SPEECH PRODUCTION AT THE NEURO-MUSCULAR LEVEL*

S. ÖHMAN*—A. PERSSON**—R. LEANDERSON***

The phonological output of a generative grammar may be regarded as an abstract description of certain subjectively observable aspects of speech production. The question as to what form the outputs of the phonological component should take is thus dependent on the nature of the physical parameters controlling speech production as well as on the properties of subjective phonetic observation. The purpose of the present paper is to emphasize the rather high degree of abstraction involved in even the lowest level phonological representations of human utterances

Modern acoustic phonetics has provided a firm understanding of the physical basis relating articulatory configurations to the resulting sound pressure wave. On the other hand, research at this level has demonstrated an overwhelming inconstancy in the acoustic correlates of the phonological invariants. Studies of articulatory dynamics suggest, however, that much of this variability may be due to built-in physiological properties of the articulatory organs and of the neural circuits controlling them. A model of speech production adequate for phonological purposes should thus incorporate a neuro-motor level exhibiting these properties.

An attempt to devise a partial model of this sort is summarized in Fig. 1. The upper left corner shows schematized sound spectrograms of the VCV utterances */ø:gø:/*, */ø:ga:/*, */a:gø:/*, and */a:ga:/* as spoken by a male Swedish talker. These data have been extracted from a larger study described elsewhere.[1]

Note that the formant transitions from the initial */ø/* into the medial */g/* are different when */ø/* and */a/* occupy the final position of the utterance. Similarly, the transitions from the medial */g/* into the final */a/* are different when */ø/* and */a/* occupy the initial position of the utterance. Thus, as was also observed by Menzerath and Lacerda,[2]

* Speech Transmission Laboratory, Dep. of Speech Communication, Royal Institute of Technology, Stockholm.

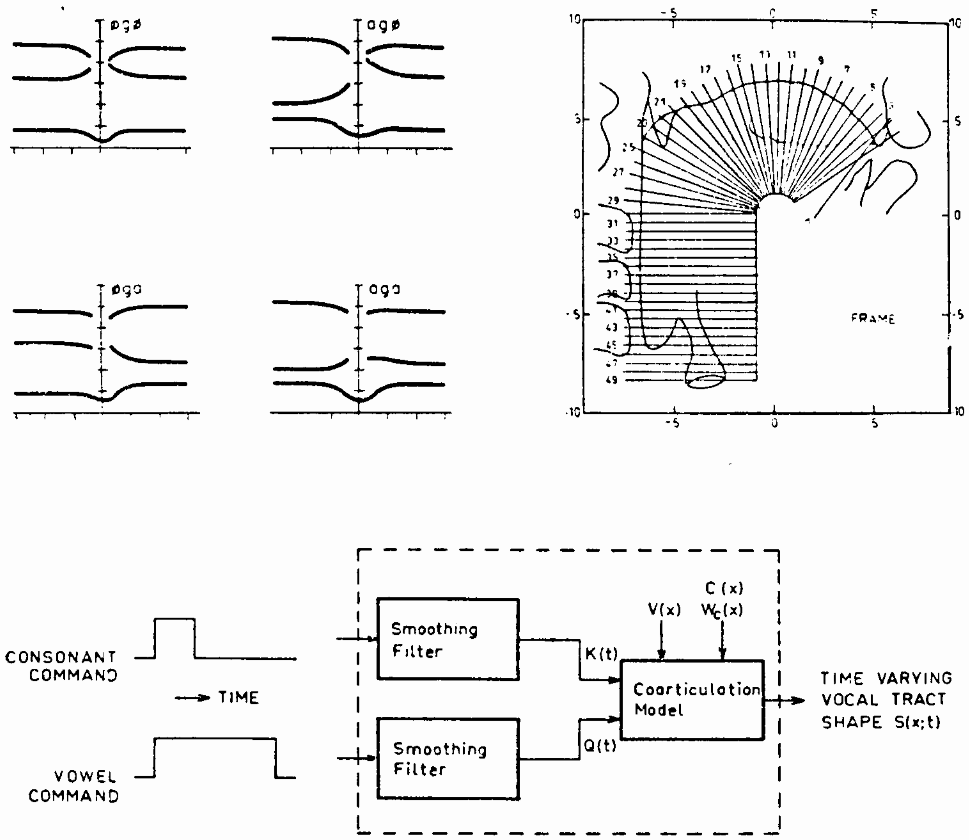
** Central Neurophysiological Laboratory, Karolinska Sjukhuset, Stockholm.

*** Phoniatric Clinic, Karolinska Sjukhuset, Stockholm.

[1] Öhman, S. E. C.: "Coarticulation in VCV Utterances: Spectrographic Measurements", *J. Acoust. Soc. Am.*, 39 (1966), 151—168.

[2] Menzerath, P. and de Lacerda, A.: *Koartikulation, Steuerung und Lautabgrenzung* (Berlin 1933).

the production of an intervocalic consonant is greatly modified by the vowel context. In fact, the variability of the formant transitions contained in the initial vowel suggests that the speaker starts a gesture towards the final vowel while he is making the consonant gesture. It is as if the consonant gesture is superimposed on a diphthong movement.



$$S(x;t)=V(x)+K(t)[C(x)-V(x)]W_c(x)$$

$$V(x)=(1-Q(t))V_1(x)+Q(t)V_2(x)$$

Fig. 1. Numerical Model of Coarticulation.

The block diagram of the lower part of Fig. 1 summarizes a strategy for the synthesis of time varying vocal tract shapes that reproduce coarticulation. This strategy has been implemented on a digital computer and tested against data collected from X-ray motion pictures of a human talker. Without going into details that have been published elsewhere,[3] I should like only to draw attention to the separate representations of the commands for the timing of the vowels and the consonants, and the commands specifying *which* consonants and vowels are to be synthesized. Target configurations for the initial vowel, the medial consonant, and the final vowel are fed to the model in the form of sets of numbers representing distances along the coordinate lines shown in the upper right corner of the figure. The consonant and vowel timing pulses shown on the left of the block diagram are then passed through the smoothing filters and enter the coarticulation model as one-dimensional signals marked K(t) and Q(t). A VCV gesture complex is then calculated according to the formulas shown below the block diagram. In this process the vowel diphthong ge-

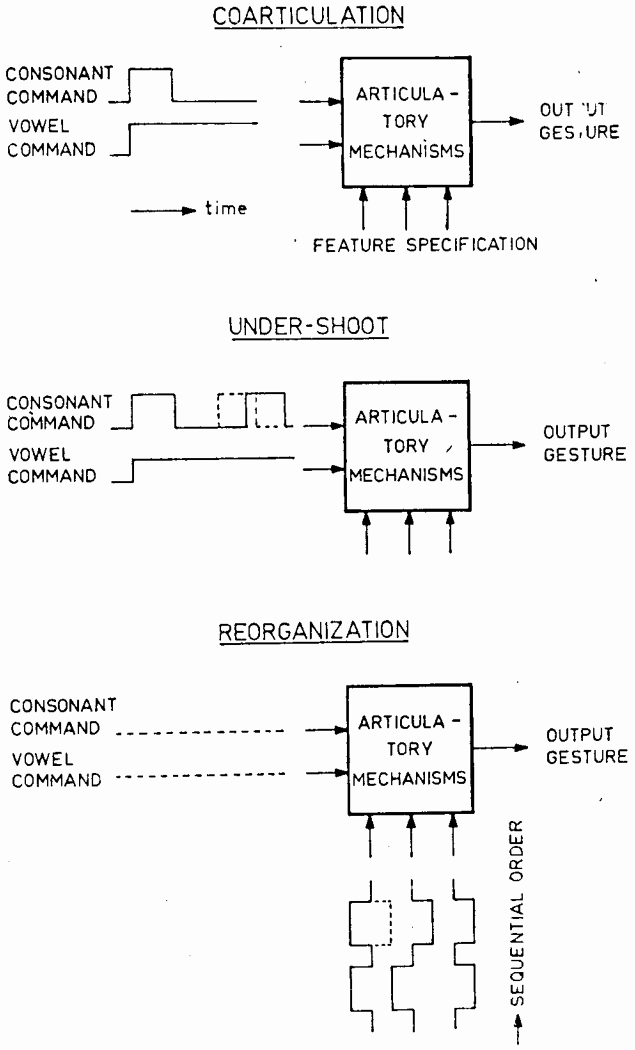


Fig. 2. Causes of Phonetic Variability.

[3] Öhman, S. E. G.: "Numerical Model of Coarticulation", *J. Acoust. Soc. Am.*, 41 (1967) 310—320

sture is governed by Q(t) and the superimposed consonant gesture is governed by K(t). The calculation is done in such a way that, at the moment of consonantal closure, a residue of the underlying vowel gesture is always present in the vocal-tract configuration, so that the consonant becomes colored by the vowel environment. Hence variance is reproduced at the output while invariance is preserved at the input.

The articulatory mechanisms enclosed by dashed lines in Fig. 1 have been represented by a single "black box" in Fig. 2. Here, again, the timing commands are fed over two separate channels corresponding to consonants and vowels, and they are responsible together for the temporal integration of the syllable. The feature commands, on the other hand, specify the target configurations that the initial, medial, and final gestures of the syllable are aiming at.

In terms of this model the phonetic variance of phonological entities as observed at the acoustic level may be related to three types of physiological processes: *coarticulation*, *undershoot*, and *reorganization*.

We have already discussed coarticulation and this concept is illustrated again in the uppermost part of Fig. 2. Coarticulation results when the articulators are moving in response to distinct but temporally over-lapping commands.

Undershoot is indicated in the middle part of the figure by the difference in timing of the second consonant pulse. Undershoot thus results when an incomplete articulatory gesture is interrupted by a neural command that brings about the next gesture of the utterance. Examples of undershoot are found in the neutralization of vowels and consonants under increased rates of speech, as demonstrated by Lindblom.[4] This type of variance may be accounted for quantitatively in terms of the numerical model discussed in connection with Fig. 1.

Reorganization, finally, is the result of a context dependent change of the feature specification, as illustrated in the bottom part of Fig. 2. This sort of effect is found, for example, in the devoicing of final voiced consonants in German and Russian, and in the quality alternations of vowels under vowel-harmony in a great many languages.

With respect to the underlying neural control there is thus an essential difference between the phonetic variance due to reorganization on the one hand and the variance typified by coarticulation and undershoot on the other hand. The sequence of feature specifications may be viewed as a phonological signal that modulates a phonetic carrier consisting of the standard timing pattern of the syllable. From this point of view reorganization is a perturbation of the modulating signal while coarticulation and undershoot are due to the structure of the carrier.

The model discussed so far has grown out of analyses of acoustic records and X-ray motion pictures. It is therefore of considerable interest to compare the picture of speech production summarized by the model with data obtained at the peripheral neural level by means of electromyographic methods. We shall here examine a few examples from a study in which thin concentric needle electrodes were used to record the motor unit activity from the facial muscles of a Swedish subject.

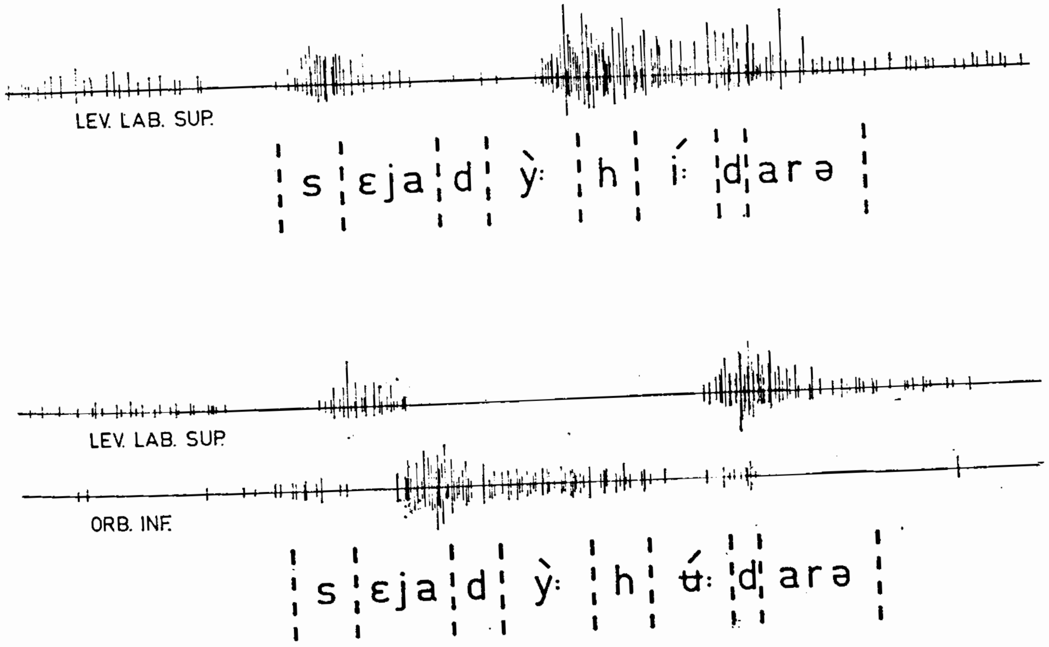


Fig. 3.

The top part of Fig. 3 shows a trace from a muscle that lifts the upper lip. In this record each motor unit potential is represented by a vertical line that indicates the amplitude of the spike. The VCV utterance */y:hi:/* is embedded in the frame */seja dỳ:hí:dare/*.

Note that this muscle is tonically active *between* the utterances of the list that the subject read in the recording session. *During* the utterances, however, this background activity is depressed or enhanced in synchrony with the rounding and spreading gestures of the lips.

This behavior is typical and shows in speech the articulators assume a basic and apparently fixed posture on top of which excitatory and inhibitory motor commands are superposed as a modulating signal. This phonetic modulation of the basic speech posture occurs, of course, at a lower level than the phonological modulation of the syllable timing carrier, discussed earlier.

The phonetic modulation is also shown in the lower part of Fig. 3. The upper trace derives from the muscle just mentioned that lifts the upper lip, and the lower trace was picked up from the muscle that shortens and protrudes the lower lip. The utterance contains the VCV sequence */y:hʉ:/*.

Note the reciprocal nature of these two signals. Whenever the lower trace shows

[4] Lindblom, B.: "Spectrographic Study of Vowel Reduction", *J. Acoust. Soc. Am.*, 35 (1963)a, pp. 1773—1781.

activity for rounding the upper trace is depressed, and vice versa. In this way the labial configuration as a whole is made to fluctuate about a constant average posture, so to speak.

Do we find invariance of motor commands at the peripheral neural level? Fig. 4 gives a negative answer to this question. The upper trace shows the motor unit activity

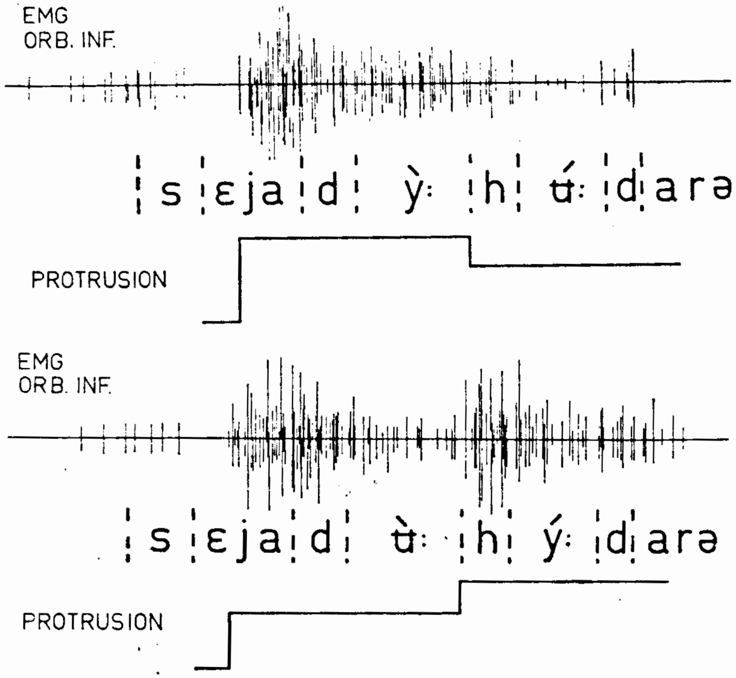


Fig. 4.

of the lower lip muscle referred to earlier. Here the VCV sequence */y:hʉ:/* is embedded in the standard frame. In the utterance of the lower trace, recorded from the same lower lip muscle, the vowels of the VCV sequence have the opposite order, */ʉ:hy:/*. The degree of protrusion for these two vowels is indicated schematically by the straight lines below the electromyographic records.

When the vowel /ʉ/ is preceded by the less protruded vowels of the frame, a transitional overshoot appears at the beginning of the motor command. This overshoot is absent when /ʉ/ follows the vowel /y/ which is more protruded. Hence the commands are not invariant. The steady state activity levels of these vowels seem to be less variable, however.

It is evident that the peripheral motor commands are calculated by the brain not only with respect to the constant effort necessary to maintain a target configuration, but also with respect to the variable effort needed to *move* the articulators from wherever they are to the desired target. It is quite likely that the last mentioned phase of the calculation takes place at a rather peripheral level in solving feedback from the many sensory receptors in the oral region.

To sum up, a model of speech production—adequate for phonological purposes—should incorporate a sequence of stages of modulation as suggested in Fig. 5. As was emphasized by Dudley[5] the acoustic sound pressure wave results from the modulation

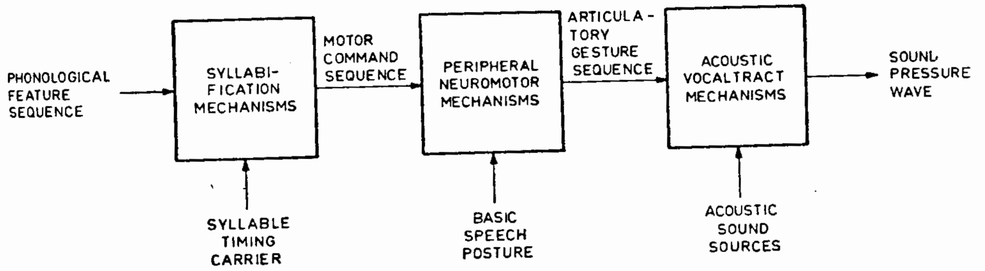

PHONOLOGICAL FEATURE SEQUENCE
SYLLABI-FICATION MECHANISMS
MOTOR COMMAND SEQUENCE
PERIPHERAL NEUROMOTOR MECHANISMS
ARTICULA-TORY GESTURE SEQUENCE
ACOUSTIC VOCALTRACT MECHANISMS
SOUND PRESSURE WAVE
SYLLABLE TIMING CARRIER
BASIC SPEECH POSTURE
ACOUSTIC SOUND SOURCES

Fig. 5. Stages of Modulation in Speech Production.

of quasi-periodic or noisy sound sources by the relatively slow articulatory gestures. The latter gesture sequence comes about through the modulation of a basic speech posture by a sequence of excitatory and inhibitory peripheral commands. These commands, finally, may be regarded as the outputs of a set of more-central neural circuits through which a complex timing carrier is channeled by means of a certain phonological signal. This signal determines specific phonetic features of the various phases of the resulting syllables.

DISCUSSION

Fant:

1. Are the muscular movements programmed in advance only or is sensory feedback of basic importance?

2. Would it be possible to discover if unstressed vowels are modification of specific phonemes or independent allophones?

3. Why do you choose the term "undershoot" instead of reduction?

Lehiste:

I have been bothered for some time by the undershoot model of vowel reduction in unstressed syllables; it appears to me that the phenomenon can be language-bound and not predictable by the model. For example, post-tonic unstressed /o/ is realized as [ə] in Russian, but as [u] in Bulgarian. The undershoot model does not explain the latter type of vowel reduction.

Tatham:

1. Give details of experimental procedure, please.

2. What relationship is there between the neural command arriving at the muscle fibre and the E.M.G. signal as measured, which is a result of the muscle fibre's reaction to the command?

[5] Dudley, H.: "The Carrier Nature of Speech", Bell System *Techn. J.*, 19 (1940), 495—515.