RECONSIDERING THE MCGURK EFFECT

Vesna Mildner and Arnalda Dobrić

Faculty of Humanities and Social Sciences, Department of Phonetics, University of Zagreb, Croatia
vmildner@ffzg.hr and adobric@ffzg.hr

ABSTRACT

The McGurk effect was studied on 3 groups of subjects differing in age: pre-schoolers (mean: 6 years), elementary-school 4th-graders (mean: 11 years) and adults (mean: 24 years). The stimuli included audio-visual combinations of syllables /pa/, /ta/, /ka/, /ba/, /da/, /ga/. The combination of auditorily presented (A) bilabials with visually presented (V) velars, was reported by some subjects as a third (dental, e.g. /da/) plosive, and in the reversed-modality combination as a bilabial+velar response (e.g. /bga/), manifesting the McGurk effect. In all groups, dental–velar pairs and Adentals combined with Vbilabials elicited close to 100% responses corresponding to the A stimuli. Vdentals combined with Abilabials elicited an unexpectedly high proportion of responses corresponding to the visual stimuli. The McGurk effect seems less robust than commonly reported and exhibits individual variation. Group differences were inconsistent and the effect strength could not be attributed to age.

Keywords: McGurk effect, age, multisensory perception, audio-visual processing

1. INTRODUCTION

Speech perception involves attending simultaneously or in very close sequence to numerous cues, frequently in less than perfect conditions (e.g. in noise, over distances, in a less familiar language, with poor hearing). Apart from the high degree of redundancy in the speech signal itself this is aided by the multisensory nature of communication process. Although the auditory modality is primary in speech perception, it is well established that visual information can greatly contribute to intelligibility and processing speed, particularly in unfavourable conditions. In foreign language learners and hearing-impaired individuals communication is also facilitated by combinations of different modalities, primarily auditory and visual [8]. It has developmental significance as well – typically developing children benefit from the congruence of these modalities in the process of speech acquisition [11]. On the other hand, age differences in the development of multisensory integration have been reported. Burr and Gori [3] discuss the results of previous studies on audio-visual integration indicating that children are responsive to non-speech illusions by age 5, but take longer to develop their ability to perceive audio-visual integration in speech stimuli, and corroborate them with their own research, concluding that children reach adult-like cross-modal integration ability around age 10–11. Sekiyama *et al.* [13] reported age-dependent differences even in adulthood, with older adults relying on the visual channel more than the younger ones.

Research methods, such as event related potentials (ERP), functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG), positron emission tomography (PET) and transcranial magnetic stimulation (TMS), have enabled search for the site(s) of multisensory integration or crossmodal processing in our brain. Prevailing evidence points to the left superior temporal sulcus as the crucial area in audio-visual integration [1, 5, 7, 10]. That area is also part of a network involving other brain regions whose components are differently specialized for integration of different modalities [4].

A behavioral example of crossmodal (specifically audio-visual) integration is the McGurk effect [9]. It is a perceptual phenomenon that occurs when subjects are presented with combinations of incongruous auditory-visual stimuli: (a) A video clip of a human face pronouncing a syllable that represents the combination of visually presented /ga/ simultaneously with the auditorily presented /ba/ elicits /da/ responses (fusion); (b) A combination of V/ba/ simultaneously with the A/ga/ elicits /bga/ responses (combination). Disorders, such as dyslexia, specific language impairment, language-learning disabilities, Alzheimer's disease, aphasia, may diminish the effect. Its strength may vary across languages and cultures and may be affected by experience in watching dubbed programs [2, 6].

The McGurk effect is supposedly a robust phenomenon that occurs automatically and persists despite possible onset asynchronies: the timeframe within which the effect is present is between 60 ms preceding and 180 – 250 ms or more lagging of the auditory stimulus with respect to the visual one [1, 5]. However, recently there has been some evidence

that the McGurk effect is neither as robust nor as automatic as it has been generally claimed [1, 3, 10].

The aim of this study was to examine the occurrence and strength of the McGurk effect in 3 age groups (mean age: 6, 11 and 24 years). Absence of the effect in the youngest group and/or significant increase in effect strength with increasing age would support the developmental hypothesis.

2. METHOD

2.1. Speaker and stimuli

A male speaker (age 42) was recorded (camera Sony DSR PD150P; MiniDV cassette) pronouncing syllables /pa/, /ta/, /ka/, /ba/, /da/, /ga/ in a sound-proof studio, directly facing the camera. Video clips were rendered with a digitization rate of 25 frames per s, with a 720x576 resolution. Sound was digitized at 48,000 Hz, 16-bit resolution. The material was edited on Dell Precision T3500 computer by Adobe Premier Pro CS4 software and exported in Microsoft DV avi format. A total of 36 test clips were produced, which comprised all possible audio-visual combinations of the six syllables. There were 6 audio-visually congruent stimuli (e.g. A/ga/–V/ga/), 6 stimuli sharing the same place of articulation, but differing in [±voice] (e.g. A/ga/–V/ka/) and 24 audio-visually incongruent stimuli (6 types of combination with 4 different combinations in each:

- Adental–Vvelar (e.g. A/ta/–V/ka/),
- Avelar–Vdental (e.g. A/ga/–V/ta/),
- Adental–Vbilabial (e.g. A/da/–V/pa/),
- Abilabial–Vdental (e.g. A/ba/–V/da/),
- Abilabial–Vvelar (e.g. A/ba/–V/ga/),
- Avelar–Vbilabial (e.g. A/ka/–V/pa/).

Each clip was approximately 1.8 s long and contained one stimulus, with 500 ms before the start of the stimulus and 1 s after, rendering signals approximately 300 ms long. The clips were compiled in random order with 4 s pauses between them. A 10-item practice session preceded the test.

2.2. Subjects

Three groups of subjects participated in this study: 20 pre-school children (PS), mean age: 6 years; 21 elementary-school 4th graders (S), mean age: 11 years; and 23 adults (G), mean age: 24 years. The two younger groups were recruited in their (pre)school institutions upon approval of (pre)school authorities and obtaining of parental consent. Adults were university students of phonetics and individuals affiliated with the department. All participants reported normal hearing and normal or corrected vision. All children were typically developing.

2.3. Procedure

Groups G and S were tested in groups of ten and entered their responses in a response sheet provided by the experimenters. The stimuli were projected on the screen mounted on the wall in a classroom, with loudspeakers placed below the screen. Due to their young age the PS group were tested individually. Stimuli were presented on computer screen directly in front of them, with the sound source aligned with the screen. Their oral responses were written down by the experimenter and audio-recorded for later confirmation. All participants were instructed to look at the screen and listen, and then write down (or repeat) "what the man in the video said". They were encouraged to respond no matter how 'weird' the stimulus was. The training session lasted 1 m 1 s, the test 3 m 43 s.

2.4. Analysis

In order to be certain that the responses to incongruent stimuli were in fact a result of the test condition we included in the analysis only participants who had 100% correct responses to the 6 congruent stimuli. We also omitted participants whose responses were not clear or who gave more than one response per stimulus. Of the original 28 adult participants 2 had to be omitted because of less than perfect score on congruent stimuli and 3 for other reasons. Of the original 37 school-age children 13 did not have 100% correct responses to congruent stimuli and 3 were omitted for other reasons. Of the original 59 pre-schoolers we could use only 20 based on the 100% correct response to congruent stimuli criterion.

Level of significance was set to 95% ($p < 0.05$) and all confidence intervals were given on the 95% level. In all instances, two-tailed tests of statistical significance were used. Wherever the samples were smaller than $n = 30$ exact or Monte Carlo tests of statistical significance were used instead of asymptotic ones. Normality of distribution was tested with Shapiro–Wilk test. As the measure of central tendencies median and interquartile range were used wherever the distribution statistically significantly deviated from the normal one. Analysis of differences in medians was analyzed by Kruskal Wallis and Mann–Whitney U test for independent or Friedman and Wilcoxon test for dependent variable relations. All analyses were carried out using SPSS 17.0 (SPSS Inc., Chicago, IL, USA) statistical software package.

3. RESULTS AND DISCUSSION

As mentioned in the section on methods, the qualifying requirement for further analysis was 100% correct responses on the six A–V congruent stimuli, so we can reiterate here that all participants had 100% score on A–V congruent stimuli.

With respect to the six stimuli that comprised A–V pairs with shared place of articulation differing only in [±voice], the two older groups (G and S) had 100% A responses. In the youngest group, 7/20 children had one non-A response each, with no noticeable consistency among their responses: 3 missed place of articulation and 4 gave V responses across different places of articulation. Of these 4, two preferred the V channel when it was [-voice] and two preferred it when it was [+voice] compared with the A channel.

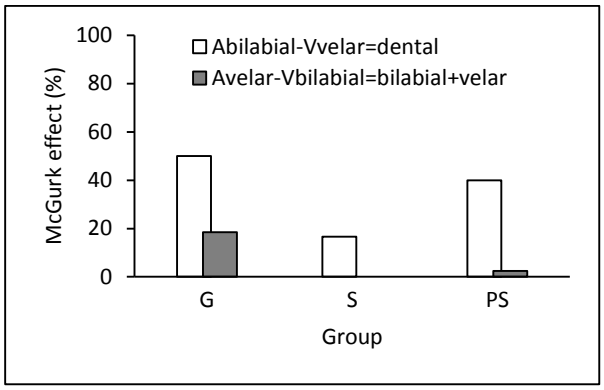
In the A–V incongruent stimuli comprising Adental–Vvelar and Avelar–Vdental combinations, all subjects preferred the auditory channel. Their A responses ranged from 97% to 100%.

The two possible bilabial–dental A–V incongruent combinations behaved somewhat differently. The responses to the Adental–Vbilabial combination revealed a preference for the auditory channel (between 73% and 95% A responses), albeit not as strong as in the above combinations. This increase in V responses may be attributed to the visibility of bilabial articulation which competes for attention with the auditorily presented, but less clearly discernible dentals. Due to small variance, the significance of differences among groups could not be determined, but it may be seen that the two younger groups rely more on the A channel than the oldest group (73% A and 12% V responses in G group, vs 94% A and 5% V responses in PS group and 95% A and 4% V responses in S group). This is a much higher proportion of A responses than the 10% reported by Rouger *et al.* [11] It should be stressed here that 3/23 adults in the G group contributed to the total of 15% "other/0" responses, i.e. neither A, nor V. All 3 of them exhibited the "combination" variant of the McGurk effect, reporting bilabial+dental, i.e. /bda/ or /pta/. Similarly to [9], the feature [±voice] corresponded to the A channel in all responses (e.g. in case of A/ta/–V/ba/ the responses were /pta/, and in case of A/da/–V/pa/ the responses were /bda/. This occurrence of the McGurk effect in adults (albeit in only 13%) and in neither of the younger groups may be an indication of the age dependence of crossmodal integration. However, it is considerably lower than the 82% reported by Rouger *et al.* [11].

The Abilabial–Vdental combination elicited a significantly higher proportion of V responses than other combinations and, accordingly, significantly fewer A responses (both $p = 0.000$). A *post hoc* pairwise test of between-combination differences revealed the same level of significance for all compared combinations in the V channel ($p = 0.000$). In the A channel, only the difference between this and the Abilingual–Vvelar combination was not significant ($p = 0.738$). All other between-combination differences were significant ($p = 0.000$). In this combination, in both channels there were significant group differences (A channel: $p = 0.01$; V channel: $p = 0.025$). *Post hoc* analysis revealed that the difference in both modalities was significant between G and S groups ($p \leq 0.010$). In the two younger groups the A channel was still preferred, but in a considerably lower proportion than in other combinations (S: 73% A and 26% V responses; PS: 59% A and 36% V responses). Among the "other/0" responses there were none that could be classified as examples of McGurk effect. The adult group actually preferred the visual channel (38% A vs 58% V responses). Two G participants were responsible for the 4% "other/0" responses. One had one "0" response, and the other had three responses that could be classified as the "combination" variant of the McGurk effect, i.e. /bda/ or /pta/. Again, feature [±voice] corresponded to the A channel. These results differ considerably from the 80% reported McGurk effect and 20% auditory preference by Rouger *et al.* [11]. It seems counterintuitive that subjects should report so many V responses given the fact that the less visible dentals were presented through this channel and the clearly visible bilabials were presented through the A channel. A possible explanation may be that hearing a bilabial was not a strong enough incentive to report it when it was clearly not present in the V channel. The dominance of the V channel and the occurrence of the traces of McGurk effect in adults but not in younger groups indicate stronger crossmodal integration, as opposed to children's still higher reliance on the A channel. Besides, both A–V incongruent bilabial–dental combinations, regardless of channel, corroborate the notion that the effect does not occur only in the bilabial–velar combination [11, 14].

The A–V incongruent combination commonly referred to as the source of the McGurk effect is the bilabial–velar one, manifested as a reported dental (i.e. fusion) in case of Abilabial–Vvelar combination and a reported bilabial+velar (i.e. combination) in case of Avelar–Vbilabial stimulus. Literature range of the effect occurrence is between 26% and 98% [3, 10]. Given the wide range, it is not surprising that our data fit. McGurk effect proportions for these two combinations are presented in Fig. 1.

Figure 1: Proportion of McGurk effect across groups in bilabial–velar combinations.



In the Abilabial–Vvelar combination, 16/23 G (70%) gave the dental response in 25% to 100% occurrences (5 had 100%). PS had the next highest score: 13/20 (65%) reported a dental in 25% to 100% occurrences (1 had 100%). Only 8/21 S (38%) responded with a dental (no 100% responses). Inspection of all supplied responses across groups reveals that in the G group, McGurk effect responses were the highest (50%), followed by slightly fewer A responses (48%). In both younger groups, A responses predominated (S: 70% A vs 17% dental; PS: 51% A vs 40% dental). The proportion of V responses was between 1% (G and S) and 4% (PS). The proportion of "0" responses was between 1% (G) and 12% (S), with PS between them at 5%. The only significant group difference was in proportion of dental responses ($p = 0.012$); *post hoc* analysis revealed significant differences between S and G ($p = 0.005$) and between S and PS ($p = 0.028$).

In the Avelar–Vbilabial combination, all 3 groups reported predominantly A responses (G: 79%; S: 100%; PS: 98%). McGurk effect, manifested as bilabial+velar response, was found only in 2 groups (G: 18%, PS: 3%). In the G group, 6/23 (26%) participants reported the effect (2 had 100% response) and in the PS group only 1/20 (5%) reported the effect in 50% of combination occurrences. Due to insufficient variability within groups statistical significance could not be calculated. However, it can be seen from both these combinations, possible sources of the McGurk effect, that adults are more susceptible to it than children, and that there are no consistent differences between the 2 groups of children – if anything, the younger group behaves more like the adult group than the older children, which is contrary to the age-dependence hypothesis.

In a similar study, run with comparable subjects, we tested the responses to the same stimuli as in this study in V-only and A-only conditions. In A-only condition, all age groups recognized with above 75% accuracy place of articulation of all stimuli (the two older groups' responses ranged between 98% and 100%). Actual stimuli (including [±voice]) were recognized with above 84% accuracy by the 2 older groups, but with considerable drop in accuracy in the youngest group, who clearly preferred the voiceless variants, that were recognized with above 81% accuracy as opposed to 25–62% accuracy for the voiced stimuli. In V-only condition the bilabials were recognized with above 76% accuracy by all ages. For the 2 older groups, dentals were somewhat easier than velars (between 63% and 86% vs 41–88% correct, respectively). Both categories were equally difficult for the youngest group, and the proportion of correct responses ranged between 30% and 39% across categories. As expected, recognizing stimuli correctly including [±voice] was even more difficult. Across all groups, the proportion of correct responses ranged between 10% and 62%. Clearly, auditory modality is preferred in speech perception and success in visual modality is, not surprisingly, highly dependent on stimulus visibility. Consistently higher scores in older groups (i.e. PS < S < G) may be attributed to longer experience with visual processing of speech and this may be taken as indication that the contribution of the V channel increases with age and exposure to A–V stimuli, as discussed elsewhere [13, 15].

4. CONCLUSIONS

This study suggests that the McGurk effect is not entirely automatic and requires some attention, as suggested by Alais *et al.* [1]. We have also found that it is not as robust as reported by some authors (e.g. [3, 11]) and that there is considerable individual variability – responses vary from subjects who never report it to those who report it always (compare with [10]). Its age dependence has been only partly confirmed, with adults being generally more susceptible to the effect than children, but with no clear age-related differences between the two younger groups, who tend to rely more on the auditory channel (compare with [3, 10, 13]).

5. ACKNOWLEDGEMENTS

We wish to thank parents, children and educators in pre-school *Sunce* and elementary school *Brezovica*, especially D. Pinter, M. Uvalić, B. Pauković and M. Varžić for their participation. Thanks to J. Bićanić, K. Ivanković and G. Koletić for help in recording, preparation of the stimuli and statistics.

6. REFERENCES

[1] Alais, D., Newell, F. N., Mamassian, P. 2010. Multisensory processing in review: from physiology to behavior. *Seeing and Perceiving* 23, 3–38.

[2] Boersma, P. 2011. A constraint-based explanation of the McGurk effect. http://www.fon.hum.uva.nl/paul/papers/McGurk.pdf (retrieved January 15, 2015).

[3] Burr, D., Gori, M. 2012. Multisensory integration develops late in humans. In: Murray, M. M., Wallace, M. T. (eds), *The Neural Bases of Multisensory Processes*. Boca Raton, FL: Taylor & Francis Group, 345–362.

[4] Calvert, G. A. 2001. Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex* 11, 1110–1123.

[5] Calvert, G. A., Thesen, T. 2004. Multisensory integration: methodological approaches and emerging principles in the human brain. *Journal of Physiology–Paris* 98, 181–205.

[6] Hisanaga, S., Sekiyama, K., Igasaki, T., Murayama, N. 2009. Audiovisual speech perception in Japanese and English: Interlanguage differences examined by event-related potentials. http://www.isca-speech.org/archive_open/avsp09/papers/av09_038.pdf (retrieved January 15, 2015).

[7] King, A. J., Calvert, G. A. 2001. Multisensory integration: Perceptual grouping by eye and ear. *Current Biology* 11, R322–R325.

[8] Kushnerenko, E., Tomalski, P., Bailleux, H., Ribeiro, H., Potton, A., Axelsson, E. L., Murphy, E., Moore, D. G. 2013. Brain responses to audiovisual speech mismatch in infants are associated with individual differences in looking behaviour. *European Journal of Neuroscience* 38, 3363–3369.

[9] McGurk, H., MacDonald, J. 1976. Hearing lips and seeing voices. *Nature* 264 (5588), 746–748.

[10] Nath, A. R., Beauchamp, M. S. 2012. A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *Neuroimage* 59, 781–787.

[11] Rouger, J., Frayse, B., Deguine, O., Barone, P. 2008. McGurk effects in cochlear-implanted deaf subjects. *Brain Research* 1188, 87–99.

[12] Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., Barone, P. 2007. Evidence that cochlear-implanted deaf patients are better multisensory integrators. *PNAS* 104, 7295–7300.

[13] Sekiyama, K., Soshi, T., S. Sh. 2014. Enhanced audiovisual integration with aging in speech perception: a heightened McGurk effect in older adults. *Frontiers in Psychology* 5, Article 323, 1–12.

[14] Tiippana, K. 2014. What is the McGurk effect? *Frontiers in Psychology* 5, Article 725, 1–3.

[15] Tremblay, C., Champoux, F., Voss, P., Bacon, B. A., Lepore, F., Theoret, H. 2007. Speech and non-speech audio-visual illusions: A developmental study. *PLoS ONE* 2(8): e742.