Reprinted from

Life Sciences Research Report 5

Theodore H. Bullock

Editor

Recognition of Complex Acoustic Signals

Dahlem Konferenzen

Universals in Phonetic Structure and Their Role in Linguistic Communication

M. G. Studdert-Kennedy
Haskins Laboratories Inc.
270 Crown Street, New Haven, CT 06511, USA

Abstract. All known languages display duality of patterning, phonological system and structure. All spoken languages are syllabic and constrain ~~perceptual~~ structure in terms of consonants and vowels. The syllable is a unit of timing in articulation, of contrast and compression in perception. These functions arise from the temporal structure of an acoustic signaling system and are fulfilled by spatial structure in the visual signaling system of American Sign Language. Syllabic compression poses a problem for the perceiver, if he is to recover discrete message units from an unsegmented signal. A possible mechanism based on acoustic feature detectors, for accomplishing the segmentation, and an alternative process, based on perceptual contrast and continuous tracking of the signal, are considered.

INTRODUCTION

If we are to compare the acoustic signaling systems of man and other animals, we must begin by distinguishing between the physical signal and the perceived message. To understand the importance of this distinction, compare the approaches of a phonetician and a cryptographer to the spectrographic display of a spoken utterance. The phonetician comes armed with preconceptions as to how the utterance is to be segmented - into phonemes, syllables, words - and, naturally enough, seeks in the spectrogram the acoustic correlates of those segments. The fact that he finds, on the one hand, more segments than he wants and, on the other hand, many acoustically indivisible segments that he knows must correspond to two or more perceived segments, has been a main reason for the

Manuscript received: 1976, August 17; received: 1976, November 24.

hypothesis that speech perception engages specialized decoding mechanisms. By contrast, the cryptographer, as the student of birdsong, will begin by dividing the utterance into acoustically distinct segments. Then, with all the means at his disposal - systematic distributional analysis of other utterances, segment transposition, informant tests, and so on - he will try to determine the groupings and cuts in the acoustic signal necessary to derive the set of functional (as opposed to physical) segments that constitute the message. Not surprisingly, phoneticians have shrunk from this task. They have preferred to accept the segments of abstract linguistic analysis and to search for their correlates in the signal. For the present, we have little choice but to continue this tradition and most of what I have to say circles around the resulting problem of segmentation. But before I come to this, a few further general points must be made.

SOME LANGUAGE UNIVERSALS

All known languages (and perhaps some animal communication systems) display "duality of patterning" [6]. Utterances can be described, at the syntactic level of the message, as sequences of lexical and grammatical segments (words or morphemes); at the phonological level, as sequences of meaningless segments (phonemes). Words leave their faint traces only in prosodic features of the signal, if at all, and our present concern is entirely with the "lower" meaningless segments.

All known spoken languages have a sound system, or phonology, based on feature opposition. Sound units, the phonemes of a language, are a relatively small set (usually a few dozen) of meaningless segments that serve to distinguish among its words. For example, the words "bed" and "red" are distinguished by their initial, but not by their medial or final phonemes. The phonemes are not randomly selected. Each has a characteristic internal structure that may be described in terms of the small set of phonetic features (usually, a dozen

or so) used in a particular language. The phonemes may be classified according to their shared phonetic features, and the resulting classes contrasted with one another on the basis of their feature differences, or oppositions. For example, English /b/ and /p/, formed by closure of the vocal tract at the lips, share the feature "labial". They contrast with /d/ and /t/, formed by closure of the tract at the gum ridge behind the upper teeth, and termed "alveolar". At the same time, /b/ and /d/ share the phonetic feature "voiced" and contrast with "voiceless" /p/ and /t/. Taken together these four phonemes constitute a little system of feature similarities and oppositions, such that /b/:/d/ = /p/:/t/ and /b/:/p/ = /d/:/t/. Phonemic, or feature, oppositions are presumed to be reflected in the signal by acoustic contrast.

All known spoken languages display phonological structure; that is, they constrain the arrangements in which phonemes may be combined to form words. These constraints reflect and, in part, define the feature classifications within the system: the domain of a phonotactic rule is a feature class. For example, in English, a stop consonant following initial /s/ is always voiceless (as in "spy", "sty" or "sky"); or, initial voiced stop consonants cannot be followed by nasal consonants (/bn-/ or /dn-/, for example, is not permitted). Phonological constraints of this kind are reflected in the signal by the types of sequential acoustic contrast permitted in a language.

I have rehearsed these generalities in order to raise the question of form and function. The particular form taken by the phonology of a language results from complex historical and social forces, as well as from phonetic, syntactic and semantic forces within the language itself. The general form taken by the phonologies of all languages, that is, the phonetics of language, is constrained by human anatomy and physiology: the elements must be drawn from, though probably do not exhaust, the (uncharted) sampling space bounded by what we can articulate and what we can perceive.

Several likely preconditions of linguistic communication suggest plausible functions for the "lower" level of the dual pattern: (1) learnability: a limited set of feature oppositions and a limited set of permissible sequences must facilitate language acquisition and retention; (2) lexical productivity: a system of minimal sound opposition between meaningful segments increases potential semantic range and flexibility; (3) memorability: both long-term and short-term memory must be facilitated. The ability to store meaningless phonemic sequences, pending syntactic and semantic processing, as we listen and perhaps as we speak, may be a condition of complex syntactic processing. The form of this store is a matter of considerable interest and has been the target of many short-term memory studies (see [15] for a review).

All that I have said so far has been intended to draw attention to the crucial role that segmentation, whether by feature or by phoneme, plays in language and speech, and to set the stage for discussion of a structure common to all languages, the consonant-vowel syllable.

A UNIVERSAL OF PHONETIC STRUCTURE: THE CONSONANT-VOWEL SYLLABLE

All spoken languages are syllabic. All languages constrain syllable structure in terms of consonants and vowels. All languages permit the consonant-vowel (CV) syllable. For all languages this is the canonical speech gesture.

Articulatory Function of the Syllable

The reason is not hard to find. Nothing is easier than to open the mouth. The CV syllable is fundamentally an articulatory unit. The consonant-vowel feature opposition is between a constricted and an open vocal tract, the two shapes being combined in a single "ballistic gesture" [12, p. 4]. The articulatory function of the syllable is probably as a unit of neural timing in the control of speech [5,9]. Control of the speech musculature - of the muscles for breathing,

phonating and articulating - is obviously a very complicated affair, requiring precise coordination of its parts. The time window over which coordination is accomplished might, in principle, be fixed, with new specifications appearing automatically as earlier specifications are implemented. The process would then be continuous and unsegmented, with new specifications adjustable to earlier, but not to later, commands. However, studies of speech production have shown that both anticipatory and perseverative coarticulation occur [6]. For this to be so, two or more phonetic segments must be programmed simultaneously. The decision as to how many units will be programmed at a time may then be determined at the lowest level by the composition of the syllable (which in many languages may include several consonants). Although no detailed model of the articulatory process has been worked out, this view of the syllable is not incompatible with the fact that coarticulation occurs across syllable boundaries, since this may be presumed to arise at a higher level of temporal coordination. That several levels of coordination may be required is suggested by the precision with which speech rhythm can be controlled over a lengthy utterance, despite considerable internal variation in segment duration.

Perceptual Function of the Syllable

The acoustic consequences of syllabic gestures can be observed in the undulations of an oscillogram. The perceptual function of these variations is to carry contrasts in stress, rhythm and intonation. None of these contrasts would be possible without the vowel, their carrier. Yet a language that consisted entirely of vowels would soon exceed the acceptable limits of homonymity. The addition of various forms of vowel onset or "attack" (that is, of consonants) to the phonetic repertoire provides the acoustic ground of perceptual contrast and releases the potential of dual patterning.

A second consequence of the syllabic gesture is evinced in the standard spectrogram by the apparent absence of acoustic

segmentation within many stretches of the signal known to correlate with syllables. The effect of this compression is to increase the rate of signaling phonetic segments (or phonemes). Precise limits on signaling rate are not known. But, on the perceptual side, an upper limit must be set by the ability of the ear to resolve a succession of more or less discrete segments, and a lower limit must be set by our short-term memory for unparsed stretches of speech. If the lower limit exceeds the upper, a possible solution is to merge units into larger units and so reduce the segment rate. This is exactly what is accomplished by the syllable.

However, syllabic compression has the apparent disadvantage of destroying the acoustic basis for recovering the phonetic segments on which dual patterning is premised. The difficulty is exacerbated by the spread across an entire syllable of acoustic properties presumed to characterize only one of the component phonetic segments. This effect, although often treated separately as a manifestation of "the invariance problem", is equally puzzling for an account of segmentation: to segment a syllable we must not only separate the pieces that do not belong together, but we must also connect those that do. Before examining this problem in more detail, let us consider how syllabic functions are fulfilled in a language that uses another modality.

SOME COMPARISONS WITH AMERICAN SIGN LANGUAGE

In recent years several laboratories in the United States have been intensively investigating American Sign Language (ASL) [7,8]. ASL is the language of many thousands of deaf persons in the United States. Its medium is manual (primarily), facial and bodily gesture. Here the arbitrary structure of ASL signs is of particular interest. Each sign is a meaningful unit, translatable into one or more English words, but signs are neither perceived nor remembered as "wholes". Intrusion errors in short-term memory reflect formational properties of the signs, analogous to the

phonological intrusions of reading and hearing studies [2]. Manual signs vary along the dimensions of shape, location, orientation and movement. Within each dimension there appears to be a limited set of oppositions. Although it is not yet clear precisely what these oppositions are, clustering and scaling analyses of errors in the perception of hand shapes in visual noise [10] have achieved at least the quantitative validity of comparable speech studies [11]. Furthermore, permissible feature combinations are a small set of the possible combinations, and rules for such standard formational ("phonological") processes as deletion and assimilation have already been described [1, 4]. It is probably a mere matter of time before a "phonology" of ASL is derived, and before work on the "universal phonetics" of sign language begins. What we have then in ASL (as, no doubt, in the sign languages of China, England, France and many other communities) is a system of communication that displays syntax, formational system and formational structure: in short, duality of patterning. How, we may now ask, are the syllabic functions of contrast and compression fulfilled in this language?

Sign uses a spatial rather than a temporal medium. Although signs may differ in movement, they are primarily distinguished by simultaneous, not sequential contrasts. They are linked sequentially to form "utterances", so that coarticulation over time does occur (for example, a hand may adopt the shape of a following sign before the preceding sign has been completed). But temporal coarticulation is not intrinsic to sign as it is to speech, because its units are units of spatial rather than temporal coordination. The functions of contrast and compression are therefore fulfilled simultaneously. The analogue of the syllable is the sign itself.

We may point up the analogy, and the difference, by noting that the feature oppositions of both speech and sign are articulatory, and that fine motor control is typically vested

in the same cerebral hemisphere for both dominant hand and mouth. (One hand is designated the "more active" in ASL formational rules [1], but which one depends on the handedness of the signer.) We are then led to wonder whether speech, viewed through transparent skin rather than heard as the acoustic consequences of its articulation, could be "read" as directly as sign. Probably not: for it is precisely in the temporally organized gestures of speech that the phonetic segments are lost. The segmentation problem may therefore be peculiar to speech.

THE ROLE OF ACOUSTIC FEATURE DETECTORS

One approach to the problem is implicit in the work of Stevens [13, 14] who has argued "... that there is some justification on a purely physical basis for characterization of phonemes in terms of discrete properties or features." [13, p. 53]. He shows that there are configurations of the vocal tract for which relatively large changes in articulation lead to relatively small changes in acoustic output. These regions are bounded by others for which precisely the reverse relation holds: small changes in articulation lead to large changes in acoustic output. He illustrates this principle with instances of articulatory-acoustic distinctions among important distinctive feature classes used in many languages, and suggests that the phonetic inventory of all languages is assembled from these "quantal" regions.

At first sight, one might take Stevens to be offering a straight-forward description of articulatory-acoustic parameters that could be used for speech synthesis or automatic speech recognition. However, he also states that a requirement for selection of quantal regions for phonetic use is "... that the attributes of the signal be relatively insensitive to articulatory perturbations *after the signal is transformed by the auditory mechanism*" ([13, p. 64, italics in the original]). In other words, not only must the acoustic signal be relatively insensitive to articulatory perturbation,

but the auditory system must be relatively insensitive to acoustic perturbation: it must perceive categorically. To meet this requirement Stevens, in a later paper [14], posits the existence of "auditory property detectors", tuned (or tunable) to the acoustic properties of speech. One explicitly stated reason for positing these detectors is that they provide a mechanism by which the infant might latch onto the phonetically relevant properties of speech. A second reason, not mentioned by Stevens but, if I understand him, essential to his account is that they might segment the flow of speech. For it matters little that speech is "quantal", if we have no device for sifting the quantal properties from the flow. In short, far from discovering the message in the signal, as it were, Stevens is offering an explicitly physiological account, to which the existence of discrete property detectors is essential.

A great deal of research in the past few years has been directed toward isolating such detectors. Eimas [3] introduced an "adaptation" procedure, modified from visual research. He and his colleagues showed that repeated exposure to a stimulus possessing a particular acoustic feature (for example, the rising formant transitions of some labial stop consonants) reduced a subject's sensitivity to that feature and relatively increased his sensitivity to an opponent feature (the falling transitions of some apical stop consonants). The procedure has been effectively used with many types of speech stimuli, contrasting in minimal acoustic features known to be associated with phonetic feature oppositions. The results have generally been interpreted as evidence for detectors tuned to the manipulated feature.

Let us suppose that this interpretation is correct and that banks of acoustic feature detectors, or analyzers, are neatly sprung by the syllabic flow. What then has been gained? First, a degree of segmentation. Second, a recoding of the signal from continuous to discrete form, so as to allow

short-term storage of information without "echoic" decay. These are important gains, but they do not move us very far toward a phonetic interpretation of the signal. One reason for this is that they can, at best, perform one half of the segmentation: they may separate what must be separated, but they cannot connect what must be connected.

Consider, for example, how feature detectors would analyze the existential injunction, "Be!" (/bi/). The following features might be detected: (i) silence, (ii) a rapid upward shift of the spectrum, voiceless for, say, 10 msec., voiced for 30 msec., in the vicinity of the second and third formants, (iii) a brief delay (10 msec.) between the onset of the spectral shift and the onset of glottal pulsation, (iv) a rapid (30 msec.), small upward shift of the first formant at the onset of glottal pulsation, (v) a relatively sustained formant pattern. The phonetician knows that the first four features are typical of voiced labial stops before high front vowels. But the cryptographer (that is to say, the auditory system) does not. What auditory principle groups the first feature, silence, with the next three, but fails to group the fourth with the fifth? In other words, what auditory principle integrates the acoustic features of the consonant and separates them from those of the vowel?

If we must rely on feature analyzing systems, there seems, in fact, to be none. The proposed detectors thus lead us into an impasse from which we can only escape by invoking some non-auditory principle of perceptual organization - precisely the impasse they were intended to avoid.

PERCEPTUAL CONTRAST AND CONTINUOUS TRACKING

The source of the difficulty is the desire to match our percepts with both phonological description and the acoustic signal. But perhaps we have been misled in attempting to model perceptual performance after the linguist's model of phonological competence. We cannot evade the dual pattern.

But must the perceptual segments be static? Why, if speech is acoustic and if the essence of an acoustic event is its temporal organization, are features and phonemes commonly defined as points in space, static configurations of the vocal tract, or as stationary auditory qualities?

If we return to the canonical speech gesture, two facts stand out. First, the articulatory poles of the syllable - constricted vs. open - provide maximal perceptual contrast. Second, the contrast is always and only manifested over time. The syllable is a unitary event of which the auditory quality (or phonetic manner) changes as it occurs. If we perceived the contrast directly, as a development, much as we perceive the attack and sustention of a musical note, without benefit of specialized detectors to "stop the image", the contrast would be the ground of our perceptual segmentation.

Our percepts would not then be segments, but acoustic events for which we happen to have a segmental notation (arrayed in space). From this point of view, the perceptual process is a continuous tracking of an acoustic signal, isomorphic, point for point, with the continuously changing articulation. The perceptual elements of the dual pattern would not then be the timeless entities of current phonology, but dynamic events, jointly shaped by the timing mechanisms of motor control and by the demands of the auditory system for perceptual contrast and compression.

REFERENCES

[1] Battison, R. 1974. Phonological deletion in American Sign Language. Sign Language Studies 5: 1-19.

[2] Bellugi, U.; Klima, E.S.; and Siple, P.A. 1975. Remembering in signs. Cognition 3: 93-125.

[3] Eimas, P.D., and Corbit, J.D. 1973. Selective adaptation and linguistic feature detectors. Cog. Psych. 4: 99-109.

[4] Frishberg, N. 1975. Arbitrariness and iconicity: historical change in American Sign Language. Language 51: 696-719.

[5] Fry, D.B. 1964. The function of the syllable. Zeitschrift für Phonetik, Sprachwissenschaft u. Kommunikationforschung 17: 215-221.

[6] Harris, K.S. In press. The study of articulatory organization: Some negative progress. In Research on Dynamics of Speech Production, Annual Bulletin of the Research Institute of Logopedics and Phoniatrics. Tokyo, Japan.

[7] Hockett, C.F. 1958. A course in modern linguistics. New York: MacMillan.

[8] Klima, E.S., and Bellugi, U. (eds.) In press. The Signs of Language. Cambridge: Harvard University Press.

[9] Kozhevnikov, V.A., and Chistovich, L.A. Rech' Artikuliatsia i vospriiatie. Moscow, Leningrad. Translated as Speech: Articulation and Perception. Washington: Clearinghouse for Federal Scientific and Technical Information, J.P.R.S., 1965: 30.

[10] Lane, H.; Boyes-Graem, P.; and Bellugi, U. 1976. Preliminaries to a distinctive feature analysis of hand shapes in American Sign Language. Cog. Psych. 8: 263-289.

[11] Miller, G.A., and Nicely, P.E. 1955. An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Amer. 27: 338-352.

[12] Stetson, R.H. 1951. Motor phonetics. Amsterdam: North-Holland.

[13] Stevens, K.N. 1972. The quantal nature of speech: evidence from articulatory-acoustic data. In David, E.E. and Denes, P.B. Human communication: a unified view. New York: McGraw Hill, 51-66.

[14] Stevens, K.N. 1975. The potential role of property detectors in the perception of consonants. In Fant, C. G.M. and Tatham, M.A.A. (eds.) Auditory analysis and perception of speech. New York: Academic Press, 303-330.

[15] Studdert-Kennedy, M. Speech Perception. In Lass, N.J. (ed.) Contemporary Issues in Experimental Phonetics. New York: Academic Press, 1976, 243-293.

Biology, Behavioral Sciences, Linguistics, Communication, Computer Science, Physics, Neurology, Physiology

Life Sciences Research Reports (ISSN 0340–8132)
Vol. 5

Recognition of Complex Acoustic Signals

Report of the Dahlem Workshop on Recognition of Complex Acoustic Signals Berlin 1976, September 27 to October 2

Editor: **Theodore H. Bullock**, University of California, San Diego

Acoustic Signals may reach and influence many individuals simultaneously and over a wide area. The mechanisms of these signals and their nature have come to be the center of worldwide research. The sharp increase in findings and claims pertaining to decipherment of animal vocalizations and human speech, and processing by the central nervous system and by machine systems makes the need for a review of the present state of knowledge evident.
This report is the result of combined efforts by physicists, communication engineers, information scientists, linguists, behavioral biologists, neurologists, and physiologists to provide such a review.

CONTENTS

April 1977; 406 pages. $ 25.00 / £ 14.70 / DM 64,00
Berlin: Dahlem Konferenzen. ISBN 3-8200-1206-0

Dahlem Konferenzen

CONTENTS continued:

Background Papers *by E. F. Evans, M. Konishi, A. M. Liberman and D. B. Pisoni, P. R. Marler, J. D. Miller, W. D. Neff, G. Neuweiler, H. Scheich, M. R. Schroeder.*

Seminar Papers *by W. A. Ainsworth, C. J. Darwin, E. E. Douek, E. F. Evans, R. Plomp, A. Risberg, P. A. Tallal.*

Disorder of Hearing und Language. Seminar Report, *E. F. Evans, Rapporteur.*

Glossary – Subject Index – Author Index

ORDER FORM

Please send to your regular bookseller
or to Dahlem Konferenzen, Delbrückstraße 4 C, D-1000 Berlin 33,
or to Heyden & Son, Spectrum House, London NW4 3XX
or to Koehn + Schneider, 50 Sound Beach Avenue,
Old Greenwich, Conn. 06870, U.S.A.

Please send

Life Sciences Research Reports (ISSN 0340–8132)

☐ Vol. 1: The Molecular Basis of Circadian Rhythms
J. W. Hastings and H.-G. Schweiger, eds.
May 1976; 464 pages. $ 27.50/£ 15.50 / DM 75.00.

☐ Vol. 2: Appetite and Food Intake
T. Silverstone, ed.
May 1976; 498 pages $ 26.50 / £ 15.00 / DM 72.00.

☐ Vol. 3: Hormone and Antihormone Action at the Target Cell
J. H. Clark, W. Klee, A. Levitzki, and J. Wolff, eds.
August 1976; 226 pages. $ 16.50 / FF 78.00 / DM 45,–.

☐ Vol. 4: Organization and Expression of Chromosomes
V. G. Allfrey, E. K. F. Bautz, B. J. McCarthy, R. T. Schimke, A. Tissiéres, eds.
November 1976; 349 pages. $ 24.00/£ 14.80/DM 62.00.

☐ Vol. 5: Recognition of Complex Acoustic Signals
Theodore H. Bullock, ed.
April 1977; 406 pages. $ 25.00 / £ 14.70 / DM 64.00.

☐ Vol. 6: Function and Formation of Neural Systems
G. S. Stent, ed.
September 1977 approx. 360 pages, approx. $ 24.00.

☐ Enter my continuation order for Life Sciences Res. Reports

Name _____

Address _____

_____

_____

☐ payment enclosed (no postage) ☐ bill me (plus postage)

Date/Signature _____