YAŞAR UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

MASTER THESIS

# SPEECH RECOGNITION ALGORITHM IMPLEMENTATION FOR REMOTE CONTROLLING UNMANNED AERIAL VEHICLES (UAVS)

ILHAN SOFUOĞLU

THESIS ADVISOR: ASST. PROF. DR. İBRAHİM ZİNCİR

COMPUTER ENGINEERING MSC.

PRESENTATION DATE: 26.12.2017

BORNOVA / İZMİR
DECEMBER 2017

We certify that, as the jury, we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

**Jury Members:** **Signature:**

Asst.Prof. Dr. İbrahim ZİNCİR
Yaşar University ……………………

Assoc.Prof. Dr. Mehmet Süleyman ÜNLÜTÜRK
Yaşar University M.S. ……………………

Asst.Prof. Dr. Samsun BAŞARICI
Adnan Menderes University ……………………

-----------------------------------------------------------------------

Prof.Dr. Cüneyt GÜZELİŞ
Director of the Graduate School

# ABSTRACT

## SPEECH RECOGNITION ALGORITHM IMPLEMENTATION FOR REMOTE CONTROLLING UNMANNED AERIAL VEHICLES (UAVS)

Sofuoğlu, İlhan
MSc. Computer Engineering
Advisor: Asst. Prof. Dr. İbrahim Zincir
December 2017

This study aims to develop a system that directs the Bebop Drone (Unmanned Aerial Vehicle) using Turkish isolated voice commands, isolated words of a voice recognition algorithm. Some of the algorithms and methods used in the previous studies have been examined. Some algorithms such as Gaussian Mixture Models, Hidden Markov Models, Dynamic Time Warping, Artificial Neural Networks and Google's Deep Learning Algorithm have been examined in the literature with the studies mentioned as isolated word recognition. In addition, preliminary processing such as sampling, windowing, framing, noise reduction and subtraction of audio features before using these classifier algorithms has been examined, and the effects of these processes on the isolated word recognition process have been compared.

**Key Words:** UAV, isolated word recognition, Artificial Neural Network, Hidden Markov Model, Gaussian Mixture Model, Dynamic Time Warping, MFCC, LPCC

# ÖZ

## İNSANSIZ HAVA ARAÇLARININ UZAKTAN KONTROLÜ IÇIN SES TANIMA ALGORITMASI IMPLEMENTASYONU

Sofuoğlu, İlhan
Yüksek Lisans Tezi, Bilgisayar Mühendisliği
Danışman: Yrd. Doç. Dr. İbrahim Zincir
Aralık 2017

Bu çalışma, bir ses tanıma algoritmasının Türkçe izole ses komutları, izole sözcükler kullanılarak Bebop Drone İnsansız Hava Aracını yönlendiren bir sistem geliştirilmesini amaçlar Daha önce yapılmış olan çalışmalarda kullanılan algoritmaların ve yöntemlerin bazıları incelenmiştir. Gaussian Karışım Modeli, Saklı Markov Modelleri, Dinamik Zaman Bükme, Yapay Sinir Ağları ve Google Derin Öğrenme Algoritması gibi bazı algoritmalarının literatürde izole kelime tanıma olarak bahsedilen çalışmaların incelemesi yapılmıştır. Ayrıca bu sınıflandırıcı algoritmaların kullanılmadan önce ses sinyallerinin örnekleme, pencereleme, çerçeveleme, gürültü azaltımı ve öz niteliklerin çıkarılması gibi hangi ön işlemlerden geçtiği incelenmiş ve bu işlemlerin izole sözcük tanıma sürecine etkileri kıyaslanmıştır.

**Anahtar Kelimeler:** İHA, izole kelime tanıma, Yapay Sinir Ağı, Gizli Markov Modeli, Gauss Karışım Modeli, Dinamik Zaman Bükme, MFCC, LPCC

# ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor Asst. Prof. Dr. İbrahim ZİNCİR for his guidance and patience during this study.

I would like to express my enduring love to my parents, who are always supportive, loving and caring to me in every possible way in my life.

I would like to thank all my research assistant friends at Yaşar University who always supported me and helped me with best of their abilities.

İlhan SOFUOĞLU
Izmir, December 2017

# TEXT OF OATH

I declare and honestly confirm that my study, titled "SPEECH RECOGNITION ALGORITHM IMPLEMENTATION FOR REMOTE CONTROLLING UNMANNED AERIAL VEHICLES (UAVS)" and presented as a Master's Thesis, has been written without applying to any assistance inconsistent with scientific ethics and traditions. I declare, to the best of my knowledge and belief, that all content and ideas drawn directly or indirectly from external sources are indicated in the text and listed in the list of references.

İlhan SOFUOĞLU  
Signature  
…………………………………..  
December 26,2017

# TABLE OF CONTENTS

# LIST OF FIGURES

## SYMBOLS AND ABBREVIATIONS

ABBREVIATIONS:

| | |
|---|---|
| HMM | Hidden Markov Model |
| GMM | Gaussian Mixture Model |
| ANN | Artificial Neural Network |
| DTW | Dynamic Time Warping |
| MFCC | Mel-Frequency Cepstral Coefficients |
| LPC | Linear Predictive Coding |
| DFT | Discrete Fourier Transform |
| DCT | Discrete Cosine Transform |
| FFT | Fast Fourier Transform |
| UAV | Unmanned Aerial Vehicles |

# CHAPTER 1
# INTRODUCTION

## 1.1 Brief History of Speech Recognition

First studies in the field of speech were started in the 1930s. In 1936, AT & Bell Labs developed the first speech synthesizer which is called Voder, but the development of speech recognition systems using computers has been in the 1950s and 60s. The system which is called Audrey, developed in 1952 at Bell Laboratories, it only recognizes digits. Ten years later, in 1962, IBM's Shoebox machine was able to understand 16 English words. In the 1970s, speech recognition gained momentum with the help of United States Department of Defense (DARPA) studies in this field. From 1971 to 1976, “Harpy” system, developed by Carnegie Mellon University, was able to recognize 1011 words. In the 1980s, a different statistical algorithm such as the Hidden Markov model began to be used, allowing voice recognition systems to recognize more words. The possibilities of voices have been used instead of sound patterns or templets to recognize the terms. In 1990, Dragon launched the first consumer speech recognition product. This system knows how to talk continuously, and after 45 minutes of training, it can recognize 100 words a minute. In 2010, Google launched a voice search app for Google Android, aiming to develop a more accurate conversation model with users' speech data. Today, the companies have developed systems such as Google Now, Microsoft Cortana, Apple Siri, and Amazon Alexa, we can not only control our computers but also use them to manage all our mobile devices.

## 1.2 Unmanned Aerial Vehicles

In recent years, people have been working and encountering more often with computers, robots, and next-generation vehicles. Drones are the top of this list which is also known as unmanned aerial vehicles. As with many other technologies, the vehicles have been developed exclusively for the military industry and has served this industry for many years. These vehicles, which were used primarily for observation purposes before, then they were started to use for assault purposes. Thus these vehicles were widely known.

The most common use of unmanned aerial vehicles is in the military industry, but in recent years the commercial and individual use of drones has become increasingly widespread. Unlike military drones, the commercial drones are made with cheaper materials. Therefore, the companies can produce lightweight cheap and portable drones. This also shows that variety of use of these vehicles that are used will increase rapidly.

These unmanned aerial vehicles are usually managed by remote controllers, but developers can control these vehicles in many ways in their hobby projects or academic studies; by using gesture, image, sound recognition, etc.

This study focuses on remote controlling unmanned aerial vehicles by speech recognition algorithm implementation. The second chapter focuses on general information about speech and steps in speech recognition such as speech pre-processing feature extraction and acoustic models. The third chapter focuses on literature review and applications using speech recognition. The fourth chapter focuses on the implementation of the system. The fifth chapter contains a conclusion and future work.

# CHAPTER 2
# SPEECH RECOGNITION

## 2.1. BACKGROUND

### 2.1.1. Definition of Sound

The sound is called the pressure waves that the objects vibrate. These pressure waves can propagate in a gas, liquid or solid environment. As the vibrations created by these waves in the eardrum are processed by the brain, the process of hearing occurs. A sound wave can be described as a sine wave. These are some of the main features of sound waves.

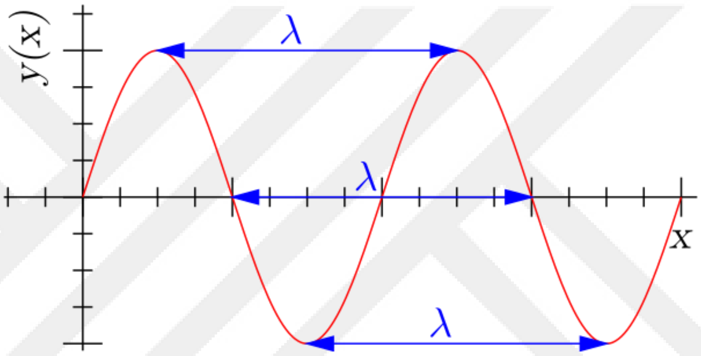


***Figure 2-1*** *SoundWave (Wikipedia Inc., 2017)*

Frequency: The frequency of the observed wave peak/pit in one second. Because of this definition, it is inversely proportional to the wavelength. The frequency unit is Hz. The human ear can detect sounds in the 20 Hz to 20000 Hz frequency range. In the human ear, high-frequency sounds represent a thinner sound, while low-frequency sounds represent a thicker sound.

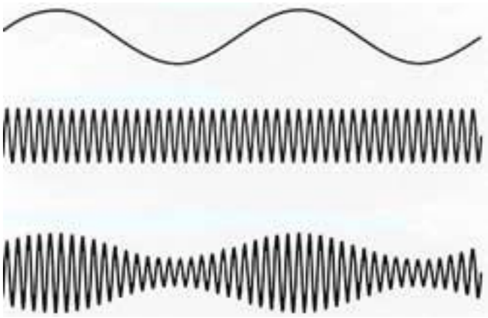


***Figure 2-2*** *Waves with Different Frequency (Sesin genligi ve frekansı, 2017)*

Amplitude: Amplitude represents intensity in the sound wave. The amplitude of the sound is related to the energy of the sound wave. The unit is dB. A high amplitude sound in the human ear represents a higher sound, while a low amplitude sound represents a lower sound.

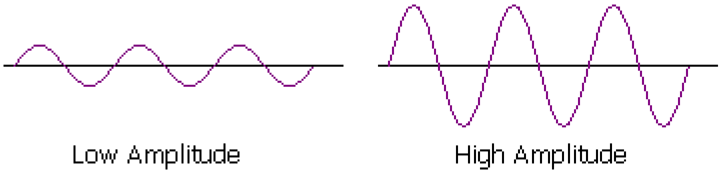


***Figure 2-3*** *Waves Different Amplitude (Introduction to Music Production, 2017)*

### 2.1.2. Definition of Speech

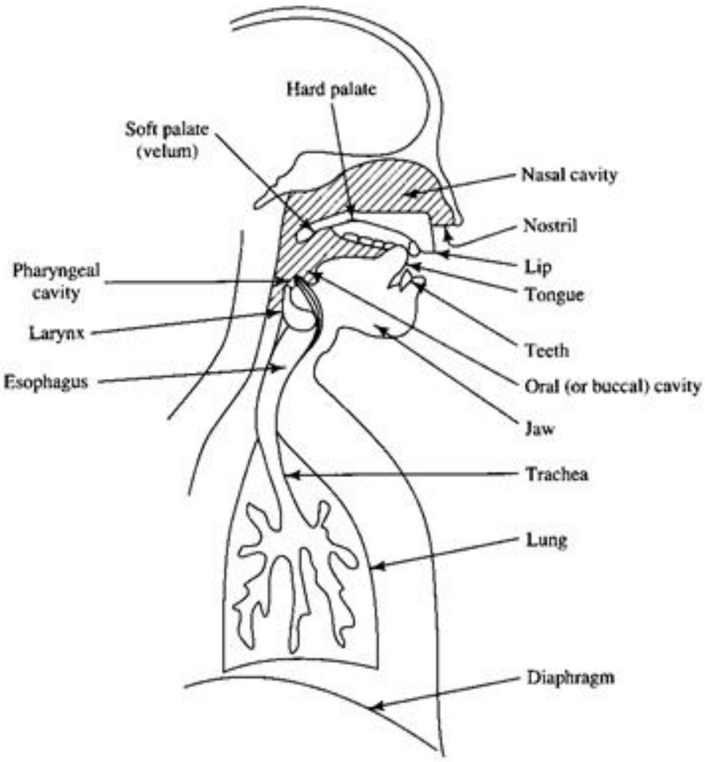


***Figure 2-4*** *Human Vocal System (Michigan State University, 2017)*

Speech is a bodily reaction based communication system in which traditional symbols are used for connection and relationship of people. Speech symbols are in the conversation. These symbols are the result of air molecules passing through the sound path of air held in the lungs.

Vocal cords are two structures, right and left, at the top of the pale. As they breathe away from each other and open the airway; they close and close each other during swallowing and sounding. As the two vocal cords approach each other and air passes through them, vibrations occur, periodically cutting and opening the air flow. Thus, a certain frequency sound is produced. This is called the fundamental frequency. In males, the basic frequency is lower than females and children, which is why the male voice is darker. The average is 120 Hz. (105-140 Hz). The basic frequency values of women and children are high. Women average is 210 Hz. (180-240 Hz). In children speech, this value can be higher than 240 Hz depending on age and sex.

### 2.1.3. Perception of Sound

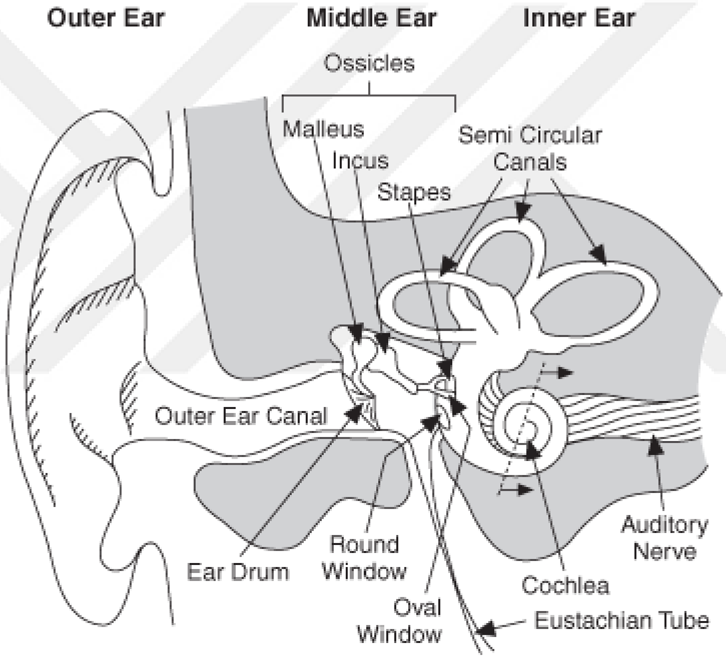


***Figure 2-5*** *Human Hearing System (Human Perception of Sound , 2017)*

The ear structure contains three parts; outer ear, middle ear, inner ear. The outer ear is the part of the ear that looks outward. The outer ear is wide and curved also it is flexible since it is made from cartilage. Ear folds help to sum up the sound waves and find the direction of the sound comes. The lower part of the ear bucket ends with an earlobe filled with oil. The middle ear is attached to the outer ear with eardrum and the with inner ear with an oval window. In the middle ear, there are earbuds, malleus-incus-stapes bones and the inside of the middle ear is filled with air. In the inner ear, there is an oval window, cochlea, semicircular canals, hearing sensation cells, hearing nerves, balance-related nerves and the inner ear is filled with liquid.

Sound waves are gathered by the earbuds and transmitted to the ear canal and vibrates the eardrum. When the eardrum vibrates, the malleus-incus-stapes bones here also vibrate, and these bones increase the intensity of the sound vibrations and give these vibrations to the oval window. The sound vibrations in the oval window are transmitted to the cochlea. Sound vibrations from Cochlea are taken up by the hearing sensory cells in the cortical organ and are transferred to the auditory nerves and carried to the hearing center of the brain with the help of the nerves. The stimuli are evaluated by the brain, and the hearing occurs.(Junqua & Haton, 1996)

## 2.2. SPEECH PREPROCESSING

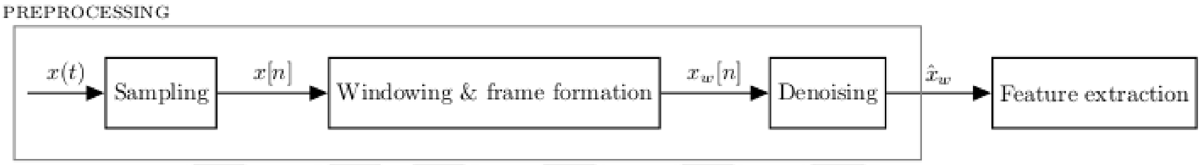


***Figure 2-6*** *Speech Pre-processing (Preprocessing, 2017)*

The speech preprocessing is an important step in speech recognition systems which directly affects on feature extraction of the sound signal is generally used before acoustic model training. Parameters changes or different algorithms in noise elimination, sampling rate selection, windowing and frame size, voice activity detection techniques have notable effects on feature extraction of the speech signal. Thus, these preprocessing operations are commonly used before feature extraction and acoustic model training

### 2.2.1. Sampling

Sampling is the process of converting a continuous analog signal to a digital signal by recording at certain intervals. The frequency of recording the analog signal indicates the sampling frequency.According to Claude E. Shannon to reproduce the signal losslessly into the analog world, the sampling rate must be twice as fast as the rate of change of the signal.This is also called Nyquist rate. Telephone quality speech is sampled at 8 KHz by using 8-bit data. Speech recognition systems generally use 10-16 KHz, 12-16 bit sample rate. (Sampling, 2017)

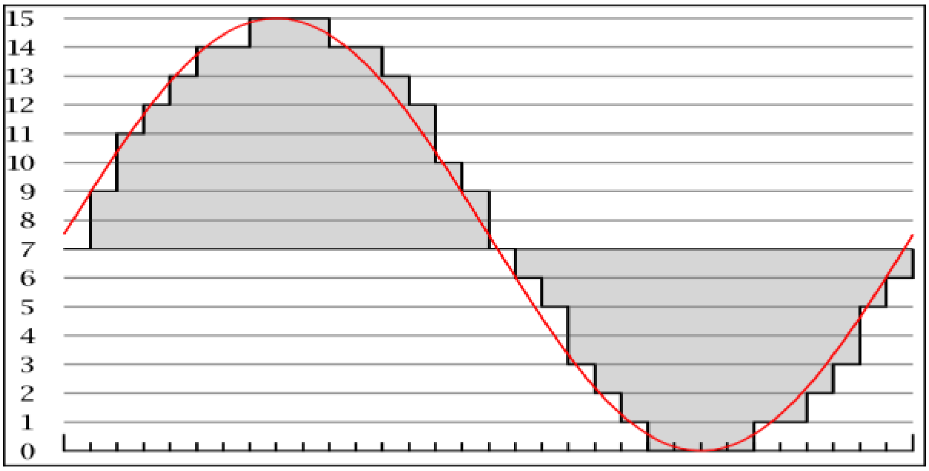
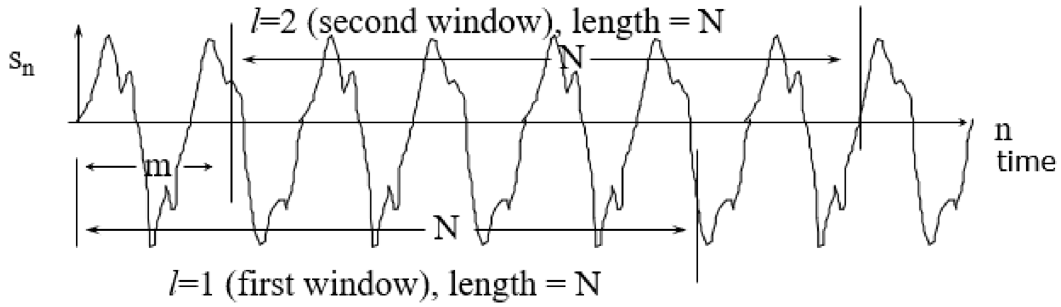


***Figure 2-7*** *Sampling (Macquarie University, 2017)*

### 2.2.2. Framing and Windowing

Audio framing and windowing have an important role for feature extraction of speech signals. The speech signal is a continuous signal over time. The windowing method is applied because it is difficult to work on a non-stationary signal. By applying this method, the frequency and spectral components of the signal that change over time is fixed. Before the signal is processed, fragments containing a certain number of samples are separated. Each of these parts is called frames. Since human ear cannot respond to the very fast change of speech data content, the speech signal is normally cut into frames before the analysis. In most studies frame size is between 10ms and 30ms. Frames can be overlapped, normally the overlapping region ranges from 0 to 75% of the frame size. The audio signal is separated adjacent frames according to the number of samples. After the framing operation, windowing functions are used to windowing speech signal.



$l$=2 (second window), length = N
N
$s_n$
n
time
m
N
$l$=1 (first window), length = N

***Figure 2-8*** *Windowing (Al-Sabbagh, 2017)*

Window functions are functions that multiply signal parts. With these functions, it is ensured that the middle parts of the signal parts are emphasized. There are many types of windowing techniques such as Rectangular, Hamming, Hanning, Blackman, Barlett, Kaiser.(Podder, Khan, Khan, & Rahman, 2014) One of the most commonly used soft windowing technique is Hamming Windowing. The Hamming window is described by

$$w(n) = 0.54 - 0.46\,cos\left(\frac{2\pi n}{M}\right)$$

(1) (Labbookpages, 2017)

### 2.2.3. Denoising

Speech signals contain many unwanted signals and noises due to the microphone, electrical noises, and environmental noise. These unwanted signals in speech cause speech recognition to become difficult. .According to their frequency, there are many noise; white noise, band-limited noise narrow band noise, colored noise.(Noise Distortions, 2000) Filtering techniques and speech enhancement methods are often used to obtain clean speech signals. Adaptive Weiner Filtering, Adaptive Line Enhancer, Multiple microphone usage, spectral subtraction, Gamma Tone Filter are some of these techniques.(Garg et al., 2016)

Adaptive Weiner Filter: In this method, two microphones should be used. One of these is for noise, one of these is for speech. After the record finish, noise elimination operation occurs.

Adaptive Line Enhancer: This method discriminate pitch location of the speech signal and noise signal seems like an error. Therefore, this method can separate speech and noise signal.

Multiple Microphone Usage: This method can be implemented in two ways. First, directly close to the speaker and other noises are eliminated according to spectral characteristics. Another way is microphone separate noise spectral characteristics.

Spectral Subtraction: This method estimates noise signal features when there is no speech signal. After that noise signal can be subtracted from noisy speech signals.

Gamma Tone Filter: This method works like human auditory system. This method use bank of the overlapped bandpass filter.

## 2.3. FEATURES OF SPEECH

### 2.3.1. Linear Predictive Coding Coefficient

The basic idea in Linear Predictive Coding (LPC) is to estimate a sound sample using its linear combination of previous sound samples. The error between the actual sound samples and the predicted samples is minimized to obtain parameter values consisting of prediction coefficients. Unlike other models that Linear Predictive Coding Coefficients are modeled the structure of human vocal structure, it adopts a single-inlet and single-outlet filter model rather than a filter group.

### 2.3.2. Mel -Frequency Cepstrum Coefficients

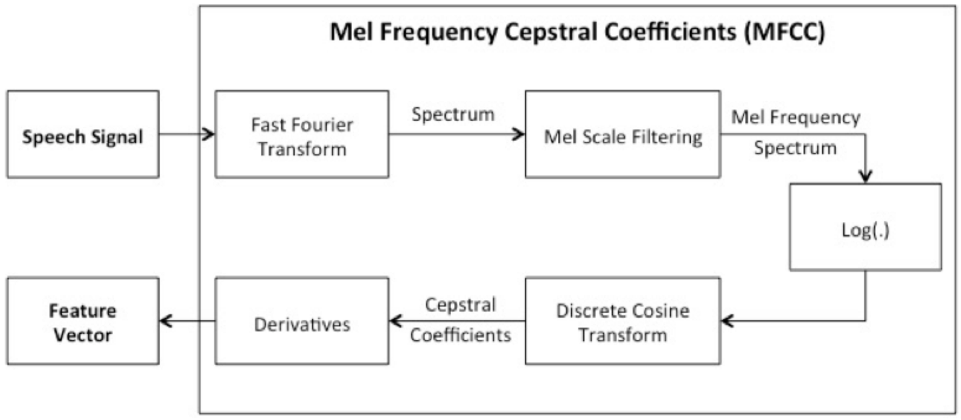


***Figure 2-9*** *MFCC Process (MFCC, 2017)*

MFC (Mel-frequency cepstrum), simulates the behavior of the human ear and uses an FFT-based digital analysis technique. As it seems in Figure, to obtain MFCC feature vectors five essential steps are required. MFC analysis the numbers are called MFC coefficients (MFCC).

The Mel unit is a unit that is subjectively designed to mimic the human ear. In other words, the Mel unit is developed not according to a linear frequency axis but according to how the human ear perceives it. The non-linear series formed by this unit is called the Mel scale. The transformation between the Mel and the frequency gauge is provided by the equality given below.

$$\mathrm{Mel}(f) = 2595 \log\left(1 + \frac{f}{700}\right)$$

(2) (Scientific Research Publishing, 2017)

Accordingly, the Mel-scale is linear for frequencies lower than 1000 Hz, higher than 1000 Hz for frequencies, logarithmic values will be displayed in this perceptual spectrum. One way of implementing is to design filters that will display the distribution according to the Mel-scale. Mel-frequency is triangle and bandpass

## 2.4. ACOUSTIC MODELS

### 2.4.1. Hidden Markov Models

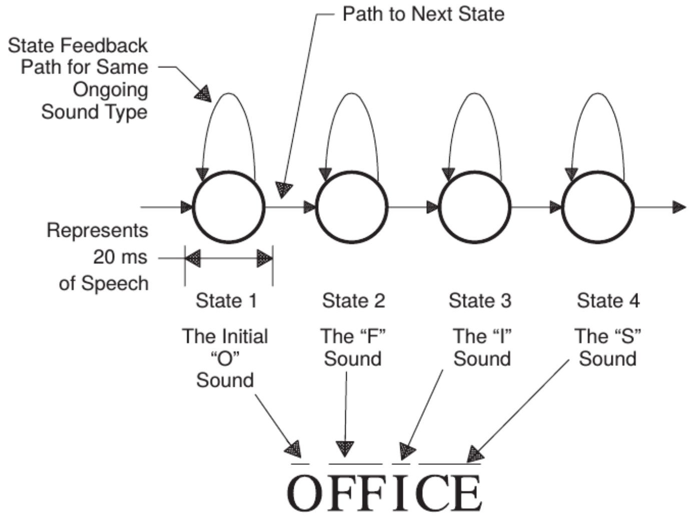


***Figure 2-10*** *Hidden Markov Model (Language Model, 2017)*

The Markov Chain is a stochastic process. In Markov models, future states are independent of past states. Future states can be reached according to the information and probabilistic value of current states. Each state can go another state or remain the same state according to their transition probability values between states. One of the examples of Markov chain is a simple random walk. In simple random walk probability of going from current corner to another neighboring or from neighbor corner to current corners is always the same and equal to each other. (Markov Chain, 2017)

In the Normal Markov Model, states are visible to the observer, and so a single parameter, state transition, can be seen, and so the only parameter is state transition probabilities. In the Hidden Markov Model, the situation is directly related to the possibilities. Moreover, in the Hidden Markov Model, the state is not directly visible, but the state-dependent outputs are visible. However, state-dependent outputs are visible. The probability distribution on each state is the probability distribution over the possible exit signs. So, the series of signs is about the order of the situations. So, the series of markers are produced by HMM, which gives us some information about the order of the states. HMM is especially useful for the recognition of sounds, handwriting, etc. (Rabiner, 1989)

### 2.4.2. Gaussian Mixture Models

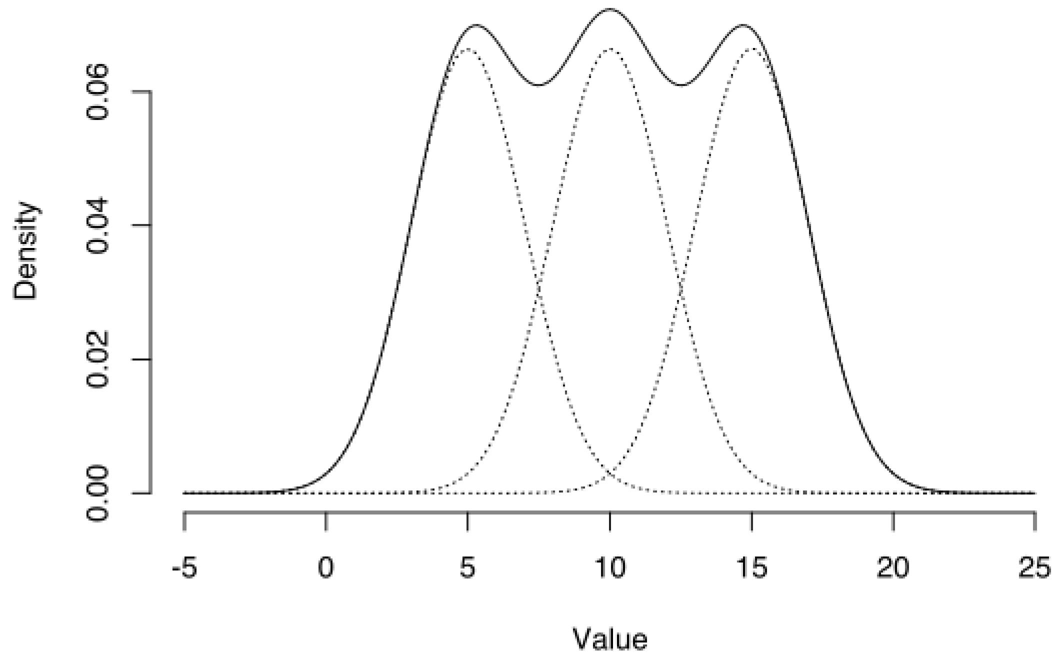


***Figure 2-11*** *1- D Gaussian Mixture Model (GMM, 2017)*

A Gaussian mixture model (GMM) is a probabilistic model which contains multiple Gaussian distributions that are obtained from feature vectors. The parameters for Gaussian mixture models are getting from either from maximum a posteriori estimation or an iterative expectation-maximization algorithm from a prior model which is well trained. Gaussian mixture models are very valuable when it comes to modeling data, especially data which comes from several groups. Mixing collects the results from each Gaussian in the mixture, that is passed each data point to all distributions in the mixture one by one, and to sum the results and weight.

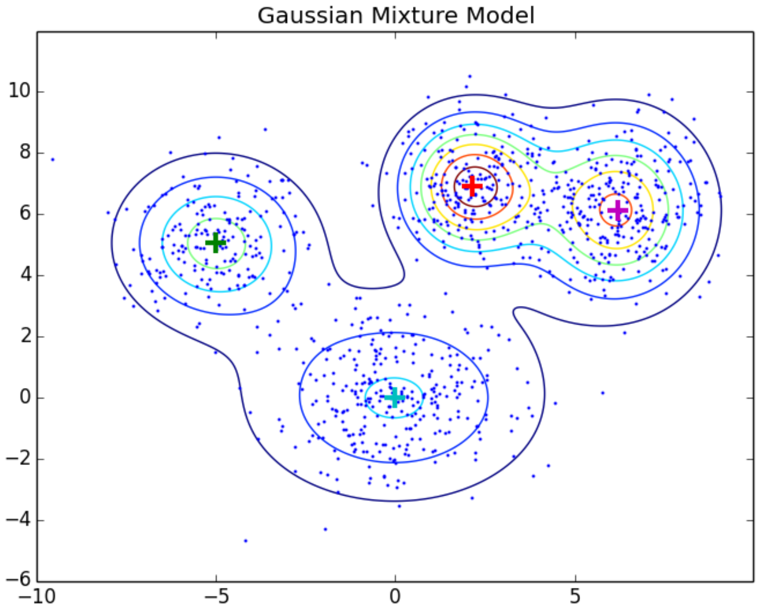


***Figure 2-12*** *2D Gaussian Mixture Models (Gaussian Mixture Models, 2017)*

Gaussian Mixture Models are also a powerful statistical algorithm for clustering. GMMs are generally used in biometric systems such as speech recognition systems. GMM is a very useful technique to measure and cluster spectral features of human vocal systems.

### 2.4.3. Dynamic Time Warping

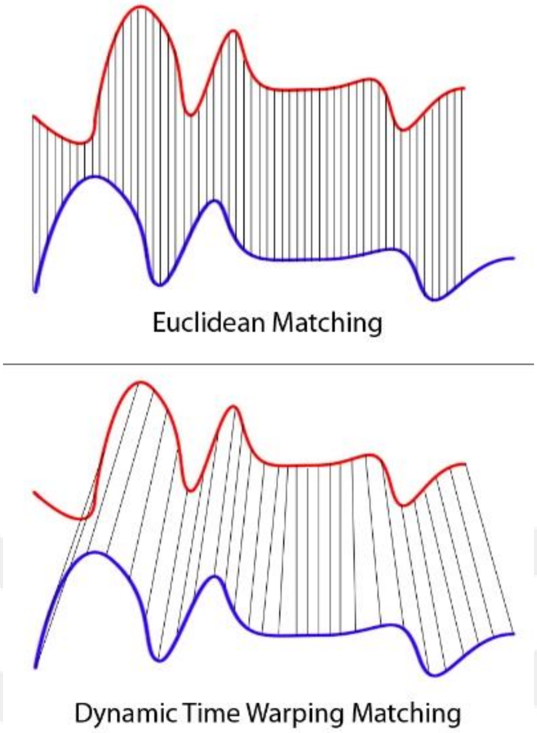


***Figure 2-13*** *Dynamic Time Warping Matching (Wikiwand, 2017)*

Dynamic time warping (DTW) algorithm is a mapping method used in the similarity measurement of time series. The method has different uses. For example, this method can be applied when comparing feature vectors in a speech recognition system. Time series can be thought of as number arrays of elements whose values are time-dependent. For example, a sampled audio signal over a certain time period forms a time series. It is not possible to match such sequences with strict time-dependent methods.

Similarity measurement can be done by finding the sum of the Euclidean distances between the elements of two series. The closer this sum is to the zero, the more similar the series are to each other.

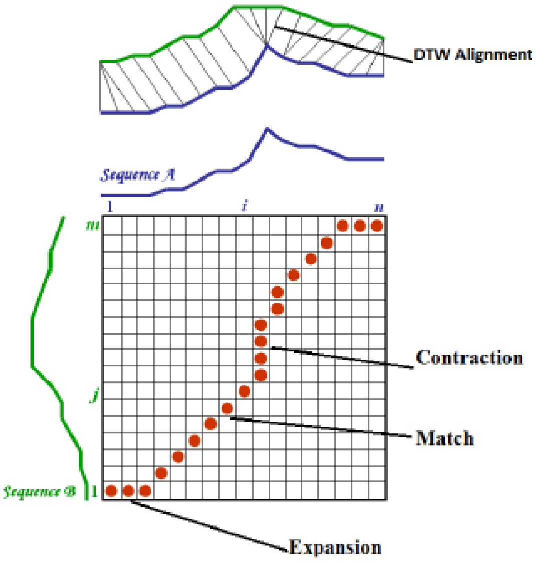


***Figure 2-14*** *DTW Warp Function (kNN with DTW, 2017)*

Dynamic Time Warping method aims at solving the problem of time dependency mentioned above. By matching the nearest elements of vectors to each other, it is ensured that the sum of distances can be kept at the smallest possible value. With this method, similarity measurements can be made for vectors of different sizes.

The cumulative Euclidean distances are first calculated when the algorithm is poured into the code. The Euclidean distance between two values in one dimension is found by taking the absolute value of the square root of the difference squared, that is, in fact, the difference between these two values. The cumulative distance is the sum of the distance to the equals of an element and the distances to the mappings of all the elements that match it. When the cumulative distance is calculated, the calculated Euclidean distance between each pair of elements is added to the smallest cumulative distances computed from these elements for a unit of backward element pairs. Thus, when the last matching elements are reached, the calculated cumulative distance is the smallest possible value of the distance. (Brown & Rabiner, 1982)

### 2.4.4. Artificial Neural Networks

ANN is a logical software developed to imitate the working mechanism of the human brain and to perform basic functions such as brain learning, recall generalization, and deriving new information. ANNs are synthetic constructs that mimic biological neural networks.

Inputs: Element which is entered the neuron from the environment or other layers.

Weights: The information is based on the weights on the links they enter the function element. Therefore, the weights determine the effect on the processing element.

Addition Function: A calculator that calculates the net input from a cell and usually net input, the corresponding weighted products of inputs the total.

Activation Function: Activation, also known as transfer function is a function of the net return from the aggregate function which determines the cell output and is usually a non-linear function.

Input Layer: The nerve cells in the input layer biological nerve networks, in artificial neural networks, they are called artificial neurons, and cells are also called processing elements. Operation in this layer they transfer information from the outside of the system to the intermediate layers

Intermediate Layers (Hidden Layers): Search input layer as the information from the layer is processed and sent to the output layer. Processed information is carried out in intermediate layers. If there is more than one, it can be an intermediate layer.

Output Layer: The processing elements in the layer processing, produces for the input set presented by the network input layer they produce the required output. The output produced is sent to the outside system.

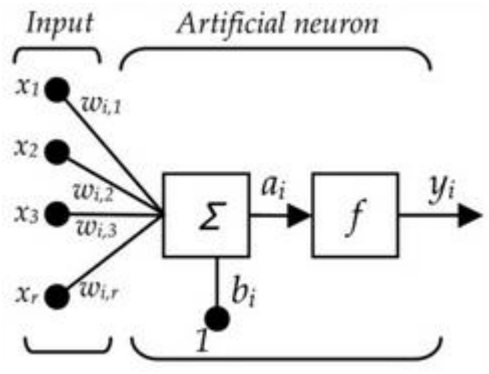


***Figure 2-15** Artificial Neuron (45 Questions to test a data scientist on basics of Deep Learning (along with solution), 2017)*

Feedforward Neural Networks: Feedforward Neural Networks are neural network model which have one-way information flow. In the network model, the information received from the input layer is transmitted to the hidden layer. The output value is determined by processing the information from the hidden and output layers.

Back Propagation Neural Networks: A feedback network is a network structure in which the outputs at the output and intermediate layers are fed back to the input units or previous intermediate layers. Thus, inputs are passed both forward and backward. There are dynamic memories of this kind of ANN, and one output reflects both that input and the previous input. They are therefore particularly suited for predicting applications. Feedback networks have been quite successful in predicting various types of time series. For instance, these networks are Hopfield, SOM (Self Organizing Map), Elman and Jordan networks.

The activation function provides the curvilinear relationship between the input and output units (layers). Correct selection of the activation function influences network performance significantly. The activation function is usually selectable as unipolar (0 1), bipolar (-1 +1) and linear. It is the component that enables the network to learn to be nonlinear.

# CHAPTER 3
# LITERATURE REVIEW

In 1952, Bell Lab began to develop speech recognition, which was basically started with a small vocabulary recognition effort. In the 1960s, isolated word recognition was implemented in a simple way with algorithms that began to develop. The methods used for voice recognition such as LPC, DTW, HMM, MFCC, LM, Neural Networks, Kernel Based Classifiers and Dynamic Bayesian Networks contributed to speech recognition in different ways.(O'Shaughnessy, 2008) These algorithms are generally combined with hardware systems such as robotic systems or mobile vehicles. One of the similar studies like this thesis that manages the voice command and the unmanned aerial vehicle is the study of (Bernadin, Patel, & Smith, 2015) in this study, the accuracy of phoneme recognition in the software model using embedded voice recognition system is estimated to be about 40%. One of the main purposes of speech recognition is communication with computers and robotic devices. In this study, similar speech recognition algorithms were used to direct the HOAP-2 humanoid robot. MFCC was used as the feature vector, and backpropagation neural network was used as the classification method. There are 15 examples for each command with 5 English commands for the system. 10 of these samples were used for the training and 5 were used for the test. The recognition rate of 93% was observed. (Joshi, Kumar, Chakraborty, & Kala, 2016)Today smart home systems are the systems that speech recognition systems are used extensively. There is a study that HMM and GMM algorithms were used together, the speech signal was converted into text and signals to the control module to perform operations in the smart house, and this study had been reached 93% accuracy.(Zhang et al., 2016)

The DTW algorithm is one of the most commonly used dynamic programming technique in speech recognition. DTW algorithm implemented in different languages. According to this study (Shinde, 2012) a speech database consisting of 21 speakers and 105 samples in the Marathi language was used on the basis of LPC coefficient and DTW algorithm. Speaker tried to develop an independent system. A comparison was made on the 3-word DTW scores of the 5 speakers. Beside, DTW was implemented in this study with combination of LPCC, speech recognition algorithm was developed for robot direction. The LPC coefficients are used as the feature vector. Dynamic Time Warping algorithm is used as the pattern matching algorithm. 90% correct recognition rate is reached as a result of this experiment. (Zhizeng & Jingbing, 2004) Similar studies with high accuracy cause new studies in new branches. Another use of speech recognition is the field of medicine. A speech recognition algorithm has been developed on a laparoscopic robot arm according to a study performed in this field. DTW technique is used as a the pattern matching algorithm. (Sauer, Kozłowski, Pazderski, Waliszewski, & Jeziorek, 2005)

The LPCC feature vectors are always compared in many studies with other feature vectors by using different classifier algorithms. In this study, MLP and SOM algorithms were used as classifier and LPCC, MFCC were compared. In result, it was examined both LPCC and MFCC has advantages and disadvantages. Although LPCC has 94.25 % accuracy MFCC has 89.14 % accuracy in clean speech. In the noisy speech, MFCC vectors provide more recognition accuracy. (U. Bhattacharjee, 2013) In another study, MFCC and spectrum based feature methods are compared according to the study done. It has been observed that the proposed method gives better results than the studies made with the MFCC vector. (Liu, Ge, Jiang, & Goh, 2016) In a study that was implemented the speech algorithm on UAVs, 4 commands were also implemented using MFCC vectors. It has been observed that this feature extraction method has achieved successful results.(Lavrynenko, Konakhovych, & Bakhtiiarov, 2016) In contrast to LPCC, MFCC gives better accuracy results in literature. According to Bangla language based study, where 20 speakers were used, 98% accuracy was achieved in a noiseless environment and an accuracy of 93% was achieved in a noisy environment. MFCC feature extraction technique and Vector Quantization feature matching technique were used.(A. Bhattacharjee et al., 2017) Using different algorithms in studies can affect accuracy. According to this study, a system has been developed in MATLAB environment using MFCC vectors and HMM algorithm. According to this study, in this study, the accuracy of the study is 40% and 50% in a noisy environment and 70% and 80% in a noiseless environment. (Mohamad, Jamaludin, & Isa, 2017)

The speech pre-processing steps performed before the pattern matching a step in the speech recognition increase the accuracy of the speech recognition. Noise reduction technology plays a very important role. In this comparative study(Garg & Jain, 2016) it was found that Adaptive Wiener filter applied various filters and Gamma tone filter at low decibels gave better results at high decibel values. For noisy environments, different algorithms can be used for better recognition rate. In this study, which is carried out in the Vietnamese language, a wavelet-based algorithm is proposed which gives a better recognition rate for a robot arm in noisy environments. Noise-free speech recognition is better than MFCC-based recognition, while noise reduction is better. (Nghia & Vinh, 2008) Moreover, different noise reduction filters can be used for speech enhancement process. According to this study, the noise reduction filters used before the automatic speech recognition algorithm were compared to increase the system accuracy. Spectral subtraction, Wiener filtering, MMSE, logMMSE Wiener-TSNR e Wiener-HRNR filters are compared. It has been observed that the filters exhibit different characteristics according to the noise level in the environment.(Prodeus, 2015)

Zero Crossing Rate technique is used to increase accuracy rate by eliminating speechless data. In this comparative study (Gamit & Dhameliya, 2015) after a series of preprocessing of the audio signal, the patched matching operation was tried to be performed with the aid of artificial neural networks. ZCR technique was used to remove the silent parts of the speech signal. According to another study, a speech recognition system has been developed to move a robot arm using Adaptive Input Neural Network in Microsoft Windows 3.1 platform. This system is based on the Learning Vector Quantization artificial neural network developed by Kohonen. Short-Term Energy, Zero Crossing Rate Linear Predictive Coefficient and LPC Cepstrum parameters are used in this study. The user-dependent accuracy rate was 82% while the independent performance was measured as 64%. (Zhou, 1994)

ANNs and HMMs are also the commonly used acoustic models. In this study in the year 2013(Md, Md, Kumar Prodhan, & Md, 2013) , a study was performed by a large number of speakers of 10 digits in the Bangla language. An accuracy of 92% was achieved in the study using the recycled artificial neural network. Sphinx which is HMM-based speech recognition tool is very useful in different platform and different branches. The Java-based Sphinx-4 library was used for this study. Using the HMM algorithm, a high accuracy of 95% is achieved. (Watanabe, Izumi, Ohshima, & Ishii, 2006) In this study HMM-based, CMU Sphinx software was used on the endoscope robot. The most convenient way to use an endoscope is to use limbs, which is the result of this. (Zinchenko, Chien-Yu Wu, & Kai-Tai Song, 2017) HMM, tools can be used with different speech recognition tools. According to this study, it is revealed that a robot system can distinguish simultaneous speech using different multiple acoustic models. In this system, HMM-based HTK toolkit and Japanese automatic speech recognition software Julian are used.(Nakadai, Okuno, & Kitano, 2003) Furthermore, HMM give better results when compared with other algorithms. The accuracy of the system was examined that the recognition rate of a previously developed HMM-based system using Kalman filter increased. Compared with DTW and Crossing Zero rate, the HMM-based model shows better results.(Hamza, Fezari, & Bedda, 2009) A speech recognition system consisting of Arabic Spotted Words has been developed for the robot arm. In this system consisting of 12 words, Hidden Markov model-based algorithm is used. According to the results obtained, the rate of recognition of a word is at least 76%.(Fezari, Hamza, & Bedda, 2008)

Different techniques, algorithms and hardware systems can be used in speech recognition. In this study, PCA (Principal Component Analysis) and FFT based speech recognition algorithm are proposed to guide a robot in the Thai language. According to this algorithm, the correct recognition rate is 67% in males and 71% in females. (Phanprasit, 2014) In another comparative study, three algorithms were compared. HMM, Feature Weight DTW and the proposed algorithm, the performance of the system is changed. (Lei, Gan, Jiang, & Dong, 2014) According to this study, a clustering algorithm was used to classify word (Patil, Shirbahadurkar, & Paithane, 2017) used KNN classifier as feature matching in MFCC technique. This study in the Hindi language has reached 94% accuracy. According to this study, instead of using HMM or artificial neural networks, a system was developed using embedded DSP platforms. (Kuljić, János, & Tibor, 2007)

Speech recognition systems can be combined with another recognition systems such as gesture recognition. In this study, mouth gesture recognition and speech recognition algorithms are used together. For the mouth gesture, HMM was used for the GMM voice recognition algorithm. In this study, a fast system which is returning at a high level was tried to be developed.(Ǵomez, Ceballos, Prieto, & Redarce, 2009) The system which was developed on UAVs can recognize gesture and voice. In this study, it is observed that the two methods compared to the recognition time and accuracy of voice recognition based system have lower success rate than gesture recognition based. (Yao, Xiaoling, Zhiyuan, Huifen, & Pengcheng, 2017) Another study is about the combination of image and voice recognition. In a study conducted in 2004, image and voice recognition were used together. ViaVoice developed by IBM was used as voice recognition software. The color words recognized by this software are then routed to a robot arm that follows the recognized colors after image processing.(Izumi, Tamano, & Nose, 2004)

The systems which used powerful application or services were developed on UAVs. These systems succeed high accuracy rates. According to this study for handicapped people, AMR Voice application was tried to simulate wheelchair. With only 4 voice commands, 95% accuracy is achieved. (Upase, 2016) In another study, the voice-based drone controlling system was developed. 7 voice commands were used to direct the drone to reach the 80% accuracy at least.(Fayjie, Ramezani, Oualid, & Lee, 2017) Some systems were used cloud services to recognize speech signals. A system which controls the robot with voice commands are receiving voice signals translated by online cloud server, and the robot is routed by sending signals to the robot via Bluetooth. This study has developed object tracking robot with voice commands.(Kumar, Nandan, Mishra, Kumar, & Mittal, 2016) In another study, the system benefits the powerful aspects of Google Speech API. According to this study, researchers have developed such a system using the Google Speech API, because the API is better than platform flexibility, semantically better than other systems, faster development is more reliable, higher performance.(Stefanovic, Cetic, Kovacevic, Kovacevic, & Jankovic, 2012)

# CHAPTER 4
# METHODS AND RESULTS

## 4.1. DEVELOPMENT ENVIRONMENTS

### 4.1.1. Bebop Drone

***Figure 4-1*** *Parrot Bebop Drone (Parrot Bebop Drone, 2017)*

In this study, Parrot Bebop Drone was used as an unmanned aerial vehicle. The reason for choosing this drone is that it is cheaper and easier to access. The drone weighs 400 g and is 8 times stronger than the previous generation. MIMO dual-band with 2 double-sets of dipole antennas for 2.4 and 5 GHz is used as the Wi-Fi antenna for connection, and the maximum signal range is 250 meters. It has 4 propeller drone and magnesium and abs body. Its maximum speed is 13m / s. It has a full HD camera with 1080p resolution. Drone batteries are lithium polymer. It has the Parrot drone

Linux SDK and can be guided through the app by mobile devices. During the experiments, the commands were sent from PC to drone via Wi-Fi.

### 4.1.2. Google Speech API

The reason for using the Google Speech API in this system is that it has a higher accuracy rate than other acoustic model algorithms and does not require extra noise filtering in noisy environments.

Google Speech API is speech recognition cloud service which is using deep neural networks. The Speech API service is charged monthly. 0 to 60 minutes free, after 60 minutes 15 seconds is $ 0.006. The Speech API supports over 80 languages. Since the Speech API is based on deep neural networks, its stability is increasing day by day. API does not need any noise elimination process.(Google Inc., 2017)

### 4.1.3. Katarina Framework

Katarina Framework is a python library which is written and published by Martin Dlouhy in 2015. In this study, Katarina Framework was used to control Parrot Bebop Drone according to voice commands. The framework contains bebop.py, commands.py, navdata.py, video.py, play.py, capdet.py, apyros folder, demo.py files and behaviors folder.

bebop.py is the main file that should be imported for each implementation. The Bebop class includes ports and IP addresses to connect with the drone. Functions that update the data sent to the drone and update the data are in this class. Main activities of the drone functions that allow the forward movement, the takeoff, landing, and the suddenly landed for emergency situations in this class. Also, image capture, photo taking functions available in this class.

commands.py - here you discover rundown of some fundamental ARDrone3 summons changing over name into a couple of bytes parcel.

navdata.py - parsing of route information.

video.py - giving of parts of H264 video outlines.

play.py - play video utilizing OpenCV2 (utilizes video.py as change schedule, if fundamental).

capdet.py - two shading Parrot top identification try.

apyros organizer - logging instruments, shared among a few Python driven robots.

demo.py - a mix of picture handling with flight control. (Github-Katarina, 2017)

### 4.1.4. Python

Python is a programming language that can run on many operating systems and platforms. Having a large number of libraries supported by developers makes it a powerful language. Python was used in this thesis as a language with many scientific computing libraries and because of Katarina framework is written in Python 2.7.x in the thesis is also in the Python language.

## 4.2. IMPLEMENTATION

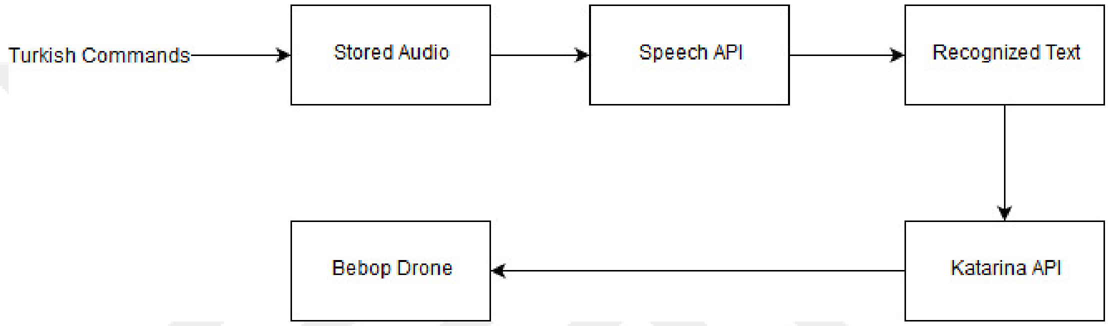


***Figure 4-2*** *Architecture of System*

Implementation of this thesis consists of two main steps. These are controlling unmanned aerial vehicle and voice recognition process. For voice recognition step, the Google speech API, which is based on a deep learning algorithm, was used. This API is a cloud-based API, so it needs to be online during experiments. For voice recognition, voice recognition must first be recorded in the computer environment. PyAudio library, which is useful Python package for audio processing, is used for this process.

```
CHUNK =1024
FORMAT =pyaudio.paInt16
RATE =16000
RECORD SECONDS = 4
WAVE_OUTPUT _FILENAME= “output.wav”
```

4 seconds of Turkish voice command 16000 KHz sample rated, 16-bit, mono channel audio signal is recorded in a computer environment as .wav file. The recorded audio file is then sent to the cloud-based API for processing with the help of web services. The Turkish audio signal processed by the Google Speech API eventually returns the alternative result to PC as text, which has high possibility to be among similar alternative voices. The result value returned by Google Speech API is expected to match the Turkish commands previously defined in the thesis.

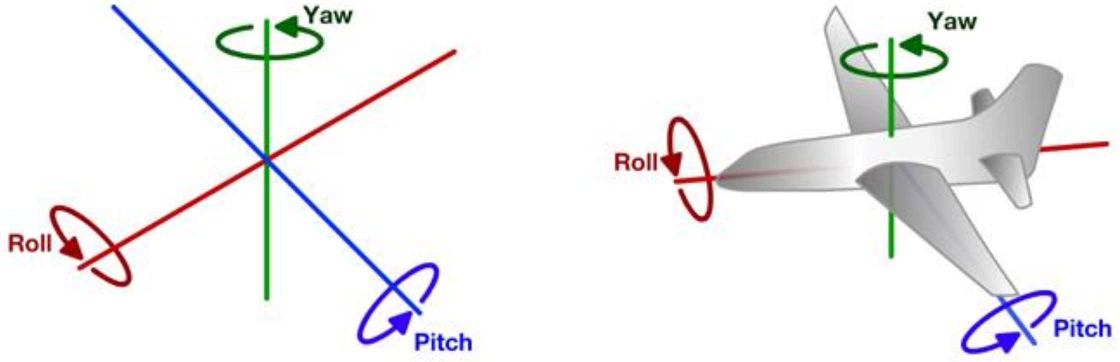


***Figure 4-3*** *UAV Movement Axis (Quadrocopter DIY, 2017)*

Most of the aerial vehicles have 3-axis movements.

Roll: Roll movement allows the vehicle to maneuver to the left in negative values and to the right in positive values.

Pitch: Pitch movement allows the vehicle to maneuver to the front in positive values and to the back in negative values.

Yaw: Yaw movement allows the vehicle to turn clockwise in positive values and counterclockwise in negative values.

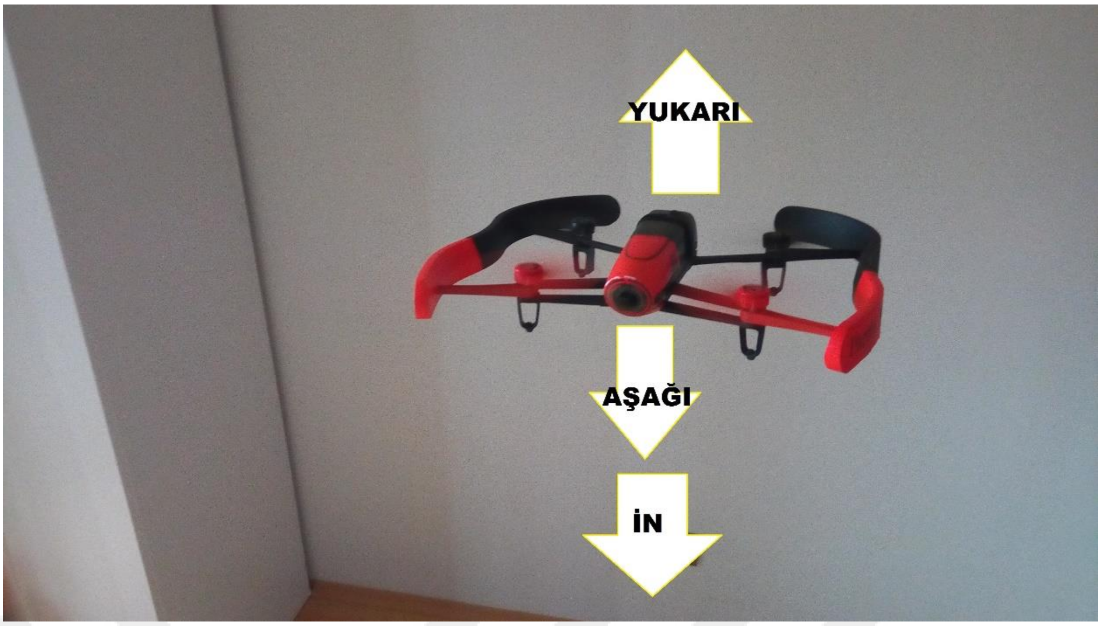


***Figure 4-4*** *Turkish Commands 1*

"YUKARI": Allows UAV to rise at a predefined value.

"AŞAĞI": Allows UAV to fall in at a predefined value.

"IN": Command that enables an unmanned aerial vehicle to land safely in the air.

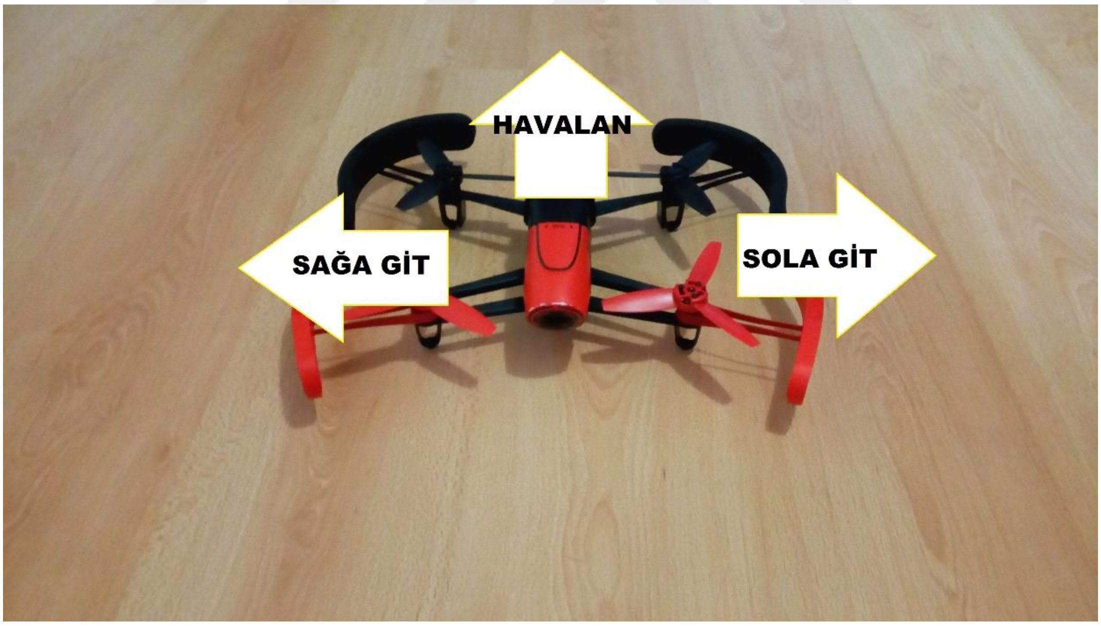


***Figure 4-5*** *Turkish Commands 2*

"HAVALAN": A command that allows the aircraft to take off at a predefined value in a fixed position on the ground.

"SAGA GIT": With a predefined positive roll value, the unmanned aerial allows the vehicle to move to the right.

"SOLA GIT" With a predefined negative roll value, the unmanned aerial allows the vehicle to travel to the left.

"SAGA DON" With a predefined positive yaw value, the UAV turns clockwise around its axis.

"SOLA DON" With a predefined negative yaw value, the UAV turns clockwise around its axis.

"DUR": is a command that allows the drone to hang in the air by resetting the yaw, roll and pitch values of the vehicle.

"ACİL": UAV in the air in the event of an emergency, allows the all engines to stop suddenly.

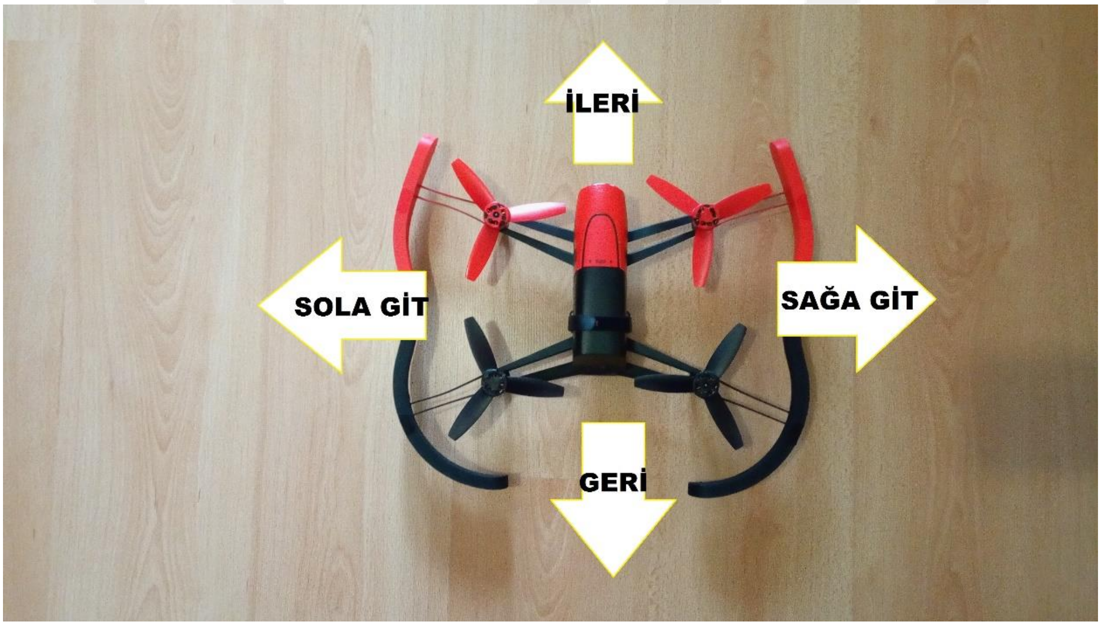


***Figure 4-6*** *Turkish Commands 3*

"İLERİ": UAV with a predefined positive pitch value allows the vehicle to move forward.

"GERİ" UAV with a predefined negative pitch value allows the vehicle to go backward.

# CHAPTER 5
# CONCLUSIONS AND FUTURE WORK

In prototype experiments, the Google Speech API was used for speech recognition and a free license Katarina Navigation API which is written using the Python programming language to navigate Parrot Bebop Drone. Some of these commands are the Turkish commands which are used for navigate Bebop Drone; “İLERİ”, “GERİ”, “SAGA GIT”, “SOLA GIT”, ”YUKARI”, ”AŞAĞI”,”İN” , “HAVALAN”,”SAĞA DÖN”,” SOLA DÖN”, ”DUR” ,“ACİL”.

Which means in English; “FORWARD”,”BACKWARD”, “TO RIGHT”,”TO LEFT”, “UPWARD”, “DOWN”,”TAKE OFF”, “LAND”, “TURN RIGHT“, “TURN LEFT“, “STOP”,”EMERGENCY”.

The speech controlled UAV system was developed and drone navigate according to Google Speech API response successfully.

Speech recognition systems have been around for almost 70 years. These systems will continue to be developed with various algorithms and techniques in various languages. There are factors that affect these systems from signal processing to attribute extraction, from the noisy environment to the language used. The use of speech recognition systems in other systems such as drone navigation will improve the speech recognition rate of these systems.

In the future, other algorithms may take the place of the statistical and artificial intelligence methods used today. For future work, communication with drones or robots can be done with more semantic speech recognition methods. Moreover, such systems can also be developed on a mobile platform.

Future work will use automatic speech recognition algorithms instead of isolated voice commands. Instead of existing acoustics model algorithms, a more sophisticated algorithm can be developed. The development of machines capable of understanding human thoughts with the development of artificial intelligence will increase the stability of the developed speech recognition systems. Deep learning algorithms, one of the machine learning methods, will make our communication with robots and computers more personalize.

Furthermore, the development of hybrid systems using a number of recognition algorithms in human-computer communication has also been shown to improve the reliability and accuracy of communication between human and computer.

# REFERENCES

Bernadin, S. L., Patel, R., & Smith, E. (2015). Work-in-progress: Evaluating the performance of voice recognition approaches for autonomous vehicular systems. *Conference Proceedings - IEEE SOUTHEASTCON*, *2015–June*(June), 0–1. https://doi.org/10.1109/SECON.2015.7132877

Bhattacharjee, A., Khan, A. I., Haider, M. Z., Fattah, S. A., Chowdhury, D., Sarkar, M., & Shahnaz, C. (2017). Bangla voice controlled robot for rescue operation in noisy environment. *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, 3284–3288. https://doi.org/10.1109/TENCON.2016.7848659

Bhattacharjee, U. (2013). A comparative study of LPCC and MFCC features for the recognition of Assamese phonemes. *International Journal of Engineering Research & Technology*, *2*(3), 1–6. https://doi.org/10.4236/jis.2012.34041

Brown, M., & Rabiner, L. (1982). Dynamic time warping for isolated word recognition based on ordered graph searching techniques. *ICASSP ’82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, *7*, 1255–1258. https://doi.org/10.1109/ICASSP.1982.1171695

Fayjie, A. R., Ramezani, A., Oualid, D., & Lee, D. J. (2017). Voice enabled smart drone control. *International Conference on Ubiquitous and Future Networks, ICUFN*, 119–121. https://doi.org/10.1109/ICUFN.2017.7993759

Fezari, M., Hamza, A., & Bedda, M. (2008). Arabie spotted words recognition system based on HMM approach to control a didactic manipulator. *AICCSA 08 - 6th IEEE/ACS International Conference on Computer Systems and Applications*, 927–928. https://doi.org/10.1109/AICCSA.2008.4493648

Gamit, M. R., & Dhameliya, K. (2015). Isolated words recognition using MFCC, LPC and neural network, 2319–2322.

Garg, K., & Jain, G. (2016). A Comparative Study of Noise Reduction Techniques for Automatic Speech Recognition Systems. *Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE*, 2098–2103. https://doi.org/10.1109/ICACCI.2016.7732361

Garg, K., Sciences, S., Delhi, N., Jain, G., Sciences, S., & Delhi, N. (2016). A Comparative Study of Noise Reduction Techniques for Automatic Speech Recognition Systems, 2098–2103.

Ǵomez, J. B., Ceballos, A., Prieto, F., & Redarce, T. (2009). Mouth gesture and voice command based robot command interface. *Proceedings - IEEE International Conference on Robotics and Automation*, 333–338. https://doi.org/10.1109/ROBOT.2009.5152858

Hamza, A., Fezari, M., & Bedda, M. (2009). Wireless voice command system based on kalman filter and HMM models to control manipulator arm. *2009 4th International Design and Test Workshop, IDT 2009*. https://doi.org/10.1109/IDT.2009.5404140

Izumi, K., Tamano, Y., & Nose, Y. (2004). Tracking of Colored Image Objects with a Robot Manipulator Controlled by Japanese Speech, 1319–1322.

Joshi, N., Kumar, A., Chakraborty, P., & Kala, R. (2016). Speech controlled robotics using Artificial Neural Network. *Proceedings of 2015 3rd International Conference on Image Information Processing, ICIIP 2015*, 526–530. https://doi.org/10.1109/ICIIP.2015.7414829

Junqua, J.-C., & Haton, J.-P. (1996). *Robustness in Automatic Speech Recognition : Fundamentals and Applications*.

Kuljić, B., János, S., & Tibor, S. (2007). Mobile robot controlled by voice. *5th International Symposium on Intelligent Systems and Informatics, SISY 2007*, 189–192. https://doi.org/10.1109/SISY.2007.4342649

Kumar, K., Nandan, S., Mishra, A., Kumar, K., & Mittal, V. K. (2016). Voice-controlled object tracking smart robot. *Proceedings of 2015 International Conference on Signal Processing, Computing and Control, ISPCC 2015*, 40–45. https://doi.org/10.1109/ISPCC.2015.7374995

Lavrynenko, O., Konakhovych, G., & Bakhtiiarov, D. (2016). Method of voice control functions of the UAV. *2016 IEEE 4th International Conference Methods and Systems of Navigation and Motion Control, MSNMC 2016 - Proceedings*, 47–50. https://doi.org/10.1109/MSNMC.2016.7783103

Lei, Z., Gan, Z. H., Jiang, M., & Dong, K. (2014). Artificial robot navigation based on gesture and speech recognition. *Proceedings 2014 IEEE International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, 323–327. https://doi.org/10.1109/SPAC.2014.6982708

Liu, X., Ge, S. S., Jiang, R., & Goh, C. H. (2016). Intelligent speech control system for human-robot interaction. *Chinese Control Conference, CCC*, *2016–Augus*, 6154–6159. https://doi.org/10.1109/ChiCC.2016.7554323

Md, A. H., Md, M. R., Kumar Prodhan, U., & Md, F. K. (2013). Implementation of back-propagation neural Network for isolated bangla speech recognition. *International Journal of Information Sciences and Techniques*, *3*(4), 1–9. https://doi.org/10.5121/ijist.2013.3401

Mohamad, S. N. A., Jamaludin, A. A., & Isa, K. (2017). Speech semantic recognition system for an assistive robotic application. *Proceedings - 2016 IEEE International Conference on Automatic Control and Intelligent Systems, I2CACIS 2016*, (October), 90–95. https://doi.org/10.1109/I2CACIS.2016.7885295

Nakadai, K., Okuno, H. G., & Kitano, H. (2003). Robot recognizes three simultaneous speech by active audition. *2003 IEEE International Conference on Robotics and Automation (ICRA '03)*, *1*, 398–405. https://doi.org/10.1109/ROBOT.2003.1241628

Nghia, P. T., & Vinh, T. Q. (2008). A novel fast noise robust Vietnamese speech recognition applied for robot control. *2008 10th International Conference on Control, Automation, Robotics and Vision, ICARCV 2008*, (December), 821–826. https://doi.org/10.1109/ICARCV.2008.4795623

Noise, T., Noise, S., Noise, E., Distortions, C., & Noise, M. (2000). Noise and distortion 2.1, *9*.

O'Shaughnessy, D. (2008). Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, *41*(10), 2965–2979. https://doi.org/10.1016/j.patcog.2008.05.008

Patil, U. G., Shirbahadurkar, S. D., & Paithane, A. N. (2017). Automatic Speech Recognition of isolated words in Hindi language using MFCC. In *International Conference on Computing, Analytics and Security Trends, CAST 2016* (pp. 433–438). https://doi.org/10.1109/CAST.2016.7915008

Phanprasit, T. (2014). Controlling Robot using Thai Speech Recognition Based on Eigen Sound, 57–62.

Podder, P., Khan, T. Z., Khan, M. H., & Rahman, M. M. (2014). Comparative Performance Analysis of Hamming, Hanning and Blackman Window. *International Journal of Computer Applications*, *96*(18), 975–8887. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.681.6082&rep=rep1&type=pdf

Prodeus, A. M. (2015). Performance measures of noise reduction algorithms in voice control channels of UAVs. *2015 IEEE 3rd International Conference Actual Problems of Unmanned Aerial Vehicles Developments, APUAVD 2015 - Proceedings*, 189–192. https://doi.org/10.1109/APUAVD.2015.7346596

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286. https://doi.org/10.1109/5.18626

Sauer, P., Kozłowski, K., Pazderski, D., Waliszewski, W., & Jeziorek, P. (2005). The robot assistant system for surgeon in laparoscopic interventions. *Proceedings of the Fifth International Workshop on Robot Motion and Control, RoMoCo'05, 2005*, 55–62. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-33845509014&partnerID=tZOtx3y1

Shinde, R. B. (2012). Isolated Word Recognition System based on LPC and DTW Technique, *59*(6), 3–6.

Stefanovic, M., Cetic, N., Kovacevic, M., Kovacevic, J., & Jankovic, M. (2012). Voice control system with advanced recognition. *2012 20th Telecommunications Forum, TELFOR 2012 - Proceedings*, 1601–1604. https://doi.org/10.1109/TELFOR.2012.6419529

Upase, S. U. (2016). Speech recognition based robotic system of wheelchair for disable people. *Proceedings of the International Conference on Communication and Electronics Systems, ICCES 2016*. https://doi.org/10.1109/CESYS.2016.7889851

Watanabe, K., Izumi, K., Ohshima, A., & Ishii, S. I. (2006). An action decision mechanism using fuzzy-neural network in voice commanded fuzzy coach-player system for robots. *2006 SICE-ICASE International Joint Conference*, 5120–5125. https://doi.org/10.1109/SICE.2006.315379

Yao, C., Xiaoling, L., Zhiyuan, L., Huifen, W., & Pengcheng, W. (2017). Research on the UAV Multi-channel Human-Machine Interaction System, 190–195.

Zhang, W., An, Z., Luo, Z., Li, W., Zhang, Z., Rao, Y., … Duan, F. (2016). Development of a voice-control smart home environment. *2016 IEEE International Conference on Robotics and Biomimetics, ROBIO 2016*, 1697–1702. https://doi.org/10.1109/ROBIO.2016.7866572

Zhizeng, L. Z. L., & Jingbing, Z. J. Z. (2004). Speech recognition and its application in voice-based robot control system. *2004 International Conference on Intelligent Mechatronics and Automation, 2004. Proceedings.*, (August), 960–963. https://doi.org/10.1109/ICIMA.2004.1384339

Zinchenko, K., Chien-Yu Wu, & Kai-Tai Song. (2017). A Study on Speech Recognition Control for a Surgical Robot. *IEEE Transactions on Industrial Informatics*, *13*(2), 607–615. https://doi.org/10.1109/TII.2016.2625818

Zhou, R., Ng, K. P., & Ng, Y. S. (1994). A voice controlled robot using neural network. In Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on (pp. 130-134). IEEE.

45 Questions to test a data scientist on basics of Deep Learning (along with solution). (2017, 12 12). https://www.analyticsvidhya.com/:

Al-Sabbagh, M. (2017, 12 12). Speech Signal Processing. Retrieved from https://www.slideshare.net/l: https://www.slideshare.net/lucky43/speech-signal-processing

Github-Katarina. (2017, 12 12). https://github.com/robotika/katarina

Google Inc. (2017, 12 12). Cloud Speech API. https://cloud.google.com/

Human Perception of Sound . (2017, 12 12). http://zone.ni.com/:

Introduction to Music Production. (2017, 12 12). http://introduction-to-music-production.blogspot.com.tr/.

Labbookpages. (2017, 12 12). FIR Windowing. http://www.labbookpages.co.uk: http://www.labbookpages.co.uk/audio/firWindowing.html

Parrot Bebop Drone. (2017, 12 12). https://www.oyuncakhobi.com: https://www.oyuncakhobi.com/urun-kategori/akilli-oyuncaklar/prd-parrot-bebop-drone-kirmizi

Parrot Inc. (2017, 12 12). Bebop Drone. http://global.parrot.com/: http://global.parrot.com/au/products/bebop-drone/

SPEECH ACOUSTICS .(2017, 12 12). http://clas.mq.edu.au/: http://clas.mq.edu.au/speech/acoustics/frequency/analog_digital.html

Speech Production. (2017, 12 12). https://msu.edu/.

Sampling. (2017, 12 12). http://support.ircam.fr/:
http://support.ircam.fr/docs/AudioSculpt/3.0/co/Sampling.html

Scientific Research Publishing. (2017, 12 12). http://file.scirp.org. http://file.scirp.org: http://file.scirp.org/Html/4-6801198_34183.htm

Scientific Research Publishing. (2017, 12 12). http://file.scirp.org/Html/4-6801198_34183.htm.

Sesin genligi ve frekansı. (2017, 12 12). www.karmabilgi.net/: www.karmabilgi.net/sesin-genligi-ve-frekansi/

Wikipedia Inc. (2017, 12 12). Sine wavelength. Wikipedia: https://commons.wikimedia.org/wiki/File:Sine_wavelength.svg

Wikiwand. (2017, 12 12). Dynamic Time Warping. http://www.wikiwand.com/: http://www.wikiwand.com/th/Dynamic_time_warping

# APPENDIX 1 –Recording Speech And Google Speech API Recognition

```
def record_audio():
    """PyAudio example: Record a few seconds of audio and save to a
WAVE file."""
    CHUNK = 1024
    FORMAT = pyaudio.paInt16
    CHANNELS = 1
    RATE = 16000
    RECORD_SECONDS = 4
    WAVE_OUTPUT_FILENAME = "output.wav"

    p = pyaudio.PyAudio()

    stream = p.open(format=FORMAT,
                    channels=CHANNELS,
                    rate=RATE,
                    input=True,
                    frames_per_buffer=CHUNK)

    print ("* Listening...")
    frames = []

    for i in range(0, int(RATE / CHUNK * RECORD_SECONDS)):
        data = stream.read(CHUNK)
        frames.append(data)

    print ("* Done Listening")

    stream.stop_stream()
    stream.close()
    p.terminate()

    # Write the data to a wav file
    wf = wave.open(WAVE_OUTPUT_FILENAME, 'wb')
    wf.setnchannels(CHANNELS)
    wf.setsampwidth(p.get_sample_size(FORMAT))
    wf.setframerate(RATE)
    wf.writeframes(b''.join(frames))
    wf.close()

def run_recognition_system():

    record_audio()
    with io.open('./output.wav', 'rb') as audio_file:
        content = audio_file.read()
        audio = types.RecognitionAudio(content=content)
    config =
types.RecognitionConfig(encoding=enums.RecognitionConfig.AudioEncodi
ng.LINEAR16,
                                     sample_rate_hertz=16000,
                                     language_code='tr-TR')
    response = client.recognize(config, audio)
    for result in response.results:
        print (result.alternatives[0].transcript)
        return result.alternatives[0].transcript
```

## APPENDIX 2 –UAV Controlling Codes Katarina Framework

```
while(1):
            try:
                a=run_recognition_system().lower().replace(" ","")
                #print a.encode('ascii','ignore')
            except:
                print 'Komut girin’

            if (joystick.get_button(0)==1 or a=='havalan'):

                drone.takeoff()
                drone.flyToAltitude(0.3)
                drone.hover()

            if(joystick.get_button(1)==1 or a=='in'):

                drone.land()

            if(joystick.get_button(2)==1 or a=='acil'):
                drone.emergency()

            if(joystick.get_button(3)==1):
                drone.land()

            if(joystick.get_button(4)==1 or a=='yukar'):
                drone.flyToAltitude(drone.altitude+0.1)

            if(joystick.get_button(6)==1 or a=='aa'):
                drone.flyToAltitude(drone.altitude-0.1)

            if(joystick.get_hat(0)==(0,1)or a=='ileri'):
                pitch=20

            if(joystick.get_hat(0)==(0,-1)or a=='geri'):
                pitch=-20

            if(joystick.get_hat(0)==(-1,0)or a=='solagit'):
                roll=-20

            if(joystick.get_hat(0)==(1,0)or a=='saagit'):
                roll=20

            if(joystick.get_hat(0)==(0,0)or a=='dur'):
                roll=0
                pitch=0
                yaw=0

            if(joystick.get_button(5)==1):
                yaw=25
                textPrint.printx(screen,"RB")
            if(joystick.get_button(7)==1):
                yaw=-25
                textPrint.printx(screen,"RTE")
            else:
                print "no command"

        drone.update( cmd=movePCMDCmd( True,roll,pitch, yaw  ,0 ) )
```

# CURRICULUM VITAE

İlhan Sofuoğlu was born in 1989, Izmir – Turkey. He holds a Bachelor of Science degree from Yaşar University, Computer Engineering Department after graduating in 2014. After starting his Master of Science (MSc.) education at Yaşar University. He simultaneously started as a research assistant at Yaşar University in 2014, and he continues his job since then**.**