The McGurk Illusion in the Oddity Task

Tobias S. Andersen

1 Department of Informatics and Mathematical Modeling, Technical University of Denmark
ta@imm.dtu.dk

Abstract

Despite many studies of audiovisual integration in speech perception very few studies have addressed the issue of cross-modal response bias. Using synthetic acoustic speech, the current study demonstrates the McGurk illusion in the oddity task which is not prone to cross-modal response bias. This new method may be an useful tool for determining under which conditions audiovisual integration of speech occurs.

Index Terms: McGurk illusion, response bias

1. Introduction

One of the most striking demonstrations of audiovisual integration in speech perception is the McGurk illusion in which the sight of the talking face change the phoneme perceived in the acoustic speech signal. The typical example of this illusion is when a voice saying /ba/ dubbed onto the video of a face saying /da/ is perceived as /ga/ [1]. The McGurk illusion has been demonstrated in scores of studies, many trying to determine under which conditions audiovisual integration of speech occurs (see e.g. [2, 3] for reviews).

Given that many studies have sought to determine when perceptual audiovisual integration occurs in speech perception, it is remarkable that few have addressed the issue of cross-modal response bias. For incongruent audiovisual speech, cross-modal response bias could work so that the phoneme perceived from lip-reading would bias the responses that observers give identifying the acoustic speech signal. There are several good reasons to suspect that watching the incongruent face could bias observers’ responses independent from any perceptual effect. For instance, a response bias is induced across visual submodalities in the Stroop task in which observers are presented with names of colors written in colored ink [4]. Their task is to name the ink color, but they often, particularly when responses are speeded, fail to do so and report instead the written color name. Clearly, the written color name does not influence the perceived color so the effect is due to a response bias effective across sub-modalities. Cross-modal bias effect also occurs across audition and vision. In visual detection tasks, the presence of an auditory signal can change observer’s response bias so that they are more likely to report the presence of the visual signal both when it is, in fact, present but also when it absent [5]. Notably, in this case, the cross-modal response bias can occur in addition to a perceptual effect as the auditory signal also increases the cross-modal response bias. The typical McGurk illusion on observers’ responses in an identification task and it is therefore important to be able to distinguish the two effects in order to assess their respective strengths.

Signal Detection Theory (SDT) offers a well-founded method to separate response bias from perceptual effects [8] but this approach is almost never used in studies of audiovisual speech. One exception is a study by Kislyuk et al. [9]. They studied observers’ ability to identify speech sounds /va/ and /ba/. In the auditory condition the speech sounds were presented without a video of a talking face and the proportion of correct responses was very high. In the McGurk condition the speech sounds were dubbed onto a video of a talking face saying /va/ and the proportion correct was much lower mostly because /ba/ was misidentified as /va/ due to the McGurk illusion. Kislyuk et al. calculated the sensitivity measure, d’; and showed that it was significantly lower in the McGurk condition. This result provides strong evidence of perceptual audiovisual integration which cannot be ascribed to a change in response bias.

Using the paradigm, which Kislyuk et al. used, it is, however, possible to construct scenarios where perceptual integration and cross-modal response bias would have the same effect on observers’ responses, and hence also on d’. First, take the example of full visual capture of perfectly discriminable speech sounds. In the auditory condition the two speech sounds are perfectly discriminable, so the proportion correct is maximal, i.e. equal to one. In the McGurk condition, the speech sounds are dubbed onto a talking face congruent with one of the speech sounds. If perceptual integration causes full visual capture, both speech sounds may be identified as one and the same, namely the one spoken by the face. However, a strong cross-modal response bias could have the same effect: It could bias observers to always respond according to what they lip-read from the face. Hence cross-modal response bias and perceptual integration are indistinguishable in this scenario. As another example, take two speech sounds played back at a lower signal-to-noise ratio. In the auditory condition, performance would be less
than perfect but better than chance so the proportion correct would lie somewhere between 0.5 and 1 given only two response options. In this case, perceptual integration in the McGurk condition would have two effects. First, the McGurk illusion would lower the proportion correct. Second, for the other speech token, the face would be congruent and increase the proportion correct [10]. Whether these two opposite effects on the proportion correct would cancel is unknown and therefore we do not have a clear prediction of whether the McGurk effect would decrease sensitivity. For both types of stimuli, the effect of viewing the talking face would be to increase the number of responses given according to the face. The effect of a cross-modal response bias could be very similar. Hence, also in this example, perceptual integration could be difficult to distinguish from cross-modal response bias.

To avoid confounding sensitivity and response bias in the current study we used the oddity task in which the response categories are orthogonal from the stimulus attributes of interest (here, phonemic category). In the oddity task three stimuli are presented consecutively. Two of the stimuli are identical and the observer has to determine which of the three stimuli stands out from the other two. In the McGurk condition we compared congruent and incongruent audiovisual speech stimuli which differed acoustically but were visually identical. As auditory and visual stimuli we used /aba/ and /ada/. When we used /ada/ as visual stimulus, we expected it to cause auditory /aba/ to be perceived as /ada/ and hence harder to distinguish from a veridical auditory /ada/ than in the auditory condition because no articulating face was shown. One problem with this approach is that observers might be confused with respect to the task when seeing the face in addition to hearing the voice. If they try to pick the odd stimulus based on the face rather than the voice, their performance would suffer as the visual stimulus did not change within trials. To control for this we also used visual /aba/ which we expected either to cause auditory /ada/ to be perceived as /aba/ [11] or no illusion [11] neither of which should be hard to distinguish from a veridical auditory /aba/. Hence this control condition of using visual /aba/ should check whether observers base their judgment on the visual stimulus as performance in the McGurk condition should only be poorer than in the auditory condition if they do so.

These predictions rely on observers comparing the three stimuli solely on the basis of their phonetic content. Observers were, however, free to base their performance in the oddity task on any acoustic cue, phonetic or otherwise. Therefore, in order to maximize observers’ reliance on phonetic cues, we minimized other cues by using synthetic speech stimuli, for which acoustic features could be controlled with great precision.

2. Methods

9 native Danish speakers (20-29 years of age, 6 males) participated in the study. The stimuli were videos of a native Danish speaker (the author) articulating /ada/ or /aba/ and acoustic dipphone synthetic speech tokens (also /ada/ or /aba/). Audiovisual synchrony between the face and the voice was optimized as follows: First, synthetic speech tokens were generated using the Carsten voice from Mikro Værkstedet, Odense, Denmark. These were played back repeatedly while recording video and audio of the talker trying to articulate the same tokens synchronously. The synthetic and recorded natural speech sounds were then compared found not to match in duration. Therefore, a new set of synthetic speech stimuli were created with vowel and consonant durations similar to the recorded speech sounds. Only these synthetic speech stimuli were used in the actual experiments. The audiovisual stimuli were created by dubbing the synthetic speech stimuli onto the video of the talker using video editing software, so that it was synchronized with the recorded acoustic speech signal.

The experiment consisted of five blocks, all of which were preceded by written instructions. The blocks are described in the following:

1. In Block 1 the two unimodal auditory, two incongruent audiovisual and two congruent audiovisual stimuli were presented 20 times in randomized order. The task of the observer was to identify the consonant using a keyboard. The observer was free to type in several consonants.

2. In Block 2 the two unimodal visual stimuli were presented 20 times in a repeated stimulus paradigm. The task of the observer was to identify the consonant from lip reading. Responses were given as in the previous block.

3. Block 3 introduced the auditory oddity task. Three auditory speech stimuli were presented consecutively with a short break in between. Two of the three stimuli were identical while one stood out (e.g. /ada/, /aba/, /ada/). The task of the observer was to indicate which of the three stimuli differed from the other two by pressing 1, 2 or 3 on a standard keyboard. A still image of the talker with mouth closed was displayed concurrently and observers were instructed to look at it. The block continued until the observer gave five consecutive correct responses.

4. Block 4 introduced the visual oddity task. Three silent videos of the talker were presented consecutively with a short break in between. Two of the three stimuli were identical while one stood out. The task of the observer was to indicate which of the three videos differed from the other two by pressing 1, 2 or 3 on a keyboard. The block continued until the observer gave five consecutive correct responses.

5. Block 5 contained three adaptive staircase procedures (accelerated stochastic approximation [12, 13]) aimed at finding the 67% threshold. For all three staircases the task was to discriminate between auditory /aba/ and /ada/. The only difference between the staircases I-III was the visual stimuli:

I. Auditory stimuli dubbed onto the face articulating /ada/

II. Auditory stimuli dubbed onto the face articulating /aba/

III. Unimodal auditory (with a still image)

In addition to the staircase procedures, Block 5 also contained trials in which the oddity task was visual. The purpose of these catch trials was to intermix auditory and visual oddity tasks so that observers had to look at the face throughout the experiment to perform well. The observers were instructed beforehand that the odd stimulus might stand out acoustically, visually or both. Visual catch trials were presented with a probability of 25%. In the remaining trials, a stimulus was picked with equal probability from the staircase procedures that had not yet reached 20 reversals. The block ended when all staircases had reversed 20 times. The observers’ thresholds were calculated as the average sound intensity during the last five trials in each staircase.
In Blocks 1-4 the sound level was constant at 65 dB(A). The staircase procedures in Block 5 also started at this level, which they could not exceed. The sound intensity of the visual catch trials in Block 5 was the average of the current sound intensity in the three staircase procedures. All stimuli were presented through desktop loudspeakers in constant 40 dB(A) background noise from computer ventilation. The visual stimuli were 16 cm wide and 25 cm high and were presented on a computer monitor. The experiment was controlled with the Psychophysics Toolbox [14, 15]. Observers were sitting 57 cm from the monitor and loudspeakers with their head resting in a chin rest.

3. Results

The results from the identification tasks in Blocks 1-2 are tabulated in Table 1 as mean ± standard error of the mean (SEM) response percentages. The unimodal auditory and unimodal visual stimuli were identified correctly in most trials although one observer tended to confuse auditory /ada/ with /aga/. Unsurprisingly, the percentage correct for congruent audiovisual stimuli was even higher than that for unimodal audiovisual stimuli. As expected, auditory /aba/ was perceived as /ada/ when dubbed onto visual /aba/. However, contrary to expectations, auditory /ada/ was not perceived as /abda/ when dubbed onto visual /aba/. Only two observers responded hearing /aba/ and each only did so once, so we excluded those two trials from the analysis. Still, observers were influenced strongly by the visual component of this stimulus type as they reported to her /aba/ very often. This pattern of responses is atypical and we cannot explain why it occurred.

The training tasks of Blocks 3-5 were completed by all observers within 60-70 seconds except one observer who used 104 seconds to complete the auditory training task. Since a single trial lasted approximately 12 s this means that all observers needed only few trials to learn the auditory and visual oddity tasks.

In Block 5, the mean ± SEM of the thresholds in the auditory condition was 51 ± 1 dB(A). In the McGurk condition, the threshold was much higher, 64 ± 1 dB(A). When the visual stimulus was /aba/. Since the staircases were limited to 65 dB(A) this means that observers’ performance was very poor even at this clearly audible sound intensity. This result is strong evidence of visual capture by the McGurk illusion and cannot be ascribed to cross-modal response bias. When the visual stimulus was /aba/ the results were less clear: The mean threshold was 57 ± 3 dB(A) reflecting that for four observers, the threshold was within 1 dB of the threshold in the unimodal auditory condition indicating no visual influence, for three observers it was within 1 dB of their threshold in the other McGurk condition indicating strong visual influence and for two observers it was somewhere in between indicating moderate visual influence.

Comparing performance in the identification and oddity tasks we found that all observers were strongly influenced by visual /aba/ in both tasks while visual /aba/ seemed to influence some observers in the identification task and some observers in the oddity task. To see if the influence from visual /aba/ in the two tasks was correlated we compared two measures of visual influence: The difference between the thresholds in the McGurk and auditory conditions measures visual influence in the oddity task and the proportion of incorrect responses measures the visual influence in the identification task. The two measures are plotted in Figure 1. We found that the two measures did not correlate well ($R^2=0.03$). Three subjects (only two visible in Figure 1 due to data overlay) had a strong McGurk illusion in the identification task but not in the oddity task. This could be due to these observers not perceptually integrating this audiovisual stimulus while being prone to cross-modal response bias. All of the observers who were strongly influenced by vision in the oddity were also strongly influenced by vision in the identification task.

![Figure 1](image.png)
It is not the first to use the oddity in multisensory research although we believe it is the first to apply it to the study of audiovisual speech. Hillis et al. used the oddity task to study integration of visual and haptic information when estimating the slant of a surface [16]. Contrary to our findings, they found that observers were able to distinguish the stimuli in the oddity task on the basis of cues combined across sensory modalities or single cues from either modality. When the multisensory stimulus was congruent, observers tended to combine information across vision and touch leading to improved performance compared to when only unisensory information was available. When the multisensory stimulus was incongruent, observers tended to base their judgment on the modality in which the stimuli were more easily discriminable. This shows that although observers are able to integrate information about surface slant across vision and touch, they still have access to the unisensory representations. This was not the case in our study. If observers had had access to the unisensory auditory representation they should have performed the oddity task equally well in the McGurk and auditory conditions, which they, in general, did not.

Using the oddity task to demonstrate the McGurk illusion rests on using very homogenous speech sounds that do not differ markedly on cues not relevant to identifying the consonant. In the current study we used synthetic speech for this purpose. With the speech synthesizer we could set the fundamental frequency, duration and intensity of each phoneme to be the same for /ada/ and /aba/. Had we not used synthetic speech, the oddity task could perhaps be solved based on cues not influenced by the McGurk illusion, such as the fundamental frequency, duration or intensity of the vowels. Hence it is possible to have a strong and truly perceptual McGurk illusion, which still would be easy for observers to distinguish from the congruent stimulus mediating the same phoneme if some trivial acoustic cue, such as the overall loudness, differs between the two stimuli. The oddity task is thus a tough test of audiovisual integration. One could say that poor discrimination performance in the oddity task is sufficient, but not necessary, evidence of audiovisual integration. Poor performance in the identification task, i.e. being unable to identify the acoustic component of audiovisual speech, is a more tolerant test. It probes only the relevant phonetic percept without being influenced by phonetically irrelevant acoustic cues but it is also prone to response bias. Hence it provides only necessary, but not sufficient evidence of audiovisual integration. Applying both task to the study of audiovisual speech perception may thus help us separating the effects of cross-modal response bias from those of perceptual integration in future studies.