MULTI-SENSORY PERCEPTION IN VIRTUAL WORLD ANIMATION USING MCGURK EFFECT

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

BABAK ALLAHGHOLIPOUR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF MULTIMEDIA INFORMATICS

DECEMBER 2019

Approval of the thesis:

MULTI-SENSORY PERCEPTION IN VIRTUAL WORLD ANIMATION USING MCGURK EFFECT

Submitted by BABAK ALLAHGHOLIPOUR in partial fulfillment of the requirements for the degree of **Master of Science in Modeling and Simulation Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics** _____

Asst. Prof. Elif Sürer
Head of Department, **Modeling and Simulation** _____

Assoc. Prof. Dr. Cengiz Acartürk
Supervisor, **Cognitive Science Dept., METU** _____

Examining Committee Members:

Asst. Prof. Elif Sürer
Multimedia Informatics, METU _____

Prof. Dr. Alptekin Temizel
Multimedia Informatics, METU _____

Assoc. Prof. Dr. Cengiz Acartürk
Supervisor, Cognitive Science Dept., METU _____

Dr. Murat Perit Çakır
Cognitive Science Dept., METU _____

Doç. Dr. Hacer Karacan
Computer Engineering Dept., Gazi University _____

Date: **09/12/2019**

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: BABAK ALLAHGHOLIPOUR

Signature : _____

ABSTRACT

MULTI-SENSORY PERCEPTION IN VIRTUAL WORLD ANIMATION USING MCGURK EFFECT

Allahgholipour, Babak

MSc., Department of Multimedia Informatics

Supervisor: Assoc. Prof. Dr. CENGİZ ACARTÜRK

December 2019, 42 pages

Information about multisensory perception and brain cognition mechanism lead to a better virtual world creation. Brain store visual and auditory information to use in predicting future events. Image memory takes information from visual clues which during speech recognition happens by recording visual information and its auditory mappings. Based on this process, brain create prediction mechanism for future events. In speech recognition brain record sound and its visual presentation. When brain encounter difficulty in perception an auditory stimulus, it looks for visual mappings to guess the result. Considering the fact that some completely unrelated similarity may exist, some mapping may not lead to better understanding the world around. When previous information misleads the brain to understand the situation cognitive biases happen.

McGurk Effect discusses how different lip movements affect auditory perception to get disparate information. Variety of factors influence amount of dependency on visual clues. Tracking speaking, especially in challenging environments, can be increased by visual clues such as face expressions, tongue and lips movements. Considering these elements in virtual world and game designing lead to improving believability. Perception processes, brain nature, designing factors, graphical elements, and visual structures are discussed to find out improved ways in designing realistic auditory and visual components.

Keywords: Multisensory Perception, Auditory and Visual Process, Virtual World Animation, McGurk Effect

ÖZ

ÇOK SENSÖRLÜ ALGILAMA SANAL DÜNYANIN ANİMASYONUNDA MCGURK ETKİSİNİ KULLANARAK

Allahgholipour, Babak

Yüksek Lisans, Çoklu Ortam Bilişimi Bölümü

Tez Yöneticisi: Doç. Dr. CENGİZ ACARTÜRK

Aralık 2019, 42 sayfa

Çok sensörlü algı ve beyin biliş mekanizması hakkında olan bilgiler, daha iyi bir sanal dünya yaratılmasına yol açar. Beyin, gelecekteki olayları öngörmek için kullanılacak görsel ve işitsel bilgileri saklar. Beyin görüntü belleği, konuşma tanıma sırasında görsel bilgileri ve işitsel haritalarını kaydederek meydana gelen görsel ipuçlarından bilgi alır. Bu sürece dayanarak, beyin gelecekteki olaylar için tahmin mekanizması oluşturur. Konuşma tanımında, beyin ses ve onun görsel sunumunu kayıt eder. Beyin bir işitsel uyarıcıyı algılamakta güçlükle karşılaştığında, sonucu tahmin etmek için görsel haritalamalar arar. Bazen tamamen alakasız benzerliklerin var olmasından kaynaklı, bazı haritalandırmalar daha iyi anlamaya yol açmayabilir. Önceki bilgiler durumu anlamak için beyni yanlış yönlendirdiğinde bilişsel önyargılar meydana gelir.

McGurk Effect, farklı dudak hareketlerinin farklı işitsel algıyı dolaysıyla farklı bilgi elde etmeyi nasıl etkilediğini tartışıyor. Çeşitli faktörler görsel ipuçlarına bağımlılık miktarını etkiler. Özellikle zorlu ortamlarda konuşma algılama, yüz ifadeleri, dil ve dudak hareketleri gibi görsel ipuçları ile artırılabilir. Bu unsurları göz önünde bulundurmak, sanal dünyada ve oyun tasarımında inanılırlığın arttırabilir. Gerçekçi işitsel ve görsel bileşenlerin tasarımında gelişmiş yollar bulmak için algı süreçleri, beyin yapısı, tasarım faktörleri, grafik öğeler ve görsel yapılar araştırılmıştır.

Anahtar Sözcükler: Çok Sensörlü Algılama, İşitsel ve Görsel İşlem, Sanal Dünya Animasyonu, McGurk Etkisi

To My Helpful Supervisor

ACKNOWLEDGMENTS

I would like to thank my thesis supervisor, Assoc. Prof. Dr. Cengiz Acartürk who worked patiently with me to finish this thesis. My questions get answers immediately and he never hesitates to give ideas to improve the research and experiments. His tendency and knowledge motivated me through all steps to fulfill the ideas. Whenever there was a problem, he presented understanding behavior and assigned a lot of time and energy to solve it properly. On the other hand, my supervisor accepted to work with me on this thesis which I will always remember and appreciate, while he could reject it instantly.

I would also like to thank participants in survey who spent more than ten minutes to watch the videos and answer the questions.

Finally, I would like to thank Examining Committee Members, who prepared to my thesis defense in short amount of time. They answered our emails instantly and we could plan events in shortest amount of time.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

A knowledge of brain cognitive mechanism and psychology of human perception can lead to better understanding of how visual and auditory communication works, create virtual designs accordingly and reach perceptual fluency. Environment provide ambiguous information which would take huge amount of time to process and interpret. To make the procedure faster and sometimes to guess the event, brain shapes a higher cognitive structure based on stored knowledge and past experiences [1]. According to such information hypotheses are formed to make perception which can be presented as a balance between stimuli and stored samples in memory [3]. Some cells in brain are constantly work on matching incoming sensory information with expected ones and create predictions which actually makes brain a prediction machine [2]. Due to such predictions human beings could survive in history since they could map information, predict the results and act accordingly. Not all mappings lead to better understanding the world considering the fact that some completely unrelated similarities may exist. Cognitive biases refers to situations where, based on previous experiences, information mapping process mislead the brain and causes wrong results.

Optical illusion happens when brain is unable to find enough visual cues to interpret surroundings or when brain guesses are different from reality. A psychological method

called Gestalt principles which is a powerful tool for interpreting visual designs can help designer understand people approach in intrepeting visuals elements. Gestalt psychology consist of five principles including figure-ground relationship, proximity, similarity, continuity, closure. These principles explain sensory organization of information in the mind. Millions of pieces of objects and things around us are organized as units and patterns to become perceptional to brain. These principles make people understand visual presentation faster and unconsciously enjoy what they are seeing. Auditory perceptions works somehow similar which clues are used to understand the whole audio message. In noisy areas where hearing and differentiating the words become hard, brain uses the words that been heard and clues which can be caught, to guess missing part and understand the situation.

Visual components are important in speech recognition both in noisy and in normal situations. McGurk effect present the idea that mismatching between visual and auditory information, whenever it is available, can unconsciously mislead the brain in speech recognition [7]. Tracking individuals speaking, especially in challenging environments, can be boosted by visual clues like face expressions, tongue and lips movements [9]. By presenting visual clues understanding threshold with young adults recorded as being -14dB while with only auditory information that level decreases to -9dB [10]. Visual component role become even more important with people having hearing impairment or in noisy environments.

Another important factor is first impression which is not straightforward to overcome especially bad first impression [8]. Previous experiences and knowledge in combinition with ability to percept a design affect how people feel about particular situation [4]. Designing realistic elements and considering foveal vision which is psychologically the most accurate point of seen, can help viewer to understand the design faster and achieve better first impression. The time duration in which brain spends to understand a particular object influences the way people feel about particular situation. This is one of the most effective element in first impression and called cognitive load which is a difinition for brain efforts to learn the situation and keep track of reasons for them which represents brain capacity in handling information on working memory [17]. Making visual amd auditory communication effective requires keeping this load time low. Even though this duration is never zero, lowering it will make people percept faster.

Image memory which consist of visual design of things and self-created mental images takes in information from visual clues immediately [4]. In speech recognition process, people store images of talking characters and the way that images map with voice information. Brain always try to interpret faster, so it filters out unnecessary information

and keep other ones. Image memory keep information of lips, tongue and teeth in expression a word and blurs information of hair and eye movements. When brain gets less information of voice and unable to interpret the situation, it looks for expected imagery information to map. To do so, it focuses on foveal parts and seek familiar factors on what it sees. People have tendency to filter out unfamiliar or at the time unnecessary information. Even though this process restricts visual perception, it is done on daily basis and all the time. People usually try to predict what would happen or being said. As a result only information about lips, tongue movement and teeth maybe available to interpret the voice being heard.

Improvement of drawing and modeling tools set forth ideas of designing rather than its techniques. Suggestions of complexity and charming features of a character was presented to make it feel real for players [8]. Considering brain process in visual-auditory communication can set ideas to make virtual design even more fluent for perception. Most virtual worlds and games are about interactions of characters which bring out the importance of character design both psychologically and visually. McGurk Effect provide ideas about looking for visual components to understand auditory speech. Considering such features in designing process will make people feel being in realistic world and enjoy more.

Computer games and virtual world gained wide range of equipment to surpass designing limits due to enhanced hardware features. Visual and auditory communication became easier as designing reached realistic presentation. Good design provides functionality, user friendliness, and consistency while otherwise huge efforts on a game would vanish in a matter of days. Games with similar artificial intelligence, physics, and game play get completely different range of success. This phenomenon originated by a hidden factor which usually stay unnoticeable and can be found in visual and auditory communication. Ideas about visual designs come from social, technological, and intellectual context of public communications, their history, and the experiences of their makers [11]. Importance of visual sense has driven great number of researchers to scrutinize it. Economic circulation of computer games and amount of time that people spend by playing it has increased over the past decade [11] which introduced new field of media. Considering that to some extent all age groups are engaged with this field [11] and it fosters problem-solving, scientific inquiry and scientific breakthrough discovery [12], developing visual and auditory communication become more important. Considering that computer games influence visual-processing skills positively [13] visual design can accelerate that and use it in effective way. Massive amount of time spent with games can be presented as opportunity instead of threat if it is valorized appropriately. Designing visual elements considering its psychological, cultural, and educational effects will satisfy player and leave them with new ideas for social and educational life which can lead to unconsciously education in parallel with entertainment. Introduction part of this thesis

present visual discussions in different majors and fields such as Psychology, Business, Art, Industrial Design, Graphic Design, Computer Science, Architecture, Literature, and Neuroscience, to bring together ideas that can improve game visual communication which are useful in both serious and entertaining game designs. Then they will be discussed more deeply in the following chapters.

Multidisciplinary approaches in visual thinking are needed to find out complex visual processes in brain which are also influenced by studies in psychology and visual arts [4]. Regarding the fact that mistakes in the design of control room and control panel of a power plant once led to nuclear accident in the United States [14], makes visual design a crucial subject to deal with. Considering visualization as communication presents idea that there can be a specific guideline of visualization to improve presentation [15]. Virtual world designs and computer games can use such guideline to present its visualized content and make perception faster. People are substantially different in their expectation, education, culture, intelligence, communication and needs which make designing process a complicated task. This present a challenge for designers who usually have to design a single project for everyone and it is even more intricate when all people are expected to use the product [14]. When auditory factors added, ignoring phenomenon like McGurk effect would end up in unrealistic and confusing results. On the other hand considering it would make the virtual world design much more pleasurable and useful especially for people with hearing impairments.

According to huge number of psychological studies people treat computer games and other media like television as real people and places [16]. As a result character creation, virtual world modeling and game design required to have an amount of believability, which otherwise would cause to lose significant number of audiences. Some people prefer to leave their knowledge of reality for the sake of enjoyment, like when they believe "Spider-man" character, but a complete novelty would overflow such capacity. Due to the fact that these are hidden factors which mostly even players themselves do not know why they quit the game, considering details to increase believability will lead to better visual design of games and virtual world.

People of all ages and both gender use virtual world and play computer games. If we suppose games started in 1980, first gamer generation are in middle age now. These people are getting older and by age physical and visual abilities will decrease, agility of body and speed of some operations will be lessened. Visual contrast is diminished and material scattered in eyes [14]. These people will still need entertainment and there is no study to design games for elder people. Although designing games for old people is beyond the scope of this study, information on speech cognition, designing approaches, visual

perception, and psychology can provide ideas of designing for elderly people and keep them entertained and healthy.

Main goal of this thesis is to propose visual design ideas considering auditory factors in virtual world designing to improve believability of a virtual world or a game by reaching realistic design ideas. Interdisciplinary studies such as perception processes, brain nature, designing factors, graphical elements, and visual structures covered to find out improved ways in designing a realistic visual and auditory designs. Details will be discussed in chapters as follow:

- Chapter 2 literature review considering different studies on visual auditory integration and McGurk Effect.
- Chapter 3 methods and design process and examines on provided animation.
- Chapter 4 results of the study.
- Chapter 5 discussion and conclusion considering the results and future works.

CHAPTER 2

LITRETURE REVIEW

McGurk and MacDonald (1976) found out an illusion caused by multisensory interpretation which happens when same voice presented by different visual clues. It can be defined as changes in auditory perception due to different visual information, so perception of the voice is something different than what it is. The powerful effect is a significant demonstrator of multisensory integration and shows that visual and auditory information unified into merged perception. Although individual differences are influential [20], it is extremely strong effect that even when people know the process they cannot interfere or prevent the illusion. It is language-universal effect and its qualitative production rules exist in experimented languages; however, some quantitative differences may exist dependent on the structure of the language [21]. While it is extremely powerful in English, in noise-free conditions Japanese reported small effect. However, when environment was noisy and hearing became hard, the effect was robust. Japanese adults are faster with audio conditions, while English adults are faster when just visual conditions fulfilled [26]. German and Spanish subjects show strongly rely on visual stimuli especially when they encounter foreign words [22]. Arabic language also showed presence of the effect [21] while Chinese natives demonstrate a weaker effect [22]. In Finnish language visual stimuli were strongly influential on subjects where more than 90% of participant reported biased by the effect [22]. In Italian language the effect reported to be significant with some vocals and a bit weaker with some others [26]. Some argue that languages like English, Spanish, and Italian are more complex, therefore need stronger visual clues, than Japanese and Chinese which are simpler [26]. Such strength presents the strength of audiovisual integration in perception and provides a highly useful tool in research [18].

Sensory integration is automatic, mandatory and uncontrollable since perception remains strong regardless of full knowledge of the situation [24]. Minimal visual presentation also

supports the effect like when researchers used point-light representation of mouth and jaw motion leaving other facial expressions aside [24].

McGurk Effect is also known as Fusion Effect due to the fact that when clues are not enough to perceive the event, brain fuse received information to perceive the situation [18][19]. Auditory stimuli with an incongruent visual presentation cause a perception of third syllable which is different from both auditory and visual syllable [20]. As an example, if visual stimuli present “gi” and auditory is “bi”, a lot of people report hearing “di”. Combination Effect happens when auditory stimuli and visual presentation precept as one. For example, auditory “ga” with visual stimuli of “ba” may be heard as “bga” [24].

Variety of versions have been examined and the results were somehow confirming the audiovisual integration and the illusion. The most famous variant is the voice saying “Ba” with different visual presentation which were percept as “Fa” or dependent on the visual presentation were percept as “Ga”. Another examination was done by presentation visual “d” and acoustic “b” which were heard as “d”.

Facial structure manipulation through inversion, rotation, translation into moving dots and rearrangement may reduce McGurk Effect [24]. Such changes which causes reduction of the effect can be interpreted as a change in intelligibility. Auditory supplemented by visual speech and familiar voice speeds up learning process. Familiar voices activate visual areas of the brain, regardless of the fact that such activation is not clear to be presented by what part [24]. Doing McGurk Effect on oneself which is also can be called self-McGurk Effect presented the idea that facial identity is independently processed and auditory identity affects speech processing.

Visual information influence starts in prelinguistic age where presence of McGurk fusion demonstrated [26]. However, in children between ages of 3 to 5 and 7 to 8 McGurk Effect is weak due to the fact that visual model patterns and their perception are less than adults. As age increases the influence of visual stimuli raises due to the fact that brain develops speech sound and their visual presences patterns. On the other hand, older people rely more on visual clues are hearing ability decreases.

Temporal differences reported to be effective on the McGurk Effect. A strict synchronization is not required in McGurk effect, however a small delay in audio

increases visual influence strongly [19]. As a result, time-varying aspects affect perceivers but temporal coincidence is not required.

The word itself and structure of it rarely influenced the effect which happens in nonsense syllables and words very similarly to other words. Furthermore, frequency of the effect is not significantly influenced by meaning of the word. As a result, visual and auditory integration starts to occur at early stages [22]. Some other researchers believe that signals are processed independently and then they integrated at later stage to form perception [23].

Emotional McGurk Effect has been studied and presented that if the video spoken in an emotion dubbed with voice of another emotion, third emotion will be perceived. When a video with angry spoken in a modality and happy spoken in another modality, subjects shifts answers to happy and vice versa [25].

Several authors argued that while vocals “i” and “a” have stronger McGurk effect, vocals “u” have weaker effect. Lip movement during pronouncing can be influential in the effect when “u” provides less quality of visual information [26].

Studying the unisensory component role in perception of the speech requires carefully designed experiments (Bertelson etal., 2003; Alsius etal., 2005), computational modeling (Massaro,1998; Schwartz, 2010), and scrutinizing brain mechanisms (Sams etal.,1991; Skipper et al.,2007). Dependent of how coherent and reliable information provided by each visual of auditory stimuli, the extension of influences changes [18].

The questions about how, where and when brain performs audiovisual integration remains debatable [26]. Different methods in examining the process can make cleared understanding of the process.

CHAPTER 3

METHODS AND DESIGN PROCESS

3-1 Introduction

Main purpose of this study is examining the McGurk Effect in virtual world. Making virtual world more realistic, believable and useful depends on details that are hidden in unnoticeable factors. Brain working process and its clues in understanding different situation can be used in designing virtual assets to make perception more fluent. Such ideas can have significant roles in teaching, helping impairments and other fields. McGurk Effect proves that brain use integrated senses to understand and process received information. Visual clues have important roles in processing audio waves. McGurk Effect provides tools to examine the brain tools. To do so, a realistic 3D character and its rigging prepared. As a 3D software, Autodesk Maya and as render software Mental Ray are used. Maya is one of the pioneers in animation industry and since 1998 become first choice of animators. In 2005 Autodesk bought the software and the name changed to Autodesk Maya. Mental Ray is one of the most famous render programs integrated into other third-party applications to achieve high quality images from 3D models. It was bought by Nvidia in 2007 and was used in a number of films. Although it was discontinued in 2017, a lot of applications still support it and can be used.

Due to the fact that lips movement are vital in McGurk Effect, lips texture, its rigging and movement became the most important part of the preparation period. Based on a real-world video and its duration the animation prepared and dubbed by its sound. Lip movements designed to be the exact same as the videos. Inspired by the video and its duration, 32 frames rendered as first frame to be the starting point of the animation and the middle one to be the widest open mouth in pronouncing “fa” or “ba”. Last frame designed to be near to first frame to make the animation seamless as it repeats several times in pronouncing the words. Each part repeated a number of times to provide brain enough time to process.

To make 3D render faster frame size considered as width to be 1280px and height to be 720px (16:9 frame size rate). Data rate is 10mbps and total bitrate is 100mbps. Frame rate considered to be 24 fps. For audio render bitrate considered as 314kbps and stereo sound channel prefered for audio channels. Audio sample rate is 48kHz. Each video length is 12 seconds. For video H.264/MPEG-4 AVC and codec avc1 are prefered and for audio mp4a: MPEG-4 AAC LC codec selected which both are most commong video render setting in industry.

The animation is prepared based on real world video. 30 frames were needed to pronounce “fa” or “ba” and 2 frames is considered to be transition frames which is silence between two pronunciations.

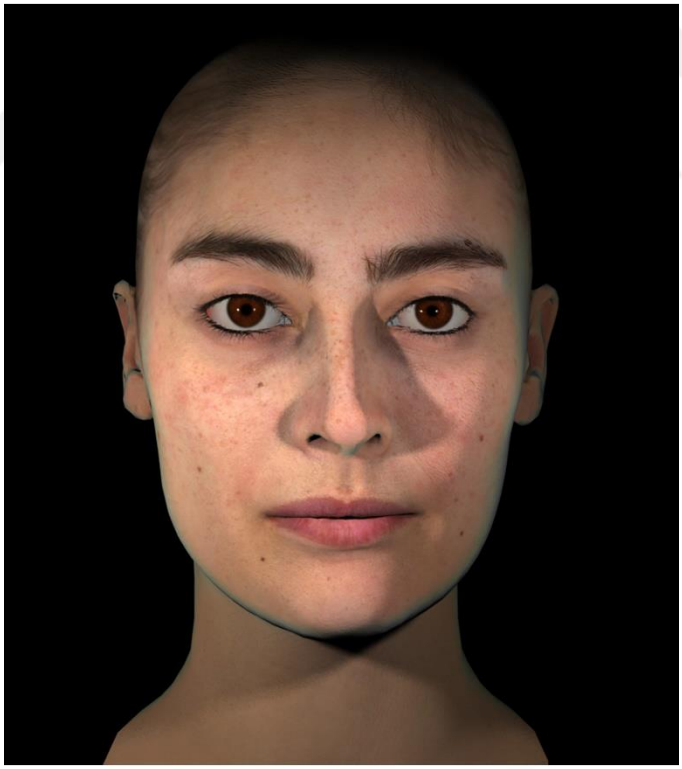

Figure 1: The 3D character used in the animation. Modeled in Maya and rendered with Mental Ray *.

* https://www.eisko.com/

Figure 2 : Animation is prepared by considering the frames in video, pronouncing “ba”, first frame

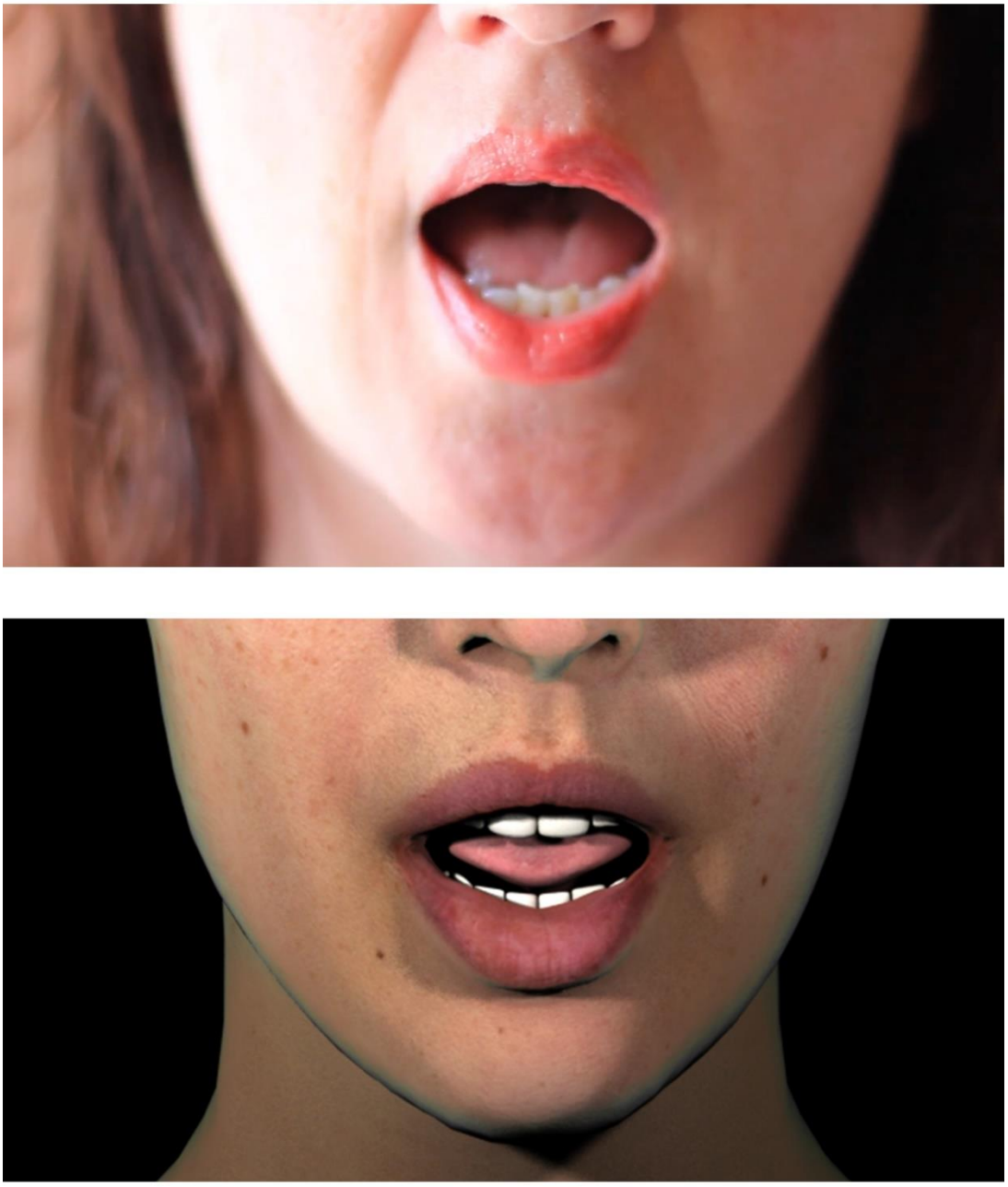

Figure 3 : Animation is prepared by considering the frames in video, pronouncing “ba”, 15th frame.

Figure 4 : Lips movement frame by frame from a video pronouncing “ba”. It is 30 frames with 2 frames for transition.

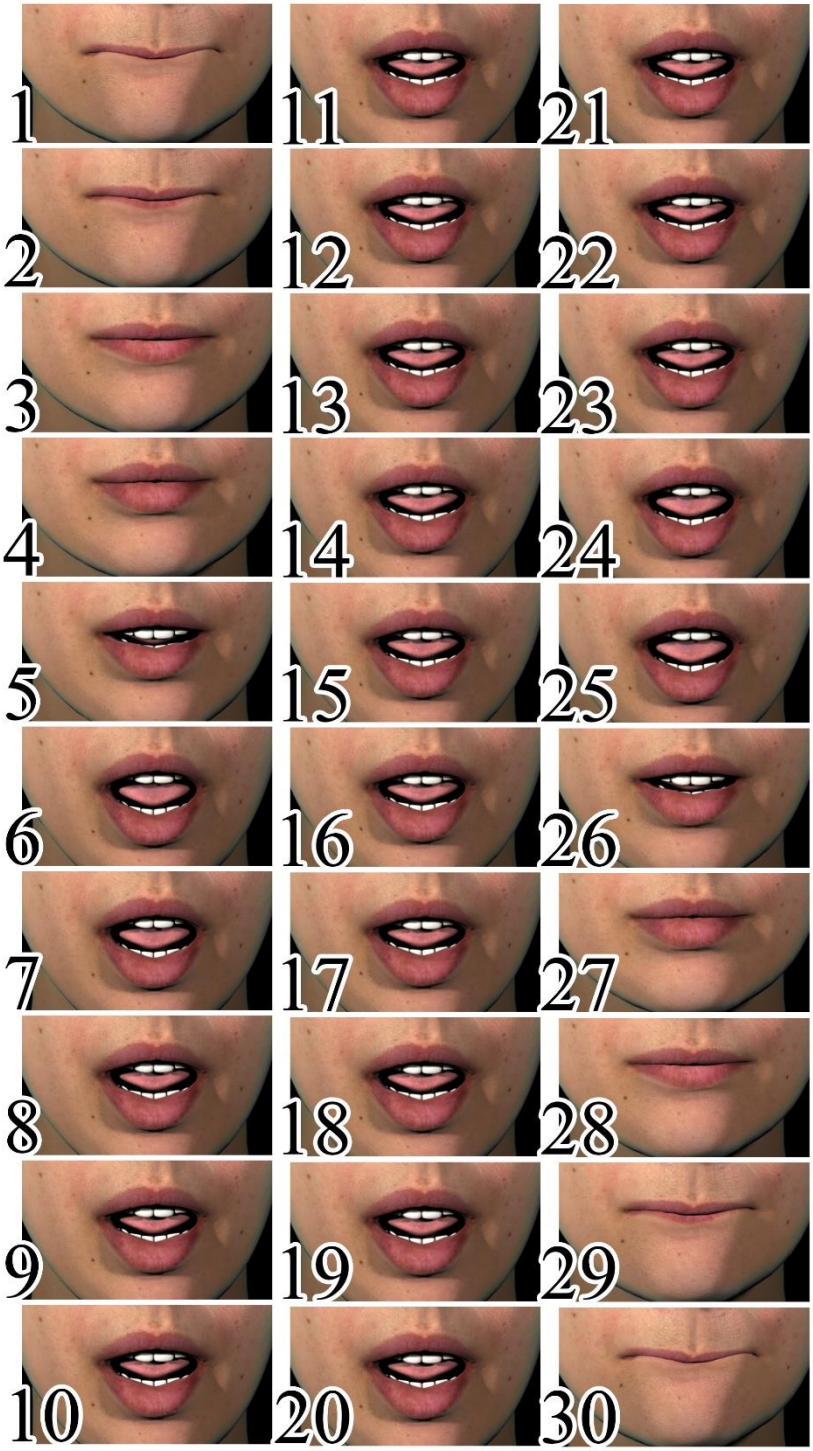

Figure 5 : Animation frame by frame according to video frames, pronouncing “ba”, close to lips version.

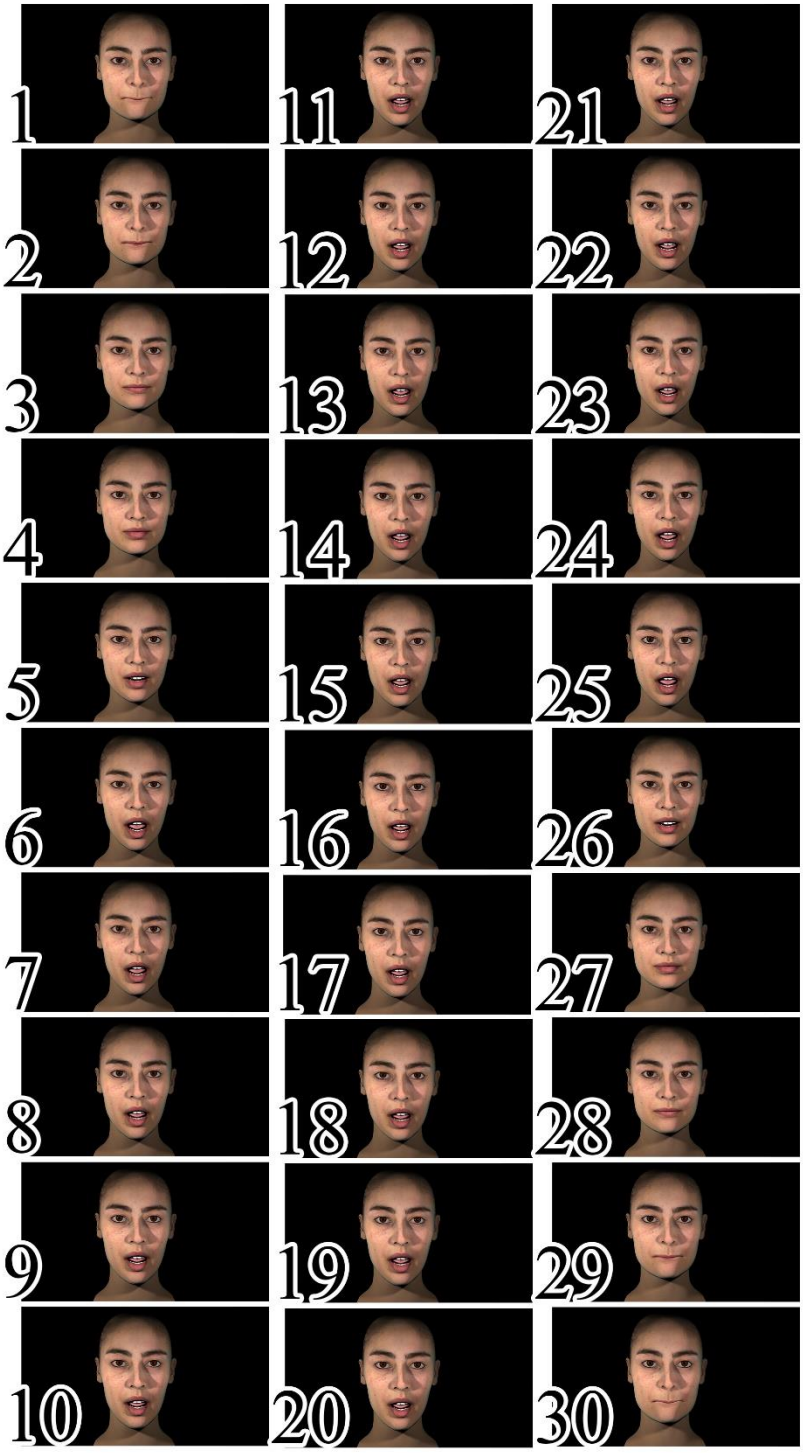

Figure 6 : Animation frame by frame according to video frames, pronouncing “ba”, full face version.

Figure 7 : Animation is prepared by considering the frames in video, pronouncing “fa”, first frame.

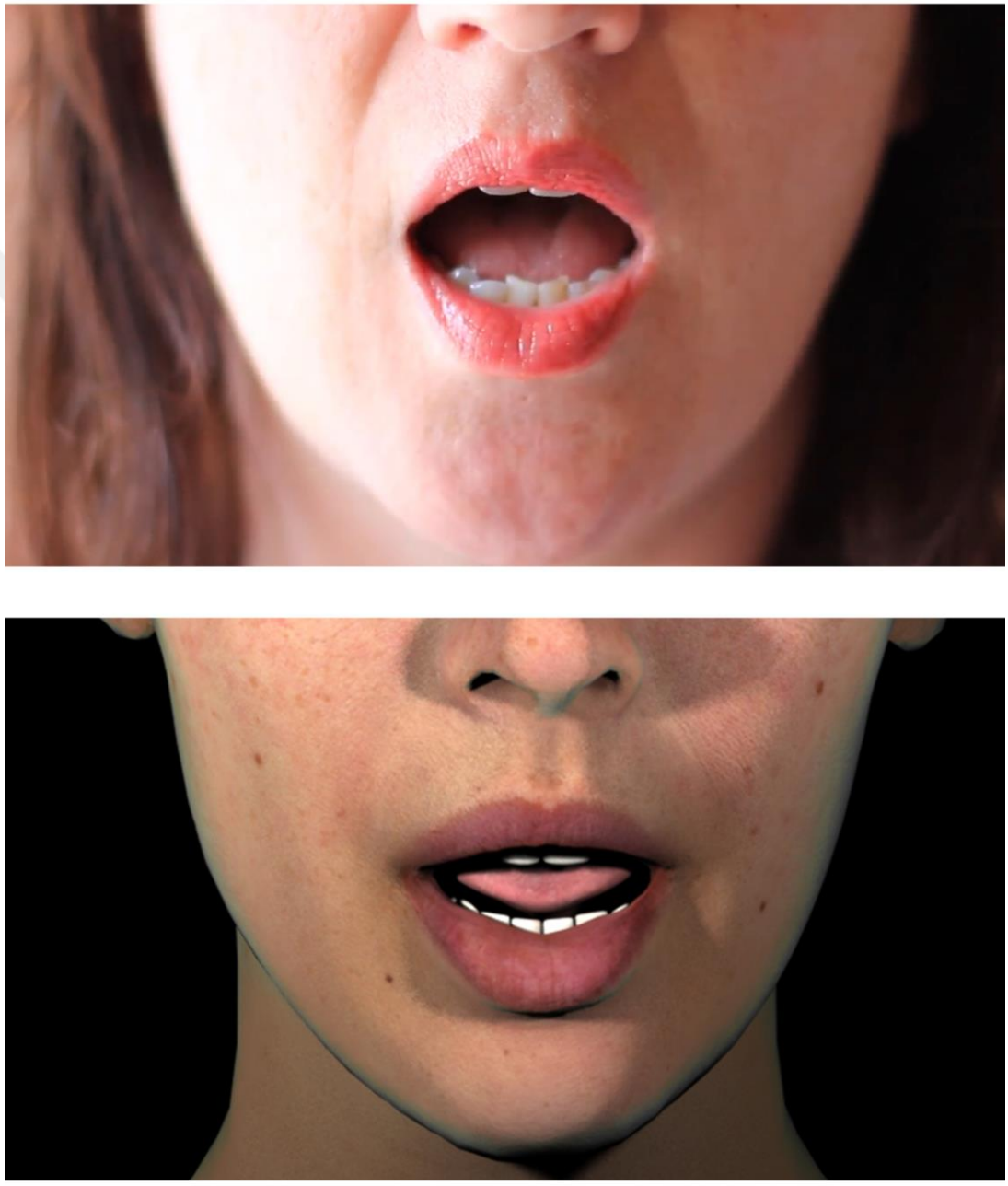

Figure 8 : Animation is prepared by considering the frames in video, pronouncing “fa”, 15th frame. It is almost the same in both “ba” and “fa” when mouth is open.

Middle frames, when the month is open, somehow are similar in both “ba” and “fa” pronunciation. Key frames are first 10 frames when shape of lips are different in different pronunciation. Most important parts are first three frames when starting points are completely different and brain analysis and maps with previous imagery information.

Raw data and tests on limited number of participants showed that McGurk Effect exists in virtual world and animation designs. Prepared animation with be shown to different people and results will be analysis if McGurk Effect plays important role in virtual world and animation designing. Different versions on McGurk will be prepared and examined to reach clear idea.

Brain training using McGurk Effect will be another test. Brain gets visual clues and maps it with received auditory information. The process can be used to train the brain to record certain visual information which are not real presentation of auditory information presented. The test, if it works according to theories, will show that brain can be taught or untaught or mis-taught.

3-2 Survey Design

To prepare survey, we worked on different variations to test what happens when factors changes in the videos. According to literature, researchers worked on classic McGurk Effect, desynchronization and noisy environment which all were on real videos and environment. This thesis tried to test these ideas on 3D animation and virtual world. Beside of these factors, human and robotic sounds were tested which were new to the literature. Variations generated by five different factors which gave us $2^6 = 64$ different videos.

There are two different videos and sounds saying “ba” or “fa”. First variation were videos saying “ba” with sounds saying “ba” or “fa” and vice versa. All videos generated considering with or without third factor which is the noise. All these factors prepared with four different amount of delay which gave us desynchronization. 32 videos generated by these factors, were prepared with human and robotic sound.

Human sound was recorded in office environment and robotic sound generated using speech recognition web site*.

* https://text-to-speech-demo.ng.bluemix.net/

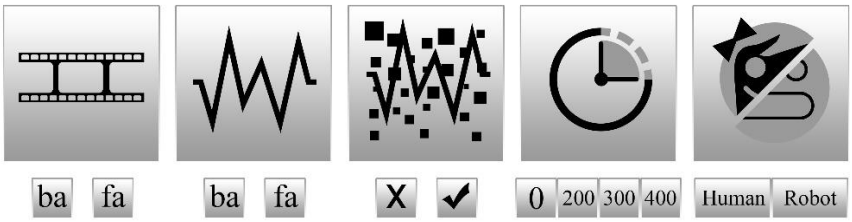

Figure 9 : Variation of factors in videos which present video, sound, noise, delay and kind of sound.

Figure 10 : List of variations and their numbers

All videos are prepared and uploaded to Youtube, video sharing website. Video titles are just numbers to avoid any kind of misleading in survey. Participants could see title of the videos but they did not have any idea about the numbers. There was just one question to decide if the heard “ba”, “fa”, “da” or “va”, and videos were presented randomly. So, order of the video would not affect their decision.

Survey prepared by an online survey tool* and presented in my personal website using a Wordpress plugin. Survey link was hidden from ordinary people and just presented to people who were volunteer in taking the survey. To do so, an announcement presented in social media networks to ask people if they would like to take part. If a person would like to participate, he has to send a private message and get the link. Some participants came to office in city center and took the survey with my supervision. Due to time restrictions and long list of videos, some participants left the survey in middle of the test, so limited number of people could finish all questions which however were enough to get results.

Some unexpected behaviors and answers gave us ideas about new research topics. People reported weird sounds and complained about not hearing some sounds while there were just two sounds.

3-3 Technical Information

A list of tools was used to prepare videos for survey. Below is a list of software used in this thesis:

1. Autodesk Maya
2. Zbrush
3. Mental Ray
4. Adobe Photoshop
5. Adobe Premiere Pro
6. Adobe Audition
7. Adobe Media Encoder
8. Adobe Illustrator
9. Wordpress and Crowdsignal

* https://www.crowdsignal.com/

3-3-1 Autodesk Maya

Since the thesis is about 3D animations and virtual environments, a 3D design tool was needed to model, animate and render the character. To design a 3D character and blueprint of a character is needed, which is a photo of a character from side, front and top. In Autodesk Maya, these photos can be imported from view, image plane, import image. In four view each view can be used for proper side of the photo, and perspective is used to rotate through the model and check the progress.

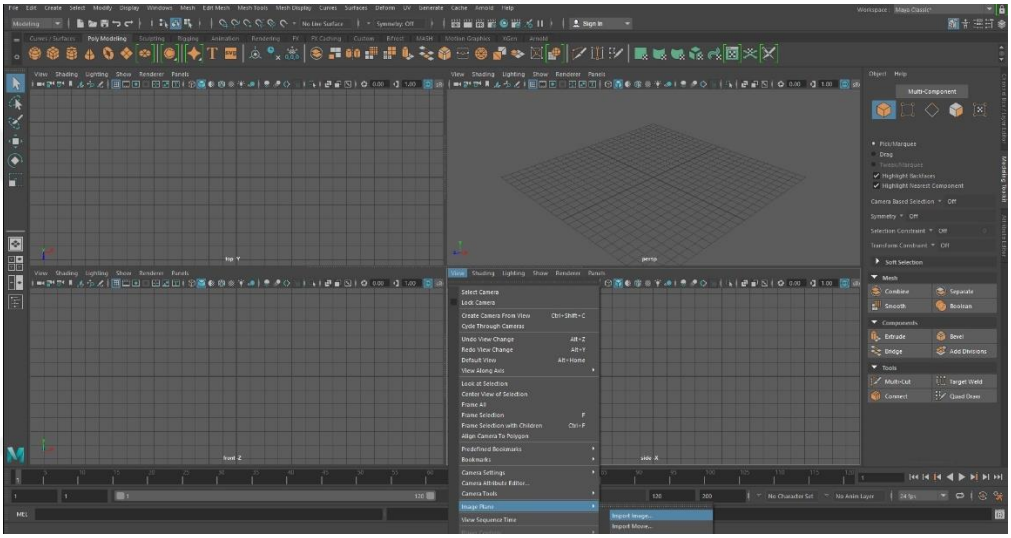

Figure 11 : Autodesk Maya, importing photos to model the character.

Wide variety of methods can be used to model a 3D character but the most common one is starting by a simple plane drawing in front or side view. Then each side can be extruded and continued in modeling using Crtl+E. Then moving tool can move the extruded face to match the photos in different view. It is an exhausting process but basically it is the whole idea in modeling an organic 3D object. A character design could take some weeks to be prepared properly. Other modeling methods, like curves and commands related to curves, usually are not used in character or animal designs.

After modeling the character, rigging is needed to give animation to the character. Rigging is giving bones to models. Without rigging a model and its animation may shape weirdly. In a survey like McGurk Effect, believability is vital to get proper results. Otherwise people would ignore visuals and focus on audios.

Autodesk Maya has an auto rig tool which can recognize the 3D object and rig it properly. However, to get precise result, this process should be done manually. After drawing bones, binding is done to direct how rigs can move. Binding prevent rigs from weird moves. As

an example, an arm can be opened at most, 180 degrees, more than that is not a natural move. Binding prevent these moves and limits it to 180 degrees or whatever amount is needed.

A 3D model in Autodesk Maya is usually colored in grey and to cover it properly texturing process is needed. Texturing is done using photoshop which is a 2D visual editing software. To prepare a 3D model for texturing in Photoshop, unwarp process should be done in Autodesk Maya. In unwarp window there are tools in open up 3D model and optimize it for Photoshop.

Figure 12 : Simple photo presenting the main idea of unwarping process in 3D modeling (unknown origin)

3-3-2 Zbrush

Zbrush is a 2.5D sculpturing software and is used to make details on 3D models. Details like winkles on 3D characters can be done in Zbrush which is used mostly in designing old characters or creatures with a lot of details in modeling. Zbrush rarely used in this project, however, details on models are done through this software.

3-3-3 Mental Ray

After modeling process, the 3D model must be rendered to be used in real world environment. A single frame or an animation is done using a render tool which in this case we used Mental Ray which were a Nvidia rendering tool and was free. Nvidia stopped supporting Mental Ray so people start immigrating to other tools.

Lightening is done using plane light tools. A photography studio is usually simulated in 3D software environment to make it believable for people due to the fact that these kind

of visual are presented to people in various visual presentation. So, people deal with these visual as realistic since their brain usually see these kinds of visuals. Therefore, 3D artists and animators have to simulate such environments to reach realistic designs. Main principles in rendering realistic visuals are lightening and texturing.

A frame by frame rendering process is used to get the animation. This process is the most recommended one since render can be continued if any interruption happens during long period of rendering process.

3-3-4 Adobe Photoshop

A photo editing tool is necessary in designing any kind of visual related project. Photoshop is used in editing textures, preparing unwrapped files and designing visuals needed in the thesis.

Clone tool in Photoshop is used to sample a texture in a photo and copy it onto new visual which is widely used in texture preparing process. Skin texture which is needed in 3D model is prepared by photographing real model and using its photo to copy and paste the texture. This process in done widely in animation and 3D game design industry. Masking is another tool in Photoshop to remove any part of the photo which is not needed. Hue/Saturation tool can be used to change colors of a visual whenever needed.

3-3-5 Adobe Premiere Pro

Premiere Pro is used to animate frames of render from Maya. Frame make a video and using nest tool in Premiere Pro bring them together to make a single line of video. Then by copying it animation can be repeated. By considering first and last frame of the animation, this repetition can be seamless which otherwise a skip happens that can cause distraction.

Sounds are placed using Premiere Pro which has also sound editing tools. Desynchronization needed in the thesis is also done by Premiere Pro. There are different standards in animation fps rate but most common one is 24 fps. Left/Right arrow tool skips one frame in Premiere Pro, so each push on right arrow affect the video 1/24 seconds which almost equals to 0.04 seconds. So 200ms equals to five times of pressing the shortcut and so on.

A lot of visual and auditor layers can be created in Premiere Pro. By enabling and disabling some layers, it is possible to render a visual with different sound and vice versa.

3-3-6 Adobe Audition

Even though there are sound editing tools in Premiere Pro, Audition present more tools to get needed results from sounds. As an example, there were some need to make a sound longer so that it fills the duration of the animation. Time span can easily extend by cutting and dragging audio so that the duration become even in all sound and it can be matched with the video.

Multisession tool in Audition gives access to multiple sound tracks to be worked together. So, it is possible to bring a sound and arrange others according to its duration and volume. Recorded sounds are placed in different tracks and original sound is placed on top. With editing tools durations are matched and volumes increased to match each other's sound waves. Each track can be rendered separately from other tracks. By putting point in line inside tracks, volumes of sounds increased as time passed.

Noise reduction tools in Audition allows users to remove extra sounds from recorded track, so unpleasant environment sounds removed completely from recorded sounds.

3-3-7 Adobe Media Encoder

Adobe Media Encoder provides possibility to render a queue of videos. A list of videos can be sent to Media Encoder and can be rendered collectively. This is a vital feature in working with high number of videos which otherwise would take tremendous amount of time in waiting and rendering one by one.

To use Media Encoder, just pressing "queue" button in exporting window of Premiere Pro will be enough. The software starts automatically and has a green play button to start the process. It is possible to set the format and location to save the file. Media Encoder is widely used with Premiere Pro and After Effects (advanced video editing tool).

3-3-8 Adobe Illustrator

Adobe Illustrator is a vector designing software and can be used in designing icons, logos and these kinds of visual presentation. The icons and visual presentations in this thesis are designed using Adobe Illustrator.

These icons can be designed using geometrical shape existing in the software and by doing mathematical operation such as adding, subtracting etc. on them. There is pen tool which is used to draw a shape from scratch.

3-3-9 Wordpress and Crowdsignal

Wordpress is a CMS (Content Management System) and widely used to prepare server-side websites. Page creating tools are available in Wordpress and can be simply set in a matter of seconds. To do so, a domain and a host is needed. These items should be bought from a website that presents host services. Then Wordpress can be installed and prepared. After installation a template is selected and proper colors are set. After these steps, creating pages, menus etc. would take a short amount of time. However, I used my website which were designed and published some years ago. To hide survey from random people to access, the page included the survey did not added to the menu list, so just people with the link could access it. The possibility to access the link from other people were near to zero if however, we did not consider it as zero.

Crowdsignal is a tool developed by Wordpress to do online surveys. It can be added as plug-in to the wordpress website. There are tools to create different kinds of question in the Crowdsignal. This thesis just needed multiple choice questions in the survey steps. In demographic information part, free text and date types are also used.

Multiple choice question type gives access to bring codes from other websites so that an embed code from Youtube sharing system can be place here. Then video appears in the survey. From multiple choice part, different answers entered and title is used to write the question which in this case was a similar question. The cloning tool is used to clone the questions and videos are replaced with new ones. There is a randomize option which allow users to present their questions randomly which was vital in this thesis.

After preparing the survey, it can be added to a Wordpress website using short code generated from the tool. Crowdsignal plugin must be added to the website to get access to the survey.

CHAPTER 4

RESULTS

Survey is done during five days and participants were usually Turkish people resident in Ankara. 34 participants finish the survey and more than 15 people left the survey unfinished due to some technical problems (some videos could not be loaded due to low internet speed to some wi-fi connectivity problems) or compact daily schedule. Participants are usually between 25-35 age. 58% of participants were female and 42% were male. 76% of participants were university graduated,15% had Master degree and 9% were high school graduated. People from variety of backgrounds took part in the survey. Among participants there were photographer, architecture, economist, electric engineer, English teacher and so on. All participants mother language (mother tongue) are Turkish. Most participants mentioned intermediate level of English as their second language and there were some cases of German, Italian and Persian language to be third or fourth language.

Videos and questions are presented randomly and participants are asked to select whatever they hear from videos. In this study people were free to answer as they like, however, some participants are witnessed not watching videos as they were trying to answer, which could lead to unexpected results. Some people watched videos several times to decide what they heard. Some of participants were watching monitor but focusing on hears instead of eyes, so they weren't looking away but they weren't watching too. These factors prevented complete intended results. Regardless of such efforts McGurk effect still happened and people act as they were expected according to McGurk Effect.

Videos numbering from 9-12 showed McGurk effect happens, when video presents "fa" and audio presents "ba", 88% of people heard sound of "va" which showed Confusion Effect of McGurk Effect. When brain cannot decide if "fa" or "ba" is presented, it generates third concept which is not any of them and in this case, it is "va". As delay increases from video number 9 to video number 12, people start to ignore visual and focus on audio, still McGurk Effect happens as 79% still select "va" and 21% select "ba" which were what the sound were presenting.

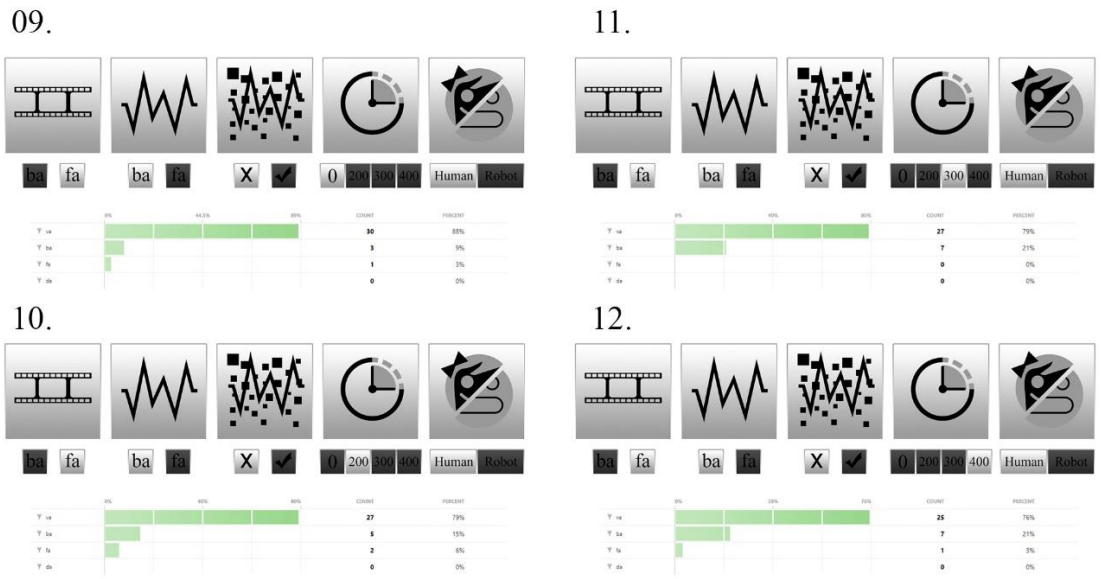

Figure 13 : Results of the survey, videos from 9 to 12.

In video numbering from 25-28 same results happens, with noise added to previous version of the videos. As an unexpected result, number of people hearing “ba” increased in comparison with previous results. While it was expected to decreased due to the fact that noise causes people to focus on visuals. Participants who watched away as they were doing survey could cause such results. So, it is better to say to people that it is obligatory to watch as they do the survey.

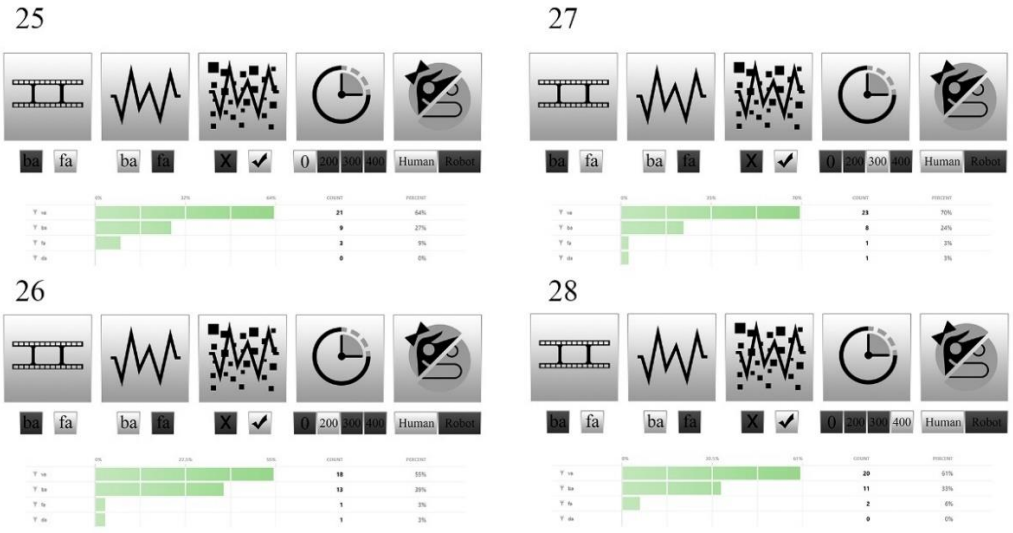

Figure 14 : Results of the survey, videos from 25 to 28.

Results from same videos with robotic sounds presents the idea that people focus on audio more that visuals when they are sure about the sound. However, as noise added to the audio, people start to select mostly "va" which shows even if the sound were clearly saying "ba" other sounds from environment push people to focus on visual and trust it. In this case cultural factors are also play an important role. This may cause different result with a survey on English people.

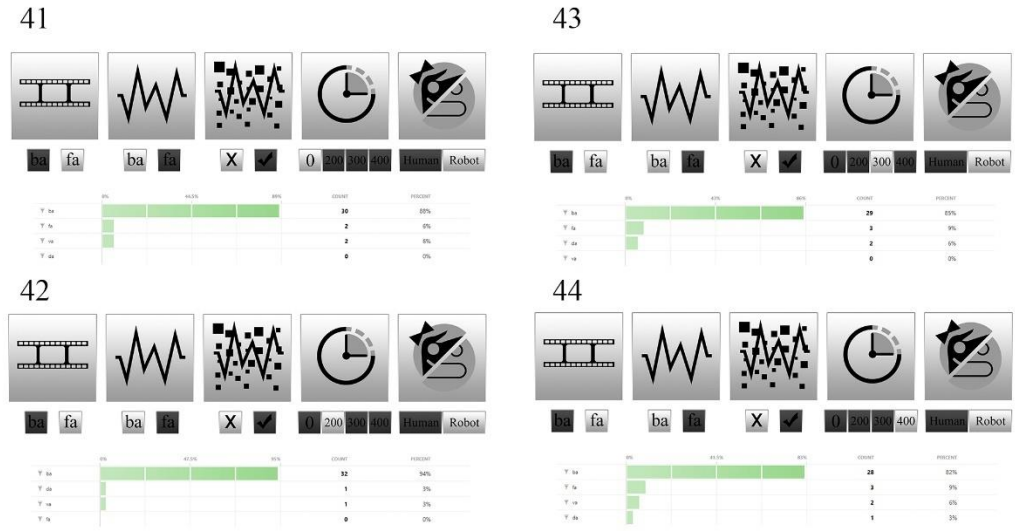

Figure 15 : Results of the survey, videos from 41 to 44.

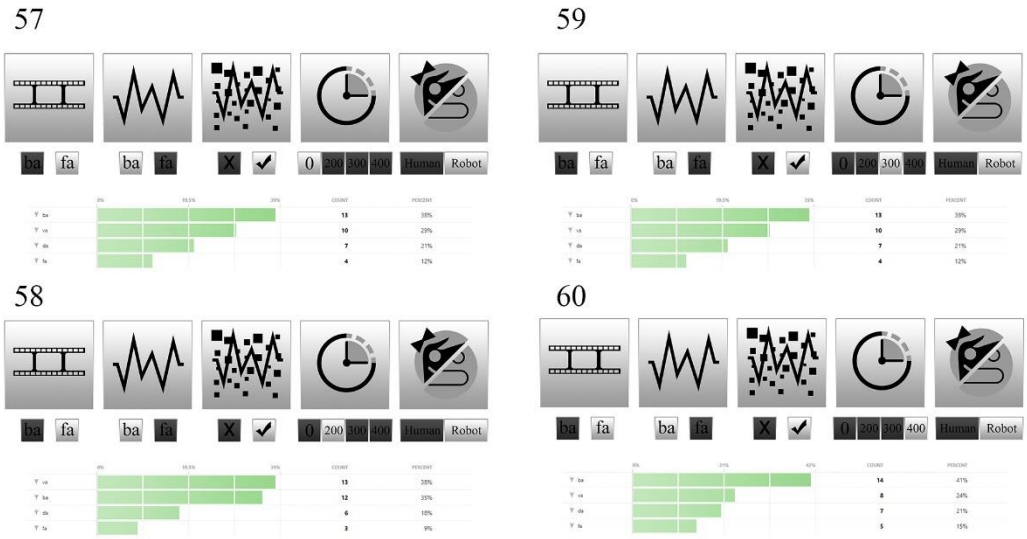

Figure 16 : Results of the survey, videos from 57 to 60.

A whole list of results is presented as table in the attachments with details.

There were some unexpected results and behavior from participants. A participant reported just hearing “va” from male voice, and said female voice always said “fa” and “ba”. This case is completely weird since there was not male voice in any of 64 videos. It can be theorized as brain confusion whenever it cannot decide about the result of an incoming signals and decides completely different thing rather than any of the incoming results.

A number of people reported to select just according to the voice. They said they completely ignored visual elements and just focus on the sound. There must be a method preventing people from ignoring the visual elements.

Some of the participants watched a single video several times. This behavior was unexpected so there was not a prevention. Rules should not be changed during a survey so we avoid changing it but note it as a role-playing factor.

A group of participants paused video immediately and select the answer. These people most probably focused of audio and ignored the video. This behavior also was unexpected and there was not a prevention.

Some people took part from their mobile phones, so visual quality and size would affect their attention and most likely they focus on sounds so the results may become different than real expected one.

Weird results happened when video presents “ba” and audio was also “ba”, some participants selected “va”. It becomes even more weird when these participants select “ba” when there was noise with the sound.

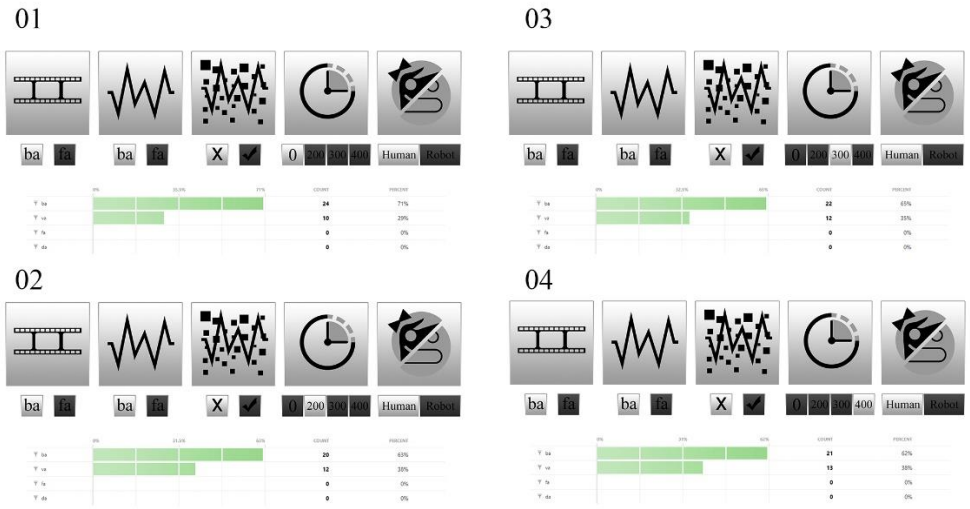

Figure 17 : Results of the survey, videos from 1 to 4.

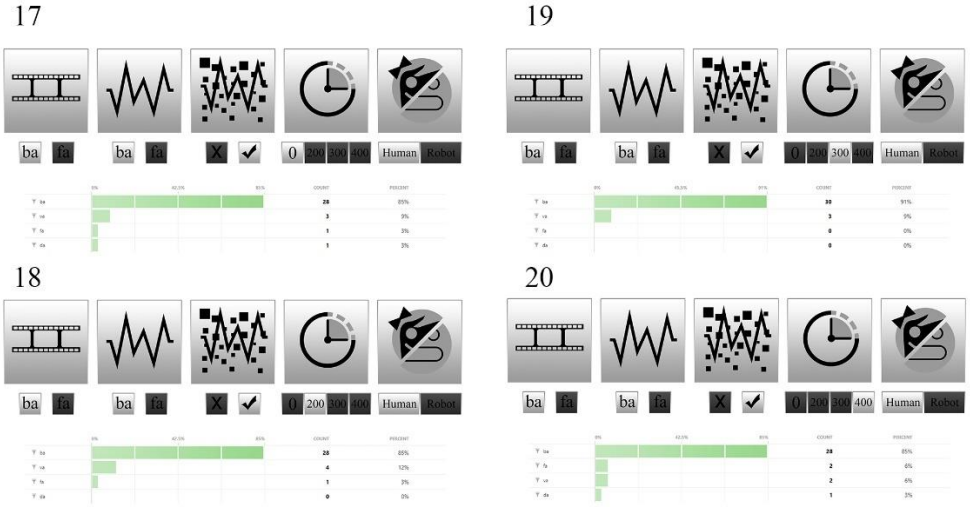

Figure 18 : Results of the survey, videos from 17 to 20.

CHAPTER 5

DISCUSSION AND CONCLUSION

McGurk Effect is such a powerful effect that even when people know the idea, they cannot prevent it from happening. It happened even when some participants saw same sound file being copied under different video files. It is one of the best methods in understanding brain process during analyzing environment. Such understanding from brain behavior can help different fields of study to approach accordingly.

This thesis and experiment showed that McGurk Effect also exist in 3D animation and virtual world. Variety of factors changed to understand brain behavior in different cases. It is found that desynchronization causes people to ignore visual after 400ms. At 200ms McGurk still works and 300ms is the threshold of changing point.

It is also discovered that noise in sound push brain to look for other clues to understand signals. In video containing noise, results were showing McGurk appearing as Confusion Effect when brain cannot decide what was signals about.

Some unexpected results give ideas about future works. There must be some methods to prevent people from watching away as they take the survey. Secondly participant should be expected to watch whole video and once. To do so, survey questions can be prepared to pop up after video ended, so people cannot return to watch the video again, and they should not be able to pause the video.

Some participants complained about not hearing “da” sound, so there can be some videos containing “da” sounds to be placed among other videos.

Cultural factors also affect the results. This survey can be done with some English people to get ideas about how culture and mother language (mother tongue) would affect the result.

Participating using small screen like mobile phone increased role of audio, so all people should be pushed to participate using a big screen computer, so they were pushed to watch the video.

Behavioral study can be done on prepared videos. When people encounter difficulty in understanding and answering question, they start to show different physical movements. Such movements can be interpreted as brain trying to find some clues.

The experiment of this thesis was unique in some case in literature. McGurk Effect has never done on 3D animation and never done with a computer-generated voice. 64 variation of different factors is an original idea and is a comprehensive study in this case. By removing some distracting factors mentioned above, this study can present more detailed results comparing to any other studies in this field. A developed version of this study can be done on behavior and emotional cases. Such studies will provide ideas about brain process with more details.

REFERENCES

[1] McLeod S. (2018) "Visual Perception Theory" Retrieved from https://www.simplypsychology.org/perception-theories.html [Accessed July-2019]

[2] Clark A. (2013) "Whatever next? Predictive brains, situated agents, and the future of cognitive science" School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Scotland, United Kingdom.

[3] Sloos M. (2015) "Bias in Auditory Perception" Interacting Minds Centre, Aarhus University, Aarhus, Denmark, Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5016825/ [Accessed July-2019]

[4] Hilligoss S. and Howard T. (2002) "Visual Communication: A Writer's Guide" Clemson University, Published as Longman Publishers, Printed in the USA.

[5] Manhartsberger M. and Zellhofer N. (2005) "Eye tracking in usability research: What users really see" Interface Consult GmbH, Vienna, Austria.

[6] Demiralp C. et al (2015) "The VERP Explorer: A Tool for Exploring Eye Movements of Visual-Cognitive Tasks Using Recurrence Plots" Institute for Visualization and Interactive Systems, Stuttgart, Germany.

[7] Burnham D. and Dodd B. (2018) "Language–General Auditory–Visual Speech Perception: Thai–English and Japanese–English McGurk Effects" MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, Australia.

[8] Isbister K. (2006). "Better Game Characters By Design: A Psychological Approach" Elsevier's Science & Technology Department Oxford, UK.

[9] Giordano B. L. et al (2017) "Contributions of local speech encoding and functional connectivity to audio-visual speech perception" Institut de Neurosciences de la Timone UMR 7289, Aix Marseille Université – Centre National de la Recherche Scientifique, France.

[10] Blackburn C. L. et al (2019) "Visual Speech Benefit in Clear and Degraded Speech Depends on the Auditory Intelligibility of the Talker and the Number of

Background Talkers" Department of Psychology, Nottingham Trent University, UK.

[11] Donald F. Roberts. et al (2005). "Generation M: Media in the Lives of 8–18 Year-olds" Stanford University, USA.

[12] Asbell-Clarke J. and et al (2013). "Assessment Design for Emergent Game-Based Learning" University of Terc, Massachusetts, USA.

[13] "Understanding Media and Culture" (2013). Publisher: Saylor Academy, USA.

[14] Norman D. A. (1990). "The Design Of Everyday Things" Doubleday Publishing, New York, USA.

[15] Ware C. (2004). "Information Visualization: Perception For Design" Morgan Kaufmann publications, an imprint of Elsevier, San Francisco, USA.

[16] Reeves B. and Nass C. (2013). "The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places" Stanford University, California, USA.

[17] Chi-Hsun C. (2013). Verification Of Theory Based Design Features For Designing Online Instruction For Students With Learning Disabilities And Other Struggling Learners (Doctoral dissertation) Retrieved from https://pqdtopen.proquest.com/doc/1353184391.html?FMT=AI [Accessed July-2019]

[18] Tiippana K. (2014) "What is the McGurk effect?" Division of Cognitive Psychology and Neuropsychology, Institute of Behavioral Sciences, University of Helsinki, Helsinki, Finland.

[19] Munhall K. G. et al (1996) "Temporal Constraints on the McGurk Effect" Queen's University and ATR Human Information Processing Research Laboratories Kingdom, Ontario, Canada.

[20] Mallick D. B. et al (2015) "Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type" Department of Psychology, Rice University, Houston, USA.

[21] Ali A. N. et al (2005) “McGurk Fusion Effects In Arabic Words” University of Huddersfield and Al Akhawayn University, UK.

[22] Sams M. et al (2012) “Audiovisual Fusion In Finnish Syllables And Words” Department of Psychology, University of Tampere, Tampere, Finland.

[23] Sporea I. and Gruning A. (2010) “Modelling the McGurk effect” University of Surrey, Department of Computing, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, Surrey, UK.

[24] Aruffo C. (2009) “Can you McGurk yourself? Self-face and self-voice in audiovisual speech” (Master of Science Thesis). Boston University, USA.

[25] Fagel S. (2006) “Emotional McGurk Effect” Institute for Speech and Communication Technical University Berlin, Berlin, Germany.

[26] Bovo R. et al (2009) “The McGurk phenomenon in Italian Listeners” Audiology and Phoniatrics Department, University Hospital of Ferrara, Ferrara, Italy.

APPENDICES

APPENDIX A

SURVEY RESULTS WITH NUMBERS

Visual	Auditory	Noise	Delay	Human/Robot	ba	fa	da	va
ba	ba	x	0	Human	24	0	0	10
ba	ba	x	200	Human	20	0	0	12
ba	ba	x	300	Human	22	0	0	12
ba	ba	x	400	Human	21	0	0	13
ba	fa	x	0	Human	1	33	0	0
ba	fa	x	200	Human	2	32	0	0
ba	fa	x	300	Human	2	32	0	0
ba	fa	x	400	Human	2	32	0	0
fa	ba	x	0	Human	3	1	0	30
fa	ba	x	200	Human	5	2	0	27
fa	ba	x	300	Human	7	0	0	27
fa	ba	x	400	Human	7	1	0	25
fa	fa	x	0	Human	0	34	0	0
fa	fa	x	200	Human	0	33	0	1
fa	fa	x	300	Human	0	33	0	1
fa	fa	x	400	Human	0	32	1	1
ba	ba	✓	0	Human	28	1	1	3
ba	ba	✓	200	Human	28	1	0	4
ba	ba	✓	300	Human	30	0	0	3
ba	ba	✓	400	Human	28	2	1	2
ba	fa	✓	0	Human	8	23	0	2
ba	fa	✓	200	Human	9	23	0	1
ba	fa	✓	300	Human	8	22	1	2
ba	fa	✓	400	Human	11	19	1	2
fa	ba	✓	0	Human	9	3	0	21
fa	ba	✓	200	Human	13	1	1	18
fa	ba	✓	300	Human	8	1	1	23
fa	ba	✓	400	Human	11	2	0	20
fa	fa	✓	0	Human	0	31	0	2
fa	fa	✓	200	Human	1	29	1	2
fa	fa	✓	300	Human	0	31	0	2
fa	fa	✓	400	Human	1	31	0	1
ba	ba	x	0	Robot	32	2	0	0
ba	ba	x	200	Robot	30	2	1	1
ba	ba	x	300	Robot	30	1	1	2
ba	ba	x	400	Robot	29	1	2	2
ba	fa	x	0	Robot	3	28	0	3
ba	fa	x	200	Robot	1	32	0	1
ba	fa	x	300	Robot	6	27	0	1
ba	fa	x	400	Robot	1	31	1	1
fa	ba	x	0	Robot	30	2	0	2
fa	ba	x	200	Robot	32	0	1	1
fa	ba	x	300	Robot	29	3	2	0
fa	ba	x	400	Robot	28	3	1	2
fa	fa	x	0	Robot	3	28	0	3

Visual	Auditory	Noise	Delay	Human/Robot	ba	fa	da	va
fa	fa	x	200	Robot	3	29	1	1
fa	fa	x	300	Robot	3	27	1	3
fa	fa	x	400	Robot	2	29	0	3
ba	ba	✓	0	Robot	26	2	4	0
ba	ba	✓	200	Robot	28	1	2	2
ba	ba	✓	300	Robot	27	1	5	1
ba	ba	✓	400	Robot	25	2	6	1
ba	fa	✓	0	Robot	11	21	1	1
ba	fa	✓	200	Robot	9	21	1	3
ba	fa	✓	300	Robot	6	27	0	0
ba	fa	✓	400	Robot	13	18	1	1
fa	ba	✓	0	Robot	13	4	7	10
fa	ba	✓	200	Robot	12	3	6	13
fa	ba	✓	300	Robot	13	4	7	10
fa	ba	✓	400	Robot	14	5	7	8
fa	fa	✓	0	Robot	1	28	1	4
fa	fa	✓	200	Robot	4	27	0	2
fa	fa	✓	300	Robot	5	28	0	3
fa	fa	✓	400	Robot	0	28	1	5

APPENDIX B

ALL THE RESULT WITH VISUAL PRESENTATION AND PERCENTAGE


ba
fa
X
0
200
300
400
Human
Robot


ba
fa
X
0
200
300
400
Human
Robot


ba
fa
X
0
200
300
400
Human
Robot