THE ROLE OF SPATIAL SEPARATION ON VENTRILOQUISM AND
MCGURK ILLUSIONS

M. RADEAU (1) (2) & C. COLIN (1)

(1) Free University of Brussels, Belgium
(2) National Fund for Scientific Research, Brussels, Belgium

ABSTRACT

This study aimed at assessing whether two audiovisual interactions (the McGurk and the ventriloquism effects) are affected by different degrees of spatial separation between the auditory and the visual signals. The materials consisted in trains of three audiovisual monosyllables. They were visually displayed in front of the participants’ head on a TV screen and auditorily played through one of nine hidden loudspeakers (placed every 20°, from straight ahead to 80° to the left or to the right). There were two different conditions in which the speaker’s face was presented either upright or inverted. Each condition included an identification task (to measure the McGurk effect) and a localization task (to estimate the ventriloquism effect). The ventriloquism effect was maximal at 20° and decreased as the loudspeaker location moved away to the left or to the right, but it was not affected by inverted presentation of the speaker’s face. The McGurk effect exhibited the reverse pattern.

INTRODUCTION

In case of moderate spatial conflict between auditory and visual signals, participants are likely to ignore or to underestimate the spatial separation between the two sources. This illusion, known as the ventriloquism effect, concerns the localization of auditory and visual information. Audiovisual interactions also occur in the identification of speech. As shown by McGurk & MacDonald [1], when confronted to discrepant auditory and visual speech information, people often report hearing phonetic segments that are not presented auditorily or visually. Two kinds of illusions have been reported: fusions and combinations. For example, when an acoustic /ba/ is dubbed onto a visual /ga/, a fused response /da/ is likely to occur, but with the reverse presentation, the subsequent perception is a combination such as /bga/.

One may wonder whether these two audiovisual interactions are subserved by the same perceptual mechanisms or by different ones. First, their domains are different: ventriloquism concerns localization while audiovisual speech concerns identification of speech events. In the visual modality, there is evidence for the “what” and “where” problems being subserved by different neural pathways [2]. Identification is considered to be mediated by a ventral pathway and localization by a dorsal pathway.

One of us [3, 4] previously suggested that the system underlying the ventriloquism effect is modular and governed by its own rules, which would be the gestalt principles of common fate and proximity. Indeed, several studies showed that the ventriloquism effect is affected by temporal desynchronisation [5] and that it decreases as the spatial separation between the signals increases [6]. Moreover, the system underlying the ventriloquism effect seems to be cognitively impenetrable. In particular, it was shown to be unaffected by cognitive manipulations like the use of a realistic face-voice situation [7]. In audiovisual speech, the gestalt principle of spatial proximity would play a less critical role [3,4], but a realistic face-voice situation seems to be necessary. Bertelson et al. [8] showed that, contrary to the ventriloquism effect, the McGurk illusion is indeed independent of the spatial separation between the sources but is reduced by face inversion. This dissociation between the two illusions suggests that they may be subserved by different mechanisms.

The aim of the present work was to assess to what extent both phenomena are affected by different degrees of spatial separation between the signals. Whereas previous studies of the ventriloquism effect used rather small separations (0° to 25° [6] and 0° to 32° [8]), we extended the separation till 80° to the left and to the right of straight ahead (by steps of 20°). In addition, the susceptibility of both illusions to the realism of the situation was examined by presenting the speaker’s face upright in one condition and upside-down in another condition.

METHOD

Participants

Thirty-two participants (19 to 33 years), 16 per condition, were paid for taking part in the experiment. They were all French speaking, without reported history of hearing disorder and with normal or corrected-to-normal vision.

Materials

The materials consisted of four CV monosyllables (bi, gi, pi, ki) articulated by a man whose only the lower part of the face was filmed (from the top of the nose till the chin).

The stimuli were constructed on a Panasonic AG-A770 editing controller. We edited four incongruent stimuli (auditory /bi/ and /bi/ dubbed onto visual /bi/ and...
consonants (/ki/, and conversely) and four congruent ones (bi, gi, pi, ki) that were used as control items for the McGurk effect. The experimental stimuli were created by replacing the original audio signal of a given item by the audio signal corresponding to another item of about the same length in frames. The sound was synchronized with the last picture preceding mouth opening. Each item consisted in a train of three repetitions (without interval) of the same monosyllable.

Four blocks were edited, each one including 24 trials (eight different stimuli repeated three times each) presented in random order. The presentation order of the stimuli varied from one block to another. Each block was repeated four times on a same tape and four tapes were edited this way.

Procedure
The participants seated in front of a table, at 1m from a Sony color screen (width: 44 cm; height: 31 cm). They had the head on a chin rest during the whole experiment. The stimuli were played, at an average level of 40dB, through one of nine loudspeakers (Monarch SPH-60X) hidden under the table, one of them being straight ahead and the other ones being displayed every 20° from straight ahead to 80° to the left and to the right. In one half of the cases, the sound was played through the loudspeaker located straight ahead. In the other cases, it came from one of the remaining eight loudspeakers. Each auditory stimulus was played the same number of times through each loudspeaker. A scale graduated from 1 to 18 in steps of 10°, from 90° left to 90° right of straight ahead, was displayed in a circular arc on the table just above the loudspeakers.

Before beginning the experiment, the auditory localization abilities of our participants were assessed with the screen switched off. The participants heard three series of 24 trials, each auditory syllable being presented twice in each loudspeaker location. After each trial, they had to tell where they thought the sound came from.

After this pre-test, the participants had to carry out two tasks, identification and localization. In the identification task (devoted to the measure of the McGurk effect), they had to repeat aloud the syllable they had heard. In the localization task (aimed at estimating the ventriloquism effect), they had to tell whether the auditory and visual stimuli came from the same location or not by saying “same” or “different”.

Each task was preceded by a 32 trials training block. The whole experiment was achieved in two sessions of one hour each. During each session, half the time was devoted to the identification task and the other half to the localization task.

There were two different conditions in which sixteen different participants took part. In the first one, the TV screen was in a normal position; in the second one, it was inverted from 180°, in such a way that the speaker’s face was upside-down.

RESULTS

Pre-test
For the four kinds of auditory stimuli and the nine loudspeakers locations, the participants made on average an error of 8.1° (range: 2.8° to 17.6°) in localization. The error rates were rather congruent from one stimulus to another and from one loudspeaker location to another.

Identification task
For each item, the participants were likely to give an answer corresponding to the auditory information, or to the visual information, or an audiovisual response (that is a combination or a fusion). We considered as illusory responses, the audiovisual responses and also the visual responses. Errors on control trials were subtracted from the total number of illusion responses for each participant.

The results are presented in Table 1 as a function of Type of illusion, Type of consonant, Screen position and Loudspeaker location. In order to facilitate the description of the results, the loudspeakers were labeled L1 to L9. L1 corresponding to 80° left, L5 to straight ahead and L9 to 80° right.

There were on average 79% of fusions for the upright screen condition and 51% for the inverted screen condition. For the combinations, these percentages were respectively 63% and 47%.

We carried out an Anova with Type of consonants (voice vs voiceless), Loudspeaker location (nine levels) and Type of illusions (fusions vs combinations) as within-participants factors and Screen position (upright vs upside-down) as between-participants factor.

<table>
<thead>
<tr>
<th></th>
<th>comb. fusions</th>
<th>voiced</th>
<th>voiced</th>
<th>upright</th>
<th>inverted</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>80°L</td>
<td>63.6</td>
<td>63.0</td>
<td>66.7</td>
<td>59.9</td>
<td>76.6</td>
<td>50.0</td>
</tr>
<tr>
<td>60°L</td>
<td>54.7</td>
<td>63.6</td>
<td>62.5</td>
<td>55.8</td>
<td>67.2</td>
<td>51.1</td>
</tr>
<tr>
<td>40°L</td>
<td>52.1</td>
<td>69.3</td>
<td>64.6</td>
<td>56.8</td>
<td>72.9</td>
<td>48.4</td>
</tr>
<tr>
<td>20°L</td>
<td>56.3</td>
<td>64.6</td>
<td>59.4</td>
<td>61.5</td>
<td>69.8</td>
<td>51.1</td>
</tr>
<tr>
<td>centre</td>
<td>55.2</td>
<td>67.7</td>
<td>63.0</td>
<td>59.9</td>
<td>71.4</td>
<td>51.6</td>
</tr>
<tr>
<td>20°R</td>
<td>52.6</td>
<td>66.2</td>
<td>55.7</td>
<td>63.0</td>
<td>74.0</td>
<td>44.8</td>
</tr>
<tr>
<td>40°R</td>
<td>63.1</td>
<td>64.1</td>
<td>67.2</td>
<td>59.9</td>
<td>72.9</td>
<td>54.2</td>
</tr>
<tr>
<td>60°R</td>
<td>40.1</td>
<td>56.8</td>
<td>48.4</td>
<td>48.5</td>
<td>59.4</td>
<td>37.5</td>
</tr>
<tr>
<td>80°R</td>
<td>58.4</td>
<td>69.3</td>
<td>68.8</td>
<td>58.9</td>
<td>74.5</td>
<td>53.1</td>
</tr>
<tr>
<td>mean</td>
<td>55.1</td>
<td>64.9</td>
<td>61.8</td>
<td>58.2</td>
<td>71.0</td>
<td>49.1</td>
</tr>
</tbody>
</table>

Table 1: Percentages of illusory responses for both kinds of illusions (comb. = combinations), both types of consonants (voiced vs voiceless) and both screen positions (upright; inverted) as a function of Loudspeaker location (L= left; R= right).

Screen position was significant (F(1,30)=9.05, p<.01). We observed 22% illusions more when the speaker’s face was upright than when it was upside-down. Type of illusion was also significant (F(1,30)=5.23, p<.05), the percentage of fusions being 10% higher than that of combinations, as well as Loudspeaker location (F(8,240)=4.65, p<.01). The number of illusions were actually rather similar for all locations except for L8 (60° right) where the illusions were significantly less numerous than for the two adjacent locations (L8 vs L7: F(1,30)=25.48, p<.01; L8 vs L9: F(1,30)=23.61, p<.01). Among interactions, only the triple interaction between Type of consonants, Type of illusion and Screen position was significant (F(8,240)=4.65, p<.01).
Loudspeaker location was significant ($F(8,240)=2.05$, $p<.05$). Contrasts involving the screen position factor all went in the same direction as the main effect (more illusions when the screen was upright). For the contrasts involving the consonant, voiceless consonants gave rise to more illusions when the screen was inverted and when the sound came from L6 ($F(1,30)=6.06$, $p<.05$).

**Localization task**

The number of “same” vs “different” responses were calculated for each loudspeaker location. The participants gave a “same” response when they thought that the sound came from the speaker’s mouth. The percentages of “same” responses thus represent the ventriloquism effect for each loudspeaker location, except L5 (for which “same” was the expected and correct answer). Results are displayed in Table 2 as a function of Type of illusion, Type of consonant, Screen position and Loudspeaker location.

<table>
<thead>
<tr>
<th>Location</th>
<th>Illusions</th>
<th>Voiced Illusions</th>
<th>Voiceless Illusions</th>
<th>Upright</th>
<th>Inverted</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>80°L (L1)</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>60°L (L2)</td>
<td>8.6</td>
<td>6.5</td>
<td>7.5</td>
<td>7.5</td>
<td>6.7</td>
<td>8.3</td>
</tr>
<tr>
<td>40°L (L3)</td>
<td>19.8</td>
<td>17.5</td>
<td>14.3</td>
<td>22.9</td>
<td>31.0</td>
<td>6.3</td>
</tr>
<tr>
<td>20°L (L4)</td>
<td>44.6</td>
<td>42.5</td>
<td>41.7</td>
<td>45.3</td>
<td>62.5</td>
<td>24.5</td>
</tr>
<tr>
<td>Centre (L5)</td>
<td>96.4</td>
<td>97.7</td>
<td>96.4</td>
<td>97.7</td>
<td>96.6</td>
<td>97.4</td>
</tr>
<tr>
<td>20°R (L6)</td>
<td>84.0</td>
<td>47.9</td>
<td>42.7</td>
<td>49.2</td>
<td>43.8</td>
<td>48.2</td>
</tr>
<tr>
<td>60°R (L8)</td>
<td>4.5</td>
<td>1.3</td>
<td>0.3</td>
<td>1.6</td>
<td>0.8</td>
<td>1.0</td>
</tr>
<tr>
<td>80°R (L9)</td>
<td>0.5</td>
<td>0.3</td>
<td>0.8</td>
<td>0.0</td>
<td>0.8</td>
<td>0.0</td>
</tr>
<tr>
<td>Mean</td>
<td>24.7</td>
<td>25.0</td>
<td>23.6</td>
<td>26.2</td>
<td>27.9</td>
<td>21.8</td>
</tr>
</tbody>
</table>

Table 2: Percentages of “same” responses for both kinds of illusions (comb. = combinations), both types of consonants (voiced; voiceless) and both screen positions (upright; inverted) as a function of Loudspeaker location (L= left; R= right).

The Anova was the same as for the identification task. **Loudspeaker location** was significant ($F(8,240)=138.69$, $p<.01$). The “same” responses decreased significantly from straight ahead to the far left location (L4 vs L5: $F(1,30)=105.85$, $p<.01$; L3 vs L4: $F(1,30)=43.07$, $p<.01$; L2 vs L3: $F(1,30)=11.71$, $p<.01$; L1 vs L2: $F(1,30)=13.00$, $p<.01$). The illusions percentages also decreased from straight ahead to L8, L9 and L9 being non significantly different from each other (L6 vs L5: $F(1,30)=75.37$, $p<.01$; L7 vs L6: $F(1,30)=59.40$, $p<.01$; L8 vs L7: $F(1,30)=15.91$, $p<.01$). Besides, we found a significant difference between L3 and L7 ($F(1,30)=5.96$, $p<.05$): illusions were 9% more numerous to the left. **Type of consonant** was also significant ($F(1,30)=12.24$, $p<.01$), voiceless consonants giving rise to 3% of “same” responses more than voiced ones. The double interaction between **Type of consonant and Loudspeaker location** was significant ($F(8,240)=2.70$, $p<.01$) as well as that between **Type of consonant and Type of illusions** ($F(1,30)=7.97$, $p<.01$). For both interactions, all significant contrasts involving the consonant were congruent with the main effect. **Screen position and Loudspeaker location** interacted significantly ($F(8,240)=7.30$, $p<.01$). The “same” responses were more numerous with the screen upright than inverted for only two loudspeakers: L3 ($F(1,30)=11.73$, $p<.05$) and L4 ($F(1,30)=12.76$, $p<.01$). When the screen was upright, illusions were more numerous to the left (L4 vs L6: 18.7%, $F(1,30)=6.61$, $p<.05$; L3 vs L7: 22.1%, $F(1,30)=18.52$, $p<.01$). However, when the screen was inverted, there was a tendency toward the opposite pattern: more illusions to the right (L4 vs L6, 23.7%: $F(1,30)=10.56$, $p<.01$). The triple interaction between **Type of illusions, Screen position and Type of consonant** was significant ($F(1,30)=5.27$, $p<.05$), so was the **quadruple interaction** ($F(8,240)=4.28$, $p<.05$). For both interactions, the Consonant effect was still congruent with the main effect (more illusions with voiceless consonants) and, as regards the effect of Screen position, “same responses” were more numerous for the upright screen in some conditions only, especially for L3 and L4.

**CONCLUSION**

In summary, our data showed that the McGurk effect was globally unaffected by the spatial separation between the auditory and the visual signals, but it decreased when the visual stimuli were presented upside-down (49% relative to 71% for the normal presentation). The opposite pattern was observed for the ventriloquism effect. The number of illusions decreased as spatial separation increased but it was not affected by the inverted presentation of the visual stimuli. With the upright screen, there were more ventriloquism illusions when the auditory stimuli came from the participant’s left side, whereas with the upside-down screen, the ventriloquism illusions were more numerous on the right side.

The ventriloquism effect reached 45% (“same” responses for a 20° separation between the visual and auditory sources). This percentage decreased to 14% at 40°, and for larger separations, there was almost no illusion at all. When the screen was upright, there were more illusions with the loudspeakers located at the participant’s left side, whereas when the screen was inverted, the illusions were more numerous on the right side. It should be noted that the TV loudspeaker, although it was switched off, was apparent and located on the left side of the screen, in the normal position. It was of course on the right side when the screen was inverted. The left-right effect can probably be accounted for in terms of conceptual factors: attraction of the sound by the dummy TV loudspeaker. This explanation does not contradict the cognitive impenetrability assumption of the ventriloquism effect because this assumption only concerns the aftereffects of ventriloquism, and not immediate effects which have long been shown to be affected by response biases [3]. Thus, the left-right effect found here is probably a response bias, not a perceptual effect. Let us note that even if more numerous on the side of the dummy loudspeaker, illusions nevertheless occurred, in substantial number, on the other side.
We found no difference between voiceless and voiced consonants for the McGurk illusions. In our previous studies however [9] [10], we observed more combinations with voiceless consonants, whereas a tendency toward the reverse pattern was obtained for fusions. For the ventriloquism effect, on the contrary, illusions were more numerous with voiceless consonants. This result does not seem to have ever been reported in the literature. Let us note, however, that the present voiced-voiceless difference was very small. Thus the effect found here has to be re-examined in further studies.

The realism of our experimental situation did not influence the ventriloquism effect but affected the McGurk illusions. The inversion of the speaker’s face probably distorted the spatial arrangement of the articulators and hence reduced the visual saliency of the syllables. This result is in line with that of Bertelson et al. [8] and with other studies investigating the effect of a face inversion on the McGurk effect [11] [12].

We also replicated Bertelson et al. (1994)’s results concerning the effect of spatial separation. Even with spatial separations far more large than those used by these authors, the McGurk illusions were rather similar in size from one loudspeaker to another. The ventriloquism effect, on the contrary, was strongly affected by the spatial separation, no illusion occurring beyond 40° of separation. The experimental dissociation found here between these two audiovisual interactions and the fact that they also concern different domains suggest that they are subserved by different systems.

REFERENCES


Acknowledgments: this work has been supported by grants A.R.C. (96/01-203) and F.R.F.C. (8.4519.96). We thank W. Serniclaes and D. Demolin for their fruitful advice and we are grateful to F. Goossens and the C.A.V of the Free University of Brussels for their help in editing the stimuli.