The pursuit of invariance in speech signals

Leigh Lisker
Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06511-6695 and Department of Linguistics, University of Pennsylvania, Philadelphia, Pennsylvania 19104

(Received 24 February 1984; accepted for publication 30 October 1984)

The search for the acoustic properties useful to the listener in extracting the linguistic message from a speech signal is often construed as the task of matching invariant physical properties to invariant phonological percepts; the discovery of the former will explain the latter. These phonological percepts are essentially the phonemes of pregenerative phonology, and they are more or less faithfully reflected in standard alphabetic writing. Thus English *deep* and *doom* are supposed to be perceptually identical in their initial /d/s; the orthographic similarity is in agreement with the linguist's "representation" of these forms. The partial identity in spelling is only weak evidence for perceptual invariance, however. First, while some phonemes may comprise a single "sound," others are said by linguists to include phonetically distinct ones. Thus English /p/ includes both aspirated and unaspirated voiceless labial stops. The view that it is not the phoneme, but rather the phonetic feature, to which an acoustic invariant might be attributed, raises two questions: (a) Since segments sharing a feature are rarely judged to constitute a single sound, the search for a feature-specific invariant, whose function is to explain perceptual constancy, is deprived of its essential motivation, and (2) there is no more reason to expect the acoustic cues to a feature to be context-independent than is the case with the phoneme. What seems more likely is to find that some phonemes, and some features, are more invariantly marked in the speech signal than others.

PACS numbers: 43.70.Fq, 43.70.Hs, 43.71.An, 43.72.Ne

The auditory analysis of speech into sequences of elementary speech sounds long antedates the development of our present methods for the instrumental recording and analysis of acoustic signals. The alphabetic registration of speech, and, in particular, its phonetic and phonological spellings by linguists, embody a once generally accepted model for signals produced and perceived in the speech communication process: Speech is articulated, i.e., jointed, so that a sequence of discrete vocal tract shapes gives rise to a sequence of similarly discrete sounds, which, in turn, is interpreted as some specific linguistic message. In some part, this view still prevails. Speech is now regarded as being both articulated and fluent, and we continue to look for acoustic properties by which each category of phonetic segments, or the phonological unit to which it is assigned, may be characterized. We persist, moreover, in thinking of these sought-after properties as attributes of discrete and acoustically delimitable intervals to which the names of our phonetic/phonological categories are directly applicable, thereby conflating the rather different units designated by the terms "phonetic segment" (or "speech sound") and "acoustic segment" (see, e.g., Repp, 1981).

Surveys of the modern literature addressing the invariance question (e.g., Cooper, 1980); Darwin, 1976; Liberman and Studdert–Kennedy, 1978; Wickelgren, 1976) suggest that neither the definition of invariance nor the type of linguistic unit to be specified by physical invariants has held constant. Invariance has been posited, sometimes to be dismissed, but sometimes perhaps demonstrated with convincing plausibility, at several levels of abstraction—as a temporal interval having a "typical" waveform (Fletcher, 1929), a particular spectral property (Stevens and Blumstein, 1978) or a given dynamic pattern (Kewley-Port, 1983), by a set of "target" format frequencies (Lindblom and Studdert–Kennedy, 1967), or by so-called "locus" frequencies (Delattre *et al.*, 1964). Moreover, there does not seem to be entire agreement as to either the size or level of abstractness of the linguistic elements for which invariant acoustic properties (given some definition of "invariant") are to be sought: Should they be phonetic features, segments, demisyllables or syllables? For any one of these entities, at what level of abstractness should they be construed? Clearly, unless there is agreement on these matters, we cannot pose the problem of invariance so that it can be resolved. Even with such agreement it is by no means self-evident that a single answer will ever be forthcoming, one that is valid for all elements of the same size and level of abstractness.

In considering the invariance question, we must remember that the original motivation of the search for acoustic invariants was to explain why speech signals can be perceived as sequences of "sounds" drawn from a limited inventory of such elements, whose freedom to occur in a virtually unlimited number of combinations makes human speech and language possible. The perceptual invariance that presumably characterizes each sound type is of a special kind—it is not auditory invariance, but only invariance with respect to those auditory properties that have what we might call potential linguistic significance, or perhaps phonetic significance. In short, the members of a sound type share the property of phonetic invariance, and one way of construing the invariance problem is to specify it as a task of determining what acoustic invariants, if any, can be associated with each of the elements for which phonetic invariance is posited. In recent years, however, emphasis has been shifted from

 0001-4966/85/031199-04$00.80

the segment to the phonetic feature as the linguistic element to be paired with an acoustic invariant. This shift, although it faithfully reflects the practice of current phonological analysis, has at least one serious drawback—namely, that, even if a feature can be associated with an acoustically invariant property, the feature is a component of a phonetic segment (which is not abolished), and segments sharing this feature do not constitute a perceptually invariant set unless they are identical in respect to *all* their constituent features. But the "bundle" of all these features *is* the segment. Thus the smallest size unit for which (phonetic) perceptual invariance can be claimed is not the feature, but the segment, and the most abstract category level of this size and perceptual status is the phoneme of pregenerative phonology.

In the discussion of a possibly invariant relation between phonetic and acoustic properties, we must bear in mind that the first question for the linguist is not one of evaluating the similarity relations among segments, but of deciding, with respect to the speech events observed in a language community, which of them, taken pairwise, are perceived by community members to be repetitions of each other, and which are not. If their behavior leads the linguist to suppose that two events are functionally the same, then the linguist may decide that they are phonologically identical, i.e., composed of the same segments in the same order. But if two events are judged to be functionally and perceptually different for the language community, then the linguist cannot on the same basis decide whether they are *in part* the same for speakers of the language. Because there can be no experimental verification of the perceptual identity or nonidentity of two phonetic segments in different contexts that is nearly as direct as can be applied in deciding the relation between speech events, the establishment of a collection of segments abstracted from different events as a phonetic or phonological category rests on auditory and linguistic judgments by the linguist, judgments that include hypotheses about the native speaker's perceptions of the segments. Thus the linguist can readily decide by test that the English forms *deep* and *doom* are phonetically distinct, but not whether, for the native speaker, they are identical in their initial consonants and different in their vowels and final consonants.

It might be supposed that the similarity in the linguist's spellings of *deep* and *doom* reflects a perceptual invariant for which an acoustic invariant awaits discovery. A partial identity in spelling, however, is a doubtful basis for anticipating acoustic invariance, for we might suppose the asserted identity of the two words to be as much dependent on the difference in their contexts (on the analogy of a modified Mueller–Lyer Illusion) as on the presence of a common acoustic property. The words *calf* and *cough* are also alike in the phonological spelling of their initial consonants and different in their vowels, i.e., / kæ f/ and /k ɔ f/. A speaker of Arabic, however, might dispute this way of representing the nature of the contrast, equating *calf* with Arabic كاف and *cough* with قاف , and claiming that the difference resides ("contrastively") in the initial consonants and not in the vowels. The observing linguist, equally conversant in or perhaps equally ignorant of both languages, would say that, in the two word pairs, the phonetic differences involve both the consonants and the vowels. Thus the speech researcher, in quest of acoustic invariants matching the phonological units represented in spelling, whether standard orthographic or phonemic, could define the task variously, depending on whether he wanted to account acoustically for the phonologically defensible spelling behavior of the English speaker, the Arabic speaker, or the linguist. The latter would not only be of the opinion that the words in both languages differ in the initial consonants and in the vowels, but that English *cough* and Arabic قاف are far from being the same in their initial consonants. From all this, then, we are entitled to believe that the degree of invariance by which the onsets of *deep* and *doom* are connected is not the same as that linking the two initial consonants of *calf* and *cough*. [We may recall from these examples the findings of Liberman *et al.* (1952) and Schatz (1954), that indicate that English /d,t/ are more nearly invariant in their burst than either /b,p/ or /g,k/.]

Additional examples from English can be cited that do not encourage us to expect to find invariant acoustic properties marking the phonological categories commonly recognized. The ability of listeners to distinguish the words *beeper* and *peeper* is ascribed entirely to the /b/–/p/ contrast, /b/ being characterized usually as [+voice] and /p/ as [−voice]. The medial /p/ of both words is, of course [−voice]. But, while it is no doubt correct to say that initial /b/ is more voiced than initial /p/, it is not so clear that it is regularly more voiced than medial /p/. Thus in a phrase *this beeper* the two labial stop consonants need differ not at all in degree of voicing, certainly never as much as do the stops in *this peeper*. Moreover, a pair of expressions, *this beaker* and *the speaker*, if they are said to include a /b/ and a /p/, respectively, can certainly not be distinctively marked by invariant acoustic properties associated with the stop voicing contrast.

The notorious *writer–rider* pair of many varieties of American English is another case that poses a problem. If the phonemes /t/ and /d/ are to be associated with invariants marking, respectively, the word sets *tear toll heat rote* and *dear dole heed road*, then the inclusion of *writer* in the first set and *rider* in the second must be at the cost of any claim that /t/ and /d/ are distinctively and invariantly marked. (Since some British English speakers use a voiceless aspirated stop in *writer* we must accept as fact that in American English the /t/–/d/ contrast, if it operates to separate *writer* and *rider*, is marked in a less than maximally invariant fashion.) When I asked linguistically untrained speakers their opinion as to the basis on which they distinguished the two words, I failed to elicit answers consistent enough to justify a conclusion that (1) the first vowels are different perceptually and the medial consonants are identical, or (2) the vowels are the same and the consonants distinct, or (3) both vowels and following consonants are perceived as different. Under this kind of questioning, moreover, those listeners who first opted strongly for some one view soon enough showed all the uncertainty that experienced linguists have expressed over the many years that this troublesome pair of words has been a subject of dispute (see, e.g., Fischer–Jørgensen, 1975; Hymes and Fought, 1975).

The *writer–rider* example might be faulted as irrelevant

to the present discussion precisely on the ground that listeners do not agree on what they hear as different when they distinguish auditorily between the two words. Absent such agreement, we may continue to posit an acoustic basis for connecting *writer* with *write* and *rider* with *ride*, but we need not assume that the identification of the flap in *writer* with /t/ and the one in *rider* with /d/ is based on segment-specific invariant properties. The phonemic encodings of *writer rider* as, e.g., /raytər/ /raydər/ are dictated by considerations that include no strong claim about the perceptual status of the alveolar flaps in those words. Hence, the motivation for seeking invariant properties connecting them "correctly" with /t/ and /d/ is weak, if not entirely lacking.

Another case involving the voicing contrast does have more relevance to the invariance question; this is the case of the post-/s/ stops in word–initial position in English. If we believe that the linguist's spelling of *spin* is evidence that the stop is perceived as a member of /p/, then we might describe the effect of replacing the /s/-noise with silence as one of shifting /p/ to /b/(see Lotz *et al.*, 1960). On the other hand, replacing the closure voicing in a token of the word *ruby* with silence of a certain (i.e., greater) duration will often cause listeners to report having *rupee* instead (Lisker, 1957a). Thus silence in one context is a "cue" to /b/, in another to /p/. There are, one would agree, other ways of describing this situation, but none will entirely explain away the problem it poses for a claim that the /p/–/b/ contrast is correlated with an acoustically invariant difference.

It may be appropriate to recall that the phonological literature was once alive with controversy as to whether the English stops are distinctively voiced and voiceless, with aspiration a redundant feature of some members of the voiceless category, or whether, instead, they are distinctively weak (*lenis*) and strong (*fortis*) in force of articulation, with voicing a redundant feature of the weakly articulated category (see, e.g., Jakobson and Waugh, 1979). If the voicing of /b,d,g/ is disposable in initial and some other positions, and if aspiration is positively unnatural except initially and preceding the stressed vowel of a word, then we may claim that the /b,d,g/–/p,t,k/ contrast is signaled only by "redundant" features. If such a claim is dismissed as simply too "radical" to be considered seriously, the claim that membership in the /b,d,g/ and /p,t,k/ sets is definable in terms of acoustic invariants seems to revive a notion that is widely thought to have been conclusively demolished by the generative phonologist—namely, the biuniqueness relation between phonetic sement and phonological category (Chomsky and Halle, 1968).

The case of stop voicing involves the relation between acoustic and linguistic/perceptual aspects of the speech signal. A similar relation between articulation and linguistic percept can also be suggested. The two events represented as /iwi/ and /uyu/ in English involve the glides /w/ and /y/, the first described as tongue backed and lip rounded, the second as tongue fronted and lip unrounded. It is possible, however, to produce a recognizable /iwi/ without moving the tongue from an /i/ position, and to produce an /uyu/ without moving the lips from a posture appropriate to /u/. The vocal-tract shapes to and from which the glides are articulated are the same for these perhaps unusual ways of producing /iwi/ and /uyu/; that configuration is the one used in pronouncing the French front rounded glide of the word *huit* [ɥit]. I confess that I have not been able to produce these sequences so that the two lowest formants show exactly the same frequencies at the midpoints of the glides, and my claim as to the articulations should be checked by x-ray monitoring. However, my claim is no more doubtful, I would submit, than many another description of articulation for which no evidence other than proprioceptive introspection by the linguist speaker is provided. There are, moreover, "harder" data from experiments in synthesis to show that the same set of formant frequencies in different vowel-like contexts will be reported as more than one member of the /w,r,l,y/ set, e.g., as *iri ala uyu* (Lisker, 1957b).

In conclusion, it can be said that the search for acoustic properties by which linguistic messages are signaled in speech should and will continue to be vigorously pursued, for this enterprise is, after all, a central one in phonetics. To the extent that invariant correlates of those linguistic units having the status of perceptually defined elements turn up, fine. In some cases these elements may well be the phonemes of pregenerative phonology. But these phonemes, which linguists and the rest of us recognize in our various spelling practices, are not all perceptual constants, and we must therefore be prepared to find that some phonemes are less invariantly marked than others. If the site of acoustic invariance is postulated to be the phonetic feature rather than the phoneme, then we must still reckon with the likelihood that some features, e.g., voicing, are acoustically less stable across contexts than others, e.g., nasality. In other words, we should be prepared to live with the finding that acoustic invariance is itself a variable.

ACKNOWLEDGMENT

Preparation of this paper was supported by NICHD Grant HD-01994 and BRS Grant RR-05596 to Haskins Laboratories.

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

Cooper, F. S. (1980). "Acoustics in human communication: Evolving ideas about the nature of speech," J. Acoust. Soc. Am. **68**, 18–21.

Darwin, C. J. (1976). "The perception of speech," in *Handbook of Perception, Vol. VII: Language and Speech*, edited by E. C. Carterette and M. P. Friedman (Academic, New York).

Delattre, P. C., Liberman, A. M., and Cooper, F. S. (1964). "Formant transitions and loci as acoustic correlates of place of articulation in American fricatives," Stud. Linguist. **18**, 104–121.

Fischer–Jørgensen, E. (1975). *Trends in Phonological Theory* (Akademisk Forlag, Copenhagen).

Fletcher H. (1929). *Speech and Hearing* (Van Nostrand, New York).

Hymes, D., and Fought, J. (1975). "American structuralism," in *Current Trends in Linguistics, Vol. 13: Historiography of Linguistics*, edited by T. A. Sebeok (Mouton, The Hague).

Jakobson, R., and Waugh, L. (1979). *The Sound Shape of Language* (Indiana University, Bloomington, IN).

Kewley-Port, D. (1983). "Time-varying features as correlates of place of articulation in stop consonants," J. Acoust. Soc. Am. **73**, 322–335.

Liberman, A. M., Delattre, P. C., and Cooper, F. S. (1952). "The role of

selected stimulus variables in the perception of the unvoiced stop consonants," Am. J. Psychol. **65**, 497–516.

Liberman, A. M., and Studdert–Kennedy, M. (**1978**). "Phonetic perception," in *Handbook of Sensory Physiology, Vol. III: Perception*, edited by R. Held, H. W. Leibowitz, and H.-L. Teuber (Springer-Verlag, New York).

Lindblom, B. E. F., and Studdert–Kennedy, M. (**1967**). "On the role of formant-transitions in vowel recognition," J. Acoust. Soc. Am. **42**, 830–843.

Lisker, L. (**1957a**). "Closure duration and the intervocalic voiced–voiceless distinction in English," Language **33**, 42–49.

Lisker, L. (**1957b**). "Minimal cues for separating /w,r,l,y/ in intervocalic position," Word **13**, 256–267.

Lotz, J., Abramson, A. S., Gerstman, L. J., Ingemann, F., and Nemser, W. J. (**1960**). "The perception of English stops by speakers of English, Spanish, Hungarian and Thai: a tape-cutting experiment," Lang. Speech **3**, 71–77.

Repp, B. H. (**1981**). "On levels of description in speech research," J. Acoust. Soc. Am. **69**, 1462–1464.

Schatz, C. (**1954**). "The role of context in the perception of stops," Language **30**, 47–56.

Stevens, K. N., and Blumstein, S. E. (**1978**). "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. **64**, 1358–1368.

Wickelgren, W. A. (**1976**). "Phonetic coding and serial order," in *Handbook of Perception, Vol. VII: Language and Speech*, edited by E. C. Carterette and M. P. Friedman (Academic, New York).