VISUAL LIPREADING OF VOICING FOR FRENCH STOP CONSONANTS

C. COLIN (1), M. RADEAU (1) (2), D. DEMOLIN (1) & A. SOQUET (1)
(1) Free University of Brussels, Belgium
(2) National Fund for Scientific Research, Brussels, Belgium

ABSTRACT

This study examined whether visually presented bilabials consonants are better identified than velars in a CV (C = consonant; V = vowel) or VCV context. We also investigated whether voiced and voiceless consonants sharing a same place and manner of articulation could be differentiated from each other with visual cues only. Although it is generally assumed that voicing is mainly mediated by the auditory modality, one cannot discard the possibility that the production of a voiced stop consonant produces a pattern of facial cues that could be detectable visually. Two pairs of stop consonants (/b/-/p/ and /g/-/k/) were articulated by a man and by a woman speaker in two syllabic contexts (CV monosyllables or VCV bisyllables). The bisyllables were uttered according to three speaking rates: slow, medial and fast. The materials were edited on a videotape and presented on a TV screen without sound. After each trial, participants had to choose between several written possibilities what they had perceived. Percentage of correct identifications reached 42% on average for the four consonants. Errors mostly consisted in voicing confusions (37%). Place of articulation confusions occurred in only 8% of the cases. Correct identifications were more numerous for bilabials than for velars but more particularly for monosyllables. Voiced consonants were better identified than voiceless in both syllabic contexts, but especially for velars. This suggests that some voicing distinction is possible on the basis of visual cues.

INTRODUCTION

It is well known that seeing the face of the speaker enhances the intelligibility of speech in noisy environments (Sumby & Pollack, 1954) and also in normal conditions (Summerfield, 1979). One of the most convincing demonstrations of lipreading’s importance in speech perception is the McGurk effect. McGurk and McDonald (1976) showed that when confronted to discrepant auditory and visual speech, participants often report hearing a percept that does not correspond to the auditory information but integrates features from the visual input. Two kinds of illusions have been described: 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 response such as /bga/. It has been suggested that perceptual salience principles could underlie the occurrence of one or the other kind of illusion (Summerfield, 1987). For combinations, the visual information from the bilabial would be so salient that it could not be ignored. So is it also for the phonetic information provided by the velar, the burst being more intense as the point of occlusion moves back in the mouth (Dorman, Studdert-Kennedy & Raphael, 1977). Consequently, listeners report hearing both the visual and the auditory components. For fusions, the visual information for velars and the acoustic cues for bilabials would be more ambiguous, thus leading to a percept intermediate between sound and sight. In terms of this account, the two McGurk illusions would thus mainly differ as a function of the place of articulation of the visual and auditory consonants.

There are some data arguing for better visual identification of bilabials over velars in English. This was shown for consonants presented in a CV context (Walden, Prosek & Montgomery, 1977) and in a VCV context (Bengherel & Pichora-Fuller, 1982).

The present experiment aimed at assessing whether, in French, visually presented bilabial consonants are also easier to identify than velars in both CV and VCV contexts. It is generally assumed that whereas the auditory modality provides information about voicing and manner of articulation,
the visual channel predominantly conveys information about place of articulation and cannot enable to discriminate voicing contrast (Summerfield, 1987). However, facial cues, other than lip movements, could play a role in voicing identification. Participants could (maybe involuntary) use such cues in the visual identification of consonants. This potential effect of voicing could moreover interact with place of articulation. We thus asked participants to identify visual CV and VCV syllables, whose consonant was bilabial (voiced /b/ and voiceless /p/) or velar (voiced /g/ and voiceless /k/). The VCV bisyllables were uttered following three speaking rates (slow, medial and fast) in order to assess whether the slowing down of the visual speaking rate might increase identification scores.

METHOD

Participants
Sixteen students (14 women), 17 to 22 years old (mean 18.7 years old), participated in the experiment as part of an introductory psychology course. They were all French speaking, without reported history of hearing disorder and with normal or corrected-to-normal vision.

Materials
The materials consisted of eight CV monosyllables (ba, ga, pa, ka, bi, gi, pi, ki) and eight VCV bisyllables (aba, aga, apa, aka, ibi, igi,ipi,iki). They were articulated by two native speakers of French, a 35 years old woman and a 24 years old man. Only the lower part of their face was filmed (from the top of the nose till the chin), the lips playing the most important role in decoding visible speech (Benoît, 1996). The bisyllables were uttered following three speaking rates: slow, medial and fast.

The stimuli were constructed on a Panasonic AG-A770 editing controller. In a trial, the lip movements were preceded and followed by 800 ms video display of the speaker’s face with a neutral facial expression in order to avoid brutal transitions. Each item lasted about 2 to 3 seconds on average. Two tapes were edited, one for each speaker. Each tape consisted of four blocks, each corresponding to one of the three speaking rates of the bisyllables or to the monosyllables. A block was made up of 24 trials (eight different stimuli repeated three times each) presented in a random order.

Procedure
The participants seated in front of a table, at 75 cm from a Panasonic color screen (width: 33 cm; height: 25 cm) with the sound off. They had to choose between several possibilities (/b/, /g/, /p/, /k/, /t/, /d/, /bg/, /pk/ and /other/) the consonant they had perceived. Those answers were the same as in a previous McGurk study (Colin, Radeau & Deltenre 1998), using the same visual materials. Two items were separated by a three seconds interstimulus interval that consisted of a black screen period and during which participants had to write their answer. Two groups of participants were presented the two tapes in a different order. The session began by an eight trials training block and lasted about half an hour.

RESULTS

The participants gave, on average, 42% correct responses. Most errors consisted in voicing confusions (for example, /b/ instead of /p/ or /k/ instead of /g/) and occurred in 37% of the cases. Place of articulation confusions reached about 8%. Table 1 displays the percentages of the three kinds of responses as a function of speaker gender, item length and speaking rate.

<table>
<thead>
<tr>
<th></th>
<th>Bisyllables</th>
<th>Monosyllables</th>
<th>Global Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Slow</td>
<td>Medial</td>
<td>Fast</td>
</tr>
<tr>
<td>Correct Responses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/b/</td>
<td>44</td>
<td>39</td>
<td>53</td>
</tr>
<tr>
<td>/p/</td>
<td>53</td>
<td>47</td>
<td>47</td>
</tr>
<tr>
<td>/g/</td>
<td>52</td>
<td>56</td>
<td>60</td>
</tr>
<tr>
<td>/k/</td>
<td>29</td>
<td>32</td>
<td>27</td>
</tr>
<tr>
<td>tot</td>
<td>44</td>
<td>43</td>
<td>47</td>
</tr>
<tr>
<td>Voicing confusions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/b/</td>
<td>44</td>
<td>48</td>
<td>32</td>
</tr>
<tr>
<td>/p/</td>
<td>33</td>
<td>33</td>
<td>43</td>
</tr>
<tr>
<td>/g/</td>
<td>31</td>
<td>26</td>
<td>31</td>
</tr>
<tr>
<td>/k/</td>
<td>49</td>
<td>52</td>
<td>56</td>
</tr>
<tr>
<td>tot</td>
<td>39</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>Place of articulation confusions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/b/</td>
<td>4</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>/p/</td>
<td>8</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>/g/</td>
<td>6</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>/k/</td>
<td>7</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>tot</td>
<td>6</td>
<td>7</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 1: Percentages of correct responses, voicing confusions and place of articulation confusions as a function of item length (bisyllables vs. monosyllables)
and, for bisyllables, as a function of speaking rate (slow, medial or fast).

Percent of correct responses were submitted to two separate ANOVAs on each item length (monosyllables vs. bisyllables). Correct responses were averaged across speakers because they were very similar for both of them. Voicing (voiced vs. voiceless), Place of articulation (bilabial vs. velar) and, for bisyllables, Speaking rate (slow, medial or fast) were the within-participants variables.

Bisyllables

The correct responses did not differ as a function of Speaking rate or Place of articulation. Voicing, however, was significant ($F(1,15)=6.37$, $p<.05$), voiced consonants gave rise to 11% correct responses more than voiceless consonants. The interaction between Voicing and Place of articulation was also significant ($F(1,15)=24.33$, $p<.001$) and the bilabial /p/ being better identified than the velar /k/ ($F(1,15)=11.94$, $p<.01$).

Monosyllables

Both effects of Voicing and Place of articulation were significant but did not interact. Voiced consonants produced 15% correct responses more than voiceless ones ($F(1,15)=6.31$, $p<.05$) and bilabials elicited 17% correct responses more than velars ($F(1,15)=20.24$, $p<.001$).

Comparison between bisyllables and monosyllables

The results of the monosyllables were compared to the mean results of the three speaking rates of the bisyllables in an ANOVA with three within-participants factors (Item length, Voicing and Place of articulation) conducted on the correct responses (again averaged across speakers). Main effect of Item length was significant ($F(1,15)=12.12$, $p<.01$). Bisyllables produced 5% correct responses more than monosyllables. Voicing ($F(1,15)=7.36$, $p<.05$) and Place of Articulation ($F(1,15)=24.13$, $p<.001$) were significant too. Voiced consonants gave rise to 13% correct responses more than voiceless ones and bilabials to 11% more than velars. The interaction between Item length and Place of articulation was significant ($F(1,15)=7.37$, $p<.05$), showing that the main Item length effect was due to velars ($F(1,15)=11.82$, $p<.01$). As previously mentioned, bilabials elicited more correct responses than velars in monosyllables. Finally, the triple interaction between Item length, Voicing and Place of articulation also reached significance ($F(1,15)=6.72$, $p<.05$). This interaction enables us to precise that the Item length effect (more correct responses for bisyllables) was mainly caused by the consonant /g/.

Errors analysis

We conducted an ANOVA on errors as dependent variable with Type of errors (voicing confusions vs. place of articulation confusions), Voicing and Place of articulation as within-participants variables. Type of errors was highly significant ($F(1,15)=25.83$, $p<.0001$), voicing confusions being 29% more numerous than place of articulation confusions. Voicing was also significant ($F(1,15)=5.32$, $p<.05$) while Place of articulation did not reach significance ($F<1$). Voiceless consonants produced 5% more than voiced consonants. The interaction between Voicing and Place of articulation ($F(1,15)=5.11$, $p<.05$) indicates that the voiceless velar /k/ gave rise to more errors than the voiced velar /g/ ($F(1,15)=14.41$, $p<.01$), and that it also elicited more errors than the voiceless bilabial /p/ ($F(1,15)=5.35$, $p<.05$). In other words, /k/ was the most often misidentified consonant as regard both voicing and place of articulation (it was mostly confused with /p/, /t/ and /d/). Other interactions were not significant.

DISCUSSION

Two main issues underlay this study. First, we aimed at verifying the hypothesis according to which consonants with a frontal place of articulation are visually easier to identify than consonants with a back place of articulation. We also tested the possibility for participants to distinguish between two consonants differing only in voicing, and this on the basis of the visual information. Two pairs of consonants were examined: /b/-/p/ and /g/-/k/. They were uttered by two speakers (a man and a woman) according to two item lengths (monosyllables and bisyllables) and, for the bisyllables, according to three speaking rates (slow, medial and fast).

For the monosyllables, we found a very clear place of articulation effect, correct identifications being much more numerous for bilabials than for velars. However, the effect was less clear for bisyllables. This latter observation runs counter the data of Bengherel and Pichora-Fuller (1982) who found, in a VCV context, better identifications scores for bilabial consonants than for velar consonants. However, for CV context, our results are in accordance with the
confusion matrices of Walden et al. (1977), showing poorer consonant identification as the place of articulation moves back in the mouth. It may be that in a CV context, the initial position of the consonant facilitates the detection of the occlusion at the labial position (whereas for velars, the occlusion could be almost as difficult to detect in a VCV than in a CV context).

As regards voicing, we found more correct identifications for voiced than for voiceless consonants for both places of articulation in monosyllables. For the bisyllables, the same tendency was found but it was significant for velars only. Overall the results indicate that voicing can be partly extracted from the visual modality. The visual cues used to distinguish between voiced and voiceless consonants remain to be determined.

There was no difference in the correct identification scores between the three speaking rates of the bisyllables. Thus, a slow speaking rate does not seem to facilitate consonant identification in a VCV context. This is consistent with Ijsselstjik (1992) results showing that slowing down speaking rate by two, three or four times relative to a normal rate does not affect lipreading performance of words, phrases and sentences in hearing-impaired children. On the other hand, the lack of effect of fast rate in the present experiment is consistent with the results found by Massaro, Cohen and Gesi (1993) for CV syllables.

VCV items produced more correct responses than CV ones in the case of velars and especially /g/. It may be assumed that in the case of VCV items, the transition between the first vowel and the consonant could provide extra information helping to identify the consonant. However, it remains to explain why this should apply to velars only.

As shown in the errors analysis, voicing confusions occurred much more often than place of articulation confusions. This result seems logical because place of articulation information is provided mainly by the visual modality while voicing is mostly conveyed by the auditory channel. Both voicing and place of articulation confusions were the most numerous for /k/. The difficulty in identifying the phoneme /k/ might be partly responsible for the voicing effect found for velars.

In conclusion, our results show that bilabials are easier to identify than velars, but especially for monosyllables. Concerning voicing, we found that voiced consonants gave rise to more correct responses than voiceless consonants for velars in both syllabic context and for bilabials in CV context only. This voicing effect and the kind of visual cues that could be at work need of course to be re-examined in further research.

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 for his fruitful advice.