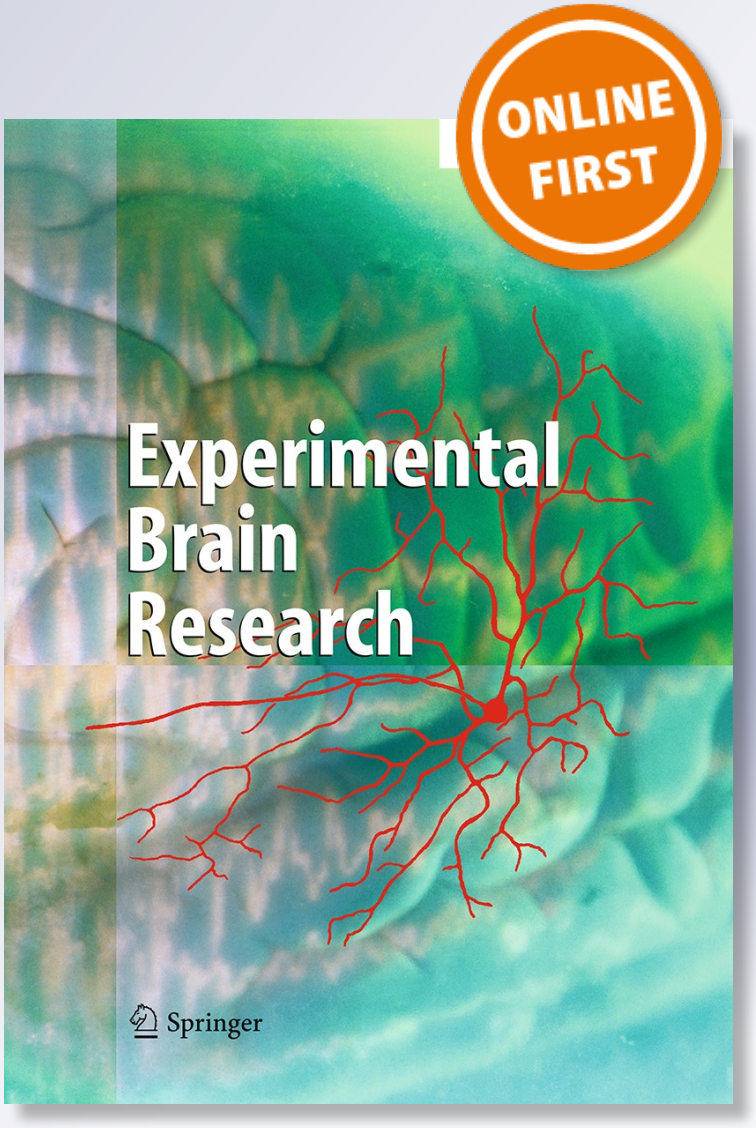
# *Silent articulation modulates auditory and audiovisual speech perception*

**Marc Sato, Emilie Troille, Lucie Ménard, Marie-Agnès Cathiard & Vincent Gracco**

**Experimental Brain Research**

ISSN 0014-4819

Exp Brain Res
DOI 10.1007/s00221-013-3510-8




Springer

Exp Brain Res
DOI 10.1007/s00221-013-3510-8

RESEARCH ARTICLE



# Silent articulation modulates auditory and audiovisual speech perception

**Marc Sato · Emilie Troille · Lucie Ménard · Marie-Agnès Cathiard · Vincent Gracco**

Received: 21 November 2012 / Accepted: 3 April 2013


**Abstract** The concept of an internal forward model that internally simulates the sensory consequences of an action is a central idea in speech motor control. Consistent with this hypothesis, silent articulation has been shown to modulate activity of the auditory cortex and to improve the auditory identification of concordant speech sounds, when embedded in white noise. In the present study, we replicated and extended this behavioral finding by showing that silently articulating a syllable in synchrony with the presentation of a concordant auditory and/or visually ambiguous speech stimulus improves its identification. Our results further demonstrate that, even in the case of perfect perceptual identification, concurrent mouthing of a syllable speeds up the perceptual processing of a concordant speech stimulus. These results reflect multisensory-motor interactions during speech perception and provide new behavioral arguments for internally generated sensory predictions during silent speech production.

**Keywords** Speech perception · Speech production · Silent speech · Audiovisual speech perception · Internal forward models · Sensory-motor interactions · Efference copy · McGurk effect

Aspects of this work were presented at the 2008 International Conference on Auditory-Visual Speech Processing, Tangalooma, Australia.

M. Sato (✉) · E. Troille
GIPSA-LAB, UMR CNRS 5216, Département Parole and Cognition, Grenoble Université, 1180, Avenue centrale, BP 25, 38040 Grenoble Cedex 9, France
e-mail: marc.sato@gipsa-lab.inpg.fr

M. Sato · L. Ménard · V. Gracco
Centre for Research on Brain, Language and Music, McGill University, Montreal, Canada

E. Troille · M.-A. Cathiard
Centre de Recherche sur l'Imaginaire, Université Stendhal, Grenoble, France

L. Ménard
Département de Linguistique, Université du Québec à Montréal, Montréal, Canada

V. Gracco
School of Communication Sciences and Disorders, McGill University, Montreal, Canada

V. Gracco
Haskins Laboratories, New Haven, CT, USA

## Introduction

Speech production is a complex multistage process that converts an intended linguistic message, through specific articulatory movements, into an acoustic speech signal that can be perceived and understood by a listener (Levelt 1989). From higher-order linguistic conceptualization of the intended message, speech production requires phonemic encoding of the articulatory plans, initiation and coordination of sequences of movements produced by the combined actions of the respiratory system, the larynx and the supra-laryngeal vocal tract. Online auditory and somatosensory feedback control mechanisms also play a key role in speech production. During the phonemic encoding stage of the intended linguistic message, it is proposed that segmental speech movements are programmed to reach phonemic auditory and somatosensory goals (for reviews, Perkell et al. 2000; Perkell 2012; Perrier 2005, 2012), which in turn are used to estimate actual sensory inputs during speech production. This feedback monitoring process is thought to be essential for native and foreign

Published online: 17 April 2013

language acquisition during which the relationship between speech motor commands and sensory feedback signals is progressively learned by the central nervous system in relation to the intended speech goals.

The hypotheses that sensory feedback plays an important role in tuning the speech motor control system and that speech goals are defined in multidimensional auditory and somatosensory spaces are also empirically supported in adult/fluent speech production. For instance, unexpected dynamical mechanical loads of supra-laryngeal lip and jaw articulators during speech production, leading to transient transformations of both the auditory and somatosensory feedback of the speech motor act, also result in online and rapid articulatory adjustments (Folkins and Abbs 1975; Abbs and Gracco 1984; Gracco and Abbs 1985). Similarly, compensatory changes in speech production are also observed when auditory feedback is altered in its pitch (Elman 1981; Burnett et al. 1998; Jones and Munhall 2000), vowel formant frequencies (Houde and Jordan 1998; Houde et al. 2002; Jones and Munhall 2005; Purcell and Munhall 2006a, b), or fricative first spectral moment (Shiller et al. 2009, 2010). Importantly, while auditory information is often assumed as the dominant sensory modality, the integration of auditory and somatosensory information in the achievement of speech movements has been demonstrated (Feng et al. 2011; Lamettti et al. 2012) and mechanical loads that alter jaw movements without perceptible effect on acoustic output yield robust compensation (Tremblay et al. 2003; Nasir and Ostry 2006). Finally, while manipulation of the auditory and/or somatosensory feedback during speech production leads to rapid motor corrections to counteract the effect of perturbation, an aftereffect or perceptuo-motor adaptation can also be observed when the perceptual manipulation is removed (Houde and Jordan 1998; Houde et al. 2002; Tremblay et al. 2003; Jones and Munhall 2005; Nasir and Ostry 2006; Purcell and Munhall 2006b; Shiller et al. 2009). Interestingly, the fact that motor compensatory adjustments do not disappear immediately likely reflects a global remapping, or recalibration, of the sensory-motor relationships.

Taken together, all these studies reveal a key role for online auditory and somatosensory feedback control mechanisms and highlight the sensory nature of speech goals. In order to explain how the central nervous system rapidly reacts to perturbations and adjusts fine-grained motor parameters, in spite of the intrinsic temporal limitations of the biological feedback systems, the concept of efference copy (von Holst and Mittelstaedt 1950) and internal models (Francis and Wonham 1976; Kawato et al. 1987) has been introduced in the speech motor control literature (Guenther 1995; Perkell et al. 1997; Guenther et al. 1998; Houde and Jordan 1998; for recent reviews, see Perkell et al. 2000; Perkell 2012; Perrier 2012). Perceptuo-motor goals that define successful speech motor acts are thought to be gradually acquired in the initial stage of the speech motor learning process. When reaching sensory speech goals, the relationships between speech motor commands and actual sensory feedback signals are then progressively learned by the central nervous system and stored in the form of an *internal forward model*. Once a robust forward model has been acquired, it allows for the prediction of the sensory consequences of speech motor movements in relation to an intended sensory speech goal. These *internal sensory predictions*, generated prior to the actual execution and feedback from the intended speech movements, partly compensate for the delay inherent in feedback systems and can assist in speech motor control. On the other hand, sensory input from the speech mechanism is still available and can be used, notably in case of external perturbations, in the comparison between internal sensory predictions from the forward model and actual sensory inputs for online corrective motor adjustments.

Recent brain imaging studies and neurobiological models of speech production support online auditory and somatosensory feedback control mechanisms and the existence of internal forward models. For instance, reduced responses of the auditory cortex have been reported during self-produced overt speech, compared to utterances recorded and replayed to the subjects. This so-called speaking-induced response suppression has been observed using functional magnetic resonance imaging (fMRI; Christoffels et al. 2007, 2011), positron emission tomography (PET; Hirano et al. 1996, 1997), electroencephalography (EEG; Ford et al. 2001; Ford and Mathalon 2004) and magnetoencephalography (MEG; Numminen and Curio 1999; Numminen et al. 2000; Curio et al. 2000; Houde et al. 2002; Heinks-Maldonado et al. 2006; Ventura et al. 2009). Conversely, compared to normal auditory/somatosensory feedback, increased activity during overt speech production has been observed in the auditory cortex with, continuous or transient, altered or delayed auditory feedback (Hashimoto and Sakai 2003; Heinks-Maldonado et al. 2006; Christoffels et al. 2007, 2011; Tourville et al. 2008), as well as in the anterior supramarginal gyrus during unexpected somatosensory feedback (Golfinopoulos et al. 2011). More detailed accounts of forward models have also been proposed in recent neurobiological models of speech production. For instance, in the DIVA model of speech production (Directions Into Velocities of Articulators; Guenther 2006; Guenther and Vladusich 2012), modulated responses within the auditory and somatosensory cortices are thought to reflect online corrective control mechanisms in which auditory and somatosensory consequences are estimated internally through efference copy of the planned motor commands (for similar models derived from feedback motor control and internal forward model of speech

production, see also Ventura et al. 2009; Tian and Poeppel 2010; Hickok et al. 2011; Price et al. 2011; Hickok 2012).

Evidence for forward models has also been shown during silent articulation and inner (imagined) speech (Yetkin et al. 1995; Numminen and Curio 1999; Wise et al. 2001; Shergill et al. 2002; Hickok et al. 2003; Shuster and Lemieux 2005; Tian and Poeppel 2010). Crucially, speaking-induced response suppression has been observed during silent articulation (Numminen and Curio 1999). A recent MEG study also demonstrates auditory activity around 170 ms in the presence of covert speech and the absence of auditory feedback (after articulatory imagery of a syllable) (Tian and Poeppel 2010). These results appear in line with internal forward models and the existence of internal motor-to-sensory predictions in the absence of auditory feedback. More generally, they also support the hypothesis that motor simulation relies, at least partly, on neural mechanisms common to those for motor execution (for reviews, see Jeannerod 1994, 2001). Because silent articulation modulates the activity of the auditory cortex, presumably by means of auditory prediction of the simulated speech motor act, one prediction is that it might interfere with or bias speech perception. Consistent with this hypothesis, Sams et al. (2005) showed that the identification of an acoustically presented syllable, embedded in continuous white noise, is enhanced or, conversely, reduced when listener silent articulates a similar/different syllable synchronously. According to Sams et al. (2005), these results can be explained by auditory (and somatosensory) predictions of the speech motor commands sent in parallel to the auditory cortex which, in turn, influence in a phoneme-specific manner the perception of the acoustic stimulus.

In order to further extend this finding, we examined in the present study whether silently articulating a syllable in synchrony with the perception of a similar, or different, auditory or audiovisual syllable might improve and speed up or, conversely, reduce and slow its identification. In order to vary syllable identification, they were embedded or not with continuous white noise and, in case of the audiovisual syllables, with the visual track spatially degraded or not. In addition, we tested the possible effect of silent articulation on the perception of incongruent audiovisual syllables, which were expected to produce a strong McGurk effect (i.e., when a visual /aga /dubbed with an acoustic /aba/ is perceived as /ada/; McGurk and MacDonald 1976). These stimuli were presented also with the visual track spatially degraded or not and the acoustic track embedded or not with continuous white noise. Mouthing a concordant syllable was expected to improve and speed up the identification of the auditory and audiovisual syllables, especially in case of acoustically and/or visually degraded ambiguous stimuli. Regarding the incongruent audiovisual syllables, mouthing a syllable concordant with the McGurk percept was expected to bias and speed up the perception toward this illusory syllable.

## Methods

### Participants

Sixteen voluntary healthy subjects participated in the study (eleven females; mean age ± SD: 28 years ± 7). All were native Canadian French speakers, right-handed (Oldfield 1971), with normal or corrected-to-normal vision and no reported history of speaking or hearing disorders. Written informed consent was obtained for all subjects and they were paid for their participation. The protocol was approved by the McGill University Faculty of Medicine Institutional Review Board.

### Stimuli

Multiples utterances of /aba/ and /aga/ syllables were individually recorded by a male native Canadian French speaker, using a high-quality digital video camera. The speaker produced each syllable naturally, maintaining an even intonation, tempo and vocal intensity. Video digitizing (the speaker's full face being presented against a gray background) was done at 30 frames per second with a resolution of 720 × 480 pixels. Audio digitizing was done at 44.1 kHz with 16-bit quantization recording. One clearly articulated /aba/ and /aga/ tokens were selected. The two syllables were temporally aligned, according to the first vocalic and consonantal onsets (mean value ± SD: 177 ms ± 12 and 616 ms ± 16, respectively) and matched for global acoustic duration (mean value ± SD: 731 ms ± 13) and intensity (mean value ± SD: 69 dB ± 1). The speaker initiated and finished each utterance from a neutral closed-mouth position, with each movie being 36 frames long (1,200 ms.). From these syllables, 16 stimuli were then created using Adobe Premiere software (Adobe Systems, Inc., San Jose, USA). The 16 stimuli included:

- A and $A_n$: Four auditory stimuli consisting of /aba/ or /aga/ syllables presented together with a static face of the speaker. Each stimulus pair (/aba/, /aga/) was presented in clear audio (A) or embedded with continuous white noise at a −6 dB signal-to-noise ratio ($A_n$).
- AV, $AV_n$, $A_nV$ and $A_nV_n$: Eight congruent audiovisual stimuli consisting of /aba/ or /aga/ syllables presented with the speakers moving face. The stimuli were either unedited (AV), with the visual track spatially degraded (the visual track of the speaker's face being quantized by a mosaic transform of 36 × 24 pixels—see Fig. 1) ($AV_n$), the audio track embedded in continuous white

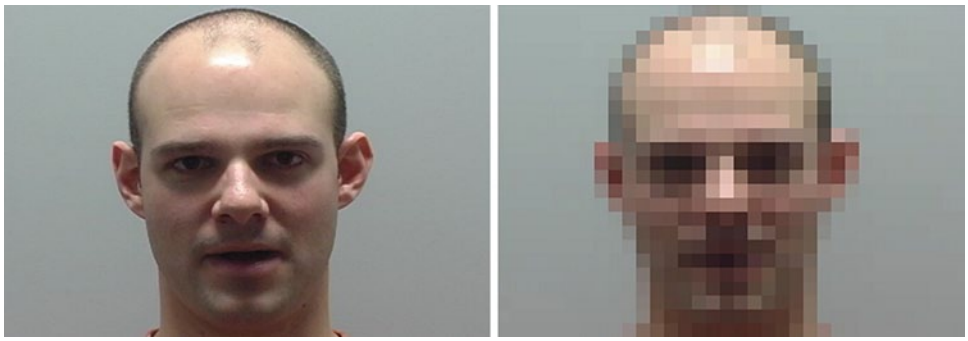


**Fig. 1** Still picture of the face stimulus used in the audiovisual conditions. *Left* original video recording (720 × 480 pixels), *Right* spatial quantization (36 × 24 pixels)

noise at a −6 dB signal-to-noise ratio ($A_nV$) or both audio and visual tracks degraded/embedded ($A_nV_n$).

- iAV, $iAV_n$, $iA_nV$ and $iA_nV_n$: Four incongruent audiovisual stimuli corresponding to auditory /aba/ and visual /aga/ syllables with no manipulation of the audio and visual tracks (iAV), with the visual track spatially degraded ($iAV_n$) (see above), the acoustic track embedded in continuous white noise at a −6 dB signal-to-noise ratio ($iA_nV$) or both audio and visual tracks manipulated ($iA_nV_n$).

## Experimental procedure

The experiment was carried out in a sound-attenuated room. Participants sat comfortably in front of a 19 in. computer monitor at a distance of approximately 50 cm. The acoustic stimuli were presented at a comfortable sound level through earphones. During the experiment, the subject's face was recorded via a digital video camera, together with the acoustic track of the stimulus, to ensure the correctness of the subject's silent articulation.

The subjects' task was to carefully listen to and/or watch auditory and audiovisual stimuli of a speaker and to report the perceived syllable (see Fig. 2). A speeded three-alternative forced-choice identification task was used, in which participants had to give a motor response as fast and accurately as possible, by pressing one of three keys corresponding to either /aba/, /ada/ or /aga/ with their right fore finger, middle finger or ring finger (the key designation was counterbalanced across participants). On some trials, they were instructed to silently articulate a syllable in synchrony with the stimulus presentation. Each trial started with either /aba/ or /ada/ presented at the center of the screen for 350 ms, followed by either a green or a red circle for 350 ms and then the auditory or audiovisual stimulus presentation for 1,200 ms. In case of a green circle, participants silently articulated the written syllable and identified the presented auditory or audiovisual syllable; otherwise they just identified the auditory or audiovisual presented syllable.

A go/no-go procedure (no articulation vs. silent articulation) was used to dissociate the possible influence of the

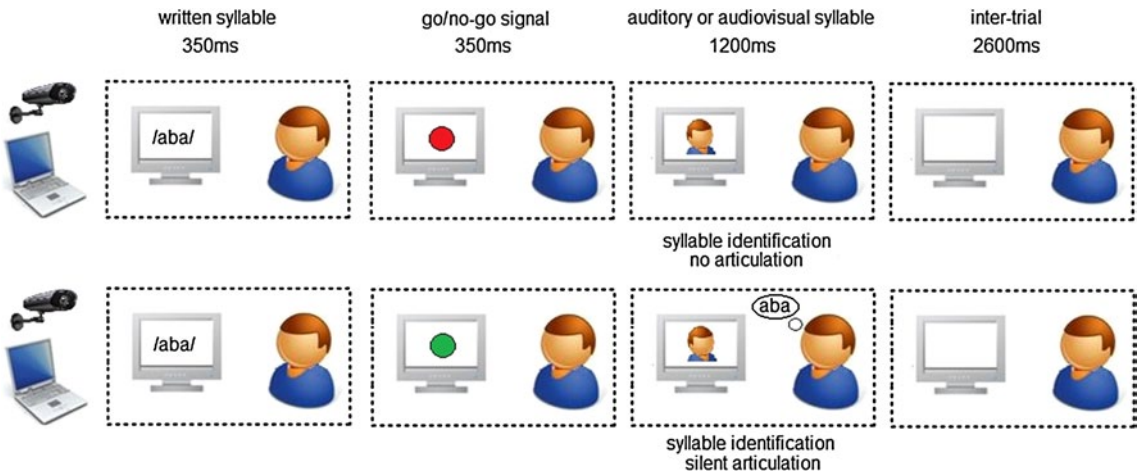



**Fig. 2** Experimental design. Participants had to carefully listen to and/or watch auditory (/aba/ or /aga/), congruent audiovisual (/aba/ or /aga/) or incongruent audiovisual syllables ($/aba/_A$-$/aga/_V$) and to manually report the perceived syllable (/aba/, /ada /or /aga/). On some trials (*go trials*, *bottom*), they were instructed to silently articulate a visually presented syllable (/aba/ or /ada/) in synchrony with the stimulus presentation. Otherwise (*no-go trials, top*), they just had to identify the auditory or audiovisual presented syllable without articulation

written cue syllable to that of silent articulation on auditory or audiovisual perception. Indeed, it has been shown that a written cue presented together with an acoustic stimulus might bias speech perception (Fowler and Dekle 1991; Sams et al. 2005). The intertrial interval was 2,600 ms. During this period, the computer screen remained blank.

The /aba/ and /ada/ written syllables were chosen in order to directly test our two hypotheses on the auditorily and audio visually presented /aba/ syllables as well as on the incongruent /aba/$_A$-/aga/$_V$ syllables. However, due to the duration of the experiment (around 45 min), we did not include a condition in which participants had to mouth an /aga/ syllable. For this reason, the auditorily and audio visually presented /aga/ syllables here served as a control condition. Regarding our first hypothesis, silently articulating a syllable in synchrony with the perception of a similar, or different, auditory or audiovisual syllable might improve and speed up or, conversely, reduce and slow its identification. For example, mouthing an /aba/ syllable compared to mouthing an /ada/ syllable might improve and speed up the identification of /aba/ presented syllables. For /aga/ syllables, we expect no identification differences when mouthing /aba/ or /ada/ syllables. Finally, regarding our second hypothesis on the incongruent audiovisual syllable, mouthing a syllable concordant with the McGurk percept (i.e., /ada/) might bias and speed up the perception toward this illusory syllable.

Because of the relative complexity of the experimental procedure (go/no-go paradigm, silent articulation of the written syllable in synchrony with the presented auditory or audiovisual syllable, three-alternative forced-choice identification task), participants were progressively trained before starting the experiment. To this aim, three training sessions of 32 trials were performed. For the first session, participants were trained to repeat aloud the written syllable /aba/ or /ada/ in synchrony with the stimulus presentation. On 1/3 of the trials, the stimulus was preceded by a red circle indicating no response from the participants. The second training session was identical to the first one except that participants had to silently articulate the written syllable in the presence of a green circle. In order to facilitate the synchrony between the subject's articulation and the stimulus presentation, only audiovisual stimuli, without background noise or spatial quantization, were presented in these two sessions. The last training session was identical to the second one except that participants had also to report the perceived syllable and all stimuli types were presented. In each training session, the investigator ensured that the subject performed the task correctly; notably that he/she pronounced the written syllable in synchrony with the stimulus presentation only in the presence of a green circle and without any phonation in the two last sessions. The three training sessions lasted around 10 min. The training session was repeated for few participants who had difficulty correctly performing the task.

For the experiment, each of the sixteen stimuli was presented 30 times. Each stimulus was preceded either by the written syllable /aba/ (10 trials) or /ada/ (10 trials) and the green circle, or by the written syllable /aba/ (5 trials) or /ada/ (5 trials) and the red circle. Because of the duration of the experiment and our focus was primarily on the silent articulation condition, we chose to employ more trials with silent articulation than without (i.e., 20 trials with silent articulation and 10 trials without articulation). For each subject, the stimuli were presented in a fully randomized sequence for a total of 480 trials. The experiment lasted around 45 min and was divided in three experimental sessions of equal durations. Short breaks were offered between sessions.

## Data analyses

Audiovisual inspection of all trials was first carried out to ensure the correctness of subject's articulation. Four types of articulation errors were observed: silent articulation in case of a red circle or, in case of a green circle, no articulation, wrong articulation (e.g., /aba/ instead of /ada/) and asynchrony between articulation and the stimulus presentation. In addition, outlier response times (i.e., responses faster than 200 ms and slower than 2,000 ms) were also considered as errors. Three participants were removed from the analysis because their errors exceeded 10 % of all trials. For the remaining thirteen participants, the mean percentage of error was 4 % (SD $\pm$ 3 %).

Median reaction times (RTs) and percentage of correct responses were the dependent variables. For /aba/$_A$-/aga/$_V$ stimuli, given the hypothesis that mouthing a syllable concordant with the McGurk percept would bias the perception toward this illusory syllable (i.e.,/ada/), the percentage of "correct" responses for /aba/$_A$-/aga/$_V$ were calculated according to the number of /ada/ responses (i.e., corresponding to an audiovisual fusion effect; McGurk and MacDonald 1976). RTs were calculated from the consonantal onset of the auditory syllable (i.e., [b] for /aba/ or [g] for /aga/). For each dependent variable, repeated-measures three-way analyses of variance (ANOVA) were performed separately for /aba/, /aga/ and /aba/$_A$-/aga/$_V$ stimuli. For /aba/ and /aga/, the within-subject independent variables corresponded to the type of articulation (written /aba/ followed by a red circle, without articulation; written /ada/ followed by a red circle, without articulation; written /aba/ followed by a green circle, with silent articulation; written /ada/ followed by a green circle, with silent articulation), to the type of auditory presentation (without acoustic noise; with acoustic noise) and to the type of visual presentation (static face; visual movie without spatial quantization;

visual movie with spatial quantization). For /aba/$_A$-/aga/$_V$, the independent variables were the same except for the type of visual presentation (visual movie without spatial quantization; visual movie with spatial quantization). For all the analyses, the significance alpha level was set at $p < .05$ and Greenhouse–Geisser corrected when appropriate. When required, post hoc analyses were conducted with Newman–Keuls tests (all comparisons reported being significant). All statistical analyses were done using Statistica software (StatSoft Inc., Tulsa, USA).

## Results

Reaction times and proportions of correctly identified /aba/, /aga/ and /aba/$_A$-/aga/$_V$ syllables in the different experimental conditions are presented in Fig. 3a, b, c and Tables 1, 2, 3, respectively.

### Perception of /aba/—identification scores

Irrespective of the articulation conditions, the identification of /aba/ syllables replicates a number of well-known effects in auditory and audiovisual speech perception (see section "Discussion"), with fewer correct responses observed in the presence of acoustic noise ($A_n$) and visually degraded stimuli ($A_nV_n$). First, the main effect of auditory presentation was significant ($F(1,12) = 144.6$, $p < .001$), with more correct responses when the stimulus was presented without than with acoustic noise (on average, 94 vs. 66 %). Second, the type of visual presentation also produced significant differences ($F(2,24) = 99.1$, $p < .001$), with more correct responses without spatial quantization than with spatial quantization (on average, 94 vs. 88 %), as well as with spatial quantization than with the static face (on average, 88 vs. 57 %). Finally, the interaction between the type of auditory and visual presentation was also significant ($F(2,24) = 82.4$, $p < .001$). Post hoc analyses showed that there were fewer correct responses in the auditory-only presentation in presence of noise (An) than when both audio and visual presentations with noise were available ($A_nV_n$). More correct responses were observed in all the other conditions (on average, $A_n$: 22 % > $A_nV_n$: 82 % > A: 92 % = $A_nV$: 94 % = $AV_n$: 94 % = AV: 95 %).

Of more interest is the significant effect of articulation ($F(3,36) = 5.7$, $p < .005$), with more correct responses when reading and/or silently articulating /aba/ (on average, 83 and 86 %) than when reading and/or silently articulating /ada/ (on average, 76 and 75 %). This indicates that reading or mouthing a concordant syllable improves the identification of the presented /aba/ syllables. However, it appears that this effect was only observed in the presence of acoustic noise ($A_n$) and in case of visually degraded stimuli ($A_nV_n$). Indeed, as shown by a significant articulation × auditory presentation interaction ($F(3,36) = 3.9$, $p < .05$), this effect was only observed in the presence of auditory noise (on average, reading /aba/: 70 % = mouthing /aba/: 76 % > mouthing /ada/: 58 % = reading /ada/: 59 %), no effect of articulation being observed without auditory noise (on average, reading /aba/: 95 % = mouthing /aba/: 95 % = reading /ada/: 92 % = mouthing /ada/: 92 %). Similarly, the articulation × visual presentation interaction was also significant ($F(6,72) = 5.0$, $p < .001$), with no effect of articulation observed with and without spatial quantization, but significantly more correct responses with the static face when silently articulating and/or reading /aba/ than when reading and/or silently articulating /ada/ (on average, mouthing /aba/: 72 % = reading /aba/: 60 % > reading /ada/: 49 % = mouthing /ada/: 48 %).

Finally, the three-way interaction was also found to be significant ($F(6,72) = 4.9$, $p < .001$). There was no effect of articulation observed in the AV, $AV_n$, A and $A_nV$ conditions. In $A_nV_n$ condition, there were more correct responses when reading /aba/ than when silently articulating /ada/ (on average, reading /aba/: 89 % > mouthing /ada/: 75 %). In $A_n$ condition, there were more correct responses when mouthing /aba/ than when reading /aba/, and when reading /aba/ than when silently articulating or reading /ada/ (on average, mouthing /aba/: 49 % > reading /aba/: 25 % > mouthing /ada/: 8 % = reading /ada/: 6 %).

### Perception of /aba/—reaction times

The main effects of auditory ($F(1,12) = 23.2$, $p < .001$) and visual ($F(2,24) = 18.5$, $p < .001$) presentations were significant. RTs were faster without than with acoustic noise (on average, 732 vs. 868 ms), and without spatial quantization than with spatial quantization (on average, 732 vs. 804 ms) and with spatial quantization than with the static face (on average, 804 vs. 865 ms). The interaction between these two variables was also significant ($F(2,24) = 12.9$, $p < .001$) with slower responses in $A_n$ condition than in $A_nV_n$ condition and in $A_nV_n$ condition than in $A_nV$, A, $AV_n$ and AV conditions (on average, $A_n$: 978 ms > $A_nV_n$: 866 ms > $A_nV$: 760 ms = A: 751 ms = $AV_n$: 742 ms = AV: 704 ms).

Crucially, the main effect of articulation was significant ($F(3,36) = 6.5$, $p < .005$), with slower RTs when silently articulating /ada/ compared to all other conditions (on average, mouthing /ada/: 856 ms > mouthing /aba/: 807 ms = reading /ada/: 778 ms = reading /aba/: 760 ms). In addition, the articulation × auditory presentation interaction was also significant ($F(3,36) = 4.8$, $p < .01$), with no effect of articulation observed with auditory noise

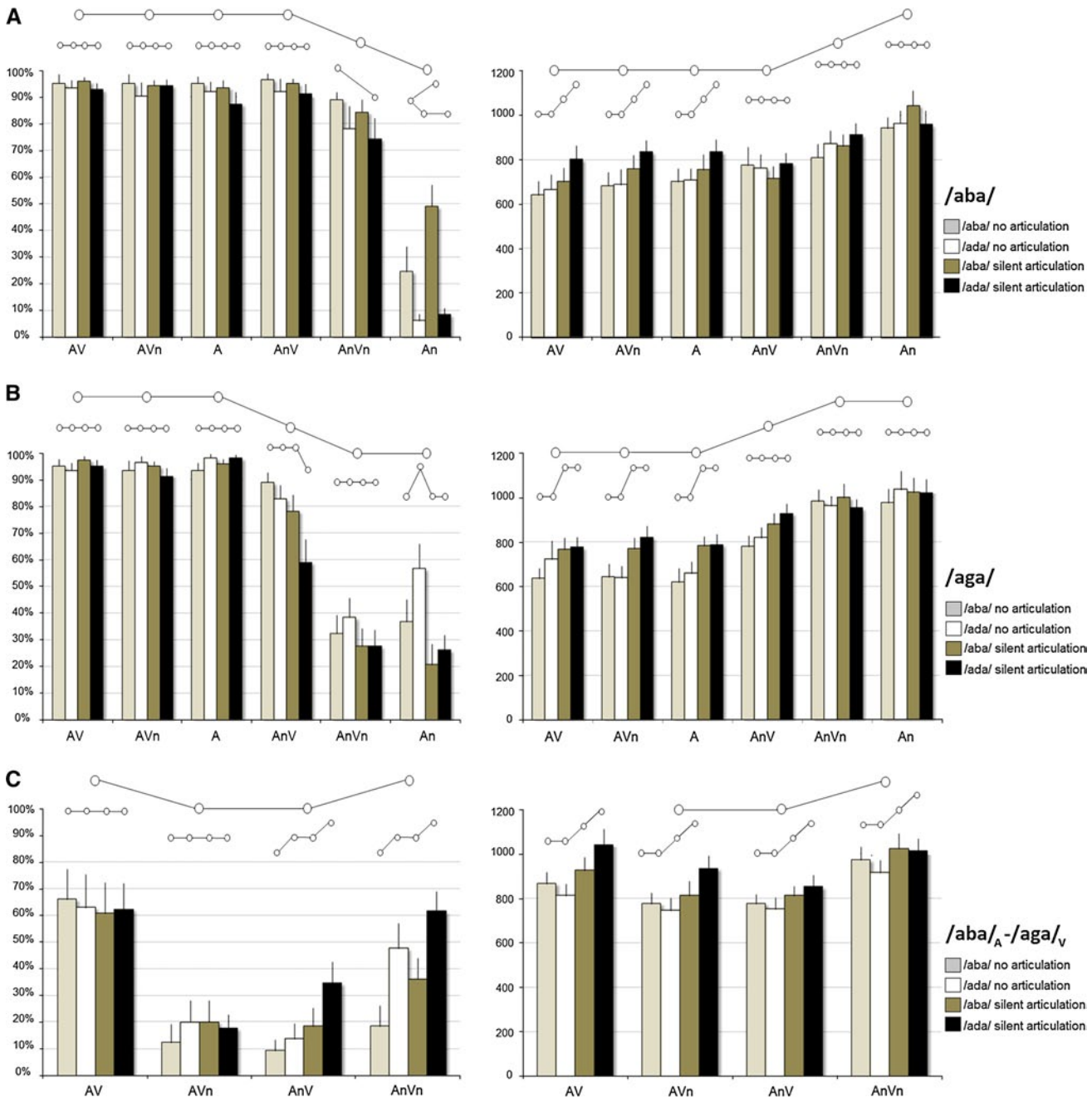



**Fig. 3** **a** Percentage of correct identification (*left*) and median RTs (*right*) for /aba/ stimuli (*y axis*) for each modality of presentation (*x axis*) and articulation condition (see legend). *Error bars* represent standard errors of the mean. Oblique *connection lines* on superimposed graphs (*top*) denote statistical difference related to the modality of presentation (*big circle*) and to the type of articulation (*small circle*). *Flat lines* denote no statistical difference. **b** Percentage of correct identification (*left*) and median RTs (*right*) for /aga/ stimuli (*y axis*) for each modality of presentation (*x axis*) and articulation condition (see legend). *Error bars* represent standard errors of the mean. Oblique *connection lines* on superimposed graphs (*top*) denote statistical difference related to the modality of presentation (*big circle*) and to the type of articulation (*small circle*). *Flat lines* denote no statistical difference. **c** Percentage of correct identification (*left*) and median RTs (*right*) for /aba/$_A$-/aga/$_V$ stimuli (*y axis*) for each modality of presentation (*x axis*) and articulation condition (see legend). *Error bars* represent standard errors of the mean. Oblique *connection lines* on superimposed graphs (*top*) denote statistical difference related to the modality of presentation (*big circle*) and to the type of articulation (*small circle*). *Flat lines* denote no statistical difference

but, when the stimulus was presented without auditory noise, with slower RTs when silently articulating /ada/ than when articulating /aba/ and when articulating /aba/ than when reading /ada/ or /aba/ (on average, mouthing /ada/: 825 ms > mouthing /aba/: 740 ms > reading /ada/: 689 = reading /aba/: 676 ms).

**Table 1** Percentage of responses for /aba/ stimuli for each modality of presentation and each articulation condition

| Response | Visual target | Articulation condition | AV (%) | $AV_n$ (%) | A (%) | $A_nV$ (%) | $A_nV_n$ (%) | $A_n$ (%) |
|---|---|---|---|---|---|---|---|---|
| /aba/ | /aba/ | No articulation | 95 | 95 | 95 | 97 | 89 | 25 |
| | /ada/ | No articulation | 94 | 91 | 92 | 92 | 78 | 6 |
| | /aba/ | Silent articulation | 96 | 95 | 94 | 95 | 85 | 49 |
| | /ada/ | Silent articulation | 93 | 95 | 88 | 92 | 75 | 8 |
| /ada/ | /aba/ | No articulation | 2 | 3 | 2 | 2 | 2 | 23 |
| | /ada/ | No articulation | 0 | 8 | 5 | 5 | 12 | 38 |
| | /aba/ | Silent articulation | 1 | 2 | 2 | 1 | 5 | 28 |
| | /ada/ | Silent articulation | 2 | 1 | 8 | 3 | 20 | 58 |
| /aga/ | /aba/ | No articulation | 0 | 0 | 0 | 0 | 5 | 46 |
| | /ada/ | No articulation | 2 | 0 | 2 | 0 | 3 | 45 |
| | /aba/ | Silent articulation | 1 | 1 | 1 | 1 | 2 | 16 |
| | /ada/ | Silent articulation | 3 | 1 | 0 | 2 | 2 | 27 |

**Table 2** Percentage of responses for /aga/ stimuli for each modality of presentation and each articulation condition

| Response | Visual target | Articulation condition | AV (%) | $AV_n$ (%) | A (%) | $A_nV$ (%) | $A_nV_n$ (%) | $A_n$ (%) |
|---|---|---|---|---|---|---|---|---|
| /aba/ | /aba/ | No articulation | 0 | 0 | 2 | 2 | 31 | 32 |
| | /ada/ | No articulation | 0 | 2 | 0 | 0 | 9 | 5 |
| | /aba/ | Silent articulation | 0 | 2 | 1 | 5 | 34 | 49 |
| | /ada/ | Silent articulation | 1 | 3 | 0 | 5 | 9 | 17 |
| /ada/ | /aba/ | No articulation | 0 | 0 | 2 | 6 | 31 | 25 |
| | /ada/ | No articulation | 2 | 0 | 0 | 9 | 46 | 34 |
| | /aba/ | Silent articulation | 2 | 0 | 0 | 11 | 35 | 26 |
| | /ada/ | Silent articulation | 1 | 2 | 1 | 34 | 56 | 50 |
| /aga/ | /aba/ | No articulation | 95 | 94 | 94 | 89 | 32 | 37 |
| | /ada/ | No articulation | 94 | 97 | 98 | 83 | 38 | 57 |
| | /aba/ | Silent articulation | 98 | 95 | 96 | 78 | 28 | 21 |
| | /ada/ | Silent articulation | 95 | 92 | 98 | 59 | 28 | 26 |

**Table 3** Percentage of responses for $/aba/_A$-$/aga/_V$ stimuli for each modality of presentation and each articulation condition

| Response | Visual target | Articulation condition | AV (%) | $AV_n$ (%) | $A_nV$ (%) | $A_nV_n$ (%) |
|---|---|---|---|---|---|---|
| /aba/ | /aba/ | No articulation | 28 | 78 | 0 | 25 |
| | /ada/ | No articulation | 25 | 77 | 3 | 11 |
| | /aba/ | Silent articulation | 26 | 77 | 3 | 36 |
| | /ada/ | Silent articulation | 24 | 76 | 3 | 19 |
| /ada/ | /aba/ | No articulation | 66 | 12 | 9 | 18 |
| | /ada/ | No articulation | 63 | 20 | 14 | 48 |
| | /aba/ | Silent articulation | 61 | 20 | 18 | 36 |
| | /ada/ | Silent articulation | 62 | 18 | 35 | 62 |
| /aga/ | /aba/ | No articulation | 0 | 2 | 88 | 51 |
| | /ada/ | No articulation | 12 | 2 | 82 | 38 |
| | /aba/ | Silent articulation | 8 | 0 | 76 | 25 |
| | /ada/ | Silent articulation | 8 | 2 | 58 | 15 |

Perception of /aga/—identification scores

For the perception of /aba/, fewer correct responses were observed in the presence of acoustic noise ($A_n$) and in the visually degraded condition ($A_nV_n$). The main effects of auditory ($F(1,12) = 153.8$, $p < .001$) and visual ($F(2,24) = 38.5$, $p < .001$) presentations were significant. There were more correct responses when the stimulus was presented without than with acoustic noise (on average, 96 vs. 48 %), and when it was presented without spatial quantization than with spatial quantization or with the static face (on average, 87 vs. 63 and 66 %). The interaction between these two variables was also significant ($F(2,24) = 40.8$, $p < .001$). Post hoc analyses showed that there were fewer correct responses in $A_n$ and $A_nV_n$ conditions than in $A_nV$ condition, and in $A_nV$ condition than in A, $AV_n$ and AV conditions (on average, $A_n$: 35 % = $A_nV_n$: 32 % < $A_nV$: 77 % < $AV_n$: 94 % = AV: 96 % = A: 97 %).

The main effect of articulation was significant ($F(3,36) = 5.7$, $p < .005$), with more correct responses when reading /ada/ than when mouthing /aba/ or /aba/ (on average, reading /ada/: 76 % > mouthing /aba/: 69 % = mouthing /ada/: 66 %). The articulation × auditory presentation interaction was also significant ($F(3,36) = 6.5$, $p < .005$), with no effect of articulation observed without auditory noise but, when the stimulus was presented with auditory noise, significantly more correct responses when reading /ada/ or /aba/ than when silently articulating /aba/ or /ada/ (on average, reading /ada/: 59 % = reading /aba/: 53 % > mouthing /aba/: 42 % = mouthing /ada/: 38 %). The articulation × visual presentation interaction was also significant ($F(6,72) = 3.2$, $p < .01$), with no effect of articulation observed with spatial quantization. Without quantization, significantly fewer correct responses were observed when silently articulating /ada/ than when reading /aba/ or /ada/ or when silently articulating /aba/ (on average, mouthing /ada/: 77 % < reading /ada/: 88 % = mouthing /aba/: 88 % = reading /aba/: 92 %). With static face, fewer correct responses were observed when reading /aba/ or when silently articulating /aba/ or /ada/ than when reading /ada/ (on average, reading /aba/: 65 % = mouthing /ada/: 62 % = mouthing /aba/: 58 % < reading /ada/: 78 %).

Finally, the three-way interaction was also significant ($F(6,72) = 2.2$, $p < .05$) with no effect of articulation in AV, $AV_n$, A and $A_nV_n$ presentations. In $A_nV_n$ condition, there were fewer correct responses when mouthing /ada/ compared to all other conditions (on average, mouthing /ada/: 59 % < mouthing /aba/: 78 % = reading /ada/: 83 % = reading /aba/: 89 %). In the $A_n$ condition, there were more correct responses when reading /ada/ compared to all other conditions (on average, reading /ada/: 57 % > reading /aba/: 37 % = mouthing /ada/: 26 % = mouthing /aba/: 21 %).

Perception of /aga/—reaction times

The main effects of auditory ($F(1,12) = 41.6$, $p < .001$) and visual presentations were significant ($F(2,24) = 4.9$, $p < .05$), with RTs faster without than with acoustic noise (on average, 719 vs. 947 ms), and without spatial quantization than with spatial quantization or with the static face (on average, 789 vs. 847, 864 ms). The interaction between these two variables was also significant ($F(2,24) = 5.7$, $p < .05$) with slower responses in $A_n$ and $A_nV_n$ conditions than in $A_nV$ condition and in $A_nV$ condition than in A, $AV_n$ and AV conditions (on average, $A_n$: 1,015 ms = $A_nV_n$: 975 ms > $A_nV$: 852 ms > AV: 726 ms = $AV_n$: 719 ms = A: 712 ms).

The main effect of articulation was significant ($F(3,36) = 10.4$, $p < .001$), with slower RTs when silently articulating /ada/ or /aba/ than when reading /ada/ or /aba/ (on average, mouthing /ada/: 881 ms = mouthing/aba/: 871 ms > reading /ada/: 808 ms = reading /aba/: 773 ms). Finally, the articulation × auditory presentation interaction was also significant ($F(3,36) = 4.5$, $p < .01$), with no effect of articulation observed with auditory noise but, without auditory noise, slower RTs when silently articulating /ada/ or /aba/ than when reading /ada/ or /aba/ (on average, mouthing /ada/ 795 ms = mouthing /aba/: 774 ms > reading /ada/: 674 ms = reading /aba/: 633 ms).

Perception of /aba/$_A$-/aga/$_V$—identification scores

The interaction between the auditory and visual presentation was significant ($F(1,12) = 29.0$, $p < .001$) with more /ada/ responses in iAV than in $iA_nV_n$ conditions, and more /ada/ responses in $iA_nV_n$ than in $iA_nV$ and $iAV_n$ conditions (on average, iAV: 63 % > $iA_nV_n$: 41 % > $iAV_n$:18 % = $iA_nV$: 19 %).

The main effect of articulation was significant ($F(3,36) = 5.8$, $p < .005$), with more /ada/ responses when silently articulating /ada/ than when reading /aba/ (on average, 44 vs. 27 %). The articulation × auditory presentation interaction was also significant ($F(3,36) = 5.9$, $p < .005$), with no effect of articulation observed without auditory noise but, with auditory noise, significantly more /ada/ responses when silently articulating /ada/ than when reading /ada/ or when silently articulating /aba/, and when reading /ada/ or when silently articulating /aba/ than when reading/aba/(on average, mouthing /ada/: 48 % > reading /ada/: 31 % = mouthing /aba/: 27 % > reading /aba/: 14 %).

Perception of /aba/A-/aga/V—reaction times

The interaction between the auditory and visual presentation was significant ($F(1,12) = 19.2$, $p < .001$) with slower

responses in $iA_nV_n$ than in $iAV_n$ and $iA_nV$ conditions (on average, $iA_nV_n$: 983 ms > $iAV_n$: 818 ms = $iA_nV$: 799 ms).

The main effect of articulation was also significant ($F(3,36) = 12.02$, $p < .001$), with slower responses when silently articulating /ada/ than when articulating /aba/ and when articulating /aba/ than when reading /aba/ or /ada/ (on average, mouthing /ada/: 961 ms > mouthing /aba/: 894 ms > reading /aba/: 850 ms = reading /ada/: 808 ms).

## Discussion

In a previous study, Sams et al. (2005) showed that the identification of an acoustically presented syllable, embedded in continuous white noise, is enhanced or, conversely, reduced when listeners mimic a similar/different syllable synchronously with their perceiving. In the present study, these findings were fully confirmed and extended to audiovisual syllables. Furthermore, mouthing /ada/ syllable biased the perception of /aba/$_A$-/aga/$_V$ stimuli toward this illusory syllable. Our results also demonstrate that, even in the case of perfect perceptual identification, concurrent mouthing of a syllable speeds up the processing of a concordant speech stimulus, or of that presented acoustically in case of incongruent audiovisual stimuli. Taken together, these results thus appear in line with the two hypotheses that first mouthing a concordant syllable improves and speeds up the identification of the auditory or audiovisual presented syllables and, second, regarding the incongruent audiovisual syllable, mouthing a syllable concordant with the McGurk percept biases and speeds up the perception toward this illusory syllable. The observed modulation of perception is likely to reflect multisensory integration mechanisms, with speech perception mediated by both auditory and visual speech signal, and appears in line with internal forward models and the existence of internally generated motor-to-sensory predictions in silent production. From this view, sensory predictions of the speech motor commands would be sent in parallel to the auditory and somatosensory cortices which, in turn, would influence in a phoneme-specific manner the auditory perception of the acoustic stimulus. Alternate explanations of the present results based on multisensory-motor interactions and a functional coupling between speech perception and production systems will also be discussed.

### Effects of the auditory and visual masking

Irrespective of the articulation conditions, the present study replicates a number of well-known effects in auditory and audiovisual speech perception. As expected, both RTs increased and accuracy decreased in the presence of acoustic masking noise. Furthermore, in accordance with the principle of inverse effectiveness (i.e., multisensory enhancement is greatest when unimodal stimuli are least effective), concordant visual information improved speech intelligibility in noisy auditory conditions (Sumby and Pollack 1954; MacLeod and Summerfield 1987; Benoît et al. 1994). For both dependent measures, this improvement was also observed when facial information was spatially degraded. This result indicates that visual information processing depends on the level of visual specificity of the presented syllable (Campbell and Massaro 1997; MacDonald et al. 2000). Finally, incongruent visual speech information produced a significant number of /ada/ illusory percepts (McGurk and MacDonald 1976; Cathiard et al. 2001), notably in the AV and $A_nV_n$ conditions. Complementing previous findings (Klucharev et al. 2003), RTs were also found to be slower in these conditions, as compared to $AV_n$ and $A_nV$ conditions in which subjects' identification primarily relied on one modality, with /aba/ and /aga/ responses being mostly reported, respectively.

### Effects of the written cue syllable

Of more interest are the effects of various articulation conditions. It is first important to dissociate the influence of the written cue syllable to that of silent articulation on perception. Indeed, as previously mentioned, a written cue presented together with an acoustic stimulus can bias perception (Fowler and Dekle 1991; Sams et al. 2005), with subjects apparently using all possible cues to constrain the number of response alternatives. According to the percentage of responses, there was no "priming" effect of the written syllable when the stimuli were presented without acoustic masking noise (i.e., in the AV, $AV_n$ and A conditions). Indeed, without acoustic masking noise, no differences between articulation conditions were observed for /aba/, /aga/ and /aba/$_A$-/aga/$_V$ stimuli. On the other hand, there were clear effects of the written syllable in the $A_n$ condition for /aba/ and /aga/ stimuli in the absence of any articulation. Significantly more correct responses were observed for /aba/ when preceded by /aba/ as compared to /ada/ written syllables. More intriguing is the larger number of correct responses for /aga/ stimuli in the case of /ada/ as compared to /aba/ written syllables. Although there is a substantial effect here, we have no clear explanation for it. It should be, however, noted that adding white noise at a −6 dB signal-to-noise ratio to both /aba/ and /aga/ syllables dramatically increased the proportion of /aga/ and /ada/ responses in the auditory-only condition without articulation (see Tables 1, 2, 3). This effect might partly be due to the strong masking of the lowest part of the spectrum where acoustic cues for /aba/ are predominant (given that the upper part of the spectrum and the frequency transitions of the second and third formants

for /aba/ appear closer to that of /aga/ compared to that of /ada/).

Finally, in line with a specific effect of the written syllable, fewer /ada/ responses were reported for /aba/$_A$-/aga/$_V$ stimuli in the case of /aba/ as compared to /ada/ written cue syllables. When considering RTs, whatever the stimuli and the modality of presentation, no differences were observed when the stimuli were preceded by written /aba/ or /aga/ in the absence of silent articulation. Globally, these results thus confirm and extend previous findings (Fowler and Dekle 1991; Sams et al. 2005) by showing that a written syllable might bias accuracy but not RTs of an auditory or incongruent audiovisual syllable when masked with acoustic noise.

### Effects of the silent articulation

Despite the influence of the written syllables, silent articulation specifically biased speech perception. Considering the perceptual scores, a larger number of correct responses were observed for /aba/ stimuli in the $A_n$ condition when subjects mouthed /aba/ compared to all other articulation conditions. In the $A_nV_n$ condition, more errors were also reported during silent articulation of /ada/ than in the two conditions in which the written /aba/ syllable appeared (with or without subsequent articulation). As previously noted, the results for /aga/ stimuli are less easily interpretable, with more errors in the case of silent articulation of /ada/ in the $A_nV$ condition. Finally, silent articulation of /ada/ produced a stronger McGurk effect for /aba/$_A$-/aga/$_V$ stimuli in the presence of an acoustic masking noise compared to the other articulation conditions. This is of particular interest because the produced syllable differs acoustically and visually from the illusory syllable. Globally, these results demonstrate that the influence of the written cue and the silent articulation on perception is clearly dissociable. Hence, for identification scores, these results confirm and extend the results by Sams et al. (2005) and support our two hypotheses that mouthing a concordant syllable improves the identification of the auditory or audiovisual syllables and mouthing an /ada/ syllable biases the perception of /aba/$_A$-/aga/$_V$ stimuli toward this illusory syllable.

However, subjects had to keep the to-be-articulated syllable in memory until/during the stimulus presentation in the two silent articulation conditions. One possibility is that working memory might strengthen the "priming" effect of the written syllable. Therefore, the observed perceptual modulation might not be due to motor-to-sensory predictions and feedback control mechanisms but rather to short-term memory processes related to the written cue. However, while we cannot reject a possible priming effect of the written cue on perception, the specificity of RT modulation for /aba/ and /aga/ stimuli when presented without acoustic masking noise argues against this interpretation. Indeed, faster responses were observed for /aba/ stimuli when subject silently pronounced /aba/, as compared to /ada/, in the AV, $AV_n$ and A conditions, but not in the $A_nV$ condition despite similar global RTs in these presentation modalities. With respect to discordant articulation, this supports our first hypothesis that mouthing a concordant syllable speeds the identification of the auditory or audiovisual syllables. Furthermore, silent articulation of /aba/ and /ada/ did not produce RT differences for /aga/ stimuli. These results suggest that, despite perfect identification, silent articulation speeds the auditory and audiovisual identification of a similar presented syllable compared to a dissimilar one. It is also worthwhile noting that the absence of RT differences for /aga/ stimuli between the two pronounced syllables reinforces this interpretation by eliminating the possibility that this effect was due to variations of the speed of articulation between these syllables. Regardless of the written syllable and for both /aba/ and /aga/ stimuli, RTs were faster in the absence of silent articulation possibly due to additional cognitive effort associated with the concurrent articulation task. Finally, no difference in RTs was found for /aba/ and /aga/ stimuli compared to the different articulation conditions in the presence of masking noise (i.e., in the $A_nV$, $A_nV_n$ and $A_n$ conditions). This can be attributed to slow RTs due to masking noise, eliminating any difference between the articulation conditions. Regarding /aba/$_A$-/aga/$_V$ stimuli, faster RTs were observed when subjects articulated /aba/ as compared to /ada/. Importantly, this effect appears regardless of the reported syllable: /ada/ being mostly reported in the AV and $A_nV_n$ conditions but not in the $AV_n$ and $A_nV$ conditions in which more /aba/ and /aga/ responses were observed. Hence, although for incongruent audiovisual stimuli silent articulation of a similar syllable to that presented acoustically speeds perceptual processing, our results do not support our second hypothesis of shorter identification time when silently articulating a syllable congruent with the illusionary percept (i.e., /ada/). A possible explanation is that the final perceptual categorization of the conflicting auditory and visual inputs into a unitary illusionary percept arises in a late phonetic decision stage (Massaro 1998; see also Sato et al. 2010), therefore minimizing the effect of the silently produced syllable onto the late illusionary percept.

### Multisensory-motor interactions during speech perception and internally generated sensory predictions during silent speech production

As previously noted, the observed modulation of perception is likely to reflect multisensory integration mechanisms, with speech perception mediated by both auditory and visual speech signals. It also suggests the existence of

internal motor-to-sensory (auditory and/or somatosensory) predictions in silent production. However, it is difficult to dissociate the effects of action on perception from the effects of perception on action. Indeed, recent behavioral and neurophysiological studies suggest that speech perception is partly mediated by reference to the motor actions afforded in the speech signal. For instance, seeing a video of an articulating mouth influences the production of similar or dissimilar articulations (Kerzel and Bekkering 2000; Gentilucci and Cattaneo 2005). Recent brain imaging and neurophysiological studies also provide supporting evidence for sensorimotor interactions during speech perception. Research on mirror neurons in the macaque, and on a putative mirror neuron system in humans, has demonstrated that action observation and listening to action-related sounds partly involve the same neural circuits than those that are used in action performance (for a review, see Rizzolatti and Craighero 2004). Consistent with the view that sensorimotor interactions might play a fundamental role in speech processing (Schwartz et al. 2002, 2012), brain areas involved in the planning and execution of speech gestures have been found to be activated in processing speech sounds, beyond the involvement of superior temporal auditory regions (Fadiga et al. 2002; Watkins et al. 2003; Wilson et al. 2004; Pulvermuller et al. 2006; Skipper et al. 2007; Callan et al. 2010; Sato et al. 2010; Tremblay and Small 2011; Alho et al. 2012; Grabski et al. in press). From these results, it is therefore possible that motor representations activated by the presented speech stimulus also interfered with those activated by the silent articulation task. Although we cannot reject this possibility, it is to note that the observed specificity of RT modulation does not appear fully in line with previous studies that demonstrate sensorimotor interactions during speech perception. First, while previous brain imaging studies have shown that visual and audiovisual speech perception involve brain regions associated with speech production to a greater extent than auditory-only speech stimuli (Skipper et al. 2005, 2007), silent articulation of /aba/ compared to /ada/ speeded /aba/ identification in the A, $AV_n$ and AV conditions. Similarly, the fact that silent articulation speeded the perceptual processing based on acoustic rather than visual information seems also to minimize possible sensory-to-motor effects, notably for /aba/$_A$-/aga/$_V$ stimuli where faster RTs were observed when subjects articulated /aba/ as compared to /ada/ syllable irrespective of the reported syllable. Finally, previous studies provide evidence that the motor system is strongly recruited in speech perception under adverse conditions in order to resolve phonetic ambiguity (Binder et al. 2004; Zekveld et al. 2006; Sato et al. 2009; d'Ausilio et al. 2011), while, in the present study, no RT differences were observed across the articulation conditions when the auditory and congruent audiovisual syllables were acoustically and/or visually degraded.

In conclusion, this study demonstrates that our own silent articulation may alter speech perception. Our results suggest multisensory-motor interactions during speech perception and provide support for the existence of internal motor-to-sensory predictions in order to anticipate the consequence of the action. Adding to a large number of behavioral and neurophysiological studies, the present results support a functional coupling between sensory and motor systems in speech production.

**Acknowledgments** We thank Jérome Aubin for his help in this study.

## References

Abbs JH, Gracco VL (1984) Control of complex motor gestures: orofacial muscle responses to load perturbations of lip during speech. J Neurophysiol 51:705–723

Alho J, Sato M, Sams M, Schwartz J-L, Tiitinen H, Jääskeläinen IP (2012) Enhanced early-latency electromagnetic activity in the left premotor cortex is associated with successful phonetic categorization. Neuroimage 60:1937–1946

Benoît C, Mohamadi T, Kandel SD (1994) Effects on phonetic context on audio-visual intelligibility of French. J Speech Lang Hear Res 37:1195–1203

Binder JR, Liebenthal E, Possing ET, Medler DA, DouglasWard B (2004) Neural correlates of sensory and decision processes in auditory object identification. Nat Neurosci 7:295–301

Burnett TA, Freedland MB, Larson CR, Hain TC (1998) Voice F0 responses to manipulations in pitch feedback. J Acoust Soc Am 103:3153–3161

Callan D, Callan A, Gamez M, Sato MA, Kawato M (2010) Premotor cortex mediates perceptual performance. Neuroimage 51:844–858

Campbell CS, Massaro DW (1997) Perception of visible speech: influence of spatial quantization. Perception 26:627–644

Cathiard M-A, Schwartz J-L, Abry C (2001) Asking a naive question to the McGurk effect: why does audio [b] give more [d] percepts with usual [g] than with visual [d]? In: Proceedings of the auditory visual speech processing, AVSP'2001. Aalborg, Copenhague, pp 138–142

Christoffels IK, Formisano E, Schiller NO (2007) Neural correlates of verbal feedback processing: an fMRI study employing overt speech. Hum Brain Mapp 28:868–879

Christoffels IK, van de Ven V, Waldorp LJ, Formisano E, Schiller NO (2011) The sensory consequences of speaking: parametric neural cancellation during speech in auditory cortex. PLoS One 6:e18307

Curio G, Neuloh G, Numminen J, Jousmaki V, Hari R (2000) Speaking modifies voice-evoked activity in the human auditory cortex. Hum Brain Mapp 9:183–191

d'Ausilio A, Bufalari I, Salmas P, Fadiga L (2011) The role of the motor system in discriminating degraded speech sounds. Cortex 48:882–887

Elman JL (1981) Effects of frequency—shifted feedback on the pitch of vocal productions. J Acoust Soc Am 70:45–50

Fadiga L, Craighero L, Buccino G, Rizzolatti G (2002) Speech listening specifically modulates the excitability of tongue muscles: a TMS study. Eur J Neurosci 15:399–402

Feng Y, Gracco VL, Max L (2011) Integration of auditory and somatosensory error signals in the neural control of speech movements. J Neurophysiol 106:667–679

Folkins JW, Abbs JH (1975) Lip and jaw motor control during speech: responses to resistive loading of the jaw. J Speech Hear Res 18:207–219

Ford JM, Mathalon DH (2004) Electrophysiological evidence of corollary discharge dysfunction in schizo-phrenia during talking and thinking. J Psychiatr Res 38:37–46

Ford JM, Mathalon DH, Heinks T, Kalba S, Faustman WO, Roth WT (2001) Neurophysiological evidence of corollary discharge dysfunction in schizophrenia. Am J Psychiatry 158:2069–2071

Fowler CA, Dekle DJ (1991) Listening with eye and hand: cross-modal contributions to speech perception. J Exp Psychol Hum Percept Perform 17:816–828

Francis BA, Wonham WM (1976) The internal model principle of control theory. Automatica 12:457–651

Gentilucci M, Cattaneo L (2005) Automatic audiovisual integration in speech perception. Exp Brain Res 167:66–75

Golfinopoulos E, Tourville JA, Bohland JW, Ghosh SS, Nieto-Castanon A, Guenther FH (2011) fMRI investigation of unexpected somatosensory feedback perturbation during speech. Neuroimage 55:1324–1338

Grabski K, Schwartz JL, Lamalle L, Vilain C, Vallée N, Baciu M, Le Bas JF, Sato M Shared and distinct neural correlates of vowel perception and production. J Neurolinguistics (in press)

Gracco VL, Abbs JH (1985) Dynamic control of the perioral system during speech: kinematic analyses of autogenic and nonautogenic sensorimotor processes. J Neurophysiol 54:418–432

Guenther FH (1995) Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. Psychol Rev 102:594–621

Guenther FH (2006) Cortical interactions underlying the production of speech sounds. J Commun Disord 39:350–365

Guenther FH, Vladusich T (2012) A neural theory of speech acquisition and production. J Neurolinguistics 25:408–422

Guenther FH, Hampson M, Johnson D (1998) A theoretical investigation of reference frames for the planning of speech movements. Psychol Rev 105:611–633

Hashimoto Y, Sakai KL (2003) Brain activations during conscious self-monitoring of speech production with delayed auditory feedback: an fMRI study. Hum Brain Mapp 20:22–28

Heinks-Maldonado TH, Nagarajan SS, Houde JF (2006) Magnetoencephalographic evidence for a precise forward model in speech production. Neuroreport 17:1375–1379

Hickok G (2012) Computational neuroanatomy of speech production. Nat Rev Neurosci 13:135–145

Hickok G, Buchsbaum B, Humphries C, Muftuler T (2003) Auditory-motor interaction revealed by fMRI: speech, music, and working memory in area Spt. J Cogn Neurosci 15:673–682

Hickok G, Houde JF, Rong F (2011) Sensorimotor integration in speech processing: computational basis and neural organization. Neuron 69:407–422

Hirano S, Kojima H, Naito Y, Honjo I, Kamoto Y, Okazawa H, Konishi J (1996) Cortical speech processing mechanisms while vocalizing visually presented languages. Neuroreport 8:363–367

Hirano S, Kojima H, Naito Y, Honjo I, Kamoto Y, Okazawa H, Konishi J (1997) Cortical processing mechanism for vocalization with auditory verbal feedback. Neuroreport 8:2379–2382

Houde JF, Jordan MI (1998) Sensorimotor adaptation in speech production. Science 279(5354):1213–1216

Houde JF, Nagarajan SS, Sekihara K, Merzenich MM (2002) Modulation of the auditory cortex during speech: an MEG study. J Cogn Neurosci 14:1125–1138

Jeannerod M (1994) The representing brain. Neural correlates of motor intention and imagery. Behav Brain Sci 17:187–245

Jeannerod M (2001) Neural simulation of action: a unifying mechanism for motor cognition. Neuroimage 14:S103–S109

Jones JA, Munhall KG (2000) Perceptual calibration of F0 production: evidence from feedback perturbation. J Acoust Soc Am 108:1246–1251

Jones JA, Munhall KG (2005) Remapping auditory–motor representations in voice production. Curr Biol 15:1768–1772

Kawato M, Furukawa K, Suzuki R (1987) A hierarchical neural network model for the control and learning of voluntary movements. Biol Cybern 56:1–17

Kerzel D, Bekkering H (2000) Motor activation from visible speech: evidence from stimulus response compatibility. J Exp Psychol Hum Percept Perform 26:634–647

Klucharev V, Möttönen R, Sams M (2003) Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. Cogn Brain Res 18:65–75

Lamettti DR, Nasir SM, Ostry DJ (2012) Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. J Neurosci 32:9351–9358

Levelt WJM (1989) Speaking: from intention to articulation. The MIT Press, Cambridge, MA

MacDonald J, Andersen S, Bachmann T (2000) Hearing by eye: how much spatial degradation can be tolerated? Perception 29: 1155–1168

MacLeod A, Summerfield Q (1987) Quantifying the contribution of vision to speech perception in noise. Br J Audiol 21:131–141

Massaro DW (1998) Perceiving talking faces: from speech perception to a behavioral principle. The MIT Press, Cambridge, MA

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

Nasir SM, Ostry DJ (2006) Somatosensory precision in speech production. Curr Biol 16:1918–1923

Numminen J, Curio G (1999) Differential effects of overt, covert and replayed speech on vowel-evoked responses of the human auditory cortex. Neurosci Lett 272:29–32

Numminen J, Salmelin R, Hari R (2000) Subject's own speech reduces reactivity of the human auditory cortex. Neurosci Lett 265:119–122

Oldfield RC (1971) The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9:97–114

Perkell JS (2012) Movement goals and feedback and feed forward control mechanisms in speech production. J Neurolinguistics 25:382–407

Perkell JS, Matthies ML, Lane H, Guenther FH, Wilhelms-Tricarico R, Wozniak J et al (1997) Speech motor control: acoustic goals, saturation effects, auditory feedback & internal models. Speech Commun 22:227–250

Perkell JS, Guenther FH, Lane H, Matthies LM, Perrier P, Vick J, Wilhelms-Tricarico R, Zandipour M (2000) A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. J Phon 28:233–272

Perrier P (2005) Control and representations in speech production. ZAS Pap Linguist 40:109–132

Perrier P (2012) Gesture planning integrating knowledge of the motor plant's dynamics: a literature review from motor control and speech motor control. In: Speech planning and dynamics, collection: speech production and perception

Price CJ, Crinion JT, MacSweeney M (2011) A generative model of speech production in Broca's and Wernicke's areas. Front Psychol 2:237

Pulvermuller F, Huss M, Kherif F, del Prado Martin FM, Hauk O, Shtyrov Y (2006) Motor cortex maps articulatory features of speech sounds. Proc Natl Acad Sci USA 103:7865–7870

Purcell DW, Munhall KG (2006a) Compensation following real-time manipulation of formants in isolated vowels. J Acoust Soc Am 119:2288–2297

Purcell DW, Munhall KG (2006b) Adaptive control of vowel formant frequency: evidence from real-time formant manipulation. J Acoust Soc Am 120:966–977

Rizzolatti G, Craighero L (2004) The mirror-neuron system. Annu Rev Neurosci 27:169–192

Sams M, Möttönen R, Sihvonen T (2005) Seeing and hearing others and oneself talk. Brain Res Cogn Brain Res 23:429–435

Sato M, Tremblay P, Gracco V (2009) A mediating role of the premotor cortex in phoneme segmentation. Brain Lang 111:1–7

Sato M, Buccino G, Gentilucci M, Cattaneo L (2010) On the tip of the tongue: modulation of the primary motor cortex during audiovisual speech perception. Speech Commun 52:533–541

Schwartz JL, Abry C, Boë LJ, Cathiard MA (2002) Phonology in a theory of perception-for-action-control. In: Durand J, Lacks B (eds) Phonology: from phonetics to cognition. Oxford University Press, Oxford, pp 240–280

Schwartz JL, Ménard L, Basirat A, Sato M (2012) The perception for action control theory (PACT): a perceptuo-motor theory of speech perception. J Neurolinguistics 25:336–354

Shergill SS, Brammer MJ, Fukuda R, Bullmore E, Amaro E Jr, Murray RM, McGuire PK (2002) Modulation of activity in temporal cortex during generation of inner speech. Hum Brain Mapp 16:219–227

Shiller DM, Sato M, Gracco VL, Baum S (2009) Perceptual recalibration of speech sounds following speech motor learning. J Acoust Soc Am 125:1103–1113

Shiller DM, Gracco VL, Rvachew S (2010) Auditory-motor learning during speech production in 9–11 year-old children. PLoS One 5:e12975

Shuster LI, Lemieux SK (2005) An fMRI investigation of covertly and overtly produced mono- and multisyllabic words. Brain Lang 93:20–31

Skipper JI, Nusbaum HC, Small SL (2005) Listening to talking faces: motor cortical activation during speech perception. Neuroimage 25:76–89

Skipper JI, Van Wassenhove V, Nusbaum HC, Small SL (2007) Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. Cereb Cortex 17:2387–2399

Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26:212–215

Tian X, Poeppel D (2010) Mental imagery of speech and movement implicates the dynamics of internal forward models. Front Psychol 1:166

Tourville JA, Reilly KJ, Guenther FH (2008) Neural mechanisms underlying auditory feedback control of speech. Neuroimage 39:1429–1443

Tremblay P, Small SL (2011) On the context-dependent nature of the contribution of the ventral premotor cortex to speech perception. Neuroimage 57:1561–1571

Tremblay S, Shiller DM, Ostry DJ (2003) Somatosensory basis of speech production. Nature 423:866–869

Ventura MI, Nagarajan SS, Houde JF (2009) Speech target modulates speaking induced suppression in auditory cortex. BMC Neurosci 10:58

von Holst E, Mittelstaedt H (1950) Das Reafferenzprinzip. Wechselwirkungen zwischen Zentralnervensystem und Peripherie. Naturwissenchaften 37:464–476

Watkins KE, Strafelle AP, Paus T (2003) Seeing and hearing speech excites the motor system involved in speech production. Neuropsychologia 41:989–994

Wilson SM, Saygin AP, Sereno MI, Iacoboni M (2004) Listening to speech activates motor areas involved in speech production. Nat Neurosci 7:701–702

Wise RJ, Scott SK, Blank SC, Mummery CJ, Murphy K, Warburton EA (2001) Separate neural subsystems within "Wernicke's area". Brain 124:83–95

Yetkin FZ, Hammeke TA, Swanson SJ, Morris GL, Mueller WM, McAuliffe TL, Haughton VM (1995) A comparison of functional MR activation patterns during silent and audible language tasks. Am J Neuroradiol 16:1087–1092

Zekveld AA, Heslenfeld DJ, Festen JM, Schoonhoven R (2006) Top–down and bottom–up processes in speech comprehension. Neuroimage 32:1826–1836