EXPLORING THE INFORMATION SUPPORT FOR SPEECH*

J. A. Scott Kelso+ and Betty Tuller++

Abstract. A well-established feature of speech production is that talkers, faced with both anticipated and unanticipated perturbations, can spontaneously adjust the movement patterns of articulators such that the acoustic output remains relatively undistorted. Less clear is the nature of the underlying process(es) involved. In this study we examined five subjects' production of the point vowels /i, a, u/ in isolation and the same vowels embedded in a dynamic speech context under normal conditions and a combined condition in which (a) the mandible was fixed by means of a bite block, (b) proprioceptive information was reduced through bilateral anesthesia of the temporomandibular joint, (c) tactile information from oral mucosa was reduced by extensive application of topical anesthetic, and (d) auditory information was masked by white noise. Analysis of formant patterns revealed minimal distortion of the speech signal under the combined condition. These findings are unfavorable for central (e.g., predictive simulation) or peripheral closed-loop models, both of which require reliable peripheral information; they are more in line with recent work suggesting that movement goals may be achieved by muscle collectives that behave in a way that is qualitatively similar to a nonlinear vibratory system.

The remarkable generativity of human movement is a mystery that continues to resist explanation. Within limits, people (and animals) can achieve the same 'goal' through a variety of kinematic trajectories, with different muscle groups and in the face of ever-changing postural and biomechanical requirements. This phenomenon—variously referred to as motor equivalence (Hebb, 1949) or equifinality (von Bertalanffy, 1973)—has been demonstrated again quite recently by Raibert (1978), who showed writing patterns to be characteristic of the same individual even when produced by structures (such as the foot or mouth) that had never previously been used for the act of writing.

*A preliminary version of this paper was presented at the 101st meeting of the Acoustical Society of America, May 18-22, 1981.
+Also Departments of Psychology and Biobehavioral Sciences, The University of Connecticut.
++Also Department of Neurology, New York University Medical Center.

Acknowledgment. This research was supported by NIH Grants N313617 and Biomedical Research Support Grant RR05596 to Haskins Laboratories and NIH Post-Doctoral Fellowship 1-F32-NS-06718-01 to the second author. We thank Thomas Gay and the University of Connecticut Health Center at Farmington for use of facilities. We are especially grateful to Dr. Robert Gross of Louisiana State University Medical Center for performing anaesthetization procedures.

Human language is generative in a qualitatively similar way: We seem to have a potentially infinite number of ways of constructing sentences. Nor is it trivial that language, even when stripped of its symbolic component, is a creative or generative activity. Articulatory maneuvers for producing speech sounds can be effected in spite of continuously varying initial conditions. Often the same phonetic segment in different environments can be achieved by very different movement trajectories and end-states. One commonly used experimental paradigm for examining equifinality in speech takes the form of placing a bite block between the teeth, thus fixing the position of the mandible. Under such conditions, so-called "steady state" vowels can be produced apparently without the need for on-line acoustic feedback. Normal range formant patterns are obtained even at the first glottal pitch pulse (Gay, Lindblom, & Lubker, 1981; Lindblom, Lubker, & Gay, 1979; Lindblom & Sundberg, 1971). Moreover, speakers are capable of such "compensatory articulation" with little (if any) articulatory experimentation. Recent work on bite-block speech has shown that response times to produce vowels of the same acoustic quality under normal and bite-block conditions are nearly identical. In addition, the degree of "compensation" (as indexed by deviations from normal formant frequencies) remained unchanged as a function of practice (Fowler & Turvey, 1980; Lubker, 1979). The evidence, then, favors an interpretation that articulatory adjustments to novel contextual conditions are relatively immediate.

What kind of control process could account for the adaptive, generative nature of speech production? An open-loop control system in which commands for producing a given vowel prescribe in detail the activities of relevant muscles can be dispensed with because, by definition, such systems are insensitive to changing contextual conditions. On the other hand, closed-loop control does offer the advantage of adjustment to initial conditions. In peripheral closed-loop, feedback systems, a sensory goal in the form of a spatial (MacNeilage, 1970) or auditory target (Ladefoged, DeClerk, Lindau, & Papyun, 1972; MacNeilage, 1980) is paired with an appropriate set of commands for accomplishing the goal. Resulting sensory consequences are then compared with the sensory goal so that corrections can be made. A potential problem with peripheral closed-loop control is that the corrective process requires time (at least one cycle around the corrective loop). However, if the adjustment to novel conditions is indeed immediate—thus excluding the need for trial and error methods—then a closed-loop mechanism tied to the peripheral motor system fails to capture the phenomenon of interest.

An alternative account favored by Lindblom and colleagues (e.g., Lindblom et al., 1979) replaces the peripheral feedback loop by a central simulation process that derives the expected sensory consequences from a simulated set of motor commands before the actual efferent signals are sent to the periphery. An internal comparison between the simulated and 'target' sensory consequences yields an error signal on the basis of which new (and correct) commands can be emitted. In this manner, adjustments to changes in context can be made in the internal simulation without incurring erroneous effects at the periphery.

It is important to note that the models discussed thus far make the explicit assumption that reliable peripheral information about the articulators' initial conditions is available before motor commands (simulated or actual) are generated. In the peripheral closed-loop model, for example,
sensory input must be compared to the internal referent before the output of command signals. In the internal loop model, simulated motor commands are generated for the initial conditions that currently exist (Lindblom et al. 1979). It is not clear in the latter formulation what would happen if contextual conditions changed between the time that simulated and actual motor commands were generated. A more efficient system would be continuously sensitive to, and be capable of modulation by, contextual conditions. For the sake of argument, however, let us assume with Lindblom et al. that one benefit of the internal loop is its speed of correction; possibly the loop is so fast that appropriate output can be generated before contextual conditions have changed.

In any case, for both closed-loop models, elimination or reduction of peripheral information about initial conditions should drastically affect the system's ability to adjust to the novel situation created by a bite block. There are very limited data on this point. Gay and Turvey (1979) found that a single subject (a phonetician) made several attempts before producing 'normal' formant frequencies for the vowel /i/ under conditions in which a bite block was combined with topical anaesthesia of the oral mucosa and bilateral nerve blockage of the temporomandibular joint. Although this result has suggested to some (cf. Perkell, 1979) that joint and tactile information is used to establish an "orosensory frame of reference," we believe there are grounds for caution. One problem is that it is unclear how—given the considerable reduction of peripheral information—Gay and Turvey's subject was capable of adaptive adjustment at all. One possibility, which we consider here, is that auditory information may have played a potentiating role. Although acoustic information does not appear to be a necessary condition for compensatory articulation (e.g., Lindblom et al., 1979), the Gay-Turvey experiment does not preclude an auditory contribution in "recalibrating" the speech system when information from motor structures is rendered unreliable.

The present experiment was designed to examine the role of peripheral information (auditory and somesthetic) in accounts of "immediate adjustment" by asking naive subjects to produce vowels under normal conditions and under bite-block conditions in which somatosensory information was drastically reduced (if not eliminated) and audition was masked by white noise. In addition we address the question of whether the so-called "steady-state" paradigm for bite-block vowels reflects normal dynamic speech motor processes. By examining the production of vowels embedded in a dynamic speech context as well as in isolation, we can discover what differences there are, if any, in observed acoustic patterns. As we shall see, the availability of peripheral information from neither auditory nor peripheral motor structures appears to be crucial to immediate adjustment. We take this result as non-supportive for extant models of the phenomenon. In their place, we offer a class of model—emerging in other areas of motor control (Bizzi, 1980; Fel’dman, 1966, 1980; Kelso, 1977; Kelso & Holt, 1980; Kelso, Holt, Kugler, & Turvey, 1980; Polit & Bizzi, 1978) as well as in the recent speech production literature (cf. Fowler, 1977; Fowler, Rubin, Remez, & Turvey, 1980)—that identifies functional groupings of muscles as exhibiting properties qualitatively similar to a nonlinear oscillatory system. The bottom line of this model and of the present paper is that the equifinality characteristic of vowel production may not be prescribed by closed-loop servomechanisms of the peripheral or central kind. Rather, we argue that it may be a consequence of the parameterization...
of a dynamical system whose design is intrinsically self-equilibrating. That is, a design in which equilibrium points are a natural by-product of the stiffness and damping specifications for the vowel-producing system.

Subjects. Four female volunteers were paid to participate in this experiment. All were naive to the purpose of the experiment. A fifth subject (male) who was phonetically trained and had prior experience in a similar experiment (see Gay & Turvey, 1979) was also included.

Stimuli. The subjects' task was to say the point vowels /i, a, u/ in isolation and in a /p/-vowel-/p/ context. The /pVp/ syllables were spoken in the carrier phrase "A ___ again." Utterances were produced in three groups of three tokens of a particular vowel or phrase. The subjects were instructed to produce all tokens of a given utterance in exactly the same fashion, with a clear pause after each token. They were also told not to talk between experimental conditions or to practice the production task.

Conditions. The bite block used was a small acrylic cylinder with wedges carved out of each end so that it could fit snugly between the teeth. A 5 mm bite block was used to restrict the normally low jaw position for production of /a/ and /pap/. Either a 17 mm or a 23 mm bite block was used (depending on the individual subject's oral dimensions) for production of /i, u, pip, pup/, which normally involve a high jaw position.

All anaesthetic procedures were performed by Dr. Robert Gross, a specialist in oral and maxillofacial surgery who had collaborated with us in earlier work (Tuller, Harris, & Gross, 1981). Tactile information from the oral mucosa was reduced by spraying the surface of the tongue and oral cavity with a 2% Xylocaine solution. The effectiveness of the topical anaesthesia was tested by pricking the surfaces with a needle until the subject no longer reported sensation. A few catch trials were also included in an attempt to insure honest reporting on the part of the subject. Information from mechanoreceptors in the jaw was reduced by injecting percutaneously a 2% Xylocaine solution directly into left and right temporomandibular joint capsules to achieve auriculotemporal nerve blockage. Chemical blockage of this nerve drastically impairs perception of joint position and movement (cf. Thilander, 1961). This condition will be referred to as the TMJ block.

In order to restrict the availability of auditory information, white noise was presented to the subject over headphones at approximately 90 dB. The subject was told to monitor the amplitude of her or his productions by watching a VU meter and to restrict the excursion of the needle to approximately 55 dB or under.

All subjects spoke with and without the bite block prior to the application of anaesthesia and under all experimental conditions. Two of the four naive subjects received the TMJ block before the topical anaesthesia, and the other two subjects underwent topical anaesthesia first. In each of these pairs, one subject spoke under conditions of auditory masking and the other subject was allowed normal auditory information. The phonetically trained subject received topical anaesthesia before the TMJ block and spoke with masking noise in combination with these two procedures.
Measurement procedure. Individual utterances were input through a Ubiquitous spectrum analyzer to a Honeywell DDP-224 computer, using a 12.8 msec window and 40 Hz frequency resolution. The first and second formants of each utterance were measured from a spectral section display. As in previous experiments (e.g., Lindblom et al., 1979; Fowler & Turvey, 1980), acoustic measures of the isolated vowels were made at the first glottal pulse. For many English speakers the isolated vowels may not be truly static, that is, they may show some articulatory movement and thus some shifting of formant frequencies; nevertheless, the adopted procedure was to measure formant frequencies at the first glottal pulse. For the /p/-vowel-/p/ syllables, F1 and F2 values were taken from the point within the vowel at which F2 was most extreme. This point was chosen as the closest approximation to the "target" vowel formants.

Results. The main interest of the present experiment rests on a comparison of speech under normal conditions and conditions of reduced peripheral information. Figure 1 shows the mean values for F1 and F2 for each subject. The top half shows the mean formant values for the isolated vowels, and the bottom half the mean formant values for the /p/-vowel-/p/ syllables. The conditions of speaking are coded as follows: "M" means the subject produced the utterances under conditions of masking noise, "J" is the TMJ block, "T" corresponds to topical anesthesia, and "BB" is the bite block. Each subject's nine normal productions of a given utterance were compared using t-tests with his or her productions under the most extreme condition of sensory deprivation. None of the subjects (except subject 1) showed any differences in formant frequency values between normal and deprived conditions. Such was the case regardless of whether vowels were spoken in isolation or in a consonantal frame; t(S) values ranged from .05 to 1.79, ps > .1. For Subject 1, a significant mean difference occurred between the normal and most deprived condition only for F1 and F2 of the vowel /u/. We are hard put to account for these anomalies: The effect on /u/, though substantial (a 97 Hz difference for F1 and a difference of 363 Hz for F2) is in the direction opposite to expectation. Specifically, the presence of a bite block might be expected to raise all formant frequencies when producing /u/ because of possible structural limitations on lip protrusion and constriction. In contrast, however, this subject's productions of /u/ with a bite block actually showed lower F1 and F2 frequencies than when there was no bite block. Neither can the effect be attributed only to masking (which might implicate higher formant frequencies). Notice that the formant values for combined sensory deprivation conditions are very similar with and without masking. It is also worth remarking that S5 in Figure 1 is the phonetically trained subject whose results conform to the general pattern shown by naive subjects.

Before concluding that these results reflect immediate adjustment in the deprived condition, it is necessary to exclude the possibility that systematic changes in formant values occurred over trials. Figure 2 shows the F1 and F2 values for individual tokens, in order, for the vowel /i/ produced by one subject under the most extreme conditions (i.e. topical anesthesia, TMJ block, a 23 mm bite block and masking noise). Also shown are the mean F1 and F2 values for this combined condition, and the mean value of the subjects' "normal" formants. The slope of the line formed by tokens one through nine does not differ significantly from the line formed in the (non-bite block) control condition. Evidently, there does not appear to be a systematic
Figure 1. Mean values of $F_1$ (x-axis) and $F_2$ (y-axis) for nine repetitions of the indicated utterances. Data for the five subjects are presented separately. Top: Isolated vowels. Bottom: /pVp/ syllables. M = masking noise, J = TMJ block, T = topical anaesthesia, BB = bite block.
Figure 2. \( F_1 \) and \( F_2 \) values for individual tokens, in order, of the vowel /i/ produced by one subject under the most extreme experimental condition.
learning effect occurring over trials. We confirmed this statistically for all subjects by performing linear regression analyses across trials of each subject's productions under normal and deprived conditions. Correlations were converted to z-scores and t-tests performed to determine whether the slopes differed between the two conditions. Of the sixty analyses performed (5 subjects by 6 utterance types by 2 formants), not a single one showed a difference in slope; t(7) values ranged from .00 to .84, ps > .1.

DISCUSSION

The present data are not easily explained by current models of movement control that have been proposed to account for the remarkable context sensitivity of the speech production system. Closed-loop models---of the central or peripheral kind---both entail an availability of reliable sensory information about the initial conditions of the articulators. However, our experiment shows that acoustically normal vowels can be produced not only when the normal relationships among the articulators are changed by a bite block, but also when sensory information from auditory, joint, and tactile sources is drastically reduced as well. Furthermore, and as other recent work also suggests (cf. Fowler and Turvey, 1980), "articulatory compensation" appears to be achieved immediately and with little or no practice; none of our naive subjects' data provided any evidence of short-term adaptation. In support of the latter claim, the data displayed in Figure 2 are actually from the same subject that appeared to display motor learning effects in an earlier study (Gay & Turvey, 1979).1

Before offering an alternative interpretation of our data in terms other than closed-loop models, two caveats may be in order. The first is that our results do not necessarily refute closed-loop control when the system is in its normal mode, that is, when all sources of information are available. The second is that our paradigm in all likelihood does not completely eliminate peripheral information, and hence a closed-loop simulation model cannot be ruled out completely. Nevertheless, given the drastic reduction in proprioceptive information we (and surely the proponents of closed-loop models) might have expected much more severe distortion of the acoustics than was observed here.

In spite of these caveats, we believe that a more parsimonious account of the phenomenon can be forwarded, though it is less well known in speech research than the servoengineering model. The account that we shall consider does not, in fact, depend on whether sensory input about the initial positions of articulators is available or not. Thus there is no requirement for one model when sensory afference is available and another quite different model when it is absent.

The view that we shall express for the present data has been laid out in some detail elsewhere (Fowler et al., 1980; Kelso, Holt, Kugler, & Turvey, 1980). In brief, it argues that functional groupings of muscles---sometimes called synergies (cf. Gelfand, Gurfeinkel, Tsetlin, & Shik, 1971) or coordinative structures (cf. Turvey, Shaw, & Mace, 1978)---exhibit behavior qualitatively similar to a (nonlinear) mass-spring system. Such systems are intrinsically self-equilibrating in the sense that the "end-point" of the
system, or its "target," is achieved regardless of initial conditions. Thus in normal and deafferented animals (Bizzi, Dev, Morasso, & Polit, 1978), it can be shown that desired limb positions are attainable without starting position information, and even when the limb is perturbed on its path to the target. Similarly, the localization ability of functionally deafferented humans (Kelso, 1977; Kelso & Holt, 1980) and individuals with the metacarpophalangeal joint capsule surgically removed (Kelso, Holt, & Flatt, 1980) is not affected by altered initial conditions or unexpected perturbations. These data have led to the view that the "target" of the system is not achieved by means of conventional closed-loop control; rather it is a consequence of the system's dynamic parameters (mass, stiffness, damping). In such a model, the only parameters that need be specified for voluntary movement are stiffness and resting length: Kinematic variations in displacement, velocity, and acceleration are consequences of the parameters specified, rather than controlled variables, and sensory "feedback"—at least in the conventional computational sense—is not required (cf. Fitch & Turvey, 1978; Kelso, Holt, & Flatt, 1980; Kelso, Holt, Kugler, & Turvey, 1980).

It is worth noting that the view expressed above is equally applicable to disruptions that are static and anticipated (as in the present bite block experiment), and those that are time varying and unanticipated. For example, recent studies of the latter kind have shown that "compensatory responses" of short latency are observed in perturbed articulators as well as in others that contribute to the same "vocal tract goal" (cf. Abbs, 1979, for review). Current theorizing, however, offers two distinct mechanisms to explain the system's reaction to perturbation: A predictive simulation mechanism for anticipated disruptions (Lindblom et al., 1979) and a closed-loop peripheral feedback mechanism for unanticipated disruptions (Abbs, 1979).

The analysis offered here views such a distinction as redundant. Immediate adjustment to either type of perturbation is a predictable outcome of a dynamical system in which muscles function cooperatively as a single unit. If the operation of certain variables is fixed, as in the bite block, or unexpectedly altered, as in online perturbation, linked variables will automatically assume values appropriate to the constraint relation (as long as biomechanical limitations are not violated). In short, dynamical systems (of which speech is a member) always operate in a mode that one can describe as "compensatory."

Although we cannot offer a detailed description of the muscles of the vocal tract in terms of the style of control outlined above, we believe there are some grounds for optimism. Fujimura and Kakita (1979), for example, have performed a three-dimensional simulation of the tongue that uses quantitative control of contractile forces of the muscles actually involved. By treating the tongue muscles (in this case the posterior and anterior portions of genioglossus) as a cooperative unit and maintaining the relative magnitude of contractile inputs to each muscle, it can be shown that the acoustic pattern for the vowel /i/ is obtainable with a wide variety of absolute force values. Thus, as long as the contractile balance among linked muscles is preserved, the exact magnitude of muscle contraction (beyond a critical value) does not matter (see also Kakita & Fujimura, 1977). The generality of this model is limited, at this time, to a single point vowel. Nevertheless, the nonlinear relationship between muscle forces and acoustic pattern allows, or rather
provides for, a context-conditioned production system. As in recent accounts of limb localization, invariant "targets" can be attained with different stiffness specifications, as long as the balance in stiffness among relevant muscles is preserved.

As a final point, the analysis offered here suggests a commonality in function between the system capable of producing vowels and that involved in the attainment and maintenance of limb postures. Both systems are materially distinct from each other but share behaviors qualitatively like a nonlinear mass spring. The nontrivial claim, then, is that speech and limb movements are dynamically alike in sharing a common solution to the equifinality problem.

REFERENCES


FOOTNOTE

1 Indeed it was after observing the bite block performance of our naive subjects under reduced information conditions that this person offered to participate in the present experiment. This was a magnanimous gesture for which we express our gratitude.