Integration of Audiovisual Speech and Priming Effects

Azra N. Ali

School of Computing and Engineering, University of Huddersfield, Huddersfield, England

Abstract

Humans report fusion /da/ when presented with audio /ba/ aligned with visual /ga/. Over the last three decades most researchers have neglected the velar McGurk fusion: audio /ba/ with visual /da/ eliciting /ga/ fusion, and audio /pa/ with visual /ta/ resulting in /ka/ fusion. Cathiard [1], claimed that these latter types of fusion are laboratory curiosities which do not occur embedded in French VCV syllables. We conduct two experiments; in the first experiment, incongruent segment is embedded in real English words and in the second experiment we use a priming approach to bias the perception towards either the audio channel or the visual or the expected fusion. Results show that velar fusion perceptions are not just a product of isolated nonsense syllables, but are robust perceptions formed by integration of audio and visual channels. Thus, using both types of McGurk fusion has potential to be used as a probing tool for exploring the phonological organization of lexical entities.

Index Terms: Speech perception, McGurk effect, Priming

1. Introduction

People use both the acoustic and the visual modality to understand speech. Many people with a hearing impairment can understand speech by lip-reading, which shows that linguistic information, can be conveyed by vision [2, 3]. When the audio channel is noisy, information from the visual channel significantly improves the accuracy of speech perception, as was demonstrated quantitatively by Sumby and Pollack [4]. Thus, congruent audio and video speech channels not only provide two independent sources of information, but do so complementarily: each is strong when the other is weak. Furthermore, the complementarity makes accurate speech perception more resistant to channel noise [4,5]. Evidence for strong interaction between audio and visual speech channels in human speech perception is found in the well-known McGurk effect [6]. If humans are presented with temporally aligned but conflicting audio and visual stimuli – now known as 'incongruent stimuli' - the perceived sound may differ from that present in either channel. The McGurk effect is robust, occurs in many languages and in presence of noise etc..

2. McGurk fusion

McGurk and MacDonald asked their recording technician to create a videotape with the audio syllable /ba/ dubbed onto a visual /ga/, most normal adults reported hearing /da/ or /ba/. When the participants were presented with only one modality (visual or audio), or stimuli without audiovisual incongruity, they reported the syllables correctly.

2.1. The systematics of McGurk fusion

In this paper, the following bracket notation (Figure 1, or alternatively as detailed in Table 1) is used to describe incongruent stimuli and the percepts that they evoke. Taking the above incongruent stimulus as an example:

\[ \lambda(\text{ba} || \text{ga}) \rightarrow (\text{da}) \] (1)

Figure 1: Incongruent audiovisual stimulus notation

In cases where a group of participants reported alternative fusion responses, say some reporting (da), and others (ba), we adapt the generative grammar rule format to accommodate this and state the relevant fusion rates:

\[ \lambda(\text{ba} || \text{ga}) \rightarrow (\text{da})_F(0.64), (\text{ba})_F(0.27), (\text{ga})_F(0.04) \] (2)

The equation in (2) represents a situation in which 64% of participants report /da/, 20% report /ba/, 12% the acoustic option /ba/, and 4% the visual option /ga/.

In 1978, MacDonald & McGurk [7] carried out a second study, extending their experiments with additional incongruent syllables of the CV type: /ta, da, ma, na/. They retained /ba, ga, pa, ka/ from their earlier study, and investigated all possible auditory-visual combinations of these eight syllables. The results revealed that when participants are presented with conflicting audio and visual channel, place fusion is greatest with a labial sound in the audio channel than when a labial mouth shape was evident in the visual channel. In summary, McGurk and MacDonald found two prevalent types of fusion, detailed below.

\[ \lambda(\text{ba} || \text{ga}) \rightarrow (\text{da})_F(0.64) \ldots \text{alveolar fusion} \] (3)

\[ \lambda(\text{pa} || \text{ka}) \rightarrow (\text{ta})_F(0.50) \ldots \text{alveolar fusion} \] (4)

\[ \lambda(\text{ba} || \text{ga}) \rightarrow (\text{ga})_F(0.27) \ldots \text{velar fusion} \] (5)

\[ \lambda(\text{pa} || \text{ta}) \rightarrow (\text{ka})_F(0.50) \ldots \text{velar fusion} \] (6)

The first, traditional McGurk fusion involves a labial plosive in the audio channel, paired with a velar gesture in the visual channel, to give an alveolar percept, as in examples (3) and (4) above. The second fusion involves a labial plosive in the audio channel paired with alveolar gesture in the visual channel to give a velar percept, as in (5) and (6) above. It appears from the literature review that the second type of fusion has largely gone unnoticed by rest of the audiovisual community since the study of MacDonald and McGurk in 1978 [7]. Although, in 2001, Cathiard et al.[1], acknowledges the velar fusion (ga) in but fails to state or acknowledge the voiceless fusion (ka) in that was reported in MacDonald &
McGurk’s work [7:255]. Cathiard et al. conducted a research with French nonsense VCV syllables (voiced consonants only) and showed no velar fusion occurring. It is possible that the voicing contrast in French is phonetically very different to that of English in terms of timing of voicing onset. Thus, this could affect the degree of fusion expected. The research in this paper will show that the velar fusion are robust percepts formed by integration of audio and visual channels even when the priming biases towards the audio or visual channel.

3. Experiment 1

3.1. Stimuli

We used read English words to create the incongruent stimuli, as shown in Table 1. The video recordings were done using a standard 8 mm digital (Sony) Camcorder with built-in microphone for audio. The words (audio, visual and expected fusion) were randomized before they were recorded. From the resultant video recording, each word uttered was captured into *.avi files. For the creation of incongruent stimuli, standard editing software (Adobe Premier 5.5) was used. Words were paired up such as /prəɪd/ (video 1) and /trəɪd/ (video 2) which differed in one phonemic segment. The audio channel from video 1 of the pair was imported and aligned with the visual channel of video 2 of the pair. The experimenter made fine judgments of proper alignment manually, after previewing the video clip. After coarse alignment, the audio of video 2 (/trəɪd/) channel was erased. Incongruent stimuli and a few natural, fully congruent controls were then saved as *.avi files with a frame rate of 30 per second and a frame size of 640 x 512 pixels.

<table>
<thead>
<tr>
<th>Stimuli Type</th>
<th>Audio</th>
<th>Visual</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voiceless</td>
<td>54.5 (50)</td>
<td>22.7 (0)</td>
<td>22.7 (50)</td>
</tr>
<tr>
<td>Voiced</td>
<td>28.4 (19)</td>
<td>20.4 (27)</td>
<td>51.4 (27)</td>
</tr>
<tr>
<td>Total</td>
<td>41.5 (38.5)</td>
<td>21.6 (13.5)</td>
<td>36.9 (38.5)</td>
</tr>
</tbody>
</table>

Table 1. Incongruent stimuli.

The results do highlight that even with this small sample of real word stimuli the velar McGurk fusion does occur and it is not an isolated case. However, the stimuli that were most vulnerable to velar fusion were stimuli 3, 5, 6 and 7. Only one stimulus did not generate a fusion. $A_{(prəɪd || trəɪd)}$, and in Section 8.0 we discuss this further.

![Figure 1: Percentage of responses for incongruent stimuli.](image)

The results are compared with MacDonald and McGurk [7] in Table 2. Our study shows greater velar fusion for voiced consonants than for voiceless consonants. But, for the combined stimuli (voiced and voiceless) the fusion rates were similar to MacDonald and McGurk findings, 38.5% and 36.9% respectively.

Table 2. Comparison of response rates.

3.2. Participants and procedure

Forty-four participants took part in the experiment, a mixture of both females and males, with an age range between 21 to 54 years. None had hearing problems and all either had normal vision or wore prescribed corrective lenses. Also, none of the participants had any linguistic knowledge. The participants were provided with an open choice response form to minimize any metalinguistic issues. There was no time limits set to complete the experiment, and no feedback was given to the participants and the experiment was conducted double-blind to eliminate experimental effects.

3.3. Results of Experiment 1

For congruent channels, over 97% of the participants accurately reported what the speaker was saying and for incongruent channels, there were distinctly different vulnerabilities to fusion amongst the stimuli, Figure 1. Not all participants experience the fusion, and the bar-chart simply shows the response rates in percentages across the stimuli.

We embedded the prime (text -- in lowercase) into a single frame in the visual channel that either biased towards the
audio channel or the visual channel or towards the expected fusion. The prime appeared on the screen for 40ms followed by the incongruent audiovisual target. The duration of the prime was set just above the threshold of perceivability. The incongruent with the prime stimuli were randomized amongst some fully congruent stimuli with no primes.

4.2. Participants and procedure

A total of 36 participants took part in the experiment: There were 3 separate groups of stimuli and participants: 12 participants in Group A (bias towards the audio channel), 13 participants in Group B (bias towards the visual channel) and 11 participants in Group C (bias towards the expected fusion). A mixture of both females and males, with an age range between 18 to 49 years took part in each group. None had hearing problems and all either had normal vision or wore prescribed corrective lenses. The participants were provided with an open choice response form. There were no time limits set to complete the experiment.

5. Results 2

For congruent stimuli, 100% of the participants accurately reported what the speaker was saying. For the incongruent stimuli with the prime, the results were quite striking, the velar fusion still occurred despite the bias towards the audio or the visual channel. The results in Table 4, clearly show, that when the prime biased towards the expected velar fusion, there was a slight increase in the fusion rate. But when the prime biased towards the audio or towards the visual channel, the fusion rates only marginally dropped.

Table 4. Response rate for prime experiment.

<table>
<thead>
<tr>
<th>Prime Bias</th>
<th>Audio</th>
<th>Visual</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A: towards audio</td>
<td>27.1%</td>
<td>14.6%</td>
<td>58.3%</td>
</tr>
<tr>
<td>Group B: towards visual</td>
<td>25.0%</td>
<td>15.4%</td>
<td>59.6%</td>
</tr>
<tr>
<td>Group C: towards fusion</td>
<td>22.7%</td>
<td>13.6%</td>
<td>63.7%</td>
</tr>
<tr>
<td>Group D: No prime*</td>
<td>29.5%</td>
<td>9.1%</td>
<td>61.4%</td>
</tr>
</tbody>
</table>

*from experiment 1, for the 4 stimuli

Exploratory chi-squared test (categorical analysis) revealed no significant difference in velar fusion rates between the four groups (χ² = .305, df = 3, p = .959). The data was analyzed with the Kruskal-Wallis Analysis of Variance test which also showed similar finding, no significant difference in fusion rates amongst the four groups (H = .121, df = 3, p = .990). We further tested whether the non significant finding were due to a particular stimulus from the four stimuli. Thus, we tested a two-way Friedman test (non-parametric test) which showed no significant finding (χ² = .40, df = 2, p = .861), which indicated that the fusion rates were not stimuli and group dependent.

6. Further Analysis – Experiment 1 & 2

6.1. Audio modality and visual modality

Studies on unimodal - auditory modality and visual modality have shown that the perceptual intelligibility of consonants varies. Perceiving a place of articulation contrast, such as that between labial /b/ and alveolar /d/, is difficult via sound but relatively easy via sight. On the other hand, a contrast of voicing state, such as that between voiceless /s/ and voiced /z/ is easy to hear but scarcely visible [3, 5, 9]. A number of studies have provided evidence that adults with no hearing problems can be just as skilful in visual lip-reading as adults with hearing impairment [3, 4, etc.]. Binlin et al [10] derived nine viseme groups for consonants: {f, v}, {p, b, m}, {w}, {l, n}, {j, 3}, {t, 0, ð}, {l, d, s, z} and {k, g}. Although, a review of the literature shows that visemes /d/ and /g/ and /t/ and /k/ can belong to the same viseme group. Thus, the purpose of this experiment was to ensure that the observed McGurk fusion reported in Experiment 1 and 2 were not due to poor audio or visual stimulation. Therefore, in separate experiments the stimuli were tested in audio-mode only (with no visual channel) and visual-mode only (no audio channel, participants required to lip-read).

6.1.1. Method – audio only

In the audio only experiment, the visual channel was digitally removed from stimuli thus leaving only an audio (.wav) file. Ten different participants from the audiovisual experiment took part in the experiment and were asked to record what they had heard via headphones. No response forms were provided, participants were asked to write down what they heard. None of the participants had any hearing problems.

6.1.2. Method – visual only

In the visual only experiment, the audio channel was digitally removed from stimuli thus leaving only the visual channel. Ten different participants from the audiovisual experiment took part in the experiment and were asked to lip-read what the speaker was saying to the first consonant of the word. Participants did not have any prior experience nor were they given any training sessions in lip-reading. The same response sheets were used as in the audiovisual experiments.

6.1.3. Results – audio only and visual only

The results were analyzed in terms of percentage of correct identification for audio only and visual only which showed 100% accuracy rate for audio only and 58% accuracy rate for visual only. In the visual only, no significant difference were found in correct identification score between velar viseme (/g, k/) and alveolar viseme (/d, t/), ANOVA: F = 1.75, df = 1, p = .235. At first glance it appears that the correct identification score (58%) for consonants were relative high for visual only, as participants were not given any training for lip-reading. This was not an experimenter effect because double-blind experimentation was used and response forms were non-directive (included an open choice). But, our findings are in line with the work of Massaro [11] with CV syllables and also with Dekle et al. [12] and Sams et al. [13] with English and French words. Correct identification score for the target consonant in these studies were: for audio only 100%, 97% and 94% respectively, and for visual only 55%, 69% and 62% respectively. However, in the work of Cathiard et al. [1] their participants showed greater accuracy rate for visual only (~88%) than for audio only (~78%).

6.2. Word frequency and vowel context effects

The second part of the further analysis was to investigate word frequency effects to establish whether the expected fusion word had a higher frequency count than the audio word or the visual word. We used British National Corpus (BNC) to obtain word frequencies for the words listed in Table 1. For example, in the stimulus /breɪn/ | /bɹɛn/, → /greɪn/}_{0.63}.
the expected fusion word /gæn/ is a low frequency word than the audio word /bæn/ or the visual word /dren/ and yet it generated a relatively high fusion rate of 63%. Working with relative frequency rather than absolute frequency, the results showed that velar fusion rates are not statistically significant for high and low frequency words (χ² = 1.88, df = 1, p = .665).

Also, there were no coarticulation effect that could have resulted in a velar fusion occurring. The consonant that followed the incongruent segment was always with an /t/ which minimized any coarticulation effect occurring.

7. Discussion

The traditional McGurk effect was first reported in 1976 and then in 1978 the velar McGurk fusion was reported. However, the velar fusion has largely gone unnoticed in the audiovisual community until 2001 by Cathiard et al [1] but they did not find any velar fusion occurring in their nonsense VCV (voiced consonant) syllable. Also, they fail to report the velar fusion in voiceless consonants in CV syllables that was reported by MacDonald and McGurk [7:255].

In this paper we have shown that the velar McGurk fusion does occur even in real words, it occurs both in voiced and voiceless consonants which confirms the original finding of MacDonald and McGurk in 1978 [7:255]. We have further shown that when a prime biases either towards the audio word or the visual word the velar fusion was still persistent. These velar perceptions were not due any poor audio or visual quality as this was demonstrated in unimodal experiments. Participants accurately perceived the audio word and were able to distinguish velar visemes /g, k/ from viseme /d, t/.

Munhall et al. [14] provided some account of variability of the (da) fusion across talkers. They measured the opening of the mouth (vertical lip separation) in pixels, showing that certain talkers produced greater McGurk fusion than others. Currently, we are carrying out measurements of articulatory movement (vertical lip separation) of viseme /d/ and /t/ and viseme /g/ and /k/ from our stimuli. The aim is to establish whether the visual consonant /d/ or /t/ resembled more closely to the articulatory movement of velar /g/ or /k/, which may explain why stimulus (pru: d || tru: d), did not result in a velar fusion. In addition we intend to explore other visual gestures measurements, for e.g. width of the lips, roundness etc.

In all of our experiments we have always opted for an open choice response form, allowing participants to freely record their true responses to incongruent and congruent stimuli. Our future studies will record participant’s real-time eye-movement data whilst watching the stimuli and reporting their percepts. The aim is to correlate true regions of visual interest with reported percepts. This may explain the variation of fusion across stimuli and participants.

7.1. Phonological models

Since the seminal work of McGurk and MacDonald in 1976, the traditional McGurk effect (ba || ga)→(da): has been studied extensively in psychophysical and cognitive fields. Most of the studies have used nonsense CV or VCV syllables. The broad pattern of results from these studies has been that fusion is qualitatively rather robust against a wide range of source variability. We have used both types of McGurk fusion for probing phonological organization of lexical entities. In our studies, we use incongruent audiovisual stimuli with real words (in isolation and embedded in sentence context) that probes phonological syllabic structure, syllable boundaries and morphological binding in English and Arabic [15, 16].

8. Conclusions

The current research has shown that the velar fusion is robust and is not a laboratory curiosity. For future studies we propose brain imaging methods, which have the potential to locate fusion events (both types of McGurk fusion) in the neural substrate of subjects from a wide range of language cultures and establish a degree of language-universality.

9. References