Perception & Psychophysics
2004, 66 (8), 1341–1350

Perception of place-of-articulation information in natural speech by monkeys versus humans

JOAN M. SINNOTT and CASEY S. GILMORE
University of South Alabama, Mobile, Alabama

Four monkeys and 6 humans representing five different native languages were compared in the ability to categorize natural CV tokens of /b/ versus /d/ produced by 4 talkers of American English (2 male, 2 female) in four vowel contexts (/i, e, a, u/). A two-choice "left/right" procedure was used in which both percentage correct and response time data were compared between species. Both measures indicated striking context effects for monkeys, in that they performed better for the back vowels /a/ and /u/ than for the front vowels /i/ and /e/. Humans showed no context effects for the percentage correct measure, but their response times showed an enhancement for the /i/ vowel, in contrast with monkeys. Results suggest that monkey perception of place of articulation is more dependent than human perception on the direction of the *F*2 onset transitions of syllables, since back-vowel *F*2s differentiate /b/ and /d/ more distinctively. Although monkeys do not provide an accurate model of the adult human in place perception, they may be able to model the preverbal human infant before it learns a more speech-specific strategy of place information extraction.

For the past 25 years, researchers have been interested in animal perception of speech sounds in order to assess the extent to which human perception is based on "general" versus "special" mechanisms (e.g., Kuhl, 1986; Trout, 2001). It is commonly acknowledged that "general" mechanisms are those inherited from ancestral primate, mammalian, or vertebrate psychoacoustic systems and that "special" mechanisms are those that do not arise from simple psychoacoustics but are in some way language-learning–specific, or possibly even species-specific, to humans.

It is likely that both general and special mechanisms contribute to human speech perception, depending upon the particular phenomenon studied. For example, there is at least one phonetic contrast for which animals appear to provide excellent models: English voice onset time (VOT). Specifically, if the focus is on perception of the English phoneme boundary along synthetic VOT continua spanning the English (positive) range, the comparative data strongly suggest the involvement of a generalized psychoacoustic mechanism for temporal order judgments (for a review, see Kuhl, 1986). Furthermore, this mechanism seems to be spontaneously available to all animals tested so far and to require no particular learning on the part of the animal. The comparative behavioral evidence is now so persuasive that auditory physiologists are fruitfully using animal models to infer what the human auditory system does in (English) VOT processing (e.g., Eggermont, 1995; Sinex & McDonald, 1988).

These VOT results are important because they indicate that a general mechanism hypothesis can be validly considered with respect to some important aspects of human speech perception. However, for other types of human speech phenomena, the relative extent of general versus special mechanisms involved is still very much an open question (see, e.g., Sinnott, 1998).

The Place-of-Articulation Feature and the "Invariance Problem"

The present goal is to further explore monkey versus human perception of the place-of-articulation contrast. This contrast continues to be of intense interest to speech researchers because, relative to other contrasts involving voicing or manner, "invariant" acoustic cues for perception of /b/–/d/–/g/ or /p/–/t/–/k/ across vowel contexts have been quite difficult to find (e.g., Dorman & Loizou, 1996; Jongman & Miller, 1991; Kewley-Port, 1983; Kobatake & Ohtani, 1987; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Nossair & Zahorian, 1991; Repp & Lin, 1989; Smits, ten Bosch, & Collier, 1996; Stevens & Blumstein, 1978; Sussman, McCaffrey, & Matthews, 1991).

In fact, some research has even put in question whether invariant cues actually exist in the physical structure of these consonants (Fowler, 1994). Despite the elusive acoustic nature of place contrasts, all human languages seem to contrast labial and coronal stop phonemes in some form, although the phonetics of the actual contrasts may

This research was supported by NIH Grant PHS R01 DC 00541-12 to J.M.S. C.S.G. was the recipient of a Carol P. Sinnott Graduate Student Research Award, and these data are based on his master's thesis. We thank Chris Gonzalez, Kelly Mosteller, and Laura McArdle Powell for producing the stimuli. We thank Ambrin Masood, Jazmin Camchong, Lakshmikar Kuravi, and Toshimasa Ishihara for participating as subjects. We thank Kelly Mosteller and Jazmin Camchong for assistance in testing monkeys. Send correspondence to J. M. Sinnott, Psychology Department, University or South Alabama, Mobile, AL 36699 (e-mail: jsinnott@jaguar1.usouthal.edu).

vary (Lahiri, Gewirth, & Blumstein, 1984; Sussman, Hoemeke, & Ahmed, 1993).

We believe that animal comparisons can shed a particularly interesting light on the invariance problem for place of articulation. If animals perceive this feature as humans do, the implications are that invariant cues are in fact hidden somewhere in the signal and are accessible via general auditory mechanisms, although perhaps our technology is not yet advanced enough to find them. In contrast, if animals do not perceive this feature as humans do, a special mechanism for place processing may exist only in humans (e.g., Liberman et al., 1967).

The latter result would open myriad additional questions, such as: Is the place extraction mechanism an innate genetic one that differentiates humans (even human infants) from animals, or is it rather a learned one that differentiates verbal from nonverbal subjects? To what extent does articulation enter the process? What do human infants perceive regarding place? Could animals learn to perceive place like adult humans, given the proper training? (These issues will be addressed in the Discussion section.)

We do not wish to speculate in the present article about the exact nature of the possible special human mechanism for place extraction hypothesized by Liberman et al. (1967). Our present goal is instead data oriented: When aspects of place perception are studied comparatively, is there evidence that animals use a *similar* or a *different* mechanism when compared with adult humans? So far, five studies have examined various aspects of place perception in animals. Of these, three conclude in favor of similar and two in favor of different mechanisms, as discussed below. For the studies that tested human controls, all used American-English listeners.

The first of these studies, Sinnott, Beecher, Moody, and Stebbins (1976), examined discrimination on a synthetic three-formant /ba/–/da/ continuum using a repeating-standard procedure. The researchers compared monkey and human difference limens (DLs) and response times (RTs) measured from each endpoint of the continuum. The basic DL measure (based on 50% correct hits) showed only minor quantitative differences, suggesting that the species utilize similar psychoacoustic mechanisms. However, the RT measure found much longer RTs for humans in comparison with monkeys for within-category discriminations, a finding reminiscent of a "categorical perception" effect operating for humans, and thus of different mechanisms in the two species.

Second, Kuhl and Padden (1983) used a categorical perception paradigm to compare monkey and human discrimination of pairwise comparisons on a two-formant /bae/–/dae/–/gae/ synthetic continuum. Using a percentage correct measure, they found similar enhanced discrimination for both species at both the /b/–/d/ and /d/–/g/ boundaries, implying a similar mechanism for perception of phoneme boundary effects along this particular continuum.

Third, Kluender, Diehl, and Killeen (1987) examined quail perception of place in natural speech. The stimuli were consonant–vowel (CV) + /s/ tokens produced by one male talker using different vowels. The birds were trained to peck to syllables beginning with /d/ and to refrain from pecking to syllables beginning with /b/ or /g/. Results showed that the birds learned to classify the /d/ syllables correctly using the training vowels /i,æ, a, u/ and also generalized to several new vowels and diphthongs. Thus, the authors concluded that the birds formed a "phonetic category" for /d/ in natural speech, implying a similar mechanism in both quail and humans.

Fourth, Lotto, Kluender, and Holt (1997) studied Japanese quail "perceptual compensation for coarticulation." Their stimuli consisted of a synthetic /da/–/ga/ continuum preceded by /al/ versus /ar/. The birds' pecking rates indicated that they perceived ambiguous stimuli in the center of the continuum as /d/ when stimuli were preceded by /r/ but as /g/ when they were preceded by /l/. The authors attributed the effect to a similar mechanism of "spectral contrast" in both quail and humans.

Finally, Sinnott and Williamson (1999) used macaque monkeys to conduct the most recent study of animal place perception. The stimuli were synthetic CVs consisting of both place contrast /b/–/d/ and control manner contrast /z/–/d/. The monkeys were taught to categorize /ba/–/da/ (or /za/–/da/) using a two-choice procedure and were then tested for transfer to new vowel contexts /i, e, o/. The monkeys had no trouble learning the manner contrast, which obviously contained a highly salient invariant acoustic cue in the form of the strident frication of /z/. More important, the monkeys all easily generalized the manner contrast to all of the new vowels. Thus, these results provided no evidence that the monkeys were using a different mechanism than humans do in categorizing and generalizing the manner contrast.

However, the same study found that the monkeys all had various problems in both learning to distinguish and generalizing the place contrast. After learning to distinguish the /b/ and /d/ stimuli in combination with /a/, the monkeys generalized well to combinations with /o/, but not to those with either /e/ or /i/. The authors attributed the problems to the lack of a strident burst cue in the stimuli to signal /b/ versus /d/, since the two phonemes were synthesized using formant transitions only. Nevertheless, since the control human listeners had no trouble identifying these "burstless" stimuli, the overall results suggested different mechanisms in monkeys and humans for extracting place information from formant transitions, in contrast to the results for the manner information.

To summarize, the comparative data for various aspects of place perception indicate both similar (Kluender et al., 1987; Kuhl & Padden, 1983; Lotto et al., 1997) and different (Sinnott et al., 1976; Sinnott & Williamson, 1999) mechanisms in humans and animals. Two of these studies in particular appear to contradict each other: The monkey results of Sinnott and Williamson (1999) seem different from the quail results of Kluender et al. (1987), because the monkeys did not perceive place contrast in an invariant manner across different vowel contexts, but the quail presumably did.

However, a number of potential reasons could explain the different results of the two studies.

(1) *Different stimuli*. The rich natural speech stimuli used in the quail study could result in reduced vowel context effects for all animals, in contrast with the more sterile synthetic speech stimuli used in the monkey study.

(2) *Different training procedures*. The quail were trained right from the start in multiple vowel contexts, whereas the monkeys were trained in only one vowel context, /a/.

(3) *Different response measures*. The quail pecking rate measure may not be as precise a measure as percentage correct to tap potential context-sensitive place perception in animals.

(4) *Different subjects*. Quail could be more like humans in place perception than monkeys, since it is commonly acknowledged that some birds have extremely sophisticated vocal-auditory communication systems (e.g., Kroodsma, Miller, & Ouellet, 1982).

Purpose of the Present Study

Therefore, the purpose of the present study is to combine some positive aspects of both the Kluender et al. (1987) quail study and the Sinnott and Williamson (1999) monkey study, in order to compare more precisely the extent to which vowel context–sensitive place categorization occurs in animals versus humans and (it is hoped) to reconcile the seemingly divergent results of the two studies. Specifically, we propose to derive stimuli from rich natural speech and to expose monkeys to multiple vowel contexts right from the start, in line with the Kluender study, but to present these stimuli using a more precise procedure that can be used with both monkeys and humans, just as the Sinnott study did.

If no qualitative differences in monkey versus human categorization were to emerge as a function of vowel context, this result would provide evidence for similar mechanisms of place perception in humans and monkeys. However, if qualitative differences were to emerge, this result would suggest that human and monkeys use different mechanisms. We leave open the possibility that quantitative differences in overall sensitivity might occur between monkeys and humans, because speech stimuli are highly overlearned for humans, but we argue that such differences alone, in the absence of qualitative differences, would be irrelevant to the "special" versus "general" mechanisms debate.

The present study uses two different measures to compare monkey and human performance. First, the traditional measure, percentage correct (PC), is analyzed in order to determine if monkeys are capable of forming place categories with a level of accuracy similar to humans. Although we expect that humans will perform at PC levels near 100%, how high monkey PC scores will go is an open question. A basic PC measure should give us additional information that is not available from a study such as Kluender et al. (1987), in which birds' peck rates to various stimuli were measured, but it will be difficult, if not impossible, to equate peck rates with a PC measure.

Second, a more molecular measure, RT, is analyzed in order to compare the relative efficiency with which the monkeys and humans spontaneously categorize place contrasts for the different vowels. We proceed from the general assumption that RT reflects the efficiency with which the central nervous system performs the task required. Faster RTs indicate an easy, automatic categorization, whereas slower RTs indicate a harder, less automatic categorization. Note that of the comparative "phoneme boundary" studies reviewed in the introduction, the one that used an RT measure (Sinnott et al., 1976) uncovered differences between animal and human response modes, whereas those that did not use an RT measure (Kuhl & Padden, 1983; Lotto et al., 1997) did not. Therefore, we propose that the RT measure gives an added "window" on response processes not tapped by the PC measure. In addition, another important reason to use RT data in comparing animal with human performance is that PC data for humans are often at ceiling levels, and thus the ceiling effect could mask certain perceptual strategies operating at a more molecular level.

METHOD

Subjects

The monkey listeners were 4 male Japanese macaques (*Macaca fuscata*). Two (Dart and Harry) were 15 years old, and over the course of 12 years in the lab they had participated in a variety of other speech perception studies, including Sinnott and Williamson (1999). Two younger monkeys (Bongo and Jocko) were 3 years old and had previously participated in one speech perception study using synthetic VOT stimuli, but they were completely naive to place stimuli. All monkeys had normal hearing, as measured by hearing tests. All animal housing and testing procedures were approved by the University of South Alabama Institutional Animal Care and Use Committee.

The 6 human participants were: 2 American English (AE) listeners (authors J.M.S. [female, age 53] & C.S.G. [male, age 29]), 1 Urdu listener from Pakistan (A.M.B., female, age 35), 1 Spanish listener from Ecuador (J.A.Z., female, age 28), 1 Hindi listener from India (L.A.K., male, age 23), and 1 Japanese listener (T.O.S., male, age 53). None reported any speech or hearing problems. A.M.B., J.A.Z., L.A.K., and T.O.S. were all international students who worked in the lab, and all spoke English as their second language. The rationale behind including them was to ensure that any potential monkey/human differences that occurred would not be attributable to some peculiarity of AE listeners only.

Apparatus

The apparatus has been previously described (Sinnott & Williamson, 1999). Briefly, test sessions were conducted free-field in a double-walled IAC booth lined with sound absorbing material. Audio signals were presented through a Genesis loudspeaker positioned in a corner of the booth, approximately 84 cm from the listener's head. Stimulus presentation, experimental contingencies, and response recording were controlled by a Dell computer and TDT equipment.

During testing, a monkey sat in a primate restraint chair and responded by contacting a metal lever mounted on the chair and moving it left or right. A cuelight attached above the lever signaled the

onset of a trial. A food cup was also attached to the mounting, allowing the monkey to obtain 190-mg banana reward pellets. The humans were tested while seated in an ordinary chair and used a lever and cuelight apparatus, similar to that for the monkeys, mounted on a stand in front of them.

Stimuli

Four AE talkers (females J.M.S. and L.A.M. and males C.L.G. and K.W.M.) each recorded two tokens of eight CV syllables /bu, ba, be, bi; du, da, de, di/ in a room with low noise (<30 dB SPLA) using an AKG B18 microphone. The CVs were digitized, analyzed and edited using CSRE (Canadian Speech Research Environment, AVAAZ Innovations, ON). All four talkers listened to all the stimuli, and all agreed that the stimuli were good clear instances of the phonemes. The final set of stimuli thus consisted of 64 CVs (4 talkers × 8 syllables × 2 tokens) that were organized into eight different stimulus set files (SSFs), each containing eight stimuli. The eight SSFs were: CLG1, CLG2, JMS1, JMS2, KWM1, KWM2, LAM1, LAM2. During testing, stimuli were presented at a normal conversational level of approximately 65 dB SPLA (about 55 dB SL), as calibrated by a B&K SPL meter placed in the position of the listener's head.

Procedure

Two-choice "left/right" procedure. The go-left/go-right procedure used was identical to that of Sinnott and Williamson (1999). At the start of a trial, the listener manually contacted the metal lever positioned below the flashing cuelight. Upon contact, the cuelight turned on and a CV started to repeat (1 per sec) until a response was made. A correct response was to move the lever to the left for /b/ CVs and to the right for /d/ CVs. Stimuli requiring left versus right responses were randomly presented on each trial. A correct response was immediately followed by a 2-kHz 100-msec tone pip as feedback and an intertrial interval (ITI). An incorrect response was followed by a 5-sec timeout during which the light extinguished and a 300-Hz tone sounded. After each incorrect response, a correction procedure took effect in which the missed stimulus was repeated on successive trials until a correct response was made. Correction procedure trials occurring after the initial miss were not counted in the data analysis.

Each daily test session terminated after 120 correct responses, resulting in about 15 trials per CV. The procedure was identical for monkeys and humans, except that the monkeys received a food pellet after each correct response (in addition to the 2-kHz tone pip). Also, the monkey ITI was set at 5 sec to allow time to eat the pellet, but the human ITI was only 1 sec.

Monkey testing procedure. The testing sequence for the monkeys for the entire experiment was as follows: The monkeys were introduced to each SSF using a "fading" procedure in which a highly salient intensity cue (−10 dB) initially reduced the level of the left-side /b/ CVs relative to the right-side /d/ CVs. Thus, the monkeys were initially "instructed" to "go left" or "go right" by means of this cue. When the monkey reached approximately 90% correct identifications for both the left- and right-side stimuli, the cue was reduced by 1 dB per session until both left- and right-side stimuli were at equal intensity. Final data for monkeys were obtained by presenting 10 sessions with stimuli at equal intensity for each SSF, and then averaging over the last 5 sessions. In general, a monkey worked for about 1 month (30 daily test sessions) on each SSF. Thus, it required about eight months for a monkey to go through all eight SSFs in the following order: CLG1, JMS1, KWM1, LAM1, CLG2, JMS2, KWM2, LAM2.

Human testing procedure. The testing sequence for the humans was as follows: First, a practice session was conducted in which the listener was introduced to the left/right procedure, using SSF CLG1. The listener was verbally instructed to "go left" for /b/ CVs and to "go right" for /d/ CVs. No human had any problems with the practice session, and all performed at >95% correct for all stimuli. Actual testing started after the initial practice session. A different SSF (repeating CLG1) was tested on each day, using a different random order for each listener. Thus, it took only 8 testing days for a human to go through all eight SSFs.

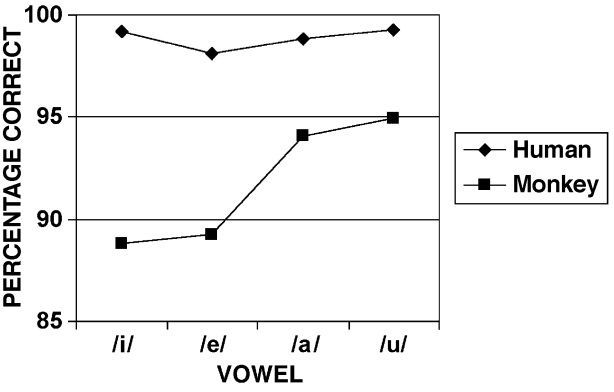

Figure 1. Mean percentage correct data for humans and monkeys as a function of vowel context, showing reduced accuracy for monkeys with the front vowels of /i/ and /e/.

RESULTS

Percentage Correct Data

The first question of interest in this study is: What levels of accuracy can monkeys attain in categorizing /b/–/d/ contrasts based on natural speech? Our PC data averaged over all SSFs indicate that the monkeys were quite accurate, but that they did not reach the ceiling levels attained by the humans. Figure 1 shows the PC data compared for humans versus monkeys. The humans averaged >98% correct for all vowel contexts. The monkeys averaged 89%–95% correct and also showed the first indications of performance differences as a function of vowel context, performing better with the two back vowels /a/ and /u/ than with the two front vowels /i/ and /e/.

To analyze the PC data, we used the arcsine transformation, because normality assumptions may not hold at high PC levels. Table 1 shows both the raw and the transformed PC data for the humans and the monkeys. A 2 (species: monkey, human) × 4 (vowels: /i, e, a, u/) analysis of variance (ANOVA) using the transformed PC data revealed a significant main effect of species [$F(1,8) = 84.749$, $p = .000$], indicating that monkeys were overall less accurate than humans. The main effect of vowel was also significant [$F(3,24) = 6.293$, $p = .003$]. Of primary

Table 1
Human and Monkey Raw Percentage Correct and Arcsine-Transformed Scores

	Raw PC				Arc PC			
	Humans		Monkeys		Humans		Monkeys	
Vowel	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
/i/	99.2	0.7	88.8	3.8	1.47	0.08	1.10	0.08
/e/	98.1	1.3	89.2	1.4	1.39	0.07	1.10	0.03
/a/	98.9	1.0	94.1	1.7	1.44	0.08	1.23	0.05
/u/	99.3	0.8	94.9	2.0	1.48	0.08	1.26	0.06

interest, the interaction was significant [$F(3,24) = 3.259$, $p = .039$], verifying that the monkeys, unlike the humans, did not categorize with equal accuracy across the four vowel contexts.

To further analyze the monkey data, an additional ANOVA on the transformed PC data compared the two older, more experienced monkeys (Dart, Harry) with the two younger monkeys (Bongo, Jocko). There was no effect of age [$F(1,2) = 0.318$, $p = .629$], a significant effect of vowel [$F(3,6) = 8.116$, $p = .016$], and no interaction [$F(3,6) = 0.330$, $p = .805$], verifying that vowel context effects were present for all the monkey listeners.

To further analyze the human data, an additional ANOVA on the transformed PC data compared the two AE listeners (C.S.G., J.M.S.) with the four non-AE listeners (A.M.B., J.A.Z., L.A.K., T.O.S.). There was no effect of AE [$F(1,4) = 0.081$, $p = .790$], no effect of vowel [$F(3,12) = 1.303$, $p = .319$], and no interaction [$F(3,12) = 0.333$, $p = .801$], verifying that vowel context effects were absent for all the human listeners.

Response Time Data

The second question of interest in this study is: How does vowel context affect monkey versus human performance in categorizing /b/–/d/ contrasts? Because of the ceiling effects that operate in the human PC data, this question is more legitimately answered by using the RT measure. RTs were analyzed for correct responses only.

Figure 2 shows the RT data for the humans and the monkeys for the four vowel contexts. The human RT functions were basically flat across /e/, /a/, and /u/, but slightly faster for the /i/ context. In contrast, the monkey RTs were faster for the two back vowels /a/ and /u/ than for the two front vowels /i/ and /e/, thus paralleling the above monkey PC data.

For analysis purposes, the raw RT data were converted to logs, in order to equate proportional changes in RT for subjects with overall different response speeds. (For example, some of the older subjects had longer RTs than did the younger ones.) Table 2 shows the raw and log-transformed RT data.

Table 2
Human and Monkey Raw and Response Times (RT, in Milliseconds) and Log-Transformed Scores

	Raw PC				Log RT			
	Humans		Monkeys		Humans		Monkeys	
Vowel	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
/i/	579	133	1,032	534	2.754	0.098	2.971	0.221
/e/	620	127	1,025	514	2.785	0.087	2.969	0.221
/a/	616	119	815	362	2.783	0.084	2.879	0.191
/u/	637	126	827	369	2.797	0.085	2.884	0.196

A 2 × 4 ANOVA on the log RTs revealed no significant main effect of species [$F(1,8) = 2.488$, $p = .153$]. There was a significant main effect of vowel [$F(3,24) = 6.630$, $p = .002$]. Of primary interest, there was a significant species × vowel interaction [$F(3,24) = 15.375$, $p = .000$], verifying that the monkey pattern of RT diverged from the human pattern as a function of the four vowel contexts.

To further analyze the monkey RT data, an additional ANOVA on the log RT data compared the two older monkeys with the two younger ones. There was a significant main effect of age [$F(1,2) = 20.392$, $p = .046$], indicating that the younger monkeys had overall faster RTs than the older ones. There was a significant effect of vowel [$F(3,6) = 4.955$, $p = .046$]. Of primary interest, there was no interaction [$F(3,6) = 0.214$, $p = .883$], verifying that vowel context effects were similar for all monkeys.

To further analyze the human data, an additional ANOVA on the log RT data compared the two AE listeners with the four non-AE listeners. There was no effect of AE [$F(1,4) = 1.181$, $p = .338$], but a significant effect of vowel [$F(3,12) = 36.092$, $p = .000$], reflected faster RTs with the vowel /i/. There was no interaction [$F(3,12) = 2.089$, $p = .155$], verifying that the /i/ vowel context effect was similar in all human listeners.

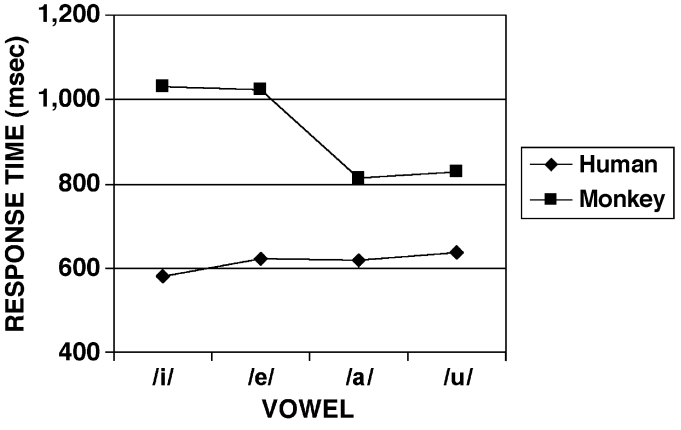

Figure 2. Mean response time data for humans and monkeys as a function of vowel context, showing increased RTs for monkeys with the front vowels of /i/ and /e/.

DISCUSSION

Percentage Correct Analysis

Our PC data indicate that monkeys can be trained to quite high levels of accuracy in categorizing /b/–/d/ contrasts produced by four different talkers using four different vowels. Although they did not reach the ceiling levels of the humans, they did reach approximately 85% correct or higher for most individual stimuli tested. It is possible that they might have gotten more accurate with continued training, but note that the two "experienced" monkeys (Dart and Harry, subjects in Sinnott & Williamson, 1999) were no more accurate than the two "inexperienced" monkeys (Bongo and Jocko, who were totally naive to place testing). This comparison suggests that the present study probably brought the monkeys to their asymptotic levels of PC for these particular stimuli. The PC data also give the first indication of vowel context effects in monkeys, since they performed less accurately on front than on back vowel contexts.

The human subjects all operated at ceiling levels in the PC measure. All reported that their rare mistakes were not perceptual ones, but simple hand-motor errors, since the task was quite boring and repetitious. The 4 non-AE listeners (Japanese, Hindi, Spanish, Urdu) performed as accurately as the two AE listeners, even though all the stimuli were articulated by AE talkers. The /b/–/d/ contrast is essentially universal among human languages, although the actual phonetics of coronal /d/s vary. For example, Hindi and Urdu use dental and retroflex stops rather than the alveolar English stop (e.g., Werker, 1994). Despite the different phonetics, our Hindi and Urdu listeners seemed to "perceptually assimilate" the AE /d/s into their native categories, such that no deficits in perception emerged (e.g., Best, 1994). In fact, it would probably be very hard, if not impossible, to find an adult human from any language who would not perform as well as the present subjects, given the ubiquitous nature of the /b/–/d/ contrast in human languages and the ability of humans to perceptually assimilate it.

RT Data Analysis

The most important aspect of the RT data is that they yield an exact measure of vowel context effect that is not clouded by ceiling effects in human performance. The vowel context effect was very striking for the monkeys, who exhibited faster RTs for back /a/ and /u/ than for front /i/ and /e/. This RT effect parallels the monkey PC data by indicating relatively more problematic perception for front vowel contexts. The vowel context effect for the humans was different from that of the monkeys: All humans from all the different languages showed faster RTs for the /i/ vowel context. The human RT data are therefore a further testimonial to the efficiency with which "perceptual assimilation" operates among fellow humans listening to nonnative talkers articulating similar, if not identical, phonemes (Best, 1994).

Comparing the Present Monkey Data and the Kluender Quail Data

As discussed above, Kluender et al. (1987) reported no problems by quail in generalizing from one vowel to another. The authors inferred from this result that their birds were using a mechanism similar to the human one for categorizing place. How can this quail result be reconciled with the present results indicating that our monkeys did not use a humanlike mechanism? The fact that both studies used natural speech seems to rule out potential differences in the stimuli, although this factor cannot be ruled out with absolute certainty. Another possibility is that quail (being acoustically "sophisticated" birds) are more like humans in place perception than are monkeys.

A final possibility is that the Kluender peck rate procedure was not precise enough to tap potential vowel context effects in quail perception. This possibility prompted us to reanalyze the Kluender quail data for context effects, which we did by transforming the birds' absolute peck rates into /d/ : /b/ "peck ratios" for each vowel and plotting these on a log scale for normalization purposes. The results are shown in Figure 3, in which vowels on the *x*-axis are ordered from front to back, according to decreasing height of *F*2. Diphthongs are plotted relative to the initial vowel, which would determine the shape of the initial *F*2 onset transition (see below).

Note that the peck rate ratios tend to be higher, indicating better categorization, for back than for front vowels. Thus, these reanalyzed data suggest that the quail were in fact differentiating between front and back vowels in their pecking rates to novel stimuli. If this particular analysis is correct, then the quail might in fact be qualitatively similar to the present monkeys in place perception.

Monkey Versus Human Perception of Place

Why would back vowels be "favored" by animals in place perception? One answer may lie in the pattern of the *F*2 onset transition, as first documented by Liberman, Delattre, and Cooper (1955). Their classic figure is partially replotted in Figure 4. For example, while *F*2 always rises for /b/, independent of vowel context, the direction of the /d/ *F*2 transition is variable. Because back vowels (e.g., /a, o, u/) have low *F*2s, *F*2 normally decreases from the /d/ locus at about 1,800 Hz to the start of the vowel. However, because front vowels (e.g., /i, e/) have high *F*2s, *F*2 may either stay flat or increase from the /d/ locus to the start of the vowel. The upshot is that there is a much clearer acoustic cue differentiating /b/ and /d/ for back vowels (rising vs. falling *F*2), but not for the front vowels (both rising *F*2s).

These classic patterns were later verified by Kewley-Port (1982) in extensive analyses of natural speech, and also appeared in analyses of our own stimuli. For example, Figure 5 shows four CVs from talker K.W.M. analyzed using both the spectrogram and the formant tracking option of CSRE. Note how /bi/ and /di/ both have similar rising *F*2 transitions, while /ba/ and /da/ have differently

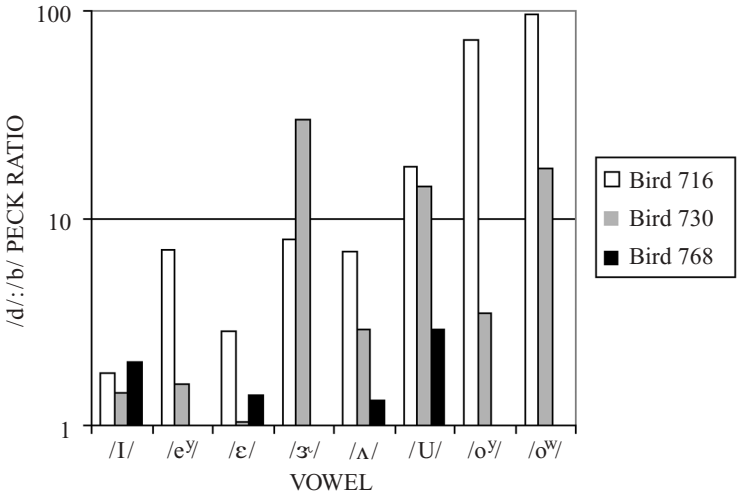

Figure 3. Reanalysis of the quail peck rate categorization data from Kluender et al. (1987) showing higher /d/:/b/ peck ratios for back vowels /U, oʸ, oʷ/ in comparison with front vowels /ɪ, eʸ, ɛ/. Diphthongs are ordered relative to the *F*2 of the initial vowel, which would determine the pattern of *F*2 onset transition.

shaped *F*2 transitions: slightly rising for /ba/ and clearly falling for /da/.

Thus, one hypothesis to explain our monkey data is that they focused on directional differences in the *F*2 onset transition in categorizing /b/ versus /d/. This strategy worked well with back vowels, but not so well with front vowels. However, our humans did not follow this pattern. In fact, their RT data actually showed an enhancement for the front vowel /i/, which has a similar rising *F*2 for both /bi/ and /di/. The present RT data therefore imply that our monkeys and humans did not use the same mechanism in categorizing /b/–/d/ contrasts.

Our results would be consistent with some kind of motor theory (for humans) proposing that common mediating articulations somehow link together the different percepts for /b/, /d/, and /g/ with variable vowels (Liberman et al., 1967). Motor theory could not account for the RT enhancement for /i/ seen in our data, however.

What do other theories of human speech perception concerned with the "invariance problem" have to say about human versus animal perception of place? Both Stevens and Blumstein (1978) and Kewley-Port (1983) have proposed that invariant cues in the initial short-term onset spectra can differentiate /b/ versus /d/ across vowel contexts: /b/ can be matched to a "diffuse-falling" template with relatively more low-frequency emphasis, and /d/ can be matched to a "diffuse-rising" template with more high-frequency emphasis. Although these authors make no

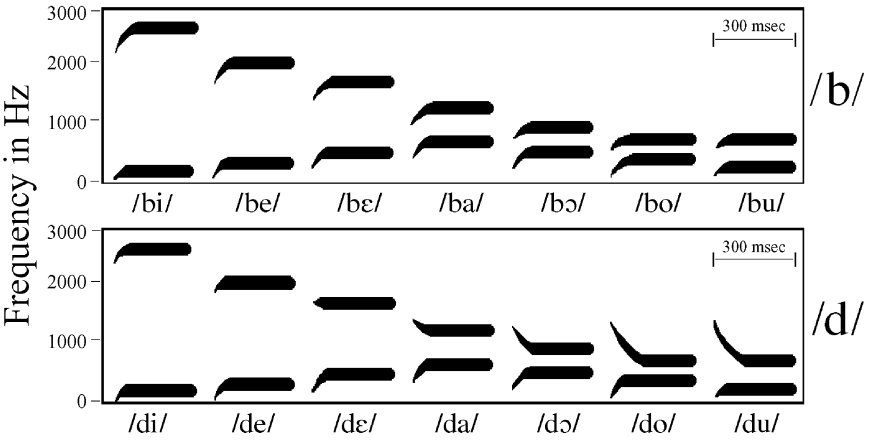

Figure 4. Schematized spectrograms depicting the *F*1 and *F*2 onset transitions for /b/ versus /d/ in the context of several different vowels. From "Acoustic Loci and Transitional Cues for Consonants," by A. Liberman, C. Delattre, and F. Cooper, 1955, *Journal of the Acoustical Society of America*, *27*, 769-773. Copyright 1955 by Acoustical Society of America. Adapted with permission.

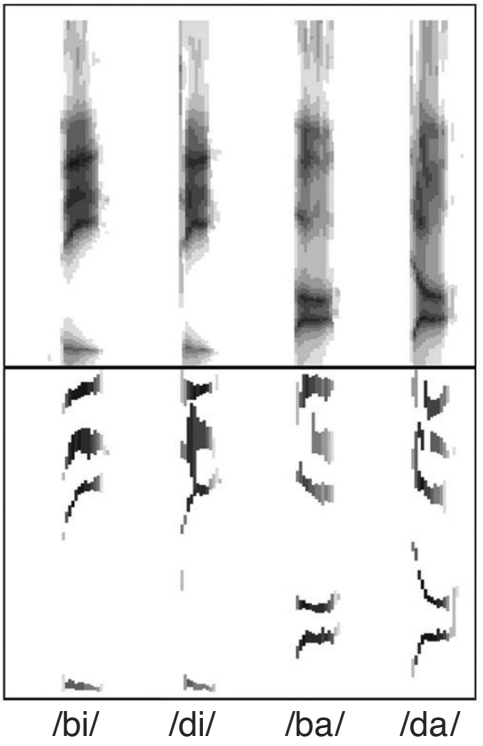

Figure 5. Spectrograms and formant plots of four stimulus tokens by talker K.W.M., analyzed using CSRE. The /bi/ **and** /di/ **tokens both show similar rising *F*2 transitions. The** /ba/ **and** /da/ **tokens show different patterns of the *F*2 transition.**

explicit claims about whether a monkey could use the templates, templates do seem to be based on purely psychoacoustic principles with no mediating articulation. Thus, it seems that monkeys should be able to use the templates to extract place cues as well as humans do, and thus that onset spectra theories should not predict differences in human versus animal perception of place such as those we found in the present study.

The most recent attempt to solve the invariance problem has been put forth by Sussman et al. (1991). Their "locus equations" involve comparing the onset frequency of the *F*2 transition with its steady-state frequency at vowel midpoint. When enough of these are plotted for different vowel contexts, a linear relationship emerges that is quite distinctive for /b/ versus /d/ versus /g/. The equations for different places are presumably stored in memory and referred to when extracting place information. The authors make no explicit claims about whether a monkey could use these equations, but again their method seems to be strictly auditorily based, with no adjunct mechanism of articulation, possibly implying once again that monkeys could have access.

However, one aspect of locus equation theory is very intriguing with regard to the present data, because it implies that our monkeys might be using such equations, but not our humans. The locus equations for /b/ versus /d/ are more distinct for back vowels and tend to converge for those in the front, implying that some perceptual confusion might result for front vowels (Sussman et al., 1991, p. 1322, Figure 7). In this sense, our monkey data are quite congruent with locus equation theory. On the other hand, common sense would argue against a monkey using locus equations, because they are complicated both to implement neurophysiologically (Sussman et al., 1991, p. 1324, Figure 9) and to learn (Deng & Braam, 1994). Thus, why would a monkey bother to use them, when it is so much simpler to attend to the absolute direction of the *F*2 transition?

Also note that our human data are not congruent with locus equation theory. Our human RT data indicate that the most fronted vowel /i/ is the "speediest" vowel context, despite the fact that it should be the most confused according to locus equation theory. To summarize, both our monkey and human data reveal problems with locus equation theory.

Can a Monkey Model a Human Infant's Perception of Place?

If we accept that monkeys are different from (adult) humans in place perception, it is of interest to ask if they might be more similar to human infants or children. If there were developmental data to show that young humans have more difficulty differentiating /b/ and /d/ with front than with back vowels, this finding would suggest that younger, less experienced humans also use a strategy of weighting formant transition direction to differentiate /b/–/d/ contrasts, just as monkeys do. Such a finding would also bode well for using the monkey as a model of the human infant before it tunes into a more speech-relevant, context-free adult mode of perceiving place contrasts.

Several recent developmental studies offer suggestive evidence with regard to this question. Bertoncini, Bijeljac-Babic, Blumstein, and Mehler (1987) found that newborn infants tested with a sucking habituation procedure and short synthetic stimuli showed "marginally" less ability to differentiate /bi/ and /di/ in comparison with /ba/ and /da/. Also, Ohde and colleagues (Ohde & Haley, 1997; Ohde, Haley, Vorperian, & McMahon, 1995) reported a series of studies comparing 3- to 11-year-old children with adults in /b/–/d/–/g/ perception using various types of synthetic stimuli with different vowels. They found that the children had some major problems with /b/ identification in the /i/ vowel context. However, since the adults tested as controls tended to exhibit the same types of errors, it is not clear if the children were exhibiting behavior that was qualitatively, or simply quantitatively, different from the adults.

Finally, Eimas (1999) made some very intriguing observations on 3- to 4-month-old infant perception of natural /b/–/d/ contrasts using a visual habituation–type "categorization" procedure. Here, the infant was presented with a series of repetitive /bV/ syllables with variable vowels. After adaptation, the infant was presented with a series of /dV/ syllables. Presumably, if the infant had perceived the /b/ category in the initial series, it should show an orienting response specifically to the /d/ syllables in the second series. However, the data showed that the infants displayed no tendency to orient specifically to the novel /d/ stimuli following /b/ adaptation. These data

might indicate that infants do not perceive the /b/ and /d/ phoneme categories as adults do.

At present, it thus does not appear possible to answer the question of whether a preverbal infant might use a monkeylike context-sensitive strategy in place perception.

Can a Monkey Be Trained to Model Human Adult Perception of Place?

Assume for the moment that the preverbal human infant does initially show a generalized monkeylike mode of /b/–/d/ perception that involves attending to *F*2 transition directional differences. As speech develops, this mode is then replaced at some point in time by the adult mode. In this case, we might ask: Could a monkey also be trained somehow to an adult mode of place perception? For example, new animal speech research has shown that certain songbirds (specifically, starlings) exposed to the proper experiential stimulus input can learn to perceive vowels as either English or Swedish adult listeners do (Kluender, Lotto, Holt, & Bloedel, 1998).

Note, however, that the present study has already presented monkeys with a wide selection of natural tokens of /b/ and /d/ produced by different talkers with different vowels and has given them extensive and consistent feedback with regard to correct or incorrect categorizations. Consider as well our two "experienced" monkeys Dart and Harry: Both have been in the lab for about 12 years and have had much "informal" exposure to human speech from many different male and female human technicians. In fact, Dart and Harry have been listening to conversational AE speech longer than the international students tested in the present study. In addition, both Dart and Harry were subjects in Sinnott and Williamson (1999), where they received approximately one year of more "formal" place training using synthetic tokens of /b/ and /d/ with both front and back vowels. Despite all of this informal and formal exposure to /b/ versus /d/, they did not attain humanlike place perception in the present study.

Could our monkey training regime be revised to differentially reinforce responses for front versus back vowel contexts? Perhaps a monkey could be given two banana pellets instead of one when he makes a correct response to /b/ versus /d/ with a front vowel. Or perhaps the monkey RTs could be differentially reinforced so that the monkey receives two pellets instead of one if he makes a faster RT to /b/ versus /d/ with a front vowel. Such manipulations might be possible, but differential reinforcement of different vowel contexts is certainly not the way the human infant develops a context-free mode of place perception. Even if a monkey could be taught to model the human adult perception of place via a complicated method of training, most likely the mechanism engaged would not be comparable to that which emerges naturally in the human infant.

CONCLUSION

If animal perception of place of articulation is linked to the directional pattern of *F*2 transitions, then human (adult) perception does not appear to be, at least to the extent that our measures were able to tap. Nevertheless, it is probably safe to say that both "general" (similar) and "special" (different) mechanisms are involved in human place perception, depending upon the level of analysis invoked.

At a molar level, general mechanisms are certainly involved, simply because the ability to form place categories is within the perceptual capacities of nonhumans. Nowhere has this fact been more apparent than in the present study, where monkeys attained levels of 90% correct categorization of /b/ versus /d/, averaged over all talkers and vowel contexts. However, on a more molecular level of analysis, a different mechanism appears to enter the human perceptual mode at some point, allowing a human to encode a place contrast in a more context-free manner, which would be of obvious advantage in processing running speech.

Of course, we also leave open the possibility that creative methods of training may be able to induce an adult-human-like "special" mode of place perception into animal subjects.

REFERENCES

Bertoncini, J., Bijeljac-Babic, R., Blumstein, S., & Mehler, J. (1987). Discrimination in neonates of very short CVs. *Journal of the Acoustical Society of America*, **82**, 31-37.

Best, C. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. Goodman & H. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167-223). Cambridge, MA: MIT Press.

Deng, L., & Braam, D. (1994). Context-dependent Markov model structured by locus equations: Applications to phonetic classification. *Journal of the Acoustical Society of America*, **96**, 2008-2025.

Dorman, M., & Loizou, P. (1996). Relative spectral change and formant transitions as cues to labial and alveolar place of articulation. *Journal of the Acoustical Society of America*, **100**, 3825-3830.

Eggermont, J. (1995). Representation of a voice onset time continuum in primary auditory cortex of the cat. *Journal of the Acoustical Society of America*, **98**, 911-920.

Eimas, P. (1999). Segmental and syllable representations in the perception of speech by young infants. *Journal of the Acoustical Society of America*, **105**, 1901-1911.

Fowler, C. A. (1994). Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. *Perception & Psychophysics*, **55**, 597-610.

Jongman, A., & Miller, J. D. (1991). Method for the location of burst-onset spectra in the auditory-perceptual space: A study of place of articulation in voiceless stop consonants. *Journal of the Acoustical Society of America*, **89**, 867-873.

Kewley-Port, D. (1982). Measurement of formant transitions in naturally produced stop consonant-vowel syllables. *Journal of the Acoustical Society of America*, **72**, 379-389.

Kewley-Port, D. (1983). Time varying features as correlates of place-of-articulation in stop consonants. *Journal of the Acoustical Society of America*, **73**, 322-334.

Kluender, K., Diehl, R., & Killeen, P. (1987). Japanese quail can learn phonetic categories. *Science*, **237**, 1195-1197.

Kluender, K., Lotto, A., Holt, L., & Bloedel, S. (1998). Role of experience for language-specific mappings of vowel sounds. *Journal of the Acoustical Society of America*, **104**, 3568-3582.

Kobatake, H., & Ohtani, S. (1987). Spectral transition dynamics of voiceless stop consonants. *Journal of the Acoustical Society of America*, **81**, 1146-1151.

Kroodsma, D., Miller, E., & Ouellet, H. (1982). *Acoustic commu-*

nication in birds: Vol. 1. Production, perception, and design features of sounds. New York: Academic Press.

KUHL, P. (1986). Theoretical contributions of tests on animals to the special-mechanisms debate in speech. *Experimental Biology*, **45**, 233-265.

KUHL, P., & PADDEN, D. (1983). Enhanced discrimination at the phoneme boundary for the place feature in macaques. *Journal of the Acoustical Society of America*, **73**, 1003-1010.

LAHIRI, A., GEWIRTH, L., & BLUMSTEIN, S. (1984). A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study. *Journal of the Acoustical Society of America*, **76**, 391-404.

LIBERMAN, A., COOPER, F., SHANKWEILER, D., & STUDDERT-KENNEDY, M. (1967). Perception of the speech code. *Psychological Review*, **74**, 431-461.

LIBERMAN, A., DELATTRE, C., & COOPER, F. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, **27**, 769-773.

LOTTO, A., KLUENDER, K., & HOLT, L. (1997). Perceptual compensation for coarticulation by Japanese quail. *Journal of the Acoustical Society of America*, **102**, 1134-1140.

NOSSAIR, Z., & ZAHORIAN, S. (1991). Dynamic spectral shape features as acoustic correlates for initial stop consonants. *Journal of the Acoustical Society of America*, **89**, 2978-2991.

OHDE, R., & HALEY, K. (1997). Stop-consonant and vowel perception in 3- and 4-year-old children. *Journal of the Acoustical Society of America*, **102**, 3711-3722.

OHDE, R., HALEY, K. VORPERIAN, H., & MCMAHON, C. (1995). A developmental study of the perception of onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America*, **97**, 3800-3812.

REPP, B., & LIN, H. (1989). Acoustic properties and perception of stop consonant release transients. *Journal of the Acoustical Society of America*, **85**, 379-398.

SINEX, D., & MCDONALD, L. (1988). Average discharge rate representation of voice-onset-time in the chinchilla auditory nerve. *Journal of the Acoustical Society of America*, **83**, 1817-1827.

SINNOTT, J. (1998). Comparative phoneme boundaries. *Current Topics in Acoustical Research*, **2**, 135-138.

SINNOTT, J., BEECHER, M., MOODY, D., & STEBBINS, W. (1976). Speech sound discrimination by monkeys and humans. *Journal of the Acoustical Society of America*, **60**, 687-695.

SINNOTT, J., & WILLIAMSON, T. (1999). Can macaques perceive place-of-articulation from formant transition information? *Journal of the Acoustical Society of America*, **106**, 929-937.

SMITS, R., TEN BOSCH, L., & COLLIER, R. (1996). Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants: I. Perception experiment. *Journal of the Acoustical Society of America*, **100**, 3852-3864.

STEVENS, K. & BLUMSTEIN, S. (1978) Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, **90**, 1309-1325.

SUSSMAN, H., HOEMEKE, K., & AHMED, F. (1993). A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation. *Journal of the Acoustical Society of America*, **94**, 1256-1266.

SUSSMAN, H., MCCAFFREY, H., & MATTHEWS, S. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America*, **90**, 1309-1325.

TROUT, J. (2001). The biological basis of speech: What to infer from talking to the animals. *Psychological Review*, **108**, 523-549.

WERKER, J. (1994). Cross-language speech perception: Development change does not involve loss. In J. Goodman & H. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 93-119). Cambridge, MA: MIT Press.

(Manuscript received February 16, 2003;
revision accepted for publication January 26, 2004.)