Neuropsychologia 63 (2014) 85–91

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

Vision of tongue movements bias auditory speech perception

Alessandro D'Ausilio [a,*], Eleonora Bartoli [a], Laura Maffongelli [a], Jeffrey James Berry [a], Luciano Fadiga [a,b]

[a] Robotics, Brain and Cognitive Sciences Department, The Italian Institute of Technology, Via Morego, 30, 16163 Genova, Italy
[b] Section of Human Physiology, University of Ferrara, Via Fossato di Mortara, 17/19, 44100 Ferrara, Italy

ARTICLE INFO

Article history:
Received 20 May 2014
Received in revised form
15 July 2014
Accepted 15 August 2014
Available online 27 August 2014

Keywords:
Speech
Audiovisual speech perception
McGurk
Sensorimotor map
Motor theories
Action perception

ABSTRACT

Audiovisual speech perception is likely based on the association between auditory and visual information into stable audiovisual maps. Conflicting audiovisual inputs generate perceptual illusions such as the McGurk effect. Audiovisual mismatch effects could be either driven by the detection of violations in the standard audiovisual statistics or via the sensorimotor reconstruction of the distal articulatory event that generated the audiovisual ambiguity. In order to disambiguate between the two hypotheses we exploit the fact that the tongue is hidden to vision. For this reason, tongue movement encoding can solely be learned via speech production but not via others' speech perception alone. Here we asked participants to identify speech sounds while matching or mismatching visual representations of tongue movements which were shown. Vision of congruent tongue movements facilitated auditory speech identification with respect to incongruent trials. This result suggests that direct visual experience of an articulator movement is not necessary for the generation of audiovisual mismatch effects. Furthermore, we suggest that audiovisual integration in speech may benefit from speech production learning.

1. Introduction

Stable and reliable multisensory representations can be achieved by the natural alignment of information, from different modalities, related to the same event. Asynchrony in audio-visual temporal alignment can be detected in a variety of multimodal stimuli (speech, music and object action; Vatakis & Spence, 2006), indicating that we are particularly sensitive to violations in the temporal correlations. Intriguingly, participants also infer causal relationships from temporal correlation between audio and visual events (Parise, Spence, & Ernst, 2012). During every day communication the auditory information produced by a speaker is often temporally coupled with the visual information arising from visible articulators, such as lips. In most cases, building such stable correlations between speech audio–visual signals can aid perception in ecological scenarios. For instance, vision of the articulators enhances accurate auditory perception in noise (Sumby & Pollack, 1954). More generally, visible speech influences perception both by integrating under-specified acoustic information and by making perception more robust through redundancy (Campbell, 2008).

Otherwise, perturbation of the normal spatio-temporal alignment between audio and visual cues can induce illusory percepts. For example, in the ventriloquism effect, when auditory and visual information come from different spatial sources we tend to illusorily displace sounds towards the visual source (Pick, Warren, & Hay, 1969). On the other hand, if auditory (i.e. /ba/) and visual (i.e. /ga/) information do not match, an illusory perception such as the McGurk effect (McGurk & MacDonald, 1976) may arise. In this case, participants perceive a third syllable (/da/ or /tha/). The McGurk illusion is generally seen as a landmark demonstration of how previous learning affects the analysis and integration of multimodal speech stimuli.

Generally speaking, this effect is due to a perturbation of the learned auditory and visual speech-related association. This illusion is quite robust to even large temporal asynchronies (Munhall, Gribble, Sacco, & Ward, 1996) or spatial manipulations (Jones & Munhall, 1997), and is elicited without participants being aware of the task (Alsius & Munhall, 2013). However, one key question since the pioneering work of McGurk and McDonald was how much this effect is a by-product of being exposed to a stable multisensory environment providing repeated and reliable audiovisual correlation. In fact, during development, the repeated co-occurrence and match of audio and visual information was thought to build reliable statistics of the environment. In this sense, different age-spans were investigated (McGurk & MacDonald, 1976; Massaro, 1984) leading to the finding that the effect in children is somewhat weaker than in adults. These initial observations suggested that the McGurk effect was at least partially driven by a form of experience-dependent learning of the audiovisual statistics of the environment.

[*] Corresponding author. Tel.: +39 10 71781975; fax: +39 10 7170817.
E-mail address: alessandro.dausilio@iit.it (A. D'Ausilio).

http://dx.doi.org/10.1016/j.neuropsychologia.2014.08.018
0028-3932/© 2014 Elsevier Ltd. All rights reserved.

In the following years a series of studies showed that infants are indeed affected by the visual cues present during speech perception. Four-month-old infants, show preference for the face that matches an auditory vowel (Kuhl & Meltzoff, 1982; Patterson & Werker, 1999). Similarly, two-month-old infants detect the correspondence between the auditory and visually perceived speech information (Patterson & Werker, 2003). This could be explained by the fact that audiovisual matching could arise from at least three partially independent feature sets, including temporal cues (Vatakis & Spence, 2006), energetic cues (Grant, van Wassenhove, & Poeppel, 2004) and phonetic cues (Kuhl et al., 2006). Infants do not seem to rely on phonetic cues (Baart, Vroomen, Shaw, & Bortfeld, 2014; Jusczyk, Luce, & Charles-Luce, 1994) whereas temporal and spectral ones might be employed for early audiovisual correspondence detection in speech. Nevertheless, all these studies confirm that some form of multimodal matching of audiovisual speech already exist in pre-linguistic children (Rosenblum, Schmuckler, & Johnson, 1997; Burnham & Dodd, 2004), thus suggesting a partial independence with respect to their linguistic environment.

Generally speaking, the literature seems to suggest that some basic form of pre-linguistic audiovisual statistical association can be acquired very early in life. The critical problem concerns the nature of this audiovisual association. Specifically, the question is if active vocal exploration plays some role in the acquisition of these associations or if passive (rather limited) exposure to environmental audiovisual speech statistics is able to provide enough information. Along this line, a recent study used a sort of reversed McGurk effect. Adult participants heard speech sounds and at the same time had to judge the shape of “mouth-like” ellipses (Sweeny, Guzman-Martinez, Ortega, Grabowecky, & Suzuki, 2012). The clever use of ellipsoidal visual stimuli should in theory avoid the automatic and direct association with the visual representations of mouth shapes stored in memory. However, the authors suggest two alternative hypotheses for the biasing effect that auditory syllables had on shape judgments. One is that participants were able to grasp the statistical association that exists between speech sounds and the mouth visual shapes to produce that sound (from now on called audiovisual hypothesis). These audiovisual associations, favored by the perceptual similarities between mouth configurations and ellipse shapes, are substantially analogous to the previously outlined auditory and visual correlation, we are tuned to detect since infancy. Such a hypothesis thus predicts that passive exposure to environmental audiovisual speech statistics can cause the effect.

Alternatively, the effect could be due to the correspondence emerging from the automatic transformation of auditory and visual information, into articulatory movements in the motor system (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; from now on called sensorimotor hypothesis). The general mechanism could be that suggested by the analysis by synthesis approach (Stevens & Halle, 1967). This model proposed that the perception is derived from the computational re-creation of the input (Bever & Poeppel, 2010). Such a synthetic process regenerates the input by means of an abstract motor code without specifying detailed acoustic, visual or motor correlates. Liberman's motor theory, instead, insisted on a more specified motor program. The main difference being that the reconstruction envisioned by the motor theory, implies an internal representation of actual vocal movement. Here the driving factor might be the capability to extract, from abstract visual stimuli such as ellipses, basic sensorimotor primitives learned from active vocal production.

Unfortunately, in this study as well as in many audiovisual integration studies, both accounts are equally probable. No conclusion can be drawn in favor of either the audiovisual or sensorimotor hypotheses (Spence & Deroy, 2012). One possible solution to discriminate between the two hypotheses might instead be the study of adults' behavior on material for which no reliable audiovisual statistics is present. We propose that watching articulators, for which we have no visual experience such as the tongue could be the key aspect.

The tongue is indeed a critical articulator, hardly visible in its full motion and target configurations. During speech perception we at most barely see just the anterior tip of the tongue and thus a rather loose temporal association with the auditory effects of tongue movements. In fact, even if the most anterior part of the tongue can be partly exposed during speech production (if the jaw opening is somewhat exaggerated), the tongue back motion, which is the critical component for the /ga/ or /ka/ syllable (velar constriction), is in contrast always occluded. However, it is still possible that some correlation could be picked up between the posterior motion (inferred) and the anterior tongue information (partly visible). Nevertheless, such a correlation is almost absent since the anterior tip motion, for velar sounds, is much more variable than the tongue back, as for any non-critical articulatory feature (Papcun et al., 1992; Canevari, Badino, Fadiga, & Metta, 2013). During speech production, instead, we exploit tightly coupled proprioceptive, tactile, motor and auditory cues associated with tongue motor control. Therefore, we can get access to accurate tongue kinematics knowledge solely through tongue movement learning. It is important to stress that such sensorimotor knowledge does not necessarily need to be learned via speech production but rather can also emerge from non-speech tongue motor control.

In the present study, we capitalize on the fact that tongue motion is concealed from vision by running two behavioral experiments. More specifically, participants had to identify auditory syllables while we visually presented real tongue movements recorded with an ultrasound imaging technique. Visual stimuli showed a sagittal profile of a tongue producing a syllable that was either matching or mismatching with the auditory stimuli. Our prediction is that if the audiovisual hypothesis is true then we should see no difference between matching and mismatching audiovisual presentations. In fact, there is no statistical audiovisual association between speech sounds and visible tongue shapes. Furthermore, there is no perceptual similarity between visible mouth configurations and tongue movement (as it was the case for Sweeny et al., 2012). On the other hand, if the visual presentation of tongue movements induces a significant bias on auditory perception then, the sensorimotor hypothesis is more likely to be true. In fact, it would demonstrate that in principle, learning the statistics of the audiovisual environment cannot account for such a multimodal effect but rather we need to get access to knowledge that can be acquired solely via tongue movement control.

2. Materials and methods

2.1. General methods

Visual stimuli consisted of short video clips showing the sagittal profile of a tongue (See Fig. 1) articulating different syllables. Since ultrasound images can be very noisy, the tongue dorsal profile was enhanced by drawing a red line on top of it. Video clips were utterances of a female speaker producing /ba/, /ga/, /pa/, and /ka/ syllables. Frame-to-frame differences in pixel intensity (0 for black and 1 for white pixels) were measured to check for a possible bias in global visual motion between stimuli (sum of absolute differences: /ba/=52,105; /ga/=57,020; /pa/=61,185; /ka/=60,588). In fact, the ultrasound probe captures information (i.e. background or tongue body) that is not necessarily conveying information about articulatory gestures. In this sense, we computed whole image statistics about global motion, to control for spurious (non-gestural) movement differences and thus exclude the contribution of low-level global visual feature identification. Paired *t*-tests between anterior and posterior articulated stimuli were not significant (mean motion and standard deviation: /ba/=1914.72 std=694.62; /ga/=2097.03 std=853.9; /pa/=2225.83 std=710.67; /ka/=2220.84 std=782.16;

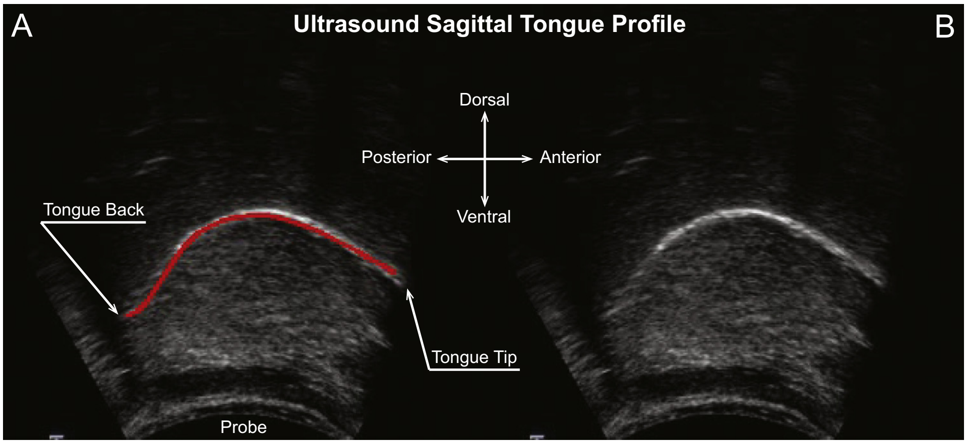

Fig. 1. Stimuli. Panel A and B show two snapshots taken from the video-clips. Pictures show the ultrasound image of the tongue, with and without the superimposed red line on the tongue surface. Participants have to respond only to stimuli with the red line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

t-test: /ba/ Vs. /ga/: $t(27)=-1.138369$, $p=0.265348$; /pa/ Vs. /ka/: $t(27)=0.130398$, $p=0.897255$). We conclude that stimuli do not differ in terms of global visual motion. Stimuli lasted 28 frames and thus had a length of 1120 ms.

An auditory syllable (/ba/, /ga/, /pa/, and /ka/, was mixed with white noise to avoid ceiling effects, SNR ratio: 4.17 dB) was temporally aligned with each video clip. The stimuli (ultrasound data and audio) were extracted from a larger database acquired and synchronized in our lab, for automatic speech recognition research (Castellini et al., 2011). The auditory syllable was either the same syllable as the one depicted in the video clip or a different one, leading to matching and mismatching audiovisual stimuli. Synchronization between audio and visual streams for the mismatching condition was obtained by aligning the vowel onset of the mismatching sound to the already synchronized matching sound. In addition, a neutral video clip with no information regarding tongue movements was obtained by scrambling pixels of the original video clips. Pixels were randomly scrambled in space (each frame separately, taken from all clips), to eliminate both configuration and dynamical information. This procedure reliably deletes spatio-temporal information about the tongue motion, but maintains critical information such as general image motion over time. This information still signals the participant about the length of the visual event, its onset and offset as well as its luminance variations over time. This latter information was necessary to control for unspecific low-level visual attention variations, and thus keeping participants similarly engaged. Based on the possible combinations of visual and auditory stimuli, there were three experimental conditions: 'Congruent' (matching auditory and visual stimuli), 'Incongruent' (mismatching stimuli) and 'Control' (scrambled video of the tongue associated with either auditory stimuli).

2.2. *Methods experiment 1*

2.2.1. *Design*

Stimuli consisted of short video clips showing the sagittal profile of a tongue articulating different syllables (/ba/, /ga/, /pa/, /ka/). At the same time, two of these syllables (/ba/, /pa/) were auditorily presented. In this experiment we intended to measure whether the visual presentation of velar (/ga/ and /ka/) syllables could interfere with the perception of bilabial sounds (/ba/ and /pa/). Therefore, participants' task was to identify sounds according to the consonant of the sound, which differed according to voice onset timing (VOT). It is important to note that here the implicit response dimension (VOT) was orthogonal to the audiovisual manipulation (place of articulation). In this design we thus implicitly ask participants to focus on one articulatory dimension while we present visual stimuli that may or may not match on a different and independent speech feature.

Experimental stimuli consisted of the combination of visual and auditory information into 3 classes: matching audio–video stimuli (video /ba/ and audio /ba/; video /pa/ and audio /pa/), referred to as the 'Congruent' condition; mismatching audio–video stimuli (video /ga/ and audio /ba/; video /ka/ and audio /pa/), referred to as the 'Incongruent' condition. Finally, the 'Control' condition, including scrambled pixels videos together with audio /ba/ or /pa/. Therefore, the experimental design was 3 (Condition: Congruent, Incongruent, Control) x 2 (Syllable: audio /ba/, audio /pa/) factorial design.

2.2.2. *Participants*

Fifteen right-handed participants participated in the first experiment (9 males; mean age: 27.8 ± 1.9). All participants had normal hearing abilities and written informed consent was obtained from them. The Ethical Committee of IIT approved all the procedures.

2.2.3. *Procedure*

Participants were seated comfortably in front of a computer screen showing the visual stimuli and wore headphones to deliver the auditory ones. Audio intensity was set to a comfortable level before the beginning of the experiment. They were informed that they were going to view ultrasound recordings of tongue movements as well as scrambled-pixels versions of the same clips. They were not instructed about the presence of a mismatch between audio and visual stimuli and they were required to pay attention to both video and audio presentation throughout the experiment. In a short debriefing after the experimental session, none of the participants reported to have noticed the presence of the mismatch between audio and video stimuli. The full experiment lasted for about 20 min.

2.2.4. *Task*

Stimuli were presented by means of E-Prime software (Psychology Software Tools, Inc., USA, v2.0.8.22). Participants were asked to perform 2 alternative forced-choice task: they were asked to identify the heard syllable (/ba/ or /pa/) by pressing the associated button on a response pad. The response was given by pressing with the right index finger one out of two buttons associated with either the /ba/ or /pa/ auditory syllables (button-stimuli association was counterbalanced across participants). One finger was used to avoid any bias induced by differences in mean RT. The index finger was positioned between the two buttons.

In order to check that participants were paying attention to video clips, random rare (10% of total number of trials) catch trials were introduced. These catch trials were the same ultrasound videos used as stimuli, with the only difference that no red line was superimposed on the tongue dorsal profile. Participants were required to suppress any response when a catch trial was presented. Participants first completed a short training phase to get used to the response pad and stimuli (12 trials, 2 for each of the six stimuli) with accuracy displayed after each trial. Upon successful completion of training, they entered the experimental phase. In this phase, they completed 36 Congruent trials, 36 Incongruent trials and 36 Control trials as well as 12 catch trials (2 for each stimulus), leading to a total of 120 trials.

2.2.5. *Data acquisition*

Each video clip contained the auditory syllable on the left audio channel, and a short pulse, at the beginning of the syllable, on the right channel. The left channel was split and fed to both channels of the participants' earphones. The right channel was sent to an A/D acquisition board (CED Power1401 MkII, Cambridge Electronics, UK and Signal 4 software) together with the analog signal from the custom-made response pad (acquisition at 5 kHz). Therefore, RTs were acquired independently of the computer generating the stimuli to enable better temporal precision, whereas the E-prime script measured only accuracy of responses.

2.3. *Methods experiment 2*

2.3.1. *Design*

In the second experiment, we changed the auditory and visual stimuli maintaining the same design as in Experiment 1, with a two-fold aim. First of all, we wanted to replicate the previous effect on a different set of stimuli to check for robustness. Second, and more important, we intended to test whether asking the participants to implicitly respond on the same dimension (place of articulation) of the experimental stimuli manipulation was still able to elicit same effects. In brief, we again asked the participant to identify the consonant of the syllable, which in this experiment differed in terms of place of articulation and we presented visual stimuli that may or may not match on the same speech feature. Therefore, each trial presented tongue movements for either a bilabial or a velar place of articulation (bilabial /ba/ and velar /ga/) accompanied by an acoustic syllable that either matched or mismatched for place of articulation. Thus, the Congruent condition consisted of matching audio–video stimuli (video /ba/ and audio /ba/; video /ga/ and audio /ga/), whereas the Incongruent condition was formed by mismatching associations (video /ba/ and audio /ga/; video /ga/ and audio /ba/). The Control condition contained scrambled pixels videos associated with /ba/ or /ga/ auditory stimuli.

2.3.2. *Participants*

Eighteen right-handed participants participated in the second experiment (7 males; mean age: 26 ± 3.7), none of the participants had participated in the first experiment. All participants had normal hearing abilities and written informed consent was obtained from them. The Ethical Committee of IIT approved all the procedures.

2.3.3. *Procedure*

Procedure, task and data acquisition were the same as for Experiment 1.

2.4. *Analysis*

For both experiments, the same analysis procedure was applied. Reaction times (RTs) were analyzed by means of a within-participants ANOVA model, using as within factors the *Condition* (3 levels: Congruent, Incongruent, Control) and the *Syllable* (2 levels: /ba/, /pa/ in Experiment 1 and /ba/, /ga/ in Experiment 2). Outliers were removed by excluding RTs values exceeding two standard deviations from the participants' general mean. RTs associated with incorrect responses were also excluded from subsequent analysis. Accuracy was analyzed separately by calculating the proportion of correct responses and using this as the dependent variable in the same within-participants ANOVA model described above.

3. Results

3.1. *Reaction times*

3.1.1. *Experiment 1*

RTs were analyzed with 3 × 2 within-participants ANOVA (factor *Condition*: Congruent, Incongruent, Control; factor *Syllable*: /ba/, /pa/). The results show a significant main effect of *Condition* ($F(2,28)=19.35$, $p<0.0001$, effect size $\eta_p^2=0.58$; See Fig. 2a). No other effects are present (main effect of *Syllable*: $F(1,14)=0.885$, $p=0.36$; interaction *Syllable*Condition*: $F(2,28)=0.347$, $p=0.71$). Post-hoc comparisons (two-tailed paired *t*-test with Holm-Bonferroni method to correct for multiple comparisons) showed that RTs in the Incongruent condition (528.55 ± 27.27 ms, mean ± standard error of mean, S.E.M.) are slower with respect to RTs in Congruent condition (513 ± 26.94; $t(14)=-3.4967$, $p=0.00211$) and are faster than RTs in Control condition (553 ± 26.89; $t(14)=-4.1134$, $p=0.00031$).

3.1.2. *Experiment 2*

RTs were analyzed with 3 × 2 within-participants ANOVA (factor *Condition*: Congruent, Incongruent, Control; factor *Syllable*: /ba/, /ga/). A main effect for *Condition* is found ($F(2,34)=16.26$, $p<0.0001$, effect size $\eta_p^2=0.49$; See Fig. 2b), whereas both the main effect of *Syllable* ($F(1,17)=1.057$, $p=0.32$) and the interaction *Syllable*Condition* ($F(2,34)=0.015$, $p=0.98$) are not significant. Post-hoc comparisons (two-tailed paired *t*-test with Holm–Bonferroni method to correct for multiple comparisons) revealed that RTs in the Incongruent condition (528.77 ± 14.44 ms, mean ± S.E.M.) are slower than RTs of the Congruent condition (516.09 ± 14.25; $t(17)=-2.5442$, $p=0.02096$) and are faster than Control trials (552.09 ± 15.82; $t(17)=-3.4436$, $p=$ 0.0062).

3.2. *Accuracy*

3.2.1. *Experiment 1*

In the first experiment, accuracy was generally high during the whole task (92%) although Congruent trials lead to a non-significantly worse performance (90%), with respect to Incongruent (93%) and Control trials (93%). A within-participant ANOVA (as described in the Section 2.4) using accuracy as dependent variable showed a significant main effect for *Syllable* ($F(1,14)=12.03$, $p<0.01$), but no significant effects for *Condition* ($F(2,28)=2.15$, $p=0.136$) and for the interaction ($F(2,28)=1.06$, $p=0.36$) between the two factors. The difference in accuracy between the two syllables, evidenced by the result of the ANOVA analysis, is due to worse performance in the identification of the syllable /pa/ (87%) with respect to /ba/ (97%). Interestingly, such difference in accuracy did not influence RTs, which were not significantly different for the two syllables.

3.2.2. *Experiment 2*

In the second experiment, accuracy was very high (97%) and similar across conditions (Congruent: 98%, Incongruent: 97%, Control: 96%). The proportion of correct responses was not affected by experimental manipulation, as revealed by the absence of significant effects in the ANOVA (main effect of Condition: $F(2,34)=1.562$, $p=0.22$; main effect of Syllable: $F(1,17)=0.06$, $p=0.79$; interaction: $F(2,38)=0.48$, $p=0.61$). In this experiment, two syllables lead to a similar proportion of correct responses (/ba/ : 97%; /ga/: 97%).

The overall high accuracy obtained in both experiments and no differences across the three conditions assures that the difficulty in identification did not affect the result on reaction times. The difference in accuracy between the two syllables in the first experiment can be explained by the difference in VOT of the auditory stimuli. We can interpret this result in terms of the acoustic properties of the stimuli, where the identification of voiceless consonants could be more compromised by the presence of white noise with respect to voiced ones. This interpretation is further supported by the results of the accuracy in the second experiment, where both syllables (/ba/ and /ga/) were voiced and lead to the same identification rate (97%).

4. Discussion

In the present experiments, we find faster identification of auditory syllables when associated with both Congruent and Incongruent visual tongue movements with respect to the Control condition. Furthermore, congruent audio–visual matching stimuli led to an additional speed advantage with respect to the incongruent couples. The audiovisual matching effect we show, in the face of a lack of visual experience with tongue movements, argues against a purely audiovisual association hypothesis and suggests the recruitment of sensorimotor processes in speech perception.

4.1. *Sensorimotor speech maps*

Visual information present in all video clips (except Control movies) specified the temporal profile of a plosive event followed by the same /a/ vowel. Plosion is produced by completely blocking the airflow and then releasing it with the lips. Thus, visual information always belonged to the same hierarchically higher phonetic category (same vowel and same manner of articulation) as the auditory information. This fact may explain the general speed advantage with respect to the Control condition, which did not specify any vowel or plosive event. More importantly however, we did show a significant difference between audiovisually congruent and incongruent trials. Incongruent trials introduced a mismatch in the place of articulation (velar Vs. bilabial) leading to an increase in reaction times with respect to Congruent trials where such an audiovisual difference was not present.

Crucially, the difference between Congruent and Incongruent conditions supports the idea that place of articulation, in visually presented tongue movements, is implicitly recognized by the participants. This result is strengthened by the fact that we

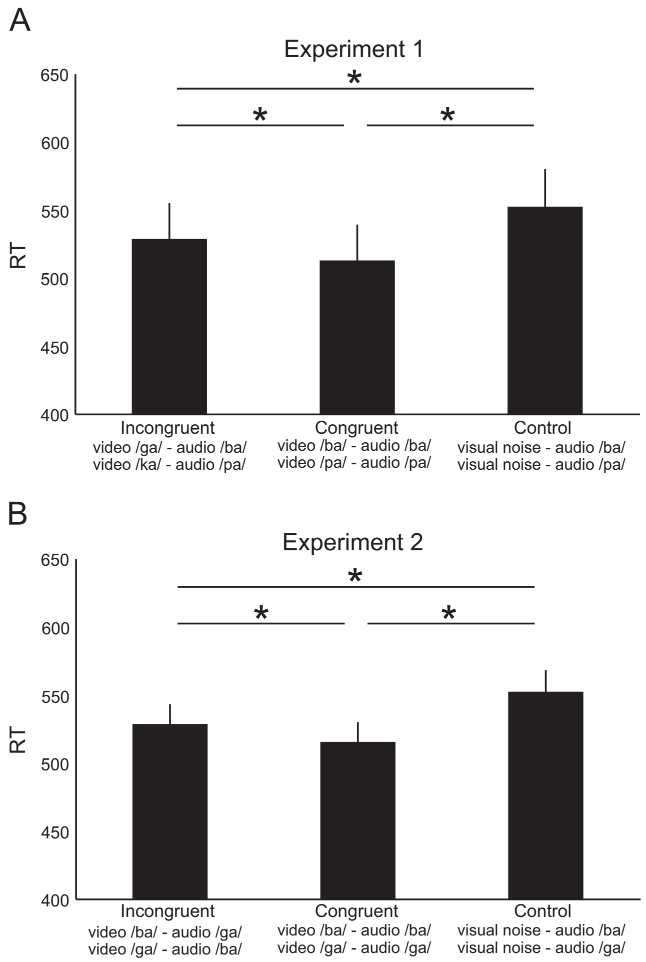

Fig. 2. Experiments 1 & 2. In panel A the mean reaction times for the two experimental and the control conditions, in the first experiment. In panel B, instead, the mean reaction times for the two experimental and the control conditions, in the second experiment. Asterisks show significant differences ($p < 0.05$). Thin bars above histograms depict standard error of the mean.

replicated it in two different stimulus sets. In fact, both tasks requiring the identification of the place of articulation (Experiment 2) or the voice onset timing (Experiment 1), lead to the same pattern of results. In one case we implicitly asked participant to pay attention to the very same articulatory dimension that we were also manipulating in the audiovisual matching/mismatching (Experiment 2). In the other case we kept the participants' task orthogonal to the experimental manipulation (Experiment 1). The fact that in both experiments we show the same pattern of results, testifies to the robustness of the effect and its independence with respect to task requirements. Therefore, our results seem to exclude the audiovisual hypothesis since there is no statistical association and no perceptual similarity between the speech sounds and visible tongue movements we used. Rather, our results provide compelling support for the claim that multimodal speech maps include sensorimotor information (Spence & Deroy, 2012).

4.2. *The role of familiarity in building sensorimotor speech maps*

Additional empirical support for the existence of sensorimotor speech maps comes from different behavioral paradigms. One of them consists of testing how much visual perception taps into implicit motor competence (Viviani, 2002). Along these lines, Viviani, Figliozzi, and Lacquaniti (2011) investigated the perception of visible speech by applying a series of temporal manipulations. These authors demonstrated that speech-related mouth movements are processed more accurately than temporally distorted speech movements (played backward), which indeed violate speech production rules. Otherwise, another approach is that of testing perceptuo-motor compatibility effects in speech (Prinz, 1990). In this sense, Kerzel and Bekkering (2000) as well as Galantucci, Fowler, and Goldstein (2009) showed that speech production might be selectively affected by the concurrent presentation of task-irrelevant visual speech.

However, in both approaches it is difficult to separate familiarity from sensorimotor competence. In fact, we are continuously exposed to speech, and thus any kind of unusual audiovisual mismatch (as in the McGurk effect), unusual visuomotor mismatch (Kerzel & Bekkering, 2000; Galantucci et al., 2009) or unusually reversed visual speech stimuli (Viviani et al., 2011) may lead to worsened performance, just because stimuli are not familiar. In this sense, two or more modalities can be temporally associated to offer a familiar multimodal stimulus. The more parsimonious hypothesis of familiarity expects that familiar stimuli be easier to identify than unfamiliar ones – independently from the unimodal content of the stimuli. In the present study however, we can exclude such possibility since visual stimuli are all equally unfamiliar (visually speaking) and not directly associated to their auditory counterpart, but yet induce specific bias on auditory identification performance.

4.3. *The role of experience in building sensorimotor speech maps*

In line with our results is also the demonstration that sensory information can affect perception whenever it is informative about its articulatory causal source rather than just associated with it. Skin stretch, usually experienced by lips and mouth during speech production, can exert a specific influence on perceptual discrimination tasks too (Ito, Tiede, & Ostry, 2009). Interestingly, tactile information related to mouth movements, experienced with hands (using the Tadoma lip reading that consists in placing the thumb on the lips and the other fingers on the jaw and throat), improves correct perception of the auditory stimulus, and can cause McGurk-like effects when the two sources are mismatching (Fowler & Dekle, 1991). Similarly, rather basic tactile sensations such as cutaneous air puffs mimicking airflow from the mouth, delivered to different body locations (hand and neck), affect the perception of concurrent speech sounds (Gick & Derrick, 2009). Altogether, these results corroborate the assumption that multisensory integration occurs between sources of information that are causally related to the same distal-motor- event.

Crucially, the latter two studies also imply that direct experience with a particular stimulus is not necessary to bias perceptual speech identification (Fowler & Dekle, 1991; Gick & Derrick, 2009). In fact, skin stretches (Ito et al., 2009), air puffs (Gick & Derrick, 2009) or Tadoma lip-reading (Fowler & Dekle, 1991) offer tactile signals that the participant experiences every time during speech production. Instead, passive speech perception does not readily allow the mapping of audio signals onto these somatosensory representations of speech sounds. Our result also adds to this body of evidence that is mainly based on a direct realist approach (Fowler, 1986). In fact, our results, by suggesting a sensorimotor component in audiovisual speech perception, cannot directly speak in favor of the existence of a motor representation or direct perception. Rather, our results show that (somatosensory and motor) information usually experienced during speech production can affect auditory identification even if presented into another un-experienced sensory modality (such as vision). This fact strengthens the claim that experience with a particular stimulus is not necessary to bias perceptual speech identifications as long as it refers to the same distal- motor- event (Gick & Derrick, 2009).

4.4. Separating the role of perceptual learning in building sensorimotor speech maps

Classically, the role of experience in building multimodal maps has been studied on infants since they have not had the time to acquire such knowledge. For example, haptic and visual shape invariances of objects were shown for 29-days old infants (Meltzoff & Borton, 1979). Similarly, imitation of facial gestures is already present few minutes after birth (Meltzoff & Moore, 1977). This latter result shows that even if infants have not yet seen their tongue (or other facial gestures) they recognize the recruited effector as well as the associated action and correctly match it with an appropriate imitation. Therefore, developmental research suggests that experience-dependent associative mechanisms cannot account for multisensory association effects in general (Meltzoff & Borton, 1979; Meltzoff & Moore, 1977), and during audiovisual speech processing (Coulon, Hemimou, and Streri, 2013).

However, we still lack a formal model to disambiguate the role of visual and motor experience in adults. In this sense, an alternative to the study of very young infants is to engage adult participants in a new learning scenario where the experimenter can manipulate the amount of prior experience in each modality. Using such approach, it was recently shown that motor learning has a direct and highly selective influence on visual action recognition that is not mediated by visual learning (Casile & Giese, 2006).

In the present study, instead we used tongue movement to disentangle the relative contribution of visual and motor experience in adults. The tongue is indeed the only effector, with enough kinematic complexity, that is concealed from vision and thus could represent a good model to dissociate the role of visual learning. Along these lines, we show that making visible the normally hidden tongue movements can affect the identification speed of speech sounds.

5. Conclusions

In conclusion, our results support the sensorimotor hypothesis formulated in the introduction, claiming that knowledge arising from motor production experience is exploited during perception. The use of motor knowledge allows us to make inferences about the distal sources that cause the events that we sense. In fact, information about the tongue movement can be integrated with the auditory modality only during explicit speech production. Audiovisual statistical learning via passive exposure to speech cannot explain our findings, since tongue-related information is not directly available for learning. We propose that the driving factor in building these audiovisual maps is speech production knowledge (Pulvermuller & Fadiga, 2010; D'Ausilio et al., 2009). In fact, internal motor models seem the most likely candidate to allow access to tongue kinematic knowledge. The activation of specific internal models during speech identification tasks, may offer effective generative methods to synthesize missing sensory information (Friston, Mattout & Kilner, 2011) such as the visual coding of the tongue during speech production. The advantage of a rich sensorimotor mapping might be that of constraining the top-down search for specific sensory features. An active sensory feature search may allow a faster and more effective confirmation of one perceptual hypothesis among others.

Acknowledgments

This work was supported by European Community grants SIEMPRE (ICT-FET project number 250026) and POETICON++ (STREP Project ICT-288382). We thank Giacinto Barresi and Elisabetta Ferrari for helping with data collection.

References

Alsius, A., & Munhall, K. G. (2013). Detection of audiovisual speech correspondences without visual awareness. *Psychological Science*, *24*(4), 423–431.

Baart, M., Vroomen, J., Shaw, K., & Bortfeld, H. (2014). Degrading phonetic information affects matching of audiovisual speech in adults, but not in infants. *Cognition*, *130*(1), 31–43.

Bever, T. G., & Poeppel, D. (2010). Analysis by synthesis: a (re-)emerging program of research for language and vision. *Biolinguistics*, *4*(2–3), 174–200.

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: perception of an emergent consonant in the McGurk effect. *Developmental Psychology*, *45*(4), 204–220.

Campbell, R. (2008). The processing of audiovisual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society B*, *363*, 1001–1010.

Canevari C., Badino L., Fadiga L., Metta G., (2013) Relevance-weighted-reconstruction of articulatory features in deep-neural-network-based acoustic-to-articulatory mapping. In *Proceedings of Inter Speech*, Lyon, France.

Casile, A., & Giese, M. A. (2006). Non-visual motor learning influences the recognition of biological motion. *Current Biology*, *16*(1), 69–74.

Castellini, C., Badino, L., Metta, G., Sandini, G., Tavella, M., Grimaldi, M., et al. (2011). The use of phonetic motor invariants can improve automatic speech discrimination. *PLoS One*, *6*(9), e24055.

Coulon, M., Hemimou, C., & Streri, A. (2013). Effects of seeing and hearing vowels on neonatal facial imitation. *Infancy*, *18*(5), 782–796.

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., & Fadiga, L. (2009). The motor somatotopy of speech perception. *Current Biology*, *19*(5), 381–385.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *The Journal of Phonetics*, *14*, 3–28.

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: cross-modal contributions to speech perception. *The Journal of Experimental Psychology: Human Perception and Performance*, *17*(3), 816–828.

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, *104*(1-2), 137–160.

Galantucci, B., Fowler, C. A., & Goldstein, L. (2009). Perceptuomotor compatibility effects in speech. *Attention Perception and Psychophysics*, *71*(5), 1138–1149.

Gick, B., & Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature*, *462*(7272), 502–504.

Grant, K. W., van Wassenhove, V., & Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditoryvisual (cross-modal) synchrony. *Speech Communication*, *44*, 43–53.

Ito, T., Tiede, M., & Ostry, D. J. (2009). Somatosensory function in speech perception. *Proceedings of the National Academy of Sciences of the United State of America*, *106*, 1245–1248.

Kerzel, D., & Bekkering, H. (2000). Motor activation from visible speech:evidence from stimulus response compatibility. *The Journal of Experimental Psychology: Human Perception and Performance*, *26*, 634–647.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, *218*(4577), 1138–1141.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*(2), 13–21.

Jones, J. A., & Munhall, K. G. (1997). The effects of separating auditory and visual sources on audiovisual integration of speech. *Canadian Acoustics*, *25*, 13–19.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*(5), 630–645.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431–461.

Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child Development*, *55*, 1777–1788.

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, *198*, 75–78.

Meltzoff, A. N., & Borton, R. W. (1979). Intermodal matching by human neonates. *Nature*, *282*, 403–404.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, *58*(3), 351–362.

Papcun, G., Hochberg, J., Thomas, T. R., Laroche, F., Zacks, J., & Levy, S. (1992). Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *Journal of the Acoustical Society of America*, *92*(2), 688–700.

Parise, C. V., Spence, C., & Ernst, M. O. (2012). When correlation implies causation in multisensory integration. *Current Biology*, *22*(1), 46–49.

Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior & Development*, *22*, 237–247.

Patterson, M., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, *6*, 191–196.

Pick, H. L., Warren, D. H., & Hay, J. C. (1969). Sensory conflicts in judgments of spatial direction. *Perception & Psychophysics*, *6*, 203–205.

Prinz, W. (1990). A common coding approach to perception and action. In: O. Neumann, & W. Prinz (Eds.), *Relationships between perception and action: current approaches* (pp. 167–201). New York: Springer.

Pulvermüller, F., & Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience*, *11*(5), 351–360.

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception* & *Psychophysics*, *59*(3), 347–357.

Spence, C., & Deroy, O. (2012). Hearing mouth shapes: sound symbolism and the reverse McGurk effect. *I-Perception*, *3*(8), 550–552.

Stevens, K. N., & Halle, M. (1967). Remarks on analysis-by-synthesis and distinctive features. In: W. Wathen-Dunn (Ed.), *Models for the perception of speech and visual form*. Cambridge, MA: MIT Press.

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

Sweeny, T. D., Guzman-Martinez, E., Ortega, L., Grabowecky, M., & Suzuki, S. (2012). Sounds exaggerate visual shape. *Cognition*, *124*(2), 194–200.

Vatakis, A., & Spence, C. (2006). Audiovisual synchrony perception for music, speech, and object actions. *Brain Research*, *1111*(1), 134–142.

Viviani, P. (2002). Motor competence in the perception of dynamic events: a tutorial. In: W. Prinz, & B. Hommel (Eds.), *Attention* & *performance XIX: common mechanisms in perception and action* (pp. 406–442). Oxford: Oxford University Press.

Viviani, P., Figliozzi, F., & Lacquaniti, F. (2011). The perception of visible speech: estimation of speech rate and detection of time reversals. *Experimental Brain Research*, *215*(2), 141–161.