Fusion Confusion: Exploring Ambisonic Spatial Localisation for Audio-Visual Immersion Using the McGurk Effect

ABUBAKR SIDDIG, University College Dublin
ALESSANDRO RAGANO, University College Dublin
HAMED Z. JAHROMI, University College Dublin
ANDREW HINES, University College Dublin

Virtual Reality (VR) is attracting the attention of application developers for purposes beyond entertainment including serious games, health, education and training. By including 3D audio the overall VR quality of experience (QoE) will be enhanced through greater immersion. Better understanding the perception of spatial audio localisation in audio-visual immersion is needed especially in streaming applications where bandwidth is limited and compression is required. This paper explores the impact of audio-visual fusion on speech due to mismatches in a perceived talker location and the corresponding sound using a phenomenon known as the McGurk effect and binaurally rendered Ambisonic spatial audio. The illusion of the McGurk effect happens when a sound of a syllable paired with a video of a second syllable, gives the perception of a third syllable. For instance the sound of /ba/ dubbed in video of /ga/ will lead to the illusion of hearing /da/. Several studies investigated factors involved in the McGurk effect, but a little has been done to understand the audio spatial effect on this illusion. 3D spatial audio generated with Ambisonics has been shown to provide satisfactory QoE with respect to localisation of sound sources which makes it suitable for VR applications but not for audio visual talker scenarios. In order to test the perception of the McGurk effect at different direction of arrival (DOA) of sound, we rendered Ambisonics signals at the azimuth of 0°, 30°, 60°, and 90° to both the left and right of the video source. The results show that the audio visual fusion significantly affects the perception of the speech. Yet the spatial audio does not significantly impact the illusion. This finding suggests that precise localisation of speech audio might not be as critical for speech intelligibility. It was found that a more significant factor was the intelligibility of speech itself.

CCS Concepts: • **Human-centered computing → Virtual reality**.

Additional Key Words and Phrases: Ambisonics, McGurk effect, Virtual Reality

ACM Reference Format:
Abubakr Siddig, Alessandro Ragano, Hamed Z. Jahromi, and Andrew Hines. 2019. Fusion Confusion: Exploring Ambisonic Spatial Localisation for Audio-Visual Immersion Using the McGurk Effect. 1, 1 (November 2019), 10 pages. https://doi.org/10.1145/3304113.3326112

1 INTRODUCTION

Virtual Reality (VR) is a three dimensional computer generated environment that simulates or creates a new version of the physical world. Individuals immersed in a VR experience interact and explore the virtual environment like it was real through the use of appropriate devices. Recently, VR has delivered important advances in several sectors such as scientific and data visualisation, education, surgical training [24]. A typical VR application stimulates parts

Authors' addresses: Abubakr Siddig, abubakr.siddig@ucd.ie, University College Dublin, Dublin, Ireland, Alessandro Ragano, alessandro.ragano@ucdconnect.ie, University College Dublin, Dublin, Ireland, Hamed Z. Jahromi, hamed.jahromi@ucdconnect.ie, University College Dublin, Dublin, Ireland, Andrew Hines, andrew.hines@ucd.ie, University College Dublin, Dublin, Ireland,

Manuscript submitted to ACM

of the human sensory system such as vision, sound and tactile feedback. From a technological perspective VR has been focusing more on video, however 3D audio reproduction needs to be considered to address quality of experience (QoE) limitations [11] which can be quantified in terms of fidelity, immersion and presence [22]. It has recently been observed that "directional, 3D Sound will kick virtual reality up to a new level of vivid realism" [11]. One of the biggest challenges for VR streaming applications is guaranteeing satisfactory QoE in presence of the bandwidth limitations given the considerable amount of information to be transmitted [8, 22]. Typically, lossy audio/video compression for streaming applications relies on modelling perceptual phenomena in order to remove redundant information. Therefore, understanding perceptual phenomena in spatial audio is necessary for compression algorithms to deliver QoE for VR. Particularly, 3D spatial audio methods such as Ambisonics consist of capturing the acoustic scene (encoding) and then reproducing it using a set of loudspeakers or headphones (decoding). Sound source localisation accuracy in both encoding and decoding stages are important factors for QoE in Ambisonics applications.

In this study we investigate the relationship between a perceptual phenomena called the McGurk effect and the spatial localisation of Ambisonics. The McGurk effect is a multisensory integration phenomena which demonstrates that visual information affects speech perception even with perfect hearing conditions [18]. The discovery of the McGurk effect had impact on understanding audiovisual stimulus as it confirms that the perception of speech fuses audio and visual stimuli. The “unity assumption” [5] describes how multisensory cues can be treated as originating from the same source and interpreted as one unit of information. In this paper we explore the phenomenon of McGurk effect for an Ambisonics spatial audio scenario. Based on the “unity assumption”, our hypothesis is that spatial separation will impact perception of the McGurk effect. Understanding this relationship is important for streaming VR applications and will allow perceptual 3D audio coding algorithms to be developed both to improve QoE and use bandwidth efficiently.

In this paper we address the question: does the localisation provided by Ambisonic spatial audio influence audio-visual fusion? We examine this using an experiment based on the McGurk effect and compare the results to a similar experiment conducted with loudspeakers. To facilitate comparison with previous works and to eliminate various variables other than 3D spatial audio, we use a 2D video display and spatial audio rendered over headphones rather than a full VR setup.. This will provide insights into the influence of localisation accuracy on speech QoE in VR applications with immersive spatial audio.

The paper is structured as follows. Section 2 gives an overview of the related work which includes an explanation of the McGurk effect, an analysis of the spatial separation in such a phenomena and an overview of Ambisonics. In Section 3 we describe the proposed methodology for collecting data of the McGurk effect in Ambisonics scenarios. Section 4 provides a statistical analysis of the results followed by discussions in Section 5 and conclusions in Section 6.

2 RELATED WORK

The McGurk effect is a perceptual phenomena that demonstrates that hearing and vision are related in speech perception. When a speech sound is dubbed with a visual stimulus related to a second speech sound, we perceive a third sound. In "Hearing lips and seeing voices" [18], the authors dubbed the sound of /ga/ over the visual of /ba/ and vice versa, this led to the illusion of hearing the sound of /da/ for over 90% of the participants.

This illusion shed the light on how human brain perceives speech as a multisensory information and not only as an auditory process. This finding supports some previous researches [25] where it has been discovered that in face-to-face situation people can perceive speech even if the audio signal was in a noisy environment. Overall, McGurk effect is measured through subjective tests in which participants are given a set of recorded videos with dubbed audio that is not for the visual expression.

The significance of the McGurk effect is determined based on the responses received from the participants on what they think they heard. If the response does not match the actual auditory signal, then it is considered to be a McGurk effect. For example, if the original auditory is /ba/ and a response is /tha/ then this is a McGurk effect response. Most researches report the correct identification percentage and an f-test to provide statistical significance for the results.

2.1 Factors affecting the tests of McGurk Effect

Several studies have been conducted into the factors that affect the test of the McGurk effect. In [16, 18] age of participants has been analysed. Results show that adults (18 years and above) are more likely to experience the McGurk effect. Other studies focused on the problem of synchronising the visual signal of a syllable with the audio of a different syllable which is necessary for testing the McGurk effect. Munhall et al. [20] reported that a lagging of up to 360 ms still produces illusion for 40% of the participants while Massaro and Cohen [17] reported that the illusion occurred for up to 200 ms. Talker quality has been also investigated. Results show that the illusion was reported within a large range (from 17% to 58%) across different talkers [15]. The reasons behind this phenomena are not clearly known and potential causes could be clarity of articulation and speech rate as discussed in [2]. The gender of talkers has no effect on the perceived speech [15].

Audio and video quality have been also manipulated in order to find relevant features that affect the McGurk effect. Weak auditory consonant turned out as a key factor implying a stronger presence of the illusion. This factor has been investigated by decreasing sound intensity [6], increasing acoustic noise [1], and manipulating talker intelligibility [9]. Contrary results were discovered regarding visual stimulus. When visual information degrades the perception of the illusion decreases. These results were explored by adding noise [9], using spatial quantization [14], and spatial filtering [26].

2.2 Spatial Separation in McGurk Effect

The effect of the spatial segregation between the audio and the visual information in the McGurk effect has been poorly explored. Investigating the spatial separation consists of changing the direction of arrival (DOA) of the auditory information with respect to the visual stimulus to establish any changes in perception of the illusion. Bertelson et al. [4] concluded that the spatial separation up to 37.5° does not affect the audiovisual integration. A small effect of the separation has been discovered by Sharma [23] where azimuth of 60° to the left and to the right have been assessed. However these results were judged as not consistent by Jones and Munhall [13] where the authors proposed a more reliable method using multiple loudspeakers. In this work the auditory signal was presented as an azimuth of 0°, 30°, 60° and 90° with respect to the visual signal source. Jones and Munhall concluded that increasing the spatial separation has little impact on the McGurk effect. A later study by Jones and Jarick [12] explored the relation effect of the time combined with spatial audio on the McGurk effect. Their study concluded that the time separation has significant effect, where spatial separation will only be significant if the sound is originated from behind. However no significance is reported for DOA of up to ± 90° of the participant location. Tiippana et al. [19] explored the effect of spatial separation on McGurk effect by changing the permutations of sound DOA angles tested. They found that the McGurk effect is influenced only by spatial attention and not spatial separation except for separation test scenarios where the majority of sound DOAs were co-incident. To the best of our knowledge the above-mentioned works are the only ones that investigated the relation of the McGurk effect and the spatial audio separation. Some limitations in the experimental design of Jones and Munhall [13] should be noted. Participants were divided in three groups of 12 where they identified

consonants in audio-only, video-only, audiovisual conditions respectively. This approach is inconsistent given that the McGurk effect is different for each individual.

2.3 Ambisonics

Ambisonics is a method that reproduces a real soundfield previously recorded with microphones or synthesised. It virtually creates a 3D sound sphere around the listener [3] giving the impression of a real acoustic scene. Ambisonics consists in two stages. First the acoustic scene is captured by decomposing the soundfield in spherical harmonics (encoding). Then the acoustic scene is reproduced without relying on a specific loudspeaker setup (decoding) which makes it suitable for virtual reality and augmented reality [21, 22].

The accuracy of the reproduced soundfield strongly depends on the order of the spherical harmonics determined in the encoding stage. First-order Ambisonics (FOA) encodes the audio field into a format of 4 signals called B-Format: omni-directional signal (the gain) plus three additional difference signals X, Y, and Z, oriented along the three spatial axes. The decoding stage aims to reproduce the soundfield in a listening area called *sweet spot* surrounded by loudspeakers. Also, the soundfield can be reproduced for two channel scenarios like headphones through appropriate rendering methods. FOA has some limitations for real application scenarios given that sweet spot size, frequency accuracy and localisation accuracy depend on the order employed. Therefore higher-order Ambisonics (HOA) can be used to provide better localisation in the virtual scene. However, using higher order implies using more data to encode the signal.

In this study we use third order Ambisonics, to re-create the 3D audio compared to the loudspeakers setup previously done by [13]. Third order Ambisonics provides a good trade-off between localisation of sound sources and complexity with respect to the processing time and data utilisation for real time applications [21]. For headset based VR, Ambisonics needs to be rendered binaurally via headphones so the sound can be aligned with the visual display. One way to reproduce accurate binaural rendering in headphones is through the use of the Head Related Transfer Function (HRTF). In order to provide 3D spatial audio in headphones the encoded Ambisonics signals can be filtered with HRTF. For general purpose solutions, e.g. the Google Resonance Audio tool [10], audio is processed with a generic HRTF in order to provide spatial perception.

2.4 Summary

This section discussed the aspects of the McGurk effect by showing that speech is perceived through both hearing and vision. Studies have explored this phenomenon from various perspectives, including audio-visual spatial separation. This study reproduces the findings from previous studies. Additionally, it explores the limitations that might arise with implementing a virtually rendered spatial audio. To the best of our knowledge, this paper is the first to study the perception of speech in form of McGurk effect using Ambisonics rendered binaurally over headphones.

3 METHOD

Our method follows the methodologies described in [13] and [15] with some adjustments based on suggestions provided in [2] to improve reliability. This section outlines the procedure and setup.

3.1 Subjects

Thirty four volunteers (N=34: 20 males, 14 females, mean age = 29 years, range 23–37 years) from University College Dublin participated in the study. They are all postgraduate students from the School of Computer Science. They are

	Attribute	Original	Modified
Video	**fps**	29.97	29.15
	codec	h264	h264
	size	640x480 px	640x480 px
Audio	**codec**	aac	pcm_s16le
	sampling freq	48000 Hz	48000 Hz
	channels	2	1

Table 1. Table summarises the changes made to original McGurk stimulus files

English speakers (native or 2nd language), having normal or corrected to normal vision and reporting no hearing problems.

3.2 Stimulus materials

The stimulus videos used in [15] were shared by the authors of the study and consist of 4 males and 4 females. It was not reported whether the talkers are native English speakers. In this study, 4 out of the 8 stimuli were used: 2.3, 2.5, 2.7, 2.8. When selecting materials we made the following considerations:

(1) We tried to eliminate the quality of the talker issue as discussed by [2]. In their research they found that the McGurk effect under the same circumstances varies significantly among the participants based on the talker, therefore this research measures the illusion over various talkers (2 males and 2 females).
(2) When rendering Ambisonics, Resonance Audio tool requires the audio signal as input without the video signal. Therefore we separated the two signals and provided two different inputs to our web application, one containing the video and one containing the audio processed by Resonance Audio. We repeat each audiovisual stimulus 3 times. The Resonance Audio API introduced a lag between the audio and video signals on repetition. We selected videos to minimise distracting discontinuities in shot stitching for smooth transitions between repetition.

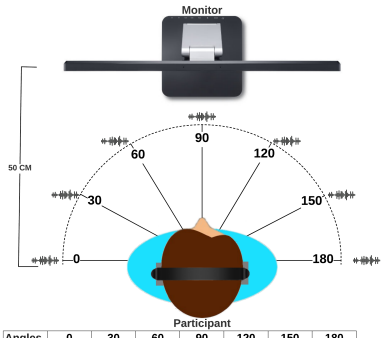
The stimuli are consonant-vowel (C-V) pair of /ba/ for audio and /ga/ for visual. This pairing generally leads to the illusion of /da/. This combination was originally used in [18], and many other researches. To concatenate the video files and isolate the audio and video files, ffmpeg tool was used. For the purpose of the research, some changes were made to the audio file in order to run with Ambisonics mainly the use of mono channel and the use of WAV files. Table 1 lists the original files attributes and the modified attributes when used to run the experiment.

The length of original files were 2.58 s, 2.58 s, 2.07 s and 2.058 s for stimulus 2.3, 2.5, 2.7 and 2.8 respectively. The modified audio/video files have 3 times the length i.e. 7.62 s, 7.62 s, 6.11 s and 7.62 s respectively. Hereafter the modified audio/video files will be referenced as stimulus 1, 2, 3 and 4 respectively.

3.3 Experimental procedure

The experiments were carried in a quiet lab room within a normal office setup where participants were asked to sit facing a 27" PC monitor (Dell S2719H, 1920x1080 px) with a viewing distance of approximately 50 cm. High quality studio headphones were used (audio-technica model no. ATH-M70x).

Spatial audio using Ambisonics is generated by playing the audio using the Google Resonance Audio tool for web [10]. The room was modelled as 4x4 m room with the participant at the centre. The sound is played from 7 different azimuth values starting from 0° which represents the talker is located to the left of the participant, to 180° which represents the



Angles	0	30	60	90	120	150	180
Setup							
2	A			-	B		
3	A		B			A	
4	A	B					A

Fig. 1. Top view of experiment setup. Virtual rendering of voice with respect to the test participants. The talkers were displayed on a video monitor located directly in front of them (90°. Speech was spatially rendered using third order Ambisonics with Resonance Audio at 0°, 30°, 60°, 90°, 120°, 150° and 180° of azimuth and 0° of elevation. This follows the loudspeaker configuration used in the experimental design of [13]. The table shows the analysis setups used in Sec 4.2

talker is located to the right of the participant. Figure 1 shows the setup layout. Throughout the experiment, third order Ambisonics were used.

Each participant hears each talker from all of the selected DOAs, therefore each participant did 28 audio-visual tests in total (4 talkers x 7 DOAs). Each participant also did 4 audio only tests, 1 for each talker where the azimuth is set to 90° which represents that both audio and video are from the same DOA. The purpose of the audio only tests is the anchor of the intelligibility of the speech perception of the participant.

3.4 Task structure

The experiment includes three tasks:

(1) **Introduction and Training task:** Each participant had 2 training tests (based on pilot tests), where the results were not part of the analysis. The purpose of these tests is to get the participant familiar with the setup. After each stimuli, the participant reports what is heard using radio buttons (ba, ga, da or other). If "other" is selected, then the participant writes what is heard in a text box.
(2) **Audio-visual McGurk tests:** The various clips were randomly selected from 4 talkers, where audio is played randomly from the any of the 7 azimuth values. However, the talker-azimuth combination is played once only.
(3) **Audio only tests:** The aim of these tests is to help including or excluding the results of one participant based on the ability to clearly identify what is heard. These are used to identify non-perceivers of the McGurk effect and to calibrate the results as discussed by [6].

4 RESULTS

4.1 Audio Only Results Analysis

The audio only test was conducted fixing the DOA in front of the listener, i.e. an azimuth of 90°, without the presence of the video. With this experiment we want to know if there could be intelligibility problem of sound independently of

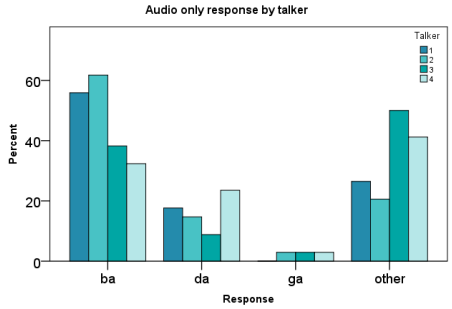


Fig. 2. Percentage of the responses of the Audio Only tests for each talker

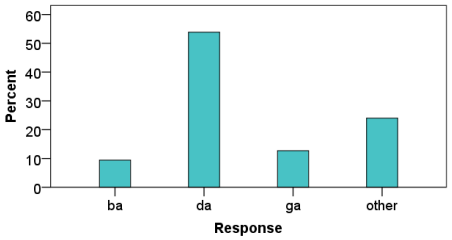


Fig. 3. Percentage of the responses of the Audio-Visual tests for all azimuth values

the McGurk effect and on the DOA. We decided to do not assess the intelligibility depending on the azimuth as it was already explored by Jones and Munhall [13] which reported that no particular DOA was affecting intelligibility.

Figure 2 shows the percentage of various stimuli that participants heard from each of the four talkers. The correctly identified stimuli percentages are 56%, 62%, 38% and 32% for talkers 1,2,3 and 4 respectively. It means that talkers 3 and 4 have impact on the results of McGurk effect since audio only was incorrectly classified.

We excluded participants who failed to classify all the stimuli in audio only test (N=8) since they do not provide reliable results for the McGurk effect.

4.2 Analysis by Virtual Audio Source Location

Overall McGurk effect analysis: By analysing the audio-visual responses, we found that only 13% could identify the /ba/ stimuli with the presence of visual content over the whole range of audio sources used in this research. Figure 3 shows the overall reported /ba/ for all azimuth values.

Directional audio analysis: We carried out this analysis assuming 4 different setups as shown in Figure 1. In all these setups, the video source is fixed at 90° facing the participant, only the DOA of the audio source is changed.

In setup 1, an analysis that includes all individual DOAs was performed. Figure 4 shows that the change of the azimuth angle does not affect the average number of hearing the correct/incorrect stimuli. Accordingly, no significance was found between the correct response /ba/ and the DOA of the rendered audio [$F(6, 182) = 0.148, p > 0.5$] for all azimuth values. This figure also shows the difference between audio only and audio-visual responses where it clearly reflect the McGurk effect when the video is introduced.

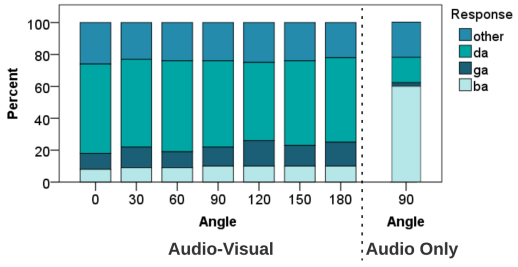


Fig. 4. Responses over various audio DOAs (setup 1) along side audio only responses at 90°

Figure 4 shows slightly higher correct responses to the right of the participant, this was also observed by Jones et al. [13]. Consequently, for setup 2 we looked into the effect of the right vs left audio source by creating two groups; A: [0°,30°,60°] and B: [120°,150°,180°]. We found that there is no significance effect on speech intelligibility if the audio is originated from right or left of the participant $[F(1, 624) = 0.719, p > 0.4]$. We concluded that we could "fold" L-R to examine spatial separation effect more deeply by examining results by angular distance independent of direction left or right.

Accordingly, we analysed the results by separating the angles tested into 2 groups: middle and sides. Bertelson et al. [4] discussed that the spatial separation up to 37.5° does not affect the audiovisual integration. We examined the results for setup 3 where group A is (0°,30°,150°,180°) and group B is (60°,90°,120°) and then for setup 4 where group A is (0°,180°) and group B is (30°,60°,90°,120°,150°). In both setup 3 and 4 we did not find any significance for the DOA on the perception of the McGurk effect i.e. speech intelligibility, the significance is found to be $[F(1, 728) = 0.113, p > 0.5]$ and $[F(1, 728) = 0.26, p > 0.5]$ respectively.

The above results show that no interaction is observed between the McGurk effect and the spatial audio using Ambisonics, which does confirm the previous results by Berltson et al. [4] and Jones et al. [13].

Talker analysis: As discussed in section 4.1, variations in responses were observed with the various talkers. Therefore we decided to further investigate this issue and to study its significance to the McGurk effect.

For all setup 1, 2, 3 and 4 above we found that the talker plays a significant role in the illusion, where the reported significance by setup are $[F(3, 728) = 2.367, p < 0.1]$, $[F(3, 156) = 2.261, p < 0.1]$, $[F(3, 182) = 2.25, p < 0.1]$ and $[F(3, 182) = 2.74, p < 0.05]$.

4.3 Discussion

We believed the that McGurk effect would change with an increase/decrease of DOA separation between audio-visual information. The results presented were in line with those presented by Jones et al. [13] indicating that the “unity assumption” is valid for binaurally rendered Ambisonics for separations up to ±90°.

However, the utilisation of Ambisonics not only repeat the results, but also overcome a shortcoming of their work where the audio only participants were not part of the audio-visual experiment. This might cause a bias in the results. This experiment also highlighted the issue of the talkers where we found that the McGurk effect can vary significantly upon various talkers, although this was reported by Mallick et al. [15], yet they did not justify the reason. In this paper we examined the audio-only intelligibility of the talkers where we found that for some of the talkers it was not an issue with the McGurk effect, but with the quality of the pronunciation itself. Since Mallick et al. [15] have not conducted

audio-only tests for the same group, our experiment suggests that the pronunciation quality can be the reason why participants reported higher McGurk effect for some talker over the other talkers.

One of the important factors in real time VR systems is the compression algorithm. These algorithms try to optimise the bandwidth while maintaining an acceptable level of QoE [22]. A better understanding of the importance of precise localisation of a talker audio with respect to the visual stimulus on speech intelligibility will be important for adaptive transmission algorithm development. For example, decisions regarding the Ambisonics order (first or the need for a higher order) and the level of compression to apply.

5 CONCLUSIONS AND FUTURE WORK

This paper explored the perception of the McGurk effect when the DOA of speech sounds changes according to Ambisonics rendering. The primary aim was to assess the impact of separated 3D audio from a talker video. To explore this issue, we studied and tested the influence of spatial separation on the McGurk effect. Our primary finding was that a significant number of participants do not correctly identify the correct phonemes from the audio signal making them unreliable subjects for assessing the McGurk effect when a visual stimulus is introduced. This is inline with [6, 7]. These results highlight the importance of assessing the intelligibility of audio stimuli independently of the audio visual or spatial factors. The experiment found that participants misidentified for talkers 3 and 4 significantly more than for talkers 1 and 2 (Figure 2). This finding may explain the higher McGurk effect reported by [15] as there was no auditory only study using this test material unlike other studies (e.g. [13]).

We concluded that the McGurk effect is not affected by the DOA for sound that is binaurally rendered using Ambisonics and that confusion caused by speech fusion is not influenced by DOA with respect to the video source. This result confirms the previous observation by [13].

Although it was not part of this research, it was observed that the mother tongue of the participants may have influenced the phoneme intelligibility. Participants from some countries reported hearing similar "other" sounds such as "taa" for Arabic speakers, and "pa" for Iranian speakers. This points to the effect of mother tongue language or possibly the accents of the talkers in the test material.

The results obtained will guide future work related to speech fusion and localisation for VR. Firstly, feedback from some participants highlighted their perception that stimuli were repeated. To eliminate possible bias a wider pool of speech should be used such as /ga/ and /ka/ to add variety to the samples for each talker. Secondly, a pilot study with audio only pre-screening tests should be completed to ensure high levels for speech intelligibility prior to mixing the McGurk effect videos. Finally, in this paper we investigated the azimuth angle, however a useful exploration can be also done on the elevation angle and the radial distance. The ability to simulate elevation angle in Ambisonics gives the opportunity to add perception of height and investigating the McGurk effect on different combinations of elevation, azimuth and radial distance for streaming VR scenarios. As the main aim of this study is to understand the effect of the spatial sound DOA on speech perception and speech intelligibility and how it may affect the QoE for VR, a follow on study will replace the video monitor with a VR headset giving the participant a greater feeling of immersion.

ACKNOWLEDGMENTS

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) and is co-funded under the European Regional Development Fund under Grant Number 13/RC/2077 and Grant Number SFI/12/RC/2289.

REFERENCES

[1] Magnus Alm, Dawn M. Behne, Yue Wang, and Ragnhild Eg. 2009. Audio-visual identification of place of articulation and voicing in white and babble noise. *The Journal of the Acoustical Society of America* 126, 1 (2009), 377–387. https://doi.org/10.1121/1.3129508

[2] Agnès Alsius, Martin Paré, and Kevin G Munhall. 2018. Forty years after hearing lips and seeing voices: the McGurk effect revisited. *Multisensory Research* 31, 1-2 (2018), 111–144.

[3] Xavier Amatriain, Jorge Castellanos, Tobias Höllerer, JoAnn Kuchera-Morin, Stephen T. Pope, Graham Wakefield, and Will Wolcott. 2007. Experiencing Audio and Music in a Fully Immersive Environment. In *Computer Music Modeling and Retrieval. Sense of Sounds*. Springer Berlin Heidelberg, Berlin, Heidelberg, 380–400. https://doi.org/10.1007/978-3-540-85035-9_27

[4] Paul Bertelson, Jean Vroomen, Geert Wiegeraad, and Beatrice de Gelder. 1994. Exploring the relation between McGurk interference and ventriloquism. In *Third International Conference on Spoken Language Processing*. ISCA, Yokohama, Japan, 559–562.

[5] Yi-Chuan Chen and Charles Spence. 2017. Assessing the role of the 'unity assumption'on multisensory integration: A review. *Frontiers in psychology* 8 (2017), 445.

[6] Cécile Colin, Monique Radeau, Paul Deltenre, Didier Demolin, and Alain Soquet. 2002. The role of sound intensity and stop-consonant voicing on McGurk fusions and combinations. *European Journal of Cognitive Psychology* 14, 4 (2002), 475–491.

[7] Sheetal Desai, Ginger Stickney, and Fan-Gang Zeng. 2008. Auditory-visual speech perception in normal-hearing and cochlear-implant listeners. *The Journal of the Acoustical Society of America* 123, 1 (2008), 428–440.

[8] Ching-Ling Fan, Jean Lee, Wen-Chih Lo, Chun-Ying Huang, Kuan-Ta Chen, and Cheng-Hsin Hsu. 2017. Fixation prediction for 360 video streaming in head-mounted virtual reality. In *Proceedings of the 27th Workshop on Network and Operating Systems Support for Digital Audio and Video*. ACM, Taiwan, 67–72.

[9] Eric Fixmer and Sarah Hawkins. 1998. The Influence of Quality of Information on the McGurk Effect. In *International Conference on Auditory-Visual Speech Processing*. ISCA, Australia, 27–32. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.399.5377{&}rep=rep1{&}type=pdf

[10] Google. 2016. Resonance Audio - Discover Resonance Audio. https://resonance-audio.github.io/resonance-audio/discover/overview.html Accessed on 29.01.2019.

[11] Mathias Johansson and Eddie Guy. 2019. VR For Your Ears: Dynamic 3D audio is key to the immersive experience. *IEEE Spectrum* 56, 02 (2019), 24–29. https://doi.org/10.1109/MSPEC.2019.8635813

[12] Jeffery A Jones and Michelle Jarick. 2006. Multisensory integration of speech signals: the relationship between space and time. *Experimental Brain Research* 174, 3 (2006), 588–594.

[13] Jeffery A Jones and Kevin G Munhall. 1997. *THE EFFECTS OF SEPARATING AUDITORY AND VISUAL SOURCES ON AUDIOVISUAL INTEGRATION OF SPEECH*. Technical Report 4. Canadian Acoustics / Acoustique Canadienne. 13–19 pages. https://jcaa.caa-aca.ca/index.php/jcaa/article/viewFile/1106/836

[14] John MacDonald, Søren Andersen, and Talis Bachmann. 2000. Hearing by eye: How much spatial degradation can be tolerated? *Perception* 29, 10 (2000), 1155–1168. https://doi.org/10.1068/p3020

[15] Debshila Basu Mallick, John F Magnotti, and Michael S Beauchamp. 2015. Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. *Psychonomic bulletin & review* 22, 5 (2015), 1299–1307.

[16] Dominic W Massaro. 1984. Children's perception of visual and auditory speech. *Child development* 55, 5 (1984), 1777–1788.

[17] Dominic W Massaro and Michael M Cohen. 1993. Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communication* 13, 1-2 (1993), 127–134.

[18] Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264, 5588 (1976), 746–748. https://doi.org/10.1038/264746a0

[19] Riikka Möttönen, Kaisa Tiippana, Mikko Sams, and Hanna Puharinen. 2011. Sound location can influence audiovisual speech perception when spatial attention is manipulated. *Seeing and perceiving* 24, 1 (2011), 67–90.

[20] K. G. Munhall, P. Gribble, L. Sacco, and M. Ward. 1996. Temporal constraints on the McGurk effect. *Perception & Psychophysics* 58, 3 (1996), 351–362. https://doi.org/10.3758/BF03206811

[21] Miroslaw Narbutt, Andrew Allen, Jan Skoglund, Michael Chinen, and Andrew Hines. 2018. AMBIQUAL-a full reference objective quality metric for ambisonic spatial audio. In *10th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, Italy, 1–6.

[22] Miroslaw Narbutt, Seán O'Leary, Andrew Allen, Jan Skoglund, and Andrew Hines. 2017. Streaming VR for immersion: Quality aspects of compressed spatial audio. In *23rd International Conference on Virtual System & Multimedia (VSMM)*. IEEE, Ireland, 1–6.

[23] D Sharma. 1989. *Audio-visual speech integration and perceived location*. Ph.D. Dissertation. University of Reading.

[24] Mel Slater and Maria V. Sanchez-Vives. 2016. Enhancing Our Lives with Immersive Virtual Reality. *Frontiers in Robotics and AI* 3, December (2016), 1–47. https://doi.org/10.3389/frobt.2016.00074

[25] William H Sumby and Irwin Pollack. 1954. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america* 26, 2 (1954), 212–215.

[26] Amanda H. Wilson, Agnès Alsius, Martin Paré, and Kevin G. Munhall. 2016. Spatial frequency requirements and gaze strategy in visual-only and audiovisual speech perception. *Journal of Speech, Language, and Hearing Research* 59 (2016), 601–615. https://doi.org/10.1044/2016