Perception Deception: Audio-Visual Mismatch in Virtual Reality Using the McGurk Effect

AbuBakr Siddig[1], Pheobe Wenyi Sun[1], Matthew Parker[2], Andrew Hines[1]

[1]School of Computer Science, University College Dublin, Ireland
[2]Texas Tech University, USA
abubakr.siddig@ucd.ie, wenyi.sun@ucdconnect.ie,
matthew.parker@ttu.edu, andrew.hines@ucd.ie

Abstract. The audio-visual synchronisation is a big challenge in the Virtual Reality (VR) industry. Studies investigating the effect of incongruent multisensory stimuli will have a direct impact on the design of immersive experience. In this paper, we explored the effect of audio-visual mismatch on the sensory integration in a VR context. Inspired by the McGurk effect, we designed an experiment addressing a few critical VR content production concerns today including sound spatialisation and unisensory signal quality. The results confirm previous studies using 2D videos where audio spatial separation has no significant impact on the McGurk effect, yet the findings raise new thoughts regarding the future compression and multisensory signal design strategies to optimise the perceptual immersion in the 3D context.

Keywords: Ambisonics, McGurk effect, Virtual Reality

1 Introduction

A virtual reality (VR) system places a participant in a computer-generated 3D environment that mimics or expands the physical world. Creating a better immersive experience has been an ongoing goal in VR research. Humans immersive experience is a result of multisensory integration. VR researchers are developing technologies that deliver accurate sensory cues providing increasingly natural sensorimotor contingencies [17] increasing the plausibility of virtual events. The sensory cues in state of the art VR systems (e.g. Oculus, Vive) primarily focus on visual and auditory immersion with some haptic feedback in hand controllers. Although the technology to create immersive visual content has become more powerful, accessible, and affordable, attention on auditory content creation and the audio-visual synchronisation has lagged which impacts the overall fused immersive experience [13].

To ensure the immersive feeling, audio and vision should match. If the sounds and audio cues are not plausibly aligned with the associated visual experience, the resulting incongruity causes the virtual immersion to collapse [13]. Multi signal mismatch is still a common problem due to issues related to latency, head-tracking technology, headphone, and immersive media content production.

Despite the potential of breaking the immersive feeling due to unavoidable multisensory disparity, an interesting phenomenon saying that humans can perceive a unified precept in the event of incongruent stimuli [11] first presented by psychologists McGurk and McDonald in 1976 inspires new concerns of audio-visual design for VR. This phenomenon proposes that human minds can join together mismatched sensory information to form an acceptable conclusion about an experience, and is studied using a classical research tool called the McGurk effect experiment. The McGurk effect highlights the importance of knowing the qualities of the unisensory stimuli (i.e., the clarity of the visual components, and the resolution of auditory components) and the coherence between the input sensory information before analysing the overall integrated experience [19]. The immersive experience in VR relies on multisensory integration. In this paper, we focus on the spatial qualities of audio-visual stimuli, specifically the relationship between where sounds are perceived to be coming from relative to their corresponding visual source. A better understanding of the importance of spatial audio localisation and image resolution on the immersive experience can guide the development and application of compression from a quality of experience perspective.

In this paper, we explore the interactions between disparities between the auditory and visual signals in a VR context. We experiment on speech perception inspired by the McGurk effect and we investigate how different factors (audio directionality and quality of visuals) affect the strength of audiovisual integration. The results are anticipated to guide to the candidacy of audio localisation for data compression in VR as a means to limit the bandwidth used by streaming media in a virtual environment.

2 Background

2.1 Immersive Experience in VR

Immersive experience is a result of multisensory processing. Sensory inputs for immersive experience commonly include vision, audio, touch and force feedback and less often smell and taste. In the context of VR, the goal is to simulate as many human natural sensory inputs as possible with the help of computer-based technology. The critical tools developed to reach this goal include wide field-of-view vision, stereo, head tracking, low-latency from head move to display, and high-resolution displays [17]. However, due to the multisensory property of the immersive experience, a single improvement in one aspect is not sufficient to improve the overall immersive experience in VR. When a virtually generated multisensory illusion gives an impression of a high plausibility, it is thus believed that the designed scene is actually happening [17]. Any mismatch between different senses would make a scene less convincing and resulting in the immersive sense of presence in a virtual scene collapse [13].

2.2 Spatial Audio

In practice, designers have been increasingly adopting 3D sound techniques to enrich the scenes contained in the auditory information to overcome the shortcomings of the traditionally-used stereo recordings. However, quality 3D audio is technically challenging to deliver for a variety of reasons. These include compromises in current content capture, production and delivery. For example, many existing affordable 360° cameras still use mono or stereo as their audio signals [14]. Without extra effort spent in capturing and rendering the spatial audio, the camera-recorded audio-visual content is incapable of creating a truly immersive sense of being there in the VR environment [14]. Dynamically updating both the visual and the auditory signal based on head-position and orientation pose further challenges to both network and compression technologies used to deliver immersive VR.

2.3 Spatial Separation in VR

The disparity between modalities can occur as a result of a timing mismatch between network and tracking technology. It can also be due to a mismatch of localisation information as a result of audio production and rendering technology. A detailed discussion of hardware and network factors that can result in disparities is beyond the scope of this paper. This paper will focus on the perception of spatial separation. The ramifications of audio-visual spatial separation are still not definite [16]. A misalignment of auditory and visual signals can negatively influence VR immersive experience [13]. However, psychologists highlight the human mind's capacity to form illusions given incongruent stimuli. For example, the 'ventriloquism effect' says given a visual effect, the location of the auditory sources can be overlooked, therefore forming an illusion as if the sound source is coming from the same place as the visual [18]. A similar idea was also reflected in the 'unity assumption' effect [2]. These theories drive the motivation to investigate whether spatial separation is perceptually detrimental in VR. The extent that spatial separation is tolerated will be important to both VR content and technology developers.

2.4 The McGurk Effect

To investigate whether a fusion effect exists in the event of audio-visual spatial separation in a VR context, this study builds upon a classical research tool called the McGurk effect that explores a phenomenon of an altered perception of auditory speech signal given incongruent audiovisual pairing [11] [4]. It is chosen due to the strength of the McGurk effect to 'reflect the strength of audiovisual integration' [19]. In a classical McGurk effect experiment, a participant is usually presented with an auditory and a visual signal simultaneously where each signal carries different speech tokens (i.e., audio of /ba/ together with a vision of /ga/). The participant's reported subjective auditory precept is expected to deviate from what is actually presented acoustically as a result of unconscious integration

of phonetic information (i.e., hearing /da/ when being presented with audio /ba/ and visual /ga/). Such categorical change of speech perception is described as the McGurk effect [19].

Since the initial publication of this lab controlled illusion experiment in 1976, the McGurk effect has attracted a lot of attention in the field of cognitive science. Psychologists believe the 'laws of common fate' and 'spatial proximity' are the two underlying theories for this phenomenon. These theories are based on the fundamental principles of perceptual information fusion [4]. Further down the line, researchers have been actively investigating factors that contribute to this phenomenon. Aside from the interpersonal difference and age factors [11][9], those factors describing stimuli including visual degradation [7], talker voice[8], choice of utterance [6], and time lag in synchronisation [10][12] all have been proven to have an impact to the strength of McGurk effect. All these influencing factors lead to a further abstraction of the conditions for the McGurk effect to occur: the quality of the unisensory signal and the coherence of the multiple sensory signals [19].

2.5 The McGurk Effect in VR

The spatial separation of audio and the visual information, as discussed in 2.3, is one of the most prominent concerns in the VR content creation process. Based on the current findings of the conditions for the occurrence of the McGurk effect, factors including the quality of sound localisation information, the quality of the visual information, and how far apart the separation between the auditory and the visual scenes are all critical concerns to provide users with an immersive experience. Ideally, the sound localisation information should be designed in the utmost detailed and accurate manner to match the visual scene so that the combined effect can generate a higher plausibility (Psi) to trick a participant into believing a virtual scene is real. Meanwhile, the 'ventriloquism effect', familiar to children where a puppet appears to speak, implies some tolerance of the coherence of the multisensory signals for the fusion effect to take place.

The existing literature, however, presents inconsistent conclusions regarding the ramifications on the integrated speech perception in the event of audio-visual disparity. There was no obvious impact found on the audiovisual integration when the spatial separation was up to 37.5 degrees [1]. A weak separation effect was found when the separation angle reached 60° [15]. Jones and Munhall found little difference in the impact on the McGurk effect when increasing the spatial separation of the auditory and the visual scenes [5] [4]. These results were also obtained by Siddig et al. [16] using 3D audio virtually binaurally rendered spatial audio for up to 90°. Jones and Jarick also concluded that the effect of the spatial separation was only significant when the sound was 180° away from the visual [4]. All of the above experiments were conducted based on 2D visuals only, although spatial sound was used to generate the auditory signals. We are interested in investigating whether the occurrence of the fusion phenomenon would change in a 3D scenario where participants are placed in a surrounding environment stimulated by both 3D audio and 3D visuals. Because in a 3D environment a

participant has an expectation of the source of the audio according to the visual stimuli, and the mind in this scenario is more likely to fuse multisensory signals to trick them into believing a scene is real. Therefore we postulate that a change of visual scenario from 2D to 3D could lead to a change in the integration effect. Following the current reasoning of the McGurk effect discussed in Section 2.4, it is reasonable to hypothesise that the strength of the McGurk effect in a VR context are a result of intelligibility of speech, resolution of the visual signals, and different degrees of spatial disparities between the sources of the auditory and visual signals.

If the psychological theories regarding perceptual phenomena hold true in a VR context, we will tolerate an amount of audio-spatial separation resulting from streaming and digital content processing as it will not result in a negative integrated experience. Building on the experiment described in [16] we aim to provide insights on multisensory data compression factors to consider for VR streaming applications.

3 Methodology

3.1 Experiment Design

The study was carried out in a quiet meeting room with a meeting room setup. Participants were asked to sit in a chair wearing a VR headset (HTC Oculus Quest) and stereo closed-back headphones (Audio-Technica ATH-M70x). Multiple virtual scenes of a speaker sitting across them were presented to the participants in sequence. The participants were asked to report what they heard after each playback. The experiment consisted of two tests: the audio-only utterance discrimination test and the audio-visual utterance discrimination test.

The collection of the perceived utterance takes the form of self-reported measures. A VR controller was used for participants to choose what they perceived given 8 possible confusing options. Namely /ba/, /ga/, /da/, /ma/, /ka/, /pa/, /va/, and /ta/.

3.2 Control

To detect the genuine McGurk effect in the experiment, we have to ensure the quality of the test unisensory stimuli controlled before we experiment. From the literature, the choice of utterance, the talker voice, the intelligibility of utterance, and the synchronisation of multisensory signals are factors contributing to the McGurk effect. We adopted what was used in the classical McGurk effect experiment, the simplest consonant-vowel combination (/ga/, /ka/, /ma/, and /pa/) as test utterances, and used different techniques to keep the levels of intelligibility of each enunciation of test utterances consistent. We selected talkers' voices that generate the highest intelligibility result among the multiple recorded talkers. In addition, the resolutions of each video clip capturing the lip movements are controlled. To set the same expectation in all participants that

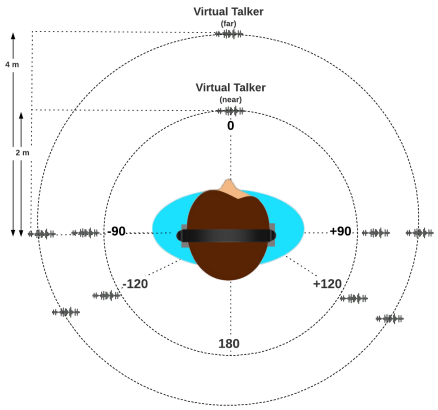

Fig. 1. Top view of experiment setup. Virtual rendering of voice with respect to the test participants. A virtual talker was displayed at azimuth 0° in a VR headset. Speech was spatially rendered using third order Ambisonics with Resonance Audio at 0°, 120°, 90°, and 180° of azimuth, and 60° of elevation.

they are placed in an immerse environment, all participants were asked to be seated, and the VR scene was preloaded before the headset was given to the participants. The same headset, headphones, and controllers were used for all participants.

The primary factor of interest in this experiment is the degree of spatial separation between the auditory and the visual signals. Therefore we fixed the position of the video to the front of the participants and rendered auditory signals with multiple directions of arrival (DOA). To further investigate the strength of the McGurk effect, two different qualities of the visual presentations (camera placed at 0.8 m and 2.3 m away from the talker) were used to test the robustness of the finding.

3.3 Stimuli

Audio and visual recordings The goal of designing a McGurk effect is to have 'incongruent' stimuli where 'each modality would present different speech tokens' [6]. We have recorded two male talkers uttering /ba/, /ma/, /pa/, /ga/, /ka/ and captured 360° videos of the same talkers mouthing /ga/ and /ka/. During the experiment, different combinations of the stand-alone visual and audio files were used to generate different pairs of audio-visual presentations. In addition to a complex of possible confusing combinations of visual and audio presentation, the effect of distance was also taken into account using the near and far scenes respectively. A close-up scene was taken at a distance of 2 metres, and the far scene was taken from 4 metres away.

Virtual environment creation The 360° visual stimuli featuring a talker's lip movements were produced using the Insta 360 One X camera. The spatial effect of the recorded auditory stimuli was created using Google ambisonics API. The virtual scenes of a talker uttering different syllables were dubbed with various conflicting auditory utterances in HitFilm. Unity was used to develop and build the VR experiment on Oculus Quest. To present the participants with a higher standard of immersive experience, the experiment was designed to be carried out using over-the-ear closed-back headphones in addition to a VR headset.

Fig. 2. Audio-Visual Utterance Test Scene

3.4 Participants

A total of seventeen participants from University College Dublin took part in this experiment. All participants spoke English fluently, having normal or corrected to normal vision and reporting no hearing problems.

4 Experiment Analysis

4.1 Audio-Only Utterance Discrimination Test

Procedure In the first test, the stimuli were played to the participants without visual display to learn participants' ability to discriminate the auditory utterances without any visual interference. To put the audio discrimination test in a spatial context, the audio rendered as if it is coming right in front of the participants. A fixed DOA, an azimuth of 0 degree, was used because previous literature showed that speech intelligibility did not change with azimuth angle [5][16]. Table 1 shows a list of the audio stimuli. After listening to each audio utterance, the participants were then asked to choose what they heard from the 8 options (namely /ba/, /ga/, /da/, /ma/, /ka/, /pa/, /va/, and /ta/) given on the screen using the VR controller.

Data Processing The purpose of this test is two-fold. It is used to filter out unreliable responses for the following test. Those who had difficulty discriminating between utterances from the audio will not be able to reflect the hearing

No.	Utterance	Talker
1	/ba/	1
2	/ma/	1
3	/pa/	1
4	/ba/	2
5	/ma/	2
6	/pa/	2

Table 1. Audio-Only Test Stimuli

confusion as a result of the visual interference [3] in the Audio-Visual Utterance Discrimination Test. This test is also used to gauge the quality of unisensory input signal. The accuracy of utterance discrimination based on the unisensory signal here will be taken into account when analysing the integration effect in the later test to gauge the genuinity of the McGurk effect and the strength of it [19].

Results Fifteen out of seventeen participants could discriminate most of the spoken token (/ba/, /ma/, /pa/) from the auditory utterances recorded from two different talkers. Two participants fell below 50% and were excluded from the second experiment due to consideration of the reliability of the responses [3] as discussed above. There was a slight difference in the rate of successful recognition between different utterances (see Table 2). Some of the languages do not have both sounds of /b/ and /p/, this is reflected on the lower accuracy rate for these sounds as seen in Table 2. Authors think this confusion is the main reason for two participants to fell below 50% in audio-only tests. Further investigation can be done to cover this issue. This insight was used in evaluating the strength of the McGurk effect in the second test.

Utterance	Accuracy Rate
/ba/	73.5%
/ma/	88.2%
/pa/	70.6%

Table 2. Unisensory Signal Quality

4.2 Audio-visual Utterance Discrimination Test

Procedure In the second test, both audio and visual stimuli were presented to the participants. The audio and video recordings of different utterances were put together in a synchronised manner to test the McGurk effect. The visual

stimuli were presented at the fixed azimuth position, as if the virtual talker were sitting in front of the participant (see Fig 2). However, the far and near scenes capturing two different levels of details of the talker's lip movements were used to test factors' impact on the strength of the McGurk effect. The audio stimuli were played at different azimuth angles to test the effect of spatial separation on the occurrence of the McGurk effect. Every audio-visual test stimuli was played only once for each participant (see Table 3).

Paring Type	Audio Signal	Visual Signal	Audio Directionality	Talker Voice	Visual Distance
1	/ba/	/ga/	azi 0°, ±90°, ±120°, 180°	1	near
1	/ba/	/ga/	azi 0°, ±90°, ±120°, 180°	2	near
1	/ba/	/ga/	azi 0°, elevation ±60°	1	far
1	/ba/	/ga/	azi 0°, elevation ±60°	2	far
2	/ma/	/ka/	azi 0°, ±90°, ±120°, 180°	1	near
2	/ma/	/ka/	azi 0°, ±90°, ±120°, 180°	2	near
3	/pa/	/ka/	azi 0°, ±90°, ±120°, 180°	1	near
3	/pa/	/ka/	azi 0°, ±90°, ±120°, 180°	2	near

Table 3. Audio-Visual Test Stimuli

Data Processing In this paper, the McGurk effect is labelled as positive when the reported percept deviates from the auditory signal. In the case where the reported percept is a third consonant other than the auditory or the visual signal, the response is also labelled as a strong McGurk effect. Because hearing a third consonant is a result of the fusion effect where the brain merges the speech tokens from both the auditory signal and the visual signal [19].

The test stimuli were grouped in various ways in the ANOVA test to evaluate the effect of sound directionality on the integration effect. Specifically, we were interested in exploring the front-rear difference, directionality difference, and the elevation difference in contributing to the McGurk effect in a VR scenario. the left-right difference was not explored because the previous research using 2D visual signals [16] concluded there was no significant effect between the left and right auditory signals.

Results In this experiment, the percentage of the reported the McGurk effect given front 3D visuals and audio at azimuth 0° was 30%, much lower than what was reported in the literature for experiments that were carried out with 2D videos. This result indicates that it is harder for the McGurk effect to take place in the event of mismatching multisensory signals in an immersive environment. Among all the reported occurrence of the McGurk effect, 97.7% showed signs of a strong McGurk effect. Such a high percentage of reported strong McGurk effect gave more confidence to say that the reported confused answers were as

a result of a fusion effect. Jones and Jarick previously concluded the spatial separation effect on the McGurk effect was only significant when the sound was coming from the rear [4]. However, in this experiment, there showed no significant difference between the signals generating from the front (0°) and the rear (180°) [$F(1, 270) = 0$, $p > 0.5$]. Unlike some results drown from the spatial separation experiment in a 2D environment, the change of DOAs seemed to have little effect on the formation of illusion in a 3D environment [$F(5, 408) = 1.9$, $p > 0.1$]. No talker's effect was found in this experiment [$F(1, 408) = 0.72$, $p < 0.5$].

Test	Independent Variables	p
Front & rear	azi 0° & 180°	>0.5
Azimuth angles	azi ±60°, ±90°, ±120°	>0.1
Talkers	talker 1 & talker 2	<0.5

Table 4. Test Result

5 General Discussion

Several recent studies did not find a significant difference in the effect of sound directionalities on the audio-visual immersion effect [16]. Following previous researchers findings, this paper specifically analysed the difference between the sound coming from the front and from the back. However, no difference was found between the front and rear sounds in this experiment. This could be accounted for by poor externalisation from binaural ambisonic rendering but requires more experimental testing. This paper also controlled the quality of unisensory input signal, attempting to see how much contribution a single modal signal can contribute to the strength of the McGurk effect. The controls used in this experiment were the resolution of talkers' utterance video and the recognition rate of the utterance sounds. The result did not show a significant statistical difference that altered the McGurk effect.

Comparing the occurrence of the McGurk effect in 3D with the findings from the 2D experiments [7], contrary to our expectation, we realised the conditions for a McGurk effect to occur is stricter in the immersive environment. This finding could be due to the change of expectation. Participants in a VR setting are by default expecting a higher level of coherence of the multisensory signals to form a belief that a real event occurs, thus hindering the chances for an illusion to occur given mismatched audio-visual signals.

6 Conclusions and Future Work

The experimental results in the VR environment are in line with the results seen by [5] using loudspeakers and with binaurally rendered spatial audio over headphones [16]. In an immersive environment with the source behind your head,

no statistical difference in the McGurk effect was seen although the experimental protocol did not capture the participants feedback on sound externalisation which can be covered in future work. The visual lip cue distance did not show a significant statistical difference point to an altered McGurk effect experience. The future work can further investigate the audio-visual mismatch along the elevation level to complement the current study on the effect of spatial separation of multisensory cues.

Spatial separation is considered an important quality issue for VR environments. However, these results show that people may not be very sensitive to separation for multisensory inputs – the fusion phenomenon was not significantly impacted by the changes to separation or visual cue resolution in VR. Our current finding indicates that, unless the audio and the visual are perfectly synchronised, a considerable extent of spatial separation between auditory and visual signals can be tolerated as the level of mismatch does not influence the level of immersion. As the 3D content streaming services rise in popularity, given limited bandwidth in a streaming scenario, a less strict audio localisation requirement can be adopted without deteriorating the overall immersive experience. As VR has gained popularity in today's society, this finding might be helpful for the VR content creators to optimise the usage of bandwidth when streaming 3D media content in a virtual environment. However, more experiments are needed to draw a final conclusion.

7 Acknowledgements

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/2289 and Grant Number SFI/12/RC/2077. The experimental work was supported by the University College Dublin STEM Summer Research Project. Thanks to Hamed Z. Jahromi and Alessandro Ragano for assistance with statistical analysis of the results.

References

1. Bertelson, P., Vroomen, J., Wiegeraad, G., de Gelder, B.: Exploring the relation between McGurk interference and ventriloquism. In: Third International Conference on Spoken Language Processing. pp. 559–562. ISCA, Yokohama, Japan (1994)
2. Chen, Y.C., Spence, C.: Assessing the role of the unity assumptionon multisensory integration: A review. Frontiers in psychology **8**, 445 (2017)
3. Colin, C., Radeau, M., Deltenre, P., Demolin, D., Soquet, A.: The role of sound intensity and stop-consonant voicing on McGurk fusions and combinations. European Journal of Cognitive Psychology **14**(4), 475–491 (2002)
4. Jones, J.A., Jarick, M.: Multisensory integration of speech signals: The relationship between space and time. Experimental Brain Research **174**(3), 588–594 (2006). https://doi.org/10.1007/s00221-006-0634-0, https://link.springer.com/content/pdf/10.1007%2Fs00221-006-0634-0.pdf

5. Jones, J.A., Munhall, K.G.: The effects of separating auditory and visual sources on audiovisual integration of speech. Canadian Acoustics - Acoustique Canadienne **25**(4), 13–19 (1997), https://jcaa.caa-aca.ca/index.php/jcaa/article/viewFile/1106/836
6. MacDonald, J.: Hearing Lips and Seeing Voices: The Origins and Development of the 'McGurk Effect' and Reflections on Audio-Visual Speech Perception over the Last 40 Years. Multisensory Research **31**(1-2), 7–18 (2018). https://doi.org/10.1163/22134808-00002548, https://brill.com/view/journals/msr/31/1-2/article-p7_2.xml
7. MacDonald, J., Andersen, S., Bachmann, T.: Hearing by eye: How much spatial degradation can be tolerated? Perception **29**(10), 1155–1168 (2000). https://doi.org/10.1068/p3020
8. Mallick, D.B., Magnotti, J.F., Beauchamp, M.S.: Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. Psychonomic bulletin & review **22**(5), 1299–1307 (2015)
9. Massaro, D.W.: Children's perception of visual and auditory speech. Child development **55**(5), 1777–1788 (1984)
10. Massaro, D.W., Cohen, M.M.: Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. Speech Communication **13**(1-2), 127–134 (1993)
11. McGurk, H., MacDonald, J.: Hearing lips and seeing voices. Nature **264**(5588), 746–748 (1976). https://doi.org/10.1038/264746a0, http://www.nature.com/articles/264746a0
12. Munhall, K.G., Gribble, P., Sacco, L., Ward, M.: Temporal constraints on the McGurk effect. Perception & Psychophysics **58**(3), 351–362 (1996). https://doi.org/10.3758/BF03206811, http://www.springerlink.com/index/10.3758/BF03206811
13. Narbutt, M., O'Leary, S., Allen, A., Skoglund, J., Hines, A.: Streaming VR for immersion: Quality aspects of compressed spatial audio. In: 23rd International Conference on Virtual System & Multimedia (VSMM). pp. 1–6. IEEE, Ireland (2017)
14. Rana, A., Ozcinar, C., Smolic, A.: Towards Generating Ambisonics Using Audio-visual Cue for Virtual Reality. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2012–2016 (2019)
15. Sharma, D.: Audio-visual speech integration and perceived location. Ph.D. thesis, University of Reading (1989)
16. Siddig, A., Ragano, A., Jahromi, H.Z., Hines, A.: Fusion confusion: exploring ambisonic spatial localisation for audio-visual immersion using the McGurk effect. In: Proceedings of the 11th ACM Workshop on Immersive Mixed and Virtual Environment Systems. pp. 28–33 (2019)
17. Slater, M., Sanchez-Vives, M.V.: Enhancing Our Lives with Immersive Virtual Reality. Frontiers in Robotics and AI **3**(December), 1–47 (2016). https://doi.org/10.3389/frobt.2016.00074
18. Stein, B.E., Meredith, M.A.: The merging of the senses. The MIT Press (1993)
19. Tiippana, K.: What is the McGurk effect? Frontiers in Psychology **5**, 725 (jul 2014). https://doi.org/10.3389/fpsyg.2014.00725, http://journal.frontiersin.org/article/10.3389/fpsyg.2014.00725/abstract