SPEECH STRUCTURE DECISIONS FROM SPEECH MOTION COORDINATIONS

Marie-Agnès Cathiard & Christian Abry

Université Stendhal - BP25 - 38040 GRENOBLE Cedex 9 FRANCE

marieagnes.cathiard@u-grenoble3.fr, christian.abry@u-grenoble3.fr

ABSTRACT

Argument. Ten years ago we wrote a caveat against the view that speech is essentially a kinematic phenomenon, implying exclusively motion representations [13]. Since, mainstream studies have steadily forgotten the evidence, coming from static or stationary phases in the "elastic speech" flow, that these phases can give direct access to speech structures at their best. Moreover whenever the *Structure-from-Motion* paradigm, or the Point-Light technique, were used, they did not seem to lead to the awareness that the very name of this Structure-from-Motion problem is a telltale sign that motion is just there for recovering structures, in cases where they could be underspecified or undersampled. We will show that when combining two classical paradigms in perception, this Structure-from-Motion plus *Multistability*, we reinforce the claim that changes in the perceiver's mind, regarding stationary or repetitive audio(visual) speech moving displays, are perceptual decisions on changes in structure, rather than simple low-level decisions on changes in motion *per se* (say direction). Finally, the outcome of the quest for speech structure recovery is that, contrary to other perception domains, where scientists are struggling in search of stabilizing biases, the very time unfolding of speech coordinations –and non-human primate calls– gives for free neural control biases within their natural integrative time-windows.

Keywords: Audiovisual Speech Production and Perception, Multistability, Working Memory.

1. INTRODUCTION TO SECTIONS 2-5

The backbone of our contribution is the link between audio-visual speech *structures in motion* and speech *perceptual decisions*, building upon two long-standing paradigms imported later in speech research: *Structure-from-Motion* (SfM), with Point-Light technique, and *Multistable Perception*. Our "shape-from-motion" (SfM) stance for AV Speech has been steadily defended since the mid 90's [13]: structure recovery needs motion *only when structure (shape) is underspecified* (section 3). A new McGurk effect, the "Power McGurk illusion" will reinforce the claim that, in order to change the representation of a purely *moving* glide to a fully represented consonant, a variation toward exemplars with a more stable constriction *state* phase is needed (section 4).

Our interest in multistable speech perception appeared at the beginning of this century [5]. From this paradigm we argued that the best possible soft entry in speech perception with motor theoretical issues, is an *enactment* approach (as fostered by [34]), which allows to stabilize swiftly low-level speech structures (as well as representations or memories sensitive to these low levels). And this is an advantage compared to the controversial state of the elder vision field, where neuroscientists are still in the quest for stabilizing classical multistable patterns. Regarding more specifically the bimodality issue of this session, we will read fresh results with the claim that AV speech does not help to evidence the best synergetical motion coordinations, the ones which bias speech structures toward the simplest control stability (section 2). But vision offers in speech a powerful bias for "destructuring" an initially stable audio structure towards a less stable one, giving a new AV (meta)stability, which could not happen without visibly salient articulatory events. Finally we will reemphasize that speech structure, with its natural unfolding of coordinated events, is the best cradle for the time course of perception. And we will propose that the anticipatory nature of certain non-human primate calls –with a seamless *co-modality*, *e.g.* display of vision first then sound, of

oro-facial visible articulation, then inner phonation, letting ultimately hear inner articulation (via formants), *i.e.* oro-laryngeal coordination– is fully relevant for behavioural and neural human studies of AV integration under natural timing and configurational congruence (section 5).

2. (MULTI)STABILITY OF AV SPEECH STRUCTURE-FROM-MOTION : THE "STABIL-LOOP"

The Nobel prize Francis Crick outlined a year before his death "A framework for consciousness" [15], where stimulus blending and rivaling were the stars. Sudden awareness of a change in *perceptual decision* about the structure in depth of the same stimulus (e.g. a Necker's cube), occurs when an illusory dot cylinder or sphere switches in its rotation direction. Neuroscientists are able to manipulate stimuli in a motion disparity continuum, and they place their electrodes in the MT complex or STPa, where structure from motion has been extensively studied ([8] [27], etc.). But even for such a controlled motion illusion, as for binocular rivalry, their neural models became more and more complex [22] [37].

Beyond these most mastered phenomena, nobody would have bet to find the neural secret for what occurs when you repeat continuously "life" and suddenly find a "fly" in your mind, a classical case of enactment. This is verbal bistability, exemplified 30 years before the Warren & Gregory's [40] *Verbal Transformation (Effect)*, by Stetson [35]. We decided to add two issues to this VT. First take seriously the proposal coming from Reisberg [34], that the network recruited for VT would be the one and the same as for *verbal working memory*. Reisberg was right. In an fMRI study [29] we demonstrated for the first time the use of the *articulatory loop* circuitry for the VT (a finding basically replicated by an oncoming experiment [19]). Our second addition was to explore the unaddressed asymmetry phenomenon: if "fly" is more reluctant to give back "life", why? Selecting properly the *articulatory loop* as the locus of a neural control bias in speech motion coordination, we were able to reject all other possible explanations of the finding that the recurrent winner-take-all of the 6 possible syllables made of one vowel (schwa) and two consonants ([p] and [s]), was [psə] [31]. Briefly said, [psə] is the most *in-phase* coordination of the vowel gesture with the consonants (already coarticulated before [p] release), and [s] is ready-made to hiss within [p] (compare [əsp], the most out-of-phase as to the vowel and the consonants). This is not to say that *psi* is the optimal syllable worldwide: [p] is here obviously not acoustically salient. But from a control point of view, if a human being acquired this skill (neither English, nor Spanish), you can test it, and find the optimum coordination is [psə], over [səp], [pəs], [spə], [əps], and [əsp], all structures attested in French phonotactics (like you can test the optimum gait among the 3 regimes of a pony, 2 in a human, or among the many skills of a pianist, etc.).

In a study to be published, Sato et al. [30] addressed again this paradigm, adding vision. In comparing [psə] *vs.* [səp], the first remained the winner in stability. Congruent AV did not enhance audio VTs (as expected in a perceptual task, with no overt enactment). But when dubbing visual [səp] on audio [psə], it was possible to destabilize this latter most stable structure, i.e. to drive it with the help of the visual support toward [səp]: for doing that, the AV timing recoherence of the [p] release event was decisive.

Since Leopold et al. [20] addressed the issue of stable perception for multistable visual patterns, many proposals have challenged the common nature of sustained multistability *vs.* initial dominance (see recently [11]). In spite of eye movement control, and of the reputedly decisional fronto-parietal FEF-LIP link, vision is not as obviously enacted as speech can be. The demonstration just reported of an easy driving from a stable to a less stable (metastable) structure, by just retiming the coherence of AV decisive events, is an advantage over non basically enacted vision phenomena. And this fits well with the explanation of VT asymmetries –in the framework of an articulatory loop working memory, which can host motion coordinations– by a control stance (optimal phasing). That is what we dubbed the *Stabil-Loop* [6], as a to-be-worked-out piece in a still too general speech Perception-for-Action Control Theory [32] [33].

3. MOTION AND STATIONARITY FOR SPEECH STRUCTURE RECOVERY

As regards specifically the bimodal vowel timing, one of the strongest pieces of natural counterevidence against the claim that «time-

varying information is primary» [28, p. 76] comes from the very temporal organization of visible and audible vowel information in speech [13]. What we observed in our French articulatory-acoustic data was that, in initiating an utterance, after a pause, typically with an initial vowel, the first glottal pulse occurred at or nearly at the point where the articulatory setting of the desired vocalic configuration of the vocal-tract was achieved. This means that it is only when this state is reached that the acoustics of the vowel is triggered. If the featural/gestural information of the vowel had to take advantage of the dynamics of the gesture towards its target, the glottal excitation would have to be initiated as soon as possible, that is during the transitional gliding phase, just in order for it to be heard. But this is clearly not what the natural temporal organization of speech reveals. This is why speech can be seen before it is heard.

The extent of the perceptual benefit of this on-gliding phase is predicted by our *Movement Expansion Model (MEM)*, which was based primarily on anticipatory rounding data for French, recently extended to English speaking adults [24], and tested along its acquisition by French children [25]. Abry et al. [4] discussed the corresponding perceptual predictions of the MEM. What was clear was that the identification curves on the anticipatory phase showed a motion benefit only for front views, but not for profile ones [13], where rounding is *fully specified*. The explanation we gave was that rounding in front views had to be recovered by *shape-from-shading* and mostly via *shape-from-motion-in-depth*. It is important to underscore again that at the completion of the climax phase, i.e. with a sufficiently small lip area –which can be held as a static phase–, the identification scores are at their ceiling values.

This natural configurational and temporal coherence of anticipation has been tested via a step-by-step desynchronization procedure [4]. We wanted to avoid global desynchronization, for which AV speech is known to be very robust (since [10]; but see below [23], for the consequences on frontal brain activity). And we were aware of preceding results showing seemingly no sensitivity of vowels to desynchronization ([21]; again see below [26], for the same frontal consequences). Hence we demonstrated that a categorical switch from [y] to [i] can be obtained when making the acoustic [y] vowel begin ahead of the visual anticipatory [i]/[y] boundary, which occurs about the center of the on-gliding phase (for a discussion see [2]).

4. A NEW "POWER" McGURK: EVIDENCE FOR A STATIC PHASE WEIGHT IN THE BIRTH OF A CONSONANT REPRESENTATION

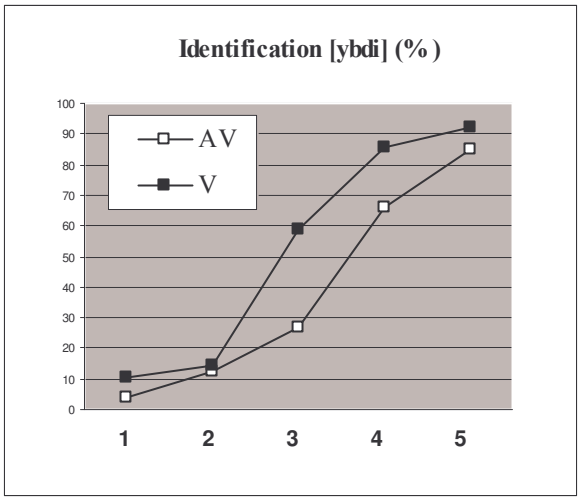
How do previously unnoticed glides bloom as consonants into the mind? How do they stabilize as phonological controls? As a reflex from Latin *potere*, via Old French *poër*, English *power* has not developed a true labiodental fricative (see the diphthongal glide variants, still illustrated by *flour* and *flower*), unlike Modern French *pouvoir*. This change is ubiquitous in languages. But under what static conditions such a transitional (epenthetic) glide becomes integrated as a new consonant into speech structure representation? This is the issue we addressed experimentally, within what we dubbed the "Power McGurk" illusion. While analyzing [y] to [i] transitions in French with the *ICP Lip-Shape-Tracking System*, we were able to observe the three following phases, typically: (i) the lips are constricted and pursed in the [y] steady phase; (ii) then they retract *with a resulting narrowing of the constriction* (between-lips area values, possibly as small as ½ mm2, were accurately measured with our "Deep Blue make-up"); (iii) before the lips finally open and reach the [i] steady retraction state. The second transition phase gives the evidence of a true labiopalatal glide [ɥ] (for more details see [3]). This phenomenon was explained within our *2-Component-Vowel Model* [1], where a glide appears when one of the two constitutive commands of the vowel (here the pursed *shaping* of the constriction for [y]) is relaxed *before* the change in position (dubbed *placing*), here mainly visible on the lips.

We used this phenomenon in the following design within the McGurk paradigm. When a visual [aba] is dubbed on an audio [ada], the classical coherent percept is [abda] (rather than [adba]). We guessed that if a visual [yɥi] (with a more or less close to zero constriction glide) were dubbed on an audio [ydi], the result would be the combination [ybdi].

We recorded from the same talker several sequences containing [i] to [y] transitions in a carrier sentence, like "T'as dit: UDI ise?" (Did you say: UDI ise [pseudo-verb]?). After image

processing we selected four sequences [yi], [yɥi], [ybi] and [ydi]. In order to obtain a pseudo-continuum we ranked them in the following way: [ydi], [yi], [yɥi], [ybi] (with the closure phase reduced by one 40 ms image suppression) and the original [ybi].

Figure 1: [ybdi] identification % for 5 visual stimuli (1 = [ydi], 2 = [yi], 3 = [yɥi], 4 = [ybi] shortened, 5 = [ybi]) presented in visual only condition (V) and audiovisual condition (AV; in this case dubbed on an audio [ydi]). Since this is not a real continuum, we ran an ANOVA (instead of a Probit fitting) [12]



15 French subjects with normal hearing and vision were tested in three conditions (each stimulus repeated randomly 10 times). The audio condition allowed to check that the [ydi] stimulus, on which the visual conditions would be dubbed, was 100% identified as [ydi], when contrasted with audio [ybdi]. In the visual only condition, subjects saw the 5 stimuli of the pseudo-continuum and had to decide whether it was [ydi] or [ybdi]. Combination [ybdi] answers increased monotonically from [ydi] motion (10%) to [ybi] motion (90%). In the audiovisual condition, with the 5 visual stimuli dubbed on the same [ydi] audio, we expected that, as a whole, the [ybdi] scores would be significantly decreased respective to the visual ones. That is what was found.

The ANOVA with 2 factors (condition and stimulus) revealed a significant effect of stimulus ($F[4, 56] = 91.5$, $p<.01$), of condition ($F[1,14] = 7.9$, $p<0.014$) and a significant interaction effect ($F[4,56] = 4.36$, $p<0.01$). Post-hoc comparisons indicated a significant difference between visual and audiovisual scores exclusively for the [yɥi] stimulus (n°3). In vision only, this [yɥi] was identified at 58% as [ybdi]: so this glide was perceived above chance as a [b] consonant. But in the audiovisual condition, the same glide, with 26% [ybdi] responses only, was no more integrated as a consonant. To sum up, combination percept was clearly possible only with the shortened (n°4) and the original visual [ybi] (n°5), which are above 50%. This means that a transitional movement, a glide, between two vowels can give birth to a consonant, only if it displays a sufficient closure duration (n° 4-5). In other words only if it is held in a static phase which occurs only for [ybi] sequences, for the original and even for the shortened one. Hence in the audiovisual condition, the static phase has to be more salient in order to give rise to a consonant percept, which occurs for the original [ybi].

5. PERCEPTION IN THE NATURAL TIME-UNFOLDING OF PRIMATE ORO-LARYNGEAL COORDINATIONS

Charlie is a young pigtail macaque (*macaca nemestrina*) who has been trained in the team of Leonardo Fogassi at the *Istituto di Fisologia Umana* in Parma. Thanks to the kindness of Gino Coudé we had recently the privilege to attend a training session. Charlie had to produce a “coo” call in order to be fed (voluntary vocalization operant conditioning). First he protruded and constricted the lips; then he had his vocal folds vibrate via a chest pulse (actually we had no information about a possible tongue configuration; and it seems that the onset of his “coos” was more a glottal attack, than any k-type release: more a blowing-a-candle sound [18]). Interestingly, sometimes he failed to get a Bernoulli effect and he produced just a puff of air instead of a “coo”. Coudé et al. [14] found that "Lateral F5 [Broca’s homologue] contains a population of neurons that can control voluntary vocal productions". This mirror neuron quest is still in progress (no connectivity with [36] for AV perception).

Ghazanfar et al. ([16], p. 5007) took advantage of this anticipatory lip motion coordination over the laryngeal signal: “Coos are long-duration, tonal calls produced with the lips protruded […]. As in human speech [4], the onsets of mouth movements during production of rhesus monkey vocal signals precede the auditory component.”

We will take the place to quote the important paragraph where they raised the seminal debate

regarding failures in the literature to integrate AV signal when Gestalt principles in the natural speech structures were violated. "Previous neuroimaging studies of multimodal speech suggested that suppression is especially prominent when the speech tokens from the two modalities are incongruent in identity [9]. The present data, along with recent human neuroimaging data ([41] [7] [38]), suggest that identity incongruence is not a requirement for response suppression. Recently, two human evoked-potential studies have reported that face plus voice integration is represented only by suppressed auditory N100 responses ([7] [38]). This is not supported by our LFP [local field potential] data, in which we found both suppression and enhancement (in fact, more frequently enhancement) to our congruent face plus voice stimuli relative to voice alone. We suggest that the consistently suppressed responses in the N100 component in these human studies are attributable to the very long time interval between the presentation of the face and the voice signal ([7] [38]). In both studies, the time between the appearance of the face and the onset of the auditory signal typically exceeded 500 ms. In our data, enhanced responses were primarily seen when this time interval was <100 ms, and suppressed responses were primarily seen at intervals >200 ms." ([16], p. 5010). Which is close to the natural time lag of the voice in our lip rounding anticipation data [13]. This caveat against the violation of the Gestalt grouping timing law was recalled also some ten years ago by Wallace, Wilkinson and Stein: "Generally, multisensory interactions were evident when pairs of stimuli were separated from one another by <500 ms, and the products of these interactions far exceeded the sum of their unimodal components." [39]. However Sugihara et al. [36] raised some doubts about this issue. In their study (like [23] and [26]), they found more activity for incongruent stimuli than for congruent, especially in the prefrontal region, or in Broca's area (IFG). According to Sugihara et al. [36], there seems to be no unique and straightforward explanation for their very disparate stimuli, hence types of congruence. Ojanen et al. [26] tested configurational vowel incongruence. Whereas Miller & D'Esposito [23] tested temporal incoherence with judgments of AV synchronicity in VCV. Instead of categorical vowel incongruence due to desynchronization of a natural anticipation pattern [4]. Hence, before arguing that frontal understanding of actions (mirror-neuron like) is less sensitive to timing coherence, different coherence types have to be clearly parcellated (for a multisensory brain parcellation together with the state of knowledge on connectivities, see [17]).

Acknowledgments: To Leonardo Fogassi and Gino Coudé, from the Parma group. To our colleagues working on the VT, especially to Marc Sato.

6. REFERENCES

[1] Abry, C., Cathiard, M.-A., Laboissière, R., Loevenbruck, H., Payan, Y., & Schwartz, J.-L. 1999. Dynamics in vowel and glide: A double-component account of vowel gestures. In Proc. of the Satellite session ("*Dynamics of the Production and Perception of Speech*") of the XIVth ICPhS, San Francisco, 29-36.

[2] Abry C., Cathiard M.-A., Robert-Ribès J. & Schwartz J.-L. 1994. The coherence of speech in audio-visual integration. Commentary on target paper 'Auditory-visual spatial interaction and modularity' by M. Radeau. *Current Psychology of Cognition*, *13:1*, 52-59.

[3] Abry, C., Cathiard, M.-A., Vilain, A., Laboissière, R. Loevenbruck, H., Savariaux, C. & Schwartz, J.-L. 2007. Some insights in bimodal perception given for free by the natural time course of speech production. In E. Vatikiotis-Bateson, P. Perrier & G. Bailly (Eds.), *Advances in auditory-visual speech processing*, Cambridge: MIT Press

[4] Abry, C., Lallouache, M.-T., Cathiard, M.-A. 1996. How can coarticulation models account for speech sensitivity to audio-visual desynchronization? In D. Stork & M. Hennecke (Eds.), *Speechreading by Humans and Machines*, NATO ASI Series F: Computer and Systems Sciences, *150*: 247-255, Springer-Verlag.

[5] 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*, *26:6*, 728-729.

[6] Abry, C., Vilain, A. & Schwartz, J.-L. 2004. "Vocalize to Localize"? A call for better crosstalk between auditory and visual communication systems researchers: From meerkats to humans. In C. Abry, A. Vilain, J.-L. Schwartz (Eds) Special issue: "Vocalize to Localize". *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, 5(3), 313-325.

[7] Besle, J., Fort, A., Delpuech, C. & Giard, M.-H. 2004. Bimodal speech: early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, *20*, 2225-2234.

[8] Bradley, D.C., Chang, G.C. & Andersen, R.A. 1998. Encoding of three-dimensional structure-from-motion by primate area MT neurons. *Nature*, *392*, 714-717.

[9] Calvert, G.A., Campbell, R. & Brammer, M..J 2000 Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr. Biol.*, *10*, 649–657.

[10] Campbell, R. & Dodd, B. 1980. Hearing by eye. *Quarterly Journal of Experimental Psychology, 32*, 85-99

[11] Carter, O. & Cavanagh, P. 2007. Onset rivalry: Brief presentation isolates an early independent phase of perceptual competition. *PLoS ONE*, *4*, *e343*, 5 p.

[12] Cathiard, M.-A., Gedzelman, S., Abry, C. & Loevenbruck, H. 2004. Naissance de la représentation d'une consonne entre les voyelles : Les conditions d'une intégration audiovisuelle. *Actes des XXVèmes Journées d'Etudes de la Parole*, Fès, Maroc, 117-120.

[13] Cathiard, M.-A., Lallouache, M.-T., & Abry, C. 1996. Does movement on the lips mean movement in the mind? In D. Stork & M. Hennecke (Eds.), *Speechreading by Humans and Machines*, NATO ASI Series F, *150*, 211-219, Springer-Verlag.

[14] Coudé, G., Ferrari, P.F., Rozzi, S.; Borelli, E., Bonini, L., Veroni, V., Rodà, F., Rizzolatti, G. & Fogassi, L. 2005. Motor control sensorimotor integration and cognitive functions of the cortical ventral motor areas of the macaque monkey: A mapping study. *Soc. Neurosci. Abs*, *194.2*.

[15] Crick, F. & Koch, C. 2003. A framework for consciousness. *Nature Neuroscience*, *6:2*, 119-126.

[16] Ghazanfar, A.A., Maier, J.X., Hoffman, K.L., and Logothetis, N.K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *J. Neurosci*., *25*, 5004–5012.

[17] Ghazanfar, A.A. & Schroeder, C.E. 2006. Is neocortex essentially multisensory ? *TICS*, *10:6*, 278-285.

[18] Ghazanfar, A.A. Turesson, H.K., Maier, J.X., van Dinther, R., Patterson, R.D. & Logothetis, N.K. 2007. Vocal-tract resonances as indexical cues in rhesus monkeys. *Current Biology*, *17*, 425-430.

[19] Kondo, H.M. & Kashino, M. 2007. Neural mechanisms of auditory awareness underlying verbal transformations. *NeuroImage* (in press).

[20] Leopold, D.A., Wilke, M., Maier, A. & Logothetis, N.K. 2002. Stable perception of visually ambiguous patterns. *Nature Neurosciences*, *5:6*, 605-609.

[21] Massaro, D.W. & Cohen, M.M. 1993. Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communication, 13*, 127-134.

[22] Meng, M. & Tong, F. 2004. Can attention selectively bias bistable perception? Differences between binocular rivalry and ambiguous figures. *J. Vis*., *4*, 539-551.

[23] Miller, L.M. & D'Esposito, M. 2005. Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *The Journal of Neuroscience*, *25:25*, 5884-5893.

[24] Noiray, A., Ménard, L., Cathiard, M.-A., Abry, C., Aubin, J. & Savariaux, C. (2006). Extending the Movement Expansion Model (MEM) for rounding from French to English. *Proc. of the 7th Int. Seminar on Speech Production*, Ubatuba, Brazil, 319-326.

[25] Noiray, A., Ménard, L., Cathiard, M.-A., Abry, C. & Savariaux, C. (2004). The development of anticipatory labial coarticulation in French: A pionneering study. In *Proc. of the 8th ICSLP*, Jeju Island, Korea, *1*, 53-56.

[26] Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, L.P., Joensuu, R., Autti, T. & Sams, M. 2005. Processing of audiovisual speech in Broca's area. *NeuroImage*, *25:2*, 333-338.

[27] Parker, A.J., Krug, K. & Cumming, B.G. 2002. Neuronal activity and its links with the perception of multi-stable figures. *Phil. Trans. R. Soc. Lond. B*., *357*, 1053-1062.

[28] Rosenblum, L.D. & Saldana, H.M. 1998. Time-varying information for visual speech perception. In R. Campbell, B. Dodd &D. Burnham (Eds.), *Hearing by Eye II*, pp. 61-81, Hove: Psychology Press.

[29] Sato, M., Baciu, M., Lœvenbruck, H., Schwartz, J-L., Cathiard, M-A., Segebarth, C. & Abry, C. 2004. Multistable representation of speech forms: A functional MRI study of verbal transformations. *NeuroImage*, *23*, 1143-1151.

[30] Sato, M., Basirat, A. & Schwartz, J-L. (2007). Visual contribution to the multistable perception of speech. *Perception & Psychophysics* (accepted).

[31] Sato, M., Schwartz, J-L., Cathiard, M-A., Abry, C. & 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.

[32] Schwartz, J.-L., Abry, C., Boë, L.-J. & Cathiard, M.-A. 2002. Phonology in a theory of perception-for-action-control. In J. Durand & B. Laks (Eds), *Phonetics, Phonology, and Cognition*, pp. 254-280, Oxford: Oxford University Press.

[33] Schwartz, J.L., Boë, L.J. & Abry, C. 2007. 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 (in press).

[34] Smith, J.D., Reisberg, D. & Wilson, M. 1995. The role of subvocalization in auditory imagery. *Neuropsychologia*, *11*,1433-1454.

[35] Stetson, R.H. 1928. Motor phonetics. A study of speech movements in action. *Archives néerlandaises de phonétique expérimentale*, *t.3*.

[36] Sugihara, T., Diltz, M.D., Averbeck, B.B. & Romanski, L.M. 2006. Integration of auditory and visual communication information in the primate ventrolateral prefrontal cortex. *The Journal of Neuroscience*, *26:43*, 11138-11147.

[37] Tong, F., Meng, M. & Blake, R. 2006. Neural bases of binocular rivalry. *TICS*, *10:11*, 502-511.

[38] van Wassenhove, V., Grant, K.W. & Poeppel, D. 2005. Visual speech speeds up the neural processing of auditory speech. *PNAS*, *102:4*, 1181-1186.

[39] Wallace, M.T., Wilkinson, L.K. & Stein, B.E. 1996. Representation and integration of multiple sensory inputs in primate superior colliculus. *J. Neurophysiol*., *76:2*, 1246-1266.

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

[41] Wright, T.M., Pelphrey, K.A., Allison, T., McKeown, M.J. & McCarthy, G. 2003. Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral Cortex*, *13*, 1034-1043.