Philip Lieberman

Towards a Unified Phonetic Theory*

0. Introduction

In recent years there has been renewed interest in both the physiological basis of speech production and the perception of speech. Concurrently, linguists have begun to pay attention to a universal theory of grammar. We shall attempt in this paper to outline what appears to be a plausible framework for a universal phonetic theory that takes account of human physiology and human perception as well as the requirements imposed by abstract linguistic analysis. The two topics that we will discuss in detail are the nature of the ensemble of phonologic features and the relation between these features and the acoustic speech signal.

1. The Physical Bases of the Phonologic Features

Phonologic features obviously must have a concrete basis in the sound pattern of a language.[1] This does not preclude phonologic features from acting as abstract counters that are manipulated by the rules of the grammar. Features thus are "meaningful" and concrete at the phonetic level and abstract as we go into the rule structure of the phonologic component of the grammar.

What are the physical bases of the phonologic features? We have seen much work on both the acoustic correlates of phonologic features and the articulatory maneuvers that are involved in the production of speech. Jakobson, Fant and Halle (1952), for example, mapped out a hypothetical feature set in terms of acoustic correlates. Each feature was defined in terms of an invariant acoustic correlate. The feature *compactness*, for example, was defined in terms of ". . . the relative predominance of one centrally located formant region . . ." (p. 27). The feature voicing, in turn, was defined in terms of ". . . the appearance of a strong low component which is represented by the voice bar along the base line of the spectrogram" (p. 26).

* Some aspects of this paper were discussed in a preliminary manner in a paper, "On the Physical Bases of Phonologic Features," that was read at the 43rd meeting of the Linguistic Society of America, December 28, 1968.

[1] Just as semantic features must have some close relevance to categories of "meaningfulness," phonologic features must have some physical basis in the sound pattern. The phonologic level of the grammar ultimately must result in a specification of speech. Phonologic features that were unrelated to the physical aspects of speech would be as vacuous as semantic features that were unrelated to the notion of "meaning."

Reprinted from Linguistic Inquiry, Vol. 1, No. 3, July, 1970.

The features in *Preliminaries to Speech Analysis*, of course, had corresponding articulatory correlates. Voicing, for example, was the result of the articulatory maneuvers that are involved in "... the concomitant periodic vibrations of the vocal bands...." The primary emphasis in *Preliminaries*, however, is on the acoustic correlates of the distinctive features.[2] There is an implicit assumption throughout this work that it is possible to find some one-to-one mapping between a set of phonologic features and a set of invariant acoustic correlates.[3] We shall not here review the research projects that attempted to isolate invariant acoustic correlates, but the results have not been consistent with the notion of an invariant one-to-one mapping between all of the features and their acoustic correlates.

Chomsky and Halle in *The Sound Pattern of English* (1968) now emphasize the relationships between features and invariant articulatory maneuvers. The features still, of course, have acoustic consequences, but the primary emphasis is on articulatory mechanisms, and there is an explicit one-to-one mapping between some features and articulatory maneuvers. For example, the feature *flatness* in *Preliminaries* had the invariant acoustic correlate of "... a downward shift of a set of formants or even of all the formants in the spectrum" (p. 31), which could be effected by means of two alternate articulatory maneuvers, that is, either lip rounding or pharyngealization. In *The Sound Pattern of English* we find instead of the acoustically defined feature of *flattening* the articulatory defined features of *velarization*, *pharyngealization*, and *rounding*.

[2] This emphasis on the acoustic correlates of the features has important effects on the feature ensemble of *Preliminaries*. The feature *voicing* was defined in terms of "... the appearance of a strong low component which is represented by the voice bar along the base line of the spectrogram" (p. 26). In other words, a segment was +*voiced* only if the acoustic signal was periodic. Stops like /p/ and /b/ in English were, therefore, differentiated by the feature *tenseness*, since for many speakers of English the acoustic signal is not periodic during the release of the stop /b/ in initial position. If the feature of *voicing* is also defined in terms of its articulatory correlates, then it is clear that the distinction bewteen /p/ and /b/ in English is always one of −*voiced* versus +*voiced*. The articulatory implementation of +*voiced* involves, among other things, moving the vocal cords inwards by adducting the arytenoid cartilages. Lisker and Abramson (1964 and 1967), in a series of studies, have shown that when /b/ occurs in initial position the glottis is always more constricted, i.e. closer to the final position necessary for +*voiced* than it is for /p/. The constricted glottis characteristic for /b/ results in a greater air pressure drop at the larynx. Buccal air pressure is, therefore, smaller at the moment of the stop release for /b/ than it is for /p/ and the stop burst for /p/ thus has more energy than the burst for /b/. The different responses that Slavic and English listeners gave to the artificial stimuli noted in *Preliminaries* (p. 38) probably follow from timing differences characteristic of Russian and English. Russian speakers start to move their vocal cords together earlier relative to the release of the stop than English speakers do. Russian +*voiced* stops in initial position, thus, tend to produce acoustic signals that are periodic at the moment of the stop release. The difference between the stops /b, d, g/ and /p, t, k/ for Russian and English would appear to be in differences in the timing of the laryngeal articulatory maneuvers relative to the articulatory maneuvers that occur in the supralaryngeal vocal tract as the stop is released. It is not necessary to invoke an additional feature of *tenseness* to explain the differences between English and Russian stops. The difference resides in the way that the feature of *voicing* is implemented in these languages. We will return to this point.

[3] It should be stressed that the invariant acoustic correlates in this system could and were often relativistic, derived measures that, in effect, normalized the acoustic signal to take account of speaker variation and dependencies within the feature ensemble. The acoustic correlate for *gravity* thus involves a relative energy balance and as such is immune to variations in the overall speech intensity, while the acoustic correlates of compactness differ in degree for vowels and consonants.

Matching of Features to the Constraints of Speech Production and Perception

The notion of an invariant one-to-one mapping between phonologic features and particular muscles, muscle groups, or anatomical structures is quite reasonable for some features, and it would be appealing to extend it to all features. However, though this simple one-to-one mapping between a feature and its physical attributes leads to a consistent system for some features, it has proved unmanageable when features like *stop*, *vocalic*, *consonantal*, or *prominence* are examined. It is possible that these features are not appropriate ones, that is, we have selected variables that are not relevant to the phonologic level of language. There is, however, another possibility which we would like to explore here.

We propose that the physical bases of phonologic features involve both articulatory and acoustic factors and that features represent "matches" along these two dimensions. Some articulatory maneuvers are inherently easier to produce. Gestures like closing the lips or velum are "all or nothing" maneuvers, where the final position of the articulatory apparatus is automatically determined without the necessity of invoking fine positional control. The lips or velum are simply closed and it is impossible to overshoot the final intended position. Some acoustic signals are inherently easier to perceive and identify. The presence or absence of the acoustic signal is, for example, simpler to perceive than the presence of a local energy minimum in the acoustic spectrum at 700 Hz.[4] Man, like other animals, also appears to possess neural mechanisms that can detect certain acoustic signals. These neural mechanisms may be the bases of other features.

We are proposing that some phonologic features exist because their acoustic correlates "match" a particular neural acoustic detector. Other features exist because it is easy to produce a particular articulatory maneuver with the human vocal apparatus; they "match" an articulatory constraint. Still other features may represent an optimization along both of these dimensions; they may have articulatory correlates that are "easy" to produce that result in acoustic correlates that are readily perceptible. These features thus would represent an "optimal match" between the constraints imposed by the human vocal apparatus and the human perceptual system. They would be the stablest and most "central" features in a hierarchy of phonologic features structured in terms of these physical criteria.

Studies of animals like the bullfrog show that meaningful acoustic signals, like their mating calls, have characteristic acoustic properties that are determined by the vocal apparatus of these animals (Capranica 1965). These animals also have receptors in their auditory system that selectively respond to these acoustic signals (Frishkopf

[4] Local energy minima in the acoustic spectrum are difficult to detect. It is, for example, difficult to perceive when a vowel like /a/ is nasalized since the perceptual judgment then depends on the presence of local energy minima. When /a/ vowels are produced on speech synthesizing machines, the "nasal" circuits can be left in place and listeners will not notice any nasal quality. The presence or absence of the acoustic signal is, in contrast, a simple judgment that can readily be simulated by electronic devices.

and Goldstein 1963). The frog's mating call thus is a "behavioral" construct that represents an "optimal match" between the acoustic properties that his vocal apparatus imposes on the call and acoustic detectors that exist in his auditory system. We are proposing that human phonologic features are "linguistic" constructs that may be structured in terms of the properties of both the human vocal apparatus and the human perceptual system. They "match" the constraints of either or both of these physical systems.

We will look at the articulatory and acoustic aspects of a number of phonologic features to see how they may reflect on two related topics, (a) one-to-one mapping between features and articulatory maneuvers and (b) the matching of features to the constraints of the human vocal tract and auditory perceptual mechanisms.

Nasality. Let us first consider the feature *nasality*. The articulatory correlate of this feature is quite simple. The velum is relaxed as the levator palatini is laxed for +*nasal*. It is closed for −*nasal*. A "simple" articulatory maneuver appears to be the basis for this feature. In contrast, the acoustic correlates of +*nasal* are not very discernible. It is quite difficult to isolate the acoustic correlates of nasality. In extreme conditions like those that occur for pathologic cleft palate it becomes extremely hard to judge perceptually the degree or even the presence of nasality. In Figure 1 we have sketched the relationship between the phonologic feature and its muscular and articulatory correlates by means of the solid line. Note that a one-to-one relationship does exist for this feature.

Rounding. If we were to consider *rounding* as a phonologic feature, it would also have a fairly straightforward articulatory basis. The lips would always maneuver towards a vocal tract shape that was longer and had a smaller orifice at the mouth. The degree of rounding that is the consequence of this feature is, however, by no means invariant. The degree of jaw opening determines the extent to which the vocal tract will be constricted by lip rounding. Unlike the situation for *nasality*, where the same muscle always executes the relevant articulatory maneuver, different talkers appear to use different muscles to effect this articulatory maneuver. Harris and her associates (1969) in an electromyographic study find that some talkers actively use the muscles of their upper and lower lips to both initiate and release rounding while others use only the muscles of the lower lip to release rounding. In other words, different talkers use different patterns of muscular activity to effect the same articulatory maneuver. We can, therefore, schematically indicate a one-to-one relationship between rounding and an articulatory maneuver involving the lips in Figure 1, though we really cannot make any such claims at the level of muscular commands. At the muscular level we could say that the alternate patterns of muscular activity all result in particular articulatory "states" (lip rounding and unrounding). Thus we could introduce the notion of the feature's being a "state function" at the articulatory level.

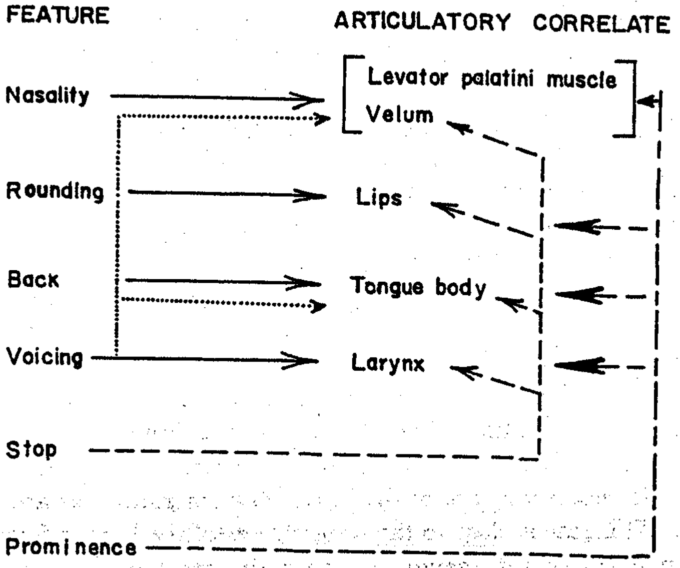

Figure 1. Features and their articulatory correlates. The solid lines indicate cases where some aspect of a feature can be related to an articulatory maneuver involving a particular muscle (in the case of nasality), *muscle group, or anatomical structure. The interrupted lines indicate cases where the feature at the articulatory level must be regarded as a "state" function since its implementation involves muscles, muscle groups and anatomical structures that are also involved in the implementation of other features. Note that a feature like* +stop *has a "simple" acoustic correlate though it has no "simple" articulatory correlate. The reverse is true for the feature* nasality.

Stop. Consider the feature *stop*. The acoustic basis of +*stop* is clear and simple and consists of an abrupt cessation or diminution of the sound pressure level. However, any one of a number of different muscles may be involved in the articulation of the feature +*stop*—the muscles of the lips, tongue, velum, etc. Note that these muscles are also involved in the articulation of other features, e.g. *rounding* and *back*. We have indicated the mapping between this feature and its articulatory correlates in Figure 1 by means of the dashed lines. We no longer have an invariant one-to-one mapping. The articulatory basis of the feature *stop* is thus a "state" function rather than a maneuver of a particular articulator. The "state" is to abruptly occlude the vocal tract. The feature *stop* at the articulatory level, is thus a binary state function.

Voicing. Note that we have not said that the muscular gestures that are involved in implementing a feature are always binary. The phonologic feature and its articulatory state function may be inherently binary, even if the articulatory maneuvers and muscular commands that are necessary to execute the feature are gradual in nature. Consider the feature *voicing*. The articulatory maneuvers associated with maintaining the state of voicing do not always act in an "all or nothing", i.e. binary mode. Voicing is initiated by adducting the arytenoid cartilages and by placing an appropriate amount of medial and longitudinal compression on the vocal cords. The initial inwards motion of the arytenoid cartilages could perhaps be regarded as a "binary" gesture. However, the tensions of the muscles that adduct the vocal cords at the start of phonation must gradually change to maintain the degree of medial compression that is necessary to sustain phonation when the longitudinal tension changes at the end of a marked breath-group or when either tension or subglottal air pressure change as a result of the feature *prominence* (Lieberman 1967; Ohala and Hirano 1967; Van den Berg 1960 and 1968; Harris *et al.* 1969). The binary basis of the feature *voicing* may be the presence of a voicing detector in man. The human auditory system has the ability to detect pitch periodicity for speech and speech-like signals.[5] The human "pitch detector" is matched to the acoustic range of the human larynx. When listeners are asked to make judgments on the "pitch" of pulse trains, they perceive a fundamental frequency that is equal to the repetition rate of the pulse train for rates that are less than 500 pulses per second. This rate is close to the upper range of the human larynx.

The articulatory maneuvers of the larynx are only the most obvious of the articulatory maneuvers that may be involved in voicing. The evidence cited by Chomsky and Halle (1968) as an articulatory correlate of *tenseness* can, for example, equally be interpreted as an articulatory correlate of *voicing*. Chomsky and Halle cite an X-ray study by Perkell (1965) where,

> In analyzing the behavior of the pharynx in the nontense words [hətˈɛ] and [hədˈɛ] as spoken by American subjects, Perkell found that during the period of closure there was a significant increase in the pharynx width when the nontense [d] was articulated but not when the tense [t] was articulated. This increase in pharynx volume in the nontense obstruent was also accompanied by the presence of voicing during the period of oral closure, which, however, died off toward the end of the stop gap (p. 325).

Chomsky and Halle believe that the motion of the pharyngeal wall during /d/ follows from the "nontense" state of the muscles of the pharyngeal wall. The movement of the pharyngeal wall is supposed to be passive and is supposed to follow from the air pressure build up above the larynx during the obstruent. The sustension of voicing during /d/ is thus explained as a concomitant result of the nontense nature of /d/. It

[5] Flanagan (1965) reviews studies of the perception of fundamental frequency which seem to indicate that man has a "detector" that responds to pitch. Human beings can readily detect the fundamental frequency or "pitch" of a speech signal in adverse acoustic environments, where it is impossible to measure fundamental frequency by means of electronic devices. Human listeners can, for example, detect the fundamental frequency of phonation when the acoustic signal is degraded by telephone circuits that filter out the low frequency components of speech signal and distort the waveform of the speech signal. The facility with which humans can detect fundamental frequency is in sharp contrast to the problems that have been encountered in the design of electronic "pitch extractors". It still is not feasible to extract "pitch" on telephone circuits by electronic means, even though important commercial applications have engendered extensive research since the invention of the Vocoder (cf. Flanagan 1965) by Dudley of Bell Telephone Laboratories in 1937. These unsuccessful electronic devices indicate that the human pitch detector is a rather special device that is particularly well adapted to detecting fundamental frequency.

is, however, just as likely that the motion of the pharyngeal wall during /d/ is an overt gesture that takes place in order to prolong voicing in /d/ by maintaining transglottal airflow during the closed phase of the stop. Abramson and Lisker (1967), for example, note that,

> In the absence of precise knowledge of the muscular action involved, one could even speculate that the pharynx is actively widened for the voiced stops.

Pharyngeal widening appears to be only one of the maneuvers that correlate with voicing during the closed phase of /d/. Rothenberg (1968), for example, shows that velo-pharyngeal leakage also takes place during the closed phase of /d/. Pharyngeal widening and velar opening may thus be articulatory correlates of /d/ under these conditions. We have entered the dotted lines in Figure 1 to indicate these possibilities. Note that the particular articulatory maneuvers that implement the feature voicing depend on the values of the other features for a particular segment. The articulatory manifestations of *+voicing* therefore would be different for /a/ and /d/.

Halle has proposed that oral sensory detectors that might respond to parameters like air pressure and flow could be relevant in defining features. Receptors have been isolated that do respond to "states". Kirchner and Suzuki (1968) have, for example, found muscle spindles in the human larynx that apparently sense the rate at which the vocal cords open and close during phonation. These spindles could play a role in defining the "state function" of phonation. Campbell (1968) has reported that human subjects can sense and control airflow by means of subglottal sensors and regulatory mechanisms that can maintain constant airflow in the presence of abrupt (50 msec.) mechanical obstructions. These airflow receptors may play a role in defining the state function that is the articulatory correlate of the *breath-group* (Lieberman 1967). We are mentioning these receptors in connection with *voicing* because it is clear that airflow is not conserved for *−voiced* segments. Klatt, Stevens and Mead (1968) measured airflow for different consonants and vowels. They found that the greatest average airflow occurred for /h/. Less airflow occurred for /f, s, š, and θ/, least airflow for the vowels. Other consonants had intermediate airflows. Their data, of course, show that airflow is not maintained at a uniform rate during speech, despite the fact that humans apparently have the ability to regulate airflow (Kirchner and Suzuki 1968). Although efferent pathways in the nervous system that monitor physiologic parameters may play a role in defining some phonologic features, they are not the only factors that define features.

The airflow data of Klatt, Stevens, and Mead may have some further relevance with respect to the feature *voicing*. One inference that can be drawn from this data is that the airflow through the supralaryngeal vocal tract is adjusted to maintain the turbulence that is necessary to excite the vocal tract during *−voiced* segments. The consonant /h/ has the maximum cross sectional area supralaryngeal vocal tract constriction (Flanagan 1968). In order to generate turbulent noiselike energy, it is necessary to exceed a critical air velocity in the vocal tract. It is therefore necessary to increase the airflow through the vocal tract as the cross sectional area of the most constricted part of the supralaryngeal vocal tract increases. The Klatt, Stevens, and Mead data thus suggest that one of the articulatory correlates of the state *−voiced* is a glottal adjustment that permits an airflow that is sufficient to generate turbulence in the supralaryngeal vocal tract. The degree of glottal opening that is necessary is a function of the particular *−voiced* segment. The articulatory correlates of *−voicing* are thus scarcely less complex than those of *+voicing*. They appear to be directed towards producing a stable acoustic correlate, i.e. noiselike excitation of the vocal tract.

Prominence. The feature *+prominent* is especially interesting. Its acoustic correlates are increases in fundamental frequency, sound pressure level and duration, and reduction of formant coarticulation (Lieberman 1967; Lindblom 1968). The perceptual effect is one of increased "loudness". The feature may reflect a "match" with an auditory mechanism (Lieberman 1967). At the articulatory level the correlates are very complex. Fundamental frequency increases can be effected by either laryngeal activity or by increases in subglottal air pressure, The two mechanisms, of course, involve different muscles. Recent experimental evidence indicates that the feature *+prominent* may result in greater activity in virtually any of the muscles of the vocal tract. The lip muscles may show increased activity in prominent segments for some speakers (Harris *et al.* 1969). Other speakers may simply increase the duration of a prominent segment. Still other speakers move their tongues more forcefully while they keep the duration constant. At the articulatory level *+prominent* thus must be regarded as a state function in which activity increases throughout the vocal tract. In Figure 1 some of the known consequences of *+prominent* have been entered as interrupted lines. Note that it is not possible to associate any particular anatomical structure or muscle group with the feature *prominence*. It likewise is impossible to associate any known efferent sensor with the feature *prominence*. Although the "archetypal", i.e. primary, articulatory correlate of the state *+prominent* appears to be an increase in subglottal air pressure,[6] which could be monitored by an oral sensor that monitored

[6] Lieberman (1967) proposed a theory that accounts for some aspects of intonation in terms of two phonologic features, the *breath-group* and *prominence*. Acoustic and physiologic correlates of these features were derived by experimental procedures that made use of subglottal air pressure and flow measurements as well as acoustic analysis. Perceptual data indicated that listeners "decoded" certain intonational signals by means of "motor theory perception" structured in terms of the "archetypal", i.e. primary, articulatory correlates of these features. In a recent study (Lieberman *et al.* 1970) this theory was tested by recording the electrical activity of the crico-thyroid muscle of the larynx for a set of 480 short statements and yes-no questions that sometimes had nonterminal *+prominent* syllables. Independently derived data of Fromkin and Ohala (1968) also were examined. The data were consistent with the theory proposed by Lieberman (1967) except that *+prominent* syllables in *unmarked breath-groups* had crico-thyroid activity. In yes-no questions where the crico-thyroid was active at the end of the *marked breath-group*, nonterminal *+prominent* syllables had no crico-thyroid activity. The archetypal articulatory correlate of the *marked breath-group* is an increase in laryngeal tension, whereas the archetypal articulatory correlate of *+prominent* is an increase in subglottal air pressure. Implementation rules that must take into account the "remote" context of a *+prominent* syllable determine its articulatory correlates.

pressure, the secondary articulatory correlates of *+prominent* would each involve "monitors" of time and muscle force. *Prominence* thus could not be associated with a particular efferent sensor.

Consonantal. We have referred to neural "feature detectors" that respond to acoustic parameters like fundamental frequency. We have not discussed the details of the neural mechanisms that must be involved in the perception of speech, because we simply do not know very much about these mechanisms at present. However, we would also propose that the constraints imposed by the neural mechanism involved in speech perception must structure the phonologic feature ensemble just as the constraints of the articulatory mechanism that is involved in speech production must be inherently structured into the phonologic feature system. Recent experiments conducted at Haskins Laboratories, the Montreal Neurological Institute, and elsewhere show that the perception of consonants that involve formant transitions is quite different from the perception of vowels (and all other sounds). These consonants are "decoded" by means of a process that takes account of the constraints of the speech production apparatus (Fry *et al.* 1962; Liberman *et al.* 1963 and 1967). This process, which has been called "the motor theory of speech perception," results in "*categorical*" perception (Eimas 1963). People can normaly discriminate between stimuli much better than they can categorize them. Human beings, for example, can discriminate between hundreds of shades of color. They can instantly detect a subtle color mismatch. Human beings, however, when they categorize colors, can reliably sort them into only seven to nine labelled categories. In general, people can *discriminate* a hundred to a thousand times better then they can *categorize*.

The categorical perception of consonants that involve formant transitions is thus quite striking. Whereas we can discriminate between hundreds of variations in vowel quality, we can not discriminate between consonants any better than we can categorize them. We are apparently optimally adapted to the perception of consonants. The most recent experiments, moreover, show that the perception of consonants takes place outside of the auditory system. It instead occurs only in the dominant hemisphere of the brain (Kimura 1961 and 1964; Shankweiler and Studdert-Kennedy 1967; Haggard 1969; Darwin 1969). In other words, we perceive consonants in a part of the brain that is adapted to the perception of speech. This decoding apparently involves "motor theory" analysis of formant transitions. We therefore propose that the physical basis of the phonologic feature *consonantal* is special neural processing of formant transitions.[7]

[7] Note that this definition of the feature *consonantal* would result in both the liquids, /r/ and /l/, and glides, /y/ and /w/, being marked *+consonantal*. The definition of a feature obviously will have important consequences throughout the phonologic component and new definitions should not be casually adopted. Our proposal for a new definition of the feature *consonantal* is thus just that, a proposal, and it should be carefully explored before new phonologic rules are devised. The same comment obviously applies to any other "new" or revised feature.

"Motor Theories of Perception" and Optimal Matching

Theories like "analysis-by-synthesis" and the decoding of the speech signal in terms of articulatory modeling, which fall into the general class of "motor theories of speech perception" are consistent with this notion of features as articulatory-perceptual matches. Stevens (1969), in a computer-implemented study that makes use of an analog of the human vocal tract, advances the concept of articulatory-perceptual matches. He proposes that the most highly valued supralaryngeal vocal tract configurations are ones where errors in articulation will produce minimum perturbations in the acoustic signal.[8] That is, it is "easy" to use these vocal tract configurations to produce a linguistic output, because the speaker can be sloppy and still produce a stable acoustic signal. Stevens set up approximations to particular consonants and vowels on his computer model of the vocal tract. He, for example, set up an approximation to the vowel /a/ and perturbed, i.e. moved the "point of articulation" (the position of the major constriction of the tongue) through the vocal tract analog. The computer analog calculated the formant frequencies that corresponded to the different points of articulation. Stevens found that in certain parts of the vocal tract small changes in the position of the tongue constriction resulted in large variations in the formant frequencies, whereas in other regions, which corresponded with the "natural" points of articulation, small variations in the position of the major tongue constriction had a very small effect on the formant frequencies. In other words, the natural "points of articulation" occur in regions of the vocal tract where it is not necessary to be overly precise. "Errors" in articulation in these regions have minimum acoustic effect. Stevens' study can thus be viewed as an example of an articulatory to perceptual match structured into the ensemble of phonologic features.

2. The Phonetic Output and the Speech Signal

We have discussed several phonologic features that must be regarded as state functions at the articulatory level rather than as specific, invariant, muscular commands or articulatory maneuvers. We must therefore posit a complex phonetic component that has enough apparatus to yield a speech signal with these phonologic features as an input. Although a great deal of attention has been directed at the phonologic, syntactic and semantic levels of language in recent years, the phonetic level has not attracted much attention. In part, this inattention reflects a naive assumption regarding the nature of the phonetic level, that the process of speech perception can be described by merely listing the "simple" sounds that we hear and that speech production,

[8] Lindblom and Sundberg (1969), in an independent study of vowel production, systematically explore the acoustic consequences of the range of possible human tract configurations. They propose that the most highly valued supralaryngeal vocal tract configurations are those in which the relation between vocal tract configuration and formant frequencies is unambiguous. The particular vocal tract configurations that best meet this criterion also meet the criterion proposed by Stevens. Lindblom and Sundberg also discuss the problem of language-specific implementation rules.

at most, involves "simple" articulatory maneuvers that are strung together like, "beads on a string" to form speech. The traditional "phonetic" solution of a linguistic problem is essentially a regularized orthography, and phonetic theories usually are glossed symbol inventories.

In Figure 2 a schematized view is presented of the phonetic component of the grammar. Note that the input of phonetic features is far from the level of the actual articulatory maneuvers that generate the acoustic speech signal. The feature bundles

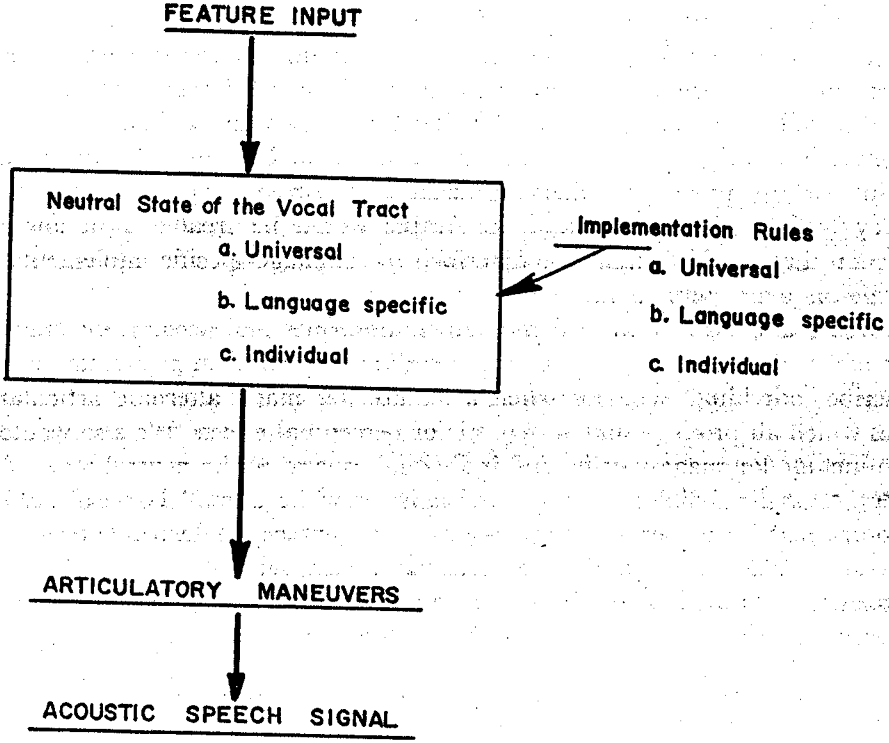

Figure 2. Schematic of the phonetic component. The features serve as the input to the phonetic component. They result in articulatory maneuvers that perturb the vocal tract from its neutral state. The phonetic component's implementation rules determine what articulatory maneuvers will follow from a phonetic feature in a given context as well as the extent of the articulatory maneuver. For example, in Arabic +flat *would result in either pharyngeal or lip maneuvers as a result of the following language specific implementation rules:*

1. $\begin{bmatrix} +\textit{Vocalic} \\ +\textit{Flat} \end{bmatrix} \longrightarrow [+\textit{Round}]$

2. $\begin{bmatrix} +\textit{Cons} \\ +\textit{Flat} \end{bmatrix} \longrightarrow [+\textit{Pharyngeal}]$

3. $[+\textit{Voc}] \longrightarrow [+\textit{Pharyngeal}] \Big/ \begin{Bmatrix} \rule{1.5em}{0.4pt}\,[+\textit{Pharyngeal}] \\ [+\textit{Pharyngeal}]\,\rule{1.5em}{0.4pt} \end{Bmatrix}$

that are the output of the phonologic component are *controls* for the phonetic component. The features thus are neither actual articulatory gestures, nor are they actual motor commands. Through the mediation of the phonetic component of the grammar they result in articulatory maneuvers that generate speech. The phonologic features thus have both articulatory and acoustic correlates, but these correlates are also determined, in part, by other aspects of the phonetic component. Some of these aspects of the phonetic component have been noted before. Halle and Chomsky (1968), in particular, make use of the concept of a "neutral speech state" as well as certain universal implementation rules. However, the more abstract nature of the phonologic features that we have posited calls for additional implementation rules that determine the articulatory maneuvers that implement a feature in a particular context. We could, for example, retain the Jakobsonian feature of *flatness* in the universal feature inventory and employ implementation rules like those noted in Figure 1 to account for the process of pharyngealization in Arabic. Thus, in contrast to McCawley (1967), pharyngealization in Arabic would be treated as a low-level phenomenon that would formally be described by language-specific implementation rules in the phonetic component.

We would also want to include implementation rules that account for universal phonetic effects like those of jaw angle or rounding, as well as implementation rules that describe individual behavior when a feature has many alternate articulatory correlates which all produce similar acoustic or perceptual effects. We also would be forced to include language-specific and individual aspects of the neutral state of the vocal tract. A nasalized dialect of American English in which a small degree of nasalization accompanied all segments that are *−nasal* in the underlying lexical entries might have a neutral vocal tract configuration that had a velar leak.

A language or dialect would have to be specified for its implementation rules and neutral position,[9] as well as for the subset of features that it employed from the universal

[9] One very reasonable question that comes to mind when we postulate the existence of a "neutral" state for speech production is how to determine experimentally what the neutral state is? This is not a trivial question, since individual speakers differ from each other. Individual speakers may, furthermore, make errors when they produce an actual utterance. The well-known distinction between "competence" and "performance" hinges on the fact that a speaker may depart from the form that he "knows" is correct when he produces an actual utterance. How can we then determine what the "neutral" position of the vocal tract is for a particular speaker? Moreover, how can we determine the "universal" and "language-specific" aspects of the neutral position?

The only answer is that we must form hypotheses and test them using many sources of data. Obviously, the selection of the data that we take to be pertinent will implicitly shape these hypotheses. In this connection, we propose that the universal aspects of the speech neutral position reflect the most basic aspects of speech production in both the ontogenetic and the phylogenetic sense. The data that we will consider in deriving and testing our hypotheses concerning the nature of the neutral position thus will include phylogenetic and ontogenetic studies of the cries of nonhuman primates and human infants. Nonhuman primates (Lieberman 1968; Lieberman, Klatt, and Wilson 1969) produce schwa-like cries that always are restricted to uniform to slightly flared supralaryngeal vocal tracts.

Human neonates, like the nonhuman primates, have supralaryngeal vocal tracts that are restricted to uniform, or slightly flared, cross sections (Lieberman, Harris, Wolff, and Russell 1969). The larynx in human neonates is positioned quite high relative to its position in an adult. Human neonates, like the nonhuman primates, lack a pharyngeal region that can change its cross sectional area independent of the cross sectional area set of phonologic features.[10] Particular languages or dialects might indeed use the same subset of features with different implementation rules or neutral positions. Öhman (1968), for example, in a study of the accents of Swedish, has accounted for many of its accents by means of two features and simple implementation rules that merely involve fixed time delays for different dialects. Similar results have been found by Lisker and Abramson (1964 and 1967) across different languages for voiced and unvoiced stops. The manner in which the articulatory apparatus moves also seems to differ for different languages. Stevens and Klatt (1969) have found that differences between Spanish and English vowels reside not only in the target formant frequencies, but also in the shape of the F_1–F_2 contour, i.e. in the dynamics of the vowel. Different implementation rules or different features would, therefore, be necessary for Spanish and English vowels. We have only begun to explore cross-language differences quantitatively.

3. Concluding Comments

A unified phonetic theory should be structured in terms of our knowledge of the anatomic, physiologic and neural mechanisms of speech production and perception. We have discussed some factors that may be important, like the existence of acoustic feature detectors, articulatory maneuvers that are "easy" to effect, articulatory maneuvers that produce stable acoustic outputs and efferent sensors of physiologic events. These factors may be the bases of particular phonologic features. The presence of a "voicing" detector may, for example, be the reason why *voicing* functions as a phonologic feature. The facility with which the velum can be opened and closed may

of buccal region. The only articulatory maneuvers that neonates execute during their cries involve up and down motions of the larynx that shorten or lengthen the overall supralaryngeal vocal tract.

X-ray studies of the adult human vocal apparatus (Perkell 1965) show that the entire vocal tract moves into a "speech" position immediately before the speaker begins to talk. The supralaryngeal vocal tract configuration appears to approximate the uniform tube, schwa, configuration, though it is difficult to be absolutely certain from lateral X-rays, There are also differences between individual speakers. One difference between different speakers may be in the state of the velum. The adult speaker of English observed by Perkell, (1965) closed his velum in the neutral position. The neonates observed by Truby and his associates (1965) kept their velums open during their cries for the first three days of life and thereafter closed their velums during their cries. These infants had the muscular ability to close their velums from birth onwards, since they always closed their velums during swallowing. This data, therefore, does not indicate whether the "simplest", most basic, i.e. "universal", aspect of the neutral speech position involves an open or closed velum.

All of this research supports the concept of a "neutral" schwa-like configuration of the vocal tract. This aspect of the neutral position would appear to be universal. Deviations from this position for particular languages or individual speakers would, according to this view, involve additional effort. It is not clear whether the "universal neutral position" involves an open velum, i.e. nasality. If this were the case, nasality would be unmarked and the marked state would be *+oral*. Particular languages like English, might however have a language-specific neutral position that involves a closed velum. The acoustic consequences of nasality, which involve energy losses in parts of the acoustic frequency spectrum, may have favored the development of a universal neutral position that involved a closed velum.

[10] There is no general agreement on the specific set of features that is sufficient to specify all languages. Ladefoged (1967) discusses this problem in detail. Ladefoged (1967) and Fromkin (1968) relate phonetic features to articulatory maneuvers in terms of a speaker's "competence" and his "performance."

be the basis of the feature *nasality*. We have proposed that the phonologic features are thus signalling units that take advantage of various innate mechanisms in the human vocal apparatus and auditory perceptual system. It seems clear that features cannot generally be regarded as one-to-one correspondences with particular muscle commands or particular articulatory maneuvers. The physical bases of most phonologic features are, moreover, still speculative. It would also be foolhardy to assume that we have isolated all the factors that may play a role in structuring the ensemble of phonologic features or the phonetic component that we have devised in order to produce an acoustic speech signal from a feature specification. Much of what we have said concerning acoustic feature detectors and the articulatory bases of some features is implicit in earlier studies like Jakobson, Fant and Halle (1952) and Chomsky and Halle (1968). However, a quantitative and explicit phonetic theory has yet to be developed; this paper may serve as a preliminary framework.

References

Abramson, A. S. and L. Lisker (1967) "Laryngeal Behavior, the Speech Signal and Phonological Simplicity," *Proceedings of the 10th International Congress of Linguistics*, Bucharest, August–September, 1967.

Campbell, E. J. M. (1968) "The Respiratory Muscles," *Annals of the New York Acad. Sc.* 155, 135–139.

Capranica, R. R. (1965) *The Evoked Vocal Response of the Bullfrog*, MIT Press, Cambridge, Mass.

Chiba, T. and M. Kajiyama (1958) *The Vowel, Its Nature and Structure*, Phonetic Society of Japan, Tokyo.

Chomsky, N. and M. Halle (1968) *The Sound Pattern of English*, Harper and Row, New York.

Darwin, C. J. (1969) "Laterality Effects in the Recall of Steady State and Transient Speech Sounds," *Proceedings of the 77th Meeting of the Acoustical Society of America* 44.

Eimas, P. D. (1963) "The Relation Between Identification and Discrimination along Speech and Non-speech Continua," *Language and Speech* 6, 206–217.

Fant, C. G. M. (1960) *Acoustic Theory of Speech Production*, Mouton, The Hague.

Flanagan, J. L. (1965) *Speech Analysis, Synthesis and Perception*, Springer-Verlag, Berlin, New York.

Frishkopf, L. S. and M. H. Goldstein, Jr. (1963) "Responses to Acoustic Stimuli from Single Units in the Eighth Nerve of the Bullfrog," *J. Acoust. Soc. Am.* 35, 1219–1228.

Fromkin, V. (1968) "Speculations on Performance Models," *J. Linguistics* 4, 47–68.

Fromkin, V. and J. Ohala (1968) "Laryngeal Control and a Model of Speech Production," *Working Papers in Phonetics, UCLA* 10, 98–110.

Fry, D. B., A. S. Abramson, P. D. Eimas, and A. M. Liberman, (1962) "The Identification and Discrimination of Synthetic Vowels," *Language and Speech* 5, 171–189.

Haggard, Mark P. (1969) "Perception of Semivowels and Laterals." *Proceedings of the 77th Meeting of the Acoustical Society of America* 45.

Harris, K. S., T. Gay, G. N. Sholes and P. Lieberman, (1969) "Some Stress Effects on Electromyographic Measures of Consonant Articulation," *Status Report on Speech Research* 13/14, Haskins Laboratories, New York City.

Jakobson, R., C. G. M. Fant and M. Halle (1952) *Preliminaries to Speech Analysis*, Technical Report No. 13, Acoustics Laboratory, MIT: reprinted by the MIT Press, Cambridge, Mass.

Kimura, D. (1961) "Some Effects of Temporal-lobe Damage on Auditory Perception," *Canad. J. Psychol.* 15, 166–171.

Kimura, D. (1964) "Left-right Differences: the Perception of Melodies," *Quart. J. Exp. Psychol.* 16, 355–358.

Kirchner, J. A. and M. Suzuki (1968) "Laryngeal Reflexes and Voice Production," *Annals of the New York Acad. Sc.* 155, 98–110.

Klatt, D. H., K. N. Stevens and J. Mead (1968) "Studies of Articulatory Activity and Airflow during Speech," *Annals of the New York Acad. Sc.* 155, 42–54.

Ladefoged, Peter (1967) "Linguistic Phonetics," *Working Papers in Phonetics, UCLA* 6.

Liberman, A. M., F. S. Cooper, K. S. Harris, and P. F. MacNeilage (1963) "A Motor theory of Speech Perception," *Proceedings of the Speech Communication Seminar*, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden (August, 1962).

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy (1967) "Perception of the Speech Code," *Psychol. Review*, 74, 431–461.

Lieberman, P. (1967) *Intonation, Perception and Language*, MIT Press, Cambridge, Mass.

Lieberman, P. (1968) "Primate Vocalizations and Human Linguistic Ability," *J. Acoust. Soc. Am.* 44, 1574–1584.

Lieberman, P., K. S. Harris, P. Wolff, and L. H. Russell (1969) "Newborn Infant Cry and Nonhuman Primate Vocalizations," *J. Speech and Hearing Research* (forthcoming).

Lieberman, P., D. H. Klatt, and W. A. Wilson (1969) "Vocal Tract Limitations of the Vocal Repertoires of Rhesus Monkey and other Non-human Primates," *Science* 164, 1185–1187.

Lieberman, P., M. Sawashima, K. S. Harris, and T. Gay (1970) "The Articulatory Implementation of the *Breath-group* and *Prominence*: Cricothyroid Muscular Activity in Intonation," *Language* (forthcoming).

Lindblom, B. (1968) "Temporal Organization of Syllable Production," *Speech Transmission Laboratory Report 2/3*, Royal Institute of Technology, Stockholm, Sweden.

Lindblom, B. and J. Sundberg (1969) "A Quantitative Model of Vowel Production and the Distinctive Features of Swedish Vowels," *Speech Transmission Laboratory Report 1*, Royal Institute of Technology, Stockholm, Sweden.

Lisker, L. and A. S. Abramson (1964) "A Cross-language Study of Voicing in Initial Stops: Acoustical Measurements," *Word 20*, 384–422.

McCawley, J. D. (1967) "Le Role d'un Systeme de Traits Phonologiques dans une Théoric du Langage," *Langage* 8, 112–123.

Müller, J. (1848) *The Physiology of the Senses, Voice and Muscular Motion with the Mental Faculties*, translated by W. Baly, Walton and Maberly, London.

Ohala, J. and M. Hirano (1967) "Studies of Pitch Change in Speech," *Working Papers in Phonetics*, UCLA, November, 1967.

Öhman, S. (1968) "A Model of Word and Sentence Intonation," *Speech Transmission Laboratory Report 2/3*, Royal Institute of Technology, Stockholm, Sweden.

Perkell, J. S. (1965) "Cineradiographic Studies of Speech: Implications of a Detailed Analysis of Certain Articulatory Movements," *Reports to the Fifth International Congress of Acoustics 1*, A32, Universite de Liège.

Rothenburg, M. (1968) *The Breath-Stream Dynamics of Simple- Released-Plosive Production*, S. Karger, Basel (Switzerland) and New York.

Shankweiler, D. P. and M. Studdert-Kennedy (1967) "Identification of Consonants and Vowels Presented to Left and Right Ears," *Quart. J. Exp. Psychol.* 19, 59–63.

Stevens, K. N. (1969) "The Quantal Nature of Speech: Evidence from Articulatory-acoustic Data," in *Human Communication: A Unified View*, E. E. David, Jr. and P. B. Denes, eds., McGraw-Hill, New York (in press).

Stevens, K. N. and M. H. Klatt (1969) "Analysis of Vowels Produced by Spanish and English Speakers," *J. Acoust. Soc. Am.* 45 (abstract).

Van den Berg, J. W. (1960) "Vocal Ligaments Versus Registers," *Current Problems in Phoniatrics and Logopedics* 1, 19–34.

Van den Berg, J. W. (1968) "Sound Production in Isolated Human Larnyges," in *Sound Production in Man, Annals of the New York Acad. of Sc.* 155, 18–27.

Department of Linguistics
University of Connecticut
Storrs, Connecticut 06268