546

4

Dynamic Modeling of Phonetic Structure*

Catherine P. Browman
Louis M. Goldstein

1. INTRODUCTION

Much linguistic phonetic research has attempted to characterize phonetic units in terms of measurable physical parameters or features (Fant 1973; Halle & Stevens 1979; Jakobson, Fant, & Halle 1969; Ladefoged 1971). Basic to these approaches is the view that a phonetic description consists of a linear sequence of static physical measures, either articulatory configurations or acoustic parameters. The course of movement from one such configuration to another has been viewed as secondary. We have proposed (Browman & Goldstein 1984) an alternative approach, one that characterizes phonetic structure as patterns of articulatory movement, or gestures, rather than static configurations. While the traditional approaches have viewed the continuous movement of vocal-tract articulators over time as "noise" that tends to obscure the segment-like structure of speech, we have argued that setting out to characterize articulator movement directly leads not to noise but to organized spatiotemporal structures that can be used as the basis for phonological generalizations as well as accurate physical description. In our view, then, a phonetic representation is a characterization of how a physical system (e.g., a vocal tract) changes over time. In this chapter, we begin to explore the form that such a characterization could take by attempting to explicitly model some observed articulatory trajectories.

Although we want to account for how articulators move over time,

* This research was supported by Grant No. NS-13617 from the National Institute of Neurological and Communicative Disorders and Stroke and Grant No. HD-01994 from the National Institute of Child Health and Human Development.

ISBN 0-12-268990-9

this does not mean that time per se must appear as a dimension of the description. In fact, a dimension of time would be quite problematic because of temporal variations introduced by changes in speaking rate and stress. For example, suppose our phonetic description were to specify the positions of articulators at successive points in time. As speaking rate changes, the values at successive time points are all likely to change in rather complex ways. Such a representation would not, therefore, be very satisfactory. It would be preferable to describe phonetic structure as a system that produces behavior that is organized in time but which does not require time as a control parameter (as has been suggested, e.g., by Fowler 1977, 1980). Like conventional phonetic representations, such a system does not explicitly refer to time. Unlike these representations, however, it explicitly generates patterns of articulator movement in time and space.

The dynamical approach to action currently being developed, for example, by Kelso and Tuller (1984) and Saltzman and Kelso (1983) provides the kind of time-free structure that can characterize articulatory movement. The approach has been applied to certain aspects of speech production (Fowler, Rubin, Remez, & Turvey 1980; Kelso, Tuller, & Harris 1983; Kelso, Tuller, & Harris in press), as well as to more general aspects of motor coordination in biological systems (e.g., Kelso, Holt, Rubin, & Kugler 1981; Kugler, Kelso, & Turvey 1980). Previous approaches to motor coordination have emphasized the importance of a time-varying trajectory plan for the muscles and joints to follow in the performance of a coordinated activity and require an intelligent executive to ensure that the plan is followed. In the dynamical approach taken by these investigators, actions are characterized by underlying dynamic systems, which, once set into place, can autonomously regulate the activities of sets of muscles and joints over time.

A physical example of a dynamic system is a mass–spring system, that is, a movable object (mass) connected by a spring to some rigid support. If the mass is pulled and the spring stretched beyond its equilibrium length, the mass will begin to oscillate. In the absence of friction, the equation characterizing motion is seen in Eq. (1), and the trajectory of the object attached to the spring can be seen in Figure 4.1.

(1) $m\ddot{x} + k(x - x_0) = 0$

where m = mass of the object

k = stiffness of the spring

x_0 = rest length of the spring

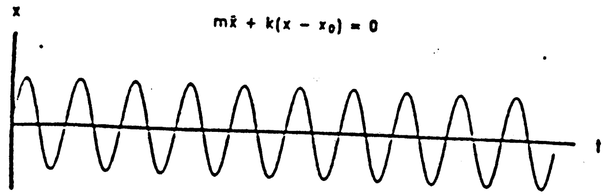


Figure 4.1 Output of undamped mass–spring system. Time (t) is on the abscissa and displacement (x) is on the ordinate.

x = instantaneous displacement of the object

$\ddot{x}$ = instantaneous acceleration of the object

Notice that an invariant organization, that in Equation (1), gives rise to the time-varying trajectory in Figure 4.1. No point-by-point plan is required to describe this pattern of movement, and time is not referred to explicitly. Only the parameter values and the initial conditions need be specified. This undamped mass–spring equation is a very simple example of a dynamic system. It is important to note that this system can give rise to a whole family of trajectories, not just the one portrayed in Figure 4.1. Different trajectories can be generated by changing values of the system's parameters. For example, changing the stiffness of the spring will change the observed frequency of oscillation. Changes in the rest length of the spring and the initial displacement of the mass will affect the amplitude of the oscillations.

This simple mass–spring equation (generally with a linear or nonlinear damping term added), exemplifies the dynamical approach to coordination and control of movement in biological systems in general and of speech articulators in particular. The appeal of this approach lies both in its potentially simple description of articulatory movements (i.e., only a few underlying parameters serve to characterize a whole range of movements) and in its physical and biological generality. In order to be useful for phonetics and linguistics, however, such a dynamic system must be related to phonetic structure. In one early attempt to specify this relationship, Lindblom (1967) proposed that a dynamic description could be used to account for speech-duration data. More recently, Kelso, Vatikiotis-Bateson, Saltzman, and Kay (1985) and Ostry, Keller, and Parush (1983) have analyzed stress and speaking rate variation in terms of the parameters of a dynamic model. In this chapter, we explore a basic linguistic issue that arises in the attempt to couch phonetic representations in the language of dynamics, namely, the definition of the articulatory gesture.

To begin to relate phonetic description to a dynamic system, let us consider a very simple example. Figure 4.2b shows the vertical position of a light-emitting diode (LED) on the lower lip of a speaker of American English as she produces the utterance ['babəbab] in the frame *Say* ——— *again*. The acoustic closures and releases marked on the articulatory trajectory are determined from the acoustic waveform, shown in Figure 4.2a. Note that the lower lip is raised (toward the upper lip) for the closures and lowered for the vowels. How can this observed lower-lip trajectory be described in terms of a dynamic system? Clearly the lower lip is showing an oscillatory pattern, that is, it goes up and down in a fairly regular way, but it does not show the absolute regularity of our mass–spring system in Figure 4.1. For example, the lip is lower in the full vowels than in the schwa. Thus, a mass–spring organization with constant parameter values will not generate this lower-lip trajectory.

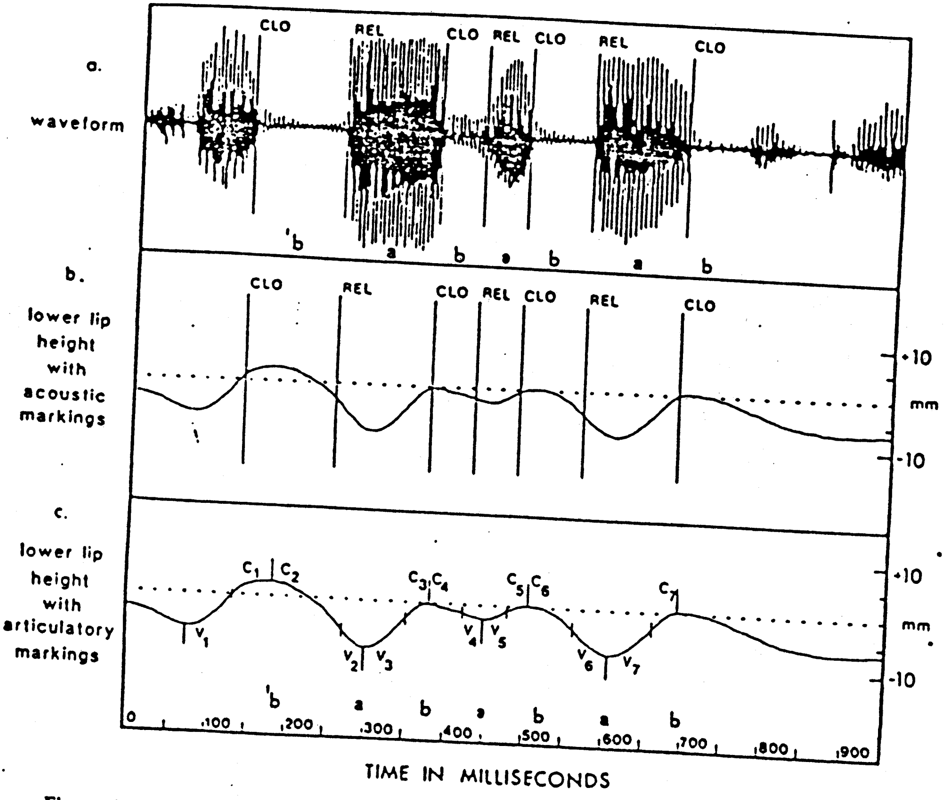


Figure 4.2 Lower-lip height and waveform for single token of ['babəbab].

However, it might be possible to generate this kind of trajectory if the parameter values were changed in the course of the utterance. The underlying dynamic organization, together with the particular changes of the parameters, would then serve to characterize the phonetic structure of the utterance.

It is, of course, obvious that a characterization of lower-lip position over time is not a complete phonetic representation. Nonetheless, in very simple utterances containing only bilabials and a single vowel, it comes quite close to being an adequate phonetic description. Browman, Goldstein, Kelso, Rubin, and Saltzman (1984) have shown that an alternating stress ['mama'mama . . .] sequence can be adequately synthesized using a vocal-tract simulation controlled by only two mass–spring systems—one for lip aperture (the distance between the two lips) and one for lip protrusion. Clearly, however, more complex utterances will require additional dynamic systems and relations among these systems; such interrelations and their implications for phonology are discussed in Browman and Goldstein (1984). Even for the restricted utterances considered here, we simplify the phonetic characterization by considering only the vertical position of the lower lip. We ignore horizontal lip displacement, the upper lip, and the fact that the movements of the lower lip can be decomposed into movements of the jaw and movements of the lower lip with respect to the jaw. The general framework we are operating within (the task dynamics of Saltzman & Kelso 1983) allows us to describe the coordination of multi-articulator gestures, but this is irrelevant to the present chapter, in which we consider only how to describe a particular articulator trajectory as the output of a dynamic system.

The undamped mass–spring system with constant parameter values generates sinusoidal trajectories with constant frequency and amplitude. We show that observed trajectories can be directly modeled as sinusoids whose frequency and amplitude vary at particular points during the utterance. Of particular interest is how to define these points at which the values are changed. Since time is not a parameter of the system, they are defined not with respect to some reference clock but in terms of the inherent cyclic properties of the dynamic system.

One set of inherently definable points at which parameter values can be modulated are the points of minimum and maximum articulator displacement. Modulation at these points is suggested by studies of articulator movement that characterize trajectories in terms of opening and closing gestures (e.g., Kuehn & Moll 1976; Parush, Ostry, & Munhall 1983; Sussman, MacNeilage, & Hanson 1973). Alternatively, points of peak velocity (both positive and negative) can also serve as dynamically definable markers for modulation. In a simple mass–spring system, velocity peaks occur at the resting, or equilibrium, position. These different points of change imply different phonetic organizations, as can be seen with the help of Figure 4.2c. Here we see the same articulatory trajectory as in Figure 4.2b, with the addition of tick marks that indicate the displacement and velocity extrema. These points divide the utterance into intervals, each of which has been labeled with either a C (for consonant) or a V (for vowel). The consonant intervals are those on either side of a displacement peak, and the vowel intervals are those on either side of a displacement valley. Points of peak velocity, indicated by the smaller tick marks on the slopes, separate consonant intervals from vowel intervals. For example, V_1 is the interval from the minimum position of the lower lip in the frame vowel [ey] to the point of peak velocity as the lip starts to raise for the initial [b]. C_1 is the interval from this latter peak velocity point to the center of maximum lower-lip height during the [b]. C_2 is the interval from this displacement peak to the peak velocity as the lower lip lowers for the following vowel [a].

If we change our model parameters only at displacement peaks and valleys, then successive VC or CV intervals (e.g., V_1C_1, C_2V_2) will be characterized with the same set of parameters. This constitutes a phonetic hypothesis that the articulatory trajectories can be modeled as successive CV and VC transition gestures, each with its characteristic values for the dynamic parameters. The parameters for these opening and closing gestures must take into account both the particular consonant and the particular vowel. Thus, this hypothesis provides a phonetic structure rather different from that commonly assumed in linguistics, in that it does not provide a physical characterization of individual consonants or vowels.

An alternative division of the articulatory trajectories is clearly possible if we change parameters at velocity extrema rather than displacement extrema. In this way, successive C intervals will have a single characterization, as will successive V intervals (e.g., C_1C_2, V_2V_3). These new intervals, then, correspond roughly to consonant and vowel gestures, rather than to CV and VC transition gestures. Under this hypothesis, the relationship between the dynamic characterization and more conventional phonetic representations is somewhat more transparent than it is under the transition hypothesis. Note, however, that even under this hypothesis, consonants and vowels are defined in terms of dynamic structures rather than as spatial targets.

In this chapter, then, we present the results of some preliminary modeling of articulatory trajectories with sinusoids (the output of an undamped mass–spring system) under the C–V and transition hypotheses outlined above. In particular, the two hypotheses are contrasted with respect to how the frequency parameter of the sinusoidal model is modulated. The frequency parameter (proportional to the square root of the stiffness of the underlying mass–spring system, assuming a unit mass) is of particular interest, because it controls the duration of a given gesture and thus holds the key to how temporal (durational) regularities can be accommodated in a descriptive system that does not include time as a variable. Therefore, we examine how the frequency of an articulatory gesture varies as a function of stress, position within the item, and vowel quality.

2. METHOD

2.1. Articulatory Trajectories

The trisyllabic nonsense items shown in Table 4.1 were chosen for analysis. Stress is either initial or final, with the second syllable always reduced, and the vowels are either [i] or [a]. The items were recorded by a female speaker of American English in the carrier sentence *Say* ——— *'again*. Table 4.1 indicates the number of tokens of each of the items that were analyzed.

Movements of the talker's lips and jaw were tracked using a Selspot system that recorded displacements, in the midsagittal plane, of LEDs placed on the nose, upper lip, lower lip, and chin. The Selspot output was recorded on an FM tape recorder and was later digitally sampled at 200 Hz for computer analysis. To correct the articulator displacements for possible movements of the head, the Selspot signal for the nose LED was subtracted from each of the articulator signals. Each resulting articulator trajectory was then smoothed, using a 25 ms triangular window. For the present purpose, only the vertical displacement of the lower lip was analyzed.

Displacement maxima and minima were determined automatically using a peak-finding algorithm. Instantaneous velocities were computed by

Table 4.1
NONSENSE ITEMS USED IN ANALYSIS

Utterance	Number of tokens
bibə'bib	11
'bibəbib	14
babə'bab	10
'babəbab	11

taking the difference of successive displacement samples. The maxima and minima of the resulting velocity curves were determined using the same program as for the displacements. Displacement and velocity extrema were used to divide each token into seven C and seven V intervals, as shown in Figure 4.2c.

2.2 Modeling

Each successive interval of each token was modeled as the output of a simple mass–spring system by fitting sinusoids to the articulatory trajectories. We generated the model trajectories using a sine-wave equation directly, Equation (2), in order to emphasize the inherent cyclic properties of dynamic systems. Recall that frequency is related to stiffness and amplitude to rest length and maximum displacement. Thus, we controlled frequency, amplitude, and equilibrium position (rest length). (Phase is discussed below.) The individual model points, $x'(i)$, for an interval were generated according to Equation (2) for the ith point in the interval (one point every 5 ms):

$$(2) \quad x'(i) = x_0 + A \sin (wi + \phi)$$

where x_0 = equilibrium position

A = amplitude

w = frequency (in degrees per sample point)

ϕ = phase

Frequency varied every two-interval gesture, where the gestures were defined according to the two hypotheses outlined in the previous section. For the C–V hypothesis, a gesture included the two intervals between successive velocity extrema (e.g., C_1C_2, V_2V_3). For the transition hypothesis, a gesture included the two intervals between successive displacement extrema (e.g., V_1C_1, C_2V_2). We posit that a gesture constitutes a half cycle. Therefore, the frequency was computed as the reciprocal of twice the combined duration of the two intervals comprising a gesture. For example, the frequency used to model intervals C_1 and C_2 under the C–V hypothesis was 1 / (2 * (duration of C_1 + duration of C_2)). Similarly, the frequency for intervals C_2 and V_2 under the transition hypothesis was computed as 1 / (2 * (duration of C_2 + duration of V_2)).

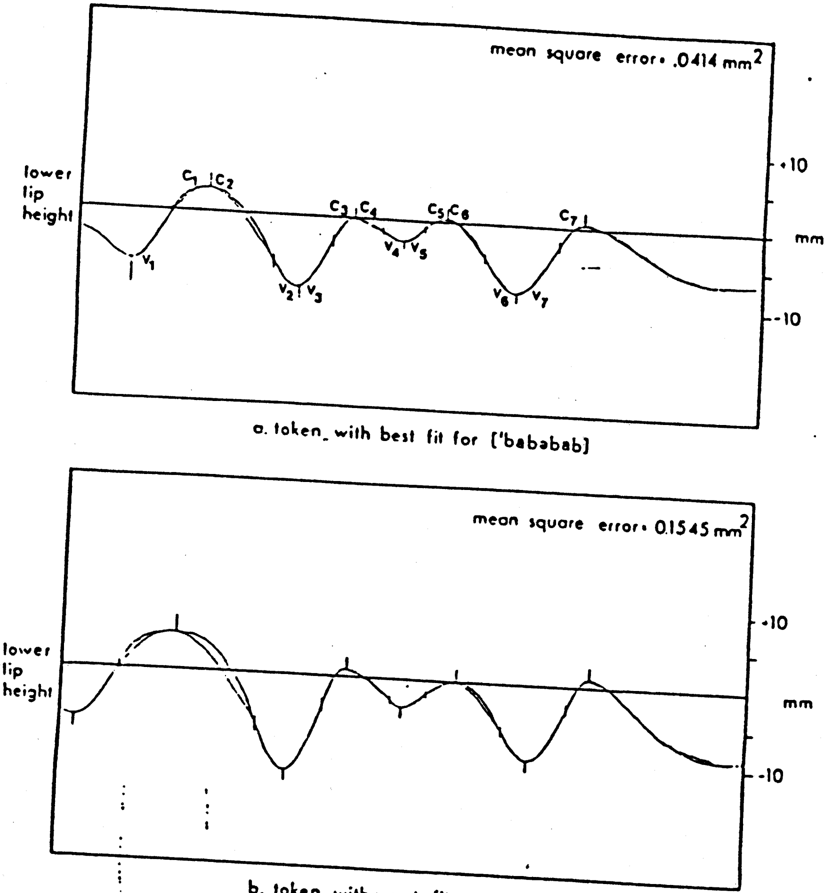
Since our primary interest in this study was in the frequency parameter, we allowed the values of the equilibrium position and amplitude to change every interval. The values were determined from the initial and final displacement of the interval, adjusted for phase. The phase angle for a sine wave is 90 degrees at maximum displacement (the peaks) and 270 degrees at minimum displacement (the valleys). Amplitude and equilibrium position values were determined by the constraint that model and data agree exactly at these points, both in phase and in displacement. That is, the observed peaks and valleys were assumed to be the displacement extrema generated by the underlying model. The analogous assumption was not possible for the velocity extrema, however, since often the velocity extrema were not midway between the displacement extrema, as they would be if the parameter values were not changing (see Figure 4.1). Thus, the observed velocity extrema did not correspond to 0 and 180 degrees in the modeled trajectories. Rather, the phases for these points in the model were permitted to vary according to the constraint that model and data agree exactly here as well as at the displacement extrema.

3. RESULTS

Sinusoidal models are strikingly successful in fitting the articulatory data. Figure 4.3a shows the model trajectory generated for the C–V hypothesis superimposed on the real trajectory for our sample token of ['babəbab]. The curves lie almost completely on top of one another, diverging substantially only during the C_1, C_2, and C_6 intervals. This particular token is the best modeled of all ['babəbab] tokens, as measured by the mean square error of the modeled points. The token with the worst fit, not only for this utterance but for all the utterances, is shown in Figure 4.3b. Again the curves lie almost completely on top of one another, diverging substantially only in the same places as in Figure 4.3a.

In general, the modeled trajectories for both hypotheses and for all utterances fit comparably to the trajectories shown in Figure 4.3a. Table 4.2 gives the mean square error averaged across all tokens for each of the four utterances under the C–V and transition hypotheses. The two hypotheses differ by only a small amount, but the C–V hypothesis appears to be consistently better. Comparison of individual tokens supports this slight superiority of dividing the trajectory into consonant and vowel gestures.

The contribution of different intervals of the trajectories to the error can be seen in Figure 4.4. The four curves show the model and data superimposed for the best tokens of each of the utterance types under the C–V hypothesis. Utterances with [i] are shown on the left (a and b), and utterances with [a] are shown on the right (d and e). The graphs at the bottom of the figure show the mean square error for the individual



a. token with best fit for ['babəbab]

b. token with worst fit : ['babəbab]

Figure 4.3 Sample comparisons of superimposed model (C–V hypothesis) and data trajectories.

Table 4.2
MEAN SQUARE ERROR AVERAGED ACROSS TOKENS

Utterance	Mean square error (mm^2)	
	C–V hypothesis	Transition hypothesis
bibə'bib	.0154	.0171
'bibəbib	.0466	.0615
babə'bab	.0358	.0471
'babəbab	.0907	.1220

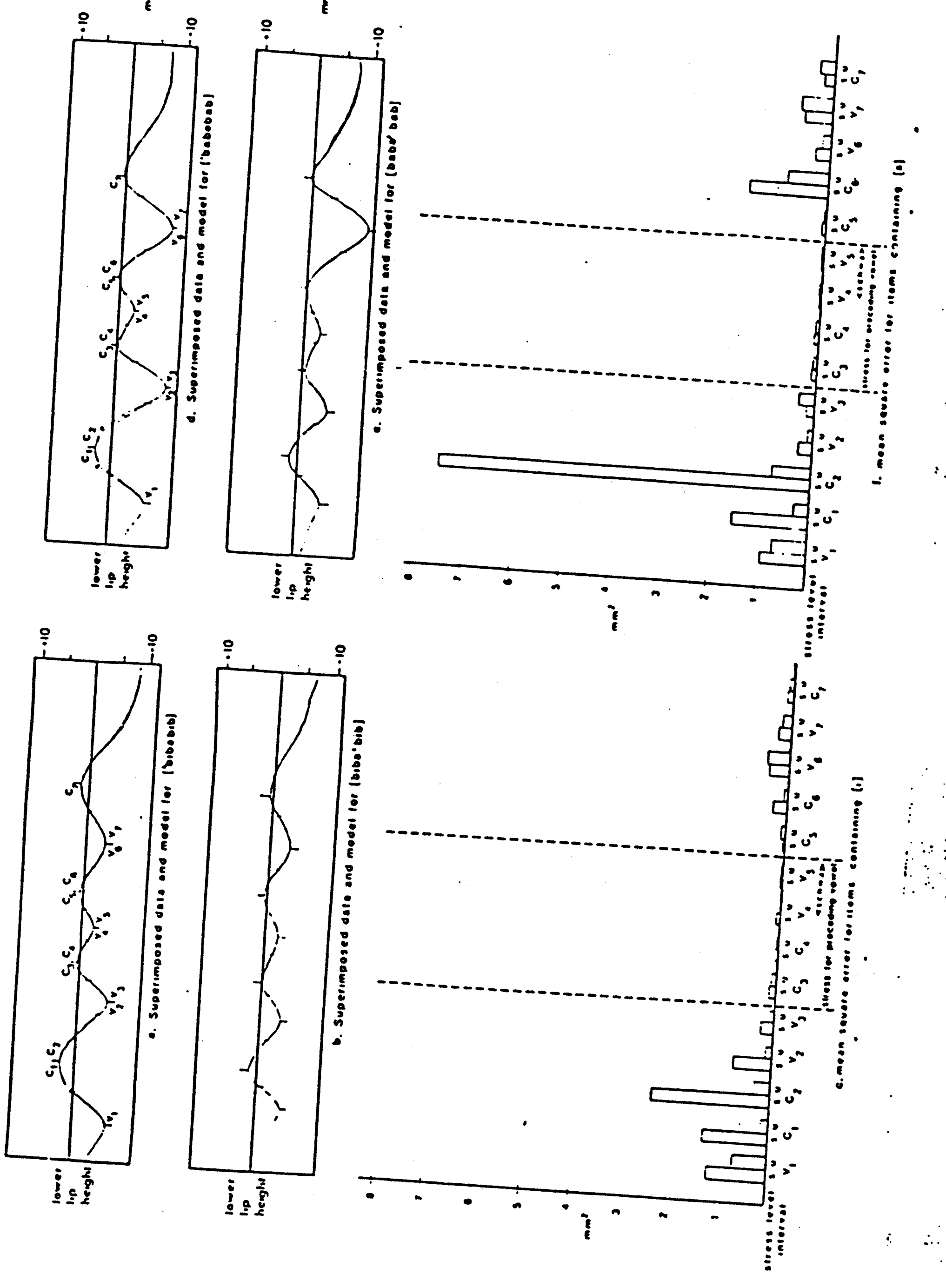
intervals from V_1 to C_7. These are averages across all tokens of a given utterance. Again, results for utterances with [i] are shown at the left and with [a] at the right. Intervals occurring in stressed and unstressed syllables are shown separately.

The error distributions show that the worst fit is found for item-initial stressed consonants, for both [a] and [i] utterances. In particular, interval C_2, the release of this initial stressed consonant, is poorly modeled relative to the other intervals. The release of the stressed consonant is also relatively poorly modeled in final syllables containing [a]. Examining the trajectories in the poorly modeled regions of ['babəbab] in Figure 4.4d, we can see that the actual consonant trajectory (indicated by arrows) shows a flatter top than that predicted by sinusoidal trajectories. This can, perhaps, be explained by noting that it tends to occur in regions in which the lower lip is raised quite high against the upper lip. The flattening may be the result of some limit on the compressibility of the lips. Alternatively, it may be that there is some tendency for initial stressed consonants to be held, suggesting a somewhat different kind of dynamic system (e.g., a damped mass–spring).

The error distributions also show a clear tendency for the reduced syllables to have the smallest error. This may partly be due to the fact that the actual displacement differences between the beginning and end of such intervals tend to be very small, and, given that the ends are perfectly modeled, there simply is not much room for error. Similarly, there is some tendency for utterances with [i] to show less error than utterances with [a]. Again, the lower lip shows less movement with [i] than [a], leaving less room for error. However, the smaller amplitude of movement does not completely account for the better fit. Correlations between amplitude of movement and error are not high, for example, .242 for [babə'bab]. Thus, the straightforward mass–spring model we have chosen to investigate appears to be adequate for the unstressed and reduced syllables but needs to be modified for stressed, item-initial consonants.

In addition to goodness-of-fit considerations, a dynamic phonetic structure can also be evaluated with respect to how well it can elucidate systematic variation. For example, we can examine how the values of the model parameters vary as a function of context. Given the preliminary nature of our modeling, we simply show some easily observable trends, rather than present a detailed statistical analysis.

The bars with solid lines in Figure 4.5a show the mean value of the frequency parameter for the consonant gesture under the C–V hypothesis, as a function of the consonant's stress and position within the item for the two vowel contexts. Only the three consonants preceding vowels



a. Superimposed data and model for ['bibəbib]

b. Superimposed data and model for [bibə'bib]

c. mean square error for items containing [i]

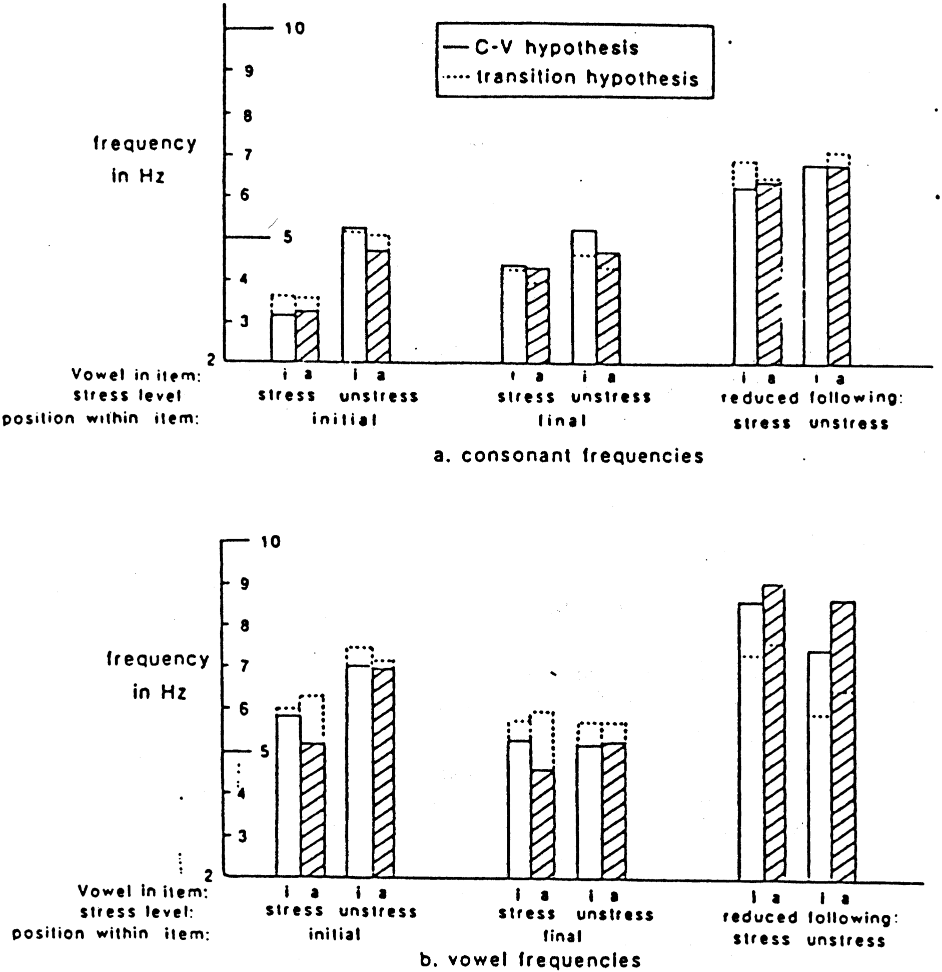
d. Superimposed data and model for ['babəbab]

e. Superimposed data and model for [babə'bab]

f. mean square error for items containing [a] are shown. The first thing to note about the data is that the nature of the vowel in the item ([i] or [a]) has little effect on the consonant frequency, (although unstressed [b] has a lower frequency before [a] than before [i]). That is, consonant frequency is relatively independent of vowel. Stress, however, clearly shows a systematic influence on the frequency of the [b] gesture. The consonant has a higher frequency in unstressed syllables than in stressed syllables for both the initial and final syllables in the item. In the medial reduced syllable, the consonant has the highest frequency of all. Kelso et al. (1985) also found unstressed gestures to be stiffer than stressed gestures, which is equivalent to an increase in frequency. This pattern of variation is completely consistent with the lengthening effect of stress as measured acoustically (e.g., Klatt 1976; Oller !973). Additionally, there is variation according to position. Item-initial stressed consonants are lower in frequency than consonants in the final syllable of the item. Again, this is consistent with observed acoustic word-initial consonant lengthening (Oller 1973).

The vowel gestures are analyzed in a similar way in Figure 4.5b. Reduced vowels have higher frequencies than full vowels, as expected from the consonant data. Full vowels, however, do not behave quite as systematically as the consonants. For unstressed full vowels, there is little or no difference between [i] and [a] in frequency. Stressed full vowels, however, show a slight difference depending on whether the item contains [i] or [a]. Stressed [i] has a slightly higher frequency than stressed [a], which corresponds to the measured acoustic duration difference noted, for example, by Umeda (1975). (Reduced vowels show a possible compensatory effect, in that reduced vowels in items containing [i] have a lower frequency than those in items with [a].) The effect of stress for the full vowels is also not completely regular but rather depends upon position. Only vowels in initial syllables show lower frequencies when stressed. Note, however, that vowels in final syllables are lower in frequency than those in initial syllables, which is in agreement with the acoustic effect of final lengthening (Klatt 1975). It may be, then, that the final-lengthening effect washes out temporal differences between stressed and unstressed vowels in the final syllable. (At least one of Oller's 1973 subjects shows this kind of pattern). Looked at in another way, when the initial vowel is stressed, it has about the same frequency as the unstressed final vowel in the same item. That is, the final lengthening effect is similar in magnitude to the stress effect. This is consistent with

Figure 4.4 Best fit and error distributions by intervals for comparison of model (C–V hypothesis) and data trajectories; s = stressed interval, u = unstressed interval; for reduced intervals C_3–V_3, s = preceding vowel stressed, u = preceding vowel unstressed. Arrows indicate data.



a. consonant frequencies

b. vowel frequencies

Figure 4.5 Consonant and vowel frequencies generated by C–V and transition hypotheses, according to vowel in item, stress level, and position within item. For reduced consonants and vowels, always in medial position, stress or unstress refers to the preceding syllable.

acoustic and perceptual investigations of stress patterns (Fry 1958; Lea, 1977).

The bars with dotted lines superimposed on the solid-line bars in Figure 4.5 show the mean frequencies obtained under the transition hypothesis. For reasons to be discussed in the next section, the CV transitional gestures have been superimposed on the corresponding C gestures, and the VC transitions on the corresponding V gestures. For example, the consonant in initial position, which represents the consonant closing and release (C_1C_2) under the C–V hypothesis, represents, under the transition hypothesis, the consonant release and movement to the following vowel (C_2V_2). Similarly, the initial vowel represents V_2V_3 under the C–V hypothesis and V_3C_3 under the transition hypothesis. Comparison of the dotted lines and solid bars shows substantial similarity. The only important differences are in the frequencies of the reduced vowels, which in the transition hypothesis are not higher than the full vowels. This is perhaps not surprising, given that the VCs that constitute the reduced syllables (V_5C_5) include the initial consonant interval (C_5) of the following unreduced syllable.

To summarize, both the C–V hypothesis and the transition hypothesis fit the data quite well (except for stressed item-initial consonants) and generate very similar frequencies. The two hypotheses differ slightly in that the C–V hypothesis provides marginally better fit and they predict differing patterns of frequencies for reduced vowels. Only stressed and reduced vowels show a difference in the frequencies generated for items containing [a] and items containing [i]. Stress level, however, has a generally consistent effect, with stressed syllables having the lowest frequency, unstressed syllables somewhat higher, and syllables containing reduced vowels having the highest frequency. This stress effect fails only for full vowels in final syllables, which in addition display lowered frequencies relative to initial syllables. Consonants, in contrast, have lower frequencies in initial syllables than in final. These stress and position effects are consistent with acoustic duration effects noted in the literature. Thus, well-known aspects of the temporal organization of speech can be accounted for in a model that does not explicitly refer to time.

4. IMPLICATIONS AND PROSPECTS

The success of a very simple dynamic system in modeling the observed trajectories of individual gestures gives important empirical support to the dynamic approach to phonetic structure. The approach is theoretically appealing because it provides a way of explicitly generating articulator trajectories from a time-free sequence of parameter specifications for consonants and vowels. This is made possible by recognizing, as suggested by Fowler (1977, 1980) and Fowler et al. (1980), that a phonetic structure is not just a linear sequence of parameter, or feature, values but also must be described as some particular dynamic organization that the parameter values serve to modulate. The successive changes in parameter values can be linked to particular points in the underlying dynamic organization. This differs from conventional phonetic representations that do not provide any explicit way of generating articulatory trajectories from a sequence of parameter specifications.

The present model is only a preliminary validation of the general approach. A number of improvements need to be made before it can be claimed to have predictive power. In particular, the interval-by-interval specification of amplitude, with endpoints exactly matched, needs to be replaced with a procedure that allows amplitude to be specified over longer stretches. The determination of frequency should be made in a way that is less vulnerable to experimental (and theoretical) error in determining the endpoints of the gestures. Both frequency and amplitude should ultimately be determined by general linguistic parameters, for example, stress level and position, rather than by item-specific trajectory matching. These improvements can be carried out using the present simple undamped mass–spring dynamic model. In addition, alternative dynamic models need to be explored in order to account for the poorly matched item-initial stressed consonants as well as the interarticulator compensation effects discussed in Saltzman & Kelso (1983).

Another area to be investigated further is the organization of the underlying phonetic structure. This chapter compared two organizational hypotheses: consonant–vowel gestures and transitional gestures. While both hypotheses fit the data quite well in this preliminary test, there is some indication that additional organizational hypotheses should be explored in future modeling attempts.

In the comparison of the two hypotheses, the CV transition was equated with the C, and the VC transition with the V. This was a post hoc decision, based on the similarity of frequencies when the two hypotheses were so equated. In fact, the frequencies would not appear similar at all if the CV transitions were equated with Vs, rather than with Cs, and the VC transitions were likewise switched. Why the frequencies should line up this way is not clear. It may simply be the case that the intervals immediately following the displacement extrema, which are the intervals common to the C (or V) gestures and their equated transition gestures, are those in which frequency is crucially controlled. This interpretation is supported by results from an additional analysis in which frequency was determined interval by interval rather than by using two contiguous intervals. In this analysis, exactly those intervals following the displacement extrema displayed the stress and position patterns discussed in the preceding section, while the alternate intervals showed no clear relationship to the linguistic variables. However, there is also a more interesting account. This involves positing a structure in which frequency is fixed over a larger span of at least three intervals, for example, $C_1C_2V_2$ and $V_2V_3C_3$.

These longer gestures constitute a kind of overlapping organization (V_2 appears in both above) that is independently motivated by the kinds of coarticulatory phenomena typically observed in speech (see the overlapping segment analysis of coarticulation presented by Fowler 1983).

Some such concept of overlapping gestures is also suggested by another regularity observable in the frequency patterns. The frequency of a consonant gesture under the C–V hypothesis is lower than the frequency of the vowel that follows it. This is counter to the common assumption that consonants involve short, rapid movements, while full vowels correspond to longer movements. It might, of course, be a startling new result. Or it may simply be that the choice of endpoints needs to be improved. But such a counterintuitive result may also be indicative of a basic flaw in the hypothesis generating the result. One obvious candidate is the assumption, in both hypotheses investigated, of independent, sequential gestures. Such an assumption was useful as a starting point, but is unlikely to be accurate. Rather, some form of overlap of the gestures—coarticulation—would likely give a better picture and will be permitted in future modeling attempts. A possible overlapping structure is one in which consonantal gestures are phased relative to ongoing vowel gestures (Tuller, Kelso, & Harris 1982).

Finally, the comparison of the C–V hypothesis with the transitional hypothesis carries certain implications, not only for future research into phonetic organization, but also for the interpretation of past studies. Investigations into the nature of speech articulator movements have tacitly assumed the transition hypothesis (e.g., Kuehn & Moll 1976; Parush et al. 1983; Sussman et al. 1973) and have consequently couched the description of their results in terms of opening and closing gestures. The present study, however, shows that the C–V hypothesis provides an organization that captures all of the same generalizations in the data as the transitional hypothesis; one that fits the data as well as or better than the transitional hypothesis; and moreover, one that is more immediately relatable to traditional linguistic units. In addition, while the two hypotheses generally produce equivalent frequency analyses, in at least one case—that of reduced vowels—they appear to differ substantively. The present study does not constitute evidence for one hypothesis over the other, given the overall similarity in fit. However, it does constitute evidence that the C–V organization, or some variant thereof, warrants serious consideration in the interpretation of speech articulator-movement data. In general, we think that bringing dynamic principles to bear on problems of linguistic organization will lead to more linguistically relevant accounts of speech production as well as to a much richer, yet simple, conception of phonetic structure. The structure comprises an underlying dynamic system with associated parameter values. Together, the system and its parameters explicitly generate patterns of articulator movement. In addition, as we have demonstrated, such structures can retain the useful descriptive properties of more conventional phonetic representations.

REFERENCES

Browman, C. P., Goldstein, L., Kelso, J. A. S., Rubin, P., & Saltzman, E. (1984). Articulatory synthesis from underlying dynamics. *Journal of the Acoustical Society of America, 75,* S22–23. (abstract).

Browman, C. P., & Goldstein, L. (1984). *Towards an articulatory phonology.* Unpublished manuscript.

Fant, G. (1973). Distinctive features and phonetic dimensions. In G. Fant, *Speech sounds and features* (pp. 171–191). Cambridge, MA: MIT Press. (Originally published 1969)

Fowler, C. A. (1977). *Timing control in speech production.* Bloomington, IN: Indiana University Linguistics Club.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing control. *Journal of Phonetics, 8,* 113–133.

Fowler, C. A. (1983). Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet. *Journal of Experimental Psychology: General, 112,* 386–412.

Fowler, C. A., Rubin, P., Remez, R. E., & Turvey, M. T. (1980). Implications for speech production of a general theory of action. In B. Butterworth (Ed.), *Language production* (pp. 373–420). New York: Academic Press.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech, 1,* 126–152.

Halle, M., & Stevens, K. N. (1979). Some Reflections on the theoretical bases of phonetics. In B. Lindblom & S. Ohman (Eds.), *Frontiers of speech communication research* (pp. 335–353). New York: Academic Press.

Jakobson, R., Fant, C. G. M., & Halle, M. (1969). *Preliminaries to speech analysis: The distinctive features and their correlates.* Cambridge, MA: MIT Press.

Kelso, J. A. S., Holt, K. G., Rubin, P., & Kugler, P. N. (1981). Patterns of human interlimb coordination emerge from the properties on nonlinear limit cycle oscillatory processes: Theory and data. *Journal of Motor Behavior, 13,* 226–261.

Kelso, J. A. S., & Tuller, B. (1984). A dynamical basis for action systems. In M. S. Gazzaniga (Ed.), *Handbook of neuroscience* (pp. 321–356). New York: Plenum.

Kelso, J. A. S., Tuller, B., & Harris, K. S. (1983). A 'dynamic pattern' perspective on the control and coordination of movement. In P. MacNeilage (Ed.), *The production of speech* (pp. 137–173). New York: Springer-Verlag.

Kelso, J. A. S., Tuller, B., & Harris, K. S. (in press). A theoretical note on speech timing. In J. S. Perkell & D. Klatt (Eds.), *Invariance and variation in speech processes.* Hillsdale, N.J.: Erlbaum.

Kelso, J. A. S., Vatikiotis-Bateson, E., Saltzman, E. L., & Kay B. (1985). A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling. *Journal of the Acoustical Society of America, 77,* 266–280.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics, 3,* 129–140.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America, 59,* 1208–1221.

Kuehn, D. R., & Moll, K. L. (1976). A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics, 4*, 303–320.

Kugler, P. N., Kelso, J. A. S., & Turvey, M. T. (1980). On the concept of coordinative structures as dissipative structures: I. Theoretical lines of convergence. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior* (pp. 3–47). New York: North-Holland.

Ladefoged, P. (1971). *Preliminaries to linguistic phonetics*. Chicago: University of Chicago Press.

Lea, W. A. (1977). Acoustic correlates of stress and juncture. In L. M. Hyman (Ed.), *Studies in stress and accent*. Los Angeles: University of Southern California.

Lindblom, B. (1967). Vowel duration and a model of lip mandible coordination. *Speech Transmission Laboratory Quarterly Progress Report, STL-QPSR-4*, 1–29.

Oller, D. (1973). The effects of position in utterance on speech segment duration. *Journal of the Acoustical Society of America, 54*, 1235–1246.

Ostry, D. J., Keller, E., & Parush, A. (1983). Similarities in the control of speech articulators and limbs: Kinematics of tongue dorsum movement in speech. *Journal of Experimental Psychology: Human Perception and Performance, 9*, 622–636.

Parush, A., Ostry, D. J., & Munhall, K. G. (1983). A kinematic study of lingual coarticulation in VCV sequences. *Journal of the Acoustical Society of America, 74*, 1115–1125.

Saltzman, E. L., & Kelso, J. A. S. (1983). Skilled actions: A task dynamic approach. *Haskins Laboratories Status Report on Speech Research, SR-76*, 3–50.

Sussman, H. M., MacNeilage, P. F., & Hanson, R. J. (1973). Labial and mandibular dynamics during the production of labial consonants: Preliminary observations. *Journal of Speech and Hearing Research, 16*, 397–420.

Tuller, B., Kelso, J. A. S., & Harris, K. S. (1982). Interarticulator phasing as an index of temporal regularity in speech. *Journal of Experimental Psychology: Human Perception and Performance, 8*, 460–472.

Umeda, N. (1975). Vowel duration in American English. *Journal of the Acoustical Society of America, 58*, 434–455.