ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF ARTS AND SOCIAL SCIENCES

OBSERVING THE LIMITS OF VISUAL BIAS ON SOUND SOURCE LOCALIZATION IN VENTRILOQUISM EFFECT: A CASE STUDY ON SOUND ENGINEERS VS. NAIVE LISTENERS

M.A. THESIS

Naci TEPEDELEN

Department of Music

Music M.A. Programme

DECEMBER 2017

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF ARTS AND SOCIAL SCIENCES

OBSERVING THE LIMITS OF VISUAL BIAS ON SOUND SOURCE LOCALIZATION IN VENTRILOQUISM EFFECT: A CASE STUDY ON SOUND ENGINEERS VS. NAIVE LISTENERS

M.A. THESIS

Naci TEPEDELEN
(409151113)

Department of Music

Music M.A. Programme

Thesis Advisor: Yrd. Doç. Dr. Taylan ÖZDEMİR

DECEMBER 2017

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ SOSYAL BİLİMLER ENSTİTÜSÜ

GÖRSEL UYARANLARIN SES KAYNAĞI LOKALİZASYONUNA ETKİSİ (VANTRİLOK ETKİSİ): SES MÜHENDİSLERİ VE NORMAL DİNLEYİCİLERİN KARŞILAŞTIRILMASI ÜZERİNE ÖRNEK ÇALIŞMA

YÜKSEK LİSANS TEZİ

Naci TEPEDELEN
(409151113)

Müzik Anabilim Dalı

Müzik Yüksek Lisans Programı

Tez Danışmanı: Yrd. Doç. Dr. Taylan ÖZDEMİR

ARALIK 2017

Naci Tepedelen, a M.A student of ITU Graduate School of Arts and Social Sciences student ID 409151113 successfully defended the thesis entitled "OBSERVING THE LIMITS OF VISUAL BIAS ON SOUND SOURCE LOCALIZATION IN VENTRILOQUISM EFFECT: A CASE STUDY ON SOUND ENGINEERS VS. NAIVE LISTENERS", which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Yrd. Doç. Dr. Taylan ÖZDEMİR** Istanbul Technical University	.....
Jury Members :	**Doç Dr. Can KARADOĞAN** Istanbul Technical University	.....
	Yrd. Doç. Dr. Yahya Burak TAMER Bahçeşehir University	.....

Date of Submission : 22 November 2017

Date of Defense : 12 December 2017

FOREWORD

People perceive the events around them with their sensory systems. These events can be examined temporally or spatially. By using temporal and spatial information, they determine the location of those events in space with their time. To able to this, their sensory system communicates with each other. In some cases, sensory system integration can cause getting wrong information.

For instance, when the auditory system of a person localized a sound source, visual stimuli of that sound source can affect the perception of hearing. Observing this perceptual shift have been subject of many researchers. Also in the recent thesis, an experimental setup was designed for analyzing this issue.

I want to thank my thesis advisor Yrd. Doç. Dr. Taylan Özdemir for help and guidance of this thesis at all stages. Also, I want to thank Doç. Dr. Can Karadoğan for opening my horizon about this subject.

November 2017 Naci TEPEDELEN

TABLE OF CONTENTS

Page

FOREWORD ..... **vii**
TABLE OF CONTENTS ..... **ix**
ABBREVIATIONS ..... **xi**
SYMBOLS ..... **xiii**
LIST OF TABLES ..... **xv**
LIST OF FIGURES ..... **xvii**
SUMMARY ..... **xix**
ÖZET ..... **xxi**
1. INTRODUCTION **1**
1.1 Ventriloquism Effect 1
1.2 Literature Review 1
1.3 Purpose of Thesis 4
1.4 Methodology 5
1.4.1 Experiment 5
1.4.2 Calculations 6
1.5 Hypotheses 8
2. OBSERVING THE LIMITS OF VISUAL BIAS ON SOUND SOURCE LOCALIZATION IN VENTRILOQUISM EFFECT: A CASE STUDY ON SOUND ENGINEERS VS. NAIVE LISTENERS **9**
2.1 Which Factors Influence the Results of the Experiment 9
2.1.1 The experience of subjects 10
2.1.2 Spatial and temporal disparity 10
2.1.3 Compelingness 10
2.1.3.1 Concrete or abstract relation (conceptual relation) 12
2.1.3.2 Static or dynamic stimuli (motion relation) 13
2.1.4 Attention 15
2.1.5 Audio-only localization performance 16
2.2 Experimental Setup 17
2.2.1 Premature thoughts and thoughts after analyzing the constraints 18
2.2.2 Stimulus and grids 20
2.2.3 Subjects and questions 25
2.2.4 Placement of loudspeakers 25
2.2.5 Applying the experiment and collecting data 25
3. RESULTS AND DISCUSSION **29**
3.1 Visual Bias On Sound Source Localization 29
3.1.1 Audio visual task 30
3.1.2 Enveloped audio-animated visual task 34
3.2 Audio Only Performances 40
4. CONCLUSION **49**
REFERENCES **53**
APPENDICES **57**
CURRICULUM VITAE **65**

ABBREVIATIONS

AO	**:** Audio-only
AV	**:** Audio-visual
En. A-An. V	**:** Enveloped audio-Animated Visual
HDTV	**:** High definition television
ITD	**:** Interaural time difference
ILD	**:** Interaural level difference
ms	**:** Milisecond
OTAM	**:** Otomotiv Teknolojileri Araştırma Geliştirme San. ve Tic. A.Ş.
MIAM	**:** Müzik İleri Araştırmalar Merkezi
est.	**:** Entered estimations in Excel table
err	**:** Error amount of subjects in Excel table
div.	**:** Angular divergance
%div.	**:** % Angular divergance
Subj	**:** Subject

SYMBOLS

μ: Discrimination angle
α : Maximum angular differences between two speakers in specific grid section
β : Minimum angular differences between two speakers in specific grid section
γ : Subject's angle of view

LIST OF TABLES

Page

Table 2.1: Grouping the influences factors on visual bias in azimuth. **9**
Table 2.2: Calculations for each grid setups **24**
Table 2.3: Question distribution for subjects and tasks **25**

LIST OF FIGURES

Page

Figure 2.1: Hierarchic grouping of audio-visual relations in compelingness main title..... **14**
Figure 2.2: Possible results of audio-visual relations and their effects on the visual bias. **14**
Figure 2.3: First experimental design idea (top view). When the subject can understand the sound source is not coming from the real sound source, the degree of α is obtained. **18**
Figure 2.4: Grid design for the current study. Stable speaker positions. **19**
Figure 2.5: Upper pictures are from MIAM and the bottom one is from OTAM. ... **19**
Figure 2.6: 2000 Hz sine wave burst waveform and "Operator" settings..... **20**
Figure 2.7: White noise signal burst waveform and "Operator" settings. **21**
Figure 2.8: White noise signal enveloped waveform and "Operator" settings..... **21**
Figure 2.9: Top view of 5-grid setup. **22**
Figure 2.10: Projected 5-grid and static visual stimuli. **23**
Figure 2.11: Grids, visual stimuli and calculations in 3d space..... **23**
Figure 2.12: 17-grid setup, one specific example of the spatial disparity between audio and visual stimuli..... **25**
Figure 2.13: A part of Excel table for collecting estimations and calculating data for 5-grid setup. **27**
Figure 2.14: A part of Ableton project that includes noise and sine stimulus channels, video of the grids and stop track for 5-grid setup **28**
Figure 2.15: Routing for noise stimulus..... **28**
Figure 3.1: %Error amount for noise and sine waves in 4 different grid setups..... **29**
Figure 3.2: %Error amount in audio-only and audio-visual tasks. **30**
Figure 3.3: %Error amount of sound engineers and naive listeners in audio-only and audio-visual tasks. **30**
Figure 3.4: %Error amount of sine and noise in audio-only and audio-visual tasks. **31**
Figure 3.5: Visual biases in grid level (Sound engineers vs. Naive listeners)..... **31**
Figure 3.6: Visual biases in grid level (Sine vs. Noise)..... **32**
Figure 3.7: AO and AV estimations for sine wave (visual at 7th grid) **32**
Figure 3.8: AO and AV estimations for noise signal (visual at 7th grid) **33**
Figure 3.9: Visual biases percentage in AV task (Sine vs. Noise)..... **33**
Figure 3.10: Visual biases percentage in AV task (Sound engineers vs. Naive listeners). **34**
Figure 3.11: AO vs. AV vs. En. A-An. V %error (all subject groups for noise signa13 grid)..... **35**
Figure 3.12: AO vs. AV vs. En. A-An. V %error (Sound engineers vs. naive listeners noise-13 grid)..... **35**
Figure 3.13: Visual biases in grid level (AV vs. En. A-An. V) for noise signal **36**
Figure 3.14: Visual biases in grid level (AV vs. En. A-An. V) for noise signal **36**
Figure 3.15: AO and En. A-An. V estimations for noise signal (visual at 7th grid).. **37**

Figure 3.16: Visual biases percentage (AV vs. En. A-An. V) for noise signal..... **37**
Figure 3.17: Visual biases percentage in En. A-An. V task (Sound engineers vs. Naive listeners). **38**
Figure 3.18: En. A-An. V estimations for noise signal (visual at 6th grid) **38**
Figure 3.19: En. A-An. V estimations for noise signal (visual at 8th grid) **39**
Figure 3.20: %Errors in audio-only task, sound engineers vs. naive listeners..... **40**
Figure 3.21: %Errors in audio-only task, sound noise vs. sine. **41**
Figure 3.22: Requested Sound Positions (All questions in AO test) **42**
Figure 3.23: All estimations in audio-only task for sine questions (All Subjects) ... **43**
Figure 3.24: All estimations in audio-only task for noise questions (All Subjects) . **43**
Figure 3.25: Mean angular divergence comparison between sine and noise signal . **44**
Figure 3.26: Estimations for Sine Questions (Naive Listeners)..... **45**
Figure 3.27: Estimations for Sine Questions (Sound Engineers)..... **45**
Figure 3.28: Estimations mean, Sound Engineers vs. Naive Listeners (Sine)..... **46**
Figure 3.29: Estimations for Noise Questions (Naive Listeners) **47**
Figure 3.30: Estimations for Noise Questions (Sound Engineers) **47**
Figure 3.31: Estimations mean, Sound Engineers vs. Naive Listeners (Sine **48**
Figure A.1: Experiment area **57**
Figure A.2: Projected grids **57**
Figure A.3: General view of experiment area **58**
Figure B.1: AO and AV estimations for sine wave (visual at 6th grid)..... **59**
Figure B.2: AO and AV estimations for noise signal (visual at 6th grid)..... **59**
Figure B.3: AO and AV estimations for sine wave (visual at 8th grid)..... **59**
Figure B.4: AO and AV estimations for noise signal (visual at 8th grid)..... **60**
Figure B.5: AO and AV estimations for sine wave (visual at 8th grid)..... **60**
Figure B.6: AO and AV estimations for noise signal (visual at 8th grid)..... **60**
Figure B.7: AO and AV estimations for sine wave (visual at 10th grid)..... **61**
Figure B.8: AO and AV estimations for noise signal (visual at 10th grid)..... **61**
Figure B.9: AO and AV estimations for sine wave (visual at 4th grid)..... **61**
Figure B.10: AO and AV estimations for noise signal (visual at 4th grid)..... **62**
Figure B.11: AO and AV estimations for sine wave (visual at 7th grid)..... **62**
Figure B.12: AO and AV estimations for noise signal (visual at 7th grid)..... **62**
Figure B.13: AO and En. A-An. V estimations for noise signal (visual at 7th grid) . **63**
Figure B.14: AO and En. A-An. V estimations for noise signal (visual at 8th grid) . **63**
Figure B.15: AO and En. A-An. V estimations for noise signal (visual at 8th grid) . **63**
Figure B.16: AO and En. A-An. V estimations for noise signal (visual at 10th grid) **64**
Figure B.17: AO and En. A-An. V estimations for noise signal (visual at 4th grid) . **64**
Figure B.18: AO and En. A-An. V estimations for noise signal (visual at 7th grid) . **64**

OBSERVING THE LIMITS OF VISUAL BIAS ON SOUND SOURCE LOCALIZATION IN VENTRILOQUISM EFFECT: A CASE STUDY ON SOUND ENGINEERS VS. NAIVE LISTENERS

SUMMARY

Visual stimuli and its sound stimuli are judged by the human brain together and perceived them as a single event, even if both are assessed by different sensory systems. Also even if these two stimuli are realized in different spatial positions, a human can perceive both stimuli at the same spatial position. Generally, depending on the properties of the stimuli, the visual sensory system dominates the auditory sensory system, and the sound perceptually is heard from the position where the visual is located. This domination of visual system over auditory one is called ventriloquism effect.

Past studies have investigated ventriloquism effect with various experiments. While neuroscientists have mostly focused on its effect on human perception, sound engineers have mostly focused on its effects on sound localization.

In order to observe the ventriloquism effect, an experimental setup was designed for this thesis, under the guidance of the past studies. The primary goal was to understand the effects of visual stimuli on sound localization, just as it is in the sound engineers' studies. For this, the experiment was composed of two main parts. The first part included only sound localization questions, while the second part had visual and sound stimuli together to investigate the effects of visual stimuli on sound source localization. The results from these two sections were assessed separately and interdependently. As a result, it was observed that the sound estimations were given under the effect of visual stimuli as in the previous studies. Furthermore, this effect occurred at different rates depending on the influences of experiment variables as previously predicted in the light of past studies. Some of these influencing factors are; characteristic of audio-visual togetherness, subjects' experiences, spatial discordance amount between audio and visual stimuli, directing subjects with the instruction of the experiment, localization accuracy of sound stimuli and reliability of visual stimuli.

Variables of this experiment were also prepared considering the influencing factors those mentioned above. The noteworthy results of this experiment were obtained when these variables were compared. For instances, estimations of sound engineers and naive listeners were evaluated as two separate subject groups, taking into account the experience of the subjects on the sound source localization. Audio stimuli with two different frequency characteristics (2000 Hz sine and white noise), was chosen for comparing them in localization accuracy. Also by changing the relationship between audio and visual stimuli, characteristic of togetherness was differentiated and those were compared too.

If the results of the audio-only test are given at first, there was no significant difference between the sound engineer and the naive listeners. While sound engineers determined the location of the sounds with an error of 1.9º and 3.4º standard deviation, naive listeners determined with an error 2º and standard deviation of 3.2º. Sine wave were localized by both subject groups with a slightly larger angular difference than the noise signal. While the location of the sine wave can be determined with an average of 4,1 º error and 2,3 º standard deviation, the noise signal was determined with a 3.2 º error and a standard deviation of 1.8º. One of the interesting results of the experiment was that the estimations of the sine wave tend to shift towards the center when the actual sound position is taken into account, whereas the noise signal estimates tend to be out of the center.

When the audio and visual were presented together with spatial disparity, the first salient thing that was to observe the maximum visual bias rate when the visual stimuli were presented at the center. When the average bias effect of all visual positions is taken into account, sound localization estimations of the sound engineers shifted towards the visual by 15.2%, while the estimations of the naive listeners were 13.7%. Estimations made for sine shifted towards the visual by 18.1%, while the estimations for noise signal were 10.8%. However, only when it is considered that the visual is presented at the central position; sound engineers' sound localization estimations 69,7%, naive listeners 46%; localization estimations of the sine wave 62.9%, and noise signal shifted towards the visual by 52.8%.

In the audio-visual test, another experimental variable was created by adding motion relation to the audio-visual togetherness. To put it briefly, the visual stimulus was made to have the characteristic of slowly fade in and fade out, rather than suddenly appearing and disappearing. Along with this, attack and release times were given to audio signal according to the movement of the visual. As a result, visual and audio stimuli, which tend to move together, had more effects on subjects' perception than static audio-visual relation. So estimations were shifted towards the visual position more. When all visual positions and all estimations are taken into account, the visual bias rate created by stimuli with static features was 10.8%, while the visual bias created by motion-related stimuli was measured as 28.7% (this comparison was made only on noise signal estimations). Furthermore, for centrally presented visual stimuli, when the audio engineer's noise signal estimations in the previous test shifted towards the visual by 59.5, in the second it increased to %86,5, also naive listeners’ estimations increased from %46,5 to %79,6.

GÖRSEL UAYARANLARIN SES KAYNAĞI LOKALİZASYONUNA ETKİSİ (VANTRİLOK ETKİSİ): SES MÜHENDİSLERİ VE NORMAL DİNLEYİCİLERİN KARŞILAŞTIRILMASI ÜZERİNE ÖRNEK ÇALIŞMA

ÖZET

Bir görsel uyaran ve onun işitsel uyaranı, farklı duyu sistemleri tarafından değerlendirilse bile algısal olarak bu iki uyaranı insan beyni birlikte değerlendirir ve tek bir olay gerçekleşiyormuş gibi algılar. Hatta bu iki uyaran farklı mekansal konumlarda gerçekleşse bile, belli konum açısı farklılıklarına kadar, insan algısında görsel-işitsel olayın aynı konumda gerçekleştiği hissi uyanır. Genellikle uyaranların özelliklerine bağlı olarak görsel duyu sistemi işitsel duyu sistemine baskın çıkar ve sesin görselin bulunduğu konumdan geldiği algısı uyanır. Görselin ses üzerindeki bu üstünlüğüne vantrilok etki denir.

Geçmiş çalışmalar çeşitli deneylerle bu etkiyi araştırmışlardır. Genellikle sinirbilimi alanında çalışmalar yapan araştırmacılar bu durumun insan algısı üzerindeki etkisine odaklanırken, ses mühendisleri ise sesin lokalizasyonu üzerindeki etkilerine odaklanmışlardır.

Vantrilok etkiyi gözlemlemek adına, geçmiş çalışmalardan yola çıkarak, bu tez için bir deney düzeneği hazırlandı. Öncelikli amaç ses mühendislerinin çalışmalarında olduğu gibi, görsel uyaranların ses lokalizasyonuna etkilerini ölçebilmekti. Bunun için deney iki ana bölümden oluşturuldu. İlk bölüm sadece ses lokalizasyonu soruları içerirken ikinci bölümde işitsel ve görsel uyaranlar birlikte sunularak görselin ses lokalizasyonuna etkileri incelendi. Bu iki bölümden elde edilen sonuçlar ayrı ayrı ve birbirlerine bağlı olarak değerlendirildi. Sonuç olarak önceki çalışmalarda olduğu gibi ses tahminlerinin görsel etki altında verildiği gözlemlendi. Fakat bu etki geçmiş çalışmaların ışığında önceden öngörüldüğü gibi deneyin değişkenlerine bağlı olarak farklı oranlarda vuku buldu.

Önceden araştırılan bu değişkenleri kısaca sıralamak gerekirse; işitsel-görsel olayın inandırıcılığı, deneklerin konuyla ilgili tecrübeleri, ses ve görselin kaç derece açı farkıyla sunulduğu, deneklerin deney sırasında nasıl yönlendirildiği, kullanılan sesin konumunun kolay veya zor belirlenebiliyor oluşu, kullanılan görselin kolay veya zor algılanabiliyor oluşu gibi değişkenleri sayabiliriz.

Bu çalışmada kurulan deney düzeneği de yukarıda bahsedilen değişkenler göz önüne alınarak hazırlandı ve birbirleri arasında kıyaslamalar yapıldı. Tezin kayda değer sonuçları bu kıyaslamaların sonucunda elde edildi. Ses mühendisleri ve normal dinleyicilerin tahminleri, deneklerin ses lokalizasyonu üzerindeki tecrübeleri göz önüne alınarak iki ayrı grup olarak değerlendirildi. İki farklı frekans karakteristiğine sahip ses sinyali kullanılarak (2000Hz sinüs ve beyaz gürültü), lokalizasyon kesinliği farklı ses uyaranlarının tahmin sonuçları kıyaslandı. Ya da görsel ve işitsel arasındaki ilişki değiştirilerek, birlikteliğin inandırıcılığına müdahale edildi ve iki durumun sonuçları kıyaslandı.

Öncelikle sadece ses testindeki sonuçlar verilecek olursa; ses mühendisi ve normal dinleyicilerin performansları arasında kayda değer bir farklılık gözlemlenemedi. Normal dinleyiciler seslerin konumunu 2º hata ve 3,2º standart sapma ile belirlerken, ses mühendisleri 1,9º hata ve 3,4º standart sapma ile belirledi Sinüs sinyali tahminlerinin gürültü sinyaline göre biraz daha büyük bir açı farklılığıyla saptandığı gözlemlendi. Sinüs sinyalinin yeri ortalama 4,1º hata ve 2,3º standart sapma ile belirlenebilirken, gürültü sinyali 3,2º hata ve 1,8º standart sapma ile belirlendi. Deneyin ilginç sonuçlarından biri de sinüs sinyali tahminlerinin gerçek ses pozisyonu dikkate alındığında merkeze doğru kayma eğilimi göstermesine karşılık, gürültü sinyali tahminlerinin merkezden dışarı doğru eğilim göstermesiydi.

Ses ve görselin birlikte sunulduğu testte göze çarpan ilk sonuç, görselin merkezden (deneklerin bulunduğu konumun tam karşısından) gönderildiği zaman, görsel önyargının en üst seviyeye çıkmasıydı. Bütün görsel pozisyonların ortalama etkisi dikkate alındığında, ses mühendislerinin ses lokalizasyonu tahminleri %15,2 oranında görsele doğru kayarken, normal dinleyicilerin tahminleri %13,7 oranında kaydığı; sinüs için yapılan tahminler %18,1 kayarken, gürültü sinyali için yapılan tahminlerin %10,8 oranında görsele doğru kaydığı tespit edildi. Ancak sadece görselin merkez pozisyondan gönderildiği durumlar dikkate alındığı zaman ise; ses mühendislerinin ses lokalizasyon tahminleri %69,7, normal dinleyicilerin %46; sinüs sinyalinin lokalizasyon tahminleri %62,9, gürültü sinyalinin %52,8 oranında görsel uyarana doğru kaydığı gözlemlendi.

Ayrıca ses ve görsel testinde görsel-işitsel birlikteliğine hareket ilişkisi eklenerek başka bir deney değişkeni oluşturuldu. Kısaca anlatmak gerekirse, görsel uyaran bir anda görünüp yok olmak yerine, yavaşça belirip yavaşça yok olma karakteristiğine sahip kılındı. Bununla birlikte ses sinyali de görselin bu hareketine uygun olarak *attack* ve *release* zamanları verildi. Sonuç olarak birlikte hareket etme eğiliminde olan görsel ve işitsel uyaranlar, hareket ilişkisi barındırmayan uyaranlara göre deneklerin algısını daha fazla etkiledi ve tahminler görselin bulunduğu konuma doğru daha fazla kaydı. Bütün görüntü pozisyonları ve bütün tahminler dikkate alındığında, durağan özelliklere sahip uyaranların yarattığı görsel önyargı oranı %10,8 iken, hareket ilişkisine sahip uyaranların yarattığı önyargı %28,7 olarak ölçüldü (bu kıyaslama sadece gürültü sinyali tahminleri üzerinden yapıldı). Ayrıca görsel merkezden gönderildiği zaman ses mühendislerinin gürültü sinyali tahminleri bir önceki testte %59,5'lik görsele doğru kayma yüzdesine sahipken bu testte %86,5'e, normal dinleyicilerin tahminleri ise %46,5'den %79,6'ya kadar yükseldiği gözlemlendi.

1. INTRODUCTION

The location of any object that has enough specifications such as being touchable, audible or visible, can be perceived by using visual, haptic and auditory sensory modalities. Two or more of these modalities can interpret the stimuli together to create location information, and this is called cross-modal perception. However, vision dominates the perception when these modalities are in conflict. This is known as "visual capture" (Grunwald, 2008). When vision dominates the audition, Howard and Templeton specifically called this phenomenon "ventriloquism effect" (as cited in Warren et al., 1981).

1.1 Ventriloquism Effect

As Connor mentioned that in his book the name is coming from an ancient art, and ventriloquism based on convincing people that the demonstrator's voice is coming from somewhere else instead of its mouth. Roman and Greek seers took advantages of it for seeming more mysterious (as cited in Alasis and Burr, 2004). People who did this trick denominated as a ventriloquist. For instance, people who make the talking puppet show without moving their lips are one of the today's examples of ventriloquists. In daily life, another and most common example of this effect is our televisions. Although the location of the loudspeakers is in a different place, we are tending to perceive that the voice comes from lips of people. Generally speaking, visual information dominates the auditory information. Indeed, this domination depends on visual and auditory cues reliability and academicians have been researching ventriloquism effect and its visual-auditory cues almost since half of the previous century.

1.2 Literature Review

Researchers, mostly neuroscientist and sound engineers, have been designing experimental setups for measuring what the angle threshold between audio and visual stimuli that can be detected by the test subject in various ways is. In another word,

which angle difference between audio and visual stimuli cause that people can understand the sound actually is not coming from its original source. The results show that there is no constant degree threshold between them. These differences depend on researchers' methods and what they searched for. For instance, in Jackson' (1953) experiment, two different experimental setups were prepared. Their differences depended on how the audio and visual stimuli interact with each other. For the first experiment, when the disparity between sound and visual was 20º-30º, 43% of the subjects noticed that sound is still coming from the visual. However, for the second experiment result was different for different interaction. When the disparity between audio and visual was 20º-30º, this time %97 of the subjects reported that sound is coming from the visual and basically it was related with subjects' perception that how audio-visual event was perceived (details in title 2.1.3, par.2). When the audio and visual behaves like single-event (depending on their interaction), the subject can hardly define the audio-visual disparity (Hendrickx 2015).

Warren et al. (1981) used single-event, dual-event conditions in their experiment. They created the condition by orienting the subjects instead of using different interactions between stimuli. Their results showed that ventriloquism effect, in another word larger visual bias was observed 3,5 times much more in dual-event situation (title 2.1.4, par. 2).

Also in experiments, specifically focusing on disparities between audio and visual changes the subjects' perceptions. Komiyama's (1989) experiment that was depended on acoustic engineers' vs. naïve listeners' reaction on audio-visual disparity for HDTVs is one of the particular examples of changing subject perception via changing their attention (title 2.1.4, par.3). Another important point of this experiment was comparing experienced and non-experienced subjects. Their results show that acoustic engineers had more attention on audio-visual discrepancy (title 2.1.1, par.1).

Beyond that academicians have been researching ventriloquism effect in new types media too. Kytö et al. (2015) published their article based on augmented reality and ventriloquism effect. In augmented reality world, the results were different from real-world experiments. It was reported that visual bias in augmented reality approximately 5º-10º much higher than the real world experiments. Furthermore, in augmented reality, it needed 30º discrepancy between audio and visual sources to be able to understand they are two different sources. Moreover, the other recent study in a new

type of media which is André et al.'s (2014) research based on stereoscopic-3D videos that commonly used technology in today's world.

Some other studies investigated that how subject's perception on sound localization is affected when the sound localization method is changed in ventriloquism effect experiment. For instance, Török et al. (2015) compared visual biases, when sound localization is obtained by stereo speaker setup or single speaker setup that is depending on its spatial location (title 2.1.5, par.3). As additional information, in that study, they noticed that when visual stimuli were located on the vertical axis, amount of visual bias was much more than the horizontally located one. Montagne and Zhou (2016), used different approaches for changing the localization method. They used just stereo setup then the location of the sound was determined by time, and intensity differences between these two speakers and visual bias comparison between these two methods were presented in their research (title 2.1.5, par.4).

A part of Werner et al.'s (2013) experiment was depending on how different types of sound stimuli affects the result of visual biasing. For making this comparison, they used noise burst and 6-second saxophone samples. They noticed that Saxophone was chosen because of its sound characteristic is similar to human voice. The results showed when two stimuli were compared, their differences in ventriloquism effect were sufficiently small (title 2.1.5, par.5).

Not only sound cue reliability has influences on ventriloquism effect but also visual cure reliability too. Even when visual stimuli have weak reliability, sound can affect the position of visual. Alais and Burr (2004), reported that if the weakly localizable visual stimuli are used in the experiment, audio dominates visual. In their experiment, blobs were used as visual stimuli and click as a sound. The test system produced consecutively two audio-visual events. In the events, audio-visual stimuli were in synchrony at different locations, and it was asked is the blob on the left or right respected to first blop location. When the blobs were bigger, identifying the center point location of the blop got harder, so subject got help from the audio to identify the visual location.

Additionally, some studies concentrated on vertical planes rather than horizontal plane. For instance, Werner et al. (2013), reported that in elevated position has larger degree bias was observed than horizontally designed experimental setup. It was

because of sound localization performance is being better in azimuth plane (Makous and Middlebrooks 1990). Furthermore, Hendrixkx et al. (2015) had similar results as Werner et al.'s observations.

Finally, in specific conditions, visual dominates the auditory location yet perception resolution is different for two modalities. Cavonius and Robbins mentioned that people could determine the visual position spatially in a 1/60 of degree arc (as cited in Hendrickx et al., 2015). Furthermore, as many studies reported that sound could be spatially localized relatively or absolutely in a maximum 1º arc (Recanzone et al. 1998; Perrott and Saberi, K. 1990; Yost 2016).

1.3 Purpose of Thesis

The thesis has three purposes. First and the main one is about determining the sound localization performances of subjects with and without visual biasing in azimuth. The second purpose is to compare the sound engineers and naive listeners' experiment results, and the third one is observing the effects of using a different kind of stimulus on ventriloquism effect. Also, past studies have this kind of purposes. However, all these studies have differences when they are specialized. So the small differences between recent thesis and the past studies were explained in the next three paragraphs.

As it is mentioned before visual bias on sound localization was observed in many studies. Even more studies can be given as an example, like Bertelson and Aschersleben's (1998), Wallace et al.'s (2004) Battaglia et al.'s (2003) studies. In that thesis, the relationship between only sound localization performance and sound localization performance with visual cues will be investigated separately, and the result will be evaluated respectively, instead of making exact bias degree inferences.

Similar to Komiyama's (1989) experiment, sound engineers and naive listeners will be compared. Nonetheless, it will be differed in another point, since Komiyama's experiment was depending on psychological implications of audio-visual discrepancy, this study will be based on audio localization tasks.

Past studies proved that audio or visual stimuli qualities have huge effects on ventriloquism effect. However, trying to establish new relationships between audio and visual could make the recent study slightly different. By that reason, two types of visual and three types of audio stimuli will be used in experiments, and these are static

black circle or animated black circle and sine wave, broadband noise burst or broadband noise signal with envelope characteristic[1].

1.4 Methodology

The methodology was explained in two subtitles which are "experiment" and "calculations," and their details will be given in chapter 2. Such as, audio or visual stimuli characteristics, angles, the position of the curtain and subject, etc.

1.4.1 Experiment

Experiments will be applied in 2 major and 1 minor section and the major ones are audio-only task and audio-visual[2] task, the minor one is audio-visual task with animated visual and enveloped sound. All these sections will be applied both sound engineers and naive listeners.

Firstly, in the audio-only task subjects will encounter with projected vertical grids on an acoustically transparent visually opaque curtain. Behind the curtain, loudspeakers will be located, and their center position will be behind the center of the grids. Noise or sine wave will be used as audio stimuli.

To measure the spatial precision of the audio localization, firstly less number of grids will be presented to subjects then the number of the grids will be increased (so the angle between them decreases). The process will continue until when the results[3] show that mean estimation error threshold reached more than %50. When it reaches, grid number will have been determined for either noise and sine wave, with more than %50 mean estimation error threshold. Then audio localization estimations will have been collected for two different types of audio signals and two type of subjects.

Secondly, in the audio-visual task 1 grid section will be applied which is specified in the audio-only task that has more than %50 estimations error. In that stage, noise or sine wave and visual stimuli will be given and these will be in synchrony at a different

[1] It has different attack, decay and release time than noise burst and it will be explained in detail in chapter 2.

[2] Audio-only and audio-visual terms are used as Montagne and Zhou used as in their article (2016).

[3] Since the localization of the sine wave is blurrier than the noise signal, the noise signal will be considered as reference (Yost, 2016).

location. The aim will be making a comparison between previously determined audio-only localization estimations and newly determined audio localizations with visual stimuli.

Thirdly the audio-visual task with animated visual and enveloped sound procedure will be same like audio-visual task, but the stimulus characteristic will be different and will be based on motion percept.

Finally, results will have been collected which have enough data for making comparison between;

- Localization precision of sine and noise signal,
- Localization estimations of sound engineers and naive listeners,
- Sound localization estimations before visual stimuli and after it,
- Sound localization estimations before animated visual-enveloped sound and after it.

1.4.2 Calculations

Firstly, as it is mentioned before, number of grids will be determined by calculating the mean estimation error threshold in audio-only task. In every grid sections, several amounts of localization estimation question will be given. Whether wrong estimations are close to the real sound source or not, it will be evaluated as an error. When that grid section was finished for all subject, it will be analyzed whether it is reaching enough threshold or not. If it is not, the subjects will be called again, and the new section will begin with more number of grids. For the calculation in that specific grid section, all subject's number of errors will be summed, and it will be found out the percentage of it in total questions amount. When the equation 1.1 will be achieved, audio-only task will have been finished (reference is noise stimuli). At that point also sine stimuli localization errors will be calculated too. (t=number of total questions, n=number of total subjects).

$$\frac{Subject_1\ number\ of\ errors + \cdots + Subject_n\ number\ of\ errors}{t} \times 100 \geq \%50 \quad \textbf{(1.1)}$$

Furthermore, for two types signals, divergence angles will be calculated in percentage value, depending on which grid is chosen by subjects and what is the angle difference between it and the real position of the sound. Also, it is depending on what is the total

projected area that depending on subject position and projected area length. So equation 1.2 explains the % mean divergence for the first subject. (Real position⁰ will be accepted as reference point, and when if estimated grid is on left side of the real position it will take negative value otherwise positive value).

$$\%Subj_1\ mean\ div.$$

$$= \frac{(|Estimated\ grid_1{}^0| - Real\ sound\ Position_1{}^0) + \cdots (|Estimated\ grid_t{}^0| - Real\ sound\ Position_t{}^0)}{Total\ Projected\ Area^0 \times t} * 100 \quad \textbf{(1.2)}$$

Equation 1.3 is for calculating % total mean divergence.

$$\%\ Total\ mean\ div.$$

$$= \frac{\%Subj_1\ mean\ div. + \cdots + \%Subj_n\ mean\ div.}{n} \quad \textbf{(1.3)}$$

Secondly, in the audio-visual task, the experiment will be achieved on a previously specified number of grids. Visual stimuli will be given in synchrony with audio stimuli however its position will be different. To succeed the comparison between two task, the location of the audio stimulus will be exactly same as audio-only task location. But order of them will have variability. Error threshold and %total mean divergence angles will be calculated in new task again with a same equation that was given above 1.1, 1.2, 1.3.

Besides these, as Alasis and Burr (2014) noticed in his article, when if audio stimuli have low spatial location precision, then in audio-visual task their perceptual spatial location is affected by visual stimuli more or if stimuli have high spatial location precision, their perceptual spatial location is affected by visual stimuli less. So to be able to figure out the relationship between them equation 1.5 will be used. (Real sound position will be reference). Also, equation 1.4 which will be used for calculating the visual biases. This equation has the same approach like Warren et al. (1981) used in their experiment.

$$\%\ div.\ in\ audio\ only\ task$$

$$= \frac{|Estimated\ grid^0| - Real\ sound\ Position^0}{Total\ Projected\ Area^0}$$

$$\%Amount\ of\ visual\ bias\ on\ perceptual\ sound\ position$$

$$= \frac{|Estimated\ grid^0| - Real\ sound\ position^0}{|Visual\ grid^0| - Real\ sound\ Position^0} * 100 \textbf{ (1.4)}$$

$$Relative\ amount\ of\ ventrloquism\ effect\ on\ perceptual\ sound\ position\ depends\ on\ its\ spatial\ location\ precision$$

$$= \frac{\%Amount\ of\ ventriloquism\ effect\ on\ perceptual\ sound\ position}{\%\ div.in\ audio\ only\ task} \textbf{ (1.5)}$$

By using this equation, how visual relatively affects sine and noise perceptual spatial location depending on their spatial location precision will be tried to figure out. For the comparison, mean of every subject and every question will be calculated.

Finally, to making error threshold comparison between audio-visual task and enveloped audio-animated visual task, the equation (1.1) will be used separately for each task.

1.5 Hypotheses

Experiment have several different variables and consequently, there are several predictions which are based on expectations of estimation errors and these are:

- Higher estimation errors in audio-visual task than audio-only task,
- Higher estimation errors in enveloped audio-animated visual task than audio-visual task,
- Higher estimation errors for naive listeners than sound engineers,
- Higher estimation errors for sine wave than noise.
- Moreover, the last one is based on of establishing a proportional relationship between sine wave and noise burst by using equation 1.4, and the prediction is, mean results of sine wave and noise burst should be equal[4].

[4] Mean results of equation (1.4) for sine wave = Mean results of equation (1.4) for noise burst.

2. OBSERVING THE LIMITS OF VISUAL BIAS ON SOUND SOURCE LOCALIZATION IN VENTRILOQUISM EFFECT: A CASE STUDY ON SOUND ENGINEERS VS. NAIVE LISTENERS

Previous studies show that visual biasing of perceptual sound localization has affected by how the experiment was designed. In this chapter firstly, it will be analyzed that which factors influenced results[5] of the previous experiment in azimuth and respectively what are the differences or similarities between them and present experiment. Secondly, all the details that the present experiment has will be explained.

2.1 Which Factors Influence the Results of the Experiment

Influencing factors were similarly grouped as Hendrickx et al. (2015) used in their research. Additionally, one extra group title was added, one title name was changed, and one of the titles was divided into two subcategories, to approach to influencing factors more specifically. The differences are presented in table 2.1.

Table 2.1: Grouping the influences factors on visual bias in azimuth.

Hendrickx et al. (2015)	Recent Study
1. The experience of subjects	1. The experience of subjects
2. Temporal disparity	2. Temporal and spatial disparity
3. Compellingness	3. Compellingness
-	3.1. Concrete or abstract (conceptual relation
-	3.2. Static or dynamic (motion relation)
4. Attention	4. Attention
-	5. Audio-only localization performance

[5] It was explained in literature review part shortly. In that chapter, details will be given.

2.1.1 The experience of subjects

It is expected that sound engineers or people who are dealing with sound localization have better spatial sound localization than the others. Additionally, subjects those are trained for the specific experimental task had better estimations too. Results of Komiyama's (1989) experiment on HDTV, showed that experienced people (acoustic engineers) had more attention on audio-visual discrepancy. In this physiological experiment, acoustic engineers annoyed when the discrepancy was about 11º, and naive listeners annoyed when it was 20º. An example of trained subjects is Andeol et al.'s (2015) experiment. They noticed that in one of the sound localization task, subjects after trained showed better results.

In the present experiment, none of the subjects were trained. Their differences are based on sound localization experiences. The experienced group that is sound engineers have experiences on phantom sound source localization which is created by two speakers by using ITD and ILD. On the other hand, in the experiment, single speaker determines the sound location that is provided by horizontally set speaker array.

2.1.2 Spatial and temporal disparity

Wallace et al. (2004) reported that temporal and spatial disparity between sound and visual stimuli affects the percentage of localization bias. When the stimulus was presented in 5º spatial disparity and 200ms temporal disparity, the percentage of localization bias was %95. However, when the temporal disparity increased to 800ms, this percentage dropped to %88. Also keeping the temporal disparity constant at 200ms and changing the spatial disparity 5º to 15º affected the visual bias percentage from %95 to %90.

In the present experiment, the spatial disparity is constant in every audio-visual question, and there is no temporal disparity, the stimulus is presented in synchrony.

2.1.3 Compelingness

Warren et al. (1981), analyzed that, whether audio-visual stimulus was in harmony with one another or not and the results of visual bias effect were reported depending on this. For instance, in their experiment, three specified tasks were used, and these

were high-medium-low compelingness tests. In the high compelingness test, person's lips movement on TV screen were used as visual stimuli, and loudspeakers of the TV was deactivated. In a separate loudspeaker, the speech of that person was used as sound stimuli, and it was located a different location to create a spatial disparity. For the medium compelingness test person voice was delayed 150ms (temporal disparity), and in the low-compelingness test the screen was illuminated, but the picture of the person was not there, and the small tape was replaced in its mouth position. The results showed that percentage of visual bias on audition was %78 for high compelingness, %40 for medium compelingness and %32 for low compelingness tasks.

Jackson's (1953) experiment also related with compelingness. His experiment based on using two different kinds of audio-visual relationship. Consequently, he got two different results from two different tasks. In first task (which can be thought as a low compellingness task), he used bell sound, and light stimuli and these were rung and shone in the same time, and the locations of these could be different or the same. For the second task (this can be thought as high compellingness task), he used a puff of steam from kettle whistle as a visual source, and the whistle sound as a sound source and like the previous experiment the location of the sound source was variable. For the bell and light experiment, when the disparity between sound and visual was 20º-30º, 43% of the subjects noticed that sound is still coming from the visual. However, for the second experiment result was different. When the disparity between kettle and whistling was 20º-30º, this time %97 of the subjects reported that sound is coming from the visual.

Vatakis and Spence (2007), used a different kind of method to analyze whether audio-visual cues have compelling evidence or not in the perception of the subjects. Their aim was to find whether subjects can discriminate the temporal disparity between audio and visual or not. To do this, two tasks were designed. In the first task, the talking male or female videos were used as visual stimuli, and their own voices were used as sound stimuli (match or high compelingness task). In the second task, the voices were changed, female voice was used with male video or vice versa (mismatched or low compelingness task), and stimulus were presented with the temporal disparity in all tasks. The results showed that, when the mismatched task was presented, subjects could have easier discrimination and understand whether audio or visual had come first. Actually, this experiment is not directly related to ventriloquism effect. However,

to able to determine that single event or dual event is occurring or cues have strong compelling evidence are influencing factors of both situations.

In the direction of these examples, “compelingness” title was divided into two subgroups and these subgroups contains examples those depending on whether audio and visual stimulus have “concrete or abstract” relationship, or they have “static or dynamic” relationships. The purpose of doing this, to understand the influencing factors specifically and to explain the “enveloped audio-animated visual task” of the present thesis clearly.

2.1.3.1 Concrete or abstract relation (conceptual relation)

Concrete relationships have stronger evidence that audio and visual stimulus can behave like single event and subjects can follow the event without discrimination. In the abstract relations side, weaker evidence may cause the audio-visual relation to be perceived as dual events. While Warren et al.’s (1981) first task (high compelingness task) is the example of a concrete relationship, their third task (low compelingness task) is the example of an abstract relationship. Because in the first task speaker’s mouth and the sound has a usual relationship like real life events and consequently subjects firstly followed the event before discriminating. On the contrary, in the third task, instead of using real visual on the screen, just tape was used as a representation of its mouth, and it created the abstract relationship between them because the relationship was not like real life events and subjects were forced to make own relation between them. Therefore, following the events got harder and strong evidence between them got weaker, consequently discriminating got easier for the subjects.

The other two examples also had same relations between their individual tasks. While Jackson’s (1953) first task that light and bell were used, have an abstract relation, the second task that kettle visual and related sound stimuli were used, have a concrete relation. In the experiment of Vatakis and Spence (2007), using female visual with female voice or male visual with male voice put the task in concrete relation, on the contrary, changing the gender relations put the task in abstract one.

2.1.3.2 Static or dynamic stimuli (motion relation)

Any audio or visual stimuli could have static or dynamic motion. For the sound, if its dynamics or frequency information is changing over time (like speech, music, etc.), it has been called dynamic sound stimuli, or if its dynamics or frequency information are not changing perceptually over time (like a noise burst), it has been called static audio stimuli. For the visual, if the information of visual (brightness, color, position, scale etc.) is changing over time (like video of the speaker's mouth on TV screen or any animated visual etc.), it has been called dynamic visual stimuli or if it has not any movement over time (like tape on TV screen or stable images, stable lights etc.), it has been called static visual stimuli. After these definitions, if visual stimuli have a dynamic motion and sound stimulus follows it, it is expected that visual bias of sound localization should be more than the other possibilities (static visual-static sound and static visual-dynamic sound or vice versa). Because, the subject gives more attention to the event when the visual stimuli have dynamic motion (Bur et al., 2007). As a result of this, subject possibly percept the audio-visual event as it were a single event.

In past studies, motion relation was not analyzed specifically. However, Warren et al.'s (1981) high compelingness task (speaker's visual and its voice in synchrony) and low compelingness task (tape and speaker's voice) can be compared in "motion relation" category. While the high compelingness task has dynamic visual (mouth) and dynamic audio (voice), the other one has static visual (tape) and dynamic audio (voice). However, at first sight, it is hard to understand what caused the visual bias differences between two task. Was it because of their conceptual relationship differences (concrete or abstract) or because of motion relationship differences (dynamic or static)?

In the present thesis, two task was designed for analyzing motion relation and its influences on visual bias effect. One of them is "Enveloped audio-animated visual" task which have dynamic visual (appearing and disappearing black circle in time) and dynamic audio (has attack, decay and release time depending on visual movement) and the other one is "audio-visual task" which have static visual (suddenly appearing and disappearing black circle) and static audio (almost no attack and release time for fitting to visual). In addition, using black circle and noise or sine waves make both task abstract, so which helps the making comparison between influences of static and dynamic stimulus on visual bias purely.

Figure 2.1 and figure 2.2 is visual representation of the things those were explained in "2.1.3 Compelingness" title and also explained in its two subtitles.

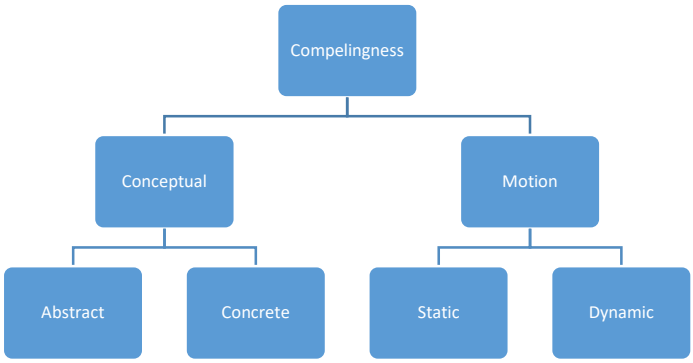


Figure 2.1: Hierarchic grouping of audio-visual relations in the compelingness main title.

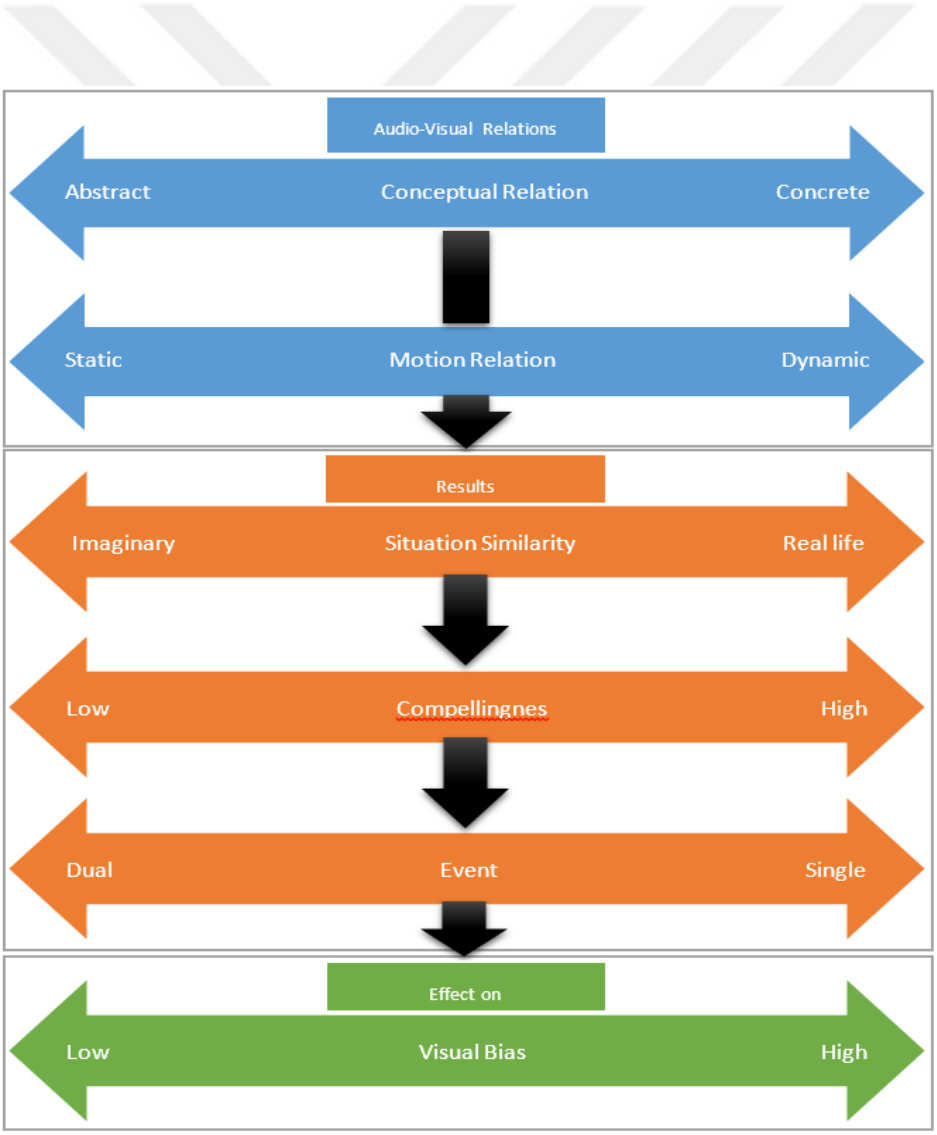


Figure 2.2: Possible results of audio-visual relations and their effects on the visual bias.

2.1.4 Attention

Past studies show that directing people by questions or instructions has affected the results of experiments. When it was asked the subject if there was any discordance between audio and visual stimuli, 50% threshold of audio-visual spatial discrimination got lower than the subjects were forced to find the specific location of audio (Hendrickx 2015).

Warren et al.'s (1981) single-event, dual-event conditions can be given an example for directing people by instructions. They created the conditions by orienting the subjects. For the single-event situation, they instructed the trials that there will be no real spatial discrepancy between speaker's voice and the talking person's video on the screen. They said to the subject, if any discrepancy will occur, it is because of using the optical device before the screen. Then subjects' location estimations were collected depending on that. For the dual-event situation, they explained that the audio and video could be in different places. Even tough same amount spatial disparity was used in both task, the results showed that larger bias was observed 3,5 times much more in the single-event situation.

Komiyama's (1989) experiment can be given as an example for how questions are affected the perception of the subject. His experiment was depended on the audio-visual discrepancy on HDTV, and the purpose was asking the subjects that are the discrepancy between audio and visual make them annoyed while they were watching HDTV. This experiment had two group of people one of them was naive listeners the other one was the acoustic engineers group. After the experiment, some of the acoustic engineers noticed that if discrepancy judgment were not asked, probably they would not have been disturbed.

In this experiment, unfortunately, subjects are oriented by questions and instructions as well. However, the orientations were tried to design to allow subjects to move away from a specific focal point. For instance, subjects are instructed that sound can come from any grid, also from the same grid as well. Specifically, for the audio-visual and enveloped audio-animated visual task, they are oriented that sound and visual position can be same or different.

2.1.5 Audio-only localization performance

Depending on characteristics of the audio and the environment, audio location precision could differ for a person. For example, when the localization performance is compared, the signal with richer frequency content (2000 Hz center frequency and 2-octave width) can be better localized than the signal with limited frequency content (2000 Hz center frequency and $1/10^{th}$-octave width). Also changing the center frequency and keeping the octave wideness constant also affects the localization performance too. The signal with 250 Hz center frequency and $1/10^{th}$-octave wideness can be localized better than the signal with 4000 Hz center frequency and same octave wideness. Additionally, the broadband noise signal can be better localized than all these band-limited signals (Yost, 2016). Localization performances are just not only depending on the frequency characteristics but also changing the plane orientation of loudspeakers array effects the results too (Perrott and Saberi, 1990). For instance, subjects had better performances in azimuth than the elevation (Morrongiello and Rocca, 1987; Su and Recanzone, 2001).

This kind of audio localization performance varieties affects the impact of visual biases on perceived audio location. Ernst and Bültholf (2004) mentioned that sensory cues (e.g., audio and visual stimulus) are interpreted depending on their reliabilities, and the blurry cue is dominated by reliable cue and result of this it affects the perception. Also, they noticed that determining the what kind of cues are reliable is not exactly known because there is no exact definition of the cue. However, after estimation process, the results can show which cue is reliable, relatively. Through this information, it is expected that if the sound stimuli have lower estimation precision, the visual cue affects the sound perceptual sound position more (Montagne and Zhou, 2016).

Török et al. (2015) observed that when the sound location obtained by stereo setup, audio localization blur much more than single speaker setup and respected to that visual bias was much more in a stereo setup. In detail, seven speakers were used in their experiment, and they were placed on 2,99 radius circle, and each of them had $10{,}5^{o}$ separations. First and the last speakers were used for stereo setup and the middle ones for the single speaker setup. Sound position in stereo speakers were matched with single speaker ones.

Montagne and Zhou (2016), used only stereo speaker setup and they located the sound via using ITD and ILD separately. When these two types of localization cues were compared in audio-visual discrepancy test, located sound with ITD had a weaker influence on ventriloquism effect. Also, they add that, it was because of time differences between two speakers was not enough cue to determine the location of the sound, and as a result of this, visual stimuli affected the perception of the subjects more.

Montagne and Zhou's (2016) and Török et al.'s (2015) experiments pointed out; sound localization cues influence ventriloquism effect. So it has been searched that when sound, with different frequency characteristics, is used, how these differences impact on ventriloquism effect. Thus, two candidate studies have been found. One of them which is Werner et al.'s (2013) experiment that noise burst and 6-second saxophone samples were used to compare their impacts on ventriloquism effect. However, although two different types of sound stimuli were used, saxophone have not stable spectral characteristic when we compared it with noise stimuli. Because the saxophone does not have constant sound characteristic for localization. It has different pitches (frequencies), and different attack-decay-release characteristic that is depending playing style, so making a comparison between two sound in ventriloquism effect experiment was not looking for at first. Beyond that, as it is mentioned in first paragraph Yost (2016) used different types of sound stimuli which are desired for comparison, but he used these in just audio localization task, not in ventriloquism effect experiment.

In the present experiment, 2000 Hz sine wave and the white noise signal are used. They were chosen depending on Yost's (2016) study, and it is showed that almost there is 10^{o} estimation error between them. So the aim is making a comparison between hardly-localized sound (2000 Hz sine, blurry cue) and easily-localized sound (white noise, reliable cue) and understanding how visual changes the perceived location of these signals respectively.

2.2 Experimental Setup

In this section, all the details of the present experiment will be given. Also, ideas that previously were thought about the experiment and why these ideas were changed,

will be discussed. Also, some images from experiment environment can be seen in appendix-a.

2.2.1 Premature thoughts and thoughts after analyzing the constraints

Using one speaker behind the curtain which can able to slide by the automatic control system and using real visual source (e.g., a real person) in front of the curtain was the first idea. So the person would wear a headphone and hear the metronome then move its mouth as it was studied before. The loudspeaker would give the voice of that person which would have been recorded previously. When the subjects follow the event, the loudspeaker will begin to slide in azimuth. The aim was determining specific angular differences between person mouth and the loudspeaker that subject can understand the sound is actually not coming from its original source (dual-event). Figure 2.3 shows top view of that experimental design

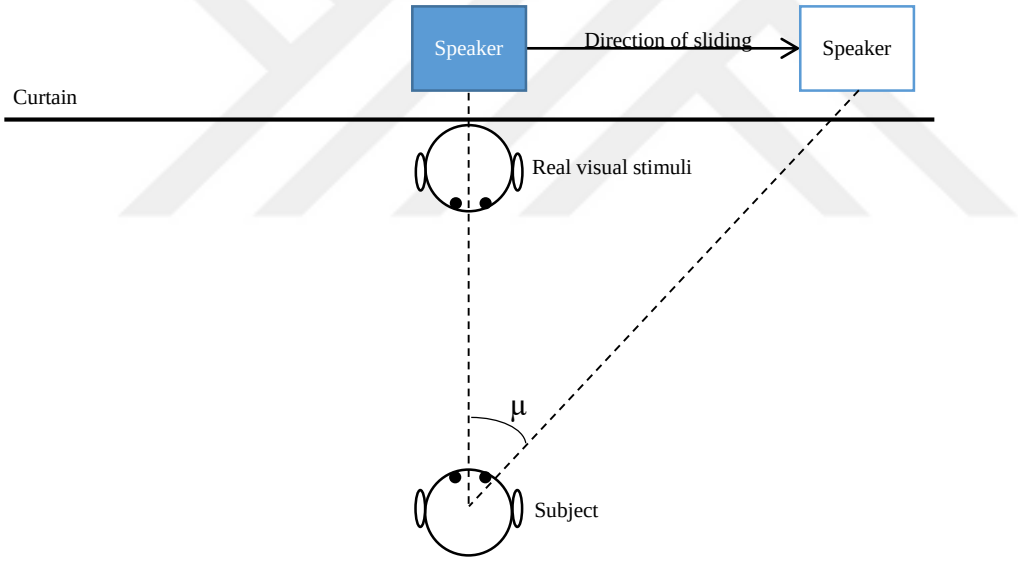


Figure 2.3: First experimental design idea (top view). When the subject can understand the sound source is not coming from the real sound source, the degree of α is obtained.

However, matching the sound source to visual source was almost impossible. Even if accepted that person can move its mouth with the sound source in synchrony, recording and the reproducing the sound causes frequency matching problem between possible sound expectation from visual source and speaker responses. Consequently, it was decided to use a grid system and projected visual stimuli. That is showed in Figure 2.4.

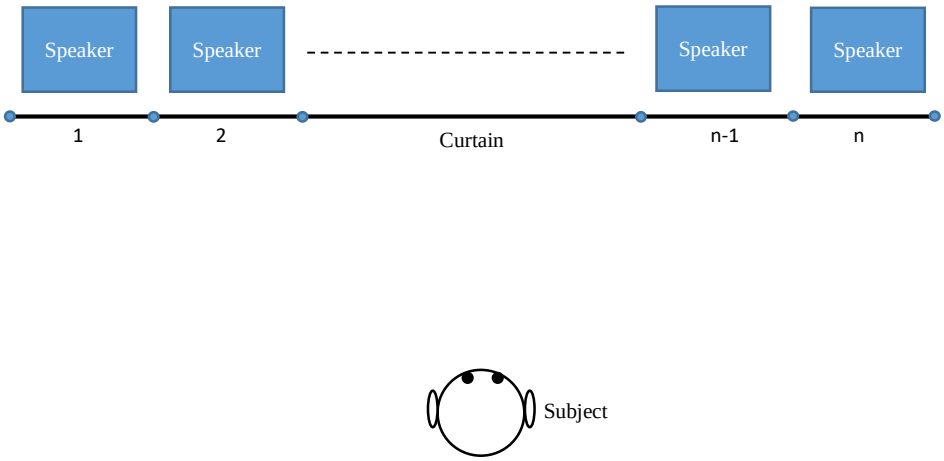


Figure 2.4: Grid design for the current study. Stable speaker positions (see appendix-a).

The other premature taught was to compare estimations errors between in two types of room those have different reverberant characteristics. The candidates were the semi-anechoic room of Automotive Technologies Research and Development Company (OTAM) in Istanbul Technical University (ITU) and studio control room of Center for Advanced Studies in Music (MIAM) in ITU. Then it was decided that this comparison will add more variables in to experiment and those are not useful for the thesis aim. Also, usage of semi-anechoic room was limited, and control room of MIAM was small for the experiment. So final decision was to use MIAM studio's live room. Figure 2.5 shows these facilities.

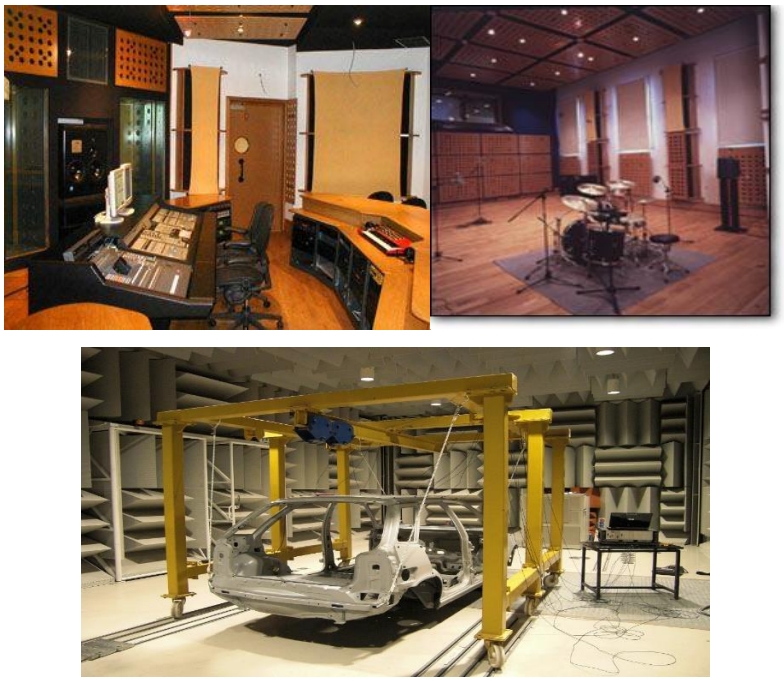


Figure 2.5: Upper pictures are from MIAM, and the bottom one is from OTAM.

2.2.2 Stimulus and grids

Two types of audio frequency characteristic and two types of audio envelop characteristic were obtained for sound stimuli, and two types of motion were obtained for the visual stimuli.

Audio stimulus: for succeeding two types of the frequency characteristic, white noise, and 2000 Hz sine wave were produced in Ableton Live (9.7) digital workstation. Specifically, "Operator" that is own synthesizer of Live was used. Also, they have 62,5 ms duration with 3 ms attack and 3 ms release time. 3 ms ramp time was given[6] due to avoid the click sounds. For succeeding two types of envelop characteristic, white noise signal was separated to "burst" and "enveloped sound." Duration of "burst" stimuli set 62,5 ms with 3 ms attack and 3 ms release time (same as frequency characteristics group). Duration of "enveloped sound" set 250 ms, with 125 ms attack, 62,5 ms decay and 62,5 ms release time. Figure 2.6, 2.7 and 2.8 show waveform of these sounds and their synthesizer settings.

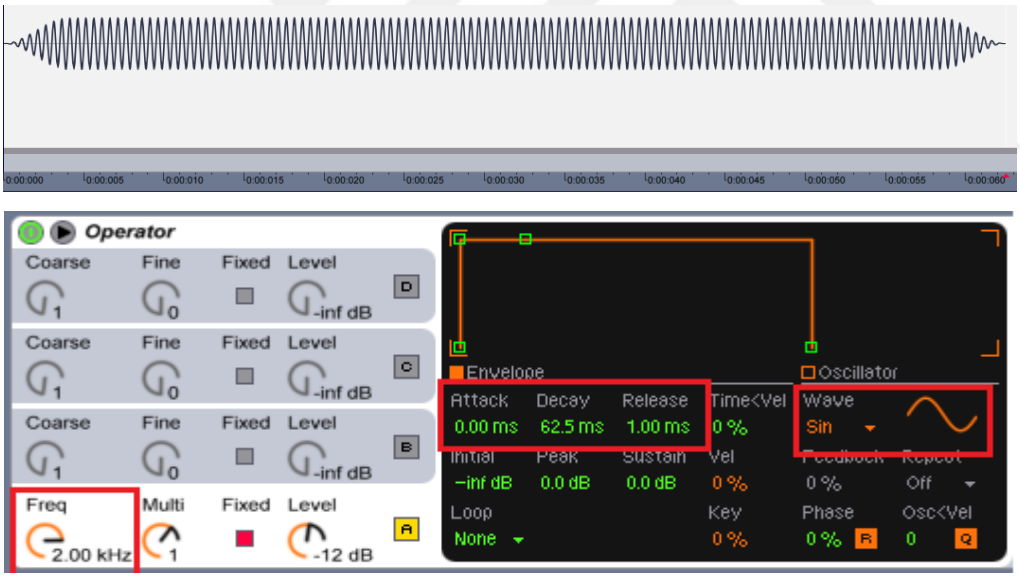


Figure 2.6: 2000 Hz sine wave burst waveform and "Operator" settings

[6] It was given by using fade-in and fade-out function of Live.

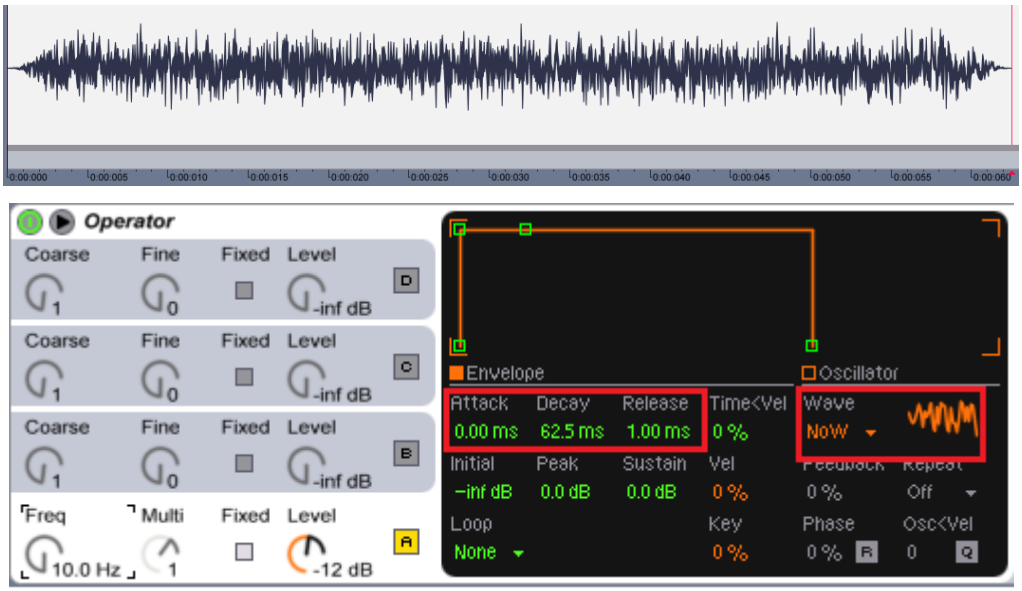


Figure 2.7: White noise signal burst waveform and "Operator" settings.

Figure 2.8: White noise signal enveloped waveform and "Operator" settings

In addition, it was explained in title 2.1.3.2 and 2.1.5 that why these sound stimuli were chosen.

Visual Stimulus: Cinema 4D R18 graphic design software were used for producing the visual stimulus and the grids. Two types of visual stimuli were designed, and these

were separated depending on their motion characteristic (static and dynamic)[7]. Static visual stimuli have approximately 67,5 ms duration, and it has no fade-in or fade-out time almost like audio stimuli. The aim is to try to create single event perception. Dynamic visual has almost 250 ms length, and its scale movements imitate envelope of the sound stimuli. This was tried to achieve by using "sound effector" tool of Cinema 4D. In both static and dynamic situations same black circle is used. In static one black circle radius is 7 cm, in the dynamic one it fades in and reaches almost 8 cm radius than fades out together with a release time of audio to 0 cm.

Grids: Four grid type were designed, and these are 5-grid, 9-grid, 13-grid and 17-grid setups[8]. In this part, grids will be explained in detail. Figure 2.9 represents the 5-grids setup from the top view. Hemispherical audio-visual areas have been used in past studies. In the present experiment straight projected area is used because of the limited possibilities. It causes angular differences between speakers when we consider the subject listening position. Also, it causes distance differences (dB differences for the subject) too. However, estimation error angles are calculated independently for each question, and also dB differences have not significant role in audio localization (Yost, 2016).

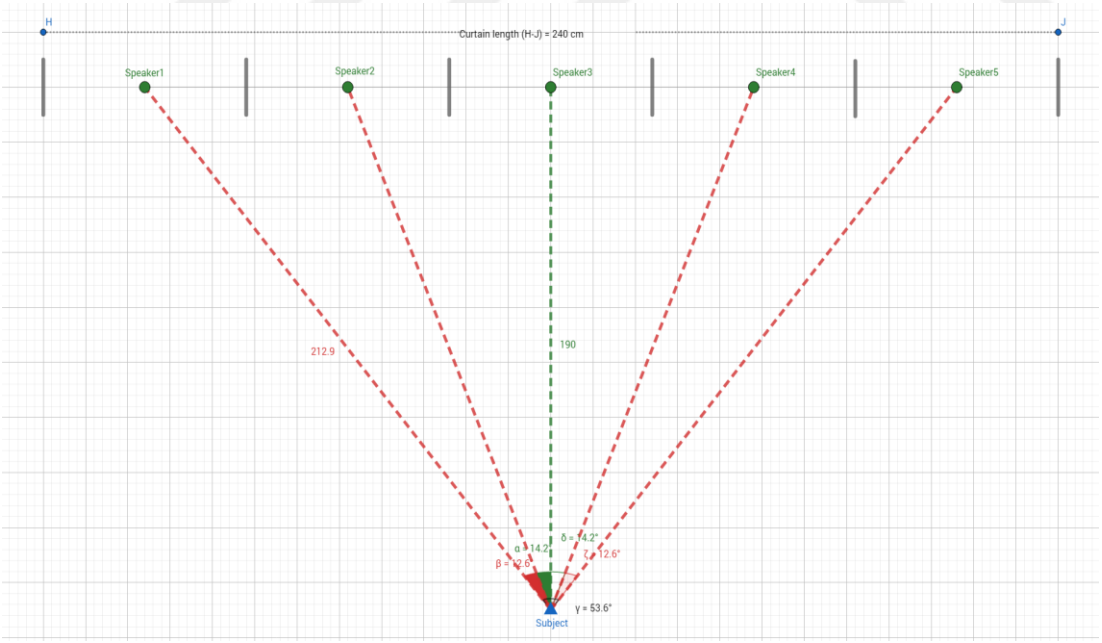


Figure 2.9: Top view of 5-grid setup. Curtain length = 240 cm, maximum angle difference = α-β = 14,2º-12,6º = 1,6º, maximum distance difference = length of

[7] These were explained in the third paragraph of title 2.1.3.2.

[8] The aim was explained under the third paragraph of title 1.4.1.

subject to speaker1 - length of subject to center speaker = 212,9-190 = 22,9 cm, total point of angular area = γ = 53,6º, mean angles between speakers = 13,4^{o}

Also figure 2.10 shows drawing of the projected 5-grid screen, and static visual stimuli and Figure 2.11 is the representation of grids, visual stimuli, and calculations together.

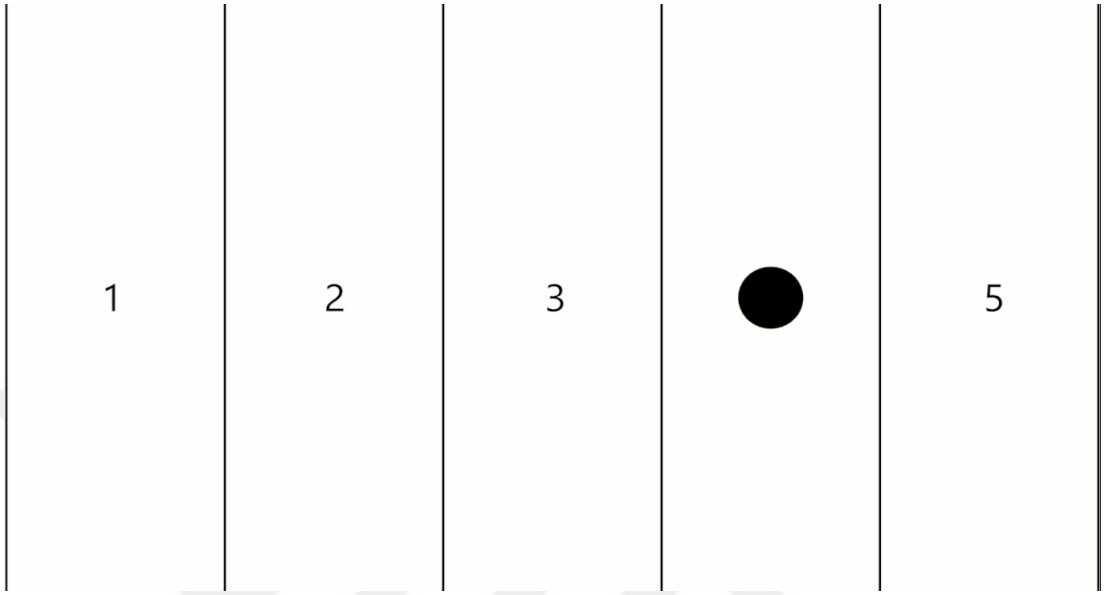


Figure 2.10: Projected 5-grid and static visual stimuli.

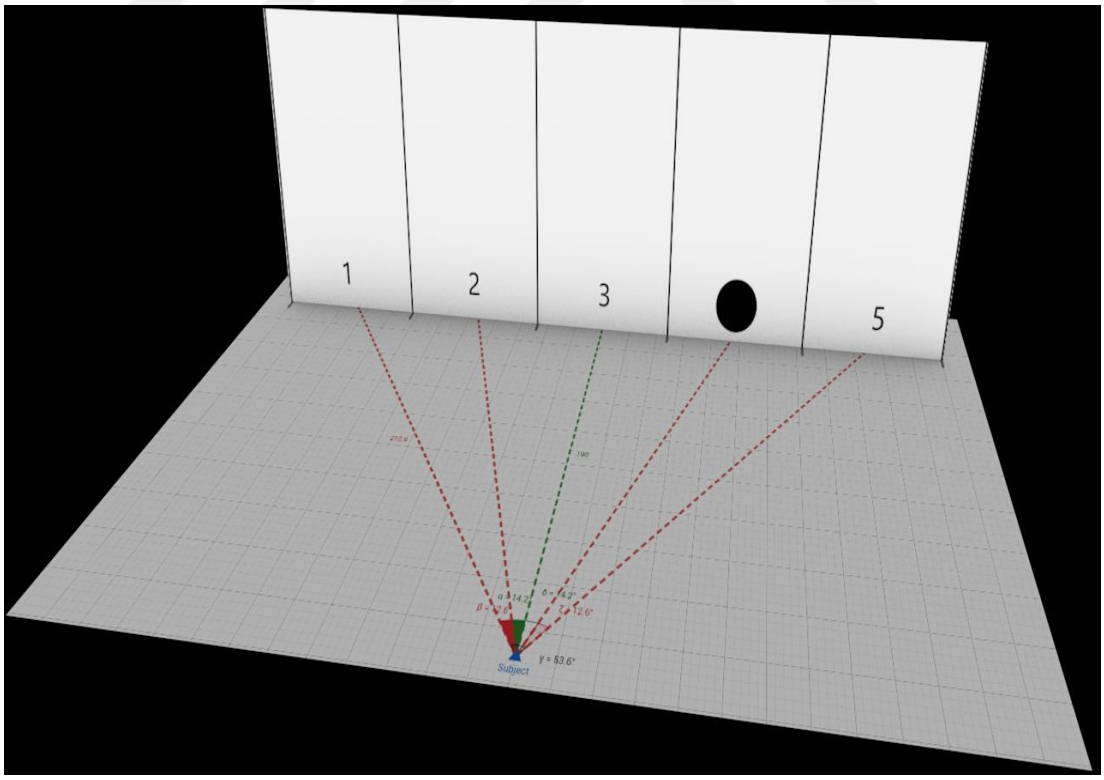


Figure 2.11: Grids, visual stimuli and calculations in 3d space.

All calculations for 5-grid, 9-grid, 13-grid and 17-grid in the setups table 2.2 are given below.

Table 2.2: Calculations for each grid setups

	5-grid	9-grid	13-grid	17-grid
Curtain length	240 cm	240 cm	240 cm	240 cm
Maximum angle difference	1,6º	1,5º	1,2º	1º
Maximum Length difference	22,9 cm	27,9 cm	29,9 cm	31 cm
Total point of view	53,6º	58,6º	60,5º	61,5º
Mean angles between speakers	13,4º	7,3º	5º	3,8º

Visual stimuli and sound stimuli disparity on grids: Depending on which grid setup is used in audio-visual task, grid differences between audio-visual is changed for keeping the similar angular disparity between them. For instance, in 5-grid setup, disparity is 1 grid (mean angular disparity = 13,4º), in 9-grid setup, disparity is 2 grid (mean angular disparity = 14,7º), in 13-grid setup disparity is 3 grid (mean angular disparity = 15,1º) and in 17-grid setup disparity is 4 grid (mean angular disparity = 15,4º). Figure 2.12 shows the specific one example for audio-visual position on 17 grid setup.

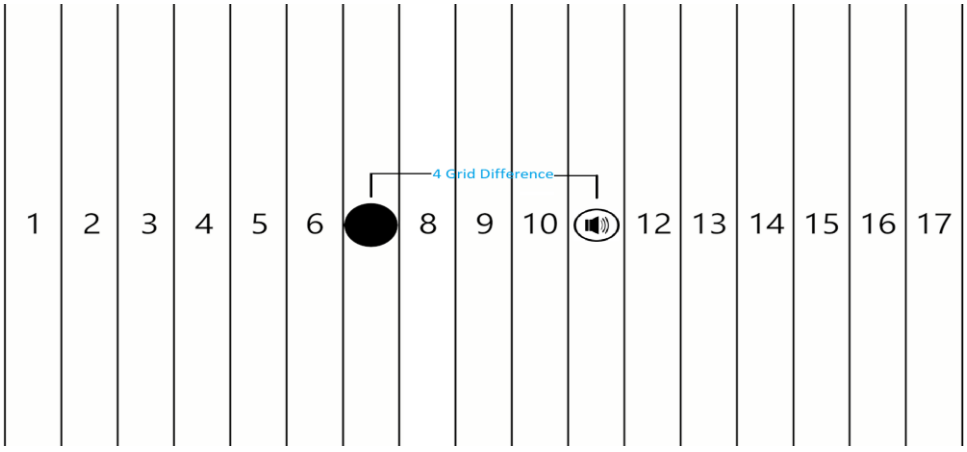


Figure 2.12: 17-grid setup, one specific example of the spatial disparity between audio and visual stimuli.

2.2.3 Subjects and questions

Five naive listeners and five sound engineers or sound engineering students participate in the experiment. All of them are performed all task individually. In every task, 12 estimation questions are asked that sound position is requested. In audio-only task, subjects are instructed that sound can come from any grid and in audio-visual tasks they are instructed that sound and visual can come with zero spatial disparity or come with any spatial disparity. Table 2.3 shows the tasks, subject groups and question amount.

Table 2.3: Question distribution for subjects and tasks

	Audio only task								Audio-visual task		Enveloped audio-animated visual task
	5-grid setup		9-grid setup		13-grid setup		17-grid setup		Obtained grid setup		Same grid setup in audio-visual task
Sound characteristic	Sine	Noise	Sine	Noise	Sine	Noise	Sine	Noise	Sine	Noise	Noise
One audio engineer or student question amount	6	6	6	6	6	6	6	6	6	6	6
One audio naïve listener question amount	6	6	6	6	6	6	6	6	6	6	6

Total question amount for all 10 subject = 660

2.2.4 Placement of loudspeakers

For each task, six pieces of Adam F5 loudspeakers are used (except in 5-grid audio-only task). They are located just behind the acoustically transparent visually opaque curtain. The height of tweeters is adjusted to ear level of subjects.

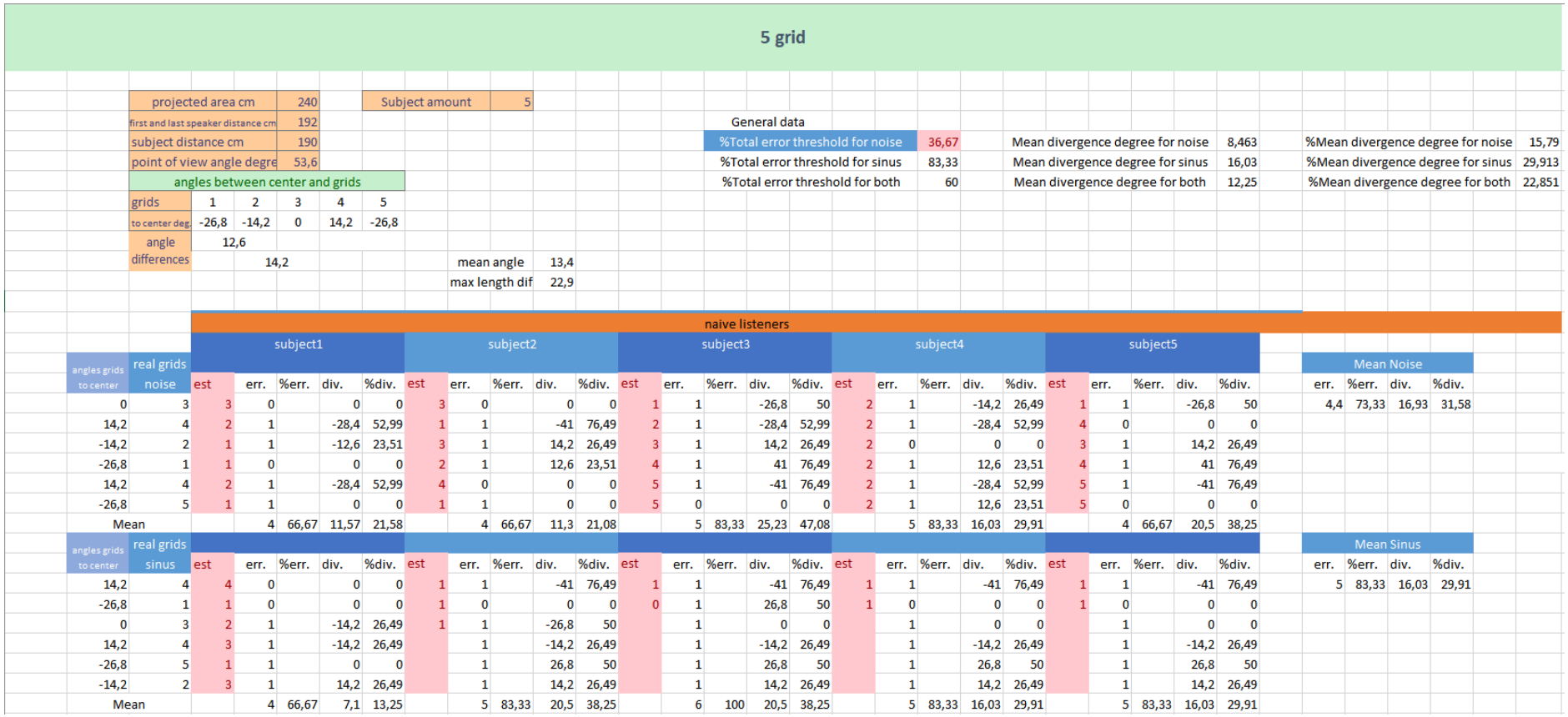
2.2.5 Applying the experiment and collecting data

For all task, subjects are called randomly. In the 5-grid audio-only task, for the one particular subject, six noise and six sine questions are asked (via using prepared

Ableton Live project), and estimations are collected at the same time (via using prepared Excel project). After it finished, the other subjects are called, and it lasts until when all subjects participate in the 5-grid audio only task. Next, the location of the loudspeakers is arranged for the 9-grid setup and subjects are called again. This process is resumed until 13-grid, and 17-grid setups is done.

After the audio-only task finishes, the results are evaluated and looked for in which grid setup, estimation errors rate is reached to %50 (noise questions are taken into account). Furthermore, to see the results easily and quick, the Excel project has formulas based on momentarily entered estimations. Figure 2.13 shows Excel project and 2.14 shows Live project. Figure 2.13 includes information about 5-grid setup and sound engineers group. All the necessary information and calculations can be seen in it. For instance, on the left, light blue colored cells show the actual grid position of the noise and sine stimulus, and on the right-hand side red cells (est.) shows the estimated grids which are blank in the experiment session and filled after getting estimations. When estimations are entered in it, all the data are calculated. For example, error amount (err.), %error (%err.), angular divergence (div.) and %angular divergence (%div.) are calculated for each subjects. On the right side of the figure, total results for sine and noise can be seen separately. Moreover, on the upper right side, under the general data tab, mean data for all subjects can be seen. Also in that tab, %total error threshold for noise cell shows the % estimation error threshold for 5-grid setup and it is useful to see %50 threshold momentarily.

Figure 2.14 is one part of Live project. It includes fix media noise and sine stimulus and grids video channels with stop tracks. In static movement channels, first number of each differently colored clips show the grid number that sound will be presented from that grid. Also as it is seen, real grids cells (light blue, on the left) in Excel table and these number are matched. When an audio signal is sent from Live, stop track is automatically stopped the playing, after getting the estimations, the estimated grid is entered into excel cells (est.). So in this way, Live and Excel project works together, and estimations are evaluated momentarily. Additionally Figure 2.15 shows the noise stimulus and its routing.



5 grid

projected area cm	240
first and last speaker distance cm	192
subject distance cm	190
point of view angle degre	53,6
Subject amount	5

angles between center and grids

grids	1	2	3	4	5
to center deg.	-26,8	-14,2	0	14,2	-26,8
angle differences		12,6			
			14,2		

mean angle	13,4
max length dif	22,9

General data

%Total error threshold for noise	36,67	Mean divergence degree for noise	8,463	%Mean divergence degree for noise	15,79
%Total error threshold for sinus	83,33	Mean divergence degree for sinus	16,03	%Mean divergence degree for sinus	29,913
%Total error threshold for both	60	Mean divergence degree for both	12,25	%Mean divergence degree for both	22,851

naive listeners

angles grids to center	real grids noise	subject1					subject2					subject3					subject4					subject5					Mean Noise			
		est	err.	%err.	div.	%div.	est	err.	%err.	div.	%div.	est	err.	%err.	div.	%div.	est	err.	%err.	div.	%div.	est	err.	%err.	div.	%div.	err.	%err.	div.	%div.
0	3	3	0		0	0	3	0		0	0	1	1		-26,8	50	2	1		-14,2	26,49	1	1		-26,8	50	4,4	73,33	16,93	31,58
14,2	4	2	1		-28,4	52,99	1	1		-41	76,49	2	1		-28,4	52,99	2	1		-28,4	52,99	4	0		0	0				
-14,2	2	1	1		-12,6	23,51	3	1		14,2	26,49	3	1		14,2	26,49	2	0		0	0	3	1		14,2	26,49				
-26,8	1	1	0		0	0	2	1		12,6	23,51	4	1		41	76,49	2	1		12,6	23,51	4	1		41	76,49				
14,2	4	2	1		-28,4	52,99	4	0		0	0	5	1		-41	76,49	2	1		-28,4	52,99	5	1		-41	76,49				
-26,8	5	1	1		0	0	1	1		0	0	5	0		0	0	2	1		12,6	23,51	5	0		0	0				
Mean			4	66,67	11,57	21,58		4	66,67	11,3	21,08		5	83,33	25,23	47,08		5	83,33	16,03	29,91		4	66,67	20,5	38,25				
angles grids to center	real grids sinus	est	err.	%err.	div.	%div.	est	err.	%err.	div.	%div.	est	err.	%err.	div.	%div.	est	err.	%err.	div.	%div.	est	err.	%err.	div.	%div.	err.	%err.	div.	%div.
14,2	4	4	0		0	0	1	1		-41	76,49	1	1		-41	76,49	1	1		-41	76,49	1	1		-41	76,49	5	83,33	16,03	29,91
-26,8	1	1	0		0	0	1	0		0	0	0	1		26,8	50	1	0		0	0	1	0		0	0				
0	3	2	1		-14,2	26,49	1	1		-26,8	50		1		0	0		1		0	0		1		0	0				
14,2	4	3	1		-14,2	26,49		1		-14,2	26,49		1		-14,2	26,49		1		-14,2	26,49		1		-14,2	26,49				
-26,8	5	1	1		0	0		1		26,8	50		1		26,8	50		1		26,8	50		1		26,8	50				
-14,2	2	3	1		14,2	26,49		1		14,2	26,49		1		14,2	26,49		1		14,2	26,49		1		14,2	26,49				
Mean			4	66,67	7,1	13,25		5	83,33	20,5	38,25		6	100	20,5	38,25		5	83,33	16,03	29,91		5	83,33	16,03	29,91				

Figure 2.13: A part of Excel table for collecting estimations and calculating data for 5-grid setup.

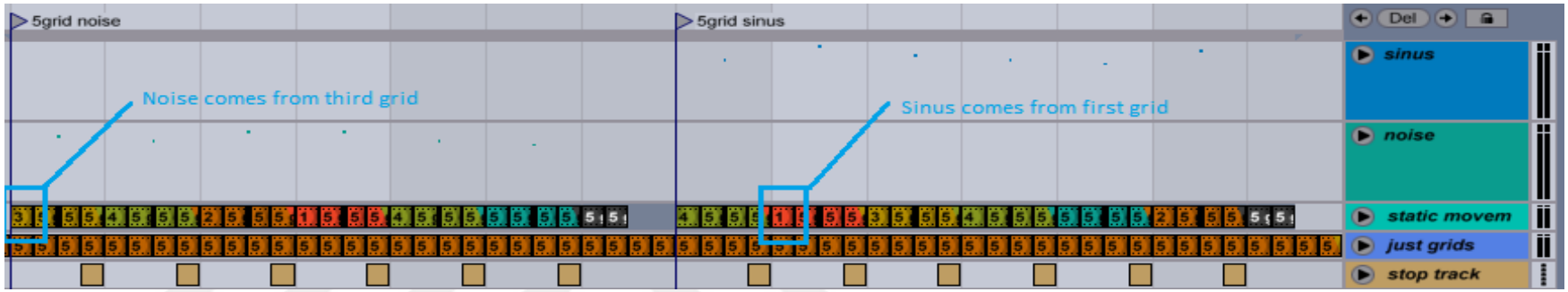


Figure 2.14: A part of Ableton project that includes noise and sine stimulus channels, video of the grids and stop track for 5-grid setup

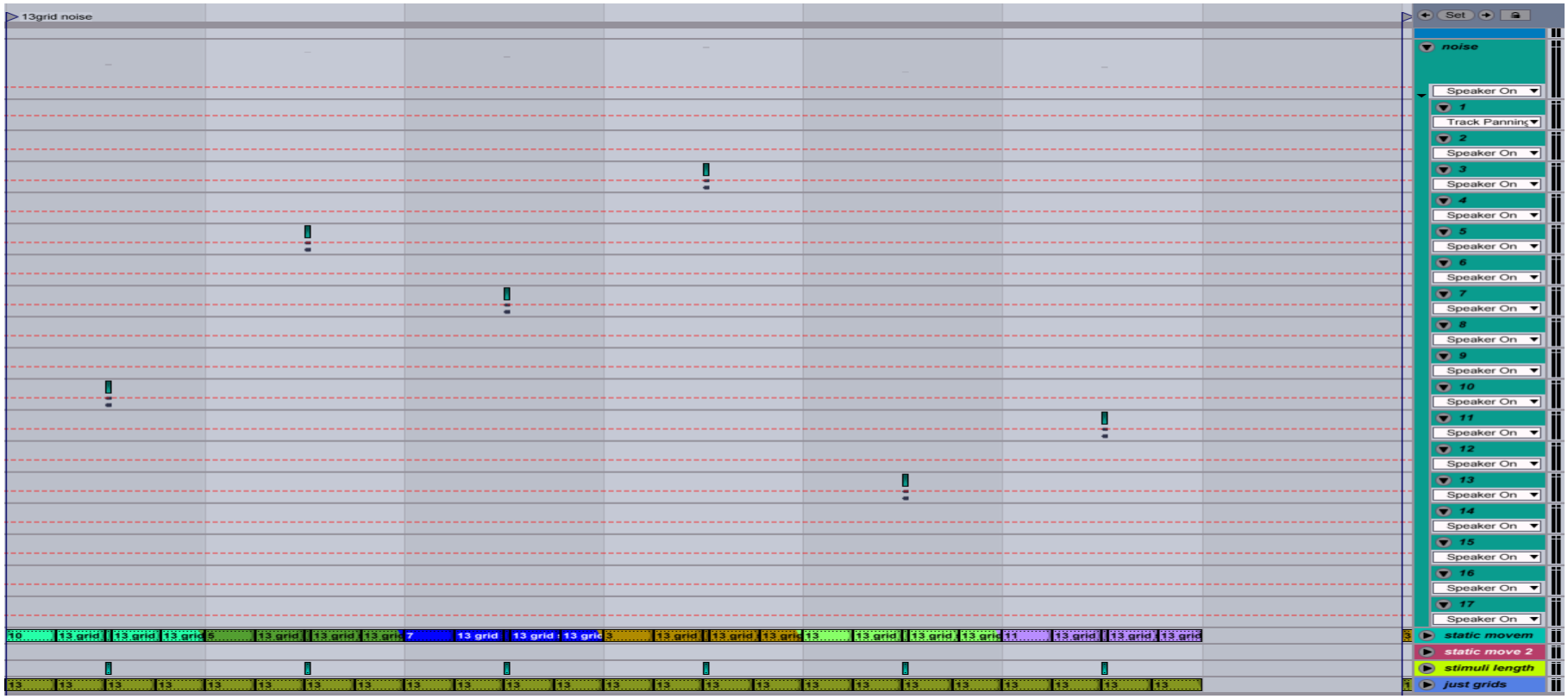


Figure 2.15: Routing for noise stimulus

3. RESULTS AND DISCUSSION

In Audio only task subjects reached the %50 error threshold for noise signal and also for sine wave in 13 grid setup (mean angular differences between grids was 5º, angular differences have variety in 4,3º to 5,5º due to the flat placement of speaker array). Audio-visual task and enveloped audio-animated visual task was applied to subjects in that grid setup. Figure 3.1 shows the %error amounts in all grid setup.

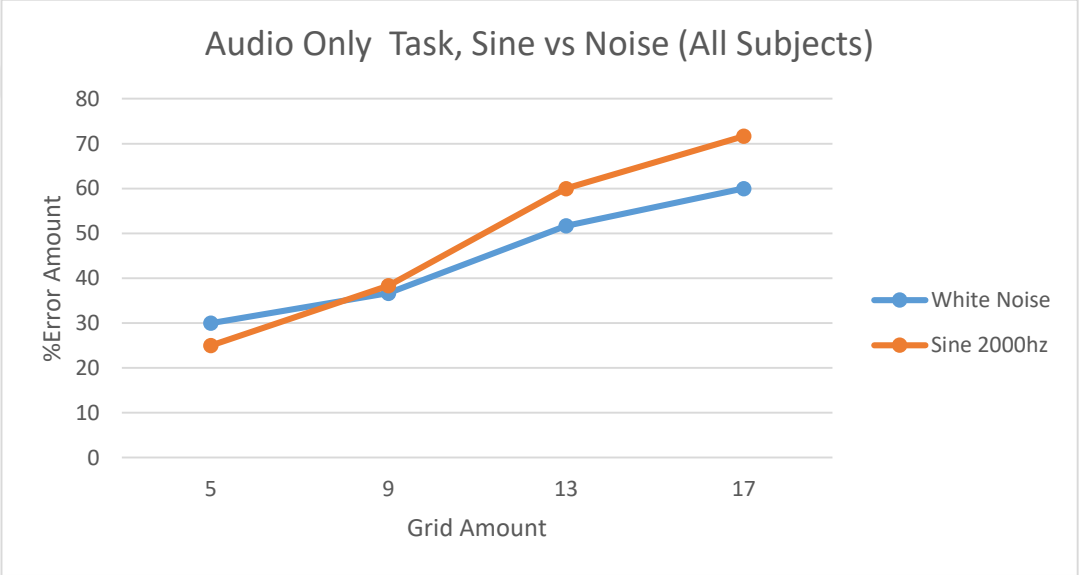


Figure 3.1: %Error amount for noise and sine waves in four different grid setups.

In this chapter, firstly, the results that how visual stimulus was affected the audio localization performances will be given which was the first aim of this thesis. Secondly, the results of the audio-only performances will be analyzed in detail. In both chapter estimations of sound engineers and naive listeners and also sine wave and noise signal comparison will be given.

3.1 Visual Bias On Sound Source Localization

Visual bias will be analyzed in two subtitles which are audio-visual task and enveloped audio-animated visual task titles. Briefly, in enveloped audio-animated visual task more visual bias was observed and it confirmed the one of the hypothesis. Also in both task more visual bias was observed when visual stimuli were presented at the center

position. Presented visual positions those are out of center; a specific experimental setup can be designed for analyzing it in detail. So generally comparison between noise and sine or sound engineers and naive listeners was made for the centrally located visual position.

3.1.1 Audio-visual task

The visual stimulus had affected the sound source localization performances. This was analyzed by looking at %error amounts and how visual stimulus changed the previously localized sound. In that grid setup, there was three grid spatial disparity between visual and audio stimuli in every sound position.

%Error amounts in 13 grid audio-only and 13 grid audio-visual tasks are given in figure 3.2. Also, the comparison between sound engineers and naive listeners are given in figure 3.3 and comparison between noise and sine waves in figure 3.4.

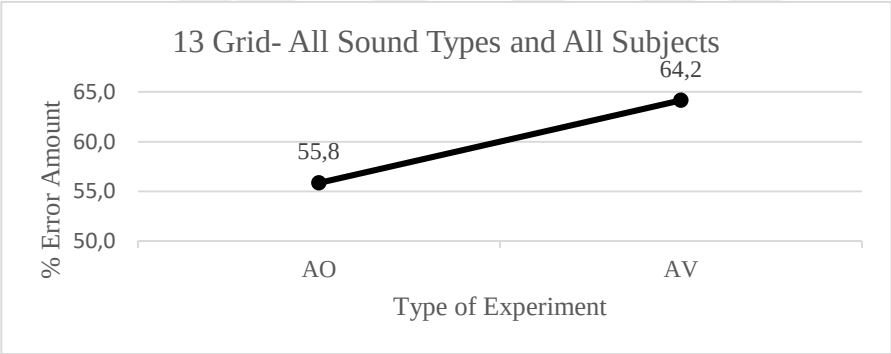


Figure 3.2: %Error amount in audio-only and audio-visual tasks.

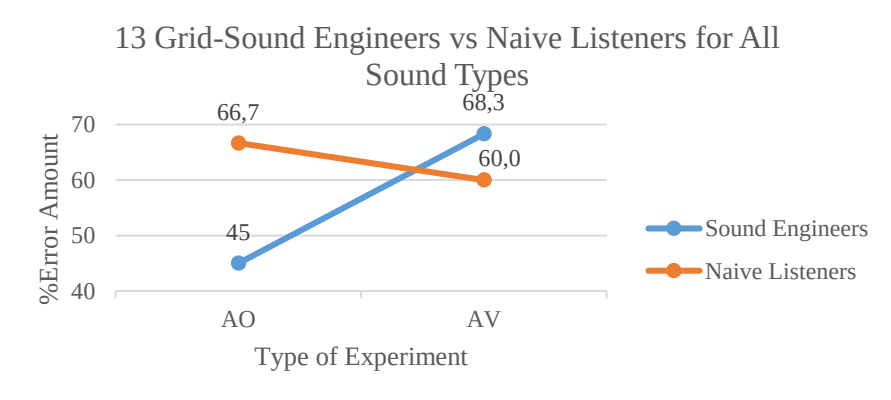


Figure 3.3: %Error amount of sound engineers and naive listeners in audio-only and audio-visual tasks.

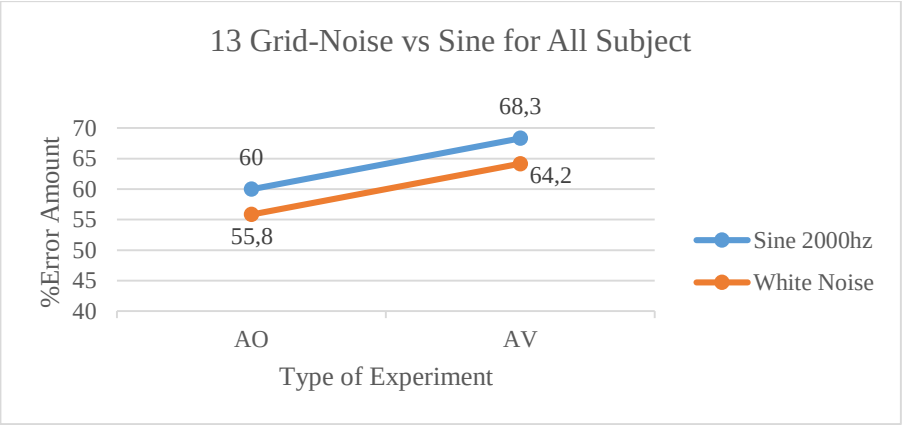


Figure 3.4: %Error amount of sine and noise in audio-only and audio-visual tasks.

As it is seen in Figure 3.3, naive listeners have lower error ratio in audio-visual task. Their %errors of sine estimations had changed %73 to %60, and noise estimations had stayed same at %60 in both task, and it was not expected. However, directly looking at the %error amount is not the correct way to understand visual biases on sound localization source.

Relative visual bias: How the visual stimulus affected the previously localized sounds was analyzed. Figures 3.5 and 3.6 show the number of grids changing in audio-only estimations when visual stimuli presented. X-axis of graphics refers to where visual stimuli presented. Y-axis refers to how much grids that visual stimuli pulled toward itself previously localized sound in audio-only task. Also, the 7th grid is the center point of 13 grid setup. In Audio-visual task, the 8th grid had two question, in one of them the sound was on the left side of the visual and in the other question on the right side (on 5th and 11th grids). The only first question was taken into account while these graphics were created for equalizing the data.

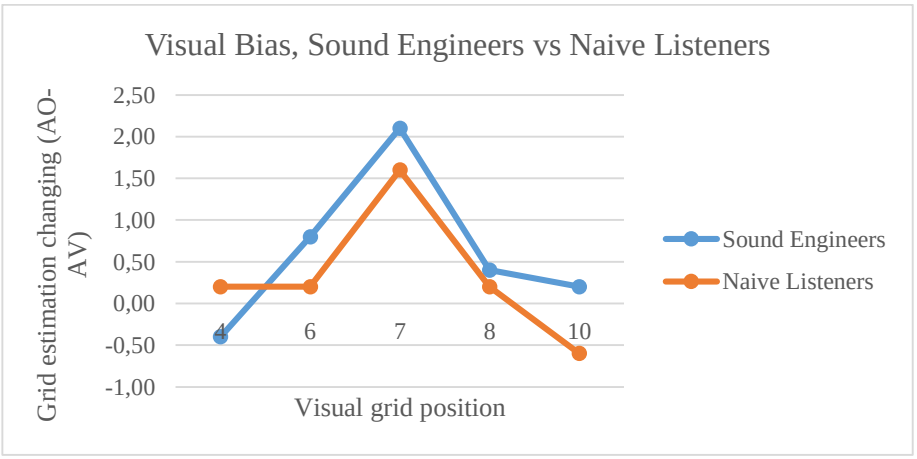


Figure 3.5: Visual biases in grid level (Sound engineers vs. Naive listeners).

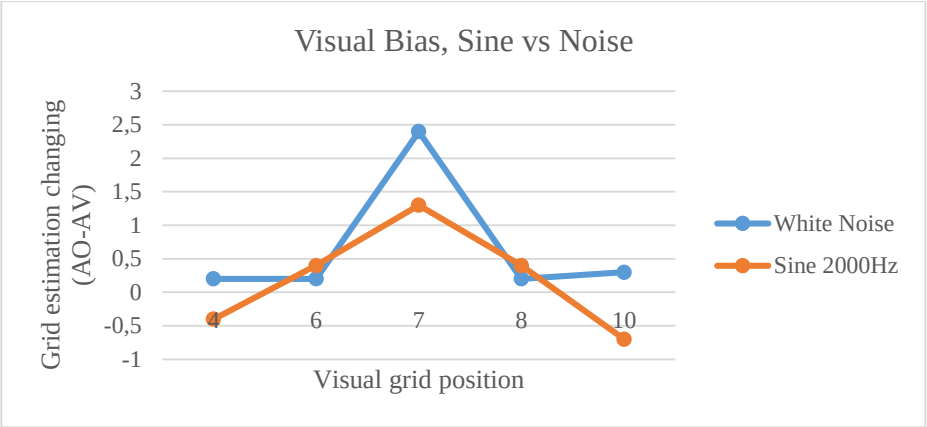


Figure 3.6: Visual biases in grid level (Sine vs. Noise).

When visual stimuli were presented in the center of the screen, ventriloquism effect got much stronger. On the sides, the effect was dramatically decreased. Even though when visual stimuli were presented one grid next to center position (5.5º difference) ventriloquism effect lost its effectiveness. Also, sine wave and naive listeners were not affected by visual stimuli (relatively) as much as noise signal and sound engineers when visual stimuli were presented in the center.

Figure 3.7 and 3.8 shows answers of ten subject for sine and noise stimulus when visual stimuli presented in the center. The orange line refers to visual position, the blue line refers to audio position, black circles refer to estimations in audio-only task, orange circles refers to estimations in audio-visual task. Subjects 1-5 are naive listeners, and 6-10 are sound engineers. Also, all graphics for all visual position could be found in appendix-b.

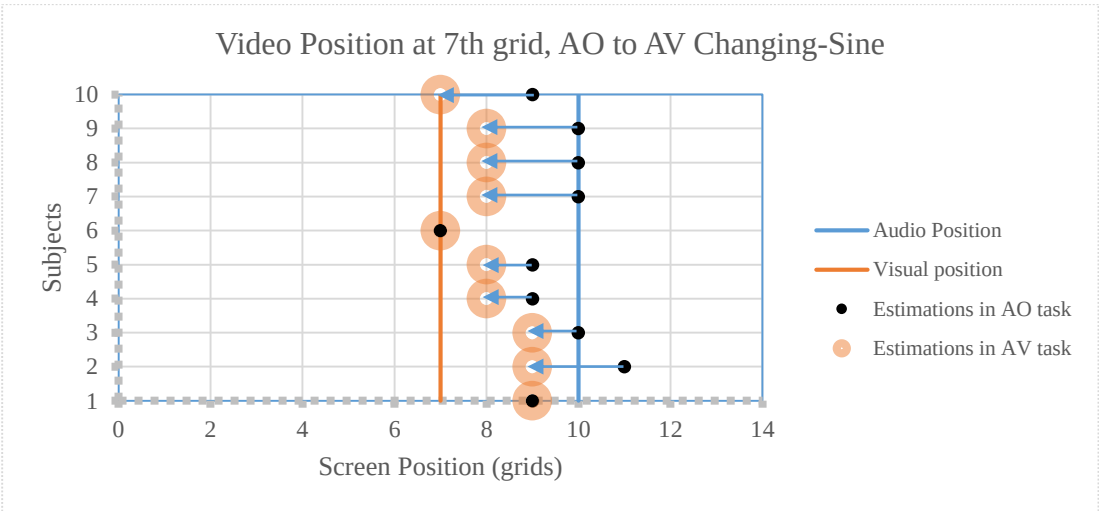


Figure 3.7: AO and AV estimations for sine wave (visual at 7th grid)

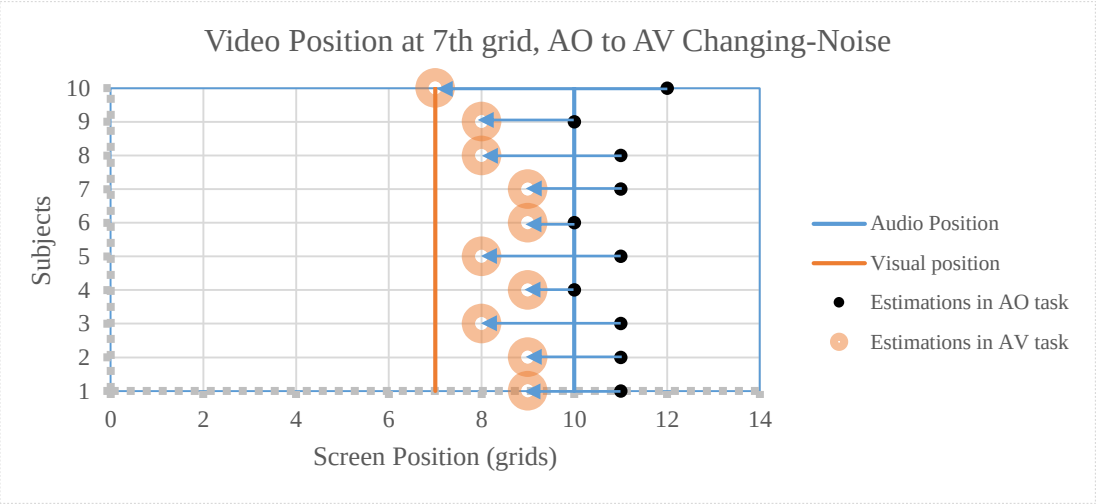


Figure 3.8: AO and AV estimations for noise signal (visual at 7th grid)

As it is understood from the graphics 3.8 and 3.9, visual stimuli in the center changed almost every subjects answer those approached the visual position. While sine wave has less relative changes, noise signal has more changes.

Absolute visual bias: If the results are compared without audio only task estimations, sine wave estimations are closer to the visual source (just see orange circles). Figure 3.9 and 3.10 show the visual bias percentage on sound localization when we just consider the audio-visual task. Graphics were drawn via using equation 1.4.

$$\%Amount\ of\ visual\ bias\ on\ perceptual\ sound\ position = \frac{Estimated\ grid^0 - Real\ sound\ position^0}{Visual\ grid^0 - Real\ sound\ Position^0} * 100 \quad \textbf{(1.4)}$$

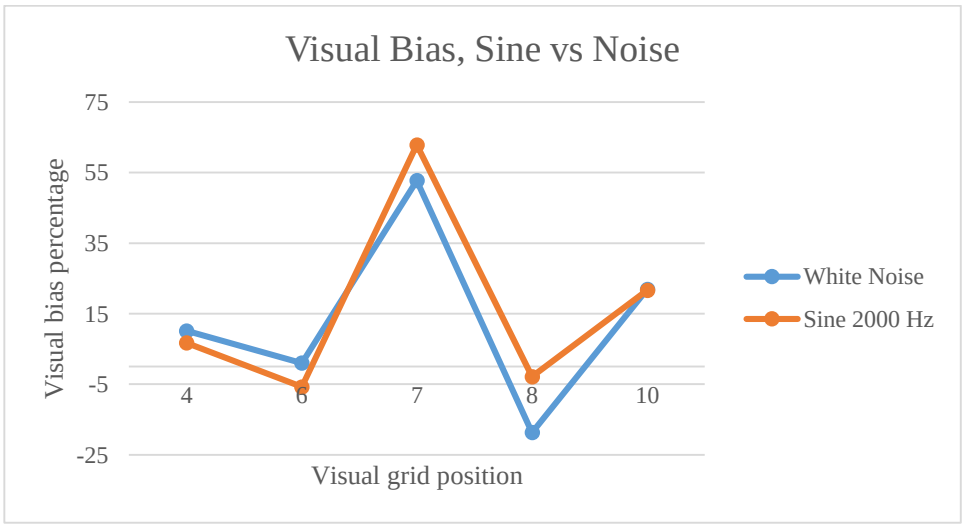


Figure 3.9: Visual biases percentage in AV task (Sine vs. Noise).

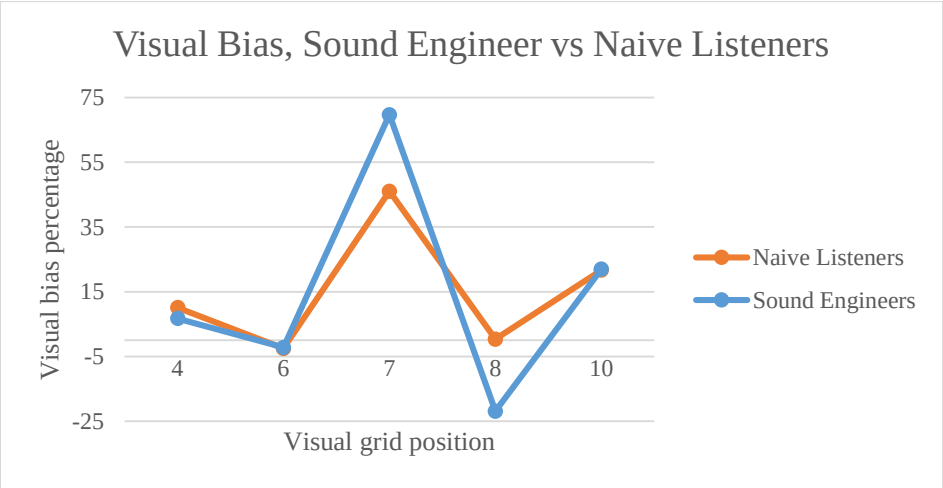


Figure 3.10: Visual biases percentage in AV task (Sound engineers vs. Naive listeners).

The results show that without considering the audio only task estimations, sine wave was affected by visual stimuli %10 more than the noise signal when visual stimulus was on the center. Furthermore, relative and absolute visual bias differences occurred because estimations of noise signal tended to be located more out of center than the sine wave estimations in audio-only task (see black dots in figure 3.7 and 3.8). So previously localized noise signal had more potential to have more relative changes. Also for the center position, naive listeners were significantly less affected by visual than sound engineers.

3.1.2 Enveloped audio-animated visual task

When audio and visual had a dynamic-abstract relation which was applied in enveloped audio-animated visual task, visual stimuli were quite effective on sound source localization. Figure 3.9 shows the %error amount of subjects in three tasks (AO – AV – En. A-An. V) and figure 3.10 shows the comparison between sound engineers and naive listeners. Because of enveloped audio-animated visual task was applied for only noise signal, graphics contains just data for noise signal estimations.

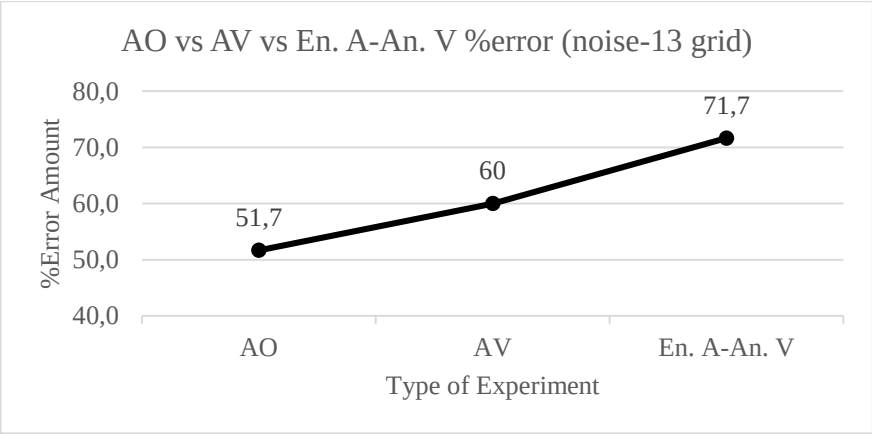


Figure 3.11: AO vs. AV vs. En. A-An. V %error (all subject groups for noise signa13 grid)

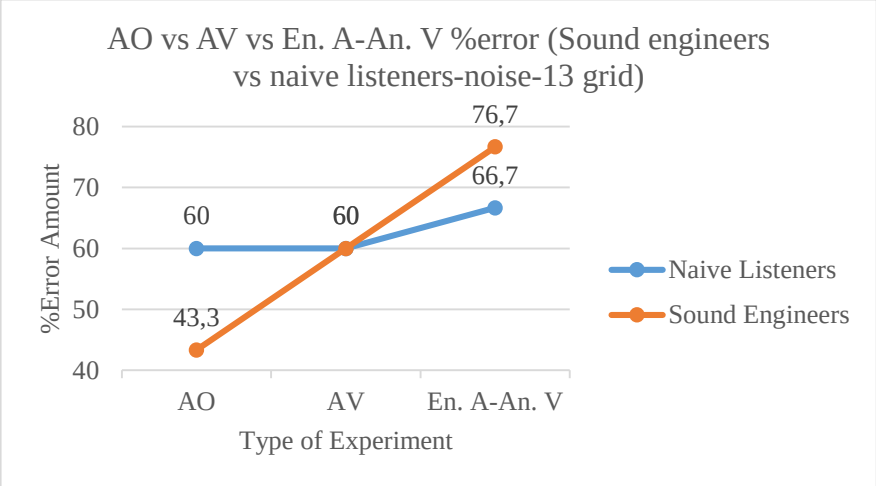


Figure 3.12: AO vs. AV vs. En. A-An. V %error (Sound engineers vs. naive listeners-noise-13 grid)

Figure 3.11 confirms the hypothesis that %error amount will be increased AO to AV and AV to En. A-An. V tasks. However, as it is mentioned before, it was understood that looking at the %error amount was not the correct way to analyze visual bias on sound localization.

Figure 3.11 shows the comparison between audio-visual task and enveloped audio-animated visual task regarding their visual biases on previously localized sounds in audio-only task.

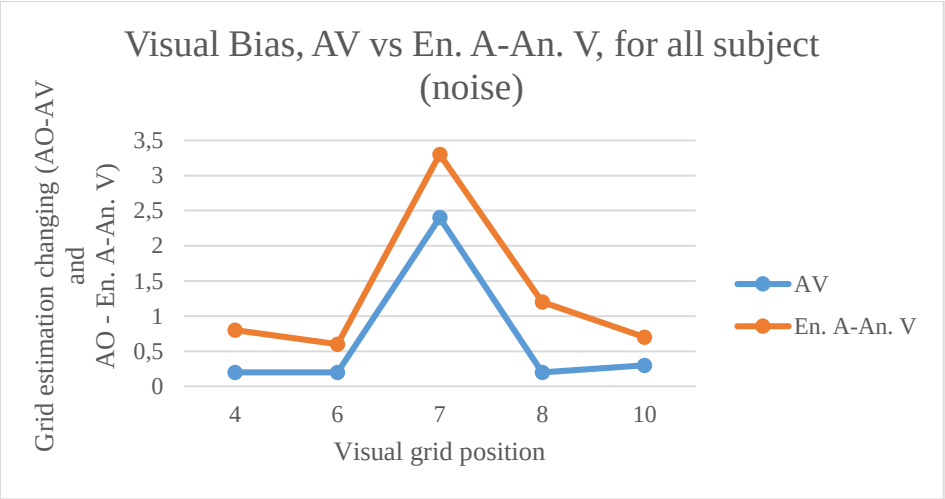


Figure 3.13: Visual biases in grid level (AV vs. En. A-An. V) for noise signal

Also, figure 3.12 shows the results comparison between naive listeners and sound engineers at enveloped audio-animated visual task.

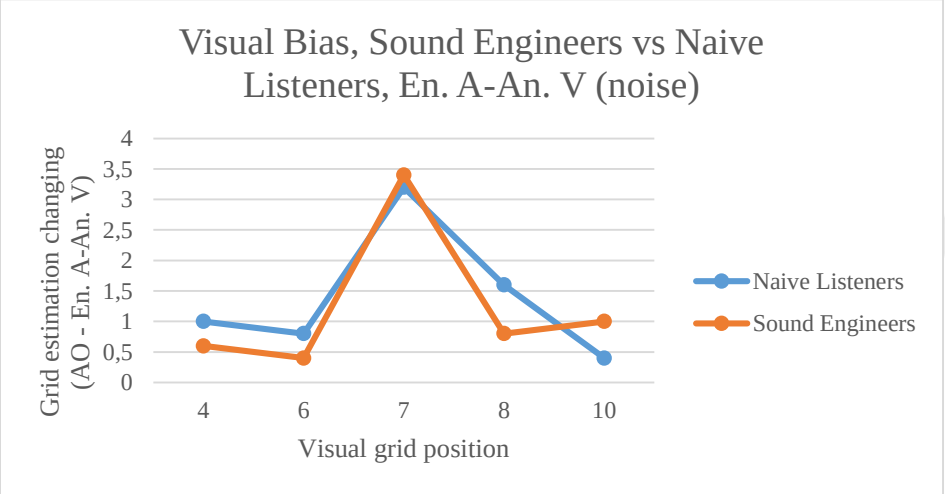


Figure 3.14: Visual biases in grid level (AV vs. En. A-An. V) for noise signal

It is clearly seen in figure 3.13 when audio and visual have a dynamic-abstract relation, more visual bias was observed. Even tough spatial disparity was three grids between audio and visual stimulus, in enveloped audio-animated visual task, visual changed the previously localized sound 3,3 grid when visual stimulus was presented in the center. Also as it is seen in figure 3.14, there were not noticeable estimation differences between sound engineers and naive listeners.

It can be clearly seen in figure 3.15 how animated visual was affected the previously localized sound positions when visual was at the center position. Graphic for the other visual position can be seen in appendix-b.

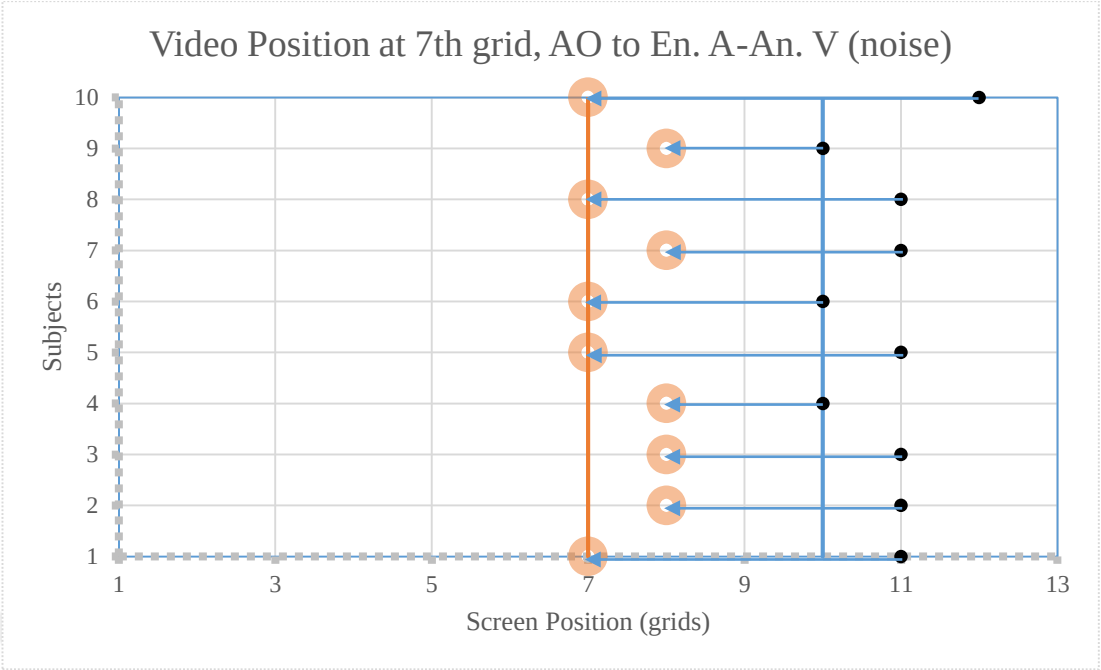


Figure 3.15: AO and En. A-An. V estimations for noise signal (visual at 7th grid)

If it is considered absolute visual biases by using equation 1.4, while the percentage of visual bias on sound source localization was %57,8 in audio-visual task, this percentage increased to %83,3 in enveloped audio-animated visual task, when the visual was presented on the center. Figure 3.15 shows the bias percentage comparison between audio-only task and enveloped audio-animated visual task. Also, Figure 3.16 shows the comparison between sound engineers and naive listeners.

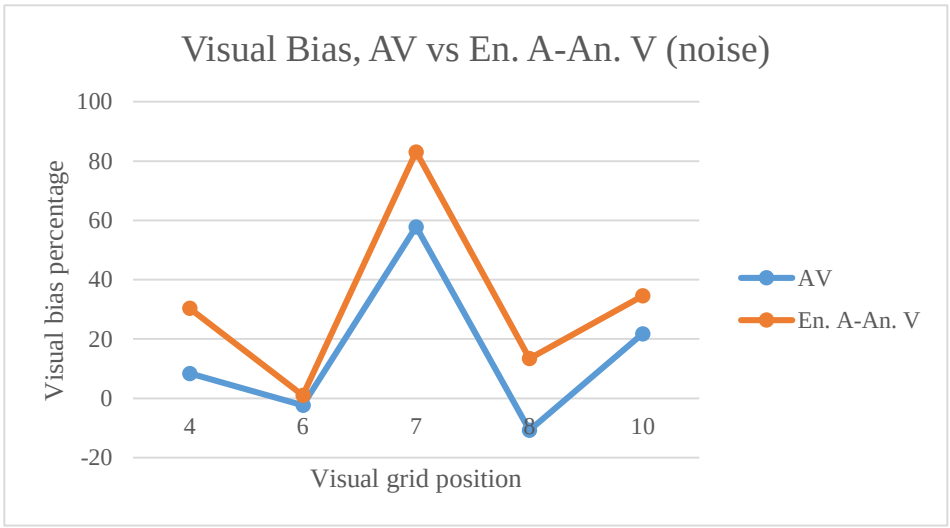


Figure 3.16: Visual biases percentage (AV vs. En. A-An. V) for noise signal

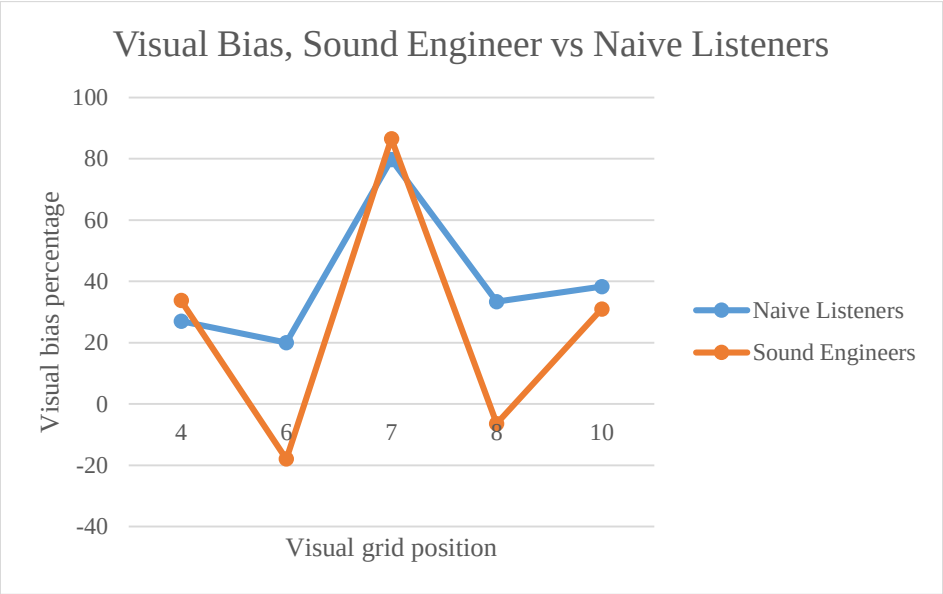


Figure 3.17: Visual biases percentage in En. A-An. V task (Sound engineers vs. Naive listeners).

In figure 3.17, interestingly, noticeable differences were observed between sound engineers and naive listeners when visual stimuli were presented at 6th and 8th grids. Thus, figures 3.18 and 3.19 are given to see individual answers at those specific visual positions. Because of figure 3.17 containing just En. A-An. V task information (absolute visual bias), just orange circles were considered to interpret the data in figures 3.18 and 3.19.

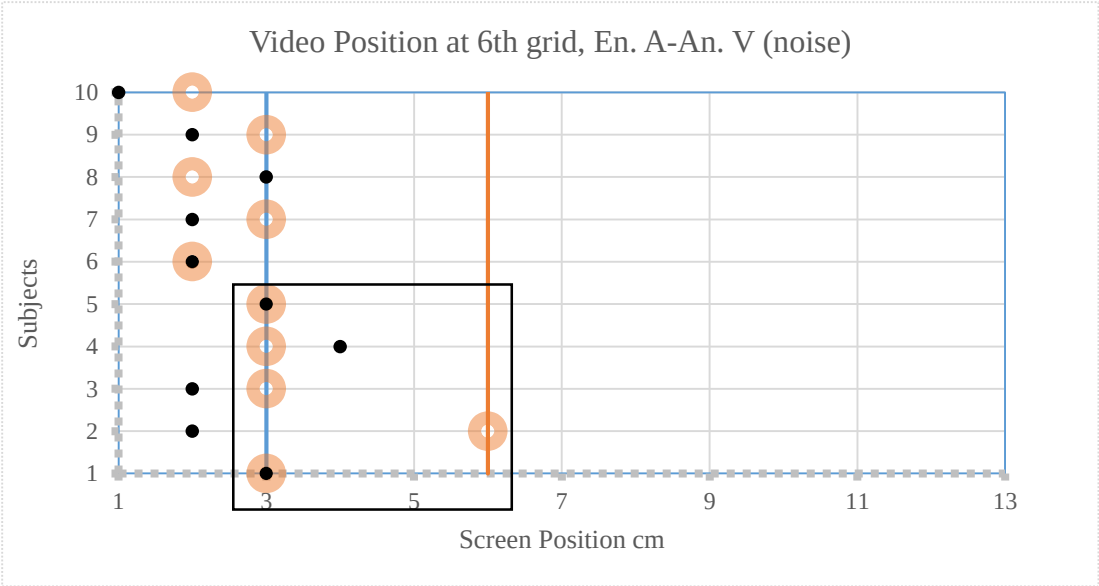


Figure 3.18: En. A-An. V estimations for noise signal (visual at 6th grid)

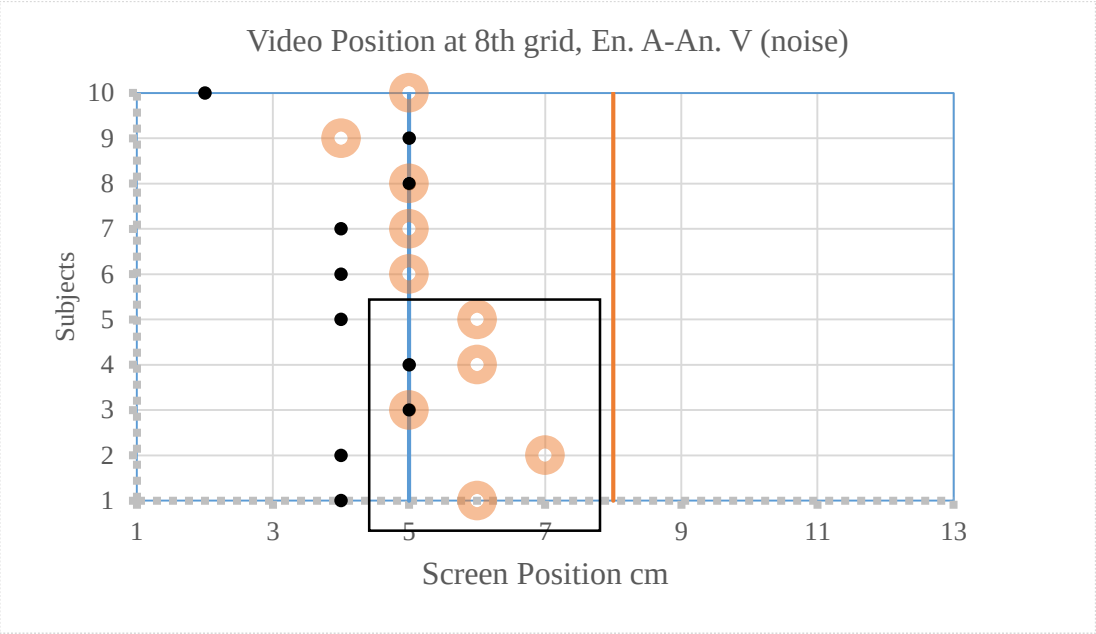


Figure 3.19: En. A-An. V estimations for noise signal (visual at 8th grid)

Black square covers the estimations of naive listeners in enveloped audio-animated visual task. As it is seen, their estimations were closer to the visual position. So this caused the obvious differences between sound engineers and naive listeners in the figure 3.17. However, it is difficult to make the inference that why it was observed. To make clear inferences in that specific point, the experiment needs more subject and estimations.

To sum up, more visual bias was observed in enveloped audio-animated visual task than audio-visual task. In both task centrally located visual stimuli affected the visual position more. Beyond that, for the centrally located visual stimuli, estimations of sine wave were affected less by visual stimuli when they are compared with previously localized sound. On the contrary, sine wave was affected more than noise signal when the absolute visual bias was considered. One of the hypotheses of this thesis was depending on the article of Alasis and Burr (2014), and they mentioned that when if audio stimuli have low spatial location precision, then in audio-visual task their perceptual spatial location is affected by visual stimuli more or vice versa. So it was hypothesized that sine wave would have lower location precision than noise signal in audio-only task than in the audio-visual task, it will be affected by visual stimuli more than noise signal. However, in that study, the results do not allow us for finding the clear ratio between two tasks and two sound types (see equation 1.5). Experiments need to be specialized for calculating this relationship.

In the subject groups comparison, the results were different from Komiyama's (1989) experiment. His results showed that acoustic engineers were discriminated easier than naive listeners and got annoyed quicker when visual and sound was presented at a different position (HDTV experiment). In this thesis, it was observed that experienced subjects actually could not discriminate audio-visual events easier than the naive listeners. Even they were affected by centrally located visual stimuli more than inexperienced subjects. However, as it is mentioned before while his experiment was depending on psychological effects of audio-visual discrepancy, this thesis was depending on audio localizations task. Even so, finding out this results were noteworthy.

3.2 Audio Only Performances

In audio-only task, total 480 estimation question was asked. There was four grid setup (5-9-13-17 grids). In every grid setup five sound engineers and five naive listeners answered six sine and six noise questions.

Figures below shows %errors when the grids amount was increased. Comparison between sound engineers and naive listeners is presented in figure 3.20, noise and sine in figure 3.21.

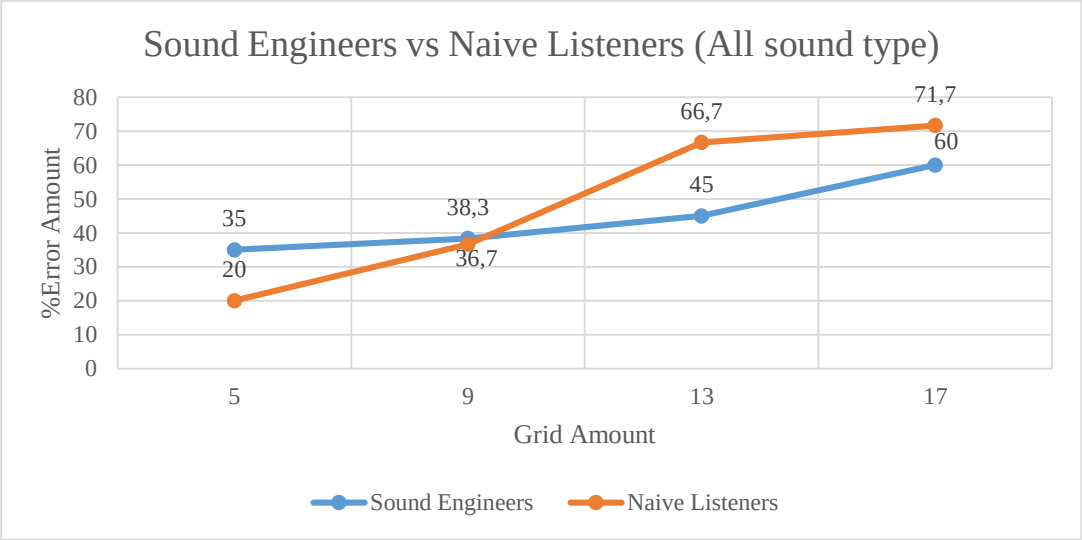


Figure 3.20: %Errors in audio-only task, sound engineers vs. naive listeners.

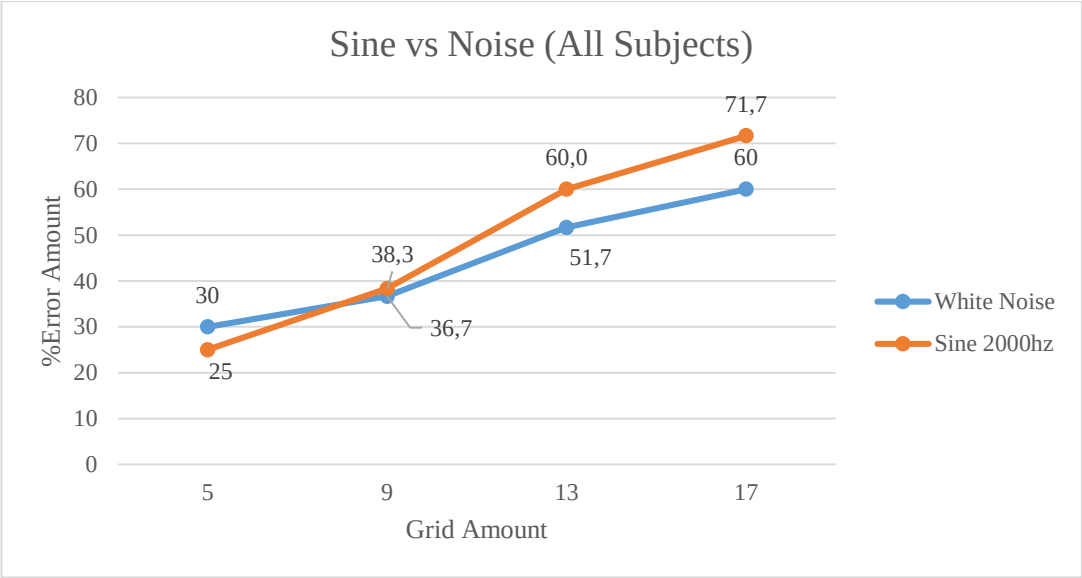


Figure 3.21: %Errors in audio-only task, sound noise vs. sine.

In 5 grid setup, sine wave questions were estimated in better success rate, and it was not expected. The reason of this could be; experiment began with noise signal questions then sine wave question was asked for each subject. So subjects recognized the system with noise signal questions and they could be better in sine wave questions. After all, in the other grid setups, they had already known the system, so error amounts for noise and sine returned the expected rates, and this was the one of the hypothesis that estimations error of sine wave would be more than noise signal. Moreover, angular divergences of sine wave were higher than the noise signal. Also, results of comparison between sound engineers and naive listeners was not expected. Even though in 13 and 17 grid setups, sound engineers have a better success rate, it does not totally support the hypothesis that sound engineers will have better success rate than naive listeners. Also, angular divergences comparison between sound engineers and naive listeners does not support the hypothesis.

So to better understand the results, in the audio-only test, sound estimation questions asked on all grid setups were presented in a single graph, taking into account their positions on the screen. While figure 3.22 shows the real positions of the sounds, figure 3.23 and 3.24 show the subjects' estimations for noise and sine waves. Also, It can be seen in those graphics, estimations those have how many degrees divergences depending on real sound positions. If the divergences have minus values that means subjects estimations are on the left side of the real sound position or vice versa. Besides these, the darkness of circles increases with if the estimations have same angular

divergences in that specific real sound position. The red dots show the mean divergence of the subjects at that real sound position. (For the center 30, for the remaining position 10 question were asked).

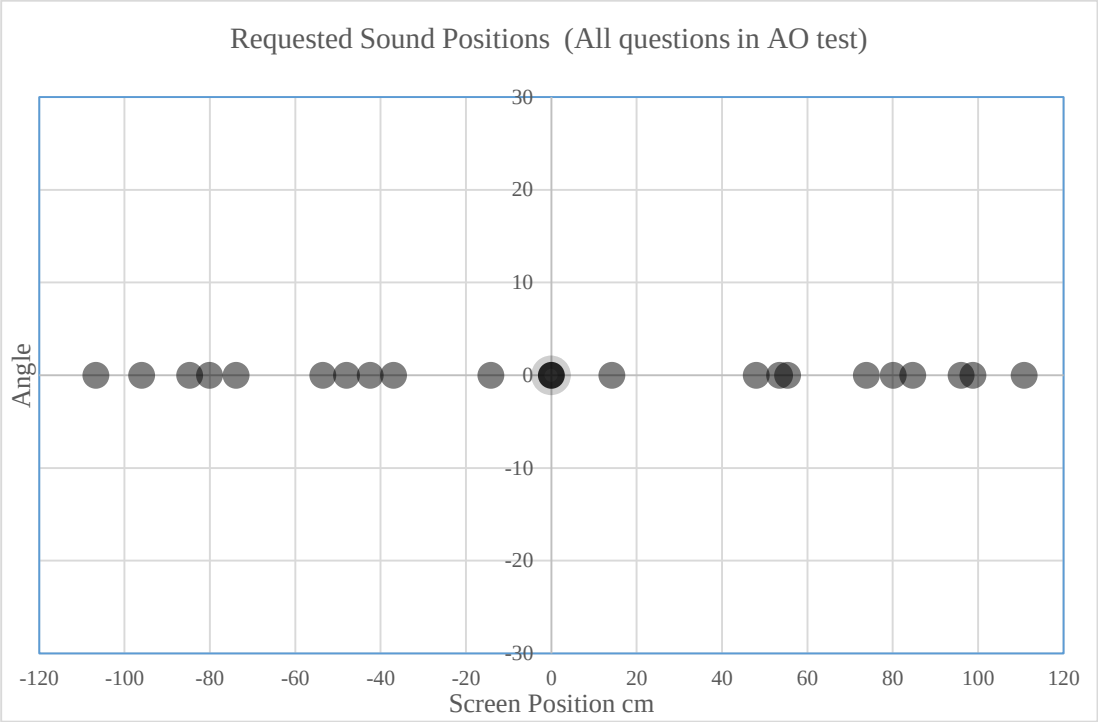


Figure 3.22: Requested Sound Positions (All questions in AO test)

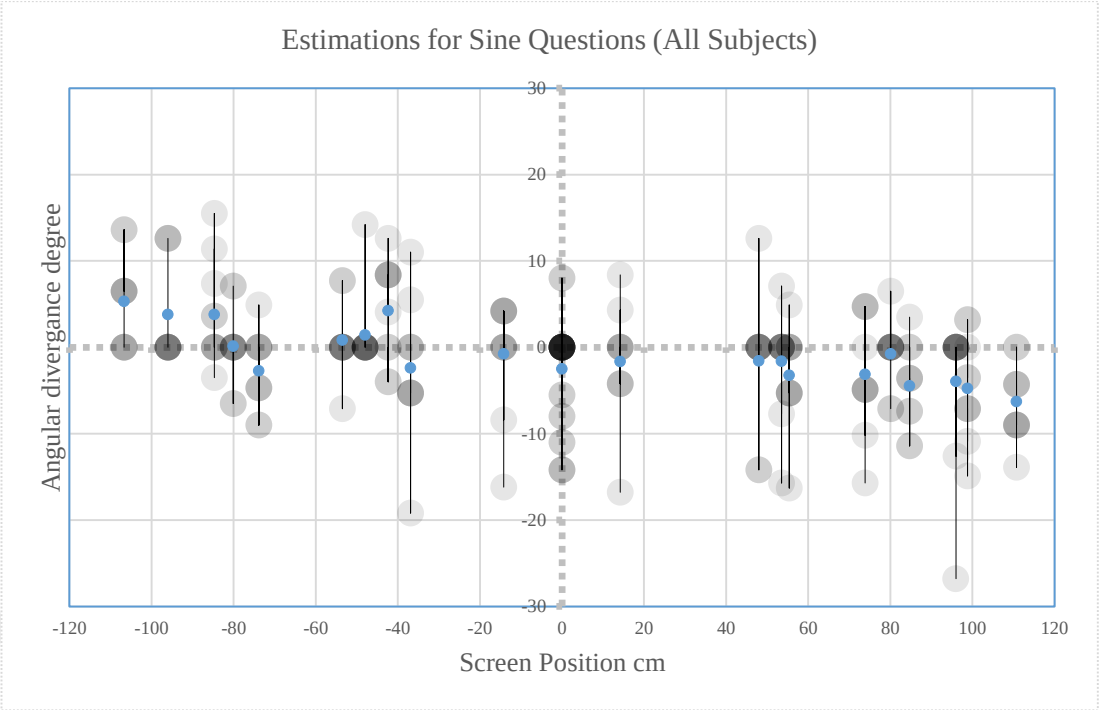


Figure 3.23: All estimations in audio-only task for sine questions (All Subjects)

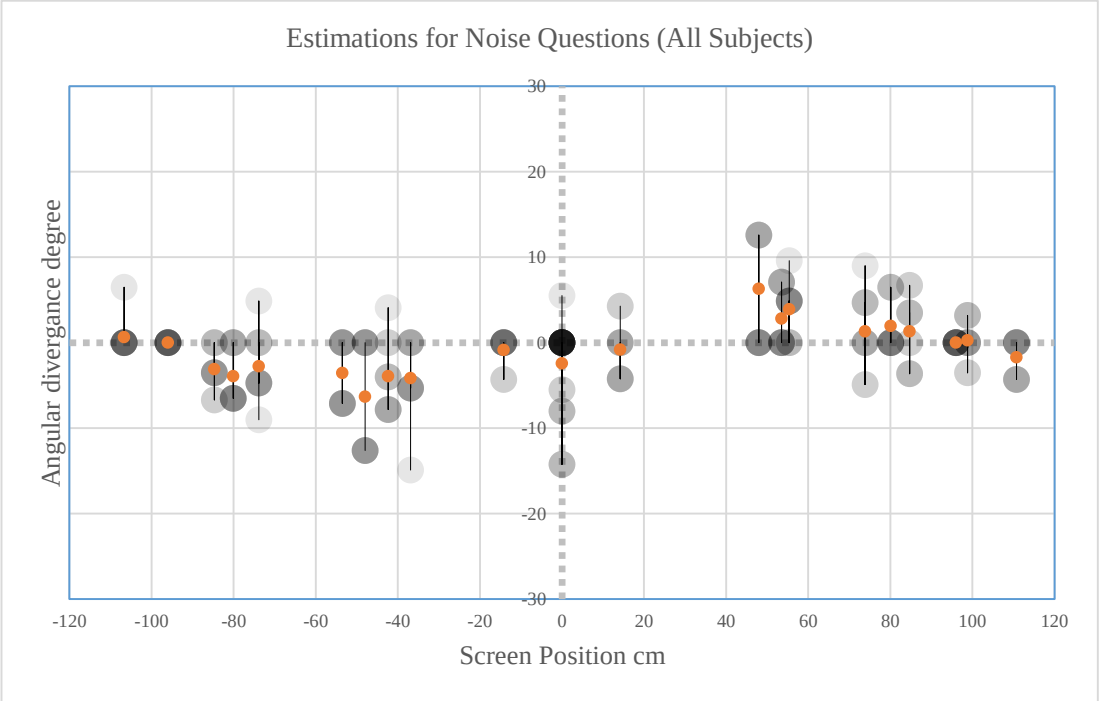


Figure 3.24: All estimations in audio-only task for noise questions (All Subjects)

As it is seen in the figures, sine wave estimations have more angular variety. It can be said that noise signal has better localization accuracy. Also, another important thing that can be observed in these graphics is angular divergence tendencies. When it is considered the left side and right side of the screen separately, sine wave estimations have tendency to be closer to the center (on left side positive, on the right side negative

angular divergence values), on the contrary noise signal estimations have tendency to be far away from the center (on left side negative, on the right side positive angular divergence values). Figure 3.25 shows this comparison via using mean values of angular divergences.

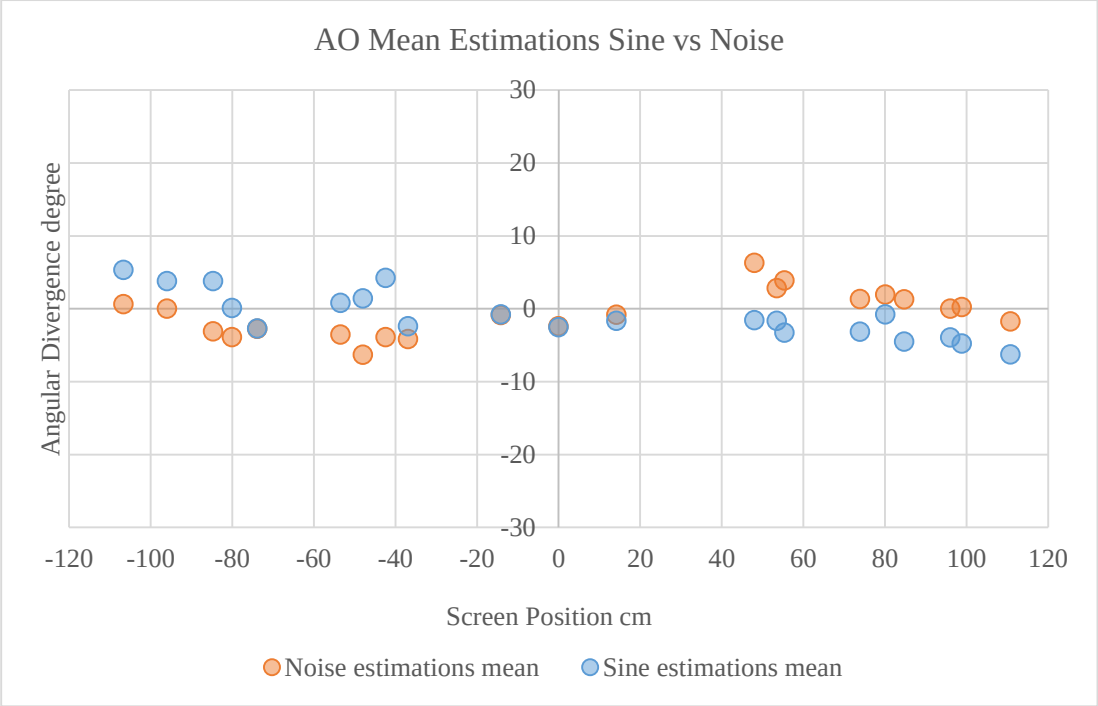


Figure 3.25: Mean angular divergence comparison between sine and noise signal

Also as it was mentioned before, this tendency affected the relative visual biases in audio-visual task (see figure 3.7 and 3.8). Basically, previously localized noise signal in audio-only test potentially is affected more by centrally located visual stimuli when it is compared with sine wave. Also around the center estimations for noise and sine close to each other, however at the end of the projected area, estimations of noise signal have better success. These results differ from Yost's (2016) experiment. In his study, as the sound signal moved away from the center, it was observed that the angle of divergence increased and all the tested sound types had this tendency (2 or $1/10^{th}$ octave wide, 250, 2000 and 4000 Hz center frequency signals). In that study, it's hard to find that kind of relation. The reason for these differences can be, most probably not to stabilize head of the subjects during the experiment or another possibility can be using a flat surface for the projected area (angular differences between speakers).

All the figures below show the comparison between sound engineers and naive listeners. Figure 3.26 and 3.27 show the performances of naive listeners and sound engineers on sine wave. Also in figure 3.28, mean estimations of both are compared.

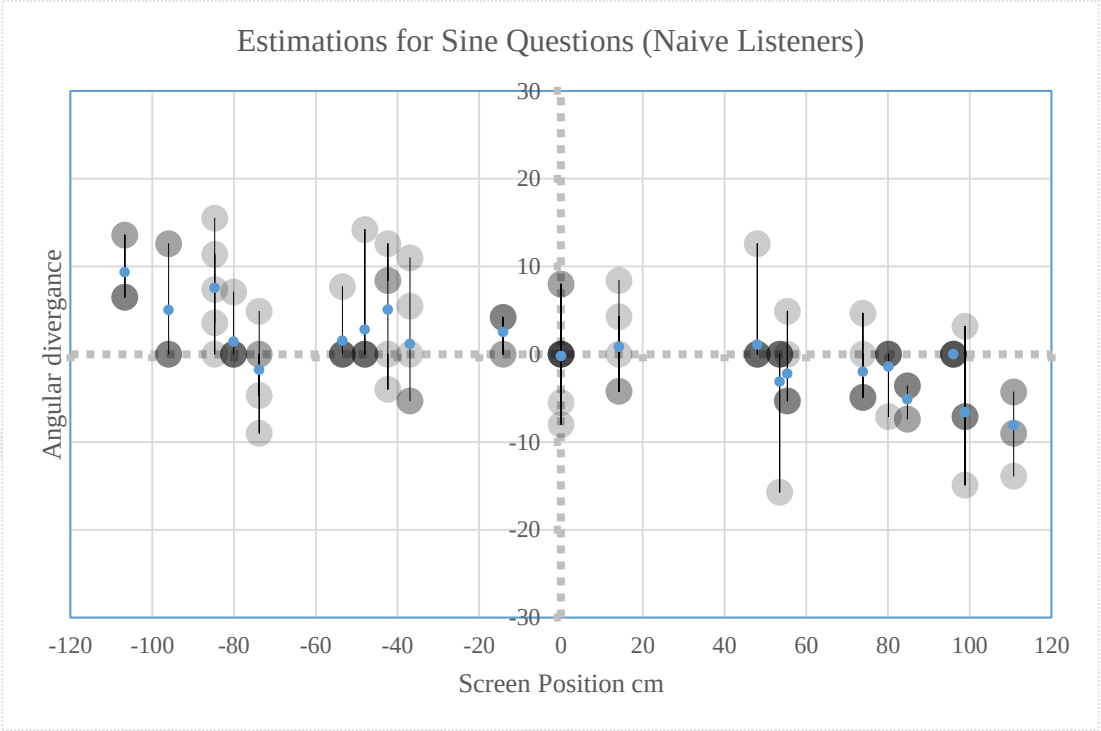


Figure 3.26: Estimations for Sine Questions (Naive Listeners)

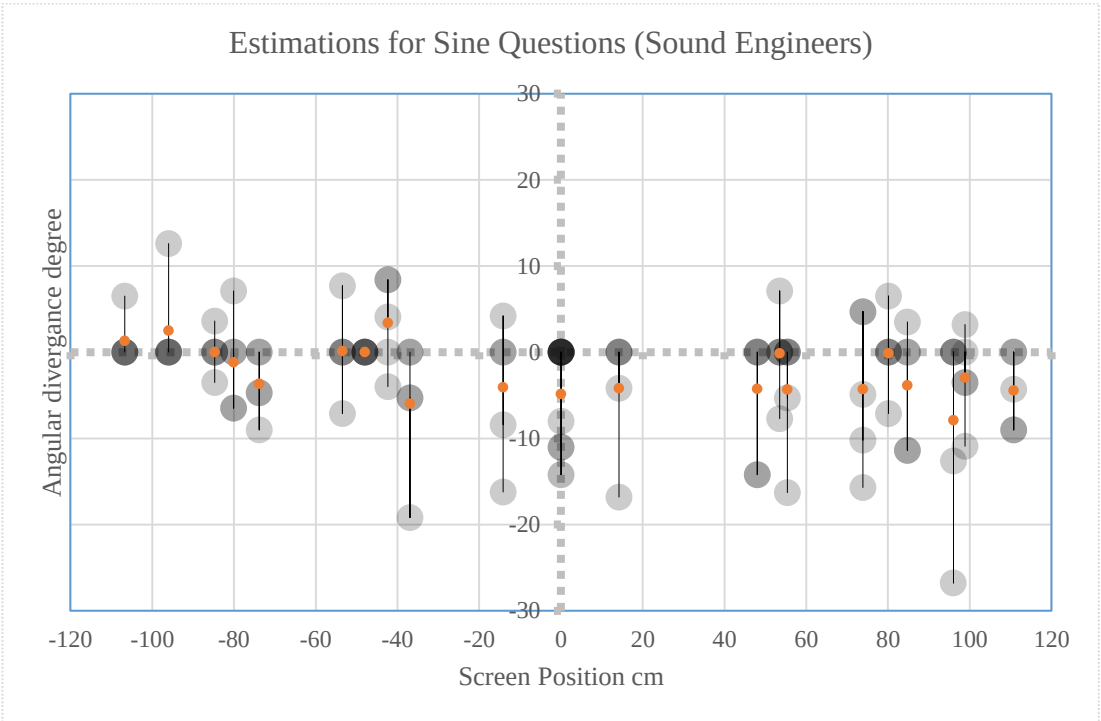


Figure 3.27: Estimations for Sine Questions (Sound Engineers)

Figure 3.27 and 3.28 show that there aren't significant differences between sound engineers and naive listeners. For sound engineers, while angular variety tendency of estimations is on the left side of the real sound position, for naive listeners is on the righter side respectively.

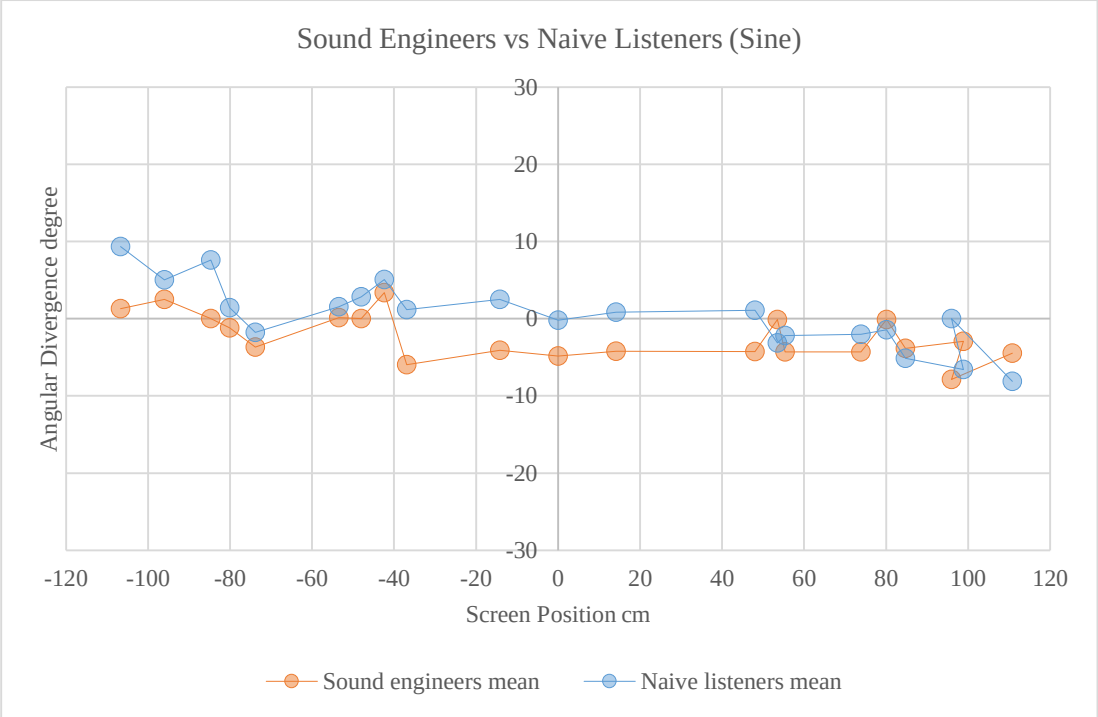


Figure 3.28: Estimations mean, Sound Engineers vs. Naive Listeners (Sine)

As it is understood from figure 3.28, both subject groups have a similar slope. However, on the left side of the screen similarity of the slopes changes and enters each other. Also while naive listeners have better results around the center, sound engineers have better results at various area of the screen.

Figure 3.29, 3.30 and 3.31 shows the noise signal comparison between two subject groups.

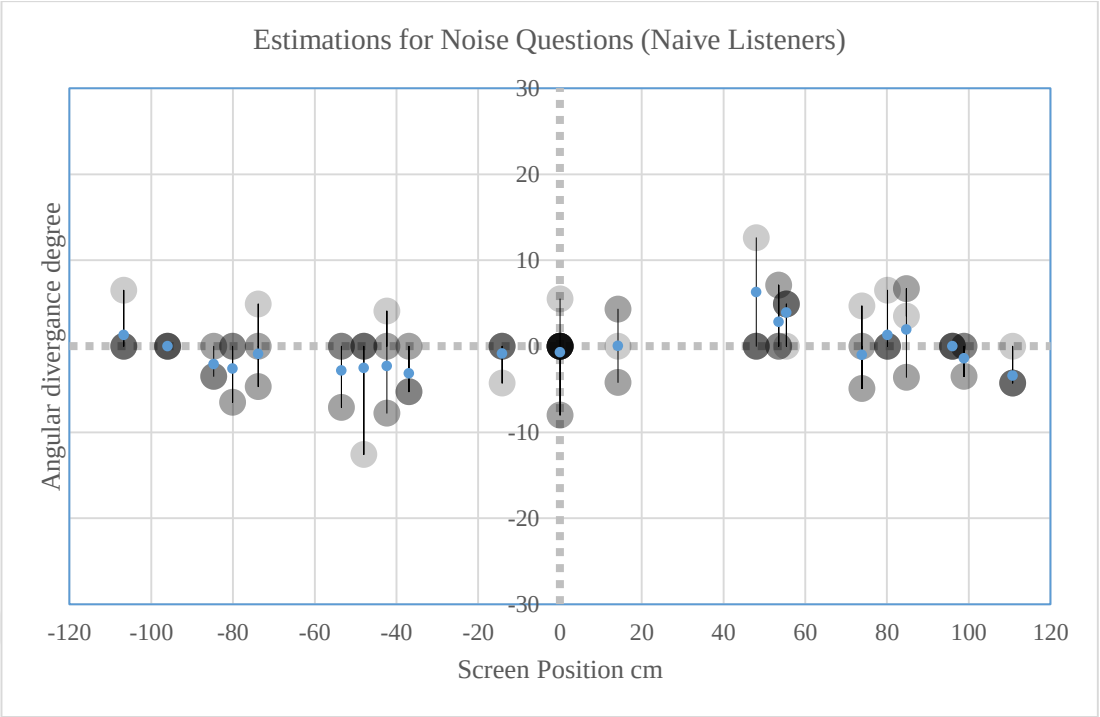


Figure 3.29: Estimations for Noise Questions (Naive Listeners)

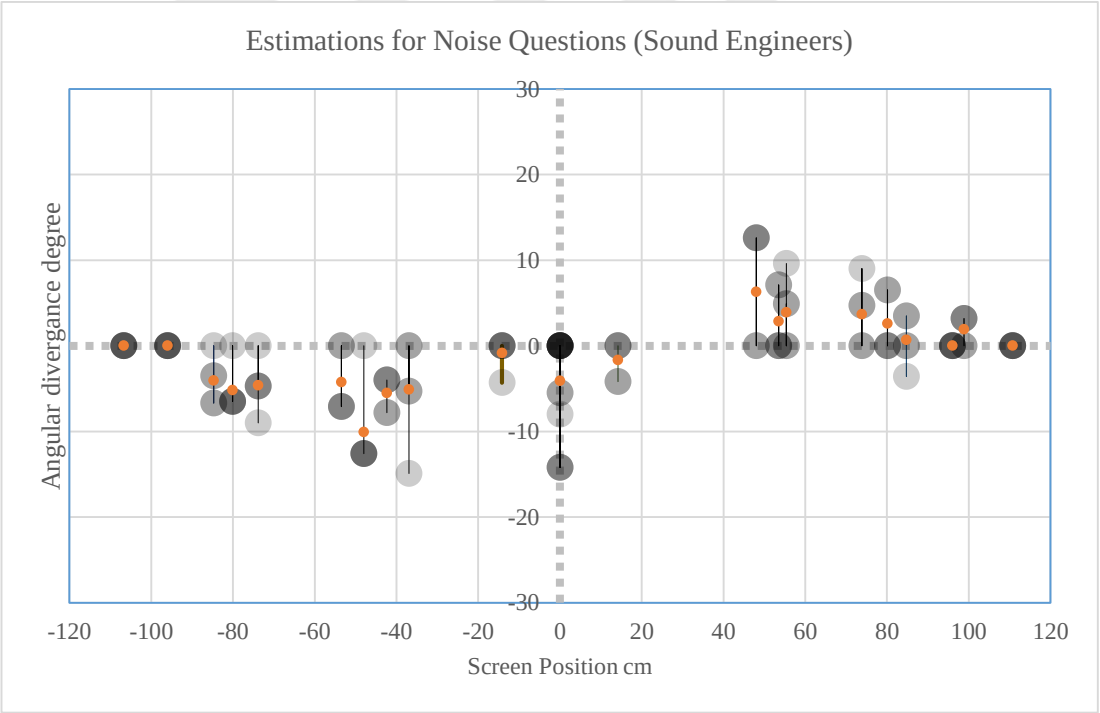


Figure 3.30: Estimations for Noise Questions (Sound Engineers)

Figure 3.29 and 3.30 show that there is no significant angular divergences variety between two subject groups. However, estimations of sound engineers for the center position have always minus angular divergences (left side of the real sound position).

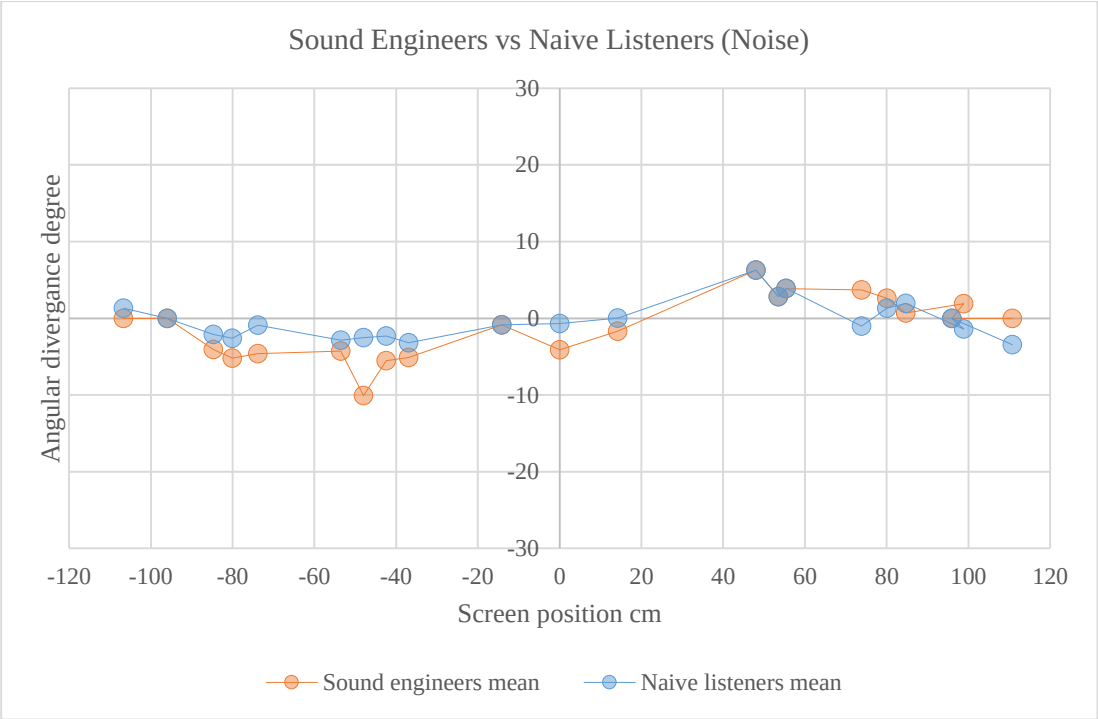


Figure 3.31: Estimations mean, Sound Engineers vs. Naive Listeners (Sine

In the figure 3.31, again both subject groups have a similar slope, and again naive listeners have slightly better performances in around the center position.

To sum up, estimations of the noise signal and sine wave have different characteristics. While sine wave estimations have more angular variety, noise signal have not. While estimations of sine wave tend to be prone to center, noise signal have opposite characteristic. Furthermore, estimations of sound engineers and naive listeners have similar characteristics. Only small differences were observed that while naive listeners have slightly better performances around the center, sound engineers have slightly better performances on the leftmost and rightmost side of the screen.

Thus, one of the hypotheses of this thesis argued that sound engineers will have better performances in audio-only task. Actually results proved that this is not correct for this study. Maybe more subject participants probably change the results. Because each subject groups have one or two people that have different performances when their estimations compared with their own groups.

4. CONCLUSION

This study mostly focused on making a comparison between experiment variables instead of making exact inferences. For able to compare variables, three separate task were designed. These were audio-only task (AO), audio-visual task (AV) and enveloped audio-animated visual task (An. A-En. V).

In audio-only task, there were four different grid setups (5-9-13-17 grids). When the grids amount increased, the angular differences between speakers reduced. Five sound engineers and five naive listeners estimated six 2000Hz sine wave and six white noise localization questions in every grid setup (total 480 question in AO task). After all, subjects participated those grid setups, it was checked that in which grid setup, %50 error threshold was reached. The aim of doing this is to obtain which angular difference between speakers causes localization uncertainty and also to use obtained grid setup in audio-visual and enveloped audio-animated visual tasks. Thus, in 13-grid setup, %50 error threshold was reached (mean angular differences between speaker was 5º). Also in AO tasks, estimations in all grid setups were analyzed together for taking advantages of using more data and making more general inferences. In that way, estimations of noise and sine or estimations of sound engineers and naive listeners was compared. Results showed that estimations of sine wave have more angular varieties than the noise signal estimations, and also past studies proved this. However, one interesting results of this study was while sine estimations tended to be close to the center, noise estimations tended to be away from the center. Another result of the audio-only task was not observing the significant performance differences between sound engineers and naive listener.

In audio-visual task was applied at 13 grid setup and previously asked sound positions in audio-only task were presented again in a different order with static visual stimuli (black circle with no fade in and fade out time). The comparison between sound engineers and naive listeners or sine and noise signal was analyzed. The results showed that, when visual stimuli were presented in the center position, more visual bias was observed for both subject groups and sound types, and the comparison generally was

made depending on centrally located visual stimuli. Also, two types of visual bias calculations were used, and these were relative and absolute visual biases. When it was considered the amount of estimation location changes from audio-only task to audio-visual task it was called relative visual bias or when it was considered how visual stimuli changed the position of estimation just in audio-visual task, it was called absolute visual bias. In both calculation method, estimations of sound engineers were affected by centrally located visual stimuli more than naive listeners' estimations. Beyond that, estimations of sine wave were affected less than noise signal's estimations when it was considered relative visual bias. On the contrary, they were affected more when it was considered absolute visual bias for the centrally located visual stimuli. As it is mentioned before, this was observed because of most probably localization tendencies of sound types were being different in audio-only task.

In enveloped audio-animated visual task, a new relationship between audio and visual stimuli was tried to be established. While in audio-visual task they had an abstract-static relationship, in this task they had an abstract-dynamic relation. Black circle as visual stimuli had fade-in fade-out time and the audio stimulus had attack and release time which helped them to act together. The experiment was applied in 13 grid setup with using just noise signal. When this task was compared with audio-visual task, in every visual position, and for every subject groups more visual bias was observed. Absolute visual bias percentage increased from %57,8 to %83,1 for the centrally located visual stimuli. Furthermore, in this task, estimations of both subject groups were affected by visual stimuli almost in the same amount.

This study had possible deficiencies when it was compared with past studies. It is because of using flat projected area which caused angular asymmetry between speakers or having not enough questions or subjects for observing the specified things in deeply. However, it had superficial inferences about sound localization and ventriloquism effect which noteworthy to study in detail. For instance, instead of researching spatial localization precision of different sound type (different frequency contents), researching their perceptual vectorial tendencies in space can be studied as a different research area in sound localization issue. Another thing that audio-visual motion relation can be studied in detail for abstract multimedia artwork with ventriloquism effect, to create a perceptual illusion by changing their togetherness (dynamic to static or vice versa). Because, results of recent thesis showed that when

audio and visual, are differently located in a spatial plane, have a dynamic relation (motion relation), subjects can more focus on the event instead of discriminating.

REFERENCES

Agganis, B. T., Muday, J. A., & Schirillo, J. A. (2010). Visual biasing of auditory localization in azimuth and depth. *Perceptual and Motor Skills, 111*(3), 872-892. doi:10.2466/22.24.27.PMS.111.6.872-892

Alasis, D., Burr, D. (2004). The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology, Vol. 14,* 257-262. doi: 10.1016/j.cub.2004.01.029

Andeol, G., Savel, S., Guillaume, A. (2015). Perceptual Factors Contribute More Than Acoustical Factors to Sound Localization Abilities with Virtual Sources. *Frontiers in Neuroscience, 8(451),* 1-17 doi:10.3389/fnins.2014.00451

André, C., Corteel, E., Embrechts, J.-J., Verly, J., and Katz, B. F. G. (2014). Subjective Evaluation of the Audiovisual Spatial Congruence in the Case of Stereoscopic-3D Video and Wave field Synthesis. *Int. J. Human-Computer Studies, 72,* 23–32.

Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A, 20*(7), 1391-1397. doi:10.1364/JOSAA.20.001391

Bertelson, P., & Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. *Psychonomic Bulletin and Review, 5*(3), 482-489.

Bur, A., Wurtz, P., Müri, R.M., Hügli, H. (2007). Dynamic visual attention: competitive versus motion priority scheme. *5th International Conference Computer Vision Systems (ICVS).* Germany: Bielefeld University. URL: http://biecoll.ub.uni-bielefeld.de/volltexte/2007/75/pdf/WCAA2007-160.pdf

Demir, A. (n.d). Müzik İleri Araştırmalar Merkezi büyüyor. NTV archive website. URL: http://arsiv.ntv.com.tr/news/159522.asp

Ernst, M. O., Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8(4),* 162-169. doi:10.1016/j.tics.2004.02.002

Grunwald, M., Dr. Phil. (2008). Human Haptic Perception: Basics and Applications. Basel;Boston;: Birkhäuser

Hendrickx, E., Paquier, M., Koehl, V., Palacino, J. (2015). Ventriloquism effect with sound stimuli varying in both azimuth and elevation. *The Journal of the Acoustical Society of America, 138(6),* 3687-3697. doi: 10.1121/1.4937758

Komiyama, S. (1989). Subjective Evaluation of Angular Displacement between Picture and Sound Directions for HDTV Sound Systems. *Journal of Audio Engineering Society Vol. 37 No: 4,* 210-214.

Kytö, M., Kusumoto, K., Oittinen, P. (2015). The Ventriloquist Effect in Augmented Reality. *IEEE International Symposium on Mixed and Augmented Reality* pp. 49-53.

Makous, J. C., & Middlebrooks, J. C. (1990). Two-Dimensional Sound Localization by Human Listeners. *Journal of the Acoustical Society of America, 87*(5), 2188-2200. doi:10.1121/1.399186

Montagne, C., Zhou, Y. (2016). Visual Capture of a Stereo Sound: Interactions Between Cue Reliability, Sound Localization Variability, and Cross-Modal Bias. *The Journal of the Acoustical Society of America, 140(1),* 471-485 doi: 10.1121/1.4955314

Morrongiello, B. A., Rocca, P. T. (1987). Infants' Localization of Sounds in the Median Vertical Plane: Estimates of Minimum Audible Angle. *Journal of Experimental Child Psychology 43,* 181-193.

Noise/Vibration Testing Laboratory. (2012). OTAM website. URL: http://www.otam.com.tr/SayfaDetayD.aspx?ItemID=111

Perrott, D. R., & Saberi, K. (1990). Minimum audible angle thresholds for sources varying in both elevation and azimuth. *Journal of the Acoustical Society of America, 87*(4), 1728-1729. doi:10.1121/1.399421

Recanzone, G. H., Makhamra, S. D. D. R., & Guard, D. C. (1998). Comparison of relative and absolute sound localization ability in humans. *Journal of the Acoustical Society of America, 103*(2), 1085-1097. doi:10.1121/1.421222

Su, T. K., Recanzone, G. H. (2001). Differential Effect of Near-Threshold Stimulus Intensities on Sound Localization Performance in Azimuth and Elevation in Normal Human Subjects. *Journal of the Association for Research in Otolaryngology, 2(3),* 246-256. doi:10.1007/s101620010073

Török, A., Mestre, D., Honbolgo, Ferenc., Mallet, Pierre., Pergandi, J., Csépe, V. (2015). It Sounds Real When You See It. Realistic Sound Source Simulation in Multimodal Virtual Environments. *Journal on Multimodal User Interfaces, 9,* 323-331. doi: 10.1007/s12193-015-0185-4

Vatakis, A., Spence, C. (2007). Crossmodal Binding: Evaluating the Unity Assumption Using Audiovisual Speech Stimuli. *Perception & Psychophysics. 69,* 744–756.

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., & Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Experimental Brain Research, 158*(2), 252-258. doi:10.1007/s00221-004-1899-9

Warren, D. H., Welch, R. B., & McCarthy, T. J. (1981). The Role of Visual-Auditory "Compellingness" In The Ventriloquism Effect: Implications for Transitivity Among the Spatial Senses. *Perception & Psychophysics, 30*(6), 557-564. doi:10.3758/BF03202010

Werner, S., Liebetrau, J., Sporer, T. (2013). Vertical Sound Source Localization Influenced by Visual Stimuli. *Signal Processing Research 2(2*), 29-38. URL: http://www.seipub.org/spr/paperInfo.aspx?ID=2701

Yost, W. A., (2016). Sound Source Localization Identification Accuracy: Level and Duration Dependencies. *The Journal of Acoustical Society of America 140(1),* EL14-EL19.

APPENDICES

Appendix A

Figure A.1: Experiment area

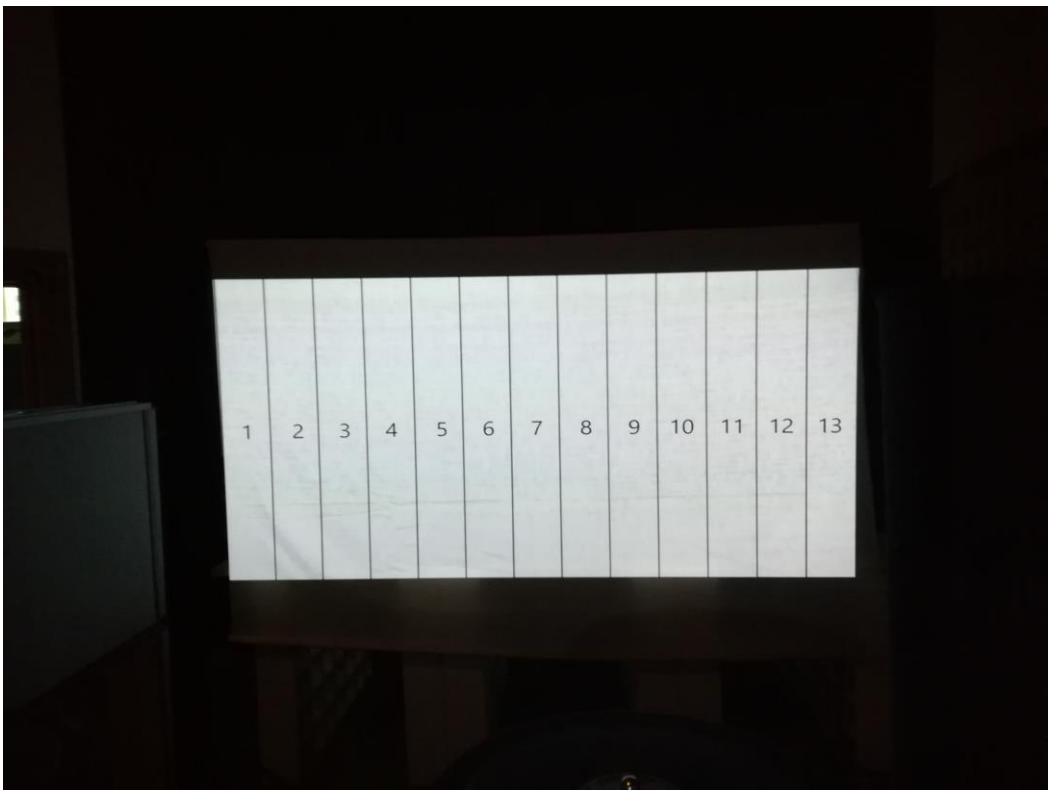


Figure A.2: Projected grids

Figure A.3: General view of experiment area

APPENDIX B

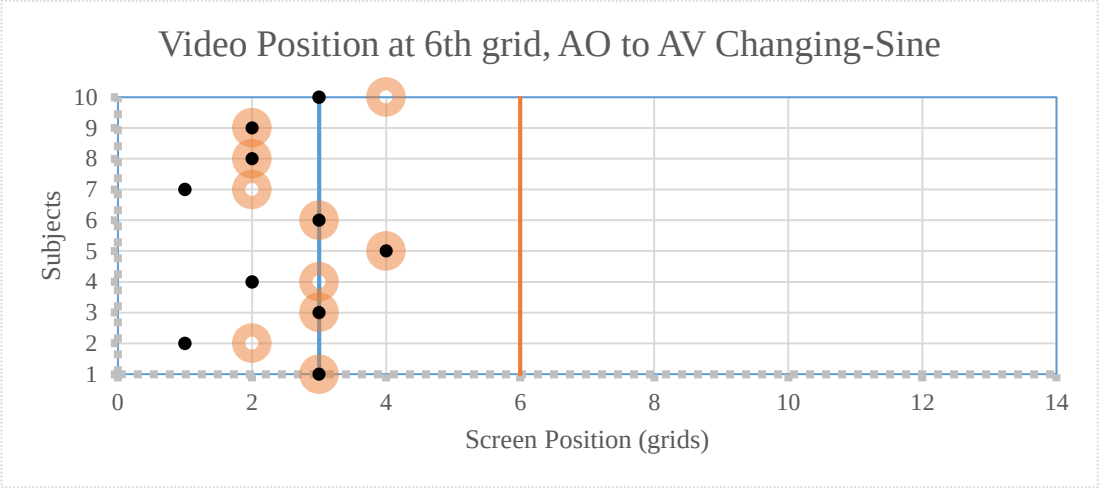


Figure B.1: AO and AV estimations for sine wave (visual at 6th grid)

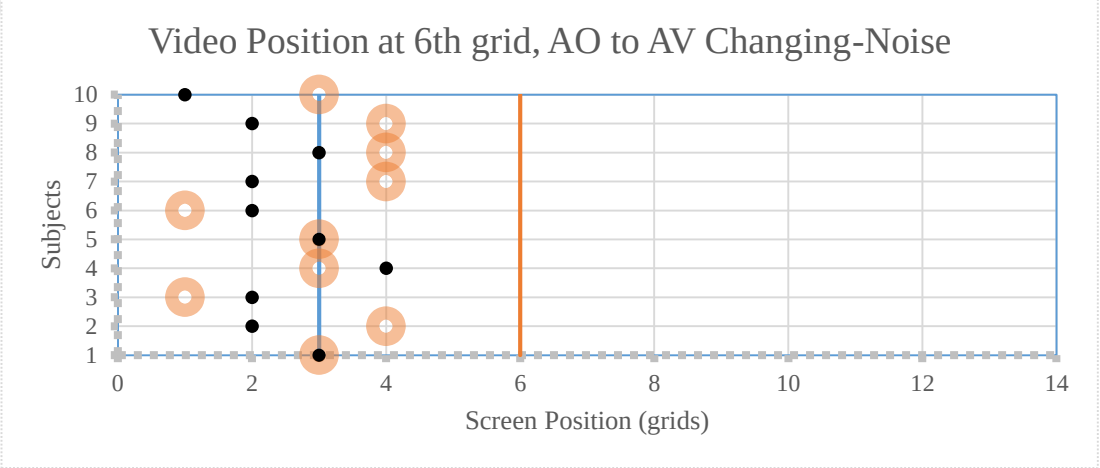


Figure B.2: AO and AV estimations for noise signal (visual at 6th grid)

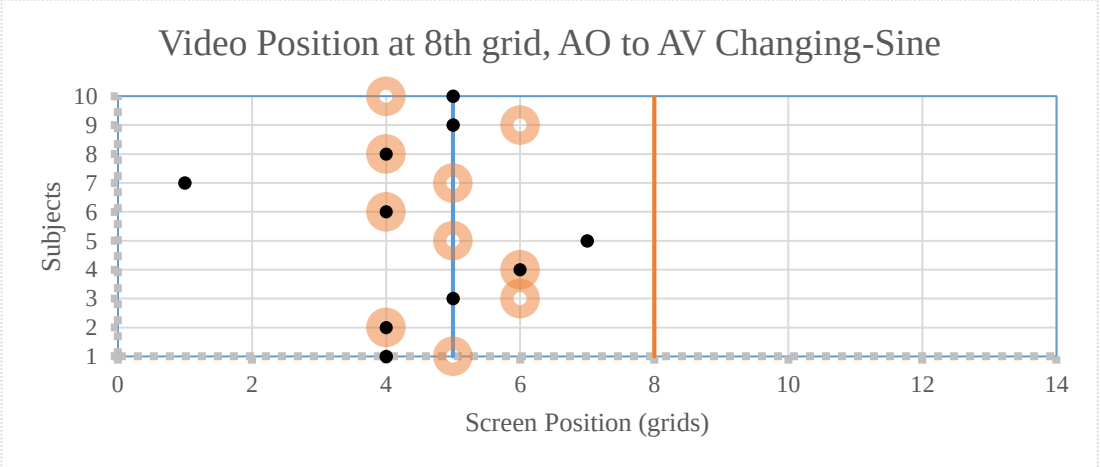


Figure B.3: AO and AV estimations for sine wave (visual at 8th grid)

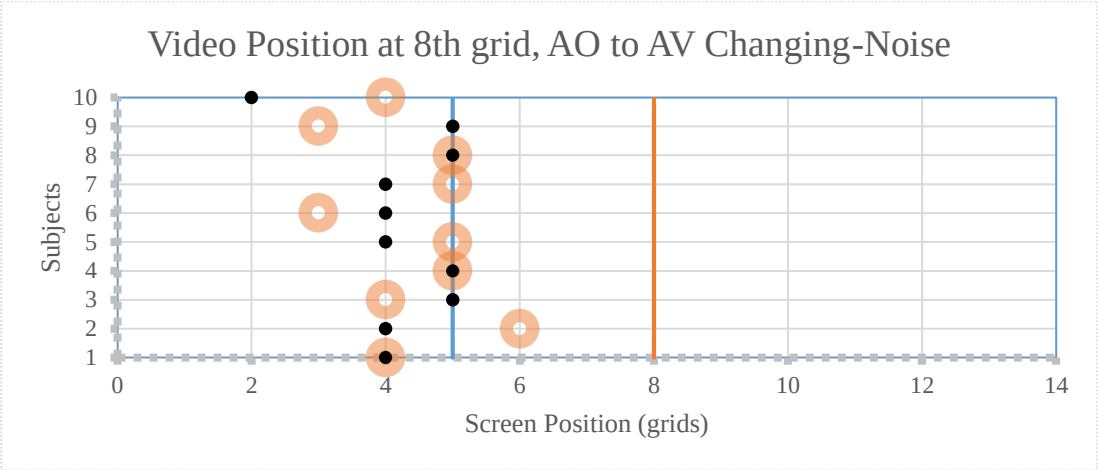


Figure B.4: AO and AV estimations for noise signal (visual at 8th grid)

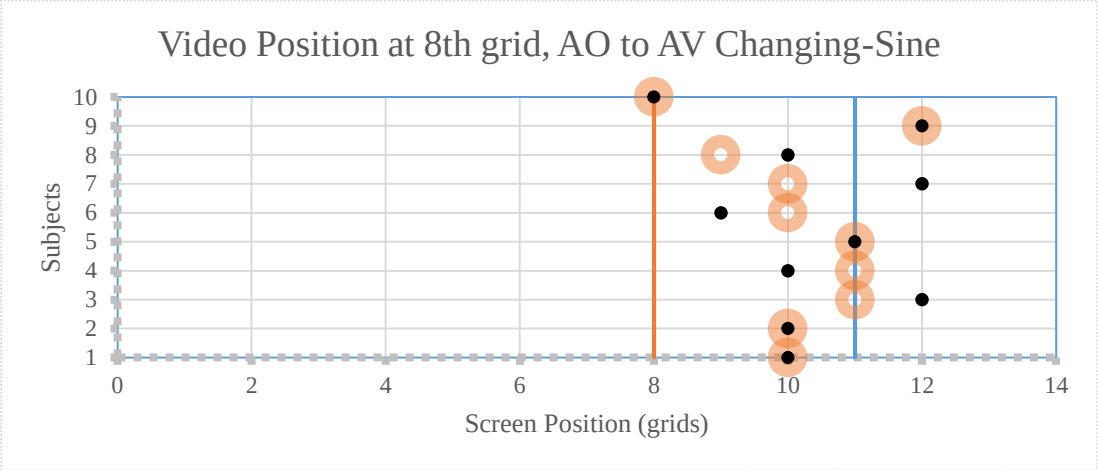


Figure B.5: AO and AV estimations for sine wave (visual at 8th grid)

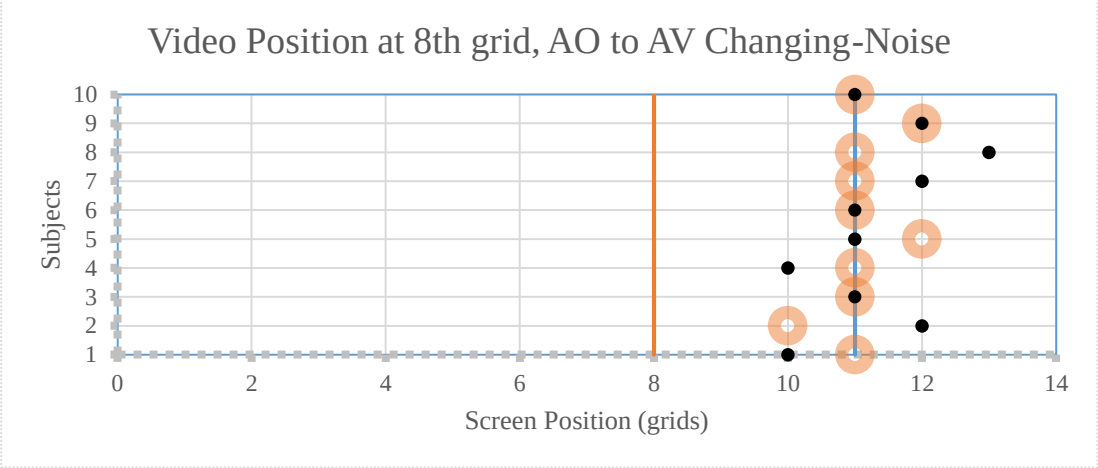


Figure B.6: AO and AV estimations for noise signal (visual at 8th grid)

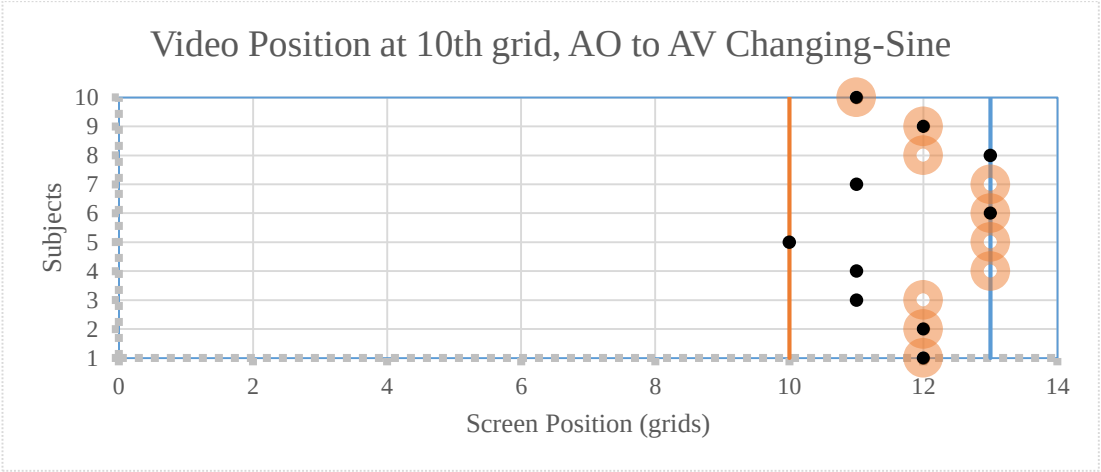


Figure B.7: AO and AV estimations for sine wave (visual at 10th grid)

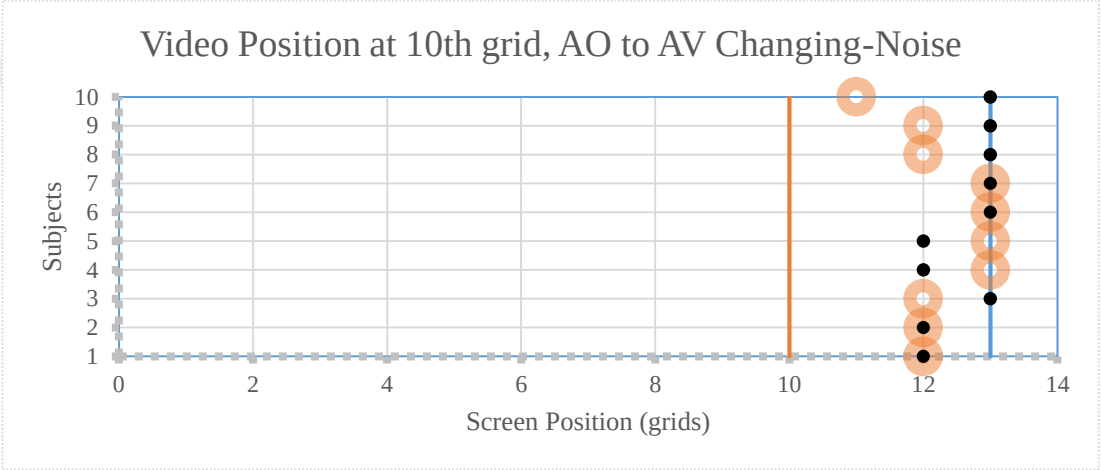


Figure B.8: AO and AV estimations for noise signal (visual at 10th grid)

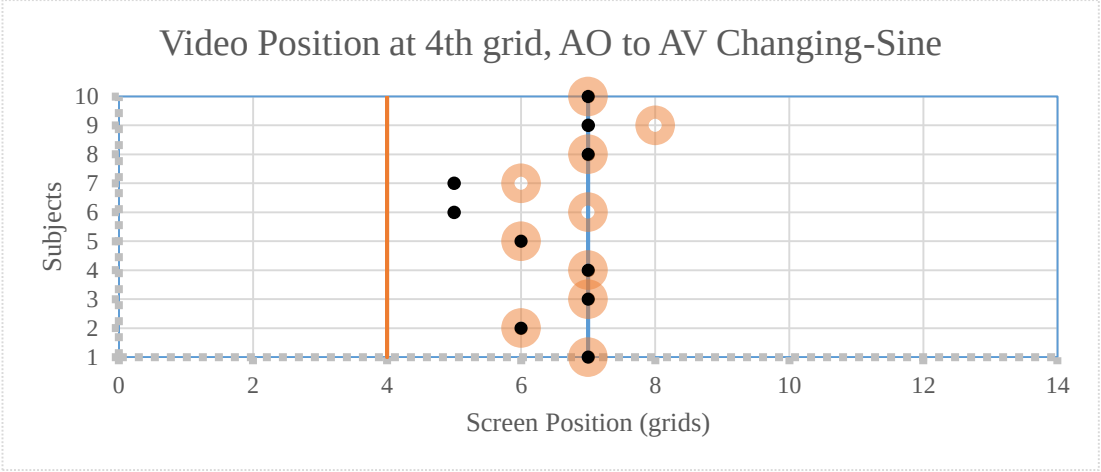


Figure B.9: AO and AV estimations for sine wave (visual at 4th grid)

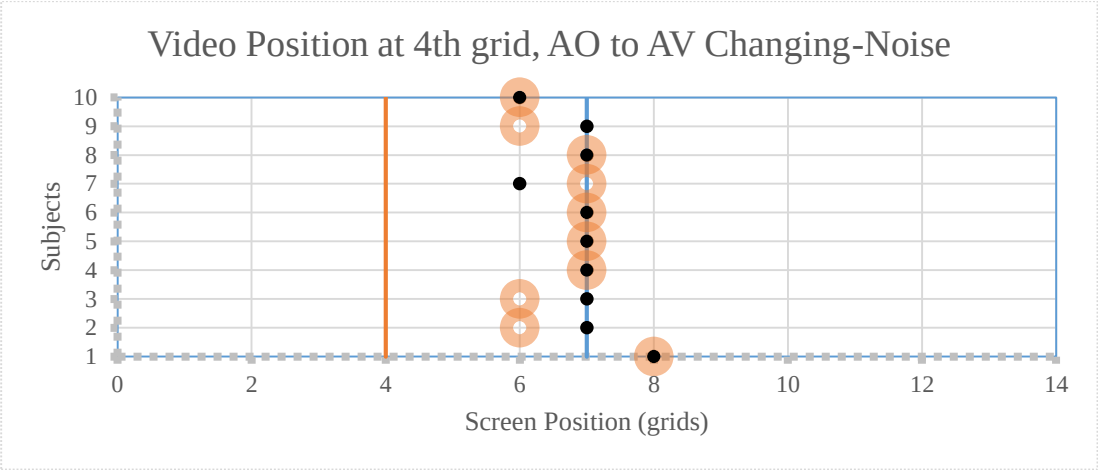


Figure B.10: AO and AV estimations for noise signal (visual at 4th grid)

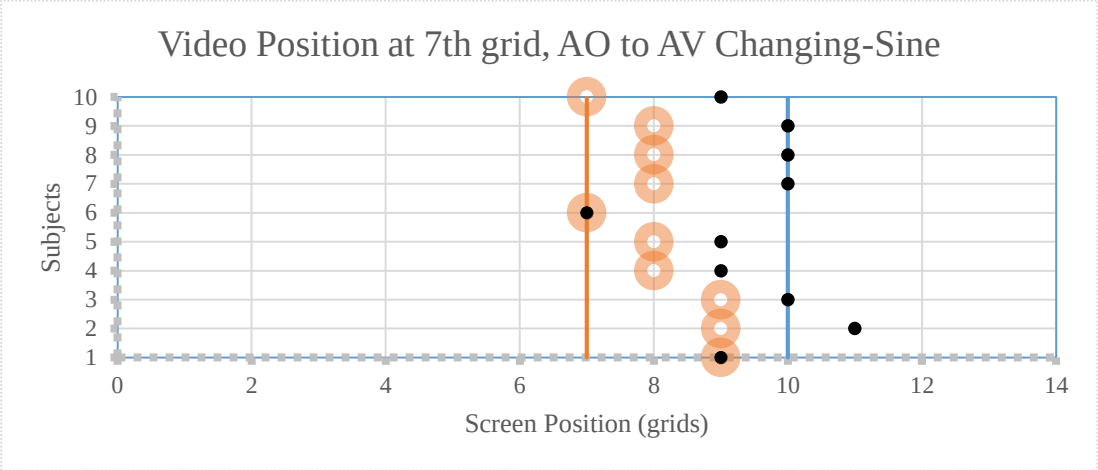


Figure B.11: AO and AV estimations for sine wave (visual at 7th grid)

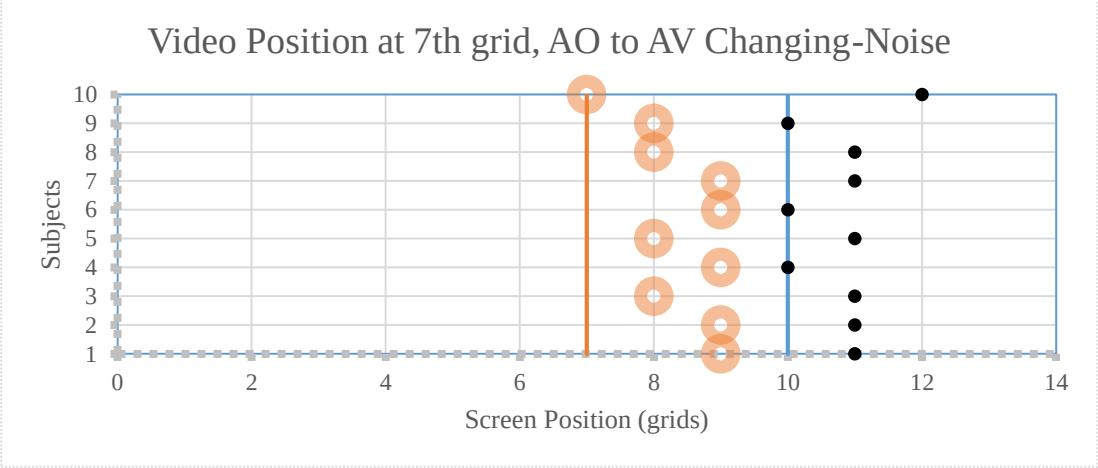


Figure B.12: AO and AV estimations for noise signal (visual at 7th grid)

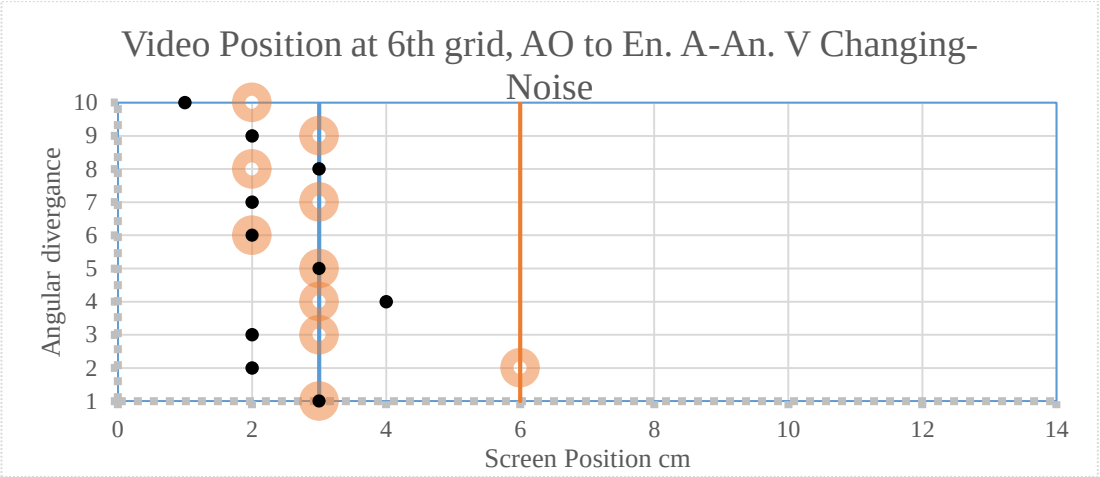


Figure B.13: AO and En. A-An. V estimations for noise signal (visual at 7th grid)

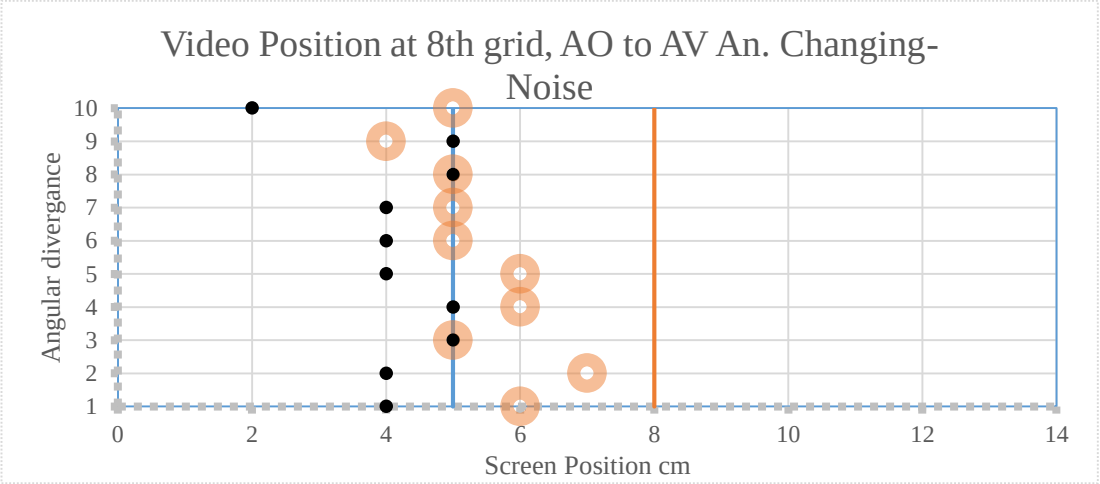


Figure B.14: AO and En. A-An. V estimations for noise signal (visual at 8th grid)

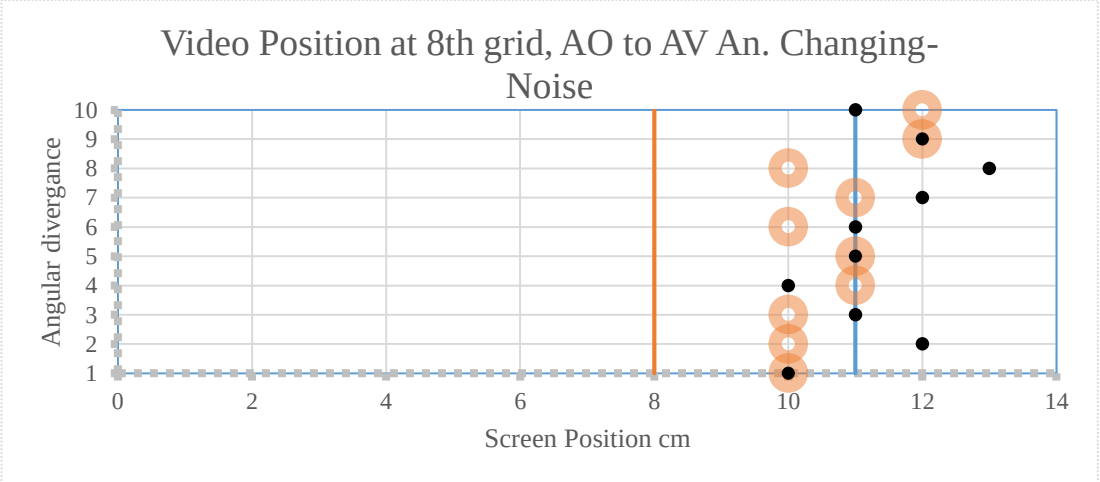


Figure B.15: AO and En. A-An. V estimations for noise signal (visual at 8th grid)

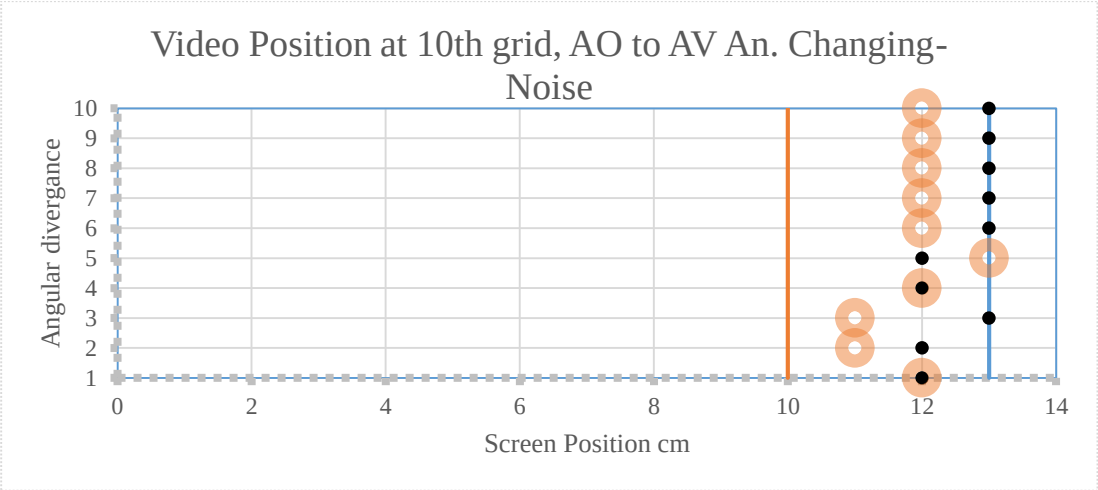


Figure B.16: AO and En. A-An. V estimations for noise signal (visual at 10th grid)

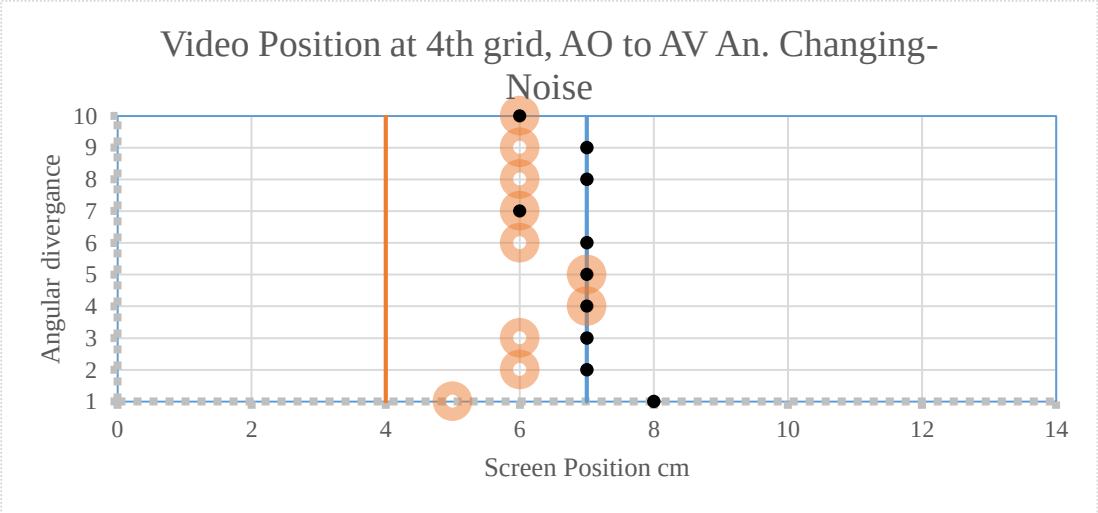


Figure B.17: AO and En. A-An. V estimations for noise signal (visual at 4th grid)

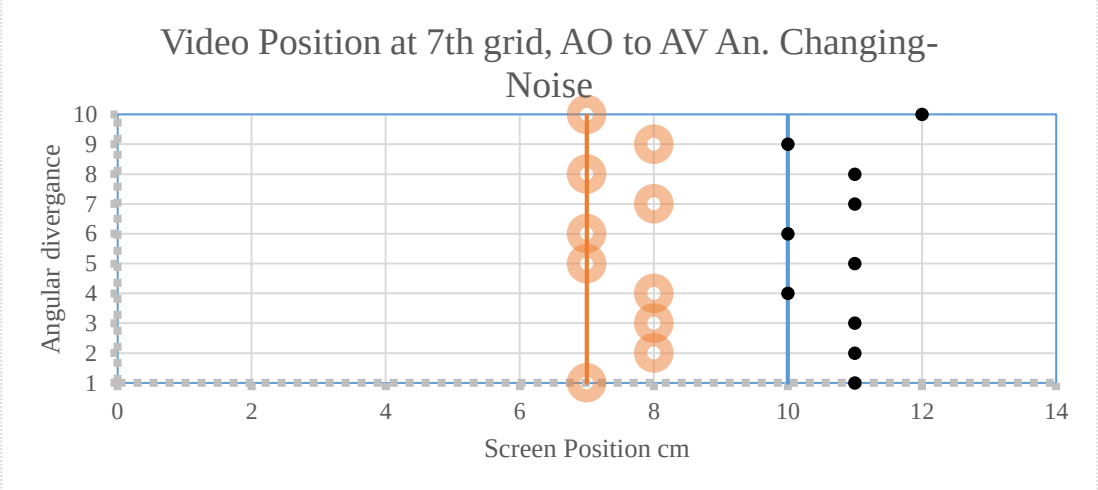
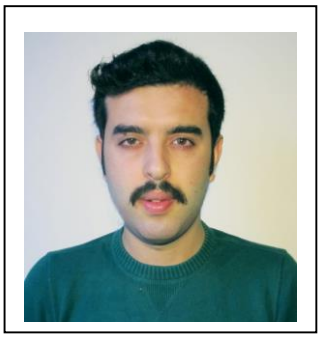


Figure B.18: AO and En. A-An. V estimations for noise signal (visual at 7th grid)

CURRICULUM VITAE



Name Surname **:** Naci Tepedelen

Place and Date of Birth **:** İstanbul-22.01.1990

E-Mail **:** Nacitepedelen@gmail.com

EDUCATION **:**

- **B.Sc.** **:** 2013, ITU, Electrical Engineering