Acoustical segmental duration or articulatory inter-targets as an indicator of speaker specific kinematic properties

Martine TODA and Shinji MAEDA (LTCI - CNRS)

Introduction

The segmental duration is an easily measurable speech parameter on the acoustic signal. Recent studies have shown that segmental duration is speaker specific (Pfitzinger, 2002), and it can be used for the automatic speaker recognition exploiting this speaker specificity (Ferrer et al., 2003).

In this report, we discuss the interest in the temporal aspects of speech production in the context of the acoustic-to-articulatory inverse. In fact, its characteristic of speaker specificity suggests its possible link with the kinematic and underlying bio-mechanical properties specific to individual speakers. Everything else being equal, a longer segmental duration can be regarded as the manifestation of either a longer path length between two successive articulatory targets or a slower articulator's speed suggesting a weaker stiffness of the related muscles in bio-mechanical terms. Turning our attention to the inverse problem, the derived kinematic properties may allow us to adapt the control sequence to a specific speaker in connection with a generic articulatory model already adapted to the morphology of that speaker. Moreover, the acoustically derived bio-mechanic properties can provide a reasonable constraint on the possible articulatory trajectories in the speech inversion.

In this sturdy, we shall focus our attention to unvoiced sibilant fricatives, /s/ and /ʃ/, because their segmental duration can be automatically and reliably measured on a large speech database. Actually we have formulated a robust segmentation method with high accuracy.

1. Inter-speaker variations in the segmental duration of /s/ and /ʃ/

In this section, we illustrate how is the segmental duration of sibilants, /s/ and /ʃ/, speaker dependent. Figure 1 plots the data borrowed from published works (Fuchs and Toda, 2010), indicating averaged segment duration of /s/ in the abscissa and of /ʃ/ in the ordinate for each of eight German and eight English speakers. The duration of each of these paired sibilants was measured at the initial position in a nonsense word, [CaCa], embedded in a carrier sentence. Each speaker repeated 30 utterances for each sibilant.

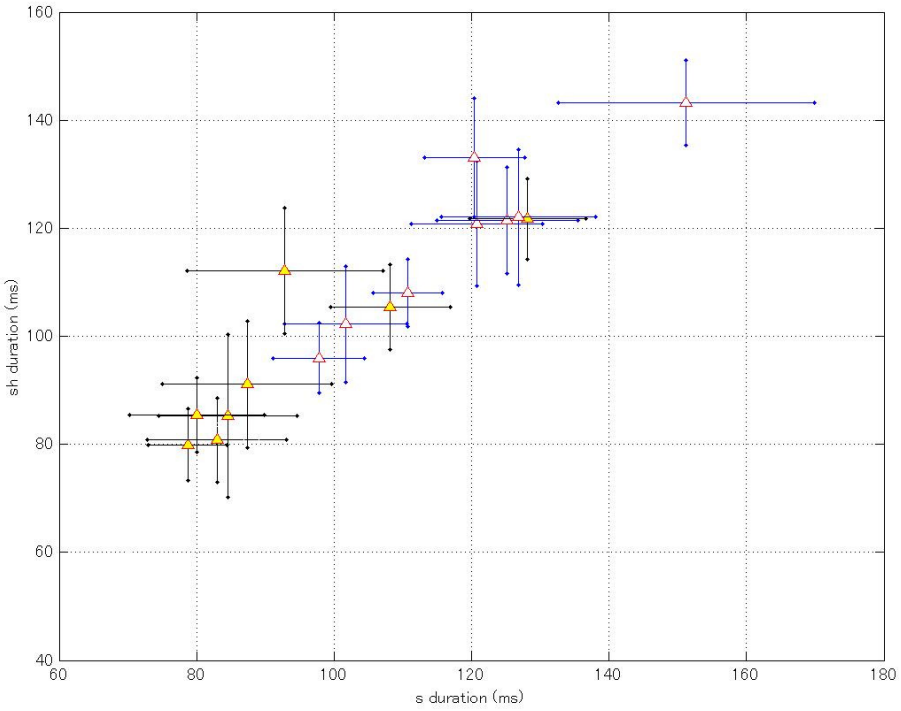

Figure 1. Averaged duration (n=30) of sibilant /s/ and of /ʃ/ at the initial position in a nonsense word [CaCa] embedded in a carrier sentence "Ich habe [CaCa] gesagt" for eight German speakers (indicated by the yellow triangles) and "I say [CaCa] again" for eight English speakers (by white triangles): The length of arms of the crosses indicates the unit standard deviation.

In the case of the controlled context in this experiment, the dispersion of data points in Figure 1 seems to indicate the following two things: a) the acoustical segment duration of /s/ and /ʃ/ systematically co-vary in function of individual speakers, suggesting common underlying mechanisms, and b) the range of the speaker-dependent variations is quite large for both sibilants by factor 2. These two indications encourage us for further investigations.

2. Speaker-dependent duration of /ʃ/ and possible contributions of the kinematic properties specific to speakers

Now, let us show some evidence for possible link between the acoustical segment duration of /ʃ/ and kinematic properties with a help of articulatory movements data acquired using EMA (Electro-Magnetic Articulography).

2.1. Acoustical segment duration

Figure 2 illustrates the segmental duration of each of three successive phonemes, /aʃə/, uttered by four German speakers (see the caption for more details). Although, these four speakers are different from the previous eight German speakers, the large range of the dispersion for the sibilant /ʃ/ in Figure 2, again by factor of two, is comparable to and corroborates the previous data shown in Figure 1, as far as /ʃ/ is

concerned. It may be noted that speakers F2 and M2 exhibit relatively short /ʃ/ durations, whereas F1 and M1 have long durations, in terms of both absolute in (a) and normalized values in (b). This speaker grouping might be objected, since the duration of F1 (long) and that of M2 (short) don't differ much. In fact however, this grouping will be justified in the followings.

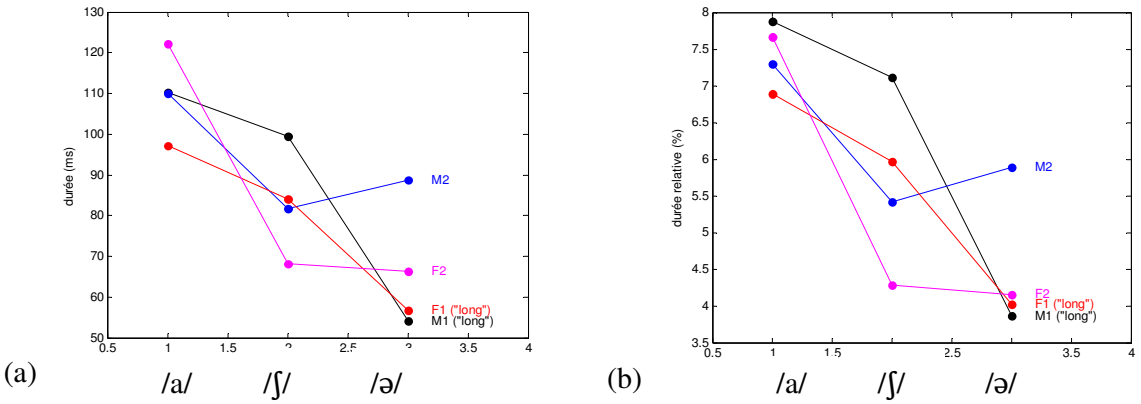

*Figure 2. Acoustical segment duration of /a/, /ʃ/, and /ə/ uttered within the sentence "Ich wasche ([va**ʃ**__**e**__])* Hagi / Hagu / Haku *im Garten." (I wash Hagi / Hagu / Haku in the garden.) by four German speakers, two females (F1 and F2) and two males (M1 and M2): The segmentation was done manually and provided to us by Miss. Weirich at ZAS, Berlin. (a) Absolute duration in ms, and (b) Normalized duration in % calculated by dividing the measured duration of each phoneme by the total duration of the sentence.*

It is also interesting to note that the duration of two vowels adjacent to the sibilant /ʃ/ tends to compensate its large variation so that the total duration of the word [vaʃə] is relatively constant across four speakers. Here, we interpret the variations of segmental duration along /aʃə/ seen in Figure 2 as the indication of temporal compensation that give rise to a relatively constant total duration despite of large individual variations, especially that of /ʃ/. Specifically, for the group with long /ʃ/, including F1, the duration of final vowel is considerably shortened relative to the short-/ʃ/ group M2 and F2. Conversely, for the short-/ʃ/ group, the final vowel duration, especially for the border speaker M2, is lengthened. We shall suggest in the next section that this group-dependent difference in the temporal compensation appears to be related to the difference in the articulatory organization.

2.2 Articulatory targets and the contribution of the jaw *vs.* the tongue

Successive articulatory targets along the target word are identified on the trajectory of each EMA's position sensor coil with the help of a method proposed by Ananthakrishnan and Engwall (2008). This method seeks in the trajectory's turning points where the movement speed becomes slow, as shown in Figure 3 for the EMA's sensor coil glued on the tongue tip. We define those identified points as articulatory targets, which are close to "via-points" proposed by Vatikiotis-Bateson *et al.* (1994) for the purpose of a data reduction on movement curves by keeping only pertinent points.

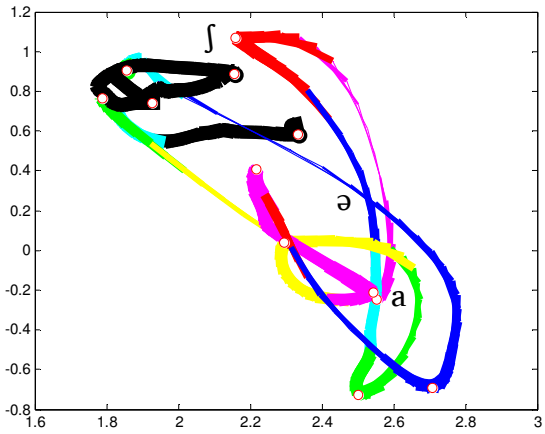

Figure 3. Trajectory of the EMA's sensor coil glued on the tongue tip region (Speaker M1) along the utterance "Ich wasche Haku im garten": ***[# ɪç vaʃə hakʊ ɪm gartən #].*** *The thickness of the trajectory path indicate the speed; the thicker the slower. Phonemic acoustical segments are coded in colors and the white circles indicate the identified articulatory targets (for the whole sentence).*

Figure 4 illustrates the anterior part of the vocal-tract profiles defined by EMA's position sensors. The markers of the three sensor coils on the tongue surface are connected by lines to show the gross tongue position and shapes. The profiles corresponding to target points within the initial vowel /a/ are indicated by the blue markers and lines, within the sibilant /ʃ/ by the reds, and within the final vowel /ə/ by the blacks.

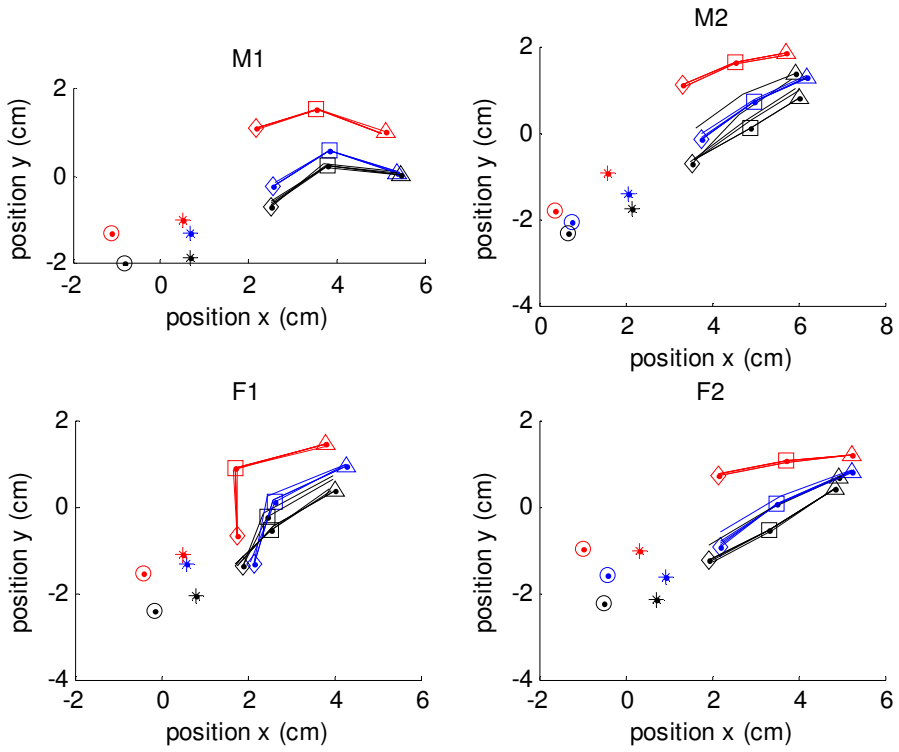

Figure 4. Positions of EMA's sensor coils at target points, indicated by the markers, from left to right, circles for the lower lip, stars for the lower jaw (more precisely lower incisor), diamonds for tongue tip, squares for the tongue-body, and triangles for the tongue-dorsum: In each of these four speakers' data, the blue color indicates that the targets are within /a/-, the red within /ʃ/-, and the black within /ə/-segment along the sequence [aʃə].

Generally in speech production, a short segmental duration implies an articulatory reduction due to the lack of time for an articulator to reach its target position. Despite of the large variations of segmental duration of those two vowels across speakers, articulatory reduction is not observed in the data shown in Figure 4 concerned with our four speakers. The tongue position, which is the most relevant articulatory parameter for speech production, is expected to show some reduction due to the shortened duration of the two vowels as far as the "long-/ʃ/" speakers, F1 and M1, are concerned. This is not the case, because their tongue positions are not particularly higher than those of the "short-/ʃ/" speakers, F2 and M2. We see here no indication that could distinguish the two groups of speakers in articulatory terms. Because of the requirement for the formation of narrow constriction during /ʃ/, the tongue profiles takes highest positions apart from those of vowels for all the four speakers. The height is similar for the two vowels, but the initial vowel /a/ is slightly but systematically higher than the final vowel /ə/, with perhaps exception of F1. In this regard, we can state that the tongue position is not distinctive between these two speaker groups.

In detail however, Figure 4 indicates the difference in the lower jaw position across these two speaker groups. Notice that both long-/ʃ/ speakers, F1 and M1, exhibit that the jaw (indicated by star markers) in the initial vowel /a/ (in blue) and that in the

sibilant /ʃ/ (in red) are very close to each other. Note moreover that the jaw position at the target points of /ʃ/ is relatively constant across speakers as claimed in Lee, Beckman, and Jackson, 1994) and also supported in Section 3.3 in this report. This relatively small distance in the jaw target positions from the /a/ to /ʃ/ suggests the anticipation of the high jaw position for the /ʃ/ during the preceding vowel /a/. This is not the case in the short-/ʃ/ speakers, F2 and M2, indicating group-dependent articulatory organization for the production of the sibilant /ʃ/.

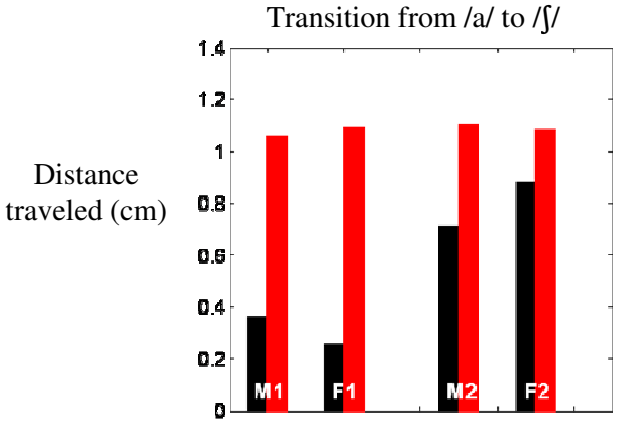

Figure 5. Traveled distance of the sensor coils glued on the lower incisor (jaw) in black and the sensor on the tongue-body in red, during the transition from /a/ to /ʃ/: These distances are calculated on the trajectory as shown in Figure 3 for the long-/ʃ/ speakers, F1 and M1, and the short-/ʃ/ speakers, F2 and M2.

The difference in the articulatory organization can be seen more clearly in Figure 5 where traveled distance of the jaw sensor-coil (lower incisor) and that of the tongue-blade coil during the transition from the initial vowel /a/ to the following sibilant /ʃ/ are compared for the four speakers. It is evident, on one hand, that the traveled distances of the lower jaw are much shorter for the long-/ʃ/ speakers, F1 and M1, than for the short-/ʃ/ speakers, F2 and M2. On the other hand, there is not much difference among the traveled distances of the tongue-body.

It is important to recall that tongue position (and as well as shape) varies in function of the jaw position and of the intrinsic activities of the tongue muscles. Moreover, the jaw position and the intrinsic tongue activities can compensate each other to articulate the target tongue position, at least for vowels (Maeda, 1991). This compensation ability between the jaw and the tongue has been supported by Laprie and Ouni (2005) in the works concerning the acoustic-to-articulatory inversion. The traveled distance of the tongue body shown in Figure 5 must be interpreted on the background of the inter-articulator compensation.

From the above discussion, it is not so unreasonable, as a gross approximation, that the calculated distance traveled of the tongue-blade is the sum of the distance traveled by the jaw and that by the intrinsic tongue muscle activities. If this simple sum principle is valid, then the intrinsic traveled distance of the tongue becomes equal to the observed distance of the tongue (the red bar in Figure 5) minus the corresponding distance of the jaw (the black bar). In our interpretation therefore, the articulatory strategy taken by the long-/ʃ/ speakers is to dominantly use the tongue

with the anticipatory use of the jaw during the preceding vowel, which allows a relatively small jaw motion during the /a/ to /ʃ/ transition. Conversely, the short-/ʃ/ speakers employ relatively small tongue actively and large jaw motion during the transition.

Moreover, it may be interesting to note, now, that the observed traveled distance of the tongue blade is the sum of the jaw motion and the intrinsic tongue motion, as mentioned above. Then about the same distance traveled across all the four speakers roughly means all the speaker expended about the same amount of total effort in the transition from /a/ to /ʃ/, but with different weights on the jaw and on the tongue.

3. Kinematic properties of the articulators, inter-target intervals, and acoustical segment duration: Perspective in large database

In order to verify whether kinematic properties of the articulators allow us to predict the acoustical segment durations or inter-articulatory target intervals across different phonetic contexts, we turn our attention to the MOCHA database (Wrench, 1999). This database contains, among others, EMA derived articulatory data accompany by the acoustic speech signals for 459 phonetically balanced English sentences uttered by two native speakers.

3.1. Profiles of articulator's speed: Comparison on the two speakers

Articulatory speed can be considered as kinematic variable that is intrinsic to individual articulators and, therefore to individual speakers. In our data analysis, the histogram of speed of every articulator exhibits a semi-bell curve as shown in Figure 6 for the lower jaw motions.

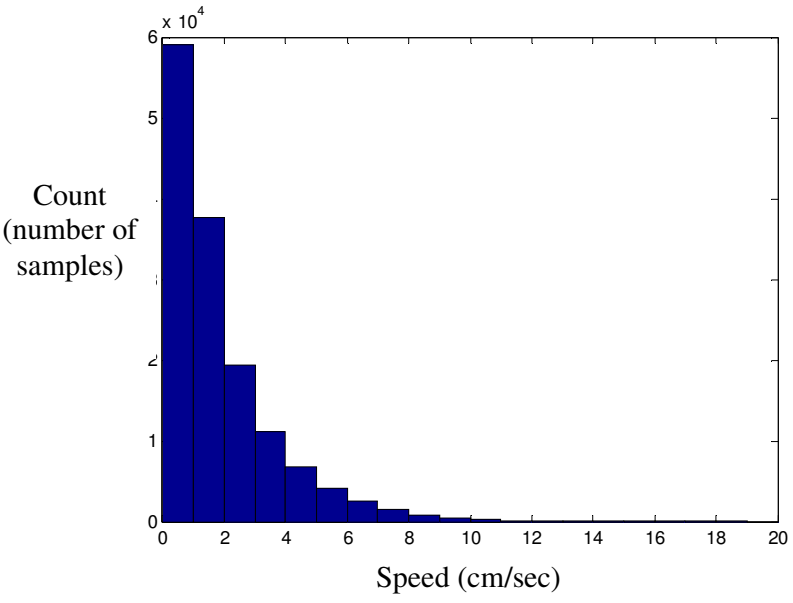

Figure 6. Histogram of speed of jaw motions (more specifically the lower incisor) derived from the data of a female speaker, fsew, in 1 cm/sec bins.

The averaged values and standard deviations of the histograms on speed vary in function of articulators as seen in Table 1. Table 1 summarizes the characteristics of the distribution at the 50th and 95th percentiles in function of the articulators for the two speakers.

Table 1. Descriptive statistics of the speed (in cm/s) of the articulators derived on 459 sentences.

	Fsew (percentiles)		Msak (percentiles)	
	50th	95th	50th	95th
Lower incisor	1.2740	5.5955	0.8913	3.6936
Upper lip	0.8900	4.3234	0.9868	4.1376
Lower lip	3.1387	13.2487	2.5752	11.1703
Tongue tip	4.8562	17.8281	4.3752	16.9814
Tongue body	4.0604	12.2656	3.6032	11.8150
Tongue dorsum	3.4605	11.4810	3.1958	11.1008
Velum	0.5148	2.7575	0.9504	3.8224

It is striking, in Table 1, that average speeds of the articulators, or more specifically the speed of EMA's sensor coils glued of the articulators' surface, vary one articulator to another of which the patterns are common to these two speakers. For the two speakers, the fastest articulator is the tongue-apex, which is followed by the second group, the tongue-blade, tongue-body and lower-lip. The slowest group includes the upper lip, lower-jaw, and velum. Although the number of speakers is only two, this result doesn't contradict with the finding in the previous section 2 in that speaker-dependency seems to manifest more on the articulatory organization rather than basic kinematic properties of individual articulators. In fact, the averaged speed of the jaw of the male speaker msak is considerably slower than the female speaker fsew by factor of (3.7/5.6=) .66. For the speed of the tongue-body of the speaker-msak, for example, rather catches up however, because the factor is now (11.8/12.3) = .96, i.e. about the same seed as the speaker fsew. Although the result here concerns with only two speakers, it appears to corroborate with the finding in the previous Section 2 in that the speaker-dependency is not so much on the basic bio-mechanic properties of the articulators, but rather on the strategies of articulation.

3.2. Accurate automatic measurement of acoustical segment duration of /s/ in the MOCHA database

In order to compare the acoustical segment duration and the kinematic properties, the accurate duration of /s/ is automatically measured on a large number of its occurrences in the MOCHA database using the zero-crossing rate of the acoustic signal as illustrated in Figure 7. This automatic detection procedure for fricatives was strictly applied to intervocalic utterances of fricatives. The /s/ being discussed in this Section 3 is thus limited to intervocalic /s/'s. The result of /s/ duration measurements will be used in Section 3.4 comparing the segment durations and inter-target intervals.

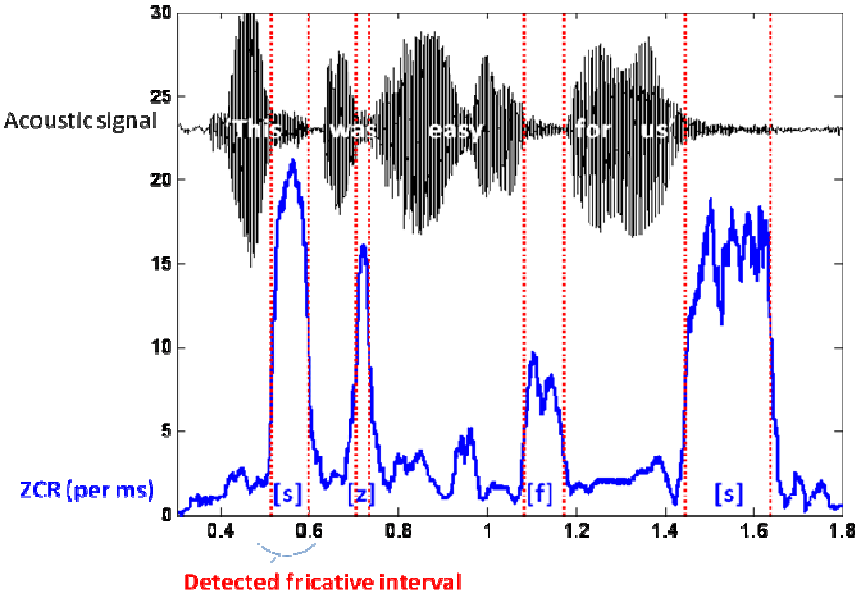

Figure 7. Speech (audio) signal (in black), the calculated zero-crossing rate (in blue), and the detected fricative-vowel boundaries in dotted red lines: The boundaries at the onset and offset of a fricative determine its acoustical segment duration.

3.3. Constrained articulators' positions during /s/ segments and transitions

Figure 8 illustrates the superposition of the trajectory of all the seven articulators during 225 sentences (one half of the database) in gray dots, in which the trajectories during the production of the intervocalic /s/-segments are colored in blue. It is interesting to note that the jaw is the articulator of which position is most narrowly constrained during /s/ segments in comparison with the other six articulators, followed by the lower lip, the tongue-tip, and then the velum. The position of the tongue-body and the tongue-dorsum is distributed vertically, in the high-low dimension, and is constrained in the anterior-posterior dimension.

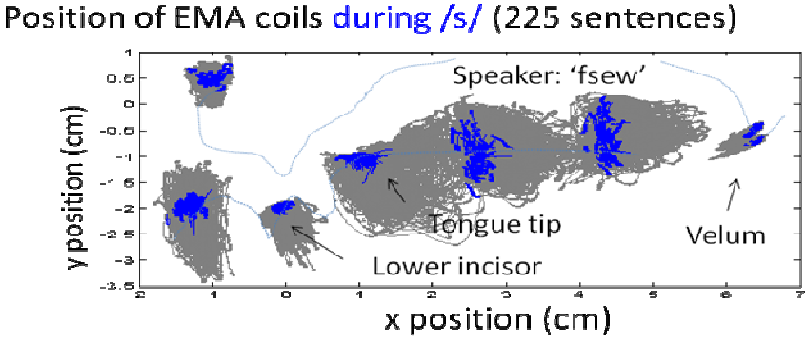

Figure 8. Trajectories of the EMA's sensor coils glued on the seven articulators, the upper-lip, lower-lip, jaw (lower incisor), tongue-tip, tongue-body, tongue-dorsum, and velum, are shown with grey dots and the parts corresponding to all the intervocalic /s/-segments in the 255 sentences with blue dots.

Even though the jaw is the articulator that is most constrained, it is unlikely that the acoustical segment duration of /s/ could be predicted from the kinematic properties of the jaw-muscle system because its positions rarely reach physical limits. Moreover, we cannot expect their predictability, if we consider the stationary nature of the sibilant fricatives, in which the articulators don't move much. It is not unreasonable to expect, however, the effects of kinematic properties could manifest better during transitions between a vowel and the fricative, where there the kinematic activities are high. In fact, as illustrated in Figure 9 lower panel, the acceleration of the lower jaw during transitions is relatively high.

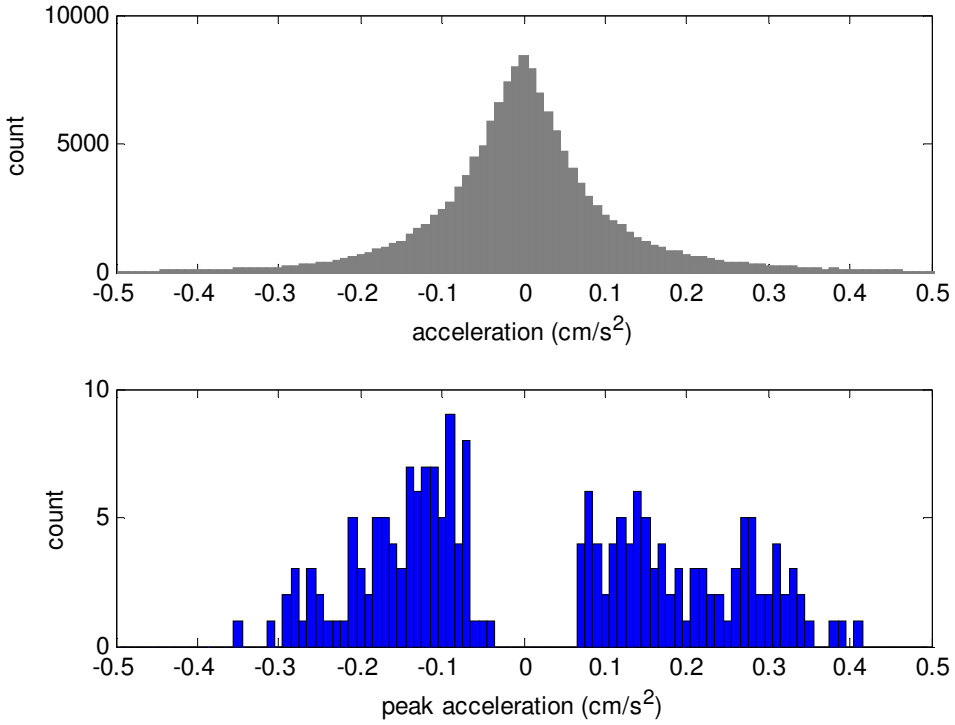

Figure 9. Histogram of acceleration of the jaw (the lower incisor) calculated on the EMA data recorded for the female speaker, fsew: In the top plate, the histogram is derived from the whole data. The bottom plate shows the histogram of peak deceleration, indicated by the minus sign, during transition from the preceding vowel to /s/ and that of peak acceleration from /s/ to the following vowel (in 0.01 cm/s bins).

In the bottom plate, the histogram exhibits a hole in the vicinity of the zero-acceleration, whereas there is a peak around zero for the whole data in the upper plate. This is because only the acceleration and deceleration *peaks* have been taken into account. These peaks are located around the acoustical phoneme edges (during transitions). The higher acceleration (or deceleration) levels means the higher kinematic activities, then in order to find the link between the kinematic properties and the temporal characteristics of the segments, speaker-dependent or not, we must focus our attention at a variable related to articulatory movements during transitions. This is the main reason for why we examine the time interval between successive articulatory target points defined in Section 2.2. We shall call this articulatory interval (in time) as inter-target interval.

3.4. Traveled distance, inter-target interval, and acoustical segment duration

In the light of the basic question whether the intrinsic segmental duration of phonemes are determined by the kinematic constraints of the individual articulators, “intrinsic timing”, or by a “metronome” assumed to exist in a higher level of the speech production chain, “extrinsic timing”, the terms employed by Fowler (1980), it must be interesting to assess whether the kinematic constraint of the tongue allows us to explain the acoustical segment duration or rather the inter-target interval. Here we have chosen the tongue and not the jaw of which the kinematic properties seemed unable to predict segmental duration of /s/ as already discussed in Section 3.3. In this experimental paradigm, we take the traveled distance as an indication of a global or macro kinematic property, because the distance traveled would vary in function of the articulator’s speed for a given motion time.

3.4.1. On the female speaker fsew

We quantify here the degrees of the link between the kinematic properties, and the inter-target interval or the segmental duration discussed in Section 2.2 by calculating the correlations. As described before, the traveled distance is calculated on the xy-trajectory, as shown in Figure 3, delimited by two successive articulatory targets. In this sense, the traveled distance can be called “inter-target distance traveled”. Temporarily, we used the time span defined by target points of the jaw trajectories to calculate the traveled distance of the tongue-body. This means that the inter-target interval of the tongue-body equals that of the jaw. Table 2 summarizes the results.

Table 2. Correlation between traveled distance, and inter-target interval of the articulator (the jaw or tongue-body) and the acoustical segmental duration, for the female speaker fsew

		All utterances		Utterances whose deceleration or acceleration peak stand outside the 90th percentile	
		r	p	r	p
Acoustical interval of /s/	Jaw	0.59	< 0.001	--	--
	Tongue body	0.29	< 0.005	--	--
Inter-target intervals	Jaw; interval preceding the /s/ target	0.53	< 0.001	0.58	< 0.001
	-- following the /s/ target	0.58	< 0.001	0.71	< 0.001
	TB, preceding	0.67	< 0.001	0.63	< 0.001
	TB, following	0.73	< 0.001	0.80	< 0.001

It is remarked in Table 2 that the correlation is, in general, higher for inter-target interval than acoustically determined segmental duration of /s/. This result is expected, as already discussed before, because the inter-target interval (T) is inherently and directly related to the traveled distance (D). If the correlation were perfect, i.e., = 1, then we would have the simple proportional relation as $T = aD$, where ‘a’ denotes a correlation slope. Now, we rewrite the equation as $D = (1/a)T$. Evidently, (1/a) can be interpreted as the average speed of that articulator. In reality, the data points are not distributed along the strait line, but distributed around that line,

depending on the significance of the correlation. Moreover, the lower correlation of the acoustical segment duration, especially of the tongue-body, suggests that the traveled distance of a single articulator cannot predict well the segmental duration. We suspect that the acoustical segment duration is not determined by the activities of a single articulator, but the combination, or more smartly the coordination, of activities of different articulators, although inter-target interval of the individual articulators might be well predicted by their traveled distance. It appears then the result support the intrinsic timing of acoustical segment duration rather than the extrinsic timing. Particularly, inter-target interval from the /s/-target points to those of the following vowel manifests the highest correlation. In other word, this target interval is best predicted by the traveled distance. Moreover, for the inter-target interval, the tongue-blade presents a higher correlation than the jaw, which favors the hypothesis as inter-target interval are more constrained by the kinematic properties of that articulator.

By the way, the bottom plate in Figure 9 indicate that the deceleration and acceleration (i.e. forces) do not reach their maximum level (as observed in the whole dataset) in most of cases, suggesting that only in the minority of cases the articulators (the jaw, in the figure) employ high-level forces, if not their maximum forces. We could assume that the correlation between distance traveled and inter-target interval could become higher in the high-level activities than in the whole data condition, including the majority of low-level activity cases. As mentioned before, the kinematic properties of articulators should manifest better in the distance traveled as the output of the activities during high activity than low activity.

In order to confirm our assumption, we calculated the correlation in two conditions, one using the whole data and the other in which the peak deceleration/acceleration exceeds the 90th percentile. The values of the correlation are always higher in the selected data than in the whole data for the tongue blade, as indicated in Table 2. Note that the selected data of the traveled distance and inter-target interval concerned with the tongue blade from /s/ to the following vowel exhibits the highest correlation coefficient value, 0.80, whereas the value of correlation from the whole date is somewhat lower yet still honorable 0.73. The dispersion of those data points are graphically illustrated in Figure 10. The dispersion of the selected data points covers the same area as that of the whole data points indicated by the white triangles, but with much smaller numbers of points indicated by the red triangles, which presumably leads to its higher correlation value.

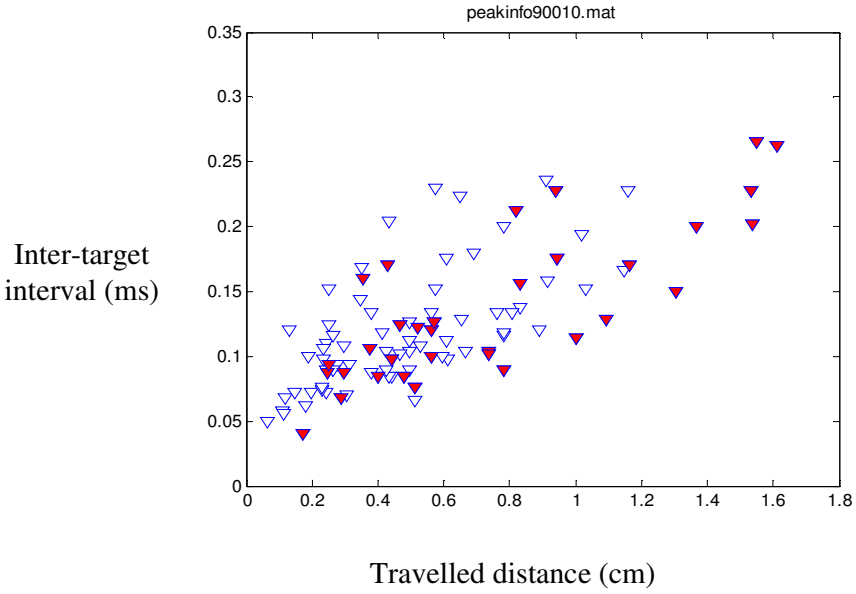

Figure 10. Inter-target interval plotted in function of the traveled distance by the tongue-blade, from the /s/ to a following vowel, concerning with the speaker fsew: The whole data points are plotted with white triangles. Among them, data points in which the peak acceleration (deceleration) exceeds the 90 percentile are colored in red.

3.4.2. On the male speaker msak

The results for the male speaker msak are fairly different from the previous speaker. The histograms of the peak deceleration/acceleration of the lower jaw are plotted in Figure 11, in the same manner as Figure 9. Contrary to the speaker fsew, the histogram of the peak deceleration and acceleration during, respectively, from the preceding vowel to /s/ and from /s/ to following vowel exhibits peaks, one in the deceleration and the other in the acceleration stage.

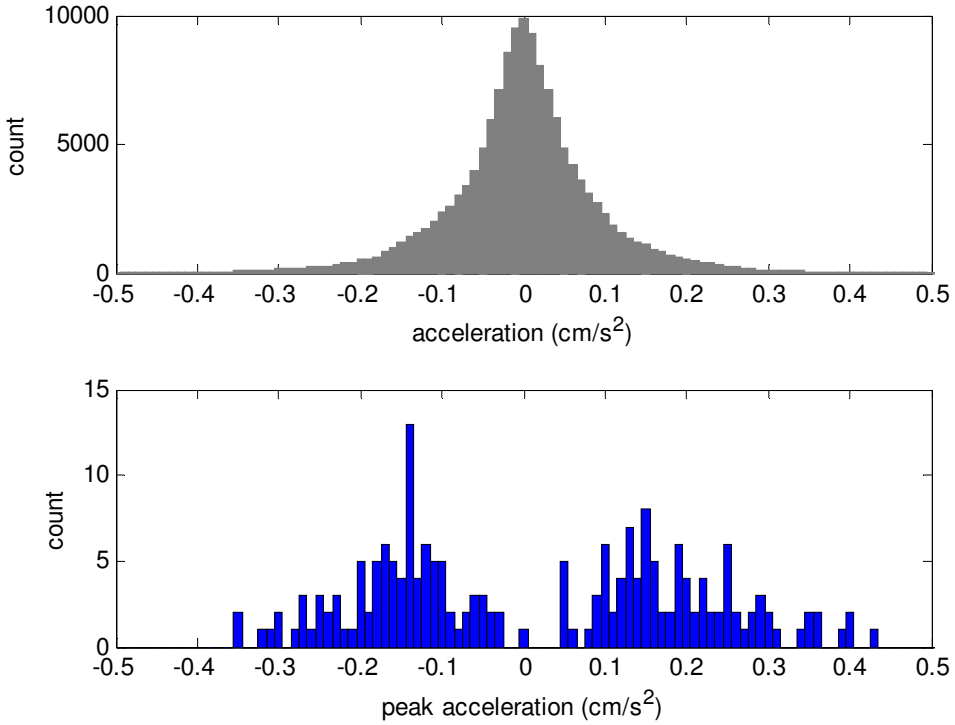

Figure 11. Same as Figure 9, except for the male speaker msak

The presence of these two peaks means the peak acceleration (and deceleration) of this speaker has straight forward statistical properties in that peak acceleration is distributed around a fairly well defined average value, and in turn, implies fairly well defined average speed, which is the integrated acceleration. It is not so unreasonable, then, to expect a reasonably good predictability of the inter-target interval by the traveled distance. In fact, the calculated correlation coefficients between the distance and the interval for the jaw for both from a preceding vowel to /s/, and from /s/ to a following vowel are fairly high as shown in the third and fourth line from the bottom in Table 3.

Table 3. The same as Table 2, except for the male speaker msak

		All utterances		Utterances whose deceleration or acceleration peak stand outside the 90th percentile	
		r	p	r	p
Acoustical interval of /s/	Jaw	0.54	< 0.001	--	--
	Tongue body	0.23	< 0.05	--	--
Inter-target intervals	Jaw; interval preceding the /s/ target	0.74	< 0.001	0.68	< 0.001
	-- following the /s/ target	0.75	< 0.001	0.77	< 0.001
	TB, preceding	0.80	< 0.001	0.74	< 0.001
	TB, following	0.79	< 0.001	0.70	< 0.001

Table 3 also indicates that the inter-target interval can be well predicted from the traveled distance in the case of the tongue body. In fact, the value of correlation coefficients is slightly, but higher than that of the female speaker fsew. Curiously however, selection of data point with the 90th percentile did not improve the scores. In fact, in Table 8 the values of the correlation coefficients calculated on the selected data points are smaller than those on the whole data points, except one case of the jaw movements from /s/ to the following vowel. We suspect that somewhat arbitrarily chosen 90 percentile was too high for this speaker and that all the transitions from one target to another in a short distance and in a short time simultaneously are removed by the selection. This later point is evident in Figure 12, showing the absence of the red triangle in the vicinity of the axis origin, except the one data point, in comparison with the dispersion seen in Figure 10. The value of the threshold, 90th percentile, was somewhat arbitrarily chosen. We need to find out an appropriate threshold value for this speaker, if possible, in a rational way.

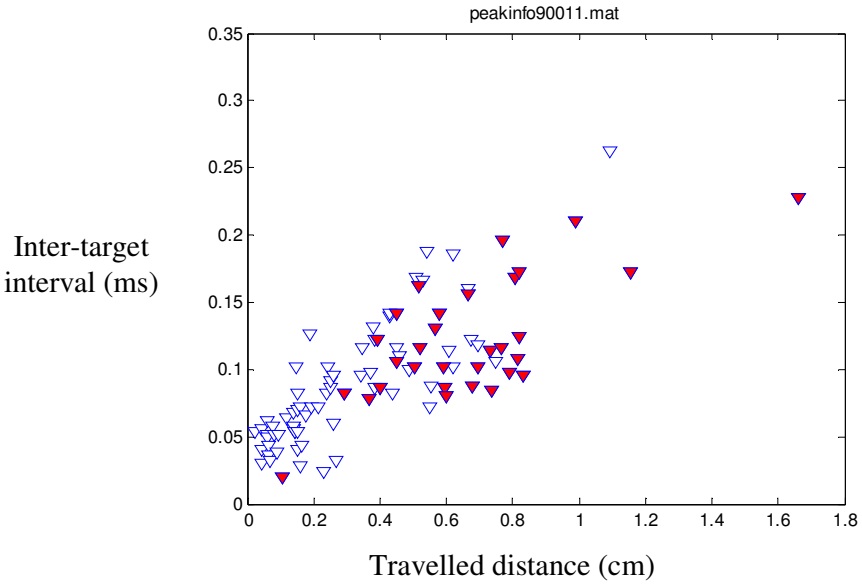

Figure 11, except for the male speaker msak

Conclusion

We have shown that the inter-target interval is fairly well correlated with the traveled distance between articulatory target point in /s/ and that in the preceding or following vowel by the articulator in question. This result is encouraging because it tends to suggest that the articulatory rhythm during speech production is predictable, at least in part, from the kinematic properties of the articulators that could be specific to individual speakers. Inversely, the articulatory rhythm could, therefore, allow us to infer articulatory coordination of tongue and jaw, and the speaker-dependent kinematic properties. The acoustical segment duration of fricative, however, is not correlated in an important way with the traveled distance. The temporal relationships between the articulatory targets that depends, more of less, on the kinematic properties and the acoustic targets that could situate in the vicinity of the center of each phoneme are needed to be established, which might open a route to the link

between the segmental duration, more specifically the intrinsic segmental duration (Klatt, 1974), of phoneme and the speaker-dependent kinematic properties of the articulators.

Références

Ananthakrishnan, Gopal and Olov Engwall, 2008. “Important regions in the articulator trajectory”. *Proc. 8th ISSP*, 305-308.

Ferrer, Luciana, Harry Bratt, Venkata R.R. Gadde, Sachin S. Kajarekar, Elizabeth Shriberg, Kemal Sonmez, Andreas Stolcke, and Anand Venkataraman, 2003. “Modeling duration patterns for speaker recognition”, *EUROSPEECH 2003*, 2017-2020.

Fowler, Carol, 1980. “Coarticulation and theories of extrinsic timing”, *Journal of Phonetics* 8, 113-133.

Fuchs, Susanne and Martine Toda, 2010. “Do differences in male versus female /s/ reflect biological or sociophonetic factors?”. In Fuchs, Toda and Zygis (eds.), *Turbulent sounds. An interdisciplinary guide.* Mouton de Gruyter, 281-302.

Klatt, Dennis H., 1974. “The duration of [s] in English words”, *Journal of Speech and Hearing Research* Vol.17, p. 51-63.

Klatt, Dennis H., 1979. “Synthesis by rule of segmental durations in English sentences.” In Lindblom and Öhman (eds.), *Frontiers of speech communication research*, Academic Press, 287-300.

Laprie, Yves and Slim Ouni, 2005. “Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion”. *J. Acoust. Soc. Am*. 118 (1), 444-460.

Lee, Sook-Hyang, Mary E. Beckman and Michel Jackson, 1994. “Jaw targets for strident fricatives”, *proc ICSLP*, 37-40.

Maeda, Shinji, 1991. “On articulatory and acoustic variabilities”, *Journal of Phonetics* 19, 321-331.

Pfitzinger, Hartmut R., 2002, “Intrinsic phone durations are speaker-specific”, *proc. ICSLP 2002*, 1113-1116.

Vatikiotis-Bateson, E., M. Tiede, Y. Wada, V. Gracco, and M. Kawato, 1994. “Phoneme extraction using via point estimation of real speech”, *proc. ICSLP 1994*, 45-48.

Wrench, Alan, 1999. The MOCHA-TIMIT articulatory database. http://www.cstr.ed.ac.uk/research/projects/artic/mocha.htm.