Corrected Human Vision System and the McGurk Effect

Ladislav Kunc and Pavel Slavík

Department of Computer Graphics and Interaction, Faculty of Electrical Engineering,
Czech Technical University in Prague
kuncladi@fel.cvut.cz, slavik@fel.cvut.cz

Abstract. The McGurk effect test is a method to evaluate articulation of talking heads. Our work addresses the issue of corrected vision influence on the McGurk effect perception. We conducted an experiment to investigate this influence. Measured data shows different perception of participants with corrected vision in some cases. The results could help talking head evaluators to compare talking head implementations each other with elimination of the influence of corrected vision.

Keywords: Talking head, McGurk effect, vision correction.

1 Introduction

Voice based user interface is one of possible human-computer interaction methods. Spoken dialog systems provide new interaction modalities for user. This usually means verbal (speech) part of communication. We should not forget that the verbal part is closely tied to nonverbal part of our language in traditional human-to-human interaction. Humans listening to speech are used to focus on words. But during complex assessment of speaking person we process both parts of speech: nonverbal and verbal one [1].

Inclusion of some form of face-to-face interaction into spoken dialog technology enables the system to express nonverbal part of communication. Seeing virtual faces further humanizes computer user interfaces and makes them more acceptable for a common user [2]. For example talking heads provide means for additional nonverbal communication.

Perceptually realistic visual articulation of a talking head is very important because humans are highly sensitive to perception of face muscles movements. The McGurk effect test is a method to evaluate speech articulation. Our work addresses the issue of corrected vision influence on the McGurk effect perception. The McGurk effect shows that humans use both hearing and vision modalities in parallel to perceive and understand speech. The first experiment was presented in [3]; there was dubbed videotape of visual *ga* syllable with audio *ba* syllable. Experiment's participants thought that *da* syllable was pronounced.

Our hypothesis is that people with corrected vision will judge the McGurk effect sequences differently than people with non-corrected normal vision. We conducted an experiment to validate our hypothesis. We prepared synthetic McGurk video sequences for participants and evaluated participants' responses.

C. Stephanidis (Ed.): Posters, Part II, HCII 2011, CCIS 174, pp. 345–349, 2011.

The next section gives an overview of related works on evaluation of talking heads articulation.

2 Related Work

The primary goal of talking head articulation research is to produce the realistic visual articulation which is indistinguishable from real human. This task includes implementation methods and in the end evaluation of results.

The quality of talking heads' visual articulation has been measured using various methods. These comprise subjective evaluation, perception of speech in noisy environments and others.

The most common method is subjective evaluation [4]. This method is based on getting comments and rating from naïve and expert participants. This provides information on quality of talking head articulation but does not give possibility to compare various talking head implementations.

The second method – perception of speech in noisy environments improves the quality of results and facilitates comparisons of miscellaneous talking head implementations. Participants try to listen to talking head footage in noisy room and the ability of understand the words indicates how much they are able to improve the intelligibility of speech by lip-reading talking head [5].

Good results provide methods based on forced choice. Participants see videos and identify which animation is real or synthetic [6].

Extensive experiments of the McGurk effect using talking head were done in [7]. Interesting method based on perception of this effect was proposed in [8]. Participants are given synthetic talking head McGurk sequences and their confusion responses are measured.

The last method was proved by judgments of participants with normal hearing and vision; not mentioned whether participants had corrected vision or not. Our work focuses on how corrected-to-normal vision influences perception of speech.

3 Experiment

This section describes details of perceptual experiment conducted with aim to evaluate how corrected-to-normal vision influences perception of the McGurk effect on a talking head. The experiment should validate or reject our hypothesis:

- *People with corrected vision will judge the McGurk effect differently than people with non-corrected normal vision.*

Participants saw stimuli videos with synthetic McGurk effect and control sequences without McGurk effect. Totally 33 stimuli videos were prepared for participants. Videos contained possible combinations of artificial face, human face, text-to-speech audio and human audio modeling syllables *ba – da – ga* (see Table 1).

The video sequences in human category were recorded by the means of usual web camera in resolution 640x480. Figure 1 illustrates what portion of face was recorded.

Table 1. Stimuli audio-video combinations prepared for participants. TTS stands for text-to-speech audio.

Visual	Sound *ba*	Sound *da*	Sound *ga*	TTS *ba*	TTS *da*	TTS *ga*
Human	Visual *ba*	Visual *ba*	Visual *ba*			
	Visual *da*	Visual *da*	Visual *da*			
	Visual *ga*	Visual *ga*	Visual *ga*			
Talking head	Visual *ba*	Visual *ba*	Visual *ba*	Visual *ba*	Visual *ba*	Visual *ba*
	Visual *da*	Visual *da*	Visual *da*	Visual *da*	Visual *da*	Visual *da*
	Visual *ga*	Visual *ga*	Visual *ga*	Visual *ga*	Visual *ga*	Visual *ga*
Black screen	N/A	N/A	N/A	N/A		N/A

Audio tracks were recorded using computer microphone headset. Recorded visual video sequences were mixed with particular audio tracks in video editing software[1].

Talking Head Application. Preparation of stimuli video sequences required application that is able to animate talking head model. ECAF toolkit application was used [9]. This toolkit displays and animates 3D talking head model of woman and produce output video sequences synchronized with either external audio file or integrated text-to-speech engine. The synchronization of audio track and phonemes animation of McGurk sequences was modified manually.

Experiment Procedure. We prepared 33 video sequences (see Table 1) generated either by ECAF toolkit or as recorded human sessions. Each sequence is about 7 seconds long.

The talking head or human repeats 6 times particular syllable *ba, da* or *ga.* The experiment was conducted remotely. Every participant observed the video sequences in his/her own computer in different but quiet environment.

The test was handled by PHP web application with embedded Adobe Flash player for video playback (see Figure 1). It was not possible for participant to scale down the video from original resolution (640x480).

The participants were both males and females in between ages of 23 to 26, university students of computer science. The whole experiment was anonymous and participants were instructed how to proceed the experiment verbally. At the beginning the web application asked participants whether they have or not have some vision correction. After this question a participant was navigated to the first video sequence. For each participant random order of video sequences was generated. The first three video sequences were training ones and the data from them was not used. Participant observed each video and fill-in the textbox with the text that he/she perceived. Each participant observed 36 (3 training ones + 33) video sequences.

3.1 Experiment Evaluation

In total 32 participants took part in our experiment (41% had corrected vision). The data from experiment were manually normalized. For example: One participant

[1] Adobe Premiere CS5.

Fig. 1. Experimental web application – a) example of human and talking head video sequence snapshots; b) the screenshot of web page with observation no. 3 which was seen by some participant. There is a text box for participant's answer under the video player window.

answered 'vaaa vaa va' and the second participant answered 'bababa'. The correct answer based on audio track should be 'ba-ba ba-ba ba-ba'. After normalization the first answer was marked as correct and the second as incorrect.

The graph in Figure 2 shows results for the McGurk sequences with the highest number of incorrect answers. Confusion video sequences are our main interest. The mean of incorrect responses for participants with normal vision is 3.89 (standard deviation 3.2). The mean of incorrect responses for participants with corrected-to-normal vision is 1.56 (standard deviation 1.85). This result denotes some differences in the McGurk effect perception by normal vision participants and participant with corrected vision. To make the results complete sequences with no confusion give mean of incorrect responses 1.2 (standard deviation 1.1).

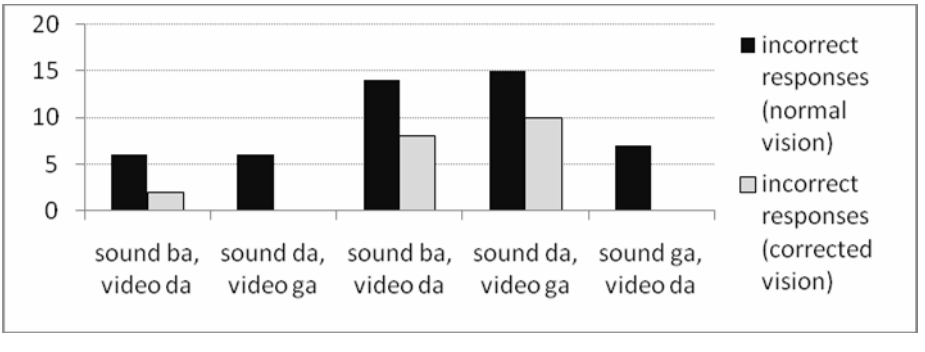

Fig. 2. Results of talking head McGurk sequences with the highest number of incorrect answers. Last three results used text-to-speech generated audio.

4 Conclusion

We conducted an experiment with the McGurk confusion audio-video sequences using talking head and human head. Our hypothesis was that people with corrected vision will judge the McGurk effect differently than people with non-corrected normal vision. The previous section presented results of our experiment and it is clearly visible that there are some differences between group of participants that had corrected-to-normal vision and participants with non-corrected normal vision.

Results of our work could help researchers that want to evaluate their developed talking heads using the McGurk perception test. They should consider the vision correction of their participants in their results and report both groups separately to be comparable.

Acknowledgement. This research has been partially supported by the MSMT under the research program MSM 6840770014. This paper has been also partially supported by the EC funded project VERITAS, contract number FP7 247765.

References

[1] Knapp, M., Hall, J.: Nonverbal Communication in Human Interaction, 7th edn. Wadsworth Publishing, Belmont (2009)
[2] Yee, N., Bailenson, J.N., Rickertsen, K.: A Meta-Analysis of the Impact of the Inclusion and Realism of Human-Like Faces on User Experiences in Interfaces. In: Proc. of CHI 2007, pp. 1–10 (2007)
[3] McGurk, H., MacDonald, J.: Hearing Lips and Seeing Voices. Nature 264, 746–748 (1976)
[4] Cosatto, E., et al.: Lifelike Talking Faces for Interactive Services. IEEE Special Issue on Human-Computer Multimodal Interface 91(2), 1406–1429 (2003)
[5] Massaro, D.W.: A Framework for Evaluating Multimodal Integration by Humans and a Role for ECAs. In: Proc. of the 6th ICMI Conference, pp. 24–31 (2004)
[6] Hack, C., Taylor, C.J.: Modelling 'Talking Head' Behavior. In: Proc. of British Machine Vision Conference, pp. 122–132 (2003)
[7] Massaro, D.W., Stork, D.G.: Speech Recognition and Sensory Integration. American Scientist 86(3), 236–244 (1998)
[8] Cosker, D., Paddock, S., Marshall, D., Rosin, P., et al.: Towards Perceptually Realistic Talking Heads: Models, Methods and McGurk. In: Proc. of the 1st Symposium on Applied Perception in Graphics and Visualization, pp. 151–157 (2004)
[9] Kunc, L., Kleindienst, J.: ECAF: Authoring Language for Embodied Conversational Agents. In: Proc. of Text, Speech, Dialog Conference, pp. 206–213 (2007)