

SPEECH MAPPING

AS A FRAMEWORK FOR AN INTEGRATED APPROACH TO THE SENSORI-MOTOR FOUNDATIONS OF LANGUAGE

Christian Abry & Pierre Badin

Institut de la Communication Parlée
UPRESA CNRS 5009, INPG – Université Stendhal
46, Av. Félix Viallet, F-38031 Grenoble Cedex 01, France
Email: (abry,badin)@icp.grenet.fr - Fax: (33) 76.57.48.26

Résumé

Le concept d'une cartographie sensori-motrice de la parole semble viable à la fois technologiquement, pour l'inversion, et pour dépasser les théories polarisées sur l'un des versants de cette correspondance. Les données psychologiques, neuropsychologiques et les expériences d'imagerie cérébrales montrent que les aires de réception et de production de la parole peuvent être mutuellement activées, ce qui nous amène à discuter les théories unilatérales. La primauté du mouvement comme format représentationnel et comme connaissance procédurale est examinée à la lumière de la parole visuelle et de sa neuropsychologie. Concernant l'équivalence motrice considérée comme équivalence acoustique, on montre que lorsqu'il y a véritablement perturbation de la liaison articulatoire-acoustique, une *recartographie* s'impose. Une possible "histoire naturelle" de l'ontogenèse de cette liaison sera donnée pour la voyelle cardinale extrême [u].

Abstract

Beyond its technological spin-offs, is the *speech mapping* concept viable as a framework for an integrated approach to the sensori-motor foundations of language? Psychological experiments and brain imaging data show that both speech production and reception areas can recruit one another: we will consequently argue against one-sided theories. The primacy of movement over shape in implicit knowledge or representational format is questioned in the light of visual vowel identification tasks and neuropsychological data. For motor-equivalence and sound-equivalence, it is shown that, when the articulatory-acoustic organisation is crucially perturbed, it is difficult not to adopt a *remapping* stance. A

tentative two-sided story of the ontogenetic achievement of such an articulatory-acoustic organisation is given for the point vowel [u].

1. Introduction

Beyond the European project *Speech Maps* (Mapping of Action and Perception in Speech), dealing with sound-to-gesture inversion, the aim of the present contribution is to foster the *speech mapping* concept as a framework for an integrated approach to the sensori-motor foundations of language. Of course, this concept could actually be useful at two very different levels (they were both addressed in a special ASA session, *JASA* 99(3), 1996, pp. 1680-1741). First, it allows to posit the essentiality of the *link* between production and perception without an a priori oath of allegiance to any one-sided theory. Secondly, it allows to federate within a *robotic* framework many scattered trends in speech research, from articulatory modelling and synthesis, articulatory feature-based recognition, to multi-modal human-machine interfaces, through inversion and computational theories of control (cf. the constant interest for speech of the group of a prominent robotician, Kawato, in Wada *et al.*, 1995).

This latter aspect will not be addressed in this paper due to space restrictions. We just briefly suggest in this paragraph that in order to remain as close as possible to speech technology, our general idea is to obtain a maximum coherence of speech audio and visual synthesis taking advantage of an articulatory device in which this coherence is built in, i.e. a talking head which sounds and reflects light, and is even touchable, not a multimodal pasting. One of the main challenges for speech recognition is now to constrain the statistical models in order to spare time in the learning phase. Conse-

quently, there are two complementary ways to feed an articulatory platform for speech synthesis and recognition. The first calls for better biomechanical models: for the moment, the trend is to use partial models making available for the future a kind of toolbox of articulators, a tinkering approach which will lead to more integrated systems, when powerful machines will be cheap on the market and allow people in the field to develop less time-consuming algorithms. The second way is to improve control models, including their use in solving the various inverse problems. In fact, there is a need for both approaches to keep an eye on each other's progress, since an increase of biomechanical complexity will imply simpler control strategies.

Concentrating henceforth uniquely on the first level – i.e. the *link* between production and perception –, we will temperate the arguments of the most influential one-sided theories, like Motor Theory of Speech Perception and Direct Perception, and as concerns more intrinsically relational approaches, such as Quantal Theory or Hyper-Hypo Theory, we will emphasise the need to re-balance them in the realm of speech production control (we will not even mention auditory-only theories).

To support our claim that the task of speech scientists is to substantiate the diverse motor and sensory *maps* that enable humans to access fully or partially to speech, be they deaf, blind or deaf-blind, we consider (ii) & (iii), some relevant experimental data, based on psychological experiments, and (i) brain imaging techniques. For the moment, the message of this brain and (neuro)psychological data set is that both speech production and reception areas can recruit the other one (i). We consequently argue against both one-sided views. Within the motoric side, most experiments take as granted the primacy of movement over shape in implicit knowledge or representational format: we question such a primacy at the light of visual vowel identification tasks (ii). In order to test through perturbation the production/ perception link, most experimenters take as granted motor-equivalence as sound-equivalence: it is shown that when the articulatori-acoustic organisation is crucially perturbed, results are rather difficult to interpret in a such a one-sided view (iii).

Finally, we sketch some guidelines for the generation of sensori-motor maps and

put forward computational and developmental arguments which take into account that, if any kind of representation were to be recovered from reception only, it would surely lead to culs-de-sac, i.e. prevent subjects from finding the generally selected way to produce the intended speech sounds.

2. *Speech mapping and the theories of the production-perception link*

Strictly speaking, there is no speech theory that deals uniquely with the *link* between perception and action. Not to speak of auditory only theories (e.g. Kluender, 1994; Greenberg, 1995), even relational approaches such as the Quantal Theory (Stevens, 1989) or the Hyper-Hypo Theory (Lindblom, 1990) have a clear preference for promoting one side of the link as the *representational* format. In this case Stevens as well as Lindblom have steadily decided in favour of the sound side. This, in reaction to the Motor Theory of Speech Perception (Liberman & Mattingly, 1985) and the Gibsonian or Direct Perception Theorists (Fowler, 1995), who both advocate – the specificity of speech apart – the primacy of articulatory dynamics in commands. We cannot review here the classical arguments given in favour of a general auditory processing, say Japanese quail's performances in categorisation generalisation. Nor the pros and cons of "Motor Theory", within the paradigms of duplex perception or sine-wave speech, in this latter case with the recent proposal of a related but different perception theory of phonetic objects (Remez *et al.*, 1994).

Instead, as just said, we will put forward some forgotten or new relevant studies. But apart from these informational arguments, the core of our contribution tries to disentangle two syncretic views, coming from the most wrong part of each one-sided theory, namely the trend to take as logical implications of their initial assumptions, proposals that are not implicated at all. As concerns the motoric side, the emphasis on the contribution of motor control must not mean *ipso facto* that the claim could be that everything in speech lies ultimately in the movement format (kinematics or dynamics). As concerns the acoustic side, we claim that it could be untrue that every possible gesture is equivalent, provided it leads to the desired sound and satisfies to the coarticulatory requirements. The two sections following our sketchy review of brain data aim at these clarifications.

Provisionally, let us quote as an introductory guideline, the words of an approach recently promoted by Kent. Such an approach contrasts basically with one-sided views, and lies very close to what we called *speech mapping* (Abry *et al.*, 1993): “Speech may be represented as a number of neuronal “maps” that combine different kinds of sensory and motor information. In its global nature, speech is defined by the totality of these maps, and their interactions. More narrowly, speech can be defined by interactions among selected maps. Therefore, speech is auditory-visual (as the McGurk effect demonstrates), and tactuo-motor (as in the haptic communication system employed by users of Tadoma, who can understand speech from tactile cues gathered from a hand placed over the talker’s face and neck)” (Kent, 1995, p.482).

Within *Speech Maps*, i.e. built in the architecture of an audio-visual *Articulotron*, such an approach has been decidedly implemented in the framework of Morasso and Sanguineti’s *SoBos* (Self-organizing Body-schemas, 1995a) by Bailly (1995, 1996), in connection with the λ -commands of the Equilibrium Point for jaw and tongue (Morasso & Sanguineti, 1995b, Laboissière *et al.*, 1996, Ostry *et al.*, 1995, Payan *et al.*, 1995, Payan & Perrier, 1996).

3. *Speech mapping, brain imaging and (neuro)psychological architectures*

In this section, we do not refer to Penfield’s legacy in brain stimulation, called brain mapping (Ojeman, 1991), but only to recent techniques in brain imaging, particularly PET (Positron Emission Tomography; for a review of language studies, see Poeppel, in press) and fMRI (Functional Magnetic Resonance Imaging; for language, see Binder, 1995).

As concerns *speech mapping*, the role of Broca’s area in speech *perception* is particularly relevant. Zattore *et al.* (1992) have found an increased activity in this left area, corresponding to a phonetic discrimination task. More specifically, Demonet *et al.* (1994) have found this area to be active, in a phoneme monitoring task, only when the instruction was *sequential* (“detect [b] only if preceded by [d]”) and the phoneme to detect set in a phonetic environment that made it *ambiguous*. Their interpretation is that, in such a case, a specific part of the *working memory* is recruited, namely the

articulatory loop, which uses the *arcuate fasciculus* path which connects the POT (parieto-occipito-temporal area) to Broca’s area. However, it must be emphasised that this latter area is rather involved in sequencing and hierarchising tasks and recruited for speech inasmuch as it calls for such organisational properties.

Another finding is of some relevance to the connection of cortical maps. Hinke *et al.* (1993) reported coherently that, during a silent speech generation task, thirteen of their fourteen subjects displayed a higher activity in the left Broca’s area. But for six subjects, for whom the field of view was large enough to observe Wernicke’s area, this part was also activated during silent speech production. According to Latchaw *et al.* (1995, p. 201): “This experiment suggests that speech is a complex phenomenon, and that paradigms to localise motor speech also demonstrate activation in areas considered responsible for receiving and understanding that speech.”

Some years ago, Monsell (1987) established a careful inventory of the architectures available for linking or not production and perception in a lexical access framework. His conclusions – partly based on shadowing and repetition evidence – were in favour of a model with connections between sub-lexical input and output separate phonologies, but not with a common production-perception lexicon. More interestingly he referred to priming experiments to support his choice : “Silent generation, or preparation for generation of phonology [output] influences later auditory identification of words [phonology input]” (Monsell, 1987, p. 304). Though his reference to his own silent mouthing data, with M.T. Banich, was not relevant for a comparison with Hinke’s findings of an activation of Wernicke’s area, *preparation for generation* data could be (Gordon & Meyer, 1984).

As concerns intersensory speech perception, we will mention some recent cortical imaging data in the following section.

4. *Motor representations : Does movement in the ears and the eyes mean movement in the mind ?*

For most tenants of motor theories, the primacy of movement seems to be a logical implication. After all, motor control and movement control are pure synonyms for the same field of research. In fact, there are

posture controls and there exist theories of movement generation by position changes (e.g. Feldman & Levin, 1995); not to speak of movement generation with static control parameters as in steady-state phonation, for the vocal folds, or trill production, for the uvula, tongue blade or the lips (e.g. Pelorson *et al.*, 1994). In short: control of dynamics does not mean “dynamic control”.

4.1 Motor knowledge influences shape discrimination

Joining speech motor theorists, Viviani is a strong defender of perceptuo-motor interactions. His main views in the field are the following (as summarised by Kandel *et al.*, 1994). Viviani & Stucchi (1992) observed that the visual perception of the shape of a trajectory was influenced by implicit knowledge about the laws of motor production. In one of their experiments, a point-light stimulus traced elliptic trajectories with a kinematic pattern corresponding to the one observed in hand tracing movements. This kinematic pattern corresponded to the so-called *two-thirds power law*, which fits to observed data a function between shape and motion, i.e. between tangential velocity and the radius of curvature (Viviani & Schneider, 1991). The most relevant result was that when the spot traced a circle with a velocity profile corresponding to an ellipse, subjects tended to perceive an ellipse rather than a circle. Thus, motor variables influenced visual perception to such an extent that when the kinematics of the trajectory were in disagreement with the laws of motor production, geometric distortions were elicited.

According to the authors, this demonstrates that the perception of the shape of a trajectory implies knowledge about the laws characterising biological motion: hence they adopt a Motor Theorist’s view. In our opinion, the influence of this procedural knowledge does not mean necessarily that the representation of an ellipse must ultimately be in a motor format. It may be simply that, in the case of such an undersampled tracing like a moving spot, motion is necessary for processing shape. This shape-from-motion stance (see below) is classically rejected by the authors with RT arguments. In identifying moving light displays in the Johansson’s vein, 100 ms (for adults) to 200 ms (for children) are supposed to be too short for the available signal processing algorithms (but remember that more than 800 ms are at least necessary

in a masked point light walker, Cutting *et al.*, 1988).

4.2 Motion : A common metric for audio and visual speech ?

As concerns speech, this issue of a common metric for acoustic and optic speech – say motoric, like in the *Speech Maps* integration platform (Robert-Ribes *et al.*, 1995) – is orthogonal to the motion stance. We will show in the section following the present one, that it is also orthogonal to the representational/processing issue.

Many researches in auditory speech perception are in favour of dynamic cues for vowel identity. This is the topic of the silent-centre paradigm, where the steady-state phase of the vowel is reduced to silence, in order to test if the transitional cues alone can provide equal or even better results for the identification of the vowel. Whether such cues are due to coarticulation of vowels with consonants or inherent to vowel production remains largely controversial (see Bohn & Strange, 1995, *vs.* Nearey, 1995).

As concerns specifically audiovisual speech integration, Summerfield (1987) proposed at first a representation in terms of static configurations of the vocal tract. He indicated that the perceptual relevance of such configurations could be easily justified for consonants, since the articulators generally achieve their target positions. But the situation is not so clear for vowels. In fact, if vowel targets are generally reached when vowels are articulated in isolated syllables, it is not always the case in running speech, where changes in stress, tempo and phonetic context cause important variations. So, pushing further proposals coming from articulatory phonologists (Browman & Goldstein, 1985), Summerfield (1987) proposed finally that the listener-viewer could identify the relevant parameters of acoustic and optic signals in terms of time-varying patterns, advocating that this proposal provides a natural *common metric* between the kinematics of acoustics and optics.

This proposal was recently supported by Rosenblum and Saldaña (in press). They claimed that the shapes of the mouth in static configurations (frozen images) were insufficient for audiovisual speech integration when associated with audio signals. They used the classical moving light paradigm launched for face expression by Bassili. A talker was filmed in the dark with 28 point-lights fixed on his face, particularly on lips,

teeth, tongue tip, chin, cheeks, jaw and nose tip. When such point-light images were presented statically, they were not recognised as faces. But, in dynamic presentation, subjects often reported a McGurk effect: they perceived [va] for a moving [va] display, in spite of the fact that it was dubbed on an audio [ba] syllable.

However, the illusion was significantly different from that observed with *static* displays of the *full* face only when subjects had a chance to *first* see the *real* face in *motion*.

4.3 More on the dissociation between motion, shape, and shape-from-motion, in visual speech: A shape for shape, and a shape-with-motion for motion account

In reporting the results of a series of experiments, Cathiard *et al.* (1996) addressed the same question in the frame of another paradigm than moving dots. Full images of the face were presented in moving and static conditions along a naturally silent vowel-to-vowel *rounding* gesture. In order to compare the benefit of motion *vs.* shape, the change in angle paradigm for face perception was used for feature/gesture identification. It was hypothesised that a viewer-centred optimisation of the projection of the rounding gesture would override motion benefit. The relatively poor performance of dynamic presentations which were variable and never better than the best static ones – those for profiles – led to question the format of representations for speech. In fact such a saturation effect found for profile views supported rather a *shape-from-shading* processing for front views, enhanced by the *shape-from-motion* system. Both systems could be recruited in case of undersampling (point-lights) or for out of shape projections.

Such a computational approach would avoid calling for a direct perception account, i.e. motion-from-motion in all cases. Moreover it would leave open the alternative for representational formats: motion or shape from a non optimal, or even non recognisable, shape put in motion. So shape could remain a possible format for static as well as for dynamic stimuli, at least for a set of category features of speech sounds, in this case the main visible vowel gesture, *rounding*.

In spite of an increasing knowledge in the visual architecture of the cortex, it remains putative whether these computa-

tional algorithms (Pentland, 1989; Terzopoulos & Waters, 1990) are neurally implemented and we mentioned debates about available RTs.

Some times ago, Campbell (1992) reported cases supporting a neuropsychological dissociation between the ability to identify labial postures and the ability to lipread dynamic stimuli. Recently she added more evidence from McGurk experiments with normal subjects and neurological patients, and also from cortical imaging studies (Campbell, 1996a, b). Subjects who have damage to areas V1, V2, and V4 – say shape processing areas – cannot lipread from photographs. But they can speechread moving faces *inasmuch as their lesions have preserved some parts of the shape areas* (subject HJA, but not WM). For the same moving faces they need of course to have V5 (motion direction) and V3 (shape-from-motion) spared (HJA). A subject (DF) with only V5 and V3 spared can report directional movement, but not identify moving point-light displays, and neither speechread. Conversely a subject (LM) with lesions to V5 (but spared V1, V2, V3, V4) cannot experience motion, but is able to retrieve shapes from moving arrays; in fact as concerns speech, she can only identify in a movie the initial and final mouthshapes. Hence these “studies with patients who have damage to either the visual form [V1, V2, V4] or visual movement systems [V3, V5] show us that neither, alone, can deliver effective speechreading. This tells against a theory of speechreading as a form of direct visual movement perception: a theory implied by the point-light experiments of Rosenblum and colleagues” (Campbell, 1996b).

This supports the visual experiments reported above (Cathiard *et al.*, 1996). More specifically, these experiments have challenged the favour given in the literature to dynamic representations of vowels. We cannot predict if this could question the auditory perception of vowel dynamics. There are some suggestions that results in the silent-centre paradigm could be explained by perceptual interpolation of the missing “stable” central part of the vowel, using offset (“nucleus”) and onset (“offglide”) of the remaining transitions (Nearey & Assmann, 1986). And as concerns diphthongs, the question is not solved whether their perception can be accounted for by intrinsic dynamics or by a multiple nuclei

approach (Gottfried *et al.*, 1993; for visual nuclei, see Jackson *et al.*, 1976). The same for glides, like [w], which are supposed to be intrinsically dynamic representations.

To summarise, our claim is clearly *not* that movement is not useful in speech perception. It is of course very useful in many instantiations of casual, not overclear, speech, as shown especially in the common case of vowel reduction (Schwartz *et al.*, 1992). But there are, in our opinion, no clear data forcing us to infer that perceived movements are finally stored in a movement format in the mind, rather than shapes or targets. Indeed, our brains have a kind of general purpose built-in shape-from-motion processor (Oram & Perrett, 1994) working together with shape processors in speech, and in addition possibly a specific processing for lipreading in the *right* superior temporal region (analogous to the left Wernicke's reception site, Campbell, 1996b), not to speak of the system for extracting speech sound patterns from the auditory stream.

5. Sound-equivalence mapping : “All gestures which sound the same are equal, but some are more equal than others”?

By plagiarising Orwell's last commandment of *Animal farm*, we intend more than just pointing on individual differences in the skill to recover acoustic targets in perturbed speech, as evidenced by the experiment reported hereafter (Savariaux *et al.*, 1995). In our opinion, such an experiment supports both an articulatory coding of an auditory representation and, for all speakers – be they successful or unskilled in target recovery –, enough memory of the encoding (learning) phase not to attempt ill-directed manoeuvres, according to their sensori-motor procedural knowledge. In short, acoustic-equivalence even in the seemingly most free cases (in the present case of [u], see the discussion about the pseudo-precursor [w], in the following section), together with coarticulatory pressure, does not select a free varying articulatory target, but a rather specific one.

First, one must emphasise that a true perturbation of the sound-to-gesture link cannot be actually achieved through the bite-block technique of the jaw – in the Lindblom, Lubker & Gay's vein – which does not perturb the acoustically relevant end-effectors, the tongue and/or the lips. Consequently a labial perturbation of the French rounded vowel [u] was used. A

20-mm diameter lip-tube was inserted between the lips of the speakers. Acoustic and X-ray articulatory data were obtained for isolated vowel productions by eleven native French speakers in normal and lip-tube conditions (20 trials). Compensation abilities were evaluated through accuracy of the F1-F2 pattern. One must recall that such compensation is predicted by the simulations on an articulatory model. When this model – the SMIP, Maeda's model coupled with a control model – is instructed to produce an [u] and it has its lips open, the result is a lowering of the tongue in the upper pharynx and more constriction at that place to produce acoustic equivalence. So, this experiment contrasts with perturbation of a vowel like [y] – in the Ohala & Riordan's vein – for which it is not at all granted by the articulatory modelling that the lowering of the larynx could recover the acoustic pattern.

For the first perturbed trial, immediately after the insertion of the tube, no speaker was able to produce complete compensation. And clear differences between speakers were observed. Seven of them moved the tongue – and since none of them moved it in the wrong articulatory-to-acoustic direction – they hence limited the deterioration of the F1-F2 pattern. The remaining four did not show any significant articulatory change. These first reactions support the idea of speaker-specific internal representations of the articulatory-to-acoustic relationships. The results for the following 19 perturbed trials indicate that speakers used the acoustic signal in order to elaborate an optimal compensation strategy. One speaker achieved complete compensation, by changing his constriction location from a velo-palatal to a velo-pharyngeal region of the vocal tract. Six others moved their tongues in the right direction, achieving partial acoustic compensation. While the remaining four did not compensate. Again no one moved his tongue constriction location in the wrong direction.

The control of speech production thus seems to be directed towards achieving an auditory goal. But completely achieving the goal may be impossible because of speaker-dependent articulatory constraints. It is suggested that these constraints are due more to speaker-specific internal representation of the articulatory-to-acoustic relationships rather than to anatomical or neurophysiological limitations. Speech control could thus be ensured partly with the use of this internal

representation, and partly – particularly under perturbed conditions – by monitoring the acoustic signal.

We will now go beyond the results of this perturbation experiment to argue that vowel-to-vowel transition cannot be specified in acoustic terms only, but needs together an articulatory target – be it proprioceptive and/or orosensory. Coherently with the results just mentioned, such a target seems deeply rooted in the memory of each subject. A tentative story of this ontogenetic achievement will be given in the following section.

6. Some guidelines in maps generation: the [u] case

A large part of the perspectives that we will propound hereabout are based on preliminary results and partial theorising. We have not at the moment a global framework for the generation of speech sensori-motor maps with clear neuronal architectures and constraints stemming from ontogenetic timing, in spite of some biological models proposed in this field (especially Kent, 1992, p. 82 sq.).

Going from the first productions, by the infant, of supra-glottal constrictions of the [əβə]-type to canonical babbling (Holmgren *et al.*, 1986), we will consider sequences such as [wawa], as exemplars of sensori-motor mapping. In this case, we suggest that the knowledge of the acoustic effect of the visible lip constriction – namely to impose a general drift in the acoustic space towards the formants of [u] – is acquired through repetitive practice on the general pattern of opening/closing rhythmic train. At this stage, there is no true mastering of the [u] vowel, which is in fact rather late. As noted by MacNeilage & Davis (1990), the muscle control of this vowel is fairly complicated, since it implies at least three main manoeuvres: constriction at the lips (*orbicularis oris superior* and *inferior*), backing of tongue body (*styloglossus* and *stylohyoid*) and deconstriction of the pharynx (*posterior genioglossus*), not to speak of finer tuning of the intrinsic muscles needed to bunch the tongue (MacNeilage & Sholes, 1964). Remember that in [wawa] productions, the position of the tongue is free to coarticulate maximally with the vowel, since the lip constriction ensures a rather invariant acoustic target (Bailly *et al.*, 1991). One could think that the [u] prototype would be at first a pure matter of auditory mapping (Kuhl, 1995).

However, even at an early stage, there is some articulatory-to-acoustic knowledge (Kuhl & Meltzoff, 1995). Remember that this point vowel is one of the objects we called *focal vowels*, i.e. vowels with formant convergences produced by manoeuvres changing the formant-to-cavity affiliations (Badin *et al.*, 1990). This is a quite specific way to produce a spectral concentration of energy to be integrated by the auditory system (Schwartz & Escudier, 1987). Anyway, simulations with an articulatory model predict that, provided the “child” starts from its quite centralised vocalic productions, the typical velar position for [u] can be reached by a pure optimisation procedure (Bailly *et al.*, 1995), i.e. with a target specified in the acoustic space only. The articulators – or degrees of freedom of the model – will naturally cooperate in synergy to reach this articulatory and acoustic target, without drifting towards the other acoustic equivalence, a kind of upper pharyngeal [o] with more constriction, both at the lips and in the pharynx. Another related behaviour of the model is also enlightening: when instructed to produce an [o]-to-[u] transition, the tongue stays very close to this upper pharyngeal constriction. Consequently, there is a demand for a specification of an *articulatory* target in order to get the tongue move its constriction place towards that of [u] (Boë *et al.*, 1996). A prediction of this model is that if the child started its [u] from its [w] knowledge or from its [o] or [ʊ] knowledge – in fact from any constriction in the pharynx from [a]-to-[o], it would fall into an articulatori-acoustic cul-de-sac, which would never enable it to acquire the proper velar [u] position. Another support to this *articulatory* target lies in the behaviour of subjects with the lip-tube perturbation (Savariaux *et al.*, 1995): most of them were unable to really abandon their acquired link between the [u] acoustic target and velar position. As for most of the sensori-motor maps, as shown by recent surgical and brain imaging data illuminating Merzenich’s legacy, it takes generally some time – even if some recoveries are surprisingly fast – to remap *speech mapping*.

So the story is more two-sided. The memory of the adult [u] vowel keeps trace of the first [u] articulatory-to-acoustic mapping, namely of the strategy of the child which does not use [w] – its earlier and closest mapping – as a precursor, and takes

advantage of the fact that it can only start from the rather centralised vowels, embedded in its variegated babbling, to develop a high back one. In this view, the universal system [i a u] is not the final result of an overall equilibrium between auditory distinctiveness, say distance, and coarticulatory ease or proximity. This is really from the beginning a timed sound *and* gesture story.

Acknowledgements

We thank Marie-Agnès Cathiard, Sonia Kandel and Christophe Savariaux, and their co-authors, for permission to cite parts of their original works. Gérard Bailly, Rafael Laboissière and Jean-Luc Schwartz, for designing and performing the robotic experiments. Christoph Segebarth for making available to us a number of recent language fMRI references ; Ruth Campbell for specific neuropsychological references, through Christian Benoît. Louis-Jean Boë for kindly demonstrating the [u]/[o] case, which started in a collaboration with Pascal Perrier, and developed in a discussion with Sidney Wood; Gérard Bailly and Frédéric Berthommier for discussing this problem in the light of their respective knowledge in speech articulatory synthesis and neuronal remapping. Carol Stoel-Gammon, Bénédicte de Boysson-Bardies and Mary Vihman, for sharing their experience in developmental speech production, through discussion and writing of illuminating books. Of course every interpretation given here above stays under our responsibility.

This work has been partly funded by the EC ESPRIT/BR project *Speech Maps* n° 6975.

References

- Abry, C., Badin, P., & Scully, C. (1994) Sound-to-gesture inversion in speech: The Speech Maps approach. In Varghese K. et al., *Advanced speech applications* (Eds.), pp. 182-196. Springer.
- Badin, P., Perrier, P., Boë, L.J., & Abry, C. (1990) Vocalic nomograms: Acoustic and articulatory considerations upon formant convergences. *JASA*, 87, 1290-1300.
- Bailly, G. (1995) Recovering place of articulation for occlusives in VCV's. *XIIIth Int. Congr. of Phonetic Sciences*, Vol. 2, 230-233. Stockholm.
- Bailly, G. (1996) Sensory-motor control of speech movements. 4th Speech Production Seminar, 1st ETRW on Speech Production Modeling: from control strategies to acoustics. May 21-24, 1996, Autrans, France.
- Bailly, G., Boë, L.J., Vallée, N., & Badin, P. (1995) Articulatory-acoustic vowel prototypes for speech production. *Proc. of EUROSPEECH'95*, Madrid, Spain, September 1995, Vol. 3, 1913-1916.
- Bailly, G., Laboissière, R., & Schwartz, J.L. (1991) Formant trajectories as audible gestures: an alternative for speech synthesis. *J. of Phonetics*, 19, 9-23.
- Binder J.R. (1995) Functional magnetic resonance imaging of language cortex. *International Journal of Imaging Systems and Technology*, 6, 280-294.
- Boë, L.J., Schwartz, J.L. & Laboissière, R. (1996). Integrating articulatory constraints in the prediction of sound structures. 4th Speech Production Seminar, 1st ETRW on Speech Production Modeling: from control strategies to acoustics. May 21-24, 1996, Autrans, France.
- Bohn, O.-S., & Strange, W. (1995). Discrimination of coarticulated german vowels in the silent-center paradigm: "Target" spectral information non needed. *XIIIth Int. Congr. of Phonetic Sciences*, Vol. 2, 270-273, Stockholm.
- Browman, C., & Goldstein, L. (1985) Dynamic modeling of phonetic structures. In V.A. Fromkin (Ed.) *Phonetic Linguistics* (pp. 35-53). Orlando, Florida: Academic Press.
- Campbell, R. (1992) The neuropsychology of lipreading. In V. Bruce, A. Cowey, A.W. Ellis & D.I. Perrett (Eds.), *Processing the facial image*, (Proc. of a Royal Society Discussion Meeting, July 1991), (pp. 39-45), Clarendon Press, Oxford.
- Campbell, R. (1996a) Seeing brains reading speech: A review and speculations. In D. Stork & M. Hennecke (Eds.), *Speechreading by Humans and Machines*, NATO ASI Series F, vol. 150 (pp. 115-133). Springer-Verlag, Berlin.
- Campbell, R. (1996b) Seeing speech in space and time: psychological and neurological findings. ICSLP'96, October 3-6, Philadelphia, PA, USA.
- 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, vol. 150 (pp. 211-219). Springer-Verlag, Berlin.
- Cutting, J.E., Moore, C., & Morrison, R. (1988) Masking the motions of human gait. *Perception & Psychophysics*, 44, 339-347.
- Demonet, J.F., Price, C., Wise, R., & Frackowiak, R. (1994). A PET study of cognitive strategies in normal subjects during language tasks: Influence of phonetic ambiguity and sequence processing on phoneme monitoring. *Brain*, 117, 671-682.
- Feldman, A.G., & Levin, M.F. (1995) The origin and use of positional frames of reference in motor control. *Behavioral & Brain Sciences* 18:4, 723-806.
- Fowler, C.A. (1995) A realist perspective on some relations among speaking, listening and speech learning. *XIIIth Int. Congr. of Phonetic Sciences*, Vol. 1, 470-477. Stockholm.
- Gordon, P., & Meyer, D.E. (1984) Perceptual-motor processing of phonetic features. *J. of Experimental Psychology: Human Perception Performance*, 10, 153-178.

- Gottfried, M., Miller, J.D., & Meyer, D.J. (1993). Three approaches to the classification of American English diphthongs. *J. of Phonetics*, 21, 205-229.
- Greenberg, S. (1995) The ears have it: the auditory basis of speech perception. *XIIIth Int. Congr. of Phonetic Sciences*, Vol. 3, 34-41. Stockholm.
- Hinke, R.M., Hu, X., Stillman, A.E. et al. (1993) Functional magnetic resonance imaging of Broca's area during internal speech. *Neuroreport*, 4, 675-678.
- Holmgren, K., Lindblom, B., Aurelius, G., Jalling, B., Zetterstrom, R. (1986). On the phonetics of infant vocalization. In B. Lindblom & R. Zetterstrom (Eds.) *Precursors of early speech* (pp. 51-63). New York: Stockton.
- Jackson, P.L., Montgomery, A.A., & Binnie, C.A. (1976). Perceptual dimensions underlying vowel lipreading performance. *JSHR*, 19, 796-812.
- Kandel, S., Orliaguet, J.P., Boë, J.L. (1994) Visual perception of motor anticipation in the time course of handwriting. In C. Faure, P. Keuss, G. Lorette & A. Vinter (Eds.) *Advances in handwriting and drawing: a multidisciplinary approach* (pp. 379-388). Paris: Europia.
- Kent, R.D. (1992) The biology of phonological development. In C.A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.) *Phonological development: Models, research, implications* (pp. 65-90). Timonium, Maryland: York Press.
- Kent, R. (1995) Toward a neurodevelopmental model of phonetics. *XIIIth Int. Congr. of Phonetic Sciences*, Vol. 3, 478-485. Stockholm.
- Kluender, K.R. (1994) Speech perception as a tractable problem in cognitive science. In M.A. Gernsbacher (Ed.) *Handbook of Psycholinguistics* (pp. 173-217). Acad. Press: San Diego – Toronto.
- Kuhl, P.K. (1995) Mechanisms of developmental change in speech and language. *XIIIth Int. Congr. of Phonetic Sciences*, Vol. 2, 132-139, Stockholm
- Kuhl, P.K., & Meltzoff A.N. (1995) Vocal learning in infants: development of perceptual-motor links for speech. *XIIIth Int. Congr. of Phonetic Sciences*, Vol. 1, 146-149, Stockholm
- Laboissière, R., Ostry, D.J., & Feldman, A.G. (1996) The control of human and hyoid movement. *Biological Cybernetics*, in press.
- Latchaw, R.E., Ugurbill, K., & Hu, X. (1995) Functional MR imaging of perceptual and cognitive functions. *Functional Neuroimaging*, 5, 2, 193-205.
- Liberman, A., & Mattingly, I. (1985) The motor theory of speech perception revised. *Cognition*, Vol. 21, pp. 1-33.
- Lindblom, B. (1990) Explaining phonetic variation: A sketch of the H&H theory. In W.J. Hardcastle & A. Marchal, *Speech production and speech modelling* NATO ASI Series (pp. 403-439). Dordrecht: Kluwer Academic Publishers.
- MacNeilage, P.F., & Davis, B. (1990). Acquisition of speech production: Frames, then content. In M. Jeannerod (Ed.), *Motor representations and control, Attention and Performance XIII* (pp. 453-476). Lawrence Erlbaum.
- MacNeilage, P.F., & Sholes, G.N. (1964) An electromyographic study of the tongue during vowel production, *JSHR*, 7, 2208-2232.
- Monsell, S. (1987) On the relation between lexical input and output pathways for speech. In A. Allport, D. MacKey, W. Price, & E. Scheerer (Eds.) *Language perception and production...* (pp. 273-311). Academic Press: London–Toronto.
- Morasso, P., & Sanguineti, V. (1995a) Self-organizing body-schema for motor planning. *J. Motor Behavior*, 26, 131-148.
- Morasso, P., & Sanguineti, V. (1995b) Kinematic invariances and body schema. Comment to A.G. Feldman & M.F. Levin, The origin and use of positional frames of reference in motor control. *Behavioral and Brain Sciences*, 18:4, 769-770.
- Nearey T.M. (1995). Evidence for the perceptual relevance of vowel-inherent spectral change for front vowels in Canadian English. *XIIIth Int. Congr. of Phonetic Sciences*, vol. 2, 678-681, Stockholm.
- Nearey T.M., & Assmann, P. (1986). Modeling the role of inherent spectral change in vowel identification. *JASA*, 80, 1297-1308
- Ojeman, G.A. (1991) Cortical organization of language. *J. of Neurosciences*, 11(8), 2281-2287.
- Oram, M.W., & Perrett, D.I. (1994) Responses of anterior superior temporal polysensory (STPa) neurons to "biological motion" stimuli. *J. of Cognitive Neuroscience*, 6, 99-106.
- Ostry, D.J., Laboissière, R., & Gribble, P.L. (1995) Command invariants and the frame of reference for human movement. Comment to A.G. Feldman & M.F. Levin, The origin and use of positional frames of reference in motor control. *Behavioral and Brain Sciences*, 18:4, 770-772.
- Payan Y., & Perrier P., (1996) Articulatory and acoustic simulations of VV transitions with a 2D biomechanical tongue model controlled by the Equilibrium-Point hypothesis. 4th Speech Production Seminar, 1st ETRW on Speech Production Modeling: from control strategies to acoustics. May 21-24, 1996, Autrans, France.
- Payan, Y., Perrier P., & Laboissière R. (1995) Simulation of tongue shape variations in the sagittal plane based on a control by the equilibrium-point hypothesis. *XIIIth Int. Congr. of Phonetic Sciences*, Vol. 2, 474-477.
- Pelorsson, X., Lallouache, T., Tourret, S., Bouffartigue, C., & Badin, P. (1994) Modeling the production of bilabial plosives: aerodynamical, geometrical and mechanical aspects. *ICSLP'94*, Yokohama, Japan, Vol.2, paper S12-8, 599-602.

- Pentland, A.P. (1989). Shape information from shading. In J.C. Simon (Ed.), *From pixels to features*, (pp. 103-113), North-Holland : Elsevier.
- Poeppe, D. (in press) A critical review of PET studies of language. *Brain Lang.*
- Remez, R.E., Rubin, P.E., Berns, S.M., Pardo, J.S., & Lang, J.M. (1994) On the perceptual organisation of speech. *Psychological Review*, 101, 129-156.
- Robert-Ribes, J., Schwartz, J.L., & Escudier, P. (1995) A comparison of models of fusion of the auditory and visual sensors in speech perception. *AI Review*, 9, 323-346.
- Rosenblum, L.D., & Saldaña, H.M. (in press). An audiovisual test of kinematic primitives for visual speech perception. *J. of Experimental Psychology: Human Perception and Performance*.
- 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. *JASA* 98(5), Pt. 1, 2428-2442.
- Schwartz, J.-L., Beautemps, D., Arrouas, Y., & Escudier, P. (1992). Auditory analysis of speech gestures. In M.E.H. Schouten (Ed.), *The auditory processing of speech – From sounds to words*, (pp. 239-252), Mouton de Gruyter: Berlin.
- Schwartz, J.L., & Escudier, P. (1987) Does human auditory system include large scale frequency integration ? In M.E.H. Shouten (Ed.) *The Psychophysics of speech perception, NATO ASI series* (pp. 284-292). Martinus Nijhoff: Dordrecht.
- Stevens, K.N. (1989) On the quantal nature of speech. *J. of Phonetics*, 17, 3-45.
- Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell, (Eds.), *Hearing by eye: The psychology of lipreading*, (pp. 3-51), London: Lawrence Erlbaum.
- Terzopoulos, D., & Waters, K. (1990). Physically-based facial modelling, analysis, and animation. *Journal of Visualization and Computer Animation*, 1(2), 73-80.
- Viviani, P., & Schneider, R. (1991) A developmental study of the relationship between geometry and kinematics in drawing movements. *J. of Experimental Psychology: Human Perception and Performance*, 17, 198-218.
- Viviani, P., & Stucchi, N. (1992) Biological movements look uniform: Evidence of motor-perceptual interactions. *J. of Experimental Psychology: Human Perception and Performance*, 18(3), 603-623.
- Wada, Y., Koike, Y., Vatikiotis-Bateson, E., & Kawato, M. (1995) A computational theory for movement pattern recognition based on optimal movement pattern generation. *Biological Cybernetics*, 73, 15-25.
- Zattore, R.J., Evans, A.C., Meyer, E., & Gjedde, A. (1992) Lateralization of phonetic and pitch discrimination in speech processing. *Science*, 256, 846-849.