On levels of description in speech research

Bruno H. Repp

Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06510
(Received 3 November 1980; accepted for publication 17 February 1981)

Many researchers use linguistic category names (consonants, vowels, syllables) to refer to observations and measurements made in records of the acoustic speech signal. The present paper serves as a reminder that linguistic categories are abstract and have no physical properties, and that, therefore, their physical correlates in the speech wave are appropriately described in acoustic terms only.

PACS numbers: 43.70.Gr, 43.70.Ve, 43.70.Dn

Every branch of science needs a precise terminology to describe the phenomena it is investigating. If there are different levels of observation, different terms must be applied at each level in order to avoid confusion. For example, the psychologist must distinguish the perceptual category "red" from the neurophysiological processes that lead to the percept; and they in turn must be distinguished from the energy and wavelength of the light that impinges on the retina. If redness were a physical property of the light wave, it would be difficult to explain why, for example, a certain wavelength is called "red" by one viewer but "orange" by another and "gray" by a third (who happens to be color blind).

Scientists concerned with speech must be especially careful because there are at least six different levels of description, each requiring its own separate set of terms: articulation, acoustic waveform, neurophysiological processes, conscious percept, nonlinguistic auditory impressions, and abstract linguistic theory. Unfortunately, the mixing of terms from different levels is a common practice of speech scientists. In particular, perceptual–cognitive (phonetic, linguistic) categories are often applied to acoustic observations. It is the purpose of the present paper to discourage this usage, as far as possible.

Terms such as "vowel duration," "fricative amplitude," "syllable onset," "/p/ duration," etc. abound in the literature. The measurements referred to by these terms are made on spectrograms or oscillograms, i.e., on graphic records of an acoustic waveform. Thus they concern (the visual correlates of) acoustic segments, such as periods of periodicity, noise, or silence. Why do so many researchers use linguistic categories (vowels, consonants, syllables) to describe these acoustic segments? Is it just carelessness, or does it reflect some incorrect assumptions about the nature of phonetic segments?

One possibility is that underlying this usage of terms is a theory of speech segmentation that considers linguistic categories as a classification system for acoustic segments that are arranged like beads on a string. This view was widely held until the advent of the sound spectrograph; however, it has long been proven to be false. There is no one-to-one (or even many-to-one) correspondence between acoustic and linguistic segments; rather, the acoustic information for successive linguistic units overlaps and interacts. This fact has been referred to as "encodedness" or "parallel transmission of information" (Liberman *et al.*, 1967). It is a consequence of the complex dynamics of articulation. Although the input to the articulatory system may consist of a sequentially arranged string of abstract linguistic units (this is a hypothesis, not a fact), the articulatory movements corresponding to these units are no longer strictly sequential, and they are subject to passive as well as planned contextual variation. While discontinuities in the acoustic output may directly reflect changes in the state of the articulators and of the larynx, it is a serious mistake to consider them as boundaries of linguistic segments (cf. Fant, 1962).

Since these facts are by now generally accepted, it seems unlikely that any serious researcher would still espouse a naive beads-on-a-string theory. However, it is important to keep in mind that this conception remains the natural choice of anyone who reflects upon the structure of speech without ever having inspected a record of its acoustic waveform. Lax use of terms by professional scientists encourages such misconceptions and impedes the task of getting the facts across to students and the interested public.

Being aware of these facts, many speech scientists nevertheless use linguistic terms (consonants, vowels, syllables) as if they were acoustic categories—a classification of speech sounds. Perhaps, this malpractice originated with the time-honored but quite misleading term, *speech sounds*. For, patently, we do not normally perceive a sequence of *sounds* when we listen to speech but a linguistic message in which phonetic segments are the smallest units. These units are *abstractions*. They are the end result of complex perceptual and cognitive processes in the listener's brain, and it is likely that, excluding certain laboratory tasks, they are in fact not perceptual primitives but are derived by cognitive analysis from larger units, such as syllables or words (cf. Foss and Blank, 1980). Moreover, it appears that their conscious perception presupposes familiarity with an alphabetic writing system (Morais *et al.*, 1979). That is, except for the rare preliterate individual who arrives at some rough approximation through intense reflection upon the nature of speech and language (witness the uniqueness of the invention of the alphabet!), awareness of the linguistic segment inventory generally derives from the experience of learning to read and write alphabetically (Lüdtke, 1969), and thus is heavily influenced by the spelling system of a language. Linguistic segments are important concepts

 0001-4966/81/051462-03$00.80

for describing and explaining language structure. However, whether units corresponding to these abstract categories play any role at a subconscious level in ongoing speech perception is an open question; certainly, they could not do so *as* abstract categories which are, by definition, post-perceptual. It seems likely that the structures utilized by the perceptual system require an entirely different (and novel) set of descriptors.

Abstract linguistic segments (the traditional "speech sounds") must be distinguished from the actual *sounds* of speech. These sounds can be described only in *auditory* terms, such as "hiss," "buzz," "silence," etc. Our vocabulary to describe these auditory impressions is rather limited (see, however, Pilch, 1979, for an attempt to organize and enrich it). These auditory qualities of the speech wave usually go unnoticed because the listener's attention is focused on the linguistic message. Considerable attention and experience are required to gain access to the auditory properties of speech, particularly to those aspects that support phonetic perception (as contrasted with suprasegmental characteristics such as intonation or voice quality that are more readily brought into awareness). Psychologists have been interested in this fact, as shown by the numerous studies of "categorical perception" which assess the (in)ability of listeners to discriminate speech stimuli on an auditory basis.

Acoustic aspects of the speech waveform do have a rather close relation to the *auditory* qualities perceived by a careful listener, but the relationship between acoustic segments and *phonetic* percepts (i.e., linguistic categories) is more complex. In general, several acoustic segments are relevant to the perception of a single phonetic segment, and each individual acoustic segment typically contains information about more than one phonetic segment. A phonetic category is not just a label attached to a particular combination of acoustic segments; for example, stop consonants in initial, medial, and final position have quite different acoustic correlates. Nor is it a label attached to the particular auditory qualities of the relevant acoustic segments, singly or in combination. Nor is it, strictly speaking, a classification of articulatory maneuvers or positions. Rather, a phonetic category is a perceptual–cognitive state resulting from the integration of diverse acoustic information into a unitary percept according to principles that are specific to phonetic perception and are best explained by reference to the articulatory origin of the speech signal. Alternatively, and perhaps more commonly, awareness of phonetic segments follows lexical access and thus results from cognitive analysis following primary perception (cf. Foss and Blank, 1980). That is to say that special perceptual and cognitive processes *intervene* between the acoustic signal and the phonetic percept. Therefore phonetic categories—consonants, vowels, and even syllables—cannot be said to be in the acoustic signal. *They have no physical properties*—such as duration, spectrum, and amplitude—and, therefore, *cannot be measured.* (The properties they do have, such as distinctive features, are equally abstract; see Parker, 1977, for an excellent discussion of this issue.) The acoustic signal only contains the *information* that supports their perception; this information *can* be described (e.g., in terms of acoustic segments or "cues") and measured along acoustic dimensions.

Some might want to argue that vowels and consonants *are* in the signal but in a shingled, interwoven fashion. In other words, a phonetic segment could be defined as the totality of all acoustic cues that support its perception. Such an operational definition, while reasonably unambiguous, still commits a category error because it ignores the perceptual and cognitive processes that intervene between acoustic cues and phonetic percept. For example, if one (e.g., in a study of the "phoneme restoration effect"—Warren, 1970; Samuel, in press) "removes a consonant" from an utterance by gating out certain portions of a speech signal, what is eliminated is the *information* that supports perception of the consonant. To state that the *consonant* has been removed from the waveform would not be proper; indeed, it might be misleading because it suggests (incorrectly) that *only* information pertaining to the consonant has been removed.

It would be unrealistic to demand that terms such as "vowel duration" and "fricative amplitude" be banned forever. However, I would like to urge researchers (1) to avoid them whenever possible, and (2) if they are to be used, to define precisely in *acoustic* terms what they are intended to refer to. It is by no means true that a seemingly innocuous term such as "vowel duration" has a generally agreed-upon interpretation in every context (see Lisker, 1974). Only if a vowel occurs in isolation is there no ambiguity. In the utterance /ba/, on the other hand, does vowel duration include the initial formant transitions which support the perception of the stop consonant? In /pa/, does it include the period of aspiration following the labial release? (If vowel duration is treated as a perceptual, not acoustic, quantity, these become legitimate empirical questions—cf. Raphael *et al.*, 1980.) In most cases, only terms such as "periodicity," "aspiration noise," "release burst," and "formant transitions" (including a suitable criterion for their beginning or end) permit an unambiguous specification of what is being measured. Once such a specification is provided by an author, and only then, the term "vowel duration" may be acceptable for the sake of convenience, although "duration of periodicity" (or whatever acoustic term is appropriate in a given context) would be preferable.

There are differences in the *degree* to which various misapplications of linguistic terms are inappropriate. This degree roughly parallels the dimension of "encodedness." For example, "fricative duration" will in most cases be unambiguously understood as referring to the duration of the noise (frication) portion of a stimulus, although the formant transitions in the surrounding acoustic segments contribute to the fricative percept (Harris, 1958; Whalen, 1981) and thus are part of the set of relevant cues. However, the noise is *not* "the fricative," and to call it so is awkward, at the least. Much more confusion is created by a term such as "stop consonant duration." While, in medial posi-

tion, many will understand the term to refer to the period of relative silence resulting from oral closure (even though this is only one of several relevant acoustic cues), in utterance-initial position it might refer to the release burst alone, or the burst plus aspiration, or the burst plus aspiration plus formant transitions; in utterance-final position, it might refer to the formant transitions only (if the stop is unreleased) or to the period of silence with or without the release burst and/or the transitions (if the stop is released); and in an utterance such as /ækt/, with the first stop unreleased, it is not clear at all where the first stop ends and the second stop begins. Therefore this term should not be used at all, not even after describing exactly what is being measured; instead, specific acoustic terms should be used throughout.

This request is not nearly as radical as it may seem. Definition of acoustic segments in purely physical terms *can* be cumbersome, e.g., "the periodic portion following the fricative noise." It is quite legitimate, therefore, to name the linguistic segment for which a given acoustic segment is the primary cue, *as long as the main term is physical in nature*, e.g., "the 'u' periodic portion," "the 'p' silence," or "the 's' noise." Consistent use of such a terminology should place only a minor burden on researchers accustomed to speak loosely of "/p/ duration" or "/s/ amplitude;" however, it would greatly increase the clarity of many research reports.

Clearly, many of these arguments have been presented before (see especially Fant, 1962; Lisker, 1957, 1974; Parker, 1977; Pilch, 1974; Zwirner and Zwirner, 1970). However, they seem to have had little impact and, therefore, are worth repeating. Examples of terminological carelessness still abound in the literature. To quote just one recent example from an otherwise excellent paper: Mills (1980) states, referring to utterance-initial consonants (and without further qualification), that "... /s/ has a lower amplitude than /b/" and "... /s/ is longer in duration than /b/" (p. 82). Similarly awkward or outright misleading statements can also be found in the pages of this *Journal* (see, e.g., Umeda, 1977). Although there are, of course, many authors who take great care to avoid such terminological confusion, I suspect that they are not in the majority. I hope the present note will draw attention to this problem and contribute to its gradual elimination.

ACKNOWLEDGMENTS

Preparation of this paper was supported by NICHD Grant HD01994 and BRS Grant RR05596 to the Haskins Laboratories. Valuable comments and criticisms were contributed by Carol Fowler, Katherine Harris, Alvin Liberman, Leigh Lisker, Virginia Mann, Ignatius Mattingly, Frank Parker, and Michael Studdert-Kennedy.

Fant, G. (1962). "Descriptive analysis of the acoustic aspects of speech," Logos **5**, 3–17.

Foss, D. J., and Blank, M. A. (1980). "Identifying the speech codes," Cognit. Psychol. **12**, 1–31.

Harris, K. S. (1958). "Cues for the discrimination of American English fricatives in spoken syllables," Lang. Speech **1**, 1–7.

Liberman, A. M., Cooper, F. S., Shankweiler, D., and Studdert-Kennedy, M. (1967). "Perception of the speech code," Psychol. Rev. **74**, 431–461.

Lisker, L. (1957). "Linguistic segments, acoustic segments, and synthetic speech," Language **33**, 370–374.

Lisker, L. (1974). "On time and timing in speech," in *Current Trends in Linguistics*, edited by T. A. Sebeok (The Hague, Mouton), pp. 2387–2418.

Lüdtke, H. (1969). "Die Alphabetschrift and das Problem der Lautsegmentierung," Phonetika **20**, 147–176.

Mills, C. B. (1980). "Effects of context on reaction time to phonemes," J. Verb. Learn. Verb. Behav. **19**, 75–83.

Morais, J., Cary, L., Alegria, J., and Bertelson, P. (1979). "Does awareness of speech as a sequence of phones arise spontaneously?" Cognition **7**, 323–331.

Parker, F. (1977). "Distinctive features and acoustic cues," J. Acoust. Soc. Am. **62**, 1051–1054.

Pilch, H. (1974). *Phonemtheorie* (Basel, Karger), 3rd ed.

Pilch, H. (1979). "Auditory phonetics," Word **29**, 148–160.

Raphael, L. J., Dorman, M. F., and Liberman, A. M. (1980). "On defining the vowel duration that cues voicing in final position," Lang. Speech **23**, 297–307.

Samuel, A. G. (in press). "Phonemic restoration: Insights from a new methodology," J. Exp. Psychol. (G).

Umeda, N. (1977). "Consonant duration in American English," J. Acoust. Soc. Am. **61**, 846–858.

Warren, R. M. (1970). "Perceptual restoration of missing speech sounds," Science **167**, 392–393.

Whalen, D. H. (1981). "Effects of vocalic formant transitions and vowel quality on the English [s]-[š] boundary," J. Acoust. Soc. Am. **69**, 275–282.

Zwirner, E., and Zwirner, K. (1970). *Principles of Phonometrics* (University of Alabama, University, AL).