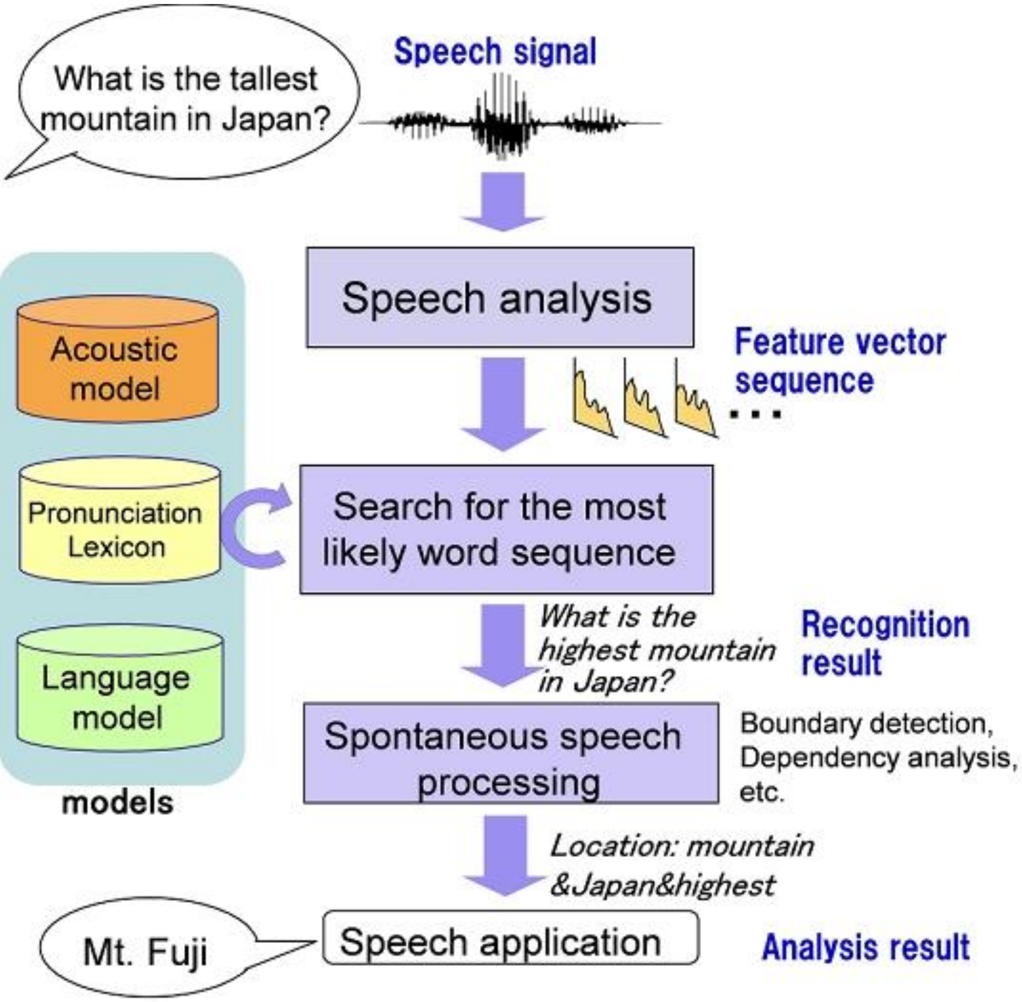

What is the tallest mountain in Japan?
Speech signal
Speech analysis
Feature vector sequence
Acoustic model
Pronunciation Lexicon
Language model
models
Search for the most likely word sequence
What is the highest mountain in Japan?
Recognition result
Spontaneous speech processing
Boundary detection, Dependency analysis, etc.
Location: mountain &Japan&highest
Mt. Fuji
Speech application
Analysis result

Figure 3.0 Sphinx Speech Recognition

SPEECH ANALYSIS

1. Choose amplitude data into frames

2. Get frequency spectrum of each frame

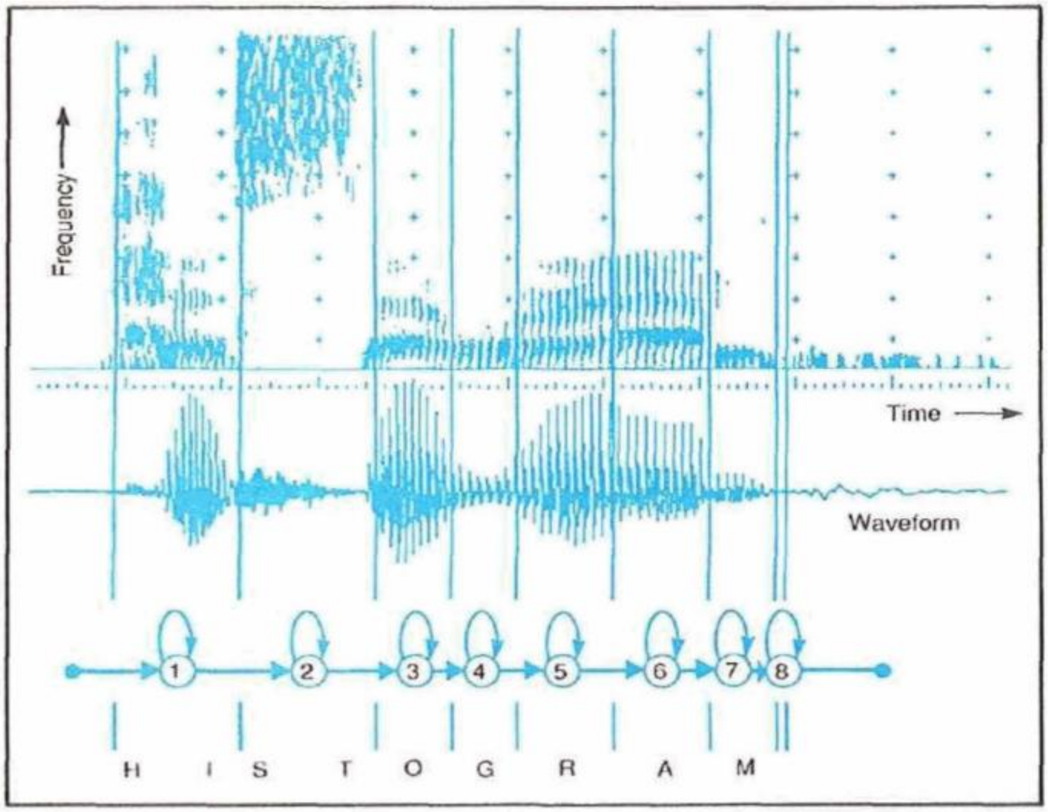


Figure 4.0 Frame Spectrum

3. Classify time frames as phonetic categories

WORD SEQUENCE GENERATION

1. Each phonetic frame will be saved and represented by an integer index

WORD MATCHING

1. Phonetic frames will be matched accordingly to possible defined Acoustic Models, Language Models and Pronunciation Lexicon

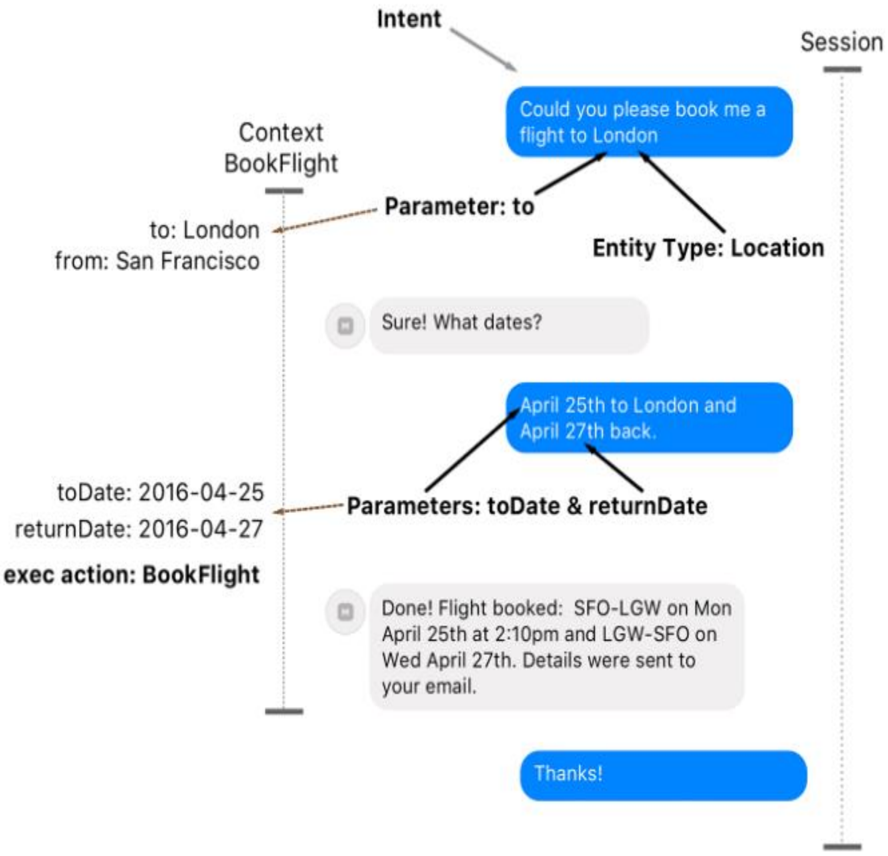
TEXT GENERATION

1. Algorithm will now generate a string or set of strings that phonetically represents the frames

2.) Adapt Pattern Based Text Intent and Entity Extraction

Adapt is an algorithm for converting natural language into machine readable data structures. Adapt is lightweight and streamlined and is designed to run on devices with limited computing resources. Adapt takes in natural language and outputs a data structure that includes the intent, a match probability, a tagged list of entities.

Overview of Adapt Intent and Entity Extraction Algorithm:



Could you please book me a flight to London

Sure! What dates?

April 25th to London and April 27th back.

Done! Flight booked: SFO-LGW on Mon April 25th at 2:10pm and LGW-SFO on Wed April 27th. Details were sent to your email.

Thanks!

Figure 5.0 Intent and Entity Extraction

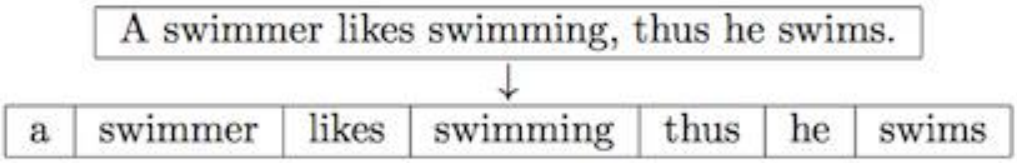
INPUT ACCEPTANCE

1. Natural Language Text Input

"Go to Starbucks tomorrow at 5 pm

TEXT ANALYSIS

1. Tokenize text to separate segments



A swimmer likes swimming, thus he swims.

↓

a	swimmer	likes	swimming	thus	he	swims

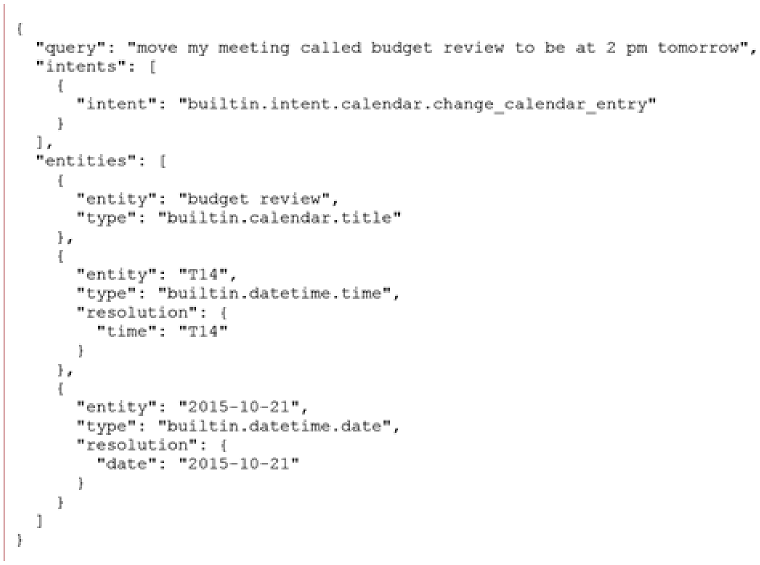
INTENT RECOGNITION

1. Identify and extract entities based from provided pattern

2. Match text input to a defined pattern in database and recognize intent

EXPORT TO MACHINE READABLE FORMAT

1. Intent and Entities will be formatted and exported in a machine readable format such as XML or JSON



```
{
  "query": "move my meeting called budget review to be at 2 pm tomorrow",
  "intents": [
    {
      "intent": "builtin.intent.calendar.change_calendar_entry"
    }
  ],
  "entities": [
    {
      "entity": "budget review",
      "type": "builtin.calendar.title"
    },
    {
      "entity": "T14",
      "type": "builtin.datetime.time",
      "resolution": {
        "time": "T14"
      }
    },
    {
      "entity": "2015-10-21",
      "type": "builtin.datetime.date",
      "resolution": {
        "date": "2015-10-21"
      }
    }
  ]
}
```

3.) <u>eSpeak Sinusoidal Speech Synthesis</u>

eSpeak is a speech synthesizer algorithm that uses a formant synthesis method, providing many languages in a small size. In addition, eSpeak can be used as a front-end, providing text-to-phoneme translation.

eSpeak Speech Synthesis Algorithm:

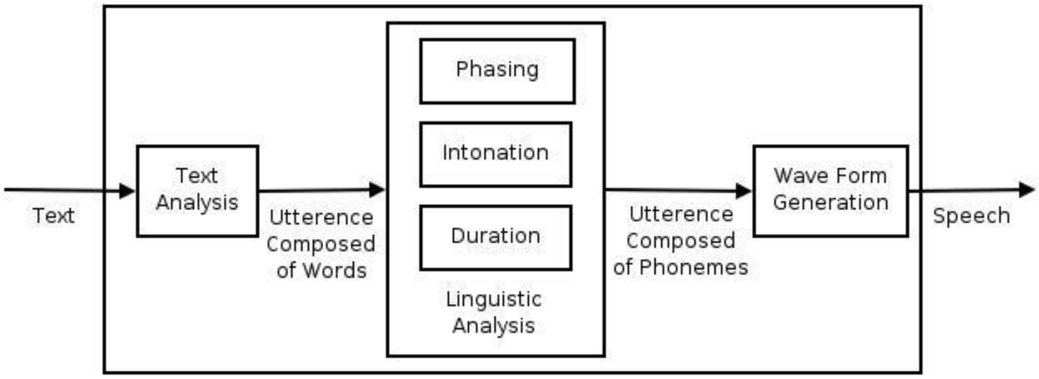


Figure 6.0 eSpeak Speech Synthesis

TEXT ANALYSIS

1. Text and symbols will be tokenized

2. After tokenization, words will be syllabicated and symbols will be evaluated to extract the phonemes

3. Phonemes will be given indices to allow programmability and matching

PHONEME MATCHING

1. Phonemes extracted from the text will be individually matched to existing sinusoidal phones in the database

PHONE CONCATENATION

1. Matched phones will then be concatenated to form the sound synthesis

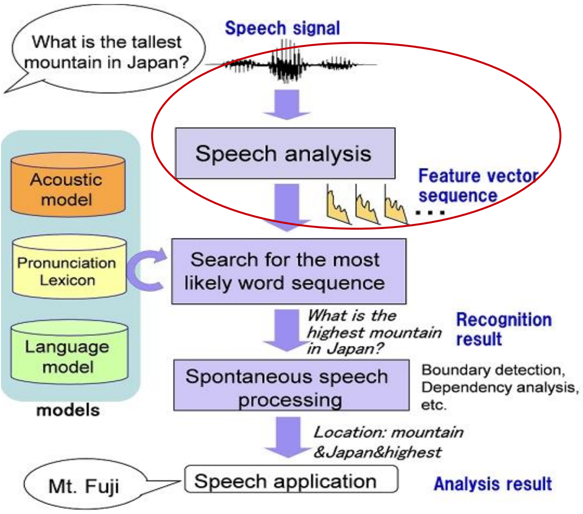
C. STATEMENT OF PROBLEMS

1.) Sphinx: Algorithm shows word boundary ambiguity when transcribing phonemes with multiple ways of grouping phones into words

Discussion:

When a sequence of groups of phones are put into a sequence of words, we sometimes encounter word boundary ambiguity. Word boundary ambiguity occurs when there are multiple ways of grouping phones into words.

Where in the Algorithm:


What is the tallest mountain in Japan?
Speech signal
Speech analysis
Feature vector sequence
Acoustic model
Pronunciation Lexicon
Language model
models
Search for the most likely word sequence
What is the highest mountain in Japan?
Recognition result
Spontaneous speech processing
Boundary detection, Dependency analysis, etc.
Location: mountain &Japan&highest
Mt. Fuji
Speech application
Analysis result

Simulation:

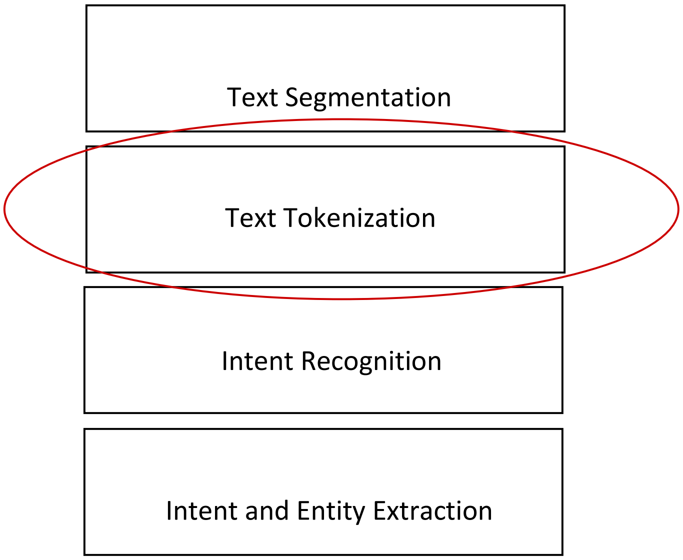
1st Run	*It's not easy to wreck a nice beach.*
2nd Run	*It's not easy to recognize speech.*
3rd Run	*It's not easy to wreck an ice beach.*

2.) Adapt: Algorithm strictly matches input to a pattern without validating words semantically

Discussion:

Entity extraction in the algorithm is achieved by matching the text input to a defined pattern or regular expression wherein parameters and location of the entity to be extracted is indicated in the said pattern or sample. The algorithm strongly matches parameters without checking and analysing semantic properties of the word.

Where in Algorithm:



Text Segmentation

Text Tokenization

Intent Recognition

Intent and Entity Extraction

Simulation:

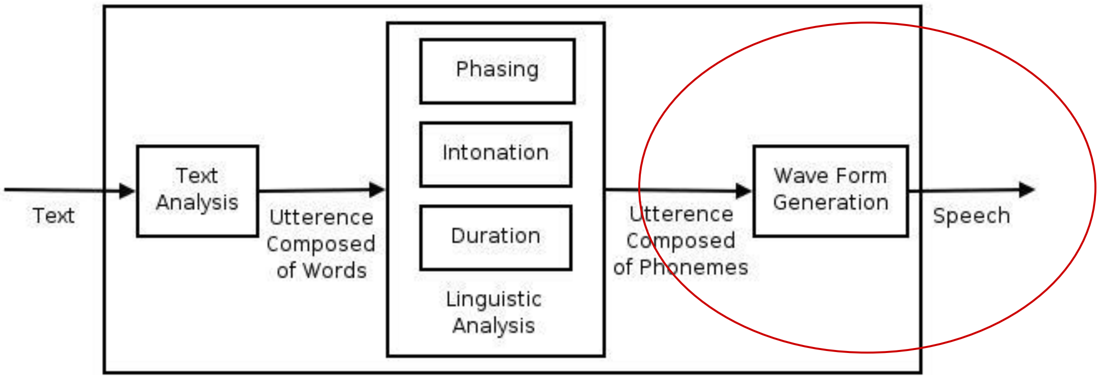
NL Text	*Set a meeting on Friday with John at 1:00pm.*
Pattern	*Set a meeting with John on Friday.*
Output Entities	*Person: Friday, Day: John, Unhandled: 1:00pm*

3.) eSpeak: <u>Algorithm randomly generates gaps in audio outputs when processing a group of speech samples</u>

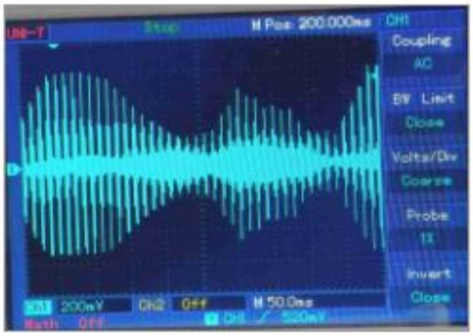
Discussion:

This problem occurs when the algorithm transcribes a group of text that takes several seconds to read. The synthesis suddenly, between groups of speech samples stutters. Random gaps are somewhat generated when speech samples are played, so due to the gaps the speech sounds stretched and slow. This is a problem of course because a speech synthesizer is supposed to be understandable and clear.

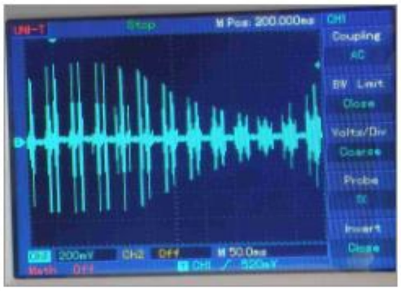
Where in Algorithm:



Simulation:



Start of the output



some seconds later, stuttering

D. OBJECTIVES OF THE STUDY

As user demand scales for intelligent personal assistants (IPAs) such as Apple's Siri, Google's Google Now, and Microsoft's Cortana, we are approaching an era automation. It is an open question how technology and developers alike should evolve to enable this emerging class of applications, and the lack of an Open Source IPA Development Platform workload is an obstacle in addressing this question.

In this paper, we present the design o Apiaro, an open end-to-end IPA Development Platform and API that can be used to build, code and extend a user's own customized IPA with features such as Speech to Text Processing, Natural Language Processing, Text to Speech, and API and Web Services Compatibility.

We aim to give developers and enthusiast alike an algorithm or platform that will enable them to build, customize and redistribute their own Intelligent Personal Assistant.

E. IMPORTANCE OF THE STUDY

Intelligent Personal Assistants is the future of computing and further development of technologies in this area will surely be the trend for the future. It was one of the strongest fields of development during 2015 and it will also be in 2016 and will certainly be one of the sectors generating most profits for years or decades to come.

1.) Intelligent machines can replace human beings in many areas of work. Robots can do certain laborious tasks. Painstaking activities, which have long been carried out by humans can be taken over by the robots. Owing to the intelligence programmed in them, the machines can shoulder greater responsibilities and can be programmed to manage themselves.

2.) Smartphones are a great example of the application of artificial intelligence. In utilities like predicting what a user is going to type and correcting human errors in spelling, machine intelligence is at work.

3.) Artificial intelligence can be utilized in carrying out repetitive and time-consuming tasks efficiently.

F. SIGNIFICANCE OF THE STUDY

This study seeks to benefit the following groups of people:

1.) To the ***computer and smartphone users***, to make computing a more interesting and worthwhile experience through the use and utilization of an A.I. Agent for automation and to ease them of the burden of executing certain computing tasks.

2.) To the ***programmers and developers,*** this study will encourage them to partake in the improvement of Artificial Intelligence and to advance and develop applications and system that utilizes this technology.

3.) To ***business owners and planners,*** A.I. is the current trend of today's technology. It exists now in almost every device. This study can help them to take their business ideas to the next level through the use of A.I. agents to cater customer's needs.

4.) Lastly, to the ***future researchers,*** that they will further this study through investing more time in finding out the effectiveness and relevance and to what extent can A.I. help in today's computing technologies.

G. SCOPE AND LIMITATIONS

This study focuses on the development of a Core Platform and API for building Intelligent Personal Assistants based from three algorithms in the field of human – computer interaction. With this algorithm or platform, IPAs can be extended to include Machine Learning, Computer, Tasks, and Home Automation, Problem Solving and Queries.

Additionally, functions and features of this project only covers topic Artificial Intelligence, Computational Knowledge, Machine Learning and Cognitive Perceptions in the field of Computer Programming and Development and not topics directly related to Medicine, Engineering and other subject areas not mentioned or related above.

APIs and SDKs that will be used in the development of this project will be strictly Open Sourced Projects, Algorithms and Technologies with General Public License Version 3, which allows developers to use, modify and redistribute code.

This study will not cover or in any way will involve proprietary technologies and features of similar Intelligent Personal Assistants such as Microsoft ‘s Cortana, Apple’s Siri, and Google’s Google Now, etc. Any similarity among these systems within this study is not intended and may be incidental.

H. DEFINITION OF TERMS

Acoustic Model is a set of words or sentences a Speech to Text Engine can understand.

Actuator is a Module or Code that executes a specific Intent.

Application Program Interface (A.P.I) is a set of routines, protocols, and tools for building software applications.

Artificial Intelligence (A.I.) is the simulation of human intelligence processes by machines.

Assistant is the term used for the Artificial Intelligence Entity, can be interchanged to Agent.

Belief is the knowledge base of the Agent about the world.

Desire is the current goal of the Agent in relation to the Belief.

Environment means the real world where the Agent takes input from.

Intelligent Personal Assistant (I.P.A.) is a software agent that can perform tasks or services for an individual.

Intent the code generated proposed “idea” of a natural language entry

Intention is the method of action or the Agent will execute based from the Belief and Desire.

Internet of Things is a proposed development of the internet in which everyday objects have network connectivity, allowing them to send and receive data.

Natural Language Processing (N.L.P.) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human.

Phoneme any of the perceptually distinct units of sound in a specified language that distinguish one word from another

Phonetic Frame is one unit of phoneme in a given word (tokenized and syllabicated)

Raspberry Pi a brand of a programmable single computer board with ARM Architectures.

Speech Recognition is the process of translating speech into human readable text

Speech Synthesis the process of transcribing text into audible computer generated voice