SILENT SPEECH PRODUCTION: ANTICIPATORY BEHAVIOUR FOR 2 OUT OF THE 3 MAIN VOWEL GESTURES/FEATURES WHILE PAUSING


Institut de la Communication Parlée
UPRESA CNRS 5009, INPG/ENSERG – Université Stendhal
46, Av. Félix Viallet, F-38031 Grenoble Cedex 01, France
Email: abry@icp.grenet.fr - Fax: (33) 76.82.43.35
*Département de Psychologie, Université Mendès-France, Grenoble

Abstract

This contribution deals with anticipatory modelling during pausing in clear speech, through a set of production and perception experiments performed in the frame of the "silent pause paradigm" for vowel-to-vowel gestures. A new model, first developed classically on upper lip protrusion behaviour for [i]-to-[y] rounding gestures through consonants, then extended to lip constriction, the MEM (Movement Expansion Model), is tested in pausing both for (i) rounding and (ii) for vowel [i]-to-[a] height gestures. (i) This procedure allows to "subtract", when it occurs, consonantal influence in the building up phase of the rounding constriction. (ii) Expansion of the opening phase is also evidenced for height dimensions. Beside its advantages, this "silent pause paradigm" introduces more directly, than juncture through consonants, to articulatory prosody of pause control, as reflected in movement time, amplitude and peak velocity, for the two main vowel visible gestures.

1. Vowel rounding

Movement Expansion Model (MEM, Abry & Lallouache, 1995a,b) is an alternative to other models available in the field of speech anticipatory behaviour: the look-ahead [LA], time-locked [TL] (now frame or coproduction) and hybrid ("LA+TL") models (from Benguerel & Cowan, 1974 to Perkell & Matthies, 1992).

Fig.1.- A sketch of the MEM model for the protrusion (left) and constriction (right) components of the rounding [i]-to-[y] gesture. Expansion from zero [iy] to 5 intervocalic consonants [ikstsky]. For [iy] and [iCy], protrusion Movement Time (MT), and Time falling+Hold (Tf+H) for constriction (lip area), have been set schematically at 140 ms. The expansion of the movement starts about 100 ms for a one-consonant [iCy] Obstruction Interval (OI). The number beside each subject refers to his slope or movement expansion coefficient.
It was developed and tested, as other models, typically on the rounding gesture.

1.1. Protrusion MEM

It was shown first, for upper lip protrusion in French, that Movement Time MT was linearly related to the duration of the string of consonants (Obstruence Interval or $OI$), produced together with an [i]-to-[y] transition. The slope of this relation was speaker-specific. MEM specifies for each subject a basic MT, typically 140 ms for [iy] (Fig. 1, left), as well as an expansion function, starting from about 100 ms for $OI$ in [iCy]. Speaker-specific parameterization is clearly not in favour of a generalisation of either the LA nor TL models or current modified versions. So in our [i]-to-[y] transitions, the anticipation of the protrusion movement was not determined by the end of the unrounded vowel [i], like in LA: in Fig. 1 (left), only one subject, Annie, displays such a behaviour, with a slope near 1. Neither was anticipation determined in a fixed way in relation to the acoustic onset of the rounded vowel [y], like in TL: no subject had a relatively constant MT, i.e. zero slope was not observed for our other two talkers, Jean-Luc and Benny, whose coefficients were about .50. Contrasting sharply with quantitative data published for English (Perkell & Matthies, 1992), correlations were all quite high, ranging from .80 to .96 ($r = .32$ at $p = .01$).

Notice than one of the 4 subjects under examination, Christophe – in spite of being French! – had a too small protrusion amplitude (a few tenth of mm) to be accurately processed. Consequently we extended this Protrusion MEM to the time course of between-lips area, providing a Constriction MEM.

1.2. Constriction MEM

Fig. 1 (right) shows a schematization of talkers’ behaviour in a Time falling ($Tf$) plus Hold ($H$) phase vs. $OI$ plot ($Tf$ goes from 90% to 10% of area amplitude; $H$ from these 10% to the following 10%). Correlation values were also very high (from .87 to .99). When considering the slopes, the same talker as for upper lip protrusion, Annie, was the only one to approach the LA model (with .93); the three others behaved in a rather small range (between .69 and .79), higher than for upper lip slopes (Fig. 1, left), but still not in the orthodoxy of LA (not to speak of TL).

1.3. Rounding and pausing

This “unavoidable” Constriction MEM – at least for subjects who don’t move their upper lip (one French... and many English speakers!) –, is behaviourally efficient, but it does not give always access to the proper control of the vowels. Contrary to upper lip protrusion, the building up phase ($Tf$) for this constriction, is more influenced by a possible speaker-specific jaw recruitment, especially for coronal consonants like [s], due to the carried lower lip. This is evidenced in Fig. 2 for Jean-Luc, where regression lines are very different for items with or without [s]. Looking at Fig 3, where the no-[s] line from Fig. 2 has been superimposed, this expansion function leads from no pause (recorded in Fig. 2), through short and medium pause intervals ($PI$), to the majority of long pauses, where more scattering occurs. So it is possible for a given speaker to model the time course of the lip control specifically for the vowel.

---

First ETRW on Speech Production Modeling, Autrans, France, May 1996

102
2. Vowel height in pausing

What about the time expansion of another basic vowel gesture, height?

For one of the same talkers, Jean-Luc, we recorded internal movements for [i]-to-[a] transitions using an Articulograph AG100. Upper teeth as a reference point; lower incisors for mandibula; tongue dorsum coil (placed about 3 cm behind tongue tip); tip was also recorded, but will not be used being too sensitive to consonant movements; no lip coil was used since we planned to use synchronous video (which was finally not practicable in this session; in 3.2, we use visible data from a preceding session). Signals were sampled at 100 Hz and smoothed. We used the same procedure as for lip constriction: a Time falling (Tf) phase was defined from 90% to 10% of movement amplitude (A).

2.1. Tongue-jaw opening

On Fig. 4, which is for the synergetic [i]-to-[a] movement (tongue dorsum coil vertical position in the sagittal plane), the trend to expansion begins in the change from short to medium pause. Long pauses are very scattered: only a minority adopts the coefficient of expansion initiated in the medium pause interval (PI), the majority displaying a lower expansion rate.

![Fig. 4](image-url)

It is interesting to mention that these six "scattered" observations group with the others, when mean or peak velocities are considered in Velocity vs. PI plots. In clear, very different Tf are correlated with movement amplitude (A); consequently mean and peak velocities are always small in this condition. So following the increase in velocity for the no-pause to the short pause change, which is due to a clear marking of the short pause - and this is performed for this talker mainly by an increase in movement amplitude (duration being rather stable from no-pause to short pause, Fig. 5) - the regular decrease of velocity from the short to the long condition (values for this last one rejoining values for the no-pause instruction) corresponds to very different strategies when the pause becomes extra long. We cannot develop here this analysis of the scattering of the long pausal condition. But such an analysis of prosodical behaviour is obviously necessary to cope with gestural variability, which was documented also for [i]-to-[y] gestures through consonant strings.

So globally one can say that there is a trend which is not exactly the same if one considers the "end effector", tongue dorsum, or the whole opening gesture, due to tongue and jaw lowering.

3. Rounding and height ahead of the sound

Visual identification tests performed with [i]-[y] stimuli for short and long pauses have shown an anticipatory benefit from 100 to 200 ms (Cathiard, 1994). We compare here 160 ms short pause stimuli taken from videos of Jean-Luc, for rounding as well as for height.
3.1. Visible rounding

For [i#y] transitions, upper lip protrusion P1 starts at the end of [i] (Fig. 6), together with the constriction, i.e. lip area S begins to decrease.

![Fig.6.- Mean visual identification results for [tadi#y] superimposed on normalized articulatory parameters -S and +P1 (scores are plotted on the last image of the gate with 40 ms steps). Visual information was explored through 10 gates. This [i]/[y] identification test was performed by 10 naive French subjects. The boundary in the [i#y] transition was measured on the mean curve of all subjects using Probit analysis. It takes place 140 ms before the acoustic onset of [y], and less than 40 ms are enough to switch from 80% [i] to 80% [y] (Fig. 6). Anticipation is earlier (140 vs. 100) and category switching is steeper (40 vs. 80) than the one we obtained with another 160 ms paused signal, whose articulatory profile was actually slower, especially in the constriction building up (Cathiard, 1994). Anyway, this 140 ms anticipation remains shorter than the maximum case we ever evidenced, within a long 460 ms pause: 210 ms (Cathiard, 1994).

3.2. Visible height

Identification tests were performed with the same subjects in the [i#a] transition shown on Fig. 7. The time courses of lip area increase and visible jaw lowering (measured by M, a point on the chin) are quite synchronous. The visual boundary takes place 160 ms before the acoustic onset of [a], and about 80 ms are necessary to switch from [i] to [a]. Compared with rounding, in spite of this small advance, more fluctuation across subjects boundaries and less overall steepness correspond both to the articulatory profile of labio-mandibular movements (for area see Fig. 6) and to the well known less visible salience of the [i] vs. [a] contrast.

![Fig.7.- Mean visual identification results for [tadi#a] superimposed on normalized articulatory parameters +S and -M.](image)

4. Lack of space for the vowel space

We are perfectly aware that we just put together production and perception data as building blocks of a vowel space, which is still lacking more extensive exploration. The timing of a building up phase (like Tf) is surely very important for gaining insight into the control side and in the timing of perceptual boundaries. So when will the MEM be tested with the last main basic vowel gesture, i.e. front-back dimension, in [y]-to-[u] transitions?

Acknowledgements: To P. Pinoit for his collaboration in editing articulatory signals and to P. Mahé who first tested the Articulograph. Work partially funded by the EC ESPRIT/BR project Speech Maps, a France Telecom contract, and a Cognitive Science Programme grant.

References