From Continuous Signal to Discrete Message: Syllable to Phoneme*

Michael Studdert-Kennedy†
Haskins Laboratories, New Haven, Conn.

All speech is syllabic. All languages constrain syllabic structure, and base their constraints on the natural phonological contrast between consonant and vowel. Certainly, the human vocal tract is capable of producing an indefinite sequence of vowels, and even of certain consonants. But the choice of all languages has been to alternate consonants, or consonant clusters, and vowels. What is the communicative function of this choice? Mattingly (this conference) has sketched the broad outline. He has pointed to the complex encoding of discrete phonetic elements into a more-or-less continuous acoustic signal; he has shown how the code permits unusually rapid and efficient transfer of information; he has drawn striking formal analogies among phonetics, grammar, and memory. My purpose is to examine two aspects of the phonetic code in more detail: first, the structure and function of the syllable as reflected in perception; second, more briefly, the underlying cerebral mechanisms. Taken together, they suggest a natural system elegantly adapted to its role in linguistic communication.

Perceptual Structure and Function of the Syllable

Let me say, to begin with, that precise definition of consonant and vowel¹ is even more difficult than the definition of syllable. For, as Sweet (1877) observed, "the boundary between vowel and consonant, like that between the different kingdoms of nature, cannot be drawn with absolute definiteness" (p. 51). Most of what I have to say deals only with stop consonants and relatively sustained monophthongal vowels. Our knowledge of speech perception comes largely from study of these two phonetic classes and their combination in a simple CV syllable. But I believe this knowledge can furnish quite general insights.

---


†Also Queens College and the Graduate Center of the City University of New York.

¹"Consonant" and "vowel" refer to elements in the phonetic message rather than to their correlates in the acoustic signal. I have used the terms for both meanings in what follows and trust that context will make clear which is intended.

Acknowledgment: I am grateful to Alvin Liberman, Ignatius Mattingly, and Donald Shankweiler for comments and criticism.

The syllable is an articulatory and acoustic integer. Greek grammarians gave us the word for it: syllabē, "a taking together," syllable. Roman grammarians gave us words for the elements that are taken together: littera vocalis, "the voiced letter," vowel; and littera consonante, "the letter that sounds with" the vowel, consonant. Into the syllable the speaker encodes, and from it the listener decodes, these discrete, phonetic segments. Among the evidence for the psychological reality of such segments is the corpus of spoonerisms gathered by Fromkin (1971). Here, one aspect of her data is of particular interest. Speakers may blunder by exchanging syllables for syllables, consonants for consonants, and vowels for vowels, but they never metathesize across classes. They may say, "I'll have a slice of [bost rif]," or even "of [rist bof]," but never "of [iost brf]." The syllable, itself a functional unit, is compounded of consonant and vowel, each fulfilling some syllabic function that forbids their metathesis.

Stetson (1951) gave us some understanding of these functions in production. He recognized the syllable as the fundamental unit of speech, the unit of stress contrast, of rhythm and meter. His motor definition of the syllable as a "chest pulse" has not stood up, at least for unstressed syllables (Ladefoged, 1967: Ch. 1). But his description of the time course of the syllable is still useful. He described the consonant-vowel-consonant (CVC) syllable as "a single ballistic movement," composed of release, nucleus, and arrest. Similarly, Pike (1943) described speech as alternate constrictions and openings of the vocal tract.

One obvious acoustic correlate of the syllabic movement can be seen in the amplitude display of a spectrogram. The onset of an open (CV) syllable tends to yield a low, but rapidly rising, amplitude, the nucleus a relatively sustained peak. For a longer utterance we can make a crude syllabic count from the amplitude peaks. The count is crude because many utterances (the word "tomorrow" [tomurro], for example) may yield a single peak (due only in part to the sluggishness of the spectrograph's amplitude integrating response), even though we know they contain several syllables. Where amplitude fails to reveal syllabic structure, formant pattern may serve: the initial pattern tends to display a rapid movement, or scatter, of energy over the frequency domain, while the later portion tends to be relatively stable and sustained (cf. Malmberg, 1955). Taken together, changes in amplitude and frequency offer an acoustic contrast between the beginning and the end of a CV syllable, between its onset and its nucleus. The event is unitary, but its character changes as it occurs. This ill-defined acoustic contrast provides the auditory ground for the perceptual consonant and vowel.

Let us turn, first, to the phenomenon of categorical perception. Experiments have repeatedly revealed differences between stop consonants and vowels in their patterns of identification and discrimination. Stimuli for these experiments consist of a dozen or so synthetic speech sounds distributed in equal acoustic steps along a continuum ranging across two or more phonetic categories (from, say, [b] to [d] to [g] or from [i] to [I] to [e]). In identification, listeners are asked to assign each of the stimuli, presented repeatedly and in random order, to one of the phonetic classes. They then assign consonants more consistently than they do vowels, particularly those tokens close to a boundary between phonetic classes. In other words, they identify consonants absolutely or "categorically," independently of the test context, while they identify vowels relatively or "continuously," with marked contextual effects (Liberman, Harris, Kinney, and Lane, 1961; Fry, Abramson, Elmas, and Liberman, 1962).
Here we have the first, and oldest, indication that listeners have a longer short-term auditory store for vowels than for consonants. Recently, Sawusch and Pisoni (1973) have elaborated with the finding that vowels, like nonspeech tones, are susceptible to psychophysical anchoring effects: the boundary between synthetic vowels along an acoustic continuum is shifted toward the vowel that occurs most frequently in the test. For consonants, the effect is absent: the listener relies on some internal standard, less readily subverted by test composition. Note, incidentally, that if a vowel, already assigned to its phonetic class, is to affect phonetic assignment of following vowels, the listener must retain an auditory image, or echo, of the vowel even after he has identified it. The process of identification does not therefore terminate auditory display: auditory store and phonetic store can exist simultaneously (cf. Wood, 1973b).

In the related discrimination task, typically administered in ABX format, a listener is called on to discriminate between pairs of tokens separated by one or more equal acoustic steps along the synthetic continuum. If these tokens are drawn from different phonetic classes, discriminative performance is high for both consonants and vowels. If they are drawn from the same phonetic class, discriminative performance drops slightly for vowels and considerably, to a point little better than chance, for consonants. In other words, a listener discriminates between vowels at a relatively high level whether he assigns them to different phonetic categories or not: his discrimination is more or less independent of identification, much as it is for nonspeech sounds (Mattingly, Liberman, Syrdal, and Halwes, 1971). But a listener can reliably discriminate between consonants only if he assigns them to different phonetic categories: his discrimination depends upon and, to a fair degree, can be predicted from his phonetic assignments (Liberman et al., 1961; Studdert-Kennedy, Liberman, Harris, and Cooper, 1970).

Early accounts of this phenomenon pointed to articulatory differences between consonants and vowels. But their acoustic differences have proved more crucial. Stevens (1968) remarked the brief, transient nature of consonantal acoustic cues, and Sachs (1969) showed that vowels were more categorically perceived if their duration and acoustic stability were reduced by CVC context. Lane (1965) pointed to the greater duration and intensity of the vowels and showed that they were more categorically perceived if they were "degraded" by being presented for discrimination in noise.

The role of auditory, or echoic, memory, implicit in the work of Stevens, of Sachs, and of Lane, was made explicit by Fujisaki and Kawshima (1969, 1970). They argued that the listener's poor auditory memory for consonants forced him to rely, for ABX discrimination, on phonetic memory. They formulated a mathematical model of the process and showed that, if they reduced vowel duration sufficiently, their model would predict quite accurately the discrimination of both consonants and vowels from listeners' phonetic identifications.

Pisoni (1971, 1973) made a direct test of this account. He varied the intratrial interval in an AX "same"-"different" task. For vowels, an increase in the A-X interval (with a presumed decrease in clarity of A's auditory store) led to a decrease in the likelihood that a listener would judge two acoustically different, but phonetically identical, tokens as "different." For consonants, there was no significant effect. In several other experiments, Pisoni (in press) has shown that the degree to which vowels are perceived categorically (measured by the degree to which phonetic class predicts discriminability) may be varied by
manipulating the degree to which auditory memory is made available in the experimental task. Note, however, that while a fair degree of categorical perception of vowels can be readily induced, continuous perception of consonants is much more difficult (Pisoni and Lazarus, in press). The listener's auditory memory for consonants is intrinsically short.

Consider next dichotic ear advantages. As is well known, Kimura (1961a, 1961b, 1967) showed that if different digits are presented to opposite ears at the same time (i.e., dichotically), those presented to the right ear are recalled more accurately than those presented to the left ear. She attributed the effect to specialization of the left cerebral hemisphere for language functions and to stronger contralateral than ipsilateral ear-to-hemisphere connections. The effect and her interpretation have been repeatedly supported.

Shankweiler and Studdert-Kennedy (1967) used Kimura's technique to probe the processes of speech perception. They showed that the right-ear advantage did not depend on higher language processes, since it could be obtained with pairs of nonsense syllables differing only in an initial or final stop consonant. Furthermore, they showed that the effect did not appear if the competing syllables were steady-state vowels or CVC syllables differing only in their vowels (Studdert-Kennedy and Shankweiler, 1970). Following Kimura and others (Milner, Taylor, and Sperry, 1968; Sparks and Geschwind, 1968), they assumed that ipsilateral ear-to-hemisphere connections were inhibited by dichotic competition, so that, while right-ear inputs reached the left (language) hemisphere by a direct contralateral path, left-ear inputs traveled an indirect route, contralaterally to the right hemisphere, then laterally across the corpus callosum to the left hemisphere. The ear advantage was due to loss of auditory information from the left-ear signal as it traveled its indirect path to the language hemisphere. The consonant-vowel difference in ear advantage could then be attributed to the same acoustic factors as their differences in degree of categorical perception: the vowel portion of the signal, being more intense and of greater duration than the consonant, suffers less loss or "degradation" on the left-ear-to-left-hemisphere indirect path, and so yields no reliable right-ear advantage. This is probably not the whole story, since nonacoustic, attentional factors have also been implicated (e.g., Spellacy and Blumstein, 1970). However, Weiss and House (1973) have played for dichotic ear advantages the role that Lane (1965) played for categorical perception: they have shown that a right-ear advantage appears for vowels if the vowels are presented dichotically in noise.

Yet another experimental paradigm yielding stop consonant-vowel differences, this time directly in short-term memory, is due to Crowder (1971, 1972, 1973). If subjects are given, one item at a time, a span-length list of digits for immediate, ordered recall, they recall the last several digits more accurately when the list has been presented by ear than when it has been presented by eye. This modality difference presumably reflects the operation of separate modality stores, the short-term auditory store being more retentive than the visual (Crowder and Morton, 1969). This interpretation is supported by the fact that the advantage to the most recent auditory items (recency effect) is reduced or eliminated if the list ends with a redundant verbal item, as a signal for the subject to begin recall (suffix effect). That the suffix interferes with auditory store is suggested by the fact that the effect occurs for either backwards or forwards speech, is reduced if list and suffix are spoken in different voices, and is absent if the suffix is a tone.
The finding of interest in the present context is simply that while all three effects (modality, recency, suffix) are observed for lists of CV syllables differing in their vowels (e.g., random repetitions of /gæ, ga, gA/ or of /bi, bo, bu/), none of them is observed for CV or VC syllables differing in their initial or final stop consonant (random repetitions of /ba, da, ga/ or of /ab, ad, ag/). Crowder (1971) concludes that vowels are included in precategorical, auditory store, but that stop consonants are excluded. Liberman, Mattingly, and Turvey (1972) agree with this conclusion, arguing further that phonetic decoding of the stops "strips away all auditory information," that phonetic classification terminates auditory display.

While this interpretation is unlikely to be correct (cf. Wood, 1973b), the difficulty of retaining auditory information about stop consonants is again suggested by results from a fourth experimental paradigm, devised by Dorman (1973). He synthesized three three-formant sounds: a 250 msec /bæ/, a 250 msec /æ/, and a "chirp" consisting of the first 50 msec of the synthesized /bæ/. The "chirp" contained all the acoustic information necessary for phonetic classification of the initial /b/, but, separated from the following vowel, no longer sounded like speech. Dorman next varied the intensity of the "chirp" and of the first 50 msec of the two speech sounds in two steps: 0, -7.5, and -9 db. He then presented these stimuli in pairs, each member drawn from the same stimulus type, to ten subjects and asked them to judge whether the initial intensities of the pairs were the "same" or "different." The results were that every subject gave close to 100% performance on the vowels and "chirps," and close to 50% performance, or chance, on the CV syllables. In other words, asked to judge acoustic differences irrelevant to segmental classification, subjects could detect those differences in vowels or nonspeech sounds, but not in stop consonants.

We must not exaggerate. Many experiments have demonstrated that listeners do retain at least some "echo," however rapidly fading, of stop consonants. All studies of categorical perception reveal some margin of auditory discriminability within stop consonant categories, and several experimenters have tested this directly. Barclay (1972), for example, showed that listeners could reliably judge variants of /d/, drawn from a synthetic continuum, as more like /b/ or more like /g/. Pisoni and Tash (in press) found that reaction times for "same" responses were faster to pairs of acoustically identical stop-vowel syllables than to pairs of phonetically identical, but acoustically different, syllables.

Furthermore, Darwin and Baddeley (in press) have recently challenged Crowder's interpretation that stop consonants are excluded from precategorical store. They have shown that a moderate recency effect may be obtained with consonants if the syllables in the list are acoustically distinct (/a, ma, ga/), and even more if the consonants are in syllable-final position. They argue that listeners cannot make use of their auditory store of the later items in a list if those items are acoustically similar and confusable, as are /ba, da, ga/. They support their argument by demonstrating that the recency effect can be reduced or abolished for vowels if the vowels are very brief (30 msec of steady-state in a 60 msec CV syllable) and occupy neighboring positions on an F1-F2 plot. They conclude that "the consonant-vowel distinction is largely irrelevant," and they propose "acoustic confusability" as the determining variable.

However, among the determinants of "acoustic confusability" are the very acoustic factors that distinguish stop consonants from vowels, namely energy and spectral stability. We have reviewed evidence from four experimental paradigms
in which consonant and vowel perception differ. In three of these (and, no doubt, if one chooses, in the fourth) the differences can be reduced or eliminated by taxing the listener's auditory memory for the vowel or by sensitizing it for the consonant. But these qualifications do not mitigate the consonant-vowel differences: they merely emphasize that the differences are there to be eliminated. There is little question that consonants are less securely stored in auditory memory than vowels. [For further discussion see Studdert-Kennedy (in press-a, in press-b).]

Plausible communicative functions for these differences are not hard to find. Consider, first, vowel duration. Long duration is not necessary for recognition. We can identify a vowel quite accurately and very rapidly from little more than one or two glottal pulses, lasting 10 to 20 msec. Yet in running speech vowels last 10 to 20 times as long. The increased length may be segmentally redundant, but it permits the speaker to display other useful information: variations in fundamental frequency, duration, and intensity within and across vowels offer possible contrasts in stress and intonation, and increase the potential phonetic range (as in tone languages). Of course, these gains also reduce the rate at which segmental information can be transferred, increase the duration of auditory store, and open the vowel to contextual effects, the more so, the larger the phonetic repertoire. A language built on vowels, like a language of cries, would be limited and cumbersome.

Adding consonantal "attack" to the vowel inserts a segment of acoustic contrast between the vowels, reduces vowel context effects, and increases phonetic range. The attack, itself part of the vowel (the two produced by "a single ballistic movement"), is brief, and so increases the rate of information transfer. Despite its brevity, the attack has a pattern arrayed in time and the full duration of its trajectory into the vowel is required to display the pattern. To compute the phonetic identity of the pattern, time is needed, and this is provided by the segmentally redundant vowel. Vowels are the rests between consonants.

Rapid consonantal gestures cannot carry the melody and dynamics of the voice. The segmental and suprasegmental loads are therefore divided over consonant and vowel—the first, with its poor auditory store, taking the bulk of the segmental load, and the second taking the suprasegmental load. There emerges the syllable, a symbiosis of consonant and vowel, a structure shaped by the articulatory and auditory capacities of its user, fitted to, defining, and making possible linguistic and paralinguistic communication.

Cerebral Specialization for Syllable Perception

The distinctive acoustic structure of the syllable into which the speaker encodes consonant and vowel seems to call for a specialized neurophysiological decoding mechanism in the listener. Evidence for the operation of such a mechanism first came from the dichotic listening studies mentioned above (Shankweiler and Studdert-Kennedy, 1967; Studdert-Kennedy and Shankweiler, 1970). One question that these studies tried to answer was whether the mechanism was specialized for both auditory and phonetic analysis of the syllable or for phonetic analysis alone. I will not review the evidence here, but simply state our conclusion that "while the auditory system common to both hemispheres is equipped to extract the auditory parameters of a speech signal, the dominant hemisphere may be specialized for the extraction of linguistic features from these parameters" (Studdert-Kennedy and Shankweiler, 1970:594).
As we shall see shortly, recent evidence suggests that this conclusion may not be correct: the left hemisphere may be specialized for both auditory and phonetic analysis. First, however, we should note that Wood (1973a; also Wood, Goff, and Day, 1971) has provided impressive support for the conclusion in a study of the evoked potential correlates of phonetic perception. He synthesized two stop-vowel syllables, /ba/ and /ga/, which differed only in the extent and direction of their second and third formant transitions, the acoustic cues to their phonetic identities. He synthesized each at two fundamental frequencies: 104 Hz (low) and 140 Hz (high). From these syllables he constructed two types of test. In the first, fundamental frequency was held constant and the syllables were presented binaurally in random order: subjects identified each syllable phonetically, as fast as possible, by pressing one of two buttons. In the second type of test, phonetic identity was held constant, while fundamental frequency varied: subjects identified the fundamental frequency of each syllable as high or low. Both types of test therefore contained tokens of the same syllables, identified by pressing the same button with the same finger. During the tests, electrical activity was recorded from a central and a temporal scalp location over both left and right hemispheres. Evoked potentials were averaged and compared at each scalp location for the prerelapse periods during presentation of identical syllables in the two tasks. Notice that the only possible source of variation in the EEG compared was in the task carried out by the subjects while the records were taken.

Statistical tests revealed significant differences between left-hemisphere records for the phonetic and fundamental frequency tasks at both locations. No significant differences appeared for either of the right-hemisphere locations. Furthermore, when the "speech" task called for identification of the isolated formant transitions of the two syllables—acoustic patterns which carry all the information necessary for phonetic identification, but which, lacking a following vowel, are not heard as speech—there were no significant left-hemisphere differences between records for "speech" and nonspeech tasks. The previously observed differences cannot therefore have been due to auditory analysis of the information-bearing formant transitions, but must presumably be attributed to phonetic interpretation of the auditory patterns. The experiments leave little doubt that different neural events occur in the left hemisphere, but not in the right hemisphere, during phonetic, as opposed to auditory, analysis of the same acoustic signal.

In other words, the language hemisphere does indeed appear to be specialized for phonetic interpretation (and, presumably, higher language functions), but not for auditory analysis of speech. This might seem to imply that the physical vehicle of the phonetic message is a matter of indifference. Superficial support for this view comes from two further sources. First, Papçun, Krashen, Terbeek, Remington, and Harshman (1974; also Krashen, 1972), have shown that experienced Morse code operators, identifying both individual letters and words presented dichotically, show a significant right-ear advantage. Second, Kimura (see Kimura and Durnford, in press) and others have repeatedly shown a right-field (left-hemisphere) advantage for tachistoscopically presented letters. If both Morse code and printed letters can invoke left-hemisphere processing, there might seem to be little reason to claim any special status for speech.

Nonetheless, there are solid grounds for making this claim. First, several studies have suggested that the left hemisphere is specialized for extracting acoustic features important in speech. Halperin, Nachshon, and Carmon (1973)
have shown that the dichotic ear advantage shifts from left to right as the number of transitions in brief, temporally patterned sound sequences increases. Among their stimuli were permutations of three long (400 msec) and short (200 msec) sound bursts similar to the patterns used in Morse code. Their results therefore fit neatly with those of Krashen (1972) who found that naive subjects have a right-ear advantage for dot-dash sequences no more than seven units long. Taken together, the two studies suggest left-hemisphere specialization for judging duration and temporal pattern. Both studies have the weakness that subjects were asked to label the sequences, a process that might well invoke left-hemisphere control.

This objection is not decisive since arbitrary labeling of isolated formant transitions in Wood's (1973) study did not evoke the left-hemisphere potentials of phonetic labeling. Nonetheless, the weakness is avoided in some recent experiments by Cutting (in press). In one of these he constructed two-formant patterns identical with patterns signaling /bV/ or /dV/ except that their first-formant transitions fell rather than rose along the frequency scale, producing a phonetically impossible sound that subjects did not recognize as speech. In a nonlabeling dichotic recognition task with these stimuli, subjects gave a right-ear advantage of the same magnitude as for the normal CV syllables also used in the study. Cutting concludes from this and other experiments that the left hemisphere may be specialized for auditory analysis of speech.

But why then did the isolated transitions of Wood (1973) yield no left-hemisphere effect? The answer to this may be that the speech auditory analyzer is engaged not simply by acoustic features, but by features distributed over a signal of some minimum duration (such as that of a stressed syllable). Here the work of Wollberg and Newman (1972) on squirrel monkey is suggestive. They made single-cell recordings from cortical neurons responsive to the species' "isolation peep." A normal pattern of neuronal response occurred only if the entire "peep" was presented. Perhaps it is not far-fetched to suppose that the human cortex is supplied with sets of acoustic detectors tuned to speech, each inhibited from output to the phonetic system in the absence of collateral response in other detectors.

Be that as it may, the evidence for specialized left-hemisphere auditory analysis is, at best, preliminary and, in any case, not essential to the claim of special status for speech. Nor, indeed, is any form of speech-specific auditory analysis, whether unilateral or bilateral. Certainly, the accumulating evidence for specialized acoustic property detectors (Cutting and Eimas, this conference) is important and may even be decisive. But the initial strength of the claim comes from the distinctive structure of the syllable. The underlying phonological elements that determine this structure are common and peculiar to all languages. And recovery of those elements, whether from alphabet, optophonic light pattern, Morse code, or the neural display of an auditory system, engages mechanisms in the language hemisphere. The syllable is the structure on which the hierarchy of language is raised.

REFERENCES

Cutting, J. E. and P. Eimas. (this conference) Phonetic feature analyzers and the processing of speech in infants. [Also in Haskins Laboratories Status Report on Speech Research SR-37/38 (this issue).]


Stetson, R. H., (1951) **Motor Phonetics.** (Amsterdam: North-Holland).


Sweet, H. (1877) **Handbook of Phonetics.** (Oxford: Clarendon). (Also, College Park, Md.: McGrath, 1970.)


