Multisensory Integration of Audiovisual Information in Avatar Faces: An Experimental Investigation of McGurk Effect

Babak Allahgholipour (bbkagp@gmail.com)
Multimedia Informatics Graduate Program,
Informatics Institute, Orta Dogu Teknik Universitesi
Ankara, 06800 Turkiye

Furkan Yaz (ffurkanyaz@gmail.com)
Statistics Department, Orta Dogu Teknik Universitesi
Ankara, 06800 Turkiye

Ayşegül Özkan (aysegul.ozkan@uj.edu.pl)
Cognitive Science Department,
Jagiellonian University
Krakow, 31-044 Poland

Cengiz Acarturk (cengiz.acarturk@uj.edu.pl)
Cognitive Science Department,
Jagiellonian University
Krakow, 31-044 Poland

Abstract

The McGurk effect is a multisensory phenomenon, an audiovisual illusion that shows how speech sounds may interfere with the visual sense. In case of an incongruency between a speech sound and a human face articulating the speech sound, human interlocutors tend to perceive an audiovisual percept different from the sound percept and the visual percept. For example, if the speech sound is *ba* and the visual presentation of the articulation is *fa*, human interlocutors usually perceive a third sound, *va*. What can we learn about audiovisual integration using an avatar face? To what extent do we observe the McGurk effect? The present study reports an empirical investigation that tested the extent of the effect in an avatar face designed for the purpose of the study. Our findings suggest that systematic patterns may obtained from the analysis of a specific avatar face, designed for the purpose of the study. The results also reveal the potential use of avatar designs for evaluating the impact of avatar designs on human multisensory integration by employing human responses to audiovisual illusions on avatar faces.

Keywords: Multisensory integration; McGurk effect; avatar face; face processing

Introduction

McGurk and MacDonald (1976) reported an audiovisual illusion, known shortly as the McGurk effect, based on a multisensory interpretation of incongruent auditory and visual information. The effect emerges when the same human voice, which utters a syllable, is presented synchronously by a different visual clue, such as a lip movement that displays another syllable. Therefore, the effect can be identified as one of the instances in which auditory perception and visual perception clash. A surprising outcome of the incongruence is the construction of a third percept, thus the audiovisual illusion, which is different from the auditory and visual information presented.

The McGurk effect has been conceived as a significant demonstrator of multisensory integration, since it shows that visual and auditory information can be merged into an emerging audiovisual percept.

In general, multisensory integration is assumed to be automatic, mandatory, and uncontrollable, as perception remains strong regardless of full knowledge of the situation. As a specific facet of multisensory integration, the McGurk effect is involuntary and is not under conscious control of a person. The effect is so evident and powerful that people cannot interfere or prevent the illusion even when they are aware of the process. Consequently, it can be recognized as an 'automatic' process in Posner and Sydner’s (1975) terminology on the automaticity of cognitive processes.

Different versions of McGurk stimuli have been examined, and the results have confirmed audiovisual integration and illusion. The effect exhibits a certain degree of variation in languages. For instance, Japanese native speakers exhibit smaller effects than English native speakers (Ali, Hassan-Haj, Ingleby, & Idrissi, 2005). However, its presence has been shown in numerous languages. A well-known variant has been the speech sound that utters *ba*, accompanied by different visual presentations, such as *da*, which were eventually perceived as *fa*, or depending on the visual presentation, as *ga*.

At a theoretical level, the effect may be an instance of a perceptually biased *best guess* of the audiovisual system to cope with the difficulty in mapping auditory and visual information. This mapping is relevant in everyday life settings, especially during communication in noisy environments, where percepts may be reconstructed by specific visual clues such as facial expressions, tongue, and lip movements.

Some of the explanations for the illusion focus on binding and fusing processes as in the fuzzy logic model of perception (FLMP; Massaro et al., 1995) and Bayesian approaches (Ma et al., 2009; Andersen, 2015; Magnotti and Beauchamp, 2017). A Bayesian approach, Causal Inference of Multisensory Speech (CIMS), focuses on binding auditory and visual stimuli based on the decision of whether they are from the same source (Berthommier, 2004; Nahorna et al., 2012). On the other hand, there are approaches that emphasize visual dominance and default hearing guesses (e see Gonzales, Backer, Mandujano & Shahin, 2021, for alternative theoretical accounts of the effect).

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

What can we learn about audiovisual integration using an avatar face? To what extent do we observe the McGurk effect? Can we use the McGurk effect to study the impact of avatar face design on human perceptual system? Some of them are not novel questions (Bregler, Covell & Slaney, 1997); however, they are all still relevant and timely questions since they aim to achieve an interdisciplinary transformation of a traditional approach in psychology research—investigating perceptual and cognitive mechanisms in borderline cases in the design of artificial agents.

Avatar faces can be designed as realistic as possible, to the extent that humans are not able to discriminate between real faces and avatar faces. For the past several years, various vendors have presented example avatar faces that impressively imitate real persons. Nevertheless, realistically looking faces are also subject to negative responses from their human interlocutors, such as the Uncanny Valley effect (Mori, MacDorman, & Kageki, 2012).

Is it likely that the McGurk effect can be employed as a methodology of assessment for the interaction with an avatar face, even possibly a minimum requirement, as long as the expectation from the avatar face is to be realistically looking? Finding answers to those questions requires extensive research on alternative designs of avatar faces and their componential structures and actions, such as lip movements. In the present study, we report an initial exploration of the opportunities for using avatar-face stimuli in an experimental setting that studied the presence of the McGurk effect on an avatar face designed for the purpose of the study. In the following, we present the relevant work that underlies the motivation of the research, the methodology, the results, and the discussion of the findings.

Relevant Work

The McGurk effect is a language-universal effect. Its qualitative production principles have been observed in numerous languages. Certain quantitative differences between languages, possibly due to differences between the structural characteristics of languages, have been reported (Ali, Hassan-Haj, Ingleby, & Idrissi, 2005), which also resulted in accumulation of divergent findings across different languages.

For example, the effect is strong in English in noise-free environments. However, relatively minor effects have been reported in Japanese. On the other hand, the effect is robust when the environment is noisy and auditory perception (hearing) becomes difficult. Adult Japanese native speakers and English native speakers also differ in reaction to audio vs. visual experimental stimuli conditions (Bovo, Ciorba, Prosser, & Martini, 2009). Native German and Spanish speakers have been reported to give more weight to visual stimuli to construct multisensory percepts, especially when they encounter foreign words (Sams, Surakka, Helin, & Kättö, 1997). The Arabic language also showed the presence of the effect (Ali, Hassan-Haj, Ingleby, & Idrissi, 2005). However, relatively weak effects have been reported in Chinese (Sams, Surakka, Helin, & Kättö, 1997). In Finnish, 90% of the participants were reported to be biased by the effect, a language in which visual stimuli had a strong impact on the results (Sams, Surakka, Helin, & Kättö, 1997). Turkish is another language in which a strong McGurk effect was observed (Erdener, 2015). In Italian, the effect was reported to be more significant with certain phonemes than with others.

The strength of the McGurk effect can be conceived as an example of the strength of audiovisual integration in the human perceptual system, making it a useful tool for research in language research (Tiippana, 2014). For instance, it has been proposed that some languages, such as English, have a deep orthography in that the relationship between phonemes and their visual (orthographic) representations is more complex than the others. Therefore, native speakers need stronger visual clues in those languages to resolve audiovisual disambiguation in noisy environments, compared to languages that have relatively shallow orthograph, such as Japanese and Chinese (Bovo, Ciorba, Prosser, & Martini, 2009).

The McGurk effect is also known as a fusion effect, since the perceptual system fuses the received information when sensory clues are insufficient to achieve coherent perception (Tiippana, 2014; Munhall, Gribble, Sacco, & Ward, 1996). The finding—that an auditory stimulus, when accompanied by an incongruent visual presentation, leads to a perception of a third percept different from auditory and visual input—is stable in various aspects (Mallick, Magnotti, & Beauchamp, 2015). For example, many participants report perceiving *di*, when the visual stimulus presented is *gi* and the auditory stimulus is *bi*. Similarly, combination effects have been reported when the auditory stimulus is *ba* and the visual stimulus is *ga*, leading to the audiovisual precept *bga* (Aruffo, & Shore, 2012).

Facial structure manipulation through inversion, rotation, translation into moving dots, and rearrangement may reduce the McGurk effect. A self-test of the McGurk effect, that is, using someone's own audiovisual stimuli, is also evident, suggesting that self-face identification processes have an impact on auditory identification of self-voice in speech processing (Aruffo, & Shore, 2012).

The presence of the effect has been shown in pre-linguistic ages (Bovo, Ciorba, Prosser & Martini, 2009). However, the effect is weak in children aged 3-5 and 7-8, possibly due to less developed visual model patterns and their perception compared to adults. As the age increases, the influence of visual stimuli increases along with the development of speech sounds and their mapping to visual patterns in the face.

On the other hand, no significant effect was observed between audiovisually congruent and incongruent articulation in bilingual infants, suggesting that increased variability in speech experience might allow greater tolerance to inconsistencies, and that the language environment in infancy might have an impact on the development of audiovisual speech perception (Mercure, Bright, Quiroz, & Filippi, 2022).

The introduction of noise in the auditory signal has been reported to increase the effect, also resulting in more gaze toward the mouth of the face stimuli compared to the absence of noise. On the other hand, adding blur to visual stimuli has been reported to reduce the effect and shorten participants' gazes in the mouth of the stimuli face (Stacey, Howard, Mitra, & Stacey, 2020).

The temporal difference between auditory and visual stimuli has been reported to have an impact on the effect. In general, a strictly controlled synchronization is not compulsory to observe the effect. However, a short delay in auditory stimuli strongly increases the influence of visual stimuli (Munhall, Gribble, Sacco, & Ward, 1996). Consequently, temporal aspects have an impact on the perceivers, since temporal coincidence is not compulsory, and the effect shows flexibility in the temporal dimension. The present study includes temporal delays as a factor in the experimental design.

Previous research on the McGurk effect suggests that audiovisual integration begins to occur in the early stages of processing in the human perceptual system (Sams, Surakka, Helin, & Kättö, 1997). This proposal is based on the findings that show that the structure of words, the use of syllables that do not exist in the language, and the use of non-words do not have a significant impact on the presence of the effect. On the other hand, alternative approaches exist, which state that modality specific sensory signals are processed independently at early stages, and then integrated later to form audiovisual percepts (Sporea, & Grüning, 2010).

The extent of the McGurk effect changes depending on how coherent and reliable information is provided by visual and auditory stimuli (Tiippana, 2014). Certain vocals, such as /i/ and /a/ have been proposed to have a stronger McGurk effect, compared to other vocals, such as /u/. Similarly, visual presentation of visual /d/ by the face and auditory presentation of acoustic /b/ have been reported to result in perceptual /d/. The underlying idea is that the lip movement formation in the pronunciation of certain vocals might be less information, thus leading to reduced visual information and a lesser effect (Bovo, Ciorba, Prosser, & Martini, 2009).

The questions about how, where, and when the brain performs audiovisual integration remain debatable (Bovo, Ciorba, Prosser, & Martini, 2009). Studying the role of the unisensory component in speech perception requires extensive testing (Bertelson et al., 2003; Alsius et al., 2005), computational modeling (Schwartz, 2010), and inspection of brain mechanisms (Sams et al., 1991; Skipper et al., 2007).

The present study reports an experimental investigation of the presence of the McGurk effect on an avatar face, designed for the purpose of the study, addressing major factors, such as the impact of auditory noise and delays between auditory stimuli and visual stimuli on the effect. In the following, we present the methodology of the study.

Methodology

Stimuli and Design

A 3D avatar (Figure 1) was prepared by a professional graphic designer, one of the authors of the present study. Several software tools were used in the design, such as Autodesk Maya for rigging and Mental Ray for rendering. Zbrush, Adobe Photoshop, Adobe Premiere Pro, Adobe Audition, Adobe Media Encoder, and Adobe Illustrator have also been used to produce final videos.

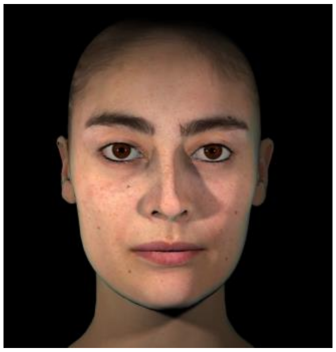

Figure 1: The avatar used in the study.

Thirty-two frames were rendered so that the first frame was the starting point of the animation, and the middle frame was the widest open mouth pronouncing *fa* or *ba*. The remaining frames presented a seamless movement of the closing mouth. The final frame was similar to the first frame, with the mouth completely closed (Figure 2).

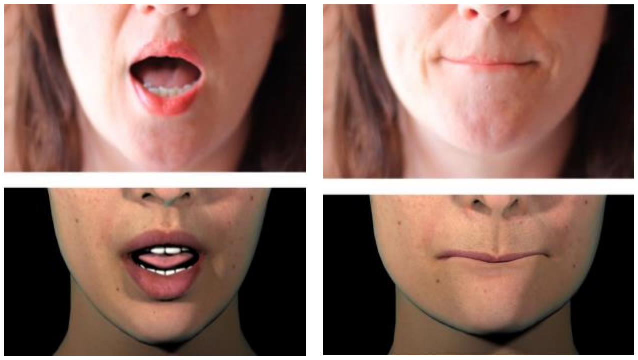

Figure 2: The middle frame (on the left) and the final frame (on the right). The top figures show the video recording frames. The figures in the bottom are for avatar frames.

Lip movements are vital to testing the McGurk effect. Therefore, the texture of the lip, its rigging and movement, the duration of the animation, and the sound dubbing have been designed based on real-world video recordings shot for the purpose of the study.

The frame size of the videos was 1280px wide and 720px height (16:9) having a 24 fps frame rate, to allow for fast 3D rendering. The data presentation bandwidth was 10 mbps and the total bitrate was 100 mbps. The audio rendering bitrate was 314kbps, with a sampling rate of 48kHz on the stereo sound channel. Each video lasted twelve seconds. We used H.264/MPEG-4 AVC and codec avc1 for video and mp4a: MPEG-4 AAC LC codec for audio.

We used recorded human voices to accompany the video stimuli. Figure 3 shows a close-up view of the movement of the lips, pronouncing a *ba* sound in 30 frames between the first frame and the last frame.

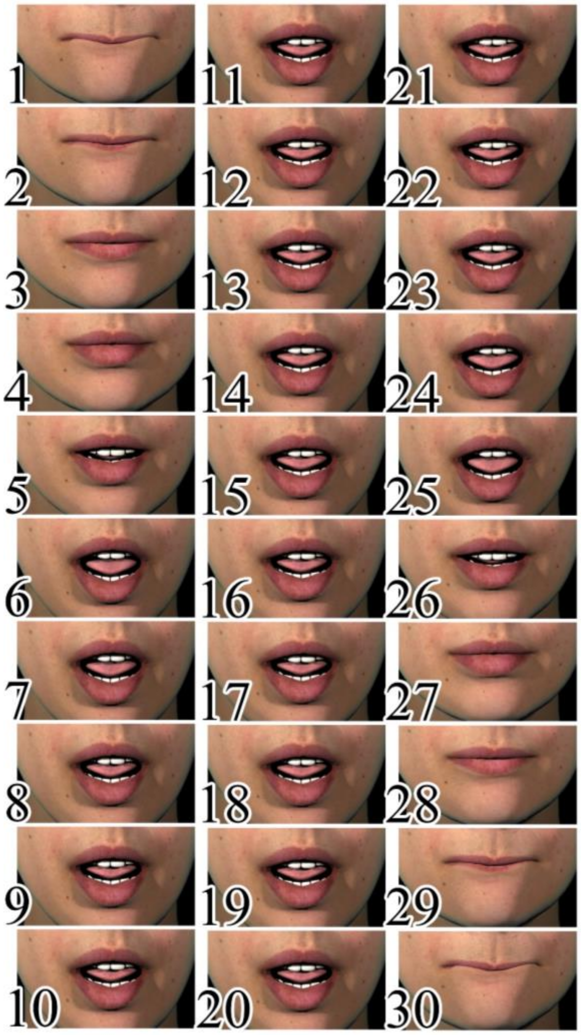

Figure 3: Close-up view of lip movements pronouncing a *ba* sound.

We designed a within-subject experiment with three factors, noise having two conditions (absence vs. presence of noise in sound), delay having four conditions (0, 200, 300, 400 ms), and the face having two conditions (real human face vs. avatar face). We used the syllables *ba* and *fa* and prepared videos representing four alternative combinations of the syllables (audio *ba* and visual *ba*, audio *ba* and visual *fa*, audio *fa* and visual *fa*, and audio *fa* and visual *ba*). The subsequent stimuli consisted of 64 videos.

Participants and Procedure

Thirty-four participants (age range 25-35, 20 females), all native speakers of the language of the experiment (Turkish), participated in the experiment on a voluntary basis. Ethics permission was granted prior to data collection.

The stimulus videos were presented online through a website designed for the purpose of the study. After playing each video, the participants selected one of the four choices: *ba*, *fa*, *da*, and *va*. The video stimuli were presented in random order. Each video lasted 12 seconds. The experiment lasted approximately 15 minutes, including the introduction and instructions. In the following, we present the results of the experiment.

Results

We report the results of the experiment in two sections. First, we provide descriptive statistics to show the presence of the McGurk effect in situations where the effect was expected (incongruent trials) and unexpected (congruent trials). Here, we present descriptive statistics for the effects of noise and delay conditions on human and avatar videos. Second, we present inferential statistics to provide a partial explanation of the results.

Descriptive Statistics

The presence of the effect. The McGurk effect is expected in situations where the audio stimulus and the accompanying visual stimulus are incongruent. Similarly, the effect is not expected when the two stimuli are congruent. Table 1 shows the distribution of responses to audiovisual stimuli as a percentage, regardless of the experiment conditions. In the table, A-V shows audio and visual stimuli in the rows, respectively, and the column names show the selected percepts of the participants.

Table 1: Distribution of responses to audiovisual stimuli.

A-V stimuli	*ba*	*fa*	*da*	*va*
ba-ba	.80	.03	.04	.13
fa-fa	.04	.89	.01	.06
fa-ba	.17	.78	.01	.03
ba-fa	.43	.06	.06	.44

The results show that the participants were able to elicit the correct responses (the initial two rows in Table 1) 80% of the time for the *ba-ba* stimuli and 89% of the time for the *fa-fa* stimuli. The human-face stimuli and the avatar-face stimuli showed slight differences in this pattern. For human-face stimuli, the correct response ratio was 76% for *ba-ba* stimuli and 95% for *fa-fa* stimuli. For avatar-face stimuli, the correct response ratio was 84% for *ba-ba* stimuli and 82% for *fa-fa* stimuli. Overall, the results show that the participants were able to recognize the presented (congruent) percept most of the time both in the human-face stimuli and in the avatar-face stimuli.

The last two rows of Table 1 show the cases in which the McGurk effect was expected. The effect was observed in *ba-fa* stimuli (the last two rows in Table 1) in 50% of the time (i.e., the participants reported other percepts, *da* or *va*, rather than *ba* or *fa*). However, the effect was virtually not observed in the *fa-ba* stimuli. Only in 4% of the cases did the participants report another percept than those presented. This result was worrying in that it could have pointed to a design issue in the avatar face. However, the results were similar between the human-face and the avatar-face stimuli (the third-percept response was 4% for the human-face stimuli and 5% for the avatar-face stimuli). It is also likely that the human-face video had design imperfections. However, the high precision of the responses to the congruent stimuli (the first two rows in Table 1) suggests that this finding is a characteristic of the language of the experiment.

In summary, in the language of the experiment, in this case, Turkish, the *ba-fa* audiovisual stimuli elicited a stronger McGurk effect than the *fa-ba* stimuli. In the following, we present the results showing the impact of adding noise to auditory stimuli.

Noise effects. Noise has been reported to be an important factor that influences the emergence of the McGurk effect. It was also part of the experiment design in the present study. We introduced auditory noise as a factor with two conditions (absence vs. presence of noise). The results reflected the impact of noise on the experimental conditions, as shown in Table 2. In the table, A-V shows audio and visual stimuli in the rows, respectively, and the column names show the selected percepts of the participants.

Table 2: Distribution of responses to audiovisual stimuli. H/A shows Human/Avatar.

A-V stimuli	*H/A*	*Noise*	*ba*	*fa*	*da*	*va*
ba-ba	*Human*	*No*	.65	.00	.00	.35
		Yes	.86	.03	.02	.09
	Avatar	*No*	.89	.04	.03	.04
		Yes	.80	.05	.13	.03
fa-fa	*Human*	*No*	.00	.97	.01	.02
		Yes	.02	.92	.01	.05
	Avatar	*No*	.08	.83	.01	.07
		Yes	.07	.81	.01	.10
fa-ba	*Human*	*No*	.05	.95	.00	.00
		Yes	.27	.66	.02	.05
	Avatar	*No*	.08	.87	.01	.04
		Yes	.29	.65	.02	.04
ba-fa	*Human*	*No*	.16	.03	.00	.81
		Yes	.31	.05	.02	.62
	Avatar	*No*	.88	.06	.03	.04
		Yes	.38	.12	.20	.30

In conditions in which the McGurk effect is not expected (i.e., the *ba-ba* and *fa-fa* stimuli), the presence of noise in the auditory stimuli mostly had a major impact on the *ba-ba* condition and a minor impact on the *fa-fa* condition, except for a few cases, such as the *va* percept in the *ba-ba* stimuli (the first raw in Table 2). Nevertheless, the overall direction of the impact was reversed in the avatar, compared to the human face stimuli.

The overall results suggest that the introduction of noise to the auditory stimuli may lead to different impacts to different stimuli (in this case, *ba-ba* and *fa-fa*). It also shows that it may lead to a discrepancy between the human face and the avatar face, suggesting that audiovisual integration may follow different patterns in the perception of avatar faces from human faces under noisy conditions. In other words, the visual system's best guess can be influenced differently when interacting with avatar faces and human faces when noise is introduced into the environment. A likely reason is that human face perception involves more than the perception of lip movements, not entirely reflected in the design of the avatar. Those aspects of avatar design might be closely related to Gestalt perspectives to face perception.

The remaining stimuli (*fa-ba* and *ba-fa*) are those for which the McGurk effect was expected. In the *fa-ba* stimuli, a minimal effect was observed. As presented above, this finding may be a characteristic of the language of the experiment. However, in *ba-fa* stimuli, in which the effect was observed (Table 1), the introduction of noise resulted in a decrease in the amount of the effect in the human face stimuli, while it resulted in a reverse impact in the avatar-face stimuli, which implies a determination of different sources of visual and auditory stimuli (Berthommier, 2004; Nahorna et al., 2012).

Delay effects. The design of the experiment also involved time delays between the presentation of the visual stimuli and the auditory stimuli (the latter preceding the former with 0, 200, 300, 400 ms delays), to address its likely influence on the McGurk effect. The results showed minimal effects of delay in virtually all conditions, generally less than 10%. The following section presents the regression models for further analysis of the results.

Regression Models

The results are presented in terms of applying multinomial logistic regression with a multilevel structure, applied to each trial from the experiment. The multinomial structure was employed to account for the probabilities of participant responses with *ba*, *fa*, *da*, and *va*. The analysis was performed to investigate the effects of five independent variables, namely *A-V stimuli* (with four levels: *ba-ba*, *ba-fa*, *fa-ba*, *fa-fa*), *Noise* (with two levels: *absent* or *present*), *Delay* (with four levels: 0, 200, 300, 400 in milliseconds), and *H/A* (with two levels: Human and Avatar) on the dependent variable with four response categories (*ba*, *fa*, *da*, *va*). The probability values for each response category were included as weights.

The model provided an overall residual deviance of 85.2 and an AIC value of 99.2 *va* response was the target McGurk effect response. We obtained significant findings in some but not all the conditions. The participants reported the *va* percept more frequently as a response to *ba-fa* stimuli than the *fa* percept ($p < .05$). No significant effect was obtained between the two noise conditions, in contrast to notable differences in descriptive statistics. Similarly, no significant effect was obtained among the four delay conditions. However, a significant effect was obtained between human and avatar face stimuli ($p = .05$). Therefore, the findings suggests that a significant part of the participant response percept *va* occurs when auditory stimuli are *ba* and visual stimuli are *fa*, and the use of avatar-face stimuli significantly eliminates the effect. In particular, the use of avatar-face stimuli muted the effect to a certain degree, compared to the use of human-face stimuli. In total, 61.8% of the variability of *va* responses can be explained by the predictors in the models.

Further analyses with regression models showed that the participants reported *fa* as a response to both *fa-fa* stimuli (auditory *fa*, visual *ba*) and *fa-fa* stimuli. In total, 96.2% of the variability of *fa* responses can be explained by the predictors. We leave further analysis of the data and the interactions between the conditions to future work. Below, we present a discussion of the findings and conclusions from the reported study.

Discussion and Conclusions

Humans usually assume that characters and places in computer games and television represent real, genuine persons and places (Mullen, 1999). We may even prefer to put the knowledge of the *real* aside for the sake of enjoyment, like wishing that the Spider-Man character would be a real one. Realistic designs have an important influence on those assumptions. Character creation, virtual world modeling, and game design assume to some degree believability to be able to keep the audience engaged. The design of avatars also provides an opportunity to expand the research findings of the past to the study of our perception in the context of current technologies.

In the present study, we tested whether it is possible to design a virtual avatar to observe audiovisual integration patterns, specifically, a well-known audiovisual illusion, the McGurk effect. Our experiment resulted in patterns about the presence of the effect, sometimes in a reverse direction between the human-face stimuli and avatar-face stimuli.

Numerous factors have an impact on the presence and strength of the effect. A change in environmental conditions (such as mandatory facemasks and reduced access to lip reading) can change a person's general disposition towards different modalities in speech communication (Chládková, Podlipský, Nudga, & Šimáčková, 2021).

However, the McGurk effect is a strong effect with numerous surprising variants. For instance, an emotional McGurk effect has been observed if visual stimuli are vocalized in a specific emotion and it is dubbed with the sound of another emotion (Fagel, 2006). These findings suggest that the effect may have many connotations in the human audiovisual perceptual system, not limited to human responses to auditory stimuli and accompanying lip movements and facial expressions. Therefore, we believe that avatar faces will provide a prolific test bed for the study of audiovisual integration under boundary conditions.

This research has limitations on several fronts. First, the experiment is limited to a specific avatar, professionally designed for the purpose of the study. The design was based on a recorded video of a human face uttering the syllables. Although the avatar was useful to achieve the objectives of the study—that it is possible to observe the McGurk effect with digital design. The generalization of the findings will be possible when avatar faces with different designs are investigated for observing the effect.

We also presented inferential statistics within a limited scope by using multinomial logistic regression. The analyses need improvement in reporting the detailed presentation of parameter values.

Another limitation is the use of a limited set of audiovisual stimuli (*ba* and *fa*), typical stimuli used in similar studies. However, both for congruent audiovisual stimuli and for incongruent ones, the response to the stimuli (percepts reported by the participants) showed variance. An assessment of a richer set of audiovisual stimuli will allow selecting the most appropriate audiovisual stimuli to observe the impact of various factors on McGurk effect.

The findings are also limited to the language of the experiment, since languages may exhibit variance in the strength of the effect. Therefore, the findings obtained in the present study should be conceived as a preliminary step to design sufficiently comprehensive experiments, currently having a limited scope in terms of its coverage of language specific findings.

Future research should address those aspects insufficiently addressed in the present study. Furthermore, future research should also address various factors that influence the presence of the McGurk effect. One of them is individual differences (Mallick, Magnotti, & Beauchamp, 2015). Future research should address an account of individual traits that may have an impact on audiovisual perception of avatar faces.

A relevant factor is the expectation of participants of a virtual reality (VR) experience. In the VR environment, participants generally expect high consistency between multisensory signals due to their expectation that a real event has occurred (Siddig, Sun, Parker, & Hines, 2019). Therefore, participants should be instructed about their expectations about the studies conducted in VR.

Avatar-face stimuli also provide an opportunity to test audiovisual integration in people with hearing difficulties, who usually have stronger audiovisual integration (Schulte et al., 2020) compared to people with normal hearing.

Acknowledgments

The research has been supported by the grant *Cognitive Aspects of Interaction and Communication in Natural and Artificial Agents* from the Priority Research Area *DigiWorld* under the Strategic Programme Excellence Initiative at Jagiellonian University, Krakow, Poland.

References

Andersen, T. S. (2015). The early maximum likelihood estimation model of audiovisual integration in speech perception. *Journal of the Acoustical Society of America, 137*, 2884–2891.

Aruffo, C., & Shore, D. I. (2012). Can you McGurk yourself? Self-face and self-voice in audiovisual speech. *Psychonomic Bulletin & Review, 19*(1), 66-72.

Berthommier, F. (2004). A phonetically neutral model of the low-level audio-visual interaction. *Speech Communication.*, 44, 31–41.

Bovo, R., Ciorba, A., Prosser, S., & Martini, A. (2009). The McGurk phenomenon in Italian listeners. Acta *Otorhinolaryngologica Italica, 29*(4), 203-208.

Bregler, C., Covell, M., & Slaney, M. (1997). Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques* (pp. 353-360).

Chládková, K., Podlipský, V. J., Nudga, N., & Šimáčková, Š. (2021). The McGurk effect in the time of pandemic: age-dependent adaptation to an environmental loss of visual speech cues. *Psychonomic Bulletin & Review*, 28, 992-1002.

Erdener, D., (2015). The McGurk Illusion in Turkish. *Türk Psikoloji Dergisi* , 30, 19-31.

Fagel, S. (2006). Emotional McGurk effect. In Proceedings of the International Conference on Speech Prosody, Dresden, Germany. https://www.isca-speech.org/archive/.

Gonzales, M. G., Backer, K. C., Mandujano, B., & Shahin, A. J. (2021). Rethinking the mechanisms underlying the McGurk illusion. *Frontiers in Human Neuroscience*, 15, 616049.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., and Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. *PLoS One*, 4, e4638

Magnotti, J. F., and Beauchamp, M. S. (2017). A causal inference model explains perception of the McGurk effect and other incongruent audiovisual speech. *PLoS Comput. Biol.*, 13, e1005229

Mallick, D. B., Magnotti, J. F., & Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review, 22*(5), 1299-1307.

Massaro, D. W., Cohen, M. M., and Smeele, P. M. (1995). Cross-linguistic comparisons in the integration of visual and auditory speech. *Memory & Cognition*, 23, 113–131. doi: 10.3758/bf03210561

Mercure, E., Bright, P., Quiroz, I., & Filippi, R. (2022). Effect of infant bilingualism on audiovisual integration in a McGurk task. *Journal of Experimental Child Psychology*, 217, 105351.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine, 19*(2), 98-100.

Mullen, C. A. (1999). The media equation: How people treat computers, television, and new media like real people and places. *International Journal of Instructional Media, 26*(1), 117.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics, 58*(3), 351-362.

Nahorna, O., Berthommier, F., and Schwartz, J.-L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *Journal of the Acoustical Society of America*, 132, 1061–1077. doi: 10.1121/1.4728187

Posner, M. I. & Snyder, C. R. R. (1975). Attention and cognitive control. In Robert L. Solso (ed.), *Information processing and cognition: The Loyola Symposium.* Lawrence Erlbaum.

Sams, M., Surakka, V., Helin, P., & Kättö, R. (1997). Audiovisual fusion in finnish syllables and words. In Proceedings of AVSP Auditory-Visual Speech Processing, (pp. 101-104), https://www.isca-speech.org/archive.

Schulte, A., Thiel, C. M., Gieseler, A., Tahden, M., Colonius, H., & Rosemann, S. (2020). Reduced resting state functional connectivity with increasing age-related hearing loss and McGurk susceptibility. *Scientific Reports, 10*(1), 1-12.

Siddig, A., Sun, P. W., Parker, M., & Hines, A. (2019). Perception deception: Audio-visual mismatch in virtual reality using the McGurk effect. In Proceedings of AICS, (pp. 176-187), https://ceur-ws.org/Vol-2563/aics_18.pdf.

Sporea, I., & Grüning, A. (2010). Modelling the McGurk effect. In *Proceedings of ESANN the 18th European Symposium on Artificial Neural Networks*.

Stacey, J. E., Howard, C. J., Mitra, S., & Stacey, P. C. (2020). Audio-visual integration in noise: Influence of auditory and visual stimulus degradation on eye movements and perception of the McGurk effect. *Attention, Perception, & Psychophysics*, 82, 3544-3557.

Tiippana, K. (2014). What is the McGurk effect? *Frontiers in Psychology*, 5, 725.