RUNNING HEAD: FACTORS AFFECTING THE McGURK ILLUSIONS

The role of sound intensity and stop-consonant voicing on McGurk fusions and combinations

Colin, C. (1), Radeau, M. (1) (2), Deltenre, P. (3), Demolin, D. (4) and Soquet, A. (4)

(1) Research Unit in cognitive Neurosciences CP 191, Université libre de Bruxelles, 50, av. Roosevelt, B-1050 Brussels, Belgium.
(2) F.N.R.S., Belgium.
(3) Evoked Potentials Unit, Brugmann Hospital, 4, Place Van Gehuchten, B-1020 Brussels, Belgium.
(4) Phonology Laboratory CP 175, Université libre de Bruxelles, 50, av. Roosevelt, B-1050 Brussels, Belgium.

Correspondence to:
Cécile Colin
Research Unit in Cognitive Neurosciences
CP 191
Université libre de Bruxelles
50, av. F. Roosevelt
1050 Brussels
Belgium

e-mail: ccolin@ulb.ac.be

Abstract

When presented with an auditory /b/ dubbed onto a visual /g/, listeners sometimes perceive a fused phoneme like /d/ while with the reverse presentation, they experience a combination such as /bg/. This phenomenon reported by McGurk and MacDonald (1976) is here investigated in French for both voiced and voiceless stop consonants, using two levels of auditory intensity (70 dB vs. 40 dB). In a first experiment, audiovisual incongruent monosyllables (A/bi/ V/gi/, A/gi/ V/bi/, A/ki/ V/pi/, A/pi/ V/ki/) uttered by a man and by a woman speakers were recorded and dubbed, using an analogical technology. In a second experiment, the same syllables articulated by the man speaker were recorded and dubbed according to digital technology. In a third experiment, the same materials as in the second experiment were used but the presentation procedure of the experimental items was changed: audiovisual incongruent trials were mixed up with congruent ones. In the three experiments, the role of voicing and of auditory intensity were investigated. Overall, combinations were much more numerous than fusions and both types of illusions tended to increase at low intensity. Voicing had a differential effect on both types of illusions. Combinations were more numerous with voiceless consonants but fusions tended to occur more often with voiced ones. The number of illusions was affected by the dubbing technique but not by the presentation procedure.

The role of sound intensity and stop-consonant voicing on McGurk fusions and combinations

Speech perception has long been considered as a mere auditory process. Auditory speech seems indeed sufficient in many communication situations when it is the only available source of information. However, in face to face communication, the speaker's lips and face provide useful information for speech perception. In 1954, Sumby and Pollack showed that the intelligibility of an auditory signal embedded in noise is considerably improved by lipreading. This increase in intelligibility may be attributed to the fact that visible information complements auditory information as regards, for example, the place of articulation, which is difficult to extract from the auditory signal alone (Miller & Nicely, 1955). Since then, numerous studies showed that visible speech is particularly useful when the auditory signal is degraded due to noise, bandwidth filtering or hearing impairment. For children suffering from hearing loss, lipreading is assumed to play an important role in spoken language acquisition (Dodd, McIntosh & Woodhouse, 1998).

The strong influence of visible speech is not limited to situations with degraded auditory inputs. It enhances the comprehension of clear speech signals that contain complicated content (Reisberg, 1987). It also gives useful information concerning the speaker's emotional state (Cerrato, Albano, Leoni & Falcone, 1998) and contributes to the acquisition of a second language (Davis & Kim, 1998). Finally, lipreading can influence speech perception when audition and vision provide incongruent signals. For example, when confronted with discrepant auditory and visual speech tokens, participants often report hearing a percept that does not correspond to the auditory information but integrates features from the visual input. This illusion, first reported by McGurk and MacDonald (1976), and generally referred to as the McGurk effect, indicates that the perceptual system makes use of the visual information even when the auditory signal is clear and unambiguous. Two different kinds of illusions have been observed: fusions and combinations. While visual presentation of a syllable containing a velar consonant like /ga/ together with an auditory syllable containing a bilabial consonant like /ba/ is likely to elicit a fused response /da/, the reverse presentation will normally give rise to a combination such as /bga/.

Proponents of the revised motor theory of speech (Liberman & Mattingly, 1985) have invoked the McGurk effect in order to support their view that speech perception is special. According to the revised motor theory, the object of speech perception is not the acoustic signal but the intended phonetic gestures of the speaker. Under this theory, a specialized perceptual Fodorian module performs the automatic conversion of received acoustic signals into intended articulatory gestures, which lead to perceived phonetic structures. The theory assumes that the module is engaged whenever an acoustic stimulus can be interpreted as linguistic. According to Mattingly and Liberman (1988), the McGurk percept is unambiguously phonetic, not bimodal, and this brings support to the assumption that speech perception is neither auditory nor visual, but results from the early conversion of received visual and acoustic information into intended articulatory gestures.

The Fuzzy Logical Model of Perception (Massaro, 1987) provides a different view about the McGurk effect. This model assumes that speech is not associated with a particular processing mode, like in the Motor Theory. On the contrary, general auditory processes are thought to be sufficient to explain the perception of speech. In case of an audiovisual conflict, both input modalities are independently evaluated and the illusory McGurk percepts result from a rather late integration process.

The McGurk effect has been replicated many times since McGurk and MacDonald's first study and it has been described as a mandatory and compulsory phenomenon. Indeed, when submitted to incongruent audiovisual stimuli, listeners do not experience the same perception when their eyes are open or shut, and even if they are fully aware of the dubbing procedure, they cannot immunize their percept from the visual influence.

Most studies have been conducted in English (Green & Gerdeman, 1995; Green, Kuhl, Meltzoff & Stevens, 1991; MacDonald & McGurk, 1978; Massaro, 1987; McGurk & MacDonald, 1976; Munhall, Gribble, Sacco & Ward, 1996; Rosenblum, Schmuckler & Johnson, 1997; Walker, Bruce & O'Malley, 1995), some others in Dutch (Bertelson, Vroomen, Wiegeraad & de Gelder, 1994), Finnish (Sams, Surakka, Helin & Kättö, 1997), German (Diesch, 1995), Japanese (Sekiyama & Tohkura, 1991; Sekiyama & Tohkura, 1993) or Chinese (Sekiyama, 1997). In French, the McGurk illusion does not appear to have been systematically investigated. To our knowledge, one of the rare

studies in which French was considered is that conducted by Werker, Frost and McGurk (1992). However, the aim of the study was to assess the effect of the degree of English linguistic experience on the illusion. For that purpose, the audiovisual syllables were recorded from a native Canadian speaker of English, and the participants were one group of English speakers and five groups of Canadian-French speakers varying in their fluency in spoken English. The materials consisted in an auditory /ba/ dubbed onto a visual /ba/, /va/, /da/, /ga/, /Za/ or onto the interdental stimulus /Da/ which occurs in English, but not in French. There was a relationship between the level of proficiency in English and the proportion of visual captures for the interdental stimulus. In response to the interdental visual stimulus, French speakers beginning in English assimilated this interdental viseme to the viseme with the closest place of articulation that has a phonemic status in French (/d/). The authors concluded that a high degree of linguistic experience with a language helps using and integrating lipreading in the perception of this language. Therefore, it appears that little is known about the McGurk effect with native French participants submitted to stimuli uttered by native speakers of French.

Some data suggest that the McGurk effect could be stronger when the auditory stimuli are made less salient, by intensity reduction or noise addition. For example, in Japanese, where the effect proved difficult to obtain, the addition of auditory noise enabled Sekiyama and Tohkura (1991) to get a much stronger effect. For native speakers of English, tested in English, Hardison (1996) found a similar result in one of her experiments. In a recent study, also conducted in English, Fixmer and Hawkins (1998) showed that the number of McGurk responses increased with auditory noise and decreased with visual noise. Kuhl and Green (1988), however, reported an increase in the number of illusory responses as the level of the auditory signal also increased (from 45 dB to 58 dB and 66 dB).

On the other hand, the influence of voicing on the size of the McGurk effect does not seem to have been investigated yet. Most studies were performed with voiced consonants, except Hardison (1996) who used voiceless stops and fricatives. Only some studies used both types of consonant but did not compare them explicitly. It is impossible to deduce from Diesch (1995) or from Sekiyama (1997) which type of consonant elicited the most illusions. MacDonald and McGurk (1978) found more fusions with voiced consonants, but for combinations, there was apparently no difference. Sekiyama and

Tokhura (1991) also used both kinds of consonant and found globally more illusions with voiceless consonants.

The goal of the first experiment was thus to study the role of a sound intensity manipulation on the McGurk illusions, using both voiced and voiceless stop consonants. Because the McGurk effect could be modulated by acoustic parameters (e.g. voice pitch) or visual parameters (e.g. size of the mouth opening), the materials were pronounced by two speakers: a woman and a man. In a second experiment, we assessed the generality of auditory intensity and voicing effects with another dubbing procedure. Finally, in a third experiment, the role of the same variables was again investigated, using the same dubbing procedure as in Experiment 2, but a somewhat different presentation procedure of the stimuli.

Experiment 1

Method

Participants.

Fifty-six students (33 women and 23 men; mean age: 21 years; range: 18-25 years) from the Free University of Brussels participated in the experiment as part of an introductory psychology course. They were all native speakers of French, without reported history of hearing disorder and with normal or corrected-to-normal vision. Thirty-six participants were exposed to the stimuli played at 70 dB and the other twenty were exposed to a sound intensity level of 40 dB.

Materials.

The same materials were used for each intensity condition. They consisted of four CV monosyllables (bi, gi, pi, ki) articulated by a woman and by a man wearing no beard who were native speakers of French.

The speakers were filmed on a turquoise background and spoke in a boom microphone (Senheiser MKH416). A color video camera (Sony A537p) and a video recorder (Panasonic AG-5700) were used for the recording. Only the lower part of the speaker's face (from the chin to the top

of the nose) was filmed in order to draw the participants' attention on the speaker's lips. Each item was pronounced three times in succession.

Editing and dubbing of the recorded stimuli were performed using two video recorders (Panasonic AG-5700) and a Panasonic AG-A770 editing desk. All experimental items were incongruent. They were created by replacing the original audio signal of a given item by the audio signal corresponding to another item that, among the three utterances, had about the same length in frames. The onset of the sound was synchronized with mouth opening. The acoustic tracks of the /bi/ and /gi/ and of the /pi/ and /ki/ items were exchanged in order to provide four incongruent audiovisual stimuli. 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. A trial thus lasted approximately two to three seconds.

In order to test the intelligibility of the auditory stimuli, control items were constructed. They consisted of the same auditory syllables as the experimental items but were recorded on a still face.

Four tapes were edited, one for each speaker and one for each kind of trials (experimental vs. control). Each tape consisted of 12 trials, four different stimuli being presented three times in random order. The auditory stimuli were low-pass filtered with a -15 dB cutoff at 10-15 kHz.

Procedure

The participants seated in front of a table, at 75 cm from a Panasonic color screen (width: 33 cm; height: 25 cm). A Trust loudspeaker was located on the top of the screen. The stimuli were played at an average level of 70 dB (which was 40 dB above the room background noise), measured on a sound-level meter (Brüel & Kjaer 2204 - A scale - fast) placed at the same distance and height as the participants' head. The students had to choose between several written answers the one corresponding to what they had heard. The possible answers were: /b/, /g/, /p/, /k/, /t/, /d/, /bg/, /pk/ and /other/ (the participants were instructed to choose this latter category as little as possible). These answers were selected on the basis of the most frequent answers found in a pilot study in which the participants had to write down what they had heard. The interstimulus interval (ISI) lasted 3 s. It consisted of a black screen period during which the participants had to give their response. The session began by an eight

trials training block. For each auditory intensity, two groups of participants were presented the two tapes in a different order. The entire session lasted about ten minutes.

Results and discussion

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. We considered as illusory responses, the audiovisual responses as well as the visual responses for which there was no voicing confusion. Indeed, because voicing is mainly conveyed by the auditory modality, visual responses without voicing confusion can not be explained in terms of pure lipreading but necessarily involve the integration of the auditory information. Let us note that, for more than 90% of the visual responses, there was no voicing confusion (e.g., /b/ was almost never confused with /p/ and so was it for /g/ and /k/). In several other studies, although voicing was not taken into account, visual responses were also included in the analyses because they were considered as particularly strong manifestations of the McGurk effect (Diesch, 1995; Rosenblum et al., 1997; Sams, Aulanko, Hämäläinen, Hari, Lounasmaa, Lu & Simola, 1991).

Illusions occurring for audiovisual stimulations the auditory part of which had been misidentified in the control condition were subtracted from the total number of illusory responses for each participant. For example, if a participant made two errors in identifying the auditory /bi/ of a control trial, these two errors were subtracted from the total number of illusory responses that occurred for the auditory /bi/, visual /gi/ experimental trial. In the auditory control condition, the confusions involving bilabial consonants appeared in only 1% of the cases and consisted mostly in a confusion between /b/ and /g/ or /p/ and /k/. Most auditory confusions (21%) however concerned velar consonants that were erroneously considered as clusters (/bg/ response instead of /g/).

Illusions percentages are displayed in Table 1 as a function of speaker gender, auditory intensity and type of pairing. The main results are that the percentage of fusions was much lower than that of combinations and that the percentage of illusions increased from 70 dB to 40 dB.

INSERT TABLE 1 ABOUT HERE

For each intensity, a variance analysis (ANOVA) was conducted on the percentages of illusions as dependent variable. Speaker gender, Type of pairing (fusion-type vs. combination-type) and Type of consonant (voiced vs. voiceless) were the within-participants variables. Presentation order of the two conditions (auditory first vs. audiovisual first) was a between-participants variable. Because this factor never reached significance or interacted with any other factor, it will no longer be mentioned.

For the 70 dB intensity, the main effect of Type of pairing was significant ($F(1,35)=93.55$, $p<.0001$), combinations being 41% more numerous than fusions. Speaker gender was also significant ($F(1,35)=15.22$, $p<.001$), the woman speaker giving rise to 14% illusions more than the man speaker, and it interacted with Type of pairing ($F(1,35)=17.16$, $p<.001$). There were more combinations for the woman speaker than for the man speaker ($F(1,35)=16.32$, $p<.001$), but no difference between both speakers for fusions ($F(1,35)=2.05$, $p<.20$). The main factor of Type of consonant was significant ($F(1,35)=15.40$, $p<.001$), voiceless consonants eliciting 13% illusions more than voiced ones. This factor interacted with Type of pairing ($F(1,35)=14.78$, $p<.001$), the main effect of Type of consonant being due to combinations ($F(1,35)=15.24$, $p<.001$). Fusions were indeed close to 0% for both types of consonant ($F<1$). The double interaction between Speaker gender and Type of consonant was not significant ($F(1,35)=1.79$, $p<.20$), nor was the triple interaction with Type of pairing ($F(1,19)=1.23$, $p<.30$).

For the 40 dB auditory intensity, Type of pairing was significant ($F(1,19)=44.65$, $p<.0001$), combinations being 34% more numerous than fusions. Type of consonant was also significant ($F(1,19)=31.87$, $p<.001$), voiceless consonants giving rise to 21% illusions more than voiced ones. Those two factors interacted with each other ($F(1,19)=33.02$, $p<.0001$) showing again that voiceless consonants produced more combinations than voiced ones ($F(1,19)=34.20$, $p<.0001$), whereas for fusions there were no difference between voiced and voiceless consonants ($F<1$). Speaker gender was not significant ($F<1$) but interacted with Type of pairing ($F(1,19)=8.70$, $p<.01$) and with Type of consonant ($F(1,19)=6.97$, $p<.05$). Fusions were more numerous with the man speaker than with the

woman speaker ($F(1,19)=14.54$, $p<.01$), whereas there were no difference for combinations ($F(1,19)=1.25$, $p<.30$). For the woman speaker, there were more illusions for voiceless than for voiced consonants ($F(1,19)=22.50$, $p<.001$) but there was no difference for the man speaker ($F(1,19)=1.86$, $p<.20$). The triple interaction did not reach significance ($F(1,19)=1.01$, $p<.40$).

On the whole, illusions were 6% more numerous at 40 dB than at 70 dB. In order to investigate if this difference was statistically significant, we conducted an ANOVA with Speaker gender, Type of pairing and Type of consonant as within-participants variables and with Auditory intensity (70 dB vs. 40 dB) as between-participants variable. Auditory intensity did not reach significance ($F(1,54)=2.57$, $p<.20$) and it only interacted with Speaker gender ($F(1,54)=8.86$, $p<.01$). There were more illusions at 40 dB than at 70 dB for the man speaker ($F(1,54)=9.37$, $p<.01$) but no difference between both intensities for the woman speaker ($F<1$).

The asymmetry between fusions and combinations was an unexpected aspect of the results. On average for the two intensity conditions, whereas the visual presentation of a bilabial stop consonant together with the auditory presentation of a velar stop consonant gave rise to a combination in 43% of the cases, the reverse presentation elicited a fusion in only 5% of the trials. This asymmetry was found for both speakers and for both voiced and voiceless consonants. Because the dubbing procedure used here was realized on analogical materials, one cannot rule out the possibility that the synchronization between auditory and visual signals was not always accurate. In order to control for this possible bias, we conducted another experiment using the same syllables but recorded and dubbed with digital technology. Only the man, for whom there was a substantial number of fusions, was chosen as speaker.

Experiment 2

Method

Participants.

Thirty-two students (23 women and 9 men; mean age: 20 years; range: 17-40 years) from the Free University of Brussels and selected according to the same criteria as in Experiment 1 participated in the experiment as part of an introductory psychology course. Half of the participants were exposed to the stimuli played at 70 dB and the other half were exposed to a sound intensity level of 40 dB.

Materials.

The syllables were the same as than in Experiment 1 (/bi/, /gi/, /pi/ and /ki/). They were uttered by the man speaker and recorded on a high speed digital camera Kodak EktaPro 1000 Imager, configured in order to capture gray scale images of 240 x 192 pixels at a rate 125 images per second. Again, only the lower part of the speaker's face was filmed. Each item was pronounced eight times in succession. The room was lit up with Arrilite 800 lighting.

Audio signals and trigger pulses corresponding to image acquisition were simultaneously recorded on a Digital Audio Tape. The trigger pulses were used to post-synchronize the audio signal with the video data. For each utterance, a movie was built so that (1) the total length of the movie was one second and (2) the burst of the stop consonant started halfway of the movie. For both couples of syllables (/bi/ and /gi/, and /pi/ and /ki/), we selected those that had the closest vowel duration in order to best match auditory and visual syllables. The acoustic tracks of the /bi/ and /gi/ and of the /pi/ and /ki/ movies were then respectively exchanged in order to provide four incongruent audiovisual stimuli.

Control items were still constructed. They consisted of the same auditory syllables as the experimental items but were paired with a still face.

Audiovisual incongruent items and control items were presented in two different sets of trials. Each set consisted of two 24 trial blocks. Each kind of stimulus was thus presented 12 times, in random order.

Procedure.

The stimuli were presented with iShell (ver. 1.2) software[1]. The participants seated in front of a table, at 75 cm from a 17' computer screen (Highscreen MS 1795p). A PC loudspeaker (Bass Booster AS 200) was located on the top of the screen. The stimuli were played at an average level of 70 dB or 40 dB, depending on the condition. The responses were given in the same way as in

Experiment 1 during a 3 s. ISI of black screen. The session began by an eight trials training block. For each auditory intensity, two groups of participants were presented the two sets of items (auditory alone or audiovisual) in a different order. The entire session lasted about 15 minutes.

Results and discussion

As in Experiment 1, we considered as illusory responses, the audiovisual responses as well as the visual responses without voicing confusion and subtracted, from this total, illusory responses the auditory part of which had been misidentified in the control condition. Similarly to the first experiment, the confusions involving bilabial consonants appeared in only 1% of the cases and consisted mostly in a confusion between /b/ and /g/ or /p/ and /k/. Most auditory confusions (22%) concerned velar consonants that were again erroneously considered as clusters (/bg/ response instead of /g/).

Illusions percentages are displayed in Table 2 as a function of auditory intensity and type of pairing.

INSERT TABLE 2 ABOUT HERE

Two separate ANOVAS were conducted on percentages of illusions as dependent variable in the two intensity conditions. Type of pairing (fusion-type vs. combination-type) and Type of consonant (voiced vs. voiceless) were the within-participants variables. Presentation order of the two sets, which was again a between-participants variable, did not reached significance and did not interact with another factor.

For the 70 dB intensity, the main effect of Type of pairing was significant ($F(1,15)=17.32$, $p<.001$). There were 46% combinations more than fusions. Type of consonant was marginally significant ($F(1,15)=4.48$, $p<.06$), voiceless consonants producing 9% illusions more than voiced ones. The double interaction did not reach significance ($F(1,15)=2.53$, $p<.20$).

For the 40 dB intensity, Type of pairing was again significant ($F(1,15)=39.49$, $p<.0001$), combinations being 50% more numerous than fusions. Type of consonant was not significant ($F(1,15)=1.47$, $p<.30$) but interacted with Type of pairing ($F(1,15)=14.77$, $p<.01$). Voiceless consonants elicited more combinations than voiced ones ($F(1,15)=16.35$, $p<.01$), whereas for fusions, there was no difference between both types of consonant ($F(1,15)=1.94$, $p<.20$).

A third ANOVA was run with Auditory intensity as between-participants variable and Type of pairing and Type of consonant as within-participants variables. Auditory intensity was significant ($F(1,30)=7.73$, $p<.05$), illusions being 18% more numerous at 40 dB than at 70 dB, but did not interact with any other factor.

On average, and considering only the man speaker, illusions percentages increased from 15% between Experiments 1 and 2. T-tests conducted, separately for each auditory intensity, on the total number of illusions, showed that this difference was significant (for 70 dB: $t(50) = 2.97$, $p<.01$; for 40 dB: $t(34) = 2.12$, $p<.05$). The two experiments only differed in the kind of technology used for the audiovisual synchronization (analogical for Experiment 1 and digital for Experiment 2). There are data arguing that the McGurk effect is unaffected by temporal desynchronizations reaching about 200 msec., at least when sound is lagging the visual signal (Massaro & Cohen, 1993; Massaro, Cohen & Smeele, 1996; Munhall et al., 1996). The synchronization procedure used with the analogical technology might on the contrary have produced an advance of the auditory signal. Indeed, in Experiment 1, the beginning of the auditory track (that is, the acoustical burst) was synchronized on mouth opening, whereas in natural speech situations, mouth opening always precedes the acoustical burst. Such a desynchronization (advance of the auditory signal) has been shown to have a strong detrimental effect on the McGurk illusions (Bertelson, Vroomen & de Gelder, 1997; Munhall et al., 1996). Consequently, the accuracy of the synchronization procedure enabled by the digital technology used in Experiment 2 might be responsible for the 15% illusions increase found in this Experiment.

Apart from the illusions percentages, results of Experiment 2 were fairly similar to those of Experiment 1, the number of illusions increasing as sound intensity decreased. Combinations were again more numerous with voiceless than with voiced consonants, whereas for fusions a tendency to the

reverse pattern was observed. Finally, the asymmetry between combinations and fusions was replicated. On average for the two intensity conditions, there were 61% of combinations but only 13% of fusions.

A particularity of the present procedure compared to that used in most McGurk studies (Diesch, 1995; Green & Gerdeman, 1995; Green, et al., 1991; Hardison, 1996; MacDonald & McGurk, 1978; Sekiyama and Tokhura, 1991; Sekiyama and Tokhura, 1993; Walker et al., 1995; Werker et al., 1992), was that there were no congruent trials. Although there is no obvious reason for the inclusion of such trials to influence the number of illusions, a third experiment was performed in which congruent trials were interspersed in incongruent ones. In order to assess the generality of the results, the same variables as in Experiments 1 and 2 (sound intensity and voicing) were again manipulated.

Experiment 3

Method

Participants.

Thirty-two students (27 women and 5 men; mean age: 19,3 years; range: 17-30 year) selected in the same way as those of the first and second experiments participated in this study. Again, half of the participants were exposed to the stimuli played at 70 dB and the other half were exposed to a 40 dB intensity.

Materials and procedure.

The materials were the same as in the second experiment, but four congruent audiovisual syllables were also edited (/bi/, /gi/, /pi/ and /ki/). The synchronization procedure was the same as in Experiment 2.

Audiovisual items and control items were again presented in two different sets. The set of control items was the same as in the second experiment. It consisted in one block of 24 trials (4 kinds of trials repeated 6 times each and presented in random order). The set of audiovisual items was made of 4 blocks of 24 trials in which audiovisual congruent and incongruent trials were mixed up.

The presentation procedure was similar to that of the second experiment. Again, for each condition of intensity, two groups of participants were presented the control and experimental sets in a different order. The entire session lasted about 20 minutes.

Results and discussion

The results were analysed in the same way as in Experiments 1 and 2 [illusory responses = (audiovisual + visual responses) – errors in control trials]. Errors on control trials reached 15% for velars and only 1% for bilabials. They consisted in the same kind of confusions as in Experiments 1 and 2. Again, the presentation order of the two sets was never significant and did not interact with another factor.

INSERT TABLE 3 ABOUT HERE

Illusions percentages are presented in Table 3. ANOVAS were conducted on the data of each intensity condition.

For the 70 dB intensity, combinations were much more numerous than fusions (difference = 46%: $F(1,15)=24.16$, $p<.001$). Type of consonant was not significant ($F(1,15)=3.24$, $p<.10$) and did not interact with Type of pairing ($F(1,15)=1.15$, $p<.30$).

For the 40 dB intensity, the asymmetry between combinations and fusions was still present (difference = 50%: $F(1,15)=30.83$, $p<.0001$). Type of consonant was not significant ($F<1$) but interacted with Type of pairing ($F(1,15)=7.28$, $p<.05$). Voiceless consonants produced more combinations than voiced ones ($F(1,15)=5.53$, $p<.05$), whereas there was no difference between both types of consonant for fusions ($F(1,15)=1.54$, $p<.30$).

Illusions were 12% more numerous at 40 dB, but the difference between both intensities was not significant ($F(1,30)=2.08$, $p<.20$) and no interaction relative to Auditory intensity reached significance.

On the whole, illusions were 10% more numerous in Experiment 3 than in Experiment 2. T-tests performed on the total number of illusions for each auditory condition however showed that this difference was not significant (for 70 dB: $t(30) = 1.61$, $p<.20$; for 40 dB: $t(30) = 0.79$, $p<.50$). Interspersing congruent items among the experimental incongruent items had thus no influence on the number of illusions.

Conclusions

The number of illusions significantly increased from Experiment 1 to Experiment 2, which differed by the dubbing procedure, but not from Experiment 2 to Experiment 3, which involved different presentation procedures of the experimental items. In addition, the pattern of results was fairly congruent across the three experiments. On the one hand, the percentages of illusions increased from 70 dB to 40 dB. On the other hand, combinations were much more numerous than fusions.

Moreover, there was about the same number of errors involving control trials across the three experiments: about 1% for bilabial consonants and 20% for velar ones. Such a high error rate for velars may be related to the kind of control items we used: a still face with a closed mouth. Indeed, the closed mouth might be considered by the participants as a cue for a bilabial place of articulation, leading thus to the perception of a cluster (a visual bilabial together with an auditory velar). In order to assess this possibility, a control experiment, in which the auditory syllables were presented without any visual signal, was performed on 16 new participants with the same digital materials as in Experiments 2 and 3. An asymmetric pattern of auditory confusions similar to that found in the three experiments was again observed. The number of confusions was 0% for bilabial consonants but reached 16% for velar consonants and still consisted in cluster responses. Such unexpected results should deserve further investigations.

The increase in the number of illusions as sound intensity decreased is in line with the results found by Sekiyama and Tokhura (1991), Hardison (1996) and Fixmer and Hawkins (1998) in which a much stronger McGurk effect was found with added auditory noise than in normal listening conditions.

Although in many studies, the listening conditions are not mentioned (Diesch, 1995; Easton & Basala, 1982; McGurk & MacDonald, 1976; Munhall et al., 1996; Walker et al., 1995; Werker et al., 1992; …), on the whole, the McGurk illusions are stronger when the visual stimuli are of good quality (Fixmer & Hawkins, 1998) and when the listening conditions are degraded. Only Kuhl and Green (1988) reported an increase in the number of illusory responses as the level of the auditory signal increased (45 dB, 58 dB and 66 dB), and they acknowledged that this was a curious result.

It could be argued that in degraded listening conditions, the visual influence did not reflect a true perceptual effect but was enhanced due to intentional lipreading by listeners who are unable to use the auditory modality. However, in the present case, care was taken for the so-called visual responses to be real audiovisual interactions. As explained in the Results section of Experiment 1, visual responses with voicing misidentification (less than 10%) were dropped from the analyses.

As regards the weak number of fusions found here relative to combinations, one can probably rule out a possible influence of the dubbing procedure (analogical vs. digital technologies) and of the presentation procedure of the stimuli (incongruent audiovisual trials alone or incongruent trials mixed up with congruent ones). Furthermore, such results are probably not related to the speaker's articulatory characteristics either, since in Experiment 1, a man and a woman acted as speakers and gave rise to comparable data. However, one cannot exclude that the McGurk effect be modulated by the phonetic properties of the language under study. Further studies involving cross-languages comparisons are needed to test such a possibility.

In addition to the quantitative asymmetry between the two kinds of illusions, we also found a differential effect of the type of consonant on the illusions. Voiceless consonants produced more combinations than voiced ones. For fusions, the consonant effect never reached significance but there was a tendency toward the reverse pattern. To our knowledge, such a differential effect has not been reported yet. However, it should be noted that most studies used only voiced consonants, except MacDonald and McGurk (1978) who used both voiced and voiceless consonants. For fusions, they observed the same pattern as in our study, but for combinations, there was no difference between both

kinds of consonant. Sekiyama and Tokhura (1991) also used both kinds of consonant and found globally more illusions with voiceless consonants.

The notion of acoustic confusability in noise might explain the tendency for a higher percentage of fusions found here with voiced consonants but not the reverse effect obtained for combinations. According to Summerfield (1987), the audiovisual percept emerging in the McGurk effect is a consonant that would be easily confused in noise with the presented auditory consonant and that is also fairly compatible with the presented visual consonant. The data on acoustic confusions reported by this author provided some support to his hypothesis. For example, an auditory /b/ is more easily misidentified in noise than an auditory /p/. Furthermore, /b/ is more often confused with /d/ than is /p/ with /t/. Consequently, an auditory /b/ is more likely to give rise to fusions than an auditory /p/. However, it seems difficult to apply such an explanation to combinations.

Another kind of explanation that allows a better account of the differential voicing effect found here for the two types of illusions is based on the same general principle of perceptual salience underlying the occurrence of either combinations or fusions. The obtaining of one or the other kind of illusion seems to depend on the more or less important perceptual weight of the acoustic and visual information. Visually, salience is greater for bilabial than for velar consonants. Phonetically, however, salience increases as the point of occlusion moves back in the mouth, the energy of the burst being more important for velar than for bilabial consonants (Dorman, Studdert-Kennedy & Raphael, 1977). For combinations, where listeners report hearing both the visible and acoustical consonants, the visual information of a bilabial /b/ or /p/ is so salient that it cannot be ignored, and so is it for the phonetic information provided by /g/ or /k/. For fusions, both the visual information from velar consonants and the acoustic cues from labial consonants are more ambiguous, thus leading to a percept intermediate between heard and seen evidence.

The superiority of voiceless over voiced stop consonants in eliciting combinations may result from the greater perceptual weight of the burst for voiceless than for voiced consonants. Burst is an important cue not only for place of articulation but also for voicing in the perception of stop consonants: it is more intense for voiceless than for voiced consonants (Calliope, 1989). The burst being more

intense for /k/ than for /g/, /k/ is auditorily more salient than /g/ and gives rise to more combinations. For fusions, since the burst is less intense for voiced consonants, these latter are less salient and more likely to be "attracted" by the visual modality and to produce an illusion such as /d/. Actually, we found a small tendency to have more illusions with voiced consonants but it was not significant. Further research is needed to determine the origin of the asymmetry between fusions and combinations and to re-examine the role of voicing on these effects.

Acknowledgements

This research has been supported by the "Communauté Française de Belgique" in the framework of the A.R.C. (96/01-203) and F.R.F.C. (8.4519.96) to Monique Radeau and of the A.R.C. (98/02-226) to Didier Demolin as well as by the “Fonds Emile Defay” of the U.L.B. to Paul Deltenre. We thank W. Serniclaes for his fruitful advice. We are grateful to M. Cluytens for her help in collecting some data and to F. Goossens for technical help.

References

Bertelson, P., Vroomen, J., & de Gelder, B. (1997). Auditory-visual interaction in voice localization and in bimodal speech recognition: The effects of desynchronisation. Proceedings of the Auditory-Visual Speech Processing Conference, Rhodes, Greece, 97-100.

Bertelson, P., Vroomen, J., Wiegeraad, G., & de Gelder, B. (1994). Exploring the relation between McGurk interference and ventriloquism. Proceedings of the International Conference on Spoken Language Processing, Yokohama, Japan, 559-562.

Calliope (1989). La parole et son traitement automatique. Paris : Masson.

Cerrato, L., Albano Leoni, F., & Falcone, M. (1998). Is it possible to evaluate the contribution of visual information to the process of speech comprehension ? Proceedings of the Auditory-Visual Speech Processing Conference, Terrigal, Australia, 141-146.

Davis, C., & Kim, J. (1998). Repeating and remembering foreign language words: Does seeing help ? Proceedings of the Auditory-Visual Speech Processing Conference, Terrigal, Australia, 121-126.

Diesch, E. (1995). Left and right hemifield advantages of fusions and combinations in audiovisual speech perception. Quarterly Journal of Experimental Psychology, 48A, 320-333.

Dodd, B., McIntosh, B., & Woodhouse, L. (1998). Early lipreading ability and speech and language development of heraing-impaired pre-schoolers. In R. Campbell, B. Dodd, & D. Burnham (Eds.), Hearing by eye II: The Psychology of Lip-Reading (pp. 229-242). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. (1977). Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. Perception and Psychophysics, 22, 109-122.

Easton, R.D., & Basala, M. (1982). Perceptual dominance during lipreading. Perception and Psychophysics, 32, 562-570.

Fixmer, E., & Hawkins, S. (1998). The influence of quality of information on the McGurk effect. Proceedings of the Auditory-Visual Speech Processing Conference, Terrigal, Australia, 27-32.

Green, K. P., & Gerdeman, A. (1995). Cross-modal discrepancies in coarticulation and the integration of speech information: The McGurk effect with mismatched vowels. Journal of Experimental Psychology: Human Perception and Performance, 21, 1409-1426.

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

Hardison, D. B. (1996). Bimodal perception by native and nonnative speakers of English: Factors influencing the McGurk effect. Language Learning, 46, 3-73.

Kuhl, P., & Green, K. P. (1988). Factors affecting the integration of auditory and visual information in speech: The level effect. Journal of the American Society of America, 83 (Suppl. 1), S86.

Liberman, A. M., & Mattingly, I. G. (1985). The motor-theory of speech revised. Cognition, 21, 1-36.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. Perception and Psychophysics, 24, 253-257.

Massaro, D.W. (1987). Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. Hillsdale, NJ: Lawrence Erlbaum Associates.

Massaro, D. W., & Cohen, M. M. (1993). Perceiving asynchronous bimodal speech in consonant-vowel and vowels syllables. Speech Communication, 13, 127-134.

Massaro, D. W., Cohen, M. M., & Smeele, P. M. T. (1996). Perception of asynchronous and conflicting visual and auditory speech. Journal of the Acoustical Sociey of America, 100, 1777-1786.

Mattingly, I. G., & Liberman, A. M. (1988). Specialized perceiving systems for speech and other biologically significant sounds. In G. M. Edelman, W. E. Gall, & W. N. Cowan (Eds.), Auditory Function: Neurobiology Bases of Hearing (pp. 775-793). New-York: Wiley.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264, 746-748.

Miller, G.A., & Nicely, P.E. (1955). An analysis of perceptual confusions among some English consonants. Journal of the Acoustical Society of America, 27, 338-352.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. Perception and Psychophysics, 58, 351-362.

Reisberg, D. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), Hearing by Eye: The Psychology of Lip-Reading (pp. 97-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Lounasmaa, O.V., Lu, S-T., & Simola, J. (1991). Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. Neuroscience Letters, 127, 141-145.

Sams, M., Surakka, V., Helin, P., & Kättö, R. (1997). Audiovisual fusion in finnish syllables and words. Proceedings of the Auditory-Visual Speech Processing Conference, Rhodes Greece, 101-104.

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. Perception and Psychophysics, 59, 73-80.

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. Journal of the Acoustical Society of America, 90, 1797-1805.

Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. Journal of Phonetics, 21, 427-444.

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

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), Hearing by Eye: The Psychology of Lip-Reading (pp. 3-51). Hillsdale: NJ: Lawrence Erlbaum Associates.

Walker, S., Bruce, V., & O'Malley, C. (1995). Facial identity and facial speech processing: familiar faces and voices in the McGurk effect. Perception and Psychophysics, 57, 1124-1133.

Werker, J. F., Frost, P. E., & McGurk, H. (1992). La langue et les lèvres: Cross-language influences on bimodal speech perception. Canadian Journal of Psychology, 46, 551-568.

Footnotes

[1] http://www.tribeworks.com

Table 1: Percentage of combinations (comb.) and fusions for voiced and voiceless consonants as a function of auditory intensity for each of the two speakers in Experiment 1. In this and subsequent tables, combinations and fusions arose exclusively from combination-type (auditory velar dubbed onto visual bilabial) and fusion-type (auditory bilabial dubbed onto visual velar) pairings, respectively.

	Man speaker				Woman speaker			
	70 dB		40 dB		70 dB		40 dB	
	Comb.	Fusions	Comb.	Fusions	Comb.	Fusions	Comb.	Fusions
b-g	10	0	23	27	47	0	15	0
p-k	44	1	55	12	64	0	83	0
Mea	27	1	39	20	56	0	49	0

Table 2: Percentage of combinations and fusions for voiced and voiceless consonants as a function of auditory intensity in Experiment 2.

	70 dB		40 dB	
	Combinations	Fusions	Combinations	Fusions
b-g	43	4	61	26
p-k	60	6	82	16
Mean	**51**	**5**	**71**	**21**

Table 3: Percentage of combinations and fusions for voiced and voiceless consonants as a function of auditory intensity in Experiment 3.

	70 dB		40 dB	
	Combinations	Fusions	Combinations	Fusions
b-g	59	17	71	32
p-k	68	19	85	25
Mean	**64**	**18**	**78**	**28**