2

"Targetless" schwa: an articulatory analysis

CATHERINE P. BROWMAN and LOUIS GOLDSTEIN

2.1 Introduction

One of the major goals for a theory of phonetic and phonological structure is to be able to account for the (apparent) contextual variation of phonological units in as general and simple a way as possible.* While it is always possible to state some pattern of variation using a special "low-level" rule that changes the specification of some unit, recent approaches have attempted to avoid stipulating such rules, and instead propose that variation is often the consequence of how the phonological units, properly defined, are *organized*. Two types of organization have been suggested that lead to the natural emergence of certain types of variation: one is that invariantly specified phonetic units may overlap in time, i.e., they may be coproduced (e.g., Fowler 1977, 1981a; Bell-Berti and Harris 1981; Liberman and Mattingly 1985; Browman and Goldstein 1990), so that the overall tract shape and acoustic consequences of these coproduced units will reflect their combined influence; a second is that a given phonetic unit may be unspecified for some dimension(s) (e.g., Öhman 1966b; Keating 1988a), so that the apparent variation along that dimension is due to continuous trajectories between neighboring units' specifications for that dimension.

A particularly interesting case of contextual variation involves reduced (schwa) vowels in English. Investigations have shown that these vowels are particularly malleable: they take on the acoustic (Fowler 1981a) and articulatory (e.g., Alfonso and Baer 1982) properties of neighboring vowels. While Fowler (1981a) has analyzed this variation as emerging from the coproduction of the reduced vowels and a neighboring stressed vowel, it might also be the case that schwa is completely unspecified for tongue position. This would be consistent with analyses of formant trajectories for medial schwa in trisyllabic sequences (Magen 1989) that have shown that F_2 moves (roughly continuously) from a value dominated by the preceding vowel (at onset) to one dominated by the following vowel (at offset). Such an analysis would also be consistent with the phonological analysis of schwa in French (Anderson 1982) as an empty nucleus slot. It is possible (although this is not Anderson's analysis), that the empty nucleus is never filled in by any specification, but rather there is a specified "interval" of time between two full vowels in which the tongue continuously moves from one vowel to another.

The computational gestural model being developed at Haskins Laboratories (e.g. Browman *et al.* 1986; Browman and Goldstein, 1990; Saltzman *et al.* 1988a) can serve as a useful vehicle for testing these (and other) hypotheses about the phonetic/phonological structure of utterances with such reduced schwa vowels. As we will see, it is possible to provide a simple, abstract representation of such utterances in terms of gestures and their organization that can yield the variable patterns of articulatory behavior and acoustic consequences that are observed in these utterances.

The basic phonetic/phonological unit within our model is the gesture, which involves the formation (and release) of a linguistically significant constriction within a particular vocal-tract subsystem. Each gesture is modeled as a dynamical system (or set of systems) that regulates the time-varying coordination of individual articulators in performing these constriction tasks (Saltzman 1986). The dimensions along which the vocal-tract goals for constrictions can be specified are called **tract variables**, and are shown in the left-hand column of figure 2.1. Oral constriction gestures are defined in terms of pairs of these tract variables, one for constriction location, one for constriction degree. The right-hand side of the figure shows the individual articulatory variables whose motions contribute to the corresponding tract variable.

The computational system sketched in figure 2.2 (Browman and Goldstein, 1990; Saltzman *et al.* 1988a) provides a representation for arbitrary (English) input utterances in terms of such gestural units and their organization over time, called the **gestural score**. The layout of the gestural score is based on the principles of intergestural phasing (Browman and Goldstein 1990) specified in the **linguistic gestural model**. The gestural score is input to the **task-dynamic model** (Saltzman 1986; Saltzman and Kelso 1987), which calculates the patterns of articulator motion that result from the set of active gestural units. The articulatory movements produced by the task-dynamic model are then input to an articulatory synthesizer (Rubin, Baer, and Mermelstein 1981) to calculate an output speech waveform. The operation of

956

*Our thanks to Ailbhe Ní Chasaide, Carol Fowler, and Doug Whalen for criticizing versions of this paper. This work was supported by NSF grant BNS 8820099 and NIH grants HD-01994 and NS-13617 to Haskins Laboratories.

	Tract variable	Articulators involved
LP	lip protrusion	upper and lower lips, jaw
LA	lip aperture	upper and lower lips, jaw
TTCL	tongue-tip constriction location	tongue tip, tongue body, jaw
TTCD	tongue-tip constriction degree	tongue-tip, tongue-body, jaw
TBCL	tongue-body constriction location	tongue-body, jaw
TBCD	tongue-body constriction degree	tongue-body, jaw
VEL	velic aperture	velum
GLO	glottal aperture	glottis

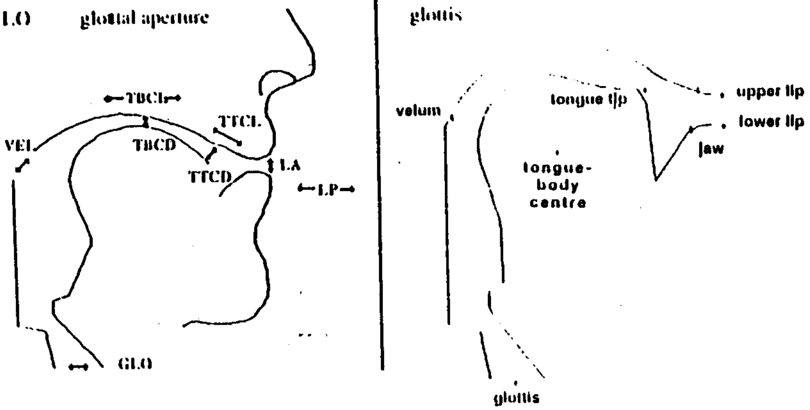

Figure 2.1 Tract variables and associated articulators

the task-dynamic model is assumed to be "universal." (In fact, it is not even specific to speech, having originally been developed [Saltzman and Kelso 1987] to describe coordinated reaching movements.) Thus, all of the language-particular phonetic/phonological structure must reside in the gestural score – in the dynamic parameter values of individual gestures, or in their relative timing. Given this constraint, it is possible to test the adequacy of some particular hypothesis about phonetic structure, as embodied in a particular gestural score, by using the model to generate the articulatory motions and comparing these to observed articulatory data.

The computational model can thus be seen as a tool for evaluating the articulatory (and acoustic) consequences of hypothesized aspects of gestural structure. In particular, it is well suited for evaluating the consequences of the organizational properties discussed above: (1) underspecification and (2) temporal overlap. Gestural structures are inherently underspecified in the sense that there are intervals of time during which the value of a given tract variable is not being controlled by the system; only when a gesture defined along that tract variable is active is such control in place. This underspecification can be seen in figure 2.3, which shows the gestural score for the utterance /pam/. Here, the shaded boxes indicate the gestures, and are superimposed on the tract-variable time functions produced when the gestural score is input to the task-dynamic model. The horizontal dimension of the shaded boxes indicates the intervals of time during which each of the gestural units is active, while the height of the boxes corresponds to the "target" or equilibrium position parameter of a given gesture's dynamical control regime. See Hawkins (this volume) for a more complete description of the model and its parameters.

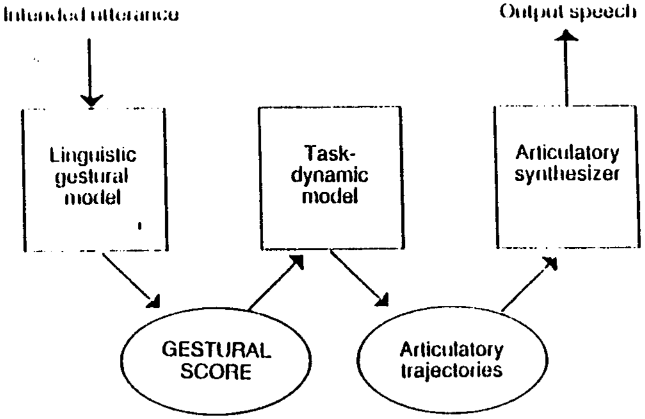

Figure 2.2 Overview of GEST: gestural computational model

Note that during the activation interval of the initial bilabial closure gesture, Lip Aperture (LA – vertical distance between the two lips) gradually decreases, until it approaches the regime's target. However, even after the regime is turned off, LA shows changes over time. Such "passive" tract-variable changes result from two sources: (1) the participation of one of the (uncontrolled) tract variable's articulators in some other tract variable which *is* under active gestural control, and (2) an articulator-specific "neutral" or "rest" regime, that takes control of any articulator which is not currently active in any gesture. For example, in the LA case shown here, the jaw contributes to the Tongue-Body constriction degree (TBCD) gesture (for the vowel) by lowering, and this has the side effect of increasing LA. In addition, the upper and lower lips are not involved in any active gesture, and so move towards their neutral positions with respect to the upper and lower teeth, thus further contributing to an increase in LA. Thus, the geometric structure of the model itself (together with the set of articulator-neutral values) predicts a specific, well-behaved time function for a given tract variable, even

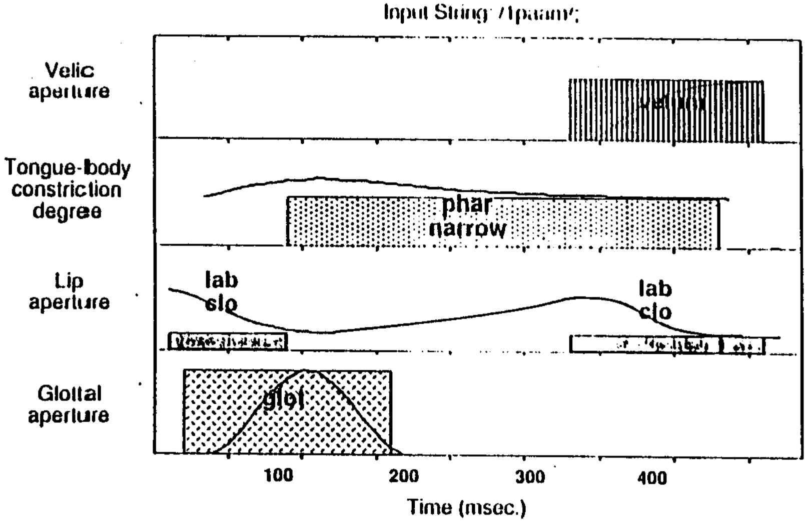

Figure 2.3 Gestural score and generated motion variables for /pam/. The input is specified in ARPAbet, so /pam/ = ARPAbet input string /paam/. Within each panel, the height of the box indicates degree of opening (aperture) of the relevant constriction: the higher the curve (or box) the greater the amount of opening

when it is not being controlled. Uncontrolled behavior need not be stipulated in any way. This feature of the model is important to being able to test the hypothesis that schwa may not involve an active gesture at all.

The second useful aspect of the model is the ability to predict consequences of the temporal overlap of gestures, i.e., intervals during which there is more than one concurrently active gesture. Browman and Goldstein (1990) have shown that the model predicts different consequences of temporal overlap, depending on whether the overlapping gestures involve the same or different tract variables, and that these different consequences can actually be observed in allophonic variations and "casual speech" alternations. Of particular importance to analyzing schwa is the shared tract variable case, since we will be interested in the effects of overlap between an active schwa gesture (if any) and the preceding or following vowel gesture, all of which would involve the Tongue-Body tract variables (TBCD, and Tongue-Body constriction location – TBCL). In this case, the dynamic parameter values for the overlapping gestures are "blended," according to a competitive blending dynamics (Saltzman *et al.* 1988a; Saltzman and Munhall 1989). In the examples we will be examining, the blending will have the effect of averaging the parameter values. Thus, if both gestures were coextensive for their entire activation intervals, neither target value would be achieved, rather, the value of the tract variable at the end would be the average of their targets.

In this paper, our strategy is to analyze movements of the tongue in utterances with schwa to determine if the patterns observed provide evidence for a specific schwa tongue target. Based on this analysis, specific hypotheses about the gestural overlap in utterances with schwa are then tested by means of computer simulations using the gestural model described above.

2.2 Analysis of articulatory data

Using data from the Tokyo X-ray archive (Miller and Fujimura 1982), we analyzed /pV1pə'pV2pə/ utterances produced by a speaker of American English, where V1 and V2 were all possible combinations of /i, ε, a, ʌ, u/. Utterances were read in short lists of seven or eight items, each of which had the same V1 and different V2s. One token (never the initial or final item in a list) of each of the twenty-five utterance types was analyzed. The microbeam data tracks the motion of five pellets in the mid-sagittal plane. Pellets were located on the lower lip (L), the lower incisor for jaw movement (J), and the midline of the tongue: one approximately at the tongue blade (B), one at the middle of the tongue dorsum (M), and one at the rear of the tongue dorsum (R).

Ideally, we would use the information in tongue-pellet trajectories to infer a time-varying representation of the tongue in terms of the dimensions in which vowel-gesture targets are defined, e.g., for our model (or for Wood 1982), location and degree of tongue-body constriction. (For Ladefoged and Lindau [1989], the specifications would be rather in terms of formant frequencies linked to the factors of front-raising and back-raising for the tongue.) Since this kind of transformation cannot currently be performed with confidence, we decided to describe the vowels directly in terms of the tongue-pellet positions. As the tongue-blade pellet (B) was observed to be largely redundant in these utterances (not surprising, since they involve only vowels and bilabial consonants), we chose to measure the horizontal (X) and vertical (Y) positions for the M and R pellets for each vowel. While not ideal, the procedure at least restricts its *a priori* assumption about the parameterization of the tongue shape to that inherent in the measurement technique.

The first step was to find appropriate time points at which to measure the position of the pellets for each vowel. The time course of each tongue-pellet dimension (MX, MY, RX, RY) was analyzed by means of an algorithm that detected displacement extrema (peaks and valleys). To the extent that there is a characteristic pellet value associated with a given vowel, we may expect to see such a displacement extremum, that is, movement towards some value, then away again. The algorithm employed a noise level of one X-ray grid unit (approximately 0.33 mm); thus, movements of a single unit in one direction and back again did not constitute extrema. Only the interval that included the full vowels and the medial schwa was analyzed; final schwas were not analyzed. In general, an extremum was found that coincided with each full vowel, for each pellet dimension, while such an extremum was missing for schwa in over half the cases. The pellet positions at these extrema were used as the basic measurements for each vowel. In cases where a particular pellet dimension had no extremum associated with a vowel, a reference point was chosen that corresponded to the time of an extremum of one of the other pellets. In general, MY was the source of these reference points for full vowels, and RY was the source for schwa, as these were dimensions that showed the fewest missing extrema. After the application of this algorithm, each vowel in each utterance was categorized by the value at a single reference point for each of the four pellet dimensions. Since points were chosen by looking only at data from the tongue pellets themselves, these are referred to as the "tongue" reference points.

To illustrate this procedure, figure 2.4a shows the time courses of the M, R, and L pellets (only vertical for L) for the utterance /pipə'pipə/ with the extrema marked with dashed lines. The acoustic waveform is displayed at the top. For orientation, note that there are four displacement peaks marked for LY, corresponding to the raising of the lower lip for the four bilabial-closure gestures for the consonants. Between these peaks three valleys are marked, corresponding to the opening of the lips for the three vowels. For MX, MY, and RX, an extremum was found associated with each of the full vowels and the medial schwa. For RY, a peak was found for schwa, but not for V1. While there is a valley detected following the peak for schwa, it occurs during the consonant closure interval, and therefore is not treated as associated with V2. Figure 2.4b shows the same utterance with the complete set of "tongue" reference points used to characterize each vowel. Reference points that have been copied from other pellets (MY in both cases) are shown as solid lines. Note that the consonant-closure interval extremum has been deleted.

Figure 2.5 shows the same displays for the utterance /pipə'papə/. Note that, in (a), extrema are missing for schwa for MX, MY, and RX. This is typical of cases in which there is a large pellet displacement between V1 and V2. The trajectory associated with such a displacement moves from V1 to V2, with no intervening extremum (or even, in some cases, no "flattening" of the curve).

As can be seen in figures 2.4 and 2.5, the reference points during the schwa tend to be relatively late in its acoustic duration. As we will be evaluating the relative contributions of V1 and V2 in determining the pellet positions for schwa, we decided also to use a reference point earlier in the schwa. To obtain such a point, we used the valley associated with the lower lip for the

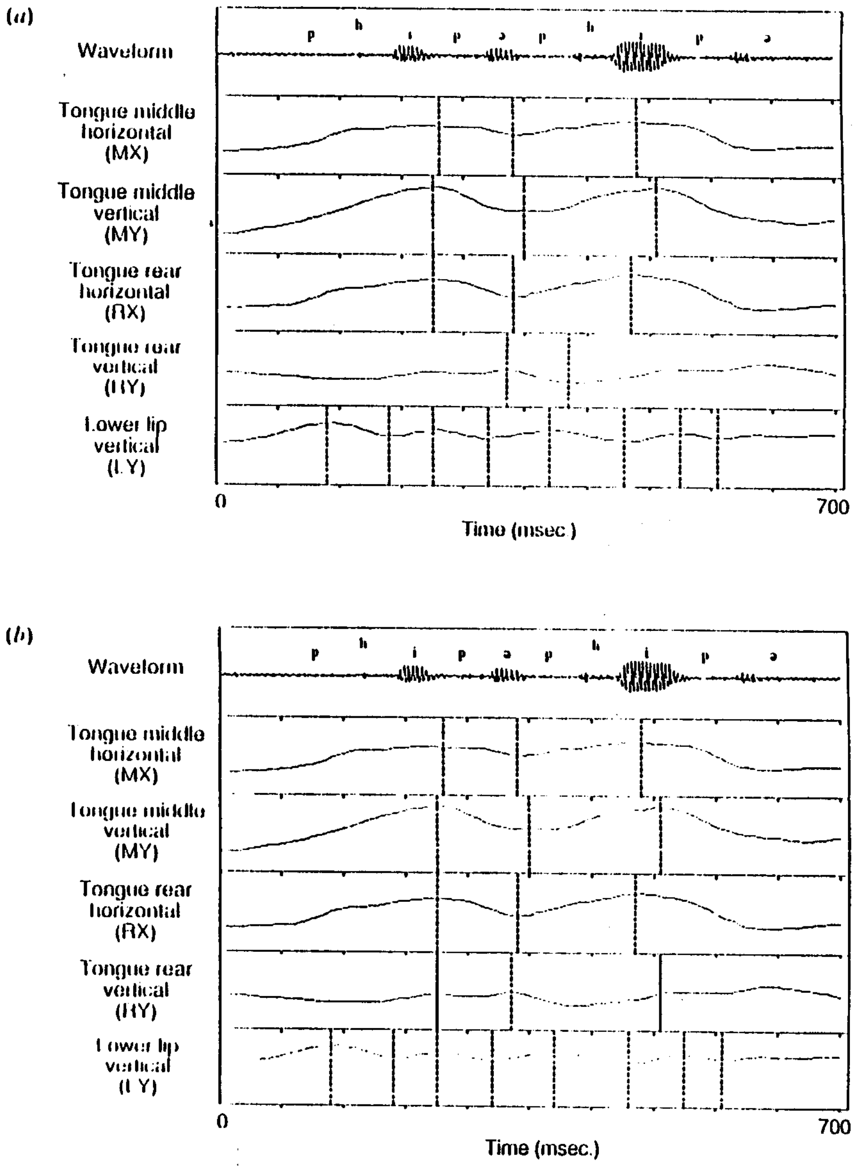

Figure 2.4 Pellet time traces for /pipə'pipə/. The higher the trace, the higher (vertical) or more fronted (horizontal) the corresponding movement. (a) Position extrema indicated by dashed lines. (b) "Tongue" reference points indicated by dashed and solid lines (for Middle and Rear pellets)

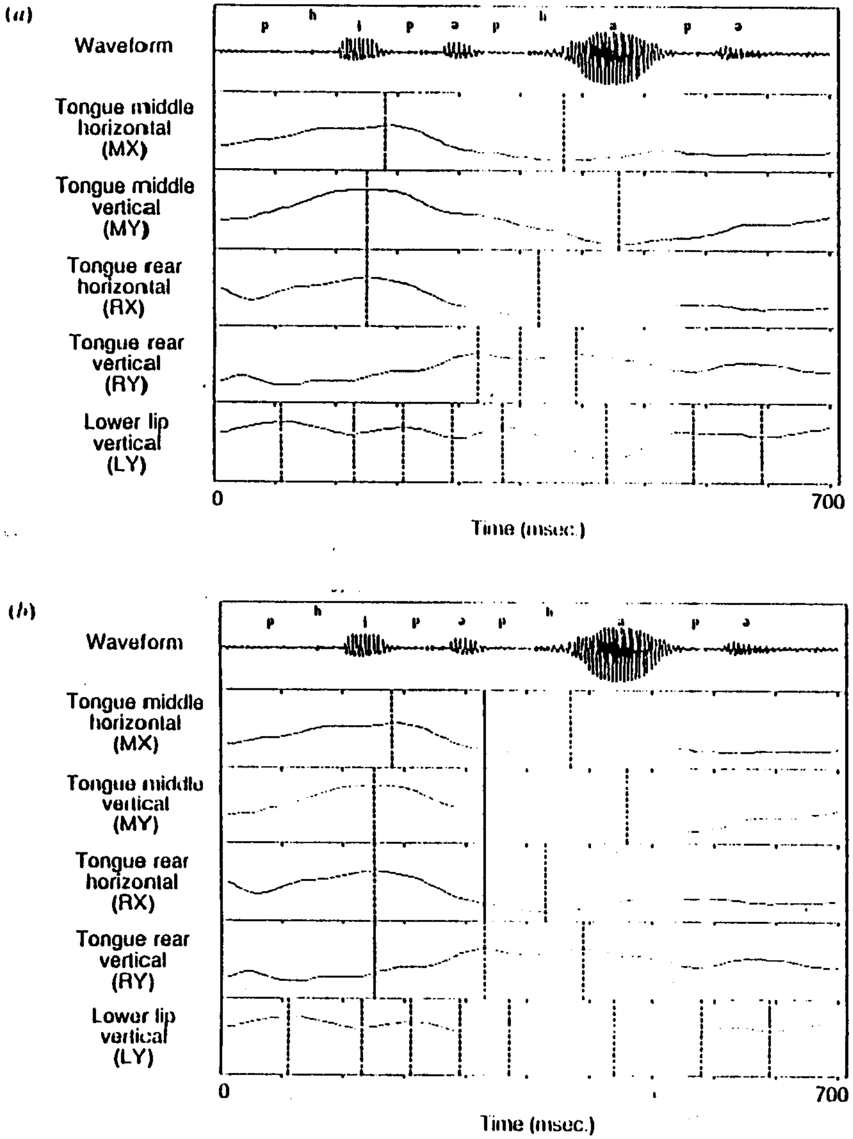

Figure 2.5 Pellet time traces for /pipə'papə/. The higher the trace, the higher (vertical) or more fronted (horizontal) the corresponding movement. (a) Position extrema indicated by dashed lines. (b) "Tongue" reference points indicated by dashed and solid lines (for Middle and Rear pellets)

schwa – that is, approximately the point at which the lip opening is maximal. This point, called the "lip" reference, typically occurs earlier in the (acoustic) vowel duration than the "tongue" reference point, as can be seen in figures 2.4 and 2.5. Another advantage of the "lip" reference point is that all tongue pellets are measured at the same moment in time. Choosing points at different times for different dimensions might result in an apparent differential influence of V1 and V2 across dimensions. Two different reference points were established only for the schwa, and not for the full vowels. That is, since the full vowels provided possible environmental influences on the schwa, the measure of that influence needed to be constant for comparisons of the "lip" and "tongue" schwa points. Therefore, in analyses to follow, when "lip" and "tongue" reference points are compared, these points differ only for the schwa. In all cases, full vowel reference points are those determined using the tongue extremum algorithm described above.

2.2.1 Results

Figure 2.6 shows the positions of the M (on the right) and R (on the left) pellets for the full vowels plotted in the mid-sagittal plane such that the speaker is assumed to be facing to the right. The ten points for a given vowel are enclosed in an ellipse indicating their principal components (two standard deviations along each axis). The tongue shapes implied by these pellet positions are consistent with cinefluorographic data for English vowels (e.g., Perkell 1969; Harshman, Ladefoged, and Goldstein, 1977; Nearey 1980). For example, /i/ is known to involve a shape in which the front of the tongue is bunched forward and up towards the hard palate, compared, for example, to /ɛ/, which has a relatively unconstricted shape. This fronting can be seen in both pellets. In fact, over all vowels, the horizontal components of the motion of the two pellets are highly correlated (r = 0.939 in the full vowel data, between RX and MX over the twenty-five utterances). The raising for /i/ can be seen in M (on the right), but not in R, for which /i/ is low – lower, for example, than /a/. The low position of the back of the tongue dorsum for /i/ can, in fact, be seen in mid-sagittal cinefluorographic data. Superimposed tongue surfaces for different English vowels (e.g. Ladefoged 1982) reveal that the curves for /i/ and /a/ cross somewhere in the upper pharyngeal region, so that in front of this point, /i/ is higher than /a/, while behind this point, /a/ is higher. This suggests that the R pellet in the current experiment is far enough back to be behind this cross-over point. /u/ involves raising of the rear of the tongue dorsum (toward the soft palate), which is here reflected in the raising of both the R and M pellets. In general, the vertical components of the two pellets are uncorrelated across the set of vowels as a whole (r = 0.020),

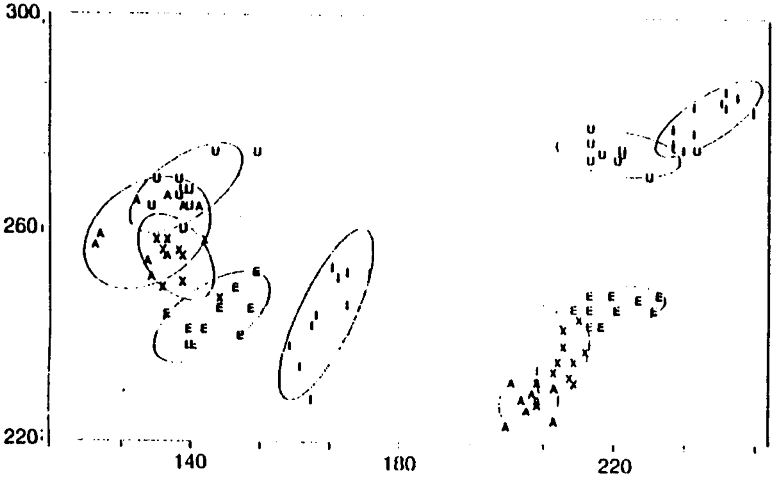

Figure 2.6 Pellet positions for full vowels, displayed in mid-sagittal plane with head facing to the right: Middle pellets on the right, Rear pellets on the left. The ellipses indicate two standard deviations along axes determined by principal-component analysis. Symbols I = IPA /i/, U = /u/, E = /ɛ/, X = /ʌ/, and A = /a/. Units are X-ray units (= 0.33 mm)

reflecting, perhaps, the operation of two independent factors such as "front-raising" and "back-raising" (Ladefoged 1980).

The pellet positions for schwa, using the "tongue" reference points, are shown in the same mid-sagittal plane in figure 2.7, with the full vowel ellipses added for reference. The points are labeled by the identity of the following vowel (V2) in (a) and by the preceding vowel (V1) in (b). Figure 2.8 shows the parallel figure for schwa measurements at the "lip" reference point. In both figures, note that the range of variation for schwa is less than the range of variation across the entire vowel space, but greater than the variation for any single full vowel. Variation in MY is particularly large compared to MY variation for any full vowel. Also, while the distribution of the R pellet positions appears to center around the value for unreduced /ʌ/, which might be thought to be a target for schwa, this is clearly not the case for the M pellet, where the schwa values seem to center around the region just above /ɛ/. For both pellets, the schwa values are found in the center of the region occupied by the full vowels. In fact, this relationship turns out to be quite precise.

Figure 2.9 shows the mean pellet positions for each full vowel and for schwa ("lip" and "tongue" reference points give the same overall means), as well as the grand mean of pellet positions across all full vowels, marked by a circle. The mean pellet positions for the schwa lie almost exactly on top of the

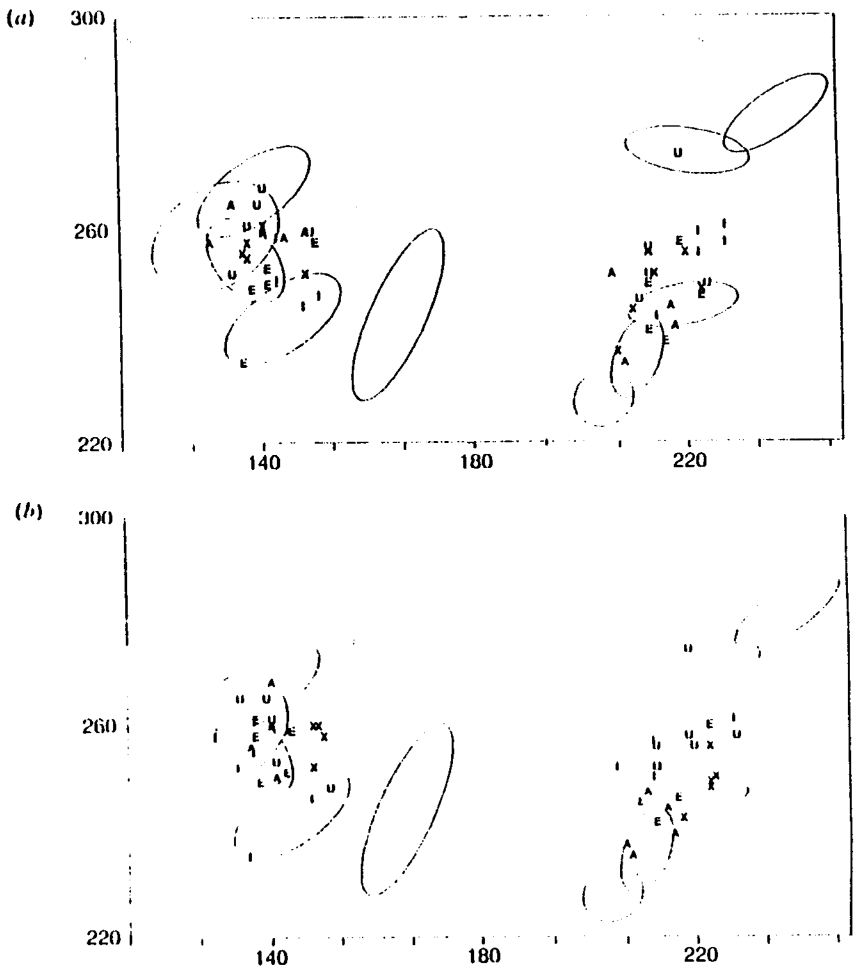

Figure 2.7 Pellet positions for schwa at "tongue" reference points, displayed in right-facing mid-sagittal plane as in figure 2.6. The ellipses are from the full vowels (figure 2.6), for comparison. Symbols I = IPA /i/, U = /u/, E = /ɛ/, X = /ʌ/, and A = /a/. Units are X-ray units (= 0.33 mm). (a) Schwa pellet positions labeled by the identity of the following vowel (V2). (b) Schwa pellet positions labeled by the identity of the preceding vowel (V1)

grand mean for both the M and R pellets. This pattern of distribution of schwa points is exactly what would be expected if there were no independent target for schwa but rather a continuous tongue trajectory from V1 to V2. Given all possible combinations of trajectory endpoints (V1 and V2), we would expect the mean value of a point located at (roughly) the midpoint of

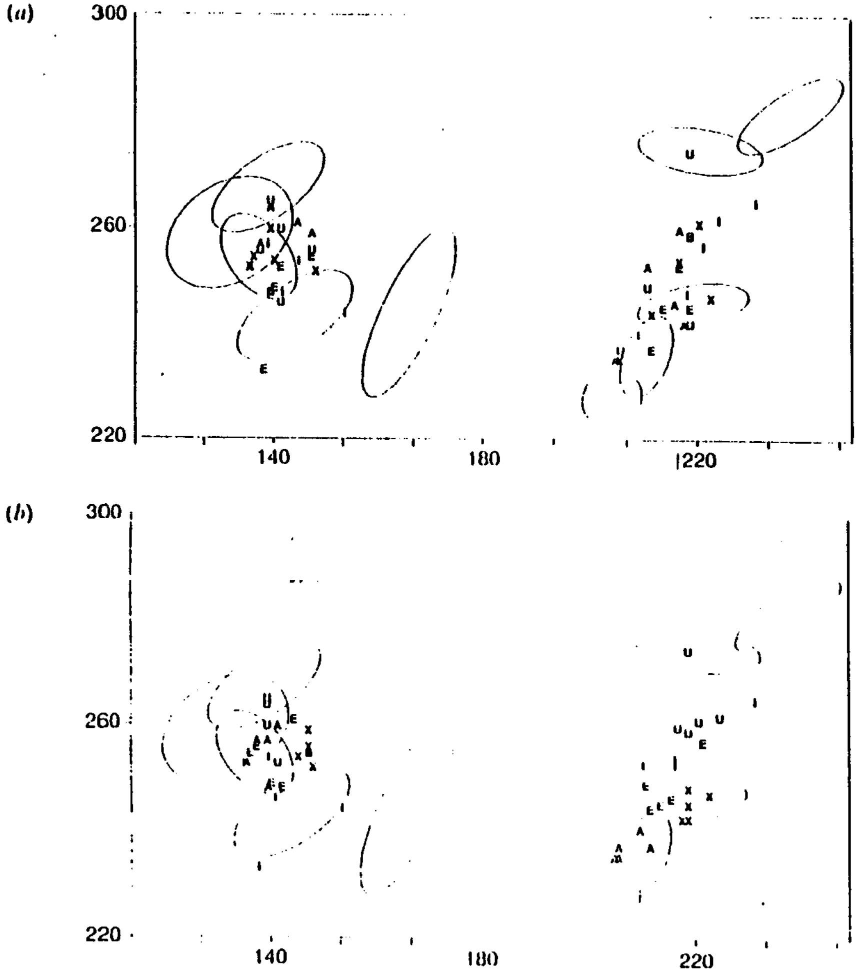

Figure 2.8 Pellet positions for schwa at "lip" reference points, displayed as in figure 2.7 (including ellipses from figure 2.6). Symbols I = IPA /i/, U = /u/, E = /ɛ/, X = /ʌ/, and A = /a/. Units are X-ray units (= 0.33 mm). (a) Schwa pellet positions labeled by the identity of the following vowel (V2). (b) Schwa pellet positions labeled by the identity of the preceding vowel (V1)

these twenty-five trajectories to have the same value as the mean of the endpoints themselves.

If it is indeed the case that the schwa can be described as a targetless point on the continuous trajectory from V1 to V2, then we would expect that the schwa pellet positions could be predicted from knowledge of V1 and V2

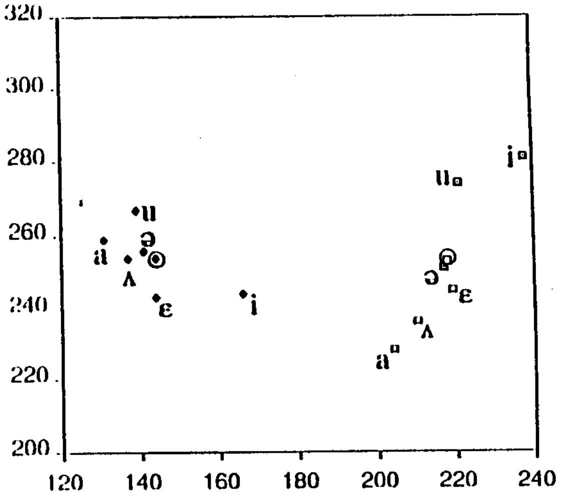

Figure 2.9 Mean pellet positions for full vowels and schwa, displayed in right-facing mid-sagittal plane as in figure 2.6. The grand mean of all the full vowels is indicated by a circled square. Units are X ray units (= 0.33 mm)

positions alone, with no independent contribution of schwa. To test this, we performed stepwise multiple linear regression analyses on all possible subsets of the predictors V1 position, V2 position, and an independent schwa factor, to determine which (linear) combinations of these three predictors best predicted the position of a given pellet dimension during schwa. The analysis finds the values of the *b* coefficients and the constant *k*, in equations like (1) below, that give the best prediction of the actual schwa values.

(1) schwa(predicted) = b1*V1 + b2*V2 + k

The *stepwise* procedure means that variables are added into an equation such as (1) one at a time, in the order of their importance to the prediction. The procedure was done separately for equations with and without the constant term *k* (using BMDP2R and 9R). For those analyses containing the constant term (which is the y-intercept), *k* represents an independent schwa contribution to the pellet position – when it is the only predictor term, it is the mean for schwa. Those analyses without the constant term (performed using 9R) enabled the contributions of V1 and V2 to be determined in the absence of this schwa component. Analyses were performed separately for each pellet dimension, and for the "tongue" and "lip" reference points.

The results for the "tongue" points are shown in the right-hand columns of table 2.1. For each pellet, the various combinations of terms included in the equation are rank-ordered according to the standard error of the schwa prediction for that combination, the smallest error shown at the top. In all cases, the equation with all three terms gave the least error (which is necessarily true). Interestingly, however, for MX, RX, and RY, the prediction using the constant and V2 differed only trivially from that using all three variables. This indicates that, for these pellets, V1 does not contribute substantially to schwa pellet positions at the "tongue" reference point. In addition, it indicates that an independent schwa component is important to the prediction, because V2 alone, or in combination with V1, gives worse prediction than V2 plus *k*.

For MY, all three terms seem to be important — removing any one of them increases the error. Moreover, the second-best prediction involves deleting the V2 term, rather than V1. The reduced efficacy of V2 (and increased efficacy of V1) in predicting the MY value of schwa may be due, in part, to the peak determination algorithm employed. When V1 or V2 was /a/ or /ʌ/, the criteria selected a point for MY that tended to be much later in the vowel than the point chosen for the other pellet dimensions (figure 2.5b gives an example of this). Thus, for V2, the point chosen for MY is much further in time from the schwa point than is the case for the other dimensions, while for V1, the point chosen is often closer in time to the schwa point.

The overall pattern of results can be seen graphically in figure 2.10. Each panel shows the relation between "tongue" pellet positions for schwa and the full vowels: V1 in the top row and V2 in the bottom row, with a different pellet represented in each column. The points in the top row represent the pellet positions for the utterances with the indicated initial vowel (averaged across five utterances, each with a different final vowel), while the bottom row shows the average for the five utterances with the indicated final vowel. The differences between the effects of V1 (top row) and V2 (bottom row) on schwa can be observed primarily in the systematicity of the relations. The relation between schwa and V2 is quite systematic -- for every pellet, the lines do not cross in any of the panels of the bottom row — while for V1 (in the top row), the relationship is only systematic for RY (and somewhat for MY, where there is some crossing, but large effects).

Turning now to the "lip" reference points, regression results for these points are found in the left-hand column of table 2.1. The best prediction again involves all three terms, but here, in every case except RX, the best two-term prediction does substantially worse. Thus V1, which had relatively little impact at the "tongue" point, does contribute to the schwa position at this earlier "lip" point. In fact, for these three pellets, the second-best prediction combination always involves V1 (with either V2 or *k* as the second term). This pattern of results can be confirmed graphically in figure 2.11. Comparing the V1 effects sketched in the top row of panels in figures 2.10 and 2.11, notice that, although the differences are small, there is more spread between the schwa pellets at the "lip" point (figure 2.11) than at the "tongue" point (figure 2.10). This indicates that the schwa pellet was more affected by V1 at the "lip" point. There is also somewhat less cross-over for MX and RX in the "lip" figure, indicating increased systematicity of the V1 effect.

Table 2.1 *Regression results for X-ray data*

	"Lip" reference point			"Tongue" reference point	
	Terms	*Standard error*		*Terms*	*Standard error*
MX	k + v1 + v2	4.6	MX	k + v2 + v1	4.5
	k + v1	5.5		k + v2	4.6
	v1 + v2	5.6		k	6.2
	k	6.3		v2 + v1	6.4
	v1	8.6		v2	10.8
MY	k + v1 + v2	4.2	MY	k + v1 + v2	4.7
	k + v1	5.0		k + v1	6.4
	v1 + v2	7.6		k	8.7
	k	10.3		v1 + v2	9.1
	v1	11.9		v1	15.1
RX	k + v2 + v1	3.8	RX	k + v2 + v1	4.0
	k + v2	3.9		k + v2	4.0
	k	4.8		k	5.8
	v1 + v2	7.0		v2 + v1	7.7
	v1	11.7		v2	10.6
RY	k + v2 + v1	4.1	RY	k + v2 + v1	4.5
	v2 + v1	4.9		k + v2	4.6
	k + v2	5.1		v2 + v1	5.3
	k	6.8		v2	6.2
	v2	7.0		k	7.1

In summary, it appears that the tongue position associated with medial schwa cannot be treated simply as an intermediate point on a direct tongue trajectory from V1 to V2. Instead, there is evidence that this V1–V2 trajectory is warped by an independent schwa component. The importance of this warping can be seen, in particular, in utterances where V1 and V2 are identical (or have identical values on a particular pellet dimension). For example, returning to the utterance /pipə'pipə/ in figure 2.4, we can clearly see (in MX, MY, and RX) that there is definitely movement of the tongue away from the position for /i/ between the V1 and V2. This effect is most pronounced for /i/. For example, for MY, the prediction error for the

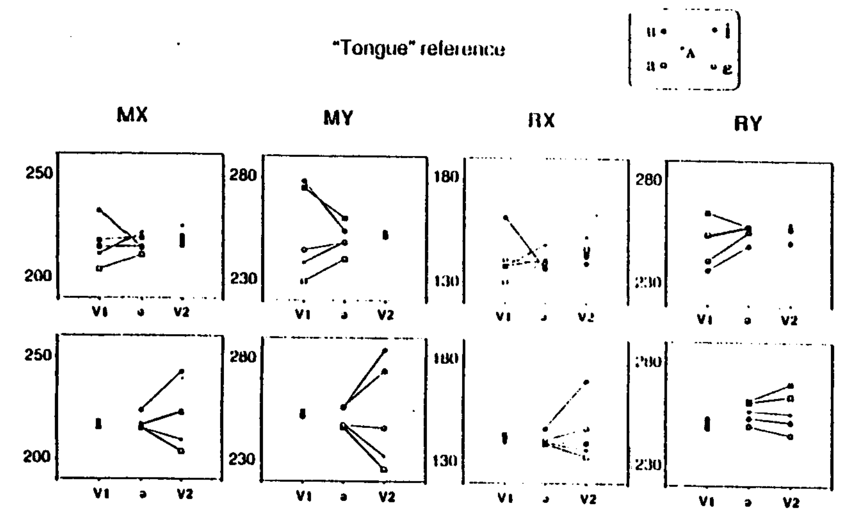

Figure 2.10 Relation between full vowel pellet positions and "tongue" pellet positions for schwa. The top row displays the pellet positions for utterances with the indicated initial vowels, averaged across five utterances (each with a different final vowel). The bottom row displays the averaged pellet positions for utterances with the indicated final vowels. Units are X-ray units (= 0.33 mm)

equation without a constant is worse for /pipə'pipə/ than for any other utterance (followed closely by utterances combining /i/ and /u/; MY is very similar for /i/ and /u/). Yet, it may be inappropriate to consider this warping to be the result of a target specific to schwa, since, as we saw earlier, the mean tongue position for schwa is indistinguishable from the mean position of the tongue across all vowels. Rather the schwa seems to involve a warping of the trajectory toward an overall average or neutral tongue position. Finally, we saw that V1 and V2 affect schwa position differentially at two points in time. The influence of the V1 endpoint is strong and consistent at the "lip" point, relatively early in the schwa, while V2 influence is strong throughout. In the next section, we propose a particular model of gestural structure for these utterances, and show that it can account for the various patterns that we have observed.

2.3 Analysis of simulations

Within the linguistic gestural model of Browman and Goldstein (1990), we expect to be able to model the schwa effects we have observed as resulting from a structure in which there is an active gesture for the medial schwa, but

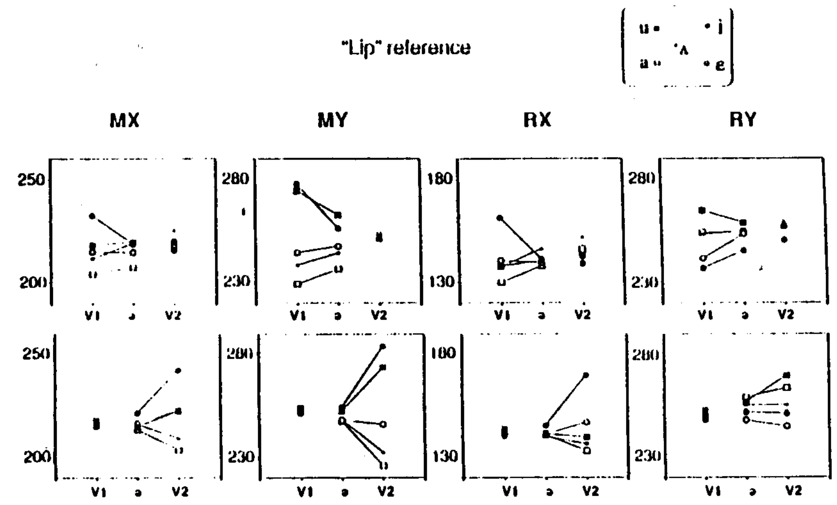

Figure 2.11 Relation between full vowel pellet positions and "lip" pellet positions for schwa. The top row displays the pellet positions for utterances with the indicated initial vowels, averaged across five utterances (each with a different final vowel). The bottom row displays the averaged pellet positions for utterances with the indicated final vowels. Units are X-ray units (= 0.33 mm)

complete temporal overlap of this gesture and the gesture for the following vowel. The blending caused by this overlap should yield the V2 effect on schwa, while the V1 effects should emerge as a passive consequence of the differing initial conditions for movements out of different preceding vowels.

An example of this type of organization is shown in figure 2.12, which is the gestural score we hypothesized for the utterance /pipə'papə/. As in figure 2.3, each box indicates the activation interval of a particular gestural control regime, that is, an interval of time during which the behavior of the particular tract variable is controlled by a second-order dynamical system with a fixed "target" (equilibrium position), frequency, and damping. The height of the box represents the tract-variable "target." Four LA closure-and-release gestures are shown, corresponding to the four consonants. The closure-and-release components of these gestures are shown as separate boxes, with the closure components having the smaller target for LA, i.e., smaller interlip distance. In addition, four tongue-body gestures are shown, one for each of the vowels V1, schwa, V2, schwa. Each of these gestures involves simultaneous activation of two tongue-body tract variables, one for constriction location and one for constriction degree. The control regimes for the V1 and medial schwa gestures are contiguous and nonoverlapping, whereas the V2 gesture begins at the same point as the medial schwa and thus completely

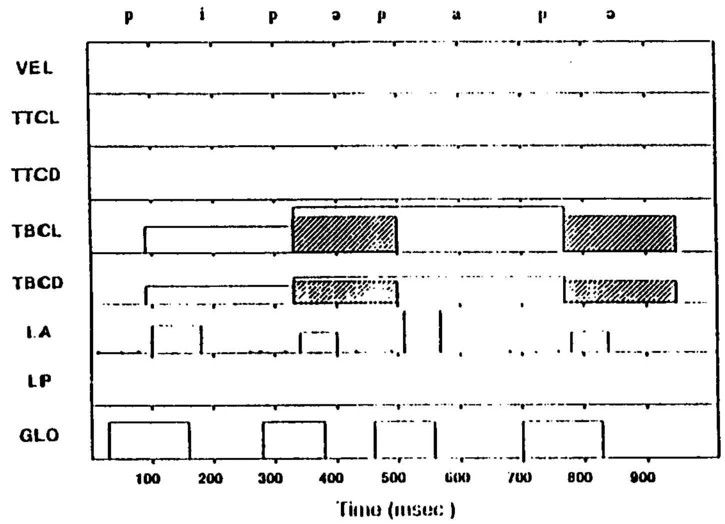

Figure 2.12 Gestural score for /pipə'papə/. Tract variable channels displayed, from top to bottom, are: velum, tongue-tip·constriction location and constriction degree, tongue-body constriction location and constriction degree, lip aperture, lip protrusion, and glottis. Horizontal extent of each box indicates duration of gestural activation; the shaded boxes indicate activation for schwa. For constriction-degree tract variables (VEL, TTCD, TBCD, LA, GLO), the higher the top of the box, the greater the amount of opening (aperture). The constriction-location tract variables (TTCL, TBCL) are defined in terms of angular position along the curved vocal tract surface. The higher the top of the box, the greater the angle, and further back and down (towards the pharynx) the constriction

overlaps it. In other words, during the acoustic realization of the schwa (approximately), the schwa and V2 gestural control regimes both control the tongue movements; the schwa relinquishes active control during the following consonant, leaving only the V2 tongue gesture active in the next syllable. While the postulation of an explicit schwa gesture overlapped by V2 was motivated by the particular results of section 2.2, the general layout of gestures in these utterances (their durations and overlap) was based on stiffness and phasing principles embodied in the linguistic model (Browman and Goldstein, 1990).

Gestural scores for each of the twenty-five utterances were produced. The activation intervals were identical in all cases; the scores differed only in the TBCL and TBCD target parameters for the different vowels. Targets used for the full vowels were those in our tract-variable dictionary. For the schwa, the target values (for TBCL and TBCD) were calculated as the mean of the targets for the five full vowels. The gestural scores were input to the task-

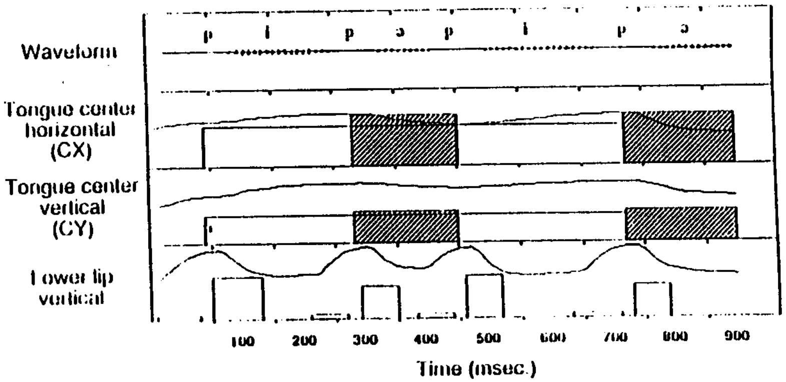

Figure 2.13 Gestural score for /pipə'pipə/. Generated movements (curves) are shown for the tongue center and lower lip. The higher the curve, the higher (vertical) or more fronted (horizontal) the corresponding movement. Boxes indicate gestural activation; the shaded boxes indicate activation for schwa. CX is superimposed on TBCL, CY on TBCD, lower lip on LA. Note that the boxes indicate the degree of opening and angular position of the constriction (as described in figure 2.12), rather than the vertical and horizontal displacement of articulators, as shown in the curves

dynamic model (Saltzman 1986), producing motions of the model articulators of the articulatory synthesizer (see figure 2.1). For example, for utterance /pipə'pipə/, figure 2.13 shows the resulting motions (with respect to a fixed reference on the head) of two of the articulators – the center of the tongue-body circle (C), and the lower lip, superimposed on the gestural score. Motion of the tongue is shown in both horizontal and vertical dimensions, while only vertical motion of the lower lip is shown. Note that the lower lip moves up for lip closure (during the regimes with the small LA value). Figure 2.14 shows the results for /pipə'papə/.

The articulator motions in the simulations can be compared to those of the data in the previous section (figures 2.4 and 2.5). One difference between the model and the data stems from the fact that the major portion of the tongue dorsum is modeled as an arc of circle, and therefore all points on this part of the dorsum move together. Thus, it is not possible to model the differential patterns of motion exhibited by the middle (M) and rear (R) of the dorsum in the X-ray data. In general, the motion of CX is qualitatively similar to both MX and RX (which, recall, are highly correlated). For example, both the data and the simulation show a small backward movement for the schwa in /pipə'pipə/; in /pipə'papə/, both show a larger backwards movement for schwa, with the target for /a/ reached shortly thereafter, early in the acoustic realization of V2. The motion of CY in the simulations tends to be similar to that of MY in the data. For example, in /pipə'papə/, CY moves down from

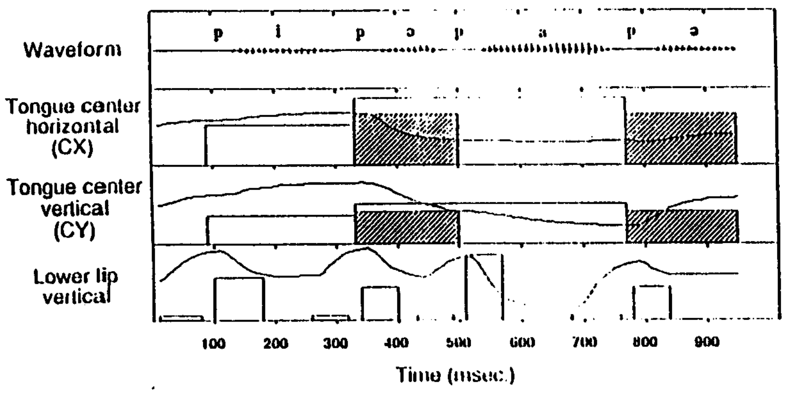

Figure 2.14 Gestural score for /pipə'papə/. Generated movements (curves) are shown for the tongue center and lower lip. The higher the curve, the higher (vertical) or more fronted (horizontal) the corresponding movement. Boxes indicate gestural activation; the shaded boxes indicate activation for schwa. Superimposition of boxes and curves as in figure 2.13

/i/ to schwa to /a/, and the target for /a/ tends to be achieved relatively late, compared to CX. Movemenťs corresponding to RY motions are not found in the displacement of the tongue-body circle, but would probably be reflected by a point on the part of the model tongue's surface that is further back than that section lying on the arc of a circle.

The model articulator motions were analyzed in the same manner as the X-ray data, once the time points for measurement were determined. Since for the X-ray data we assumed that displacement extrema indicated the effective target for the gesture, we chose the effective targets in the simulated data as the points to measure. Thus, points during V1 and V2 were chosen that corresponded to the point at which the vowel gestures (approximately) reached their targets and were turned off (right-hand edges of the tongue boxes in figures 2.13 and 2.14). For schwa, the "tongue" reference point was chosen at the point where the schwa gesture was turned off, while the "lip" reference was chosen at the lowest point of the lip during schwa (the same criterion as for the X-ray data).

The distribution of the model full vowels in the mid-sagittal plane (CX × CY) is shown in figure 2.15. Since the vowel gestures are turned off only after they come very close to their targets, there is very little variation across the ten tokens of each vowel. The distribution of schwa at the "tongue" reference point is shown in figure 2.16, labeled by the identity of V2 (in a) and V1 (in b), with the full vowel ellipses added for comparison. At this reference point that occurs relatively late, the vowels are clustered almost completely by V2, and the tongue center has moved a substantial portion of the way towards the

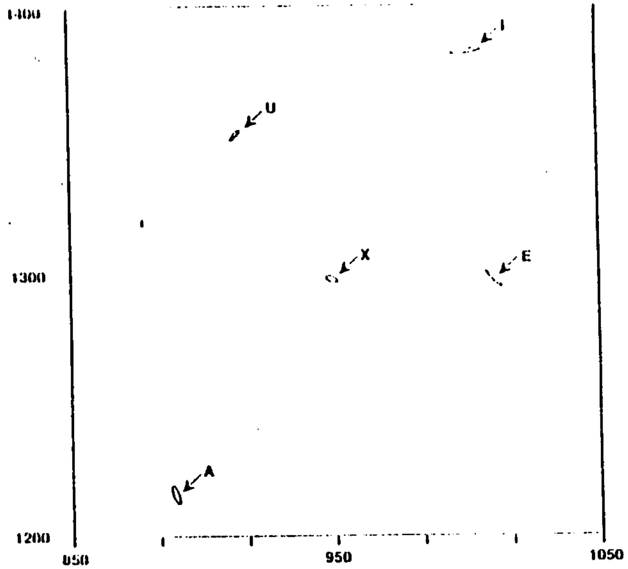

Figure 2.15 Tongue-center (C) positions for model full vowels, displayed in mid-sagittal plane with head facing to the right. The ellipses indicate two standard deviations along axes determined by principal-component analysis. Symbols I = IPA /i/, U = /u/, E = /ε/, X = /ʌ/, and A = /a/. Units are ASY units (= 0.09 mm), that is, units in the vocal tract model, measured with respect to the fixed structures.

following full vowel. The distribution of schwa values at the "lip" reference point is shown in figure 2.17, labeled by the identity of V2 (in a), and of V1 (in b). Comparing figure 2.17(a) with figure 2.16(a), we can see that there is considerably more scatter at the "lip" point than at the later "tongue" point.

We tested whether the simulations captured the regularities of the X-ray data by running the same set of regression analyses on the simulations as were performed on the X-ray data. The results are shown in table 2.2, which has the same format as the X-ray data results in table 2.1. Similar patterns are found for the simulations as for the data. At the "tongue" reference point, for both CX and CY the best two-term prediction involves the schwa component (constant) and V2, and this prediction is nearly as good as that using all three terms. Recall that this was the case for all pellet dimensions except for MY, whose differences were attributed to differences in the time point at which this dimension was measured. (In the simulations, CX and CY were always measured at the same point in time.) These results can be seen graphically in figure 2.18, where the top row of panels shows the relation between V1 and schwa, and the bottom row shows the relation between V2

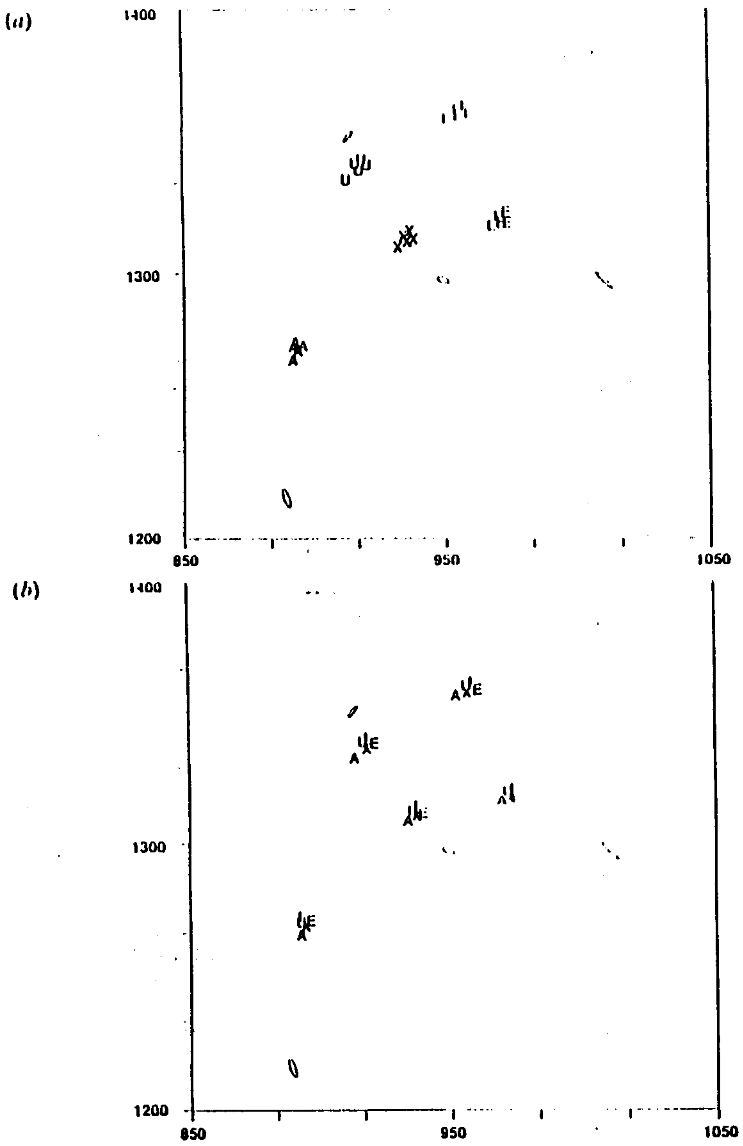

Figure 2.16 Tongue-center (C) positions for model schwa at "tongue" reference points, displayed in right-facing mid-sagittal plane as in figure 2.15. The ellipses are from the model full vowels (figure 2.15), for comparison. Symbols I = IPA /i/, U = /u/, E = /ε/, X = /ʌ/, and A = /a/. Units are ASY units (= 0.09 mm). (a) Model schwa positions labeled by the identity of the following vowel (V2). (b) Model schwa positions labeled by the identity of the preceding vowel (V1)

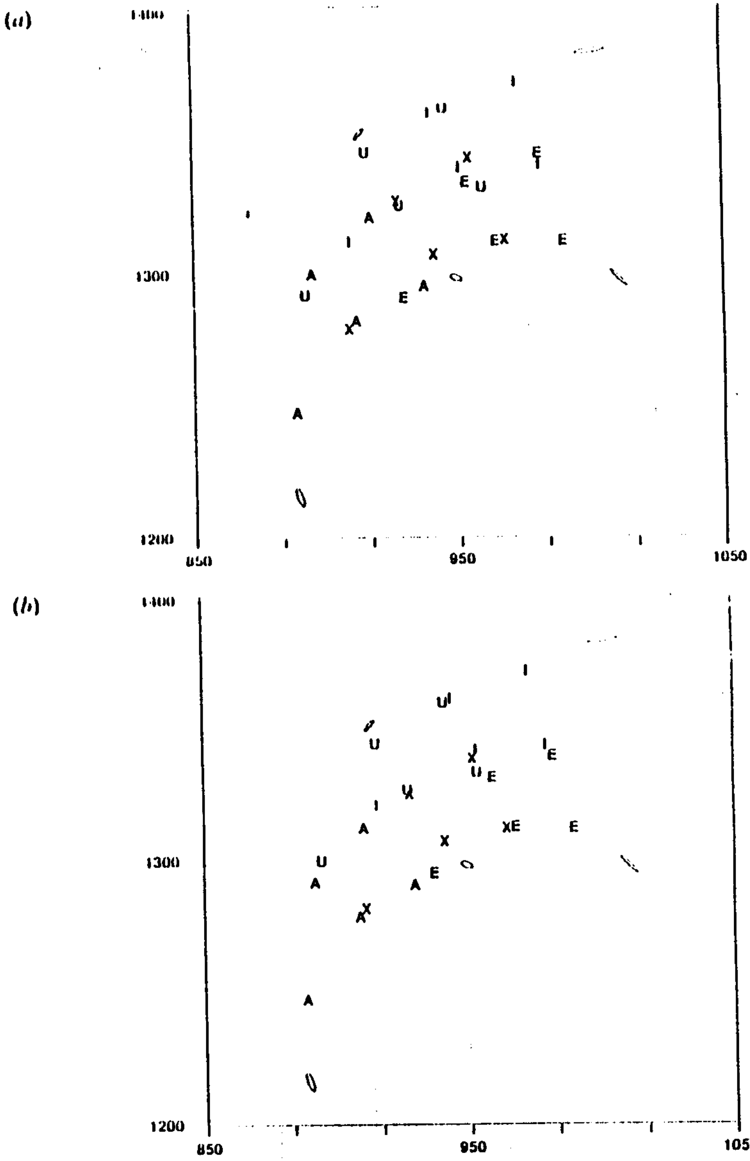

Figure 2.17 Tongue-center (C) positions for model schwa at "lip" reference points, displayed as in figure 2.16 (including ellipses from figure 2.15). Symbols I = IPA /i/, U = /u/, E = /ε/, X = /ʌ/, and A = /a/. Units are ASY units (= 0.09 mm). (a) Model schwa positions labeled by the identity of the following vowel (V2). (b) Model schwa positions labeled by the identity of the preceding vowel (V1)

Table 2.2 *Regression results of simulations*

	"Lip" reference point			"Tongue" reference point	
	Terms	*Standard error*		*Terms*	*Standard error*
CX	k + v1 + v2	7.5	CX	k + v2 + v1	5.2
	v1 + v2	9.1		k + v2	5.6
	k + v1	20.4		v2 + v1	13.2
	k	29.2		v2	20.0
	v1	33.3		k	28.6
CY	k + v1 + v2	4.7	CY	k + v2 + v1	3.2
	v1 + v2	13.5		k + v2	3.9
	k + v1	19.7		v2 + v1	18.8
	k	29.1		v2	28.0
	v1	41.6		k	30.4

and schwa. The same systematic relation between schwa and V2 can be seen in the bottom row as in figure 2.10 for the X-ray data, that is, no crossover. (The lack of systematic relations between V1 and schwa in the X-ray data, indicated by the cross-overs in the top row of figure 2.10, is captured in the simulations in figure 2.18 by the lack of variation for the schwa in the top row.) Thus, the simulations capture the major statistical relation between the schwa and the surrounding full vowels at the "tongue" reference point, although the patterns are more extreme in the simulations than in the data.

At the earlier "lip" reference point, the simulations also capture the patterns shown by the data. For both CX and CY, the three-term predictions in table 2.2 show substantially less error than the best two-term prediction. This was also the case for the data in table 2.1 (except for RX), where V1, V2 and a schwa component (constant) all contributed to the prediction of the value during schwa. This can also be seen in the graphs in figure 2.19, which shows a systematic relationship with schwa for both V1 and V2.

In summary, for the simulations, just as for the X-ray data, V1 contributed to the pellet position at the "lip" reference, but not to the pellet position at the "tongue" point, while V2 and an independent schwa component contributed at both points. Thus, our hypothesized gestural structure accounts for the major regularities observed in the data (although not for all aspects of the data, such as its noisiness or differential behavior among pellets). The gestural-control regime for V2 begins simultaneously with that for schwa and overlaps it throughout its active interval. This accounts for the fact that V2 and schwa effects can be observed throughout the schwa, as both gestures unfold together. However, V1 effects are passive consequences of the initial conditions when the schwa and V2 gestures are "turned on," and thus, their effects disappear as the tongue position is attracted to the "target" (equilibrium position) associated with the schwa and V2 regimes.

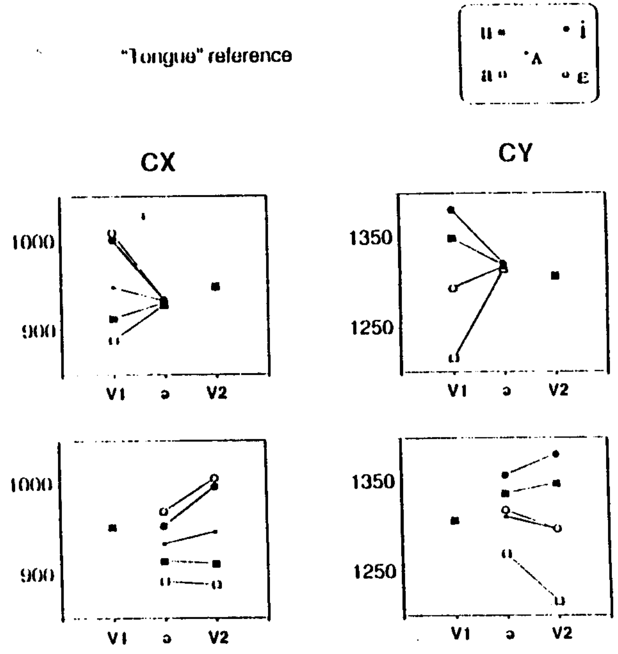

Figure 2.18 Relation between model full vowel tongue-center positions and tongue-center positions at "tongue" reference point for model schwas. The top row displays the tongue-center positions for utterances with the indicated initial vowels, averaged across five utterances (each with a different final vowel). The bottom row displays the averaged tongue-center positions for utterances with the indicated final vowels. Units are ASY units (= 0.09 mm)

2.3.1 *Other simulations*

While the X-ray data from the subject analyzed here argue against the strongest form of the hypothesis that schwa has no tongue target, we decided nevertheless to perform two sets of simulations incorporating the strong form of an "unspecified" schwa to see exactly where and how they would fail to reproduce the subject's data. In addition, if the synthesized speech were found to be correctly perceived by listeners, it would suggest that this gestural organization is at least a possible one for these utterances, and might be found for some speakers. In the first set of simulations, one of which is

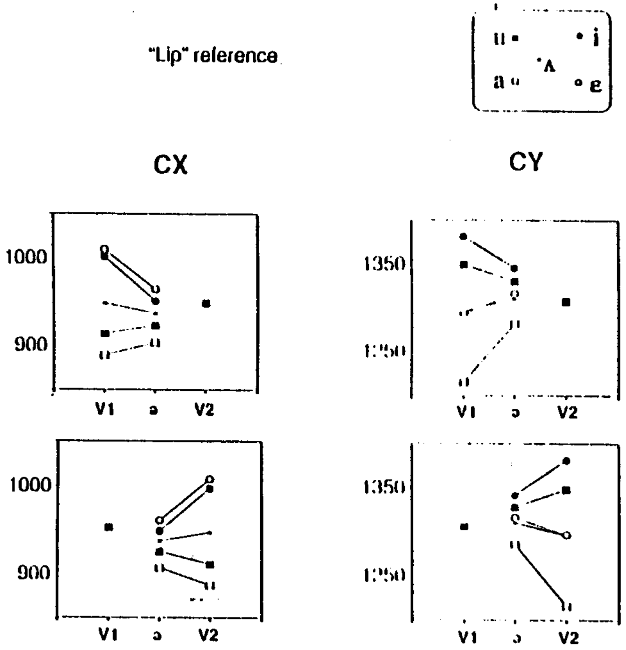

Figure 2.19 Relation between model full vowel tongue-center positions and tongue-center positions at "lip" reference point for model schwas. The top row displays the tongue-center positions for utterances with the indicated initial vowels, averaged across five utterances (each with a different final vowel). The bottom row displays the averaged tongue-center positions for utterances with the indicated final vowels. Units are ASY units (= 0.09 mm)

exemplified in figure 2.20, the gestural scores took the same form as in figure 2.12, except that the schwa tongue-body gestures were removed. Thus, active control of V2 began at the end of V1, and, without a schwa gesture, the tongue trajectory moved directly from V1 to V2. During the acoustic interval corresponding to schwa, the tongue moved along this V1–V2 trajectory. The resulting simulations in most cases showed a good qualitative fit to the data, and produced utterances whose medial vowels were perceived as schwas. The problems arose in utterances in which V1 and V2 were the same (particularly when they were high vowels). Figure 2.20 portrays the simulation for /pipə'pipə/: the motion variables generated can be compared with the data in figure 2.4. The "dip" between V1 and V2 was not produced in the simulation, and, in addition, the medial vowel sounded like /i/ rather than schwa. This organization does not, then, seem possible for utterances where both V1 and V2 are high vowels.

We investigated the worst utterance (/pipə'pipə/) from the above set of simulations further, generating a shorter acoustic interval for the second vowel (the putative schwa) by decreasing the interval (relative phasing) between the bilabial gestures on either side of it. An example of a score with the bilabial closure gestures closer together is shown in figure 2.21. At relatively short durations as in the figure (roughly < 50 msec.), the percept of the second vowel changed from /i/ to schwa. Thus, the completely targetless organization may be workable in cases where the surrounding consonants are only slightly separated. In fact, this suggests a possible historical source for epenthetic schwa vowels that break up heterosyllabic clusters. They could arise from speakers increasing the distance between the cluster consonants slightly, until they no longer overlap. At that point, our simulations suggest that the resulting structure would be perceived as including a schwa-like vowel.

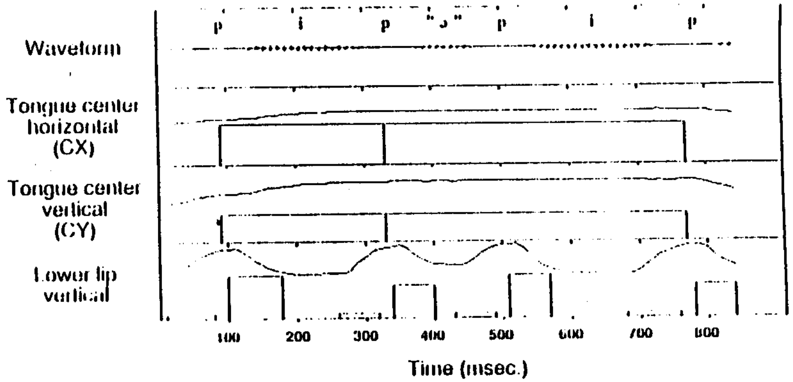

Figure 2.20 Gestural score plus generated movements for /pip_'pip_/, with no activations for schwa. The acoustic interval between the second and third bilabial gestures is perceived as an /i/. Generated movements (curves) are shown for the tongue center and lower lip. The higher the curve, the higher (vertical) or more fronted (horizontal) the corresponding movement. Superimposition of boxes and curves as in figure 2.13

The second set of simulations involving an "unspecified" schwa used the same gestural organization as that portrayed in the score in figure 2.20, except that the V2 gesture was delayed so that it did not begin right at the offset of the V1 gesture. Rather, the V2 regime began approximately at the beginning of the third bilabial-closure gesture, as in figure 2.22. Thus, there was an interval of time during which no tongue-body gesture was active, that is, during which there was no active control of the tongue-body tract variables. The motion of the tongue-body center during this interval, then, was determined solely by the neutral positions, relative to the jaw, associated with the tongue body articulators, and by the motion of the jaw, which was implicated in the ongoing bilabial closure and release gestures. The results,

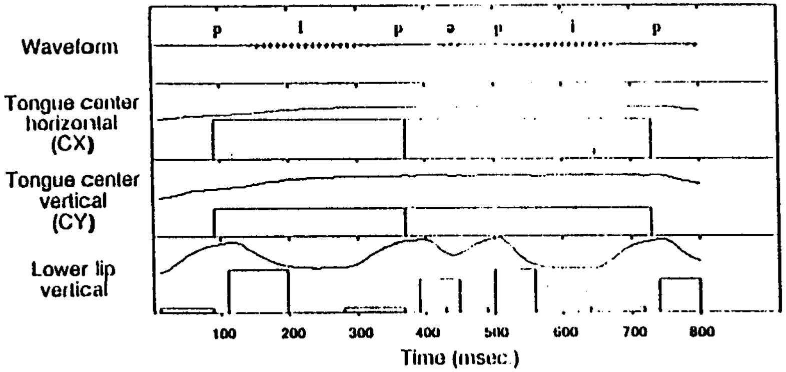

Figure 2.21 The same gestural score for /pip_'pip_/ as in figure 2.20, but with the second and third bilabial gestures closer together than in figure 2.20. The acoustic interval between the second and third bilabial gestures is perceived as a schwa. Generated movements (curves) are shown for the tongue center and lower lip. The higher the curve, the higher (vertical) or more fronted (horizontal) the corresponding movement. Superimposition of boxes and curves as in figure 2.13

displayed in figure 2.22, showed that the previous problem with /pipə'pipə/ was solved, since during the unspecified interval between the two full vowels, the tongue-body lowered (from /i/ position) and produced a perceptible schwa. Unfortunately, this "dip" between V1 and V2 was seen for all combinations of V1 and V2, which was not the case in the X-ray data. For example, this dip can be seen for /papə'papə/ in figure 2.23; in the X-ray data, however, the tongue raised slightly during the schwa, rather than lowering. (The "dip" occurred in all the simulations because the neutral position contributing to the tongue-body movement was that of the tongue-body articulators rather than that of the tongue-body tract variables; consequently the dip was relative to the jaw, which, in turn, was lowering as part of the labial release). In addition, because the onset for V2 was so late, it would not be possible for V2 to affect the schwa at the "lip" reference point, as was observed in the X-ray data. Thus, this hypothesis also failed to capture important aspects of the data. The best hypothesis remains the one tested first – where schwa has a target of sorts, but is still "colorless," in that its target is the mean of all the vowels, and is completely overlapped by the following vowel.

2.4 Conclusion

We have demonstrated how an explicit gestural model of phonetic structure, embodying the possibilities of underspecification ("targetlessness") and

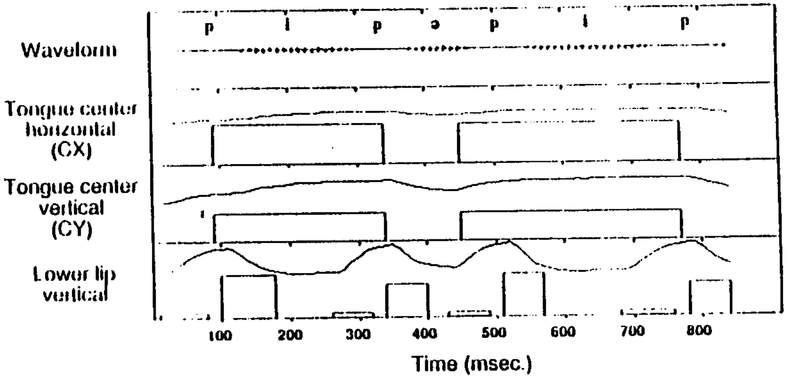

Figure 2.22 The same gestural score for /pip_'pip_/ as in figure 2.20, but with the onset of the second full vowel /i/ delayed. The acoustic interval between the second and third bilabial gestures is perceived as a schwa. Generated movements (curves) are shown for the tongue center and lower lip. The higher the curve, the higher (vertical) or more fronted (horizontal) the corresponding movement. Superimposition of boxes and curves as in figure 2.13

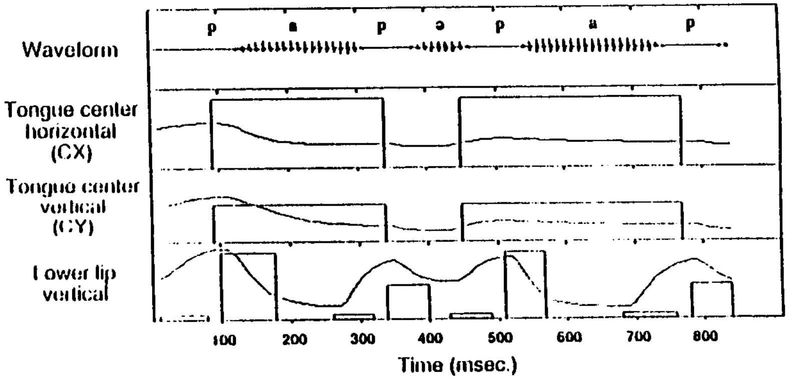

Figure 2.23 The same gestural score as in figure 2.22, except with tongue targets appropriate for the utterance /pap_'pap_/. The acoustic interval between the second and third bilabial gestures is perceived as a schwa. Generated movements (curves) are shown for the tongue center and lower lip. The higher the curve, the higher (vertical) or more fronted (horizontal) the corresponding movement. Superimposition of boxes and curves as in figure 2.13

temporal overlap ("coproduction"), can be used to investigate the contextual variation of phonetic units, such as schwa, in speech. For the particular speaker and utterances that we analyzed, there was clearly some warping of the V1 V2 trajectory towards a neutral position for an intervening schwa. The analyses showed that this neutral position has to be defined in the space of tract variables (the linguistically relevant goal space), rather than being the consequence of neutral positions for individual articulators. Therefore, a target position for schwa was specified, although this target is completely predictable from the rest of the system; it corresponds to the mean tongue tract-variable position for all the full vowels.

The temporally overlapping structure of the gestural score played a key role in accounting for the time course of V1 and V2 effects on schwa. These effects were well modeled by a gestural score in which active control for schwa was completely overlapped by that for V2. This overlap gave rise to the observed anticipatory effects, while the carry-over effects were passive consequences of the initial conditions of the articulators when schwa and V2 begin. (This fits well with studies that have shown qualitative asymmetries in the nature of carry-over and anticipatory effects [see Recasens 1987].)

How well the details of the gestural score will generalize to other speakers and other prosodic contexts remains to be investigated. There is known to be much individual variation in the strength of anticipatory vs. carry-over coarticulation in utterances like those employed here, and also in the effect of stress (Fowler 1981a; Magen 1989). In addition, reduced vowels with different phonological/morphological characteristics, as in the plural (e.g. "roses") and past tense (e.g. "budded") may show different behavior, either with respect to overlap or targetlessness. The kind of modeling developed here provides a way of analyzing the complex quantitative data of articulation so that phonological issues such as these can be addressed.

Comments on Chapter 2

SARAH HAWKINS

The speaker's task is traditionally conceptualized as one of producing successive articulatory or acoustic targets, with the transitions between them being planned as part of the production process.* A major goal of studies of coarticulation is then to identify the factors that allow or prevent coarticulatory spread of features, and so influence whether or not targets are reached. In contrast, Browman and Goldstein offer a model of phonology that is couched in gestural terms, where gestures are abstractions rather than movement trajectories. In their model, coarticulation is the inevitable consequence of coproduction of articulatory gestures. Coarticulation is planned only in the sense that the gestural score is planned, and traditional notions of target modification, intertarget smoothing, and look-ahead processes are irrelevant as explanations, although the observed properties they are intended to explain are still, of course, of central concern.

*The structure of this discussion is influenced by the fact that it originally formed part of a joint commentary covering this paper and the paper by Hewlett and Shockey. Since the latter's paper was subsequently considerably revised, mention of it has been removed and a separate discussion prepared.

Similarly, coarticulation is traditionally seen as a task of balancing constraints imposed by the motoric system and the perceptual system – of balancing ease of articulation with the listener's need for acoustic clarity. These two opposing needs must be balanced within constraints imposed by a third factor, the phonology of the particular language. Work on coarticulation often tries to distinguish these three types of constraint.

For me, one of the exciting things about Browman and Goldstein's work is that they are being so successful in linking, as opposed to separating, motoric, perceptual, and phonological constraints. In their approach, the motoric constraints are all accounted for by the characteristics of the task-dynamic model. But the task-dynamic model is much more than an expression of universal biomechanical constraints. Crucially, the task-dynamic model also organizes the coordinative structures. These are flexible, functional groupings of articulators whose organization is not an inevitable process of maturation, but must be learned by every child. Coordinative structures involve universal properties and probably some language-specific properties. Although Browman and Goldstein assign all language-specific information to the gestural score, I suspect that the sort of things that are hard to unlearn, like native accent and perhaps articulatory setting, may be better modeled as part of the coordinative structures within the task dynamics. Thus the phonological constraints reside primarily in the gestural score, but also in its implementation in the task-dynamic model.

Browman and Goldstein are less explicitly concerned with modeling perceptual constraints than phonological and motoric ones, but they are, of course, concerned with what the output of their system sounds like. Hence perceptual constraints dictate much of the organization of the gestural score. The limits set on the temporal relationships between components of the gestural score for any given utterance represent in part the perceptual constraints. Variation in temporal overlap of gestures within these limits will affect how the speech sounds. But the amount of variation possible in the gestural score must also be governed by the properties and limits on performance of the parameters in the task-dynamic model, for it is the task-dynamic model that limits the rate at which each gesture can be realized. So the perceptual system and the task-dynamic model can be regarded as in principle imposing limits on possible choices in temporal variation, as represented in the gestural score. (In practice, these limits are determined from measurement of movement data.) Greater overlap will result in greater measurable coarticulation; too little or too much overlap might sound like some dysarthric or hearing-impaired speakers. Browman and Goldstein's work on schwa is a good demonstration of the importance of appropriate temporal alignment of gestures. It also demonstrates the importance to acceptable speech production of getting the right relationships between the gestural targets and their temporal coordination.

Thus Browman and Goldstein offer a model in which perception and production, and universal and language-specific aspects of the phonology, are conceptually distinguishable yet interwoven in practice. This, to my way of thinking, is as it should be.

The crucial issue in work on coarticulation, however, is not so much to say what constraints affect which processes, as to consider what the controlled variables are. Browman and Goldstein model the most fundamental controlled variables: tongue constriction, lip aperture, velar constriction, and so on. There are likely to be others. Some, like fundamental frequency, are not strongly associated with coarticulation but are basic to phonology and phonetics, and some, like aerodynamic variables, are very complex.

Let us consider an example from aerodynamics. Westbury (1983) has shown allophonic differences in voiced stops that depend on position in utterance and that all achieve cavity enlargement to maintain voicing. The details of what happens vary widely and depend upon the place of articulation of the stop, and its phonetic context. For example, for initial /b/, the larynx is lowered, the tongue root moves backwards, and the tongue dorsum and tip both move down. For final /b/, the larynx height does not change, the tongue root moves forward, and the dorsum and tip move slightly upwards. In addition, the rate of cavity enlargement, and the time function, also vary between contexts. Does it make sense to try to include these differences? If the task-dynamic system is primarily universal, then details of the sort Westbury has shown are likely to be in the gestural score. But to include them would make the score very complicated. Do we want that much detail in the phonology, and if so, how should it be included? Browman and Goldstein have elsewhere (1990) suggested a tiered system, and if that solution is pursued, we could lose much of the distinction between phonetics and phonology. While I can see many advantages in losing that distinction, we could, on the other hand, end up with a gestural score of such detail that some of the things phonologists want to do might become undesirably clumsy. The description of phonological alternations is a case in point.

So to incorporate these extra details, we will need to consider the structure and function of the gestural score very carefully. This will include consideration of whether the gestural score really *is* the phonology phonetics, or whether it is the interface between them. In other words, do we see in the gestural score the phonological primitives, or their output? Browman and Goldstein say it is the former. I believe they are right to stick to their strong hypothesis now, even though it may need to be modified later.

Another issue that interests me in Browman and Goldstein's model is variability. As they note, the values of schwa that they produce are much less variable than in real speech. There are a number of ways that variability could be introduced. One, for schwa in particular, is that its target should not be the simple average of all the vowels in the language, as Browman and Goldstein suggest, but rather a weighted average, with higher weighting given to the immediately preceding speech. How long this preceding domain might be I do not know, but its length may depend on the variety of the preceding articulations. Since schwa is schwa basically because it is centralized relative to its context, schwa following a lot of high articulations could be different from schwa in the same immediate context but following a mixture of low and high articulations.

A second possibility, not specific to schwa, is to introduce errors. The model will ultimately need a process that generates errors in order to produce real-speech phenomena like spoonerisms. Perhaps the same type of system could produce articulatory slop, although I think this is rather unlikely.

If the variability we are seeking for schwa is a type of articulatory slop, it could also be produced by variability in the temporal domain. In Browman and Goldstein's terms, the phase relations between gestures may be less tightly tied together than at present.

A fourth possibility is that the targets in the gestural score could be less precisely specified. Some notion of acceptable range might add the desired variability. This idea is like Keating's (1988a) windows, except that her windows determine an articulator trajectory, whereas Browman and Goldstein's targets are realized via the task-dynamic model, which adds its own characteristics.

Let me finish by saying that one of the nice things about Browman and Goldstein's work is how much it tells us that we know already. Finding out what we already know is something researchers usually hope to avoid. But in this case we "know" a great number of facts of acoustics, movement, and phonology, but we do not know how they fit together. Browman and Goldstein's observations on intrusive schwa, for example, fit with my own on children's speech (Hawkins 1984: 345). To provide links between disparate observations seems to me to achieve a degree of insight that we sorely need in this field.

Comments on Chapter 2

JOHN KINGSTON

Introduction

Models are valued more for what they predict, particularly what they predict not to occur, than what they describe. While the capacity of Browman and Goldstein's gestural model to *describe* articulatory events has been demonstrated in a variety of papers (see Browman *et al.* 1984; Browman and Goldstein 1985, 1986, 1990; Browman *et al.* 1986), and there is every reason to hope that it will continue to achieve descriptive success, I am less sanguine about its predictive potential. The foundation of my pessimism is that gestural scores are not thus far constructed in terms of independent principles which would motivate some patterns of gestural occurrence and coordination, while excluding others.

"Independent principles" are either such as constrain nonspeech and speech movement alike, or such as arise from the listener's demands on the speaker. That such principles originate outside the narrowly construed events of speaking themselves guards models built on them from being hamstrung by the *ad hoc* peculiarities of speech movements. The scores' content is constrained by the limited repertoire of gestures used, but because gestures' magnitude may be reduced in casual speech, even to the point of deletion (Browman and Goldstein 1990), the variety of gestures in actual scores is indefinitely large. Further constraints on the interpretation of scores come from the task dynamics, which are governed by principles that constrain other classes of movements (see Kelso *et al.* 1980; Nelson 1983; Ostry, Keller, and Parush 1983; Saltzman and Kelso 1987). The task dynamics rather than the gestural score also specify which articulatory movements will produce a particular gesture. The gestural score thus represents the model's articulatory goals, while the specific paths to these goals are determined by entirely dynamical means. Gestural coordination is not, however, constrained by the task dynamics and so must be stipulated, and again the number of possible patterns is indefinitely large. Despite this indefiniteness in the content and coordination of scores, examining the articulation of schwa should be informative about what is in a score and how the gestures are coordinated, even if in the end Browman and Goldstein's account does not extend beyond description.

The next two sections of this commentary examine Browman and Goldstein's claim that English schwa has an articulation of its own. This examination is based on an extension of their statistical analysis and leads to a partial rejection of their claim. In the final section, the distinction between predictive vs. descriptive models is taken up again.

Table 2.3 *Variances for lip and tongue reference positions*

	MX	MY	RX	RY
Lip	910.2	2290.7	364.4	1111.2
Tongue	558.6	959.0	325.9	843.4

Does schwa have its own target?

Browman and Goldstein found that the positions of two tongue pellets (MX-MY and RX-RY) in a schwa between flanking full vowels closely match the grand mean of pellet positions in the full vowels, implying that during schwa the tongue simply has whatever position is dictated by the transition between the full vowels that flank it. Therefore, when flanking vowels are identical, the tongue should not deviate from the full vowel positions during schwa. However, Browman and Goldstein's data show the tongue does move away from these positions and back again during the schwa, implying it does have its own target.

Giving schwa its own target is supported by the stepwise regression in which including a constant factor representing effects independent of either of the flanking vowels yielded a smaller residual variance. Schwa's target only looks transitional because it is very close to the grand mean of the tongue positions of all the full vowels. The stepwise regression also revealed that the tongue position during schwa was determined more by V2 than V1, perhaps because V2 was more prominent than V1.

Influences on the tongue position for schwa were assessed by comparing standard errors for multiple-regression models containing different combinations of terms for V1, and V2, and k, the independent schwa factor. "Standard error" is the standard error of estimate (*SE*), a measure of the residual variance not accounted for by the terms in the regression models. The standard error of estimate is the term on the right, the square root of the residual mean square, in (1) (Cohen and Cohen 1983: 104);

$$(sd^2_{Y-\hat{Y}})^{1/2} = \left(\frac{(1 - R^2)\Sigma(Y - \bar{Y})^2}{(n - q - 1)}\right)^{1/2}$$

(1) Formula for the standard error of estimate

(q is the number of terms in the regression model) which shows that *SE*'s magnitude is not only a function of the proportion of variance not accounted for, $1 - R^2$, but also of the overall magnitude of variance in the dependent measure, $\Sigma(Y - \bar{Y})^2$. Since the magnitude of this latter variance will differ

Table 2.4 *Shrunken R^2s for lip reference positions*

	k+V1+V2	k+V1	k+V2	V1+V2	k	V1	V2
MX	0.879	0.848		0.846	0.818	0.752	
MY	0.957	0.945		0.917	0.882	0.864	
RX	0.750		0.731	0.517	0.654		0.157
RY	0.912		0.885	0.880	0.839		0.834

Table 2.5 *Shrunken R^2s for tongue reference positions*

	k+V1+V2	k+V1	k+V2	V1+V2	k	V1	V2
MX	0.806		0.793	0.712	0.709		0.491
MY	0.882	0.832		0.761	0.761	0.586	
RX	0.631		0.631	0.406	0.533		0.145
RY	0.872		0.863	0.842	0.778		0.778

between dependent variables, the absolute magnitude of the *SE*s for models with different dependent variables cannot be compared. Accordingly, to evaluate how well the various regression models fare across the pellet positions, a measure of variance should be used that is independent of the effect of different variances among the dependent variables, i.e. R^2, rather than *SE*. More to the point, the R^2s can be employed in significance tests of differences between models of the same dependent variable with different numbers of terms. The equation in (1) can be solved for R^2, but only if one knows $\Sigma(Y - \bar{Y})^2$ (solving this equation for R^2 shows that the values listed by Browman and Goldstein must be the squared standard error of estimate). This variance was obtained from Browman and Goldstein's figures, which plot the four tongue coordinates; measurements were to the nearest division along each axis, and their precision is thus ±1 mm for MY and RY, ±0.55 mm for RX, and ±0.625 mm for MX (measurement error in either direction is roughly equal to half a division for each of the pellet coordinates). The variances obtained do differ substantially for the four measures (see table 2.3), with the variances for vertical position consistently larger than for horizontal position at both reference points. The resulting shrunken R^2s for the various regression models at lip and tongue reference positions are shown in tables 2.4 and 2.5 (the gaps in these tables are for regression models not considered by the stepwise procedure). Shrunken R^2s are given in these tables because they are a better estimate of the proportion of variance accounted for in the *population* from which the *sample* is taken when the ratio of independent variables q to n is large, as here, and when independent variables are selected *post hoc*, as in the stepwise regression. (Shrunken R^2s were calculated according to formula (3.6.4) in Cohen and Cohen (1983: 106–7), in which q was always the total number of independent variables from which selections were made by the stepwise procedure, i.e. 3.) The various models were tested for whether adding a term to the equation significantly increased the variance, assuming Model I error (see Cohen and Cohen 1983: 145 7). Comparisons were made of k + V1 + V2 with k + V1 or k + V2 and with V1 + V2.

The resulting *F*-statistics confirmed Browman and Goldstein's contention that adding V1 to the k + V2 model does not increment the variance significantly at MX, RX, or RY at the tongue reference positions (for k + V1 + V2 vs. k + V2: MX $F_{(2,19)} = 0.637, p > 0.05$; RX $F_{(2,19)} = 0, p > 0.05$; and RY $F_{(2,19)} = 0.668$, $p > 0.05$) and also supports their observation that both V1 and V2 increment R^2 substantially for MY at the tongue reference positions (for k + V1 + V2 vs. k + V1, $F_{(2,19)} = 4.025$, $p < 0.05$).

However, their claim that for the lip reference position, the two-term models k + V1 or k + V2 account for substantially less variance than the three-term model k + V1 + V2 is not supported, for any dependent variable (for k + V1 + V2 vs. k + V1: MX $F_{(2,19)} = 2.434$, $p > 0.05$ and MY $F_{(2,19)} = 2.651$, $p > 0.05$, and for k + V1 + V2 vs. k + V2: RX $F_{(2,19)} = 0.722$, $p > 0.05$ and RY $F_{(2,19)} = 2.915$, $p > 0.05$). Comparisons of the other two-term model, V1 + V2, with k + V1 + V2 yielded significant differences for MY ($F_{(2,19)} = 8.837$, $p < 0.01$) and RX ($F_{(2,19)} = 8.854$, $p < 0.01$), but for neither dependent variable was V1 + V2 the second-best model. At MX and RY, the differences in amount of variance accounted for by V1 + V2 vs. k + V1 (MX) or k + V2 (RY) are very small (less than half of 1 percent in each case), so choosing the second-best model is impossible. In any case, there is no significant increment in the variance in the three-term model, k + V1 + V2, with respect to the V1 + V2 two-term model at MX ($F_{(2,19)} = 2.591$, $p > 0.05$) or RY ($F_{(2,19)} = 2.483$, $p > 0.05$). Thus at the lip reference position, schwa does not coarticulate strongly with V2 at MX or MY, nor does it coarticulate strongly with V1 at RX or RY. There is evidence for an independent schwa target at MY and RX, but not MX or RY. Use of R^2s rather than *SE*s to evaluate the regression models has thus weakened Browman and Goldstein's claims regarding both schwas having a target of its own and the extent to which it is coproduced with flanking vowels.

Are all schwas the same?

Whether schwa has an independent target depends on how much the schwa factor contributes to R^2 when the tokens with identical flanking vowels are omitted. If schwa's target is the grand mean of all the full vowel articulations, then the schwa factor should contribute substantially less with this omission, since the tongue should pass through that position on its path between differing but not identical full vowels. Schwas may be made in more than one way, however: between unlike vowels, schwa may simply be a transitional segment, but between like vowels, a return to a more neutral position might have to be achieved, either by passive recoil if schwas are analogous to the "trough" observed between segments which require some active articulatory gesture (Gay 1977, 1978; cf. Boyce 1986) or by means of an active gesture as Browman and Goldstein argue. (Given the critical damping of gestures in the task dynamics, one might expect passive recoil to achieve the desired result, but then why is a specified target needed for schwa?) On the other hand, if the schwa factor contributes nearly the same amount to R^2 in regression models where the identical vowel tokens are set aside, then there is much less need for multiple mechanisms. Finally, one may ask whether schwas are articulated with the same gesture when there is no flanking full vowel, on one side or the other, as in the first schwa of *Pamela* or the second schwa in *Tatamagouchi*.

A need more fundamental than looking at novel instances of the phenomenon is for principles external to the phenomena on which the modeling is based, which would explain why one gestural score is employed and not others. I point to where such external, explanatory principles may be found in the next section of this commentary.

Description vs. explanation

The difficulty I have with the tests of their gestural model that Browman and Goldstein present is that they stop when an adequate descriptive fit to the observed articulatory trajectories was obtained. Lacking is a theory which would *predict* the particular gestural scores that most closely matched the observed articulations, on general principles. (The lack of predictive capacity is, unfortunately, not a problem unique to Browman and Goldstein's model; for example, we now have excellent descriptive accounts of how downtrends in F_0 are achieved (Pierrehumbert 1980; Liberman and Pierrehumbert 1984; Pierrehumbert and Beckman 1988), but still very little idea of why downtrends are achieved with the mechanisms identified, why these mechanisms are employed in all languages with downtrends, or even why downtrends are so ubiquitous.) If V2's gesture overlaps more with the preceding schwa than V1's because it is more prominent, why should prominence have this effect on gestural overlap? On the other hand, if more overlap is always observed between the schwa and the following full vowel, why should anticipatory coarticulation be more extensive than carry-over? And in this case, what does greater anticipatory coarticulation indicate about the relationship between the organization of gestures and the trochaic structure of stress feet in English? All of these are questions that we might expect an explanatory or predictive theory of gestural coordination to answer.

The gestural theory developed by Browman and Goldstein may have all the pieces needed to construct a machine that will produce speech, indeed, it is already able to produce particular speech events, but as yet there is no general structure into which these pieces may be put which would produce just those kinds of speech events that do occur and none of those that do not. Browman and Goldstein's gestural theory is not incapable of incorporating general principles which would predict just those patterns of coordination that occur; the nature of such principles is hinted at by Kelso, Saltzman, and Tuller's (1986a) replication of Stetson's (1951) demonstration of a shift from a VC to CV pattern of articulatory coordination as rate increased. Kelso, Saltzman, and Tuller suggest that the shift reflects the greater stability of CV over VC coordination, but it could just as well be that place and perhaps other properties of consonants are more reliably perceived in the transition from C to V than from V to C (see Ohala 1990 and the references cited there, as well as Kingston 1990 for a different view). If this latter explanation is correct, then the search for the principles underlying the composition of gestural scores must look beyond the facts of articulation, to examine the effect the speaker is trying to convey to the listener and in turn what articulatory liberties the listener allows the speaker (see Lindblom 1983, Diehl and Kluender 1989, and Kingston and Diehl forthcoming for more discussion of this point).

Comments on Chapter 2

WILLIAM BARRY

In connection with Browman and Goldstein's conclusion that schwa is "weak but not completely targetless," I should like to suggest that they reach it because their concept of schwa is not totally coherent with the model within which the phenomenon "neutral vowel" is being examined. The two "nontarget" simulations that are described represent two definitions:

1 A slot in the temporal structure which is *empty* with regard to vowel quality, the vowel quality being determined completely by the preceding and

following vowel targets. This conflicts, in spirit at least, with the basic concept of a task-dynamic system, which explicitly evokes the physiologically based "coordinative structures" of motor control (Browman and Goldstein 1986). A phonologically targetless schwa could still not escape the residual dynamic forces of the articulatory muscular system, i.e. it would be subject to the relaxation forces of that system.

2 A relaxation target. The relaxation of the tongue-height parameter in the second simulation is an implicit recognition of the objection raised in point 1, but it still clashes with the "coordinative" assumption of articulatory control, which argues against one gesture being relaxed independent of other relevant gestural vowel parameters.

If an overall "relaxation target" is accepted, then, from a myofunctional perspective there is no means of distinguishing the hypothesized "targetless" schwa from the schwa-target as defined in the paper. Any muscle in a functional system can only be accorded a "neutral" or "relaxation" value as a function of the forces brought to bear on it by other muscles within the system. These forces will differ with each functional system. The rest position for quiet respiration (velum lowered, lips together, mandible slightly lowered, tongue tip on alveolar ridge) is different from the preparatory position found prior to any speech act independent of the character of the utterance onset (lips slightly apart, velum raised, jaw slightly open, laryngeal adduction).

The relaxation position may, therefore, be seen as a product of the muscular tensions required by any functional system, and implicit support for this view is given by Browman and Goldstein's finding that the mean back and front tongue height for schwa is almost identical with the mean tongue heights for all the other vowels. In other words, the mean vowel specifying the schwa "target" used by Browman and Goldstein is identical with the relaxation position of the vocalic functional system since it reflects the balance of forces between the muscle-tension targets specified for all the other vowels within the system. This accords nicely with the accepted differences in neutral vowel found between languages, and allows a substantive definition of the concept of "basis of articulation" which has been recognized qualitatively for so long (Franke 1889; Sievers 1901; Jespersen 1904, 1920; Roudet 1910).

This implies that, phonologically, schwa *can* in fact be regarded as undefined or "targetless," a status in keeping with its optional realization in many cases before sonorant consonants, and its lack of function when produced epenthetically.

One difficult question is the physiological definition and delimitation of a functional system, as it is mainly the scientific area of inquiry and the level of descriptive delicacy which defines a function. Since the same muscles are used for many different functions, a total physiological independence of on functional system from another using the same muscles cannot be expected A critical differentiation within speech, for example, is between possibl vocalic vs. consonantal functional subsystems. It has long been postulated a descriptively convenient and physiologically supportable that consonant gestures are superimposed on an underlying vocalic base (Öhman 1966 Perkell 1969; Hardcastle 1976). Browman and Goldstein's gestural score certainly in accordance with this view. A resolution of the problem within tl present discussion is not necessary, however, since the bilabial consonant context is maximally independent of the vocalic system and is kept constan

Papers in Laboratory Phonology II
Gesture, Segment, Prosody

EDITED BY GERARD J. DOCHERTY
Department of Speech, University of Newcastle-upon-Tyne

AND D. ROBERT LADD
Department of Linguistics, University of Edinburgh

The right of the University of Cambridge to print and sell all manner of books was granted by Henry VIII in 1534. The University has printed and published continuously since 1584.

CAMBRIDGE UNIVERSITY PRESS
CAMBRIDGE
NEW YORK PORT CHESTER MELBOURNE SYDNEY