ARCHIVES OF ACOUSTICS
33, 4, 447–454 (2008)

McGURK EFFECT IN POLISH LISTENERS

Wojciech MAJEWSKI

Wrocław University of Technolo1ogy
Institute of Telecommunications, Teleinformatics and Acoustics
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: wojciech.majewski@pwr.wroc.pl

(*received June 15, 2008; accepted November 4, 2008*)

The phenomenon of a multi-modal interaction between audio and video stimuli, called the McGurk effect, was investigated for the Polish subjects. Two experiments were performed. In the first experiment the occurrence of McGurk effect for Polish speech was examined. The results of the first experiment for which over 94% of listeners' answers were consistent with the audio stimuli indicated that generally it is very difficult to evoke the McGurk effect in Polish listeners. In the second experiment the stimuli for which the McGurk effect was distinctly present were subjected to different kind of audio or video distortions and presented to the listeners. The distortions in the visual signals resulted in an increase of the answers consistent with the audio stimuli while the distortions in the audio signals resulted in an increase of the answers consistent with the visual stimuli.

Keywords: McGurk effect, Polish speech.

1. Introduction

We are living in the world where the incoming information may be perceived by many senses. Our brain selects and connects the sensory information in order to create a distinctive picture of the surrounding environment. A problem arises when our senses perceive inconsistent stimuli. In such a case a new information, which is not in agreement with the perceived signals, may be generated.

In the case of speech perception this what we hear depends also on the information received by our eyes. The phenomenon of a multi-modal interaction between audio and visual stimuli has been observed by Harry MCGURK [4] as the result of the experiments on speech perception in infants, when the visual stimulus of the mother's face saying "ga" was synchronously presented with the aural stimulus "ba", what was perceived by the listeners as "da". From the discovery of McGurk effect in 1976 this phenomenon has been widely investigated [1–3, 5–7] in order to explain its mechanism, to establish the conditions under which it occurs and to find out if it appears in the same degree in any language.

The present study consists of two experiments. In the first experiment the occurrence of the McGurk effect for Polish speech was examined. In the second experiment, the stimuli for which the largest number of answers that were not in agreement with either audio or video stimuli was obtained, i.e. the stimuli for which the McGurk effect was distinctly present, were subjected to different kind of audio or video distortions to find out what kind of influence on McGurk effect such distortions may have. This should permit to determine what influence has McGurk effect in multi-modal telecommunication systems, where some distortions related to audio and video signals (e.g. a time delay between audio and video signals, echo or noise in speech signal) may appear.

2. Experimental procedure

The test material consisted of 25 utterances produced three times by 10 speakers (males and females).The list of test utterances is presented in Table 1. The consonants (positions 19–25) were spoken with the accompanying “schwa”.

Table 1. The list of test utterances.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
ba	ga	da	ka	na	ta	ag	ak	ek	eg	be	de	ke	te	aba	aka	ada	ata	p	k	t	d	w	f	g

The recordings were made in an ordinary room in a very good day light. The speakers’ faces were recorded with DVCAM SONY DSR PD-150P camera and compressed by means of Indeo® Video 5.1 codec. The speech samples were exported to separate sound files and recorded in PCM format with 16 bit resolution and 44.1 kHz sampling frequency. Next, utilizing Adobe Premiere Pro program, each video file was synchronously combined with each audio file.

In the experiments the audio-visual pairs containing the same number of phonemes were used. Since, however, some of these audio-video pairs indicated a striking inconsistence of the speech signal with the moving lips, the selected 167 audio-video pairs of stimuli, presented in Table 2, have been finally used to examine the occurrence of McGurk effect in Polish. Since each of the selected audio-visual pairs was produced by 10 speakers, a total amount of 1670 different audio-video stimuli was examined.

In the first experiment randomized pairs of audio-video stimuli were presented to a group of 12 Polish subjects of normal sight and hearing. The audio-visual sessions were carried out in an acoustic laboratory. The subject was sitting in a 75 cm distance from a 21″ monitor where a natural size head of the speaker was presented. At the same time the subject was hearing the speech sample from a good quality loudspeaker system. Each audio-visual pair of stimuli was presented three times and next there was a few seconds break to give the subject time to write down in a special answer sheet what he has heard and to mark his opinion in the case of a substantial inconsistence between sound and vision in a three dimensional scale. The experiment was carried out during 10 sessions divided into 15 minutes subsessions. After each subsession a 10 minutes

break took place to give the subjects time for a rest. Thus, each session, including breaks, lasted about four hours.

Table 2. The list of audio-visual pairs of stimuli.

No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
video	ba	ba	ba	ba	ba	ba	ba	ba	ba	ba	ba	ba	ba	ga	ga	ga	ga
audio	ga	da	ka	na	ta	ag	ak	ek	eg	be	de	ke	te	ba	da	ka	na
No.	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
video	ga	ga	ga	ga	ga	ga	ga	ga	ga	da	da	da	da	da	da	da	da
audio	ta	ag	ak	ek	eg	be	de	ke	te	ba	ga	ka	na	ta	ag	ak	ek
No.	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51
video	da	da	da	da	da	ka	ka	ka	ka	ka	ka	ka	ka	ka	ka	ka	ka
audio	eg	be	de	ke	te	ba	ga	da	na	ta	ag	ak	ek	eg	be	de	ke
No.	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68
video	ka	na	na	na	na	na	na	na	na	na	na	na	na	na	ta	ta	ta
audio	te	ba	ga	da	ka	ta	ag	ak	ek	eg	be	de	ke	te	ba	ga	da
No.	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85
video	ta	ta	ta	ta	ta	ta	ta	ta	ta	ta	ag	ag	ag	ag	ag	ag	ag
audio	ka	na	ag	ak	ek	eg	be	de	ke	te	ba	ga	da	ka	na	ta	ak
No.	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102
video	ag	ag	ag	ag	ag	ag	ak	ak	ak	ak	ak	ak	ak	ak	ek	ek	ek
audio	ek	eg	be	de	ke	te	ba	ga	da	ka	na	ta	ag	ek	ba	ga	da
No.	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119
video	ek	ek	ek	ek	ek	eg	eg	eg	eg	eg	eg	eg	eg	be	be	be	be
audio	ka	na	ta	ag	eg	ba	ga	da	ka	na	ta	ag	eg	ba	ga	da	ka
No.	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136
video	be	be	be	be	de	de	de	de	de	de	de	de	ke	ke	ke	ke	ke
audio	na	ta	ag	de	ba	ga	da	ka	na	ta	ag	be	ba	ga	da	ka	na
No.	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153
video	ke	ke	ke	te	te	te	te	te	te	te	te	aba	aka	aka	aka	ada	ada
audio	ta	ag	te	ba	ga	da	ka	na	ta	ag	ke	aka	aba	ada	ata	aka	ata
No.	154	155	156	157	158	159	160	161	162	163	164	165	166	167			
video	ata	ata	p	p	k	k	t	t	d	w	f	f	g	g			
audio	aka	ada	k	t	p	g	p	d	t	f	w	g	k	f			

After the first experiment was completed and the results analyzed, the second experiment was carried out. First, on the basis of the results obtained in the first experiment, such pairs of audio-video stimuli have been selected for which 50% or more answers of the listeners were not in agreement with both audio and video stimuli. 9 pairs of such stimuli, which are presented in Table 3, were next subjected to 20 different audio or video modifications and presented to 10 subjects, five of which were new subjects and the remaining five participated also in the first experiment. The experimental procedure

was almost the same as in the first experiment. Three repetitions of a given audio-video stimulus were presented to the subject, who next had few seconds to write down what he has heard. The only difference was that the opinion on the inconsistency between sound and vision was not collected. The second experiment was run during one session, because the selected 9 stimuli were presented to the subjects under 21 different conditions (one without any modification plus 20 modifications realized by Audacity or Adobe Premiere Pro) giving in sum 189 different audio-video stimuli.

Table 3. The test material selected for the second experiment.

speaker	no. 6	no. 6	no. 6	no. 7	no. 7	no. 7	no. 8	no. 9	no. 10
video	da	na	eg	ga	ke	ke	na	be	ta
audio	ba	ba	ba	be	ba	ag	ba	ta	ba

The experimental procedure was executed within the diploma work [8] supervised by the author of the present paper.

3. Results and discussion

The total results of the first experiment are presented in Table 4 (as the percentage of the listeners' answers consistent with video stimuli, audio stimuli or inconsistent with any of them. The inconsistent answers are considered as being caused by the McGurk effect. As it may be seen from the data presented in Table 4 the prevailing majority of answers (over 94% of 20400 stimuli) was consistent with the acoustic signal heard during the experiment. Only 1% of the answers was consistent with the visual signal. The McGurk effect was observed in roughly 4.5% of the examined cases. This result is far away from that what was expected and it suggests that under adopted measuring conditions it was very difficult to obtain the McGurk effect in Polish listeners.

Table 4. The overall distribution of answers in the first experiment (in percent).

Answers consistent with audio stimuli	Answers consistent with video stimuli	Answers inconsistent with audio and video stimuli
94.35	1.04	4.61

It is necessary to add that in the case of inconsistent answers there was a large dispersion in the listeners' answers ranging from one to six different variants of answers. The distribution of listeners' answers for given number of variants of inconsistent answers is presented in Table 5. From the data presented in this Table it may be seen that in 32% of cases the listeners perceived the stimuli in two variants, in 27% of cases in three variants and in 19% of cases in one variant only. These results show an individual

Table 5. The percent of listeners' answers for given number of variants of inconsistent answers.

Number of variants	1	2	3	4	5	6
Percent of answers	19	32	27	15	6	1

reaction to the presented stimuli and indicate that it is difficult to select one variant of answers confirming the McGurk effect in Polish listeners.

The distribution of answers for particular listeners is presented in Table 6 and the distribution of answers for particular speakers is presented in Table 7. The distribution of answers for particular listeners is very similar. The only exception was the subject no. 6 with substantially larger amount of answers consistent with the visual stimuli and inconsistent with both stimuli what may indicate that he belongs to so called "visual-izers". As far as the speakers are concerned (see Table 7), large amount of inconsistent answers was obtained for speaker no. 6, 7 and 9, and because of that the utterances of these speakers were utilized in the second experiment.

The results of the second experiment are summarized in Table 8. Since the second experiment was carried for 9 specially selected pairs of audio-video stimuli (see Table 3) the percentage of the inconsistent answers (i.e. the McGurk effect) was much higher in

Table 6. The distribution of answers for particular listeners (in percent).

Listener number	Answers consistent with audio stimuli	Answers consistent with video stimuli	Answers inconsistent with audio and video
1	94.13	0.60	5.27
2	93.29	1.38	5.33
3	95.03	0.60	4.37
4	95.81	0.36	3.83
5	93.89	0.90	5.21
6	85.27	4.97	9.76
7	96.65	0.54	2.81
8	94.19	0.42	5.39
9	97.13	0.48	2.40
10	94.19	0.78	5.03
11	96.65	0.42	2.93
12	94.49	0.90	4.61

Table 7. The distribution of answers for particular speakers (in percent).

Speaker number	Answers consistent with audio stimuli	Answers consistent with video stimuli	Answers inconsistent with audio and video
1	96.21	0.90	2.89
2	96.81	1.00	2.20
3	96.11	1.10	2.79
4	96.76	0.65	2.79
5	96.41	1.15	2.45
6	88.97	0.95	10.08
7	87.82	1.65	10.53
8	96.16	0.75	3.09
9	90.37	1.50	8.13
10	96.66	0.80	2.54

comparison to the results obtained in the first experiment and the overall mean of the inconsistent answers reached almost 44%. Substantially larger was also the number of answers consistent with the visual stimuli. Moreover, the percentage of the inconsistent answers for the audio-video stimuli without any modification was even higher (almost 48%) what indicates that an increase of the McGurk effect was mainly caused by the audio-video stimuli applied in the second experiment, i.e. by a proper selection of the phonetic material and a selection of the speakers.

Table 8. The distribution of answers for particular modifications of audio and video stimuli (in percent).

Modification	Answers consistent with audio stimuli	Answers consistent with video stimuli	Inconsistent answers
Low pass filter – 2000 Hz	32.22	8.89	58.89
Low pass filter – 600 Hz	21.11	8.89	70.00
Echo	55.00	5.00	40.00
Pink noise – amplitude 0.02	32.22	7.78	60.00
Pink noise – amplitude 0.05	25.56	10.00	64.44
Cracks	17.78	10.00	72.22
Picture turn – 90 degrees	47.78	4.44	47.78
Picture turn – 180 degrees	53.33	4.44	42.22
Dimness – 50%	53.33	4.44	42.22
Dimness – 80%	57.78	4.44	38.89
Picture duplication – 4 times	44.44	4.44	51.11
Picture duplication – 9 times	56.67	3.33	40.00
Picture reeling	68.89	5.56	25.56
Smeary picture – 10%	54.44	3.33	42.22
Smeary picture – 20%	65.56	2.22	32.22
Snowy picture – 50%	52.22	4.44	43.33
Snowy picture – 100%	70.00	4.44	25.56
Time delay audio-video – 350 ms	57.78	3.33	38.89
Time delay audio-video – 450 ms	76.67	1.11	22.22
Time delay audio-video – 500 ms	85.56	1.11	13.33
No. modification	45.56	6.67	47.78
Overall mean	51.12	5.11	43.77

As far as the influence of particular modifications is concerned some general trends may be observed in comparison to the “no modification” case. Audio distortions resulted in an increase of the inconsistent answers and in a decrease of the answers consistent with the audio stimuli. The only exception was the echo with a reverse influence. On the other hand, video distortions resulted in an increase of the answers consistent with the audio stimuli and in a decrease of the inconsistent answers. Similar trend was observed for a time delay between audio and video stimuli. For large values of time delay (450 and 500 ms) an increase of the answers consistent with the audio stimuli was especially

large and a decrease of the inconsistent answers and the answers consistent with the video stimuli was also very large.

The results obtained in the second experiment for each of 9 examined pairs of stimuli were also analyzed. The mean results for all modifications of the stimuli applied in the second experiment obtained for each pair of audio-video stimuli are presented in Table 9. Looking at these results it may be easily seen that almost all the answers are consistent with the aural stimuli or inconsistent with the aural and visual stimuli. Only for the pair "da – ba" spoken by speaker no. 6 there was a large amount of answers consistent with the visual stimuli. The answers for the 8 remaining pairs of stimuli are distributed almost evenly among the answers consistent with audio stimuli or inconsistent with audio and video.

Table 9. Distribution of answers for particular pairs of audio-video stimuli (in percent).

Speaker number	Video stimulus	Audio stimulus	Answers consistent with audio stimuli	Answers consistent with video stimuli	Inconsistent answers
6	da	ba	37.1	42.4	20.5
6	na	ba	44.3	0.0	55.7
6	eg	ba	42.9	0.0	57.1
7	ga	be	49.0	0.0	51.0
7	ke	ba	63.3	0.0	36.7
7	ke	ag	76.7	0.0	23.3
8	na	ba	48.1	0.0	51.9
9	be	ta	41.4	0.0	58.6
10	ta	ba	57.0	1.0	42.0

Finally, the variants of inconsistent answers obtained in the second experiment were analyzed. The variants of inconsistent answers and the number of answers for each of the variants are presented in Table 10. The number of these variants is restricted to 8 syllables and the number of answers for particular syllables is spread from few answers to more than 100. Thus, considering the set of examined pairs of audio-video stimuli and the applied modifications, the most distinctive cases of the McGurk effect in

Table 10. Number of answers for particular variants of inconsistent answers.

Speaker	Video	Audio	da	de	pa	ba	be	ak	ta	ga
6	da	ba	0	5	3	0	34	0	1	0
6	na	ba	65	9	0	0	40	0	3	0
6	eg	ba	62	9	1	0	44	0	1	3
7	ga	be	0	107	0	0	0	0	0	0
7	ke	ba	62	1	6	0	6	0	2	0
7	ke	ag	0	09	0	4	8	37	0	0
8	na	ba	103	2	1	0	0	1	0	2
9	be	ta	11	0	99	13	0	0	0	0
10	ta	ba	80	0	3	0	1	0	0	0

Polish listeners under examined conditions seems to be the audio-visual pair of stimuli "ga – be" perceived in inconsistent answers unanimously as "de" and the audio-visual pair of stimuli "na – ba" perceived almost unanimously as "da".

4. Conclusions

The results of the first experiment for which over 94% of answers were consistent with the audio stimuli indicate that generally it is very difficult to evoke the McGurk effect in Polish listeners. Thus, the McGurk effect in Polish seems to be a marginal phenomenon and there is no danger to distort the perception of complex audio-visual signals. Since, however, the McGurk effect depends on the interaction between aural and visual stimuli, it was possible to obtain – by a selection of phonetic material and the speakers – an increase in the McGurk effect to over 40% of the examined stimuli.

The results of the second experiment indicated also that distortions in audio and video signals have an influence on the perception of audio-video stimuli. The distortions in the visual signals result in an increase of the answers consistent with the audio stimuli, while the distortions in the aural signals result in an increase of the answers consistent with the visual stimuli.

Acknowledgments

This work was partially supported by COST Action 2102 "Cross-modal Analysis of Verbal and Non-verbal Communication" and by the grant from the Minister of Science and Higher Education (decision no. 115/N-COST/2008/0).

The extended version of the paper submitted for the 55th Open Seminar on Acoustics.

References

[1] BRANCAZIO L., MILLER J., PARE M.A., *Visual influence of the internal structure of phonetic categories*, Perception & Psychophysics, **65**, 591–601 (2003).

[2] GREEN K., *Face and mouth inversion effects on visual and audiovisual speech perception*, J. Acoust. Soc. Amer., **97**, 3286 (1995).

[3] MASSARO D.W., STORK D.G., *Visual influences on speech perception process*, American Scientist, **86**, 236–244 (1998).

[4] MCGURK H., MACDONALD J., *Hearing lips and seeing voices*, Nature, **264**, 746–748 (1976).

[5] MUNHALL K.G., GRIBBLE P., SACCO L., WARD M., *Temporal constrains on the McGurk effect*, Perception & Psychophysics, **58**, 351–362 (1996).

[6] SEKIYAMA K., TOKHURA Y., *McGurk effect in non-English listeners; Few visual effects of Japanese subjects hearing Japanese syllables of high auditory intelligibility*, J. Acoust. Soc. Amer., **91**, 1797–1805 (1991).

[7] SUKAMOTO S., MISHIMA H., SUZUKI Y., *Effects of consonance of a voice and talking face motion on the McGurk effect*, Subjective and Objective Assessment of Sound, Poznań, 1–3 September 2004, SOAS Proceedings – CD.

[8] ŻOŁYŃSKI M., *Sight and hearing integration in speech understanding process* [in Polish], Master's Thesis, Wrocław University of Technology, 2007.