ABSTRACT

Title of Thesis:	TESTING THE McGURK EFFECT WITH 3D SOUND
	Siddharth Bhagwan, Master of Science 2018
Thesis Directed By:	Associate Professor, Jennifer Golbeck, College of Information Studies

Unlike regular mono or stereo sound, 3D sound consists of an embedded omnidirectional spatial component. This makes it a lot more immersive than regular sound and can have various applications. This thesis studies an existing psychological phenomenon namely the McGurk effect while specifically replacing the auditory component with 3D sound. The data collected from the study is translated and analysed for statistical significance and the results are then interpreted in accordance with the design and the limitations of the study.

TESTING THE McGURK EFFECT WITH 3D SOUND

by

Siddharth Anil Bhagwan

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Science in Human
Computer Interaction
2018

Advisory Committee:
Dr. Jennifer Golbeck, Chair
Dr. Yla Tausczik
Dr. Beth St. Jean

© Copyright by
Siddharth Anil Bhagwan
2018

Acknowledgments

The completion of this study would not have been possible without the assistance of many people whose names may not be enumerated. Their contributions are sincerely appreciated and duly acknowledged. Nonetheless, I'd like to mention a few that I am thankful to, without whom this study might not have turned out the way it did.

First and foremost I'd like to thank the committee chair, Jennifer Golbeck, for her unwavering support and guidance throughout this project. Her positivity and doer attitude has been extremely inspiring.

Special thanks to Professor Kent Norman for his interesting insights, for constantly allowing me to bounce ideas off him and always being full of enthusiasm.

I'd also like to thank Yitzhak Paul and Preston Tobery from the John and Stella Graves MakerSpace at the McKeldin Library for their patience and guidance through the numerous iterations that led to the final version of the binaural mike.

This section would be incomplete without mentioning Raghavendra Karkera, for his unstinted support and help in recording the videos used in this study.

And finally, I thank Prianka Waghray, for believing in me.

Table of Contents

Chapter 1: Introduction

Sound as a part of technology is now ubiquitous both with and without accompanying media like video. Its uses are numerous leaving no domain untouched, be it gaming, movies, music, education, AI or even education. With progress in sound technology and the advent of 3D sound, one of the impending paradigm shifts could be in the way we consume sound.

This study is an extension of the original study on the McGurk effect from 1976, named 'Hearing lips and seeing voices' as published in the Nature Magazine, vol 264 by Harry McGurk and John Macdonald. The McGurk effect occurs when the auditory component of one sound is paired with the visual component of another sound, leading to the perception of a third sound.

Compared to regular mono or stereo sound, the audio in 3D sound consists of an embedded spatial element. This study is an attempt to test if this added dimension leads to higher immersion compared to mono sound. Participants are subjected to the already studied and well understood McGurk effect in 3D and mono variants for sound and the two distribution of responses are compared. The results show no significant difference in distribution for the study.

Chapter 2: Literature Review

What is 3D sound?

3D sound, also known as spatial sound is sound as we hear it naturally[1]. The three-dimensional aspect is due to the spatial location embedded within the sound. Pitch, tone loudness, and location - all are intrinsically preserved in 3D sound[1]. Examples from daily life include being able to hear the sound of a door being shut on the floor below or being able to listen to a conversation happening behind you. In both these scenarios, we are aware of where the sound is originating from without having to look for it specifically.

Comparison of today's sound technologies

Sound today takes the form of mono, stereo and surround for the most part. Mono sound consists, as the name suggests, of a single channel. Listening to mono audio on headphones, it can be noticed that whatever is heard in the right earbud will be heard in the left earbud. That is because the headphones are playing back the same single channel audio file into both earbuds[2]. Though not present commonly in commercially produced music, it is still used while recording sound in some cases, for example in the microphones of smartphones and in the broadcast of talk-only radio shows for reasons of a higher coverage area. Use cases where multiple channels are redundant often use mono sound. It is noteworthy that playing mono sound on multiple speakers is still mono sound. Each speaker is still playing the same copy of the sound.

Stereo sound, an upgrade to mono, is the most commonly used sound technology today, which consists of two channels while recording. Common examples of these are .mp3 tracks on cell phones, wherein one can hear the drums on the left ear and the guitar on the right. Any spatial element included while recording will have to be downgraded and the end result will always be a mix of 2 channels. Having two channels provides for mixing of audio within the left and right channels and can convey a sense of limited spatial audio[3] i.e., space and sound location can be detected only on the plane of the ear to the left and the right, but not above, below, ahead or behind.

Next in the spectrum is surround sound. It is basically engineers mixing multiple mono and/or stereo sounds via a computer program and creating a mix that can be played on multiple speakers with each speaker designed to accentuate a particular part of the sound. Examples are 5.1 (5 speakers/channels and 1 subwoofers) and similarly 7.1 systems. Companies like Dolby and DTS specialise in such technology and speaker manufacturing companies acquire a license for it. These are most prominently seen in home theater systems and cinemas.

Binaural sound differs from all the above technologies in the way it is recorded. It is captured in the way we hear audio in the real world. The spatial element is captured due to the manner in which the microphones receive the sound. The person hearing 3D sound hears it just as the person recording it hears it from the source[4]. To record binaural sound, multiple ear shaped microphones are used, that help in modeling sound exactly as it would in the real world. Variants of such microphones include a mannequin head in between the two ears and such structures with multiple ears as shown below in Image 1.

Image 1. A professional binaural microphone

Applications of 3D sound

Entertainment: Binaural audio has great potential in the entertainment industry. For instance, 3D sound has been rapidly gaining popularity in the gaming industry. As an example, consider a first person shooter game, wherein someone is walking behind the player. By virtue of 3D sound, the player is informed of the existence of another character in a direction, the proximity, and manner of the approach i.e., is the character running towards you or walking etc., without even looking. This is immense value addition and an upgrade to the gaming experience overall. Add to this the ability to track head movements, such that if a sound is emanating on the left, on turning the head to look left, the sound now emanates from in front of the player. This concept can just as easily be extended to VR gaming and VR systems in general.

Another example is listening to audiobooks. Since the experience of being spoken to closely in one ear can be replicated, audiobooks can now position the narrator differently for different characters. Similarly, watching concerts and movies and any other recorded media can now have enhanced experiences if the directionality plays a role in the given context and circumstance.

Audio Therapy: The Autonomous Sensory Meridian Response (ASMR) community has gained popularity all over the Internet by using binaural recordings to trigger physical responses that they believe can be soothing and calming. These could range anywhere from whispers directed to an ear(s), massage sounds, sounds in nature, etc. and are believed by some to be relaxing and therapeutic. Many have taken to public forums to explain their ability to induce ASMR to ease symptoms of conditions like depression, pain and chronic pain in cases where other routes of treatment may have been lacking or ineffective[5].

Aid for visual impairment: 3D sound can compensate for visual impairment to a certain extent. For instance, consider a visually challenged person being able to navigate towards the nearest exit in case of an emergency by being able to follow the direction in which the sound is being generated from. Depending on the extent of the impairment, the dependence on spatial component can vary between being an auxiliary sensory input to being the prime, as far as direction is concerned.

The McGurk effect

This study is an extension of the original study on the McGurk effect from 1976, named 'Hearing lips and seeing voices' (McGurk, 1976) as published in the Nature Magazine, vol 264 by Harry McGurk and John Macdonald[6]. The paper goes on to state that wherein normally, speech recognition is considered a purely auditory process, there is indeed an impact of vision as well, which was not recognized until then. The study reported that on seeing a woman on film say the syllable 'ba' repeatedly while having the sound dubbed to 'ga', normal adults reported hearing 'da'. In the absence of visual stimuli, the subjects reported accurately what the un-dubbed audio was relaying in both cases i.e, 'ba'/'ga'. The study conclusively shows how on an average, a majority of the responses are either fused or combined and of the remainder, the visual stimulus tends to override the auditory stimulus. A summary of the results of a few stimuli is shown below in Image 2.

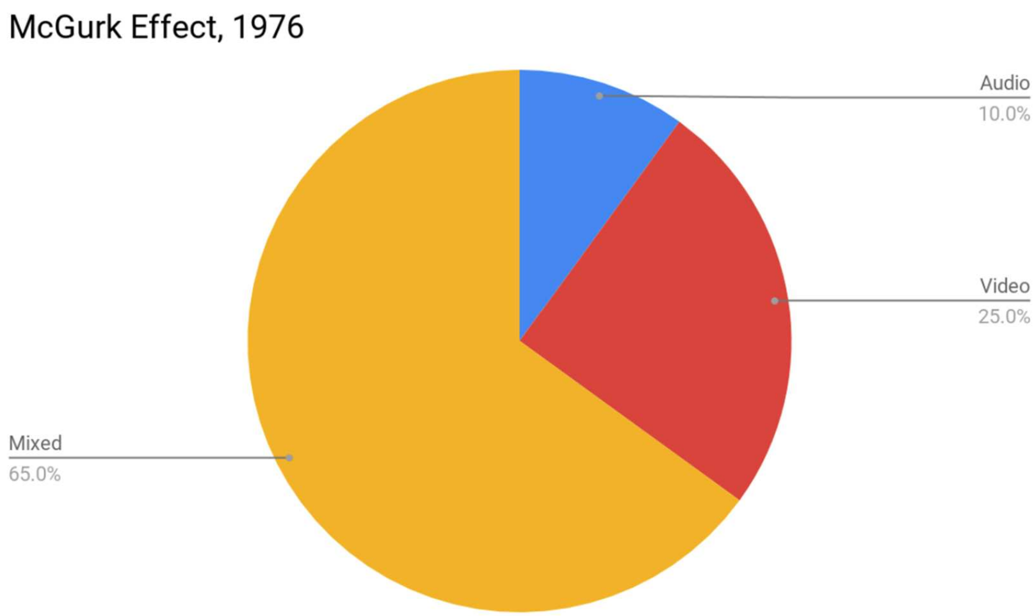


Image 2. Distribution of the McGurk effect, 1976

A formal definition of the McGurk effect is as follows: "The McGurk effect is a perceptual phenomenon that demonstrates an interaction between hearing and vision in speech perception. The illusion occurs when the auditory component of one sound is paired with the visual component of another sound, leading to the perception of a third sound".[9]

The data cleaning section of Chapter 4 goes into further detail about the meaning of each of the below mentioned categories.

Given Stimuli		No. of Adult Subjects	Total percentage of responses per category			
Auditory	**Visual**		**Auditory**	**Visual**	**Combined/ Fused**	**Other**
ga - ga	ba - ba	54	11%	31%	54%	4%
pa - pa	ka - ka	54	6%	7%	81%	6%
ka - ka	pa - pa	54	13%	37%	44%	6%

Table 1. Part summary of the original study on the McGurk effect (McGurk, 1976)[6]

The table shows the different sets of stimuli given to the subjects, and what percentage reported what stimulus. 'Auditory' refers to those participants who reported the audio component of the stimulus. Similarly, those reporting what the lip movement was conveying belong to the 'Visual' section. 'Mixed' refers to those who either reported alternating or some variant of both the auditory and visual stimuli, or an entirely new stimulus which was originally absent, formed by fusing the two.

Chapter 3: Methodology

In this chapter, we shall explore the setup used to conduct the study, specifications of the equipment used, define eligibility for the study, recruitment methods and discuss the procedure followed in detail. The last section of the study deals with the design decisions made for some aspects of the study after conducting a pilot study with a few participants.

Setup

The study is conducted inside a closed room, consisting of two chairs, a table, a laptop and a pair of over the ear headphones. There are two video clips to be shown to each participant. Each clip consists of the neck and above profile of a man looking into the camera and saying the phrase ‘va-va-va’, thrice with about a 0.5 ms gap in between each set. The audio for this video has been replaced with the phrase ‘pa-pa-pa’. The audio has been recorded in two flavors. The first is regular mono sound playing via two channels (left and right), and the second is 3D sound. The audio has been recorded in sync with the video, such that there is no gain or lag with respect to the lip movements in the video. A screenshot (Image 3) of the video has been posted below.

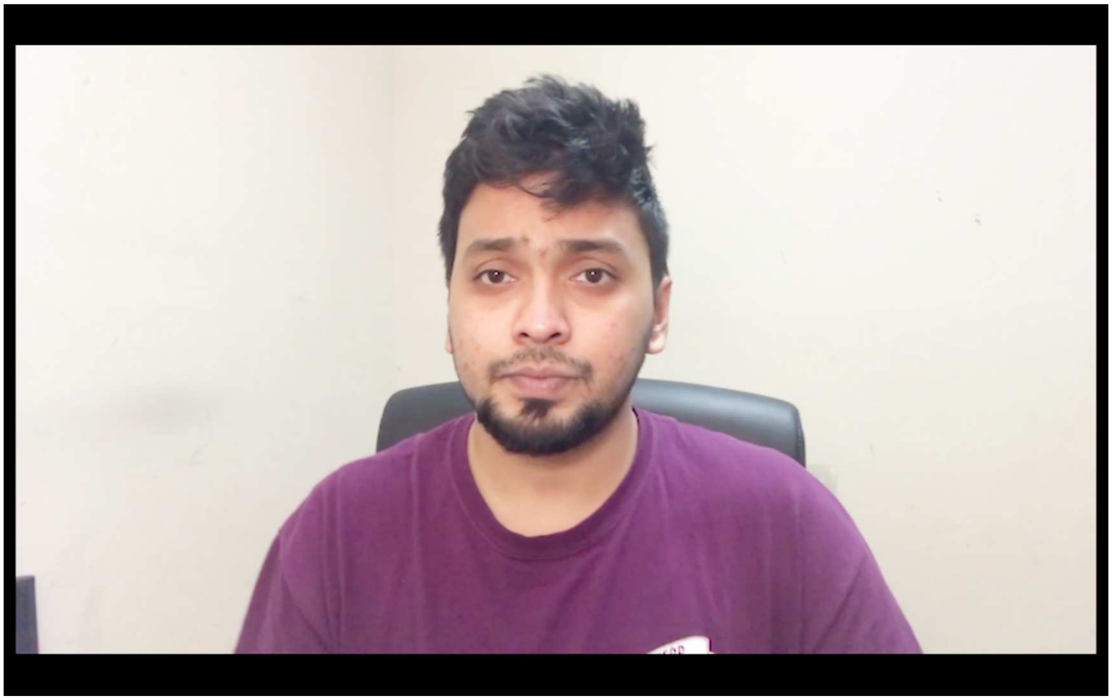


Image 3. Screenshot of the video clip shown to the participants

Equipment

- The video has been recorded using an iPhone 7 Red Special Edition
- The audio has been recorded using two RockDaMic Professional Lavalier Mics
- 3D sound was recorded using a 3D printed binaural mic, the sketch for which was borrowed from Thingiverse. The printer used was a Makerbot Replicator 2 and Zortrax m200. Flexible filament was used for the ears and ABS plastic was used for the remaining body of the mike
- Credits for the sketch go to the user Jonny, whose sketch was borrowed from Thingiverse (thing number 499001), modified and used[8]
- The audio replacement was done using Adobe Premiere Pro CC

- The videos were played on VLC player on a 15'' MacBook Pro 2013 running macOS Sierra
- The questionnaire given to the participants after each video was presented on Google's Chrome browser and created via Google Forms
- The headphones used were Sony's MDR 7506
- The sound card was Vantec NBA-200U USB External 7.1 Channel Audio Adapter

Below are a few pictures of the binaural microphone (Image 4, Image 5 & Image 6)

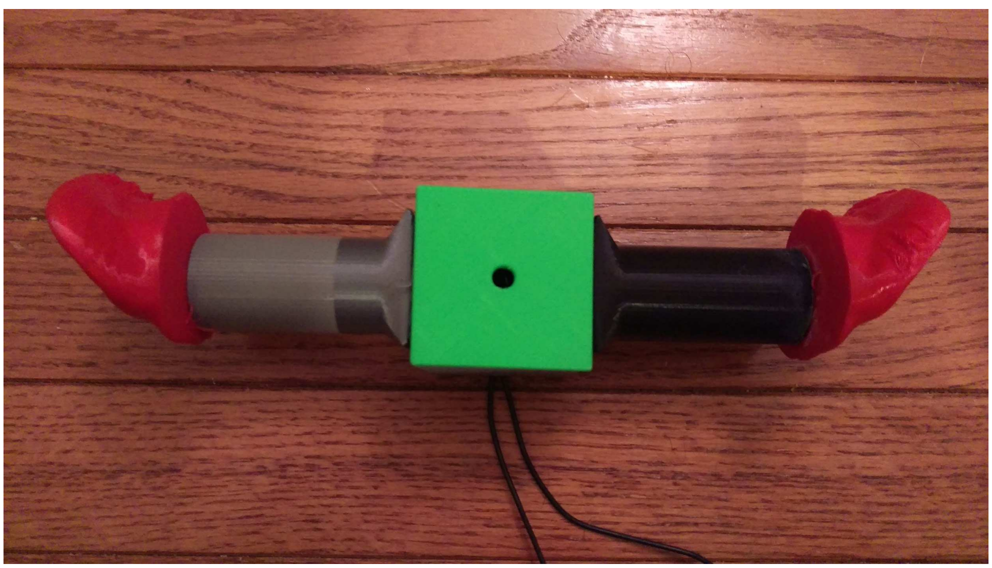


Image 4. Top view of the binaural microphone

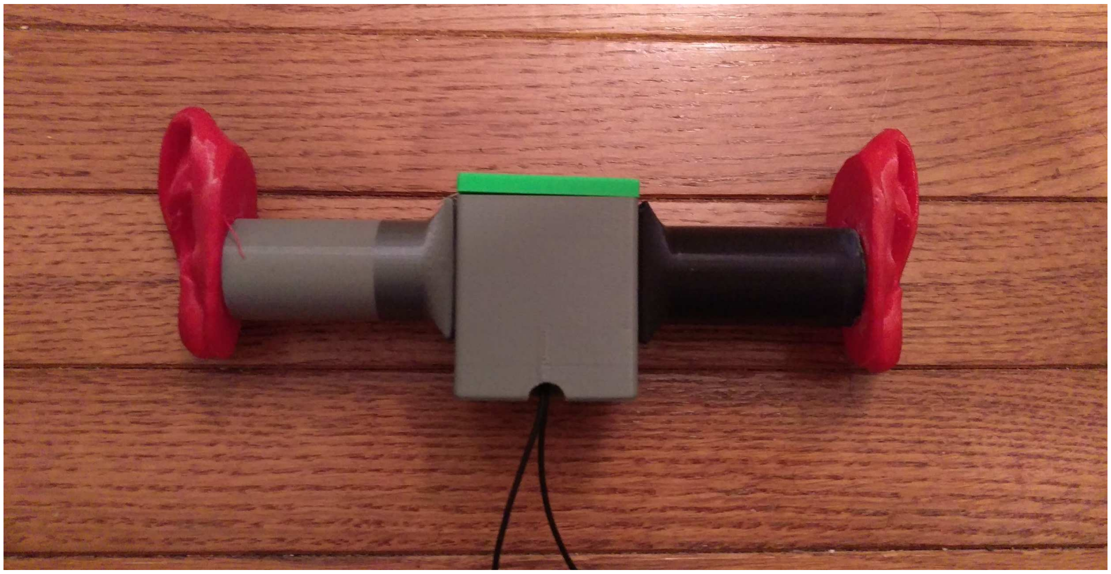


Image 5. Front view of the binaural microphone

Image 6. Side view of the binaural microphone

Eligibility & Recruitment

This study does not have any special requirements. Adults with normal vision and hearing are eligible to be a part of the study. This includes people who are using any kind of aid like contact lenses or prescription spectacles for vision corrections and/or hearing aid. As long as the end result is clear hearing and vision, the person is eligible.

Recruitment was done via convenience sampling. Most of the participants were other students from the program, the other programs in the department and some others who heard of the study via the word of mouth. Emails were also sent out via the program's (Masters of Human Computer Interaction) listserv notifying the students of the University of Maryland's iSchool (College of Information Studies) of the details of the study.

Procedure

Using the following procedure, a pilot study was conducted with 8 participants. Insights from this study were used to fine tune the procedure. and the data from the pilot was discarded and not considered while statistically analysing the data.

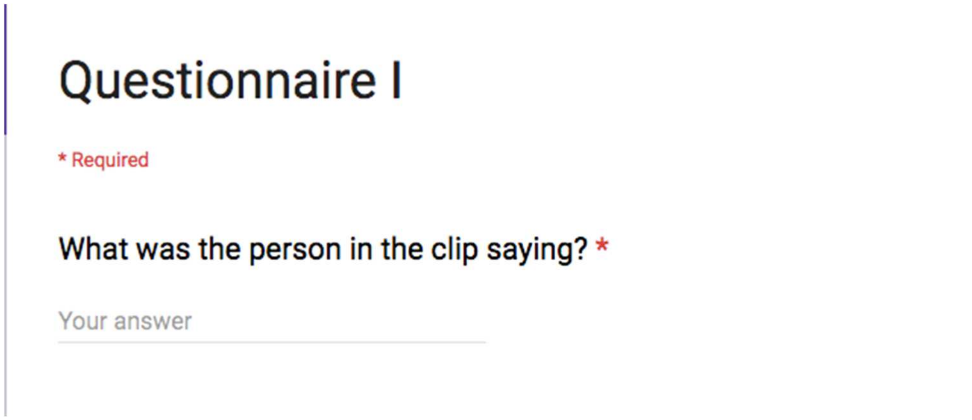
The process starts by greeting the participants and entertaining them one at a time. After they're thanked for their time, they are then informed that there is no known harm to them as a part of the study and no personal and/or identifying data is being collected during the process. For the most part, the procedure includes watching video clips and answering questions based on them. The data that is being collected would be stored securely on a password protected laptop and password protected Google Drive, and shall be deleted 6 months after the completion of the study. They are also informed that at any time through the study if they are not comfortable, they are not obligated to complete the study and are free to walk out.

Once they have signed the consent form, the process is described to them in more detail, letting them know that they will now watch two separate video clips. The clips would be short and would require immediate attention on their behalf. The clips shall only be played once. After each video clip, they shall be asked to fill out a survey containing a single, open ended question, "What was the person in the clip saying?", which they are free to answer in any manner and any number of words they see fit. The procedure would repeat for the 2nd video clip, after which they shall have to answer a closing questionnaire, with two "Yes/No" questions. That would mark the end of the study.

They are then handed over the headphones and the play button is pressed whenever they are ready. Once the video completes the browser is switched to, where the questionnaire is already open. At this point, the screen is positioned such that it is only visible to the participants to ensure an unbiased, un-affected answer. Once they are done, VLC media player is switched to and they are asked to wear the headphones in case they have been removed. Once the they signal they're ready, the video is played. Again as before, the video is paused on completion, and the browser containing the tab with the second questionnaire is switched to. As before, the screen remains visible to only the participants, and once the second questionnaire is submitted, the closing questionnaire is opened.

The closing questionnaire asks the participants if they were aware or had any knowledge of the McGurk effect as a 'Yes/No' question. It is important that this question be asked towards the end, after the completion of the study. Even for those who had no knowledge of the phenomenon, asking this question before would alert them, at the very least, that they might be tested for or against some similar effect, and that might make them more conscious and skew the results of the study. By asking the question towards the end, the result can be discarded if the answer to the question is 'Yes'.

A screenshot of the first questionnaire (which is identical to the second one) and the closing questionnaire are presented below (Image 7 & Image 8)



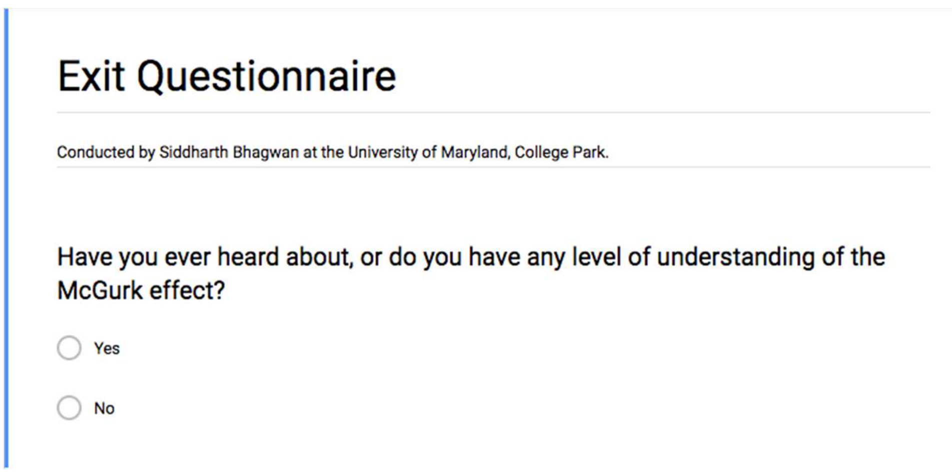
Questionnaire I

* Required

What was the person in the clip saying? *

Your answer

Image 7. Questionnaire I



Exit Questionnaire

Conducted by Siddharth Bhagwan at the University of Maryland, College Park.

Have you ever heard about, or do you have any level of understanding of the McGurk effect?

- Yes
- No

Image 8. Exit Questionnaire

Revisiting the design decisions of the study

Open ended vs Multiple Choice: The initial design of the survey consisted of giving the participants multiple choices and having them pick one of the options. The options were as shown (Image 9):

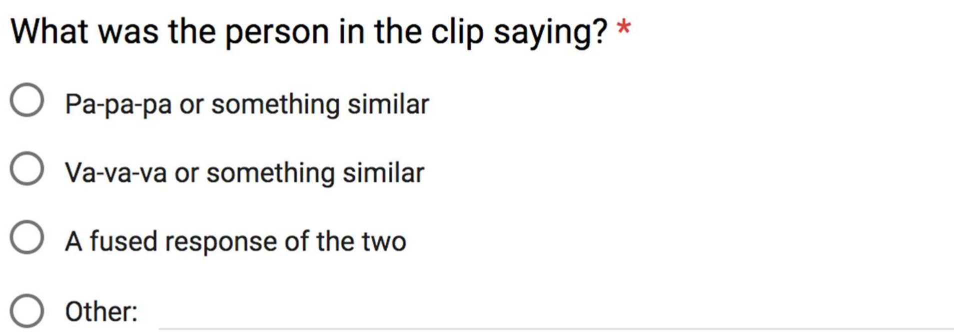


Image 9. Questionnaire I (Initial Design)

As a result, a few concerns were common amongst most of the participants:

- ‘I think I heard ‘Ba-ba-ba’ but it must have been ‘Pa’
- I was sure before seeing the options, but I am confused now, can I watch the videos again?
- Does my option have to be along the lines of the given options?
- Is ‘Va’ different from ‘Wah’? Does the detail matter?

Eventually, a majority of them ended up using the other option, even when their answer was very close to one of the options. Additionally, instead of selecting option three, ‘A fused response of the two’, the participants preferred specifying what exactly was the fused response they thought they heard.

Order of the videos: Inevitably, once the participants got familiar with the routine after the first video and questionnaire, they scrutinized the second video and were a lot more focused on what they saw and heard. It was difficult to undo this effect, so instead, the order was alternated for each participant, such that on the whole, half of the participants got the mono sound first and the other half got the 3D sound first.

Using two different videos: Initially, the video used for the mono sound and 3D sound was the same. This led to a few participants asking if they saw the same video twice. This was usually asked after they filled in the questionnaire and when checked, the responses were identical. For this reason, the videos used for the two sounds were different. They were shot back to back and are identical for the most part, but aren’t exactly the same, which seemed to quell most of the participants’ doubt. Using entirely different videos with different backgrounds seemed like over engineering and intentionally leading the participants and was avoided.

Using a single set of stimuli: Inspired by the original study, two sets of stimuli were created initially i.e., ‘va’ video - ‘pa’ audio and ‘ba’ video - ‘fa’ audio. Pilot testing revealed that the latter stimulus was somehow clear enough that most participants responded, while not necessarily understanding what was happening, that the video seemed like it was saying something different and the audio was saying something different. The study was then modified to include just the one stimulus, wherein the audio - video stimuli were not disparate enough for the study to not work and at the same time dissimilar enough for the participants to have a certain level of uncertainty.

Chapter 4: Findings

In this chapter, we start with the cleaning of the data i.e., what were the methods used followed by a brief explanation of each category, their differences and justifications for the existence of each category. We then delve into the statistics used on the cleaned data. The interpretation of the statistics is explored in the next chapter.

Data Cleaning

The data collected via the forms was transferred to Google sheets using Google Forms' inbuilt functionality, a sample of which is shown below (Image 10). Since the question was open ended, it needed to be cleaned and translated into a more usable form. The translation technique and the corresponding keys used are explained below, after which a code is assigned to each entry in column C.

	A	B	C
1	Timestamp	What was the person in the clip saying?	
2	3/8/2018 15:58:06	papapa	A
3	3/8/2018 16:13:08	Very similar to the first one, pah wah wah and then pah pah pah	V
4	3/8/2018 16:23:20	Pa Pa Pa Pa Va Pa Pa Pa Pa	A
5	3/8/2018 16:28:27	Pa Wa Wa x3 "Pa pa pa" x1	V
6	3/8/2018 17:05:19	"ba-ba-ba" repeated 3x again (first "ba" of first set had a bit of a "w" after	M
7	3/8/2018 17:29:43	bah wah bah	M
8	3/8/2018 17:40:33	Pa	A
9	3/8/2018 17:47:08	waa waa waa phaa phaa phaa baa baa baa	M
10	3/8/2018 17:59:02	wa and pA	M
11	3/8/2018 18:02:21	WAH WAH	V
12	3/8/2018 18:05:28	BAR	A
13	3/8/2018 18:09:11	BAH	A
14	3/8/2018 18:13:10	imitate a kind of animal like frog....	I
15	3/8/2018 18:16:58	Par Par Par	A
16	3/8/2018 18:20:16	He's saying "Dad, Throw, Fish" repeatedly in Thai	I
17	3/8/2018 18:23:06	pa pa pa	A
18	3/8/2018 18:33:18	Wa wa wa	V
19	3/26/2018 15:48:02	Wah	V
20	3/26/2018 15:55:39	bar bar bar	A
21	3/27/2018 16:01:47	Wah Wah Wah x3	V
22	3/27/2018 16:04:39	bah bah bah	A

Image 10. Screenshot of a section of the translated responses

Audio (A): Participants who heard the audio channel are marked so. Since the audio part was saying ‘pa-pa-pa’, answers that were ‘pa’ or variants of it like ‘paw’, ‘pah’, ‘par’ etc. were marked A. The distinguishing factor from what the video was saying i.e, ‘va-va-va’ is primarily the lip movements of ‘va’ vs ‘pa’. Keeping that in mind, responses that were on the lines of ‘ba’, ‘bah’, ‘bar’, ‘baw’ etc. were also considered equivalents of those as the variants of ‘pa’ as the sounds are very close to each other and the lip movements for both would be the same.

Video (V): Participants who reported the video channel as what they heard were marked V. This would mean any variant of 'va', 'why', 'wow', 'wah' etc. In cases where participants listed out all nine syllables and had a mixture, if the mixture was more than two-thirds of a given syllable, it was marked that particular syllable. For instance, 'wa wa wa, ba wa wa, ba wa wa' would be marked as video. Similar to the audio, approximations are made in this case as well i.e. sounds like 'va' would have extremely similar lip movements and sound as that of 'va' and are thus marked as video.

Mixed (M): Mixed stands for both fused responses, as well as other responses that do not fall under audio, video or invalid. An example of mixed would be using some combination of the auditory and visual stimulus, for instance, 'pa' and 'va' would mix to form 'pa-va-pa' or 'bah-wah-wah'. This is different from fused, where the result isn't any individual stimuli or a mixed version, but fused responses that result in the creation of an entirely new syllable in this case. For example, 'pa' and 'va' fuse to form 'pwa'. While this distinction might seem banal or even analogous, it is an important distinction in itself. While it does not affect the results of this study, understanding the difference is assuredly within the scope of the study.

Mixed responses relate to the participant being able to clearly distinguish between two different stimuli, and debating over the order, which in other terms can be described as the percentage occurrence of each stimulus. Compare this to the fused response, where the intensity of video and audio stimuli are equally strong, leading the participant to not pick one over the other, but forcing him to fuse them altogether. Fused responses are not very common, but they are more evenly distributed wherever they do occur, in comparison to mixed responses. Thus, a mixed response is likely to be less uniform compared to a fused one. For example, ‘ba-wa-wa’ vs ‘bwa-bwa-bwa’.

Invalid (I): The questionnaire being open ended, there were certain responses that had to be discarded as they were not answering the question even after approximation. Despite asking the participants to try and respond to the question to the best of their abilities, a few such cases were expected owing to subjective interpretations of both the video as well as the questionnaires. One such example that was marked invalid was ‘repeat a kind of language I don’t know’.

After marking all the responses, we had 3 invalid responses.

Sample size after cleaning (n) :

45 Total responses - 3 Invalid responses - 3 responses that knew about McGurk effect = 39

Post coding, the intermediate results are as follows:

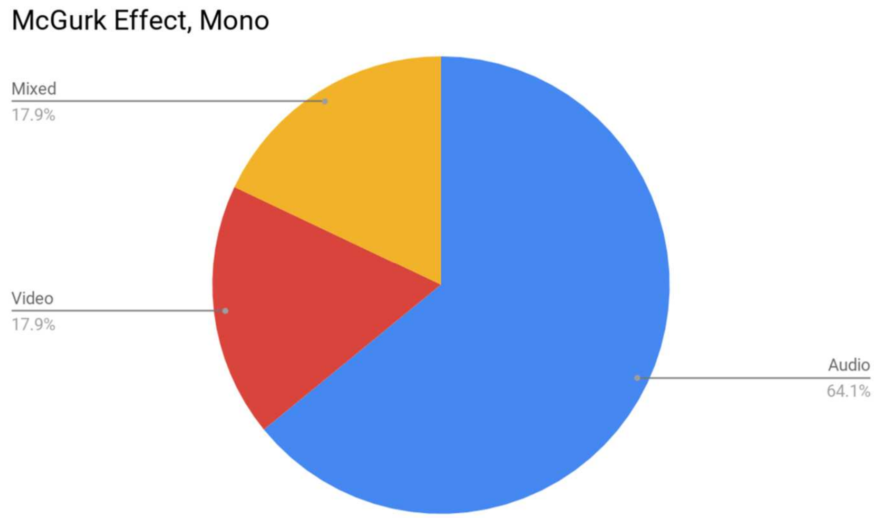


Image 11. Distribution of McGurk effect with mono sound

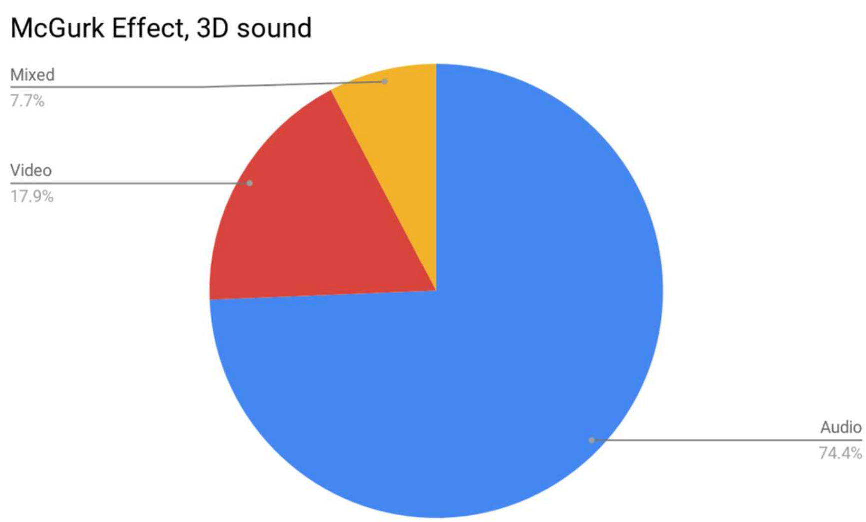


Image 12. Distribution of McGurk effect with 3D sound

It can be seen that the number of participants who reported mixed in the mono setup reduced in the 3D sound setup, with a simultaneous increase in the number that reported audio. It is noteworthy that while the number of participants who reported video remained constant, there is a shift in that category as well, since the set of participants reporting video aren't exactly identical, though there is an overlap.

The next section deals with checking this shift in responses for statistical significance.

Statistics - Mono vs 3D Sound

Hypothesis 1: There is a difference in the distribution of responses to the stimulus in the McGurk effect, which is due to the replacement of mono sound with 3D sound.

Sample size (n) : 39 Sampling method: Convenience Sampling

Design of study: Within Group Type of Variables: Categorical

Table 2. Count summary of responses for mono sound and 3D sound

Response vs Stimulus	Mono Sound	3D Sound	Total Count
Audio	25/39	29/39	54
Video	7/39	7/39	14
Mixed/Other	7/39	3/39	10
Total Count	39	39	78

Chi sq χ^2 = 1.8963 p-value = 0.3875 df = 2 α = 0.05

Effect size $\boldsymbol{\varphi} = \sqrt{\cdot\ ^2/n} = \sqrt{1.8963/39\text{x}2} \approx \mathbf{0.16}$

Power = 0.23 $\approx$ **23%**

Chi squared test was performed via R Studio. Power was calculated with G* Power.

Statistics - Original distribution vs Mono

This segment of statistics deals with testing for the successful replication of the study itself i.e., are the results of the test with mono in line with the results from the original study? For this, we run a Chi-Square Goodness of Fit test on the data from mono sound alone, using probabilities for Audio, Video and Mixed derived from the data of the original study from Table 1.

The resulting probabilities are [0.1, 0.25, 0.65] for A, V, M respectively.

We run this against the data from mono: [25, 7, 7].

The results are as follows:

Chi sq $\cdot\ ^2$ = 3 **p-value = 0.223** **df = 2** **α = 0.05**

Effect size **φ = 0.277** Power = **0.32**

Chapter 5: Discussion, Limitation, Future Work and Conclusion

In this chapter, we shall discuss the significance of the calculations made at the end of the previous chapter. We shall also delve into the limitations of the study and their role in possibly affecting the results of the study. Finally, we take a look at the future scope for the study, the adaptations, and the variations that could be made for the next level of studies on this topic and what remains to be explored. The chapter ends with a conclusion to summarize the study and its results.

Results: P - value, effect size and power for Mono vs 3D

P-value: Since the p-value is > 0.05, we fail to reject the null hypothesis i.e, there is not enough evidence to suggest that replacing mono sound with 3D sound has any significant effect on the distribution of the McGurk effect. That is to say, for a source directly in front of the subject, regardless of the nature or type of sound emanating in the case of conflicting visual and auditory stimuli, the distribution of the subject's response shall not vary significantly. Thus, while 3D audio is immersive, at this point it can't be said with assurity that it is immersive enough for this particular set of circumstances.

Effect size and power: For Chi Square Test of Independence with 2 degrees of freedom, a medium effect size is quantified as 0.3 and a small size is quantified as 0.1[10]. With the study showing an effect size of 0.155, it can be said that the effect size for the given data is less than medium and just more than small. This means that the difference between the two stimuli at this point is not enough. The results with a bigger sample size remain to be seen.

A major drawback of the study is its power, which is quite low at 0.23. This means that had there been a significant difference between the responses, there was just a 23% chance that it would have been detected.

Results: P-value, effect size and power for Original Distribution vs Mono

The P-value at 0.223 is significantly higher than the standard alpha value of 0.5, deeming that there is not enough evidence to state that the mono sound distribution in the replicated study is different from the original study.

Category vs % distribution	**% Audio**	**Video**	**Mixed**
Original	10%	25%	65%
Replicated mono	64%	18%	18%

Table 3. Comparison of the original and replicated mono distribution

Nonetheless, given the low sample size and power of 39 and 32% respectively, it is important to consider that these results might not be accurate. Considering the large variance in the original result and the replicated study shown in Table 3 above, it remains to be verified with a larger sample size if the replication of the McGurk effect was successful or not.

Significance

Inability to replicate the McGurk effect: The most noteworthy result of this study is its inability to replicate the original McGurk effect beyond doubt, as shown in the second segment of the statistics. This warrants further study of all the replications of the McGurk effect conducted since the original and trace the variation (if any) from the original. Considering that the quality of the stimuli, both audio and video, has evolved a lot since the original study, the extent of the effect itself might be different from what we expect based on the original study.

The role of spatial component in speech perception: While speech perception is surely not a purely auditory function, there might be limitations to the role of the auditory stimuli. This means it is possible that in the context of conflicting auditory and visual stimuli, the spatial component of sound, if present, does not contribute to the brain's processing of speech. If this is the case, it would be interesting to further study this process itself i.e., is the spatial component being acknowledged at all, or is it being ignored completely.

Stationary source: One possible explanation for the result lies in the nature of 3D sound. This study tests the McGurk effect, which is defined for a stationary source of sound. 3D sound is most distinguishable from regular sound and highly apparent when the source is moving. The spatial element enables for detection of the source not only on the same plane as the subject but also above, below, ahead, behind and a combination of these. Furthermore, not only is the source in the study stationary, it is also directly in front of the subject, possibly making it even more difficult to distinguish the two.

Technological limitations: It is possible that the McGurk effect is a manifestation of a side effect of human speech processing that cannot be manipulated by the technology we have today.

Limitations

Listed below are a few points that might have been responsible for skewing the results in either direction.

Sample Size: The biggest limitation of this study is the sample size. A bigger sample size might have been more potent in its ability to highlight any possible difference in distribution between the different stimuli.

Accent: Since our sample was not limited to any particular nationality, it is possible that there was some effect, however small, of the accent of the subject in the video. This effect is two way i.e, though the sounds emitted in the video have no language, the way the participants perceive the lip movements and the sounds might be affected because it is not a known word in English that might be approximated. If the sound happens to have some relevance in the participant's native language, this might have been a factor as well.

Binaural Mike: The binaural microphone used to record the 3D sound was not a professional one, but one 3D printed in a makerspace lab. While the microphone was tested and did exhibit traits expected from spatial sound, its quality might still not be as good as the professional ones that have 4 ears and are often connected via a mannequin's head. This might affect the ability of a participant to clearly distinguish between mono and 3D sound. The binaural mikes used for commercial recordings currently cost anywhere from $9000 and beyond.

Familiarity and Acknowledgment: It is possible that the effect of 3D sound is not acknowledged, because the technology hasn't yet been experienced and/or adopted by the masses. Since it isn't common knowledge that embedding of the spatial component is even possible, participants might be listening casually and not picking up on what might be in this case, a subtle difference.

Unprocessed audio: For this study, no specific processing was done on the audio, other than replacement of the original audio with mono and 3D sound. Using software to cancel out white noise and surrounding distractions like background noise might help the participants focus just on the sounds that matter and is likely to some degree of an impact on the study.

Future Scope

The most important study to better understand the McGurk effect would be to study the evolution of the results of the replications of the McGurk effect since the original while noting the specifications of the stimuli given. This is imperative given the study's inability to clearly replicate the McGurk effect.

Apart from the above, there is definitely scope for this study to be repeated with variation in different parameters as explained below.

Moving subject in the video: This step, in particular, would help greatly accentuate the effect of 3D sound. Movement of the source is perhaps the most impactful means of manifestation for the spatial component in 3D sound. The technology in itself is highly precise and would provide room for various test conditions as far as the movement is concerned. For instance, the source of the sound coming from below, above, behind, ahead, from the sides and any combination of these in addition to the speed of the source and its volume as well is apparent to a high degree in 3D sound.

Stationary angled subject in the video: This is in a way just a step above the current study. Having the subject in the video speak from not directly in front of the camera, but at an angle would help understand a few more things, while adding only a slight level of complications. It occurs often enough in the real world, where every speaker is not always directly in front of the listener and just like in the real world, we don't always need to look at the speaker to be able to hear the speaker. Please note that social norms dictate that it is polite, customary that it be done so. Looking at the speaker is also a passive means of acknowledgment. However, as far as being able to clearly hear as well as locate the speaker this is not mandatory by any means. Additional complexity for this study would stem from the fact that depending on the angle, part of the mouth and thus some lip movement might be obstructed and consequently, some other part highlighted. Combined with the enhanced 3D sound, this would make for an interesting study.

Enhanced visual stimulus: Just like 3D audio is an upgrade to just the auditory component in the McGurk effect, similar studies can be conducted upgrading only the visual stimuli. Thus, a large-sized screen can be used for the video with Ultra High Definition video quality with resolutions of 4K and 8K. Another variant of this study would be to use enhanced video in combination with 3D audio.

Professional Mike: As explained above in the limitations, replacing the mike with a professional mic used in commercial production of 3D sound would certainly be of use in further exploring this field. Some popular ones in the market as of this study are the Neumann KU100, 3Dio range and ZiBionic[7].

Conclusion

This thesis explores the replacement of mono sound with stationary 3D sound in the context of the McGurk effect and compares the distribution of responses in adults.

Subjects were presented with two similar video clips with conflicting visual and auditory stimuli, one with 3D sound, and the other with regular mono sound, and their responses gauged. The data shows no significant difference in the distribution for the given sample size.

The work presented in this thesis is exploratory in nature, in its infancy and an initial step towards better understanding both 3D sound and how it is related to the McGurk effect. Further assessment and iterations are necessary to address the given limitations.

Bibliography

[1] DiGiose, Nicole. “Surround Sound vs. 3D Sound.” *Electronic Products*, 28 Dec. 2016, www.electronicproducts.com/News/Surround_sound_vs_3D_sound.aspx.

[2] “The Difference Between Mono, Stereo, Surround, Binaural and 3D Sound.” *Hooke Audio*, 15 June 2015, hookeaudio.com/blog/2017/10/31/difference-mono-stereo-surround-binaural-3d-sound/.

[3] “The Differences Between Stereo, Virtual Surround, and 3D-Audio Headphones.” *OSSIC*, www.ossic.com/blog/2017/8/18/the-differences-between-stereo-virtual-surround-and-3d-audio-headphones.

[4] Lalwani, Mona. “Surrounded by Sound: How 3D Audio Hacks Your Brain.” *The Verge*, 12 Feb. 2015, www.theverge.com/2015/2/12/8021733/3d-audio-3dio-binaural-immersive-vr-sound-times-square-new-york.

[5] Taylor (2013) Taylor S. 2013. ‘Head orgasms’, meditation and near death experiences. The Guardian. *Available from* *http://www.theguardian.com/science/brain-flapping/2013/oct/09/head-orgasms-meditation-near-death-experiences* (accessed 30 September 2014)

Taylor (2014) Taylor V. 2014. Youtube videos trigger tingling ‘brain orgasms’ in ASMR practitioners. *Available from* *http://www.nydailynews.com/* (accessed 30 September 2014)

https://pdfs.semanticscholar.org/4921/466e051a09372ba4a449f214941a25310 ffd.pdf

[6] McGurk, Harry and MacDonald, John. “Hearing lips and seeing voices” *Nature*, Nov. 1076, http://wexler.free.fr/library/files/mcgurk%20(1976)%20hearing%20lips%20and%20seeing%20voices.pdf

[7] “Binaural Recording.” *Wikipedia*, Wikimedia Foundation, 3 Apr. 2018, en.wikipedia.org/wiki/Binaural_recording.

[8] “Binaural Microphone” *Thingiverse*, 13 Oct. 2014, https://www.thingiverse.com/thing:499001.

[9] “McGurk Effect” *Wikipedia*, Wikimedia Foundation, 10 Apr. 2018, en.wikipedia.org/wiki/McGurk_effect.

[10] “Effect Size for Chi-Square Test.” *Real Statistics Using Excel*, www.real-statistics.com/chi-square-and-f-distributions/effect-size-chi-square/