Top-down and bottom-up modulation of audiovisual integration in speech

Colin, C. (1) (2), Radeau, M. (1) (2) and Deltenre, P. (1) (3)

(1) Research Unit in Cognitive Neurosciences CP 191, Université Libre de Bruxelles, 50 Roosevelt avenue, B-1050 Brussels, Belgium.

(2) F.N.R.S., Belgium.

(3) Brugmann Hospital, 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

This research assesses how audiovisual speech integration mechanisms are modulated by sensory and cognitive variables. For this purpose, the McGurk effect (McGurk & MacDonald, 1976) was used as an experimental paradigm. This effect occurs when participants are exposed to incongruent auditory and visual speech signals. For example, when an auditory /b/ is dubbed onto a visual /g/, listeners are led to perceive a fused phoneme like /d/. With the reverse presentation, they experience a combination such as /bg/. In two experiments, auditory intensity (40 dB, 50 dB, 60 dB, and 70 dB), face size (large : 19 * 23 cm and small: 1.8 * 2 cm) and instructions (“multiple choice” and “free response”) were manipulated. Face size and instruction were between-participants variables in both experiments, whereas intensity was a within-participants variable in the first experiment and a between-participants variable in the second one. The main effect of instruction manipulation was highly significant in both experiments, the “multiple choice” condition giving rise to more illusions than the “free response” condition. Intensity was significant in the second experiment only. Illusions were more numerous at 40 dB than at the other three intensities. Finally, a small effect of face size was observed in the second experiment only, illusions being slightly more numerous with the large face. Those results indicate that the processing chain underlying audiovisual speech perception is modulated by the perceptual salience of the visual and auditory inputs as well as by cognitive variables.

Although auditory speech seems sufficient in many communication situations in which it is the only available source of information, face-to-face communication enables the listener to extract information from the speaker's lips and face. The role of visible speech is particularly obvious in situations with degraded auditory inputs (see for example, Dodd, McIntosh & Woodhouse, 1998; Erber, 1969; MacLeod & Summerfield, 1990; Middleweerd & Plomp, 1987; Mohamadi & Benoît, 1992; Sumby & Pollack, 1954) but is also of importance when audition provides a perfectly clear signal (Benoît, Mohamadi & Kandel, 1994; Burnham, 1998; Cerrato, Leoni & Falcone, 1998; Davis & Kim, 1998; Reisberg, McLean & Goldfield, 1987). One of the best demonstrations of the spontaneous and essential nature of visible speech is provided by the McGurk effect (McGurk & MacDonald, 1976). 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. Two different types of illusions have been observed: fusions and combinations. Whereas visual presentation of a syllable containing a velar consonant like /gi/ together with auditory presentation of a syllable containing a bilabial consonant like /bi/ is likely to elicit a fused response /di/, the reverse presentation will normally give rise to a combination such as /bgi/ (see Colin & Radeau, 2003; Hardison, 1996; Massaro, 1987, 1998, for reviews on the McGurk effect).

Since its discovery, the McGurk effect has been extensively applied to the study of audiovisual speech perception mechanisms. Results of this research have been used to assess and improve theories of speech perception (see Green, 1998 for a review). The Fuzzy Logical Model of Perception (FLMP; Massaro, 1987, 1998), for example, assumes that, whatever its domain, sensory information is always processed by the same mechanisms that can be organized according to three stages (evaluation – integration – decision). In case of an audiovisual conflict (e.g. A/ba/ V/ga/), each input is first independently evaluated in terms of to what extent it supports prototypes stored in memory. Then, both signals are integrated, and a response (the illusory McGurk percept) is selected (in this case, the prototype /da/ receives more support than the prototypes /ba/ or /ga/).

Contrary to Massaro (1987, 1998), proponents of the Motor Theory of Speech (Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985) have invoked the McGurk effect to support their view that speech perception is special. Under this theory, speech perception is neither auditory nor visual, but unambiguously phonetic. Audiovisual percepts emerge from the early conversion of received visual and acoustic information into intended articulatory gestures. A specialized perceptual fodorian module, which is automatically engaged whenever the stimulus can be interpreted as linguistic, achieves this conversion.

In line with the Motor Theory of Speech, the McGurk effect has usually been considered to be a compulsory phenomenon. Even if participants are fully aware of the dubbing procedure, they cannot immunize their percepts from the visual influence (Repp, Manuel, Liberman & Studdert-Kennedy, 1983; Rosenblum & Saldaña, 1996; Summerfield & McGrath, 1984). Some data suggest, however, that the McGurk phenomenon can be modulated by cognitive manipulations.

McGurk (1988), for example, found that a given audiovisual presentation may give rise to different percepts as a function of the semantic context. It is important to note that such effects, that are well known to reflect expectancy-based strategies (Posner & Snyder, 1975), have not been replicated in better controlled conditions (Sams, Manninen, Surakka et al., 1998).

Some arguments in favour of a possible role played by cognitive factors were found by manipulating face/voice gender compatibility. Results are, however, not completely coherent. Easton and Basala (1982) found less McGurk illusions when the gender of the speaker's face did not correspond to that of the speaker's voice than when they were matching. Green, Kuhl, Meltzoff and Stevens (1991), on the contrary, did not find any such effect, whereas Walker, Bruce and O'Malley (1995) found a detrimental effect for familiar faces only.

Another example of top-down influence on the McGurk effect has been observed by Sekiyama and Tohkura (1993). They asked Japanese native speakers to identify incongruent audiovisual syllables and to report whether or not they perceived any discrepancy between the auditory and visual signals. Illusion percentages were lower when the discrepancy was detected,

probably because such detection allowed listeners to isolate the auditory signal from the visual information.

Other data regarding the role of cognitive factors were obtained by manipulating visual or auditory attention. Tiippana, Sams and Andersen (2001) manipulated visual attention and showed that McGurk illusions were stronger when participants were asked to pay attention to the speaker's face than when they were asked to attend a visual distractor displayed on the speaker's face. In contrast, illusions were smaller when participants had to turn their attention to the auditory modality in order to be able to discriminate consonants (Amano & Sekiyama, 1998). An effect of auditory attention was also reported by Summerfield and McGrath (1984) by instruction manipulation. Half of the participants were informed of the dubbing procedure and had to repeat what they had heard, whereas the other half were completely naive and had to repeat what the speaker had uttered. According to the authors, the latter instructions should draw less attention to the auditory signal. Because visual influence was actually less strong for the first group, Summerfield and McGrath concluded that attentional mechanisms were at play.

Some evidence for a role of instructions was also briefly reported by Massaro (1998, p184-188) who found that combinations were more frequently elicited in a limited set of alternatives than in a completely open-ended one. There is, to our knowledge, no other study in which the role of instructions has been systematically investigated. Instructions are usually very consistent in the literature. They often consist of an invitation to report what is heard and not, for example, what is perceived, which would probably draw less attention to the auditory modality. Regarding responses, requirements are more variable : participants are asked to utter their answers aloud (e.g. McGurk & MacDonald, 1976), to write it down (e.g. Sekiyama & Tohkura, 1991), or to choose a response among several possibilities, either by pressing a button (e.g. Jones & Munhall, 1997) or by making a selection among several written answers (e.g. Colin, Radeau, Deltenre et al., 2002). It is interesting to note that in many studies in which a free response was allowed, it is not generally mentioned whether some response examples were provided or not. This point may be considered as

minor. However, being told or not what the possible answers are may lead to adopt (explicitly or not) different response strategies.

In the present study, we investigated the role of instructions on the size of the McGurk fusions and combinations with a stronger manipulation than in Summerfield and McGrath's study (1984). The instruction effect found in the latter study was probably due to the fact that half the participants were informed of the dubbing procedure. The difference between "repeat what is heard" and "repeat was is uttered" seems too weak, indeed, to elicit the reported effect. We compared two kinds of instructions: a "free response" condition in which participants merely had to write down what they heard on a sheet of paper without receiving any example regarding the kind of syllables they would be likely to hear and a "multiple choice" condition in which they had to pick out their percept between several written possibilities. The multiple choice condition thus provided participants with extra clues about possible answers like the /bg/ or /pk/ combination percepts.

If audiovisual speech integration mechanisms are really cognitively impenetrable, instructions should have no effect on illusion percentages. However, whereas lack of effect would constitute an argument for cognitive impenetrability, obtaining an effect would not necessarily argue against this principle. Indeed, the locus of the instruction effect is not necessarily the integration level but it may be a later decisional stage which is affected by generating expectancies.

Sensory factors, particularly the salience of the auditory and visual signals, are also likely to modulate the McGurk fusions and combinations. For the auditory modality, some data suggest that the McGurk effect is stronger when auditory stimuli are made less salient, by intensity reduction, acoustic blurring or by reducing the speaker's auditory intelligibility. In a research carried out in Japanese, a language in which the effect proved difficult to obtain, Sekiyama (1998) found that fusions increased by 50% between auditorily clear speakers and less intelligible speakers. Also in Japanese, the addition of auditory noise enabled Sekiyama and Tohkura (1991) to get a McGurk effect about 50% stronger than in a condition without noise. For English native speakers, similar results were found by Hardison (1996), Jordan and Sergeant (1998) and by Fixmer and Hawkins

(1998). In this latter case, three levels of auditory noise were tested. The percentages of McGurk responses increased with the severity of noise. The relation between visual influence on auditory perception and the salience of the auditory stimuli has also been shown in the !Xóõ language, an African language making use of click consonants (Traill, 1999). Visual influence was more important when auditory clicks were of weak intensity than when they were of strong intensity. However, Kuhl, Green and Meltzoff (1988) reported, for English, 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). These data published as a short conference abstract were considered to be counterintuitive by the authors themselves.

Recently, Colin et al. (2002) conducted three experiments in which, among other variables, auditory intensity was manipulated. The stimuli were presented at 70 dB or at 40 dB. On the whole, the McGurk illusions were 12% more numerous at 40 dB than at 70 dB. However, having only two levels of intensity does not allow one to determine whether the increase of McGurk illusions between 70 dB and 40 dB was an abrupt or a progressive phenomenon. In order to investigate this point, the present research compared McGurk fusions and combinations at four different auditory levels : 40 dB, 50 dB, 60 dB, and 70 dB. On the basis of the previous results on auditory salience, we expected that the McGurk illusions would increase with the reduction of auditory intensity. We made the additional prediction that, like in Fixmer and Hawkins (1998), this increase would be gradual.

The size of the McGurk effect may also be influenced by the visual salience of the stimuli, which can, for example, be modulated by manipulating the speaker's visual intelligibility, blurring the display or varying facial image size. Visual intelligibility may vary from one speaker to another as a function of speech rate or articulatory gesture accuracy (e.g. Demorest & Bernstein, 1992). Nelson and Hodge (2000) showed that the identification of audiovisual syllables was more difficult with a speaker suffering from facial paralysis than with a "normal" speaker. This difficulty was especially prominent for phonemes for which visual modality was the most informative, such as

bilabial stop consonants. A speaker's visual intelligibility can also be reduced by violating the normal configural information of the face. For example, when the speaker's face was rotated (Jordan & Bevan, 1997) or inverted (Bertelson, Vroomen, Wiegeraad & de Gelder, 1994; Colin, Radeau, Deltenre & Morais 2001; Jordan & Bevan, 1997; Massaro & Cohen, 1996), the McGurk effect was reduced by about 20%. It was, however, not face inversion in itself that was responsible for diminishing the McGurk effect but rather the inversion of the lips, which would result in a loss of configural coherence between the articulatory movements and the auditory signal (Rosenblum, Yakel & Green, 2000). Hietanen, Manninen, Sams and Surakka (2001) confirmed that, provided lips are presented upright, a normal face configuration was not necessary for a McGurk illusion to occur. They found that only an asymmetrically scrambled face stimulus (both eyes, mouth, and noise being horizontally and vertically displaced) produced strong deterioration. In visual modality, noise can also be added to the signal. Fixmer and Hawkins (1998) showed that the number of McGurk responses decreased proportionally to visual noise addition. Similarly, MacDonald, Andersen and Bachmann (2000) observed that the McGurk effect gradually diminished with visual stimulus degradation by spatial quantization (a process consisting in reducing visual resolution by local averaging of pixels).

Manipulations of the visual signal that do not affect the dynamical or the configural aspects of visual stimulation only lead to minor changes in the size of the McGurk effect. Rosenblum and Saldaña (1996) showed that illusions also occurred when the visual signal consists of reflective dots, placed on various parts of the face (lips, teeth, tongue, and jaw) that reproduced the articulatory movements. Jordan and Thomas (2001) found that performing audiovisual stimuli exhibited considerable stability from full-face, three-quarter and profile views. The McGurk effect was also well preserved when the speaker's face was displayed up to 20 m from participants (Jordan & Sergeant, 2000). However, such a result is difficult to interpret because changes in distance also affect the strength of the auditory signal and produce changes in the relative time of arrival of both signals (the loudspeaker was positioned adjacent to the monitor and auditory intensity was thus

reduced as distance increased). Other data previously reported by Jordan and Sergeant (1998) more convincingly suggested that the visual information underlying audiovisual speech processing can still be encoded when facial images are very small. The effect of facial image size on audiovisual speech identification was investigated by presenting a talking face at a normal size (21 cm high) and at four reduced sizes (20%, 10%, 5%, and 2.5% of the full-size image), all viewed at a distance of 1 m. The McGurk illusions only decreased by about 20-25% for the 5% and the 2.5% reductions. In an attempt to re-examine Jordan and Sergeant's data, we investigated the effect of facial image size, using two very different face sizes (large: 19 * 23 cm and small: 1.8 * 2 cm) viewed from the same distance. We expected visual salience to affect the McGurk illusions but to a small extent. This manipulation does not indeed affect dynamical or configural properties of the visual input.

Therefore, in order to study the modulation of audiovisual speech integration mechanisms by sensory and cognitive factors, we carried out two experiments using the McGurk effect as an experimental paradigm. In each experiment, the two sets of instructions and the two face sizes were presented to different groups of participants. In the first experiment, all participants were submitted to the four auditory intensities, whereas in the second experiment, each participant was exposed to one single auditory intensity. Moreover, in both experiments, participants were all exposed to fusion-type and combination-type trials. Indeed, in previous experiments (Colin, 2001; Colin et al. 2001, 2002), we found that, in French, combination-type trials gave rise to far more illusions than fusion-type trials. We therefore wanted to re-examine the type of illusion factor.

Experiment 1

Method

Participants

Thirty-two subjects (23 women and 9 men; mean age: 21,5 years; range: 17-25 years) participated in the first experiment as paid volunteers. All of them were French native speakers, without any reported history of hearing disorders and with normal or corrected-to-normal vision.

Materials

The materials consisted of four CV monosyllables (bi, gi, pi, ki) articulated by a man wearing no beard who was a French native speaker. The syllables were recorded on a high speed digital camera (Sony DCR-TRV 20 E), configured in order to capture colour images of 768 x 576 pixels at a rate of 24 images per second. 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 attention of the participants to the speaker's lips. Each item was pronounced three times in succession. Sound was captured by a boom Philips SBC MD 140 microphone and the acoustic track was down-sampled to 16 kHz. For both couples of syllables (/bi/ and /gi/ on the one hand, /pi/ and /ki/ on the other hand), we selected those that had the nearest vowel duration in order to best match auditory and visual syllables. Auditory and visual tracks were then synchronized on the burst of the stop consonant measured on the acoustic signal in order to provide four incongruent audiovisual stimuli (A/bi/ V/gi/, A/gi/ V/bi/, A/pi/V /ki/, A/ki/ V/pi/) and four congruent audiovisual stimuli (A/bi/ V/bi/, A/gi/ V/gi/, A/pi/ V/pi/, A/ki/ V/ki/).

Procedure

Participants were seated at a table, 75 cm from a standard 17'' colour monitor. Stimuli were played back by means of the iShell software ver. 2.5 (http://www.tribeworks.com). In the large face condition, the size of the speaker's face was 382 pixels high * 478 pixels wide (19 * 23 cm) and in the small face condition, it was 36 pixels height * 45 pixels width (1.8 * 2 cm). See Figure 1 for a comparison of the two face sizes.

The acoustic part of the stimuli was delivered at an overall intensity of 40 dB, 50 dB, 60 dB, or 70 dB through a Trust loudspeaker positioned on the top of the screen. Sound pressure level was measured on a sound-level meter (Brüel & Kjaer 2231 - A scale) equipped with a Brüel & Kjaer

4190 microphone placed at the same distance and height as the participants' heads. The room background noise reached 30 dB.

In order to test the intelligibility of the auditory stimuli, each participant was exposed to an auditory control condition in addition to the audiovisual experimental condition. Auditory trials consisted of the same syllables as audiovisual trials but were presented with the screen switched off. In order to avoid possible contamination by the illusory phonemes perceived in the audiovisual incongruent trials, the experiment began with this auditory control condition.

The four auditory intensities were presented in separate blocks. Each participant heard the four auditory intensities in the same presentation order for each condition. This presentation order was, however, different for each participant within a group and followed a Latin square. In each condition and for each auditory intensity, each type of stimulus was presented 12 times. The 12 repetitions of four different types of trials of the auditory condition (/bi/, /gi/, /pi/, and /ki/) were presented randomly in two blocks of 24 trials that were repeated four times (once for each intensity). In the audiovisual condition, the 12 repetitions of the eight different trials (four congruent ones and four incongruent ones) were randomly presented in four blocks of 24 trials which were also repeated four times (once for each intensity). So, there was a total of 48 trials in the auditory condition and of 96 trials in the audiovisual condition.

The session began with an auditory training block. Each of the four syllables was repeated four times in a random order. The first four syllables of the training block were presented at 70 dB, the following four ones at 60 dB, and so on. The training block was followed by the auditory condition. The audiovisual condition took place at the end of the session.

Participants were instructed that they would hear meaningless syllables without being provided with any examples. In the "free response" condition, they were asked to write down the syllables they had heard. In the "multiple choice" condition, they were asked to choose between several written answers the one corresponding to what they had heard, the possible answers being : /b/, /g/, /p/, /k/, /t/, /d/, /bg/, /pk/, and "other". The listed answers corresponded to the actual auditory

and visual signals and to the expected illusions. They were selected on the basis of the most frequent answers found in a pilot experiment of a previous study (Colin et al., 2002) in which participants had to write down what they had heard. Choices were presented on a sheet of paper (one line for each trial) in Geneva font (size 14) in the same order for each participant. After the auditory condition, the participants were told that they would again hear meaningless syllables but would also see the face of the speaker uttering the syllables. Response instructions for both the "free response" and the "multiple choice" conditions were the same as in the auditory condition but with the additional instruction to look at the screen.

The interstimulus interval (ISI) was 3 seconds for all types of trials (auditory or audiovisual) and both sets of instructions. The presentation screen went black for the ISI period during which participants had to issue their responses. The entire session lasted about 50 minutes.

Half of the 32 participants were exposed to the large face condition and the other half to the small face condition. Within each group, half of the participants received the "multiple choice" instructions and the other half received the "free response" instructions. All participants of those four groups were submitted to the four intensity conditions (40 dB, 50 dB, 60 dB, and 70 dB).

Data Analysis

For each item, participants were expected to give answers corresponding to the auditory information, or to the visual information, or an audiovisual response. We considered the audiovisual responses as well as the visual responses for which there was no voicing confusion as illusory responses. This procedure, already used by Colin et al. (2002), rests on the fact that voicing is difficult to see in the face (e.g. Massaro, 1998). Therefore, visual responses without voicing confusion cannot probably be explained in terms of pure lipreading but should involve the integration of auditory information.

Errors in the auditory control condition were subtracted from the number of illusory responses in the audiovisual condition. For example, if a participant made two errors in identifying the auditory /bi/ on control trials, these two errors were subtracted from the number of illusory

responses that occurred for the auditory /bi/, visual /gi/ experimental trial. However, voicing confusions made in the control condition were not taken into account in the subtraction because, for obvious reasons, they were expected to be much more numerous at 40 dB than at 70 dB. Subtracting those responses from the total number of illusory responses would thus have introduced a bias when computing the McGurk effect.

A four way analysis of variance (ANOVA) was conducted on the percentages of illusions as a dependent variable. Instruction (two levels) and face size (two levels) were the between-participants variables, whereas auditory intensity (four levels) and type of illusion (two levels) were the within-participants variables. Interactions were tested using planned comparisons.

Results and Discussion

Results of the auditory condition are displayed in Table 1 as a function of articulation place (bilabial or velar), instructions ("multiple choice" or "free response"), auditory intensity (40 dB, 50 dB, 60 dB, or 70 dB) and error type ("voicing" or "other"). It can be noted that "other" errors were fairly low and that voicing confusions mainly occurred at 40 dB.

INSERT TABLE 1 ABOUT HERE

For audiovisual trials, percentages of fusions and combinations are displayed in Table 2 as a function of instructions, face size and auditory intensity.

INSERT TABLE 2 ABOUT HERE

The main effect of instruction was significant ($F(1,28)=8.05$, $p<.01$). Illusions were 28% more numerous in the "multiple choice" condition than in the "free response" condition. The type of illusion was also significant ($F(1,28)=8.71$, $p<.01$), combinations being 22% more numerous than fusions, and interacted with intensity ($F(3,84)=2.86$, $p<.05$). All intensities gave rise to similar

numbers of fusions (17% on average) but combinations were a bit more frequent at 40 dB (42%) or 50 dB (41%) than at 60 dB (36%) or 70 dB (35%; 40 dB vs. 60 dB: $F(3,84)=3.06$, $p<.05$; 40 dB vs. 70 dB: $F(3,84)=3.32$, $p<.05$; 50 vs. 60 dB: $F(3,84)=4.33$, $p<.05$; 50 vs. 70 dB: $F(3,84)=8.09$, $p<.01$). Intensity was not significant, nor was face size (both $F<1$). None of triple interactions reached significance (all F values were close to or smaller than 1). The quadruple interaction felt short of significance ($F(3,84)=2.41$, $p=.07$).

An unexpected aspect of the results was the lack of intensity effect. A possible explanation is that since the participants were exposed to the four auditory intensities in both auditory and audiovisual conditions, they had enough time to adopt their response strategies for each of the auditory and audiovisual stimuli and, consciously or not, kept those strategies throughout the whole experiment. In order to assess such a possibility, we carried out another experiment in which auditory intensity was a between-participants variable.

Experiment 2

The design of Experiment 2 was exactly the same as that of Experiment 1 except that each participant was exposed to only one auditory intensity (40 dB, 50 dB, 60 dB, or 70 dB) in an auditory control condition and in an audiovisual experimental condition. This procedure enabled shortening the experiment duration, possibly reducing over-learning of the requested responses.

Method

Participants

One hundred and twenty-eight university students (107 women and 21 men; mean age: 19.3 years; range: 17-35 years) participated in the experiment as part of an introductory Psychology course. They were different from those of Experiment 1 but selected according to the same criteria.

Materials and Procedure

Materials and procedure were exactly the same as in Experiment 1 except that each participant was exposed to a single auditory intensity. The 128 participants were thus divided in 16 groups of eight. Eight groups were exposed to the large face condition and the other eight to the small face one. Within each of these two groups of eight, four groups were submitted to the "multiple choice" condition and the other four to the "free response" condition. Finally, within each of these four groups of four, each group was exposed to one of the four intensity conditions. The entire session lasted about 15 minutes.

Data Analysis

The percentages of illusory responses were computed exactly in the same way as in Experiment 1. A four way ANOVA was conducted on the percentages of illusions as a dependent variable. Instruction (two levels), face size (two levels) and auditory intensity (four levels) were the between-participants variables, whereas type of illusion (two levels) was the within-participants variable. As in Experiment 1, interactions were tested using planned comparisons.

Results and Discussion

Results of the auditory control condition were fairly comparable to those of Experiment 1 (see Table 3).

INSERT TABLE 3 ABOUT HERE

Percentages of fusions and combinations are displayed in Table 4 as a function of instructions, face size and auditory intensity.

INSERT TABLE 4 ABOUT HERE

The main effect of instruction was significant ($F(1,112)=31.43$, $p<.0001$). Illusions were 21% more numerous in the "multiple choice" condition than in the "free response" condition. This

factor interacted with intensity ($F(3,112)=2.73$, $p<.05$). At 40 dB ($F(1,112)=22.10$, $p<.0001$) and at 60 dB ($F(1,112)=13.49$, $p<.001$), illusions were more frequent in the multiple choice condition. The instruction factor also interacted with the type of illusion ($F(1,112)=12.09$, $p<.001$). Combinations were more frequent than fusions in the multiple choice condition only ($F(1,112)=36.59$, $p<.0001$). Intensity was also significant ($F(3,112)=13.71$, $p<.0001$), the 40 dB intensity giving rise to significantly more illusions than the other three intensities (40 dB vs. 50 dB: 28% difference, $F(1,112)=28.29$, $p<.0001$; 40 dB vs. 60 dB: 26% difference, $F(1,112)=24.37$, $p<.0001$; 40 dB vs. 70 dB: 28% difference, $F(1,112)=29.21$, $p<.0001$). Face size was significant as well ($F(1,112)=4.08$, $p<.05$). The large face condition gave rise to 8% illusions more than the small face. Finally, the type of illusion was also significant ($F(1,112)=25.77$, $p<.0001$), combinations being 16% more numerous than fusions. Most of the other interactions gave rise to F values close to, or smaller than, 1 except the three following ones which were, however, not significant : face size x type of illusion ($F(1,112)=1.95$, $p=.16$), intensity x type of illusion ($F(1,112)=1.53$, $p=.20$) and the quadruple interaction ($F(3,84)=2.57$, $p=.06$).

In summary, we replicated the asymmetry between combinations and fusions as well as the instruction effect found in Experiment 1. However, this time, we obtained a small face size effect (the large face giving rise to more illusions) and a strong intensity effect, illusions being much more numerous at 40 dB than at all other intensities.

General discussion

The most striking aspect of the results was the massive instruction effect; illusions were much more numerous in the “multiple choice” condition than in the “free response” condition. In the first experiment, the effect reached 29% and occurred for both types of illusions, both face sizes and all auditory intensities. In the second experiment, the instruction effect reached 21% and occurred again for both types of illusions, both face sizes but only for the 40 dB and 60 dB

intensities. However, although the effect was not significant for the 50 dB and 70 dB intensities, there was a trend in the predicted direction. This instruction effect is in line with that reported by Massaro (1998, p184-188). Combinations were more frequently elicited in a “multiple choice” condition than in an “open-ended” condition, which fitted well the FLMP.

In the present “free response” condition, participants were not given any example of the possible answers. Since the session began with the auditory control condition, participants may have become accustomed to the /b/, /g/, /p/, and /k/ answers. During the audiovisual condition, they may just have assumed that the answers the experimenter expected were of the same kind, which would account for the small number of illusions in this condition. This small number of illusions fits with the fact that the McGurk percepts (especially fusions) are generally weak in French (also see Colin 2001; Colin et al., 2001, 2002). In the “multiple choice” condition, on the contrary, participants were given a restricted set of possibilities, including the expected illusions for trials of the fusion and combination types. They were thus able to develop stronger expectancies regarding the answers than in the “free response” condition and, to possibly adopt strategies consisting of balancing their responses across all the possibilities.

If McGurk illusions are strongly modulated by instruction manipulation, we should note that they still arise in unfavourable instruction conditions. For example, in the present “free response” condition, illusions occurred on average in 13% of the cases, that is, in complete absence of any suggestion from the experimenter or from the experimental context. Moreover, in Sekiyama and Tohkura‘s study (1993), in which participants were asked to judge whether there was a discrepancy between auditory and visual signals, illusions did not completely disappear in case of discrepancy detection. A part of audiovisual integration mechanisms in speech can therefore be regarded as automatic, as assumed by the Motor Theory of speech which considers the computations performed by the speech perception module to be cognitively impenetrable, i.e. unaffected by information from other modules or from the experimental context. However, according to the cognitive impenetrability principle, integration processes are not supposed to be affected by cognitive factors.

Nevertheless, it is quite conceivable that those manipulations do not modulate integration mechanisms in themselves but act at a late decisional level or even at an earlier perceptual stage, through the filtering or gating functions of the efferent auditory system (for a review of the efferent auditory system, see Sahley, Nodar & Musiek, 1997).

Sensory factors, such as the salience of the auditory and visual signals, probably impinge on audiovisual integration mechanisms at an early perceptual stage, since they modulate the input signal. In the present study, we separately manipulated the visual and the auditory salience of the stimuli by respectively varying face size and auditory intensity.

In the first experiment, an intensity effect was only found for combinations that were more numerous at 40 dB or 50 dB than at 60 dB or 70 dB. In the second one, the intensity effect was highly significant. The 40 dB intensity gave rise to far more fusions and combinations than the other three intensities. We hypothesized that this difference between the two experiments might be accounted for by an “over-learning” effect. In Experiment 1, each participant was presented with 48 repetitions of the four auditory stimuli and of the eight audiovisual stimuli (instead of 12 in Experiment 2). This prolonged exposure to physically identical stimuli may have allowed participants to “over-learn” their responses and so to give the same answers repeatedly, independently of the auditory intensity. Such an “over-learning” effect had already been suggested by Jordan and Thomas (2001) to account for the lack of effect of viewing angle variations on audiovisual speech recognition. Many identical physical articulations were repeated in their experiments, which made the participants unusually familiar with the visual stimuli. According to these authors, this familiarity may have enhanced the explicit processing of the stimuli, preventing a possible viewing angle effect from occurring. The “over-learning” mechanism may therefore be regarded as a top-down influence, affecting late decision processes in the same way instruction manipulations probably do.

The strong intensity effect of the Experiment 2 could be accounted for by another top-down influence. Indeed, in degraded listening conditions, the enhanced visual influence does not

necessarily reflect a true perceptual effect but intentional lipreading by listeners who were unable to use the auditory modality. However, in the present study, as explained in the Data Analysis section of Experiment 1, care was taken for the so-called visual responses to be real audiovisual interactions. Consequently, the increase in McGurk fusions and combinations observed for the 40 dB intensity is probably due to the reduction in the salience of auditory input rather than to an explicit lipreading strategy.

The auditory salience effect found in Experiment 2 replicates the results observed in many previous studies (Colin et al., 2002; Fixmer & Hawkins, 1998; Hardison, 1996; Jordan & Sergeant, 1998; Sekiyama & Tohkura, 1991; Traill, 1999). As demonstrated by Miller and Nicely (1955), place of articulation is difficult to isolate from the auditory signal (because it is specified by rapid and weak acoustic changes) and is severely affected by noise. It can so be assumed that place of articulation is much more difficult to pick up from the auditory modality at 40 dB than at 70 dB. At lower intensities, the perceptual system would then rely more on visual modality to extract place of articulation, leading to more McGurk illusions in case of audiovisual discrepancy.

Most studies that manipulated auditory intensity failed to test whether the increased incidence of McGurk illusions with auditory salience reduction is an abrupt or gradual phenomenon. Fixmer and Hawkins (1998) were the only ones to use three levels of auditory noise and find that McGurk illusions decreased between a noise-free condition and a moderate-noise condition and between the moderate-noise condition and a severe-noise condition. Here, although intensity was linearly manipulated, illusions decreased abruptly between 40 dB and all other intensities. In the same way, identification errors in the auditory control condition were much more numerous at 40 dB than at the other intensities (see Tables 1 and 3). This is in line with the steep intensity-performance functions found for various speech samples in normally-hearing young adults (Brandy, 2002). In line with the present study, recognition scores reached their maximum around 40-50 dB, for monosyllables. Future McGurk studies could interestingly include vocal audiometric

tests in order to individually set up the auditory intensity level at which the perceptual system will be constrained to make use of the visual signal in order to identify audiovisual speech tokens.

Let us now consider what happened when the salience of the visual signal was manipulated. In the first experiment, the large face elicited on average 4% illusions more than the small face (but this difference was not statistically significant). In the second experiment, the difference between the two face sizes was significant but reached only 8% on average. In line with our prediction, the face size effect was globally not so important. In speech perception, visual modality mainly provides information about the place of articulation. Consonants belonging to two very easily distinguishable places of articulation (bilabials and velars) were used here as well as by Jordan and Sergeant (1998). In this regard, it is not so surprising to obtain only a small effect of face size. Many studies in which strong effects of the visual signal manipulations were obtained involved important deteriorations of the dynamical properties of the face (e.g. Nelson & Hodge, 2000) or of the configural aspects of the face (e.g. Jordan & Bevan, 1997). Taken together, the present data and recent literature on audiovisual speech perception indicate that, provided configural and kinematic properties of the face are not too disrupted, the visual information may be relatively coarse, since its perceptibility remains globally unimpaired even with a rather small talking face.

According to Massaro (1987, 1998), one of the major predictions of the FLMP is that the "influence of one modality will be greater to the extent that the other is neutral or ambiguous". For example, when the auditory input is impoverished (e.g. by noise), the perceptual system will rely more on the visual input. Therefore, the present sensory effects agree with the FLMP statement of a role of auditory and visual salience on audiovisual speech integration mechanisms. It should however be reminded that, as argued by Vroomen and de Gelder (1999, p37), the FLMP does not make predictions, but "fits data retrospectively by adjusting truth values until there is a satisfying fit". Moreover, according to the demonstration by Schwartz (2003), the McGurk paradigm is a specific case in which the use of the FLMP is mathematically unsound.

A particular aspect of our data was the asymmetry between fusion and combination percentages, a result replicating the asymmetry already found in French with other speakers (Colin, 2001; Colin et al., 2001, 2002). In the current experiments, combinations were on average 19% more numerous than fusions. An explanation of the asymmetry between fusions and combinations resting on the perceptual weight allocated to the acoustic and visual signals has already been suggested in one of our previous papers (Colin et al., 2002).

According to the present data and to those reported in the literature, we suggest that the perceptual processes underlying audiovisual integration are modulated by sensory and cognitive variables. Sensory factors would vary the perceptual weight that is allocated to each modality. This perceptual weight modulation may be regarded as a pure sensory phenomenon. It cannot however be excluded that modulating the salience of the input signal entails an attentional shift resulting in a top-down reweighting of the signals used in audiovisual speech perception. For example, a particularly weak auditory signal may be attributed less weight because it shifts the attention of the listener to the visual signal, retroactively modulating the weight allocated to each modality.

Whether the influence of sensory factors on the McGurk effect is merely perceptual or partly cognitive, this influence is fairly strong and should be taken into account in audiovisual speech integration studies. It is unfortunate that in many studies (e.g. Mills & Thiem, 1980; Tillman, Pompino-Marschall & Porzig, 1984; Dekle, Fowler & Funnell, 1992; Fuster-Duran, 1996; Jordan & Bevan, 1997; Sekiyama, 1998 or Cathiard, Schwartz & Abry, 2001) the presentation conditions of the auditory and visual signals are not mentioned. It is therefore difficult to assess to what extent the size of the reported McGurk effect is modulated by the particular experimental context. Auditory intensity and room background noise are generally not reported. In the same way, the size of the speaker face is rarely mentioned and the same is true for the distance at which it is displayed. The angle subtended by the talking face therefore can not be computed.

Cognitive factors such as instructions clearly modulate the size of the McGurk illusions probably by affecting a late decision level, i.e. the level at which the response is selected. We

suggest that the degree of automaticity, or cognitive impenetrability, of the processes underlying audiovisual integration in speech should depend on the situation under investigation. When the conditions of occurrence of the McGurk effect are at best, auditory and visual signals are probably automatically integrated at an early perceptual level, without the participant becoming aware of the incongruence. This would occur with auditory and visual signals seeming perfectly compatible, with clear visual signals or with weak auditory signals, for example. When the experimental context is less favourable, for example, when instructions favour the auditory modality, when the visual signal is poor or when participants are or become aware of the dubbing procedure, modality-specific processing would be enhanced and would compete with automatic integration. In that case, several responses may be possible and top-down mechanisms certainly influence response selection processes. As long as the possible effects of sensory and cognitive variables at play in an experiment are not understood, the contribution of both perceptual automatic and top-down cognitive processes on the audiovisual speech integration mechanisms may prove tricky to disentangle.

Acknowledgements

This research has been supported by a F.N.R.S. L.N. grant (8.4501.98) and by a F.E.R. grant from the U.L.B. to Monique Radeau. The authors are grateful to Alain Soquet for technical help in constructing the stimuli. We also earnestly thank Sarah Hawkins and two anonymous reviewers for their thorough examination of the manuscript and their very helpful and relevant comments.

References

Amano, J., & Sekiyama, K. (1998). The McGurk effect is influenced by the stimulus set size. Proceedings of the Auditory-Visual Speech Processing Conference. Terrigal, Australia, 43-48.

Benoît, C., Mohamadi, T., & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of speech. Journal of Speech and Hearing Research, 37, 1195-1203.

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.

Brandy, W.T. (2002). Speech audiometry. In J. Katz (Ed.), Handbook of Clinical Audiology (pp. 96-110). Lippincott Williams & Wilkins.

Burnham D. (1998). Language specificity in the development of auditory-visual speech perception. In R. Campbell, B. Dodd & D. Burnham (Eds), Hearing by eye II (pp. 27-60). Psychology Press.

Cathiard, M.-A., Schwartz, J.-L., & Abry, C. (2001). Asking a naïve question about the McGurk effect : Why does audio [b] give more [d] percepts with visual [g] than with visual [d] ? Proceedings of the Auditory-Visual Speech Processing Conference. Aalborg, Denmark, 138-142.

Cerrato, L., Leoni, F.A., & 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.

Colin C. (2001). Etude comportementale et électrophysiologique des processus impliqués dans l'effet McGurk et dans l'effet de ventriloquie. Unpublished doctoral thesis. Free University of Brussels, Brussels, Belgium.

Colin C., & Radeau M. (2003). Les illusions McGurk dans la parole : 25 ans de recherches. L'Année Psychologique, 114, 497-542.

Colin C., Radeau M., Deltenre P., Demolin D., & Soquet A. (2002). The role of sound intensity and stop-consonant voicing on McGurk fusions and combinations. European Journal of Cognitive Psychology, 14, 475-491.

Colin C., Radeau M., Deltenre P., & Morais J. (2001). Rules of intersensory integration in spatial scene analysis and speechreading. Psychologica Belgica, 41, 131-144.

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.

Dekle, D.J., Fowler, C.A., & Funnell, M.G. (1992). Audiovisual integration in perception of real words. Perception and Psychophysics, 51, 355-362.

Demorest, M., & Bernstein, L. (1992). Sources of variability in speechreading sentences : A generalizability analysis. Journal of Speech and Hearing Research, 35, 876-891.

Dodd, B., McIntosh, B., & Woodhouse, L. (1998). Early lipreading ability and speech and language development of hearing-impaired pre-schoolers. In R. Campbell, B. Dodd, & D. Burnham (Eds.), Hearing by eye II (pp. 229-242). Psychology Press.

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

Erber, N.P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. Journal of Speech and Hearing Research, 12, 423-425.

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.

Fuster-Duran, A. (1996). Perception of conflicting audio-visual speech : An examination across Spanish and German. In D. G. Stork & M. E. Hennecke (Eds.). Speechreading by Humans and Machines, NATO ASI Series F : Computer and Systems Sciences, Springer-Verlag, 150, 135-143.

Green, K.P. (1998). The use of auditory and visual information during phonetic processing: implications for theories of speech perception. In R. Campbell, B. Dodd, & D. Burnham (Eds.), Hearing by eye II (pp. 3-25). Psychology Press.

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.

Hietanen, J.K., Manninen, P., Sams, M., & Surakka, V. (2001). Does audiovisual speech perception use information about facial configuration ? European Journal of Cognitive Psychology, 13, 395-407.

Jones, J.A., & Munhall, K.G. (1997). The effects of separating auditory and visual sources on audiovisual integration of speech. Canadian Acoustics, 2, 13-19.

Jordan T.R., & Bevan K. (1997). Seeing and hearing rotated faces : Influences of facial orientation on visual and audiovisual speech recognition. Journal of Experimental Psychology : Human Perception and Performance, 25, 388-403.

Jordan T.R., & Sergeant P.C. (1998). Effects of facial image size on visual and audio-visual speech recognition. In R. Campbell, B. Dodd & D. Burnham (Eds), Hearing by eye II (pp. 155-176). Psychology Press.

Jordan, T.R., & Sergeant, P.C. (2000). Effects of distance on visual and audiovisual speech recognition. Language and Speech, 43, 107-124.

Jordan, T.R., & Thomas, S.M. (2001). Effects of horizontal viewing angle on visual and audiovisual speech recognition. Journal of experimental Psychology : Human Perception and Performance, 27, 1386-1403.

Kuhl, P., Green, K.P. & Meltzoff, A.N. (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., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M. (1967). Perception of the speech code. Psychological Review, 74, 431-461.

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

MacDonald, J., Andersen, S., & Bachman, T. (2000). Hearing by eye : How much spatial degradation can be tolerated ? Perception, 29, 1155-1168.

MacLeod, A., & Summerfield, Q. (1990). A procedure for measuring auditory and audio-visuals speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. British Journal of Audiology, 24, 29-43.

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

Massaro, D.W. (1998). Perceiving Talking Faces : From Speech Perception to a Behavioral Principle. The MIT Press.

Massaro, D.W., & Cohen, M.M. (1996). Perceiving speech from inverted faces. Perception and Psychophysics, 58, 1047-1065.

McGurk, H. (1988). Developmental psychology and the vision of speech, Inaugural Professorial Lecture, University of Surrey.

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

Middleweerd, M.J., & Plomp, R. (1987). The effects of speechreading on the speech perception threshold of sentences in noise. Journal of the Acoustical Society of America, 82, 2145-2146.

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

Mills, A.E., & Thiem, R. (1980). Auditory-visual fusions and illusions in speech perception. Linguistische Berichte, 68, 85-107.

Mohamadi T., & Benoît, C. (1992). Apport de la vision du locuteur à l'intelligibilité de la parole bruitée en français. Bulletin de la Communication Parlée, 2, 31-41.

Nelson, M.A., & Hodge, M.M. (2000). Effects of facial paralysis and audiovisual information on stop place identification. Journal of Speech, Language and Hearing Research, 43, 158-171.

Posner, M., & Snyder, C. (1975). Attention and cognitive control. In R.L. Solso (Ed.), Information Processing and Cognition: The Loyola Symposium (pp55-85). Hillsdale, NJ. Lawrence Erlbaum Associates.

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.

Repp, B.H., Manuel, S.Y., Liberman, A.M., & Studdert-Kennedy, M. (1983). Exploring the McGurk effect. Proceedings of the 24th Annual Meeting of the Psychonomic Society. San Diego, CA, 366-367.

Rosenblum, L.D., & Saldana, H.M. (1996). An audiovisual test of kinematic primitives for visual speech perception. Journal of Experimental Psychology : Human Perception and Performance, 22, 318-331.

Rosenblum, L.D., Yakel, D.A., & Green, K.P. (2000). Face and mouth inversion effects on visual and audiovisual speech perception. Journal of Experimental Psychology : Human Perception and Performance, 26, 806-819.

Sahley, T.L., Nodar, R.H., & Musiek, F.E. (1997). Efferent Auditory System : Structure and Function. Singular Publishing Group.

Sams, M., Manninen, P., Surakka, V., Helin, P., & Kättö, R. (1998). McGurk effect in finnish syllables, isolated words, and words in sentences : Effects of word meaning and sentence context. Speech Communication, 26, 75-87.

Schwartz, J.-L. (2003). Why the FLMP should not be applied to McGurk data ... or how to better compare models in the Bayesian framework. Proceedings of the Auditory-Visual Speech Processing Conference. St Jorioz, France, 77-82.

Sekiyama, K. (1998). Face or voice ? Determinant of compellingness to the McGurk effect. Proceedings of the Auditory-Visual Speech Processing Conference. Terrigal, Australia, 33-36.

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., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. Quarterly Journal of Experimental Psychology, 36A, 51-74.

Tiippana, K., Sams, M., & Andersen, T.S. (2001). Visual attention influences audiovisual speech perception. Proceedings of the Auditory-Visual Speech Processing Conference. Aalborg, Denmark, 167-171.

Tillman, H.G., Pompino-Marschall, B., & Porzig, U. (1984). The effects of visually presented speech movements on the perception of acoustically encoded speech articulation as a function of acoustic desynchronization. Proceedings of the 10th International Congress of Phonetic Sciences. Dordrecht, Holland, 469-473.

Traill, A. (1999). The McGurk effect and !Xóõ clicks. Proceedings of the International Conference of Phonetic Sciences. San Francisco, CA, 1933-1935.

Vroomen, J., & de Gelder, D. (1999). Crossmodal integration: A good fit is no criterion. Trends in Cognitive Science, 4, 37-38.

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.

Table 1 : Percentages of errors in the auditory control condition of Experiment 1 as a function of articulation place (Artic. pl.), instructions, error type and auditory intensity. In this table and table 3, errors of the “Other” category mainly contains the errors corresponding to the expected illusion for the matching audiovisual trial, e.g. /bg/ response for an auditory /b/ stimulus.

Artic. pl.	**Bilabial**				**Velar**			
Instructions	**Multiple choice**		**Free Response**		**Multiple choice**		**Free response**	
Error type	**Other**	**Voicing**	**Other**	**Voicing**	**Other**	**Voicing**	**Other**	**Voicing**
40 dB	2.3	19.8	0.0	10.2	3.6	12.0	1.8	6.3
50 dB	0.8	1.6	0.0	0.6	2.1	1.6	0.3	1.0
60 dB	1.0	0.3	0.0	0.0	3.4	0.2	0.0	1.6
70 dB	1.3	0.0	0.0	0.0	6.2	0.0	1.0	1.3

Table 2 : Percentages of combinations and fusions as a function of instructions, face size and auditory intensity in Experiment 1. In this table and table 4, 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.

Illusion	**Combinations**				**Fusions**			
Instructions	**Multiple choice**		**Free response**		**Multiple choice**		**Free response**	
Face	**Small**	**Large**	**Small**	**Large**	**Small**	**Large**	**Small**	**Large**
40 dB	59.4	63.0	32.8	14.0	19.8	39.6	0.5	2.1
50 dB	51.0	65.2	37.0	12.5	21.9	28.7	0.0	12.5
60 dB	40.1	59.9	33.8	12.0	21.9	34.9	0.0	8.4
70 dB	40.7	56.8	31.8	11.5	23.0	40.1	0.0	11.0

Table 3 : Percentages of errors in the auditory condition of Experiment 2 as a function of articulation place (Artic. pl.), instructions, error type and auditory intensity.

Artic. pl.	**Bilabial**				**Velar**			
Instructions	**Multiple choice**		**Free response**		**Multiple choice**		**Free response**	
Error type	**Other**	**Voicing**	**Other**	**Voicing**	**Other**	**Voicing**	**Other**	**Voicing**
40 dB	3.6	17.2	0.8	14.1	5.5	18.8	0.0	12.0
50 dB	2.0	2.7	0.0	0.8	1.8	0.0	0.0	1.0
60 dB	1.8	0.5	0.0	0.0	5.4	0.0	0.0	0.0
70 dB	1.0	0.3	0.0	0.0	6.5	2.4	0.3	0.0

Table 4 : Percentages of combinations and fusions as a function of instructions, face size and auditory intensity in Experiment 2.

Illusion	**Combinations**				**Fusions**			
Instructions	**Multiple choice**		**Free response**		**Multiple choice**		**Free response**	
Face	**Small**	**Large**	**Small**	**Large**	**Small**	**Large**	**Small**	**Large**
40 dB	81.3	80.2	21.9	44.3	22.9	57.8	18.2	17.2
50 dB	25.5	39.6	6.8	16.7	0.5	17.2	13.5	0.0
60 dB	21.9	50.6	0.0	12.5	33.4	14.7	0.0	0.0
70 dB	30.7	43.8	7.8	5.2	1.1	3.7	11.5	13.1

Figure 1 : Illustration of the ratio between the two face sizes used in both experiments