Reprinted from
3 October 1975, Volume 190, pp. 69-72

SCIENCE

Speech Perception by the Chinchilla: Voiced-Voiceless Distinction in Alveolar Plosive Consonants

Patricia K. Kuhl and James D. Miller

Speech Perception by the Chinchilla: Voiced-Voiceless Distinction in Alveolar Plosive Consonants

Abstract. *Four chinchillas were trained to respond differently to /t/ and /d/ consonant-vowel syllables produced by four talkers in three vowel contexts. This training generalized to novel instances, including synthetically produced /da/ and /ta/ (voice-onset times of 0 and +80 milliseconds, respectively). In a second experiment, synthetic stimuli with voice-onset times between 0 and +80 milliseconds were presented for identification. The form of the labeling functions and the "phonetic boundaries" for chinchillas and English-speaking adults were similar.*

Neither speech analysis nor speech synthesis techniques have led to a successful account of our perception of speech sounds in terms of invariant acoustic properties (*1*). For this reason and others (*1*), current theorists have hypothesized that at least some classes of speech sounds are recognized by "special processing" (*2, 3*). Speculation as to the nature of this special-to-speech processing varies: some believe that it involves "phonetic feature detectors" that presumably respond to rather complicated and abstract characteristics of the acoustic signal (*4*); others have espoused a "motor theory" of speech perception (*2, 3*), which suggests that one's tacit knowledge of the acoustic results of articulatory maneuvers somehow mediates the perception of speech. While variously described, these current theories suggest that speech perception is a species-specific behavior, and thus, in large part, a uniquely human ability. As Liberman stated, "Unfortunately, nothing is known about the way non-human animals perceive speech . . . however, we should suppose that, lacking the speech-sound decoder, animals would not perceive speech as we do, even at the phonetic level" (*2*).

We asked whether the chinchilla, a mammal with auditory capabilities fairly similar to man's (*5*), but certainly without a phylogenetic history of "phonetic knowledge," either acoustic or articulatory, could correctly classify a large number of naturally produced syllables on the basis of the voicing contrast. In experiment 1, we trained four chinchillas, using an avoidance conditioning procedure, to respond differently to a variety of spoken /t/ and /d/ consonant-vowel (CV) syllables. Once trained, these animals correctly classified novel instances of /t/ and /d/ syllables, including syllables produced by new talkers, those produced in new vowel contexts, and computer-synthesized /ta/ and /da/ syllables. In experiment 2, we presented synthetic stimuli that varied systematically from /da/ to /ta/ to the animals trained on natural speech, to animals not trained on natural speech, and to English-speaking adults for identification. The labeling functions and the "phonetic boundaries" were similar for all animal and human subjects.

The voicing feature, which distinguishes voiced (/bdg/) from voiceless (/ptk/) plosives in English, is appropriate for investigations of speech perception by animals since this distinction has been examined in cross-language studies of adults (*6, 7*) and infants (*8, 9*). The acoustic properties that distinguish voiced and voiceless plosives in absolute-initial, prevocalic, stressed position can be most readily described as a timing difference between the onset of the plosive burst and the onset of voicing (*6*), termed the voice-onset time (VOT). In synthetic speech, VOT can be varied along a continuum to produce plosives in which voicing precedes the plosive burst (prevoiced), begins nearly simultaneously with the burst (unaspirated), or lags behind the plosive burst (aspirated). The VOT is specified in milliseconds; negative values indicate that voicing leads and positive values indicate that voicing lags. In English, prevoiced and unaspirated plosives constitute the voiced phonemic category and aspirated plosives constitute the voiceless phonemic category. When English speakers identify synthetic tokens from the VOT continuum, perception changes abruptly from voiced to voiceless sounds; the VOT at which responses divide equally between voiced and voiceless is termed the "phonetic boundary." These boundaries differ with the place of articulation of the voiced-voiceless pair; the boundaries for labials (/ba-pa/), alveolars (/da-ta/), and velars (/ga-ka/) are approximately +22, +35, and +41 msec, respectively (*7, 10*). Many languages divide the VOT continuum as we do in English (*6*), but in some languages, the division between voiced and voiceless categories occurs elsewhere. For example, in languages such as Kikuyu (*6, 9*), prevoiced plosives constitute the voiced category and unaspirated plosives constitute the voiceless category (aspirated plosives do not occur). However, infants 1 to 4 months old discriminate synthetic stimuli that fall on different sides of the English /b-p/ phonetic boundary (VOT of +22 msec) whether they are reared in an English-speaking environment where this boundary is phonemically relevant (*8*) or in a Kikuyu-speaking environment where it is not (*9*). These facts lend themselves to at least two interpretations: either young infants demonstrate this perceptual boundary because their "speech processor" responds to its potential phonemic relevance or because the boundary is a natural psychophysical one that could be demonstrated by a nonhuman as well. While our results do not rule out the first interpretation, they are consistent with the second.

In experiment 1, the CV syllables used during discrimination training (/ti, ta, tu, di, da, du/), were recorded twice by each of four talkers (two male and two female), and dubbed onto a disk pack for use in a digital recorder (*11*). The VOT measurements ranged from +40 to +128 msec for /t/ syllables and −200 to +28 msec for /d/ syllables.

The animals were tested in a double-grille cage suspended below a loudspeaker in a sound-treated booth (*5, 12*). Four chinchillas were deprived of water and trained to lick a drinking tube mounted at one end of the cage for their daily ration of water. Every third lick on the tube produced a drop of water. The animals were maintained at approximately 90 percent of their original weights. It had been demonstrated (*12*) that an ongoing activity, such as drinking, reduced the number of false alarms and intratrial avoidance responses during discrimination training.

For two animals, /t/ was the positive stimulus and /d/ was the negative stimulus; for the other two animals, these roles were reversed. On positive trials, the animal was trained to flee the drinking tube and cross the midline barrier to avoid shock. When the animal crossed the barrier, lights positioned at the barrier ends were lit briefly and the stimulus was terminated. Failure to cross the barrier within the 2.5-second trial interval resulted in simultaneous presentation of buzzer and shock until the crossing response was made. On negative trials, no consequences were delivered during the 2.5-second trial interval. If the animal correctly refrained from crossing, the water valve opened, making "free" water available for 1 second, and lights above the drinking tube were lit.

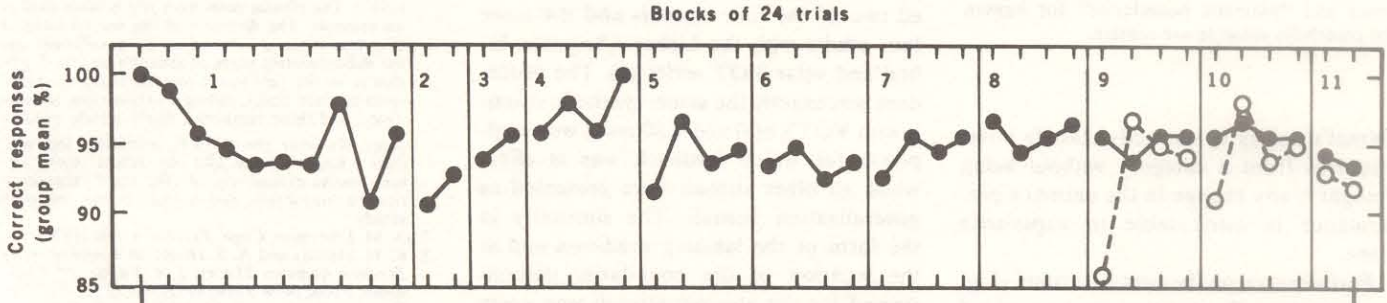
The animals were given 24 trials daily over a 7-month period. Each trial consisted of two presentations, separated by 500 msec, of the CV syllable. The time between trials was varied from 10 to 30 seconds, and the sound levels of the syllables were varied from trial to trial [52 to 66 db, sound pressure level (SPL)]. A masking noise (speech-shaped, 12 db SPL) was continuously presented in the test booth.

The experiment began with a single /tu/ and /du/ syllable; when this task was mastered, variation in the form of tokens (sec-

ond repetition by the same talker), talkers (identical phonetic token by different talkers), and vowels (different vowel contexts by a single talker) were singly introduced (*13*). The criteria for progressing from one condition to the next were that the group's percentage correct remain $\geq$ 90 percent for two consecutive days and that during this time no single animal's performance was less than 80 percent correct. Failure to make the crossing response on a positive trial and making the crossing response on a negative trial were both scored as errors. The group's percentage correct is displayed and the conditions of the experiment are described in Fig. 1. The only major problem the animals had with the task was mastering the /t-d/ distinction in new vowel contexts; however, after training on individual vowels, beginning with /ta/ and /da/, adding /ti/ and /di/, and finally /tu/ and /du/, the task was mastered. The successive addition of stimuli produced by talkers 2, 3, and 4 presented no problems (*14*).

There were three tests for generalization. The stimuli for talker generalization were the tokens /ti, ta, tu, di, da, du/ produced by four new talkers (two male and two female). The stimuli for vowel generalization were the tokens /te, tæ, to, de, dæ, dɔ/ produced by the same four new talkers. During synthetic generalization, /da/ (VOT, 0 msec) and /ta/ (VOT, +80 msec) stimuli were taken from the alveolar VOT continuum created by Lisker and Abramson (*7*).

On half of the trials during generalization tests, the stimuli presented were those already mastered by the animals. On the other half of the trials, generalization stimuli were randomly presented. In this way, performance on familiar stimuli served as a control for performance on novel stimuli. When control stimuli were presented, all feedback previously in effect was maintained. When generalization stimuli were presented, neither shock nor the buzzer were used, and all other feedback was arranged to indicate a correct response, no matter what the animal did. That is, on all generalization trials, barrier crossings during the 2.5-second trial interval resulted in lighting of barrier lights, and inhibition of the crossing response during the trial interval resulted in availability of "free" water. This procedure tests the



GRAPH PANEL	CONDITIONS	TOKENS	TALKERS	VOWELS	TOTAL STIMULI	VOT(MSEC) /d/	VOT(MSEC) /t/	$\bar{X}$ PERCENT CORRECT
1	INITIAL TRAINING (LAST 10 SESSIONS)	1	1	/u/	2	+16	+72	95.5
2	TOKEN VARIATION	2	1	/u/	4	-122 and +16	+72 and +76	91.7
3	TALKER VARIATION	1	4	/u/	8	-160 to +16	+72 to +128	94.8
-	VOWEL VARIATION: 1 TALKER	1	1	/iau/	6	-88 to +24	+64 to +76	85.2
-	RETRAINING WITH TALKER 1 AND INDIVIDUAL VOWELS							
4	VOWEL VARIATION: 1 TALKER	1	1	/iau/	6	-88 to +24	+64 to +76	97.4
5	VOWEL VARIATION: 2 TALKERS	1	2	/iau/	12	-200 to +24	+64 to +128	94.1
6	VOWEL VARIATION: 3 TALKERS	1	3	/iau/	18	-200 to +24	+64 to +128	93.7
7	VOWEL VARIATION: 4 TALKERS	1	4	/iau/	24	-200 to +24	+64 to +128	94.7
8	ALL TRAINING STIMULI	2	4	/iau/	48	-200 to +24	+40 to +128	96.0
9 TALKER	CONTROL	1	4	/iau/	24	-200 to +24	+64 to +128	95.3
9 TALKER	GEN.	1	4	/iau/	24	+10 to +28	+50 to +128	93.0
10 VOWEL	CONTROL	1	4	/iau/	24	-200 to +24	+64 to +128	96.1
10 VOWEL	GEN.	1	4	/e æ ɔ/	24	+9 to +27	+70 to +104	94.7
11 SYNTHETIC	CONTROL	1	4	/iau/	24	-200 to +24	+64 to +128	94.3
11 SYNTHETIC	GEN.	1	1	/a/	2	0	+80	92.8

Fig. 1. Conditions and major results of experiment 1. These results demonstrate that the chinchilla can be trained to discriminate /t/ from /d/ in absolute-initial, prevocalic, stressed position in spite of irrelevant changes in the sounds due to level, talker, and vowel (panels 1 to 8). Furthermore, such training generalizes to new talkers, new vowels, and synthetic stimuli (panels 9 to 11).

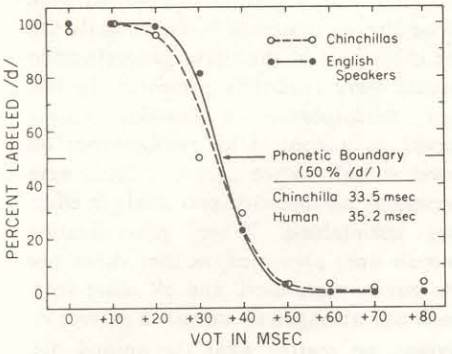


Fig. 2. Mean percentage of /d/ responses by chinchilla and human subjects to synthetic speech sounds constructed to approximate /ta/ and /da/. The animals were trained (that is, given appropriate feedback) on the two "endpoint" stimuli, VOT's of 0 and +80 msec; for all other stimuli (VOT's from +10 to +70 msec in 10-msec steps), feedback was arranged to indicate a correct response to the animal. The labeling gradients and "phonetic boundaries" for human and chinchilla subjects are similar.

animal's ability to correctly classify novel instances from a category without being "taught"; any change in the animal's performance is attributable to experience alone.

Performance on the control stimuli (Fig. 1, panels 9 to 11, closed circles) remained fairly steady throughout generalization sessions, and all generalization stimuli (open circles) were eventually classified as accurately as their controls. In fact, the continued addition of novel instances of /t/ and /d/ produced increasingly fewer errors during initual sessions (this can be seen by comparing the first open circles in panels 9, 10, and 11).

In experiment 2, two of the four animals in experiment 1 (one for whom /t/ was positive and one for whom /d/ was positive) and two new animals were trained on the "endpoints" of the synthetic continuum (VOT, 0 msec, /d/; VOT, +80 msec, /t/) until the performance of each animal was consistently greater than 96 percent correct. During the next ten sessions, these endpoint stimuli were presented in half the trials and the stimuli between them (VOT, +10 to +70 msec, in 10-msec steps) were randomly presented in the other half. During endpoint trials, all feedback was maintained, but when all other stimuli were presented, feedback was arranged to indicate a correct response. Four English-speaking adults identified the same stimuli as /da/ or /ta/. The mean percentage of "d" responses to each synthetic stimulus by human and chinchilla subjects is displayed in Fig. 2. The stimuli were labeled an equal number of times by both groups of subjects. The two curves, both of which were generated by a least-squares method, are similar both in general form and in the location of the phonetic boundaries; for English-speaking adults and chinchillas the boundaries are 35.2 and 33.5 msec, respectively. Furthermore, training on natural speech was not a necessary condition for placement of the boundary; the mean phonetic boundaries for the two animals originally trained on natural speech and the two that were trained only on the synthetic endpoints were 33.7 msec and 33.1 msec, respectively.

To examine the generality of the correspondence between the labeling functions of our animal and human subjects, we tested two of the four animals and the same four adults with the Lisker-Abramson labial and velar VOT series (*7*). The procedure was exactly the same; synthetic stimuli with VOT's of 0 and +80 msec were endpoints for which feedback was in effect while all other stimuli were presented as generalization stimuli. The similarity in the form of the labeling gradients and in the location of the boundaries demonstrated for the alveolar stimuli was again found for the labial and velar stimuli. The boundaries were about +25, +34, and +42 msec for the labial, alveolar, and velar stimuli, respectively, with standard errors of 1.75 msec; these values are in good agreement with those of Lisker and Abramson (*7*).

These experiments simply demonstrate that the voiced-voiceless distinction of plosive consonants in initial position, whether in natural speech where the critical features must be abstracted from many others, or in synthetic speech where only the critical features are varied, can be made by a mammal without phonetic mediating devices. While these findings do not rule out the possibility that human listeners process these speech sounds in a highly specialized way, they do demonstrate that uniquely human processing is not essential for these particular tasks, and they suggest that alternative explanations merit further consideration. The fact that the chinchillas respond to the synthetic speech as though an abrupt qualitative change occurs in the short voicing-lag region of the VOT continuum at precisely the place where many languages separate two phonemic categories lends support to the idea that speech-sound oppositions were selected to be highly distinctive to the auditory system. By this reasoning, one might infer that there is at least one other natural psychophysical boundary, located in the voicing-lead region of the VOT continuum, which serves as a basis for the phonemic distinction between prevoiced and voiceless-unaspirated plosives of languages such as Spanish, Thai, and Kikuyu. In any case, further experiments with animals should help to pinpoint which speech-perception tasks require "special processing."

PATRICIA K. KUHL
JAMES D. MILLER
Central Institute for the Deaf,
St. Louis, Missouri 63110

References and Notes

1. A. M. Liberman, F. S. Cooper, D. P. Shankweiler, M. Studdert-Kennedy, *Psychol. Rev.* **74**, 431 (1967). The plosive consonant /d/ is often used as an example. The direction of the second formant (F_2) transition is believed to be a sufficient cue for differentiating place of articulation (/b, d, g/); that is, in the /æ/ vowel context, rising F_2 transitions produce /bæ/, falling F_2 transitions produce /gæ/, and those remaining fairly steady produce /dæ/. However, the set of F_2 transitions that produce a single plosive, like /d/, across vowel contexts varies extensively; in /di/ the F_2 transition rises, in /du/ it falls, and in /dæ/ it remains fairly steady.
2. A. M. Liberman, *Cogn. Psychol.* **1**, 301 (1970).
3. K. N. Stevens and A. S. House, in *Foundations of Modern Auditory Theory*, J. V. Tobias, Ed. (Academic Press, New York, 1972), vol. 2, pp. 3 62; M. Studdert-Kennedy, A. M. Liberman, K. S. Harris, F. S. Cooper, *Psychol. Rev.* 77, 234 (1970).
4. P. D. Eimas and J. D. Corbit, *Cogn. Psychol.* **4**, 99 (1973).
5. J. D. Miller, *J. Acoust. Soc. Am.* **48**, 513 (1970). For differences between auditory capabilities of man and chinchilla, see p. 521.
6. L. Lisker and A. S. Abramson, *Word* **20**, 384 (1964).
7. ———, in *Proceedings of the Sixth International Congress of Phonetic Sciences Prague 1967* (Academia, Prague, 1970), pp. 563–567.
8. P. D. Eimas, E. R. Siqueland, P. Jusczyk, J. Vigorito, *Science* **171**, 303 (1971).
9. L. Streeter, thesis, Columbia University (1974).
10. At least two acoustic cues appear to contribute to the qualitative change in the perception of the synthetic stimuli as voicing increasingly lags the burst: the presence of aspiration and the absence of a rapid spectrum change at the onset of voicing [K. N. Stevens and D. H. Klatt, *J. Acoust. Soc. Am.* **55**, 653 (1974)].
11. B. F. Spenner, A. M. Engebretson, J. D. Miller, J. R. Cox, *ibid.*, p. 427(A).
12. C. K. Burdick and J. D. Miller, *ibid.* **54**, 789 (1973).
13. A new variation was introduced by adding only the positive stimuli, then only the negative stimuli, and finally both positive and negative stimuli until performance criteria were met [C. K. Burdick and J. D. Miller, *ibid.* **58**, 415 (1975)]. The data plotted in Fig. 1 are from sessions in which both positive and negative stimuli were presented.
14. Beginning with training on individual vowels, all positive and negative stimuli for a new condition were added simultaneously.
15. Supported by NIH grant NS03856 to Central Institute for the Deaf and NIH grant RR00396 to the Biomedical Computer Laboratory of Washington University. We thank A. M. Engebretson, C. K. Burdick, and R. J. Dooling for their assistance, and we acknowledge the cooperation of the Haskins Laboratories (NIH contract NIH-71-2420) in providing the synthetic stimuli.

1 January 1975; revised 22 April 1975