ACCOMPANIMENT ROBOT WITH TURKISH SPEECH SYNTHESIS AND LIP SYNCHRONIZATION

by

İbrahim Özcan

B.S., Computer Engineering, Boğaziçi University, 2013

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering
Boğaziçi University
2017

ACKNOWLEDGEMENTS

First of all, I would like to express my sincerest gratitude to my supervisor, Prof. H. Levent Akın. His guidance, support and unending patience made this thesis possible.

I am grateful to Prof. Fikret S. Gürgen and Assoc. Prof. Hatice Köse for kindly accepting to be members of my thesis committee and for their constructive suggestions.

I am thankful to my family and friends, who patiently supported me and during this journey and helped me through it.

ABSTRACT

ACCOMPANIMENT ROBOT WITH TURKISH SPEECH SYNTHESIS AND LIP SYNCHRONIZATION

With the introduction of robots to the domestic environments, robot - human relations have become a more important research area. That is why we decided to make a robot to accompany people in their homes. We were dealing with four different problems to accomplish this: creation of various robot heads, providing the robot the ability to produce speech sound, the ability to move lip positions in accordance with speech when the robot is talking and establishing a relationship between the words in the sentences which are spoken to the robots. The first of these problems was solved by the construction of three different robot heads. To solve the second problem, the Formant Speech Synthesis method was used. In the solution of the third problem, special lip positions were defined for each group of letters and during the speech these lip movements were made by robots respectively. In the last problem, root finding, annotation and relation finding tools were developed for Turkish. By the extraction of information from what was said by the robot by finding the suitability of the phrases. The mentioned statements were inserted into the patterns and the robot was able to extract information from these patterns. In the end, people and the robot were able to play together and the system was evaluated for how well it worked.

ÖZET

TÜRKÇE KONUŞMA SENTEZLEMELİ VE DUDAK SENKRONİZASYONLU REFAKATÇI ROBOT

Robotların ev içlerine girmesiyle, robot – insan ilişkileri eskisine göre daha önemli bir araştırma alanı oldu. Bu yüzden insanlara evlerinde eşlik edecek bir robot yapma kararı aldık. Bunu gerçekleştirmek için 4 farklı sorun ile uğraşıldı: çeşitli robot kafalarının oluşturulması, robotların ses üretilmesinin sağlanması, robotun konuşurken dudak pozisyonlarını konuşma ile uyumlu olarak hareket ettirebilmesi ve robotlara söylenen cümlelerdeki kelimeler arasında ilişki kurabilmesi. Bu sorunların ilki 3 farklı robot kafasının yapılması ile çözüldü. İkinci sorunu çözmek için Biçimleyici Konuşma Sentezleme yöntemi kullanıldı. Üçüncü problemin çözünde ise her bir harf grubu için özel dudak pozisyonları tanımlandı ve konuşma sırasında bu dudak hareketleri sırasıyla robotlar tarafından yapıldı. Sonuncu problem de ise Türkçe için kök bulma, ek belirleme, ilişki bulma araçları geliştirildi. Söylenen cümleler kalıplara sokuldu ve kalıplardan robotun bilgi çıkarabilmesi yapıldı. En sonunda insanlar ile robotun birlikte oyun oynaması sağlanıp, sistemin ne kadar iyi çalıştığı değerlendirildi.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

2D	Two Dimensional
3D	Three Dimensional
S(f)	Source Spectrum
T(f)	Vocal Track Transfer
R(f)	Lip Radiation
D_i	Current Duration
D_{i-1}	Previous Duration
F	Frequency
BW	Bandwidth
T	Time
N_{FE}	The number of pictures which have the feature
N_{FNE}	The number of pictures which do not have the feature

LIST OF ACRONYMS/ABBREVIATIONS

AI	Artificial Intelligence
AIML	The Artificial Intelligence Markup Language
BML	Behavior Markup Language
FPS	Frame per second
HMM	Hidden Markov model
NLP	Natural Language Processing
PCB	Printed Circuit Board
ROS	Robot Operating System

1. INTRODUCTION

The commercial activities in the service robot field are growing. Currently estimated sales values of household and entertainment robots is about 31 million US dollars [1] and also more than half of the population of the Japan want to use a domestic robot in the future [2]. These show that human-robot interaction in domestic environments is unavoidable.

Real-time human-robot interaction capability is a fundamental requirement for the robots which are to operate in domestic environments. The capability of speech recognition, speech synthesis, and having a sophisticated artificial intelligence (AI) system becomes essential when the deployment of accompanying robots is considered. Recent developments in these technologies show us that accompanying robots will soon become an inseparable member of our family. Therefore, designing such robots presents a wide range of new areas of application for robots [3].

Researchers have been working on designing robots suitable for the domestic environment. Industrial robots are generally held in cages and work in a very restricted area with no or very little human interaction. Therefore, their design may result in hurting a person who accidentally goes into the robot cage. However, for the accompanying robot case, humans should make sure that the robot is harmless. Also, the appearance of the robot should be suitable for the home environment. Therefore, in recent years the design of the robots changed from heavy and bulky to cute and small.

In addition to the design of the robot, speech synthesis systems are also an essential part. Previously, keyboard, mouse or button interfaces were used for robot - human interaction, but for a proper communication, the robots should have human like capabilities. One of the most important capabilities for humans is communication with other humans via voice. Voice includes the information about the words, but it also carries data about the emotional state of the speaker. For this reason, recently researchers have been studying on flexible speech synthesis systems [4].

The design of the robot can be perfect, speech synthesis system can be very flexible and very neutral. Yet, without processing the spoken words or constructing meaningful sentences, the accompanying robot will not be complete and the only robust way to generate those sentences is using Natural Language Processing (NLP).

In this study, we designed various robot heads with different appearances in order to determine the most suitable robot head for accompanying robots. In addition to the their appearances, various features are added to these heads such as lip synchronization. Additionally, to figure out the effect of the speech quality and the gender of speech, we developed three different voices. After that, we implemented a simple artificial intelligence based module in the robot heads to play a basic child's game with the help of the developed chatbot engine.

The rest of this thesis is organized as follows: In Chapter 2, we present the background information about the components of the developed system. This background information starts with the design types of the robot faces and continues with the speech synthesis systems and the chatbot architecture. Chapter 3 is dedicated to the introduction of the robot head types made for this thesis. In Chapter 4, the lip synchronization mechanisms for each robot head are presented with Robot Operating System (ROS). In Chapter 5, the speech synthesis system which is implemented in a flexible way is described. In Chapter 6, the chatbot system with Turkish NLP tool which is also developed for this thesis.

In Chapter 7, the experiment conducted to show how the robot behaves against the humans and how it finish a competitive task is explained.

In Chapter 8, the discussion about the results and conclusions and the future work are given.

2. BACKGROUND

Accompanying robots are currently one of the hot topics in robotics research. Many companies try to enter market with toy like robots which can help the humans via interacting with them. Even the parents have started to buy toys with an AI system instead of a static toy which cannot move or respond to the children.

Accompanying robots consist of several sub-systems. In order to interact with the humans, these robots should have speech recognition and speech synthesis systems with response generation which is powered by the natural language processing. Also, these systems should have a physical structure to be called as robot.

We focused on the human-robot interaction in this study and we only included the head related interaction without using the gestures made by the arms or body of the robot. In order to prove a necessary background information about this thesis, we divided the current developed robots analysis into the sub-fields and examine these fields separately. The fields are organized as follows:

- Robot Faces
- Speech Synthesis
- Chatbot Engines

2.1. Robot Faces

Every accompanying robot has a different design from the others. Each design is chosen based on the purpose of the robot. For example, the immobile robots which do not have any wheels or legs for movement have very small designs in order to have a small footprint in the domestic environment. On the other hand, mobile robots which have the capability of changing their locations have larger designs with many flexibility on their movement mechanism such as omni-wheels, legs etc. However, currently almost every small sized social robots look alike [5] because curvy structures is safer, pleasing

to the eye and also this is a physical structure that does not scare children for the easy interaction case [6].

This thesis is focused on human - robot interaction, so we will analyze the face and face related systems of the developed robots instead of mentioning about their other body related designs.

2.1.1. Tablet Face

One of the most common head types for robots are tablet heads [7,8]. Nowadays, tablets are easily available with a relatively low price, wide range of sizes and in addition to that, Android or IOS tablets come with many features such as Speech Recognition, Speech Synthesis and WiFi which are the important parts for an accompanying robot. Additionally, tablets have high processing powers that can allow even playing games which require OpenGL. With the power of OpenGL and rendering images with at least 30 fps, tablets become the first choice for rendering an avatar. Tablet faces have a lot of features other than these, too. One can make more human like complex faces with motions, also one can make cuter faces with different emotions without changing any hardware, with only changing the displayed model.

Buddy [9] and Zenbo [10] robots come with cute designs with various expressions. Their emotional state includes six core emotion states [11] which are sadness, anger, surprise, fear, disgust and happiness. Additionally, Zenbo has a lip synchronization mechanism to make it look like it is actually speaking.

Buddy and Zenbo robots are like cartoon characters, but there are other tablet faced robots which have avatar faces. The Personal Robot [14] which is a kick starter robot project has this type of face which can be seen in Figure 2.1. It shows the same emotional state as the Buddy and Zenbo, but it can also change the avatar model, based on the user choice. With the increased processing power and more realistic designs this type of head can show any type of face and also can display the avatar model of actual personss.

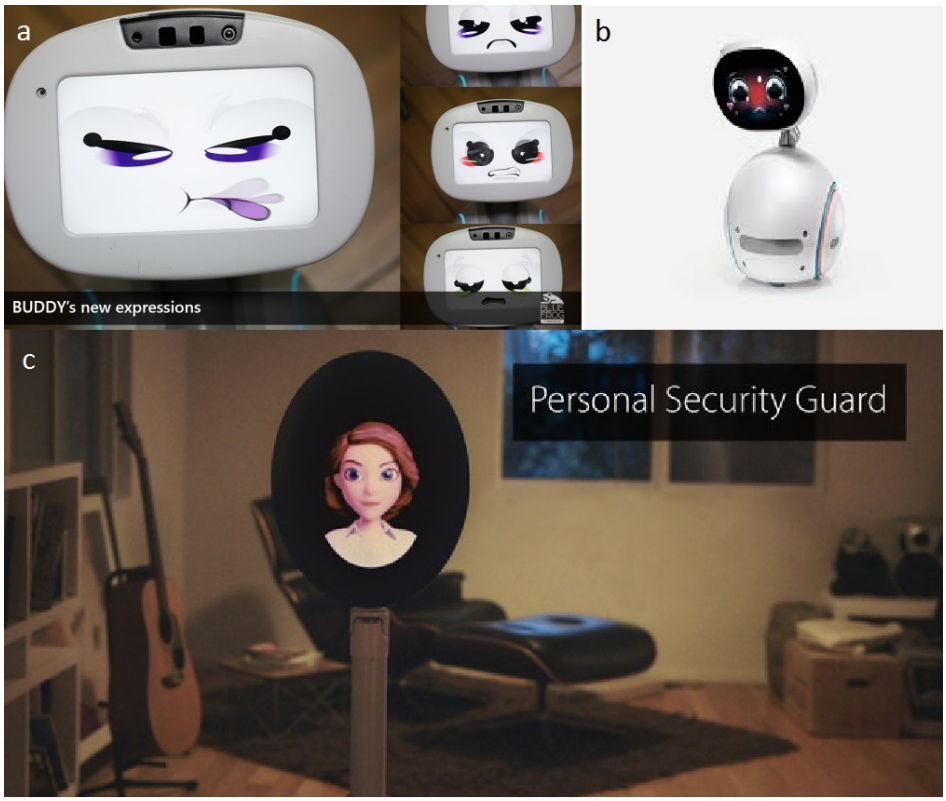


Figure 2.1. a) The Buddy Robot with different emotion states from [12], b) Zenbo Robot from [13] c) Personal Robot [14]

2.1.2. Hologram Robot

Tablet faces are displayed on 2D surfaces, still 2D is not sufficient to imitate human faces. Even if they fully design a human head and generate perfect motions, this still gives a perception of using a telepresence robot. The robot should appear with 3D structures. Without losing the flexibility to change the faces and the face expressions without changing the hardware, also constructing 3D structures to the faces bring the new challenge for the design concept. There should be a new type of head design which is in the middle of the physically 3D structure with software based 3D image. For this purpose, hologram robot head approach is emerged.

Hologram robot heads also have some varieties and some of the hologram robots are fully based on the hologram technology and some of them uses the 3D structure with hologram.

Gatebox [15] is the one of the examples which uses only hologram based approach for displaying the robot. Hologram is generated by using a projector with Rear Projection Film. Rear Projection Film allows to project any faces or images from behind the film. With this technology, Gatebox can show 3D motions and more realistic images in a small environment.

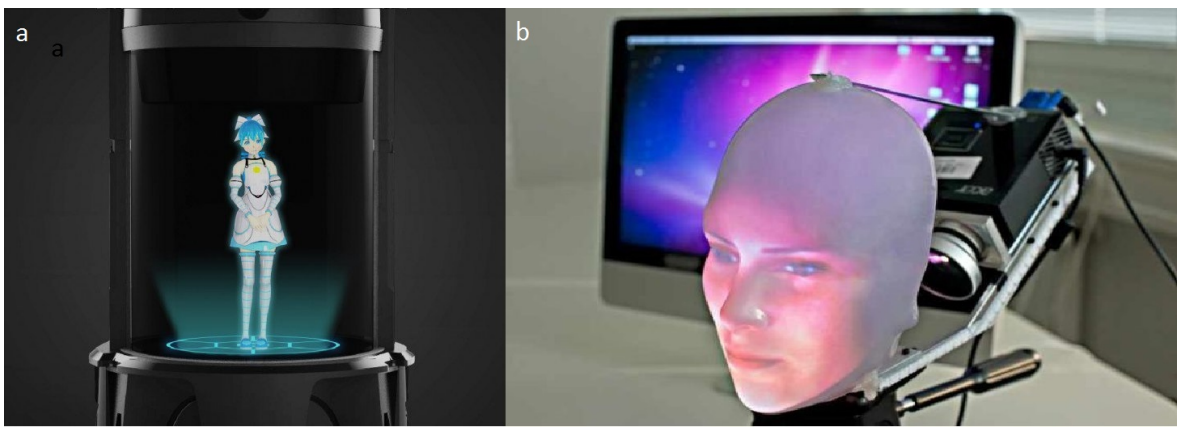


Figure 2.2. a) Gearbox from [15], b) Mask-Bot from [16]

On the other hand, Hologram robots with 3D structures use the specially made 3D model of human face. Hologram is generated similarly as in the previous case, but instead of using a flat rear projection film, a 3D model structure is constructed and created to give the same effect as the rear projection film. The model is chosen as much transparent as possible to show the projected image on the front while as white colored as to store the light. The image generation unit is also different from the previous case. The 3D structure is not flat and the image should also be warped for correctly displaying in that structure. Mask-Bot [16] is one of the examples for this kind of hologram heads.

Hologram robot head gives a lot of flexibility and compared to the tablet faces, it provides more realistic solutions. However, it also brings a huge disadvantage which is not the inability to display the image under high ambient light intensity. The projector based methods should be in a closed environment, away from direct light sources. When the light intensity of the environment is higher than the projector capability, this type of faces can not be used and if the working in that environment is not unavoidable, the only way to use such kind of head is using projectors which can provide higher intensities.

2.1.3. Animatronic Heads

Currently, the most human-like heads are Animatronic Heads. While this type does not have any features to change its face model fully without changing the hardware, still when an artificial skin which is called Frubber (Flesh Rubber) [17] or silicone is used, this kind of heads can generate human mimics and even without moving and with a make-up, it becomes harder to distinguish the robot from human. However, not all the Animatronic Heads have artificial skin, there are lots of varieties, but they can still be categorized similarly.

2.1.3.1. LED Based Heads. The LED based heads are actually between the tablet heads and animatronic heads. This type of head generally has the capability of making

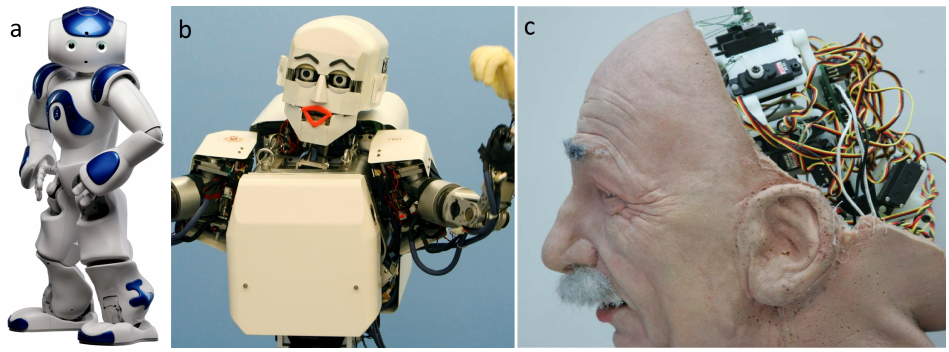


Figure 2.3. a) LED Based Heads (Nao) from [18], b) Kinematic Heads (KOBIAN-RII) from [19], c) Kinematic Heads with Skin (Albert HUBO) from [20]

neck motions, but because of the lack of Degree of Freedom (DOF) in their face parts, they cannot display any mimics. Even the displayable emotional states are limited. Mostly, the emotional states are displayed by the color of the LEDs. For example, while angry, the emotion state is represented with the red color, the white color is used for neutral state, etc. They do not have any actuators for their mouth motions. Therefore, even speaking, they do not provide any lip actions and for the human - robot interaction case this type of the animatronic head gives the minimum amount of information to the user. The most comment example for these heads are Nao [18] and Pepper [21].

2.1.3.2. Kinematic Heads. The deficiency of the expression set of the LED based head are partially solved by increasing the number of the actuators in the face. Compared with LED based heads, kinematic heads have movable parts which can allow them to express various emotions with the movements. The number of DOF is directly related to the expressible face states and this number sets the limit on the actions of the robot. Robothespian [22] like robots have a DOF in their mount to produce speaking effect on the other hand KOBIAN-RII [23] is a robot which can move its eyes, eyebrows, eyelids, lips and mouth with 27 DOF.

2.1.3.3. Kinematic Heads with Skin. Researchers try to make the robotic heads as realistic as possible by using artificial skin and producing the ear, eye and teeth parts similar to human parts. With human like motions, this type of robotic heads are the closest to the humans. Kinematic heads have a solid face part which can be made from metal or solid plastic and because of the solidness of these materials, emotions are represented by the movement of the separate parts. However in a human face all the face parts are connected via skin. Therefore, the faces of these robots are produced from elastic materials like Frubber and silicon. In order to move the part of the faces, actuators which can make pull-push motions are added to the behind of the elastic materials. Unfortunately, according to the Uncanny valley theory [24], robots should show very similar movements to humans. Therefore, these kind of heads need many DOFs in their faces like Albert HUBO [20] which has 31-DOF in its head.

2.2. Speech Synthesis

Speech synthesis is a method for converting written text to the artificial voices. Every speech synthesis system has two main phases. The phonetic information generation from the text and the sound generation from the phonetic information [25]. For this part, we will focus on sound generation. We used the Turkish language for speech synthesis and Turkish is a language that can be spoken as a written. Therefore, there is no need to convert the text to phonetic representation.

2.2.1. Articulatory Speech Synthesis

Humans have physical structures for producing voices. So, researches consider that in order to produce the sound signal, structures similar to that of the humans have should be constructed. Therefore, the first attempt for speech synthesis is using mechanical devices. These mechanical devices are generally used for producing the voice signals. These machines have two main parts. The first part is the fundamental frequency generators. The purpose of this part is generating an air flow with a specific frequency which is a hardware representation of the glottal. The second part is the formant generator. The purpose of this part is the changing the dominating frequency

of the generated fundamental frequency which is a hardware representation of the vocal track. The first part cannot be moved, but second part is consist of moving actuators. In Figure 2.4, the vocal-tract model contains a series of moving 10-mm or 15-mm thick plastic bars and moving these parts it generates different sounds.

2.2.2. Formant Speech Synthesis

After articulatory speech synthesis, the researchers tried to produce the speech signal using the electronic devices instead of using mechanical structures. The first implementations of the Formant Speech Synthesis use the analog hardware devices which can produce the speech signal. One of the most dominating speech synthesizers for this type is DECtalk [27]. The early version of the DECtalk was produced as a separate unit which can be connected via the asynchronous serial port. However, with the increase in the processing power, this formant synthesis unit has become a fully software based tool. Every electronic component in the circuit has a software implementation. By simulating the formant frequencies in the software, formants are modelled mathematically.

This is the main underlying thought for Formant Synthesis. Formant Synthesis does not store human speech or use human speech pieces. Instead, the synthesized speech outputs are generated by adding the formant frequencies to the source model. Parameters are determined by the rules and generally this synthesis method called as *Rule-Based Speech Synthesis*. Formant Speech Synthesis uses three components:

- Source Spectrum
- Vocal Tract Transfer Function
- Lip Radiation Effect

Source Spectrum is called as a source filter model. It generates the glottal source for the speech. These are six glottal source models which are Rosenberg, Hedelin, Fant, Ananthapad-Manabha, Liljencrants-Fant and Fujisaki-Ljungqvist [28] models. The more commonly used glottal models are Rosenberg and Liljencrants-Fant.

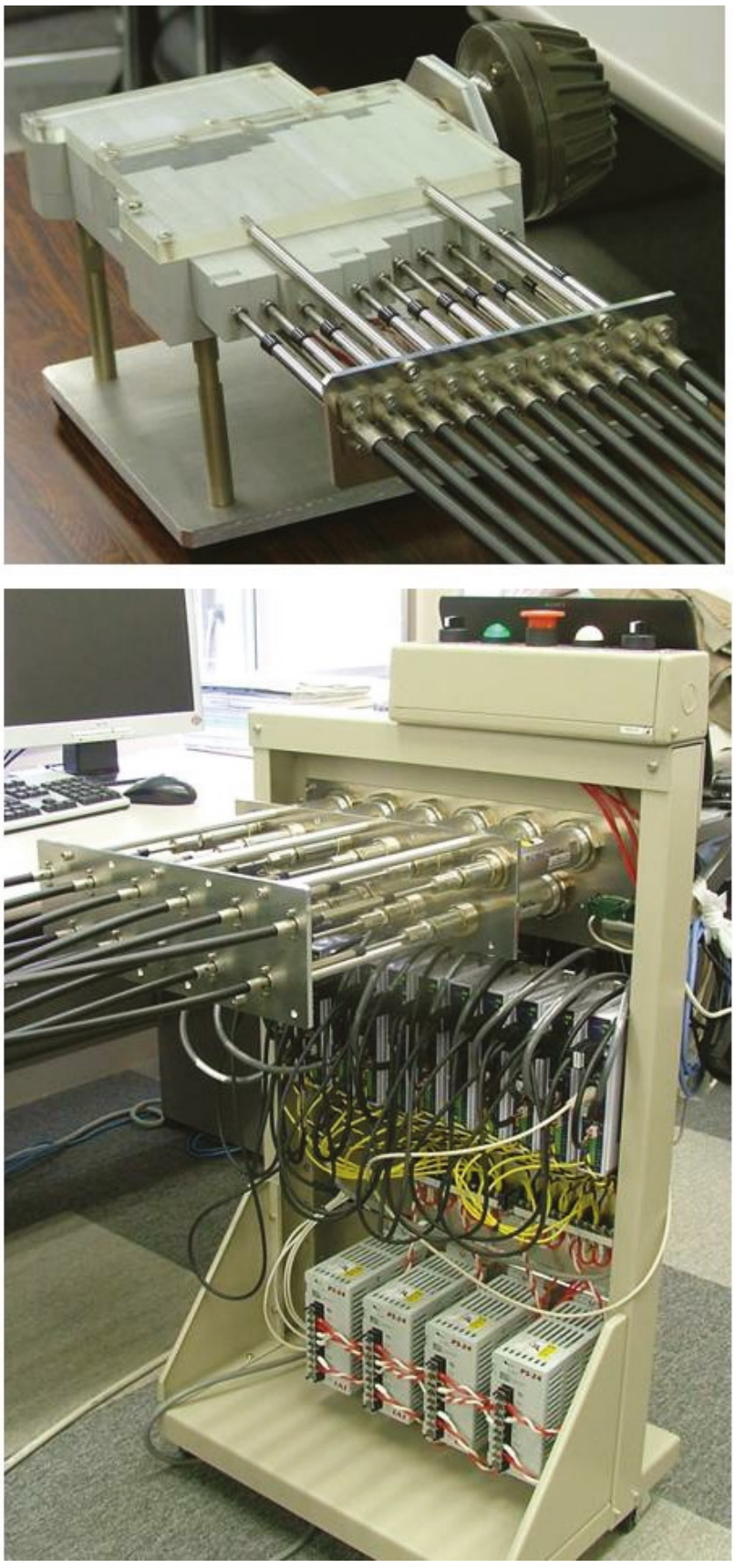


Figure 2.4. Articulatory Synthesis Example from [26]

Vocal Tract Transfer Function is for the addition of the formant frequencies. There are two ways to merge the formant frequencies which are cascade and parallel. In the cascade synthesis, every formant consists of frequency and bandwidth and in the merging method, every formant filter is concatenated the other formant frequencies. However, in the parallel synthesis method, every formant has additionally amplitude value and the merging method is the addition of the each formant filter output.

Lip Radiation Effect is used to represent the transformation of the volume of the speech at the lips. Lip Radiation Effect can be approximated with a FIR filter model with a zero pole. With these parameters, the speech output for a pronunciation can be estimated.

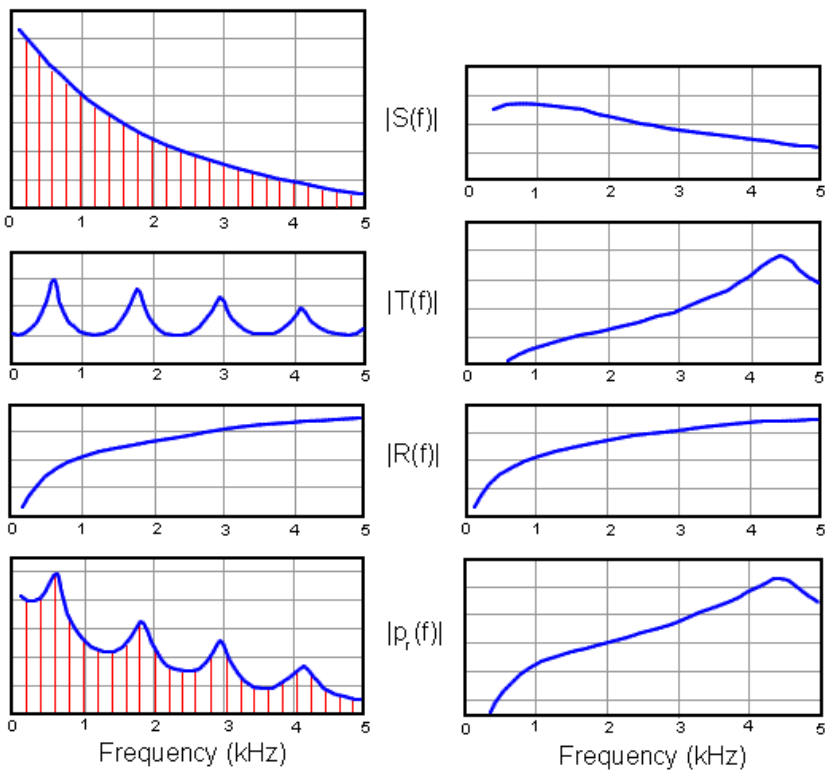


Figure 2.5. Formant Synthesis Process [29]

After vocal track estimation, sentences can be made by concatenating or dynamically chancing the vocal track information. In order to achieve a human like sound, a lot of parameters should be used on Vocal Tract Transfer Function. For example,

Klatt Formant Synthesis [30] uses 39 control parameters.

2.2.3. The Tube Resonance Model Speech Synthesis

The idea of the Tube Resonance Model Speech Synthesis is very similar to Formant Speech Synthesis. This type of synthesis does not require any recorded speech and it produces the speech by the rules. The components which are used in this model is the same as the Formant Speech Synthesis. The only difference is the implementation of Vocal Tract Transfer Function. Instead of using the formant frequencies directly, the air flow in the tube causes the formant frequencies.

In the human speech production system, air flows through the vocal track and this is the cause of the voice. In this synthesis model, the vocal track of the human is represented with the lots of number of small tube piece. If the air flow in these tubes is simulated [31], the voice output of the system can also be calculated as shown in 2.6.

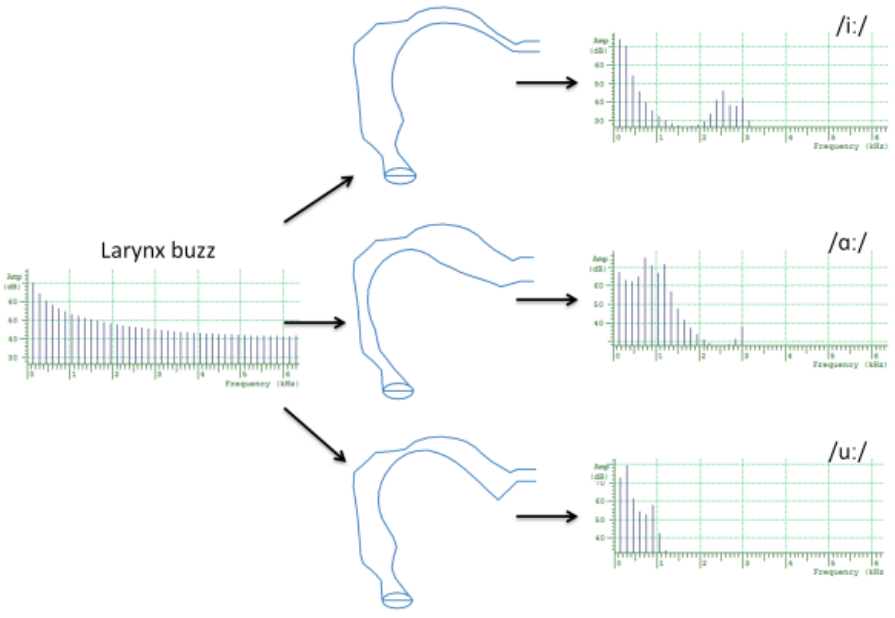


Figure 2.6. Different Tube Models for different vowels [32]

2.2.4. Concatenative Synthesis

Concatenative Synthesis is a technique that uses the recorded human speech as input. In this model recorded sounds are divided into small groups which can vary between 1 and 10 microseconds and these small pieces are concatenated in order to produce the sound signal. Modern Speech Synthesis systems use the *Concatenative approach*. The main reason for this is the high-naturalness and simple implementation. The previous speech synthesis approaches do not use human sound directly, so the noise level, formant changes and the other difficulties should be solved manually. However, in this method, the noise and the formant changes directly come with the human speech samples.

Concatenative Synthesis is a corpus based method. A large amount of sentences are read by a human and are recorded. The recorded sound duration directly affects the speech quality.

The main problem for this approach is unit selection. The recorded units should be selected correctly.

2.3. The Chatbot Engine

2.3.1. Multiple Choice

Multiple Choice is one of the basic and easy-to-build method with a high engagement rate used by chatbots [33], and it is still one of the fastest way of generating responses. This is a basic method because it decreases the complexity of the problem to a number of choices and since each choice is mapped to a different response, the actions of the robot are accordingly limited. For example, when the robot provides four options, the number of possible responses will also become four.

Multiple Choice also simplifies response recognition of the users. If a speech recognition tool is not used, the users can select the input from the touch screen or

can write the choice number from a keyboard. If a speech recognition tool is used, the number of possible words which can be said by users will decrease. By detecting only the keywords from the sentences, the selected choice can be understandable. Therefore the whole words in the dictionary are not needed to be checked and only specific words are processed.

2.3.2. The Artificial Intelligence Markup Language

A robot which uses the Multiple Choice method is not suitable for fluid real-time conversation because every time the users determine their sentences according to that of the robot instead of their own will, also this causes breaks in the dialog with the users. In order to decrease the negative effect of the Multiple Choice method, Artificial Intelligent Markup Language (AIML) [34] was formed. The AIML approach is one of the standardized methods for chatbots. It is based on pattern matching. In AIML, there is no semantic level examination, AIML tries to match the input into patterns and sends the response which is related to the matched pattern. In order to write AIML, four essential tags and five additional tags which improve the usability are used. The essential tags are:

- aiml: This is the start and end tag of AIML documents. It is also used to differentiate AIML documents from others.
- category: This is for separating the pattern and the responses from the other patterns and responses. Every category tag has only one pattern and one template tag.
- pattern: This is for checking whether the user's input is suitable for that category or not. If the pattern matches with the user's input, the response which is in the same category will be used.
- template: This tag contains the response of the category.

The additional tags are:

- random li: In a conversation, there can be various responses for the same input (for example, the response of "Goodbye" can be "Bye", "Goodbye", "See you later"). Therefore in order to increase the variety of responses, randomly a response which is in *li* tag is generated.
- that: In dialogs, every sentence is related to the previous sentence. The responses the robot gives have an effect on the responses of the user, so the robot should detect the related category from the set of categories which have the same patterns. For example, if the taken input of user is "Yes", multiple categories can be matched, but "Yes" is the response of the previous robot's question. *that* tag stores the response of the robot.
- srai: The response of different patterns can be the same. In AIML, every category has only one pattern, so in order to connect the different category templates, this tag is used.
- set: Some of the sentences include important information about the user or other things and that information should be stored in order to make a more realistic conversation. In order to store the data from previous sentences, this tag is used. For example, the sentence which begins with "My name is" should end with the user's name so the robot can get the information from the pattern.
- get: The data stored by using the *set* tag can be used in a *template* tag. This tag is to get the stored data. For example, if the user's name is known, at the end of conversation, the robot can say "Goodbye <get data="username"> </get>"

The difficulty of AIML is that many patterns are needed. Multiple Choice limits the user's input, but in the AIML case, the robot should correctly predict the user's input.

Currently, many chatbot engines use the patter matching approach for response generation. Even the API.AI [35] which is developed by Google uses the Json based chatbot engine, which tries to detect the keywords in the sentences.

2.3.3. Markup Language and The Behavior Markup Language

AIML is one of the best methods for chatbots, but this is a very basic system for robots which have actuators and ability to the move and by using these, can improve communication with the humans. Aside from giving the correct text data, one of the most important things for communication with other humans is using mimics, hands and body language. Therefore as the robot builds correctly a sentence form the user input, the robot should also make some gestures which are similar to humans. For completing this mission, a new markup language is used in Max [36]. With a markup language which is similar to AIML, Behavior Markup Language (BML) is used. In this system, the user's inputs are taken by using pattern matching. However additionally, BML not only gives text data but also provides an action. Behaviors are added in the responses. For example in the response of Figure 2.7, there is an action which is the smiling action which takes 1000 milliseconds.

```
<bml>
<speech id="speech001" start="0">
<text> I am fine and you </text>
</speech>
<gesture lexeme="smile" duration="1000" starts="speech001:start+5" />
</bml>
```

Figure 2.7. Behavior Markup Language Example

2.3.4. Question Answering

Every sentence can be considered as a question for ourselves. By using this approach, we can decrease the complexity of the NLP problem to the Question Answering problem. For example the answer to the "Hello" question is also "Hello" and the answers of the sentences which give information to ourselves are saying "Yes" or just shaking head or making a different action for showing that we understand.

This method works well and the reason for this is by using only keyword matching, the question which is asked is determined. Because of this, examining the input data becomes easier.

3. ROBOT HEADS DEVELOPED IN THIS STUDY

In this study, three robot heads were made for testing robot - human interaction efficiency. The complexity of the robot heads increases in every version with the addition of many features. The constructed heads are in the tablet head and kinematic head categories.

3.1. Requirements and Reasons for Robot Heads Development

In our robot projects, robot-human interaction holds an important role. To produce efficient communication, human-like appearance is a pivotal factor [37]. Therefore, the first requirement for the robot heads is that the heads should represent the human head.

The second requirement is the capability of the making movements in the jaw or lip mechanism, in order to perform lip synchronization when speaking. The humans produce the speech by using the vocal tract and vocal organs, so the lip motions is necessary for speech production. In order to mimic the speaking of the humans, the robot should accomplish this requirement.

The third requirement is the capability of showing different emotion states. The emotion changes and the state presentation for this are not studied on this thesis, but for the increase the usability of the robots for the task will be included at the future, we also added this specification to our requirement list.

The last requirement on the robot heads is the movement in their eyes. Also, this is not considered in this study, but the tracking an object from their eyes is needed feature for the future tasks.

Turgay robot which is developed for the Multi-Robot Tour Guide System [38] is already available for performing human-robot interaction experiments, but it does not

have a head. For this reason, we first developed a tablet based head for this robot. The reasons for choosing a tablet head type robotic head instead of using another type of robotic head is the lack of area to place the robotic head and tablet heads are easy to build. In Figure 3.1, you can see the Turgay robot with its head.

Figure 3.1. TURGAY robot with tablet head

After completing the Turgay robot project, we started to develop a new robotic platform for serving the humans in the domestic environment. This platform also will interact with the humans via its own head. Therefore, it needs a head design for its interaction and this head design has to be designed before the design of the other components because robot should have sufficient area to place its head and the head size and head appearance should be suitable for the robot. We determined that the head type for the new robotic platform should be a kinematic one and we started to design a kinematic head. The head model for the new robotic platform can be seen in Figure 3.2.

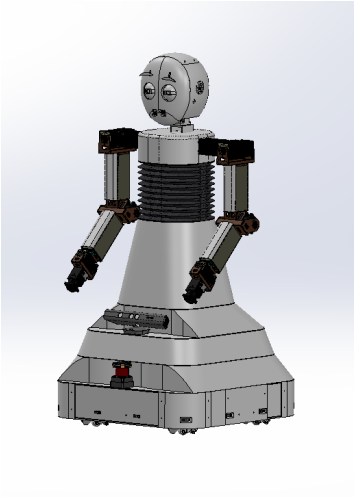


Figure 3.2. Büşra robot head with body

3.2. Tablet Head Van Gogh

As we mention in Section 2.1.1, one of the easiest robotic heads to build is the tablet head. So, we made our first robot head in that category.

Tablet Head Van Gogh displays the image of a Van Gogh self portrait [39] (Figure 3.4) and using OpenGL software [40] and Candide 3 [41] face model, we add dynamic movements to a static image.

Candide 3 is a parametrized face model which has more than 100 polygons. Muscle positions on the human face have also been taken into account in the creation of this model. The motions of the face area around the nose is more than the motions on the foreheads. For this reason, there are fewer vertices around the forehead while more vertices are used around the nose. Also, this model provides us with the *Shape Units* which can be used to fit the model to any dimensions and the *Action Units* which helps us to move the vertices of the model for animating the face.

The Candide 3D model contains 113 vertices. In order to use a picture as a texture in that model, 113 points in the picture should correspond to the points on

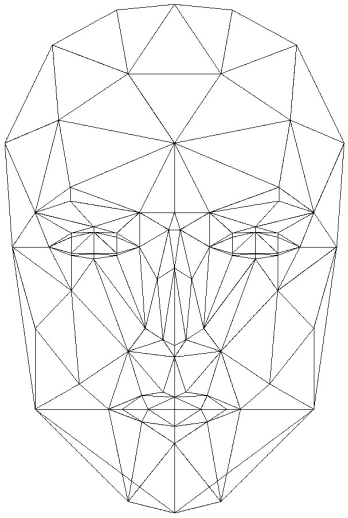


Figure 3.3. Candide 3 Face Model from [41]

the 3D model. The location of points should be almost perfectly determined and the points should represent the same description as the vertices description. For example, the 65th vertex is described as chin right corner and the point used in the picture should be located on the face at the right chin corner. In order to determine the correct points, instead of manually selecting, we used face landmark location detector Face++ [42]. The maximum number of landmarks which is given by the software was 83 points.

Face++ detected the landmarks of the Van Gogh portrait. However, the detected landmarks and vertex locations on the 3D model were not the same. Some of the landmark points could be directly used for the 3D model, but some of them did not have a match. Therefore, we implemented a mapping function to find the points for all of the vertices. While implementing this function, we assume that:

- The face is clearly seen in the picture.
- The center of the chin and the center of the upper head is almost vertically aligned.
- The center of the ears are almost horizontally aligned.

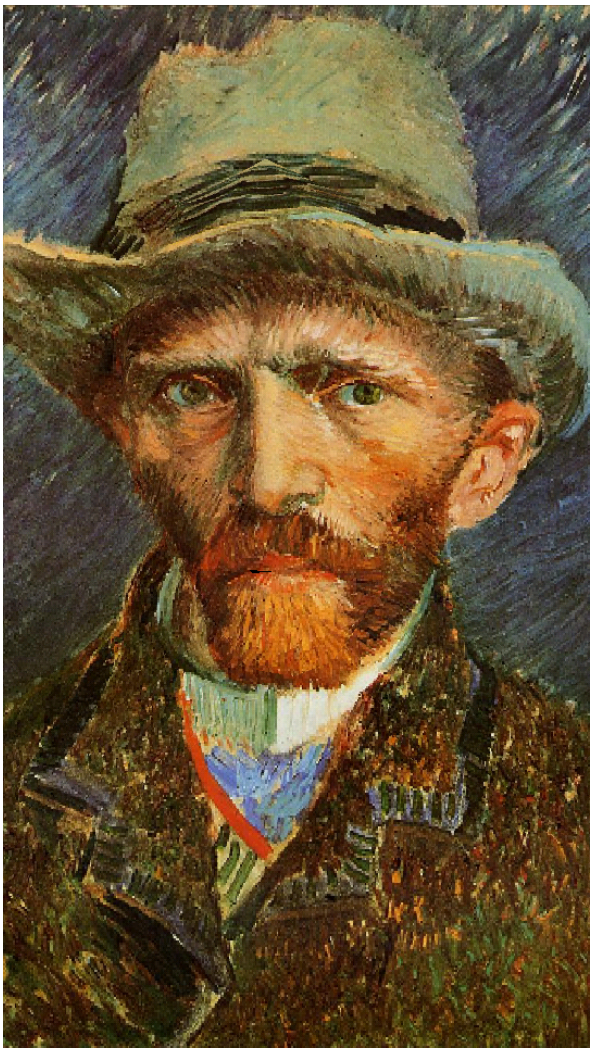


Figure 3.4. Tablet Head Van Gogh (Neutral State)

The mapping function contains simple mathematical equations such as finding the point which is between the two vertices and it is different for the each vertices. However, it is different for each points. When the 113 points are fully defined, with the help of OpenGL library, the picture is placed on the Candide 3 model as a texture.

Samsung Galaxy TabS [43] model tablet is used for the robotic platform. The main reason we use this tablet is the operating system of the tablet is Android OS. Android comes with the OpenGL ES support which allow using OpenGL and it is also supported by the Robot Operating System (ROS) [44] for the communication between the distributed systems.

3.3. Fritz Robotic Head

Our second robotic head platform is an animatronic type robotic head. In this head, we used an existing robotic head design which is Fritz [45]. Fritz robotic head was a kick starter project and the designer of this head promised that when the project is funded, the design model will be provided. The project is funded and the robotic head design files has become publicly available. The robotic head is designed for a laser cutter and the original Fritz is constructed using wood. However, with the correct thickness value, the files which are made for the laser cutter can be converted to the files for a 3D printer. The Fritz robot consist of 35 pieces (without electronic components and motors) and the all the pieces were printed using PLA filament.

The motor used in the Fritz's construction was the Tower Pro SG90 [46] which rotates up to 180 degrees and Futaba S3003 [47]. The joints except the neck joints in Fritz requires less than 180 degree rotation and less than 1 kg-cm torque. By considering these upper limits, Tower Pro SG90 is chosen to be used in these joints. However, for the neck joints, more torque is needed and in the neck, Futaba S3003 servo motors are used. The total DOF in the Fritz is 13. The distribution of the DOF is shown in Table 3.1.

Table 3.1. DOF Configuration of Fritz

Part	**DOF**
Neck	2
Eyebrow	2
Eye	4
Eyelids	2
Lip	2
Jaw	1
Total	13

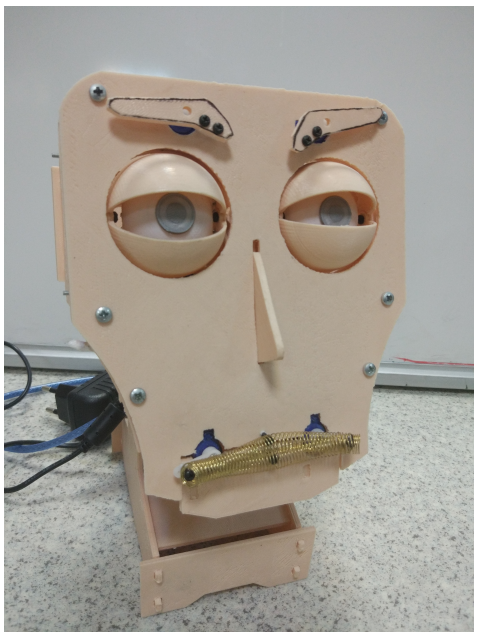


Figure 3.5. Fritz Robotic Head

The servos on the Fritz can be easily controlled by Arduino Uno [48]. The servo communication with Arduino is made by using PWM signals. Arduino has 6 PWM pins, but 20ms period is sufiicient for the communication of servos. Therefore, other digital pins of Arduino are used for servos. However, Arduino library only supports 12 servos at the same time. In order to overcome this problem, we wrote a new servo library. In that servo library, Arduino attaches the servo pins and after completing the action, Arduino releases the servo pin to attach a new servo pin whenever it is required. The limitation of the library is that 13 servos cannot be controlled at the same time. The library only controls 12 servos at the same time and the 13rd servo is rotated after the the first 12 servos.

Arduino UNO R3 is based on ATmega328P which is used with 16 MHz clock speed and 32 KB flash memory. As can be understood from these properties, Arduino has limited resources and it is not suitable for using complex algorithms and also in the distributed systems, communication with the Arduino requires Serial Communication or other shields which can add new features to the Arduino. For a safe communication and providing fast responses, we used the Robot Operating System (ROS) library which is written for the Arduino [49]. A *rosserial* packet is used for communicating with Arduino, this packet works as the middle point in the communication.

3.4. Boğaziçi University Social Robotic Assistant *Büşra*

Although Fritz is an animatronic robotic head, the number of DOF was determined to be too low for this study. The main issue, is that it does not have sufficient number of actuators in its lips for lip synchronization. In order to overcome that problem, we had to design a new robotic head which satisfies all the requirements.

3.4.1. Design Specifications

Büşra was designed using the Solidworks Software [50] as shown in Figure 3.6. The design can be printed via a 3D Printer. We used Flashforge Dreamer 3D printer [51] for printing the parts of the head. The maximum volume of a 3D printable part is 230

X 150 X 140mm and the maximum resolution is 80 microns. The head size of the *Büşra* robot is larger than that limitation. To cope with this problem, we divided the large part of the head into smaller pieces and we modified the design so that the smaller pieces can be connected by screwing all the parts.

In the early design stage of this head, we tried to consider human anatomy to determine the size of the head. Unfortunately, this attempt failed. The main reason for this problem is that reducing the size of head requires more complex design approaches and costly equipment. The first requirement for reducing the size can be handled by designing a Printed Circuit Board (PCB) for the electrical components of the head. In the current robot, we used Arduino Mega 2560 Rev3 board to drive the servo motors, but it is actually not a powerful board and other boards can also be used. However, with the ROS support and being easy to program, designing the PCB stayed out of options. The second way to reduce size say was by using smaller cameras.

The *Büşra* robot has two cameras in its eyes (Mini USB Camera). The resolution of each camera is 640 * 480 Pixels and the frame rate is 30 FPS. Its focus range is 20 mm and it is optimized for speaking 50 cm away from the robot. This camera does not have an auto-focus feature. Therefore, closer than 50cm, the image becomes blurred. We removed the cover of the camera in order to take up less space. However, the size of the camera was still large. The only way to reduce the size of the camera was buying a new camera, but these cameras are very costly components, so we removed this option, too. In the end, the only used parameters in the human anatomy is that the eyes are in the middle of the face.

The main requirement for this study is lip synchronization. To perform lip synchronization, the robot should have at least 5 DOF [52] in its lips. For this reason, we added 7 DOF in the *Büşra* head which is more than enough. The real reason for putting more DOF to lips is the jaw mechanism in its design. The jaw mechanism is divided into upper and lower lip mechanisms and instead of using the same motor for the horizontal movement of the lips, separate motors drive the lips.

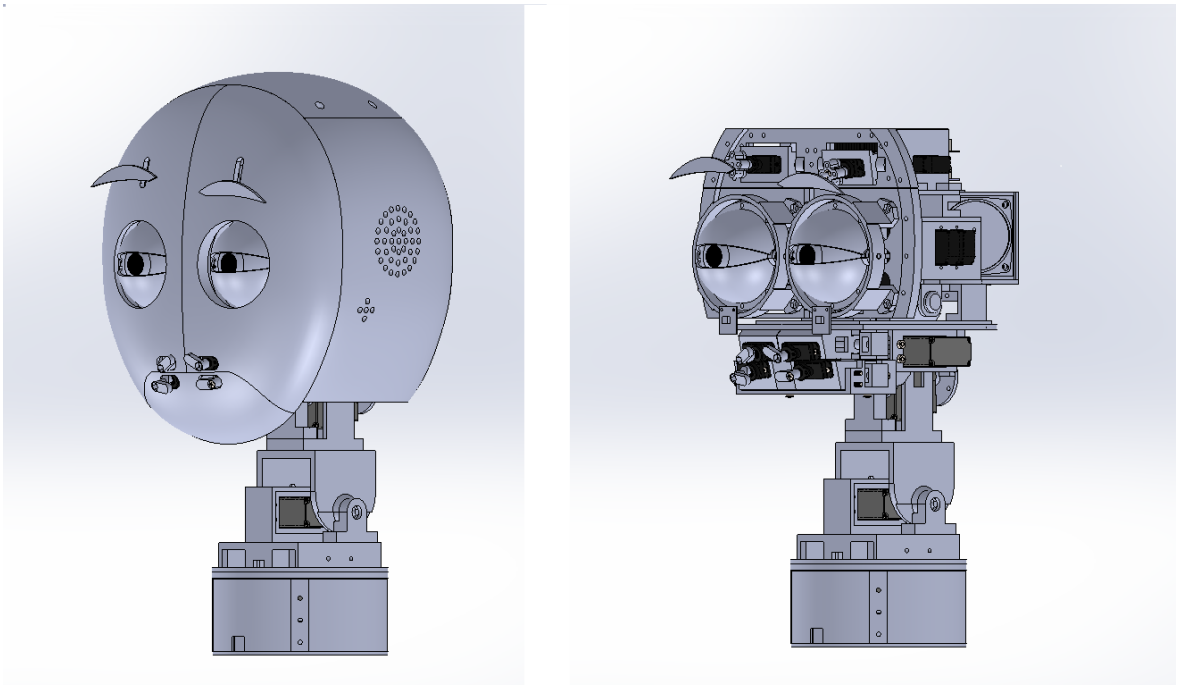


Figure 3.6. CAD model of Büşra with cover and without cover

In addition to the lip mechanism, we have a requirement that the robots should be able to show 6 core emotion states [53] as shown in Figure 3.7. To satisfy this requirement, the total number of DOF was increased to 22.

The 22-DOF of *Büşra* are given in Table 3.2. In the eyes, eyebrows, eyelids, neck, lips and jaw mechanisms, we used two types of servo. One of them is EMax ES08MAII servo motor [54] and the other one is Tower Pro MG946R servo motor [55]. These motors are chosen for their small size and low prices. For *Büşra*, the resolution of the movement and the precise motion of the joints are disregarded. The two of them can work at the same voltage which is 4.8V - 6.0V.

EMax ES08MAII is a mini type servo with 2.0 kg-cm torque. Tower Pro MG946R is a standard type servo with 13 kg-cm torque. 2.0 kg-cm torque is sufficient for most of the mechanisms in our design except the neck and jaw mechanism. The neck and jaw mechanism require more torque to carry the related systems. Therefore Tower Pro MG946R is used for these mechanisms.

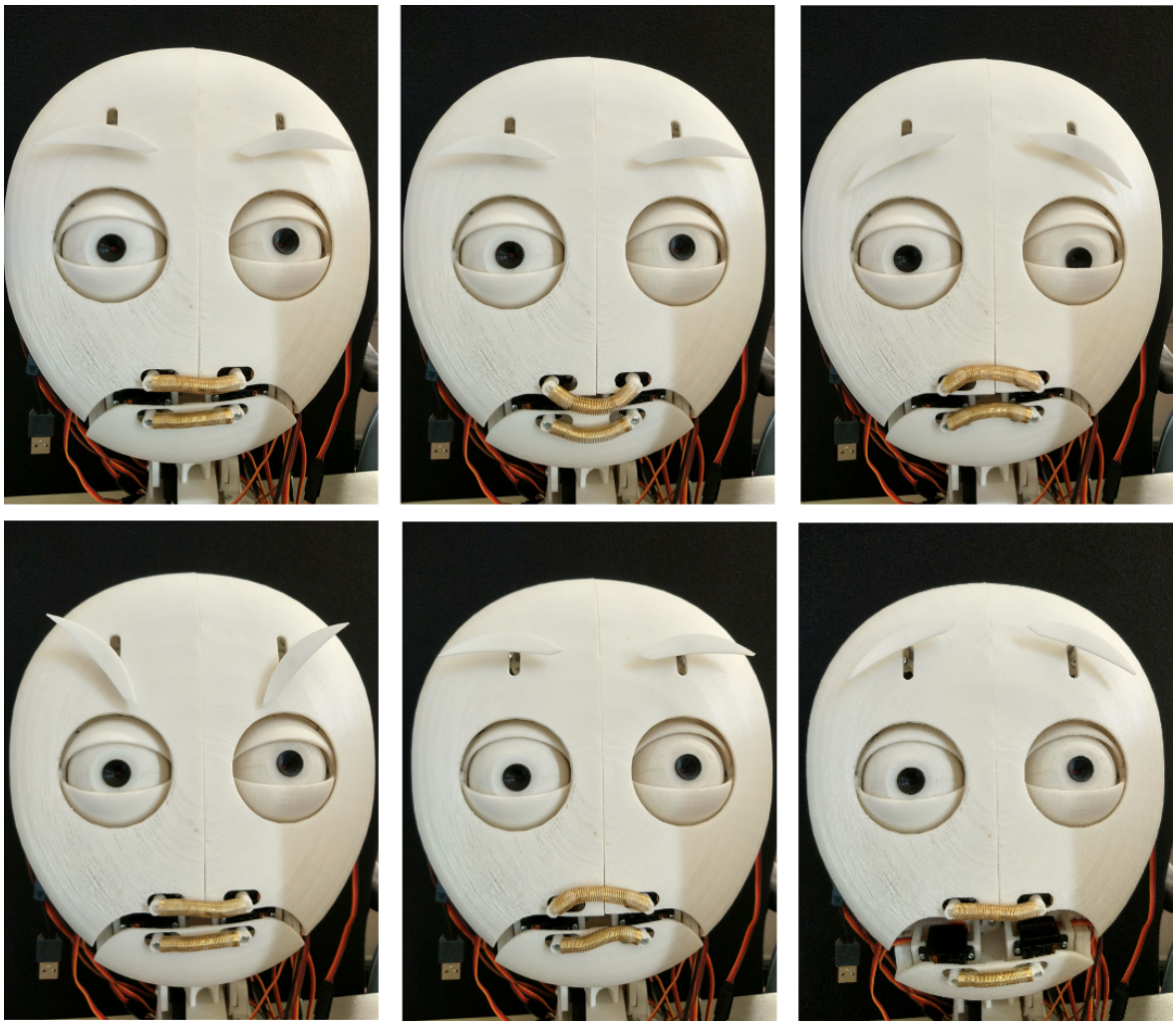


Figure 3.7. From Top-Left to Bottom-Right: Neutral, Happiness, Sadness, Anger, Fear and Surprise

Table 3.2. DOF Configuration of *Büşra*

Part	**DOF**
Neck	3
Eyes	4
Eyelids	4
Eyebrows	4
Lips	6
Jaw	1
Total	22

3.4.2. Mechanical Specifications

3.4.2.1. Neck. *Büşra*'s neck mechanism which is shown in Figure 3.8 has 3-DOF which are the pitch, the roll and the yaw axis. Instead of a harmonic drive system, we used a servo neck system. Harmonic drive systems provide more realistic neck movements. However, servo neck system is cheaper and can easily be printed. Maximum angular velocity of each axis is the same as the maximum angular velocity of servo motors which is 350 deg/s which is faster than human neck speed. Also the angular range of the axis is the same as the used servo motors which is 60 deg. The maximum torque requirement for the neck occurs for the pitch axis. Since the cover for the back of the head is not finished, the center of gravity of the head of the head is much closer than the actual design and this causes excessive stress on the motor.

As a future work, we plan to use more powerful motors for the pitch and roll axis.

3.4.2.2. Eyes. The eye mechanism (in Figure 3.9) has 2-DOF, one is for the pitch axis and the other for the yaw axis. With these DOF, the robot has the flexibility to move its eyes in a way which cannot be performed by humans. Each of the motors which are for the yaw axis can have different degree values, but that makes the robot look

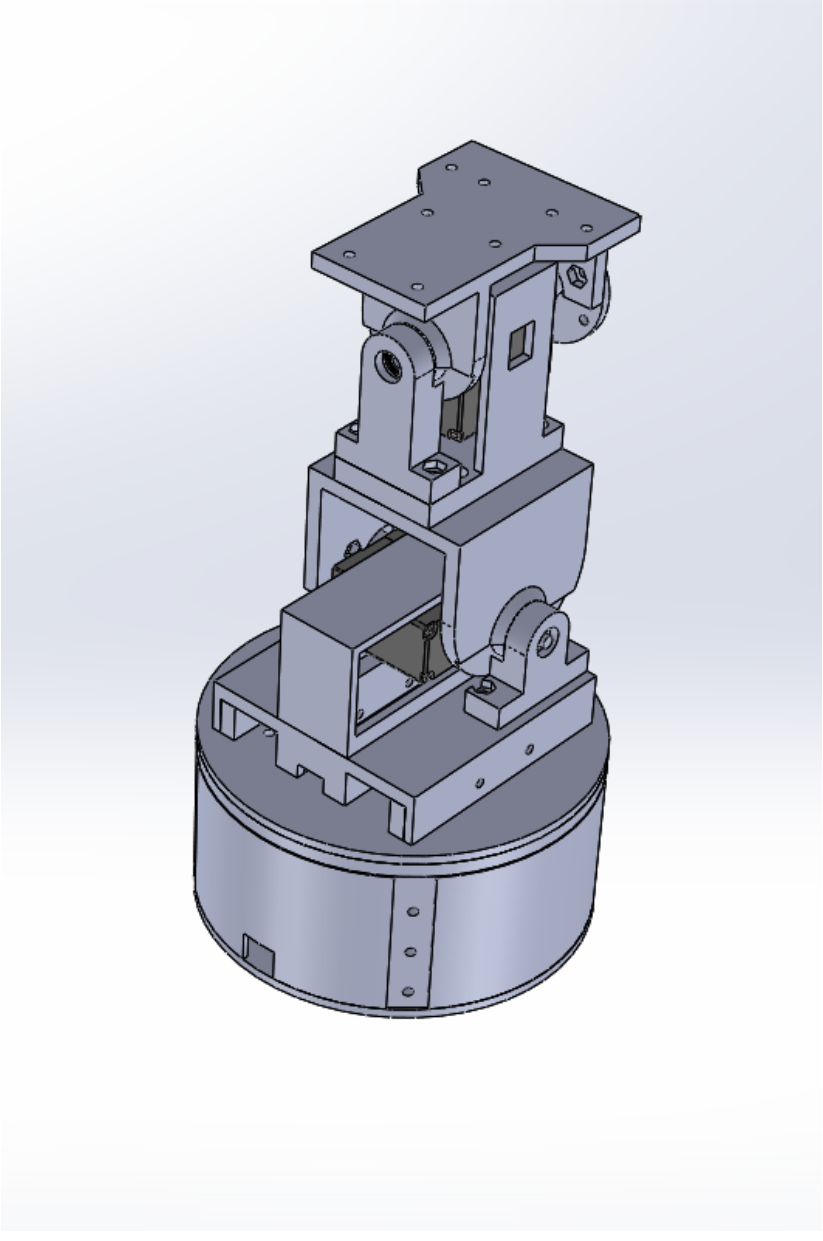


Figure 3.8. The neck mechanism model of the Büşra

creepy. Consequently, we applied the same degree values for each yaw axis.

In the eye mechanism, the pitch and yaw axes are driven by a tendon mechanism. 1mm stainless steel wires are used for push and pull actions. While performing push and pull actions, friction can cause twisting in the wires. Therefore, we designed a special 2-DOF joint in order to decrease the friction which is suitable for printing, with bearings.

3.4.2.3. Eyelids. The eyelid part consists of an upper eyelid and a lower eyelid. Each eyelid has 1-DOF for opening and closing actions which in total makes 4-DOF. This mechanism is also driven by a tendon mechanism and the same type of wire is used for all the tendon mechanisms.

The upper eyelid and lower eyelid mechanisms have different range of actions. The upper eyelid can perform more movement than lower one because humans also make that kind of movement, their upper eyelids move more than the lower ones. By doing this kind of behavior, sadness or embarrassment state for the robot can be expressed.

3.4.2.4. Eyebrows. In the eyebrows parts, a total of 4-DOF are used. These are for the roll and pitch axes. In this system, we also used a tendon mechanism, but not for the movements of the roll axis. For the roll axis, a stainless steel rod which is 27mm length and 3mm radius is directly connected to the head of the servo.

The movements for the roll axis is used to the express the anger, sadness emotions. On the other hand, the pitch movement is for the surprise emotion.

3.4.2.5. Lips. *Büşra* should show its emotional state while it should also perform lip synchronization for Turkish. For this purpose, the robot has 6-DOF (shown in Figure 3.10) in its lip mechanism. In the upper lips, 1-DOF is used for bringing the lips closer and the 2-DOF is used for rotating each lip in the roll axis. For the lower lips, the

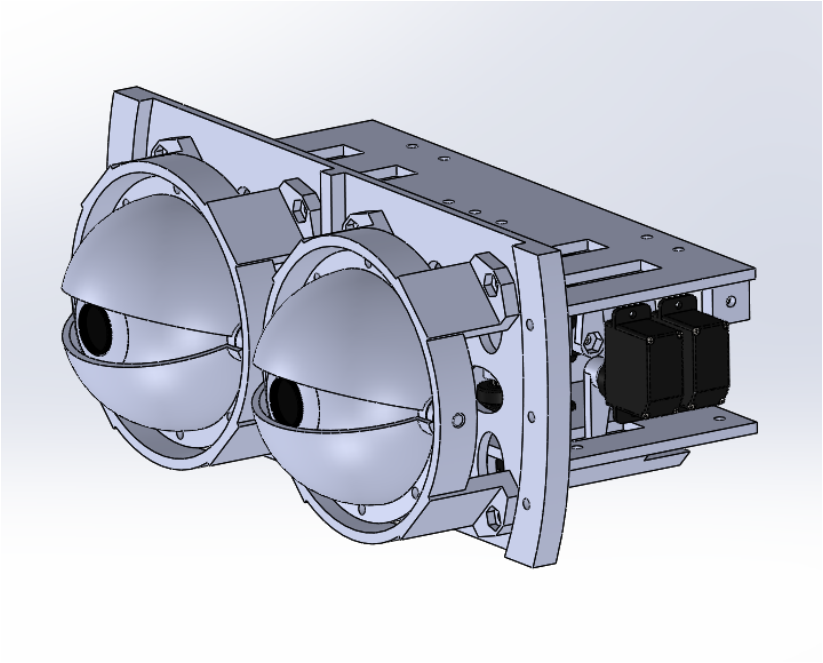


Figure 3.9. The eye and eyelid mechanism model of the Büşra

purpose of the DOFs are the same.

For the emotional states, rotating the lips is enough. However, to perform the lip action for some letter, lip locations should also be changed.

3.4.2.6. Jaw. For the jaw mechanism, only 1-DOF is used and it is for opening and closing actions. Unfortunately, due to some design error, jaw is not closing as planned. There is a small gap between the upper and lower lips when the jaw is closed.

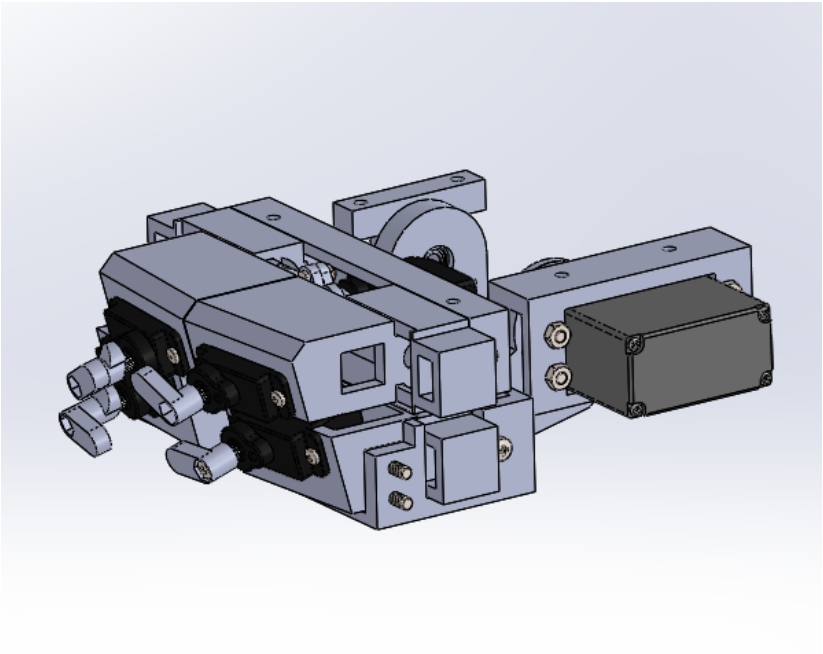


Figure 3.10. The lips and jaw mechanism model of the Büşra

4. LIP SYNCHRONIZATION

In this study, we worked on the Lip Synchronization system in order to improve the human robot interaction and communication. According to McGurk [56], when the lip action and the speech does not match with each other, the listener may recognize a different speech and that speech is neither the spoken nor the lip synchronized speech.

Most of the lip synchronization systems are based on the English language. However, the constructed robot head will be used with the Turkish language for speaking. Therefore, the lip positions should be classified for the Turkish language. In [57], the lip shapes are classified into six classes:

- Silence, Rest Position
- b, m, p
- c, ç, d, g, h, k, n, r, s, ş, t, y, z
- a, e,
- ı, i, l
- o, ö,
- u, ü,
- f, v

The algorithm which is used for all of the three heads are the same, but the way they are implemented is different. The used algorithm is simple. The speech synthesis tool while generating the speech wave form, also generates an additional message which contains the duration information of the letters. This message has four fields. The first one is the character number of the generated speech. The second one is the total duration of the speech and the third one is text data and the last field contains an array that contains the start duration indexes of all the letters. By using these parameters, the lip synchronization system just makes the transition between the letters.

The center locations of a letter in the time axis are calculated by using Equation 4.1 :

$$D_{center} = \frac{D_i - D_{i-1}}{2} + D_{i-1} \tag{4.1}$$

After calculating the center of the letter in the time axis, the transitions are added.

4.1. Van Gogh Lip Synchronization

The basic animations such as changes in the emotional states are provided by the Candide 3 model with the help of the given *Action Units*, but the *Action Units* are not sufficient to make a lip synchronization system. The model has a limited number of *Action Units* and although it included some of the lip actions for English, it still did not cover all the lip actions.

For the previously mentioned the eight lip classes, we added seven new Action Units to the model except for the silence position. Also for each *Action Unit*, the locations of the vertices are determined manually.

The frame rate of the Lip Synchronization system on the tablet is fixed at 30 fps which can be regarded as sufficient and the transition between the vertex locations are made to perform lip actions as shown in Figure 4.1.

4.2. Fritz Lip Synchronization

Unlike the tablet head, Fritz which is a Kinematic head, has hardware limitations. It only has 3 DOF for lip synchronization. Two of the DOF is located to left and right lips and the rest of the DOF is serviced the mouth motion.

Figure 4.1. Tablet Head Van Gogh at the rest position (Left) and performing the lip action for letter A (Right)

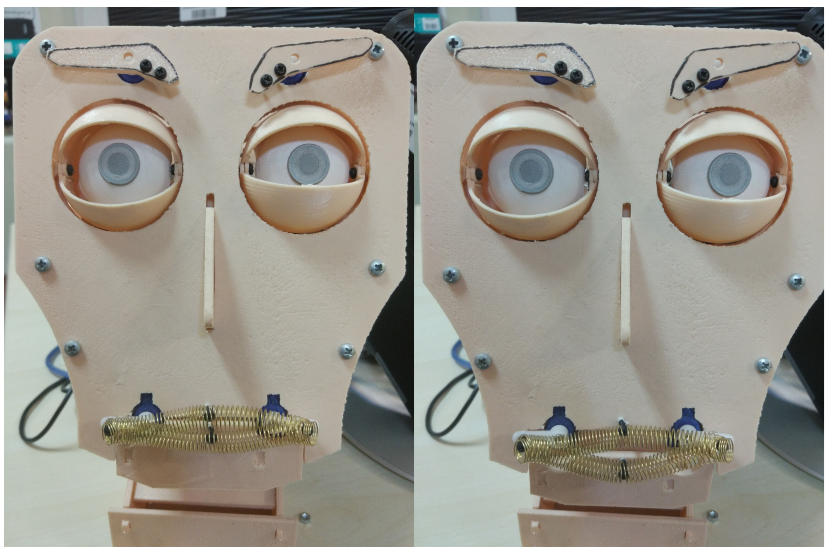


Figure 4.2. Fritz Robotic Head at the rest position (Left) and performing the lip action for the A letter (Right)

The number of the DOF are not enough for synthesizing the correct lip motions of a letter. However, we made a simple lip synchronization at least to give the speaking effect for the user as shown in Figure 4.2. This synchronization method considers only the vowels. For all consonants, it gives the same lip motions, but for the vowel case it changes the opening amplitude of mouth according to vowel.

4.3. *Büşra* Lip Synchronization

Büşra has 7-DOF for lip synchronization. As we mentioned earlier, 6-DOF is used for the lips and 1-DOF is for jaw movements. The designated speaking language for the *Büşra* is Turkish. The lip design is not only specified for the speaking the Turkish language, but also it can make the lip actions of the other languages. However, for this thesis, we made a system for the Turkish language. In Turkish, lip shapes and jaw positions are classified into eight groups, but for the *Büşra*, for the *(f, v)* letter class, we used *(b, m, p)* lip action. The reason for this change is f and v letters require front movement on the upper lips. In order to avoid that kind of movement, we merged this class with the closest class.

The lip synchronization system is very similar to that of the tablet head. The main difference is for *Büşra*, letters are not mapped to vertex positions, instead mapped to the servo angles.

The update rate of the servo angles are fixed as 50 Hz and there is no transition function implemented for *Büşra*. *Büşra* uses physical objects to move its lips positions. Therefore, each motor has a limited amount of angular velocity. Therefore, even if the different angle value is commented to the motor, the motor will not rotate to that angle instantly. Therefore, in *Büşra*, the motor limitations generates the lip position transitions in the lip synchronization as shown in Figure 4.3.

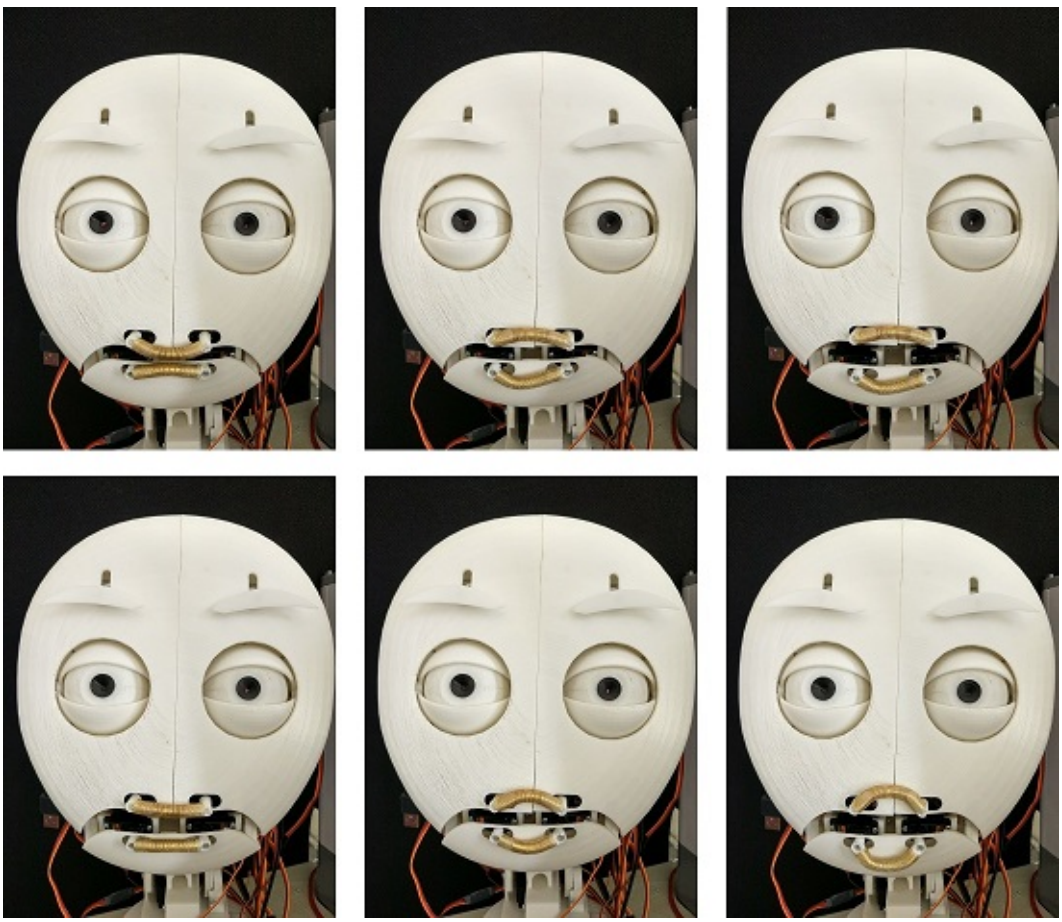


Figure 4.3. Letter Sets from Top-Left to Bottom-Right: (b, m, p), (c, ç, d, g, h, k, n, r, s, t, y, z), (a, e), (ı, i, l), (o, ö), (u, ü)

5. SPEECH SYNTHESIS

In this study, among all the speech synthesis options, the *Formant Speech Synthesis* option mentioned in Section 2.2.2 is selected. The main reason for using this option is the flexibility of the system. The other methods need huge amounts of training data and the training data should be recorded in a specific room for avoiding the noise of the outside environment and need to be labeled in order to classify the parameters of the sounds. Also the emotion state of the speaker will affect the recorded sounds and if the speaker's emotion state is not stable, the end result will not be satisfactory and the generated speech may contain different emotion states in a signal word.

Our Speech Synthesis system uses Turkish language and for Turkish language, we cannot get enough data for training the other methods. The only option for our case is using a system which requires less training data. Therefore, the *Formant Speech Synthesis* becomes the best choice with its flexibility and low number of training data requirement.

The Formant Speech Synthesis which is named as *BounTalk* in this study is divided into two parts. The first part is the wave generator and the second part is the parameter generator.

5.1. Wave Generator

The wave generator part is actually a converter which produces the speech file by using special parameters which are defined below. *BounTalk* wave generator requires 68 different parameters in order to generate one frame of speech. The overall system for Wave Generator is shown in Figure 5.1.

The 22 parameters of the 68 parameters which are shown in Table 5.1, are related to the glottal source generation and in Figure 5.2, a simple glottal source is shown.

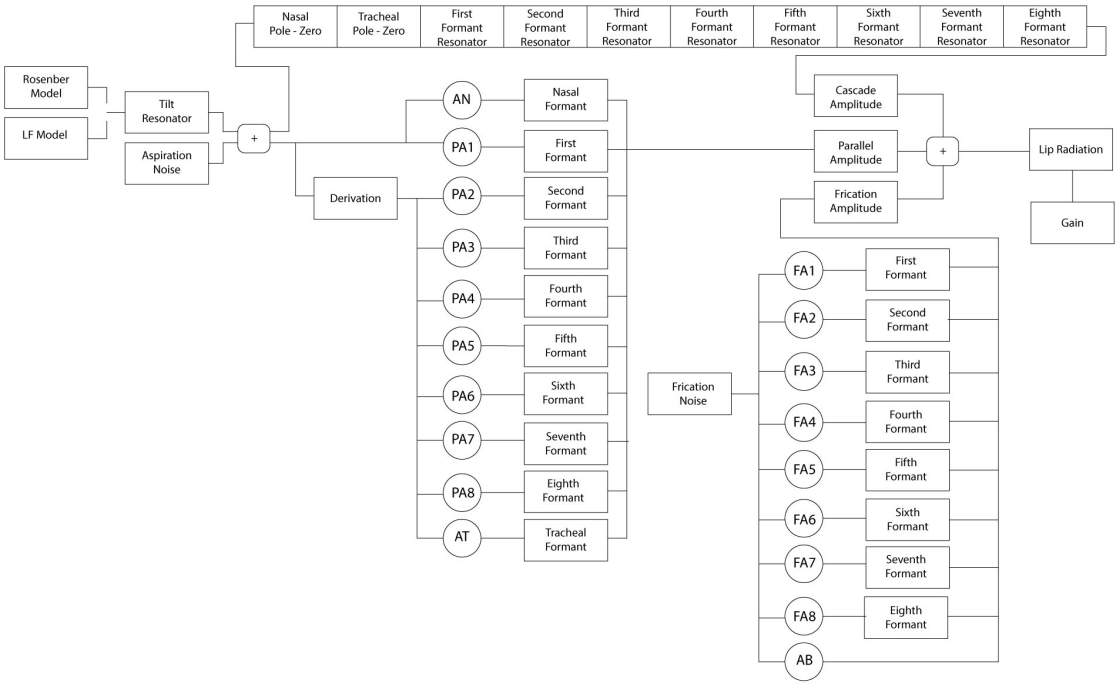


Figure 5.1. BounTalk Wave Generator

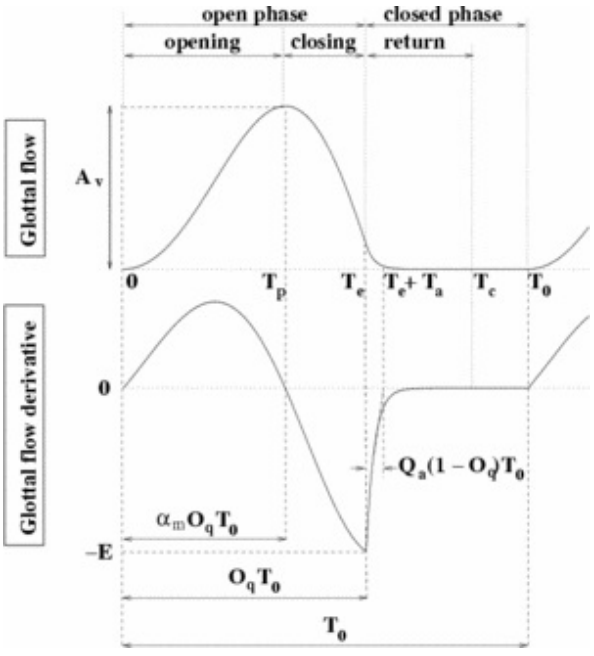


Figure 5.2. Glottal Source of the Speech from [58]

Table 5.1. Glottal Related Parameters and Description of the BounTalk Wave Generator

TIME	Frame Time
F0Hz10	Fundamental Frequency in decihertz
F0Fl100	Flutter Parameter for Fundamental Frequency
GTP100	Opening Phase Ratio in Glottal
GTC100	Opening and Closing Phase Ratio in Glottal
GLFEE	Liljencrants-Fant EE parameter
Ra1000	Liljencrants-Fant Ra parameter
Rk100	Liljencrants-Fant Rk parameter
Rg100	Liljencrants-Fant Rg parameter
TiltDB	Glottal Tilt parameter
ABreDB	Glotta Breathless parameter
AAspDB	Glotta Noise parameter
NaZFHZ	Nasal Zero Frequency
NaZBHZ	Nasal Zero Bandwidth
NaPFHZ	Nasal Pole Frequency
NaPBHZ	Nasal Pole Bandwidth
NaPDB	Nasal Parallel Amplitude
ThZFHZ	Tracheal Zero Frequency
ThZBHZ	Tracheal Zero Bandwidth
ThPFHZ	Tracheal Pole Frequency
ThPBHZ	Tracheal Pole Bandwidth
ThPDB	Tracheal Parallel Amplitude

The TIME parameter is for organizing the frames of the speech. *BounTalk* Wave Generator supports three frame size options which are 2, 4 and 5 milliseconds. Every generated speech frame can only have 2, 4, or 5 milliseconds duration and the previously mentioned 68 parameters are required for generating that long speech. Therefore, each frame should be increased by these values.

The F0Hz10 and F0Fl100 parameters are used to determine the fundamental frequency of the speech. F0Hz10 contains the fundamental frequency hertz which is stored in decihertz units. F0Fl100 is a flutter parameter which adds the noise to the fundamental frequency. In the human sound, the voice frequencies do not remain the same. The muscle on the glottal vibrates air, but the vibration amount changes because of the muscle movements and this adds a noise to the speech.

The GTP100 and GTC100 parameters are generally used for generating source of the speech. The fundamental frequency of the signal is actually the pulse frequency of the speech. However, it does not include any glottal source. In order to generate the glottal source signal, we used two main glottal source model approaches. The first one is the Rosenberg and the other is the Liljencrants-Fant glottal model. In these models, commonly glottal is divided into two phases which are the opening phase and the closing phase. The GTC100 parameter contains information about the total duration of these phases as the rate of the fundamental period. The GTP100 parameter on the other hand, keeps the opening phase ratio in the total duration. With using these two values, we can generate the Rosenberg glottal model, but for the Liljencrants-Fant model, we needed to use additional four parameters. Rosenberg source is calculated as given in Equations 5.1-5.3 [59]:

$$g(t) = \frac{A}{2}[1 - cos(\pi \frac{t}{To})] \quad for \quad 0 \leq t < T_o \tag{5.1}$$

$$g(t) = Acos(\frac{\pi(t - T_o)}{2T_c})] \quad for \quad T_0 \leq t < T_c \tag{5.2}$$

$$g(t) = 0 \qquad\qquad for \quad T_c \leq t < T \tag{5.3}$$

The GLFEE, Ra1000, Rk100 and Rg100 parameters are used for generating the glottal source when the source model is selected as Liljencrants-Fant. The Liljencrants-Fant source can be calculated as given in Equations 5.4 and 5.5 [60]:

$$g(t) = E_0 \epsilon^{\alpha t} \sin(\omega_g t), \quad 0 \leq t < T_e \tag{5.4}$$

$$g(t) = -\frac{E_e}{\epsilon T_a}[e^{-\epsilon(t-T_e)} - e^{-\epsilon(T_c-T_e)}] \quad for \quad T_e \leq t \leq T_c \tag{5.5}$$

The Ra, Rk and Rg parameters are given in Equations 5.6-5.8 [61]:

$$R_g = \frac{T_0}{2T_p} \tag{5.6}$$

$$R_k = \frac{T_e - T_p}{T_p} \tag{5.7}$$

$$Ra = \frac{T_a}{T_0} \tag{5.8}$$

TiltDB is a parameter for adjusting the returning phase of the glottal source. Liljencrants-Fant model has its own returning phase generator. However when the Liljencrants-Fant's offered version is not sufficient or Rosenberg model is used and returning phase is required, this parameter is used.

ABreDB is added to give breathy effect to the glottal. The glottal source contains three states which are breathy, normal and pressed. Actually, we did not use pressed and breathy states for the glottal, but for future use, we added this parameter.

AAspDB is noise parameter for the glottal. When the glottal source is produced because of the friction of the air in the glottal area a new noise is added glottal source.

For the formant frequencies, the resonator is used from the Klatt Formant Synthesis which is formulated as given in Equation 5.9 [30]:

$$T(f) = \frac{A}{1 - Bz^{-1} - Cz{-}2} \tag{5.9}$$

where A, B and C are given in Equation 5.10 [30]:

$$A = 1 - C - B; \quad B = 2e^{-\pi BWT}cos(2\pi FT); \quad C = -e^{-2\pi BWT} \tag{5.10}$$

For the vocal track part, 24 parameters are used which is shown in Table 5.2.

The formant frequency and the bandwidth are used for the calculation of the resonator for the cascade formant synthesis. However, for the parallel formant synthesis, this is not sufficient, and an additional amplitude value is needed. For this case, the vocal track is generated via cascade or parallel method. However, the merging the output of these two methods is done by the cascade and parallel decibel values.

We separated the frictional consonant generation from the vocal track generation. Therefore, we used additional parameters for this purpose. The resonator for the frictional consonant uses the same frequency and bandwidth with the vocal track, but instead of using parallel amplitude, it uses the friction amplitude (given in Table 5.3).

In order to generate natural voices, a noise signal should be added to the amplitudes and also the frequencies and added noise should have a special characteristic, randomly generated noises can decrease the voice quality. The used flutter noise is a sinusoidal signal with a period of 10 microseconds and the flutter values which are shown in Table 5.4 are the amplitude values of the noise.

The last parameter set (Table 5.5) is for the decibel control of the generated sound. Cascade Amplitude determines the amplitude of the cascade vocal track method and Parallel Amplitude determines the amplitude of the parallel vocal track method. Also Friction Amplitude is used for the made frictional voices. The Gain value determines the volume of the sound and Pre-emphasis is used for lip radiation.

5.2. Parameter Generator

The main purpose of the Parameter Generator is to correctly analyze the given text and convert it to the parameters. In this parameter generator phase, we labeled the training data of the users in JSON files. For each user, a separate JSON file is written according to their recorded sounds.

Table 5.2. Vocal Track Related Parameters and Description of the BounTalk Wave Generator

FO1FHZ	Formant 1 Frequency
FO2FHZ	Formant 2 Frequency
FO3FHZ	Formant 3 Frequency
FO4FHZ	Formant 4 Frequency
FO5FHZ	Formant 5 Frequency
FO6FHZ	Formant 6 Frequency
FO7FHZ	Formant 7 Frequency
FO8FHZ	Formant 8 Frequency
FO1BHZ	Formant 1 Bandwidth
FO2BHZ	Formant 2 Bandwidth
FO3BHZ	Formant 3 Bandwidth
FO4BHZ	Formant 4 Bandwidth
FO5BHZ	Formant 5 Bandwidth
FO6BHZ	Formant 6 Bandwidth
FO7BHZ	Formant 7 Bandwidth
FO8BHZ	Formant 8 Bandwidth
FO1PDB	Formant 1 Parallel Amplitude
FO2PDB	Formant 2 Parallel Amplitude
FO3PDB	Formant 3 Parallel Amplitude
FO4PDB	Formant 4 Parallel Amplitude
FO5PDB	Formant 5 Parallel Amplitude
FO6PDB	Formant 6 Parallel Amplitude
FO7PDB	Formant 7 Parallel Amplitude
FO8PDB	Formant 8 Parallel Amplitude

Table 5.3. Friction Related Parameters and Description of the BounTalk Wave Generator

FO1FDB	Formant 1 Friction Amplitude
FO2FDB	Formant 2 Friction Amplitude
FO3FDB	Formant 3 Friction Amplitude
FO4FDB	Formant 4 Friction Amplitude
FO5FDB	Formant 5 Friction Amplitude
FO6FDB	Formant 6 Friction Amplitude
FO7FDB	Formant 7 Friction Amplitude
FO8FDB	Formant 8 Friction Amplitude

Table 5.4. Flutter Related Parameters and Description of the BounTalk Wave Generator

FO1F100	Formant 1 Flutter
FO2F100	Formant 2 Flutter
FO3F100	Formant 3 Flutter
FO4F100	Formant 4 Flutter
FO5F100	Formant 5 Flutter
FO6F100	Formant 6 Flutter
FO7F100	Formant 7 Flutter
FO8F100	Formant 8 Flutter

Table 5.5. Amplitude Related Parameters and Description of the BounTalk Wave Generator

AVCasDB	Cascade Amplitude
AVParDB	Parallel Amplitude
AFricDB	Friction Amplitude
ByPasDB	By pass Amplitude
GainDB	Gain
PRE100	Pre-emphasis

The most important part for generating this JSON file is extracting the parameters from the recorded training data. Even though the Formant Synthesis method requires less training data, the data which include all the phonemes of the language is required. In this case, we used five recorded sounds while reading five different sentences. For the sentence choice, we used pangram sentences which include all the letters of a language. Turkish is a language where every written letter has its own phoneme, so the pangram sentences contain all the phoneme information about the Turkish language. However, we also need the transition information about the letters. The syllable structure for the languages contain the onset and rhyme and the rhyme contains the nucleus and coda which is shown in Figure 5.3. For the nucleus, we have eight vowels (a, e, ı, i, o, ö, u, ü), but for the onset and coda, we should get the letter transitions when the letter is used in the onset position and when the letter is used in coda position. For this case, five sentences were chosen to cover all the onset - nucleus and nucleus - coda transitions. The used sentences are as follows:

- Pijamalı hasta yağız şoföre çabucak güvendi.
- Fütursuz kaçığa göre japon şemsiyeleri bedava, hacı
- Tüh, jantı feda et; pislik böceği yormuş, vazgeç.
- Bu ganj öküzü hapis düştü yavrum, ocağı felç.
- Ne abbac ne addaf sadece yayla gag yap.

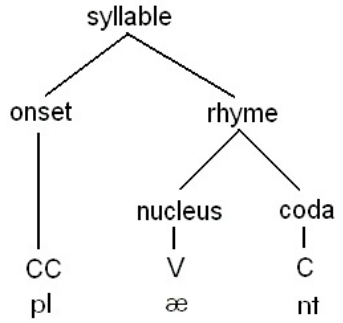


Figure 5.3. Syllable Structure from [62]

The JSON file contains six different sets of categories. The first category contains the information about the letter. The formant frequencies, bandwidths and amplitude values of the letter is written in that location. The second category contains the glottal parameters. The parameter generator uses the fixed values for the glottal. Therefore, the glottal parameters are taken from this category. The General Speech Parameters category contains the initial silent duration, with the volume of the speech. The Flutter Parameters category is where the flutter noise amount is determined. The fifth category contains the letter type parameters. For example, the start fundamental frequency of the vowel is determined in this category. The sixth category is for the transitions. Every transition information is stored in that category.

The transition codes are manually written according to the rules of the transition. Generally, the default transitions method, which makes the change of the formant frequencies and the amplitude linearly is used. However, for the plosive consonants, the transitions are not suitable for the linear changes, instead they need a burst on the amplitudes. Therefore, for these kind of transitions, we changed the amplitude of the formant exponentially.

For the frictional consonants, the parallel amplitude of the formants are not used, but in the transitions, (for example voiceless fricative - vowel transition) the parallel amplitude is only used at the end of the transition.

For speech synthesis, only one voice is used. However, we need to generate female and child voice with male voice. [63] shows that the formant frequencies are changed for the male, female and child voices with the fundamental frequency. In order to change the male voice to female and child, we also changed the formant frequencies and the fundamental frequency. The amount of the changes in the frequencies are shown in Table 5.6.

Table 5.6. The frequency changes for the female and child voices

	Fundamental Frequency	Formant Frequencies
Female Voice	+25	+200
Child Voice	+125	+250

6. TURKISH NATURAL LANGUAGE PROCESSING TOOL

Turkish language is a language with an irregular sentence structure. In a sentence the predicate can be found at the very beginning, however it can also be found at the end in other sentences. Besides this irregularity, since it is an agglutinative language, there is no limit to the number of suffixes words can take. Because of all these difficulties, before creating any conversation, sentences should be subjected to a preliminary processing. For the preliminary processing, we prepared some NLP tools which are customized for Turkish language.

6.1. Root and Suffix Finder

Before extracting any sense in the sentences, all the words should be examined separately. In this part, we try to estimate the possible root and possible suffix combinations.

In order to make the root estimation, we needed to have the Turkish root database. Firstly, we used an open source dictionary database, but this database was not enough for root finding. In Turkish, there are many exceptional cases. and without knowing these exceptions and which words are related to these exceptions, the root finder will not work properly. In order to overcome this problem, we used the dictionary which is provided from Zemberek [64]. Zemberek is an open source NLP Framework for Turkish language. However, in our study we cannot use this tool due to compatibility issues. We developed our system by using C++ language and in order to use C++ strictly, we decided not to use that library, instead we use the data of this library.

The dictionary file which is taken from Zemberek contains the roots with their exceptional cases. For example, *cehit* word has SESSIZ-YUMUSAMASI and ISIM-SESLI-DUSMESI special cases and when a suffix which starts with a vowel is added

to the this word *cehit* becomes *cehdi* (i suffix is added). By using that dictionary file, we generated every root word form with the special cases and matched with the words and with this method, we found the possible roots of the words.

Root finding is the first step for analyzing the word. The possible suffixes should also be found. For the suffixes, a dictionary file is also needed. We took the suffix file from Zemberek, too. In Zemberek, suffixes are written in the XML file. In order to make a finite state machine for the suffixes, we consider the fact that every suffix contains the next possible suffix information with regular expression with special cases. It was a very unique source which covers almost all the Turkish suffixes. With the help of that file, we extracted the possible suffix sets for the words. Unfortunately, there are many ambiguous cases in Turkish, for example *suyu* word can be analyzed as *su (y) u (hal eki)* also s*u (y) u (tamlanan eki).*

6.2. Relation Finder

After the word analysis, we implemented a tool for sentence analysis. In this tool, we give more priority on the verbs and the adjectives. The possible verb of the sentence is found by examining the last suffix. If the last suffix of the word is a suffix that can only be added to the verbs and it does not change the word type, we consider that word as a verb of the sentence. For the adjectives, we used this method too. After finding the adjectives, we added a relation between the adjective with the noun which is next to the it. After that, genitival and determinated word relations are found. The possible words which can have genitival suffix are related to the words which are determinated suffix.

Unfortunately, the method which is used in the relation finder is not a perfect method, but with this method, we can extract almost all the adjective - noun and noun - noun relations from the small regular languages.

6.3. Chatbot Engine

For answering the questions of the user, we developed a system based on the Artificial Intelligence Markup Language (AIML). AIML is prepared according to the game which will be played in the experiment. The game has limited vocabulary and a limited number of question types. Also, the answers of the questions are very deterministic. Therefore, AIML becomes the best choice for this small game.

7. EXPERIMENTS AND RESULTS

7.1. The Aim of the Experiment

We want to make a robotic platform for good quality communication with people. However, before completing its design, we need to ensure that the requirements for the human - robot interaction are met. Therefore, we need to answer the following questions:

- Do humans like to interact with the robots?
- Do humans want complex robotic heads?
- Do humans pay attention to the lip movement of robots?
- Do humans match the voice of the robot with its appearance?
- Do humans wants to play a game with the robot and which games they want to play?

In order to answer all of these questions, we developed components in this study, we established a basic scenario and this scenario includes the three robot heads, the speech synthesis system with lip synchronization and the Turkish NLP tool.

7.2. The Experimental Platform

This experiment consists of three robotic heads which are explained previously, one speech recognition device, one computer, and one gaming platform.

In the experiments, a simple child's toy called the *Guess Who* game is used (more information about the game is given in Appendix A). The gaming platform is a completely isolated platform and it does not give any input to the system or take the output of the system. This platform is only for helping the user with the game.

The platform is shown in Figure 7.1, with the participant in Figure 7.2 and the communication system for the experiment is shown in Figure 7.3.

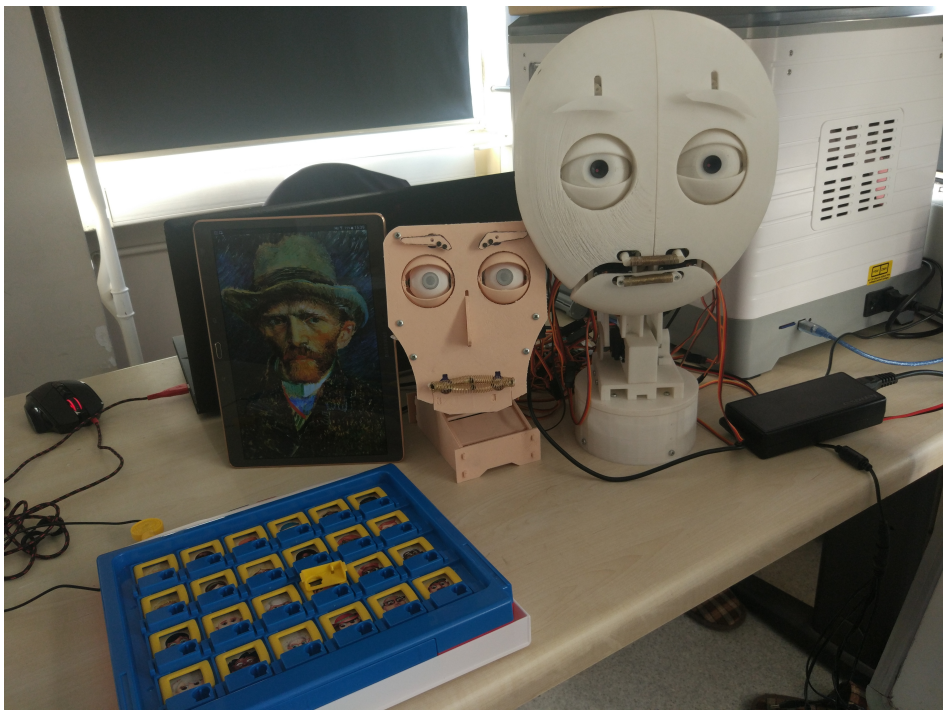


Figure 7.1. The Experimental Platform

For the software platform, Robot Operating System (ROS) is used to link all the devices and to control them synchronously. The core of the ROS is started on the computer. The computer is configured for receiving a static IP, so the *roscore* module becomes reachable across the network.

The speech synthesis system and the AI system of the robots are also commanded via the ROS and these systems run on the computer without needing extra hardware platform.

The speech synthesis system executes a ROS service server in its code and listens to the client's requests. The message which is used for this service is a custom message which consists of the standard messages shown in the Table 7.1. Whenever this service is called, it produces a response message with a WAV file which is playable by the other nodes. For this experiment, the WAV file is only played by the AI component of the

Figure 7.2. The Experimental Platform with a participant

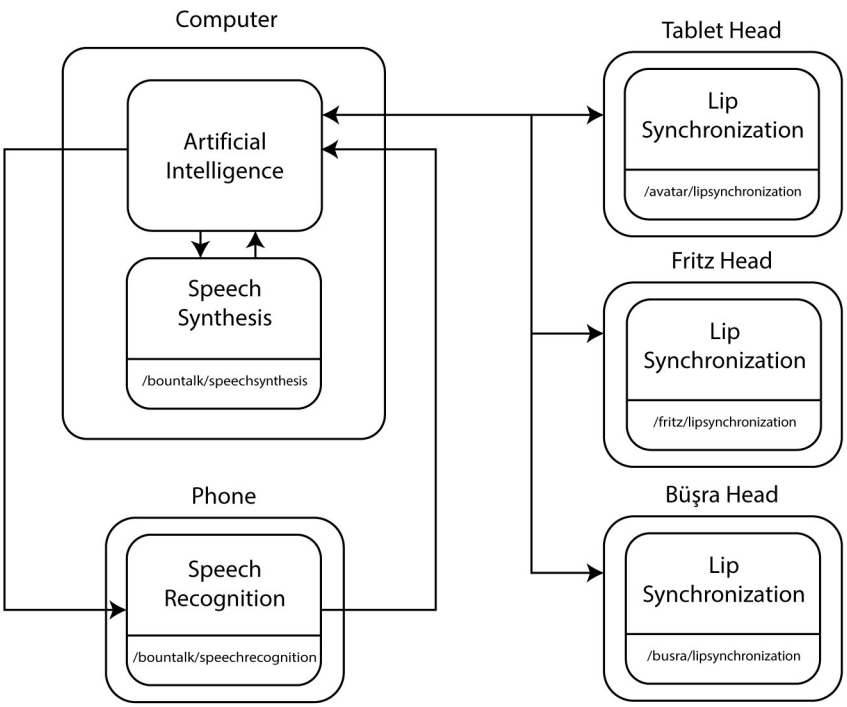


Figure 7.3. The Experimental Platform Diagram

experiment with the help of SoX library [65].

Table 7.1. The Speech Synthesis System ROS Service Message

	Message Type	Message Name
Request	string	speakText
	uint8	speakerID
Response	string	fileName
	uint32	characterNumber
	uint32	characterDuration
	string	characters
	uint32[]	characterDurationStartIndexes

For the speech recognition system, an additional hardware platform is added to the system. Xiaomi Mi5 Prime [66] phone which has the Android 6.0.1 OS with the Google Speech Recognition [67] tool is used for the recognition part. The phone is connected to the same network with the other devices and connected to the *roscore*. For the communication with the speech recognition system, a custom service server has been made. Whenever the recognition service is called, the Google Speech Recognition tool starts to record speech and until the text message is generated, the ROS service does not respond. The manual open and close options are also added to the recognition system, in order to prevent unwanted recognized texts and to give enough time to the user for thinking about the questions. The recognition system message is shown in Table 7.2.

Table 7.2. The Speech Recognition System ROS Service Message

	Message Type	Message Name
Request	uint32	counter
Response	string[]	results
	int32	exitValue

The avatar head of the robot is also connected to ROS via WiFi and also it has a ROS service server for the lip synchronization with a custom message which is shown in the Table 7.3.

Table 7.3. The Lip Synchronization ROS Service Message

	Message Type	Message Name
Request	uint32	characterNumber
	uint32	characterDuration
	string	characters
	uint32[]	characterDurationStartIndexes
Response	int32	exitValue

The Fritz and the *Büşra* robots are using the Arduino Uno and Arduino Mega boards for motor controls, respectively. In order to add these robots to the system, the Arduino library for ROS is used. Via USB port, with the Arduino serial node of the ROS, we communicated with the robotic heads. Each robotic head also has their own ROS service server and they use the same ROS service message as the avatar head for the lip synchronization. The only difference is the service namespaces of the robot. The robot servo positions are written in a hardcoded manner in the Arduino code of the robot and according to the taken message, the robot changes the angle of its own servos. The difference between these services from the other service calls (which is also true for the avatar service), these services return immediately without waiting the lip synchronization to be completed. The reason for this approach is to detect the robots available on the network, and to be sure about robots receive the lip synchronization message. The lip synchronization message should be synchronized with the speech and the output sound file for the speech is played on the same device for each robot which is the speaker of the laptop. Therefore, the synchronization of the lips should be based on the output of other devices, so this protocol is used.

7.3. The Flow Chart of the Experiment

The experiment starts with the introduction of each robotic head. Each robotic head gives information about itself to the user and wants to be the chosen for this game. The user chooses a robot head for playing the game.

After the robot head is chosen, the robot asks a question about its speech. We produced three different speeches which are constructed as male, female and child voices. The user decides the voice for the robot.

Before the game starts, the chosen robot asks the user for ensuring a picture is selected and while asking the user he/she selected a picture or not, the robot also selects a random picture and the game starts.

Each turn has two-phases. For the robot case, the first phase is asking a question (Question Phase) and the second phase is answering the questions (Answer Phase).

In the Question Phase, the robot generates a question for finding the picture which the user selected. The AI of the picture selection is explained in the Appendix B. The selected question is asked to the user and the robot waits for an answer. Each answer to the questions is given the AIML engine which has the specific pattern information for this game. When the AIML matches the answer, it gives true or false value for the asked question and according to the answer the robot changes the possible picture list.

In the Answer Phase, the recognized test is given the AIML system and the AIML system returns the pattern which is matched. When there is more than one match, the AIML returns the most matched pattern. For instance, when the user asks "Beyaz saçlı mı?", the question also matches the pattern which is added for "Beyaz mı" question, but AIML gives the pattern of the first question. After that, the question is answered according to the selected picture data.

The game goes until the user says the correct name for the picture selected by the robot or the robot says the picture which user selected. The flow chart of the experiment can be seen in Figure 7.4.

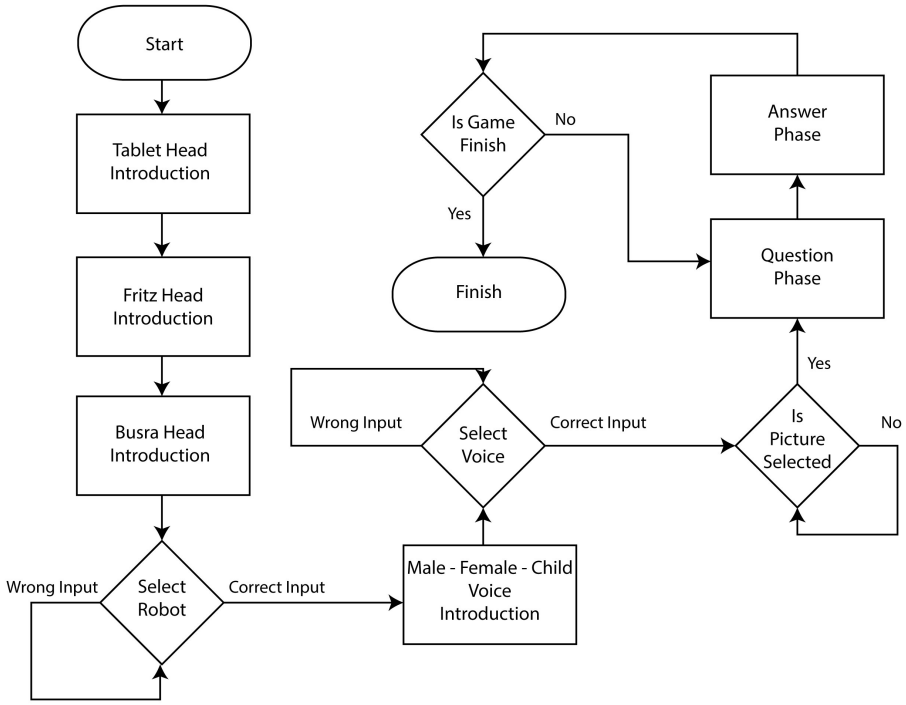


Figure 7.4. The Flow Chart of Experiment

7.4. Evaluation of the Experiment

After the experiment, the users fill a survey which is given in Appendix C. A total of 33 persons participated in the experiment. 33.3 percent of the participants saw a robot that closely for the first time, 60.6 percent came in contact with the robot for the first time.

After robots introduce themselves, the participants select a robot. The selected robot distribution which is given in Table 7.4 is very similar to the uniform distribution. However, the reasons for this selection varies, the main reasons are shown in Table 7.5.

Table 7.4. The number of participants who selected the robot heads

	Van Gogh	Fritz	Büşra
Participant Number	10	12	11

Table 7.5. The probabilities of the reasons for selecting Van Gogh, Fritz and Büşra

	Appearance	Voice	Mechanical Features	Random
Van Gogh	50%	30%	0%	20%
Fritz	66.6%	0%	0%	33.3%
Büşra	63.6%	0%	18.1%	18.1%

For 50 percent of the participants who selected the tablet head for the experiment consider, the main reason for selection was the good-looking character displayed on the tablet head. The Van Gogh portrait is used for the avatar of the tablet head which has a charismatic look and is an art related picture. So, the accompanying robots should be appealing to the eye. The second most common reason for choosing the tablet head over the other heads (30 percent) is that the speech of the tablet is clearer. Actually, the speech system uses the same voice for all the robots with the same volume but the kinematic robots add additional noise to the speech. While a kinematic robot head is speaking, because of the lip synchronization, the motors are moving and the motor and the spring which is used for the lips cause the noise. Therefore, to increase the likelihood of being selected, kinematic heads should be as silent as possible.

On the other hand, Fritz robot head is selected because of its robotic appearance. Some of the participants who selected Fritz and *Büşra* robot, do not consider the tablet as a robot. In order to be considered as a robot, accompaniment robots should have a robotic appearance with a 3D structure.

Like the tablet and the Fritz robot, the Büşra robot was also selected because of its appearance. The participant who selected Büşra robot, considers that Fritz is hard-looking, while Büşra has beautiful eyes. Only a small number of participants consider the mechanical features of Büşra.

After the experiment, we also asked the participants whether they want to use the same head for another game or they want to change, and 63 percent of the participants want to change the selected head, the rest wants to choose other heads because of wanting to try other heads, too. This shows that, aside from the first appearance effect of the robots on the participants, curiosity of the humans tends them to experience other heads. On the other hand, the participants who do not want to change the robot head, stick to the their first reasons for the selection. Only one of the participants, prefers using the same robot head because he/she lost the previous game to that head. All the robot heads use the same AI, but that participant gave a personality to that robot head.

For lip synchronization, the participants detect the lip motions equally for all of the robots. However, while 48 percent of the participants confirm that the lip motions in lip synchronization are totally correct, 33 percent says that they are partially correct. Only 9 percent of the users did not pay attention to lip motion correctness. These results show that the users generally look at the lip motions and they can detect incorrect movements.

We also tested the speech synthesis mechanism and the male, female and the child voices on the robots. The robot and speech type selections are shown in Table 7.6. According to the data, there is a correlation between the appearance of the robot and the selected voice. While the female voice is not selected for the Van Gogh, the child voice is mostly selected by the Büşra robot.

The quality of the speech synthesis is tested in two different ways. Firstly, the intelligibility of speech is asked to the users and after that male, female and child voices are evaluated. For the evaluations, the similarity of the voice is asked. For example, the

Table 7.6. The probabilities of the selected voices for each head

	Male	Female	Child
Van Gogh	80%	0%	20%
Fritz	50%	41.6%	8.3%
Büşra	36.3%	27.2%	36.3%

participants graded the female voice according to how similar it is to an actual female voice. The quality and similarity results are shown in Table 7.7. The results show that, the male voice generation of the speech synthesis system generates voices similar to male, but for the female and child case, we need to add different voice generation algorithms.

Table 7.7. The evaluation of the voices

Male Voice			
	Mean	Standard Deviation	Scale
Quality	6.38	1.46	1 - 10
Similarty	8.30	1.68	1 - 10
Female Voice			
	Mean	Standard Deviation	Scale
Quality	5.62	2.26	1 - 10
Similarty	4.21	2.49	1 - 10
Child Voice			
	Mean	Standard Deviation	Scale
Quality	7.42	1.39	1 - 10
Similarty	4.60	2.82	1 - 10

We asked to the participants whether they liked playing the game with the robot or not. Fortunately, all the participants without exceptions, liked playing the game with the robot. However, not all the participants wanted to play the game again. In

order to understand the factors for not wanting to play the game, we also asked the question about the game like *Did you win the game?*. In Table 7.8, we divided the answers for playing the game again question into three options which are yes, maybe and no. The result shows that, when the AI challenges the users, and it has proper algorithms for winning the game, the number of users who want to play the game again increases.

Table 7.8. The playing again condition with win - lose state

	Win	Lose
Will play again.	66.6%	84%
Maybe play	33.3%	4%
Will not play again	0%	12%

Lastly, we asked the "'*Which game you want to play with the robots?*"' question and according to the game types, we classify the games into three category. The first category is the games which currently robots can play with additional AI. In this category, the most dominant option is the Taboo [68] game. In the second category, the participants want to play computer games with the robot and in the last category, the games which are wanted to play with the robot needs additional actuators. In order to play the games in that category, generally the robot needs at least an arm, but for the options like Basketball, the robot need not an arm only, it needs a whole body with at least two arms and legs.

8. CONCLUSION

The need for an accompanying robot in a domestic environment is obvious. However, the complexity of these robots and the features that these robots should have was not very obvious. In order to find the desired features for the accompanying robot, we constructed three different robotic heads with different capabilities. In this thesis, in order to understand the hidden parameters for designing correct accompanying robot features, we developed a simple gaming environment. In this experiment, generally the robots are chosen with respect to their appearance and that shows that humans want to interact with a good looking robot. Also, we encountered some problems with the kinematic heads which is the noise generated from the motors. We did not consider the effect of these noises on the speech, but the experiment shows that these noises decrease the clearness of the speech and even using the same speech, the participants detect as different speeches.

In this thesis, apart from the robotic head construction, we made a speech synthesis system for the Turkish language. Despite the large amount of missing part of that system, with the help of the custom built speech synthesis, implementing the lip synchronization becomes possible. The lip synchronization system generally generates correct lip motions.

In addition to the hardware structure of the robot, the experiment shows that the AI of the robot also affects the interaction time for the accompanying robots.

In the end, the constructed experiment shows that humans want to make more interaction with the robots. Even the little interactions such as playing the game with the robots and speaking with them makes them happy and they find that it is enjoyable.

8.1. Future Work

As a future work, we plan to design more complex robotic heads with movable body. For our robot project, human - robot interaction has a vital point and in order to increase this interaction, we should more design human like and good looking robots.

For the speech synthesis case, there are a many aspects that can be improved. We can use a new noise function and new rules to increase its naturalness and also we should add emotion synthesis in the speech, too. We should add a new model in order to change frequencies of the speech according to the emotion state. Also, we need to come up a new approach for generating female and child voices.

APPENDIX A: THE "GUESS WHO?" GAME

Guess Who? is a children's game which consists of 24 different pictures. Every picture has different faces with various features. This game is played by two players and each player tries to guess the other player's selected picture. Each round, a player can ask a specific information about the picture. According to the given answer, the player eliminates pictures which are not suitable for the answer. Also, the player answers the other player's question. This game continues until one of the player guesses the correct picture.

Figure A.1. *Guess Who?* Game [69]

APPENDIX B: THE "GUESS WHO?" GAME AI

For the experiment, the robots should play the game with the humans. The selection of the picture can be random, but the questions and guesses should not be random because we try to improve the human - robot interaction. If the robot asks unrelated or unnecessary questions, the robot will probably lose every game and it will eliminate the need to play.

In the AI system, we divide the ontology into seven sub-ontologies which are gender, face color, hair color, moustache color, beard color, glasses color and the hat color. The possible colors are limited to the ten colors which are transparent, white, black, yellow, red, green, blue, auburn, brown and orange. We did not add additional information for the existence of the features. For example, if the hair color of a picture is transparent, that means that the person in the picture is bald. Every feature is written on the code for each picture. The pictures which are used in the experiment can be seen in Figure B.1.

The question answering mechanism of the robot is simple. It generates a sentence which includes the information about whether the answer is true or false. However, questions are generated according to the ontologies of the remaining pictures. For each feature, the robot calculates the multiplication of the dividend sides total number in Equation B.1 (N_{FE} is the number of pictures which have the feature and N_{FNE} is the number of pictures which do not have the feature):

$$max(N_{FE} \times N_{FNE}) \tag{B.1}$$

After the robot calculates that formula for each feature - color or gender combination, the robot asks the question about feature with the highest value.

Figure B.1. The pictures with Turkish names

The question generation uses the optimal path for finding the picture, but the AI system intestinally was not implemented perfectly for giving a chance to the humans. In order to increase the probability of losing to the human, we did not allow the AI system guess a picture when two picture are left. The AI system asks a question about the pictures without guessing the picture and this gives an opportunity to the human to win the game before the robot.

APPENDIX C: QUESTIONNAIRES

Refakatçı Robot Değerlendirme Anketi

Yaşınız :

Mesleğiniz :

1) Daha önce bu kadar yakından bir robot gördünüz mü? Evet ise daha önceden gördüğünüz robotu anlatabilir misiniz?

2) İlk defa mı bir robot ile iletişime geçiyorsunuz?

3) Deney sırasında hangi robotu seçtiniz?

4) Seçme nedeniniz ne idi?

5) Bir daha oynasaydınız yine aynı robotu mu seçerdiniz? Neden?

6) Hangi robotun konuşurken dudaklarını haraket ettirdiğini gördünüz? (Tablet, Fritz, Büşra)

7) Dudaklarını doğru haraket ettirdiğini düşünüyor musunuz?

8) Tablet Refakatçı Robotunu nasıl buldunuz? Robot ne kadar ...

eğlenceli		1 (hiç değil) - 10 (oldukça fazla)
ilginç		1 (hiç değil) - 10 (oldukça fazla)
tatmin edici		1 (hiç değil) - 10 (oldukça fazla)
sıkıcı		1 (hiç değil) - 10 (oldukça fazla)
heyecan verici		1 (hiç değil) - 10 (oldukça fazla)
kullanışlı		1 (hiç değil) - 10 (oldukça fazla)
faydalı		1 (hiç değil) - 10 (oldukça fazla)
değerli		1 (hiç değil) - 10 (oldukça fazla)

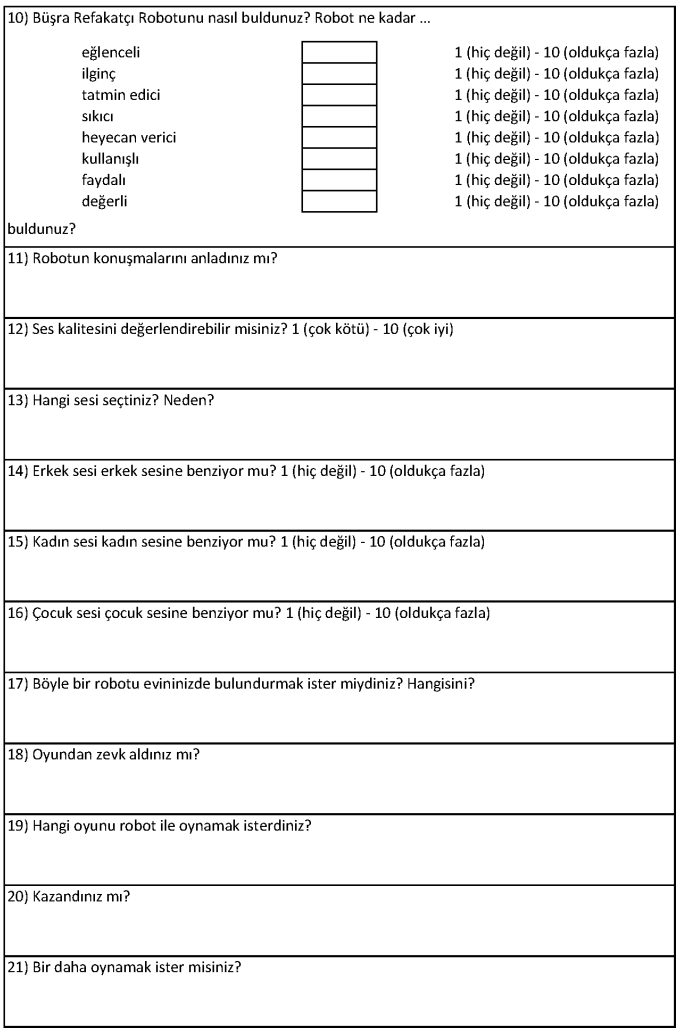
buldunuz?

9) Fritz Refakatçı Robotunu nasıl buldunuz? Robot ne kadar ...

eğlenceli		1 (hiç değil) - 10 (oldukça fazla)
ilginç		1 (hiç değil) - 10 (oldukça fazla)
tatmin edici		1 (hiç değil) - 10 (oldukça fazla)
sıkıcı		1 (hiç değil) - 10 (oldukça fazla)
heyecan verici		1 (hiç değil) - 10 (oldukça fazla)
kullanışlı		1 (hiç değil) - 10 (oldukça fazla)
faydalı		1 (hiç değil) - 10 (oldukça fazla)
değerli		1 (hiç değil) - 10 (oldukça fazla)

buldunuz?

Figure C.1. The survey applied in the experiment



10) Büşra Refakatçı Robotunu nasıl buldunuz? Robot ne kadar ...

eğlenceli		1 (hiç değil) - 10 (oldukça fazla)
ilginç		1 (hiç değil) - 10 (oldukça fazla)
tatmin edici		1 (hiç değil) - 10 (oldukça fazla)
sıkıcı		1 (hiç değil) - 10 (oldukça fazla)
heyecan verici		1 (hiç değil) - 10 (oldukça fazla)
kullanışlı		1 (hiç değil) - 10 (oldukça fazla)
faydalı		1 (hiç değil) - 10 (oldukça fazla)
değerli		1 (hiç değil) - 10 (oldukça fazla)

buldunuz?

11) Robotun konuşmalarını anladınız mı?

12) Ses kalitesini değerlendirebilir misiniz? 1 (çok kötü) - 10 (çok iyi)

13) Hangi sesi seçtiniz? Neden?

14) Erkek sesi erkek sesine benziyor mu? 1 (hiç değil) - 10 (oldukça fazla)

15) Kadın sesi kadın sesine benziyor mu? 1 (hiç değil) - 10 (oldukça fazla)

16) Çocuk sesi çocuk sesine benziyor mu? 1 (hiç değil) - 10 (oldukça fazla)

17) Böyle bir robotu evininizde bulundurmak ister miydiniz? Hangisini?

18) Oyundan zevk aldınız mı?

19) Hangi oyunu robot ile oynamak isterdiniz?

20) Kazandınız mı?

21) Bir daha oynamak ister misiniz?

Figure C.1. The survey applied in the experiment (cont.)

REFERENCES

1. Press, I., *31 million robots helping in households worldwide by 2019*, 2016, `https://ifr.org/ifr-press-releases/news/31-million-robots-helping-in-households-worldwide-by-2019`, accessed at December 2017.

2. Y-N, K., *Japanese quite open to home robots*, 2017, `https://whatjapanthinks.com/tag/robot/`, accessed at December 2017.

3. Young, J. E., R. Hawkins, E. Sharlin and T. Igarashi, "Toward acceptable domestic robots: Applying insights from social psychology", *International Journal of Social Robotics*, Vol. 1, No. 1, pp. 95–108, 2009.

4. Goy, H., M. K. Pichora-Fuller, G. Singh and F. A. Russo, "Perception of emotional speech by listeners with hearing aids", *Canadian Acoustics*, Vol. 44, No. 3, 2016.

5. Ackerman, E., *CES 2017: Why Every Social Robot at CES Looks Alike*, 2017, `http://spectrum.ieee.org/tech-talk/robotics/home-robots/ces-2017-why-every-social-robot-at-ces-looks-alike`, accessed at December 2017.

6. Jaffe, E., *Why Our Brains Love Curvy Architecture*, 2013, `https://www.fastcodesign.com/3020075/why-our-brains-love-curvy-architecture`, accessed at December 2017.

7. Nurimbetov, B., A. Saudabayev, D. Temiraliuly, A. Sakryukin, A. Serekov and H. A. Varol, "ChibiFace: A sensor-rich Android tablet-based interface for industrial robotics", *System Integration (SII), 2015 IEEE/SICE International Symposium on*, pp. 587–592, IEEE, 2015.

8. Salichs, M. A., R. Barber, A. M. Khamis, M. Malfaz, J. F. Gorostiza, R. Pacheco, R. Rivas, A. Corrales, E. Delgado and D. Garcia, "Maggie: A robotic platform for

human-robot social interaction", *Robotics, Automation and Mechatronics, 2006 IEEE Conference on*, pp. 1–7, IEEE, 2006.

9. Robotics, B. F., *Buddy The First Companion Robot*, 2017, `http://www.bluefrogrobotics.com/en/buddy/`, accessed at December 2017.

10. Asus, *Your Smart Little Companion*, 2017, `https://zenbo.asus.com/`, accessed at December 2017.

11. Ekman, P. and W. V. Friesen, *Unmasking the face: A guide to recognizing emotions from facial clues*, Ishk, 2003.

12. Crowe, S., *Buddy Robot Adds 3D Vision, Expanded Personality*, 2017, `http://www.roboticstrends.com/article/buddy_robot_adds_3d_vision_expanded_personality`, accessed at December 2017.

13. Smith, M., *ASUS' Zenbo robot walks, talks and controls your home*, 2016, `https://www.engadget.com/2016/05/30/asus-zenbo-robot/`, accessed at December 2017.

14. Robotbase, *Personal Robot*, 2017, `https://www.kickstarter.com/projects/403524037/personal-robot`, accessed at December 2017.

15. Humphries, M., *Gatebox Virtual Home Robot Wants You to Be Her Master*, 2016, `https://www.pcmag.com/news/350314/gatebox-virtual-home-robot-wants-you-to-be-her-master`, accessed at December 2017.

16. Kuratate, T., Y. Matsusaka, B. Pierce and G. Cheng, ""Mask-bot": A life-size robot head using talking head animation for human-robot communication", *Humanoid Robots (Humanoids), 2011 11th IEEE-RAS International Conference on*, pp. 99–104, IEEE, 2011.

17. Robotics, H., *Innovations / Technology*, 2017, `http://www.hansonrobotics.com/`

about/innovations-technology/, accessed at December 2017.

18. Robotics, S., *Who is NAO?*, 2017, https://www.ald.softbankrobotics.com/en/robots/nao/find-out-more-about-nao, accessed at December 2017.

19. Facebook, T. R., *KOBIAN-RII*, 2017, http://robotfacebook.edwindertien.nl/product/kobian/, accessed at December 2017.

20. Oh, J.-H., D. Hanson, W.-S. Kim, Y. Han, J.-Y. Kim and I.-W. Park, "Design of android type humanoid robot Albert HUBO", *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pp. 1428–1433, IEEE, 2006.

21. Robotics, S., *Who is Pepper?*, 2017, https://www.ald.softbankrobotics.com/en/cool-robots/pepper, accessed at December 2017.

22. Arts, E., *RoboThespian*, 2017, https://www.engineeredarts.co.uk/robothespian/, accessed at December 2017.

23. Zecca, M., N. Endo, S. Momoki, K. Itoh and A. Takanishi, "Design of the humanoid robot KOBIAN-preliminary analysis of facial and whole body emotion expression capabilities", *Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on*, pp. 487–492, IEEE, 2008.

24. Mori, M., "The uncanny valley", *Energy*, Vol. 7, No. 4, pp. 33–35, 1970.

25. Sagisaka, Y., N. Kaiki, N. Iwahashi and K. Mimura, "ATR μ-Talk Speech Synthesis System", *Second International Conference on Spoken Language Processing*, 1992.

26. Arai, T., "Education system in acoustics of speech production using physical models of the human vocal tract", *Acoustical Science and Technology*, Vol. 28, No. 3, pp. 190–201, 2007.

27. Hallahan, W. I., "DECtalk software: Text-to-speech technology and implementation", *Digital Technical Journal*, Vol. 7, No. 4, pp. 5–19, 1995.

28. Fujisaki, H. and M. Ljungqvist, "Proposal and evaluation of models for the glottal source waveform", *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.*, Vol. 11, pp. 1605–1608, IEEE, 1986.

29. Stevens, K. N., *Acoustic phonetics*, Vol. 30, MIT press, 2000.

30. Klatt, D. H., "Software for a cascade/parallel formant synthesizer", *the Journal of the Acoustical Society of America*, Vol. 67, No. 3, pp. 971–995, 1980.

31. Story, B. H., S. Fels and N. d'Alessandro, "TubeTalker: An airway modulation model of human sound production", *Proceedings of the First International Workshop on Performative Speech and Singing Synthesis*, pp. 1–8, 2011.

32. Huckvale, M., *Acoustics of Vowel Production*, 2017, `http://www.phon.ucl.ac.uk/courses/spsci/iss/week5.php`, accessed at December 2017.

33. Rosenthal, V., *Top 10 Tips To Build A Facebook Messenger Chat Bot*, 2017, `https://www.forbes.com/sites/vivianrosenthal/2017/03/20/top-ten-tips-to-build-a-facebook-messenger-chat-bot/`, accessed at December 2017.

34. Loebner, *Home Page of The Loebner Prize in Artificial Intelligence*, 2015, `http://www.loebner.net/Prizef/loebner-prize.html`, accessed at December 2016.

35. Dialogflow, *Getting started with Dialogflow fulfillment*, 2017, `https://dialogflow.com/docs/how-tos/getting-started-fulfillment`, accessed at December 2017.

36. Gesellensetter, L., N. C. Krämer and I. Wachsmuth, "A conversational agent as museum guide - Design and evaluation of a real-world application", 1 (Editor), *The 5th International Working Conference on Intelligent Virtual Agents (IVA'05)*, pp. 329–343, Springer, 2005.

37. Ishiguro, H., T. Ono, M. Imai, T. Maeda, T. Kanda and R. Nakatsu, "Robovie: an interactive humanoid robot", *Industrial robot: An international journal*, Vol. 28, No. 6, pp. 498–504, 2001.

38. Yıldırım, Y., O. Asık, B. Görer, N. E. Özkucur and H. L. Akın, "Tur Rehberi Çoklu Robot Sistemi", Türkiye Otonom Robotlar Konferansı, 2014.

39. Wikiart, *Self Portrait with a Grey Felt Hat*, 2017, `https://www.wikiart.org/en/vincent-van-gogh/self-portrait-with-a-grey-felt-hat-1887`, accessed at December 2017.

40. Group, K., *OpenGL - The Industry's Foundation for High Performance Graphics*, 2017, `https://www.opengl.org/about/`, accessed at December 2017.

41. Ahlberg, J., *CANDIDE-3 - An Updated Parameterised Face*, Tech. rep., Department of Electrical Engineering, Linköping University, 2001.

42. Face++, *Face Landmarks*, 2017, `https://www.faceplusplus.com/landmarks/`, accessed at December 2017.

43. Samsung, *Galaxy Tab S (10.5, Wi-Fi)*, 2017, `http://www.samsung.com/uk/tablets/galaxy-tab-s-10-5-t800/SM-T800NZWABTU/`, accessed at December 2017.

44. ROS, *ROS*, 2017, `http://www.ros.org/`, accessed at December 2017.

45. XYZbot, *Fritz: A Robotic Puppet*, 2016, `https://www.kickstarter.com/projects/1591853389/fritz-a-robotic-puppet`, accessed at December 2017.

46. TowerPro, *SG90 Digital*, 2017, `http://www.towerpro.com.tw/product/sg90-7/`, accessed at December 2017.

47. ServoDatabase, *Futaba S3003 - Servo Standard*, 2017, `https://servodatabase.com/servo/futaba/s3003`, accessed at December 2017.

48. Arduino, *Arduino UNO*, 2017, https://www.arduino.cc/en/Main/ArduinoBoardUno, accessed at December 2017.

49. ROS, *rosserial_arduino*, 2017, http://wiki.ros.org/rosserial_arduino, accessed at December 2017.

50. Corporation, S., *3D CAD Packages*, 2017, http://www.solidworks.com/sw/products/3d-cad/packages.htm, accessed at December 2017.

51. Flashforge, *FLASHFORGE Dreamer Dual Extrusion 3D Printer*, 2017, https://flashforge-usa.com/products/dreamer-dual-extrusion-3d-printer?variant=5344977190950, accessed at December 2017.

52. Oh, K.-G., C.-Y. Jung, Y.-G. Lee and S.-J. Kim, "Real-time lip synchronization between text-to-speech (TTS) system and robot mouth", *RO-MAN, 2010 IEEE*, pp. 620–625, IEEE, 2010.

53. Castelli, F., "Understanding emotions from standardized facial expressions in autism and normal development", *Autism*, Vol. 9, No. 4, pp. 428–449, 2005.

54. Model, E., *EMAX ES08MA II 12g Mini Metal Gear Analog Servo for RC Model*, 2017, https://www.emaxmodel.com/es08ma-ii.html, accessed at December 2017.

55. TowerPro, *MG946R*, 2017, http://www.towerpro.com.tw/product/mg946r/, accessed at December 2017.

56. McGurk, H. and J. MacDonald, "Hearing lips and seeing voices", *Nature*, Vol. 264, No. 5588, pp. 746–748, 1976.

57. Melek, Z. and L. Akarun, "Automated lip synchronized speech driven facial animation", *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, Vol. 2, pp. 623–626, IEEE, 2000.

58. B. Doval, w. N. H., C d'Alessandro, *The spectrum of glottal flow models*, 2017, https://rs2007.limsi.fr/PS_Page_2.html, accessed at October 2017.

59. Rosenberg, A. E., "Effect of glottal pulse shape on the quality of natural vowels", *The Journal of the Acoustical Society of America*, Vol. 49, No. 2B, pp. 583–590, 1971.

60. Fant, G., J. Liljencrants and Q.-g. Lin, "A four-parameter model of glottal flow", *STL-QPSR*, Vol. 4, No. 1985, pp. 1–13, 1985.

61. Pozo, A. d. and S. Young, "The linear transformation of LF glottal waveforms for voice conversion", *Ninth Annual Conference of the International Speech Communication Association*, 2008.

62. Harrington, J. and R. Mannell, *Phonetics and Phonology*, 2017, https://rs2007.limsi.fr/PS_Page_2.html, accessed at December 2017.

63. Huber, J., E. Stathopoulos, G. M. Curione, T. A. Ash and K. Johnson, "Formants of children, women, and men: The effects of vocal intensity variation", Vol. 106, pp. 1532–42, 10 1999.

64. Akın, A. A. and M. D. Akın, "Zemberek, an open source nlp framework for turkic languages", *Structure*, Vol. 10, pp. 1–5, 2007.

65. SoX, *SoX - Sound eXchange*, 2015, http://sox.sourceforge.net/, accessed at December 2017.

66. Mi, *Mi 5 Fast as light*, 2017, http://www.mi.com/en/mi5/, accessed at December 2017.

67. Google, *SpeechRecognizer*, 2017, http://developer.android.com/reference/android/speech/SpeechRecognizer.html, accessed at December 2017.

68. Wikipedia, *Taboo (game)*, 2017, http://www.wikizero.org/index.php?q=

aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnL3dpa2kvVGFib29fKGdhbWUp, accessed at December 2017.

69. Argos, *Guess Who? Board Game from Hasbro Gaming*, 2017, http://www.argos.co.uk/product/3904323, accessed at December 2017.