Plasticity of Development

edited by Steven E. Brauth, William S. Hall, and Robert J. Dooling

A Bradford Book
The MIT Press 1991
Cambridge, Massachusetts
London, England

Chapter 5

Perception, Cognition, and the Ontogenetic and Phylogenetic Emergence of Human Speech

Patricia K. Kuhl

Communication through speech and language is an exclusively human behavior. No other animal's communicative systems parallel the complexity nor the flexibility that is afforded by human language. Man's capacity for language is typically ascribed to specialized abilities that evolved for the processing of linguistic signals. These specialized linguistic abilities are hypothesized to be organized in a processing subsystem that is unique and separate from other cognitive systems. On this view language is "modularized" in an "encapsulated" and "cognitively impenetrable" processing system (Fodor 1983; Liberman and Mattingly 1985).

An alternative view is that language, and all other higher cognitive functions, are subserved by a common underlying architecture (Anderson 1983). This position attempts to formulate a unified theory of mind by asserting that all higher order cognitive functions use similar structure and similar processing strategies rather than ones that are unique and separate. On this view cognition and language deploy the same "distributed" neural machinery that interconnects diverse parts of the brain and serves many purposes (Rummelhart and McClelland 1986; Anderson 1988).

These two views of language, one holding that language stems from a fully encapsulated and independent *module*, and the other that it stems from a more generic and distributed *neural network*, are linked to different perspectives on the phylogenetic evolution of man's capacity for language. Chomsky (1980), a proponent of the modular view, argues that language is the canonical example of a sudden emergence or mutation that brought forward a fully formed and com-

Edited from a paper presented at the Biennial Distinguished Lecture Series, sponsored by the Program for Developmental Research at the University of Maryland. The preparation of this manuscript was supported by grants from NIH (HD 18286, HD 22514, and NS 26521). The author thanks A. N. Meltzoff for comments on an earlier draft of the manuscript

plex ability. Chomsky argues that until we understand how such mutations can occur, we will not fully comprehend the evolutionary biology of human language. Proponents of the alternative position argue that, regardless of its complexity, language evolved gradually from preexisting abilities (Lieberman 1984; in press). This position favors continuity in the theory of human evolution and suggests that the substrates of language are rooted in nonhuman primates.

These two views offer distinct positions on the ontogeny of language. The first view, that language is a modularized system, holds that humans' linguistic abilities are innate, that the human infant enters the world equipped with mechanisms specially evolved for the processing of linguistic signals. In effect, this view holds that infants are born with a speech module already in place (Fodor 1983; Liberman and Mattingly 1985). The alternative view suggests that infants are highly skilled at birth but that the sophistication with which they approach the acquisition of language stems from more general perceptual and cognitive abilities. On this view infants are initially capable of perceiving complex events and imposing structure on those events. Thus, while holding that infants are quite competent at birth, this position asserts that the infants' competence may well be quite general (Kuhl 1986).

The phonetic level of language—the consonants and vowels that constitute human speech—offers an ideal linguistic signal with which to test hypotheses about the phylogeny and ontogeny of the human capacity for language. The perception of speech sounds can be studied in human infants only a few hours old, well before more formal evidence of language (such as infants' first words) begins to appear. One can also examine the abilities of nonhuman animals to perceive speech sounds. No nonhuman animals are capable of human speech, in part because they lack the supralaryngeal vocal tract that is required to produce speech sounds (Lieberman 1984). The assumption made by many is that animals also have a corresponding lack of the mechanisms involved in the *perception* of speech sounds (Liberman, Mattingly, and Turvey 1972). If this were so, it would provide some evidence of human uniqueness in processing linguistic (phonetic) signals. The goal of the research reviewed here was to make direct comparisons between the speech-perception capabilities of human infants and those of nonhuman animals. By examining the set of behaviors evidenced by both groups and pinpointing where they diverge, we hoped to make inferences about the origins of human infants' abilities and to identify what, if anything, makes them unique. This in turn contributes to the more general question of the evolution of language.

The Development of Vocal Communication

Infants of many animal species are specially sensitive to the vocal signals that are critical to their survival (see Dooling and Hulse 1989 for review). Evolution also seems to have guaranteed human infants' attentiveness to their own species' communications signals. Just as the bat, the bird, the cricket, and the frog are perceptually prepared for the acquisition of species-typical vocal signals, the human baby appears to be extraordinarily well prepared to respond to the human face and the human voice. Evidence supporting interest in the face comes from studies showing that young infants prefer to look at faces rather than at other visual configurations (Frantz and Fagan 1975; Kagan et al. 1966). More surprisingly, studies show that even newborns will imitate facial actions presented to them by their conspecifics (Meltzoff and Moore 1977; 1983; 1989). In one study it was demonstrated that infants as young as 42 minutes old can imitate gestures such as mouth opening and tongue protrusion (Meltzoff and Moore 1983), thus showing that such matching behavior is part of man's basic biological endowment. This extraordinary sensitivity to human facial actions has implications for the evolution of social and communicative development as described by Meltzoff (1988).

My own work has demonstrated the human infant's exquisite sensitivity to human speech. For example, recent work in my laboratory shows that when given a choice among sounds, young infants prefer to listen to "Motherese," a highly melodic speech signal that adults use when addressing infants (Fernald 1985; Grieser and Kuhl 1988; Papousek and Papousek 1981). It is not the syntax or semantics of Motherese that holds infants' attention—it is the acoustic signal itself. When the syntax and semantics of Motherese are stripped away and only the pitch contour of Motherese remains, infants still demonstrate the preference (Fernald and Kuhl 1987). Moreover, the prosodic features of Motherese, its higher pitch, slower tempo, and expanded intonation contours, appear to be universal across language (Fernald and Simon 1984; Grieser and Kuhl 1988). We do not know what makes mothers (fathers too) speak to their infants in this way, but we do know that mothers in every language we have examined thus far produce this kind of speech and that babies demonstrate a preference for it.

The study of Motherese emphasizes the obvious impact of speech on infants' social and affective development. Infants seemingly complete absorbtion with the sound of human speech raises a different question in my mind: Does speech have any linguistic impact on infants?

I was intrigued by the problem of speech acquisition and the sudden onset of "canonical babbling" at about 6 to 8 months of life, regardless of the language environment in which the child was being reared. It caused me to wonder what went on before the onset of speech production. Were infants in any sense processing speech perceptually in a way that had linguistic relevance, even before they could produce speech? And if so, did infants' speech processing depend on listening to the sounds of their native language?

The first study published on speech perception in infants addressed this question. It demonstrated that infants exhibited a phenomenon called *categorical perception* (Eimas et al. 1971). These data provided the first evidence that infants were processing speech sounds in a linguistically relevant manner.

The Phenomenon of Categorical Perception

The phenomenon of categorical perception had been demonstrated in adults by Liberman and his colleagues at Haskins Laboratories in the 1960s (Liberman et al. 1967). Tests of categorical perception used speech sounds created by a computer. The computer created a series of sounds by altering some acoustic variable in small steps. On one end of the series the sounds were identified as the syllable /ba/; on the other end of the continuum the sounds were identified as /pa/ (figure 5.1).

The test involved asking listeners to identify each one of the sounds in the series. Researchers expected that the sounds in the series would be perceived as changing gradually from /ba/ to /pa/, with many sounds in the middle of the series sounding ambiguous. But that is not what happened. Adults reported hearing a series of /ba/s that abruptly changed to a series of /pa/s. There was no in-between. And when researchers asked listeners if they could hear the difference between two adjacent /ba/s (or/pa/s) in the series, they could not do so, even though the two /ba/s (or /pa/s) were physically different. Listeners did not hear differences between adjacent stimuli in the series until they heard a big change—the change from /ba/ to /pa/. The fact that listeners' responses were "categorical" gave the phenomenon its name.

Further research on categorical perception in adults revealed that the phenomenon was sensitive to the linguistic environment and experience of the listener (Miyawaki et al. 1975). It occurred only for sounds in an adult's native language. For example, when Japanese listeners were tested on a series of sounds that ranged from /ra/ to /la/ for American listeners, a distinction that is not phonemic in Japanese, they did not hear a sudden change at the boundary between /ra/ and /la/. They heard no change at all. (This is why Japanese speakers substitute /l/ for /r/ in speech—*flied lice* for *fried rice*.) American listeners reported hearing a series of /ra/s that changed suddenly to a series of /la/s, just as they had with the /ba/ and /pa/ stimuli.

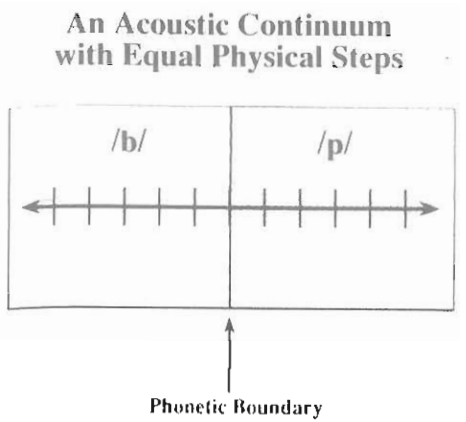

Figure 5.1
Illustration of the *categorical perception* phenomenon. An acoustic continuum is created in which changes in a physical dimension are made in small, physically equal steps. Perception of the stimuli on the continuum does not change gradually in accordance with the change in the physical dimension. Rather the stimuli are heard as a series of /ba/s that changes abruptly at the phonetic boundary to a series of /pa/s.

The finding that categorical perception was language specific suggested that it was probably learned through exposure to a specific language. This is what Eimas had set out to test in 1971. The question was: What would very young infants hear when presented with a series of /ba/s and /pa/s or /ra/s and /la/s? If hearing the sudden shift in the stimuli at the boundary between two categories was the result of experience with language (perhaps as a result of hearing their parents contrast words containing /b/ and /p/— such as *bat* and *pat*—then young infants would not be expected to show it. Older infants, on the other hand, who had experienced language, might show the categorical perception phenomenon.

Infants' responses to the sounds were monitored using a specially designed technique that relied on the measurement of sucking (Eimas et al. 1971). The results of the study revealed that infants demonstrated categorical perception. Moreover, infants demonstrated the phenomenon not only for the sounds of their own native language but also for sounds from foreign languages (Streeter 1976; Lasky, Syrdal-Lasky, and Klein 1975; Aslin et al. 1981). In all cases, infants reacted to the sounds as though they heard a sudden shift in the series

at the adult-defined boundary between the two phonetic categories. Infants appeared to be born multilingual, at least as far as phonetic perception was concerned.

Comparative Studies On Speech Perception

When the report from Eimas's lab was published, I had been reading about work on the cross-fostering of infant chimps by human adults (Gardner and Gardner 1990). It became clear from early work on cross-fostering that chimps could not learn to articulate human speech. Their vocal tracts and oral structures did not allow them to produce speech (Lieberman 1984). My question was whether animals' inability to produce speech was paralleled on the perception side. Were animals also unable to *perceive* human speech? That is, would nonhuman animals fail to demonstrate speech phenomena, such as categorical perception, that human adults and infants succeeded in demonstrating?

I began to study how nonhuman animals perceived speech sounds. The initial tests focused on the categorical perception effect (Kuhl and Miller 1975). We wanted to know whether animals heard a sudden shift in a series of stimuli at the location (for humans) of the phonetic boundary between two categories, just as humans did. Our first study resembled an identification test like those used with adult human listeners, only our test was conducted with an animal, the chinchilla (Kuhl and Miller 1975). In later tests I studied monkeys (Kuhl and Padden 1982, 1983). Both animals exhibit very good hearing and are often used in experiments on hearing because their hearing is similar to man's.

In the initial study (Kuhl and Miller 1975), animals were trained to respond differentially to computer-synthesized versions of the syllables /da/ and /ta/. The two stimuli were the endpoints of a series of stimuli that were identified (by human listeners) as /da/s and /ta/s. To one of the endpoint stimuli animals were trained to jump across a midline barrier in a cage. To the other stimulus the animal was trained to inhibit the crossing response, and this was rewarded. When performance on the endpoints was near perfect, the intermediate stimuli—those between the /da/ and /ta/ endpoints—were tested.

The critical trials were those in which intermediate stimuli were tested. The animals had not had any previous training on these stimuli and were given no feedback during the test. Each stimulus was presented and the animals' responses were monitored. These stimuli were the ones of greatest importance for theory because there were no clues telling the animal how to respond to them. The question was: How

would animals partition the continuum—would they hear the same kind of quantum leap from one category to another that humans do? None of the training they were given gave them any clue to where to draw the boundaries.

Figure 5.2 displays the results of this study (Kuhl and Miller 1975). As the data show, animals also appeared to hear the abrupt shift in the stimuli—and it occurred at precisely the location where human adults separate the /da/ and /ta/ categories. Subsequent tests on a series of stimuli ranging from /ba/ to /pa/, and tests on a series ranging from /ga/ to /ka/, were then conducted (Kuhl and Miller 1978). In all cases, the animals responded as though they heard a sudden change in the speech stimuli at the exact location where human adults perceived a shift from one phonetic category to another (see Kuhl 1986 for review).

We had thus provided some evidence supporting the evolutionary continuity hypothesis. We had shown that this aspect of the perception of human speech did not separate man from other animals. On the basis of these findings, we speculated that the boundaries for other phonetic categories might coincide with animals' *natural psychophysical boundaries*. Additional experiments were conducted to test this, our hypothesis was strongly supported (see Kuhl 1988 for review). The categorical perception of speech sounds was thus not unique to human beings.

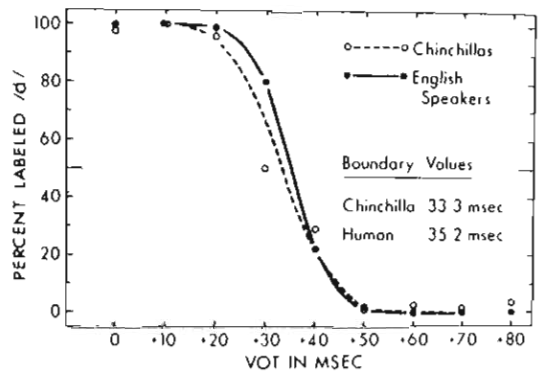

Figure 5.2
Chinchillas were trained to respond differentially to the endpoint stimuli on a continuum ranging from /da/ to /ta/. Once they were trained, the intermediate stimuli were presented and no feedback was given. The data show the mean percentage of "/d/" responses by chinchillas and humans. The phonetic boundaries between the two speech categories for the two species do not differ significantly. (From Kuhl and Miller 1978).

Infants' Perception of Speech: Beyond Categorical Perception?

Given that the phenomenon of categorical perception did not separate man from other animals, I began looking at more complex behaviors in human infants, hoping to find a place where the two species would diverge. Since that time my colleagues and I have produced four new findings on human infants' responses to speech. In one case repeated tests have failed to find the phenomenon in animals. In the other two cases we suspect that animals may not demonstrate the behaviors, but we have not as yet conducted the relevant tests.

The first phenomenon is a demonstration of *talker normalization* in infants (Kuhl 1979a, 1983). Our studies show that infants perceive vowel sounds produced by many different talkers as belonging to the same category. Why is that surprising? As adults we have no problem perceiving that a word produced by different talkers—a simple word such as *peep* produced by a man, a woman, and a child—is the same word; why is it surprising that an infant might do the same? It is surprising that infants do this because it requires a normalization process. Computers, for example, have a great deal of difficulty classifying words correctly when they are produced by a wide variety of different talkers (Kuhl et al. 1989). Yet, at least by 6 months of age, human infants accomplish this feat.

The second phenomenon focuses on the underlying basis of infants' categorization abilities. Our recent findings suggest that as early as 6 months of age, infants organize speech categories around an exemplar that adults consider to be a particularly good instance of the category, a prototype of the category (Grieser and Kuhl 1989; Kuhl in press).

The third phenomenon goes beyond the auditory processing of speech signals. This phenomenon has to do with infants' cross-modal (auditory-visual) perception of speech. We show that infants can detect correspondences between auditory speech signals and the visible articulatory movements that typically accompany them—a phenomenon linked to lip-reading (Kuhl and Meltzoff 1982, 1984a).

The fourth phenomenon is vocal imitation. Imitation examines the link between the perception and the production of speech. When infants imitate speech, they demonstrate connections between auditory perception and articulatory movements that enable them to produce speech themselves (Kuhl and Meltzoff 1982, 1988). Vocal imitation is essential to the development of speech.

Talker Normalization

The first phenomenon—talker normalization—requires what cognitive psychologists call categorization—the ability to render discriminably different things equivalent (Bruner, Goodnow, and Austin 1956).

Categorization is a phenomenon that characterizes all of perception. As stimuli typically vary along many dimensions, categorization requires that we recognize similarities in the presence of considerable variance. Often the exact criteria used to categorize are not obvious. Consider the categories *cat* and *dog*. Describing what distinguishes them, and thus what uniquely categorizes them, is not simple. They both have two eyes, four legs, fur, a tail, and so on. Configurational properties of the face probably distinguish them, but trying to describe these features is difficult. Yet we would not expect an adult to mistakenly identify a cat as a dog, or vice versa.

In speech a similar categorization problem exists. Take a simple example, such as the vowel categories /a/ as in *cot* and /ae/ as in *cat*. The differences between the two vowels are not subtle to the human ear; they are clearly different. But trying to program a computer to identify these vowels correctly when they are spoken by different individuals demonstrates it to be a very difficult program.

Figure 5.3 provides a schematic illustration of the talker normalization problem. When a single talker produces different vowels (left panel), the vowels are easily separable on some acoustic basis. The circles for each vowel enclose the utterances produced by that talker

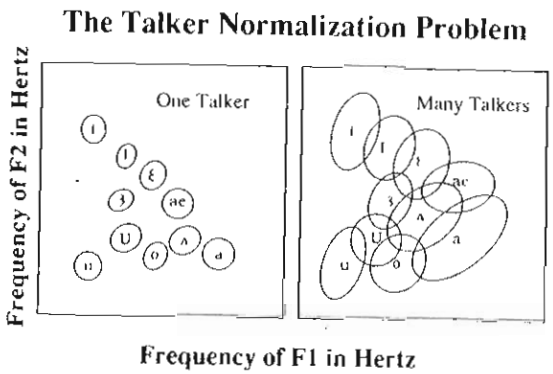

Figure 5.3
Schematic diagram illustrating the problem of talker normalization. When a single talker produces different vowels (left panel), the formant frequencies of different vowels do not overlap. However, when many talkers produce the vowels (right panel), the values of the formants for adjacent vowels overlap a great deal, making it difficult to specify the acoustic values that define any one particular vowel.

at different times. As shown, there is variability, but the circles do not overlap with one another. But when many different talkers produce vowels (right panel), there is overlap in the physical cues that underlie the two categories, and the circles overlap with one another. The explanation for this has to do with the fact that people with different-sized vocal tracts (males, females, and children) produce different resonant frequencies when they create the same mouth shape. Thus far no one has successfully described an algorithm that correctly recovers which of the vowels a speaker produced when acoustic information (the formant frequency values) is the only thing provided. In humans, various attempts to explain the processes by which we normalize the speech produced by different talkers have been offered; most of them involve computation of some kind (see Lieberman 1984 for review).

The critical question for the current discussion is whether infants recognize equivalence when the same vowel is produced by different talkers. Are all /a/s the same to the baby, regardless of the talker who produced them? It is of no small import to the child that such an ability exists early in life. Vocal-tract normalization is critical to the infant's acquisition of speech. Their vocal tracts cannot produce the frequencies produced by the adult's vocal tract, so they could not mimic the exact frequencies that an adult produces. Infants must normalize speech perceptually in order to imitate it productively.

In order to test infants' talker normalization abilities, we used a simple procedure that is shown in figure 5.4. The infant sits on a parent's lap and is visually engaged by an assistant who manipulates toys silently. A speech sound, such as the vowel sound /a/, plays repeatedly from the loudspeaker at the infant's left. The infant quickly learns that when the sound changes from the vowel /a/ to the vowel /i/ a bear playing a drum inside a black box on top of the loudspeaker is turned on. This head-turning response is the conditioned response used to test the infant's ability to normalize speech.

Once trained, the infant produces head-turning responses only when /i/ vowels occur and does not turn during presentations of the vowel /a/. The experimental question is: What will infants do when they are presented with new instances of /a/ and /i/ vowels, instances clearly different from the /a/ and /i/ stimuli heard during training? If young infants are capable of talker normalization—if they hear all /a/s (or all /i/s) as belonging to the same category—then their initial training to respond to a single /i/ sound should generalize to all members of the category. By this hypothesis, an infant trained to produce a head turn to the male's /i/ vowel, but not to his /a/ vowel, should produce head

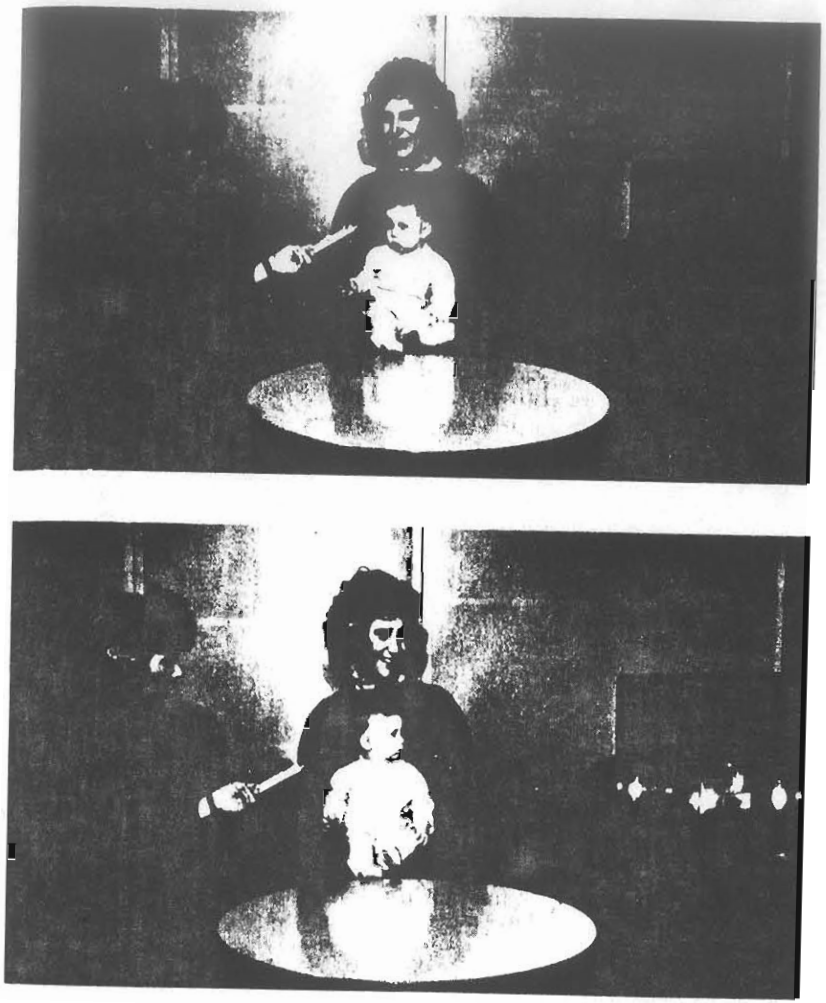

Figure 5.4
The procedure used to test infants' perception of speech. Infants who sit on a parent's lap watch toys held by an assistant (top panel). They are trained to produce a head-turning response toward the loudspeaker (located at the infant's left) when one speech sound, repeated as a background stimulus, is changed to a new speech sound. If the infant produces a head-turning response at the appropriate time, a visual reinforcer (an animated toy animal) is activated (bottom panel). The procedure is used to test infants' abilities to categorize novel speech stimuli. (See text for further details.)

turns to all novel /i/s (ones produced by females or children), but not to equally novel /a/s.

The results demonstrated that the hypothesis of talker normalization or phonetic categorization was correct (Kuhl 1979a). Infants responded correctly to the novel vowels. If the infant had been trained to turn to the male's /a/, then all novel /a/s evoked the response, while very few of the novel /i/s did. The same was true if infants were trained to turn to the male's /i/—all novel /i/s evoked the response. Figure 5.5 shows the percent head-turning responses to all of the stimuli introduced in the experiment. In the top panel infants' responses to the two stimuli used during the training phase are shown. In the bottom panel infants' responses to the stimuli presented during the test phase of the experiment are shown. Each bar in the bottom panel represents the infants' responses to the utterances of a particular talker; each talker produced one token from category 1 and one token from category 2.

As shown, infants sorted the stimuli by phonetic class, regardless of the talker producing the sounds. Infants produced high numbers of head-turning responses to the novel stimuli that were members of the phonetic category to which they were initially trained to respond (category 1 stimuli). They produced very few head-turning responses to equally novel stimuli that were members of the second phonetic category (category 2). An analysis of infants' first-trial responses showed that infants performed correctly on the very first trial. These results suggest that 6-month-old infants categorize all /a/s (and all /i/s) as the same—they appear to be capable of normalizing the speech produced by different talkers.

Kuhl (1983) extended these results to vowel categories that are much more similar from an acoustic standpoint and therefore much more difficult to categorize. The vowels were synthesized versions of /a/ (as in *cot*) and /ɔ/ (as in *caught*). In naturally produced words containing these vowels the overlap in the first two formant frequencies is so extensive that the two categories cannot be separated on this acoustic dimension (Peterson and Barney 1952). Moreover, in most dialects used in the United States talkers do not distinguish between the two vowels.

The experiment was run just as before. Infants were trained on the /a/ and /ɔ/ vowels spoken by a male talker. Then novel vowels spoken by female and child talkers, with additional random changes in the pitch contours of these vowels, were introduced. Results of the /a–/ study demonstrated that infants could still categorize the novel vowels correctly (Kuhl 1983). However, the results also showed that the task was difficult and suggested that when speech categories are very

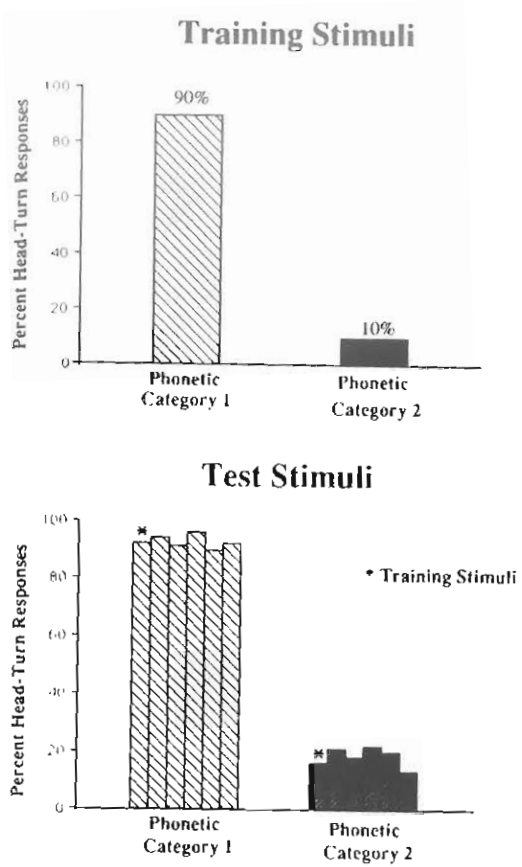

Figure 5.5
The results of tests on infants' abilities to categorize speech sounds. Infants were trained to produce a head-turning response to a single vowel from one phonetic category (either /a/ or /i/) produced by a male speaker, while refraining from producing the head-turning response to the opposite vowel produced by that same talker (top panel). Once trained, the infants' tendencies to produce the head-turning response to novel stimuli produced by men, women, and children from the /a/ and /i/ categories were tested (bottom panel). Infants' head-turning responses demonstrated that they perceptually sorted the novel stimuli by phonetic class.

similar, there is a cognitive cost associated with categorizing speech when the talker is constantly changed.

We pursued this issue further, making the experiment harder still by using many more talkers and close vowels—the /a/ in *pot* vs. the /ae/ in *pat*. This time we used vowels produced by 12 different men, women, and children. The vowels were produced naturally, rather than being computer generated, as they had been in the previous studies. We purposely chose voices that sounded very different so that extracting a constant vowel would be especially difficult. We used male talkers with deep voices, women with exceptionally high voices, even people with colds who sounded very nasal but could be understood. Adults could classify the sounds accurately. What about babies?

Figure 5.6 displays the performance on both the two training stimuli (top panel), and on the test stimuli. The results revealed two things. First, infants can categorize vowels by phonetic class when the talker is constantly changing. As shown, the percentage of head turns to novel stimuli from the two categories differed greatly. But there was another interesting finding. Although infants succeeded, the task was difficult. Switching attention from one talker to another while categorizing two vowels had a cognitive cost associated with it. These data are interesting because they are similar to data on adults showing that there are increased processing demands associated with a change in the talker producing a set of words (Mullennix and Pisoni 1990; Mullennix, Pisoni, and Martin 1989). Moreover, infants' performance on individual tokens varied. Some were classified more accurately than the training token, even though they were completely novel. Thus we had found two things. First, infants at a very young age were capable of talker normalization well before the age at which they passed any milestones in the production or in the comprehension of speech. And second, categorization of exemplars varied. Some novel instances were easier to classify than others.

This second finding came as somewhat of a surprise. Studies of categorical perception had led us to believe that, at least for speech, all members of a given category were equivalent. But these studies had been done with synthesized utterances in which all the acoustic parameters that indicate gender and those that signal a specific talker had been removed. Our studies with natural speech were replete with variation—people who were old and young, big and small, with and without colds, and all of these things led to differences in the signal that had to be contended with in the categorization task. This led to a new suggestion—that the members of a phonetic category varied qualitatively and that some might be better exemplars than others.

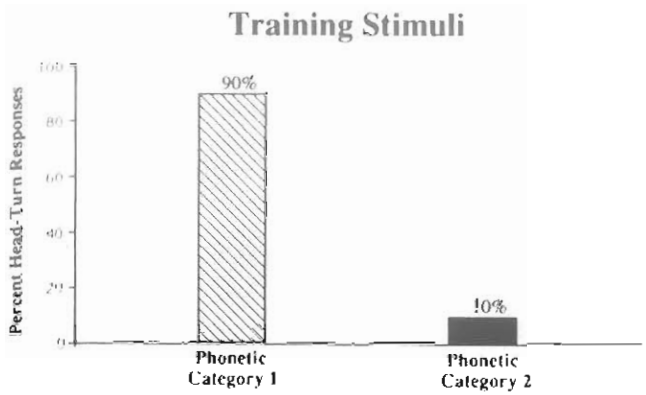

Figure 5.6
The results of tests on infants' abilities to categorize speech sounds. Infants were trained to produce a head-turning response to a single vowel from one phonetic category (either /a/ or /i/) produced by a male speaker, while refraining from producing the head-turning response to the opposite vowel produced by that same talker (top panel). Once trained, the infants' tendencies to produce the head-turning response to novel vowels produced by twelve different men, women, and children were tested (bottom panel). Infants sorted novel stimuli from the two vowel categories by phonetic class, suggesting that they *normalize* the speech produced by different talkers.

Speech Prototypes

There was some evidence in the literature that certain consonants were better exemplars than others and that they led to increased effects in certain perceptual tasks (Miller 1977; Miller and Volaitis 1989; Samuel 1982). We decided to pursue the idea that certain stimuli served as *prototypes* for speech categories. We began a new line of studies on the underlying basis of speech categorization. Our results suggest that adults and even infants organize vowel categories around an exceptionally good instance—a prototype of the category (Grieser and Kuhl 1989; Kuhl in press).

Rosch (1975) has described prototypes for physical objects as the best members of the category, the ones most representative of the category as a whole. A robin is a prototype of the category *bird*. An ostrich is not. Prototypes appear to be perceptually special. They are often processed more quickly, are more easily remembered, and are frequently preferred over others. Our question was whether there were preferred instances (prototypes) for speech categories, and if so, whether those stimuli served as cognitive reference points for speech categories.

To test the prototype hypothesis for speech, we synthesized many different instances of /i/—nearly a hundred, covering the entire range of formant values typically seen in adult speakers. We then asked adults to judge the relative goodness of each of the vowels using a scale from 1 to 7. A "7" indicated a particularly good exemplar—a perfect /i/. A "1" indicated an /i/, but a very poor one. Adults' ratings were very consistent. There was a certain location in the /i/ vowel space that always resulted in better ratings. As you moved away from that spot, the ratings became consistently worse—so adults did not perceive all members of a vowel category as equivalent. Some instances were better than others. Given that some were more striking, what was the perceptual consequence?

We developed two hypotheses. The first was that the prototype /i/ would be perceived by adults to be more similar to other /i/ vowels than the nonprototype, because it was more representative of the category as a whole. The second hypothesis added a developmental dimension. We wondered whether young infants would behave differently in a categorization test when presented with a prototype, as opposed to a nonprototype, vowel.

Two /i/ vowels were chosen from the set we had had rated by adults, one given the highest rating on average—a 6.8, and another one given a relatively poor rating— a 1.7. It is important to note that both the good and the poor exemplar were always rated as an /i/ rather than some other vowel. Both were /i/s, but the one with the 6.8 rating was perceived to be a better instance of /i/. We then computer synthesized a number of variants of /i/ around both of these two vowels.

Figure 5.7 displays the stimuli used in the experiment (Kuhl in press). Each circle on the diagram indicates an instance of a vowel. There are 32 stimuli around the prototype, represented by open circles, and 32 around the nonprototype, represented by closed circles. They form four rings around the center stimulus. An important factor about these rings is that the stimuli on them were scaled using the *mel scale* (Stevens in press). The psychophysical particulars of this scale aren't critical, but its function is to equate the distance between the center stimuli and the surrounding stimuli for the two groups (Kuhl in press). The stimuli on the first ring around the prototype are scaled to be just as discriminable from the prototype as the stimuli on the first ring around the nonprototype are from the nonprototype. One other thing to note about the stimuli is that the variants on one vector were included in both sets of stimuli. The perception of these stimuli is par-

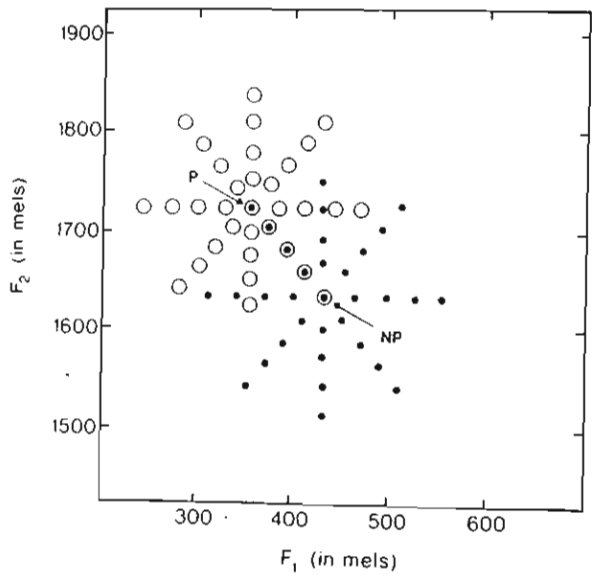

Figure 5.7
Stimuli used to test the speech prototype phenomenon. Two vowels from the /i/ vowel category were chosen, one judged by adults to be a particularly "good instance" from the category (the prototype, shown as P), and the other judged to be an /i/ with a relatively poor goodness rating (the nonprototype, shown as NP). Around each of these two stimuli, 32 variants were created by manipulating the first two formant frequencies. The stimuli were scaled using the mel scale (see text for details).

ticularly interesting because both groups of subjects were tested on them.

The hypothesis was that the prototype would be perceived as more similar to its surrounding variants than the nonprototype would be to its surrounding variants. That is, listeners would need to go further away from the prototype before they heard a difference between it and its variants than would be true of the nonprototype, even though distance was psychophysically controlled.

Two groups of 6-month-old infants were tested. The head-turning task was used. Infants heard either the prototype or the nonprototype as the reference sound during the experiment. We tested discrimination of the center stimulus from each of the surrounding stimuli and measured generalization of the head-turning response from the center stimulus to the surrounding stimuli (Kuhl in press).

We predicted that both groups, the prototype and the nonprototype, would show an effect of distance; that is, for each group, generalization from the center stimulus to the surrounding variants would be highest for those variants nearest the center vowel (those on the first ring), and generalization would decrease as you moved further away from it. This is straightforward stimulus generalization. But the prototype hypothesis predicted something more. It predicted that there would also be a significant group effect. We expected that infants in the prototype group would produce higher generalization scores at each distance because the prototype would act as a perceptual magnet and make its surrounding variants be perceived as more similar to it.

Figure 5.8 shows the mean generalization scores for each group for each ring surrounding the center vowel. As shown, there is an effect of distance for each of the two groups. Generalization scores decrease as you move further away from the center vowel. But there is also a group effect. Infants in the prototype group (the dashed line) had higher generalization scores at *each* distance. They treated many of the variants surrounding the prototype as indistinguishable from it. Infants in the nonprototype group did this to a much lesser degree. A two-way ANOVA (Analysis of Variance) examining the effect of group (prototype and nonprototype) and distance (levels 1 to 4) on infants' generalization scores showed that both of the main effects were highly significant (Kuhl in press).

There are two other results that are of interest. First, we correlated adults' ratings of the vowels' goodness for stimuli around the prototype with infants' generalization scores for stimuli around the prototype. The correlation was .95 (Kuhl in press). This suggests that the

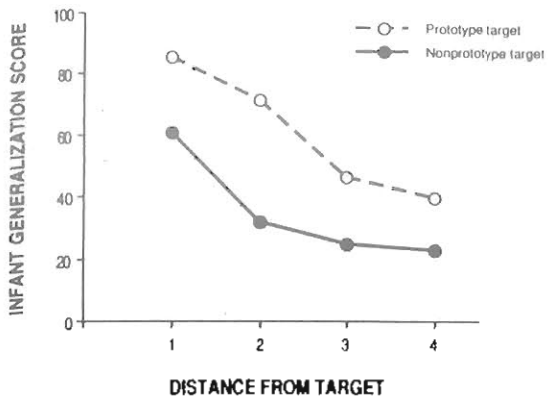

Figure 5.8
Data from a test of speech prototypes in infants. Generalization from the prototype and the nonprototype to surrounding variants (see stimuli in figure 7). Both groups show an effect of distance from the center stimulus, with generalization greater for stimuli near the center stimulus (distance 1) and poorer for stimuli far away from the center stimulus (distance 4). In addition there was an effect of stimulus group. Infants in the prototype group showed greater generalization at each distance when compared to infants in the nonprototype group.

adults' judgment of what constitutes a good or prototypic /i/ versus a nonprototypic /i/ is very closely matched to that of the 6-month-olds.

Second, recall that there was a shared vector that both groups had been tested on (figure 7). Both groups were tested on the stimuli located on the vector between the prototype and the nonprototype. The only difference was whether infants were listening to the prototype and generalizing in the direction of the nonprototype or listening to the nonprototype and generalizing in the direction of the prototype. Infants listening to the prototype generalized as far out as the third ring, whereas infants listening to the nonprototype generalizing in the direction of the prototype failed to generalize as soon as they passed the first ring. In other words, there is directional asymmetry in perception for stimuli on the common vector. This epitomizes the effect. Stimuli appear to be perceptually assimilated by the prototype. I would say that the prototype functions like a perceptual magnet—it draws other stimuli toward it, effectively reducing the perceptual distance between it and the stimuli that surround it.

These data support the notion first expressed by Stevens (1972, 1981), who argued that vowel categories were organized so as to take advantage of the quantal nature of perception. They suggest that some points in vowel space are ideal candidates for category centers, because they

are associated with perceptual stability over a broad array of category variants. Other points in vowel space are poor candidates, as perception is not stable and generalization to novel exemplars is weak. The phenomenon is consistent with prototype theory (Medin and Barsalou 1987; Rosch 1975) and is the first data that we are aware of that suggest that infants' speech categories demonstrate internal structure and organization.

The findings raise interesting questions: What makes a particular vowel a prototype? Is there some way of defining the stimulus properties of these ideal vowels? And of more interest to those of us who are attempting to explain development, how do 6-month-old babies know which vowels are prototypes? How do these ideal exemplars get into the mind of the baby?

There are two potential answers to the developmental question, and they make different predictions about the nature of the prototype. The first answer regarding development is that the prototype effect is innate. We may have tapped Platonic ideals. An alternative is that the vowel prototypes are attributable to linguistic input–that infants have already begun to form representations of the vowels in the ambient language, and they summarize this input in terms of the prototype. This second view takes the spoken language of the parents, which is still meaningless to a 6-month-old, as salient input that bathes the baby many hours a day and alters his or her perceptual space.

The two models make different predictions about infants' perception of vowels from a foreign language. The first hypothesis—that vowel prototypes are fixed—predicts that the prototype effect would exist for many vowels, even those that infants have never heard, perhaps all the vowels of all languages. The second hypothesis predicts that the prototype effect would result only when vowels in the infant's own language were used.

To test this hypothesis I designed a cross-language study wherein infants from two different language environments, English and Swedish, are each tested on the English /i/ vowel prototype and also on a prototype of the Swedish front rounded vowel (/y/ in Fant's 1973 notation). The vowel systems of the two languages are very different (Fant 1973), and adults from the two cultures rate the goodness of the exact same vowels very differently. Vowel /i/ prototypes are located in different places for American and Swedish adults, so that an English /i/ prototype is not perceived as a prototype to adult Swedes, and the Swedish /y/ prototype is not perceived as a prototype to adult Americans.

The goal is to conduct an identical study (testing both English and Swedish vowel prototypes) in two different countries. In order to achieve a situation in which an experiment conducted in two different countries was identical, I packed up my entire laboratory (everything—computer, loudspeaker, cables, reinforcers, everything down to the scissors), as well as my research team, which consisted of three testers, and sent them off to Stockholm, Sweden. All aspects of the study remained the same—the testers, the stimuli, the equipment, the reinforcers, the toys used to distract the infants, even the table mothers sat at—the only variable that changed was the language experience of the 6-month-olds who were tested. The question is, Will the 6-month-olds from the two countries resemble their adult counterparts, showing the prototype effect only for the vowels of their own language? Or will vowel prototypes be exhibited universally by infants from both cultures, in the absence of experience?

We are still in the process of testing the infants, so we do not yet know what the answer is, but there is another set of data that is relevant. A test of speech prototypes in my monkey lab has just been completed (Kuhl in press). The results showed that monkeys do *not* show the prototype effect. The test was conducted in a very similar way, the only exception being that monkeys responded by hitting a telegraph key and were reinforced with a squirt of applesauce. The results showed that monkeys demonstrated a significant *distance* effect. In other words, they demonstrated straightforward stimulus generalization around both the prototype and the nonprototype sounds. However, they did not show differential generalization, and thus no *prototype* effect. Evidently, unlike categorical perception, the prototype effect is not based on a perceptual process that is common to monkey and man (Kuhl in press).

Cross-Modal Speech Perception

Thus far I have limited the discussion of infants' perception of speech to auditory events. We typically think of speech as an exclusively auditory phenomenon. Now I extend the discussion to the detection of cross-modal equivalence for speech, wherein categorization abilities go beyond those involving auditory perception.

Recent studies on adults completed in our lab (Green and Kuhl 1989; Green and Kuhl 1991, Grant et al. 1985) and others (McGurk and MacDonald 1976; Massaro 1987; Massaro and Cohen 1983; Green and Miller 1985; Summerfield 1979) show that the perception of speech is strongly influenced by information gleaned from watching the face of a talker. This raises profound problems for a theory of speech per-

ception because it means that visual information, such as watching a talker's lips come together to produce the consonant /b/, is somehow equated in perception to acoustic information that auditorially signals the consonant /b/. (See Kuhl and Meltzoff 1988 for discussion). One important question about such complex cross-modal equivalences is how information as different as the sight of a person producing speech and the auditory speech event that is the result of production come to be related. To answer this, we decided to study the development of the ability to equate auditory and visual speech information.

We designed an experiment to pose a lip-reading problem to infants. We asked whether infants could relate the sight of a person producing a speech sound to the auditory concomitant of that event (Kuhl and Meltzoff 1982). Infants were shown two filmed faces, side by side, of a woman articulating two different vowel sounds. One face displayed productions of the vowel /a/, the other of the vowel /i/. While the infants were viewing the two faces, a single sound, either /a/ or /i/, was presented from a loudspeaker located midway between the two facial images. This eliminated any spatial cues as to which of the two faces produced the sound. The two facial images articulating the sounds moved in perfect synchrony with one another; the lips opened and closed at the exact same time, thus eliminating any temporal cues. The only way an infant could solve the problem was by recognizing a correspondence between the sound and the mouth shape that normally caused that sound. In other words, infants had to perceive a cross-modal match between the auditory and visual representations of speech.

Thirty-two infants ranging in age from 18 to 20 weeks were tested. They were placed in an infant seat facing a three-sided cubicle (figure 5.9). The experiment had two phases, a familiarization phase and a test phase. During familiarization infants saw each of the two faces for ten seconds in the absence of sound. Following this phase both faces were presented side by side, and the sound was turned on. Infants were video- and audio-recorded. An observer who was uninformed about the stimulus conditions scored the videotaped infants' visual fixations to the right or left stimulus.

The hypothesis was that infants would prefer to look at the face that matched the sound. The results confirmed this prediction; infants looked longer at the face that matched the vowel they heard. Infants presented with the auditory /a/ looked longer at the face articulating /a/. Those who heard /i/ looked longer at the face articulating /i/. The effect was strong—of the total looking time, 73 percent was spent on the matched face ($p < .001$) and 24 of the 32 infants demonstrated the effect ($p < .01$). There were no other significant

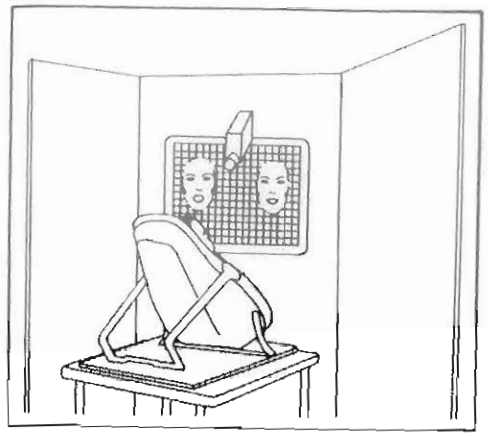

Figure 5.9
Infants' cross-modal speech-perception abilities are tested by presenting them with two facial images, one articulating the vowel /a/ and the other the vowel /i/. One or the other of the two sounds is presented from a loudspeaker midway between the two facial images. The results show that infants look longer at the face that matches the speech sound they heard. (From Kuhl and Meltzoff 1982).

effects—no preference for the face located on the infant's right as opposed to the infant's left side or for the /a/ face as opposed to the /i/ face. There was no significant difference in the strength of the effect when the matching stimulus was located on the infant's right as opposed to the infant's left. (See Kuhl and Meltzoff 1984a for full details.)

We then replicated the findings with 32 additional infants and a new research team (Kuhl and Meltzoff 1984b). All other details of the experiment were identical. The results again showed that infants looked longer at the face that matched the sound they heard. Of the total fixation time, infants spent 62.8 percent fixating the matched face ($p < .05$), and 23 of the 32 infants demonstrated the effect ($p < .01$). Recently another team of investigators has also replicated this cross-modal matching effect for speech using disyllables such as *mama* versus *lulu* and *baby* versus *zuzi* in a design similar to ours (MacKain et al. 1983).

Next we extended our tests to another vowel pair (/i-u/), thus including the third "point" vowel in the set of vowels tested. The point vowels are maximally distinct, both acoustically and articulatorily, and occur at the three endpoints of the triangle that defines "vowel space" (Peterson and Barney 1952). The test was conducted just as it had been previously, only this time infants watched faces producing the

vowels /i/ and /u/ and listened to either /i/ or /u/ vowels. The results showed that the effect could be extended to a new vowel pair. The mean percentage of fixation time to the matched face was 63.8 percent ($p < .05$), and 21 of the 32 infants looked longer at the matched face ($p < .05$) (Kuhl and Meltzoff 1984b).

Thus 4-month-olds perceive auditory-visual equivalents for speech. They recognize that /a/ sounds go with wide-open mouths, /i/ sounds with retracted lips, and /u/ sounds with pursed lips. What accounts for infants' cross-modal speech perception abilities? Have infants learned to associate an open mouth with the sound pattern /a/ and retracted lips with /i/ simply by watching talkers speak? Does some other kind of experience play a role in this ability? Our tests are now being conducted on younger infants to examine the learning account; we are specifically interested in whether or not experience in babbling plays a role in the effect (Kuhl and Meltzoff 1984a).

Vocal Imitation

Thus far in discussing the infant's detection of equivalences in speech the focus has been on the perception of speech through different sensory modalities—auditory and visual. I turn now to speech production to examine another aspect of equivalence that infants detect for speech.

As adults we can produce a specific auditory target, such as a vowel, on the first try. It is not a trial-and-error process. Auditory signals are directly related to the motor commands necessary to produce them because adults have rules that dictate the mapping between articulation and audition. This mapping is quite sophisticated. Experiments show that if an adult speaker is suddenly thwarted in the act of producing a given sound by the introduction of a sudden load imposed on his lip or jaw, compensation is essentially immediate (Abbs and Gracco 1984). The adjustment can occur on the very first laryngeal vibration, prior to the time the adult has heard anything. Such rapid motor adjustments suggest a highly sophisticated and flexible set of rules relating articulatory movements to sound.

How do auditory-articulatory mapping rules develop? Evidence suggests that at least one important mechanism for learning them is vocal imitation (Studdert-Kennedy 1986).

From Piaget on, reports have appeared that are highly suggestive of vocal imitation of at least one prosodic aspect of speech—its pitch (Kessen, Levine, and Wendrich 1979; Lieberman 1984; Papousek and Papousek 1981; Piaget 1962); however, all but one of these studies (Kessen, Levine, and Wendrich 1979) involve natural interactions between adults and infants, and as such are subject to methodological

problems (Kuhl and Meltzoff 1988). Natural observations of mothers and their infants are usually subject to the question, Who is imitating whom? The Kessen et al. study tested infants in multiple sessions over several months, giving them repeated practice and feedback, so the issue of training is unresolved in the study.

With these issues in mind we sought evidence of vocal imitation in our own experiments on infants' cross-modal perception of speech (Kuhl and Meltzoff 1982, 1988). The cross-modal studies provided a controlled setting in which to study vocal imitation. Recall our experimental set-up. Infants sat in an infant seat facing a three-sided cubicle. They viewed a film of a female talker producing vowel sounds. Half of the infants were presented with one auditory stimulus while the other half were presented with a different auditory stimulus. The stimuli were totally controlled, both visually and auditorially. There were no human interactions with the infant during the test, and thus no chance for spuriously shaping and/or conditioning a response. The room was a soundproof chamber and a studio-quality microphone was suspended above the infant to obtain clear recordings that could be perceptually or instrumentally analyzed. Finally, the stimulus on film being presented to the infant occurred once every three seconds, with an interstimulus interval of about two seconds. This was ideal for encouraging turn taking on the part of the infant. We found that infants in this setting were calm and highly engaged by the face-voice stimuli. They often listened for a while, smiled at the faces, and then started talking back. Our question was, Do infants' speech vocalizations match those they hear?

In our initial report we described data that were highly suggestive of infants' imitation of the prosodic characteristics of the signal (Kuhl and Meltzoff 1982). We observed an infant matching of the pitch contour of the adult model's vowels. Both the adult's and infant's responses are shown in figure 5.10. Instrumental analysis showed that the infant produced an almost perfect match to the adult female's rise-fall pattern of intonation. While the infant has shorter vocal folds and therefore produces a higher fundamental frequency, the pitch pattern of a rapid rise in frequency followed by a more gradual fall in frequency duplicates that of the adult. The two contours were perceptually very similar. The infant's response also matched the adult's in duration. Because vocalizations with this rise-fall pattern and of this long duration are not common in the utterances of 4-month-olds, it was highly suggestive of vocal imitation. But because we had not varied the pitch pattern of the vowel in the experiment, it was not possible to conclude definitively that infants could differentially match the pitch contour of vowels.

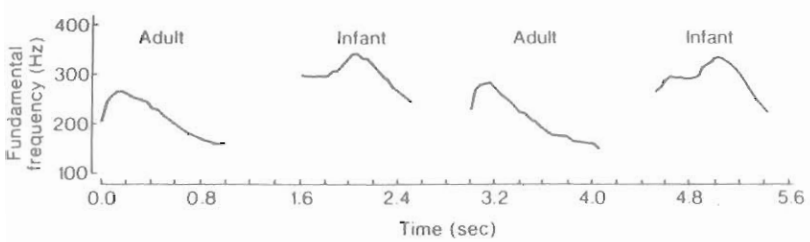

Figure 5.10
Infant vocal imitation of the adult's production of pitch. The infant duplicates the adult's pattern of change in fundamental frequency over time. Both contours show a rise in the fundamental frequency, followed by a gradual fall in the fundamental frequency. The infant's vocal cords are shorter, and thus the infant's rise-fall contour is higher in frequency.

A more rigorous test of young infants' ability to imitate relates to their matching of the phonetic segments of speech. Half of the infants in our experiments had heard /a/ vowels while the other half had heard /i/ vowels. This allowed a good test of the differential imitation of speech sounds. All of the vowel-like vocalizations produced by the infants in the /a-i/ studies were analyzed. Vowel-like sounds were defined on the basis of acoustic and articulatory characteristics typical of vowels. The sounds had to be produced with an open mouth, rather than one that was closed. They had to have a minimum duration of 500 milliseconds. They had to be voiced, that is, vocalized with normal laryngeal vibration, and could not be aspirated or voiceless sounds. They could not be produced on an inhalatory breath. Vocalizations that occurred while the infant's hand was in his or her mouth could not be reliably scored and were excluded. Consonant-like vocalizations were also scored, but they occurred rarely and were always accompanied by vowel-like sounds.

Once identified, the sounds were submitted to analysis. Perceptual scoring was done by having a trained phonetician listen to each infant's productions and judge whether, on the whole, they were more /i/-like or /a/-like. Infants at this age cannot produce perfect /i/ vowels, due to anatomical restrictions. They can, however, produce other high front vowels such as /I/ or /ɘ/. Similarly, a perfect /a/ is rare in the vocalizations of the 4-month-old, but similar central vowels, such as /æ/ and /ʌ/ are producible by infants at this age. Thus the judgment made by the observer was a forced choice concerning whether an infant's vocalizations were more /a/-like or more /i/-like.

If the observers' forced-choice decision predicted whether infants had been exposed to /a/ as opposed to /i/ based on the infant's vocalization, then there is evidence for vocal imitation. The results confirmed this prediction (Kuhl and Meltzoff 1988). Infants' vocalizations were judged to be /a/-like when listening to /a/ and /i/-like when listening to /i/; the judge's forced-choice decisions predicted accurately in 90 percent of the instances the vowel heard by the infant. These results were highly significant ($p < .01$).

We also measured infants' vowels instrumentally. Using distinctive feature theory to guide our instrumental analyses, we measured the first and second formant frequencies of the infants' vowel productions. The results demonstrated that infants' vocal responses to /a/ were significantly more "grave" in feature theory (Jakobson, Fant, and Halle 1969), than their responses to /i/. Similarly, infants' vocal responses to /a/ were significantly more "compact," that is, they had formant frequencies spaced more closely together, than their responses to /i/. Taken together, the two analyses provide evidence that 4-month-old infants are engaged in vocal imitation of the phonetic segments of speech (see Kuhl and Meltzoff 1988 for further details).

Do Animals' Abilities Extend beyond Categorical Perception?

It is now of interest to return to the question of animals' abilities. Human infants' abilities extend well beyond the skills exhibited by categorical perception. Is the same true of animals?

For the four abilities just described—talker normalization, speech prototypes, cross-modal perception, and vocal imitation—we have as yet very little data on animals' abilities. On the topic of talker normalization, there are some data on the perception of categories that include talker variation (Burdick and Miller 1975; Kuhl and Miller 1975, Dooling and Brown 1990), but the data do not include tests on vowel categories that are very similar (as in the case of /a/ versus /ɔ/), so the question of the extent of animals' abilities remains unresolved. We do have data suggesting that the prototype effect exhibited by 6-month-old human infants is not demonstrated by animals (Kuhl in press). This result is intriguing, and we are pursuing it in further studies.

On the ability of animals to detect auditory-visual correspondences for speech, and to vocally imitate speech, we do not as yet have data. We can speculate, however, that these abilities could very well go beyond the capabilities of animals. The ability to detect correspondences between auditorially and visually presented speech may depend on extensive experience in simultaneously watching and listening to spoken language. This is a quite normal occurrence for the human infant, since face-to-face communication between mother and infant begins at birth and flourishes thereafter. But it is not true for a mon-

key, even for monkeys reared in a laboratory. Alternatively it may be the case that infants' recognition of cross-modal correspondence for speech depends on the infants' recognition that the visual stimulus (the talking face) is "like me," a situation that cannot be duplicated in the monkey. Infants' detection of these correspondences may also require some degree of sound production on the part of the infant, such as that occurring during the early "cooing" stages of speech production.

Finally, the ability to imitate vocally may be beyond the monkey's competence. Monkeys do not imitate sound in the laboratory, nor, apparently, in the wild. Studies of deafened infant macaques, conducted in Japan in the seventies (Green 1975) suggested that infant macaques acquired their vocal repertoires at the same time as normal controls regardless of whether or not they were able to hear the signals that they (or other animals) produced. This is very different from the case in humans where hearing is critical to the development of normal speech production. Thus, we may have uncovered a very important difference between man and monkey.

Discussion

We return now to the question posed at the outset: Are human infants' speech perception abilities unique to the species? Five topics on infants' perception of speech have been reviewed: categorical perception, talker normalization, speech prototypes, the auditory-visual perception of speech, and vocal imitation. The data show that infants display remarkable skill in all of these tasks and that they do so remarkably early in life. Does this mean there is a speech module at work in the baby? The data tempt us to draw this conclusion, but it is probably premature to do so.

Consider first the tests on categorical perception. Here the results are very clear. Infants show the phenomenon, but monkeys and chinchillas do as well. Moreover, categorical perception results are replicable with nonspeech signals (Miller et al. 1976; Pisoni 1977; Pisoni, Carrell, and Gans 1983). It may be, then, that categorical perception for speech reflects a general and basic property of the auditory perceptual system. I have argued that this is no accident (Kuhl 1979b, 1988). In the evolution of speech, sounds were chosen for use as communicative entities—as phonemes—because they were maximally different from an auditory perspective. They fell on opposite sides of these natural psychophysical boundaries. in other words, categorical perception may not be the result of a speech module, rather, speech capitalized on an already existing tendency to carve the world of sound into certain categories (Kuhl and Miller 1978; Kuhl 1988). This would explain why there is so much regularity in the features used universally across languages.

Next, consider the studies that tie production and perception together—auditory-visual speech perception and vocal imitation. Regarding the first, we have shown that infants are capable of linking auditory and visual representations of speech. By 18 weeks they already know what an /a/, and /i/ and /u/ look like on the face of a talker. They can lip-read these sounds (Kuhl and Meltzoff 1982, 1984a, 1984b, 1988). Moreover, as early as 12 weeks of age, there is evidence of vocal imitation (Kuhl and Meltzoff 1988). These abilities are very complex, yet early imitation and cross-modal matching are not unique to the domain of speech. It has been shown that infants imitate soundless gestures such as mouth opening and tongue protrusion very early in life (Meltzoff and Moore 1977, 1983, 1989). In addition, young infants detect cross-modal relations between touch and vision (Bower 1982, Meltzoff and Borton 1979). Evidently, even these extremely sophisticated abilities do not ensure that there is a special speech mechanism at work. They may all be attributable to infants' more general cognitive abilities.

Finally, consider infants' categorization of speech sounds. Our work shows that infants are capable of categorizing speech sounds despite the variability among the members of speech categories (Kuhl 1979a, 1983). The newest results on vowel prototypes suggest that as early as 6 months of age, infants' categorization of speech may be attributable to their recognition of ideal exemplars from the category (Grieser and Kuhl 1989; Kuhl in press). We do not as yet know how prototypes get into the mind of the baby. Finding that out is our current priority. If speech prototypes are built in, then that would be evidence in support of the special mechanism hypothesis. If not, then speech categories, like categories in other domains, may be constructed through experience with exemplars, and this would obviate the need for a special mechanism.

Theoretically there is a pendulum which contrasts the two theoretical views that have been weighed here (Kuhl 1986). One view is that there exists at birth special mechanisms for the perception of speech (Fodor 1983; Liberman and Mattingly 1985). This view received strong support from the results on infants' categorical perception of speech (Eimas et al. 1971). Then when the results on animals were taken into account, the pendulum swung in the opposite direction—toward the view that the infant's behavior can be accounted for by more general auditory perceptual mechanisms (Kuhl 1986, 1987a, 1987b). Now we have the new data on infants discussed here, showing that infants

can form categories, recognize prototypes, and detect cross-modal relations for speech. One might be sorely tempted once again to attribute the infants' skills to a special mechanism for speech. But we should probably resist this tendency because there is much evidence to suggest that even these sophisticated abilities are not unique to the domain of speech.

In conclusion, the evidence in hand suggests that human infants may not begin life with a special speech module. Speech procession could well become modularized in adults with increasing experience with the phonological, semantic, and syntactic rules of the language, but it may not begin as a separate entity dedicated to the processing of speech and language. Infants bring to the task of language learning sophisticated abilities that aid them greatly in the language acquisition process. But the sophistication with which they perceive speech signals does not by itself indicate a process that is unique to speech and language. The processing strategies infants employ when acquiring the mother tongue may be rooted in quite general perceptual and cognitive skills.

References

Abbs, J. H., and Gracco, V. L. (1984). Control of complex motor gestures: Orofacial muscle responses to load perturbations of the lip during speech. *Journal of Neurophysiology* 51:705–23.

Anderson, J. A. (1988). Concept formation in neural networks: Implications for evolution of cognitive functions. *Human Evolution* 3:83–100.

Anderson, J. R. (1983). *The Architecture of Cognition*. Cambridge: Harvard University Press.

Aslin, R. N., Pisoni, D. B, Hennessey, B. L., and Perey, A. J. (1981). Discrimination of voice onset time by human infants: New findings and implications for the effects of early experience. *Child Development* 52:1135–45.

Bower, T. G. R. (1982). *Development in infancy*. 2d ed. San Francisco: W. H. Freeman.

Bruner, J. S., Goodnow, J. J., and Austin, G. A. (1956). *A study of thinking*. New York: John Wiley & Sons.

Burdick, C. K., and Miller, J. D. (1975). Speech perception by the chinchilla: Discrimination of sustained [a] and [i]. *Journal of the Acoustical Society of America* 58:415–27.

Chomsky, N. (1980). Rules and representations. *Behavior and Brain Sciences* 3, 1–61.

Dooling, R. J., and Brown, S. D. (1990). Speech perception by Budgerigars (Melopsittacus Undulatis): Spoken vowels. *Perception and Psychophysics* 47, 568–74.

Dooling, R. J., and Hulse, S. H. (1989). *The comparative psychology of audition: Perceiving complex sounds*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., and Vigorito, J. (1971). Speech perception in infants. *Science* 171:303–6.

Fant, G. (1973). *Speech sounds and features*. Cambridge: MIT Press.

Fantz, R. L., and Fagan, J. F., III. (1975). Visual attention to size and number of pattern details by term and preterm infants during the first six months. *Child Development* 46:3–18.

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development* 8:181–95.

Fernald, A., and Kuhl, P. K. (1987). Acoustic determinants of infants' preference for Motherese. *Infant Behavior and Development* 10:279–93.

Fernald, A., and Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology* 4:104–13.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge: MIT Press.

Gardner, B. T., and Gardner, R. A. (1990). Teaching sign language to cross-fostered chimpanzees. *Seminars in Speech and Language* 11:100–118.

Grant, K. W., Ardell, L. H., Kuhl, P. K., and Sparks, D. W. (1985). The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects. *Journal of the Acoustical Society of America* 77:671–7.

Green, K. P., and Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features during phonetic perception. *Perception and Psychophysics* 45:34–42.

Green, K. P., and Kuhl, P. K. (1991). Integrality of auditory voicing and visual place information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*.

Green, K. P., and Miller, J. L. (1985). On the role of visual rate information in phonetic perception. *Perception and Psychophysics* 38:269–76.

Green, S. (1975). Dialects in Japanese monkeys: Vocal learning and cultural transmission of locale-specific vocal behavior *Zeitschrift für Tierpsychologie* 38:304–14.

Grieser, D. L., and Kuhl, P. K. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in Motherese. *Developmental Psychology* 24:14–20.

Grieser, D., and Kuhl, P. K. (1989). Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology* 25:577–88.

Kagan, J., Henker, B. A., Hen-Tov, A., Levine, J., and Lewis, M. (1966). Infants' differential reactions to familiar and distorted faces. *Child Development* 37:519–32.

Kessen, W., Levine, J., and Wendrich, K. A. (1979). The imitation of pitch in infants. *Infant Behavior and Development* 2:93–99.

Kuhl, P. K. (1979a). Models and mechanisms in speech perception: Species comparisons provide further contributions. *Brain, Behavior and Evolution* 16:374:408.

Kuhl, P. K. (1979b). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America* 66:1668–79.

Kuhl, P. K. (1983). The perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development* 6:263–85.

Kuhl, P. K. (1986). Theoretical contributions of tests on animals to the special-mechanisms debate in speech. *Experimental Biology* 45:233–65.

Kuhl, P. K. (1987a). Perception of speech and sound in early infancy. In *Handbook of infant perception, Vol. 2, From perception to cognition*, ed. P. Salapatek and L. B. Cohen, 275–382. New York: Academic Press.

Kuhl, P. K. (1987b). The special-mechanisms debate in speech: Categorization tests on animals and infants. In *Categorical perception: The groundwork of cognition*, ed. S. Harnad, 355–86. Cambridge: Cambridge University Press.

Kuhl, P. K. (1988). Auditory perception and the evolution of speech. *Human Evolution* 3:19–43.

Kuhl, P. K. (In press). Human adults and human infants exhibit a "prototype effect" for speech sounds: Monkeys do not. *Perception and Psychophysics*.

Kuhl, P. K., Green, K. P., Gordon, J. W., Sanford, D. L., and Fu, C. (1989). Word recognition by humans and machines: Tests on multi-talker, multistyle database. *Journal of Acoustical Society of America* 86:Suppl. 1, S77 (A).

Kuhl, P. K., and Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science* 218:1138–41.

Kuhl, P. K., and Meltzoff, A. N. (1984a). The intermodal representation of speech in infants. *Infant Behavior and Development* 7:361–81.

Kuhl, P. K., and Meltzoff, A. N. (1984b). Infants' recognition of cross-modal correspondence for speech: Is it based on physics or phonetics? *Journal of the Acoustical Society of America* 76:Suppl. 1, S80 (A).

Kuhl, P. K., and Meltzoff, A. N. (1988). Speech as an intermodal object of perception. In *The development of perception: Minnesota symposia on child psychology* ed. A. Yonas, 235–66. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kuhl, P. K., and Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science* 190:69–72.

Kuhl, P. K., and Miller, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America* 63:905–17.

Kuhl, P. K., and Padden, D. M. (1982). Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Perception and Psychophysics* 32:542–50.

Kuhl, P. K., and Padden, D. M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *Journal of the Acoustical Society of America* 73:1003–10.

Lasky, R. E., Syrdal-Lasky, A., and Klein, R. E. (1975). VOT discrimination by four to six and a half month old infants from Spanish environments. *Journal of Experimental Child Psychology* 20:215–25.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review* 74:431–61.

Liberman, A. M. and Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition* 21:1–36.

Liberman, A. M., Mattingly I., and Turvey, M. T. (1972). Language codes and memory codes. In *Coding processes in human memory*, ed. A. W. Melton, & E. Martin. New York: John Wiley & Sons.

Lieberman, P. (1984). *The biology and evolution of language*. Cambridge: Harvard University Press.

Lieberman, P. (In press) *Speech, thought and selfless behavior*. Cambridge: Harvard University Press.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264:746–48.

MacKain, K.; Studdert-Kennedy, M.; Spieker, S.; and Stern, D. (1983). Infant intermodal speech perception is a left-hemisphere function. *Science* 219:1347–49.

Massaro, D. W. (1987) *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Massaro, D. W., and Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance* 9:753–71.

Medin, D. L., and Barsalou, L. W. (1987). Categorization processes and categorical perception. In *Categorical perception: The ground work of cognition*, ed. S. Harnad, 455–490. Cambridge: Cambridge University Press.

Meltzoff, A. N. (1988). Imitation, objects, tools and the rudiments of language. *Human Evolution* 3:45–64.

Meltzoff, A. N., and Borton, R. W. (1979). Intermodal matching by human neonates. *Nature* 282:403–4.

Meltzoff, A. N., and Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science* 198:75–78.

Meltzoff, A. N., and Moore, M. K. (1983). Newborn infants imitate adult facial gestures. *Child Development* 54:702–9.

Meltzoff, A. N., and Moore, M. K. (1989). Imitation in newborn infants: Exploring the range of gestures imitated and the underlying mechanisms. *Developmental Psychology* 25:954–62.

Miller, J. D., Wier, C. C., Pastore, R. E., Kelly, W. J., and Dooling R. J. (1976). Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *Journal of the Acoustical Society of America* 60:410–17.

Miller, J. L. (1977). Properties of feature detectors for VOT: The voiceless channel of analysis. *Journal of the Acoustical Society of America* 62:641–48.

Miller, J. L., and Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception and Psychophysics* 46:505–12.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J., and Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception and Psychophysics* 18:331–40.

Mullennix, J. W., and Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception and Psychophysics* 47:379–90.

Mullennix, J. W., Pisoni, D. B., and Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America* 85:365–78.

Papousek, H., and Papousek, M. (1981). Musical elements in the infant's vocalization: Their significance for communication, cognition, and creativity. In *Advances in infancy research*, ed. L. P. Lipsitt and C. K. Rovee-Collier, 164–224. Norwood, NJ: Ablex.

Peterson, G. E., and Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24:175–84.

Piaget, J. (1962). *Play, dreams, and imitation in childhood*. New York: Norton.

Pisoni, D. B. (1977). Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America* 61:1352–61.

Pisoni, D. B., Carrell, T. D., and Gans, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception and Psychophysics* 34:314–22.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology* 7:532–47.

Rumelhart, D. E., McClelland, J. L., and the PDP Research Group. (1986). *Parallel distributed processing: Exploration in the microstructures of cognition*. Cambridge: MIT Press.

Samuel, A. G. (1982). Phonetic prototypes. *Perception and Psychophysics* 31:307–14.

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In *Human communication: A unified view*, ed. J. E. E. David and P. D. Denes, 51–66. New York: McGraw-Hill.

Stevens, K. N. (1981). Constraints imposed by the auditory system on the properties used to classify speech sounds: Evidence from phonology, acoustics, and psychoa-

coustics. In *Advances in psychology: The cognitive representation of speech*, ed. T. Myers, J. Laver, and J. Anderson. Amsterdam: North-Holland.

Stevens, S. S., Volkman, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America* 8:185–90.

Streeter, L. A. (1976). Language perception of 2-month-old infants shows effects of both innate mechanisms and experience. *Nature* 259:39–41.

Studdert-Kennedy, M. (1986). Development of the speech perceptuomotor system. In *Precursors of early speech*, ed. B. Lindblom and R. Zetterström, 205–18. New York: Stockton Press.

Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica* 36:314–31.

Chapter 6

The Instinct for Vocal Learning: Songbirds

Peter Marler

Most of us tend to think of learning and instinct as irreconcilable opposites. Whether human or animal, behavior is construed as either learned or instinctive, but it cannot be both. According to this view animals display instincts, but our behavior is learned. We are presumed to exemplify what organisms can accomplish by the emancipation of behavior from instinctive control.

This antithesis is based on a logical fallacy. Even when we contemplate the most extreme case of purely arbitrary, culturally transmitted behavior, such as songbird dialects (Baker and Cunningham 1985) or our own patterns of speech, it is obvious on reflection that such behavior must in some sense be the result of an instinct at work. Without the bones and muscles, nerves and patterns of brain activity, and the very special capacity of nervous systems to forego existing predispositions and to reshape their activities as a result of experience, the cultural transmission of behavior would be inconceivable.

Similarly, the traditional view of instincts as fixed and immutable manifestations of purely genetic predispositions is also at fault. All behavior, whether it is viewed as instinctive or learned, develops out of an interaction between the genetic endowment of the embryo and the environment within which development takes place. Ontogenetic programs do differ in the degree to which they are open to influence by the developmental environment, however. Some are relatively closed and designed to resist or counteract displacements from a particular, specified ontogenetic trajectory. Others are genetically designed to be more open and malleable in the face of experience. It is toward this end of the closed-open continuum that culturally transmitted behavior falls. If we compare the behavior of individuals with

Research was conducted in collaboration with Susan Peters and supported by grant number MH14651. Esther Arruza prepared the figures, and Sondra O'Rourke typed the manuscript. I thank Judith and Cathy Marler and Eileen McCue for rearing the birds. I am indebted to Susan Peters, Stephen Nowicki, and Robert Dooling for discussion and criticism of the manuscript, and to the New York Botanical Garden Institute of Ecosystem Studies at the Mary Flagler Cary Arboretum for access to study areas.