Reprinted from

DYNAMIC ASPECTS OF SPEECH PRODUCTION

Current Results, Emerging Problems, and New Instrumentation

edited by
MASAYUKI SAWASHIMA and FRANKLIN S.COOPER

UNIVERSITY OF TOKYO PRESS, 1977

THE STUDY OF ARTICULATORY ORGANIZATION: SOME NEGATIVE PROGRESS

Katherine S. Harris

Department of Speech and Hearing Sciences, Graduate School and University Center, City University of New York; and Haskins Laboratories

This paper is an attempt to summarize what we now know, or rather don't know, about a vaguely defined area called "the organization of speech." In particular, the topic is what MacNeilage has called the "reality status of concepts of linguistic units" (MacNeilage, 1973).

The study of the organization of speech is the province of speech science, a rather uneasy blend of elements from phonetics and motor physiology. Perhaps I can illustrate the mixture with an anecdote quoted from Granit's (1976) biography of the great neurophysiologist, Sherrington. He is recorded as saying to his student, Wilder Penfield, "It must be nice to hear the preparation speak to you." Our hypotheses about speech organization came partly, then, from phonetics, and partly from general neurophysiology. I would like to run through four of these hypotheses, in their argument form, and discuss some evidence that has caused them to fail. The first hypothesis runs:

HYPOTHESIS 1: Speech is perceived as having invariant units; therefore, perception must operate on invariant parts of the acoustic signal.

This hypothesis, or versions of it, guided early work at the Bell Telephone Laboratories. It has two obvious problems. The first is that the acoustic signal for a given speech sound and for a given speaker depends on the size of the vocal tract.

This problem led Peterson (1952, 1961), and later Gerstman (1968) to suggest that the listener arrives at vowel judgments by some kind of perceptual normalization of the presented vowel, based on the relationship between formants. There are some problems with this theory as a dynamic hypothesis, as we will discuss below, but refinements of the theory will account for differences between steady-state formant values of the vowels for different speakers.

A second problem, and at that time an apparently more serious one, was that when convenient visual displays for the acoustic speech signal became widely known, no units corresponding to phonetic entities were obvious (Potter, Kopp, and Green, 1947). A suggestion made by them was that perception is organized to focus on the relatively steady-state aspects of the signal, skipping over the variable "transitional" stretches between those steady states. Indeed, Cyril Harris (1953), then working at Bell, attempted to synthesize speech by putting together short segments of speech clipped from the ongoing stream. The result was unintelligible.

I don't believe we have yet learned quite enough from that failure. Even at the time, a different interpretation of the transitional portions was available, namely, that these transitions were essential for speech-intelligibility, since they could be shown to have cue value, particularly for the consonants. This interpretation had, indeed, been demonstrated directly in work with speech synthesis (Cooper, Delattre, Liberman, Borst, and Gerstman, 1952). There were later attempts to "explain" the speech perceptual mechanism by more complicated hypotheses, as we will be doing below.

HYPOTHESIS 2: Speech is perceived in terms of invariant units. It can be shown that there are few steady-state segments in speech. Hence, speech perception must process the signal to extract invariant perceptual units from a time-varying signal.

This is one version of the Haskins "motor theory," stripped

of its physiological detail (Liberman, Cooper, Harris, and MacNeilage, 1963). Basically, the idea is that, in production, there is temporal and spatial smear of various low-level aspects of the articulatory process, so that the resulting acoustic output is an encipherment of the input signal; further, in perception, the perceptual apparatus somehow decodes the signal, by reference to articulation, into its underlying units. There are a number of sub-hypotheses, of varying degrees of sophistication, about what these underlying units might be (Harris, 1976).

Invariant electromyographic signals: There were some early Haskins attempts to show that the signals to the muscles were less variable than the resulting acoustic outputs (Harris, Lysaught, and Schvey, 1965; MacNeilage, 1963; Cooper, Liberman, Harris, and Grubb, 1958; Cooper, 1965). Apart from the difficulty of testing the proposition that one type of unit is less variable than another, the hypothesis suffers from the fact that, in the form stated, it ignores the variations in muscle signal size associated with the different distance through which articulators must travel when different phonetic units are juxtaposed. This point was discussed by MacNeilage (1970), who observed that coarticulation effects on muscle signals, due to this effect, are ubiquitous.

Articulatory targets: The point of view that articulatory movement "aims at" articulatory targets is the view espoused by MacNeilage in the paper cited above. He suggests that the targets are maintained by some form of feedback from the periphery, as does Abbs (1973).

Acoustic targets: A variant of this view, advanced by Ladefoged (1967) and Lieberman (1973), among others, is that speakers aim at acoustic targets, which can be realized by different articulatory maneuvers, depending on context or speaker.

Closely related views have been developed for somewhat different ends by Lindblom (1963) and by Öhman (1967). Lindblom,

in explaining vowel neutralization in rapid or distressed speech, suggested that invariant signals are sent to the articulators for a given phoneme target, but that the target is not always attained because the next signal may be sent too soon, cuasing target undershoot. Öhman (1967), in attempting to account for phonetic context effects, suggests that they arise from the temporal overlap of movements towards target positions. Lindblom has developed a very similar inertial view of speech timing effects (Lindblom, 1967) to account for differences in the inherent duration of vowels.

All these models have a common view of the speech process: peripheral encoding is believed to account for coarticulation of signals which are invariant at a central stage in the articulatory process. To the extent that these models specify a perceptual process, they assume, either explicitly or implicitly, that perception proceeds by reversing the encoding operations of production (Lindblom and Studdert-Kennedy, 1967).

Recent evidence suggests that this is a misleading picture. Strange, Verbrugge, Shankweiler, and Edman (1976) presented listeners with sets of natural vowels, either alone or in consonantal context. They found that identification was better when the vowels were in context. If perception were indeed a steady-state target-extracting process of any kind, it would be hard to explain these results. When a listener is presented with vowels in steady-state form, they are presumably already "at target." When they are presented in CVC context, the vowel target must usually be inferred, due to undershoot or similar adjustments. Since no decoding is required in the former case, listeners should be maximally accurate. The fact that they perform less well calls into question any of the "target-extraction" formulations.

HYPOTHESIS 3: Speech has invariant units. These units are maintained in adult speakers by some form of nonacoustic feedback from the periphery.

Hypotheses about the role of feedback in speech have been with us for some time, although the literature has not always been explicit about what kind, or kinds, of feedback is crucial, as between gamma-loop feedback (Abbs, 1973) or tactile and kinesthetic feedback (Ringel and Steer, 1963). However, in spite of the general importance of the topic, there seem to be substantial roadblocks in the path of finding out more by the means presently available. Three approaches have been used:

First, there have been a number of studies involving reduction of oral tactile sensation, most notably those of Ringel and his associates (e.g., Ringel and Steer, 1963). In general, these experiments show that the effects of blocking various branches of the trigeminal nerve are not overwhelming (Scott and Ringel, 1971; Borden, Harris, and Catena, 1973). Furthermore the experimental procedure causes motor, as well as sensory, effects, so that the results are difficult to interpret (Borden, Harris, and Catena, 1973; Abbs, Folkins, and Sivarajan, 1976).

A second approach has been to scan the relevant neurophysiological literature, in order to find an appropriate animal model for the human speech situation. While it is difficult for an amateur to assess the work, there does not seem to be an entirely appropriate analog for speech, and results are conflicting with regard to the importance of various types of feedback for various kinds of movement in the examples discussed. For example, animals can use deafferented limbs in learned or unlearned tasks (Taub and Berman, 1968) although some deterioration of fine motor control is generally found. On the other hand, lesions of the tract of the mesencephalic nucleus, which abolishes spindle afferent input from the masticatory muscles (Goodwin and Luschei, 1974), does not alter chewing behavior in any obvious way.

The third approach has been to study the effects of disruptions of articulation. Here again, there is no solid body of relevant experimentation and results are often conflicting.

Folkins and Abbs (1975), for example, have shown that speakers can compensate immediately for the effects of unexpected interruptions of articulator movement. In their experiment, the jaw was unexpectedly loaded during the closure for a bilabial stop consonant. Results show that the lips compensate for the jaw in completing closure on the first trial. Another often cited study by Lindblom and Sundberg (1971) reports that a speaker can duplicate his natural vowels with a bite block between his teeth, with virtually no time for relearning. However, the only citation of the study I know is an oral report, with no experimental details. Hamlet and Stone (1976), using a different experimental paradigm, find compensatory effects over fairly substantial periods.

If articulation is interfered with in some way, the speaker may use either acoustic or nonacoustic feedback to compensate for the disruption. It has been suggested by Nooteboom (1970) that compensatory articulation may well be guided by acoustic rather than articulatory equivalence. If so, we would expect devastating effects of articulatory disruption accompanied by acoustic masking. So far as I know, this line of research is unexplored. Surely the disruption experimental paradigm is eligible for far more searching exploration than it has thus far received.

HYPOTHESIS 4: Speech has units, but we would understand its organization better if we turned from phones to more appropriate units, such as:

Syllables. The evidence against Kozhevnikov and Chistovich's syllable-based model of coarticulation (Kozhevnikov and Chistovich, 1965) is, in large part, a product of the industry of Kenneth Moll and his students (e.g., Daniloff and Moll, 1968; McClean, 1973), although there has been recent interesting and important work by Benguerel (Benguerel and Cowan, 1974). These studies all show that there is little evidence that the syllable boundary, as traditionally defined, blocks coarticu-

lation. Benguerel interprets his results as supporting a feature-based model of coarticulation, such as that of Henke (see below).

Features. This is not the place for an exposition of the virtues of feature-based models in general. However, some recent experiments argue against feature-based models of anticipatory coarticulation. Tom Gay, in his paper at this conference, will be discussing evidence that phonetic entities are separately organized at the electromyographic level, even when there is no reason for it in feature terms. Perhaps even more important is evidence that specifically contradicts Henke's "scan-ahead" model for articulatory coarticulation (1967). The model proposes that a given feature will appear in the speech stream as soon as it can, by assimilative spreading. Thus, if a nasal is preceded by a series of vowels which are unspecified for nasalization, they should all be equally nasalized. This does not happen (Kent, Carney, and Sevareid, 1974; Ushijima and Hirose, 1974). The degree of nasality of a vowel, as measured by velar height during its production, depends on its proximity to a nasal consonant.

Some recent results of our own can be interpreted in the same way (Bell-Berti and Harris, 1976). We examined anticipatory coarticulation of lip rounding for /pɑsup/, pɑtup/, /pɑtsup/ and /pɑstup/, measured by the electromyographic activity of the orbicularis oris muscle. We find that the onset of electromyographic activity seems to precede the onset of acoustic activity for the vowel by a fixed temporal interval, rather than to be locked to the preceding phone or cluster of phones.

Another bit of evidence arguing against a feature-based model of coarticulation comes from an experiment on Swedish rounded vowels (McAllister, Lubker, and Carlson, 1974), again using the onset of orbicularis activity as a measure of anticipatory coarticulation of lip-rounding. They compared onsets of a series of front and back rounded vowels (both of which

occur in Swedish) in the frame /itV/. If Henke's model were correct, lip rounding should begin at the same time, relative to the offset of /i/, for all vowels, since it is the feature composition of the preceding phones which determines the onset of anticipatory coarticulation. Interestingly enough, the onset of labial activity is later for the back vowels than for the front vowels, so that the lips seem to "wait for the tongue," which must move further for back vowels than for front vowels. In short, the temporal extent of anticipatory coarticulation cannot be predicted from a knowledge of the feature composition of the phones before the target.

One explanation of these data is that articulatory gestures are programmed temporally, and not in syllable or feature units; however, we have yet to determine the influence of stress and speaking rate on this timing. In addition, we must also examine the timing relationships between movements of different articulators since we may find that sub-parts of segment gestures preserve their timing relationships.

Overall, given this rather negative review of our progress, what can we propose in a more positive direction? I can only offer a suggestion by my colleague, Michael Turvey (Turvey, Shaw, and Mace, 1977), who points out in reviewing recent Russian studies of locomotion that all skilled movements have sub-parts which tend to preserve their relationships to one another when the movement is transformed as by more rapid execution. He gives as an example the observation, by Kent, Carney, and Sevareid (1974), that velar lowering and raising in the word contract is tied to particular events in the sequence of tongue movements. Whether this particular example suggests a useful experimental paradigm or not, it emphasizes a kind of observation we have been neglecting in studies of speech production; that is, what relationships between articulatory events are preserved when context changes, whether by increased speaking rate or stress, or segmental environment? Furthermore, how do the perceptual consequences of this view

differ from those of a target extraction approach? Perhaps, when we can formulate experimental questions in terms such as these, we will be able to make progress in understanding speech organization.

Acknowledgment

This work has been supported in part by a grant from the National Institute of Dental Research.

REFERENCES

Abbs, J. H. (1973) The influence of the gamma motor system on jaw movements during speech: A theoretical framework and some preliminary observations. J. Speech Hearing Res., 16; 175-200.

Abbs, J. H., Folkins, J. W. and Sivarajan, M. (1976) Motor impairment following blockade of the infraorbital nerve: Implications for the use of anethetization techniques in speech research. J. Speech Hearing Res., 19; 19-35.

Bell-Berti, F. and Harris, K. S. (1976) EMG study of the coarticulation of lip rounding. J. Acoust. Soc. Am., 60; S 63 (A).

Benguerel, A-P and Cowan, H. A. (1974) Coarticulation of upper lip protrusion in French. Phonetica, 30; 41-55.

Borden, G., Harris, K. S. and Catena, L. (1973) Oral feedback II: An electromyographic study of speech under nerveblock anesthesia. J. Phonetics, 1; 297-308.

Cooper, F. S. (1965) Research techniques and instrumentation: EMG. ASHA Reports No.1, 153-167.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M. and Gerstman, L. J. (1952) Some experiments on the perception of synthetic speech sounds. J. Acoust. Soc. Am., 24; 597-606.

Cooper, F. S., Liberman, A. M., Harris, K. S. and Grubb, P. M. (1958) Some input-output relations observed in experiments on the perception of speech. Proceedings, 2nd International Congress on Cybernetics (Namur), 930-941.

Daniloff, R. and Moll, K. L. (1968) Coarticulation of lip-rounding. J. Speech Hearing Res., 11; 707-721.

Folkins, J. and Abbs, J. H. (1975) Lip and jaw motor control during speech: responses to resistive loading of the jaw. J. Speech Hearing Res., 18; 207-221.

Gerstman, L. J. (1968) Classification of self-normalized vowels. IEEE Trans. Audio Electroacoust. AU-16, 78-80.

Goodwin, G. M. and Luschei, E. S. (1974) Effects of destroying spindle afferents from jaw muscles on mastication in monkeys. J. Neurophysiology, 37; 967-981.

Granit, R. (1967) Charles Scott Sherrington. (Garden City: Doubleday.)

Hamlet, S. L. and Stone, M. (1976) Compensatory vowel characteristics resulting from the presence of different types of experimental dental prostheses. J. Phonetics, 4; 199-218.

Harris, C. M. (1953) A study of the building blocks of speech. J. Acoust. Soc. Am., 25; 962-969.

Harris, K. S. (1976) Physiological aspects of speech production. Haskins Laboratories Status Report on Speech Research SR-48, 21-42.

Harris, K. S., Lysaught, G. F., and Schvey, M. M.,(1965) Some aspects of the production of oral and nasal labial stops. Lang. Speech, 8; 135-147.

Henke, W. (1967) Preliminaries to speech synthesis based on an articulatory model. Conference Preprints; 1967 Conference on Speech Communication and Processing (Bedford, Mass.: Air Force Cambridge Research Laboratories), 170-177.

Kent, R. D., Carney, P. J.,and Severeid, L. R. (1974) Velar movement and timing: Evaluation of a model of binary control. J. Speech Hearing Res., 17; 470-488.

Kozhenikov, V. A. and Chistovich, L. A. (1965) Rech', Artikulyatsiya, i Vospriyatiye. Trans. as Speech: Articulation and Perception, 1966. Washington, D. C.: Joint Publications Research Service, 30, 543.

Ladefoged, P. (1967) Three Areas of Experimental Phonetics. (London: Oxford University Press).

Liberman, A. M., Cooper, F. S., Harris, K. S.,and MacNeilage, P. F. (1963) A motor theory of speech perception.

Proc. Stockholm Speech Comm. Seminar, Vol.II (Stockholm; Royal Institute of Technology).

Lieberman, P. (1973) On the evolution of language: A unified view. Cognition, 2; 59-94.

Lindblom, B. E. F. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Am., 35; 1773-1781.

Lindblom, B. E. F. (1967) Vowel duration and a model of lip mandible coordination. Quarterly Progress and Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm) STL-QPSR 4/1967, 1-29.

Lindblom, B. E. F. and Studdert-Kennedy, M. (1967) On the role of formant transitions in vowel recognition. Haskins Laboratories Status Report on Speech Research SR-9, 7.1-7.2.

Lindblom, B. E. F. and Sundberg, J. (1971) Neurophysiological representation of speech sounds. Paper presented at the XV World Congress of Logopedics and Phoniatrics, Buenos Aires, Argentina.

MacNeilage, P. F. (1963) Electromyographic and acoustic study of the production of certain clusters. J. Acoust. Soc. Am., 35; 461-468.

MacNeilage, P. F. (1970) Motor control of serial ordering of speech. Psychol. Rev., 77; 182-196.

MacNeilage, P. (1973) Linguistic units and speech production theory. J. Acoust. Soc. Am., 54; 329(A).

McAllister, R., Lubker, J. and Carlson, J. (1974) An EMG study of some characteristics of Swedish rounded vowels. J. Phonetics, 2; 267-278.

McClean, M. (1973) Forward coarticulation of velar movement at marked junctural boundaries. J. Speech Hearing Res., 16; 286-296.

Nooteboom, S. F. (1970) The target theory of speech production. IPO Annual Progress Report 5, Institute for Perception Research, (Eindhoven, Holland) 5; 51-55.

Öhman, S. E. G. (1967) Numerical model of coarticulation. J. Acoust. Soc. Am., 4; 310-320.

Peterson, G. E. (1961) Parameters of vowel quality. J. Speech Hearing Res., 4; 10-29.

Peterson, G. E. and Barney, H. L. (1952) Control methods used in a study of the vowels. J. Acoust. Soc. Am., 24; 175-184.

Potter, R. K., Kopp, G. A. and Green, H. C. (1947) Visible Speech. (New York: Van Nostrand).

Ringel, R. L. and Steer, M. D. (1963) Some effects of tactile and auditory alterations on speech output. J. Speech Hearing Res., 6; 369-378.

Scott, C. M. and Ringel, R. L. (1971) Articulation without oral sensory control. J. Speech Hearing Res., 14; 804-818.

Strange, W., Verbrugge, R. R., Shankweiler, D. P. and Edman, T. R. (1976) Consonant environment specifies vowel identity. J. Acoust. Soc. Am., 60; 213-224.

Taub, E. and Berman, A. J. (1968) Movement and learning in the absence of sensory feedback. In The Neuropsychology of Spatially Oriented Behavior, ed. by Freeman, S. J. (Homewood, Ill.: Dorsey).

Turvey, M. T., Shaw, R. E. and Mace, W. (1977) Issues in the theory of action. Haskins Laboratories Status Report on Speech Research SR-48, 1-20.

Ushijima, T. and Hirose, H. (1974) Electromyographic study of the velum during speech. J. Phonetics, 2; 315-326.

DISCUSSION

Fujisaki: I agree with the author that it is possible to find counterevidence to each of the four hypotheses stated in this paper, but I also feel that it does not mean a total failure of these hypotheses or their underlying concepts. What may possibly be true is that speech production and speech perception are based, not on either one of these oversimplified principles, but on a multiplicity of principles, which need not be mutually contradictory, but are complementary to each other.

Harris: I certainly agree that all of the hypotheses advanced (and argued against) in the paper are oversimplified. I think that the point of the paper is that the models of articulation we have been working with are demonstrably inadequate.

Strube: In spite of the highly complicated coarticulation effects, good vocoder speech can be obtained by simply interpolating pseudo-area or PARCOR parameters across quasi-stationary and transitional speech intervals (Olive, and present work at Göttingen). This might indicate that exact representation of coarticulation effects is not essential for speech quality. It might also support Gay's findings that effects outside a syllable are not essential.