VIII. SPEECH COMMUNICATION

Academic and Research Staff

Prof. K. N. Stevens
Prof. M. Halle
Prof. W. L. Henke
Prof. A. V. Oppenheim
Dr. Margaret Bullowa
Dr. J. L. Fidelholtz
Dr. K. Hashimoto
Dr. A. W. F. Huggins
Dr. A. R. Kessler
Dr. Emily F. Kirstein
Dr. D. H. Klatt
Dr. Paula Menyuk
Dr. J. S. Perkell
Dr. R. A. Stefanski
Mary M. Klatt
R. M. Mersereau
Estill Putney

Graduate Students

G. A. Akers
T. Baer
R. E. Crochiere
D. E. Dudgeon
Ursula Goldstein
S. Maeda
B. Mezrich
H. Pines
M. R. Portnoff
L. Tung
V. W. Zue

A. POTENTIAL ROLE OF PROPERTY DETECTORS IN THE PERCEPTION OF CONSONANTS

National Institutes of Health (Grant 5 RO1 NS04332-10)

K. N. Stevens

1. Introduction

Various investigations of the acoustic properties of consonant sounds and of the cues for consonant perception have shown that the acoustic correlates of a number of consonantal features are highly context-dependent. The acoustic correlates for place of articulation, for example, depend upon the vowel or other segment that follows the consonant, the stress on the adjacent vowel, the manner of articulation of the consonant, and its voicing characteristics.

In spite of this apparent lack of invariance, it is recognized that, at a stage when a child is acquiring language, he needs some basis on which to organize and classify the acoustic events in speech. Some invariant property must exist in the acoustic events associated with phonetic segments that have a particular feature in common, so that a child can identify these events as belonging to the same category or feature.

In this report we examine the acoustic properties that differentiate vowels from consonants and that identify the principal places of articulation for consonants. We show that under many circumstances there are indeed acoustic correlates for the feature [+consonantal] and for the labial, coronal and velar places of articulation that occur almost universally in language. These acoustic attributes are associated with the rapid spectrum change in the 20-odd ms following the release of a consonant into a following vowel. These properties do not identify the features in all phonetic environments, but they are effective in a large number of situations, particularly when the consonants are in prestressed position. It is postulated that the child initially utilizes property detectors to classify consonantal speech events in a canonical consonant-vowel environment and subsequently develops a more detailed internalized structure of the acoustic correlates for each phonetic category including various context-dependent secondary cues. At a later stage, these secondary cues can be used to decode the acoustic signal even when the primary properties are absent.

2. Vowel-Consonant Dichotomy

a. Results from Studies of Speech Perception

Some experimental results have suggested that the auditory processing of consonants is qualitatively different from the processing of vowels and vowel-like sounds.[1] Among the experimental findings in support of this view are data on the identification and discrimination of speech sounds, and the results of listening studies with competing syllables in the two ears. For certain consonantal stimuli there are peaks in the discrimination of small changes in the acoustic signal in the vicinity of a phoneme boundary, whereas listeners have poor discrimination for stimuli that are identified as being within the same phonetic class.[2] For vowel stimuli, on the other hand, there are essentially no peaks in the discrimination of small changes in formant frequencies in different vowel regions, and ability to discriminate stimuli far exceeds identification performance.[3,4] The dichotic listening experiments show that there is a right-ear advantage for competing consonant stimuli but not for vowels, which suggests that the left hemisphere, which is known to be involved in speech perception, somehow plays a greater or more direct role in consonant perception than in vowel perception.[5]

In a more recent series of experiments, Crowder[6] examined the ability of subjects to recall lists of syllables that differed in the vowel components in one set of experiments or in the consonant components in another. He concluded that vowels can be placed in some kind of precategorical store for a time interval of one or two seconds, where they are accessible to the listener for further processing and categorization, whereas stop consonants are classified directly and are apparently not placed in such a store before being decoded into features.

b. Acoustic-Articulatory Observations

The apparent vowel-consonant dichotomy in the perception of speech has a counterpart in the acoustic characteristics of these two classes of segments. If we examine the properties of the sound output that results from various sequences of vocal-tract and laryngeal configurations, we observe that there are two distinct types of sounds depending on the nature of the articulatory gestures that give rise to the sound.

Vowel and vowel-like segments are produced with a relatively open vocal-tract configuration and with the acoustic excitation at the glottis. The acoustic properties during these intervals when the vocal tract is relatively open persist for an appreciable time interval (at least several tens of milliseconds) or change only slowly within this time interval. For this type of sound output the short-time spectrum at a given point in time provides information concerning all relevant aspects of the vocal-tract configuration at that time. All vocal-tract resonances are directly observable in the signal, and in effect these resonances determine the vocal-tract shape.

If we regard a vowel-like segment as characterized by a target configuration of the vocal tract, then to the extent that this configuration is actually reached in running speech the properties of the acoustic signal at this time indicate all of the features of the segment. Under some circumstances, the target configuration may not be reached, because of the influence of adjacent segments.[7] When there is such a deviation from the idealized target, the features of the segment must be inferred by examining the acoustic spectrum over an interval of time. For example, if the trajectories of the first two or three formants are known, it may be possible to "compute" the target configurations of the underlying segments, even though these configurations are never reached. This would require storing certain properties of the signal in some precategorical form over a period of a few hundreds of milliseconds (or even for a second or more) before decoding the signal into phonetic units or features. Such a memory would also be needed to establish reference data concerning average formant spacing for a given speaker against which the formants of adjacent vowels can be compared.

Most types of consonant segments in most phonetic environments have the common acoustic attribute that a rapid change in the acoustic spectrum occurs at the release of the consonantal constriction at the onset of the following sound.[8] The rapid change of spectrum occurs in the frequency range above 1 kHz (i.e., in the region of the second formant and above), and the major portions of the spectrum change take place within a time interval of ~20 ms from the release. (This method of classifying consonantal sounds is similar to the acoustic feature transitional, proposed by Fant.[9,10]) The rapidity of this spectrum change is greater than that observed for the segments /w, y/, for which the changes in formant frequency occur more slowly, and for which there is no abrupt onset of energy immediately preceding the spectrum change. The contrast between the slow spectrum change for the glides /w, y/ and the rapid spectrum change for the stop consonant /d/ is illustrated in Fig. VIII-1, which shows spectra sampled at 10-ms intervals at the onsets of syllables with these initial consonants.

This type of rapid spectrum change following an abrupt onset of energy is in fact characteristic of segments with the feature [+consonantal], as defined by Jakobson, Fant, and Halle[11] and by Chomsky and Halle.[12] This acoustic characteristic is in evidence consistently when the consonantal segment is in prestressed position, i.e., in a canonical consonant-vowel syllable. There are other phonetic environments, such as in a consonant cluster or in syllable-final position, where a rapid spectrum change may not be observed, and where the feature [+consonantal] is cued by other acoustic evidence

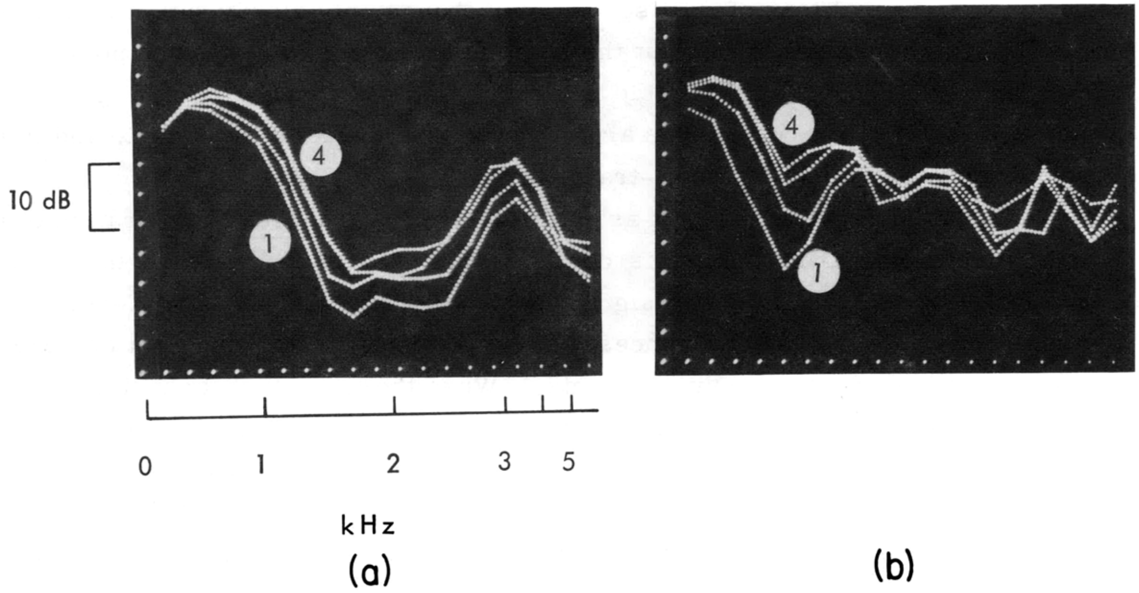

Fig. VIII-1. Short-time spectra sampled at 10-ms intervals: (a) during the formant transitions in the syllable /wa/, and (b) immediately following the consonant release in the syllable /do/. The spectra were obtained with a 19-channel filter bank at the center frequencies indicated. Filter bandwidths are 360 Hz in the frequency range up to 3000 Hz, and then increase to 1 kHz at 7 kHz. Filter outputs are rectified and smoothed with an averaging time of ~10 ms. The spectra are ordered in time from bottom to top (in the low-frequency regions). The first and fourth of the four successive sampled spectra are labeled.

or is inferred from language-specific rules.

Another result that provides some insight into the role played by rapid spectrum change at an onset in speech production and perception is the observation that in prestressed voiceless aspirated stop consonants the onset of voicing occurs only after the formant transitions are essentially complete.[13, 14] If there are substantial transitions remaining after voicing onset, the consonant tends to be perceived as voiced in English. In other words, a voiceless aspirated stop (in a consonant-vowel syllable) occurs only if the property of rapid spectrum change is absent immediately following the onset of voicing. A rapid spectrum change does occur, of course, at the onset of the aspiration phase of a voiceless stop consonant.

An indication of the duration of the rapid spectrum change associated with the perception of consonants as opposed to glides has been given by Liberman et al.[15] Through experiments on the perception of consonant-vowel stimuli with various durations for the formant transitions (requiring a response of either /b, g/ or /w, y/), they found that stop consonants were perceived when the transition durations were less than approximately 40 ms.

The acoustic events that constitute the cues for the features indicating place of articulation for consonantal segments cannot usually be manifested while the vocal tract is in the target configuration associated with the segment. These events occur when the vocal tract is in motion between the configuration in question and the target configuration for an adjacent segment. Thus, for example, for certain stop consonants no sound is generated during the closure interval, when the articulators are presumably in a target configuration appropriate for the consonant. Cues for the various features that identify the stop consonant preceding a vowel occur when sound is produced immediately following the closure interval. These acoustic events exist during the few tens of milliseconds adjacent to the stop closure. Likewise, for a nasal or fricative consonant, only some of the features can be identified from the attributes of the sound generated during the time when the vocal tract is in the constricted position. The features that identify place of articulation can only be determined from the attributes of the sound in the few tens of milliseconds after the consonantal release.

The classes of segments having the characteristic that the features are not completely specified by acoustic events in the vicinity of the target articulation include all segments for which the vocal tract has a narrow constriction in the midline. For such configurations, either the source of acoustic energy is not at the glottis or the source is at the glottis but the vocal tract from glottis to lips is not an open single tube; it is a complex tube with side branches.

c. Results from Studies of Auditory Perception

There is some evidence from psychoacoustic studies of a contrast between the perception of sounds with slow spectrum changes and sounds in which rapid spectrum changes occur within 20-odd ms of an onset of energy. In an investigation of the perception of temporal order, Hirsh[16] determined that the time needed for a subject to identify which of two successive tonal stimuli occurred first was approximately 20 ms. Nabelek and Hirsh[17] found that the discrimination of changes in transition rate for brief gliding tones or chirps was optimal when the transition durations were 20-30 ms. The tentative conclusion of these investigators was that "there are two mechanisms involved in the discrimination of the rate of change of frequency; one which is in action for fast changes over larger frequency intervals, and the other one for slow changes especially in small frequency intervals."[18]

3. Cues for Certain Consonantal Features

The evidence just presented suggests that there are well-defined acoustic and articulatory markers indicating the presence of a consonantal segment in the speech stream. These acoustic correlates may not be observable for consonants in all phonetic environments, but they are especially evident when a consonant occurs before a stressed vowel.

We examine now the acoustic correlates for certain other features of consonants, particularly those indicating place of articulation.

As indicated above, the relevant acoustic events occur within 20-30 ms following the release of the consonant. It is traditional to describe these events in terms of the formant transitions, particularly the transition of the second formant. We prefer to characterize these transitions in terms of the changes in the overall acoustic spectrum immediately following the release, rather than in terms of trajectories of individual formants. For back vowels, the principal attribute of the spectrum is a broad concentration of energy in the frequency range occupied by the first and second formants (F_1 and F_2). The third and higher formants are usually weak, and do not play a significant role in the perception of the vowel. Thus we would expect that the rapid spectrum changes that serve as cues for consonantal place of articulation would occur in the broad F_1-F_2 energy concentration at the onset of the vowel. Front vowels, on the other hand, are characterized by a broad concentration of energy formed by the relative proximity of F_2 and F_3 (and F_4, in the case of high front vowels). This energy concentration can apparently be viewed as a unit centered between F_2 and F_3 for low front vowels and in the vicinity of F_4 for high front vowels. It is appropriate, therefore, to examine how this broad spectral unit is "built up," as it were, when a consonant is released into a front vowel.

At the onset of a stop consonant, a burst of frication noise energy, often just a few milliseconds long, precedes the onset of voicing (or precedes the onset of aspiration in the case of an aspirated stop consonant). We shall assume that this burst can be considered as the initiation of the rapid spectrum change at the consonant release, if there is spectral energy in the burst in the vicinity of the major spectral peak for the vowel. Thus the initial burst of energy in syllables beginning with /g/, and the burst for syllables with a front vowel preceded by /d/ would be considered as part of the rapid spectrum change, whereas the d-burst in a syllable with a back vowel would not. (We have carried out experiments on the perception of formant transitions preceded by noise bursts, and the results of these experiments support this view of the role of the noise burst. These experiments will be reported elsewhere.) The burst at the onset of the consonant /b/ is relatively weak, and may not play a significant role in shaping the rapid spectrum change.

a. Context-Independent Cues for Place of Articulation in Canonical CV Environment: Property Detectors

We concentrate our attention initially on cues for place of articulation in stressed CV syllables. Syllables of this type occur universally in language, and children appear to learn first to discriminate between different consonant places of articulation when the consonant is in this environment. The implications are (i) that children are, in some sense, predisposed to respond to the acoustic properties that indicate place of articulation in this environment, and furthermore, (ii) that ability to discriminate different places of articulation in stressed CV syllables is a prerequisite for learning to extract these features in other phonetic environments.

A rapid spectrum change in which the onset of energy at high frequencies precedes the onset at lower frequencies characterizes a coronal consonant, i. e., a consonant produced with the tongue blade. For example, the falling second-formant transition in a syllable containing /d/ or /n/ followed by a back vowel would result in a spectrum change of this type. An example illustrating the spectrum change at the release of the initial consonant in the syllable /na/ is shown in panel f of Fig. VIII-2. The four curves represent short-time spectra sampled at 10-ms intervals, the first (and lowest) being sampled immediately before the consonantal release. The spectrum change that occurs in this interval can be described as either a falling F_2 transition or an initial onset of energy near the upper edge of the F_1-F_2 energy concentration, followed by a delayed onset of energy at frequencies below the upper edge. A similar sequence can be seen in panel b in Fig. VIII-1 for the syllable /do/. Again the energy onset at the upper edge of the broad F_1-F_2 spectral peak precedes the onset of energy at lower frequencies. These spectra were sampled after the initial high-frequency energy burst. Panels b and d in Fig. VIII-2 show spectra sampled at 10-ms intervals near the release of the consonants into front vowels in the syllables /di/ and /ne/. In these examples, the onset of energy at the upper end of the broad high-frequency peak precedes the onset at lower frequencies within the peak.

Labial consonants are generally characterized by a rapid spectrum change in which the spectrum at the initial onset of energy has an energy concentration that is lower in frequency than that in the spectrum sampled a few milliseconds later. Sampled spectra at the onset of syllables with initial labial consonants are shown in panels a, c, and e in Fig. VIII-2. In the case of the syllable /bi/ (panel a), the onset at the low-frequency end of the major energy concentration for the vowel precedes the onset at higher frequencies (but this attribute is not as clear as it is in panel c, for the syllable /me/, where the rising second formant can be seen in the two upper spectra). For the syllable /ma/ (panel e) this attribute is evident in the low-frequency F_1-F_2 peak.

Still another type of rapid spectrum change occurs for velar and other dorsal consonants. In this case, the major energy concentration at the onset is in the middle frequency range. Immediately following this onset there is a spreading or broadening of the spectral energy to frequency regions above and below this middle range. These characteristics arise from the initial noise burst that usually precedes voicing onset for velar stops, followed by the transitions of the second and third formants – F_2 usually falling and F_3 rising. We have previously[19] reported examples of this type of rapid spectrum change, as well as additional examples of spectra sampled at the onset of

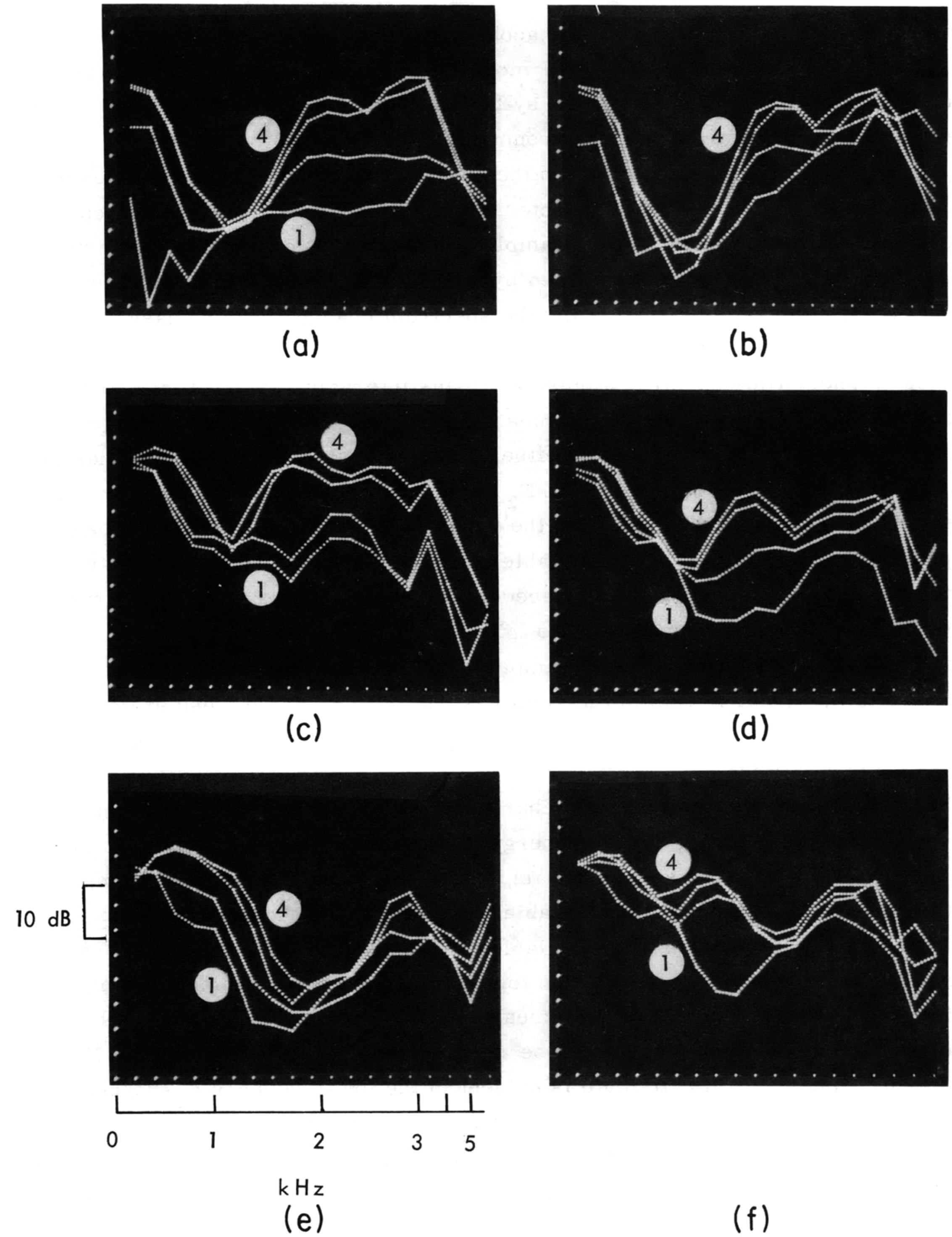

Fig. VIII-2. Spectra sampled at 10-ms intervals in the vicinity of the consonantal release for the syllables (a) /bi/, (b) /di/, (c) /me/, (d) /ne/, (e) /ma/, (f) /na/. In each case, 4 successive spectra are shown; the first and fourth are labeled. (See caption for Fig. VIII-1.)

labial and coronal consonants.

It is suggested, therefore, that detection of place of articulation of stop and nasal consonants preceding stressed vowels can be achieved if three different property detectors are available in the receiver: (i) a detector that responds when the onset of energy in one frequency region is followed by the onset of energy in an adjacent lower frequency region; (ii) a detector that responds when the onset of energy in one frequency region is followed by the onset of energy in an adjacent higher frequency region; and (iii) a detector that responds when the onset of energy in one frequency region is followed by the onset of energy in both an adjacent higher frequency region and an adjacent lower frequency region. The general form of the acoustic pattern that generates responses from each of these property detectors is shown schematically in Fig. VIII-3. It is conceivable that only two such detectors are required — one responding to rising spectral energy and the other to falling spectral energy. Both detectors respond to a velar consonant in which the energy at an onset spreads both upward and downward in frequency.

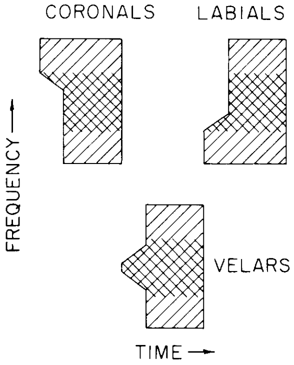

Fig. VIII-3.

Schematic intensity-frequency-time patterns for syllables with initial consonants having different places of articulation. The crosshatched areas indicate major vowel spectral prominences. The rapid spectral changes at the onset of energy for the various classes of consonants have different temporal characteristics. (After Stevens.[19])

The acoustic data cited above are for the most part consistent with the results of experiments at Haskins Laboratories[20, 21] on the role of the second- and third-formant transitions in the perception of initial stop consonants. Thus for syllables beginning with coronal consonants, the second formant before back vowels is always falling and the best responses are obtained when F_3 is also falling. When the vowel is a front vowel, there may be a slightly rising F_2 transition, but this is offset by a strongly falling F_3 transition. The transitions of F_2 and F_3 are always rising in syllables beginning with labial consonants. The synthesis experiments show that velar consonant responses are obtained when F_2 and F_3 begin close together and then separate.

The simple classes of rapid spectrum change for labials, coronals, and velars are clearest when the consonants precede the low back vowel /a/. The second formant for this vowel is in an intermediate frequency region such that it rises strongly for labials and falls for dentals, and F_2 and F_3 clearly separate after the release of a velar consonant. Spectral data consistently demonstrate the prototypical upward (for labials), downward (for coronals), and spreading (for velars) spectrum shifts in the midfrequency range for this vowel. These attributes are sometimes less consistent when the vowel is a high vowel, and in these cases other cues may be required to identify the consonants.

The contribution of the various acoustic events at the onset of stop consonants has been studied by Fischer-Jørgensen,[22] who examined listener identification of initial stop consonants under various conditions in which the initial burst or the formant transitions were removed from the stimulus through tape-cutting techniques. The clearest findings of these experiments were that the formant transitions constituted the strongest cue when the consonants preceded the vowel /a/, and the initial burst could provide an important cue in some consonant-vowel syllables with high vowels. Acoustic data show that this initial burst contributes to the shape of the rapid spectrum change in syllables like /di/ and /gi/, where the burst is in the frequency region of the broad high-frequency energy concentration for the vowel /i/.

All of these cues for place of articulation occur within the rapid spectrum change at the onsets of consonant-vowel syllables. They identify features of place of articulation without reference to acoustic events remote from this point in time. That is, these cues are absolute properties of the speech signal, and are context-independent. Determination of place of articulation from these properties does not require that attributes of the signal be placed in some kind of precategorical auditory store, since there is no need to interpret one portion of the signal with reference to another portion that occurs at a different time. That a precategorical store is not required for these consonants is consistent with the findings of Crowder.[6]

b. Context-Dependent Cues for Consonants

It is well known that strategies for identifying place of articulation for consonants are based on much more than the detection of simple types of rapid spectrum change of the type just described. For example, cues for place of articulation for certain fricative consonants can reside in the spectrum of frication noise during the consonant, as well as in the transition to the adjacent vowel.[23, 24] The spectrum of the initial burst of acoustic energy for a stop consonant (independent of a subsequent change in the acoustic spectrum) can provide a secondary cue for place of articulation, as can the spectrum of the nasal murmur in some vowel environments. Furthermore, there may be slower formant motions after the initial 20-30 ms rapid spectrum change, or the initial formant motion may be small, and such a formant movement may be a cue for consonantal place of articulation or a cue whose interpretation depends on the context in which it occurs. A classical example of an exception to a simple property-detector theory is the contrast between the two-formant synthesized syllable /di/, which has a slightly rising F_2 transition, and /du/, which has a strongly falling F_2 transition. The third and higher formants, which contribute to the net downward movement of the high-frequency energy concentration for front vowels (in spite of the slightly rising F_2 transition), are, of course, absent in these stimuli, and hence the "secondary" cue contributed by the rising F_2 in this particular vowel environment is all that is available to the listener. In phonetic environments other than initial prestressed position, formant transitions at the end of a vowel can also indicate place of articulation for the consonant.

At least some (and perhaps all) of these secondary cues require that one part of the sound stream be interpreted in the context of another part of the signal. That is, they require that certain aspects of the signal be placed in precategorical store; the utilization of these aspects for decoding the sound in terms of phonetic features must await the occurrence of additional information in the signal, or must make use of information stored from past acoustic events.

Although a variety of acoustic attributes can play a role in determining consonantal place of articulation, it appears that for a majority of consonant-vowel syllables, the labial, coronal, and velar categories have the simple primary context-independent characteristics described above.

4. Property Detectors and the Acquisition of Phonetic Distinctions

At an early stage in the acquisition of speech, a child is observed to produce simple consonant-vowel syllables and thus appears to make a distinction between vowels and consonants. That is, he distinguishes between vowel-like sounds, which have slowly varying spectrum changes, and sounds in which there is a rapid spectrum change in the vicinity of an onset of energy in the signal.

If he is able to distinguish between vowel-like acoustic events and consonantlike events, then he can develop the capability of classifying events within each of these categories. We postulate that he is equipped with two or three simple property detectors of the type described above, and utilizes these detectors for initially organizing the consonantlike sounds that are characterized by rapid spectrum change.

A substantial majority of all stressed syllables beginning with a nasal or stop consonant (and many with an initial fricative consonant) would be correctly classified in three categories, labial, coronal, and velar, if the infant were equipped with these detectors. Stop and nasal consonants preceding the vowel /a/ would have the greatest probability of being identified, since the characteristic rapid spectrum change is most evident in these cases. These are the syllables that appear first in the child's productions.[25]

A few stressed CV syllables (particularly those with high vowels) would not be correctly classified, since the cues based on the rapid spectrum changes would sometimes be ambiguous. Furthermore, these detectors would not categorize certain final consonants or consonants in pre-unstressed position. Consonants with these various places of articulation have other acoustic attributes beyond those to which the property detectors are sensitive, as noted above.

The consonants that are correctly classified by the property detectors also possess some of these secondary attributes. For example, in a V_1CV_2 utterance, the consonant may be properly classified on the basis of the rapid spectrum change at the release into V_2. But the transition from V_1 to C can also be observed by the listener, and the listener's internalized structure of the various features of the consonant C can be expanded to take into account these secondary attributes, as well as the primary properties. The findings of Brady et al.[26] and Nabelek et al.[27] on the perception of rapid frequency changes would suggest that a final vowel-formant transition is stored in memory in terms of the frequency at the end point.

It can be speculated that these secondary attributes of the consonants are available for observation in a precategorical auditory store of the type noted above, and are associated with the phonetic categories established by the property detectors. In time, these secondary cues could play a primary role in identifying the consonants, particularly in cases where the primary attribute involving the rapid spectrum change is weak or absent.

The property detectors, however, play a vital role in getting the process started. What is required is (i) the existence of a few simple property detectors, (ii) the occurrence of a sufficient number of speech events that are correctly classified by these detectors, and (iii) the capability of observing various context-dependent secondary attributes, together with the outputs of the elemental primary detectors.

Presumably, the process of organizing acoustic speech events into classes is enhanced by the experience of the child in producing sounds with his own articulatory system. Sounds that lack an obvious common acoustic attribute (/s/ and /n/, for example) can be placed in the same class, or assigned a common feature, on the basis of a similar articulatory gesture used in producing the sounds. This capability is central to the motor theory of speech perception as proposed by Liberman et al.[28]

Beyond the classification of consonants as labials, coronals, and velars, the child must acquire even finer distinctions in terms of manner of articulation and voicing, and in terms of subclassifications of place of articulation within these broad place categories. These particular phonetic distinctions can be acquired only after the initial broad consonantal categories are established, and are dependent on the language to which the child is exposed. The broad principles that guide this utilization of an expanding inventory of features have been suggested by Jakobson.[25]

5. Summary and Conclusions

The arguments presented in this report are not based on any new experimental data, and many of the points have been made by others. (See, for example, Liberman et al., 1967.[1]) The attempt here is simply to bring together a number of findings relating to speech acoustics, perception of speech and speechlike sounds, and speech acquisition. The principal conclusions are the following:

1. An acoustical basis for the phonetic feature consonantal is proposed: initial consonants that precede vowels are characterized by rapid spectrum change at the vowel onset, whereas vowels and vowel-like sounds do not have such a rapid spectrum change.

2. Place of articulation for many consonants in initial position can be identified absolutely on the basis of the nature of the rapid spectrum change at the vowel onset. The attributes of this spectrum change are determined by a combination of factors, including the transitions of the second and higher formants and, in some cases, by the spectrum of the initial energy burst for stop consonants.

3. Associated with the primary acoustic distinction between vowels and consonants there appear to be qualitatively different modes of perception for these two classes of segments. Primary cues for consonant place of articulation are context-independent, whereas cues for vowel place of articulation, as well as certain secondary cues for consonant features, are context-dependent, and the utilization of these cues requires that they be placed in a precategorical store before decoding into features can be accomplished.

4. In the acquisition of phonetic distinctions by young children, it is postulated that a few simple property detectors can provide a basis for classifying most initial consonants according to gross place of articulation. This classification in terms of context-independent attributes provides a basis for the acquisition of a set of secondary, context-dependent cues for features, and these cues are available for use when the primary cues are weak or absent.

References

1. A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the Speech Code," Psychol. Rev. 74, 431-461 (1967).
2. A. M. Liberman, K. S. Harris, H. S. Hoffman, and B. C. Griffith, "The Discrimination of Speech Sounds within and across Phoneme Boundaries," J. Exptl. Psychol. 54, 358-368 (1957).
3. D. B. Fry, A. S. Abramson, P. D. Eimas, and A. M. Liberman, "The Identification and Discrimination of Synthetic Vowels," Lang. Speech 5, 171-189 (1962).
4. K. N. Stevens, A. M. Liberman, S. E. G. Öhman, and M. Studdert-Kennedy, "Crosslanguage Study of Vowel Perception," Lang. Speech 12, 1-23 (1969).
5. D. Shankweiler and M. Studdert-Kennedy, "Identification of Consonants and Vowels Presented to Left and Right Ears," Quart. J. Exptl. Psychol. 19, 59-63 (1967).
6. R. Crowder, "Visual and Auditory Memory," in J. Kavanagh and I. Mattingly (Eds.), Language by Eye and by Ear: The Relationships between Speech and Reading (The M.I.T. Press, Cambridge, Mass., 1972), pp. 251-275.
7. K. N. Stevens and A. S. House, "Perturbation of Vowel Articulation by Consonantal Context: An Acoustical Study," J. Speech Hearing Res. 6, 111-128 (1963).
8. K. N. Stevens, "The Role of Rapid Spectrum Changes in the Production and Perception of Speech," in L. L. Hammerich and R. Jakobson (Eds.), Form and Substance: Festschrift for Eli Fischer-Jørgensen (Akademisk Forlag, Copenhagen, 1971), pp. 95-101.

9. C. G. M. Fant, "Descriptive Analysis of the Acoustic Aspects of Speech," Logos; Bull. Natn. Hosp. Speech Disorders 5, 3-17 (1962).

10. C. G. M. Fant, "Analysis and Synthesis of Speech Processes," in B. Malmberg (Ed.), Manual of Phonetics (North-Holland Publishing Co., Amsterdam, The Netherlands, 1970), pp. 173-277.

11. R. Jakobson, C. G. M. Fant, and M. Halle, Preliminaries to Speech Analysis (The M.I.T. Press, Cambridge, Mass., 1963).

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

13. K. N. Stevens and D. H. Klatt, "The Role of Formant Transitions in the Voice-Voiceless Distinction for Stops," Quarterly Progress Report No. 101, Research Laboratory of Electronics, M.I.T., April 15, 1971, pp. 188-197.

14. A. Summerfield and M. Haggard, "Perception of Stop Voicing," Speech Perception Report on Research in Progress, Department of Psychology, University of Belfast, Series II, No. 1, pp. 1-14, 1972.

15. A. M. Liberman, P. C. Delattre, L. J. Gerstman, and F. S. Cooper, "Tempo of Frequency Change as a Cue for Distinguishing Classes of Speech Sounds," J. Exptl. Psychol. 52, 127-137 (1956).

16. I. J. Hirsh, "Auditory Perception of Temporal Order," J. Acoust. Soc. Am. 31, 759-767 (1959).

17. I. Nabelek and I. J. Hirsh, "On the Discrimination of Frequency Transitions," J. Acoust. Soc. Am. 45, 1510-1519 (1969).

18. Ibid., p. 1513.

19. K. N. Stevens, "Acoustic Correlates of Certain Consonantal Features," Paper C6, Proc. 1967 Conference on Speech Communication and Processing, Massachusetts Institute of Technology, Cambridge, Massachusetts, 6-8 November 1967, pp. 177-184.

20. F. S. Cooper, P. C. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman, "Some Experiments on the Perception of Synthetic Speech Sounds," J. Acoust. Soc. Am. 24, 597-606 (1952).

21. K. S. Harris, H. S. Hoffman, A. M. Liberman, P. C. Delattre, and F. S. Cooper, "Effect of Third-Formant Transitions on the Perception of the Voiced Stop Consonants," J. Acoust. Soc. Am. 30, 122-126 (1958).

22. E. Fischer-Jørgensen, "Perceptual Studies of Danish Stop Consonants," Ann. Rept. Inst. Phon. 6, University of Copenhagen, 1972, pp. 75-176.

23. K. Harris, "Cues for Discrimination of American English Fricatives in Spoken Syllables," Lang. Speech 1, 1-17 (1958).

24. J. M. Heinz and K. N. Stevens, "On the Properties of Voiceless Fricatives," J. Acoust. Soc. Am. 33, 589-596 (1961).

25. R. Jakobson, Child Language, Aphasia and Phonological Universals (Mouton and Company, The Hague, The Netherlands, 1968).

26. P. T. Brady, A. S. House, and K. N. Stevens, "Perception of Sounds Characterized by Rapidly Changing Resonant Frequency," J. Acoust. Soc. Am. 33, 1357-1362 (1961).

27. I. Nabelek, A. Nabelek, and I. J. Hirsh, "Pitch of Tone Bursts of Changing Frequency," J. Acoust. Soc. Am. 48, 536-553 (1970).

28. A. M. Liberman, F. S. Cooper, K. S. Harris, and P. F. MacNeilage, "A Motor Theory of Speech Perception," Proc. Speech Communications Seminar, Stockholm, 1962; D3, Royal Institute of Technology, Stockholm, 1963.

B. MEASUREMENT OF VIBRATION PATTERNS OF EXCISED LARYNXES

National Institutes of Health (Grant 5 RO1 NS04332-10)

T. Baer

Excised larynxes are useful for studying the mechanism of phonation, since they can be observed and monitored more adequately than normal larynxes and since the parameters affecting phonation (such as respiratory parameters and simulated activity of most of the laryngeal muscles) can be effectively controlled and systematically varied.[1,2] There are also, of course, limitations – particularly, that activity of the vocalis muscle cannot be simulated, and also perhaps that the isolated larynx is disconnected from its vocal tract. Nevertheless, understanding the excised preparation seems an appropriate first step toward understanding the more complicated live larynx. Finally, although there have been numerous attempts to model the phonatory mechanism mathematically,[3-9] available experimental data are inadequate for evaluating in detail the performance of many of these models.

In order to obtain detailed measurements of the time-variant vocal fold shape during phonation, we have made visual observations of excised dog larynxes from the subglottal, as well as the supraglottal, aspect. Small particles, distributed on the vocal folds, are illuminated stroboscopically and tracked, using a microscope, at successive phase increments throughout a glottal cycle.

Figure VIII-4 is a schematic diagram and Fig. VIII-5 a photograph of the apparatus. The larynx preparation is similar to that employed by van den Berg.[2] Threads are used to simulate the activity of most of the laryngeal muscles. A rigid bar is attached to the cricoid cartilage and then to the apparatus by an adjustable clamp to immobilize the cartilage. The short section of trachea associated with the larynx is clamped to the pseudo-subglottal system. The respiratory source of this system currently supplies steady airflow at an adjustable rate, although the system will be changed to regulate pressure rather than flow. The rest of the subglottal system is composed of a simulated lung volume and trachea, which makes a right-angle turn in order to accommodate a subglottal window just before reaching the larynx. The output of a subglottal pressure transducer triggers an oscilloscope, which in turn triggers a stroboscope with an adjustable delay. The larynx and other apparatus are mounted on the top of a rotary indexing table that can be translated along two rectangular coordinates and rotated. The microscope can be translated vertically.

The table top is first adjusted along its rectangular axes so that its rotational axis is in focus in the middle of the microscope's field. This produces a reference position relative to which measurements are made. The location of an individual particle at a

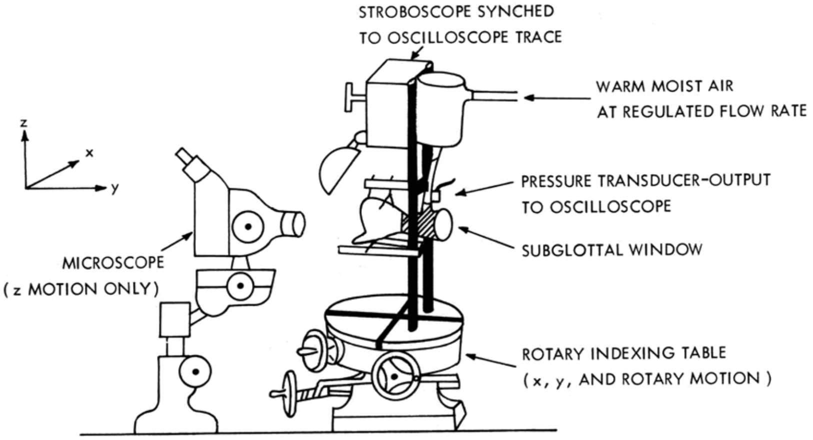
STROBOSCOPE SYNCHED TO OSCILLOSCOPE TRACE
WARM MOIST AIR AT REGULATED FLOW RATE
z
x
y
PRESSURE TRANSDUCER-OUTPUT TO OSCILLOSCOPE
SUBGLOTTAL WINDOW
MICROSCOPE (z MOTION ONLY)
ROTARY INDEXING TABLE (x, y, AND ROTARY MOTION)

Fig. VIII-4. Diagram of the apparatus.

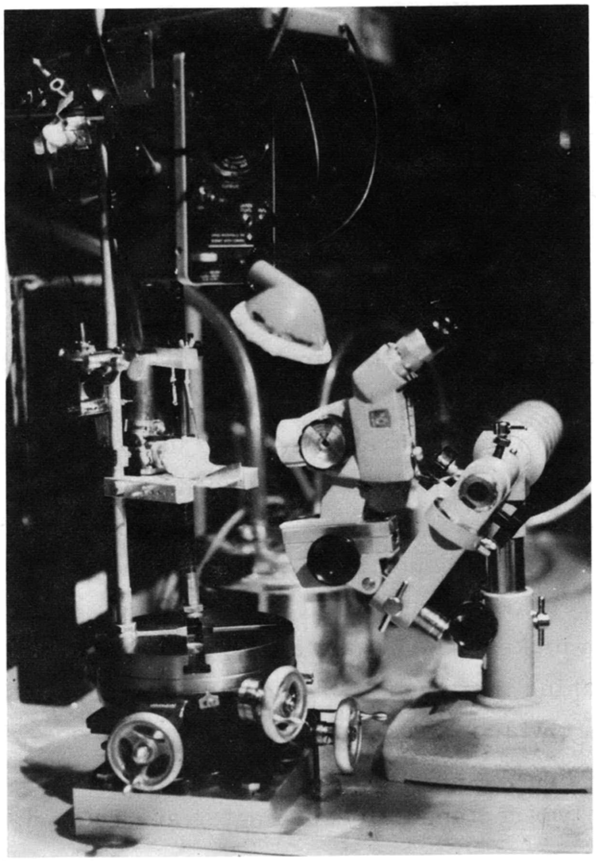

Fig. VIII-5. Photograph of the apparatus.

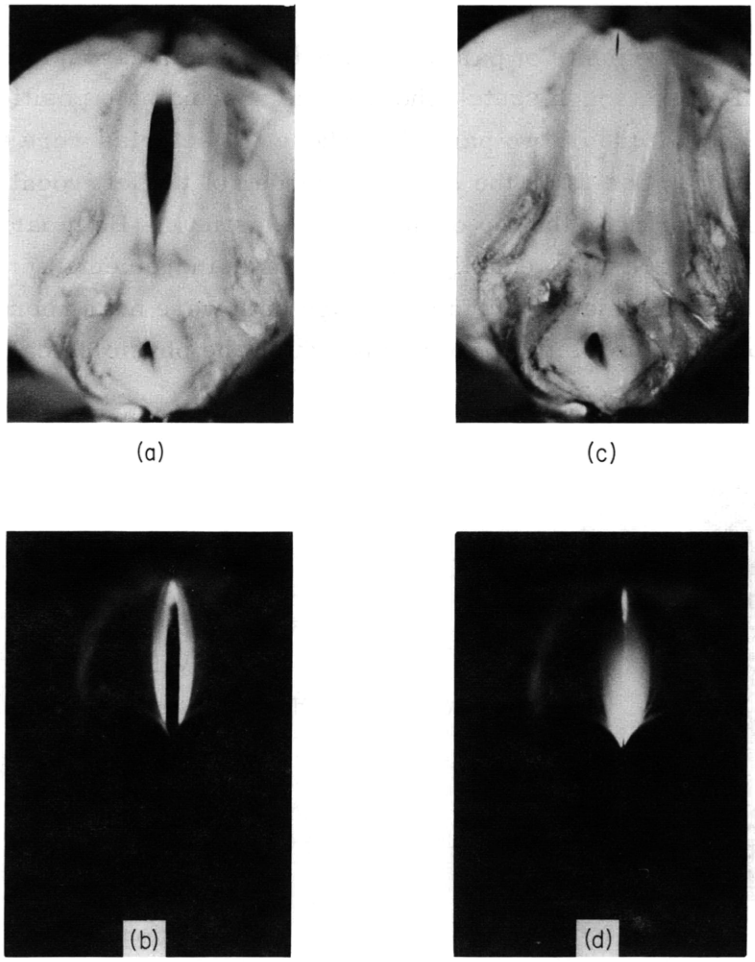

Fig. VIII-6. Vibrating larynx.
(a) Open phase, supraglottal view.
(b) Open phase, subglottal view.
(c) Closed phase, supraglottal view.
(d) Closed phase, subglottal view.

given phase in the glottal cycle is measured by translating the table top until the particle is in focus in the middle of the field. Measurements can be made from any angle, by rotating the table top, and then transformed to a coordinate system fixed with respect to the table top, and hence with respect to the larynx. With the present equipment, spatial resolution is 0.05-0.1 mm.

Figure VIII-6 shows gross views of a vibrating larynx from the supraglottal and subglottal aspects during the open and closed extremes of its cycle. The mode of vibration appears similar to that seen in photographs of the normal larynx.[10]

Figure VIII-7 is a closer view, taken through the microscope at low magnification, showing some particles on the vocal-fold surfaces. For measurements, higher microscope magnification is used.

Preliminary measurements of particle trajectories in a coronal plane have been gathered. Figure VIII-8 is a rough sketch showing the approximate positions with respect to the bodies of the vocal folds of two particles whose trajectories were tracked in a typical experimental session. One is on the superior border of the left vocal fold, while the other is on the base, or near the top of the conus elasticus. Both particles can be seen from some angle at each of 8 equally spaced phase increments throughout a cycle. Neither was on an approximating surface, though both were near approximating surfaces. The data were collected in a single run during which the physiological state of the preparation was approximately constant.

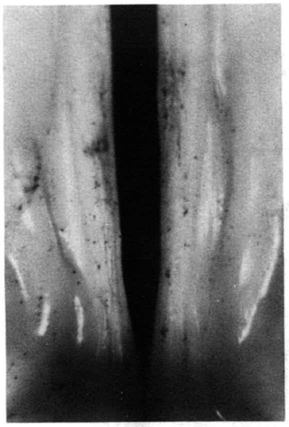

Fig. VIII-7.
Vocal folds showing particles (black spots).

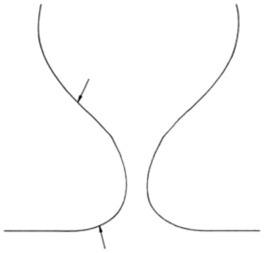

Fig. VIII-8.
Coronal section of vocal folds showing positions of two particles whose trajectories were examined.

Figure VIII-9 shows the measured trajectories of these two particles in the coronal plane. The 8 data points for each particle have been connected by a smooth curve to form the closed trajectory. (Note that, although the specimen is canine, we are consistently using human anatomical terminology.) Some of the relevant parameters of the vibrations are tabulated in the figure.

These data, at least in some ways, are typical. The superior particle traverses greater distances than the inferior particle, although that is less obvious in this case than in others. Its trajectory has typically a greater vertical than horizontal component.

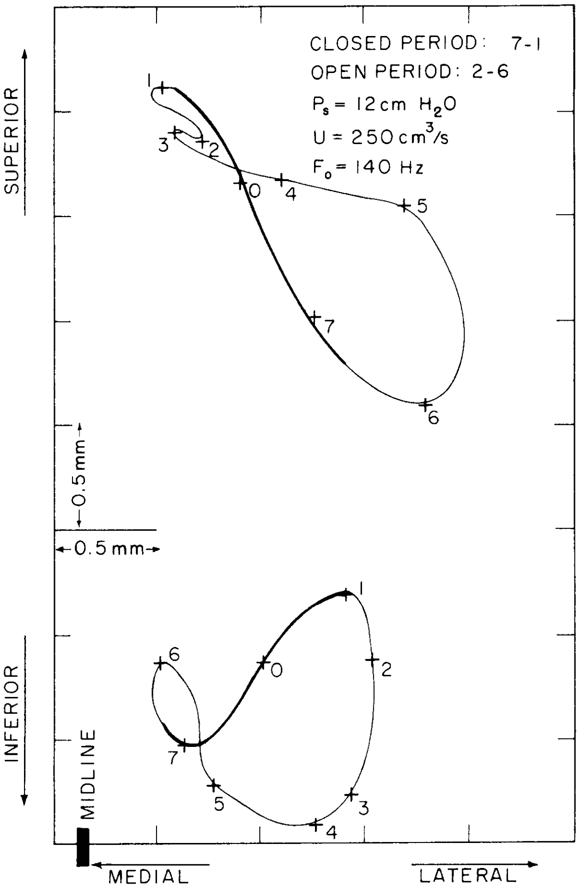

Fig. VIII-9. Trajectories of the two particles indicated in Fig. VIII-8. Positions in the coronal plane were measured at 8 phase increments throughout the cycle. (P_S = subglottal pressure; U = average volume velocity; F_O = fundamental frequency.)

The trajectory of the inferior particle has a proportionally larger horizontal component. During glottal closure, which is indicated by the bolder segments of the trajectories, the inferior particle is pushed in the superior and lateral directions, while the superior particle moves in the superior and medial directions.

Ultimately, we hope that a sufficient number of particles can be tracked so that at least the outline of the vocal folds in a coronal plane can be reconstructed. Obviously, two particles are not sufficient for this purpose. Still, it is useful to describe how the particle motions reflect the gross vibration patterns, which are typical of a mode in which the excised larynxes vibrate.

When the glottis is open, around phase 4 or 5, a wavelike ripple forms below the lower particle (perhaps in a manner resembling the formation of water waves by wind). The wave propagates superiorly and grows in amplitude. In approximately the region where the trajectory forms a small loop the wave passes the inferior particle, so that the particle is above the crest at 5, near the crest at 6, and below the crest at 7. Around phase 7, the waves from opposite sides meet and closure occurs. The area of closure, which is quite small in vertical extent, propagates upward, since tissues below are peeled apart while material above continues to move toward the midline. Then, a wave propagates laterally along the superior surface. Around phase 5, the shape is actually such that the superior particle appears on a medial rather than on a superior surface. By this time the glottis is wide open, and with the increased airflow a new wave is formed below to repeat the cycle.

For a sufficient number of points to be tracked to reconstruct coronal sections and three-dimensional surfaces, methods must be developed to increase the time during which the larynx will sustain steady vibrations. When this can be done, the shape data that can be obtained should be useful for deducing the physical mechanisms involved. In addition, however, dynamic pressure and airflow will be measured in detail near the end of the pseudo trachea. With these data, it should be possible to estimate the pressure distribution along the walls, using calculations or, if necessary, physical models that can be instrumented. With these pressure data, as well as mechanical measurements, it should be possible to deduce some mechanical properties of the system and to evaluate some parameters of models.

Dog larynxes have been used for these experiments because they are easy to obtain and are similar to human larynxes. We intend to make similar measurements on excised human larynxes. Still the measurements on dogs may be very useful. It seems practically possible to do similar experiments with live dog preparations, in which actual muscle activity could be directly controlled. With detailed data on both excised and live dog preparations, there would presumably be more basis for inferring results from excised to live human larynxes.

References

1. Jw. van den Berg and T. S. Tan, "Results of Experiments with Human Larynxes" Pract. Otol.-Rhinol.-Laryngol. (Basel) 21, 425-450 (1959).
2. Jw. van den Berg, "Sound Production in Isolated Human Larynges," Ann. N.Y. Acad. Sci. 155, 18-27 (1968).
3. R. L. Wegel, "Theory of Vibration of the Larynx," J. Acoust. Soc. Am. 1 (Suppl. No. 3), 1-21 (1930).
4. T. H. Crystal, "A Model of Laryngeal Activity during Phonation," Sc.D. Thesis, Department of Electrical Engineering, M.I.T., June 1966.
5. J. Flanagan and L. Landgraf, "Self-Oscillating Source for Vocal Tract Synthesizers," IEEE Trans., Vol. AU-16, pp. 57-64, 1968.
6. D. Dudgeon, "Multi-Mass Simulation of the Vocal Cords," S.M. Thesis, Department of Electrical Engineering, M.I.T., 1970.

7. K. Ishizaka and J. L. Flanagan, "Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Cords," Bell Syst. Tech. J. 51, 1233-1268 (1972).

8. S. Hiki, Y. Koike, and H. Takahashi, "Simultaneous Measurement of Subglottal and Supraglottal Pressure Variations," J. Acoust. Soc. Am. 48, 118-119 (1970).

9. I. R. Titze and W. J. Strong, "Simulated Vocal Cord Motion in Speech and Singing," J. Acoust. Soc. Am. 52, 123 (1972).

10. S. Smith, "Remarks on the Physiology of the Vibrations of the Vocal Cords," Folia Phoniatr. 6, 166-178 (1954).