AVSP 2003 - International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4-7, 2003

Visual and auditory perception of epenthetic glides

Marie-Agnès Cathiard, Christian Abry, Séverine Gedzelman and Hélène Loevenbruck

Institut de la Communication Parlée, INPG/Université Stendhal, Grenoble, France
cathiard@icp.inpg.fr

Abstract

The purpose of this contribution is to improve our knowledge about the time course of visual and auditory perception with regard to the representation of sound types as different in their phenomenological format as vowels and glides. Our results on the perception of Vowel-to-Vowel gesture via the production of epenthetic glides in between – according to our 2-COMP-Vowel Model – allow us to conceive of the *time-varying vs. stationary* representational issue as linked to the underlying control for moving phases: (i) while the stationary (plateau) or climax phase of a vowel can be shown as truly representative for this segment, (ii) and whereas motion in the time-varying on-gliding phase can be shown to be informative only when shape is incomplete, (iii) motion in the off-gliding phase of the same vowel can reveal itself as *misleading*, in the sense that it could prime an erroneous candidate for the following vowel, up to the end of the transitional glide in V-to-V. Thus, contrary to the dynamic specification theory for vowels, the moving on-gliding and off-gliding phases can be, respectively, less informative than stationary parts, as shown before, and even truly misleading as shown here. (iiii) Moreover the same off-gliding phase can be recovered as a true controlled glide under certain language-based constraints (as exemplified by the *power-effect*). Finally in line with our caveat against the claim that "all is dynamics in speech", we will briefly mention recent computational modeling and neural data which support our hypothesis, especially the *snapshot* neuronal computation which fits with recent brain imaging data of stilled and moving speaking mouths.

1. Introduction

Vowels can be conceived as stationary percepts whereas glides need at least a transitional phase in order to be identified as such. A well known theory [1] considers that all vowel segments are dynamically specified. *Time-varying* phases («dynamic margins» [2]) or gliding onsets and offsets are supposed to be more informative than static ones («steady states»), as exemplified by the famous silent-center paradigm for vowels [3].

In our experiments we have adopted a production and control stance, claiming that such a stance helped us conceive that perceptual processing could be given for free by the natural time course of speech production. Hence we always analyse visual and auditory perceptual results in relation to articulatory and acoustic data. And this is the case here for Vowel-to-Vowel transitions with epenthetic glides in between. Glides can be identified as true glides in certain languages. However they can automatically originate from V-to-V transitions as it is explained by phasing control asynchrony in our 2-COMP Vowel Model. Interestingly, while they can be uncontrolled (by-products), they can gain, in the course of linguistic change, the status of fully represented segments like *v* in French *pouvoir*, hence English *power* (from Old French *t* deletion of Latin *potere*): what we dubbed the *power-effect*.

Were it shown that before acquiring such a consonantal status, the off-gliding vocalic phase were of no use for identifying the on-coming vowel (contrary to the on-gliding phase, see [4]), what would be the consequences for the phonetic-phonological status of vowels, regarding specifically their psychological time-varying format?

2. The rounding gesture: its time course and the emergence of an epenthetic phenomenon, the "power-effect"

2.1. Three major visible and audible phases in the V-to-V transitions

To describe these three V-to-V phases we will take as an example an articulatory signal obtained on the lips of a male French speaker: «T'as dit (Did you say): 'UHI ise'?» [tadi#yiiz] (where UHI is a pseudo proper «Indian» name and «ise», third person of pseudoverb «iser»; for more details, see [5]). On Fig.1 the following events can numbered: (1) constriction and protrusion movement onset for [y] start both somehow in phase (for protrusion notice that it starts after a maximum retraction during preceding [i] and a plateau during the pause); (2) constriction plateau for [y] (but not narrowest constriction as we will see just below) is reached first (constriction plateau onset); (3) then protrusion maximum, index of vowel climax, together with constriction plateau; (4) shortly after protrusion decrease, a slight constriction area decrease occurs (leading to the narrowest constriction): this is what we called «off-glide epenthesis», in this case a [ɥ]-glide; this glide is produced by a retraction of the lips together with a narrowing of lip slit; (5) area of constriction increases (constriction offset) finally rejoining protrusion decrease towards the following [i] vowel.

From event 1 to event 2, phase 1-2 can be considered as the *on-gliding phase* of the vowel [y] (this phase is modelled articulatorily and perceptually by reference to our MEM Movement Expansion Model [6], which deals essentially with anticipation extent). Phase 2-4 will be labelled as the *climax phase*; it may comprise plateau phases, e.g. for [y] an area plateau, and even a protrusion plateau in other cases (these plateau phases were explored perceptually [7]). Phase 4-5 is the *off-gliding phase* of [y], during which an «*off-glide epenthesis*» occurs. This phase will be explored perceptually in this paper. Notice that the final product can be transcribed more narrowly as [iyɥi].

Let us focus on this glide epenthesis. For us the existence of this epenthesis is as evident as other consonantal or vocalic epenthetic by-products like the famous English «*Thompson phenomenon*». We proposed for remembering to coin this process as the «*power-effect*», from a French-English example, i.e. latin *potere*, Old French *poeir* (Modern French *pouvoir*), giving as a loanword English *power*, through Middle English *poër*, *pouer*. The lability or variability of the occurrence of this glide epenthesis follows the same tendencies as those evidenced for epenthesis in general. In particular, it is probable that it follows such dialectal differences in behavior as the one evidenced by [8] for American English versus South African English (with no Thompson-like epentheses). In our observations, it is not only variable amongst French speakers, but within one and the same speaker, which is common for other epenthetic cases.

2.2. Accounting for epenthesis in the *power-effect*

Glide epentheses are given for free in the framework of our double-component account of V-to-V transitions. The 2-COMP-VOWEL model [9] assumes that all speech sounds, in order to control the *geometry* necessary for their aeroacoustic regimes, can recruit the degrees of freedom of the vocal-tract in two ways. They all have a *placing* component. And some of them have an additional *shaping* component, morphing the sagittal and/or coronal geometry, i.e. a control in 2D or 3D vocal tract space. *Placing* is the *global* component. For vowels it is achieved mainly by the extrinsic muscles of the tongue as end effectors, not to speak of the lips. Intrinsic muscles finish the job, *shaping* the tongue groove for [i], bunching the tongue arch for [u]. Other proposals following Öhman's legacy [10], converge on the role of this intrinsic-muscle *local* component for consonants, which we use here for glides sometimes considered as «semi-consonants». For the lips the *shaping* job is done by the two *orbicularis* as main agonists for [u] and our French [y]. When the *placing* and *shaping* components are fairly synchronous, V-to-V transitions without glide epenthesis are produced. Glide epenthesis occurs when *shaping* is *relaxed* asynchronously respective to *placing* changes. Such an asynchrony gives rise to glide epenthesis in the transition between vowels. In summary, during the V-to-V transition there is a change from *placing* to *placing*, i.e. in the targets of the *global* component which configurates the vowel along the vocal-tract. But, during this transition, there is no control of *shaping*, i.e. of the second *local* component, which morphs the sagittal and/or coronal VT-geometry. Our claim is that glide emergence is a mere consequence of asynchrony between *placing* and *shaping*. Consequently, although such an emergence can be monitored afterwards in order to be linguistically inhibited or enhanced, glides are not *a priori* controlled.

It is now possible to read the events numbered on our time functions displayed in Fig.1 in terms of *placing* and *shaping* commands. (1) Constriction and protrusion movements for [y], starting quite in phase, are cues of *placing* initiation (*on-placing*). (2) Constriction plateau for [y] is achievement of *placing*. (3) Protrusion maximum, index of vowel climax together with constriction plateau, is achievement of *shaping* (*placing and shaping climaxes*). (4) Then protrusion decreases. This could be due at least in part to the relaxation of the *shaping* command of the lip slit. But what is the most directly related to this command of relaxation is that, shortly after, [y] *unshaping* (or *off-shaping*) results in a constriction area minimum: i.e. [ɥ] glide epenthesis. (5) Finally constriction area increases, rejoining protrusion decrease towards the following vowel, which is a clear index of [i] *placing* ([y] *off-placing*). Hence command *asynchrony* is our proposed explanation of glide epenthesis production: the [y] *shaping* command is relaxed ahead of the [y] *placing* command, and this lets the [ɥ] glide emerge.

3. Perceptual experiment

3.1. A new question

In a previous study [6] we tested the presence of a plateau to differentiate French [yi] *vs.* [ɥi]. This cue was reliable for only half of our subjects, which corresponds to the dialectal pattern of French. For these contrasting subjects, we also tested their expectation of the presence of a true (non-epenthetic) glide. We found that this glide can be identified as early as the beginning of the off-placing phase, i.e. as soon as the lip area increases towards [i], with a small benefit of expectation and this only for a true glide. Hence the question remained, for non explicitly controlled epenthetic glides, of their effect in the flow of audiovisual perception. In the present study with [yɥi] transitions, there was no glide identification task. But we tested if the epenthetic gliding phase would function as a masking, or rather *misleading*, gesture preventing from identifying the following segment [i]. Would subjects continue to identify [y], in spite of the moving off-gliding phase, or not? And at what point in the visual and acoustic flows would they start to identify the following [i] vowel? In other words would they anticipate, as soon as the [y] climax is over, or would they follow in their perception the details of the natural time course of production?

3.2. Stimuli with glide epentheses

We videorecorded a French male trained speaker uttering, in a random order, 10 repetitions of the following sentences : «Tu dis: 'UHI ise'?» [tydiyiiz], «Tu dis: 'RUHI ise'?» [tydiRyiiz] and «Tu dis: 'ZUHI ise'?» [tydizyiiz]. The articulatory analysis, essentially based on the time course of the lip area (accurately provided by the «deep blue» Chromakey preprocessing system developed in our lab, [11], [12]), evidenced a quite typical glide behaviour. We can observe in Fig. 2, for a sequence «Tu dis: 'RUHI ise'?», that the plateau constriction in 'RUHI' begins during the coarticulated initial rounded [R], with a lip area of about 90mm2, i.e. small enough to quantally contribute to the [y] acoustic characteristics [13]. And finally a minimal area constriction as small as 0.5mm2 is reached (1.47mm2, 20ms, i.e. 1 image before; 10.32mm2, 2 images before; and 3.62mm2, 1 image after; see Fig. 3), without changing the acoustic vocalic regime (i.e. the acoustic signal is not a wall vibration one).

3.3. Articulatory and acoustic analysis

For the 30 transitions, the same systematic minimal area constriction was observed: the mean value was 8 mm2 (a mean value stable for the 3 consonantal contexts preceding the [y] vowel; individual value could be as small as 0.5mm2 as in our example on Fig. 2). This minimal area value is clearly inferior to the mean value for the stable part of the [y], i.e. 50mm2 (from 26.6mm2 on average for the 10 realizations of UHI, to 44.5mm2 for ZUHI and 81.4mm2 for RUHI). The minimal constriction value of the glide was not predictable by the constriction plateau value. The acoustic analysis of the [yɥi] transitions revealed a 2 dB decrease in intensity around the minimal constriction, and an average in formant decreases of 122.5Hz for F2 and 176.5Hz for F3, relative to the target value of [y]. But F2-F3 focalization (energy concentration of close formants) remained of the [y]-type, and not of a F3-F4 [i]-type .

3.4. Method: The gating paradigm

We explored, by visual and acoustic gating (see from [14] to [15] for AV speech), the potential perceptual benefit (or disadvantage) of this constriction event. Our main question was the following: was this minimal area constriction event, accompanied by a formant decrease, perceptually *misleading* for the visual and auditory perception of the following [i] vowel?

Six sequences were chosen from the 30 studied, two from each context. The gating tests (7 steps of 20ms) delivered the sentence and stopped around the minimal constriction exploring a range of 60ms before and 60ms after that event. In the audio condition, the 20 French subjects heard the entire sequence up to the gating point. In the visual condition they both heard and saw the carrier part («Tu dis …»), and only saw the remaining portion. In all the conditions, the subject's task was to decide whether the sequence «Tu dis…» finished with [y] or [i] (for example 'HUE' [y] or 'UHI' [yi]).

3.5. Results and causal explanation

The six [i] audio and visual identification curves are displayed in Fig. 4. Three are maximally steep, given our sampling rate, switching from [y] to [i] in 20ms (two in 40 ms, one taking 60ms). They are fairly parallel. One can notice that the identification 50% boundaries in the audio condition are systematically shortly ahead (by at least 20ms) of the visual ones. A thorough examination of the possible biases of desynchronization due to video and spectral processing was carried out. The lead in audio identification therefore remains to be explained, even if it is small.

The identification boundaries, together with the slope of the curves, seem to be related to the lip area and to F3 time course of each stimulus. We propose the following articulatory-to-acoustic general causal explanation, taking as an illustration RUHI 5 (Fig. 5). The auditory switch corresponds to a very small but detectable area increase (less than 20mm2), which results in a significant change in F3. Why? Starting from [y], an affiliation model of the front cavity can be approximated by a 9cm long narrow tube (adding 4cm for lip lengthening to a 5cm tongue constriction for [i]) with a λ/2 resonance below 2kHz (above 3kHz for [i]). At the minimal area value (indicated on the time axis by 0), i.e. after the retraction of protrusion, we approximate a λ/4 resonance with a 4.5cm length (which corresponds to [i] tongue constriction, subtracting 0.5cm for lip constriction), hence remaining below 2kHz. This change in resonance mode from λ/2 [y] to λ/4 [ɥ] is to be related with small changes in F2-F3 focalization, as measured above. Of course the 0.5cm long lip closure is not fully complete, as in [ybi]. Knowing that a very small opening, like in this [b] burst, can let the [i] frication be heard (like the [s] in [ps]), the approximated 4.5cm λ/4 below 2kHz resonance can be heard as soon as reached, even as soon as within the lip constriction. It will then rapidly raise towards [i], with a 5cm λ/2 resonance above 3KHz. Contrary to the very small area increase needed for this change in acoustic regime, visual perception is lagging: a larger increase in lip area (of about 40mm2) is needed for the lip configuration to be recognized as [i]. So the story is that vision is waiting a little more for enough visible lip opening, while audition has already taken benefit of the least «leak» of the acoustic resonance, due to the invisible preshaped tongue of the vowel.

Our main concern here is that the identification boundary for [i], be it audio or visual, always comes after the minimal area constriction event. That is: (i) the acoustic identification waits for an increase in F3; (ii) and alike, the visual identification seems to follow the time course of lip area. Crucially, at the acoustic level, the small changes in focalization, after the stable part of the [y] vowel, can not at all be a cue for the identification of a following [i] vowel. And, visually, the decrease in lip area after the constriction plateau of [y] is neither taken as a cue to the oncoming [i], in spite of the achievement of a complete retraction of the lips. On the contrary, subjects wait to get sufficient increases in formant, then in lip area.

4. Discussion: Why certain vowel transitions can be misleading before leading to true consonants?

4.1. When vowel motion is misleading

How is this result relevant to the issue of a time-varying format for vowel representations? In the present study with non explicitly controlled epenthetic glides, we showed that the off-gliding phase of the [y] vowel did not lead to the identification of the following [i] segment. Thus, it cannot be considered that as soon as the apical (climax) phase for [y] is waning, i.e. while protrusion is declining, the ongoing transition (off-gliding phase) is automatically a clear cue for the identity of the following vowel. Remember that this was the case for the on-gliding phase for [y] where we found a clear *anticipation* of ceiling identifications long before its apex ([4], [5]). The time flow of the segmental information is not symmetrical in speech, which is well known for CVC, but not for V-to-V, as exemplified here for one of the main dimension of vowels, in the French [iy] (on-rounding) *vs.* [yi] (off-rounding). This fact should of course be taken into account in any dynamical stance. Our 2-COMP-Vowel modeling allows an off-shaping phase to conceal, hence delay, the identification of the following segment: this was clearly the outcome of our present experiment, for audio as well as for visual time course. That is what we dubbed *the misleading V-to-V transition phenomenon*.

The unrounding motion – lip-retraction with more constriction, i.e. a minimal area, as a consequence of *off-shaping* (rounding *relaxation*) – is not usable to take benefit of motion for the next vowel identification. But motion is used only when *off-placing* will become first audible, then visible. However this fact cannot be taken *per se* as an argument against a time-varying representation for vowels. This simply means that representing the motion path of [y] as a vowel is useless for its off-shaping phase ending in a glide, whereas it is informative for its on-shaping phase. But could such vowels be represented kinematically for roughly half of their time course? Moreover, off-shaping, following our 2-COMP-Vowel Model, produces labile glides, which can give rise to full glide representations, i.e. dynamic units for which the release phase is not more controllable as a stationary phenomenon than the transitory release of a plosive. But in the long term some of these glides can stabilize into sustainable segments like [v] in *pouvoir*. So the story seems more contrasted than uniquely dynamical.

Let us summarize. For about a decade (since 1994) we have been bringing evidence against this *uniquely* dynamic stance for speech segment representations. Our main argument came from the processing of on-gliding phases of vowel gestures. Adopting, for visual speech perception, a *view-dependent* approach, we demonstrated that non-ceiling identifications (as for the climax or apex) along this on-gliding phase could take advantage of time-varying information. But only when the shape projection was not optimal, i.e. for front views of on-rounding, but not for profile views. Profile views of course optimally characterize speech for the specific lip-controls of rounded vowels and sh-like consonants. Hence we adopted a *shape-from-shading & shape-from-motion* low level processing of front views ([15]). This clearly means that we think of motion in view dependence, as a means to recover shape, when shape is not completely given: thus, the ultimate outcome is *shape* and not motion representation. Among speech sound-types, vowels and fricatives can be sustained, whereas affricates, plosives, diphthongs and glides, cannot, at least for all their phases. In fact it is arguable that diphthongs could have only two targets and that even complete transitions, e.g. closing and release phases in plosives, can just help us recover their ultimate steady-shape, even if this target is inaudible in voiceless sounds (e.g. the lip closure steady-state for [p] is silent). But for a prolonged [j:] it is crucial to consider its dynamic release in order to identify [j] and not [i:]; and similarly for the release *slope* of affricates [ts] *vs.* plosives [t], and glides [w] *vs.* plosives [b]. So our experiments on visual glides are crucial to any program aiming at providing different representations for different sound types, and even phases.

4.2. Computational modeling and brain activation

What can be expected, as regards this issue, from recent trends in neuroscience and physiologically compatible computational models? Given the two reputed processing streams, dorsal for motion and ventral for shape (form), two recent contributions are particularly relevant. First Lorenceau & Alais [16] using different shapes, like a diamond or two herringbones (by swapping the diamond), which moved behind a grating, showed that in spite of similar *local* displacements, *motion binding*, allowing to perceive a *global* form displacement behind the grating, was constrained by a form-based veto (or green-light), depending on the *closure* of the shape, i.e. a good Gestalt for the diamond, but not for the herringbone. Moreover their results show an early (low-level) form-motion interaction; which does not discard interactions at higher levels. Notice the *gating* role of *shape* for *motion binding*. More recently Giese & Poggio [17] proposed a new computational model elaborated for the integration and testing of neural data. This model operates with the two parallel processing streams, the form pathway (ventral) and the optic-flow or motion pathway (dorsal). Apart from specific local detectors, both pathway deliver activity in motion pattern neurons. These learned patterns are encoded as sequences of *snapshots*, by *snapshots neurons* which feed each motion pattern neuron. The form pathway preserves view-dependence as an extension of Riesenhuber & Poggio's [18] model for stationary objects, which are just a special case of snapshots. «In conclusion, the available data interpreted with our model indicate that both pathways contribute to the recognition of normal biological movement stimuli […]» (p. 188). Their last question was: «whether the neural circuits for stationary object recognition and the recognition of snapshots overlap […]» (p. 190). Coincidently this answer was given by Calvert & Campbell's (C&C 2003 [19]) last fMRI study of highly identifiable frozen speech frames (at their climax or apex) *vs.* moving silent syllables, in a [v] detection task. Since brain activity for the perception of stilled speech is included in activity for dynamic speech, *they do overlap*. Stilled speech is clearly left-dominant compared to a bilateral activity for moving speech. Interestingly motion area MT/V5 is also active for frozen frames. This last result could lead to interpretations ranging from implicit motion processing to intention detection in Theory of Mind, via action understanding, as noted for Superior Temporal Sulcus activations related to body part actions (eyes, hands, mouths) by Allison et al. ([20], p. 275), even for stimuli with no implied motion at all. Since MT/V5 is situated at the upper end of the STS, being one convergence of the ventral and dorsal streams, Giese & Poggio's [17] model would allow to process about there any speech stilled snapshot as a special case of the corresponding sequence of snapshots. If the climax representative pattern for left-sided speech and language network (with the frontal-parietal bi-polarity BA44-BA40) is given (like in C&C stilled frames), this could prevent a bilateral activation during the extraction, from moving images, of the significant speech snapshot (for vowels, fricatives and plosives) or of the relevant sequence of snapshots (for true glides, possibly diphthongs, and affricates).

In other words the *snapshot neuron* approach can accommodate different representations, learned from stationary or moving visemes, and even significant phases of visemes, neglecting irrelevant or misleading parts of certain visemes. The possibility, encountered in speech, for a transition – like our epenthetic glides – to become fully representational could be preserved in learning through the language-specific constraints imposed on the phasing of placing and shaping components in our 2-COMP Vowel Model. Meanwhile Giese & Poggio [17] also allow to take into account view-dependent (with such *view-tuned snapshot neurons*) and quantify degradation or robustness, which is an agenda for future brain event experiments, with static climax *vs.* static non climax *vs.* moving non climax view-dependent shapes, as early quantified psychophysically by Cathiard et al. [15] in the time flow of speech information.

Aknowledgments: To J.-L. Schwartz, P. Badin, S. Rossato and P. Perrier for discussion; to C. Savariaux and A. Arnal for technical support; to S. Bianciotto and C. Brunière for their help in the passing of the test; to M. Odisio, M. Sato and A. Vilain for references. The perceptual study is supported by a «programme Cognitique du Ministère Français de la recherche » (Act 1b).

5. References

[1] Strange, W., Verbrugge, R., Shankweiler, D. and Edman, T., "Consonant environment specifies vowel identity", *J.A.S.A.., Vol. 60, 213-224*, 1976.

[2] Rosenblum, L.D. and Saldaña, H.M." "Time-varying information for visual speech perception", in R. Campbell, B. Dodd and D. Burnham (Eds.), *Hearing by Eye II,* 61-81, Psychology Press, 1998.

[3] Strange, W. and Bohn, O.-S., "Dynamic specification of coarticulated German vowels", *J.A.S.A.., Vol. 104(1), 488-504*, 1998.

[4] Cathiard, M.-A., Abry, C. and Lallouache, M.-T., "Does movement on the lips mean movement in the mind?", In D. Stork and M. Hennecke (Eds.), *Speechreading by Humans and Machines*, Berlin, Springer-Verlag, 211-219, 1996.

[5] Cathiard, M.-A., "*La perception visuelle de l'anticipation des gestes vocaliques: cohérence des évènements audibles et visibles dans le flux de la parole*", Doctorat de Psychologie Cognitive, Grenoble 2, 1994.

[6] Abry, C., Lallouache, M.-T. and Cathiard, M.-A., "How can coarticulation models account for speech sensitivity to audio-visual desynchronization?", In D. Stork and M. Hennecke (Eds.), *Speechreading by Humans and Machines,* Berlin, Springer-Verlag, 247-255, 1996.

[7] Cathiard, M.-A., Abry, C. and Schwartz, J.-L., "Visual perception of glides versus vowels", *International Conference on Auditory-visual Speech Processing*, Terrigal, Australia, 115-120, 1998.

[8] Fourakis, M. and Port, R., "Stop epenthesis in English", *Journal of Phonetics, 14, 197-221*, 1986.

[9] Abry, C., Laboissière, R., Loevenbruck, H., Cathiard, M.-A. and Schwartz, J.-L., "Glide production and control in the two-component vowel model", In *Proceedings of the 5th Seminar of Speech Production : Models and Data & CREST Workshop of Models of Speech Production : Motor Planning and Articulatory Modelling*, 37-40, Kloster Seeon, Bavaria, May 1-4. 2000.

[10] Öhman, S.E.G., "Numerical model of coarticulation", *J.A.S.A., Vol. 41, 310-320*, 1967.

[11] Lallouache, M.-T., "*Un poste Visage-Parole couleur. Acquisition et traitement automatique des contours des lèvres*", PhD Thesis. I.N.P. Grenoble, 1991.

[12] Audouy, M., "Logiciel de traitement d'images vidéo pour la détermination de mouvements des lèvres", Grenoble, ENSIMAG, Projet de fin d'études (génie logiciel), 2000.

[13] Abry, C., Boë, L.-J. and Schwartz, J.-L., "Plateaus, catastrophes and the structuring of vowel systems", *Journal of Phonetics, 17, 47-54*, 1989.

[14] Grosjean, F., "Spoken word recognition processes and the gating paradigm", *Perception & Psychophysics, 28, 267-283*, 1980.

[15] Munhall, K.G. and Tokhura, Y., "Audiovisual gating and the time course of speech perception", *J.A.S.A., Vol. 104(1), 530-539*, 1998.

[16] Lorenceau, J. and Alais, D., "Form constraints in motion binding", *Nature Neuroscience, 4(7), 745-751*, 2001.

[17] Giese M.A. and Poggio T., "Neural mechanisms for the recognition of biological movements", *Neuroscience, 4, 179-192*, 2003.

[18] Riesenhuber, M. and Poggio, T., "Hierarchical models of object recognition", *Nature NeuroScience, 2, 1019-1025*, 1999.

[19] Calvert, G.A. and Campbell, R., "Reading speech from still and moving faces: The neural substrates of visible speech", *J. of Cog. Neuroscience, 15(1), 57-70*, 2003.

[20] Allison, T., Puce, A. & McCarthy, G., "Social Perception from visual cues: role of STS region", *Trends in Cognitive Science, 4, 267-278*, 2000.

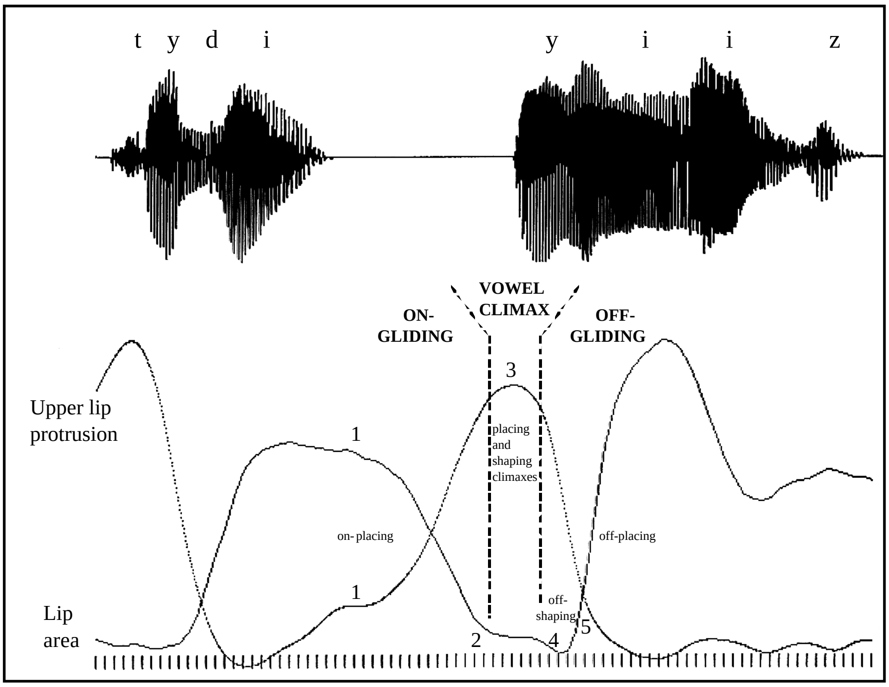

Figure 1: Acoustic signal (above) and time course (below) of upper lip protrusion and lip area for the sentence «Tu dis: UHI ise ?». On the horizontal axis, the video frames are indicated by vertical ticks every 20 ms. The following events are indicated. Event 1 corresponds to the constriction and protrusion movement onset for [y] and event 2 to the beginning of the constriction plateau. From event 1 to event 2, phase 1-2 can be considered as the on-gliding phase of the vowel [y]. Event 3 corresponds to the protrusion maximum, index of vowel climax, together with constriction plateau, i.e. event 2. Event 4 is a slight constriction area decrease leading to narrowest constriction: this is what we call «off-glide epenthesis», in this case a [ɥ]-glide. Event 5 corresponds to an increase in area of constriction together with the protrusion decrease towards the following [i] vowel. Phase 4-5 is the off-gliding phase of [y], during which an «off-glide epenthesis» occurs. It is also possible to read events on our time functions in terms of placing and shaping commands, as indicated by lower case labels (see text).

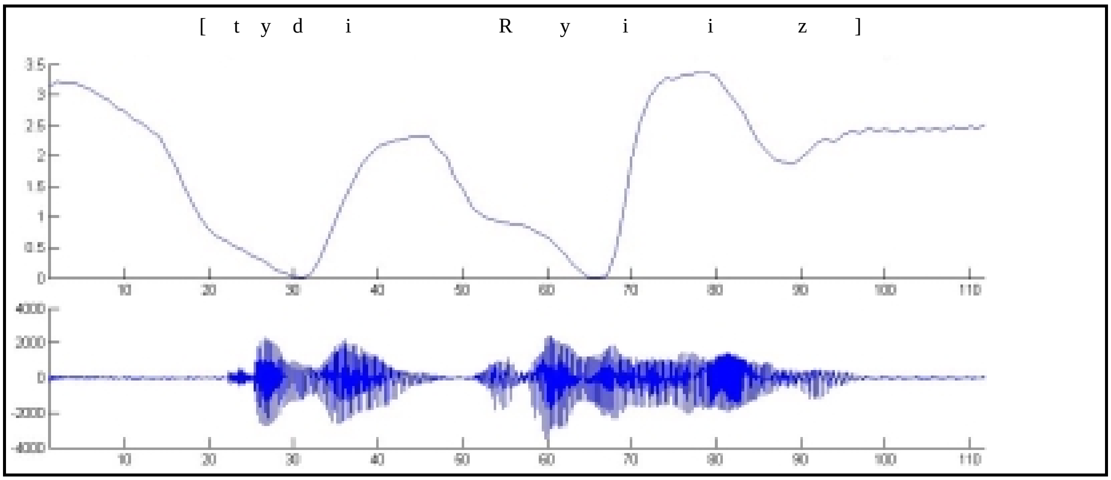

Figure 2: Acoustic signal (below; x-axis: video frame number every 20ms) and time course of lip area (above; cm2) for the sentence «Tu dis RUHI ise?». See figure 3 for area values close to 0cm2.

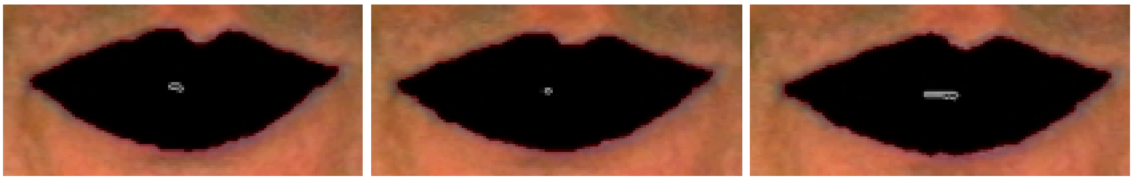

image 32A : S = 1.47 mm2 image 32B : S = 0.5 mm2 image 33A : S = 3.62 mm2

Figure 3: Front images extracted from the sequence: «Tu dis: RUHI ise?» with lip area measurements.
A minimal lip area constriction of 0.5mm2 is reached between the [y] and the [i] of «RUHI».

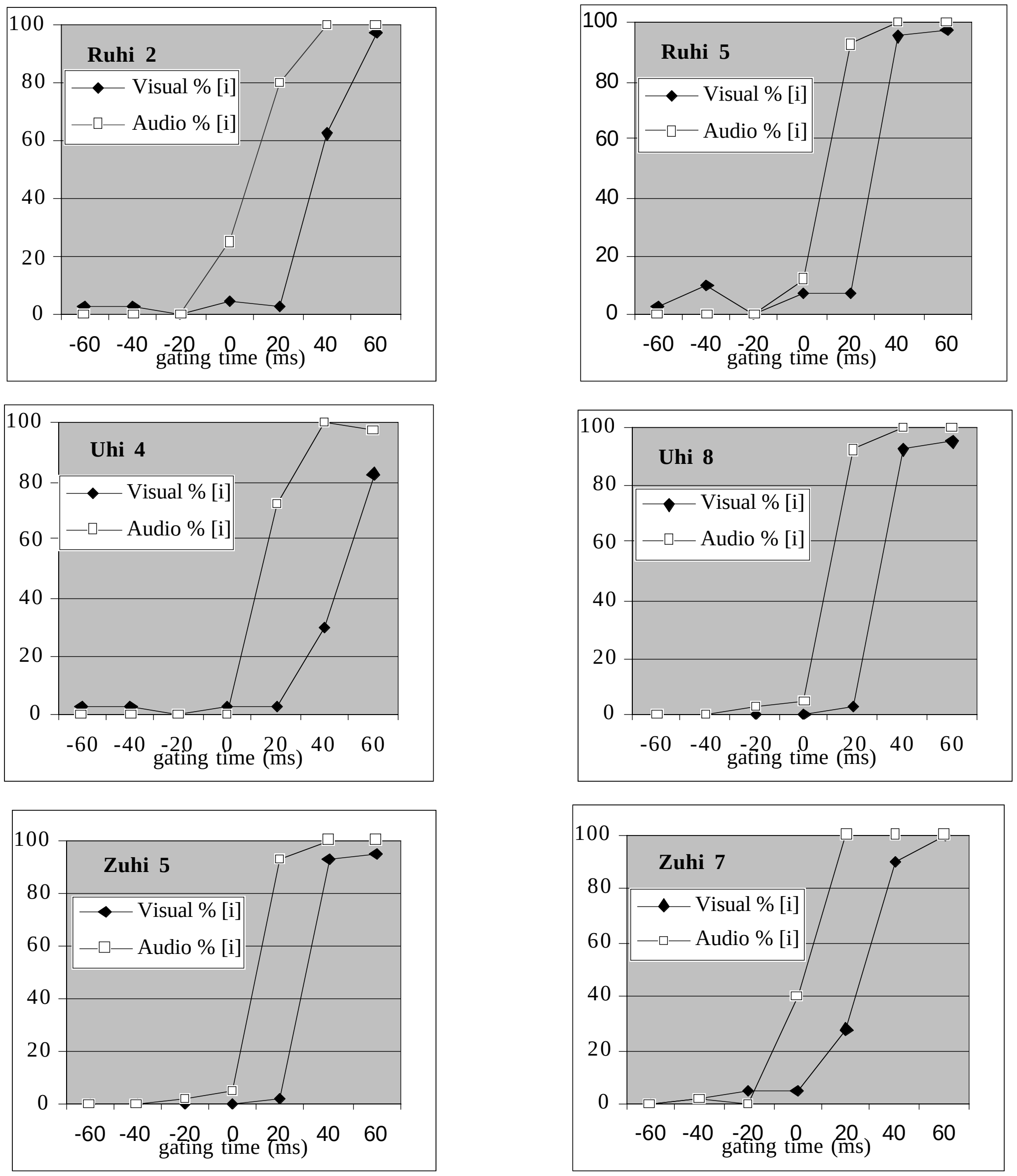

Figure 4. Audio and visual [i] identification curves for 6 stimuli (RUHI 2 & 5, UHI 4 & 8, ZUHI 5 & 7; 20 subjects for each). The horizontal axis represents the gating date (ms) relative to the minimal constriction (at zero).

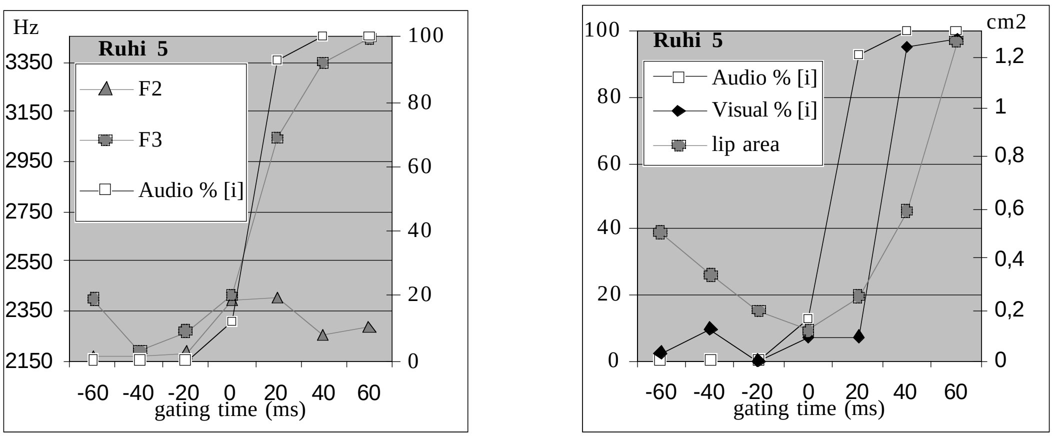

Figure 5. For a “RUHI” sequence: left, evolution of F2 and F3 formants (Hz) and audio [i] identification curve; right, time course of lip area (cm2) with audio and visual [i] identification curves.