Reprinted from
22 January 1971, Volume 171, pp. 303-306

SCIENCE

Speech Perception in Infants

Abstract. *Discrimination of synthetic speech sounds was studied in 1- and 4-month-old infants. The speech sounds varied along an acoustic dimension previously shown to cue phonemic distinctions among the voiced and voiceless stop consonants in adults. Discriminability was measured by an increase in conditioned response rate to a second speech sound after habituation to the first speech sound. Recovery from habituation was greater for a given acoustic difference when the two stimuli were from different adult phonemic categories than when they were from the same category. The discontinuity in discrimination at the region of the adult phonemic boundary was taken as evidence for categorical perception.*

In this study of speech perception, it was found that 1- and 4-month-old infants were able to discriminate the acoustic cue underlying the adult phonemic distinction between the voi ed and voiceless stop consonants /b/ and /p/. Moreover, and more important, there was a tendency in these subjects toward categorical perception: discrimination of the same physical difference was reliably better across the adult phonemic boundary than within the adult phonemic category.

Earlier research using synthetic speech sounds with adult subjects uncovered a sufficient cue for the perceived distinction in English between the voiced and voiceless forms of the stop consonants, /b-p/, /d-t/, and /g-k/, occurring in absolute initial position (*1*). The cue, which is illustrated in the spectrograms displayed in Fig. 1, is the onset of the first formant relative to the second and third formants. It is possible to construct a series of stimuli that vary continuously in the relative onset time of the first formant, and to investigate listeners' ability to identify and discriminate these sound patterns. An investigation of this nature (*2*) revealed that the perception of this cue was very nearly categorical in the sense that listeners could discriminate continuous variations in the relative onset of the first formant very little better than they could identify the sound patterns absolutely. That is, listeners could readily discriminate between the voiced and voiceless stop consonants, just as they would differentially label them, but they were virtually unable to hear intraphonemic differences, despite the fact that the acoustic variation was the same in both conditions. The most measurable indication of this categorical perception was the occurrence of a high peak of discriminability at the boundary between the voiced and voiceless stops, and a nearly chance level of discriminability among stimuli that represented acoustic variations of the same phoneme. Such categorical perception is not found with nonspeech sounds that vary continuously along physical continua such as frequency or intensity. Typically, listeners are able to discriminate many more stimuli than they are able to identify absolutely, and the dis-

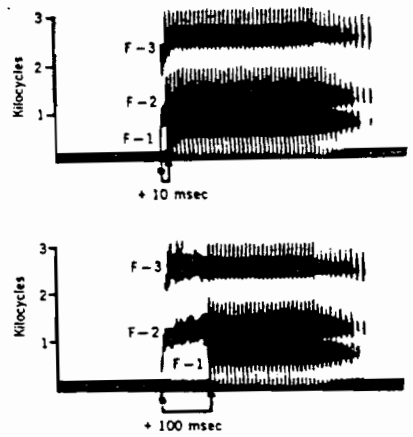

Fig. 1. Spectrograms of synthetic speech showing two conditions of voice onset time (VOT): slight voicing lag in the upper figure and long voicing lag in the lower figure. The symbols *F-1*, *F-2*, and *F-3* represent the first three formants, that is, the relatively intense bands of energy in the speech spectrum. [Courtesy of L. Lisker and A. S. Abramson]

criminability functions do not normally show the same high peaks and low troughs found in the case of the voicing distinction (*3*). The strong and unusual tendency for the stop consonants to be perceived in a categorical manner has been assumed to be the result of the special processing to which sounds of speech are subjected and thus to be characteristic of perception in the speech or linguistic mode (*4*).

Because the voicing dimension in the stop consonants is universal, or very nearly so, it may be thought to be reasonably close to the biological basis of speech and hence of special interest to students of language development. Though the distinctions made along the voicing dimension are not phonetically the same in all languages, it has been found in the cross-language research of Lisker and Abramson (*5*) that the usages are not arbitrary, but rather very much constrained. In studies of the production of the voicing distinction in 11 diverse languages, these investigators found that, with only minor exceptions, the various tokens fell at three values along a single continuum. The continuum, called voice onset time (VOT), is defined as the time between the release burst and the onset of laryngeal pulsing or voicing. Had the location of the phonetic distinctions been arbitrary, then different languages might well have divided the VOT continuum in many different ways, constrained only by the necessity to space the different modal values of VOT sufficiently far apart as to avoid confusion.

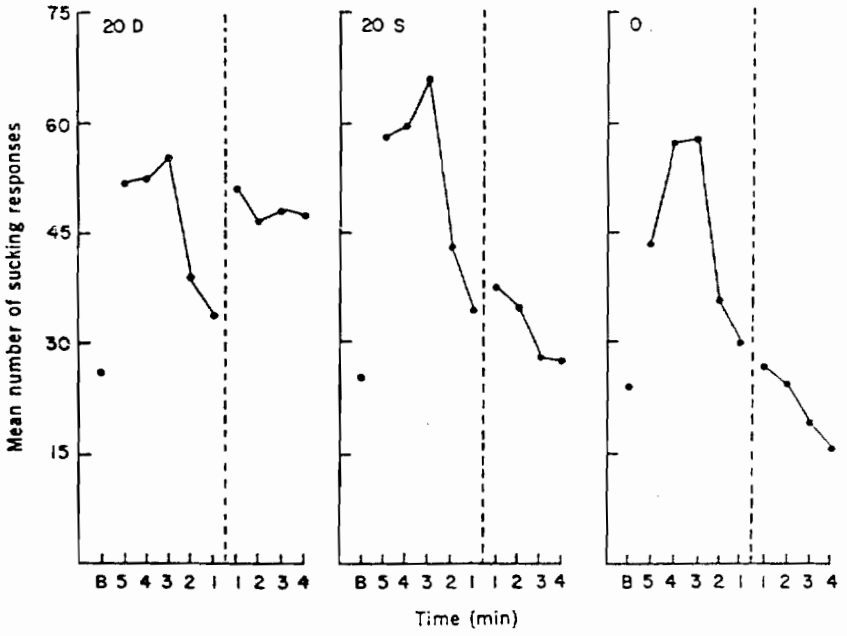

Fig. 2. Mean number of sucking responses for the 4-month-old infants, as a function of time and experimental condition. The dashed line indicates the occurrence of the stimulus shift, or in the case of the control group the time at which the shift would have occurred. The letter *B* stands for the baseline rate. Time is measured with reference to the moment of stimulus shift and indicates the 5 minutes prior to and the 4 minutes after shift.

Not all languages studied make use of the three modal positions. English, for example, uses only two locations, a short lag in voicing and a relatively long lag in voicing. Prevoicing or long voicing lead, found in Thai, for example, is omitted. Of interest, however, is the fact that all languages use the middle location, short voicing lag, which, given certain other necessary articulatory events, corresponds to the English voiced stop /b/, and one or both of the remaining modal values. The acoustic consequences for two modes of production are shown in Fig. 1; these correspond to short and long voicing lags, /b/ and /p/, respectively.

Given the strong evidence for universal—and presumably biologically determined—modes of production for the voicing distinction, we should suppose that there might exist complementary processes of perception (*6*). Hence, if we are to find evidence marking the beginnings of speech perception in a linguistic mode, it would appear reasonable to initiate our search with investigations of speech sounds differing along the voicing continuum. What was done experimentally, in essence, was to compare the discriminability of two synthetic speech sounds separated by a fixed difference in VOT under two conditions: in the first condition the two stimuli to be discriminated lay on opposite sides of the adult phonemic boundary, whereas in the second condition the two stimuli were from the same phonemic category.

The experimental methodology was a modification of the reinforcement procedure developed by Siqueland (*7*). After obtaining a baseline rate of high-amplitude, nonnutritive sucking for each infant, the presentation and intensity of an auditory stimulus was made contingent upon the infant's rate of high-amplitude sucking. The nipple on which the child sucked was connected to a positive pressure transducer that provided polygraphic recordings of all responses and a digital record of criterional high-amplitude sucking responses. Criterional responses activated a power supply that increased the intensity of the auditory feedback. A sucking rate of two responses per second maintained the stimulus at maximum intensity, about 75 db (13 db over the background intensity of 62 db).

The presentation of an auditory stimulus in this manner typically results in an increase in the rate of sucking com-

pared with the baseline rate. With continued presentation of the initial stimulus, a decrement in the response rate occurs, presumably as a consequence of the lessening of the reinforcing properties of the initial stimulus. When it was apparent that attenuation of the reinforcing properties of the initial stimulus had occurred, as indicated by a decrement in the conditioned sucking rate of at least 20 percent for two consecutive minutes compared with the immediately preceding minute, a second auditory stimulus was presented without interruption and again contingent upon sucking. The second stimulus was maintained for 4 minutes after which the experiment was terminated. Control subjects were treated in a similar manner, except that after the initial decrease in response rate, that is, after habituation, no change was made in the auditory stimulus. Either an increase in response rate associated with a change in stimulation or a decrease of smaller magnitude than that shown by the control subjects is taken as inferential evidence that the infants perceived the two stimuli as different.

The stimuli were synthetic speech sounds prepared by means of a parallel resonance synthesizer at the Haskins Laboratories by Lisker and Abramson. There were three variations of the bilabial voiced stop /b/ and three variations of its voiceless counterpart /p/. The variations between all stimuli were in VOT, which for the English stops /b/ and /p/ can be realized acoustically by varying the onset of the first formant relative to the second and third formants and by having the second and third formants excited by a noise source during the interval when the first formant is not present. Identification functions from adult listeners (*8*) have indicated that when the onset of the first formant leads or follows the onset of the second and third formants by less than 25 msec perception is almost invariably /b/. When voicing follows the release burst by more than 25 msec the perception is /p/. Actually the sounds are perceived as /ba/ or /pa/, since the patterns contain three steady-state formants appropriate for a vowel of the type /a/. The six stimuli had VOT values of −20, 0, +20, +40, +60, and +80 msec. The negative sign indicates that voicing occurs before the release burst. The subjects were 1- and 4-month-old infants, and within each age level half of the subjects were males and half were females.

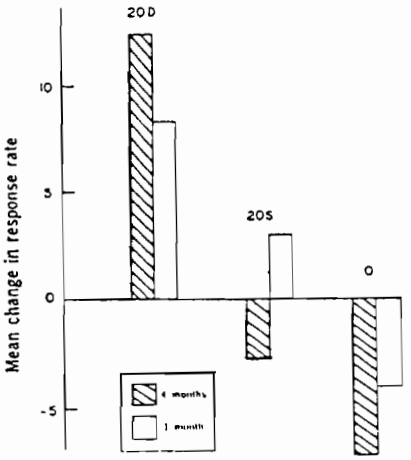

Fig. 3. The mean change in response rate as a function of experimental treatments, shown separately for the 1- and 4-month-old infants. (See text for details.)

The main experiment was begun after several preliminary studies established that both age groups were responsive to synthetic speech sounds as measured by a reliable increase in the rate of sucking with the response-contingent presentation of the first stimulus ($P < .01$). Furthermore, these studies showed that stimuli separated by differences in VOT of 100, 60, and 20 msec were discriminable when the stimuli were from different adult phonemic categories; that is, there was reliable recovery of the rate of sucking with a change in stimulation after habituation ($P < .05$). The finding that a VOT difference of 20 msec was discriminable permitted within-phonemic-category discriminations of VOT with relatively realistic variations of both phonemes.

In the main experiment, there were three variations in VOT differences at each of two age levels. In the first condition, 20D, the difference in VOT between the two stimuli to be discriminated was 20 msec and the two stimuli were from different adult phonemic categories. The two stimuli used in condition 20D had VOT values of +20 and +40 msec. In the second condition, 20S, the VOT difference was again 20 msec, but now the two stimuli were from the same phonemic category. In this condition the stimuli had VOT values of −20 and 0 msec or +60 and +80 msec. The third condition, 0, was a control condition in which each subject was randomly assigned one of the six stimuli and treated in the same manner as the experimental subjects, except that after habituation no change in stimulation was made. The control group served to counter any argument that the increment in response rate associated with a change in stimulation was artifactual in that the infants tended to respond in a cyclical manner. Eight infants from each age level were randomly assigned to conditions 20D and 20S, and ten infants from each age level were assigned to the control condition.

Figure 2 shows the minute-by-minute response rates for the 4-month-old subjects for each of the training conditions separately. The results for the younger infants show very nearly the identical overall pattern of results seen with the older infants. In all conditions at both age levels, there were reliable conditioning effects: the response rate in the third minute prior to shift was significantly greater than the baseline rate of responding ($P < .01$). As was expected from the nature of the procedure, there were also reliable habituation effects for all subjects. The mean response rate for the final 2 minutes prior to shift was significantly lower than the response rate for the third minute before shift ($P < .01$). As is apparent from inspection of Fig. 1, the recovery data for the 4-month-old infants were differentiated by the nature of the shift. When the mean response rate during the 2 minutes after shift was compared with the response rate for the 2 minutes prior to shift, condition 20D showed a significant increment ($P < .05$), whereas condition 20S showed a nonsignificant decrement in responding ($P > .05$). In the control condition, there was a fairly substantial decrement in responding during the first 2 minutes of what corresponded to the shift period in the experimental conditions. However, the effect failed to reach the .05 level of significance, but there was a reliable decrement when the mean response rate for the entire 4 minutes after shift was compared with the initial 2 minutes of habituation ($P < .02$). The shift data for the younger infants were quite similar. The only appreciable difference was that in condition 20S there was a nonsignificant increment in the response rate during the first 2 minutes of shift.

In Fig. 3 the recovery data are summarized for both age groups. The mean change in response rate (that is, the mean response rate for the initial 2 minutes of shift minus the mean response rate during the final 2 minutes before shift) is displayed as a function

of experimental treatments and age. Analyses of these data revealed that the magnitude of recovery for the 20D condition was reliably greater than that for the 20S condition ($P < .01$). In addition, the 20D condition showed a greater rate of responding than did the control condition ($P < .01$), while the difference between the 20S and control conditions failed to attain the .05 level of significance.

In summary, the results strongly indicate that infants as young as 1 month of age are not only responsive to speech sounds and able to make fine discriminations but are also perceiving speech sounds along the voicing continuum in a manner approximating categorical perception, the manner in which adults perceive these same sounds. Another way of stating this effect is that infants are able to sort acoustic variations of adult phonemes into categories with relatively limited exposure to speech, as well as with virtually no experience in producing these same sounds and certainly with little, if any, differential reinforcement for this form of behavior. The implication of these findings is that the means by which the categorical perception of speech, that is, perception in a linguistic mode, is accomplished may well be part of the biological makeup of the organism and, moreover, that these means must be operative at an unexpectedly early age.

PETER D. EIMAS
EINAR R. SIQUELAND
PETER JUSCZYK
JAMES VIGORITO
Department of Psychology,
Brown University,
Providence, Rhode Island 02912

References and Notes

1. A. M. Liberman, P. C. Delattre, F. S. Cooper, *Language and Speech* 1, 153 (1958); A. M. Liberman, F. Ingemann, L. Lisker, P. C. Delattre, F. S. Cooper, *J. Acoust. Soc. Amer.* 31, 1490 (1959). It should be emphasized that the cues underlying the voicing distinction as discussed in the present report apply only to sound segments in absolute initial position.
2. A. M. Liberman, K. S. Harris, H. S. Hoffman, H. Lane, *J. Exp. Psychol.* 61, 370 (1961).
3. P. D. Eimas, *Language and Speech* 6, 206 (1963); G. A. Miller, *Psychol. Rev.* 63, 81 (1956); R. S. Woodworth and H. Schlosberg, *Experimental Psychology* (Holt, New York, 1954).
4. A. M. Liberman, F. S. Cooper, D. P. Shankweiler, M. Studdert-Kennedy, *Psychol. Rev.* 74, 431 (1967); M. Studdert-Kennedy, A. M. Liberman, K. S. Harris, F. S. Cooper, *ibid.* 77, 234 (1970); M. Studdert-Kennedy and D. Shankweiler, *J. Acoust. Soc. Amer.*, in press.
5. L. Lisker and A. S. Abramson, *Word* 20, 384 (1964).
6. P. Lieberman, *Linguistic Inquiry* 1, 307 (1970).
7. E. R. Siqueland, address presented before the 29th International Congress of Psychology, London, England (August 1969); ——— and C. A. DeLucia, *Science* 165, 1144 (1969).
8. L. Lisker and A. S. Abramson, *Proc. Int. Congr. Phonet. Sci. 6th* (1970), p. 563.
9. Supported by grants HD 03386 and HD 04146 from the National Institute of Child Health and Human Development. P.J. and J.V. were supported by the NSF Undergraduate Participation Program (GY 5872). We thank Dr. F. S. Cooper for generously making available the facilities of the Haskins Laboratories. We also thank Drs. A. M. Liberman, I. G. Mattingly, A. S. Abramson, and L. Lisker for their critical comments. Portions of this study were presented before the Eastern Psychological Association, Atlantic City (April 1970).

14 September 1970