Perception & Psychophysics
1976, Vol. 19(4), 285-289

Speech perception by rhesus monkeys: The voicing distinction in synthesized labial and velar stop consonants

R. S. WATERS and W. A. WILSON, JR.
University of Connecticut, Storrs, Connecticut 06268

Monkeys were presented with synthetic speech stimuli in a shock-avoidance situation. On the basis of their behavior, perceptual boundaries were determined along the physical continua between /ba/ and /pa/, and /ga/ and /ka/, that were close to the human boundaries between voiced and voiceless consonants. As is the case with humans, discrimination across a boundary was better than discrimination between stimuli that were both on one side of the boundary, and there was generalization of the voiced-voiceless distinction from labial to velar syllables. Unlike humans, the monkeys showed large shifts in boundary when the range of stimuli was varied.

It has been claimed (e.g., Studdert-Kennedy, Liberman, Harris, & Cooper, 1970), although not without some disagreement (Lane, 1965), that speech perception in human beings is fundamentally different from other auditory perception, i.e., that it is categorical in nature and that it involves specialized neural processing mechanisms. Comparisons of speech and nonspeech perception (e.g., Liberman, Harris, Kinney, & Lane, 1961), analyses of perceptual mechanisms for different classes of speech stimuli (e.g., Shankweiler & Studdert-Kennedy, 1967), and electrophysiological studies of brain activity (e.g., Wood, Goff, & Day, 1971) have provided the major sources of evidence. Another kind of evidence comes from the study of nonspeaking subjects, such as human infants (Eimas, Siqueland, Jusczyk, & Vigorito, 1971) and nonhuman animals (chinchillas: Kuhl & Miller, 1975; and monkeys).

It seems particularly appropriate to study these problems in nonhuman primates, for their behavior may give some hints about the evolutionary development of human language. Morse and Snowdon (1975) and Sinnott (1974) have examined whether monkeys are able to discriminate between certain human speech sounds and whether the perceptual functions displayed are similar to those that have been seen in human subjects. In particular, they were concerned with determining whether monkeys show categorical perception of such stimuli (Studdert-Kennedy et al., 1970). Morse and Snowdon recorded the EKG as animals were presented with a series of stimuli, taking an increase in response rate when a new stimulus was presented as a manifestation of dishabituation and thus evidence of discrimination. Their results suggest that monkeys are better able to discriminate between two stimuli which fall in separate human perceptual categories (i.e., /dae/ vs. /bae/) than between two acoustically different stimuli within a class, but two stimuli within a category do indeed appear to be discriminable. Sinnott studied discrimination between variants of /b/ and /d/ using an appetitive task. Discrimination was shown, but monkeys failed to demonstrate those characteristics of behavior that the author took as evidence of categorical perception (e.g., a longer latency of response when both stimuli were within the same phonemic category).

In the present study, monkeys were trained on an avoidance task, using synthetic speech stimuli differing on the dimension of "voice onset time (VOT)." In natural speech, /ba/ and /pa/ differ in part in the time between the opening of the lips and the onset of vocal cord vibration (voicing); the synthetic syllables used here maintain the proper time intervals between the corresponding portions of the signal and produce the appropriate perceptions in humans. The basic procedure was to train animals to make different responses to each of a pair of stimuli differing in VOT and then to examine their responses to stimuli with intermediate VOT values.

This research was supported in part by USPHS Grant MH 10972 to the second author. The authors wish to thank A. M. Liberman and P. Lieberman for very helpful suggestions in the course of the research and for comments on the manuscript and A. S. Abramson and I. G. Mattingly for information and assistance. Tapes were prepared through the generosity of the Haskins Laboratories (Contract NIH 71-2420). Requests for reprints should be sent to W. A. Wilson, Jr., Department of Psychology, U-20, University of Connecticut, Storrs, Connecticut 06268.

METHOD

Subjects

The subjects were four adult rhesus monkeys. They had had

experience in food-rewarded discrimination tasks, using simple auditory stimuli and synthesized syllables. On the latter problem, none had performed consistently above chance.

Apparatus

A two-compartment shuttlebox was used. Each chamber was 50 cm square and was illuminated by a 2.8-W overhead bulb. A 30-cm-high barrier separated the chambers; the opening above it could be closed by a sliding door operated by a pulley system. A 4-in. speaker faced equally into the compartments. The chambers had a grid floor to which shock could be applied via a scrambler system.

Procedure

On "go" trials, animals were required to shuttle over the barrier within 8 sec of stimulus onset. During "no-go" trials, the monkey was required to remain in one chamber for the entire 10-sec stimulus presentation. Incorrect responses were followed by momentary closure of the barrier opening, with a brief footshock. Each stimulus was a repetitive train of synthesized human syllables (a stop consonant and the vowel /a/), with an inter-onset time of 1 sec. Within each phase of the experiment, "go" and "no-go" stimuli differed only in VOT. The stimuli were presented in a balanced order for 50 trials each day (with an intertrial interval of 10 sec).

Stimuli were composed of three-formant labial or velar syllables (see Abramson & Lisker, 1973) generated on the Haskins Laboratories parallel-resonance synthesizer. All had initial fundamental frequencies of 114 Hz that tapered off to 70 Hz during the final 120 msec. For stimuli with negative VOTs, low-frequency harmonics (centered at 154 Hz) were initially present, for a duration of up to 140 msec, to simulate glottal pulsing during the period of vocal-tract closure. At the point simulating release (opening of the vocal tract), all three formants appeared, with harmonics in the frequency bands (see below) appropriate to the consonant and vowel being simulated. For syllables with positive VOTs, F2 and F3 began first. During the interval between their onset and the beginning of voicing, they were excited by a hiss (band-limited noise) simulating the effect of aspiration through an open glottis. F1 was suppressed until the time of onset of voicing, after which the full formant pattern of fundamental-frequency harmonics continued for the remainder of the syllable. The formants contained two distinct portions: an initial transition lasting approximately 40 msec and a final steady state portion throughout the remainder of the syllable (415 msec). The center frequencies at initiation of the labial transitions were 154 Hz (F1), 921 Hz (F2), and 1,524 Hz (F3), while the corresponding values for the velar stimuli were 154, 1,386, and 1,360 Hz, respectively. The final steady state frequencies for both classes were 269, 1,232, and 2,525 Hz, respectively. The bandwidths were 60, 90, and 120 Hz, for F1 F2, and F3, respectively.

In Phases 1-4, animals were first trained to respond differentially to stimuli with labial consonants that differed widely on the VOT continuum. In Phase 1, the "go" stimulus had a VOT of +140 msec (/pa/) and the "no-go" stimulus was a prevoiced /ba/ of −140 VOT. After reaching criterion of 90% correct on two consecutive days, subjects received "boundary testing" using intermediate stimuli. Stimuli on the /pa/-/ba/ continuum were used with VOTs of +140, +70, 0, −70, and −140, respectively. Each day, subjects were retrained to criterion (8/10) on +140 and −140, and then received 40 trials with one of the 10 possible pairs of the five stimuli. The stimulus closer to +140 was the "go" stimulus, and the other the "no-go" stimulus.

During Phase 2, subjects were trained to criterion using VOT values of +100 msec (go) and −100 msec (no-go). Boundary testing was done as before, but with stimulus values of +100, +50, 0, −50, and −100 msec. Phase 3 was exactly like Phase 1, except that the reinforcement contingencies were reversed, e.g., −140 was the "go" training stimulus, etc. Phase 4 (and subsequent phases) maintained the same general response set (i.e., "go to the stimulus with less positive VOT"), but the training stimuli were 0 and +140 msec, and the test values were 0, +35, +70, +100, and +140 msec.

In Phase 5, subjects were retrained on 0 vs. +140 msec VOT labial stimuli, then transfer was tested to velar phonemes with, respectively, the same VOTs, i.e., /ga/ and /ka/. Blocks of 10 reinforced labial trials were given. When the subject made 8/10 correct on the labial trials, 10 nonshocked velar trials (5 each with 0- and 140-msec VOTs) were presented. If at least 8/10 responses indicated transfer on the basis of VOT, additional blocks of velar trials followed. Otherwise, the animal was returned to labial stimuli, again until 8/10 correct was reached. This procedure continued until the subject received a total of 40 velar transfer trials. Phase 6 was exactly like Phase 4 except that velar phonemes were used throughout.

RESULTS

All subjects met criterion on the training tasks presented. Trials to criterion ranged from 250 to 700 on the first task learned.

The data collected in the boundary-testing periods were analyzed by means of Thurstone's Case V scaling method (Guilford, 1954). This method provides a value for each stimulus on an equal-interval psychological scale. If the stimuli fall into two categories, the stimulus corresponding (by interpolation) to a scale value of zero will be one which lies between the two categories and can be taken as the boundary between the categories. The boundary values obtained in the various phases are shown in Table 1.

One striking finding was that the boundaries were all located on the positive half of the VOT continuum. Phases 1 and 3 both involved labial

Table 1
Boundary Values in msec of VOT

Phase		1	2	3	4	6
	Place	Labial	Labial	Labial	Labial	Velar
Stimuli	Go No-Go	+140, −140	+100, −100	−140 +140	0 +140	0 +140
	Others	+70, 0, −70	+50, 0, −50	−70, 0, +70	+35, +70, +100	+35, +70, +100
Monkey	No. 1	28	9	32	61	58
	No. 3	–	17	20	58	57
	No. 5	18	1	41	78	53
	No. 7	19	19	47	–	–
Mean		21.7	11.5	35.0	65.7	56.0

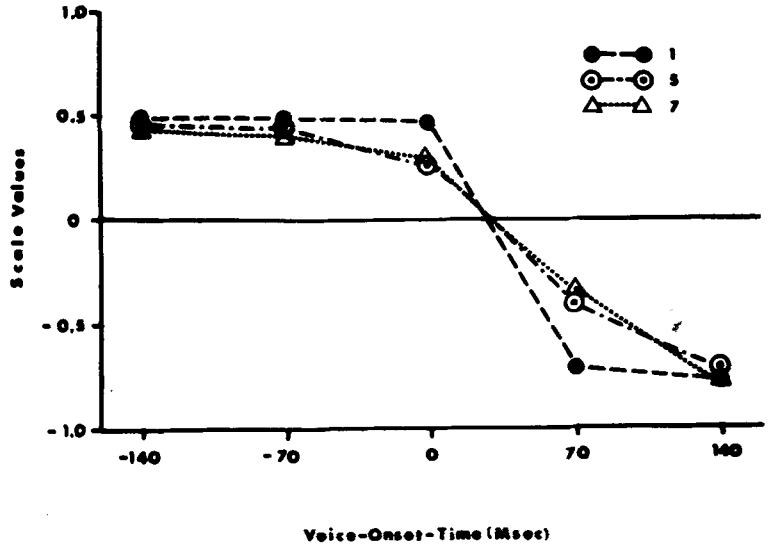

Figure 1. Scale values for /ba/-/pa/ stimuli derived from the paired comparisons in Phases 1 and 3 of the experiment.

syllables varying from −140 to +140 VOT. The fact that there was a difference in boundary value for these two phases suggests that the specific go/no-go requirements had some control over behavior, since Phase 3 was essentially a "reversal" of Phase 1. However, the boundary is positive in both phases; the overall mean on these two phases for the three relevant monkeys was 30.8 msec, which differs significantly from 0 msec, $t(2) = 43.8$, $p < .001$.

Figure 1 shows the scale values for the five stimuli, computed for individual monkeys from the responses in Phases 1 and 3 combined. As suggested above, all animals show a partition of the stimuli into two sets. Evidence that this point has a perceptual reality comes from a 2 by 2 analysis of variance which compared performance in four conditions: (a) when the pair of stimuli differed by one step (70 msec) and were both on the same side of the boundary (e.g., −140 vs. −70, or +70 vs. +140), (b) with the one-step stimulus pair that straddled the boundary (0 vs. +70), (c) with the stimulus pair that differed by two steps although both were on the same side of the boundary (−140 vs. 0), and (d) with two-step stimulus pairs that straddled the boundary (e.g., −70 vs. +70). When there was more than one comparison that fell into a class, each monkey's mean performance was calculated and entered into the analysis. Performance was significantly better when the two stimuli came from different sides of the boundary (approximately 83%) than when both were on the same side of the boundary (approximately 58%, $F(1,2) = 72.3$, $p < .05$, but neither the number of steps difference nor the interaction was significant (both Fs < 1).

For Phases 2, 4, and 6, similar comparisons were made for the between-category (boundary-straddling) and within-category performance, using measures that gave equal weight to one-step and two-step differences. In each case, mean performance was better on the between-category pairs: Phase 2, 71.5% vs. 58.6%, $t(3) = 3.76$, one-tailed $p < .025$; Phase 4, 74% vs. 73.6%, $t(2) = 0.07$, $p > .05$; Phase 6, 88.2% vs. 79.4%, $t(2) = 3.85$, $p < .05$.

The boundary locations were shifted by using stimuli that covered only a limited portion of the continuum. This is true when the center point of the stimulus set was unchanged (compare Phase 2 with Phase 1) as well as when it was shifted (compare Phase 4 with Phase 3).

On the velar transfer trials of Phase 5, the mean number "correct" (i.e., go responses to 0 msec VOT or no-go responses to +140 msec) was significantly above chance, at 33.0 out of 40, $t(2) = 4.26$, one-tailed $p < .05$. Thus /ga/ was perceived as more like /ba/, or /ka/ as more like /pa/, or both. The boundary values in Phase 6 (velar syllables) were not significantly different from those obtained with the comparable labial stimuli (Phase 4).

DISCUSSION

It has been demonstrated previously that monkeys can discriminate between synthetic speech syllables that differ in "place of articulation" (Morse & Snowdon, 1975; Sinnott, 1974); the present results show that they can also distinguish between "voiced" and "voiceless" phonemes.

The monkeys partitioned the series of labial syllables with VOT ranging from +140 to −140 msec at a value of approximately 30 msec VOT. This is similar to the behavior of the speakers of many

human languages. Thus English speakers identify such phonemes as /p/ when VOTs exceed +25 msec; otherwise they are heard as /b/. Chinchillas also show evidence of a phonetic VOT boundary similar to that of English-speaking adults, although the testing procedure did not involve as full a range of VOTs, including values less than zero (Kuhl & Miller, 1975).

Speech perception in humans is said to be categorical in nature (e.g., Studdert-Kennedy et al., 1970). More precisely, it is stated that speech perception is in general more categorical than nonspeech perception, and that some speech sounds (i.e., stop consonants) are perceived more categorically than others (i.e., vowels). One characteristic of categorical perception is that "stimuli drawn from a physical continuum are . . . perceived . . . as members of discrete categories. . . . Subjects . . . are able to discriminate between stimuli drawn from different categories, but not between stimuli drawn from the same category" (p. 234).

The monkey subjects of this experiment show behavior that tends to fit this pattern. Performance was very much better when two stimuli came from different sides of the animals' behaviorally defined phonetic boundary than when both were on the same side of the boundary, and the size of the difference between stimuli did not otherwise play a significant role. However, discrimination was not completely impossible between stimuli within the same category. The overall pattern is thus similar to that reported by Morse and Snowdon in their investigation of monkey perception of the cues for place of articulation. A more exact determination of the degree to which speech perception is categorical or continuous would require the comparison of identification data and discrimination data (cf. Liberman, Harris, Hoffman, & Griffith, 1957); this has not yet been reported for animal subjects (or for human infants). Other data indicate that categorical perception cannot serve as a unique hallmark, indicating the presence of processes similar to those found in human perception. For example, Wilson (1972) has shown that some characteristics of categorical perception may be seen in visual perception in monkeys, and Sinnott (1974) claims that strict categorical perception is not always seen in humans presented with synthetic speech sounds. It is abundantly clear, however, that perception of /ba/-/pa/ syllables is better in monkeys when stimuli lie on different sides of a definable boundary, and that this boundary is not far from one that plays a similar part in human speech perception.

There was clear evidence for generalization of the "voiced-voiceless" distinction from the labial to the velar syllables. This finding is somewhat similar to one with human beings. Greenberg and Jenkins (1964) reported that people judged /ga/ to be more like /ba/ than like /pa/, whereas /ka/ was felt to be about equidistant from /ba/ and /pa/. The difference in physical characteristics between /ba/ and /pa/ is similar to that between /ga/ and /ka/, and this simple acoustic pattern may be adequate to explain the transfer for either (or both) species.

For human subjects, the boundary location for velar stops is generally characterized by a higher VOT than that for labial stops (e.g., Abramson & Lisker, 1973). A similar finding has also been reported for chinchillas (Kuhl & Miller, 1975). In contrast, all subjects in the present study had higher VOT boundaries for labial stimuli. Although the difference was not statistically significant, there is certainly no evidence from the present experiment that monkeys and human beings are similar in this respect.

When the stimuli were restricted to only a portion of the original continuum, large shifts in boundary were seen—shifts that might be interpreted as adaptation effects. Although human beings are subject to adaptation effects in similar situations (Eimas & Corbit, 1973), the changes are not as large as that seen here, especially when the stimuli are consonants (Sawusch & Pisoni, Note 1).

The results are discussed above as if VOT were the only possible cue for distinction among the labial (or velar) stimuli. However, these data could well be interpreted in terms of the proposal that the presence or absence of a formant transition after voicing onset is an alternative (or additional) cue (Stevens & Klatt, 1974). Furthermore, they do not speak to the suggestion that the duration of the auditory memory store plays an important role in determining whether perception will be categorical or continuous (e.g., Pisoni, 1973).

These findings, nevertheless, provide a pattern of evidence about the similarities and differences between human and monkey perception of speech sounds, using a procedure in which the subject is an active participant. The fact that the boundary values were all positive, and approximated the human values rather closely in the /ba/-/pa/ case, is particularly convincing with the present procedure, because it included phases in which the training stimuli and the set of intermediate stimuli covered a wide range centered at zero. Furthermore, the superior performance that characterizes boundary-straddling discrimination provides reason to believe that the perceptual processes here are somewhat akin to those described as categorical in human speech perception. Monkeys showed stimulus-set effects that were probably larger than those that humans would display in an equivalent situation, however, and the relation between labial and velar stimulus pairs did not follow the human pattern.

The conclusion that speech-sound perception in

monkeys is in some ways similar to the same processes in human beings might be reasonably well accommodated by a generalized evolutionary view, but these data do not give much support to the proposition that human speech perception is fundamentally different from other auditory perception, nor to a species-specific motor interpretation of speech perception that has been offered to explain such a difference. Monkeys are able to discriminate /b/ from /p/, and in some ways the perception is categorical. According to Lieberman (1973), macaques have the articulatory apparatus requisite for producing labial stop consonants, even though these sounds are not heard in their natural vocalizations. The same perceptual findings are found for /g/ vs. /k/, despite the fact that these sounds are impossible for the animals to produce due to constraints of their vocal tract shape (Lieberman, 1973). Such findings do not sit well with the proposition that speech discrimination acquires its special characteristics from processes that have privileged access to motor mechanisms.

REFERENCE NOTE

1. Sawusch, J. R., & Pisoni, D. B. Category boundaries for speech and nonspeech sounds. *Research on speech perception*, Progress Report No. 1, Indiana University, 140-148, 1973-1974.

REFERENCES

ABRAMSON, A. S., & LISKER, L. Voice-timing perception in Spanish word-initial stops. *Journal of Phonetics*, 1973, **1**, 1-8.

ABRAMSON, A. S., & LISKER, L. Voice onset time in stop consonants: Acoustic analysis and synthesis. In *Proceedings 5th International Congress of Acoustics*. Liege: Imp. G. Thone, 1975.

EIMAS, P. D., & CORBIT, J. D. Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, 1973, **4**, 99-109.

EIMAS, P. D., SIQUELAND, E. R., JUSCZYK, P., & VIGORITO, J. Speech perception in infants. *Science*, 1971, **171**, 303-306.

GREENBERG, J. H., & JENKINS, J. J. Studies in the psychological correlates of the sound system of American English. *Word*, 1964, **20**, 157-177.

GUILFORD, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.

KUHL, P. K., & MILLER, J. D. Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science*, 1975, **190**, 69-72.

LANE, H. L. The motor theory of speech perception: A critical review. *Psychological Review*, 1965, **72**, 275-309.

LIBERMAN, A. M., HARRIS, K. S., HOFFMAN, H. S., & GRIFFITH, B. C. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 1957, **54**, 358-368.

LIBERMAN, A. M., HARRIS, K. S., KINNEY, J., & LANE, H. The discrimination of relative onset time of the components of certain speech and nonspeech patterns. *Journal of Experimental Psychology*, 1961, **61**, 379-388.

LIEBERMAN, P. On the evolution of language: A unified view. *Cognition*, 1973, **2**, 59-94.

MORSE, P. A., & SNOWDON, C. T. An investigation of categorical speech discrimination by rhesus monkeys. *Perception & Psychophysics*, 1975, **17**, 9-16.

PISONI, D. B. Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, 1973, **13**, 253-260.

SHANKWEILER, D. P., & STUDDERT-KENNEDY, M. Identification of consonants and vowels presented to left and right ears. *Quarterly Journal of Experimental Psychology*, 1967, **19**, 59-63.

SINNOTT, J. M. *A comparison of speech sound discrimination in humans and monkeys*. Unpublished doctoral dissertation, University of Michigan, 1974.

STEVENS, K. N., & KLATT, D. H. The role of formant transitions in the voiced-voiceless distinction for stops. *Journal of the Acoustical Society of America*, 1974, **55**, 653-659.

STUDDERT-KENNEDY, M., LIBERMAN, A. M., HARRIS, K. S., & COOPER, F. S. Motor theory of speech perception: A reply to Lane's critical review. *Psychological Review*, 1970, **77**, 234-249.

WILSON, M. Assimilation and contrast effects in visual discrimination by rhesus monkeys. *Journal of Experimental Psychology*, 1972, **93**, 279-282.

WOOD, C. C., GOFF, W. R., & DAY, R. S. Auditory evoked potentials during speech perception. *Science*, 1971, **173**, 1248-1251.

(Received for publication September 26, 1975; revision accepted January 21, 1976.)