Visual contribution to the multistable perception of speech

Marc Sato, Anahita Basirat, Jean-Luc Schwartz

Running Head: Visual contribution to the multisatble of speech - P494

Visual contribution to the multistable perception of speech

Marc Sato [1,2], Anahita Basirat [1], Jean-Luc Schwartz [1-CA]

[1] *Institut de la Communication Parlée, CNRS UMR 5009,*

Institut National Polytechnique de Grenoble, Université Stendhal

[2] *Departimento di Neuroscienze, Sezione di Fisiologia, Universita di Parma*

[CA] Corresponding author.

Institut de la Communication Parlée – UMR CNRS 5009

Institut National Polytechnique de Grenoble - Université Stendhal

46, avenue Félix Viallet, 38031 Grenoble Cedex 01 - France.

Office: +33 (0)4 76 57 47 12

Fax: +33 (0)4 76 57 47 10

E-mail address: schwartz@icp.inpg.fr

ABSTRACT

The multistable perception of speech, or *Verbal Transformation Effect*, refers to perceptual changes experienced while listening to a speech form that is repeated rapidly and continuously. In order to test whether visual information from the speaker's articulatory gestures may modify the emergence and stability of verbal auditory percepts, participants were instructed to report any perceptual changes during unimodal, audiovisual and incongruent audiovisual presentations of distinct repeated syllables. In a first experiment, the perceptual stability of reported auditory percepts was significantly modulated by the modality of presentation. In a second experiment, when presenting audiovisual stimuli consisting of a stable audio track dubbed with a video track that alternated between congruent and incongruent stimuli, a strong correlation between the timing of perceptual transitions and the timing of video switches was found. Finally, a third experiment showed that the vocal tract opening onset event provided by the visual input could play the role of a bootstrap mechanism in the search for transformations. Altogether, these results demonstrate the capacity of visual information to control the multistable perception of speech in its phonetic content and temporal course. The Verbal Transformation Effect thus provides a useful experimental paradigm to explore audiovisual interactions in speech perception.

Key Words: Multistable perception, Audiovisual speech perception, Verbal Transformation Effect, Perceptuo-Motor Interactions.

INTRODUCTION

The human ability to follow speech gestures through the visual modality can be considered a core component of speech perception. Behavioural studies have demonstrated that concordant visual information improves speech intelligibility in noisy audio conditions (e.g., Sumby & Pollack, 1954; MacLeod & Summerfield, 1987; Benoît, Mohamadi & Kandel, 1994; Robert-Ribes, Schwartz, Lallouache & Escudier, 1998) and that speech gestures can partly be followed when audition is lacking (Bernstein, Demorest & Tucker, 1998). Furthermore, even with perfect audio input, speechreading may improve speech intelligibility (Reisberg, McLean & Goldfield, 1987; Davis & Kim, 1998). On the contrary, seeing incongruent articulatory gestures can alter the perception of clear auditory input. For example, in the famous McGurk effect (McGurk & MacDonald, 1976), a visual /ga/ dubbed with an audio /ba/ is often perceived as /da/ or /tha/, while a visual /ba/ dubbed with an audio /ga/ is often perceived as /bga/ (for a review of experimental replications or refinements of the McGurk effect, see Green, 1998).

A number of theoretical models have been proposed concerning the possible cognitive basis of these findings (e.g. Summerfield, 1987; Massaro, 1987, 1989; Bernstein, Auer, & Moore, 2004)). In their reanalysis of the five architectures proposed by Summerfield (1987), Schwartz, Robert-Ribes & Escudier (1998) suggested that the “motor recoding model” was the most plausible with regard to experimental findings. In this model, audiovisual interactions take place in a common representation space in which the sensory inputs are related to compatible articulatory gestures. Though this architecture differs from the simple equivalence between percepts and gestures posited in the motor theories (e.g., Liberman & Mattingly, 1985; Fowler & Rosenblum, 1991: see Schwartz, Abry, Boë & Cathiard, 2002, Schwartz, Boë & Abry, 2006

for reviews of similarities and differences), it considers perceptuo-motor interactions as playing a key part in multisensory speech perception. Consistent with this view, it has been shown that seeing oneself articulating in a mirror improved identification of concordant acoustic syllables and deteriorated identification of discordant ones (Sams, Möttönen & Sihvonen, 2005). Neurophysiological data has provided further evidence for such perceptuo-motor interactions. Indeed, brain areas involved in the planning and execution of speech gestures (notably the left inferior frontal gyrus, the premotor and/or the primary motor cortex) have been found to be activated during audiovisual speech perception (e.g., Callan, Jones, Munhall, Callan, Kroos & Vatikiotis-Bateson, 2003; Ojanen et al., 2005; Pekkola et al., 2005; Skipper, Nusbaum and Small, 2005, 2007). In these studies, a cortical 'dorsal' circuit associating temporal and frontal regions, including the ventral premotor cortex and Broca's area, has been repeatedly shown to be involved in audiovisual speech perception, and is a likely candidate for connecting multisensory and motor representations in the human brain.

In spite of this rather coherent set of data and proposals, the precise functioning of the audiovisual speech processing and categorization system remains largely unknown, including, for instance, the content of perceptual representations, the detailed mechanisms involved in the decision-making process, and the exact way in which perceptual and motor knowledge is connected. To address these issues, a variety of experimental tools may be employed. The present study exploits a well-known paradigm, the *Verbal Transformation Effect* (Warren and Gregory, 1958; Warren, 1961), which, surprisingly, has never been explored in the context of multisensory perception. This multistable perception phenomenon refers to the perceptual changes experienced while listening to a speech form that is repeated rapidly and continuously.

Initially, a percept matching the original form is heard, but at some point another percept suddenly arises, corresponding to a change in the perceived stimulus. For example, rapid repetitions of the word ‘life’ provide a soundflow that is fully compatible with the perception of repeated instances of the word ‘life’ or ‘fly’. This transformation process persists throughout the repetition procedure, leading to perceptual transitions from one speech form to another. While verbal transformations have mainly been studied as a purely acoustical/auditory effect (e.g., Warren, 1961; MacKay, Wulf, Yin & Abrams, 1993; Pitt and Shoaf, 2001, 2002; Shoaf & Pitt, 2002), it has been shown that the effect occurs not only during a purely auditory procedure but also during an overt or covert repetition procedure (Reisberg, Smith, Baxter & Sonenshine, 1989; Smith, Reisberg & Wilson, 1995; Sato, Schwartz, Abry, Cathiard & Lœvenbruck, 2006).

In the present study, we exploited the verbal transformation paradigm, which provides multistability in speech perception, to examine whether visual information from the speaker’s articulatory gestures might modify the *emergence* and *stability* of verbal auditory percepts. Interestingly, a recent neuroanatomical study of the auditory verbal transformation effect (Sato, Baciu, Lœvenbruck, Schwartz, Cathiard, Segebarth & Abry, 2004) displayed a cortical network rather similar to the audiovisual speech perception circuit described previously, connecting temporal regions to frontal ones (including Broca’s area) through the left supramarginal gyrus in the parietal cortex. Adding this fact to the many convincing reports that audition and vision interact in most aspects of speech perception (Summerfield, 1987; Bernstein, Burnham & Schwartz, 2002), it is quite likely that verbal transformations should in fact be audiovisual.

In the present study, we are interested in three questions in particular. First, is it in fact possible to show visual influences on auditory verbal transformations? Can a congruent visual

stimulus stabilize the auditory percept and decrease the number of reported transformations? Conversely, can an incongruent visual stimulus destabilize the auditory percept and increase the number of transformations? In addition, does the visual modality alone lead to verbal transformations in speechreading? This first series of questions aims to demonstrate that verbal transformations are indeed audiovisual, and hence that they could be usefully exploited for further studies of audiovisual speech perception.

Second, if multisensoriality can indeed be introduced into the verbal transformation effect, is it possible to drive the listener's perception of a given auditory stimulus over time by adequately switching the visual input from one stimulus to another? This would provide an original and potentially powerful way to attempt to *control the perception of a multistable auditory input* by adequately changing the visual stimulus presented in synchrony with it.

Finally, the demonstration following this second question that vision could actually drive the perceptual switch over time leads to a third question: namely, how does speechreading actually drive the perceptual switch? In order to answer this question, we studied one possible component of this driving mechanism: the visual onset events associated with *jaw opening gestures*. These three questions will be successively addressed through three experiments.

EXPERIMENT 1: DOES VISION BIAS AUDITORY VERBAL TRANSFORMATIONS?

The first experiment was designed to test whether visual information from articulatory gestures might influence verbal transformations. To determine the visual contribution to the stability of verbal auditory percepts, stimuli were presented in four different modalities: audio-only, video-only, congruent audio-video and incongruent audio-video (corresponding to the audio track of a speech stimulus dubbed onto the video track of another stimulus).

METHOD

Subjects

Fifteen healthy volunteers (six males and nine females; mean age ± standard deviation, 27 ± 7 years) participated in the experiment. All were native French speakers who reported no hearing or speaking disorders and had normal or corrected-to-normal vision. None of the subjects were aware of the purpose of the experiment.

Phonetic Material

Two monosyllabic nonsense words were selected from a set of stimuli used in a previous study (Sato et al., 2006). The two syllables – /psɮ/ and /sɮp/ – consisted of the combination of the bilabial /p/ and coronal /s/ consonants with the neutral vowel /ɮ/. None of these syllables occur in the French lexicon, which minimizes lexical interference in the verbal transformation task (Shoaf & Pitt, 2002). However, the phonological types of speech sequences (i.e., /psV/, /sVp/, /Vps/, /spV/, /pVs/ and /Vsp/ - V being any oral French vowel) are all phonotactically valid in French. Lexical analyses (extracted from VoCoLex, a lexical database

for the French language, ~105000 words; Dufour, Peerman, Pallier and Radeau, 2002) showed that both the lexical type frequency (defined as the number of lexical entries incorporating a monosyllabic structure identical to that of the stimulus at any position in a word) and neighbourhood density values (defined as the number of phonologically similar words that differ from the stimulus by only a single substitution, insertion or deletion at any position in the target word; Luce, Pisoni and Goldinger, 1990) were lower for /psV/ than for /sVp/ (114 vs. 371 entries and 31 vs 59 entries, respectively).

Stimuli

Multiple utterances of the two syllables were individually recorded in a soundproof room by a trained phonetician who is a native French speaker (J.-L.S.). For the audiovisual recordings, the speaker's lips were coloured with blue make-up to allow for precise video analyses using a chroma-key process (Lallouache, 1990). The speaker was told to pronounce each syllable naturally at a speech rate of approximately one cycle per second, maintaining an even intonation and vocal intensity. The recorded clips were then edited on a computer using Adobe Premiere software (www.adobe.com). Video was sampled at 25 images/s with a resolution of 720 × 576 pixels. The audio signal was sampled at 44.1 kHz and 16-bit resolution.

One clearly articulated /psɫ/ token and one clearly articulated /sɫp/ token were selected and matched on three criteria. First, the initial and final images for each stimulus were very similar, corresponding to a neutral mid-open mouth position. Second, the consonantal acoustic onset of both stimuli (bilabial burst for /psɫ/, coronal release for /sɫp/) was aligned, with the consonantal onset occuring in both cases at the 5th image of the movie, and the stimuli were

matched in duration (640 ms, 16 frames). Third, the two utterances were matched for intensity using a spectrogram analysis of the acoustic track (Praat software, Institute of Phonetic Sciences, University of Amsterdam, the Netherlands). A display of the stimuli is presented in Figure 1. The figure includes the acoustic waveform along with the spectrogram (enabling the observation of vowel formants, plosive burst and fricative noise, intensity and fundamental frequency contours), and lip aperture variations automatically computed from the video stimulus using the chroma-key process (Lallouache, 1990; Abry, Cathiard, Vilain, Laboissière & Schwartz, 2006).

Insert Figure 1 about here

Four distinct stimuli per syllable were created by inserting the same sequence without pause in a single 96-s movie (i.e., 150 repetitions of the syllable) according to the following modality conditions: A: *audio-only*, V: *video-only*, AV: *congruent audio-video and AVi: incongruent audio-video* (i.e., consisting in a /psɬ/ audio track dubbed onto a /sɬp/ video track or vice-versa). In order to bootstrap the initial categorization of the uttered syllable for the video-only condition, a congruent synchronized audio-track was inserted for 1.92 s (i.e., three repetitions of the syllable) at the beginning of the movie. For the incongruent audio-video condition, the previously explained synchronisation of the /psɬ/ and /sɬp/ utterances ensured that the consonantal onset of the syllable in the video track was synchronized with the consonantal onset of the syllable in the audio track (at the 5th image).

Apparatus

The video track was delivered on a 19-inch computer monitor at a viewing distance of

approximately 60 cm. The audio track was presented binaurally over headphones at a comfortable sound level. Subjects' verbal responses (indicating transformations) were collected via a microphone and directly recorded as individual sound files onto the hard disk of the computer. The software ensured a precise synchronization between the presented stimulus and the recorded response. The time of occurrence of each transformation was taken to be the acoustic onset of the verbal response. Of course, there is a delay of a few hundred milliseconds between the perceptual switch and the vocal response signalling this switch. This point will be discussed in the section describing Experiment 2.

Procedure

The participants were individually tested in a quiet room. The experiment began with a lengthy briefing during which they were introduced to the verbal transformation task. They were then told that they would hear and/or watch an utterance being repeatedly articulated and were asked first to report what they heard and then to report any perceived changes in the repeated utterance. It was indicated that the changes could be subtle or very noticeable and that the stimulus could correspond to a word as well as a pseudo-word. Finally, the participants were assured that there were no correct or incorrect responses and that if they did not hear a transformation, they were to say nothing. Each 96-s stimulus was presented once to each subject. The sequence of eight 96-s stimuli (4 modalities, 2 syllables) was randomised for each subject, with short (8-s) silent breaks in between each stimulus.

Data Analysis

The data were analysed by labelling all subjects' reports in the response sound files. For each subject and each stimulus, all reported forms were extracted. A first analysis was

performed on switching frequency, that is, the number of reported transformations within the 96-s stimulus. A repeated measures analysis of variance (ANOVA) was performed to determine possible differences in switching frequency. The considered factors were: 1) the presented audio syllable (/psɫ/ and /sɫp/) and 2) the modality of the presentation (A, V, AV and AVi).

Further analyses were carried out in order to assess the perceptual stability of the reported transformations more directly than the switching frequency measure (Shoaf and Pitt, 2002). The perceptual global stability duration of each form was calculated by summing the time spent perceiving the given form before switching to another form. An analysis of the types of reported transformations was carried out. Transformations primarily included switching from /psɫ/ to /sɫp/ or from /sɫp/ to /psɫ/, substitution of a phoneme with one that was phonetically similar (e.g., /tsɫ/ for /psɫ/), auditory "streaming" in which the repeated stimulus was separated into different audio streams, giving rise to verbal transformations involving only part of the presented material as previously shown by Pitt and Shoaf (2001, 2002) (e.g., /sɫ/ for /sɫp/, the labial release for /p/ being discarded as a separate perceptual stream), and lexical transformations (e.g., 'stop' for /sɫp/). All transformations were then classified as either /psɫ/, /sɫp/ or 'other'. To further explore the nature of the perceptual grouping across /psɫ/ and /sɫp/, the difference between the mean relative stability durations (i.e., global stability duration divided by the overall 96-s duration) observed for the presented syllable and the associated ones were calculated for each participant (*delta-score*): that is, for a /psɫ/ audio syllable, the delta score was the difference between the relative durations of the /psɫ/ and /sɫp/ responses, while

for a /sɬp/ audio syllable, it was the difference between the relative durations of the /sɬp/ and /psɬ/ responses. All delta-score values will be presented as percentages (%). To statistically assess the effect of the various conditions, an ANOVA was performed on delta scores with the same factors as those used in the analysis of switching frequencies. For the following analyses, the significance level was set at $p < .05$ and Greenhouse-Geisser corrected when appropriate. When required, post-hoc analyses were conducted using Newman-Keuls tests.

RESULTS

Switching frequencies

The mean switching frequencies for the two audio syllables in the four modalities are presented in Table 1. The ANOVA showed that there was a significant effect of the modality of presentation ($F(3,42) = 4.5$, $p < .03$) but no reliable effect of the audio sequence ($F(1,14) = 0.09$) and no interaction between the two factors ($F(3,42) = 1.46$). Post-hoc analyses showed a significantly higher switching frequency in the AVi modality compared with the V modality ($p < .005$). Although the switching frequency in the V condition was smaller than that observed in the A and AV conditions, the difference was just below the significance threshold. Inter-individual variability was large (mean range: 2-23 switches), with some subjects switching much more than others, which is typical of multistability phenomena.

Global perceptual stabilities and delta scores

The mean relative stability durations for the three categories of reported transformations observed in Experiment 1 (for the two audio syllables and four modalities) are displayed in Figure 2. The reported transformations distinct from /psɬ/ and /sɬp/ (classified as ‘other’)

included the substitution of a phoneme by a phonetically similar one (e.g., /tsɬ/ for /psɬ/), auditory streaming (e.g., /sɬ/ for /sɬp/), lexical transformations (e.g., 'stop' for /sɬp/), and more complex transformations, including combinations of insertions and substitutions (e.g. /poto/ or /sapo/) and re-segmentation in larger sequences (e.g. /psɬps/).

Insert Figure 2 about here

Analysis of delta scores showed a significant effect of the modality of presentation ($F(3,42) = 22.26$, $p < .0001$), of the audio sequence (with delta scores larger for /psɬ/ than for /sɬp/ - $F(1,14) = 6.27$, $p < .03$) and of the two-way interaction between the two factors ($F(3.42) = 3.26$, $p < .04$). Post-hoc analyses revealed a significantly lower delta score in the AVi condition compared with the A, V and AV ones (all *p*s < .001), and a significantly lower delta score in the V condition than in the AV one ($p < .04$). The interaction between the two factors was largely due to the significant difference observed between the two sequences in the AVi condition, as compared with the other modalities. In this condition, while /psɬ/ remained relatively stable during the incongruent audiovisual presentation (i.e., with a positive delta score of 11%), a negative delta score (-28%) was observed for /sɬp/ ($p < .002$). The same trend, although not significant, was found within the visual modality, with a delta score 15% higher for /psɬ/ (52%) than for /sɬp/ (37%).

DISCUSSION

The first interesting result of this experiment is the finding of verbal transformations in

the visual modality. These transformations include a large number of “other” transformations, more frequently observed in the visual condition than in the audio or audiovisual conditions. This is likely due to the intrinsic ambiguity of the visual stimulus associated with the existence of visemes that correspond to multiple phonemes with similar lip shapes (e.g., /b/ for /p/ or /t/ for /s/), hence leading to a large variety of transformations (e.g. /bdɬ/, /mnɬ/, /pɬsɬ/ or /plçk/ for /psɬ/). In this context, it is somewhat surprising that the switching frequencies are relatively low in the visual modality compared with the auditory modality (though not significantly so). Altogether, visual transformations seem to involve a slower exploration of a larger number of patterns compared with auditory transformations.

The significant decrease in delta scores in the AVi modality compared with the A or AV modalities, for both audio inputs, shows that vision influences verbal transformations by modifying the perceptual stability of an audio input. This is consistent with classical perceptual experiments on conflicting audiovisual stimuli, since it is well known that vision can bias the auditory percept (e.g., McGurk & MacDonald, 1976). The trend here is that perception switches between the audio and video inputs, as displayed by the large number of /sɬp/ responses to incongruent /psɬ/ audio stimuli, and of /psɬ/ responses to incongruent /sɬp/ audio stimuli. To understand more precisely how the visual component likely modifies the percept, let us return to Figure 1. In this figure, we have displayed auditory onset events for the plosive /p/ (burst onset) and the fricative /s/ (noise onset). These two events are nearly synchronous for /psɬ/ at the syllable onset and largely asynchronous for /sɬp/, as one occurs at the syllable onset (/s/)

and one occurs at the coda (/p/). In the audio condition, most of the verbal transformations occurred according to an auditory streaming process whereby the /p/ acoustic burst was removed from the syllable and perceived as a separate stream. This typically results in /sɬ/ responses and provides a large part of the “other” responses in Figure 2. In contrast, the visual modality provides a very visible lip opening gesture for /p/ (onset event also displayed in Figure 1). In the case of congruent auditory and visual inputs, the visible /p/ lip opening gesture is synchronous with the /p/ acoustic burst and may stabilise it in the audiovisual percept. This could explain why delta scores are higher in the AV modality than in the A modality, though not significantly (likely due to a ceiling effect). In the AVi modality, the /p/ audio bursts are even more frequently removed from the primary flow since they are not synchronous with a labial gesture. In this case, the auditory /s/ onset events and the visual /p/ lip opening gestures are combined into the main speech stream (displayed by larger arrows in Figure 1). If a visual /psɬ/ is synchronous with an audio /sɬp/, the visual /p/ event is almost synchronous with the audio /s/ onset event, which biases perception towards /psɬ/. If a visual /sɬp/ is synchronous with an audio /psɬ/, the visual /p/ event is not close to the audio /s/ onset event, which biases perception towards /sɬp/. This could explain the observed systematic bias towards the visual input in the AVi modality, as shown by the low delta scores for both sequences.

Finally, it is interesting to note the significant difference in delta scores between /psɬ/ and /sɬp/ in the AVi condition. In a previous study, Sato et al. (2006) showed that in a self-production paradigm with the same /psɬ/ and /sɬp/ stimuli, there was a strong tendency for

speakers uttering /sɬp/ sequences to progressively resynchronise the /p/ and /s/ events within a single /ps/ cluster at the syllable onset, hence producing a bias in favour of /psɬ/ verbal transformations. The fact that this asymmetry appears again in the AVi modality might be due to the involvement of a covert production process in the course of audio-visual integration (Schwartz et al., 1998; Callan et al., 2003). This process would lead the listener to mentally resynchronise the /p/ and /s/ events within a single opening cycle, and hence bias the percept towards /psɬ/.

Altogether, the answer to Question 1 posed in the Introduction is positive: verbal transformations *are* audiovisual, and the visual modality drives to a certain extent the phonetic content of the displayed transformations. We shall now move on to Question 2, which deals with the possibility that the listener's perception of a given auditory stimulus *over time* can be modified by intermittently switching the video content of an audiovisual presentation

EXPERIMENT 2: CAN VISION DRIVE AUDITORY VERBAL TRANSFORMATIONS OVER TIME?

In this second experiment, we examined whether the visual modality might drive the dynamics of reported perceptual changes over time. To this aim, participants watched and listened to audiovisual clips consisting of a stable audio track of a repeated syllable mixed with a video track alternating between congruent and incongruent stimuli. A close synchrony between the timing of video switches and the timing of perceptual transitions would demonstrate the capacity of visual information to temporally drive the multistability process involved in the Verbal Transformation Effect.

METHOD

Subjects

The subjects were the same as in Experiment 1.

Stimuli

The stimuli were prepared from the same phonetic material as was used in Experiment 1, that is, the /psɫ/ and /sɫp/ 640-ms tokens described previously. For this experiment, four 96-s movies were created by dubbing a single, repeating audio syllable (either /psɫ/ or /sɫp/) without pause onto a video track corresponding to an alternation between /psɫ/ and /sɫp/ (see Figure 3). The fact that the initial and final images of the /psɫ/ and /sɫp/ utterances were indistinguishable (see Experiment 1) ensured that the subjects would not notice anything unnatural about the video tracks in spite of the stimulus alternations. In order to avoid response habituation, the durations of the /psɫ/ and /sɫp/ sections on the video track were randomly varied, with two sets

of duration values providing two distinct movies per syllable. Durations were selected from a uniform distribution ranging from 6 and 12 s in one movie and 11 and 17 s in the other. Durations of the video tracks were closely matched across stimuli.

Insert Figure 3 about here

Procedure

Experiment 2 systematically followed Experiment 1 and employed the exact same procedure (as described previously).

Data Analysis

The data were analysed by labelling all subjects' reports in the response sound files. For each subject and each stimulus, all reported forms were extracted. To assess the capacity of visual information to temporally drive the multistability process involved in the Verbal Transformation Effect, the precise response timing for the reported percepts was analysed in relation to the syllable alternation displayed on the video track. The perceptual global stability duration of each form was then calculated by summing the time spent perceiving the given form before switching to another, and all transformations were classified as a function of whether they corresponded to /psɫ/, /sɫp/ or to another speech form. The classification was made in relation to the type of audio syllable and the nature of the video track (congruent vs. incongruent). This enabled us to compute delta scores in the same way as in Experiment 1. A repeated measures ANOVA was performed on these delta scores. The considered factors were the audio syllable (/psɫ/ and /sɫp/), the visually presented syllable (congruent vs. incongruent) and the duration of the video alternation (6-12 or 11-17 s). When required, post-hoc analyses

were conducted using Newman-Keuls tests. The significance level was always set at $p < .05$. Prior to performing these statistical analyses, Mauchly's tests showed that the sphericity assumption was not violated ($p > .05$).

Finally, a comparison of delta scores was performed between Experiments 1 and 2 in order to assess the possibility that visual stimuli switching in time (in Experiment 2) could influence the current percept more than stable stimuli (in Experiment 1). The prediction was that the reported percept would correspond to the video stimulus more often in Experiment 2. For periods of time in which the video stimulus was congruent with the audio stimulus, this would result in an increased proportion of responses that correspond to the acoustic input, hence increasing the delta scores. For periods of time in which the video stimulus was incongruent with the audio stimulus, this would result in decreasing the proportion of responses corresponding to the acoustic input, hence decreasing the delta scores. In summary, the prediction was that delta scores would be higher in Experiment 2 than in Experiment 1 for congruent stimuli, and lower in Experiment 2 for incongruent stimuli. This was tested using a repeated measures ANOVA, with the experiment, the syllable and the modality as within-subjects factors. Two modalities were contrasted, namely congruent (AV in Experiment 1, periods with audio and video stimuli congruent in Experiment 2) and incongruent (AVi in Experiment 1, periods with audio and video stimuli incongruent in Experiment 2).

RESULTS

Synchrony between visual stimuli switches and perceptual switches

The histograms of delays between the visual switches and the first reported transformations coherent with the visual input indicate a high level of synchrony between the

subject's responses and the visual stimulus transformation for the two audio stimuli (/psɫ/ and /sɫp/). Indeed, 85% of the first reported forms coherent with the visual input occurred within the first two seconds following the video switch. Note that this 2-s delay includes both the time necessary for the decision-making process to realize the switch from one pattern to another and the time required to prepare and utter the response signalling the transformation. Close-shadowing experiments have shown that the time required for response preparation and delivery is in fact quite small; probably less than 200 ms (e.g., Porter & Castellanos, 1980; Porter & Lubker 1980). Considering that the stimuli in the present experiment are 640-ms long, this would suggest that 2 to 3 repetitions of the new video stimulus are necessary to induce the decision switch and hence the verbal transformation overtly signalled by the subject.

Global perceptual stability and delta scores

The mean relative stability durations for the three categories of reported transformations observed in Experiment 2 for the two audio syllables (/psɫ/ and /sɫp/) and the two video conditions (congruent vs. incongruent) are displayed in Figure 4. The values are averaged over the two movies, as the differences between durations of video alternations displayed only a marginal effect (as shown below).

Insert Figure 4 about here

Analysis of delta scores revealed a significant effect of the congruency between audio and video signals (with delta scores larger for congruent audiovisual stimuli - $F(1,14) = 81.44$, $p < .0001$) but no reliable effect of the audio sequence ($F(1,14) = 1.30$) nor an interaction

between the two factors (F(1,14) = 0.41). The effect of the duration of video alternations was not reliable (F(1,14) = 0.46) nor was there an interaction with the audio sequence (F(1,14) = 2.05) or with the congruency effect (F(1,14) = 2.53). Finally, the three-way interaction between factors was significant (F(1,14) = 4.82, $p < .05$). This interaction was largely due to the differences in delta scores observed between the two distinct movies in the incongruent condition when the /psɫ/ and /sɫp/ audio sequences were compared. While the difference observed between the two movies was minimal for /psɫ/ (-38% in movie 1 vs. -42% in movie 2), the difference was stronger (although not significantly so) for /sɫp/, with a smaller delta score for movie 1 (-57%) than for movie 2 (-28%).

Experiment 1 vs. Experiment 2

Comparison of delta scores between the two experiments revealed a significant effect of the experimental session (with delta scores being larger during Experiment 1 than during Experiment 2 - F(1,14) = 11.24, $p < .005$), a significant effect of the sequence (with delta scores larger for /psɫ/ than for /sɫp/ – F(1,14) = 5.67, $p < .04$), a significant effect of the congruency between audio and video signals (with delta scores larger for congruent audiovisual stimuli - F(1,14) = 150.20, $p < .0001$), and a significant interaction between the experimental session and the congruency between audio and video signals (F(1,14) = 4.66, $p < .05$). Post-hoc analyses showed that the interaction between the experimental session and the congruency between audio and video signals was due to a significantly lower delta score observed in the incongruent condition in Experiment 2 compared with Experiment 1 (on average, -41% vs. -8%, $p < .02$).

DISCUSSION

The comparison of delta scores between Experiments 1 and 2 basically confirms our hypothesis. Indeed, a significantly lower delta score was obtained in the incongruent audiovisual condition in Experiment 2 as compared with Experiment 1. This would suggest that the visual component of the audiovisual stimuli drives the response more when it changes (Experiment 2) than when it does not (Experiment 1). By analysing the subjects' responses in relation to the visual input rather than to the audio input, it appears that the reported percept is coherent with the visual input approximately 71% of the time in Experiment 2 compared with 59% in Experiment 1. This difference is much larger when restricted to incongruent stimuli, for which the visual input drives the response 42% of the time vs. 33% for the audio input in Experiment 1, while the values are respectively 62% vs. 21% in Experiment 2. Finally, this induction effect is confirmed by the very high level of synchrony observed between the visual stimulus transformation and the subject's responses, considering that 85% of the first reported forms coherent with the visual input occurred within the first two seconds after the video switches. Note that the delay between the internal decision about a transformation and the overt expression of this transformation, though probably small, artificially increases the amount of time during which the percept does *not* appear to correspond to the video stimulus. This artefact is not problematic, since it could only result in *decreasing* the effect observed in Experiment 2 associated with video switches. Therefore, the delta score differences between Experiments 1 and 2, though significant, are likely to have been underestimated.

This difference suggests that, in Experiment 2, the visual salience of the /p/ onset gesture is dramatically increased, hence the subject switches from the perception of

synchronous /p/$_V$ and /s/$_A$ events characteristic of /psɬ/ to the perception of an asynchronous coordination characteristic of /sɬp/. The “pop-out” effect provided by switching the visual stimuli over time might marginally depend on the temporal structure of the visual switches, as suggested by the three-way interaction between the three factors in Experiment 2, including the duration of the video alternation (6-12 or 11-17 s). Moreover, this effect seems to have eliminated the preference for /psɬ/ transformations over /sɬp/ transformations (as observed in the AVi modality in Experiment 1), likely by increasing the probability of switching towards the visual dynamic input, close to a ceiling effect.

The induction effect displayed in this experiment provides a positive answer to the second question posed above and demonstrates that vision is indeed able to control the emergence of verbal transformations over time. This leads us to the third part of this study, in which we tried to better understand how this induction mechanism might proceed, and more precisely what is the visual component of the speech gesture that actually drives the perceptual switch.

EXPERIMENT 3: STUDYING THE VISUAL INDUCTION PROCESS

The two previous experiments have shown that verbal transformations are audiovisual rather than purely auditory. The visual influence is essentially due to the visual salience of the /p/ labial onset gesture, which dramatically increased in Experiment 2. The listener was found to organize the percept around this onset gesture, hence the bias towards /psŁ/ that was observed in Experiment 1. This finding led us to the hypothesis tested in Experiment 3: that vision provides the listener with information about onset events, hence playing the role of a "bootstrap" in verbal transformations. To test the hypothesis, we replaced the monosyllabic /psŁ/ and /sŁp/ sequences with disyllabic /pata/ and /tapa/ sequences, which are characterized by two jaw opening gestures with differing degrees of visual salience: one being a highly visible gesture involving the lips for /p/, and another being a less visible gesture involving the tongue tip for /t/. Our prediction was that if /(...)patapatapa(...)/ audio sequences were combined with visual sequences in which either /p/ or /t/ gestures were visually hidden by some means, the remaining visible opening gestures would provide the subject with a perceptual bootstrap biasing perception towards either /pata/ (if only /p/ gestures were visible) or /tapa/ (if only /t/ gestures were visible). This is the focus of Experiment 3, which consists of two parts: 3A and 3B.

METHOD

Subjects

Two distinct groups of fifteen voluntary healthy subjects (eleven males and four females; mean age ± standard deviation, 24 ± 4 years - ten males and five females; mean age ±

standard deviation, 26 ± 4 years) who had not taken part in the previous experiments participated in Experiments 3A and 3B. All subjects were native French speakers, reported no hearing or speaking disorders, and had normal or corrected-to-normal vision. None of the subjects was aware of the purpose of the experiment.

Phonetic Material

The disyllables /pata/ and /tapa/, combining the bilabial /p/ and coronal /t/ plosives and the open vowel /a/, were used in this experiment. The plosives /p/ and /t/ were selected to provide two opening gestures of different amplitudes and easily distinguishable in the visual modality (Summerfield, 1983). The open /a/ vowel was selected to provide a context maximising the visibility of the CV gestures. Notice that /tapa/ exists in the French lexicon with a low word frequency while /pata/ does not exist, according to the VoCoLex lexical database (Dufour, Peerman, Pallier and Radeau, 2002). Lexical analyses showed that the lexical type and neighbourhood density values were quite similar for /pata/ and /tapa/, although lower for the later sequence (19 vs. 10 and 71 vs. 55, respectively).

Stimuli

Multiple utterances of the repeated /pata/ or /tapa/ sequence were recorded in a soundproof room by the same speaker as was used in Experiments 1 and 2, and with the same recording set-up. The speaker produced each sequence at a speech rate of 520 ms per disyllable, driven by a visual metronome. The speaker was instructed to maintain an even intonation and vocal intensity while producing the syllables. The recorded clips were then edited on a computer using Adobe Premiere software (www.adobe.com). Video was sampled at 25 images/s with a resolution of 720 × 576 pixels The audio signal was sampled at 44.1 kHz and

16-bit resolution.

For Experiment 3A, an individual /pata/ utterance was selected according to a series of acoustic and video criteria (using the Praat software for audio analysis and the ICP software for lip tracking: Lallouache, 1990): same intensity, fundamental frequency, duration and lip opening for the two vowel nuclei (respectively after /p/ and /t/). The /tapa/ utterance was obtained by just reversing the order of the two syllables. Since it was observed by a panel of reviewers that the construction of all stimuli in Experiment 3A from a single /pata/ utterance could bias perception towards /pata/, an individual /tapa/ utterance was selected for Experiment 3B using the same criteria, and a corresponding /pata/ utterance was obtained by reversing the order of the two syllables.

To hide one of the two opening gestures, /p/ or /t/, we chose to avoid tricks based on applying a visual masker (i.e., temporarily blurring or masking the lip movements), in order to avoid the kind of dynamic interruption that is likely to play a role in driving the percept (as was demonstrated in Experiment 2). We preferred to replace the gestures with a stabilized vowel nucleus. Such stimuli were prepared by taking the image around /a/ (following /p/ or /t/) and presenting it as a stable image throughout the next syllable. For both Experiments 3A and 3B, this lead to two video stimuli labelled as /pa#a/ and /ta#a/, # meaning the lack of any consonantal gesture from one vowel to the next. Starting from the course of lip height for /pata/ displayed in Figure 5A, the corresponding course of lip height respectively along the /pa#a/ and /ta#a/ stimuli is displayed in Figure 5B, C, for the video stimuli in Experiment 3A (very similar patterns are obtained for the stimuli in Experiment 3B). Notice that the stabilization of the /a/ images lasts exactly 360 ms for both /pa#a/ and /ta#a/ in both Experiments 3A and 3B, which

results in the same vowel duration for all of these stimuli.

For both Experiments 3A and 3B, four distinct stimuli per audio disyllable (/pata/ or /tapa) were created by inserting the same sequence without pause in an individual movie consisting of 150 repetitions of the disyllable, according to the following modality conditions: audio-only (labeled A), congruent audio-video (labeled AV), audio plus video /pa#a/ (labeled AV-p) and audio plus video /ta#a/ (labeled AV-t). Notice that in AV-p and AV-t stimuli, the visible gesture (/p/ or /t/) was always synchronous with the congruent sound in the acoustic flow.

Insert Figure 5 about here

Procedure

Experiments 3A and 3B were carried out in exactly in the same way, with Experiment 3B being a control for the results of Experiment 3A. The apparatus and procedures were the same as those used in Experiments 1 and 2. For each experiment, the sequence of stimuli was randomised for each subject.

Data Analysis

As in Experiments 1 and 2, the data were analysed by labelling all subjects' reports in the response sound files. The perceptual global stability duration of each form was calculated by summing the time spent perceiving the given form before switching to another one, and all transformations were classified according to whether they corresponded to /pata/, /tapa/ or to another speech form.

Next, the goal was to determine whether /pata/ or /tapa/ would be the subjects' preferred

response. Therefore, rather than focussing (as in Experiments 1 and 2) on the congruence between audio stimuli and perceptual responses, we selected an index that directly provided the difference between the amount of time spent in the /pata/ vs. /tapa/ perceptual state. For each participant and each condition, delta-scores in Experiment 3 were always computed as the time spent in the /pata/ state minus the time spent in the /tapa/ state. The prediction was that delta scores would be higher in the AV-p condition than in the AV-t condition, with the AV condition possibly being intermediate with respect to the two others. In Experiments 3A and 3B, the effect of the various conditions was assessed using an ANOVA performed on delta scores, with the stimulus (/pata/ or /tapa/, to evaluate a possible initialisation effect) and modality (A, AV, AV-p and AV-t) as within-subject factors. For the analyses, the significance level was set at $p < .05$ and data were Greenhouse-Geisser corrected when appropriate. When required, post-hoc analyses were conducted using Newman-Keuls tests.

RESULTS

Delta scores

The mean relative stability durations for the three categories of reported transformations for the two audio syllables (/pata/ and /tapa/) and in the four modalities (A, AV, AV-p and AV-t) are displayed separately for Experiments 3A and 3B in Figure 6. The reported transformations that were distinct from /pata/ and /tapa/ (i.e., ‘other transformations’) mainly included the substitution or deletion of one phoneme. Their number was larger in the AV-p and AV-t condition than in the AV condition. This is most likely due to the incongruence effect introduced by freezing the video of one consonantal gesture and associating it with the natural sound of the corresponding gesture. Indeed, a large proportion of the “other” responses in the

AV-p condition was of the form /paCa/, C being a consonant other than /t/ (often /k/), while a large proportion of the "other" responses in the AV-t condition was of the form /taCa/, C being a consonant other than /p/.

Insert Figure 6 about here

For Experiment 3A, analysis of delta scores showed a significant effect of the modality of presentation ($F(3,42) = 4.32$, $p < .01$), and of the audio sequence (with delta scores larger for /pata/ than for /tapa/ - $F(1,14) = 8.14$, $p < .05$), but no interaction between the two factors ($F(3.42) = 0.35$). Post-hoc analyses revealed a significantly lower delta score in the AV-t condition (-0.3%) as compared with the AV (24%) and AV-p (19%) conditions ($p < .02$, $p < .04$, respectively). For Experiment 3B, analysis of delta scores showed a significant effect of the modality of presentation ($F(3,42) = 4.53$, $p < .03$), but no effect of the audio sequence ($F(1,14) = 0.19$) and no interaction between the two factors ($F(3.42) = 1.05$). As in Experiment 3A, post-hoc analyses revealed a significantly lower delta score in the AV-t condition (-0.6%) as compared with the AV (18%) and AV-p (19%) conditions ($p < .02$, $p < .03$, respectively).

A further ANOVA was carried out on the data from Experiments 3A and 3B with the stimulus and modality as within-subject factors and the experiment as a between-subject factor. Only the effect of modality ($F(3,84) = 8.75$, $p < .0001$) and audio sequence ($F(1,28) = 5.51$, $p < .03$) were significant. Post-hoc analyses revealed a significantly lower delta score in both the AV-t and A conditions as compared with the AV and AV-p conditions (all p's < .01).

DISCUSSION

The significant difference between the AV-p and AV-t conditions in both Experiments

3A and 3B provides a clear confirmation of our hypothesis, demonstrating that when listeners are presented with an alternation of two syllables, the presentation of only one of these syllables in the visual channel provides a type of “bootstrap” that is very effective at driving the percept towards a disyllable beginning with the seen CV opening gesture. This has potential implications for lexical access that will be addressed in the general discussion.

The AV modality is also associated with larger delta scores than the AV-t modality, but there is no significant difference between the AV and AV-p conditions, which is contrary to our predictions. However, the time course of AV /pata/ and /pa#a/ stimuli (Figure 5A,B) provides a possible explanation. /pata/ is characterized by a major opening gesture for /pa/, while the mouth opening gesture for /ta/ appears as secondary. Therefore, it is not surprising that /pata/ provides a bootstrapping effect that is similar to /pa#a/. Furthermore, the larger number of “other” responses in the AV-p modality compared with the AV modality also contributed to a decrease in delta scores. In fact, including “other” responses of the type /paCa/ (C being a consonant other than /t/) or /taCa/ (C being a consonant other than /p/), there are altogether more responses beginning with a labial consonant in the AV-p condition than in the AV condition, which is more in line with the predictions.

The larger delta scores for the /pata/ audio sequence in Experiment 3A seems to confirm an “initialisation effect”, in which subjects remain longer on the initial syllable providing an intrinsic bootstrap for the perceptual decision. However, this effect is not confirmed by the results of Experiment 3B, suggesting that it is probably rather weak and fragile.

GENERAL DISCUSSION

The present study was designed to test whether visual information from the speaker's articulatory gestures might modify the emergence and stability of verbal auditory percepts during multistable speech perception. Two main results emerge from the three reported experiments.

First, all of the experiments confirmed that vision does penetrate into the Verbal Transformation Effect. This has been demonstrated directly in Experiment 1, through evidence that there were pure visual verbal transformations (though they occurred much less often than audio or audio-visual ones). It has also been demonstrated through the visual influence on auditory verbal transformations, in particular through the decrease in stability of the auditory percept when accompanied by an incongruent visual stimulus (in Experiment 1), and through the shifts in perception from /pata/ to /tapa/ that depended on the visual flow (in Experiment 3). Finally, Experiment 2 demonstrated the ability of the visual input to drive the percept over time, in spite of a fixed audio input.

Second, the visual influence on verbal transformations could be, at least in part, related to a possible visual "bootstrapping" mechanism, which was the focus of Experiment 3. This might be related, as we shall see, to both the general assumptions about an active search mechanism in multistable perception (Leopold, Wilke, Maier & Logothetis, 2002) and to psycholinguistic theories about the role of word onset in lexical access. These two points will be discussed in turn.

Visual and audiovisual verbal transformations: evidence for an intermodal effect

In past research, the multistable perception of speech has mainly been studied as a pure

acoustical/auditory effect (e.g., Warren, 1961; MacKay et al., 1993). The various results obtained in the three present experiments show that multistable speech perception is indeed a multisensory effect. The role played by vision during congruent and incongruent audiovisual presentations in Experiments 1 and 2 sheds some light on multisensory integration. In both experiments, perceptual stability increased or decreased depending on the congruence or incongruence of the visual input with the audio stimulus. Although not significant, the greater stability of congruent audiovisual sequences as compared with audio-only ones in Experiment 1 (77% vs. 66%) is of particular interest, considering the non-ambiguity of the stimuli in the audio-only condition (as confirmed by the fact that the first response of the subjects was always the correct one). This suggests that vision is able to reinforce the auditory stability. Considering that audiovisual perception activates frontal areas more than pure auditory perception does (Skipper et al., 2005, 2007), the greater stability of congruent audiovisual sequences could be due to a stronger involvement of the perceptuo-motor link in the audiovisual modality. This would enable the subject to better integrate all of the available sensory information into a coherent sensori-motor pattern in working memory (Abry et al., 2003; Sato et al., 2004), hence decreasing auditory streaming effects. The lower stability observed during incongruent audiovisual presentation is easily interpretable, considering that stimuli in this case are intrinsically more ambiguous and hence more susceptible to transform. Taken together, these results demonstrate that the multistable perception of speech is indeed multisensory, adding a new item to the long list of speech perception phenomena penetrated by multimodality.

The results of Experiment 2 demonstrated the ability of vision to drive the temporal pattern of speech percepts. Considering that verbal transformations belong to the general

phenomenon of multistable perception, this finding provides new insights into the phenomenon and raises general questions about the possible generalisation of multisensory interactions in this domain. For example, the 'visual induction' effect found in Experiment 2 provides some links with recent findings by Leopold et al. (2002) showing that a visually ambiguous pattern can be stabilized by incorporating periods of silence during the stimulus presentation. In the case of verbal transformations, visual sequences incorporating silences could provide a way to stabilize a speech percept as well. The ability of the visual input to drive the subjects' perceptual switches in time could also be used as a way to track such transformations in the human's brain, as has been done in the context of visual multistability (e.g., Cosmelli et al., 2004). On the other hand, binocular rivalry experiments could be carried out using audio stimuli that are correlated in time with only one of the competing visual inputs, in order to test if the visual percept can be 'driven' by the auditory modality.

The recent neurophysiological evidence in favour of a cortical "dorsal route" (Hickok & Poeppel, 2000, 2004) for speech perception, linking temporal, parietal and frontal regions, and involved both in audiovisual speech perception (Callan et al., 2003; Skipper et al., 2005, 2007) and in verbal transformations (Sato et al., 2004), provides a natural circuit for audiovisual verbal transformations. The fact that this circuit connects perceptual and motor regions, and that motor processes have been implicated both in the verbal transformation circuit (Sato et al., 2006) and audiovisual integration in speech perception (Schwartz et al., 1998; Callan et al., 2003, Skipper et al., 2005, 2007), suggests that such processes are indeed involved in multisensory verbal transformations. In order to connect the finding in the present study of a role for vision in verbal transformations with previously cited works on perceptuo-motor

interactions in speech perception (Sams et al., 2005), further studies could investigate the extent to which self-produced articulatory gestures might intervene in the Verbal Transformation Effect.

A visual bootstrapping effect in verbal transformations

The results of Experiment 3 confirmed our hypothesis about a "visual bootstrapping effect", that is that visual opening gestures contained in the video input could provide the listener with onset events that playing the role of a bootstrap in verbal transformations. This idea certainly needs to be examined in further studies, however it can already be connected to two more general questions. First, the role of the stimulus onset as a pivot for transformation search is reminiscent of the role of acoustic word onset as the pivot for lexical search in psychological models (e.g., Marslen-Wilson's cohort model, 1987). The visual bootstrapping assumption would be compatible with this kind of model if it included the visual flow, however visual input is seldom included in lexical access data and models. This suggests a potential role for visual cues in speech segmentation, which could be related to the role of lipreading in the extraction of focus (Dohen et al., 2004) and the ability to repeat and remember foreign language words embedded in sequences (Davis & Kim, 1998).

Second, this assumption fits well with the general conception of multistable perception developed by Leopold & Logothetis (1999). They suggest that perceptual reversals could be considered to be "changing views" and "related to the expression of a behaviour (rather than) to passive sensory responses" (p. 254). In particular, they propose that transformations are initiated spontaneously, often voluntarily, and under the control of active processes related to brain structures associated with planning and motor programming. This conception fits well

with the view that verbal transformations are under the control of motor processes in a dorsal cortical perceptuo-motor loop (Sato et al., 2004), as well as with the present assumption that a basic and highly visible component of speech gestures - vocal tract opening - could play an important role in this active search.

CONCLUSION

In conclusion, the results of this study show that multistable speech perception is a multisensory effect, where visual information from the speaker's articulatory gestures, together with auditory, phonological and lexical constraints, may modify the emergence and stability of verbal auditory percepts. The ability of vision to influence verbal transformations in their temporal course raises general questions about how and when the human brain integrates input from different senses.

ACKNOWLEDGMENTS

We wish to thank Douglas M. Shiller and three anonymous reviewers for their useful comments on this manuscript. This work was supported by CNRS (Centre National de la Recherche Scientifique) and MIUR (Ministero Italiano dell'Istruzione, dell'Universita e della Ricerca).

REFERENCES

Abry, C, Sato, M., Schwartz, J.-L, Lœvenbruck, H. & Cathiard, M.-A. (2003). Attention-based maintenance of speech forms in memory: The case of verbal transformations. *Behavioral and Brain Sciences*, 26: 728-729.

Abry, C., Cathiard, M.-A., Vilain, A., Laboissière, R. & Schwartz, J.-L. (2006). Some insights in bimodal perception given for free by the natural time course of speech production. In: Bailly, G., Perrier, P. & Vatikiotis-Bateson E. (Eds), *Festschrift Christian Benoît*. MIT Press, Cambridge.

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

Bernstein, L.E., Demorest, M.E. & Tucker, P.E. (1998). What makes a good speechreader? First you have to find one. In Campbell, R., Dodd, B. & Burnham, D. (Eds.), *Hearing by eye, II. Perspectives and directions in research on audiovisual aspects of language processing*. Psychology Press, Hove (U.K.), pp. 211-228.

Bernstein, L., Burnham, D., & Schwartz, J.L. (2002). Special session: issues in audiovisual spoken language processing (when, where, and how?). Proc. *ICSLP'2002*, Denver, Colorado, 1445-1448.

Bernstein, L.E., Auer, E.T. Jr & Moore J.K. (2004). Audiovisual speech binding: Convergence or association. In: Calvert GA, Spence C, Stein BE (Eds.): *Handbook of multisensory processes*, pp 203-224. MTI Press, Cambridge, MA.

Callan D.E., Jones, J.A., Munhall, K., Callan, M.A., Kroos, C. & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures.

Neuroreport, 14(17): 2213-2218.

Cosmelli, D., David, O., Lachaux, J.P., Martinerie, J., Garnero, L., Renault, B. & Varela, F. (2004). Waves of consciousness: ongoing cortical patterns during binocular rivalry. *NeuroImage*, 23: 128– 140

Davis, C. & Kim, J. (1998). Repeating and remembering foreign language words: Does seeing help? *Proceedings of AVSP'98*, Sydney, Australia, pp. 121-125.

Dohen, M., Loevenbruck,H., Cathiard, M.A., & Schwartz, J.L. (2004). Audiovisual Perception of Contrastive Focus in French. *Speech Communication*, 44: 155-172.

Dufour, S., Peereman, R., Pallier, C. & Radeau, M. (2002). VoCoLex: A lexical database on phonological similarity between French words. *L'Année Psychologique*, 102: 725-746.

Fowler, C.A. & Rosenblum, L.D. (1991). The perception of phonetic gestutres. In I.G. Mattingly and M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception*. Hillsdale, NJ: Erlbaum, 33-59.

Green, K.P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. In Campbell, R., Dodd, B. & Burnham, D. (Eds.), *Hearing by eye, II. Perspectives and directions in research on audiovisual aspects of language processing*. Psychology Press, Hove (U.K.), pp. 3-25.

Hickok, G. & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Science*, 4(4): 131-138.

Hickok, G. & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, 92: 67-99.

Lallouache, M.T. (1990). Un poste 'visage-parole'. Acquisition et traitement de contours

labiaux. *Actes des XVIIIèmes Journées d'Études sur la Parole*, Montréal, pp. 282-286.

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

Leopold, D.A. & Logothetis, N.K. (1999). Multistable phenomena: Changing views in perception. *Trends in Cognitive Sciences*, 3: 254-264.

Leopold, D.A., Wilke, M., Maier, A. & Logothetis, N.K. (2002). Stable perception of visually ambiguous patterns. *Nature Neuroscience*, 5: 605-609.

Luce, P.A., Pisoni, D.B. & Goldinger, S.D. (1990). Similarity neighborhoods of spoken words. In G.T.M. Altman (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives*. Cambridge, MIT Press, pp. 122-147.

MacKay, D.G., Wulf, G., Yin, C. & Abrams, L. (1993). Relations between word perception and production: New theory and data on the verbal transformation effect. *Journal of Memory and Language*, 32: 624-646.

MacLeod, A. & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, 21: 131-141.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word recognition. *Cognition*, 25: 71-102.

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

Massaro, D. W. (1989). A Fuzzy Logical Model of Speech Perception. In D. Vickers and P.L. Smith (Eds.), *Human Information Processing: Measures, Mechanisms, and Models* (pp. 367-379). Amsterdam: North Holland.

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

Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I.P., Joensuu, R., Autti, T. & Sams, M. (2005). Processing of audiovisual speech in Broca's area. *NeuroImage*, 25: 333-338.

Pekkola, J., Laasonen, M., Ojanen, V., Autti, T., Jaaskelainen, L.P., Kujala, T. & Sams, M. (2006). Perception of matching and conflicting audiovisual speech in dyslexic and fluent readers: an fMRI study at 3T. *NeuroImage*, 29(3):797-807.

Pitt, M. & Shoaf, L. (2001). The source of a lexical bias in the Verbal Transformation Effect. *Language and Cognitive Processes*, 16(5/6): 715-721.

Pitt, M. & Shoaf, L. (2002). Linking verbal transformations to their causes. *Journal of Experimental Psychology: Human Perception and Performance*, 28: 150-162.

Porter, R.J., Jr & Castellanos, F.X. (1980). Speech-production measures of speech perception: rapid shadowing of VCV syllables. *Journal of Acoustical Society of America*, 67(4): 1349-1356.

Porter, R.J., Jr & Lubker, J.F. (1980). Rapid reproduction of vowel-vowel sequences: evidence for a fast and direct acoustic-motoric linkage in speech. *Journal of Speech and Hearing Research*, 23(3): 593-602.

Reisberg, D., McLean, J. & Goldfield, A. (1987). Easy to hear but hard to understand: A lipreading advantage with intact auditory stimuli. In Campbell, R. & Dodd, B. (Eds.), *Hearing by eye: The psychology of lipreading*. Lawrence Erlbaum Associates, London (U.K.), pp. 97-113.

Reisberg, D., Smith, J.D., Baxter, A.D. & Sonenshine, M. (1989). "Enacted" auditory images

are ambiguous; 'pure' auditory images are not. *Quarterly Journal of Experimental Psychology,* 41A: 619-641.

Robert-Ribes, J., Schwartz, J.-L., Lallouache, T. & Escudier, P. (1998). Complementary and synergy in bimodal speech: Auditory, visual and audiovisual identification of French oral vowels. *Journal of Acoustical Society of America*, 103: 3677-3689.

Sams, M., Möttönen, R. & Sihvonen, T. (2005). Seeing and hearing others and oneself talk. *Cognitive Brain Research*, 23: 429-435.

Sato, M., Baciu, M., Lœvenbruck, H., Schwartz, J.-L, Cathiard, M.-A., Segebarth, C. & Abry, C. (2004). Multistable representation of speech forms: An fMRI study of verbal transformations. *NeuroImage,* 23(3): 1143-1151.

Sato, M., Schwartz J.-L., Abry, C., Cathiard, M.-A. & Lœvenbruck, H. (2006). Multistable syllables as enacted percept: A source of an asymmetric bias in the verbal transformation effect. *Perception & Psychophysics* 68 (3): 458-474.

Schwartz, J.L., Robert-Ribes, J. & Escudier, P. (1998). Ten years after Summerfield ... A taxonomy of models for audiovisual fusion in speech perception. In R. Campbell, B. Dodd & D. Burnham (eds.) *Hearing by eye, II. Perspectives and directions in research on audiovisual aspects of language processing* (pp. 85-108). Hove (UK): Psychology Press.

Schwartz, J.L., Abry, C., Boë, L.J. & Cathiard, M.-A. (2002). Phonology in a Theory of Perception-for-Action-Control. In J. Durand and B. Laks (Eds.) *Phonetics, Phonology, and Cognition*. Oxford: Oxford University Press, 254-280.

Schwartz, J.L., Boë, L.J. & Abry, C. (2006). Linking the Dispersion-Focalization Theory (DFT) and the Maximum Utilization of the Available Distinctive Features (MUAF) principle in a Perception-for-Action-Control Theory (PACT). In M.J. Solé, P. Beddor & M. Ohala (eds.) *Experimental Approaches to Phonology*. Oxford: Oxford University Press (to appear).

Shoaf, L. & Pitt, M. (2002). Does node stability underlie the verbal transformation effect? A test of node structure theory. *Perception & Psychophysi*cs, 64(5): 795-803.

Skipper, J.I., Nusbaum, H.C. & Small, S.L. (2005). Listening to talking faces: Motor cortical activation during speech perception. *NeuroImage*, 25: 76-89.

Skipper, J.I., van Wassenhove, V., Nusbaum, H.C. & Small, S.L. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*.

Smith, J.D., Reisberg, D. & Wilson, M. (1995). The role of subvocalization in auditory imagery. *Neuropsychologia*, 11: 1433-1454.

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

Summerfield, Q. (1983). Audio-visual speech perception, lipreading and artificial stimulation. In M.E. Lutman & M.P. Haggard (Eds.), *Hearing science and hearing disorders* (pp. 131-182). Academic Press, London.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell (Eds.), *Hearing by eye: the psychology of lipreading* (pp. 3-51). Lawrence Erlbaum Associates, London.

Warren, M.R. & Gregory, R.L. (1958). An auditory analogue of the visual reversible figure. *American Journal of Psychology,* 71: 612-613.

Warren, M.R. (1961). Illusory changes of distinct speech upon repetition – The verbal transformation effect. *British Journal of Psychology,* 52: 249-258.

FIGURES & TABLES

Table 1. Mean number of transformations observed in Experiment 1A for /psɬ/ and /sɬp/ audio sequences in each of the four presentation modalities. For each value, the standard error of the mean is indicated.

Modality	/psɬ/	/sɬp/
Audio-only	7.00 (2.02)	9.47 (3.23)
Video-only	2.80 (1.03)	3.07 (0.90)
Audio-Video	6.67 (1.84)	7.73 (2.44)
Audio-Video incongruent	13.67 (5.22)	8.07 (2.20)

Figure 1A. Experimental design in Experiment 1. Temporal alignment between audio and video tracks for /psɬ/ and /sɬp/ sequences. For both utterances, the consonantal onset occurs in the 5th image of the movie (see text for details).

Figure 1B. Audio-visual content of the stimuli. Left, acoustic waveform along with the spectrogram for /psɬ/ and /sɬp/ sequences; F0: fundamental frequency; I: acoustic intensity. Right, corresponding lip area variations. Arrows point to audio and video events associated with the consonants /s/ and /p/ (larger arrows point to the events likely to be combined in the audiovisual speech stream, see text).

Figure 2. Mean relative stability durations of the reported transformations (i.e., /psɬ/, /sɬp/ or 'others') observed in Experiment 1 for (A) /psɬ/ and (B) /sɬp/ audio sequences in each of the four presentation modalities. Error bars represent standard errors of the mean.

Figure 3. Experimental design in Experiment 1B, displaying the time course of audio and video tracks for the /psɬ/ sequence. A video track corresponding to an alternation, randomized in time, of /psɬ/ and /sɬp/ sequences is dubbed without pause onto a cycling fixed audio /psɬ/ (see text for details).

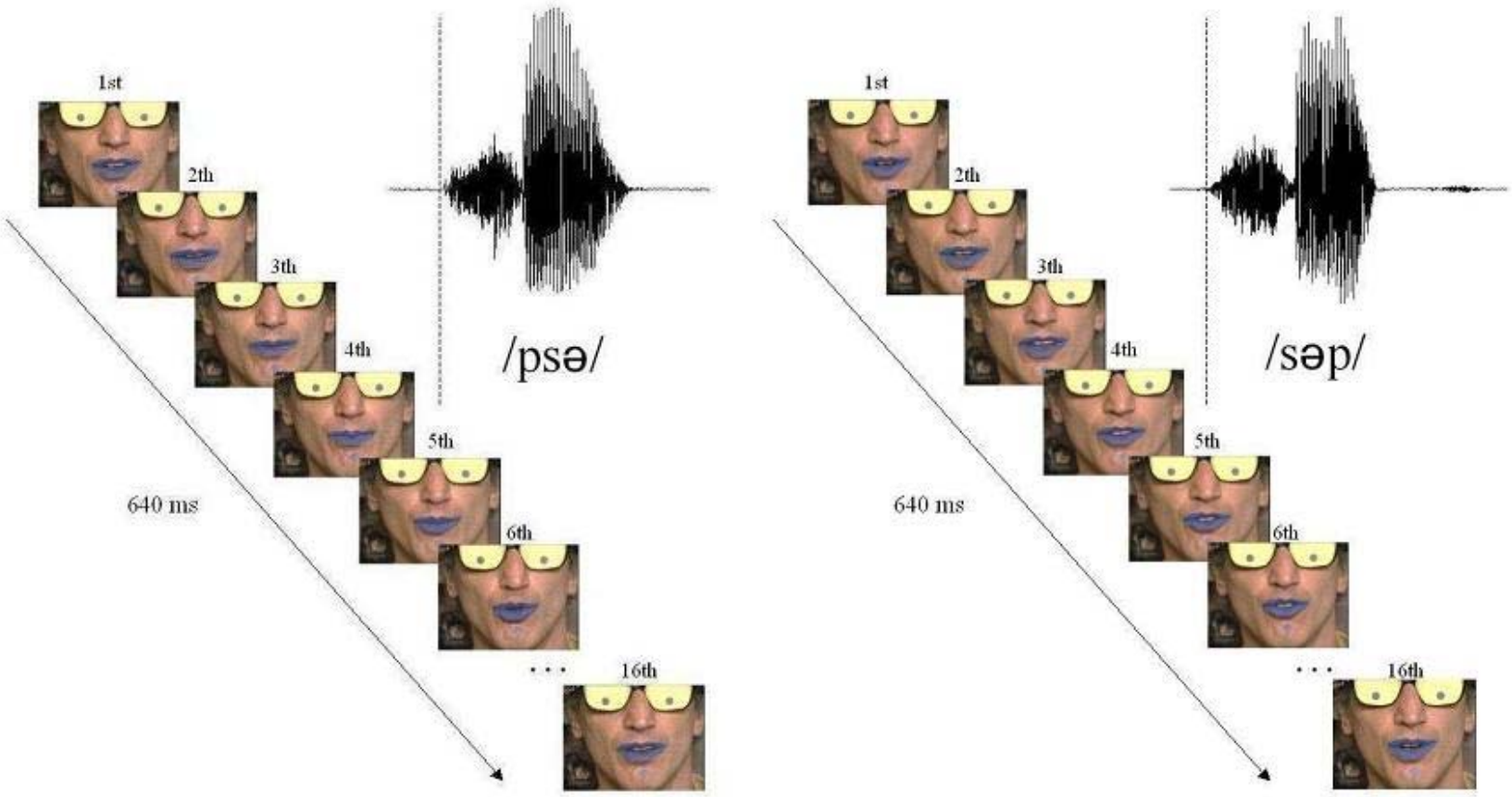
Figure 4. Mean relative stability durations of the reported transformations (i.e., /psɬ/, /sɬp/ or

‘others’) observed in Experiment 2 for (A) /psɬ/ and (B) /sɬp/ audio sequences in congruent and incongruent visual contexts. Error bars represent standard errors of the mean.

Figure 5. Lip height trajectories for the /pata/ sequence (A), and the transformed stimuli used in the AV-p (B) and AV-t (C) conditions in Experiment 3. Stimuli in the AV-p and AV-t conditions were prepared by taking the images around the /a/ after /p/ for /pa#a/ or after /t/ for /ta#a/ and stabilizing them throughout the following syllable (see text for details).

Figure 6. Mean relative stability durations of the reported transformations (i.e., /pata/, /tapa/ or ‘others’) observed in Experiment 3A for (A1) /pata/ and (B1) /tapa/ audio sequences and in Experiment 3B for (A2) /pata/ and (B2) /tapa/ audio sequences in A, AV, AV-p and AV-t modalities (see text for details). Error bars represent standard errors of the mean.

Figure 1A.


1st
2th
3th
4th
5th
6th
16th
640 ms
/psə/
1st
2th
3th
4th
5th
6th
16th
640 ms
/səp/

A

Figure 1B.

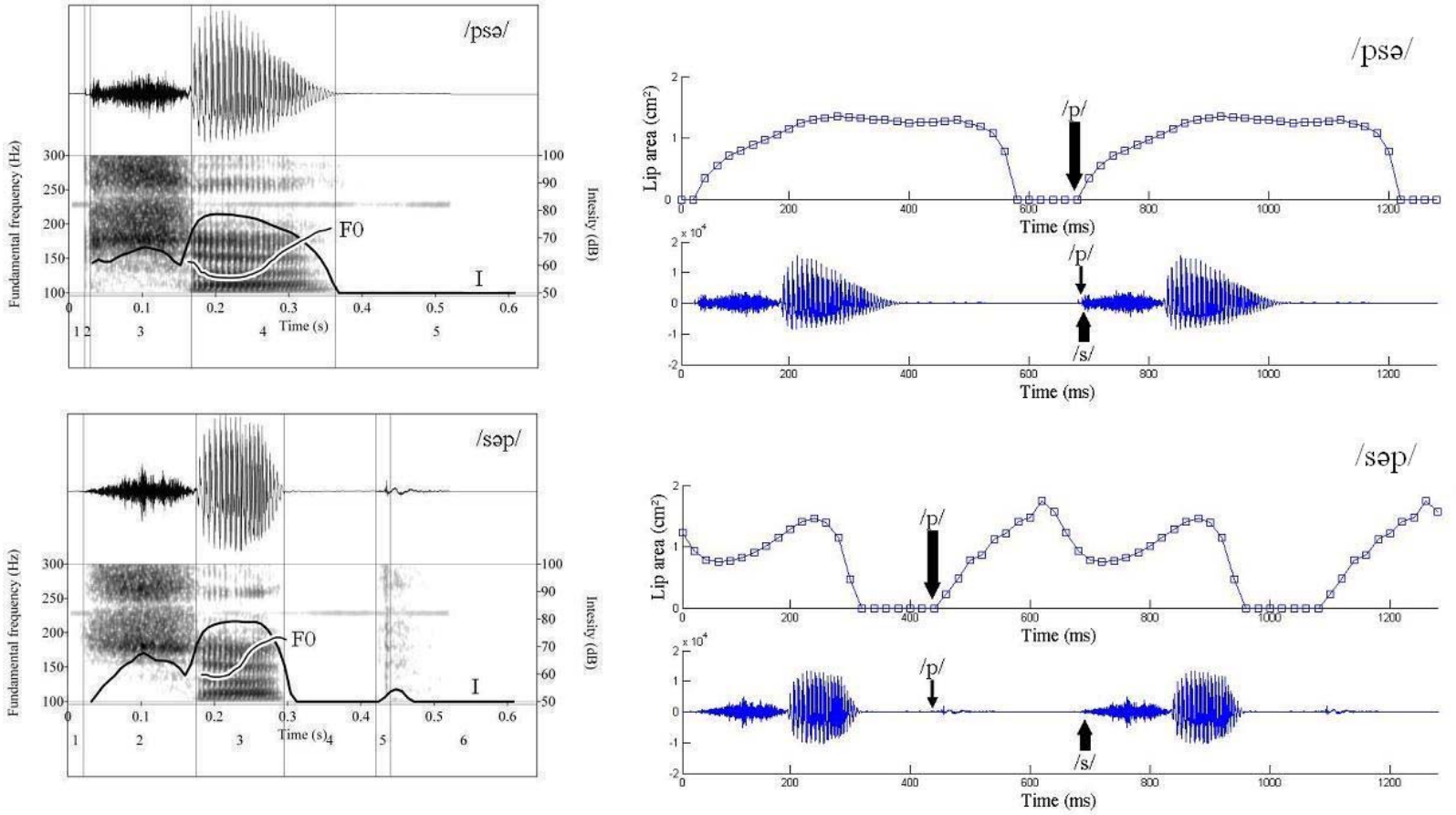

/psə/
/səp/
Fundamental frequency (Hz)
Intesity (dB)
F0
I
Time (s)
Lip area (cm²)
/p/
/s/
Time (ms)

Figure 2.

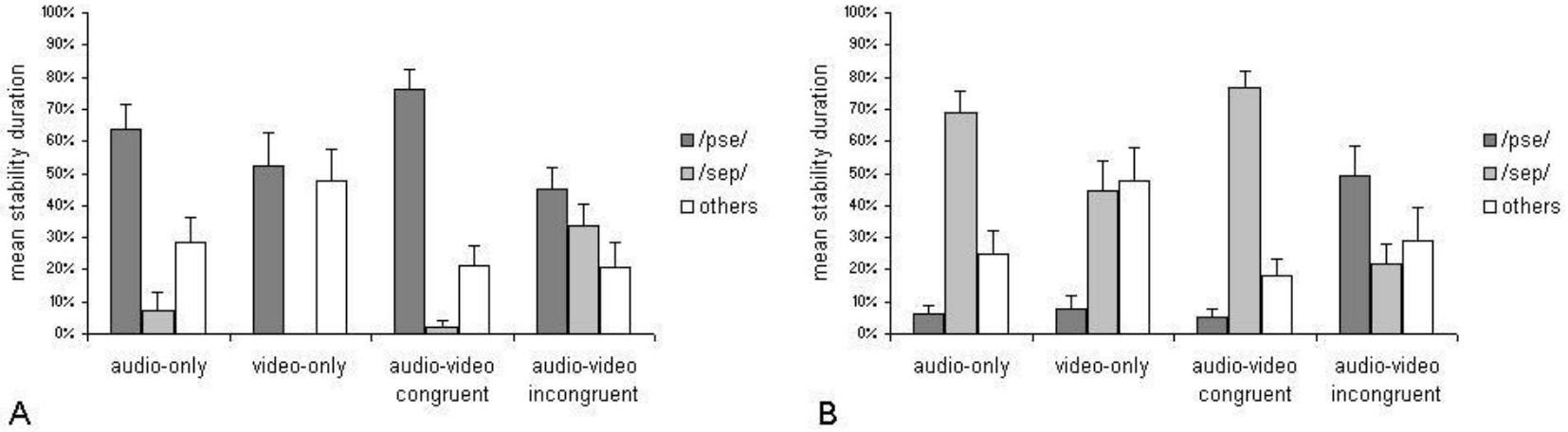

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
mean stability duration
audio-only
video-only
audio-video congruent
audio-video incongruent
/pse/
/sep/
others
A
B

Figure 3.

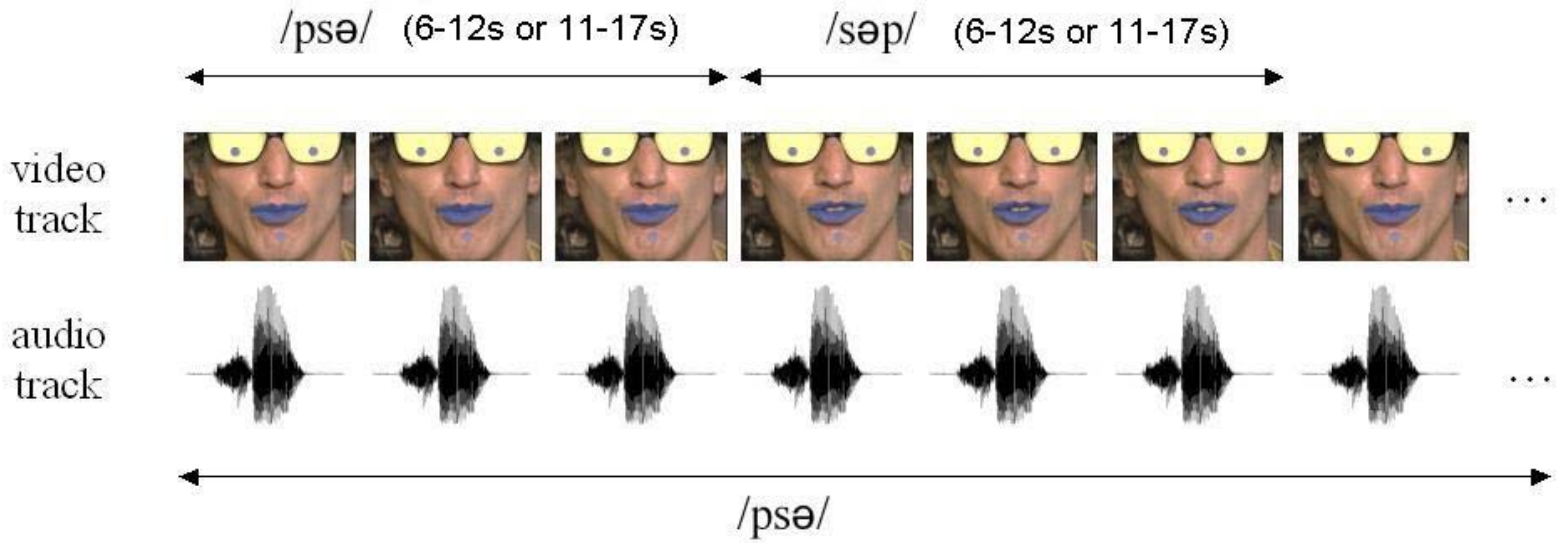

/psə/ (6-12s or 11-17s)
/səp/ (6-12s or 11-17s)
video track
audio track
/psə/

Figure 4.

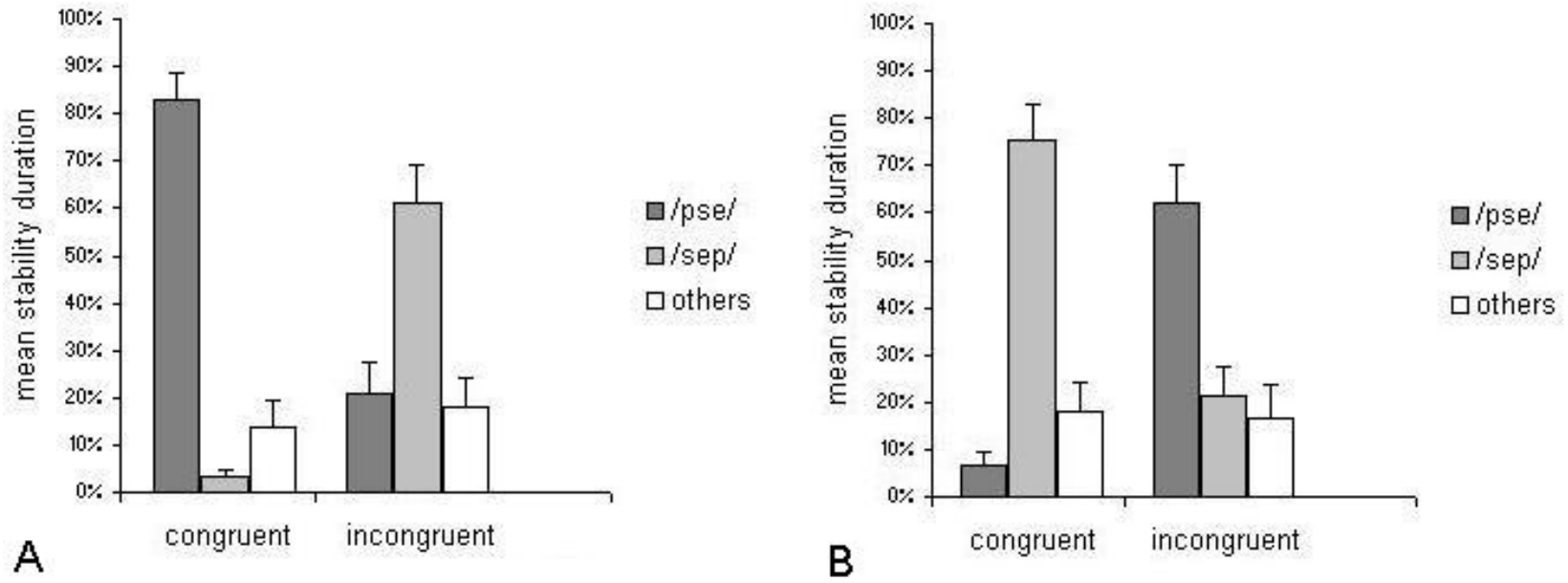

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
mean stability duration
/pse/
/sep/
others
congruent
incongruent
A
B

Figure 5.

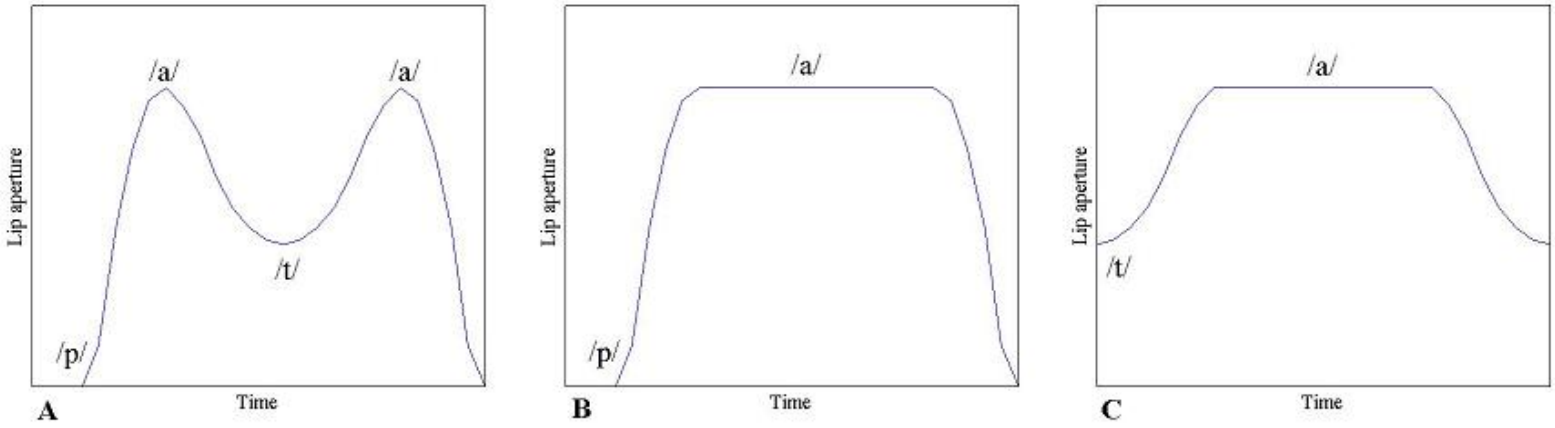
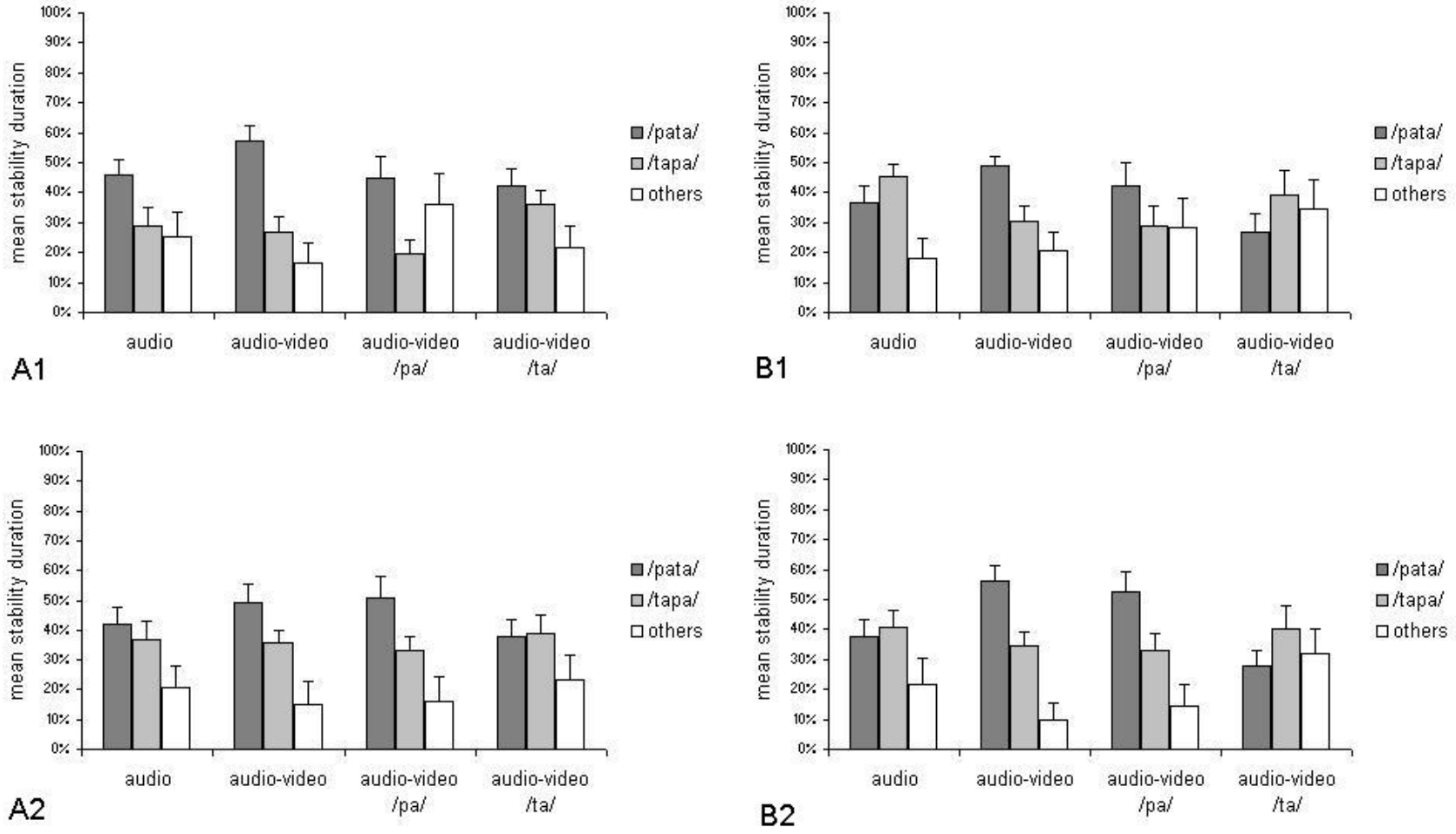

/a/
/a/
/t/
/p/
Lip aperture
Time
A
/a/
/p/
Lip aperture
Time
B
/a/
/t/
Lip aperture
Time
C

Figure 6.


mean stability duration
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
/pata/
/tapa/
others
audio
audio-video
audio-video /pa/
audio-video /ta/
A1
B1
A2
B2