OXFORD
BROOKES
UNIVERSITY

Hearing is Seeing: The Implicit McGurk Illusion - a Perceptual or Cognitive Phenomenon?

Konstantin Volkmann

Supervised by: Dr. Michael Pilling

March, 2014

The Implicit McGurk Illusion - a Perceptual or Cognitive Phenomenon?

ABSTRACT

This study addresses the question of whether audiovisual speech integration is an automatic and unconscious process or subject to attentional demands. The experimental approach utilised a variant of Garner's (1974) speeded classification task, with audiovisual stimuli comprising of disyllabic non-words. Observers had to classify the first syllable while the second syllable was experimentally manipulated. The rationale under consideration was that in a series of trials, task-irrelevant variations of the second syllable will slow down response latencies, henceforth called syllabic interference effect. This effect was produced in Experiment 1. Experiment 2 demonstrated that this effect can also be induced by virtue of a McGurk percept while Experiment 3 illustrated the elimination of such effects on the basis of a McGurk percept. Further, participants repeatedly attended the experiments over five consecutive days. Findings reinforce the claim that audiovisual integration occurs before selective attention can be allocated and support the assumption of it being an automatic process. The observed stability of the effects over time suggests that audiovisual integration is immune to top-down control often achieved through practice and thus represents a purely perceptual phenomenon. Results are discussed with regards to the question of cognitive impenetrability of audiovisual speech integration.

KEYWORDS	MCGURK ILLUSION	SELECTIVE ATTENTION	AUDIOVISUAL INTEGRATION	SPEECH PERCEPTION	COGNITIVE IMPENETRABILITY

INTRODUCTION

Speech perception involves the simultaneous processing of acoustic information presented to the ears as well as visual information visible to the eye, provided by a speaker's movements of lips, tongue and jaw. When speech can be *seen* as well as *heard*, an effective interaction between auditory and visual speech streams is found to occur in the brain, producing enhanced speech intelligibility (Sumby & Pollack, 1954; Callan, Jones, Munhall, Callan, Kroos & Vatikiotis-Bateson, 2003). This perceptual advantage of crossmodal integration of speech is reflected in objective-behavioural responses. Compared to unimodal auditory stimulation, faster reaction times to spatially concurrent and semantically congruent audiovisual stimuli are observed (Morrell, 1968; Frens & Van Opstal, 1995). If such multisensory input is made to be incongruent - e.g. when dubbing the sound of one token onto the video of a speaker articulating a different token - response times to identify the percept are slower and perceptual illusions occur (Stein, Meredith, Huneycutt & McDade, 1989; Sekuler, Sekuler & Lau, 1997).

For audiovisual speech, the "McGurk illusion" (McGurk & MacDonald, 1976) represents an intriguing example. When dubbing an articulatory movement (e.g. "ga") on an incongruent acoustic phoneme (e.g. "ba") an illusory percept (e.g. "da") is produced. In other words, the synchronised presentation of two different consonant-vowel syllables frequently means observers tend to fail to detect the conflicting modalities and instead perceive a consonant sound that was not present in either modality. This phenomenon has been applied as an 'experimental tool' in research aiming to explain the processing of speech (Green, 1998). One question, however, is still provoking an active debate: whether this integration of audiovisual speech cues occurs at a pre-attentive stage (i.e. in an automatic fashion before attention can be applied) or whether it requires attention to be specifically directed to visual and auditory stimuli.

The 'supremacy' of the McGurk illusion suggests an attention-free and automatic process. Conscious awareness of the conflicting audiovisual input does not seem to reduce the effect (Manuel, Repp, Studdert-Kennedy & Liberman, 1983). Obvious sensory discordances, such as the combination of a male face and a female voice or a high degree of spatial separation between visual and acoustic cues, also do not diminish the illusion (Green, Kuhl, Meltzoff & Stevens, 1991; Soto-Faraco & Alsius, 2009). Interestingly, observers explicitly instructed to focus their attention onto either one or both sensory modalities experience the illusion to a similar degree, regardless of their respective focus of attention (Massaro, 1987). Nevertheless, the majority of studies contemplating the robust character of the McGurk effect - even under most adverse conditions - rely on reports of phenomenological experiences (Navarra, Alsius, Soto-Faraco & Spence, 2010). Thus it has to be noted that although observers may have completed the task, e.g. by attempting to focus their attention to a single modality (as in Massaro, 1987), integration may still occur *implicitly;* i.e. without the observer's awareness (Wojciulik & Kanwisher, 1998).

A study by Gentilucci and Cattaneo (2005) addresses this issue of implicit binding, i.e. the extent to which observers are unaware of the processing of conflicting audiovisual information. They found that participants' verbal repetitions of McGurk-like syllables (presented randomly within a stream of normal syllables) were always influenced by the speaker's lip articulations even though only less than 25% reported having experienced the audiovisual illusion. Such results emphasise the challenging task of defining an appropriate method for testing the role of automaticity in audiovisual integration.

Research using electroencephalography (EEG) has shown that integration of speech occurs before the participation of attentional processes (Colin, Radeau, Soquet, Demolin, Colin & Deltenre, 2002; Colin, Radeau, Soquet & Deltenre, 2004). In these studies McGurk-like syllables elicited a specific auditory event-related potential (ERP) that is indicative of early processing of acoustic stimuli i.e. without attentional modulation. This ERP, called mismatch negativity ERP (MMN), is associated with the occurrences of an oddball (infrequent) stimulus in a continuous sequence of familiar sounds (Näätänen & Alho, 1995). However, the interpretation of these results remains matter of some debate. The authors argued that the increased magnitude of the MMN occurring as a response to deviant McGurk-like stimuli indicates early integration of audiovisual speech cues (Colin et al., 2002; Colin et al., 2004). Others see an early detection of conflicting audiovisual input (e.g. temporal disparity or misalignment) reflected in the *enlarged* MMN, thus arguing that a *reduced* magnitude is evidence for the assumption that the MMN signal reflects early audiovisual integration (Navarra et al., 2010). Indeed, findings from another study demonstrate that a *reduced* magnitude in an MMN signal is indicative of an early integration of audiovisual cues, occurring at a pre-attentive stage during information processing, i.e. on a phonetic level (Kislyuk, 2006; Kislyuk, Möttönen & Sams, 2008).

Other research indicates that attention is important in processing the McGurk effect. This work has used a visual distractibility paradigm in which the illusion is greatly reduced when a leaf moving across the screen is presented simultaneously with the McGurk illusion (Tiippana, Andersen & Sams, 2004). The task-irrelevant stimulus (the leaf) was always transparent and thus did not cover facial areas or cause other forms of masking (Öğmen, Breitmeyer, Todd & Mardon, 2006). Such results imply that visual attention can modulate the audiovisual integration process.

Research looking at brain processes has also suggested a role for attention. A study by Kaiser, Hertrich, Ackermann & Lutzenberger (2006) using magnetoencephalography (MEG) found an association between Gamma-band activity (GBA) and the detection of a deviant stimuli in incongruent audiovisual information. Significant GBA activity was identified over brain areas that are associated with the modality in which the respective deviant stimulus was presented. Since the GBA signals occurred at quite long latencies after the onset of the stimulus, the researchers argue that cognitive effort in form of a “top-down attentional-guided process” is required (Kaiser et al., 2006). In other words, the detection of conflicting audiovisual information must be followed by a “re-analysis” in early sensory areas within visual or auditory modalities, respectively. This process is dependent on attentional resources but it remains unclear whether such a re-analysis is actively applied in everyday speech perception where observers typically do not expect an audiovisual conflict.

The empirical studies discussed above manipulated attention, however none ensured that the observer’s attentional resources had been completely ‘exploited’. This issue is critical in studies based on the assumption that observers can be explicitly asked to focus their attention to one sensory modality while ignoring information from another (De Gelder & Bertelson, 2003). Lavie (2005) has argued that instructions to ignore goal-irrelevant stimuli do not prevent participant’s from unconsciously processing those. Concepts of attention, like Lavie’s model of Perceptual Load

(Lavie, 1995), claim that humans possess a fixed pool of attentional resources that are supplying processes until exhausted. Consequentially, if an observer is asked to identify relevant but ignore irrelevant stimuli he will continue processing irrelevant information as long as the task does not consume all of the available attentional resources. This ‘surplus’ may be responsible for some of the results described above in which residual attentional resources “spill over”, facilitating the processing of irrelevant stimuli (Lavie, 2005; Santangelo & Spence, 2007; Santangelo & Spence, 2008).

Interestingly, the susceptibility to McGurk stimuli is reduced when observers are concurrently performing an unrelated auditory or visual task, suggesting that audiovisual integration does falter when attentional resources are being extensively exploited (Alsius, Navarra, Campbell and Soto-Faraco (2005). In their study, participants had to verbally repeat what a speaker said under three conditions: auditory or visual speech alone vs. audiovisual speech (including the McGurk effect). Answers indicated whether participants identified the speech stimuli correctly or not. At the same time, participants were asked to divide their attention in order to detect either an auditory or visual target stimulus (co-occurring transparent objects or sounds) within the respective conditions. While the perception of speech presented in only one modality (visual or auditory) was not affected by divided attention, the perception of McGurk illusions was significantly reduced. Despite those clear results the assumption of attention playing a vital role in audiovisual integration did not become answered. Arguably, the demands of the visual distractor task may have inhibited further processing of visual speech information at an early processing stage *before* audiovisual integration could take place to evoke a McGurk percept (Alsius et al., 2005).

Therefore, testing an observer’s limit of attentional capacity does not necessarily provide direct evidence for whether integration of incongruent audiovisual information occurs automatically or not (Navarra et al., 2010). Almost any processing mechanism requires some level of attention (indeed no study to date has shown that any form of multisensory integration occurs during sleep). Thus, the question is how much *implicit* attentional resources are consumed during the process of binding crossmodal information.

This has been addressed in a study by Soto-Faraco, Navarra and Alsius (2004). They introduced an alternative method that allowed for an indirect measurement of audiovisual integration. The McGurk effect was again a ‘research tool’ but this time, the integrated percept was presented as a task-irrelevant stimulus. This contrasts with the measures used in most behavioural studies testing the observer’s direct classification of the McGurk speech sound. For this, Soto-Faraco and colleagues developed a variant of the *speeded-classification paradigm* (Garner, 1974) using disyllabic words (e.g. “tadi”). Participants were asked to identify the respective first syllable of words presented audiovisually in a series of blocks while trying to ignore the second syllable. In so called “homogenous blocks” the irrelevant second syllable remained constant (e.g. “tadi” and “todi”) while in “orthogonal blocks” the second syllable varied randomly in a trial-by-trial system (e.g. “tadi” and “tobi”). As known from an earlier study by Pallier (1994), irrelevant variations in the second syllable result in a slower identification of the first syllable; a phenomenon called “syllabic interference effect” (derived from the original term “Garner Interference”). This suggests that if participants explicitly focus their attention onto the *first* syllable alone the implicit processing of the *second* syllable remains mandatory (Pallier, 1994).

Soto-Faraco et al. (2004) did replicate this study using audiovisual material and further observed a syllabic interference effect when the irrelevant variation in the second syllable was induced by McGurk percept (even though the auditory component always remained constant). Additionally, in another experiment they were able to eliminate the syllabic interference effect by virtue of the McGurk illusion. Although the audiovisual information presented in the second syllables was alternated participants did not experience variations (and thus a syllabic interference effect) as long as the McGurk induced percept was homogenous to the second syllables from the remaining non-McGurk words. Together these results suggest that audiovisual integration cannot be overridden even when the consequences of this process negatively affect the participant's performance in a concurrent identification task. Regarding the issue of attention, such findings indicate that the integration of audiovisual speech occurs before selective attention can be allocated.

The results found by Soto-Faraco and colleagues (2004) motivated several studies (e.g. Alsius et al., 2005; Vatakis & Spence, 2006; Smith & Bennetto, 2007). However, it appears that in psychological research novelty succeeds verification and thus published empirical data may be regarded as immediate evidence. The signature strength of science is that evidence is reproducible in order for it to become a valid observation. Consequentially, there is increasing demand for direct replications of empirical studies to verify results (Simons, 2014).

Therefore, the present study had two main goals: First attempting a replication of the study by Soto-Faraco et al. (2004) by applying more precise testing conditions. Second testing the stability of those results after a period of practice. These goals were approached by creating three different experiments, repeatedly presented to participants on five consecutive days.

The first experiment introduced a syllabic interference task using disyllabic non-words where the auditory and visual information were matching. It was predicted that (task) irrelevant variations in the second syllable would elicit a syllabic interference effect. In other words, participants would be slower at identifying the first syllable in orthogonal lists where the second syllable is altered frequently compared to homogenous lists where the second syllable is constant. It was expected that this effect would be stable over time. In other words, participants would remain unable to identify the first syllable equally fast in all lists (i.e. learn to avoid the syllabic interference effect).

The goal of the second experiment was to see whether the syllabic interference effect can be obtained when irrelevant variations in the percept are produced only by the McGurk illusion. Even though the auditory speech information of the second syllable always remained constant, it was predicted that participants would be slower at identifying the first syllable as soon as the visual component is altered by virtue of a McGurk effect. In other words, the McGurk percept should evoke a syllabic interference effect. It was expected that participants would be unable to focus their attention only onto the auditory modality and therefore experience this interference effect causing reaction times similar to those observed in experiment one. If audiovisual integration occurs in a purely automatic manner repeated exposure to this experiment should not diminish the syllabic interference effect elicited by variations in the percept. Thus, it was predicted that participants would continue being unable to control their focus of attention within the time frame of five test days and subsequently fail to ignore the visual influence in oder to filter out the McGurk induced syllabic interference.

Finally, the aim of the third experiment was to test whether the syllabic interference effect can be eliminated on the basis of the McGurk illusion. No syllabic interference effect was expected to occur, since variations in the second syllable were prevented by the emergence of the McGurk percept. It was predicted that if the changes in perception evoked by the McGurk effect are of true nature, participants would not experience interference as long as the percept induced by this illusion is homogenous to the other non-McGurk syllables within that list. In other words, participants would behave as if they had only heard words with a constant second syllable. Therefore it was expected that response latencies should not differ between a list that *becomes* homogenous because of the McGurk percept and a list that *is* homogenous because of non-words that naturally possess the same second syllable. Since this perceptual process is assumed to be an automatic mechanism repeated exposure to the stimuli should not affect reaction times. On the contrary, if participants over time can learn to avoid automatic integration of audiovisual cues by means of attentional control then reaction times would indicate the reduction of a syllabic interference effect.

METHOD

Participants

Five male and five female participants with an age range from 19 to 29 (Mean=23) were recruited via opportunistic sampling (see Appendix A1). All reported normal or corrected to normal vision. None reported having any hearing impairment.

Design

Participants attended experiments on five consecutive days (repeated measures design). For each day, response times were measured in three different experiments: I. the NonMcGurk Syllabic Interference experiment (NMcSI), II. the McGurk Induced Syllabic Interference experiment (McISI) and III. the McGurk Prevented Syllabic Interference experiment (McPSI). Time required by participants to identify the first syllable of the disyllabic stimulus was measured as a dependent variable. The second syllable was manipulated differently in each experiment. In the NMcSI experiment (I) it was predicted that participants would identify the first syllable slower when the second syllable varied frequently. The McISI experiment (II) was testing the same assumption but whether this effect occurs also by virtue of the McGurk illusion. The McPSI experiment (III) did create a scenario using the McGurk illusion in which participants perceived the second syllable as being constant although its sensory input actually varied. Thus an absence of syllabic interference was expected.

All experiments were presented successively with breaks in between. The order of the experiments (I, II and III) was alternated between days. Participants were presented with the same stimuli everyday but the order of stimuli presented in different stimulus lists varied randomly between participants and test days.

Stimuli

Experiments were carried out in a noise-attenuated room. Participants were positioned one meter away from a standard pc monitor with high quality loudspeakers located next to the screen. Trials were organised using SuperLab. The loudness level for all stimuli was adjusted around 65 db using Audacity. Recorded with a high definition camcorder (Canon Vixia HF G20), a series of 1000ms long digital videos was produced.

Every video (in colour) showed the mouth of a male actor’s face in the centre (including 4cm of nose and 5cm of the chin and neck), articulating and pronouncing one exemplar of the disyllabic words “tadi”, “todi”, “tabi” and “tobi” at a time. Those four stimuli were congruent in terms of their auditory and visual information. Another two stimuli were created that carried incongruent audiovisual information, hereafter called the ‘McGurk stimuli’. Here, the lips visually formed the words “tagi” and “togi”, respectively, while a dubbed voice (from the same actor) pronounced “tabi” and “tobi”. Based on previous findings and tested during pilot sessions, participants perceived “tadi” for the combination “tagi” plus “tabi”; and “todi” for the combination “togi” plus “tobi”. For purposes of continuity, all stimuli were cross-dubbed, i.e. the auditory information was detached from one video (an exemplar) and synchronised with another video, showing the same stimulus but a different recording (another exemplar).

25 exemplars of each six stimuli (N=150) were recorded, cross-dubbed and standardised into video clips using Apple Final Cut Pro and arranged as follows: a fixation cross appeared at 900ms, accompanied by a beep sound, followed by a 200ms blank screen (white) before the audiovisual stimulus started (see Figure 1). On average, the onset of the first syllable was at 141.81ms (SD = 1.28), the onset of the second syllable was at 148.54ms (SD=4.54). Response time was measured from the moment the stimulus started to the pressing of the button on a Cedrus RB-830 response pad. Thus, participants were able to indicate responses as soon as they had identified the first syllable of the respective target stimulus. The next video started immediately after a participant’s input or a 2500ms deadline. An error-beep sounded when participants identified the wrong target syllable (or accidentally pressed the wrong button).

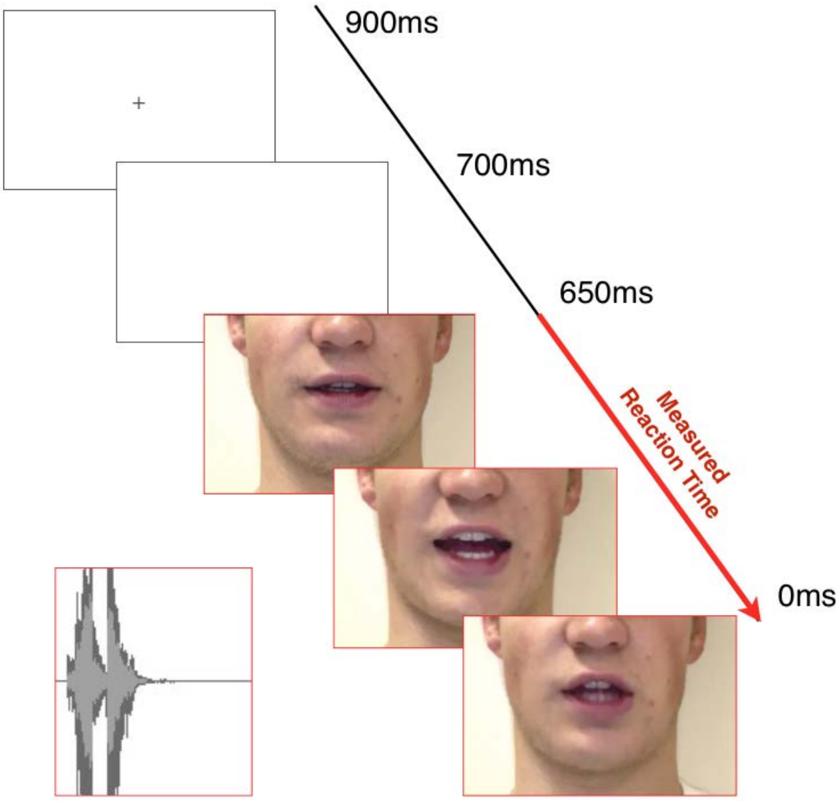

Figure 1: Schematic composition of a single trial comprising of a fixation cross (accompanied by a beep sound), a blank screen and the audiovisual stimulus.

All video clips were arranged in nine *stimulus lists* with each single list containing 72 stimuli drawn randomly from the pool of 150 exemplars. The amount of different stimuli used within each list were equal in number. Two different types of stimulus lists existed. A *homogenous* list only contained non-words with the same second syllable (matching). In other words, the second syllable remained constant while the first syllable was alternated, e.g. “tabi” plus “tobi” *or* “tadi” plus “todi”. On the contrary, an *orthogonal* list contained non-words with both syllables being different, e.g. “tabi” plus “tobi” plus “tadi” plus “todi” (see Table 1).

Table 1: Overview of stimuli and their respective audiovisual components, organised in different types of stimulus lists, presented in different conditions. Words within brackets indicate the visual information. Words within the forward dashes indicate the auditory information.

Effect	Experiment	Type of Stimulus List	Audiovisual Combinations
Syllabic Interference	I. NMcSI	Homogenous “Bi”	/tabi/ + (tabi) and /tobi/ + (tobi)
		Homogenous “Di”	/tadi/ + (tadi) and /todi/ + (todi)
		Orthogonal	/tadi/ + (tadi) and /todi/ + (todi) and /tabi/ + (tabi) and /tobi/ + (tobi)
	II. McISI	Homogenous “Bi”	/tabi/ + (tabi) and /tobi/ + (tobi)
		Homogenous “Di”	/tadi/ + (tadi) and /todi/ + (todi)
		Orthogonal McGurk	/tabi/ + (tabi) and /tobi/ + (tobi) and /tabi/ + (tagi) and /tobi/ + (togi)
Cancelling Syllabic Interference	III. McPSI	Homogenous “Di”	/tadi/ + (tadi) and /todi/ + (todi)
		Homogenous McGurk “Di”	/tabi/ + (tagi) and /tobi/ + (togi)
		Illusionary Homogenous McGurk “Di”	/tadi/ + (tadi) and /todi/ + (todi) and /tabi/ + (tagi) and /tobi/ + (togi)

Procedure

Participants were instructed to focus their attention on the first syllable of a stimulus (target), hereafter called the relevant syllable. The task was to click the left button on the response pad for “ta” and the right button for “to” immediately after having identified either one as the first syllable. After a practice trial (not included in the analyses), the experiment started. Every participant was allowed breaks as long as needed in between the three experiments. In a single experiment, participants were always presented with three types of stimulus lists, each containing 72 trials. One experiment lasted on average 9 minutes (SD= 1.64). One whole testing session containing all three experiments lasted up to 40 minutes, depending on the length of the breaks. All participants were tested on five consecutive days with no more than 29 hours in between each testing session.

RESULTS

The mean for median reaction times (RTs) of *correct responses* was calculated for *type of stimulus list* within each experiment. This was done for all five *testing days*, respectively. Most of the data were not normally distributed since reaction time distributions are more similar to *ex-Gaussian* distributions (Whelan, 2008). Further, in the majority of analyses sphericity could not be assumed. For these, statistical principles deriving from “Pillai’s Trace” (Pillai’s *F*) apply. The remaining results (sphericity assumed) report “Greenhouse Geisser” (*F*).

On average, participants responded correctly on 95% (SD = 1.7) of trials. This indicates high accuracy on the speeded classification task. Further, no evidence for a speed-accuracy trade-off was found (see Appendix A2). When comparing overall RTs there was a significant interaction between *experiment* and *type of stimulus list* used, Pillai`s *F*(2,8)=13.86, *p*<.05 η^{2p}= .776. This suggests that variations in the irrelevant syllable affected participants' performance on the speeded-classification task. In the following analyses those effects are examined in detail.

The goal of the 'non-McGurk syllabic interference experiment' (NMcSI) was to test whether a general syllabic interference effect can be elicited with this particular audiovisual material.

The goal of the 'McGurk induced syllabic interference experiment' (McISI) was to test if the same pattern of interference effects can be induced when introducing a McGurk combination in the second syllable. This manipulation was testing to what extent participants' selective attention can be mediated by implicit audiovisual integration. Here, an interference effect would result from the failure to ignore irrelevant visual variations occurring in the second syllable, transforming an *auditorily* homogenous list into a *perceptually* orthogonal list.

Finally, the goal of the 'McGurk prevented syllabic interference experiment' (McPSI) was to test if the syllabic interference effect can be eliminated by virtue of the McGurk illusion. Here, the absence of an interference effect would result from the failure to recognise a conflicting input, transforming an *audiovisually* orthogonal list into a *perceptually* homogenous list.

NonMcGurk Syllabic Interference Experiment (NMcSI)

Using a repeated-measures design, a three-way analysis of variance (ANOVA) was performed. This included the within-subjects factors *response set* ("Ta" vs. "To"), *type of stimulus list* (orthogonal "Bi" + "Di" vs. homogenous "Bi" vs. "homogenous "Di") and *test day* (1,2,3,4,5). There was a significant main effect of *type of stimulus list* on RTs, Pillai's *F*(2,8)=18.15, *p*<.05, η^{2p}=.819. As shown in Table 2, contrasts revealed that participants responded slowest to the orthogonal list and fastest to the homogenous lists. Pairwise-comparisons using Bonferroni correction demonstrated a significant mean difference (*p*< .05) of 17.4 ms (SD= 3.82) between both homogenous lists, a larger significant mean difference (*p*<.05) of 205.5 ms (SD=52.70) between the orthogonal and homogenous "Bi" list and the largest significant mean difference (*p*<.001) of 222.9 ms (SD=52.38) between the orthogonal and homogenous "Di" list.

Table 2: Mean across participant median correct response times (milliseconds). Calculated for each stimuli presented in different stimulus lists collapsed across test days. Note the larger RTs between the orthogonal and the two homogenous lists.

Experiment	**Stimulus List**	**Mean RTs (SD)**
I. Non-McGurk Syllabic Interference	Homogenous “Bi”	1283.22 ms (11.54)
	Homogenous “Di”	1265.83 ms (9.30)
	Orthogonal “Bi” + “Di”	1488.71 ms (46.90)

There was a marginally significant main effect of *test day*, Pillai’s *F*(4,6)=3.75, *p*=.07 η^{2p}=.278, suggesting that participant’s responded significantly faster during the days two and three (Figure 2) when compared to the first day (Mean difference=63.63; SD=18.79). However, pairwise-comparisons using Bonferroni correction did not reveal any further significant results.

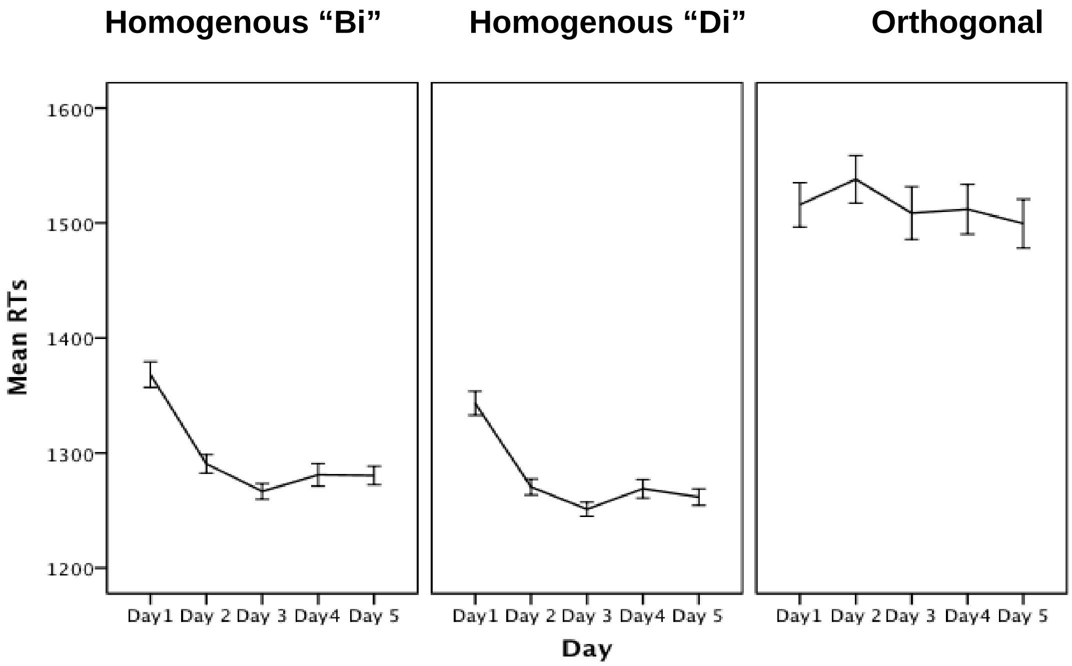

Figure 2: Mean RTs in milliseconds separated by *test day* for each *type of stimulus list* in the NMcSI experiment. From the left, the first box shows the temporal pattern for the homogenous “Bi” list, the second box for the homogenous “Di” list and the third box displays the orthogonal list. Error bars indicate 95% confidence interval around the mean.

Secondary to the goal of this experiment, there was a significant interaction between *type of stimulus list* and *response set*, Pillai’s *F*(2,8)=11.99, *p*<.05 η^{2p}= .750. Contrasts revealed that participant’s were faster at identifying the first syllable “To” as opposed to “Ta” in the speeded classification task, relative to the RTs within every *type of stimulus list*. This result may be explained by a cerebral dominance effect.

There was no significant interaction between *test day* and *response set*, Pillai’s *F*(4,6)=.36, *p*=.83. There was also no significant interaction between *type of stimulus list*, *test day* and *response set,* Pillai’s *F*(8,2)=.87, *p*=.64.

In conclusion, results support the effect of syllabic interference due to irrelevant variations in the second syllable. The impact of practice neither facilitated nor inhibited this effect. Further, this essential test can be understood as a control condition for the subsequent manipulation (McISI). Results provide a baseline curve for the strength of syllabic interference. The following results can be aligned to this curve to highlight potential similarities.

McGurk Induced Syllabic Interference Experiment (McISI)

A second three-way ANOVA was conducted. This included the within-subjects factors *response set* (Ta vs “To”), *type of stimulus list* (homogenous “Bi” vs. homogenous “Di” vs. orthogonal McGurk) and *test day* (five). Again, there was a significant main effect of *type of stimulus list* on RTs, Pillai’s $F(2,18)=14.98$, $p<.001$ $\eta^{2p}= .625$. As shown in Table 3, contrasts revealed that participants responded slowest to the orthogonal list and fastest to the homogenous lists. Pairwise-comparisons using Bonferroni correction demonstrated a nonsignificant mean difference ($p=.32$) of 6.84 ms (SD=7.20) between both homogenous lists, a large significant mean difference ($p<.05$) of 202.74 ms (SD=55.05) between the orthogonal list and homogenous “Bi” and the largest significant mean difference ($p<.001$) of 209.58 ms (SD=50.58) between the orthogonal list and homogenous “Di”. There was a significant main effect of *test day*, $F(4,36)=3.08$, $p<.05$ $\eta^{2p}= .255$, suggesting that participants became faster at the speeded classification task over the course of five days (Figure 3). However, pairwise-comparisons using Bonferroni correction revealed no further significant effect for *test day* on RTs.

Table 3: Mean across participant median correct response times (milliseconds). Calculated for each stimulus presented in different stimulus lists collapsed across test days. Note the larger RTs between the orthogonal and the two homogenous lists.

Experiment	**Stimulus List**	**Mean RTs (SD)**
II. McGurk Induced Syllabic Interference	Homogenous “Bi”	1282.04 ms (10.99)
	Homogenous “Di”	1275.20 ms (11.75)
	Orthogonal McGurk	1484.78 ms (48.31)

Secondary to the goal of this experiment, there was a significant interaction between *type of stimulus list* and *response set* Pillai’s $F(2,8)=15.63$, $p<.05$ $\eta^{2p}= .796$. Contrasts revealed that participant’s were generally faster at identifying the first syllable “To” as opposed to “Ta”.

There was a significant interaction between *test day*, *response set* and *type of stimulus list*, Pillai’s $F(8,2)=28.71$, $p<.05$ $\eta^{2p}=.991$. During day 1 participants responded slower to the orthogonal list (including the McGurk percept) and were generally faster at classifying “To” as opposed to “Ta”. During day 2 participants who responded faster to both homogenous lists responded slower to the orthogonal list when compared to day 1 but overall were again faster at classifying “To”. During day 3 participants responded even faster to both homogenous list but still significantly slower to the orthogonal list even though the response time to orthogonal list had decreased compared to day 2. Overall, participants classified “To” faster than “Ta”.

During Day 4 participants showed the same RTs as observed in day 3 except for a slightly slower response to the homogenous “Di” list and classified “To” faster than “Ta”. During day 5 participants responded slightly faster to the homogenous “Di” list, showed no change for RTs in the homogenous “Bi” list and responded slower to the orthogonal list. Again, “To” was classified faster than “Ta”. Finally, there was no significant interaction between *test day* and *response set*, Pillai’s *F*(4,6)=3.20, *p*= .10.

In conclusion, the findings strongly suggest a syllabic interference effect induced by the McGurk illusion. Results are highly similar to the findings from the NMcSI condition (see Figure 5). Thus participants perceived the integrated percept as an orthogonal syllable (even though its auditory component remained constant). This suggests that participants are likely to have failed selecting their attention to the relevant information. Again, the impact of practice and repeated exposure to the McGurk illusion neither facilitated nor inhibited this effect.

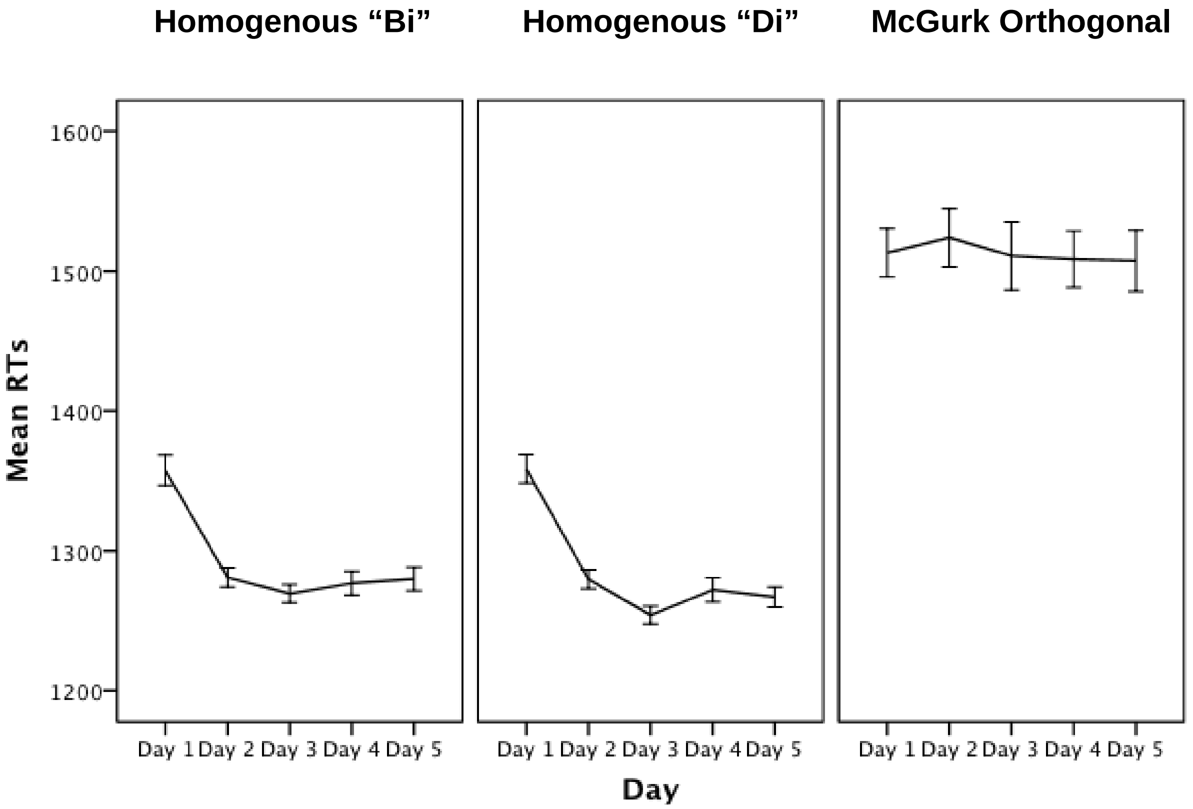
Homogenous “Bi” **Homogenous “Di”** **McGurk Orthogonal**

Figure 3: Mean RTs in milliseconds separated by *test day* for each *type of stimulus list* in the McISI experiment. From the left, the first box shows the temporal pattern for the homogenous “Bi” list, the second box for the homogenous “Di” list and the third box displays the orthogonal list including the McGurk percept. Error bars indicate 95% confidence interval around the mean.

McGurk Prevented Syllabic Interference Experiment (McPSI)

A third three-way Anova was conducted. This included the within-subjects factors *response set* (“Ta” vs “To”), *type of stimulus list* (homogenous “Di” vs. homogenous “McGurk” vs. illusionary “Di”) and *test day* (five). Interestingly, as can be seen in Table 4, there was no main effect of *type of stimulus list*, *F*(2,18)=.74, *p*=.49. Secondary to the goal of this experiment, a significant interaction between *response set* and *type of stimulus* list was found, *F*(2,18)=16.54, *p*<.001 η^{2p}= .648, showing that participants responded faster to “To” as opposed to ”Ta”. Tertiary to the goal of this experiment, a significant interaction between *response set* and *test day* was found, *F*(4,36)=3.66, *p*<.05 η^{2p}=.189. Contrasts revealed that participants showed a steady decrease in RTs for classifying "To" faster than "Ta" over the course of five days.

Table 4: Mean across participant median correct response times (milliseconds). Calculated for each stimulus presented in different stimulus lists collapsed across test days. Note the similar RTs between the three different types of stimulus lists.

Experiment	**Stimulus List**	**Mean RTs (SD)**
III. McGurk Prevented Syllabic Interference	Homogenous "Di"	1250.59 ms (8.44)
	Homogenous McGurk "Di"	1254.43 ms (8.39)
	Illusionary Homogenous "Di"	1249.31 ms (8.74)

No significant interaction between *test day* and *type of stimulus list* was found, *F*(8,72)=1.12, *p*=.22 (see Figure 4). Further, no significant interaction between *test day*, *type of stimulus list* and *response set* was found, *F*(8,72)=.84, *p*=.57.

In conclusion, results indicate an absence of syllabic interference by virtue of the McGurk illusion. Such results can only be predicted when the participant's behaviour in a speeded classification task is based on an automatic integration of audiovisual information. Although the original audiovisual components of the McGurk syllable were *orthogonal* to the audiovisual components of the remaining stimuli (in the illusionary homogenous list), the integrated percept was perceived as *homogenous*. The clear difference in the pattern of responses between this condition and both conditions described above (see Figure 5) demonstrates evidence for audiovisual integration to be mandatory. Finally, increased exposure to this condition did not cause participants to recognise the conflicting input and syllabic interference did not occur after five days of testing.

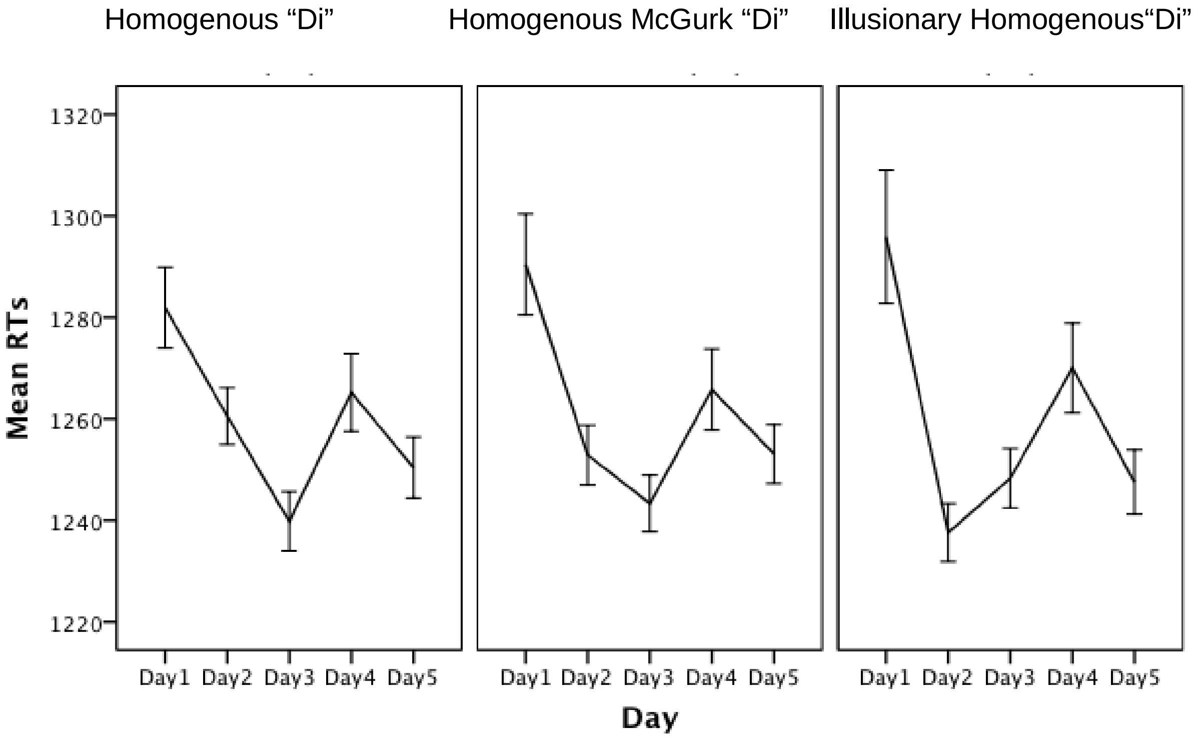

Figure 4: Mean RTs in milliseconds separated by *test day* for each *type of stimulus list* in the McPSI experiment. From the left, the first box shows the temporal pattern for the homogenous "Di" list, the second box for the homogenous "McGurk Di" list and the third box displays the orthogonal list including the McGurk percept and the matching "Di". Error bars indicate 95% confidence interval around the mean.

Analysis of Extent of Syllabic Interference Effects

The *mean difference* of RTs between the homogenous and orthogonal lists can be taken as an index of the extent of syllabic interference, i.e. the interference produced by the task-irrelevant variations in the second syllable. In order to compare the results of this study, this index was calculated separately for each day and experiment and the resulting values plotted in Figure 5. A value of zero on this index indicates no syllabic interference, positive values indicate that responses were proportionally longer in the orthogonal condition. That is because for each experiment, the mean RTs from the homogenous condition were subtracted from the orthogonal condition.

The index calculated for the NMcISI experiment shows the extent to which irrelevant variation in the second syllable reduced performance in the speeded classification of the first syllable. In line with the predictions of this study, the high values confirm a strong syllabic interference effect reflected in the participants' mean RTs.

The index calculated for the McISI experiment shows the extent to which the alternated visual stimulus induced variation in the auditory percept. As expected, the values were equally high when compared to the mean differences from NMcISI experiment suggesting a syllabic interference effect of similar magnitude. The index calculated for the McPSI experiment shows the extent to which the alternated audiovisual stimulus eliminated actual variability in the auditory percept. The low values support the assumption that as long as the integrated percept evoked by the

McGurk illusion is homogenous to the remaining auditory stimuli participants will not experience syllabic interference.

For each index the direction of the line demonstrates whether there is a trend towards the syllabic interference effect getting stronger or weaker over time. Notably, in both the NMcISI experiment as well as the McISI experiment there is a clear syllabic interference effect observable that tends to become stronger over successive test days (although this trend is non-significant). Importantly, this syllabic interference effect is largely absent in the McPSI experiment and shows no obvious change over time. When looking at the values and direction of lines plotted in Figure 5, no significant trend can be observed that would indicate a reduction of automatic integration due to practice. In other words, either reduced syllabic interference effects in the McISI experiment and/ or increased syllabic interference effects in the McPSI experiment.

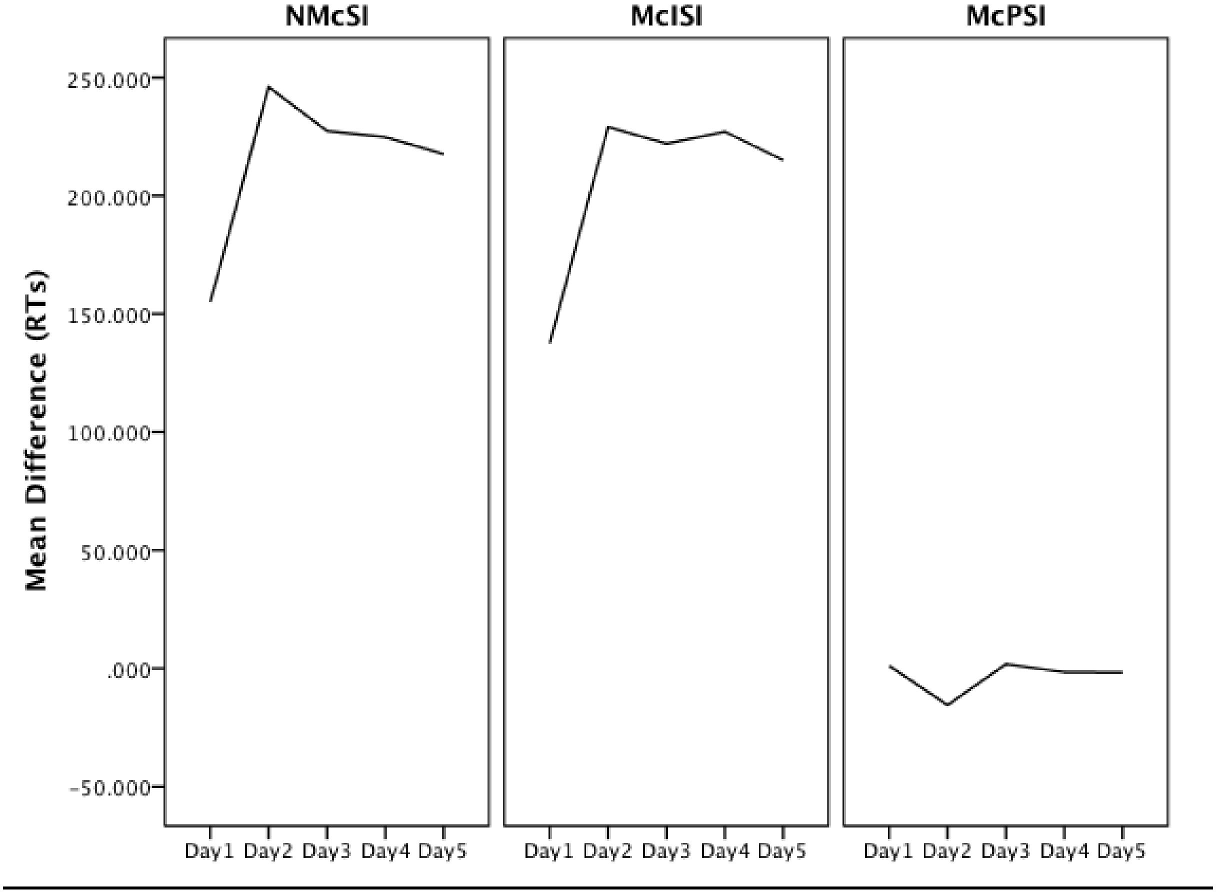

Figure 5: Index for extent of syllabic interference effects. Difference of mean RTs between types of stimulus lists in milliseconds separated by days and conditions. Plotted values reflect mean RTs from the homogenous condition that were subtracted from the orthogonal condition.

DISCUSSION

Data obtained from the present experiment successfully repeat the findings of Soto-Faraco et al. (2004) and thus strongly support the idea that integration of audiovisual speech information occurs in an automatic fashion. Here, more precise testing conditions did not falsify the results reported by Soto-Faraco et al. (2004) but produced even stronger contrasts between the evocation and elimination of syllabic interference effects on the basis of the McGurk illusion. Moreover, the absence of effective practice on syllabic interference (at least over the five-day period tested) provided additional support for accepting the role of automaticity in speech integration.

The advantage of this study was that task (i.e. speeded classification of the first syllable) and manipulation (variation in the irrelevant syllable) were completely unrelated. This method allowed for measuring selective attention *indirectly* (participants were not required to report the McGurk illusion itself but only the syllable prior to the illusion). As predicted, results from the NonMcGurk Syllabic Interference experiment (NMcSI) confirm the general syllabic interference effect (Pallier, 1994) for the audiovisual case. As further expected, the findings from the McGurk Induced Syllabic Interference experiment (McISI) and McGurk Prevented Syllabic Interference experiment (McPSI) illustrate that this effect indeed can be induced as well as eliminated, respectively, by virtue of a McGurk percept. In both experiments, NMcSI as well as McISI, responses (RTs) to orthogonal lists were significantly slower. This shared pattern of results confirms the prediction that the integrated percept causes the same syllabic interference effect on participants' speeded classification as any other syllabic variation when being presented within an orthogonal list. In other words participants seem to behave equally in their responses regardless of whether the percept is the result of an integrative process of *incongruent* or *congruent* information.

Interestingly, in the McISI experiment the perceived variation occurred only in the percept. Since the auditory components of the McGurk syllables were solely matching with 'non-McGurk syllables', the illusion could only be realised through successful integration of the audiovisually incongruent cues. Thus, participants failed to ignore variations in the second syllable based on what they had *perceived* not on what their senses had originally *received*. This suggests that audiovisual integration is mandatory and occurs before selective attention can be allocated. Moreover this demonstrates that the visual input changes auditory perception to the extent that its actual signal becomes indistinguishable. Nevertheless, as already noted by Soto-Faraco et al. (2004), participants who responded slower to orthogonal lists including the McGurk combination may not have experienced an interference effect, but simply detected the conflicting sensory input. This seems unlikely when looking at the exact similarity between the baseline RTs from the NMcSI experiment (a genuine auditory change in the second syllable) and RTs from the McISI experiment (a McGurk induced auditory change). However, such a possibility cannot be completely ruled out at this stage.

Therefore, the McPSI experiment represented a vital test for the assumption that audiovisual integration occurs automatically. Results from this experiment indicated that the syllabic interference effect can be eliminated when an 'illusionary homogenous list' is created; that is a perceived absence of variability in the irrelevant syllables induced by the McGurk effect. Non-McGurk syllables appear as matching with the McGurk percept but are orthogonal to each other in *both* their auditory as well as visual components (unlike the McISI condition which contained matching auditory components). Such results can only be explained if the integration of audiovisual speech cues is an involuntary mechanism and automatic in nature. Consequentially, this suggests that audiovisual integration cannot be overridden, even when this process negatively impacts on the participants' performance in a specific task.

Selective attention has been looked at before in research aiming to examine audiovisual integration. For example, an earlier study by Massaro (1987) demonstrated the occurrence of the McGurk illusion, although observers explicitly focussed their attention to unimodal sensory information. Others have also used a variant of the speeded classification task with audiovisual stimuli but avoided linguistic materials (Marks, 2004). Green and Kuhl (1991) used a speeded classification paradigm with audiovisual speech material. However, the role of automatic integration was not addressed appropriately. For example, in their second experiment, participants were asked directly to report their experience of the McGurk illusion. The study by Soto-Faraco et al. (2004) avoided such reports and the replication of their findings in the present study allows for a more confident interpretation of their results.

Nevertheless, a possible role of practice was not considered by Soto-Faraco et al. (2004). In their study, participants were only exposed to these series of abstract audiovisual stimuli in one short experimental session. It might be that this new task did not allow for any higher level elaboration of the audiovisual components due to the fast presentation of trials and the quick response required. Therefore one may conclude that audiovisual integration is only automatic in certain situations e.g. under stress or in ambiguous situations. Nevertheless, results from this experiment suggest that such a ‘context-specific’ model of audiovisual integration seems not applicable. Participants demonstrated an increase in proficiency in the speeded classification task (especially during the first two test days). But the magnitude of the syllabic interference effect as well its elimination - by virtue of the McGurk illusion - remained stable throughout the course of five testing days.

This suggests that this integration of crossmodal speech seems to be a low-level phenomenon based on bottom-up processing of sensory information. The absence of a high-level direction of processing was reflected in the stability of the results over repeated testing days. Practice and increased exposure to the McGurk illusion did not reverse the effects described above i.e. deautomise the binding of audiovisual information through increased cognitive effort. In fact the temporal trends if anything were in the opposite direction.

Notably, in the McISI condition a marginally significant effect of *test day* was found showing that participants became slightly faster at the identification task overall. This could reflect a minimal influence of practice. Nevertheless, it must be acknowledged that this study was testing relatively low numbers of participants (N=10) on just five consecutive days. Following the general scientific ethos “absence of evidence is not evidence of absence”, it is possible that such factors are responsible for the lack of any significant temporal effects.

However, the present findings do suggest that audiovisual integration occurs at an implicit level and therefore help explaining the ubiquitous nature of the McGurk illusion. For example, the occurrence under obvious cognitive discordances such as the combination of a male voice with a female face (Green et al., 1991) or when lips become reduced to kinematic properties such as multiple reflective dots (Rosenblum & Saldaña, 1996). Further, the present findings correspond with behavioural data such as those from the McGurk study by Gentillucci et al. (2005). They demonstrated participants’ unawareness of the incongruent visual input and the extent to which it changed their auditory perception as well as their verbal responses. In a sense their observations provide different behavioural evidence for the same underlying mental mechanism tested in the present study: the implicit, mandatory but attention-free processing of audiovisual information.

Such results contradict popular theories of attention, such as the Feature-Integration model, which propose that crossmodal integration is mediated by and thus dependent on attention (Treisman & Gelade, 1980). Results from both, the McISI and McPSI experiments, demonstrate that the fusion of audiovisual cues occurs *before* selective attention can be allocated, i.e. reflects a pre-attentive phenomenon. This is supported by findings from Colin et al. (2002) examining the mismatch negativity (MMN) ERP in audiovisual speech integration. This specific MMN signal is traditionally associated with an early, pre-attentive distinction between acoustic stimuli.

Further, results from the McISI experiment challenge the notion that selective attention occurs independently for each sensory modality (Wickens, 1984). Participants failed to select their attention only onto the auditory modality in the speeded classification task. Instead the incongruent visual cue changed the auditory perception to form a novel percept that in turn produced a syllabic interference effect (slower RTs).

The obligatory integration across modalities observed on a behavioural level coincides with modern brain imaging studies (Calvert, Bullmore, Brammer, Campbell, Williams, McGuire, Woodruff, Iversen & David, 1997). The neural synthesis of audible and visible elements of speech can be limited to specific cerebral areas. Using functional magnetic resonance imaging (fMRI), Calvert, Campbell and Brammer (2000), located increased haemodynamic activity in the superior temporal sulcus (STS) during audiovisual stimulation. The cells in this location are thought to be qualitatively distinct, i.e. fine-tuned to integrate afferent crossmodal information. This specialisation of cells has been supported by neuroanatomical studies (involving non-human primates) demonstrating that the cells within the STS receive converging inputs from e.g. visual and auditory cortices, hence also called heteromodal cortex (Jones & Powell, 1970). Finally, an fMRI navigated TMS study revealed the cortical locus for the McGurk effect, which lies within the STS (Beauchamp, Nath & Pasalar, 2010). These findings show that the capacity for audiovisual integration is a hardwired and essential part of the brain’s architecture, representing a deeply ingrained mental operation.

The occurrence of the McGurk illusion in non-human primates further reflects an ecological importance of crossmodal perception (Ghazanfar & Logothetis, 2003; Ghazanfar, Maier, Hoffman & Logothetis, 2005). Additionally, the illusion emerges earlier in life than visual word reading (Rosenblum, Schmuckler & Johnson, 1997; Kushnerenko, Teinonen, Volein & Csibra, 2008), suggesting that audiovisual integration is possibly more “deeply rooted” than the processing of visual word-forms (Lifshitz, Bonn, Fischer, Kashem & Raz, 2013).

Automatic binding seems to be a logical consequence of verbal interaction that not only enhances mere speech perception but also language comprehension (Schwartz, Berthommier & Savariaux, 2004). If extensive attentional resources are permanently required for successful integration of audiovisual cues, speech perception would not be a fast and efficient cognitive process. Results so far reveal increasing evidence for audiovisual integration being an implicit process and it becomes critical to imagine speech perception without the seemingly effortless confluence of crossmodal information.

As already discussed, audiovisual speech integration appears to be governed by a dominant role of automaticity, making it an involuntary mental process. Evidence seems to exclude the possibility for a top-down regulation of such low-level sensory integration and suggests a merely perceptual mechanism. This issue has been addressed in this study by employing an extended testing period and results illustrated that this process could not become deautomised through practice. Arguably, this 'rigid' nature characterises the McGurk illusion as being 'highly automatic' when comparing it with other perceptual illusions, e.g. the Stroop interference (Stroop, 1935), which appears to become reduced through practice (MacLeod & Dunbar, 1988). In other words, the McGurk illusion seems to be immune to top-down control. Thus the case of audiovisual speech integration seems to be somewhat special in the sense that computations carried out by the speech perception module remain *cognitively impenetrable* (Pylyshyn, 1999), i.e. stay unaffected by information from other modules or from the experimental context.

However, a study by Colin, Radeau and Deltenre (2005) suggests that the mere *quality* of *multimodal input* can evoke participation of higher-level intervention. In two experiments, they manipulated the salience of McGurk stimuli by alternating the sizes of faces and auditory intensity. Although the visual manipulation demonstrated a less strong effect, the weaker the auditory signal was the more frequent a McGurk illusion occurred. The effect of auditory salience reduction is perceived as an indication for audiovisual integration being modulated by both, sensory (as demonstrated in this study) but also cognitive variables.

According to the authors, sensory factors - such as the salience of the auditory and visual signals - may impinge on audiovisual integration mechanisms at an early perceptual stage, since they modulate the input signal. In the present study, the audiovisual stimuli were of equal salience. Thus perceptual weight that is allocated to each modality (auditory and visual) is equal. On the contrary, a weak auditory signal would elicit an attentional shift that results in a "top-down reweighing" of the audiovisual signals and thus retroactively modulate the weight to each modality (Colin et al., 2005). Such findings do not challenge the present results but clearly emphasise that the integration of speech must be a flexible mechanism nonetheless. It may be that extensive cognitive effort is required when audiovisual information are not necessarily of equal quality (or salience). Consequentially, Colin et al. (2005) do not deny the evidence for audiovisual integration being a perceptual process but *add* the possibility for potential higher-level processing in certain situations. In other words, speech integration remains automatic as long as the audiovisual stimuli are of equal salience.

Nevertheless, recent research suggests that automatic processes indeed can become deautomised when the state of mind is altered artificially. As shown by Palmer and Ramsay (2012) variations in conscious awareness can influence the convergence of auditory and visual information.

Déry, Campbell, Lifshitz and Raz (2014) expanded upon this evidence by utilising hypnosis and post-hypnotic suggestion. Employing a classic McGurk stimulus, they examined whether following a post-hypnotic suggestion to prioritise the auditory input, participants would be able to ignore the visual influence and correctly identify the auditory information presented (ignore the visual influence). They found that highly hypnotically susceptible individuals were less affected by the McGurk illusion and classified the auditory information correctly compared to less hypnotically susceptible individuals. These findings suggest that under extraordinary circumstances low-level mental functions can be mediated by top-down regulations even if deemed as highly automatic and rarely amenable to behavioural interventions. Although such findings are fascinating and raise the question what or who controls mechanisms of the mind, one has to critically evaluate the implications of such research. Can hypnosis be regarded as a natural state? If not, in what way are these findings ecologically valid? If suggestion can only effectively facilitate selective attention towards one modality when pre-stimulus brain states are altered then one may argue that these findings provide *additional* evidence for audiovisual integration being a mandatory process. In other words, in a natural state of mind audiovisual integration will always occur automatically until this state is calibrated through means of hypnotic priming. Additionally, even if hypnosis is regarded as an adequate tool to challenge the stability of automaticity in speech perception, it remains a biased one since not every individual exhibits such a high susceptibility to hypnosis as required to achieve results reported by Déry et al. (2014).

CONCLUSION

The McGurk illusion shows that auditory speech perception is not a direct reflection of information received by the ears. This study replicated the indirect experimental approach by Soto-Faraco et al. (2004) and expanded the research radius by introducing an extended testing period to assess the effects of practice. Especially results from the experiments containing the McGurk illusion (McISI and McPSI) revealed even stronger evidence for the assumption that audiovisual speech integration takes place on an implicit level. In summary, findings and the corresponding literature imply that the perception of speech is based upon an automatic, brief and involuntary integration of audiovisual information that allows humans to efficiently comprehend verbal language without the recruitment of further attentional resources. Additionally, this mechanism seems to be a deeply ingrained mental operation that is largely immune to high-level interventions such as endogenous selective attention or practice.

ACKNOWLEDGEMENTS

The author would like to thank Dr. Michael Pilling and Mr. Wakefield Carter for continuous inspiration and academic support.

REFERENCES

Alsius, A., Navarra, J., Campbell, R. and Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology,* 15, 839– 843.

Beauchamp, M. S., Nath, A. R. and Pasalar, S. (2010). fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *The Journal of Neuroscience*, *30*(7), 2414-2417.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C. and Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport*, *14*(17), 2213-2218.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., Woodruff, P. W., Iversen, S. D. and David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, 276, 593 – 596.

Calvert, G. A., Campbell, R. and Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, *10*(11), 649-657.

Colin, C., Radeau,M., Soquet, A., Demolin, D., Colin, F. and Deltenre, P. (2002). Mismatch negativity evoked by the McGurk–MacDonald effect: a phonetic representation within short-term memory. *Clinical Neurophysiology*, 113, 495–506.

Colin, C., Radeau, M., Soquet, A. and Deltenre, P. (2004). Generalization of the generation of an MMN by illusory McGurk percepts: voiceless consonants, *Clinical Neurophysiology*, 115, 495–506.

Colin, C., Radeau, M. and Deltenre, P. (2005). Top-down and bottom-up modulation of audiovisual integration in speech. *European Journal of Cognitive Psychology*, *17*(4), 541–560.

De Gelder, B. and Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in cognitive sciences*, *7*(10), 460-467.

Déry, C., Campbell, N. K., Lifshitz, M. and Raz, A. (2014). Suggestion overrides automatic audiovisual integration. *Consciousness and cognition*, 24, 33-37.

Frens, M. A., Van Opstal, A. J. and Van der Willigen, R. F. (1995). Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. *Perception & Psychophysics*, *57*(6), 802-816.

Garner, W. R. (1974). *The processing of information and structure*. Hillsdale, NJ: Erlbaum.

Gentilucci, M. and Cattaneo, L. (2005). Automatic audiovisual integration in speech perception. *Experimental Brain Research*, 167, 66–75.

Ghazanfar, A. A. and Logothetis, N. K. (2003). Neuroperception: Facial expressions linked to monkey calls. *Nature*, *423*(6943), 937-938.

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L. and Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *The Journal of Neuroscience*, *25*(20), 5004-5012.

Green, K. P. and Kuhl, P. K. (1991). Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 17, 278 – 288.

Green, K. P., Kuhl, P. K., Meltzoff, A. N. and Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics*, *50*(6), 524-536.

Green, K. P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech*, 3-25.

Jones, E. G. and Powell, T. P. S. (1970). An anatomical study of converging sensory pathways within the cerebral cortex of the monkey. *Brain*, *93*(4), 793-820.

Kaiser, J., Hertrich, I., Ackermann, H. and Lutzenberger, W. (2006). Gamma-band activity over early sensory areas predicts detection of changes in audiovisual speech stimuli. *Neuroimage*, *30*(4), 1376-1382.

Kislyuk, D.S. (2006). Visual speech affects discrimination of syllables in the auditory cortex: an MMN study, in: Paper Presented at the 7th International Multisensory Research Forum (IMRF), Dublin, Ireland, June 18th–21st.

Kislyuk, D. S., Möttönen, R. and Sams, M. (2008). Visual processing affects the neural basis of auditory discrimination. *Journal of cognitive neuroscience*, *20*(12), 2175-2184.

Kushnerenko, E., Teinonen, T., Volein, A. and Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proceedings of the National Academy of Sciences*, *105*(32), 11442-11445.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention *Journal of Experimental Psychology: Human Perception and Performance*, 21, 451–468.

Lavie, N. (2005). Distracted and confused? Selective attention under load. *Trends in Cognitive Sciences*, 9, 75–82.

Lifshitz, M., Aubert Bonn, N., Fischer, A., Kashem, I. F. and Raz, A. (2013). Using suggestion to modulate automatic processes: From Stroop to McGurk and beyond. *Cortex*, *49*(2), 463-473.

MacLeod, C. M. and Dunbar, K. (1988). Training and Stroop-like interference: evidence for a continuum of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 126.

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

Manuel, S. Y., Repp, B. H., Studdert-Kennedy, M. and Liberman, A. M. (2005). Exploring the “McGurk effect”. *The Journal of the Acoustical Society of America*, *74*(S1), S66-S66.

Marks, L. E. (2004). Cross-modal interactions in speeded classification. In Calvert, G., Spence, C. and Stein, B.E. (Eds.), *Handbook of multisensory processes* (pp. 85-105) . Cambridge, USA: MIT Press.

Massaro, D. W. (1987). *Speech perception by ear and eye: a paradigm for psychological inquiry* (pp. 66–74). Hillsdale, NJ: Lawrence Erlbaum Associates.

Morrell, L. K. (1968). Temporal characteristics of sensory interaction in choice reaction times. *Journal of Experimental Psychology*, *77*(1), 14.

Näätänen, R. and Alho, K. (1995). Mismatch negativity – a unique measure of sensory processing in audition. *International Journal of Neuroscience,* 80, 317–337.

Navarra, J., Alsius, A., Soto-Faraco, S. and Spence, C. (2010). Assessing the role of attention in the audiovisual integration of speech, *Information Fusion*, *11*(1), 4-11.

Öğmen, H., Breitmeyer, B. G., Todd, S. and Mardon, L. (2006). Target recovery in metacontrast: The effect of contrast, *Vision research*, *46*(28), 4726-4734.

Pallier, C. (1994). Role de la syllabe dans la perception de la parole: e ́tudes attentionelles. Doctoral dissertation presented at the E ́ cole des Hautes E ́ tudes en Sciences Sociales,Paris (available online at: http://www.pallier.org/papers/thesis/thsplra4.pdf, last accessed 06.01.2014)

Palmer, T. D. and Ramsey, A. K. (2012). The function of consciousness in multisensory integration. *Cognition*, *125*(3), 353-364.

Pylyshyn, Z. (1999). Is vision continuous with cognition?: The case for cognitive impenetrability of visual perception. *Behavioural and brain sciences*, *22*(03), 341-365.

Rosenblum, L. D. and Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(2), 318.

Rosenblum, L. D., Schmuckler, M. A. and Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics*, *59*(3), 347-357.

Santangelo, V. and Spence, C. (2007). Multisensory cues capture spatial attention regardless of perceptual load, *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 1311–1321.

Santangelo, V. and Spence, C. (2008). Is the exogenous orienting of spatial attention truly automatic? Evidence from unimodal and multisensory studies. *Consciousness and Cognition*, 17, 989–1015.

Schwartz, J. L., Berthommier, F. and Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, *93*(2), B69-B78.

Sekuler R., Sekuler A.B. and Lau R. (1997). Sound alters visual motion perception. *Nature*, 385, 308.

Simons, D.J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76-80.

Smith, E. G. and Bennetto, L. (2007). Audiovisual speech integration and lipreading in autism. *Journal of Child Psychology and Psychiatry*, *48*(8), 813-821.

Soto-Faraco, S., Navarra, J. and Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition*, *92*(3), B13-B23.

Soto-Faraco, S. and Alsius, A. (2009). Deconstructing the McGurk–MacDonald illusion. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 580.

Stein, B. E., Meredith, M. A., Huneycutt, W. S. and McDade, L. (1989). Behavioural indices of multisensory integration: orientation to visual cues is affected by auditory stimuli. *Journal of Cognitive Neuroscience*, *1*(1), 12-24.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, *18*(6), 643.

Sumby, W. H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212-215.

Tiippana, K., Andersen, T. S. and Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology*, *16*(3), 457-472.

Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12, 97 – 136.

Vatakis, A. and Spence, C. (2006). Audiovisual synchrony perception for music, speech, and object actions. *Brain research*, *1111*(1), 134-142.

Whelan, R. (2008). Effective Analysis of Reaction Time Data. *Psychological Record*, *58*(3).

Wickens, C. D. (1984). Processing resources in attention. In Parasuraman R. and Davies, D.R. (Eds.), *Varieties of attention* (pp. 63–102). San Diego, CA: Academic Press.

Wojciulik, E. and Kanwisher, N. (1998). Implicit but not explicit feature binding in a Balint’s patient. In Schneider, W.X. and Maasen, S. (Eds.), *Mechanisms of Visual Attention: A Cognitive Neuroscience Perspective* (pp. 157-181). Hove, East Sussex: Psychology Press.