The Perception for Action Control Theory (PACT): a perceptuo-motor theory of speech perception

Jean-Luc Schwartz, Anahita Basirat, Lucie Ménard, Marc Sato

The Perception for Action Control Theory (PACT): a perceptuo-motor theory of speech perception

Jean-Luc Schwartz (1), Anahita Basirat (1), Lucie Ménard (2), Marc Sato (1)

(1) GIPSA-Lab, Speech and Cognition Department (ICP), UMR 5216 CNRS – Grenoble University, France
(2) Laboratoire de Phonétique, UQAM / CRLMB, Montreal, Canada

Abstract

It is an old-standing debate in the field of speech communication to determine whether speech perception involves auditory or multisensory representations and processing, independently on any procedural knowledge about the production of speech units or on the contrary if it is based on a recoding of the sensory input in terms of articulatory gestures, as posited in the Motor Theory of Speech Perception. The discovery of mirror neurons in the last 15 years has strongly renewed the interest for motor theories. However, while these neurophysiological data clearly reinforce the plausibility of the role of motor properties in perception, it could lead in our view to incorrectly de-emphasise the role of *perceptual shaping*, crucial in speech communication. The so-called Perception-for-Action-Control Theory (PACT) aims at defining a theoretical framework connecting in a principled way perceptual shaping and motor procedural knowledge in speech multisensory processing in the human brain. In this paper, the theory is presented in details. It is described how this theory fits with behavioural and linguistic data, concerning firstly vowel systems in human languages, and secondly the perceptual organization of the speech scene. Finally a neuro-computational framework is presented in connection with recent data on the possible functional role of the motor system in speech perception.

Keywords: perceptuo-motor interaction; speech perception; vowel perception; perceptual organisation; multisensory interactions; neurocomputational model; dorsal route

Corresponding Author: Dr Jean-Luc Schwartz, PhD
INPG, 961 rue de la Houille Blanche - Domaine universitaire - BP 46
38402 Saint Martin d'Hères Cedex. France
Tel: (+33).4.76.57.47.12 - Fax: (+33).4.76.57.47.10 - http://www.gipsa-lab.inpg.fr/

Introduction

It is an old-standing debate in the field of speech communication to determine whether speech perception involves auditory or multisensory representations and processing, independently on any procedural knowledge about the production of speech units (for a review, Diehl et al., 2004); or on the contrary if it is based on a recoding of the sensory input in terms of articulatory gestures, as posited in the Motor Theory of Speech Perception (Liberman et al., 1962; Liberman and Mattingly 1985; Liberman & Whalen, 2000). The discovery of mirror neurons (for reviews, Rizzolatti et al., 2001; Rizzolatti & Craighero, 2004) in the last 15 years has strongly renewed the interest for motor theories. However, while these neurophysiological data clearly reinforce the plausibility of the role of motor properties in perception, it could lead in our view to incorrectly de-emphasise the role of *perceptual shaping*, crucial in speech communication. The so-called Perception-for-Action-Control Theory (PACT) aims at defining a theoretical framework connecting in a principled way perceptual shaping and motor procedural knowledge in speech multisensory processing in the human brain. In the following, the theory will be presented in details in Section I. Sections II and III will describe how this theory fits with behavioural and linguistic data, concerning firstly vowel systems in human languages, and secondly the perceptual organization of the speech scene. Section IV will consider a neuro-computational framework in connection with recent data on the possible functional role of the motor system in speech perception. The conclusion will address some perspectives in relation with language emergence.

I. The Perception for Action Control Theory (PACT) of speech perception

I.1. Action shapes perception

The Motor Theory of speech perception originated in the 50s (Liberman et al., 1952; Liberman, 1957) to tackle the problem of invariance and variability in relation with coarticulation phenomena (see Perkell & Klatt, 1986). It was developed from the following statement, based on many experiments in speech synthesis. The coarticulation-driven composition of articulatory commands during speech production is non-linearly transformed into a complex composition of acoustic features, so that the acoustic properties of speech sounds are not invariant but context dependent, and the correspondence between sounds and phonemes is far from transparent. However, speech percepts seem to be more directly related to gestures than to sounds. To recall two classical examples: the same acoustic feature, e.g. a stop burst centred at 1400 Hz, may be perceived as a labial [p] or a velar [k] depending on the following vowel, while conversely a single percept [d] may be associated to either a rising or a falling second formant transition whether the following vowel is [i] or [u]. According to Liberman and colleagues (1952,1954), there results are

driven by implicit information about coarticulation mechanisms. This led to the more general claim that speech perception involves recovery of articulatory gestures, or invariant motor commands, capitalising on the knowledge the listener has of speech production mechanisms (Liberman et al., 1967; Liberman & Mattingly, 1985). Recent reviews of the Motor Theory (e.g. Liberman & Whalen, 2000; Galantucci et al., 2006) recapitulate the major experimental arguments in favour of the Motor Theory (or its variant, the Direct Realist Theory of speech perception, Fowler, 1986). They can be summarised in two parts.

The first series of arguments deal with the possible existence of a functional connection between speech perception and production systems, thanks to which the motor system would be accessed on line during the speech perception process. A number of behavioural data support this hypothesis, beginning by two classical studies. In the "selective adaptation" task, Cooper & Lauritsen (1974) showed how reiterated perception of a given stimulus not only affects the perception of following stimuli (the classical selective adaptation paradigm, Eimas & Corbitt, 1973) but also their production. Conversely, repetitive articulation of a speech stimulus also affects the identification of following stimuli (Cooper et al., 1975, 1976). In the "close shadowing" task, Porter & Castellanos (1980) and Porter & Lubker (1980) showed that the time between perceiving a speech item and reproducing it was so short that it suggested a direct selection of the adequate gesture within the perception task itself. A number of further studies, reviewed in Galantucci et al. (2006), also provide behavioural evidence that perceiving a speech sound activates action representations compatible with the input stimulus.

Another body of evidence in favour of a link between speech perception and production comes from recent neurophysiological studies. The discovery of mirror neurons in the macaque's brain and of a putative mirror system in humans (see Rizzolatti et al., 2001; Rizzolatti and Craighero, 2004, for a review) has provided strong evidence for the existence of an action-observation matching system. Mirror neurons are a small subset of neurons, found in the macaque ventral premotor cortex and the anterior inferior parietal lobule, that fire both during the production of goal-directed actions and during watching to a similar action made by another individuals. Kohler et al. (2002) later extended the mirror neuron property to listening to the sound of an action, introducing "audiovisual" mirror neurons in monkeys. The existence of the mirror-neuron system thus suggests that actions may be understood in part by the same neural circuits that are used in action performance. In addition to action understanding, the human mirror-neuron system has been proposed to play a fundamental role in speech processing (e.g., Rizzolatti & Arbib, 1998; Arbib, 2005; Gentilucci & Corballis, 2006) by providing a neurophysiological mechanism that creates parity between the speaker and the listener. In agreement with this view, brain areas involved in the planning and execution of speech gestures (i.e., the left inferior frontal gyrus, the ventral premotor

and primary motor cortices) and areas subserving proprioception related to mouth movements (i.e., the somatosensory cortex), have been found to be activated during auditory, visual and/or auditory-visual speech perception (e.g., Möttonen et al., 2004; Wilson et al., 2004; Ojanen et al., 2005; Skipper et al., 2005; Pekkola et al., 2006; Pulvermuller et al., 2006; Wilson & Iacoboni, 2006; Skipper et al., 2007). Transcranial magnetic stimulation (TMS) studies also demonstrated that motor-evoked potentials recorded from the lips or tongue muscles are enhanced during both passive speech listening and viewing, when stimulating the corresponding area of the left primary motor cortex (Sundara et al., 2001; Fadiga et al., 2002; Watkins et al., 2003; Watkins & Paus, 2004; Roy et al., 2008). Importantly, this speech motor 'resonance' mechanism appears to be articulatory specific, motor facilitation being stronger when the recorded muscle and the presented speech stimulus imply the same articulator (Fadiga et al., 2002; Roy et al., 2008). The specificity of this speech motor resonance mechanism is also suggested by two recent functional magnetic resonance imaging (fMRI) studies showing somatotopic patterns of motor activity in the ventral premotor cortex during both producing and listening to or viewing lips- and tongue-related phonemes (Pulvermuller et al., 2006; Skipper et al., 2007). Perception in this framework would be mediated by "motor ideas" (Fadiga et al., 2000) represented in the brain by sensori-motor neurons enabling to access these "ideas" both through acting and through observing or hearing somebody else acting (though see a more critic view on this point in a recent opinion paper by Lotto et al., 2009).

The second series of arguments deal with the claim that this connection would play a critical role in speech perception. This is the "perceiving speech is perceiving gestures" argument, illustrated by a number of data in which complex links between gestures and sounds due to coarticulation effects seem to be parsed out in perception, as in the previously cited examples about plosives perception. In these various examples reviewed by Galantucci et al. (2006), it seems that the listener, who obviously "knows something" about speech production, exploits this knowledge to disentangle the complexity of the acoustic input and access functional units more directly related with motor commands. The same reasoning is proposed for multisensory perception, e.g. audiovisual (but also audio-haptic) interactions in speech perception, which are claimed to be related to the knowledge the listener has of the multisensory coherence of a speech gesture (Fowler & Rosenblum, 1991; see also Schwartz et al., 1998). This is connected to the general statement by Viviani that perception involves a "procedural knowledge of action", as nicely exemplified by his data showing how the knowledge of the laws of human movement modifies the perception of hand movements, in terms of both configurational and kinematic properties (Viviani & Stucchi, 1992).

In a recent discussion paper at the last *LabPhon* Conference, Schwartz (2009) claimed that the various studies presented in the corresponding session all contained some elements of answers to the question: "what does a listener know about a speaker's gesture, and how is this knowledge

useful for perception?". It was shown how some "procedural knowledge" on vowel reduction (for the paper by Solé & Ohala, 2009) and voicing assimilation (Kuzla et al., 2009) could actually intervene in the course of speech perception. However, one case presented a difficulty. It consisted in the observation by Mielke et al. (2009) that there may be different strategies in the production of an [ɹ] in American English, including "bunched" and "retroflex" lingual configurations (Delattre & Freeman, 1968; Tiede et al., 2004) – partly driven, according to the study by Mielke and coll., to the context in relation with articulatory ease – though the sound is quite similar in both cases. Here comes a problem for the Motor Theory of speech perception: what happens when the relationship between gestures and sounds is many-to-one (various gestures for a single sound), and hence the gesture cannot be, in theory, recovered from the sound without additional pieces of information? This means that at most some abstract representation of the gesture can be recovered, rather than a complete articulatory configuration (see Guenther et al., 1998).

The problem is not only technical, actually, but also theoretical: what makes an [ɹ] an [ɹ] appears to be related to the acoustic content of the unit rather than to its articulatory characteristics. In this case, it seems that the link between perception and action goes the other way round, perception shaping action and not the inverse.

I.2. Perception shapes action

The case of [ɹ], as many others (see e.g. Savariaux et al., 1995, 1999; Perrier, 2006) suggests that a gesture is characterised by its functional value, likely evaluated in auditory terms. This is not easy to deal with in the framework of the Motor Theory, apart if one conceives a vocal tract configuration as a global coordination oriented towards a functional phonetic goal (Mattingly, 1990).

The solution out of this difficulty is to acknowledge that gestures are not pure motor units, but rather perceptuo-motor units, gestures being *shaped* by perception. Let us take a first example, that is lip rounding. Consider what it means to round the lips starting from [i]. Decreasing the lip area from the unrounded configuration first does not change the sound almost at all. Then, it suddenly dramatically changes it into an [y]-like sound, because of both acoustic and auditory reasons (Abry et al., 1989). In this example, it should become clear that "lip rounding" does not mean anything per se, and cannot be defined in pure articulatory terms. Rounding the lips means decreasing the lip area at a value lower than the critical value under which the sound changes significantly.

Gestures are not only shaped by perception, but also *selected* in relationship with their perceptual (acoustic-auditory) value. Let us consider vowel systems in human languages. Assuming

that the Motor Theory of Speech Perception is correct, what would be the possible predictions in terms of oral vowel systems in human languages? There are basically three degrees of freedom for producing oral vowels: height, front-back position, and rounding. What would be the best three-vowel system in this space? The system [i a u] is a very good choice, in terms of articulatory dispersion, combining the high front unrounded [i], the high back rounded [u] and the open [a]. It is indeed part of most languages, as shown by the available data on vowel systems in human languages (Maddieson, 1984). However, [y a ɯ] should be as good a choice: it just consists in associating lip rounding with tongue frontness, and lip unrounding with tongue backness. It combines articulatory features differently, but the difference cannot be assessed in articulatory terms. However, this second system never appears in human languages. The reason for this is clearly auditory. Auditory perception is a kind of lateral projection of this 3-D space, in a 2-D (F1, F2) space in which [i u] is much better (in terms of dispersion) than [y ɯ]. Altogether, the prevalence of [i a u] and the absence of [y a ɯ] clearly shows that gestures are selected in relation with their perceptual value (Fig. 1).

Notice that not only does the "audibility" of a gesture intervene in the selection of speech gestures, but also its "visibility", as shown by the largely developed use of the [m]-[n] contrast in human languages, poorly audible, but quite visible (Schwartz et al., 2007).

These two examples can be related to two major theories of speech communication. Stevens's Quantal Theory (Stevens, 1972, 1989) capitalises on nonlinearities of the transformation from articulatory to acoustic/auditory features, which sometimes provide a "quantal" behaviour in the sense that some regions of the motor command do almost not change the sound, while others change it abruptly – as in the lip rounding case. Stevens suggests that this provides natural boundaries for designing phonetic contrasts exploited in human languages such as the [i]-[y] contrast in French. Lindblom's Hyper-Hypo or Adaptive Dispersion Theory (Liljencrants & Lindblom, 1972; Lindblom, 1986, 1990) exploits acoustic-auditory differentiation as a driving force shaping vowel systems (Liljencrants & Linblom, 1972) – and possibly, according to Abry (2003) and Schwartz & Boë (2007), also plosive systems. In both cases, it appears that gestures are indeed perceptuo-motor units shaped by their perceptual properties and selected for their functional/perceptual value for communication.

I.3. Integrating motor and perceptual knowledge inside PACT

What is the state of affairs at this point? If action shapes perception and perception shapes action in return, are we in a never-ending loop? Let us advance step by step.

On the one hand, the involvement of procedural knowledge about speech production in the

course of speech perception is at odds with auditory theories, relying on sensory transduction followed by categorisation processes and assuming that all the information is in the acoustical signal. On the other hand, we argue that the decision process should operate on auditory features rather than articulatory configurations, gestures being shaped by perception and selected in relation with their functional perceptual value. This is at odds with pure motor theories, which consider that perceptual representations should be gestural in nature.

A possible synthesis could be envisioned in the framework of the Theory of Event Coding developed by Prinz and colleagues (see e.g. Hommel et al., 2001) predicting that perceptual contents and action goals are cognitively coded in a common representational medium by composite codes of their distal features. However, there is a specificity of speech actions, which are *communicative actions* driven by the need to generate *communicative stimuli* that support the linguistic code. This is the reason why gestures should be considered as perceptually-shaped units in the case of speech. Altogether, these two principles – that perception involves knowledge of action, and that action aims at perceptually-shaped gestures – provide the basis of PACT.

The Perception-for-Action-Control Theory considers that speech perception is the set of mechanisms that enable not only to understand, but also to control speech, considered as a communicative process. This leads to two consequences. Firstly, perception and action are co-structured in the course of speech development, which involves both producing and perceiving speech items. In consequence, the perceptual system is intrinsically organised in reference to speech gestures, and in relation with the structure of the action system. This is why it includes, in one way or another, an implicit procedural knowledge of speech production. Secondly, perception provides action with auditory (and possibly visual) templates, which contribute to define the gesture, providing them objectives, organisation schemes and functional value. This is how PACT incorporates the "gesture shaping" component.

In PACT, the communication unit, through which parity may be achieved, is neither a sound, nor a gesture, but a perceptually-shaped gesture, that is a perceptuo-motor unit. It is characterised by both its *articulatory coherence* – provided by its gestural nature – and its *perceptual value* – necessary for being functional.

Let us take an example to make the concept hopefully clearer. In the framework of the Frame-Content Theory, MacNeilage (1998) argues that the jaw cycles pre-existing in mastication in ontogeny – and possibly in phylogeny – play a bootstrap role initiating dynamic vocalisations at the birth of speech. The key point is that jaw cyclic gestures are shaped by the acoustic/auditory transformation in a quasi quantal way in Stevens's terms (see Fig. 2). Actually, when the jaw is lowered, the sound is vocalic, while when the jaw is raised, it results in a closure alternating silence and burst, characteristic of a plosive, with possibly an intermediate stage generating a friction noise

characteristic of a fricative. These two non-linearities – from vowel to fricative through the harmonic-to-turbulent switch of acoustic mode, and from fricative to plosive through vocal tract closure – naturally generate a consonant-vowel alternation leading to syllables in all human languages. Syllables in this scenario are neither articulatory nor perceptual structures, but perceptuo-motor in essence: jaw gestures non-linearly shaped by audition in a quantal way.

We shall review in Sections 2 and 3 two series of works we did in the past years in this perspective, leading in both cases to introduce perceptually shaped gestures as a basis for explanation of perceptual processes. Then we shall discuss in Section 4 how recent data about perceptuo-motor connections in the human brain fit with PACT. In light of these data, we shall propose some elements of a neurocomputational architecture and discuss about two possible implementations, involving either implicit articulatory knowledge stored in perceptual areas, or explicit involvement of the speech production system in the perception routine.

II. Vowels as auditorily-shaped gestures

The conception of vowel systems as auditorily-optimized structures obeying perceptual dispersion principles for maximizing their efficacy of communication founded a major break in the Chomskyan conception of an autonomous "language organ". This opened the route towards what Lindblom (1986, 1990) called "substance-based" theories of human languages.

II.1. From DT to DFT, vowel systems are an auditory-driven communication system

The initial push towards such "substance-based" theories of human languages was provided by Liljencrants & Lindblom (1972) who proposed that vowel systems aimed at maximising distances between their formant sets, just as electrical charges in a bounded space would maximise their distances because of repelling Coulomb forces. Liljencrants & Lindblom defined a "Dispersion Theory" (DT) in which a vowel system with a given number of vowels is characterised by a global "energy" summing the inverse squared distances between their F1 and F2 values within an (F1, F2) triangular space bounded by articulatory constraints. They showed that this lead to rather adequate predictions of 3-vowels systems [i a u], four-vowel systems [i e a u], five-vowel systems [i e a o u], compatible with vowel inventories (e.g. Maddieson, 1984). Further evolutions of the theory were proposed by e.g. Lindblom (1984, 1986), Diehl et al. (2003), etc.

Schwartz et al. (1997) added another perceptual driving force according to which vowel acoustic configurations are also attracted towards "focal" configurations in which there is proximity between two consecutive formants, F1 and F2 (e.g. for [u] and [a]), F2 and F3 (e.g. for [y]), or F3 and F4 (typically for [i]). Focal vowels correspond to more stable auditory patterns, which actually seem to drive both infant and adult perception (e.g. Schwartz & Escudier, 1989; Schwartz et al.,

2005) and children production (e.g. Ménard et al., 2004). Focalisation would intervene in shaping vowel systems, just as "focal colours" shape the categorisation of colours in human languages (Rosch-Heider, 1972). This lead to the "Dispersion-Focalisation Theory of vowel systems" (DFT) combining dispersion and focalisation and providing a good fit of vowel systems from 3 to 9 vowel qualities, including best systems and their preferred variants (Vallée et al., 1999).

One important problem however remains in these simulations: vowel systems seem to balance the number of categories in their front unrounded and back rounded series (with e.g. [i e] and [u o] or [i e ɛ] and [u o ɔ]) and also, when they exist, in their non peripheral series such as central or front rounded (e.g. [i e ɛ] and [y ø œ] in French) while no perceptual principle ensures such an equilibrium in the DT or DFT (Schwartz et al., 1997). This regularity in the distribution of vowels within a given system may be related to a principle defined by Ohala (1979) as "Maximal Use of Available Features" (MUAF), according to which systems would combine features in a systematic way. In the case of vowel systems, height features (e.g. high, mid-high, mid-low) would be combined with tongue/lip configuration features (e.g. front unrounded or back rounded) to provide balanced systems.

II.2. The distribution of height controls, a motor-driven idiosyncrasy

In a recent paper, Ménard et al. (submitted) observed a curious idiosyncrasy in the production of vowel height in French or Canadian French. A first aspect of this idiosyncrasy is that subjects vary a lot in the way they distribute vowels along the F1 dimension (Fig. 3a). Some speakers may produce high and mid-high vowels (e.g. [i] and [e]) quite close together and mid-high and mid-low (e.g. [e] and [ɛ]) quite far apart, others widely separate [i] and [e] but have [e] and [ɛ] close together, others having [ɛ] and [a] close together. These differences exist both for French and Canadian French, and whatever the speaker's age between 4 and adulthood (the distribution being evaluated in a normalised F1 space, the F1 values of [i] and [a] playing the role of normalised references between which [e] and [ɛ] may be placed at various positions from one speaker to another).

Such inter-individual variation is not completely surprising, and compatible with a variant of the DT, called Adaptive Variability Theory (AVT) (Lindblom, 1990) according to which vowel systems do not search to *optimise* perceptual distances but to achieve *sufficient* contrast for communication. However, a striking observation by Ménard et al. is that for each subject, vowels tend to be grouped along stable F1 values, that is, if [e] is close to [i], then [ø] is close to [y] and [o] is close to [u], [e], [ø] and [o] having quite neighbour F1 values; and the same for [ɛ], [œ] and [ɔ]. This regularity, statistically assessed, is very interesting because it does not seem to obey any

perceptual law – and it seems to provide a kind of analogous to the MUAF principle, though not on logical values ("features") but on continuous cues (F1).

Ménard et al. interpreted this pattern as due to a stability of tongue height from one member of the mid-high or mid-low series to another, MUAF becoming what they proposed to call a "Maximal Use of Available Controls" (MUAC) principle. In their reasoning, speakers select a given articulatory command – vowel height – at a rather arbitrary value in an idiosyncratic way, but once selected, this value is transferred from one vowel to another. More precisely, they suggested that in the course of speech development, the young speaker would achieve a sufficient control within a given series, and then transfer the adequate (sufficient and subject-dependent) control to another tongue/lip configuration, thus achieving stable mid-high and mid-low series (with stable tongue height and thus, as shown by acoustic simulations, stable F1 values) thanks to a MUAC principle.

In this reasoning, the developmental articulatory pathway is supposed to be crucial in the achieved shape of the vowel system for a given speaker. This illustrates how vowels are, in fact, more than pure "sounds" selected for their perceptual values; but rather, vocalic gestures organised developmentally and shaped by their acoustic/auditory properties.

II.3. The categorisation of height features, a perceptual idiosyncrasy in mirror

In a later study, Ménard & Schwartz (submitted) tested a possible perceptual counterpart of this motor idiosyncrasy, by submitting the speakers of the previous study to a speech perception task aiming at determining their vocalic prototypes. For this aim, an articulatory model of the vocal tract, VLAM, was used (Boë, 1999, from Maeda, 1979). This model enables to synthesise vowel stimuli covering the vowel triangle, for various sizes of the vocal tract corresponding to various ages between birth and adulthood. A "younger" vocal tract is characterised in VLAM by both larger formants and larger F0 values (Ménard et al., 2002). Stimuli were generated for covering the vowel space corresponding to two ages: an "adult" one and a younger one, depending on the subject's age. Subjects were asked to identify the vowel stimuli within one of the 10 categories for French oral vowels [i e ɛ y ø œ u o ɔ a]. Four major results emerged from this study (Fig. 3b, c). Firstly, subjects are variable as listeners, as they are as speakers, with large inter-individual variation for category centres from one subject to another. Secondly, F1 grouping is displayed for perception as for production, mid-high and mid-low series being characterised by stable F1 values for each subject, and different from subject to another. Thirdly, the distribution of F1 values for a given subject for the two vocal tract sizes are correlated, indicating the existence of normalisation mechanisms prior to identification along the idiosyncratic prototypes. Last but not least, the distribution of F1 values for perception and production are correlated, that is, the way a given subject produces vowel height (idiosyncratically) is mirrored in the way he/she perceives vowel

height.

There is not so many evidence of a link between idiosyncrasies in perception and production (though see Bell-Berti et al., 1979, cited in Galantucci et al., 2006; Villacorta et al., 2007). The present data clearly suggest that there is a co-construction of motor and perceptual representations in the course of speech development for vowel height, as suggested in PACT.

II.4. Summary

The PACT scenario for vowel height production and perception assumes that vowel gestures are built in development through a process involving articulatory exploration and composition, and perceptual tuning (including sufficient dispersion, and focalisation), leading to a series of prototypes used by the subject for production and perception. A model presented in Fig. 4 displays how perception would function in this scenario, starting with perceptual cue extraction and normalisation, followed by categorisation based on prototypes defined in common by perception and production. Vowels in this scenario appear as auditorily-shaped articulatory gestures, displaying both an *articulatory coherence* expressed by the MUAC principle and acquired in the course of development, and a *perceptual coherence* associated to the communicative value of phonetic units, expressed by the normalised auditory cues characterising the vowel content for categorisation, and obeying dispersion and focalisation principles (DFT).

III. Speech Scene Analysis as a perceptuo-motor process

Auditory Scene Analysis (ASA) is known since the classical book by Bregman (1990) as the set of mechanisms enabling the listener to organise an auditory scene, grouping coherent ingredients within streams or sources and disentangling one stream from another thanks to various cues such as temporal, spectral, fundamental frequency or localisation coherence. 15 years later, Remez et al. (1994) noticed that a speech stream is characterised by pieces of information displaying some degree of incoherence, associated to the quick succession of silences, bursts, noises, or harmonic sections. However, the "speech scene" seems perceived in a coherent way, suggesting that specific mechanisms could be in charge of gluing the various pieces together, because of the "special" nature of speech (likely, its motor nature, in the framework of the Motor Theory which served as a background for the paper by Remez and coll.). We will show in the following that speech scene analysis is a natural framework for displaying the role of perceptuo-motor interactions in speech perception.

III.1. The Verbal Transformation Effect, a paradigm for studying Speech Scene Analysis

In the last years, Sato, Basirat, Schwartz and coll. developed a series of works exploiting the

"Verbal Transformation Effect" (VTE; Warren & Gregory, 1958; Warren 1961) as a way of assessing Speech Scene Analysis mechanisms. In the VTE, a given stimulus presented in loop may lead to switches from one percept to another (for example, the rapid repetition of the word "life" produces a perceptual transform into "fly" and "fly" back into "life" and so on). The VTE thus pertains to a family of paradigms producing multistable perception, as in the visual modality with the Necker's cube or binocular rivalry, in which the subject alternatively perceives the stimulus in one or another way, and displays more or less regular switches from one to the other. Multistability reveals various ways in which a given scene may be perceived, and has proved to be an efficient paradigm for characterising the emergence of percepts in the human brain (e.g., Leopold & Logothetis, 1999; Blake & Logothetis, 2002).

III.2. Auditory, phonetic and lexical mechanisms in the VTE

In a series of works with the VTE, Pitt and colleagues contributed to better specify the VTE methodology, exploiting such ingredients as the number of displayed forms, the time spent on each form, the frequency of switches, etc (Pitt & Shoaf, 2001, 2002; Shoaf & Pitt, 2002). They displayed three series of mechanisms at work in the VTE. The basic mechanism is a re-segmentation process changing the segmentation boundary from one position to another (e.g. from "life life life" to "fly fly fly"). A second series of phonetic/lexical/semantic mechanisms enrich the number of possible forms by introducing modifications of the input stimuli at each of these three levels, e.g.: [pold] for [pod] repetition.

A third mechanism displayed by Pitt et al. consists in auditory streaming. This is precisely what Bregman (1980) capitalised on for general Auditory Scene Analysis mechanisms, and what Remez et al. (1994) considered as unlikely for normal speech perception. For example, Pitt & Shoaf (2002) show that in repetitions of one-syllable words including a consonantal cluster with an initial fricative in the onset or a final fricative in the coda, there was a strong tendency that listeners separated the fricative from the rest of the information, verbal transformations thus involving the sequence without the fricative into a foreground stream, and the fricative constituting a kind of background stream segregated from the remaining phonemes (e.g.: for [skin] repetition, [kin] or [gin] reported as transformations and [s] as the background stream).

III.3. The role of articulatory coherence in the VTE

In a first series of experiments, Sato et al. (2006) wondered whether articulatory coherence could be shown to display a role in the VTE. For this aim, they focused on a French nonsense analogous of the "life"-"fly" transform, in which a "pse" non-word transforms into "sep" and the other way round. They suggested that in a perceptuo-motor paradigm in which the subject produces

the sequence in loop and searches for a transformation, [psə] should display a larger articulatory coherence than [səp] because of articulatory synchrony between the consonantal gestures. Indeed, in [psə], the labial gesture for [p] and the tongue gesture for [s] can be prepared in anticipation and launched almost in synchrony, the lips being opened for [p] on a vocal tract prepared for [s] with the tongue tip already in contact with the palate, so that the [s] friction can be immediately produced after the labial burst at the syllable onset. On the contrary, in [səp], the two gestures respectively happen at the syllable onset for [s] and coda for [p]. The switch from "sep sep sep" to "pse pse pse" would hence consist in a resynchronisation, and conversely, the switch from "pse pse pse" to "sep sep sep" would involve a *desynchronisation* of the two consonantal gestures. The prediction was that the synchronous [psə] would be more stable than the asynchronous [səp] and therefore, a larger number of transformations from [səp] to [psə] would be observed than the other way round (Fig. 5a). This is what was actually showed, both in an overt and in a covert mode (Sato et al., 2006).

In a further study focused on CVCV alternations such as [pata] vs. [tapa], Sato et al. (2007a) were able to produce similar effects of articulatory coherence, though in a completely auditory VTE paradigm. They capitalised on the so-called labial-coronal effect, according to which labial-coronal (LC) sequences such as [pata] (a labial plosive [p] in the first syllable, followed by a coronal plosive [t] in the second one) are more frequent in human languages (MacNeilage & Davis, 2000). An articulatory explanation of the LC effect was provided by Rochet-Capellan & Schwartz (2007) according to which the LC sequence can be realised on a single jaw cycle with an anticipation of the coronal gesture inside the labial one, the reverse being impossible (anticipating the lip closure would hide the audibility of the coronal). Therefore, the CL sequence would be less well coordinated than the LC one. A prediction was that the LC coherence would lead in the VTE to a larger stability for [pata] than for [tapa], the listener chunking "patapatapata" sequences into LC [pata] items compatible with jaw cycles, rather than into CL [tapa] items (Fig. 5b). This was actually displayed in experimental data (Sato et al., 2007a).

The fact that articulatory constraints may act on the emergence and stabilization of verbal transformations strongly suggests that they partly rely on motor neural processes. This is in keeping with several recent fMRI and intracranial electro-encephalographic (iEEG) studies demonstrating that articulatory-based representations play a key part in the endogenously driven emergence and stabilization of auditory speech percepts during a verbal transformation task (Sato et al., 2004; Kondo & Kashino, 2007; Basirat et al., 2008).

III.4. Multimodal speech scene analysis

In a further series of experiments, Sato et al. (2007b) tested whether visual speech could intervene in the VTE. They showed that lipread speech actually produced switches as auditory speech, and that the visual content of an audiovisual sequence contributed to the audiovisual switches by increasing or decreasing the stability of the sequence whether the auditory and visual stimuli were coherent or incoherent.

Furthermore, they showed that visual changes in the seen material could drive the audiovisual switches over time, as displayed when a stable audio "sepsep" sequence is dubbed on a video alternation of "sep" and "pse", resulting in switches from "sep" to "pse" perfectly synchronous with the video switches (Fig. 5c). They suggested that lip opening gestures could provide onset cues for audiovisual speech segmentation, driving the percept towards the item beginning with the most visible onset.

This is in line with a previous experiment by Schwartz et al. (2004) in which lip gestures happen to enhance the detection of audio cues necessary for categorisation. In this study, subjects had to discriminate voiced and unvoiced syllable onsets such as [tu] vs. [du] embedded in noise and presented at random temporal positions. Seeing the lip gesture provided a temporal cue useful for the auditory detection of prevoicing, resulting in increasing the audiovisual performance (Fig. 5d). This provides another example of "multimodal speech scene analysis".

III.5. Summary

In these experiments, the speech scene analysis process appears to be driven by both perceptual and motor coherence. The listener combines general auditory scene analysis (ASA) mechanisms (leading to e.g. streaming processes) with articulatory principles grouping the acoustical – and visual, if present – pieces of information in a coherent way, limiting streaming as Remez et al. suggested, and favouring articulatory coherent streams such as [psə] over [səp], [pata] over [tapa], coherent audiovisual scenes over incoherent ones, relying on visible labial onsets. Here again, speech should be conceived as what it is: neither auditory, nor motor, but perceptuo (multisensory) – motor in nature.

IV. A neurocomputational framework for PACT

PACT is built upon a central concept, that speech units are perceptuo-motor in nature, with two related main assumptions: firstly, perceptual representations are shaped by articulatory knowledge; secondly, identification proceeds from perceptual representations, which contain the communicative value of the speech gesture. We shall discuss how these assumptions fit with neurophysiological data, how they can be incorporated into a computational model, and what kinds

of experimental questions remain in this framework.

IV.1. Does the dorsal/ventral cortical network fit with PACT?

A strong lesson of the last twenty years is the very convincing accumulated evidence for sensorimotor connections between temporal auditory and audiovisual regions, parietal somesthesic/proprioceptive areas and frontal motor and premotor zones inside a dorsal cortical network as shown in Section I. The dorsal route is a candidate for establishing parity between sensory and motor representations in the framework of the mirror neurons theory (see Rizzolatti et al., 2001; Rizzolatti & Craighero, 2004 for a review). For Hickok & Poeppel (2000, 2004, 2007) it is the network involved in learning the control and processing of phonemic units for speech communication and is thought to be used strategically to assist in working memory and sub-lexical tasks. This provides a basis for one of the two major assumptions in PACT: that perceptual and motor representations of speech are learnt and structured together, and connected in a principled and functional way. However, Hickok & Poeppel (2000, 2004, 2007) consider that speech comprehension under ecologically valid conditions, i.e. when sounds are ultimately transformed into lexical/conceptual meaning, exploits another route, different from the dorsal one, and that they call the ventral route. The dorsal and ventral pathways both imply initial auditory processing in the primary auditory cortices along the supratemporal plane and subsequent phonological processing in the middle and posterior portions of the superior temporal sulcus (STS). At this point, the ventral stream however diverges from the dorsal one towards posterior middle and inferior portions of the temporal lobes involved in lexical and semantic processing Note that Scott & Wise (2004) differ on the precise temporal anatomy of the ventral route, going, in their view, from the primary auditory cortex towards the anterior part of the STS and possibly later connected with prefrontal areas for lexical, syntactic or semantic processing. Whatever the precise anatomy of the ventral route, the crucial role of temporal areas in speech comprehension is in good agreement with the second PACT assumption, that perceptual representations provide the basis for identification of phonetic units.

In summary, the dorsal route would ensure a co-structuration of perceptual and motor representations in a coherent way – and the possibility to call for motor routines in the course of perceptual processes: this remains to be further discussed in Section IV.3. And the ventral route would provide the decision/comprehension process based on perceptual (auditory, visual) representations before lexical and semantic access.

IV.2. Ingredients for a computational PACT architecture

To explore computational architectures, let us begin by asking a functional question. What is precisely the role of perceptuo-motor interactions in speech perception? Within PACT, we suggest

two kinds of answers.

Firstly, the perceptuo-motor link contributes to *structure perceptual categories* in reference to their motor content. This is the core ingredient of motor theories. Consider the [di] vs. [du] case quoted in Section I.1. Various pieces of acoustic evidence, different from one vocalic context to the other, point to the same consonantal gesture. This is taken by Liberman & Mattingly (1985) as an indication that acoustic cues are translated into articulatory features. However, another reasoning is that the articulatory link enables to define acoustic/auditory classes of equivalence, or to provide natural phonetic categories. Interestingly, these two different reasoning lead so different theoreticians as Liberman & Mattingly (1985) and Sussman et al. (1998) to call for the same neuronal analogy in auditory processing: that is, the neuronal columns for auditory localisation in the inferior colliculus of the barn owl. Wagner et al. (1987) or Konishi et al. (1988) showed how different phase relations between the two ears, related to the same orientation in the 3D space (same azimuth) for tones of different frequencies, were organised within orientation dominance columns enabling to recover the invariant character of the distal event from the variable proximal stimulus. This was considered by Liberman & Mattingly as a typical case of modular "heteromorphic" system connecting the proximal stimulus with its distal cause, just as the speech proximal acoustic stimulus is heteromorphically transformed into its distal cause, the articulatory gesture. But it was considered by Sussman et al. (1998) as a case of efficient exploitation of the available auditory information, just as their "locus equation" would provide an invariant acoustic correlate of plosive place in spite of the variable vocalic context, with no need for articulatory inversion.

Our reasoning is a synthesis of both versions: the information for decision is actually acoustic/auditory (and visual); but the equivalence classes, organising the proximal stimulus into natural categories, are defined in relationship with gestural commands. This would explain how listeners recover sources of variability related with coarticulation, reduction, etc: not through articulatory-to-acoustic inversion, but through categories based on representations and boundaries defined within the perceptual space, though organised in reference with the perceptuo-motor link, as in the experimental data on vowel production and perception presented in Section II. This is described in Fig. 6, which proposes a general computational architecture for PACT. The link between "Motor schemas" (combining prototypes and production laws) and "Auditory categorization", acquired through co-structuring of perceptual and motor representations, enables to incorporate knowledge of the speech production mechanisms inside auditory categorisation mechanisms (through sensori-motor speech maps and trajectories). For example, the fact that the speaker has "experienced" the various auditory consequences of closing the lips for "b" results in a structuring of the auditory space into a natural "closed lips" category which implements what Viviani called the "procedural knowledge of action". This is the way the "variability" due to

coarticulation is dealt with in PACT.

The second use of the perceptuo-motor link relates to what Skipper et al. (2007) call "prediction" in their model of audiovisual speech perception involving the temporo-parieto-frontal connection, though "prediction" is a principle that might appear a bit vague. Indeed, considering the speed and accuracy of the auditory system, it is not really obvious if motor-based predictions could be really and systematically helpful. However, there could be in one case: when there is a *lack of information*, say because of noise or in a foreign language. In this case, the perceptuo-motor link could enable to indicate where the lacking information could be searched. This is the case in the "speech scene analysis" problem described in Section III, in which audiovisual and perceptuo-motor links appear important for better organising the scene, segmenting it appropriately, and recovering adequate auditory pieces of information in the case of a noisy communication (as in Schwartz et al., 2004). This is incorporated in Fig. 6 in terms of "integration" that is incorporating speech production knowledge about possible sensory trajectories through "receptive field" tuned on these trajectories.

IV.3. Connecting computational principles with neurocognitive data

The two potential roles of a perceptuo-motor connection could actually rely on quite different types of implementations, and could hence be related to different sets of experimental data.

The co-structuring component suggests the possibility to connect perceptual and motor representations for speech communication in a principled way. This is what we proposed to call "sensori-motor speech maps" capitalising on previous works on speech robotics in our lab (Abry and Badin, 1986). In an experimental behavioural task on the discrimination of jaw height, tongue position or lip rounding in vowel perception, we showed (Schwartz et al., 2008) that subjects are to a certain extent able to have direct access to vowel articulatory features, which suggests that sensori-motor maps do exist, for vowels at least, in the human brain. Recent fMRI data on perceptual or articulatory maps for vowels or consonants (Obleser et al., 2003, 2004, 2006; Pulvermuller et al. 2006) confirm and provide neurophysiological correlates of such maps. Interestingly, these maps appear to be dynamic and provide a way by which changes in production can result in changes in perception. Indeed, in a recent study, Shiller et al. (2009) showed that after a speech production experiment with auditory perturbations leading to motor after-effects, perception was recalibrated in relation with this motor after-effect. They suggested that the perceptuo-motor link coordinating perceptual and motor representations is in fact plastic and susceptible to modifications relating articulatory changes to perceptual changes.

The other component, that is prediction/integration, could possibly involve on-line call to

the motor system for completing auditory speech scene analysis. There are however two caveats for this assumption. Firstly, after all, it is not necessary to call for motor simulation if the structure of the auditory representations already includes clues about perceptuo-motor trajectories. Secondly, if an explicit call for motor processes is indeed necessary, the prediction is that the call would be particularly necessary in adverse conditions, e.g. noise, audiovisual coordination, foreign language. This fits well with experimental data showing an increase in frontal activity in speech perception exactly in these conditions, that is noise (e.g., Binder et al., 2004; Zekveld et al., 2006), audiovisual integration (e.g., Jones & Callan, 2003; Ojanen et al., 2005; Pekkola et al., 2006; Skipper et al., 2007), or foreign language (e.g., Callan et al., 2004; Wilson & Iacoboni, 2006). However, no clear-cut experiment proves in an unambiguous way a decisive role of motor and premotor areas in speech perception. Hickok & Poeppel (2001, 2004, 2007) repeatedly notice that frontal lesions impairing speech production do not impede speech perception. Some weak perturbations in auditory speech tasks have been obtained by temporarily disrupting the activity of the motor system through TMS (Meister et al., 2007; d'Ausilio et al., 2009; Sato et al., 2009). These perturbations remain however small and not at the level to prove that perception *needs* action. The question hence stays rather open for future works.

IV.4. Summary

In summary, two functions could be associated with the perceptuo-motor connection. Firstly, perceptual representations in the temporal pole would be structured in relationship with motor representations in the parieto-frontal pole, through the dorsal route in the course of speech development, and later on, all along life, through dynamic adaptations in various learning conditions. Secondly, the motor system could be involved, perhaps more in adverse conditions, in order to provide a better specification of possible auditory and visual trajectories and enhance speech scene analysis. This would exploit the dorsal route on line, but comprehension would operate basically in the temporal pole within the ventral route.

V. Conclusion

PACT provides a tentative way to connect perceptual and motor representations in a principled way in line with behavioural and neurophysiological data. PACT starts from the acknowledgement that speech units are shaped by a double set of articulatory and sensory (audiovisual) constraints, and hence are characterised by both a motor and a perceptual coherence. This enables to derive a number of predictions from PACT towards the way human languages emerge in development (Serkhane et al., 2005, 2007) and phylogeny (Schwartz et al., 2007; Moulin-Frier et al., 2008). Vowels, consonants and syllables are conceived, in this process, as the

result of a set of perceptual and motor optimisation processes, in line with a late view by Chomsky himself: "*The language faculty interfaces with other components of the mind/brain. (...) Recent work suggests that language is surprisingly 'perfect' in this sense, satisfying in a near-optimal way some rather general conditions imposed at the interface*" (Chomsky, 1996, pp. 29-30). Language as an "optimal" or at least "optimised" communication system based on perceptually-shaped articulatory gestures provides in our view the framework in which a "neural theory of language" should be further explored and developed in the near future.

References

Abry, C. (2003). [b]-[d]-[g] as a universal triangle as acoustically optimal as [i]-[a]-[u]. *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona. 727–730.

Abry, C., & Badin, P. (1996). Speech mapping as a framework for an integrated approach to the sensori-motor foundations of language. In *Proceedings of the fourth speech production seminar* (pp. 175–184), Autrans.

Abry, C., Boë, L.J., & Schwartz, J.L. (1989). Plateaus, catastrophes and the structuring of vowel systems. *J. Phonetics,* 17, 47-54.

Arbib, M.A., (2005). From Monkey-like Action Recognition to Human Language: An Evolutionary Framework for Neurolinguistics. *Behavioral and Brain Sciences*, 28,105-167.

d'Ausilio,A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., & Fadiga L. (2009). The Motor Somatotopy of Speech Perception, *Current Biology,* 19, 1-5.

Basirat, A., Sato, M., Schwartz, J.L., Kahane, P., & Lachaux, J.P. (2008). Parieto-frontal oscillatory synchronization during the perceptual emergence and stabilization of speech forms. *NeuroImage,* 42, 404-413.

Bell-Berti, F., Raphael, L. J., Pisoni, D. B., & Sawusch, J. R. (1979). Some relationships between speech production and perception. *Phonetica*, 36, 373-383.

Binder, J.R., Liebenthal, E., Possing, E.T., Medler, D.A. & DouglasWard, B. (2004). Neural correlates of sensory and decision processes in auditory object identification. *Nat. Neurosci*., 7, 295-301.

Blake, R., & Logothetis, N.K. (2002). Visual competition. *Nat. Rev., Neurosci.,* 3 13–21.

Boë, L.-J. (1999). Modeling the growth of the vocal tract vowel spaces of newly-born infants and adults. Consequences for ontogenesis and phylogenesis. *Proceedings of the International Congress of Phonetic Sciences*, 3, San Francisco, 2501-2504.

Bregman, A.S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

Callan, D.E., Jones, J., Callan, A. & Akahane-Yamada, R. (2004). Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory–auditory/orosensory internal models. *NeuroImage*, 22, 1182-1194.

Chomsky, N. (1996). Language and Thought: Some Reflections on Venerable Themes. In *Powers and Prospects: Reflections on Human Nature and the Social Order* Boston: South End Press.

Cooper, W.E., Billings, D. & Cole, R.A. (1976). Articulatory effects on speech perception: a second report. *Journal of Phonetics,* 4, 219-232.

Cooper, W.E., Blumstein, S.E., & Nigro, G. (1975). Articulatory effects on speech perception: a preliminary report. *Journal of Phonetics,* 3, 87-98.

Cooper, W.E. & Lauritsen, M.S. (1974). Feature processing in the perception and production of speech. *Nature*, 252, 121-123.

Delattre, P., & Freeman, D. (1968). A dialect study of american rs by x-ray motion picture. *Linguistics*, 44, 29–68.

Diehl, R.L., Lindblom, B., & Creeger, C.P. (2003). Increasing realism of auditory representations yields further insights into vowel phonetics. *Proceedings of the 15th International Congress of Phonetic Sciences*, Vol. 2, pp.1381-1384. Adelaide: Causal Publications.

Diehl, R.L., Lotto, A.J. & Holt, L.L. (2004). Speech perception. *Annual Review of Psychology*, 74 , 431-461.

Eimas, P.D. & Corbit, J.D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, 4, 99–109.

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

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

Fowler, C.A., Rosenblum, L.D., 1991. The perception of phonetic gestures. In Mattingly, I., Studdert-Kennedy, K. (Eds.), *Modularity and the Motor Theory of Speech Perception*. Erlbaum (Lawrence), NJ, pp. 33–60.

Galantucci, B., Fowler, C.A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, 13, 361-377.

Gentilucci, M., & Corballis M.C. (2006). From manual gesture to speech: a gradual transition. Neuroscience & Biobehavioral Reviews, 30, 949-960.

Guenther, F.H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, 105, 611-633.

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

Hickok, G. & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8, 393-402.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences*, 24, 849-878.

Jones, J. & Callan, D.E. (2003). Brain activity during audiovisual speech perception: An fMRI study of the McGurk effect. *NeuroReport*, 14, 1129-1133.

Kohler, E., Keysers, C., Umiltà, M.A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: action representation in mirror neurons. *Science*, 297, 846-848.

Kondo, H.M., & Kashino, M., (2007). Neural mechanisms of auditory awareness underlying verbal transformations. *NeuroImage*, 36, 123-130.

Konishi, M., Takahashi, T., Wagner, H., Sullivan, W. E., & Carr, C. E. (l988). Neurophysiological and anatomical substrates of sound localization in the owl. In G. M. Edelman, W. E. Gall, & W. M. Cowan (Eds.), *Auditory function*, (pp. 721-745). New York: Wiley.

Kuzla, C., Ernestus, M., & Mitterer, H. (2009). Compensation for Assimilatory Devoicing and Prosodic Structure in German Fricative Perception. *Laboratory Phonology*, 10 (in press).

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

Liberman, A.M., Cooper, F.S., Harris, K.S., & MacNeilage, P.F. (1962). A motor theory of speech perception. *Proceedings of the Speech Communication Seminar*, Stockholm.

Liberman, A.M., Delattre, P.C., & Cooper, F.S. (1952). The role of selected stimulus variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, 65, 497-516.

Liberman, A.M., Delattre, P.C., Cooper, F.S., & Gerstman, L.J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs*, 68, 1-13.

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

Liberman, A.M., & Whalen, D.H. (2000). On the relation of speech to language. *Trends in Cognitive Science,* 4, 187–196.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulations of vowel quality systems: The role of perceptual contrast. *Language,* 48, 839–862.

Lindblom, B. (1984). Can the models of evolutionary biology be applied to phonetic problems? *Proceedings of the 10th International Congress of Phonetic Sciences,* Utrecht, 67–81.

Lindblom, B. (1986). Phonetic universals in vowel systems. In J.J. Ohala and J.J. Jaeger (eds.) *Experimental Phonology.* New–York: Academic Press (pp. 13–44).

Lindblom, B. (1990). On the notion of possible speech sound. *J. Phonetics,* 18, 135–152.

Lotto, A., Hickok, G., & Holt, L. (2009). Reflections on mirror neurons and speech perception. *Trends in Cognitive Sciences*, 13, 110-114.

MacNeilage, P.F. (1998). The Frame/Content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21, 499-546.

MacNeilage, P.F. & Davis, B.L. (2000). On the origin of internal structure of word forms. *Science*, 288. 527 - 531.

Maddieson, I. (1984). *Patterns of sounds*. Cambridge: Cambridge Univ. Press. 2nd ed. 1986.

Maeda, S. (1979). An Articulatory Model of the Tongue Based on a Statistical analysis. *J. Acoust. Soc. Am*, 65, S22.

Mattingly, I. G. (1990). The global character of phonetic gestures. *J. Phonetics*, 18, 445-452.

Meister, I.G., Wilson, S.M., Deblieck, C., Wu, A.D., & Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Curr. Biol*., 17, 1692-1696.

Ménard, L., & Schwartz, J.L. (submitted). Normalization effects and perceptuo-motor biases in the perceptual organization of the height feature in French vowels.

Ménard, L., Schwartz, J.L., & Boë, L.J. (2004). The role of vocal tract morphology in speech development: Perceptual targets and sensori-motor maps for French synthesized vowels from birth to adulthood. *J. Speech Language and Hearing Research,* 47, 1059-1080.

Ménard, L., Schwartz, J.L., Boë, L.J., Kandel, S., & Vallée, N. (2002). Auditory normalization: of French vowels synthesized by an articulatory model simulating growth from birth to adulthood. *J. Acoust. Soc. Am*. 111, 1892-1905.

Mielke, J., Baker, A., & Archangeli, D. (2009). Covert /ɹ/ allophony in English: variation in a socially uninhibited sound pattern. *Laboratory Phonology*, 10 (in press).

Möttönen, R., Järveläinen, J., Sams, M. & Hari, R. (2004). Viewing speech modulates activity in the left SI mouth cortex. NeuroImage, 24, 731-737.

Moulin-Frier, C., Schwartz, J.L., Diard, J., & Bessière, P. (2009). Emergence of phonology through deictic games within a society of sensori-motor agents in interaction. *Book chapter in Vocalization, Communication, Imitation and Deixis, (to appear)*.

Obleser, J., Boecker, H., Drzezga, A., Haslinger, B., Hennenlotter, A., Roettinger, M., Eulitz, C., & Rauschecker, J.P. (2006). Vowel Sound Extraction in Anterior Superior Temporal Cortex, *Human Brain Mapping* 27, 562–571.

Obleser, J., Elbert, T., Lahiri, A., & Eulitz, C (2003). Cortical representation of vowels reflects acoustic dissimilarity determined by formant frequencies. *Cogn Brain Res*, 15, 207–213.

Obleser, J., Lahiri, A., Eulitz, C. (2004). Magnetic brain response mirrors extraction of phonological features from spoken vowels. *J Cogn Neurosci*, 16, 31–39.

Ohala, J.J. (1979). Moderator's introduction to symposium on phonetic universals in phonological systems and their explanation. *Proceedings of the 9th International Congress of Phonetic Sciences*, 3, 181–185.

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, 797-807.

Perkell, J.S., & Klatt, D.H. (1986). *Invariance and Variability in Speech Processes*. L. Erlbaum, Hillsdale N. J., New Jersey.

Perrier, P. (2006). About speech motor control complexity. In Harrington, J. & Tabain, M. (Eds.) *Speech Production: Models, Phonetic Processes, and Techniques*. New York: Psychology Press (pp. 13-25).

Pitt, M. & Shoaf, L. (2001). The source of a lexical bias in the Verbal Transformation Effect. *Language and Cognitive Processes*, 16, 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., & Castellanos, F.X. (1980). Speech-production measures of speech perception: Rapid shadowing of VCV syllables. *J. Acoust. Soc. Am*., 67, 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, 593-602.

Pulvermuller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proc Natl Acad Sci U S A*, 103, 7865-70.

Remez, R.E., Rubin, P.E., Berns, S.M., Pardo, J.S., & Lang, J.M. (1994). On the perceptual organization of speech, *Psychological Review*, 101, 129–156.

Rizzolatti, G., & Craighero, L. (2004). The Mirror-Neuron System. *Annual Rev. Neurosci*., 27, 169-92.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, 2, 661-670.

Rizzolatti, G., & Arbib, M.A.. (1998). Language within our grasp. Trends in Neurosciences, 21: 188-194, 1998.

Rochet-Capellan, A., & Schwartz, J.L. (2007). An articulatory basis for the labial-to-coronal effect: /pata/ seems a more stable articulatory pattern than /tapa/. *J. Acoust. Soc. Am.*, 121, 3740-3754.

Rosch-Heider, E., 1972. Universals in color naming and memory. *J. Exp. Psychol*. 93, 10–20.

Roy, A.C., Craighero, L., Fabbri-Destro, M. & Fadiga, L. (2008). Phonological and lexical motor facilitation during speech listening: A transcranial magnetic stimulation study. *J. Physiol. Paris*, 102, 101-105.

Sato, M., Baciu, M., Lœvenbruck, H., Schwartz, J-L., Cathiard, M-A., Segebarth, C. & Abry, C.

(2004). Multistable perception of speech forms in working memory: An fMRI study of the verbal transformation effect. *NeuroImage,* 23, 1143-1151.

Sato, M., Basirat, A., & Schwartz, J-L. (2007). Visual contribution to the multistable perception of speech. *Perception and Psychophysics,* 69, 1360-1372.

Sato, M., Rousset, I., Schwartz, J-L. & Vallée, N. (2007). A perceptual correlate of the Labial-Coronal Effect. *J., Speech, Language and Hearing Research*, 50, 1466-1480.

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

Sato, M., Tremblay, P. & Gracco, V. (2009). A mediating role of the premotor cortex in speech segmentation. Brain and Language (in press).

Savariaux, C. Perrier, P., &Orliaguet, J.-P. (1995). Compensation strategies for the perturbation of the rounded vowel [u] using a lip-tube: A study of the control space in speech production *J. Acoust. Soc. Am*., 98, 2428-2442.

Savariaux, C., Perrier, P., Orliaguet, J.-P. & Schwartz, J.-L. (1999). Compensation strategies for the perturbation of French [u] using lip-tube. II. Perceptual analysis. *J. Acoust. Soc. Am*., 106, 381-393.

Schwartz, J.L. (2009). Filling the perceptuo-motor gap. *Laboratory Phonology*, 10 (in press).

Schwartz, J.L., Abry, C., Boë, L.J., Ménard, L., & Vallée, N. (2005). Asymmetries in vowel perception, in the context of the Dispersion-Focalisation Theory. *Speech Communication,* 45*,* 425-434.

Schwartz, J.L., Berthommier, F., Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition,* 93, B69–B78.

Schwartz, J.L., & Boë, L.J. (2007). Grounding plosive place features in perceptuo-motor substance. Workshop on "International Conference on Features" (N. Clements & R. Ridouane), Paris.

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* (pp. 104-124). Oxford University Press.

Schwartz, J.L., Boë, L.J., Vallée, N., & Abry , C. (1997). The dispersion-focalization theory of vowel systems. *J. Phonetics,* 25, 255-286.

Schwartz, J.L., & Escudier, P. (1989). A strong evidence for the existence of a large scale integrated spectral representation in vowel perception. *Speech Communication,* 8, 235-259.

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., Vallée, N., & Kandel, S. (2008). Hearing the tongue and lips of vowel gestures: a new differential paradigm. *J. Acoust. Soc. Am.,* 123, 3179.

Serkhane, J.E., Schwartz, J.L., Boë, L.J., Bessière, P. (2005). Building a talking baby robot: A contribution to the study of speech acquisition and evolution. *Interaction Studies,* 6, 253-286.

Serkhane, J.E., Schwartz, J.L., Boë, L.J., Davis, B.L., & Matyear, C.L. (2007). Infants' vocalizations analyzed with an articulatory model: A preliminary report. *J. Phonetics*, 35, 321–340.

Scott, S.K., & Wise, R.J. (2004). The functional neuroanatomy of prelexical processing in speech perception. *Cognition*, 92. 13 – 45.

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

Shiller, D., Sato, M., Gracco, V. & Baum, S. (2009). Perceptual recalibration of speech sounds following motor learning. *J. Acoust. Soc. Am.*, 125, 1103-113.

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, 17, 2387-2399.

Solé, M.J., & Ohala, J.J. (2009) What is and what is not under the control of the speaker: intrinsic vowel duration. *Laboratory Phonology*, 10 (in press).

Stevens, K.N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. Davis Jr. and P. B .Denes (eds.), *Human Communication: A Unified View.* New-York: Mc Graw-Hill, 51–66.

Stevens, K.N. (1989). On the quantal nature of speech. *J. Phonetics*, 17, 3–45.

Sundara, M., Namasivayam, A.K. & Chen, R. (2001). Observation-execution matching system for speech: A magnetic stimulation study. *Neuroreport*, 12, 1341-1344.

Sussman, H. M., Fruchter, D., Hilbert, J., & Sirosh, J. (1998). Linear correlates in the speech signal: The orderly output constraint. *Behavioral and Brain Sciences*, 21, 241-259.

Tiede, M.K., Boyce, S.E., Holland, C.K., & Choe. K.A. (2004). A new taxonomy of American English /ɹ/ using MRI and ultrasound. *J. Acoust. Soc. Am.*, 115, 2633-2634 (A).

Vallée, N., Schwartz, J.L., & Escudier, P. (1999). Phase spaces of vowel systems : A typology in the light of the Dispersion-Focalisation Theory (DFT). *Proc. of the XIVth International Congress of Phonetic Sciences*, 1, 333-336.

Villacorta, V.M., Perkell, J.S. & Guenther, F.H. (2007). Sensorimotor adaptation to feedback perturbations on vowel acoustics and its relation to perception. *J. Acoust. Soc. Am.*,122, 2306–2319.

Viviani, P., & Stucchi, N. (1992). Biological movements look uniform: evidence of motor-perceptual interactions. *Journal of Experimental Psychology: Human Perception and Performance*, 18, 603-623.

Wagner, H., Takahashi, T., & Konishi, M. (1987). Representation of interaural t time difference in the central nucleus of the barn owl's inferior colliculus. *J. Neuroscience*, 7, 3105-3116.

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

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

Watkins, K.E. & Paus, T. (2004). Modulation of motor excitability during speech perception: the role of Broca's area. *Journal of Cognitive Neuroscience*, 16, 978-987.

Watkins, K.E., Strafella, A.P. & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. Neuropsychologia, 41,989-994.

Wilson, S. M. & Iacoboni, M. (2006). Neural responses to non-native phonemes varying in producibility: evidence for the sensorimotor nature of speech perception. *NeuroImage*, 33, 316-25.

Wilson, S.M., Saygin, A.P., Sereno, M.I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat Neurosci*, 7, 701-702.

Zekveld, A.A., Heslenfeld, D.J., Festen, J.M. & Schoonhoven, R. (2006). Top–down and bottom–up processes in speech comprehension. *NeuroImage*, 32, 1826-1836.

Figure captions

Fig. 1 – The articulatory three-dimension space of oral vowels (a), together with its auditory (b) and visual (c) projections (see text).

Fig. 2 – Jaw cyclic gestures shaped by the acoustic/auditory transformation in a quasi quantal way

(a) When jaw position varies from low to high, the acoustic/auditory features change suddenly around two boundary values, separating plosives from fricatives and fricatives from vowels. This is a typically quantal relationship in the sense of Stevens (1972, 1989) (see text).

(b) The consequence is that inside jaw cycles, there is a natural division between a consonantal pole for high jaw positions and a vocalic pole for low jaw positions.

Fig. 3 – Perceptuo-motor distribution of vowel heights in French

Typical distributions for subjects having either close high and mid-high vowels (Type 1), or close mid-high and mid-low (Type 2) or close mid-low and low (Type 3). (a) Left column: typical production data; (b) Middle column: typical perceptual data at the same age; (c) Typical perceptual data with a smaller vocal tract (larger formant values)

Fig. 4 – A PACT model for the perception of vowels

The model comprises three stages, perceptual cue extraction, normalisation and categorisation based on prototypes defined in common by perception and production through a developmental process leading to MUAC principles.

Fig. 5 – Speech scene analysis as a perceptuo-motor multimodal process

(a) Articulatory coherence in the Verbal Transformation Effect (VTE). In the repetition of "sep sep sep" the coronal gesture for [s] and the labial gesture for [p] are asynchronous, while they are synchronous for [psə] which is hence more stable and attracts a larger number of transformations (Sato et al., 2006).

(b) Labial-Coronal effect in the VTE: in "patapata…" auditory sequences, [pata] sequences are chunked since they may be produced on a single jaw cycle (Sato et al., 2007a).

(c) Audiovisual VTE. If a "psepsepse" auditory sequence is dubbed with a video sequence switching between [psə] and [səp] the audio percept switches accordingly: the video stream binds the audio percept in the VTE (Sato et al., 2007b).

(d) Audio-visual speech scene analysis and perception in noise. The categorization of audio syllables beginning by a voiced consonant, such as [du], is improved by presentation of the video

input, thanks to the temporal cue provided by the labial gesture onset towards the rounded [u] target, and which enhances the detection of the audio prevoicing (Schwartz et al., 2004).

Fig. 6 – A PACT architecture for speech perception

The perceptuo-motor link leads to co-structuration of sensory and motor maps, and may also involve online calls to motor representations for improving the integration of speech material in the speech scene.