McGurk effect persists with a partially removed visual signal

Christian Kroos, Ashlie Dreves

MARCS Auditory Laboratories, University of Western Sydney, Australia

c.kroos@uws.edu.au

Abstract

This study investigated the impact of prolonged consonant durations on the McGurk effect and the role of static visual information in this cross-modal illusion. It was found that the McGurk effect did not break down nor diminish in strength when the consonant duration was extended up to 3.8s. This remained unchanged when the video frames in the quasi-static phase of the prolonged target consonant were substituted with copies of a single frame. Surprisingly, the McGurk effect even persisted when the visual signal in the quasi-static phase was completely removed.

Index Terms: McGurk effect, auditory-visual speech, consonant duration

1. Introduction

The discovery of the McGurk effect in 1976 [1] has opened an experimental window into the mechanism underlying the combined use of auditory and visual information in speech. In the three decades since then a wide range of research questions concerning the McGurk effect has been addressed (e.g., [2, 3, 4, 5, 6, 7, 8]), whereby the investigation how the effect is sensitive to auditory and visual temporal alignment has received the most attention. The role of static and dynamic information has been investigated to a much lesser degree (but see [9, 10, 11]) arguably due to the tacit assumption that the McGurk effect is dynamic by its very nature. The argument holds indeed insofar as quasi-stationary phases do usually not exist in the consonants and vowel transitions that elicit the effect, at least not in normal everyday-like speech. This changes, however, if the duration of the target consonant is prolonged. If the usual auditory /ba/ is used, the amount by which the consonant can be stretched in a natural way by a real speaker is very limited since the voicing cannot be maintained for longer durations; if the unvoiced counterpart counterpart /pa/ is employed, longer durations are possible but have no acoustic consequence other than a prolonged silence. The bilabial nasal /m/, however, does not pose such problems and is perceived with the McGurk effect as /n/.

By using bilabial nasals and substantially extended consonant durations, this study investigated the role of static and dynamic visual information taking advantage of the fact that the visual part of the resulting quasi-stationary phase of these consonants can be easily manipulated. In particular, the following questions were examined:

1. Does the McGurk effect break down, if an extended quasi-stationary phase is created by prolonging the duration of the target consonant?
2. Is there any change in the effect, if the movie frames in the quasi-stationary phase are replaced with a single static frame?
3. Does the McGurk effect persist, if the visual information during that quasi-stationary phase is completely removed by replacing its video frames with black frames?

If static information is not sufficient to sustain the McGurk effect, it should cease when a certain threshold duration of the visual quasi-stationary phase is exceeded. If the McGurk effect persists, it can still be reasoned that remaining very small movements might be responsible and thus it should break down if the quasi-stationary component is made strictly stationary by technical means. If the McGurk effect still persists, it would seem that static visual information is indeed sufficient. However, to exclude any other explanation it must be shown that the McGurk effect does not occur if the visual information in this phase is removed completely.

2. Method

2.1. Stimuli

The bilabial nasal consonant /m/ and the velar stop consonant /k/ were chosen as target consonants and embedded in a two-syllable word with a /maCa/ structure where C donates the target consonant. Trivially, to reliably determine the duration of the stop consonant /k/ a preceding speech segment is needed. The reason to augment the preceding speech segment to the full CV syllable /ma/ was to provide the participants with a congruent /m/ as a reference in every item making it more likely that they would corresponded with “other” (see below) if there was any unspecific influence of the visual signal on the auditory perception.

The maximum duration to be tested for the target consonant was set to 4s. Ideally the consonant duration would be increased in small steps (e.g., 25ms) starting at the upper bound of the normal range (everyday speech) of the speaker and ending at the maximum duration to be examined. However, this would result in more experimental conditions than feasible within a standard perception experiment. Since it was assumed that any change in the strength of the McGurk effect would be more likely to be observed when just exceeding the normal range [12], an exponentially increasing step size was considered appropriate allowing inclusion of very long durations while keeping a special focus on durations just exceeding the normal range of the speaker. In particular,

$$\Delta t_n = \Delta t_0 \, e^n \quad \text{with } n = 1, \ldots, 5 \text{ and } \Delta t_0 = 25\,ms \quad (1)$$

was used to compute the step size. From the stimuli recording (see below) it was determined that the duration of the target /m/ varied between 90 and 150ms for the stimuli speaker when she made no attempt to slow down or extend the target duration. Thus a starting value 150ms was chosen resulting in target consonant durations of 150, 175, 218, 335, 652, 1515 and 3860ms.

A female Australian English speaker was recorded with a Pansonic NV-MX300 video camera at 25fps in sound proof room uttering /mama/ and /maka/ with a wide range of durations with regard to the target consonant. The face of the speaker was equally lit and filled approximately 70% of the video PAL 720 x 576 pixel frame in the vertical dimension. The acoustic signal was recorded with a Bruel & Kjaer 2671 microphone placed 20cm away from the speaker. The sound signal was digitized with an Edirol UA-25 sound card with a sample rate of 44.1kHz.

The items that contained the closest match to the intended target durations were selected and trimmed or expanded to fit exactly the target durations. In the video signal single frames were either removed from the temporal centre of the consonant or duplicated. All video manipulations were accomplished with Adobe Premiere Pro 2.0. In the acoustic signal the consonant was re-synthesised using *Praat's* [13] PSOLA-based duration change function. Note that because of the many versions recorded only minor changes were necessary and these affected primarily the very long durations. After this the preceding speech segments /m/ and /a/ were carefully adjusted in the same way to temporarily match each other in each /mana/-/maka/ pair. Again the duration changes were very small, in the majority of the cases no adjustment at all was needed. To create the congruent control condition auditory /mama/ was dubbed onto visual /mama/ and to create the McGurk stimuli auditory /mama/ was dubbed onto visual /maka/.

The stimuli for the three experiment conditions resulting from the three questions put forward above were then created as follows: For the first condition (NATURAL) the stimuli were left unmodified, for the second condition (STATIC) the frames of the quasi-stationary phase (i.e., from the point in the video where the speaker had reached the target configuration of either /k/ or /m/ until the point where the transition to the following vowel started) were replaced by a single frame taken from the temporal centre of this phase, and finally for condition three (BLACK) the frames of the quasi-stationary phase were substituted with black frames. In the two shortest durations (150ms and 175ms) the closing gesture of the consonant was immediately followed by the opening gesture of the next vowel. Due to the lack of a quasi-static phase in these durations, the STATIC and the BLACK conditions consisted only of the five larger durations.

A copy of each stimulus movie was made and a black cross was superimposed onto a single frame at the temporal midpoint of the final /a/ appearing at the bridge of the nose of the speaker for 40ms. Finally the horizontal width of all stimulus movies were cropped to 576 pixels.

2.2. Participants and procedure

Twenty-two (16 female) English speaking first year psychology students at the University of Western Sydney participated in exchange for course credit. All participants reported normal hearing and normal or corrected to normal vision and gave informed consent. The participants were tested individually in a testing booth at MARCS Auditory Laboratories. The stimuli were presented on an LG LS70 Laptop using the experiment control software *Alvin* [14] with the acoustic signal played through Edirol 10A powered speakers placed directly behind the laptop. Sound pressure level was set to a comfortable level of 73dBA. The participants were seated approximately 40cm way from the laptop. They responded by pressing one of 7 buttons labeled "mama", "maka", "mana", "mupka", "munma", "munga" and "other" with the computer mouse. The labels were chosen according to the participant's orthographic expectations of the intended response categories /mama, maka, mana, mapka, manma, maŋa/ and an unspecific "other" response collecting all deviating responses.

The participants were presented with 6 repetitions of the movie stimuli at each duration and in each condition in fully randomized order. They were told to respond according to 'what the speaker said'. They were not particularly instructed to base their decision on what they heard which would implicitly reveal the fact that there might be differences between the auditory and the visual signal. Their visual attention to the stimuli was guaranteed by the secondary task to detect the cross that appeared in half of the trials (see above).

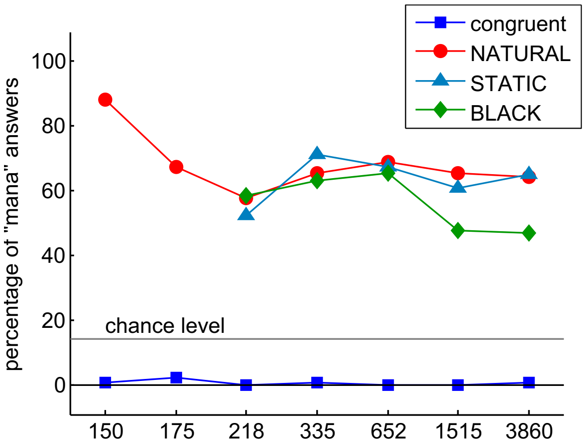

Figure 1: *Means of the percentage of "mana" responses across participants (n = 22) for the congruent /mama/ stimuli (squares) and the incongruent stimuli, i.e., auditory /mama/ dubbed onto visual /maka/ (NATURAL: circles, STATIC: triangles, BLACK: diamonds).*

3. Results and Discussion

The detection of the shortly appearing cross on the speaker's bridge of the nose was counted as correct if and only if either the cross was present and was acknowledged by the perceiver or it was not present and was not marked. The percentage of correct detection was 90.3% in the NATURAL condition, 89.5% in the STATIC condition, and 92.9% in the BLACK condition which confirmed that the participants paid attention to the visual aspects of the stimuli.

In all conditions the percentage of "mana" responses was averaged across repetitions for each participant and used to measure the experienced strength of the McGurk effect.

3.1. Condition I: Unmodified stimuli

A strong McGurk effect was found for all durations in the NATURAL condition (see Figure 1). The difference between the means of the percentages of "mana" answers for the auditory-visually congruent and incongruent stimuli was highly significant for all durations (paired-sample t-tests, df = 21) as was the differences from chance level (one-sample t-tests against chance level of 14.3%, df = 21). See Table 1, third to fifth column, for details.

To determine if there was a significant difference between the means for the seven durations, a one-way repeated measures

Table 1: *Means (and standard deviations) of the percentage of "mana" responses for the seven tested durations. The second column contains the means and standard deviations for the congruent stimuli. The following columns show the means (standard deviations) and the significance values (t- and p-value) for paired-sample t-tests (df* $= 21$*) comparing the congruent and incongruent stimuli in the NATURAL, STATIC and BLACK condition. Asterisks next to the means indicate significant differences (one-sample t-tests, df* $= 21$*) from chance level (14.3 %): **** $p = 0.000$*,* $\diamond\diamond$ *t-value cannot be computed.*

dur. in ms	congruent	incongruent								
		NATURAL			STATIC			BLACK		
	mean (std)	mean (std)	t	p	mean (std)	t	p	mean (std)	t	p
150	**0.8** (3.6) ***	**87.9** (16.4) ***	25.2	0.000	-	-	-	-	-	-
175	**2.3** (5.9) ***	**67.4** (31.5) ***	9.5	0.000	-	-	-	-	-	-
218	**0.0** (0.0) $\diamond\diamond$	**57.6** (37.0) ***	7.3	0.000	**52.3** (38.9) ***	6.3	0.000	**58.3** (34.4) ***	8.0	0.000
325	**0.8** (3.6) ***	**65.2** (34.9) ***	8.6	0.000	**71.2** (33.0) ***	10.1	0.000	**62.9** (35.2) ***	8.3	0.000
625	**0.0** (0.0) $\diamond\diamond$	**69.0** (36.1) ***	9.0	0.000	**67.4** (36.9) ***	8.6	0.000	**65.2** (36.3) ***	8.4	0.000
1515	**0.0** (0.0) $\diamond\diamond$	**65.2** (39.5) ***	7.7	0.000	**60.6** (40.7) ***	7.0	0.000	**47.7** (31.0) ***	7.2	0.000
3860	**0.8** (3.6) ***	**64.4** (38.6) ***	7.9	0.000	**65.1** (36.3) ***	8.4	0.000	**47.0** (35.5) ***	6.2	0.000

ANOVA was conducted with α set to .05. The analsis revealed a significant main effect for duration with a Greenhouse-Geisser adjustment to the degrees of freedom $F(3.14, 65.90) = 4.66$, $p = 0.005$, partial $\eta^2 = .18$ and a significant quadratic trend, $F(1, 21) = 7.31$, $p = 0.01$, partial $\eta^2 = .26$, indicating a decaying strength of the McGurk effect with longer durations. However, as can be seen in Figure 1, this appears to be caused only by the large value for the 150ms duration. Therefore post-hoc pairwise comparisons between consecutive durations were conducted using a Bonferoni adjusted α-value of 0.008 (original α = 0.5). As expected the only significant decrease in the percentage of "mana" responses was between 150ms and 175ms, $F(1, 21) = 15.17$, $p = 0.001$, partial $\eta^2 = .42$. Also, the quadratic trend disappeared when the 150ms duration was removed from the data set.

A pronounced McGurk effect was elicited as the target consonant duration was prolonged beyond durations that occur in everyday speech. The significantly stronger McGurk effect at 150ms may have occurred because the duration is still within the normal range of the speaker. Since this was the only item within the normal range, it is not possible to decide whether this fact was the cause of the higher value or whether the alternative explanation that other differences in the production of the individual stimulus (e.g., different facial motion) lead to the higher percentage of McGurk answers. However, this is of subordinate importance here, important is that the found trend is not a consistent trend across the longer durations, i.e., that it can be asserted that the McGurk effect is not incrementally diminished as duration increases. From the above it seems to follow that contrary to the findings of [10] static visual information (the long quasi-static phase in the extended stimuli) is indeed influencing auditory-visual speech information. However, minimal facial movements occurring during the quasi-static phase might still contain sufficient dynamic information to sustain the McGurk effect. If so, then the McGurk effect should break down if the remaining dynamic properties of the visual signal in the quasi-static phase are removed.

3.2. Condition II: Substitution with static frames

The McGurk effect did not vanish or weaken in the STATIC condition (see Figure 1). The difference between congruent and incongruent stimuli was highly significant for all durations (paired-sample t-tests, *df* $= 21$) and so was the differences from chance level (one-sample t-tests against chance level of 14.3%, *df* $= 21$). See Table 1, sixth to eighth column, for details.

As can be seen in Figure 1 there is little difference between the means of the "mana" response percentages coming from the NATURAL and the STATIC condition. This was confirmed by the lack of a significant main effect for condition (NATURAL/STATIC) in a repeated measure ANOVA, $F(1, 21) = .10$, $p = .75$.

There was no significant difference in the strength of the McGurk effect compared to the non-altered video. A strong McGurk effect persisted across all durations even though no dynamic visual information remained within in the non-transitional phase of the target consonant. This seems to futher indicate that static visual information is indeed sufficient to maintain the McGurk effect confirming the findings from the NATURAL condition. Clearly then, if the static visual information is removed, the McGurk effect should finally cease at least for durations where the quasi-static phase is substantially longer than the two transitional phases, i.e., starting in this study with a duration of 652ms.

3.3. Condition III: Substitution with black frames

Surprisingly, a strong McGurk effect was also found in the BLACK condition (see Figure 1). Again the difference between the congruent and incongruent stimuli was highly significant for all durations (paired-sample t-tests, *df* $= 21$) and the same was true for the differences from chance level (one-sample t-tests against chance level of 14.3%, *df* $= 21$). See Table 1, ninth to last column, for details.

A repeated measure ANOVA with factors *duration* and *condition* compared the percentages of "mana" responses in the BLACK to the to the ones in the NATURAL condition across the five durations. A significant main effect for *condition*, $F(1, 21) = 4.16$, $p = 0.001$, partial $\eta^2 = .44$ and a significant interaction between *duration* and *condition*, $F(4, 84) = 13.43$, $p = 0.001$, partial $\eta^2 = .39$, were found. To determine which durations were responsible for the significant effects, the means in the NATURAL and BLACK condition for each duration were compared employing five post-hoc paired samples t-test with a Bonferroni adjusted α-value of 0.1 (original $\alpha = 0.5$). Only for the two longest durations a significant difference was detected, i.e., at 1515, $t(21) = -3.09$, $p = 0.006$ and 3860ms, $t(21) = -4.81, p = 0.000$.

Contrary to expectations the stimuli that contained no visual information during their entire elongated quasi-static phase *did* elicit McGurk responses suggesting that the static visual information conveyed in the stationary phase of the target consonant is *not* responsible for the observed sustained McGurk effect. As the consonant duration was prolonged beyond one second the McGurk effect began to weaken slightly, though it appears that the reduction in strength leveled out immediately (see Figure 1 and Table 1)

3.4. General discussion

Taken together, the results obtained in the three conditions suggest that it is in fact the dynamic visual information in the transition phase of the target consonant that is responsible for the McGurk effect. Though when prolonging the target consonant this is accomplished by prolonging the quasi-static phase and yet the McGurk effect was found not to be diminished, it should have disappeared when the visual signal was completely removed, if in fact static visual information was the driving force for its persistence across increasingly larger durations. The findings suggest a comprehensive perceptual 'deafness' to stationary acoustic speech signals at least for nasals. This notion is partially supported by the results from [12] regarding vowels (Note that though nasals behave with regard to phonotactics in most languages like consonants they share acoustically and auditorily more properties with vowels, e.g., the importance of resonance frequencies for their identification and the sustained voicing throughout the whole phoneme). [12] found that the information contained in the transitions between consonants and vowels in CVC syllables allowed unambiguous identification of the vowel, even when the entire vowel nuclei including the 'target' information was missing (silent-centre condition). However, the alternative side of this, that is, whether the quasi-static acoustic phase on its own would preclude identification, was not tested.

To confirm the *stationary deafness hypothesis* additional auditory-only experiments are necessary in which the transitional phases are removed from acoustic /m/ and /n/ stimuli and identification is tested. To also further investigate the role of static visual information future experiments will be conducted in which the removal of visual information with regard to the stationary and transitional phases of the target consonants will be inverted: instead of replacing the video frames of the stationary phase the frames of the two transitional phases will be 'blacked out'. Based on the findings from the current study it can be hypothesised that the static visual information should not elicit the McGurk effect in this case. Note that these experiments are only possible with extended consonant durations since only they allow separating transitional and stationary phases and with this dynamic and static information.

4. Conclusions

Combining the bilabial nasal consonant /m/ as auditory stimulus with the velar stop consonant /k/ as the visual stimulus offers the possibility to investigate the impact of prolonged consonant durations on the McGurk effect and, by extension, the role of static visual information in this cross-modal illusion. It was found that the McGurk effect did not break down nor diminish in strength when the consonant duration was extended up to 3.8s. This remained unchanged when the video frames in the quasi-static phase of the prolonged target consonant were made strictly static by substituting the original frames with copies of a single frame. Surprisingly, the McGurk effect even persisted when the visual signal in the quasi-static phase was completely removed. The results indicate that static visual information is not responsible for maintaining the McGurk effect across increasingly larger durations and may not be sufficient to elicit the McGurk effect. They also point towards a comprehensive deafness to stationary acoustic speech signals at least as far as nasal consonants are concerned.

5. Acknowledgments

The authors would like to thank Chris Davis and Jeesun Kim for many helpful comments. The first author wish to acknowledge the support of NHMRC/ARC grant TS0669874.

6. References

[1] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” *Nature*, vol. 264, pp. 746–748, 1976.

[2] J. MacDonald and McGurk, “Visual influences on speech perception processes,” *Perception and Psychophysics*, vol. 24, no. 3, pp. 253–257, 1978.

[3] M. Roberts and Q. Summerfield, “Audiovisual presentation demonstrates that selective adaptation in speech is purely auditory,” *Perception and Psychophysics*, vol. 30, no. 4, pp. 309–314, 1981.

[4] D. Massaro, *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1987.

[5] K. P. Green and P. K. Kuhl, “The role of visual information in the processing of place and manner features in speech perception,” *Perception and Psychophysics*, vol. 45, no. 1, pp. 34–42, 1989.

[6] H. M. Saldaña and L. D. Rosenblum, “Visual influences on auditory pluck and bow judgments,” *Perception and Psychophysics*, vol. 54, no. 3, pp. 406–416, 1993.

[7] T. R. Jordan and K. Bevan, “Seeing and hearing rotated faces: Influences of facial orientation on visual and audiovisual speech recognition,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 23, no. 2, pp. 388–403, 1997.

[8] J. MacDonald, S. Andersen, and T. Bachmann, “Hearing by eye: How much spatial degradation can be tolerated?” *Perception*, vol. 29, no. 10, pp. 1155–1168, 2000.

[9] K. G. Munhall, P. Gribble, L. Sacco, and M. Ward, “Temporal constraints on the McGurk effect,” *Perception & Psychophysics*, vol. 58, no. 3, pp. 351–362, 1996.

[10] L. D. Rosenblum and H. M. Saldaña, “An audiovisual test of kinematic primitives for visual speech,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 22, no. 2, pp. 318–331, 1996.

[11] K. G. Munhall and Y. Tohkura, “Audiovisual gating and the time course of speech perception,” *Journal of the Acoustical Society of America*, vol. 104, no. 1, pp. 530–539, 1998.

[12] W. Strange, “Information for vowels in formant transitions,” *Journal of Memory and Language*, vol. 26, no. 5, pp. 550–557, 1987.

[13] P. Boersma, *Praat: doing phonetics by computer*, 1999. [Online]. Available: http://www.praat.org/

[14] J. M. Hillenbrand and R. T. Gayvert, “Open source software for experiment design and control,” *Journal of Speech, Language, and Hearing Research*, vol. 48, no. 1, pp. 45–60, 2005.