WHAT RELATIONSHIP BETWEEN PROTRUSION ANTICIPATION AND AUDITORY PERCEPTION?

*Rudolph Sock, Béatrice Vaxelaire, Véronique Hecker & Fabrice Hirsch

Institut de Phonétique de Strasbourg E.A. 3403 - Université Marc Bloch
†Institut de la Communication Parlée de Grenoble UMR 5009 - Université Stendhal/INP Grenoble
sock@umb.u-strasbg.fr

ABSTRACT

This investigation, while confronting data from two speakers, examines the contribution of kinematic and acoustic attributes of protrusion to early perception of a French rounded vowel. The main question addressed here is to determine the relationship between extension of auditory perception and anticipatory expansion of lip protrusion. If extension of auditory perception does increase with anticipatory expansion of lip protrusion, it would be important to determine the nature of the motor-sensory relation, and also the extent to which speaker-specific characteristics may influence the perceptual behaviour of listeners. The robustness of the temporal extent of the perceptual effects is also evaluated under increased speaking rate.

1. INTRODUCTION

The relationship between the motor level of speech production and the auditory perceptual level is highly non-linear and complex. However, robust spatio-temporal articulatory structures, associated to phonological categories that exist in a given language correspond necessarily to those that can both be articulated and correctly perceived. Consequently, production and perception must mutually impose constraints on each other (Maeda, 1999). The auditory perceptual system must impose its constraints on the selection of speech communication sounds, and the biomechanical machinery of the production system must have evolved in order to produce a zone of variation of sounds exploitable by existing auditory perceptual capacities. It is therefore judicious to want to identify regularities between the articulatory and the acoustic-auditory levels as evidence for reciprocal constraints imposed by the two systems. The emergence of coherent behaviour in the auditory domain that can directly be rationalised in articulatory terms would not only suggest that the two systems are tightly coupled but also help to reveal some aspects of the complex relationship between articulation and auditory perception. In this perspective, the major question addressed in the present study is to determine the relationship between kinematic properties of anticipatory speech gestures and any eventual precocious auditory effects of such anticipatory gestures. More specifically, the problem examined here is to find out if auditory perceptual extension follows anticipatory lip protrusion expansion. If such were the case, it would be important to determine the extent to which speaker-specific anticipatory characteristics may influence the perceptual behaviour of listeners. Such results would enable us to test the perceptual aspects of an anticipation model for French, the Movement Expansion Model — MEM. This model (Abry & Lallouache, 1995), elaborated from French data for lip protrusion, shows that the usual trend is to anticipate lip protrusion when there is enough time available (length of the pre-vocalic consonant string, for e.g.), with high correlations between this available time and movement duration. Correlations between these two factors also reveal speaker-dependent anticipation ratios.

Previous results, using the gating paradigm adopted here, have shown, for example, that an anticipatory protrusion gesture is perceptually an efficient component of the production of a rounded vowel (Lubker & Lindgren, 1982); its temporal extent is greater in the visual than in the auditory domain (Cathiard et al., 1996; Hecker et al., 1999; Sock et al., 1999, Vaxelaire et al., 1999). This investigation, while confronting data from two speakers, examines the contribution of kinematic and acoustic attributes of protrusion to early perception of a rounded vowel. The robustness of the temporal extent of the perceptual effect is also evaluated under increased speaking rate.

2. METHODS

2.1. Speaker and speech sample

Two adult native speakers of French (JLS and CS) served as subjects. The uttered speech samples consisted of sequences of 3 types: type I is V1V2, where V1 is [i] and V2 is either [e], [a] or [y]; type II is V1CnV2, where V1 is the spread vowel [i], Cn is a sequence of 2 non-labial consonants [ks], and V2 is either [i], [a] or [y]; type III is V1CnV2, where V1 is the spread vowel [i], Cn is a sequence of 5 non-labial consonants [kstk]. The speaker produced the sequences, that were inserted in a carrier sentence, several times at a normal conversational rate, and at a self-selected fast rate.

2.2. Acoustic and movement data

The movement and acoustic data were recorded simultaneously using a three-coil transmitter system, with five transducers (AG 100, Carstens Medizinelektronik). Instantaneous velocity and acceleration signals were also software calculated. A representative example of the audio and the corresponding movement signal, for each sentence was retained in both speaking rates. Two acoustic events, corresponding to the offset and onset of a clear formant structure for vowels [i] and V2 respectively, served to determine the obstruent interval of the
consonant string. Only the protrusion movement of the upper lip was retained to determine the anticipatory timing of the lip rounding gesture. The kinematic events, detected on the first and second derivatives of the upper lip position signal, are: movement onset or Minimum Protrusion; Peak Acceleration; Peak Deceleration and Maximum Protrusion. The purpose of analysing the position signal was to be able to determine the occurrence of each kinematic event with respect to the acoustic onset of the rounded vowel, thus having motoric events as landmarks for the gating decisions made on the acoustic signals. These gating decisions were then transposed to sentences that did not present protrusion, by truncating at comparable acoustic regions within the obstructed string.

2.3. Gating procedures and test tape elaboration

Gatings were made on the 36 digitally stored acoustic signals. Six gatings, located at different temporal distances from the acoustic onset of V2, were the following: Gate 1, based on Minimum Protrusion; Gate 2, based on Peak Acceleration; Gate 3, based on Peak Velocity; Gate 4, based on Peak Deceleration; Gate 5 corresponded to Maximum Protrusion; Gate 6 corresponded to the onset of the clear formant structure of the vowel. Thus each "gated" sentence consisted of the carrier sentence plus an increasing amount of the acoustic information from the consonantal string (in the case of the \V CnV2 sequences) preceding the rounded vowel. Total durations and the acoustic obstruent interval were compressed in fast speech; so also was Movement Time (which is the distance between Minimum Protrusion and Peak Protrusion). Due to these compressions of acoustic and kinematic durations in fast speech, the data were normalised. Data analyses systematically take into account the normalised data, to ensure that differences in results are not simply due to a linear compression of durations in fast speech but rather to different coarticulatory behaviours in a given prosodic condition. The test tape contained 216 gated-out stimuli (108 in each speech rate) that were then randomised, one series in normal speech and the other in fast speech. A bip, serving to alert the listeners as to the imminence of a stimulus, preceded each utterance by 1.4 seconds. Inter-stimuli interval was 4 seconds with a 10 second pause following each twelfth item. The test tape began with a training list of 12 items, then continued with the 108 stimuli in a given prosodic condition.

2.4. Listener judgements

15 adults, all native speakers of French served as listeners for the perceptual experiment. The test tape was played separately to all subjects. Score sheets were provided and subjects had to carry out two tasks for each of the 216 utterances: (1) mark with a cross which vowel he or she believed to have been "gated-out" of the particular test stimuli, and (2) check on a 5-point scale the level of confidence with which they were making each judgement.

3. RESULTS

The mean level of confidence rating, averaged over all 216 utterances by all 15 listeners (totalling 3240 responses) was 2.4 with a standard deviation of 0.91, thus showing that overall judgements were made with a good degree of confidence, in both speech rates. Moreover, the percentage of correct responses was highly correlated with confidence ratings (r=0.79 in normal speech and r=0.81 in fast speech), indicating that while being confident subjects were at the same time being highly performant with regards to the identification task required of them. Also, subjects seemed, in this specific experiment, to be particularly severe with the use of confident ratings on the subjective 5-point scale.

3.1. Auditory perceptual extent in \V1V2

As expected, no anticipatory protrusion movement is observed in the \[ie\] and \[ia\] sequences. Thus any precocious identification of the two vowels is to be attributed mainly to the anticipatory lingual gesture, in attaining the desired vocalic configuration. Results, obtained in the two prosodic conditions for both speakers, show that the percentage of correct responses for these vowels is significant (superior to 60%) at Gate 4 (Peak Deceleration). Listeners also recognise the truncated \[y\] either at this date 4 or at the previous one (date 3 or Peak Velocity). The analyses of the three formant trajectories tell us, in the absence of a labial gesture (\[e, a\]), that the date of identification corresponds, more or less, to the first inflexions of these formant trajectories (with F1 increasing, and F2 and F3 diminishing). In the case of \[y\], it is F3 that seems to account exclusively for anticipatory recognition of the vowel; identification coincides with the first sign of trajectory inflexion for F3. It is also interesting to note that the kinematic date for recognition of this protruded vowel, i.e. Peak Acceleration, is located in the vicinities of the remarkable F3 inflexion. However, identification becomes optimal when F3 merges with F2 or crosses the latter (approx. at Peak Deceleration).

3.2. Auditory perceptual extent in \[iksV2\]

In this context, no extensive anticipatory tongue gestures are observed for the unrounded vowels. The inferior limit of noise concentration in the obstruent interval is located around 6kHz, where it remains relatively stable throughout the interval. Consequently, identification of the two unrounded vowels occurs, in all cases, at onset of a clear formant structure for the vowel, or at best, at 8 ms from this formant structure. In this 8 ms coarticulatory interval that corresponds to the delay between onset of vocal fold vibrations and appearance of a clear formant structure (VOT), the energy for the fricative [s] drops in frequency and intensity. When the second vowel is the rounded vowel \[y\], recognition point is at date 3, for JLS (see Figure 1), i.e. Peak Velocity (located at 170 ms and 122 ms from onset of this vowel, in normal and fast speech respectively). Identification performances are globally high in both speech rates (r=0.93 and 0.92 in normal and fast speech respectively).
For CS, recognition occurs at date 2 (Peak Acceleration), located at 124 ms from the vowel. On the acoustic level, these recognition dates are located around onset of the spectral slope of the inferior noise limit, which is at 3.5 kHz approximately. This variation of the inferior noise limit is provoked by anticipatory lip rounding which begins at onset of friction for [s]. It is known that lip rounding increases the size of the front cavity, thus reducing pole frequencies and consequently that of the principal noise concentration. In fact, the trajectory of the inferior noise limit seems to drift towards the F3 formant value of the rounded vowel. Sensory-motor relations here suggest a correspondence between Peak Acceleration (CS) or Peak Velocity (JLS) and onset of the spectral slope of the inferior noise level.

3.3. Auditory perceptual extent in [ikstskV2]

Globally, the scenario in this context is similar to the preceding one. For both speakers, and regardless of the prosodic condition, no anticipatory recognition is possible before onset of a clear formant structure of the two unrounded vowels. Maximum loading of the obstructed interval indeed hinders anticipatory behaviour, even in the coarticulatory zone of voice onset time. As concerns the [ikstsky] sequence, recognition is at date 1 (Minimum Protrusion), in both speech rates for speaker CS (at 226 ms and at 232 ms from the vowel in normal and fast speech respectively); recognition is also at date 1 for speaker JLS (see Figure 2) in normal speech (302 ms before the vowel) and at date 2 in fast speech (202 ms from the vowel). Although performances seem to be better in normal speech at extreme gatings, results in the two speech rates become comparable as more motor-sensory information becomes available.

On the acoustic level, this very surprising identification of the rounded vowel at date 1, localised either in the transitional phase between the vowel and the [k] consonant (Voice Termination Time - VTT), or in the silent acoustic portion of the [k], seems to be provoked by a slight lowering of the F3 trajectory of the preceding vowel [i], a trajectory that “points” towards the F3 value of the distant [y]. However, it should be noted that robust identification of this vowel is related to the variation of the inferior noise limit that only begins in the fricative portion of the interval, situated at Peak Velocity (at 238 ms and 84 ms for JLS and CS, respectively).

4. Auditory perceptual extent as a function of protrusion expansion.

Accepting, in this specific investigation, that listeners were particularly severe with the use of confidence ratings, it is judicious not to take into account relatively low ratings when correct responses are strikingly high (approx. 80% of correct responses). In this case, the auditory perceptual extension indeed follows the anticipatory expansion of the protrusion movement, extending from the non-obstruent context (V1V2-rounded) to the 5-consonant obstructed context (V1CnV2-rounded). Figures 3 and 4 illustrate these findings. They show, on the y axes, the dates at which correct responses are superior to 60%, as a function of expansion of the acoustic obstructive interval (or length of the consonantal string). It can be noticed that perceptual extension increases as the interval increases too. In other terms, the tendency is to increase the anticipatory auditory efficiency of the gesture with increase in the length of the anticipatory obstructive domain. Consequently the perception data analysed here are in line with predictions of the MEM model or, at least, do not infirm this Model.
This elasticity of movement time (MT) and of the obstruent interval are thus capable of provoking noticeable differences in coarticularatory perceptual extents. In fact, results highlighted in figures 3 and 4, show a structurally similar behaviour in the anticipatory perceptual extent of listeners, but relative variations in patterns for listeners, reacting to one speaker or the other. These differences in listener behaviour are largely due to configurational differences in the façade of the protrusion gesture, and particularly to the audible portion of that gesture, being robust as from Peak Velocity.

4. REFERENCES


