Temporal coordination of articulatory gestures in consonant clusters and sequences of consonants

Hélène Loevenbruck, Michael J. Collins, Mary E. Beckman, Ashok K. Krishnamurthy, Stanley C. Ahalt

Temporal coordination of articulatory gestures in consonant clusters and sequences of consonants.

H. Lœvenbruck[1,3], M.J. Collins[2], M.E. Beckman[1], A.K. Krishnamurthy[2] and S.C. Ahalt[2]

1 Department of Linguistics - The Ohio State University, 222 Oxley Hall, 1712 Neil Avenue, Columbus, OH 43210-1298
2 Department of Electrical Engineering – The Ohio State University, 205 Dreese Lab, 2015 Neil Avenue, Columbus, OH 43210
3 Institut de la Communication Parlée – INPG/ Université Stendhal/ UPRESA CNRS 5009 – 46 av. F. Viallet – 38031 Grenoble Cedex 01 – France

Abstract

In the framework of Articulatory Phonology, linguistic structures are represented in terms of coordinated articulatory gestures that are organized into a gestural score. Syllable structure can be seen as residing in the phasing of the dynamical gestures.

The present study investigates the temporal coordination and cohesion of some articulatory gestures as a function of their respective configuration. It looks at the effect of speaking rate and accent on the articulation of the /kl/ or /skl/ clusters and on the same sequence of consonants separated by a reduced vowel (/kəl/). Articulatory-phonetic representations inspired by gestural scores were recovered independently for both the acoustic and EPG signals that were recorded for two native American English speakers. Constriction locations and degrees were inferred from the EPG data by dividing the palates into place-of-articulation regions and tabulating the percent of electrodes contacted within these regions. These measurements support a Gestural Phonology account of the difference in the production and cohesion of /kl/ and /skl/ vs. /kəl/. In a parallel modeling study, temporal events were recovered from the acoustic signal using a temporal decomposition procedure. These events and their relative timing are compared to the EPG-based gestures and their arrangement in the gestural score.

e-mail addresses :

loeven@icp.inpg.fr, mcollins@ee.eng.ohio-state.edu, mbeckman@ling.ohio-state.edu, akk@ee.eng.ohio-state.edu, sca@ee.eng.ohio-state.edu

Introduction

In many traditional non-articulatory phonologies, a hierarchy of phonological units is posited, with features as the most basic categories. The segment is viewed as a unit of co-occurrent feature specifications and the syllable as a metrical unit dominating a sequence of segments. Positing such a hierarchy can explain, for instance, certain variations in phonological units (segments) by their position within the hierarchical structure (syllable).

In contrast, within the framework of Articulatory Phonology (Browman & Goldstein, 1990), the primitives are not features but gestures; units such as segments and syllables are epiphenomena naming characteristic patterns of coordination among gestures and it is often possible to account for allophonic variations without referring to any hierarchical syllable structure (Browman & Goldstein, 1995). Linguistic structures are represented in terms of coordinated articulatory gestures that are organized into a gestural score (see figure 1). Each gesture corresponds to the formation of a constriction within one of the relatively independent articulatory subsystems of the vocal tract. The gestural score is organized into articulatory tiers— three oral tiers of lips, tongue tip and tongue body, a velic tier, and a glottal tier. These articulatory tiers are in accordance with the proposals of some autosegmental phonologists, who suggested tiers that correspond to independent articulatory systems (e.g. Clements, 1985). As shown by the gestural score in figure 1, the temporal activation interval for each gesture on its tier is represented by a rectangular box. It depends on the dynamical specification of the individual gesture itself and on the intergestural phase relations. Solid lines connect pairs of gestures that are critically phased to one another. When two gestures are critically phased, some phase of the first gesture must occur during some phase of the second gesture. The actual length of the critical common phase is not specified by the solid lines. The gestural structure (the solid lines without the boxes) is supposed to be fixed, it is a lexical property of a form. But the gestural parameters and the phase relations among gestures (the actual durations of each of the boxes and the lengths of the potential overlaps) are variable and depend on speaking conditions.

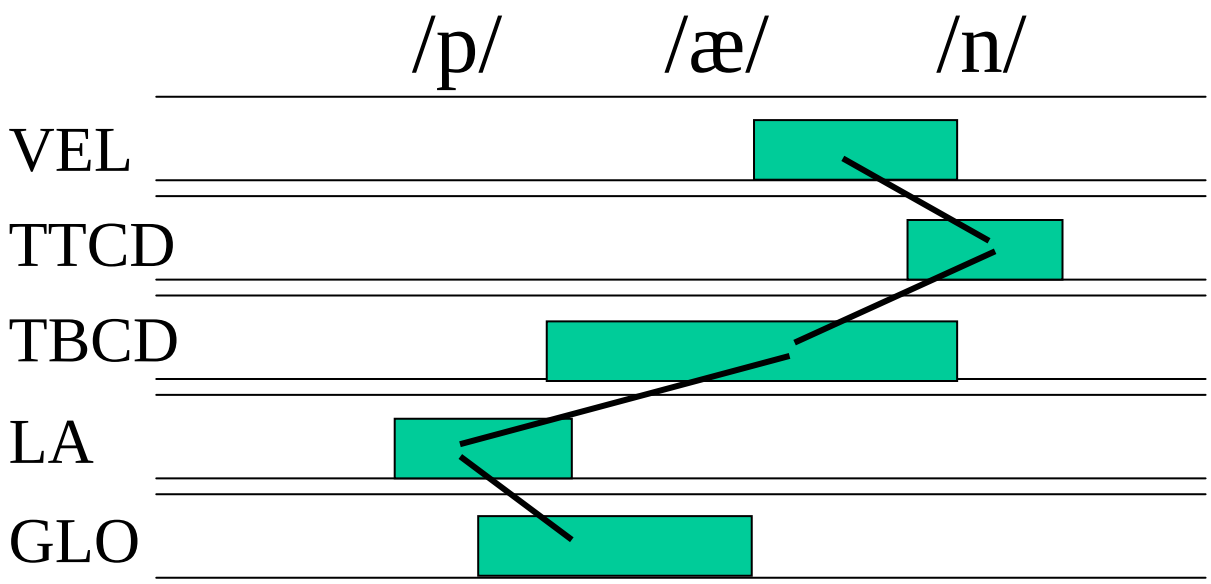

Figure 1. A Gestural Score for /pæn/. Vel: Velum, TTCD: Tongue Tip Constriction Degree, TBCD: Tongue Body Constriction Degree, LA: Lip Aperture, Glo: Glottis. The solid lines connecting pairs of gestures signify necessary phasing relationships encoded in the lexical representation (in current Articulatory Phonology, each of these would be associated with a bonding strength). The actual lengths of the necessary common phases are not specified by the solid lines. The variable boxes represent the temporal activation interval for each gesture. The temporal activation intervals and the actual overlap durations are variable and depend on speaking conditions.

In this framework, syllable structure can be seen as residing in the organization of the dynamical articulatory gestures involved. Sequences of consonants can be represented as phasing among gestures. Since the configurational properties (or the phase relations) determine the patterns of overlap among gestures, differences in configuration are automatically associated with differences in overlap, and hence, with the articulatory and acoustic consequences of such overlap. Some allophonic variations (such as nasalized vs. non-nasalized vowels or light vs. dark /l/s, see Browman & Goldstein, 1995) which are traditionally accounted for by invoking position within the hierarchical syllable structure, can be explained in the Articulatory Phonology framework using simple configurational principles.

The purpose of the present study is to investigate the temporal coordination and degree of cohesion of some articulatory gestures as a function of their respective configuration. It looks at the

effect of speaking rate and accent placement on the articulation of the /kl/ cluster and on the same sequence of consonants separated by a reduced vowel (/kəl/). Temporal information roughly similar to the gestural scores proposed by Articulatory Phonology were recovered independently for acoustic and articulatory signals that were recorded for two native American English speakers. In the first part of the paper, the articulatory-derived gestural scores are examined and a Gestural Phonology account of the difference in the production and cohesion of /kl/ and /skl/ vs. /kəl/ is proposed. In particular, the presence or absence of schwa in /kəl/ is related to the phasing of the surrounding /k/ and /l/ gestures. In the second part, the comparison of the EPG-based gestural scores with the temporal events extracted from the acoustic signal for /kəl/, suggests that the acoustic-derived events could reveal additional information about the coordination among gestures.

I. Method

I.1 Speech material

The /kl/ sequence is particularly interesting, because it involves the coordination of two articulatory subsystems: the tongue body for /k/ and the tongue tip for /l/. Three target words containing the /kl/ sequence were chosen, varying on the position of /l/ in the syllable: 'accolade' (/l/ in 1st position in the syllable), 'clay' (/l/ in 2nd position), and 'disclaimer' (3rd position).

The target words were inserted in the carrier sentence 'Say a *target_word_1* of a *target_word_2* again'. The first target word in the sentence was nuclear accented whereas the second was post-nuclear (unaccented). Each sentence was produced with five different rate instructions, categorized as slowest, slow, normal, fast, fastest.

In many traditional or non-linear phonological accounts based on the notion of "extrinsic allophone" or feature-changing "rules", the difference between the /l/s in 'accolade', 'disclaimer' and 'clay' would be explained by syllabification.

In an Articulatory Phonology account, the /skl/ cluster in 'disclaimer' is described as a Tongue Tip gesture for /s/, followed by a Tongue Body gesture for /k/ and a Tongue Tip gesture for /l/. The sequence /kəl/ in 'accolade' corresponds to a Tongue Body gesture for /k/ followed by a

Tongue Tip gesture for /l/, just as for the sequence /kl/ in ‘clay’. But in /kəl/, the schwa can be seen as a phantom acoustic product of spreading the dorsal and coronal gestures apart. According to Articulatory Phonology, the difference between /kəl/ and /kl/ lies in the configurations of the Tongue Body and Tongue Tip gestures one with respect to the other.

We should mention here however, that a detailed account of the gestures involved in the production of American English /l/s should acknowledge the presence of a Tongue Body gesture in addition to the Tongue Tip gesture. As shown by Sproat & Fujimura (1993), two basic movements are involved in the articulation of /l/: a retraction and lowering of the tongue dorsum (or body), and a forward movement of the tongue tip. Dark (post-vocalic or syllabic) /l/s have a more significant dorsal retraction gesture and a less extreme apical gesture than light /l/s (pre-vocalic). In addition to the difference in gestural extents, dark and light /l/s feature differences in the timing of the gestures. The X-ray data described in Sproat & Fujimura (1993) and the MRI data collected by Browman & Goldstein (1995) show that in dark /l/s, the dorsal gesture precedes the apical gesture, whereas in light /l/s, the apical gesture is phased before (or synchronously with) the dorsal gesture.

The presence of a tongue body gesture for /l/ could be a problem for the present study where the observed tongue body gesture could be related not only to /k/ but also to the following /l/. However, in this corpus, all of the /l/s under scope are light (prevocalic). Therefore, firstly, the potential tongue body gesture is expected to be less significant than it would be for dark /l/s. And secondly, if it is observable, it should be phased after the tongue tip gesture, so it should not prevent from examining the coordination between the tongue body gesture for /k/ and the tongue tip gesture for /l/.

I.2 Subjects and recordings

Two adult subjects (a male and a female) participated in the experiment. They are native speakers of American English and phoneticians in the Department of Linguistics at the Ohio State University. The female speaker is the third author.

The subjects were not given any reference beat and speaking rates did not fall neatly into five distinct categories that were uniform between the two speakers. Our aim was merely to collect a wide range of speaking rates, not to compare utterances across the five rate groups. That is, in any

analyses of the effect of speaking tempo, rate was treated as a continuous variable, and assessed by measuring the utterance duration directly.

Acoustic and articulatory data were recorded simultaneously, in a sound booth, using Kay Elemetrics' Palatometer. The articulatory signal consisted in linguopalatal contacts measured by electropalatography (EPG). Kay Elemetrics' Palatometer uses custom-made pseudopalates with 96 electrodes covering the entire hard palate and the inside surface of the teeth. The sampling frequency is 100Hz. The arrangement of electrodes on the pseudopalates for the two speakers is visible in figure 2. Representative frames (frames of maximum palatal contact) are shown for /s/, /k/ and /l/ in 'disclaimer'. The three successive segments correspond to distinct contact patterns: coronal for /s/ and /l/ and dorsal for /k/. Pseudo-gestural scores were derived from these articulatory data using location and amount of contact on palate.

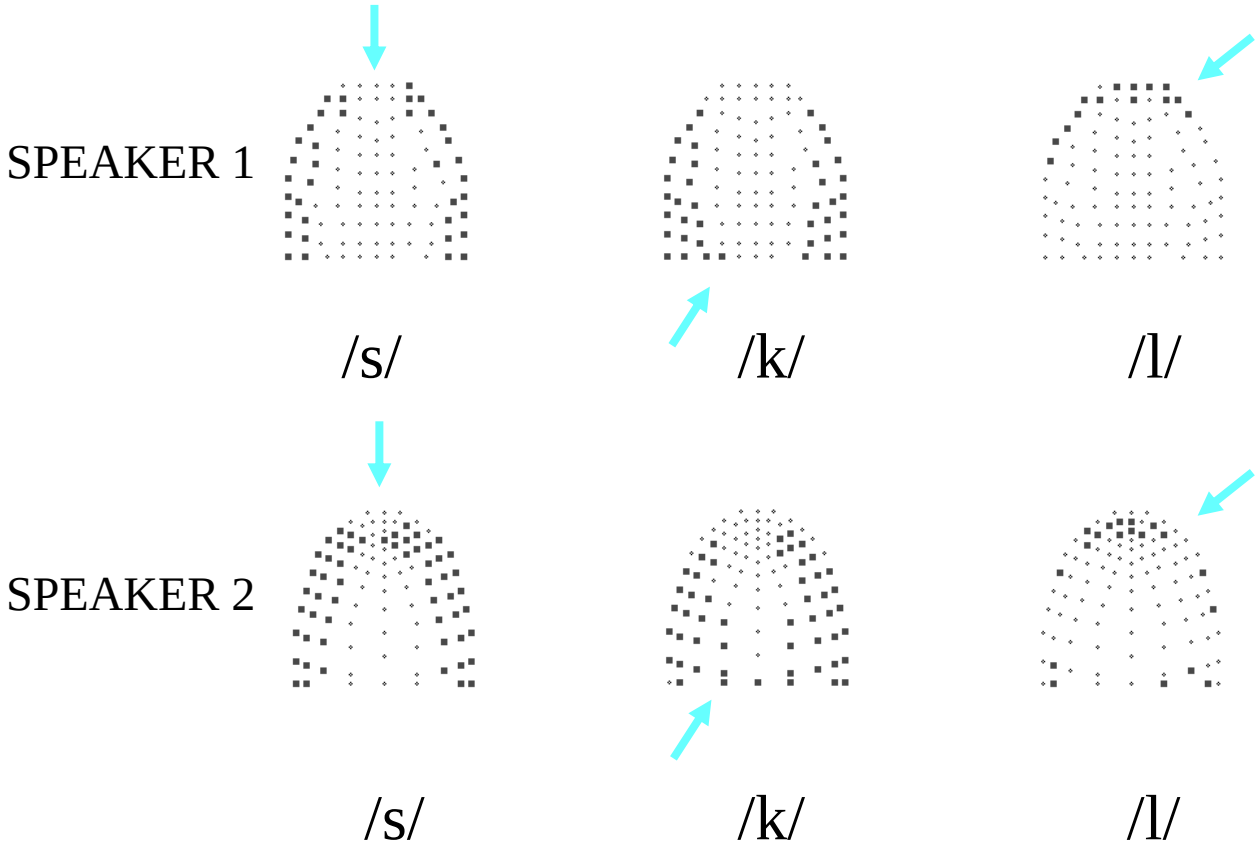

Figure 2. Representative frames for the segments /s/, /k/ and /l/ in 'disclaimer' for speaker 1 (top panel) and speaker 2 (bottom panel). Darker squares indicate contact. The arrows point to the regions of constriction (front for /s/ and /l/, back for /k/). The distance between two rows of the pseudo palate is generally 1mm (some rows are 2 to 3mm apart, see figure 4 for details).

The acoustic signal was also recorded, simultaneously with the EPG signal, via a head-mounted microphone, at a 16-kHz sampling rate. A second set of temporal events was derived from the acoustic signal using a temporal decomposition procedure. However, this analysis has been completed only for subject 1.

II. Results and Discussion

II.1 Deriving pseudo-gestural scores from EPG data

Concerning the discussion about the potential presence of a tongue body gesture in /l/ (see section I.1), it should first be noted that none of the /l/s under scope featured an observable tongue body gesture (posterior palatal contact). These /l/s being all prevocalic (light), it could be that the potential tongue body gesture was too weak to correspond to visible linguopalatal contact. In contrast, the apical gesture was rather important, as shown by the high number of anterior palatal contacts, which is in accordance with previous observations on light /l/s. Therefore all the tongue body gestures will be related to the /k/ phonemes in this study.

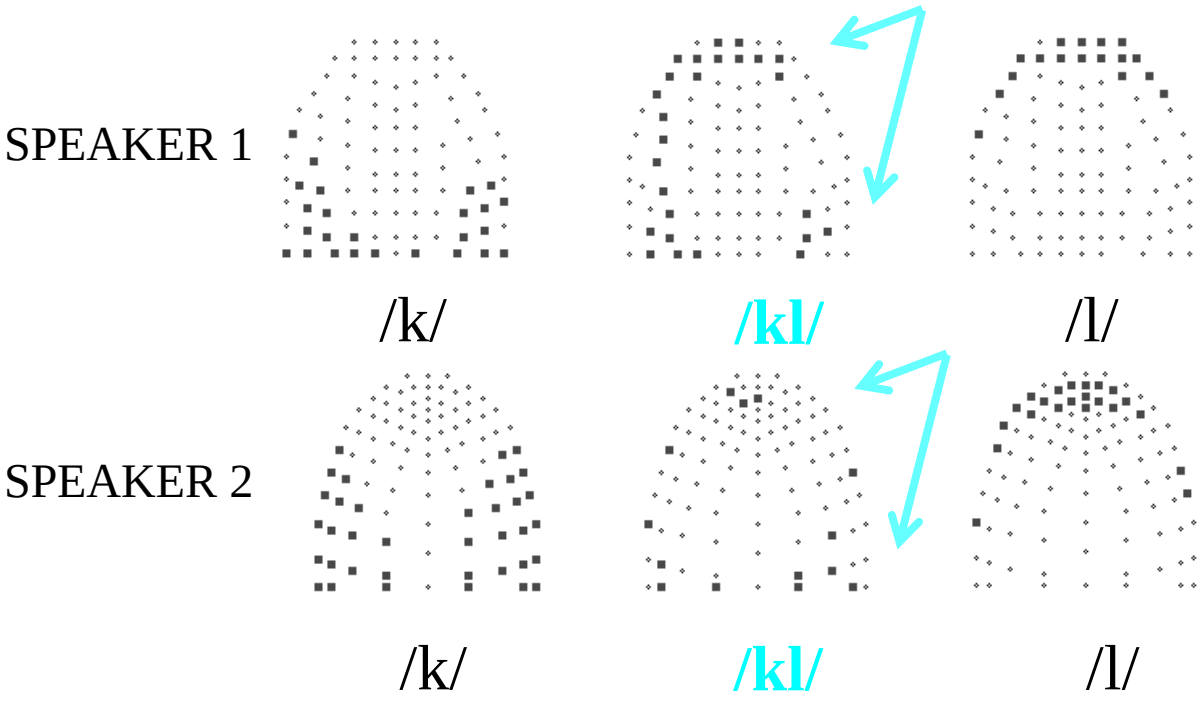

Figure 3. Representative frames for the segments /k/ and /l/ in ‘clay’ uttered at a normal rate by speaker 1 (top panel) and speaker 2 (bottom panel). Darker squares indicate contact. The arrows point to the double constriction that appears between /k/ and /l/.

A selection of three representative frames for ‘clay’ uttered at a normal rate is shown in figure 3 for the two speakers. As expected, the pattern of contact is dorsal for /k/ and coronal for /l/. An intermediate state between /k/ and /l/ is also present where two constrictions (dorsal and coronal) are visible on the palate. This corresponds to a strong overlap between the dorsal gesture for /k/ and the coronal gesture for /l/. For ‘accolade’, Articulatory Phonology predicts the same sequence of gestures, but a different intergestural organization. The dorsal and coronal gestures should not overlap as in ‘clay’, rather there should be a gap between the two gestures allowing the presence of schwa. For ‘disclaimer’, the coordination between /k/ and /l/ should be similar to that in ‘clay’.

In order to verify these predictions, gestural scores were derived from the patterns of linguopalatal contact inferred from the recorded EPG data.

II.1.1 Definition of front and back regions.

The pattern of linguopalatal contact for this corpus being either coronal (for /s/ and /l/) or dorsal (for /k/), it was decided to divide the pseudopalate into two regions: front and back. Region definition is often based on anatomical grounds rather than actual productions (see e.g. Recasens, 1984; 1990, 1991). However this method does not account for the idiosyncrasies in the articulation of specific segments by each subject, nor for the differences in the arrangement of electrodes on the pseudopalate with respect to anatomical configurations. As advocated by Byrd and Tan (1996), it was decided here to determine articulatory regions empirically for each speaker. The demarcation line (see figure 4) for each of the pseudopalates was based on the careful examination of the contact patterns for several dorsal and coronal segments (in addition to the corpus presently discussed, a number of target words involving different clusters, such as /s k r/, /s t r/, /k r/, /g r/, /g l/, /ʃ r/, /t l/, /t r/, /s l/, or /s t/, were recorded). For subject 1, the occlusion for /k/ involves central contacts from row 18 to the last (most posterior) row of the palate, whereas the production of coronals (/s/, /ʃ/, /t/, etc.) involves central contacts from row 0 to row 17 at most. For subject 2, the articulation of /k/ or

/g/ (in ‘cape’ and ‘gate’) can feature central contacts as far front as row 13, whereas the grooves for /s/ and /ʃ/ correspond to central contact before row 12. This explains the rather anterior position of the demarcation line.

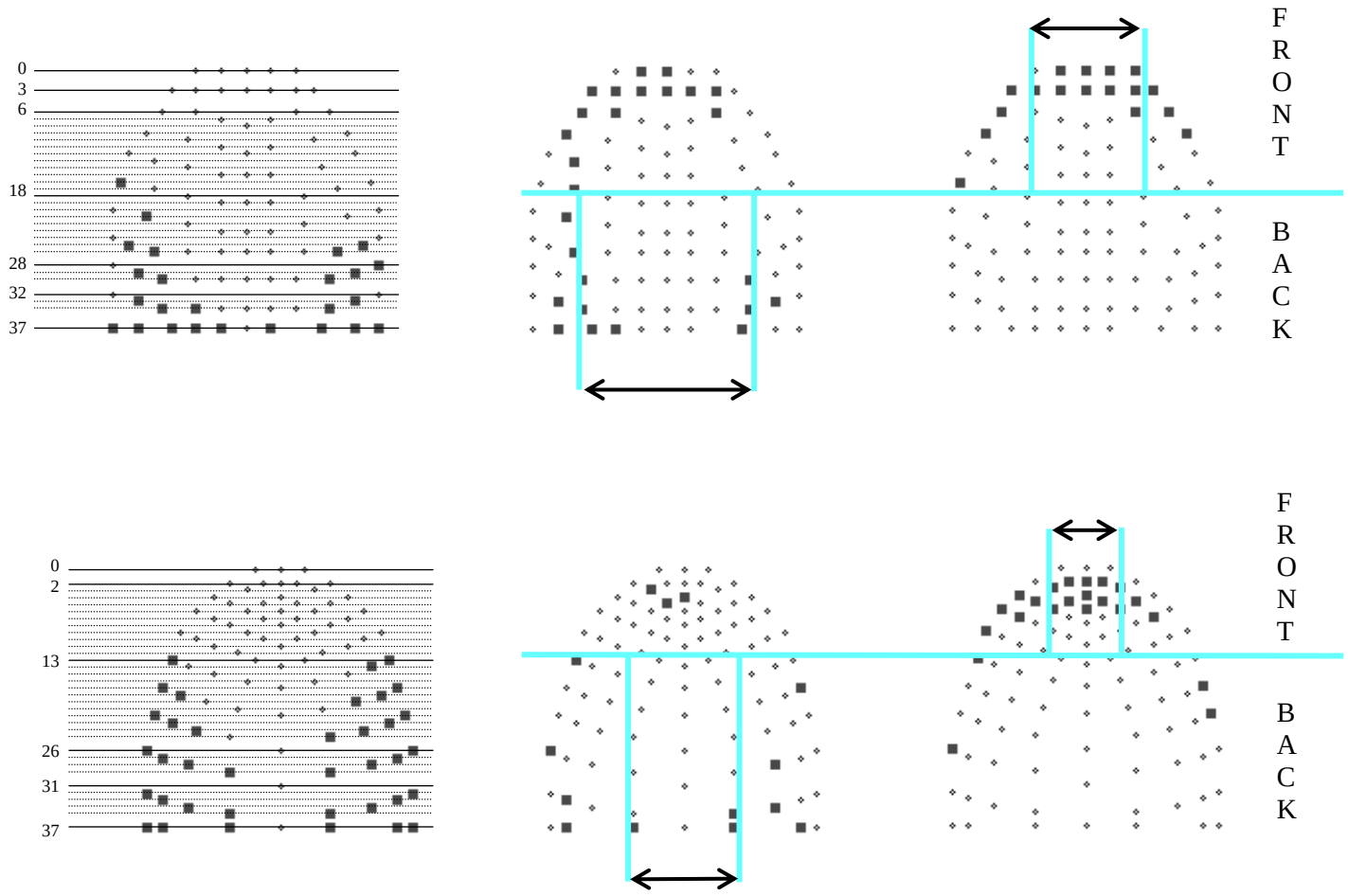

Figure 4. Definition of the front and back regions and of their central bands for speaker 1 (top panel) and 2 (bottom panel). Row numbers are detailed in the left palates, they correspond to their ordinate in mm. The interval between two rows is generally 1 mm; when it is not the case, the row number is given and the line is thicker. In addition, the first row of the back region (18 for speaker 1 and 13 for speaker 2) is highlighted.

II.1.2 Definition of a constriction

As can be seen for instance for speaker 1, in the top panel of figure 2, the occlusion for /k/ is not always fully reflected in a pattern of a full row of contacted electrodes on the pseudopalate. A

first possible reason is that the pseudopalate does not reach far enough in the rear of the mouth, so that back /k/ are not visible. The observation of the successive frames just before /k/ shows a backward displacement of the contacts. Although the midline is not necessarily contacted, there are contacts along the sides of the midline which tend to be displaced from front to back, when the articulation goes from /s/ to /k/ for instance. This contact displacement can be interpreted as a backing movement of the tongue. A second possible reason for the limited markedness of the occlusion for /k/ is that the arrangement of electrodes on the pseudopalates was devised for the articulation of coronal consonants: the density of electrodes is higher in the coronal region than in the back region. There might be zones on the palate where contact occurred, but was not recordable. A third possibility is that there was not complete contact in the /k/ in that token. However, the examination of the acoustic signal shows the presence of an occlusion which does not support this third possibility.

Furthermore, the articulation of /s/ involves the production of a sagittal channel together with strong coronal contact. This means that although the degree of contact in the front region should be high for /s/, there should be uncontacted electrodes in the sagittal zone.

Therefore, in order to account for constrictions that did not involve a full row of contacted electrodes, it was decided to acknowledge a constriction when there was contact in a central band around the midline. For each speaker, a central band was defined in each region (front and back) after close examination of the contact patterns for several coronal and dorsal consonants (i.e. /t/, /d/, /ʃ/, /s/, /l/ and /g/, /k/). Central bands are shown in figure 4. For subject 1, the mid-band's width is 40% of the maximum width in the front region and 65% in the back (wider to include hardly visible /k/s). For subject 2, the mid-band's width is 35% of the maximum width of the palate in the front region and 40% in the back region.

II.1.3 Building TTCD, TTCL, TBCD, TBCL

Once the front and back regions defined, and the criterion for presence of constriction set, the Tongue Tip Constriction Location (TTCL), Tongue Tip Constriction Degree (TTCD), Tongue Body Constriction Location (TBCL) and Tongue Body Constriction Degree (TBCD) could be computed. Their definitions (illustrated in figure 5) were based on the hypothesis that, for this particular

corpus, Tongue Tip movements correspond to contacts in the front region of the palate and Tongue Body movements to constrictions in the back region.

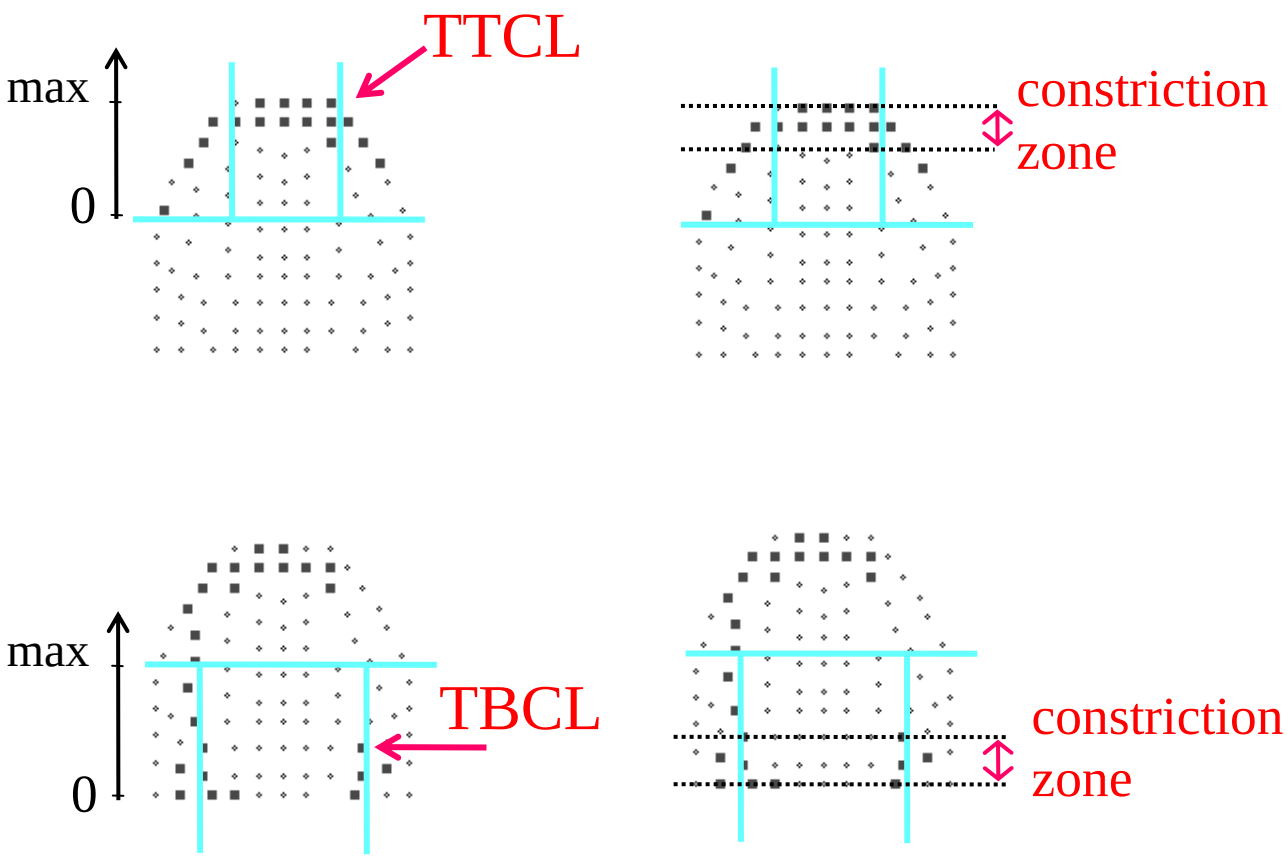

Figure 5. Definition of TTCL, TBCL and the front and back constriction zones for subject 1. TTCL and TBCL are minimum when most posterior and maximum when most anterior.

TTCL (respectively TBCL) was simply defined as the location of the most anterior of the contacted electrodes within the mid-band in the front (respectively back) region. In the absence of a constriction, TTCL (TBCL) was set to 0. TTCL and TBCL were defined so as to increase as the row is more anterior.

The anterior and the posterior constriction rows in each region were given by the most anterior and most posterior contacted electrodes within the mid-band. In each region, the *constriction zone* was then defined as the area delimited by the anterior and the posterior constriction rows (electrodes in the constriction zone therefore did not necessarily belong to the mid-band). TTCD (respectively TBCD) was then computed as the ratio of the number of contacted

electrodes in the front (respectively back) constriction zone, over the total number of electrodes in the front (respectively back) region.

This particular way of computing the degree of contact was chosen over other strategies (see e.g. Farnetani *et al.*, 1985; Gibbon, Hardcastle & Nicolaidis, 1993) because it seemed more appropriate in this particular corpus and because it allowed us to draw pseudo-gestural scores. Firstly, the involved segments (/k, l/) allowed for the definition of two contact regions only (front and back). A partition into three regions would allow us to distinguish between Tongue Tip, Tongue Blade and Tongue Body, but linguagraphic data would then be necessary in addition to palatographic data (see e.g. Dart, 1998). Secondly, the definition of a constriction zone was needed to exclude normal vocalic contacts (usually on the lateral sides of the pseudopalate) from the constrictions which are the scope of our investigation. A constriction was accounted for only when there was contact in the central part of the pseudopalate, but the side contacts were still counted in the computation of the constriction degree.

II.1.4 An example of a pseudo-gestural score derived from the EPG data

The above definitions allowed us to trace the Tongue Body and Tongue Tip constriction degrees and locations as a function of time, for the set of sentences uttered by the two speakers. These traces can be compared to the time functions of tract variables output by the Task Dynamics model, given an input gestural score. We will refer to this set of time traces as the pseudo-gestural score ("pseudo" because these derived traces are actual patterns of contact rather than a theoretical model of the intended gestures). Figure 6 gives an example of a pseudo-gestural score for the sentence 'say a clay of a Toledo again' uttered at the slow rate by subject 2. The TB 'gesture' for /k/ is marked by an increase in the TB Constriction Degree (increase in the number of contacted electrodes) and a hump on the TB Constriction Location (of a lesser amplitude than the hump for the preceding /e/ in 'say', which corresponds to a more anterior constriction). The TT 'gesture' for /l/ is marked by an increase in the Tongue Tip Constriction Degree (increase in contacts) and an increase in the TT Constriction Location (anteriority).

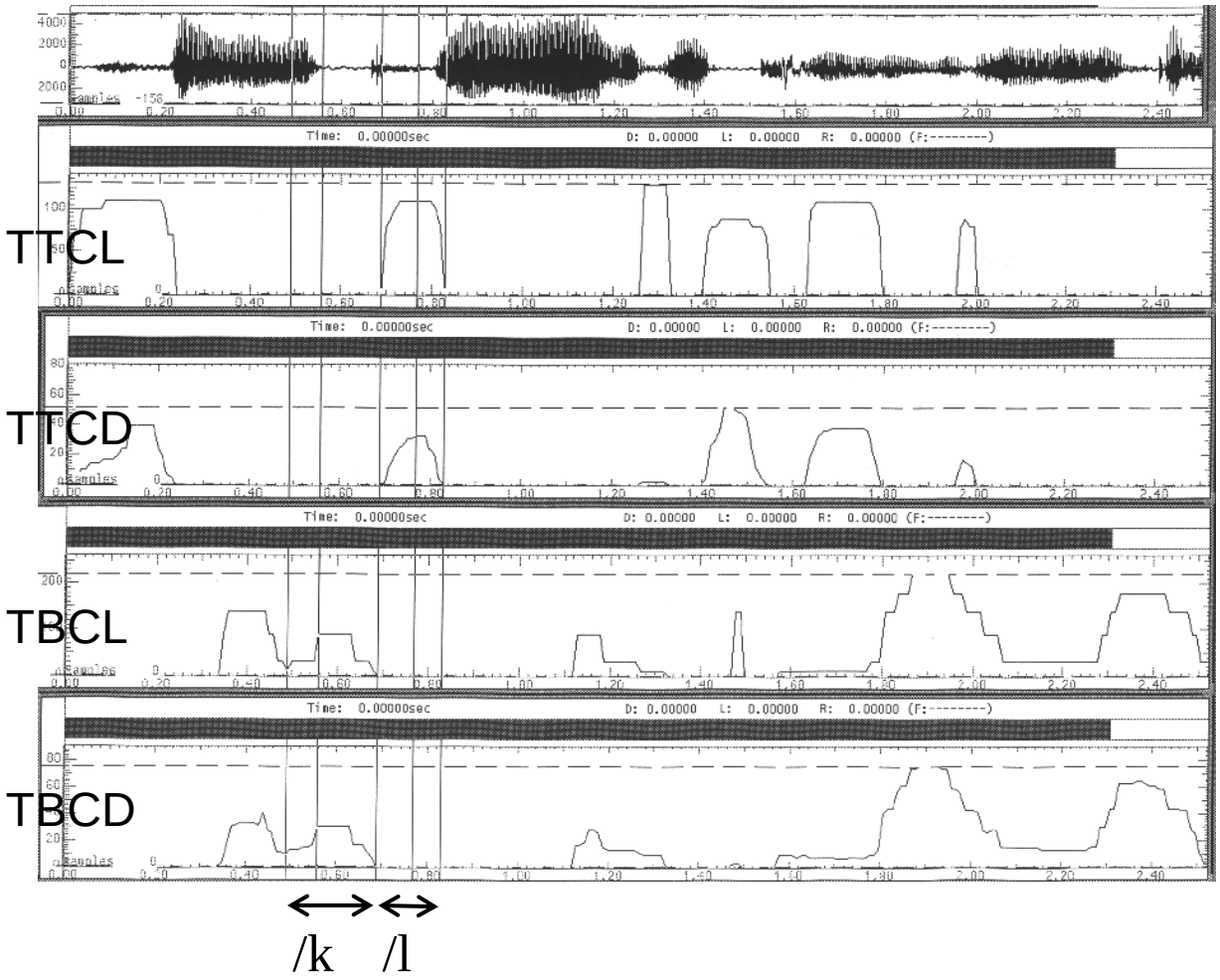

Figure 6. An example of a pseudo-gestural score derived from EPG data, for the sentence ‘Say a clay of a Toledo again’, uttered at a slow rate by subject 2. The vertical cursors delimit the beginning, peak and end of the TBCD and TTCD humps for /k/ and /l/. The peaks were taken at the earliest points in the TBCD and TTCD plateaus. In this example, as indicated by the arrows, the end of the TB ‘gesture’ for /k/ coincides with the beginning of the TT ‘gesture’ for /l/.

The primary interest of this paper being the temporal organization of articulatory gestures, their overlap or their spreading apart, we chose to discard the information given by TTCL and TBCL, which mainly concern location of constriction, and to focus only on TTCD and TBCD, which are related to the force of the linguopalatal contact. In this example, the end of the TBCD hump coincides with the beginning of the TTCD hump.

II.1.5 Comparison of the temporal organization of /kl/ and /skl/ vs. /kəl/

Times of beginning, maximum amplitude and end of the TBCD hump for /k/ and the TTCD hump for /l/ were automatically extracted using a min-max picking algorithm. Figures 7 and 8 show the time measurements for the production of ‘accolade’ in the slowest and fastest rate by subject 2. At the slowest rate (see figure 7), the beginning of the TT gesture for /l/ starts 120ms after the end of the TB gesture for /k/, whereas at the fastest rate (see figure 8), the beginning of the TT gesture starts 20ms before the end of the TB gesture.

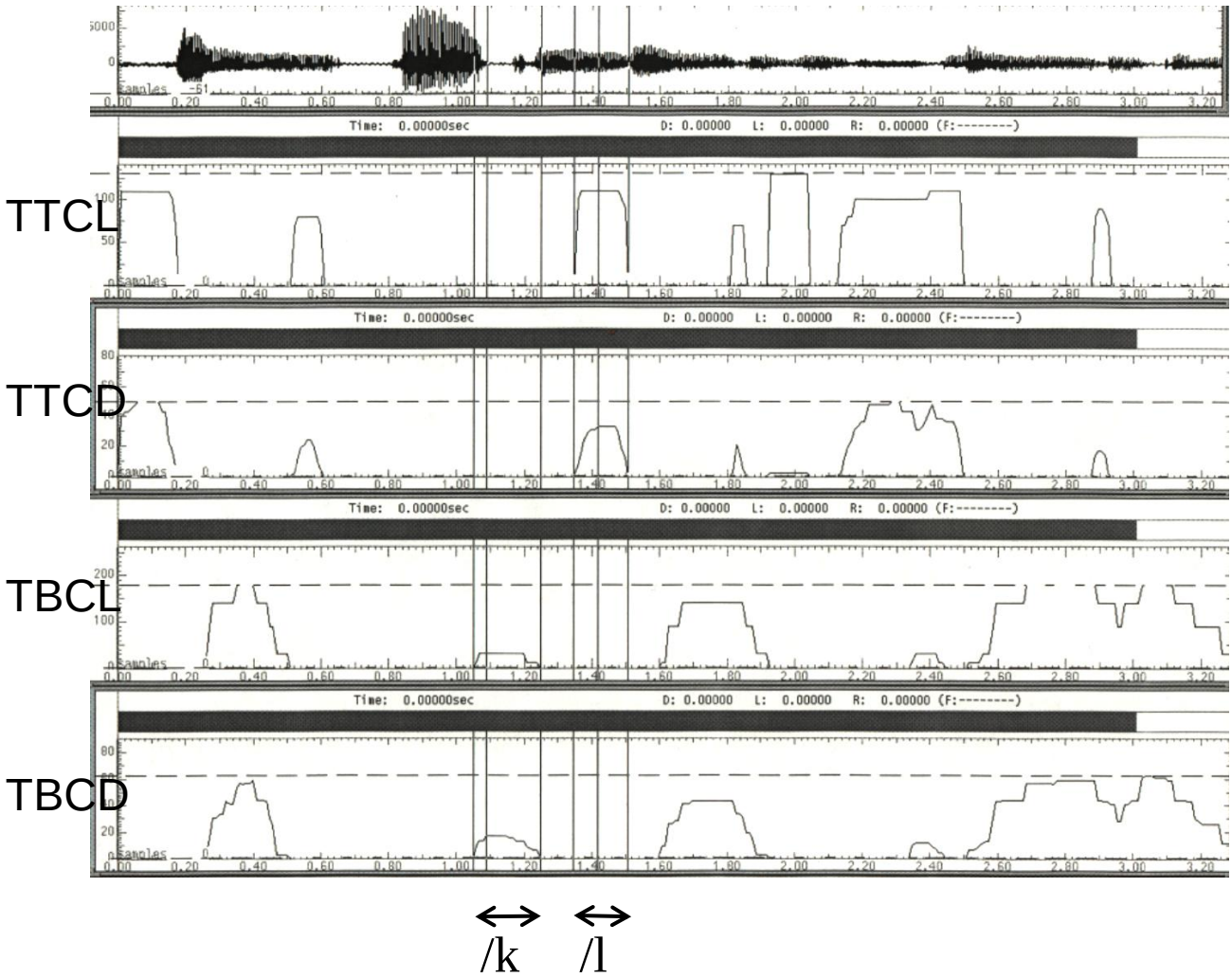

Figure 7. Pseudo-gestural score derived from EPG data, for the sentence ‘Say an accolade of a slayed again’, uttered at the slowest rate by subject 2. The vertical cursors delimit the beginning, peak and end of the TBCD and TTCD humps for /k/ and /l/. In this example, as indicated in the figure, the beginning of the TT ‘gesture’ for /l/ starts long after the end of the TB ‘gesture’ for /k/.

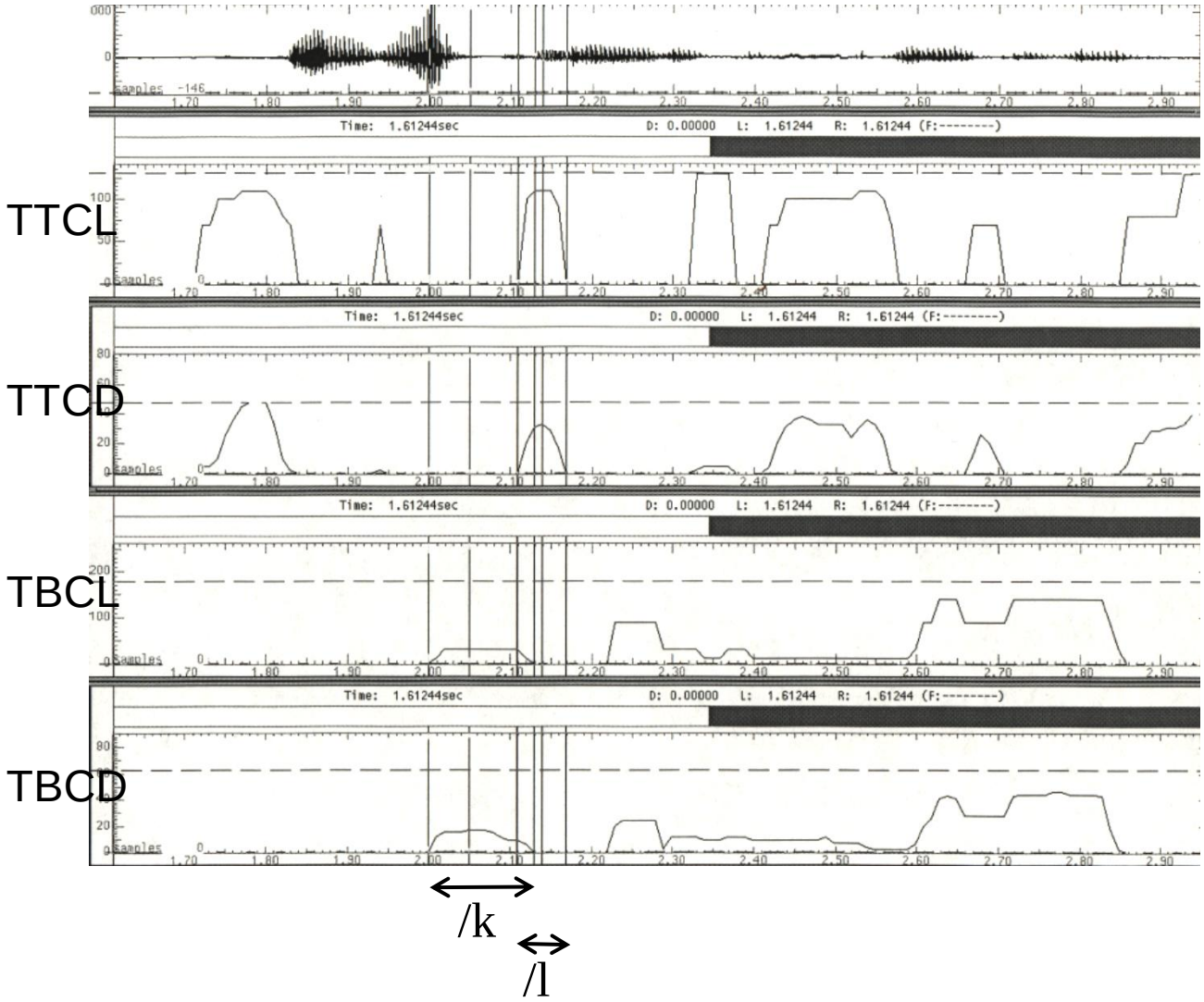

Figure 8. Pseudo-gestural score derived from EPG data, for the sentence ‘Say an accolade of a slayed again’, uttered at the fastest rate by subject 2. The vertical cursors delimit the beginning, peak and end of the TBCD and TTCD humps for /k/ and /l/. In this example, as indicated, the beginning of the TT ‘gesture’ for /l/ starts before the end of the TB ‘gesture’ for /k/.

Byrd & Tan (1996) evoke two articulatory strategies used in increasing speaking rate. The first is to decrease the duration of individual consonants. The second is to increase temporal coarticulation. The first strategy seems to be generally applied here, while the application of the second seems to depend on the phonological sequence involved.

As a measurement of speaking rate, we used the average duration of a phoneme, within the entire sentence 'Say a *target_word_1* of a *target_word_2* again'. The average phoneme duration was computed for each production by dividing the total duration of the sentence by the number of phonemes in that sentence. Tables 1 and 2 give the spans of speaking rates (in terms of minimum and maximum average phoneme durations) when rate goes from fastest to slowest for each of the speakers. Speaker 1 shows a smaller span than speaker 2: the mean variation in average phoneme duration between the fastest and the slowest rate is 29ms for speaker 1 and 117ms for speaker 2. It should be noted also that during the recordings, speaker 1 informed us that the task of controlling rate was difficult especially since the interval between utterances was quite long, in order to save each utterance to disk.

Gesture durations for TB in /k/ and TT in /l/ are given in Tables 3 and 4 for speaker 1 and 2, respectively. The rate categories correspond to the ranking of the actual rates and do not necessarily coincide with the rate instructions. When rate increases, the duration of the TB and TT gestures generally decreases. There seem to be a general rule where gesture duration decreases as a function of rate increase.

For speaker 1, there are 16 exceptions (marked with a *) to this general rule. They correspond to extreme speaking rates (fastest or slowest) in 11 cases. This could be related to the difficulty speaker 1 had of controlling rate, and especially of producing extreme rates. Furthermore, the gesture durations for the 16 exceptions are not extremely different from what they should be, had the general rule been observed. As mentioned before, the rates produced by this speaker were not markedly different from each other. So if the rule had been precisely followed, the slight increase in rate from one rate condition to the other would have been associated with only slight decreases in gesture durations. The 16 exceptions do not mean that the rule does not apply. When the expected slight gesture decrease was not observed, it could be that the other gestures in the sentence themselves followed the rule and even compensated for the 'deviant' gesture. When only the extreme fastest and slowest rates are considered, which correspond to two real categories, then there remains only two exceptions to the general rule.

For speaker 2, the decrease in TB/TT durations as a function of rate increase is more consistently observed. Among the four exceptions to this tendency, three belong to the fastest rate: TB in the accented production of /kl/ and in the unaccented production of /kəl/, and TT in the unaccented / kəl/. These exceptions could be due to the difficulty of producing a really fast rate.

Variations in the temporal coarticulation of the gestures were also studied. If the second strategy evoked by Byrd and Tan is actually applied, then there should be more overlap between TB and TT when rate increases. At very fast speaking rate then, the /kəl/ sequence – with TB and TT separated by schwa - should collapse into the /kl/ cluster – with TB and TT overlapping. Our goal being to study the behavior of the /kl/ cluster or the /kəl/ sequence when speaking rate is modified, results will only be presented for subject 2 who showed a wider range of speaking rates.

Table 1. Minimum and Maximum average phoneme durations (in ms) for subject 1.

	in /kl/		in /kəl/		in /skl/	
	acc.	unacc.	acc.	unacc.	acc.	unacc.
Min	73	78	65	60	65	71
Max	103	113	84	92	102	90

Table 2. Minimum and Maximum average phoneme durations (in ms) for subject 2.

	in /kl/		in /kʃl/		in /skl/	
	acc.	unacc.	acc.	unacc.	acc.	unacc.
Min	60	58	57	53	60	57
Max	190	224	168	206	108	153

Table 3. Gesture durations (in ms) for TB in /k/ and TT in /l/ for subject 1.

	TB /k/						TT /l/					
	in /kl/		in /kəl/		in /skl/		in /kl/		in /kəl/		in /skl/	
	acc.	unacc.	acc.	unacc.	acc.	unacc.	acc.	unacc.	acc.	unacc.	acc.	unacc.
Fastest	120	160*	50*	70	80	40	90	100*	90*	90*	50	80*
Fast	160*	110	40	70	80	50	120	90	80	70	90*	70
Normal	200*	120*	50	60*	90	70	160	90	120	90	80	100
Slow	150	110	50	30*	120	80	170	110	110*	120	80	120
Slowest	170	120			100*	80	160*	130			110	100*

Table 4. Gesture durations (in ms) for TB in /k/ and TT in /l/ for subject 2.

	TB /k/						TT /l/					
	in /kl/		in /kəl/		in /skl/		in /kl/		in /kəl/		in /skl/	
	acc.	unacc.	acc.	unacc.	acc.	unacc.	acc.	unacc.	acc.	unacc.	acc.	unacc.
Fastest	150*	120	130	190*	110	150	80	70	60	80*	70	40
Fast	110	120	130	130	150	170	80	70	70	70	70	50
Normal	160	170	140	140	150	140*	90	110	90	100	70	70
Slow	200	170	150	140	160	200	140	140	110	120	90	100
Slowest	200	170	180	180	160	210	520	210	160	200	120	140

For /kl/ and /kəl/ productions, timing data were computed relative to the /k/ gesture as follows:

Relative Time $_{\text{peak /k/}}$ = 100 (Time $_{\text{peak /k/}}$ - Time $_{\text{start /k/}}$) / Duration $_{\text{/k/}}$

Relative Time $_{\text{start /l/}}$ = 100 (Time $_{\text{start /l/}}$ - Time $_{\text{start /k/}}$) / Duration $_{\text{/k/}}$

Relative Time $_{\text{peak /l/}}$ = 100 (Time $_{\text{peak /l/}}$ - Time $_{\text{start /k/}}$) / Duration $_{\text{/k/}}$

Relative Time $_{\text{end /l/}}$ = 100 (Time $_{\text{end /l/}}$ - Time $_{\text{start /k/}}$) / Duration $_{\text{/k/}}$

Results for /kl/ are displayed in figure 9, where the relative times are plotted as a function of speaking rate, in the accented and unaccented conditions. A first result is that when speaking rate decreases, TT tends to start later relative to TB. At the fast and fastest rates, TT starts at 66 to 75% of TB (whatever the accent condition), that is the gestures strongly overlap. At the slowest rates, TT starts later (at 105% of TB at most). However a second result is that, as emphasized by the nearly constant size of the dotted ovals, whatever the accent condition and the speaking rate, the TT gesture starts before or near the end of the TB gesture. Even at the slow rate, when more time is devoted to the articulation, the TT and TB gestures of the /kl/ cluster remain cohesive. The cohesion is stronger in the accented condition (TT never starts after 100% of TB) than in the unaccented condition (105%).

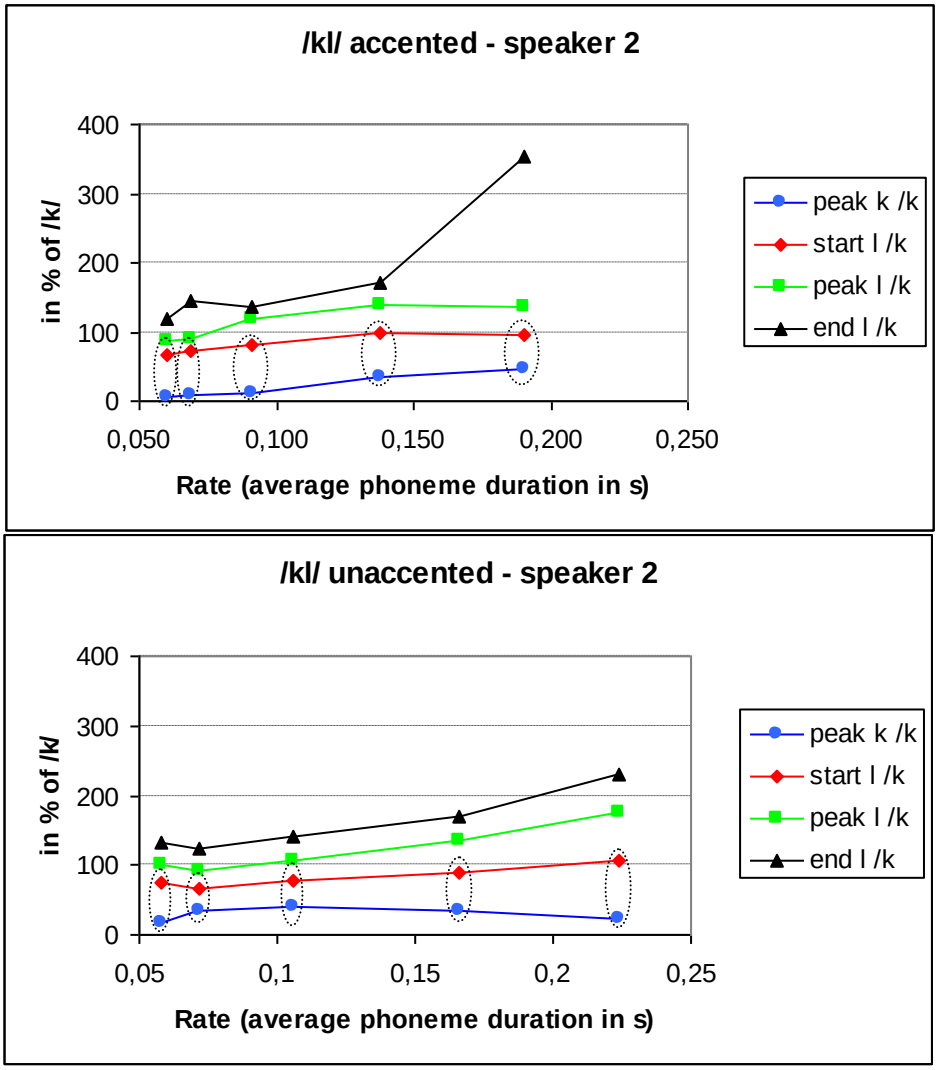

Figure 9. Time measurements for /kl/ in the accented (top) and unaccented (bottom) conditions, as a function of speaking rate, speaker 2.

In contrast, as illustrated by the variable-sized ovals in figure 10, /kəl/ shows more variation in temporal coordination, when rate is modified. At the fastest rates, TT starts before the end of TB (at 84% of TB in the accented condition and 89% in the unaccented condition), the overlap between TB and TT resembles that in /kl/. But at the slowest rates, TT starts long after TB is finished (at 166% of TB in the accented condition, 183% in the unaccented condition). As rate decreases, the overlap between the gestures gradually decreases and temporal coarticulation is reduced. The TT and TB gestures in the /kəl/ sequence are less cohesive and subject to more variation in gestural

overlap. Again, the cohesion is slightly stronger in the accented condition than in the unaccented condition.

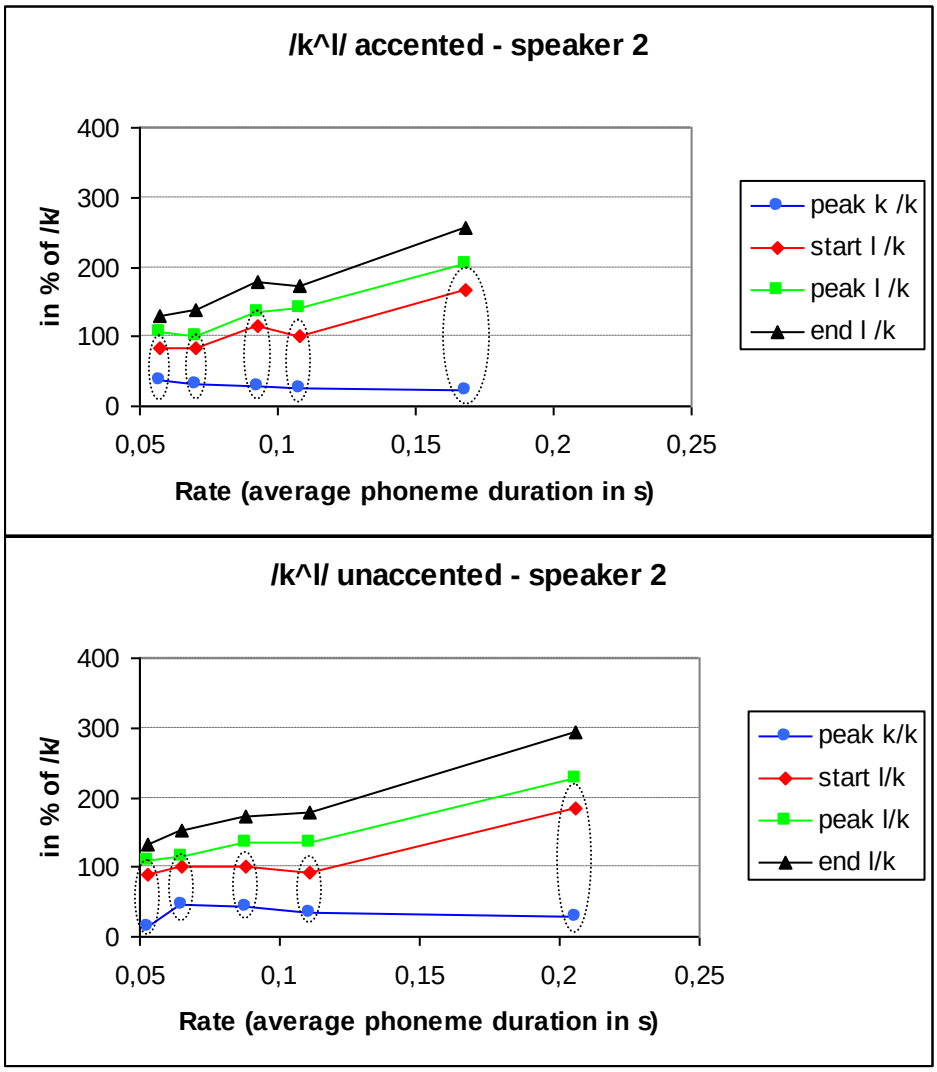

Figure 10. Time measurements for /kəl/ in the accented (top) and unaccented (bottom) conditions, as a function of speaking rate, speaker 2.

At extremely fast rates, when the overlap is equivalent to that in /kl/, Articulatory Phonology predicts the effective deletion of the schwa in /kəl/. Conversely, the schwa in “accolade” can be seen as an acoustic artifact of the separation of the two gestures. This proposal is further substantiated by the results for the /skl/ sequence.

Similar measures were taken relative to /s/ for /skl/ in the accented and unaccented conditions. Results are displayed in figure 11. As illustrated by the rather constant oval sizes, the TB gesture for /k/ starts close to the peak of the TT gesture for /s/, whatever the accent and the rate conditions. Over the entire range of rate and accent conditions, TB for /k/ starts from 21% to 66% of TT for /s/. The TT gesture for /s/ and the TB gesture for /k/ are cohesive. Similarly, whatever the accent and the rate conditions, the TT gesture for /l/ starts close to the peak and before the end of the TB gesture for /k/, just as in /kl/ (70 to 90% of TB, over the entire range of rate and accent conditions). The TB gesture for /k/ and the TT gesture for /l/ are cohesive. The TT gesture for /l/ also starts close to the end of the TT gesture for /s/, but always after. As with the /kl/ cluster and in contrast with /kəl/, a decrease in rate in the /skl/ cluster is matched with only a little decrease in temporal overlapping. Furthermore, in contrast with /kəl/, in the production of /skl/, a large amount of encroachment between the TT gesture for /s/ and the TT gesture for /l/ does not delete the TB gesture for /k/.

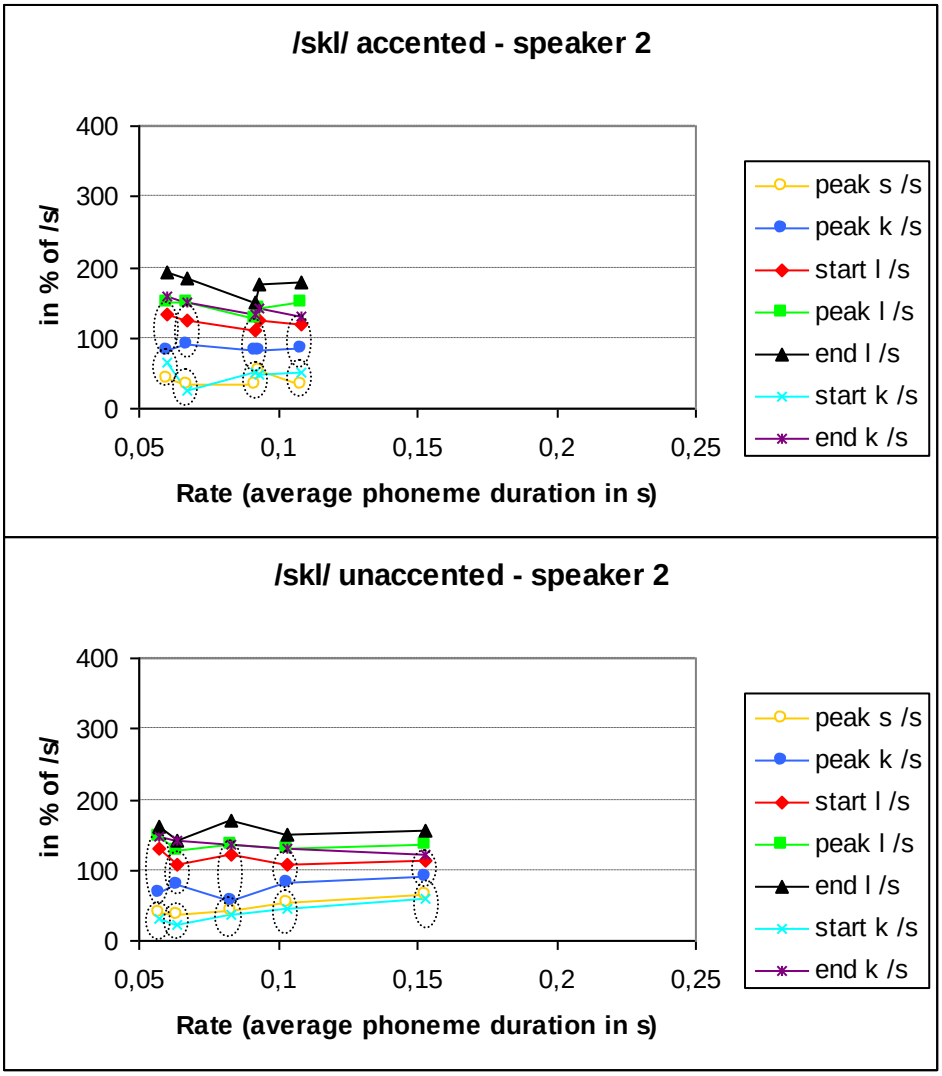

Figure 11. Time measurements for /skl/ in the accented (top) and unaccented (bottom) conditions, as a function of speaking rate, speaker 2.

This pleads for viewing the schwa in /kəl/ as not being represented in the gestural score as a separate specifically controlled constriction. Rather it is an acoustic artifact of the (typically longer) later phasing of the /l/ constriction gesture with respect to the /k/ constriction gesture. That is, the phonological category of schwa (in contrast with a CC cluster) is ontologically a derivative of the acoustic consequences of the phasing. There is no direct, invariant representation in the gestural score, where the schwa is represented "negatively" – by the absence of a tightly controlled phasing relationship between the /k/ and the /l/.

Consequently, a first conclusion from these data is that consonant gesture duration decreases with increase in rate. A second conclusion is that the decrease in gesture overlap triggered by a decrease in rate is stronger in the /kəl/ sequence than in the /kl/ and /skl/ clusters. The phasing of TB and TT seems to be more constrained in 'clay' and 'disclaimer' than in 'accolade'. This could be interpreted as a greater cohesion of the gestures in consonant clusters (/skl/ and /kl/) than in sequences of consonants such as /kəl/. The variation in temporal coarticulation as a function of speaking rate seems therefore dependent on the phonological sequence involved. The more cohesive /kl/ and /skl/ clusters are more resistant to the increase in available time; the gestures remain closely phased whatever the rate. The less cohesive /kəl/ sequence is more sensitive to variations in speaking rate; the gestures are more loosely phased one relative to the other.

These results are in accordance with recent proposals in Articulatory Phonology. In her study of articulatory timing in consonant sequences, Byrd (1996) showed that inter-gestural coordination is affected by gestural and prosodic factors and that phase relations are not equally constrained. She claimed that Articulatory Phonology must be revised to allow for linguistic and extra-linguistic influences on certain phasing relations. In order to account for Byrd's results, Browman & Goldstein (1998) have introduced the notion of "bonding strength" in Articulatory Phonology. They propose to associate every phase relation within a lexical unit with a "bonding strength" or a degree of cohesion, which defines the tightness of the configurational properties. Sources of gradual variation in gestural overlap (speaking rate, style, prosody) would affect a given pair of gestures in inverse proportion to their bonding strengths.

Using this notion then, /kl/ and /skl/ clusters being less affected by the changes in speaking rate than the /kəl/ sequences, they can be viewed as having a greater bonding strength. Conversely, the schwa in the /kəl/ sequences can be represented only negatively in the score, in terms of a lesser bonding strength between the neighboring consonant constrictions.

II.2. Deriving temporal events from the acoustic data

II.2.1 Seeded Temporal Decomposition

Jung *et al.* (1996) and Collins (1995) have examined the feasibility of generating gestural scores from articulatory data corresponding to CVC tokens, using a signal processing technique called "Temporal Decomposition". The second part of our study was an attempt to recover the temporal events suggested by the EPG analysis directly from acoustic data, using a similar approach.

Temporal Decomposition (TD), originally proposed by Atal (1983), attempts to capture the underlying temporal characteristics of a highly correlated, slowly changing, multi-channel signal. Using TD, the experimenter can model a set of speech parameters (the multi-channel data) by a set of overlapping target functions and corresponding target vectors. The target functions serve as basis functions for the speech parameters. They provide phasing information, from which a gestural score can be derived. The target vectors are the weights of the target functions necessary to reconstruct the original speech parameters.

Temporal Decomposition requires data which are multi-channel and highly correlated. The raw acoustic data do not satisfy either of these criteria and must therefore be transformed into an appropriate representation. In addition, the new representation should capture the temporal locations of the underlying gestures as accurately as possible, in order for generated gestural scores to be phonetically relevant.

Van Dijk-Kappers (1988) compared the results of Temporal Decomposition for several acoustic parameter sets. She examined the phonetic relevance of the derived basis functions, measured as the number of basis functions per phonemes and the speech quality after resynthesis. She found that the log-area parameters, derived from linear prediction analysis of the acoustic

waveform (and originally used in Atal's technique) were the most desirable, yielding a one-to-one correspondence between phonemes and basis functions 73% of the time (and one-to-one-or-two 98% of the time). A similar comparison carried out here (Collins, 1998) showed that log-area parameters appear to perform best for our purposes.

To improve the performance of temporal decomposition, which performs better with slow-changing data, the log-area parameters were low-pass filtered with a 30 point FIR filter. After several trials with different values, a cutoff frequency of 10Hz was chosen.

Temporal Decomposition was chosen over other segmentation tools because it allows overlapping segments, whereas traditional segmentation techniques yield hard boundaries between segments. However it features some disadvantages, one of them being that Temporal Decomposition is influenced by the window size and window overlap size chosen for the analysis (see Marcus and van Lieshout, 1984). Temporal Decomposition is strongly predisposed to find basis functions that are in roughly the same size and location as the specified windows. This bias means that standard TD cannot be expected to accurately locate speech events: speech segments vary in length, making it difficult to pick one TD window size which will accurately find all segments. To circumvent this difficulty, a variant of the original TD, dubbed 'Seeded Temporal Decomposition' (STD), was introduced, where the algorithm is seeded with "hard" initial segment boundaries. In implementation, STD differs from TD only at the point at which a basis function has been found and a new analysis window needs to be established. Temporal Decomposition places a fixed length analysis window (typically 150-250ms) after the current window, with a small overlap (typically 50ms). Seeded Temporal Decomposition uses the supplied hard segment boundaries to try to position the next analysis window in the location most likely to yield a correctly placed basis function. If the current basis function approximately spans the original location of the current analysis window, STD places the next analysis window with its right edge at the next specified hard boundary. This is the case where one basis function was found which roughly corresponds in location and duration with a specified phoneme. If the current basis function is sufficiently smaller, the right edge of the next analysis is the same as the original right edge of the current analysis window. This case implies that a specified phoneme is composed of more than one basis function, as in the case of burst and release for a stop consonant. If the current basis function is large enough that it spans multiple hard boundaries, the right edge of the next analysis window is set to the first hard boundary not already covered. This case results in a basis function which covers multiple

phonemes, possibly indicating a phoneme sufficiently covered by its neighbor so as to be unrecoverable as a separate entity. In all cases, the left edge of the next analysis window is placed overlapping the current window by a small amount (typically 50ms).

In the present study, the acoustic signal was therefore first hand-labeled to provide phoneme boundaries for the STD algorithm to dynamically adjust the initial window location for each basis function.

II.2.2 Comparison of EPG-based gestural scores and acoustic-derived temporal events for /kəl/

Results of a STD applied to log-area parameters low-pass filtered to 10Hz are presented in figure 12 for ‘an accolade of a slayed again’ uttered at the fast rate. Each of the humps in the third panel of the figure correspond to a basis function. EPG-derived gestural scores are included for comparison. As shown by the arrows in the figure, there is a one-to-one correspondence between the EPG-derived TB gesture for /k/ and TT gesture for /l/, and the acoustic-derived basis functions. It seems therefore that the basis function labeled as 1 in figure 12 corresponds to a gesture on the TB tier and that the basis function labeled as 2 corresponds to a gesture on the TT tier.

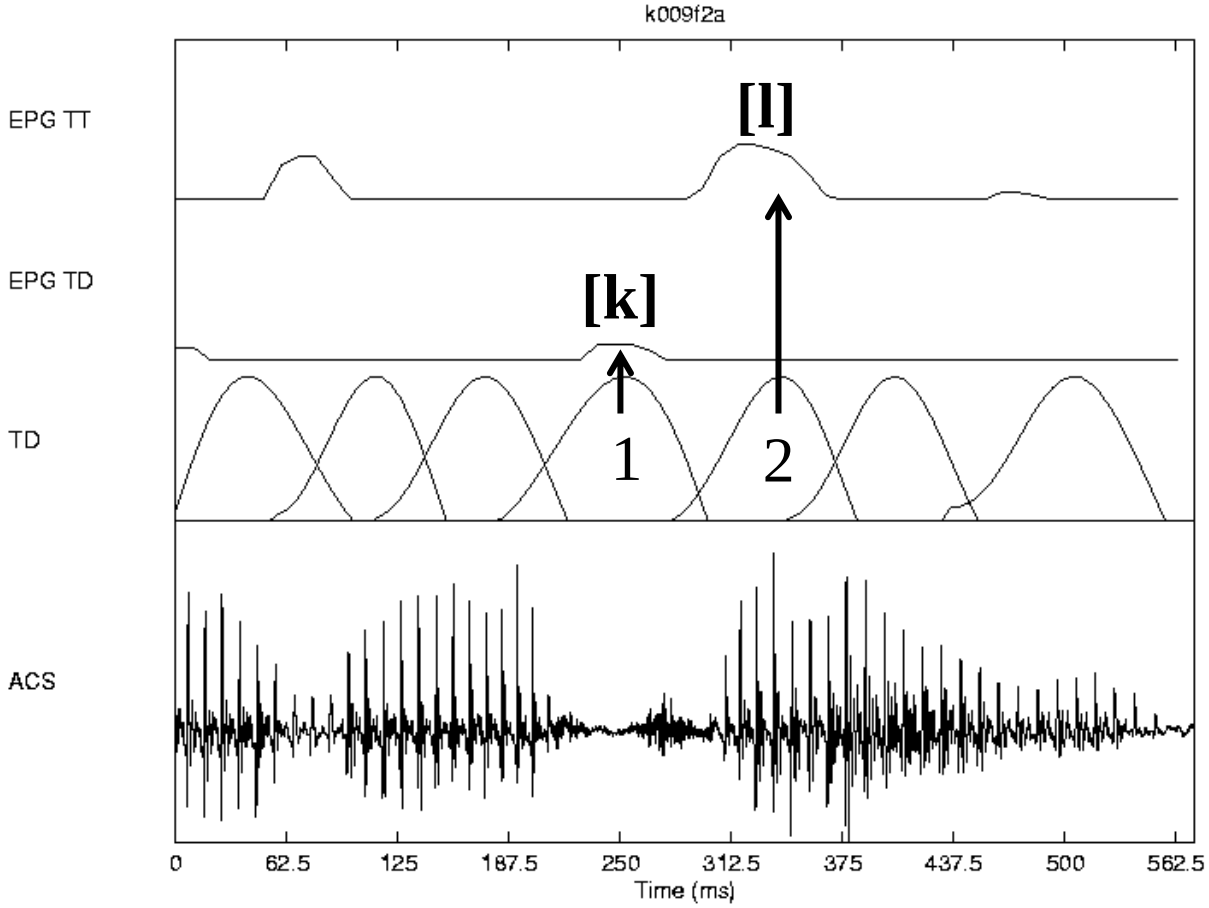

Figure 12. Deriving gestural scores for Say an accolade of slayed again', at fast rate, speaker 1. The top two panels show Pseudo Gestural Scores derived form EPG data. The third panel shows results of Temporal Decomposition of the signal into successive overlapping basis functions: the functions shown are the STD basis functions which provide phasing information about temporal events in the signal. The arrows indicate the basis functions corresponding to /k/ and /l/ which are labeled respectively 1 and 2. The fourth panel displays the original acoustic signal.

Figure 13 displays the results for the same utterance at the normal rate. The EPG-derived TB gesture for /k/ and TT gesture for /l/ have their corresponding basis functions (labeled 1 and 3), but an additional basis function (labeled 2) appears in between that for /k/ and /l/, corresponding to schwa. It also roughly corresponds to the gap in the EPG-derived scores between TB and TT. Our interpretation is that basis functions reflect distinct effective constriction degrees. Here the additional basis function corresponds to the distinct release for the /k/, that is the distinct effective constriction degree of open vocal tract in the gap.

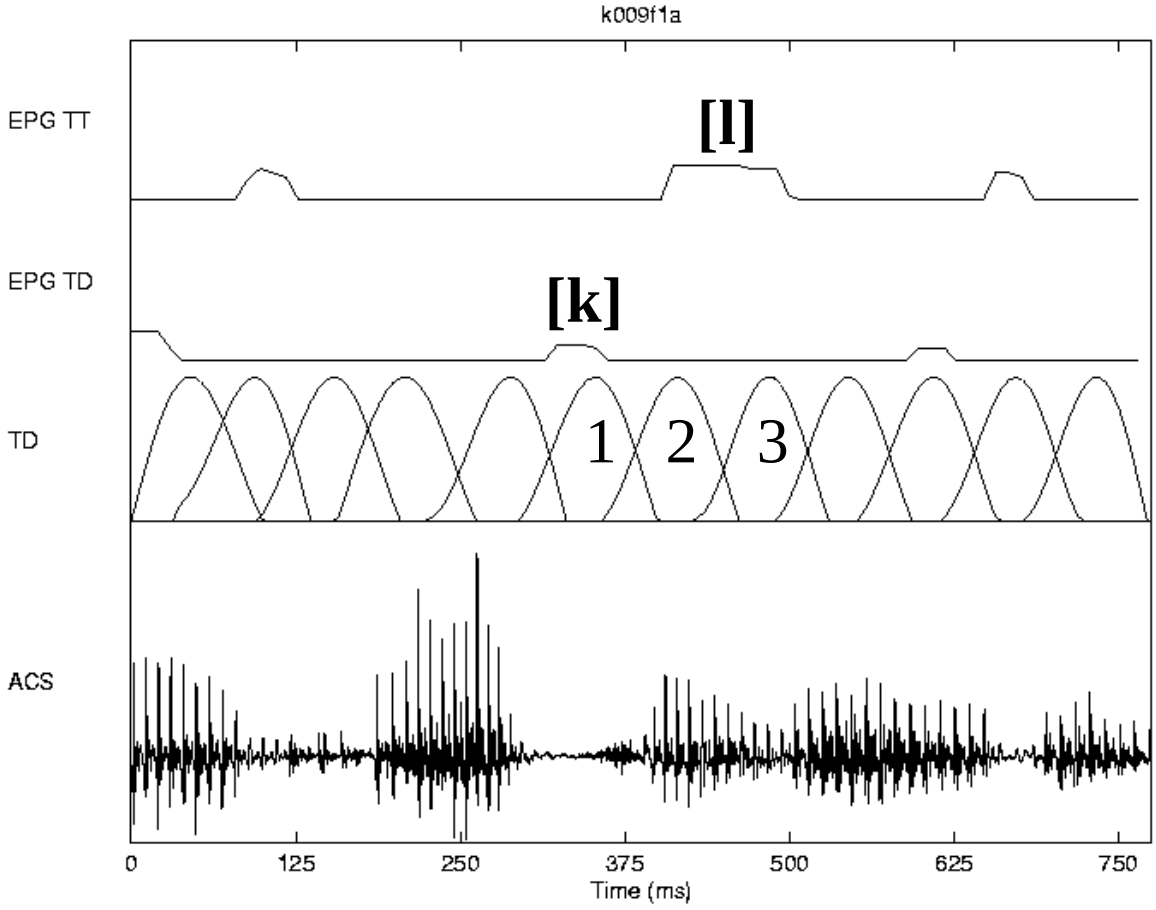

Figure 13. Deriving gestural scores for ‘Say an accolade of slayed again’, at normal rate, speaker 1. Top two panels: Pseudo Gestural Scores derived form EPG data. Third panel: Temporal Decomposition. The basis functions corresponding to /k/, schwa and /l/ are labeled as 1, 2, 3. Fourth panel: original acoustic signal.

Figure 14 displays the results of the STD at the slowest rate. As at the normal rate, a basis function (labeled 2) can be observed for schwa. But the STD also reveals two basis functions for /l/ (labeled 3 and 4), instead of one.

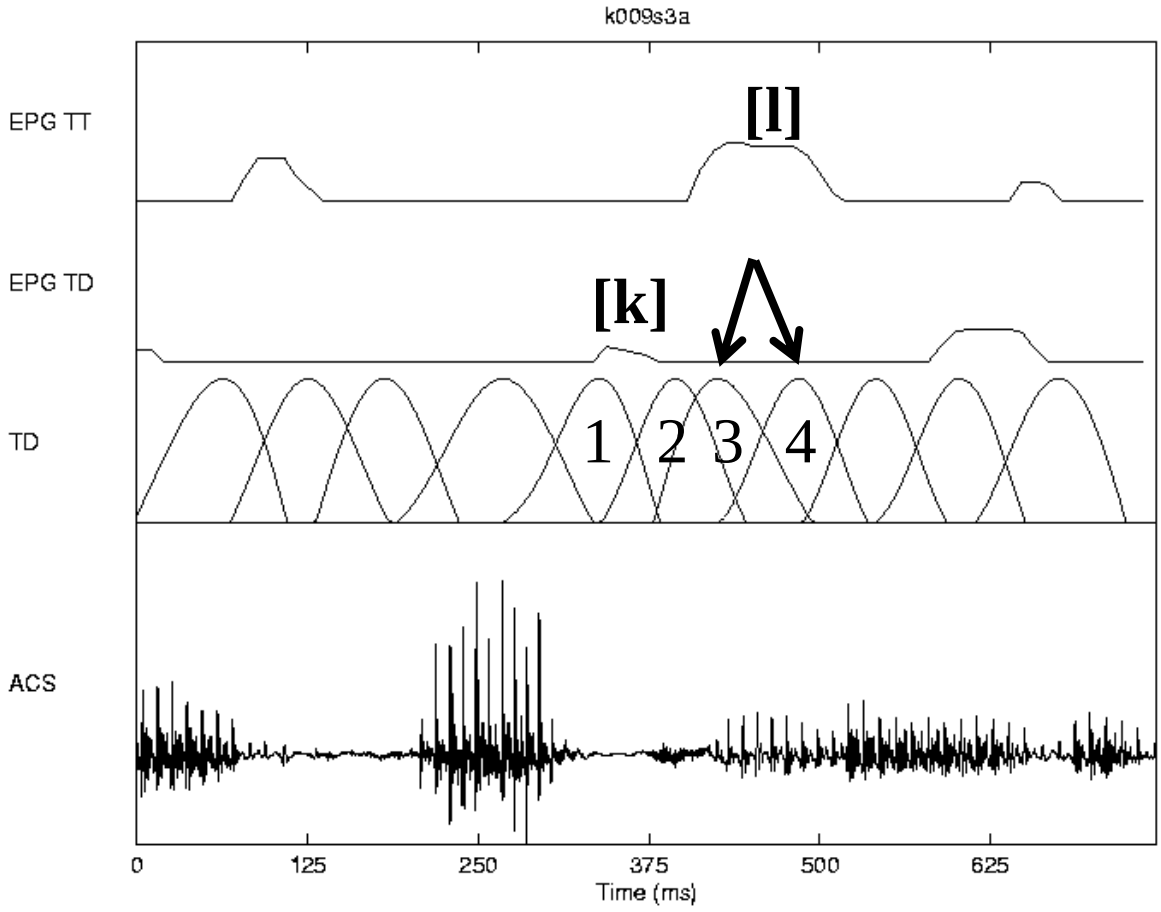

Figure 14. Deriving gestural scores for ‘Say an accolade of slayed again’, at the slowest rate, speaker 1. Top two panels: Pseudo Gestural Scores derived form EPG data. Third panel: Temporal Decomposition. The basis functions corresponding to /k/, schwa are labeled as 1 and 2. The arrows indicate the two basis functions for the /l/, which are labeled as 3 and 4. Fourth panel: original acoustic signal.

Looking at the EPG data suggests an explanation for this additional basis function (see figure 15). An intermediate state is observed between the /l/ and the following /e/, where the alveolar contacts for /l/ occur simultaneously with the raising of the sides of the tongue corresponding to the preparation of /e/. The second basis function observed for /l/ could correspond to this transient, which is characterized by a smaller diameter lateral opening – i.e. a smaller effective constriction degree for the channel that is coupled in parallel to the (closed) central channel (see Browman & Goldstein, 1989). The smaller effective constriction degree could be mirrored in the acoustic signal by an additional basis function.

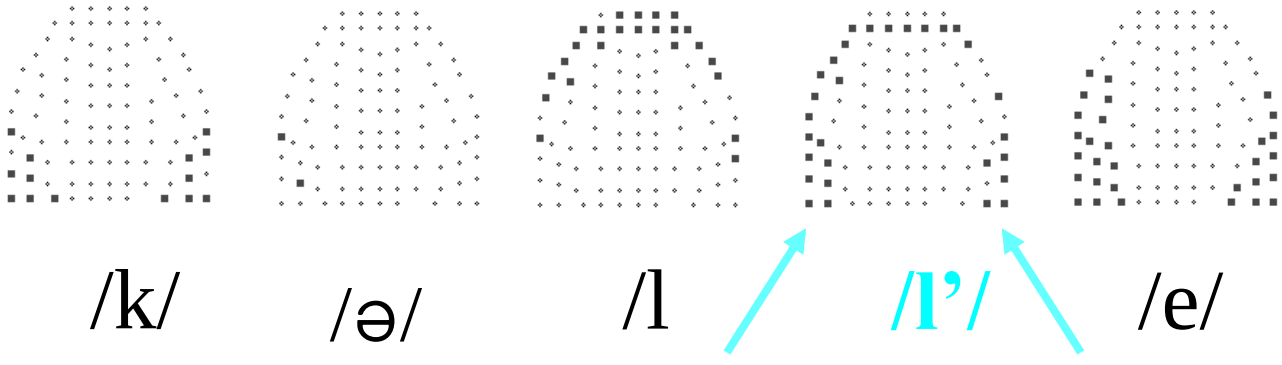

Figure 15. Palatal contacts for ‘accolade’ at the slowest rate. The intermediate state between /l/ and /e/ with simultaneous coronal and lateral contacts is reflected by an additional basis function in the STD.

III. Conclusion

The pseudo-gestural scores derived from the EPG data revealed differences in the coordination of the /kl/ and /skl/ clusters and the /kəl/ sequence. The coordination of the Tongue Body and Tongue Tip gestures in the /kl/ and /skl/ clusters appeared to be more resistant to change under variations in speaking rate than the coordination of the same gestures in the /kəl/ sequence. This greater cohesion of the /kl, skl/ clusters over the /kəl/ sequence could be interpreted as a larger bonding strength, a notion recently proposed in Articulatory Phonology. The greater bonding strength in the clusters as compared to the /kəl/ sequence means that, while the phasing of the gestures in /kəl/ can approach and even become identical to the phasing in the /kl/ cluster at the fastest rates, the phasing of the /kl/ cluster is not likely to approach that of the /kəl/ sequence at the slowest rates. That is, schwa deletion should be more common than schwa epenthesis in English (although see Jannedy, 1994, for German).

The acoustic-derived temporal events captured additional information on the organization of these clusters or sequences of consonants. The basis functions generated by STD can be seen as representing the effective constriction degree in a tube model of the vocal tract. At the fast rate, the TB gesture for /k/, which corresponds to a given constriction degree in the rear of the vocal tract, is closely followed by the TT gesture for /l/, another effective constriction degree at the front of the vocal tract. This sequence of two distinct effective constriction degrees is associated with two distinct basis functions in the STD. At the normal rate, when the two gestures are separated by a gap, a third effective constriction degree, corresponding to an open vocal tract, is associated with a third basis function. But the basis functions can also reflect the effective constriction degree in the side channels of the vocal tract. At the slowest rate, when raising of the sides of the tongue in preparation for /e/ occurs while alveolar contact is still observed, a transient effective constriction degree for the lateral channels is captured by the STD. STD can therefore account for the presence

of a transient in /l - e/, which is not codified by the phonology of English and which is not predicted by the pseudo-gestural scores derived from the EPG data.

These results show that the characterization of gestures in terms of Tongue Tip and Tongue Body is not sufficient to account for the phenomena that occur in a sequence such as /kəl/. In the TT/TB account, the size of the lateral branching tube is not accounted for and the appearance of the transient observed in 'accolade' at the slowest rate is not predicted. This suggests, as was also claimed by Stone *et al.* (1992), that a mid-sagittal description is inappropriate for the lateral /l/. The coronal shape of the tongue surface and its contribution to the geometry of the tube model should be taken into account.

As a conclusion, this work suggests that the temporal coordination of Tongue Body and Tongue Tip gestures is different in consonant clusters and in similar sequences of consonants separated by a schwa: the recorded EPG data showed that the coordination was tighter in /kl/ and /skl/ consonant clusters than it was in the /kəl/ sequence. Furthermore, the parallel investigation of the recorded acoustic signal revealed additional information about gestural coordination. Because Temporal Decomposition of the acoustic signal can reflect the effective constriction degree in central and side channels, it can account for the presence of transients in /le/ which are not codified by the phonology of English and were not predicted by the EPG-derived pseudo-gestural scores.

Acknowledgements

This work was supported by a grant from the *Fondation Fyssen* to the first author. The authors wish to thank Keith Johnson for his participation in the experiment, for his help in devising the corpus, for his guidance and many thoughtful suggestions in dealing with EPG data and for his comments on a previous version of this paper.

References

Atal B. S. (1983). Efficient coding of LPC parameters by temporal decomposition. *Proc. Int. Conf. Acoust., Speech, Signal Processing,* Boston, MA. 81-84.

Browman C.P. & Goldstein L. (1989). Articulatory gestures as phonological units. *Phonology*, **6**, 201-251.

Browman C.P. & Goldstein L. (1990). Gestural specification using dynamically-defined articulatory structures. *J Phonetics* **18,** 299-320.

Browman C. P. & Goldstein L. (1995). Gestural syllable position effects in American English. *Producing speech: contemporary issues. For K. S. Harris.* F. Bell-Berti & L.J. Raphael (Eds.). Woodbury, NY: AIP Press. 19-34.

Browman C. P. & Goldstein L. (1998). Modeling syllable structure as constraints on gestural coordination. *Segments and syllable workshop, the Ohio State University, Sept. 1998.*

Byrd D. (1996). Influences on articulatory timing in consonant sequences. *J. Phonetics,* **24,** 209-244.

Byrd D. & Tan C. C. (1996). Saying consonant clusters quickly. *J. Phonetics,* **24,** 263-282.

Clements G. N. (1985). The geometry of phonological features. *Phonology Yearbook,* **2,** 225-252.

Collins M.J. (1995). Generating Gestural Scores from articulatory data using Temporal Decomposition. *Master's thesis.* The Ohio State University.

Collins M.J. (1998). Generating Gestural Scores from acoustic data using Temporal Decomposition. *Doctoral dissertation.* The Ohio State University.

Dart S. N. (1998). Comparing French and English coronal consonant articulation. *J. Phonetics,* **26,** 71-94.

van Dijk-Kappers A. M. L. (1988). Comparison of parameter sets for Temporal Decomposition. *IPO Annual Progress Report,* **23,** 24-33, IPO, Eindhoven.

Farnetani E., Vagges K. & Magno-Caldognetto E. (1985). Coarticulation in Italian /VtV/ sequences: a palatographic study. *Phonetica,* **42,** 78-99.

Gibbon F., Hardcastle W. & Nicolaidis K. (1993). Temporal and spatial aspects of lingual coarticulation in /kl/ sequences: a cross linguistic investigation. *Language and Speech,* **36 (2,3),** 261-277.

Hardcastle W.J., Gibbon F. & Nicolaidis K. (1991). EPG data reduction and their implications for studies of lingual coarticulation. *J. Phonetics,* **19,** 251-266.

Jannedy S. (1994). Rate effects on German unstressed syllables. *Ohio State University Working Papers in Linguistics,* **44,** 105-124.

Jung T.-P., Krishnamurthy A. K., Ahalt S. C., Beckman M. E. & Lee S. H. (1996). Deriving gestural scores from articulator-movement records using weighted Temporal Decomposition. *IEEE Transactions on Speech and Audio Processing.* **4 (1),** 2-18.

Marcus S. M. & van Lieshout R. (1984). Temporal Decomposition. *IPO Annual Progress Report,* **19,** IPO, Eindhoven. 25-31.

Recasens D. (1984). V-to-C coarticulation in Catalan VCV sequences: an articulatory and acoustical study. *J. Phonetics,* **12,** 61-73.

Recasens D. (1990). The articulatory characteristics of palatal consonants. *J. Phonetics,* **18,** 267-280.

Recasens D. (1991). An electropalatographic and acoustic study of consonant-to-vowel coarticulation. *J. Phonetics,* **19,** 177-192.

Sproat R. & Fujimura O. (1993). Allophonic variation in English /l/ and its implication for phonetic implementation. *J. Phonetics,* **21,** 291-311.

Stone M., Faber A., Raphael L. J. & Shawker T.H. (1992). Cross-sectional tongue shape and linguopalatal contact patterns in [s], [sh], and [l]. *J. Phonetics,* **20,** 253-270.