Pi ka pu: The perception of speech sounds by prelinguistic infants

J. A. FODOR, M. F. GARRETT, and S. L. BRILL
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Experimentation with 14-18-week-old infants indicates that they are capable of grouping together syllables of English depending on whether the syllables share a consonant. These results indicate that infants may have access to the mechanisms that underlie certain perceptual constancies in adult speech perception.

William James thought that the experimental world of the neonate was a "Blooming, buzzing confusion." But the real world, as Robert Louis Stevenson remarked, "is full of a number of *things*" (our emphasis). Much of classical cognitive psychology is about how this gap is bridged; how development shapes the aboriginal sensory flux to produce the adult ontology of properties, persons, events, and objects. This process is sometimes called the child's construction of reality.

Many psychologists no longer believe that the classical story is right, even in outline. In particular, they doubt that the child is ever faced with an unstructured flow of sensations in which the real world must somehow be discovered. On this revised view, the basic psychological mechanisms impose perceptual structure right from the beginning; cognitive development is primarily the elaboration of these structural commitments.

Whichever view one takes, it is clearly an empirical question what the precise nature of infant perception is and how it relates to that of adults. The research we will report here touches on an aspect of this question: to what extent do infants classify speech sounds in the way that adults do? Do prelinguistic infants recognize the patterns of identity and differences in terms of which adults taxonomize the speech signal? The results of our experiments with 14- to 18-week-old infants suggest that they very likely do. Differential reinforcement of head orientations to members of a randomly repeating series which consisted of three different CV syllables (e.g., /pa/, /pi/, /ku/) produced significantly greater resistance to extinction when the reinforced stimuli shared a phone (e.g., /pa/ and /pi/) than when they did not (e.g., /pa/ and /ku/). This result is interpretable on two assumptions: (1) Infants, like adults, find "disjunctive" concepts relatively hard to learn; i.e., learning is

faster when the positive discriminative stimuli satisfy a uniform description. (2) Infants, like adults, recognize phonetic identities; i.e., the infant's judgment of similarity between *syllables* is responsive to the number of *phones* that the syllables share. On these assumptions, a learning task in which /pa/ and /pi/ are the positive discriminative stimuli ought to be mastered more readily than a learning task in which /pa/ and /ku/ (or /pi/ and /ku/) are the positive discriminative stimuli; for, while the reinforcement conditions are disjunctive in the latter cases, they are *homogeneous* in the former case (all and only reinforced signals begin with /p/). The results obtained therefore suggest that a triple of syllables like /pa/, /pi/, and /ku/ is phonetically cross-classified for the infant just as it is for the adult: /pa/ and /pi/ are heard as having more in common with one another than either does with /ku/.

The significance of this pattern of findings can best be appreciated in the context of work on the psychophysics of adult speech perception. A long series of studies (most notably at the Haskings Laboratories) have rendered it implausible that a perceived phone can be defined by a given set of contemporaneous acoustic features. That is, more often than not, researchers have failed to find acoustic invariants corresponding to the perceptual identity of the phone. The evidence indicates that: (a) a heterogeneous set of acoustic tokens may be assigned to the same phonetic type; (b) the processes involved in such assignments often involve the categorical representation of acoustic continua; (c) the acoustic features relevant to such assignments are often widely distributed across the speech signal; ordinal relations among perceived phonetic segments do not, in general, preserve temporal relations among sub-stretches of the speech stream. In particular, the character of the acoustic representative of the consonantal segment of a CV syllable is typically heavily determined by the character of the acoustic representative of the vowel segment. Acoustically, a CV syllable is likely to be a relatively fused and undifferentiated object, even though phonetically

This research was supported in part by Grants HD 05168-01, -02, -03, National Institute of Mental Health, by a Sloan Foundation grant to H.-L. Teuber, and by a Grant Foundation grant to H.-L. Teuber.

Figure 1

(and perceptually) it is analyzed as a sequence of discrete and more or less independent elements. (For extensive discussion, see Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967.)

It is, in short, plausible to think of the phonetic/perceptual representation of the speech stream as a constancy, engendered by the operation of some sort of decoding mechanism which the adult applies to the acoustic input. The complexity of the computations this mechanism performs is suggested by the extraordinary difficulties thus far encountered in attempts to develop artificial phone recognizers. If this view of adult speech perception is correct, it implies one of two theories of the ontogenesis of speech perception (paralleling the two accounts of general cognitive development mentioned above). Either the relevant perceptual constancies must be learned, or they are part of the initial equipment that the infant brings to the process of internalizing the rules of his language. Theories of the first sort have stressed the importance of the child's experience of relatively large samples of speech in working out the distributional patterns (patterns of partial contrast between meaningful utterances) on which the linguistic justification of a phonetic analysis eventually rests. Or, in the case of "motor theories," they have stressed the importance of the child's monitoring of his own verbalizations in facilitating his discovery of the relation between wave forms and phonetic strings. In either case, however, such theories suggest (as nativistic theories do not) that the mastery of the phonetic/acoustic correspondence ought to be the consequence of a relatively extensive linguistic apprenticeship; one would hardly expect to find it available to prelinguistic infants.

Recent experiments (e.g., Eimas, Siqueland,

Juczzyk, & Vigorito, 1971) have suggested, nevertheless, that infants, like adults, impose a categorization on speech signals, and, more significantly, that the category boundaries for infants and adults apparently correspond. In our experimentation, we set out to explore the independent question of whether infants, like adults, respect the existence of internal syllabic structure. Can infants, despite their lack of experience with distributional properties of the language and despite their lack of experience with the output of their own articulatory system, abstract from the *acoustic* contamination of one part of a syllable by the rest and thus recover a *linguistically* relevant representation of the syllable as a sequence of distinct phones? Does, in short, the infant hear the same pattern of phonetic identity and difference among parts of syllables that the adult hears?

DESIGN AND PROCEDURE

The paradigm we have used to test infants' perception of speech is, briefly, the following: Infants (14 to 18 weeks old) were presented with a series of 60 single-syllable sounds; three different syllables occurred in the series (e.g., /pa/, /pi/, /ku/), and there were 20 occurrences of each syllable.

Syllables were recorded in a female voice on a studio-quality recorder. Ten examples of each of the six syllable types were selected from a larger recorded set; multiple copies of these 10 were used to construct the stimulus tapes. The stimulus syllables were, by adult standards, all "good" examples of their type, and within each type, the "normal" acoustic variation found in a given speaker's repetitions of the same phone occurred.

The syllable types were presented to the subjects in a semirandom order from a sound source either to the infant's right or to his left; the locus of a particular syllable type varied randomly (except for the constraint that, summed across trials, the syllable occurred equally often left or right). For two of the three syllable types in the series, a reinforcing visual array located at the apparent sound source followed each occurrence of the syllable. The visual array consisted of a doll-mobile mounted in an internally illuminated box. The mobile rotated slowly during its display. The array was illuminated immediately after a correctly orienting head turn and remained on for 4.5 sec. If no orienting head turn occurred within a 3-sec interval following syllable presentation, the visual display was turned on automatically for 4.5 sec. See Figure 1 for a schematic representation of the test situation. The test session lasted for approximately ½ h and was repeated on 10 separate days within a 2-week period.

The variable of interest is the incidence of anticipatory head turns: those correctly orienting head turns that occur in the interval following presentation of a syllable but prior to the illumination of the visual array. This type of head turn may be taken as an index of the infant's ability to predict occurrences of the reinforcing visual array as a function of the character of the stimulus syllable. The incidence of such anticipatory head turns¹ for each of the syllable types in the stimulus series may, thus, be compared for groups in which the two syllables that are reinforced have consonants that are perceptually distinct for adults (e.g., /pi/ and /ka/) and those in which the reinforced syllables have consonants which are perceptually the *same* for adults (e.g., /pi/ and /pu/).

Consonants were counterbalanced across vowel environments, yielding three syllable sets. The two reinforcement conditions for each such set yield six experimental groups. Six infants were tested in each of these groups. The design is outlined in Table 1.

RESULTS

The performance of infants was compared in terms of the proportions of total anticipatory responses to each syllable type that occurred in the last 5 of the 10 sessions. If presentation of the visual array affects changes in the rate of anticipatory responses, reinforced syllable types will show a larger proportion of such responses in the later sessions than will unreinforced syllable types, and this difference should manifest itself regardless of the initial rates of response for particular syllable types. Of course, this will be true only if infants are capable of distinguishing among the various syllables in the stimulus series, and of connecting the occurrence of the visual array with particular syllable types. Further, on the view that the infants are *also* capable of appreciating the partial phonetic identities among the syllables, one would expect the differences between proportions of response to reinforced and unreinforced syllable types to be greater in the *same-phones* group than in the *different-phones* group. If, on the other hand, each of the syllable types in a series is heard by the infants as equally distinct from each of the other two, no differences as a function of syllable grouping should be found.

A two-factor analysis of variance (Factor 1, reinforced vs. unreinforced syllables; Factor 2, syllable grouping by same or different consonants) with repeated measures on the first factor² shows a significant effect for reinforcement ($F = 5.68$, $df = 1, 34$, $p < .05$), and for the interaction of syllable grouping with reinforcement (i.e., presentation of the visual array) ($F = 4.14$, $df = 1, 34$, $p < .05$). Tests for differences among the means show that the effects in the reinforcement factor are due to the same-phones group ($t = 2.94$, $df = 17$, $p < .01$), with no significant effects in the different-phones group ($t = .26$, $df = 17$). Table 2 gives the mean second half proportions for reinforced and unreinforced syllables in the same and different phones groups.

The strongest reflection of the effect of syllable grouping on the infants' ability to predict appearances of the visual array is, of course, the interaction. This indicates that the differences between anticipatory responses to reinforced and unreinforced syllables was significantly greater in the same-phones group than in the different-phones group, even though the direction

Table 2
Proportions of Anticipatory Head Turns in Last Five Test Sessions for Same and Different Phone Conditions

Syllable Grouping	Reinforcement Conditions			
Same Phones Condition	/pa/ & /pi/	.428	/ku/	.358
	/pi/ & /pu/		/ka/	
	/pu/ & /pa/		/ki/	
Different Phones Condition	/pa/ & /ku/	.413	/pi/	.404
	/pi/ & /ka/		/pu/	
	/pu/ & /ki/		/pa/	

of relation between the means in the two groups was the same.

DISCUSSION

These results parallel those we obtained earlier with a variation of the current paradigm and a different pair of consonants (Fodor, Garrett, & Shapero, Note 1). In the earlier experiment, we used the consonants /p/ and /g/ in a design similar to one described for the present study.³ The results, like those described above, were that infants in the same-phones condition showed significantly greater retention of the head-turning response across test sessions than did infants in the different-phones condition.

The stops we have used in these experiments have the virtue of yielding cases of what might be called "acoustic overlap" across vowel environments. Thus, for example, the noise burst appropriate to the acoustic representation of the percept /p/ in the environment /i/ is the same as that appropriate to /k/ (or /g/) in the environment /a/ (see Liberman, Delattre, Cooper, & Gerstman, 1954, and Schatz, 1954).

There are, in fact, three types of acoustic cues which seem to play a significant role in distinguishing among stops that differ in place of articulation: burst frequency, aspiration, and formant transitions. Of these, the indications are that burst frequency and aspiration are most important for distinguishing stops in the vowel environments /i/ and /u/, while formant transitions are of greater importance in the vowel environment /a/ (Fischer-Jørgensen, Note 2). Thus, the types of acoustic cues which distinguish initial stops are relativized to their vowel environment. Moreover, *within* types there is also a conditioning of the cue value of a given acoustic event as a function of vowel environment. The most striking such case is the pronounced failure of acoustic invariance for burst and aspiration cues in the stops /g/ and /k/. If, for example, one transposes the noise burst and aspiration (first 50 msec) from the initial portions of the syllable /ki/ to the vowel /u/, one gets identification of the composite as the syllable "pu." In general, a number of tape-splicing experiments (Cole

Table 1

Syllable Series	Visual Array Presented	
	"Same Phones" Conditions	"Different Phones" Conditions
A	/pi/, /ka/, /pu/	for /pi/ & /pu/ for /pi/ & /ka/
B	/pa/, /ku/, /pi/	for /pa/ & /pi/ for /pa/ & /ku/
C	/pu/, /ki/, /pa/	for /pu/ & /pa/ for /pu/ & /ki/

& Scott, 1974; Schatz, 1954; Fischer-Jørgenson, Note 2) and synthetic speech experiments (Delattre, Liberman, & Cooper, 1955; Liberman, Delattre, & Cooper, 1952; Liberman, Delattre, Cooper, & Gerstman, 1954) have found that the effect of these cues on the identification of stops must be relativized to their vowel environment.

If we apply these observations to the analysis of the infants' task in the present experiments, it is clear that their perceptual responses cannot turn on a sensitivity to one or two simple acoustic parameters which partition the stimuli into the reinforced and unreinforced categories. For example, if infants were assumed to be responding simply to the burst-frequency cue in order to group /pi/ with /pu/ and exclude /ka/ (in Group A), we would have to assume that they ignore that cue as a basis for classification in the "pu ki pa" case (Group C), for, (a) in the case of /pa/, the salient cues (for adults) are the formant transitions and (b) the /k/ before /i/ burst is "p-like" (e.g., gives rise to /p/ judgments when spliced onto the vowel /u/).

Note that even the apparently simple case of voiced vs. voiceless stops (e.g., the g/p contrasts of Fodor, Garrett, & Shapero) is less straightforward than it might appear, for there is evidence that the principal cue involved, VOT (voice onset time), varies in its value as a function of vowel environment. Thus, a VOT of 30 msec, which might be adequate to shift perception from +vcd to -vcd before the vowel /a/, will not be sufficient before the vowel /i/ (see, Cooper, 1974, for some experimental evidence bearing on this issue).

In short, if infants were responding just to the presence of some value of a specific acoustic parameter in order to perform their perceptual sorting, the "wrong" classifications would have emerged (where "wrong" is defined by adult perceptual sorting of the stimuli). There must, of course, be some set of acoustic parameters which will appropriately sort the syllables in these experiments; that is to admit no more than that the infants' performance is not occult. What is significant is that the parameters which the infants apparently *do* select are those which yield the adult classification of percepts. In general, we do not therefore consider that the existence of such cues could have rendered the infants' perceptual task (that of selecting among the various available acoustic parameters available just those which yield the perceptually relevant adult taxonomy) less exacting.

The results of the current experiment and the earlier one both indicate that 14- to 18-week-old infants are capable of a segmental analysis of syllables (i.e., they respond to their internal structure as a sequence of phones). It appears that infants with no experience in producing articulate speech and with no

experience of the distributional features of language nevertheless can appreciate the perceptual identity of stop consonants across vowel environments, and to this extent, at least, their naive perceptual analysis of speech signals corresponds to the sophisticated adult one.

REFERENCE NOTES

1. Fodor, J. A., Garrett, M. F., & Shapero, D. B. Discrimination among phones by infants. Quarterly Progress Report No. 96, Research Laboratory of Electronics, M.I.T., 1969.
2. Fischer-Jørgensen, E. Tape cutting experiments with Danish stop consonants in initial position. Annual Report VII, University of Copenhagen, Institute of Phonetics, 1972.

REFERENCES

- COLE, R. A., & SCOTT, B. The phantom in the phoneme: Invariant characteristics of stop consonants. *Perception & Psychophysics*, 1974, 15, 101-107.
- COOPER, W. E. Contingent feature analysis in speech perception. *Perception & Psychophysics*, 1974, 16, 201-204.
- DELATTRE, P. C., LIBERMAN, A. M., & COOPER, F. S. Acoustic loci and transitional cues for consonant. *Journal of the Acoustical Society of America*, 1955, 27, 769-773.
- EIMAS, P., SIQUELAND, E., JUSCZYK, P., & VIGORITO, J. Speech perception in infants. *Science*, 1971, 171, 303-306.
- LIBERMAN, A. M., DELATTRE, P. C., & COOPER, F. S. The role of selected stimulus variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, 1952, 65, 497-516.
- LIBERMAN, A. M., DELATTRE, P. C., COOPER, F. S., & GERSTMAN, L. J. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs*, 1954, 68, 1-13.
- LIBERMAN, A. M., COOPER, F. S., SHANKWEILER, D. P., & STUDDERT-KENNEDY, M. Perception of the speech code. *Psychological Review*, 1967, 74, 431-461.
- SCHATZ, C. D. The role of context in the perception of stops. *Language*, 1954, 30, 47-56.

NOTES

1. No more than a single such turn was counted per presentation of a syllable. If the infant oriented to the sound, looked away, and then returned, only *one* anticipatory turn was counted. This is relevant only for the *unreinforced* trials, since for reinforced trials the initial head turn toward the stimulus triggered the visual array.

2. This analysis simplifies some details of the experiment. The variable of "syllable series" (resulting from counterbalancing consonants across the vowel environments) is not represented. There is no indication, however, of a vowel grouping effect (e.g., of greater effects of reinforcement when /i/ and /u/ are paired—regardless of consonants—than when /i/ and /a/ are paired). There are differences among the three syllable series in the degree to which the syllable grouping effects appear, however; the means for the same phones condition were higher than those for the different phones condition in all three syllable series, but the effects were greatest in the pa-ku-pi and the pu-ki-pa groups. The pi-ka-pu group showed a superiority of reinforced over unreinforced syllables for both same and different phone conditions. Interpretation of differences among the three syllable series is risky, given that it is a between-subjects comparison with only six subjects in each group. Our own guess is that with larger subject groups the differences among the syllable series would diminish.

One might also analyze the contrast between reinforced and unreinforced syllables in greater detail than first half of the sessions

vs. second half. A priori, there is no reason to expect that infants would not be able to make the appropriate sorting of the syllables sometime during the first two or three sessions. For some of the infants, there is indication that this is the case. Our analysis simply focuses on those trials in which the contrast is sharpest across the full set of infants.

3. The earlier experiment used visual scoring of infants' head turns and a different reinforcing visual stimulus (motion pictures).

Infants were also seated on their mother's laps during the test sessions rather than in an infant seat.

(Received for publication December 1974;
revision accepted April 25, 1975.)