HUMAN EVOLUTION Vol. 3 - N. 1-2 (19-43) - 1988

P. K. Kuhl

Department of Sprecht Hearing Sciences
University of Washington
Seattle
Washington 98195
USA

Auditory Perception and the Evolution of Speech

Among topics related to the evolution of language, the evolution of speech is particularly fascinating. Early theorists believed that it was the ability to produce articulate speech that set the stage for the evolution of the «special» speech processing abilities that exist in modern-day humans. Prior to the evolution of speech production, speech processing abilities were presumed not to exist. The data reviewed here support a different view. Two lines of evidence, one from young human infants and the other from infrahuman species, neither of whom can produce articulate speech, show that in the absence of speech production capabilities, the perception of speech sounds is robust and sophisticated. Human infants and non-human animals evidence auditory perceptual categories that conform to those defined by the phonetic categories of language. These findings suggest the possibility that in evolutionary history the ability to perceive rudimentary speech categories preceded the ability to produce articulate speech. This in turn suggests that it may be audition that structured, at least initially, the formation of phonetic categories.

Key words: speech perception, language evolution, infants, nonhuman primates

Among the special skills that infants of a species are evolutionary prepared for is the perception of vocal signals that are critical to their survival. Bats, birds, crickets, frogs — all come into the world especially prepared to detect and respond to species-typical vocal signals. Vocal signals are species' signatures, immediately and unambiguously identifying an individual animal as a member of a particular species. Evolution seems to have guaranteed infants' behavioral attentiveness to them (KUHL, 1988).

So it is with the human infant. 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 is 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 similarly complex visual stimuli (FANTZ & FAGAN, 1975). More surprisingly, recent studies show that even newborn infants within 72 hours of birth will imitate facial actions presented to them, duplicating simple oral gestures such as mouth opening and tongue protrusion (MELTZOFF & MOORE, 1977, 1983).

My own work has demonstrated the human infant's exquisite sensitivity to speech. Work in my laboratory shows that when given a choice young infants prefer to listen to «Motherese», a highly melodic speech signal that adults use when addressing infants (FERNALD, 1985; GRIESER & KUHL, 1988). It is not the syntax or semantics of Motherese that holds infants' attention — it is the acoustic signal itself (FERNALD & KUHL, 1987). Moreover, the prosodic features of Motherese, its higher pitch, slower tempo and expanded intonation contours are universal in the maternal speech addressed to infants (GRIESER & 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 produce this kind of speech. We also know that babies attend to Motherese more than they attend to the speech that adults use when addressing one another. Such

«matches» between infants' perceptual proclivities and important environmental stimuli are striking and unexplained. Is it simply fortuitous that such matches occur, or are they the result of specific evolutionary pressures?

Our work has led to the discovery of another match, this one also involving infants' underlying auditory capacities and the speech sounds used in the world's languages. But while the work on Motherese revealed a match that has social significance for the infant, this is a match that has linguistic significance. At a very young age infants demonstrate a match between their auditory perceptual skills and the phonetic categories of language (KUHL, 1979a, 1985a). Work in our laboratory shows that infants have the ability to differentiate speech sounds, even those that are very similar, like the vowel in the word «cot» versus the vowel in «caught» (KUHL, 1983), or the vowels in «mop» and «map» (KUHL, WOLAK & GREEN, in preparation). More importantly, we have shown that infants are able to «sort» perceptually different instances of these same vowels, ones spoken by different talkers in varying tones of voice, into the two phonetic categories. In short, they «normalize» across talkers (LIEBERMAN, 1984). In categorizing these vowels correctly, infants respond as though the differences *between* the two vowel categories are large while the differences that occur *among* members of any one of the categories are small. In other words, infant perception seems to work in such a way as to maximize between-category differences while minimizing within-category differences. The surprising thing is that from a physical standpoint, just the opposite is true. Physically, the differences between close vowels are extremely small, while the differences between two different people saying the same vowel are quite large. This is why computers cannot correctly categorize vowels produced by different people: it is a classic problem in computer speech recognition (ATLAS, 1987).

Perceptual categorization feats such as these, ones that are intractable for computers, but babies find easy, suggest that from the very beginning of life infants hear differences where they are «supposed to» linguistically (KUHL, 1987a). Thus, infants' abilities to perceive the sound units that languages use to convey meaning go beyond a keen ability to *hear* the sounds of human speech. While infants indeed demonstrate a keen ability to detect and differentiate among the sounds of speech, they do something more than this. Infants appropriately succeed and appropriately fail to respond to physical differences between two speech syllables depending upon whether those physical differences are important linguistically (EIMAS, SIQUELAND, JUSCZYK & VIGORITO, 1971; EIMAS, 1974, 1975). In essence, infants' auditory peception is «phonetically relevant»; there is a match between auditory perception and linguistic categories such that auditory perception partitions the spectrum of sound, effectively dividing acoustic space into phonetic categories (KUHL, 1987a).

What this means is that before infants have any need to pay attention to words, indeed before they even know that things in the world are named by the acoustic events we know as words, infants have a special proclivity to categorize the speech sounds that make up words. How is it that the young infant's auditory abilities are so perfectly suited to the detection of the acoustic events that signal phonetic differences? Why are their auditory abilities linguistically appropriate? Did hearing evolve to mirror the needs of language? Did a «special mechanism» evolve to process linguistic stimuli, one necessitated by the production of speech?

My work has focused on this problem for the last 15 years. I have conducted experiments on young human infants in an attempt to find out why this match exists. The original and prevailing theory of speech perception is that it is due to a «special mechanism», one that evolved especially for speech. This theory, a «motor theory», holds

that the special mechanisms are based on an articulatory representation of speech sounds (LIBERMAN, COOPER, SHANKWEILER & STUDDERT-KENNEDY, 1967; LIBERMAN & MATTINGLY, 1985). An alternative point of view is developed here. I have tested the perception of speech by infrahuman species to find out whether infants' abilities are uniquely human. The outcomes of these experiments were surprising. Since early man did not have the ability to produce articulate speech (LIEBERMAN, 1984), theorists have assumed that the ability to process speech would have been similarly absent (MATTINGLY, 1972; MATTINGLY & LIBERMAN, 1986). But the data reviewed here argue differently. Here, two lines of evidence, one from human infants and one from non-human animals, will show that the perception of speech does not necessarily require the ability to produce speech. Neither infants nor animals produce articulate speech. Yet both provide evidence of sophisticated auditory perceptual categories that conform to the phonetic categories of language. Those data have consequences for current theories of human speech perception, but of more interest to the present discussion, they have implications for theories of the evolution of human linguistic ability.

My purpose here is to detail the results of these tests on infants and animals, and demonstrate how they in turn led to the formulation of a hypothesis regarding the role that auditory perception played in the evolution of language. First, the acoustic properties that define the phonetic features of language are described. Then, the perception of these acoustic events by human adults, human infants, and nonhuman animals is considered. Finally, the implications of these findings for the evolutionary history of the human capacity for speech are discussed.

Phonetic Features of Language: Universal Acoustic Properties

One very interesting fact about human language is that its phonetic inventory is restricted across languages (LIEBERMAN, 1984; LINDBLOM, 1986; STEVENS, 1972). While many different sounds can be produced by the human vocal tract, certain articulatory maneuvers never appear as phonetic units in the world's languages. Others are very popular. Similarly, sound can be varied continuously. Yet among all of the potential acoustic signals that could have been used to signal meaningful differences, only a limited set of acoustic forms were in fact used in the world's languages. To illustrate this point, we will examine the range of acoustic values used to represent phonetic features.

Studies demonstrate that the perception of phonetic features is controlled by subtle differences in acoustic signals. In general, perception of a particular speech sound depends on the locations of its «formants». Formants are created when the airstream from the larynx passes through the vocal tract. Depending upon the shape of the mouth and the placement of the tongue, particular frequencies will pass freely while others will be damped. The resulting sound pattern has certain resonant frequencies; these freely passed resonant frequencies are called formants. The frequency locations of formants govern what one hears when listening to a speech sound.

Each time the vocal tract configuration changes (primarily due to the placement of the tongue), the formants change. Because the configuration of the mouth for the vowel /a/ in «cop» is different than the configuration of the mouth for /i/ as in «keep» or /u/ in «coop», the formants of these vowels differ. *Figure 1* displays the first two formant frequencies for these three English vowels. While more than two formants appear in natural reproductions of speech, studies have shown that the locations of the first two formants are the critical ones for vowels. When the two formants shown in *Figure 1* are

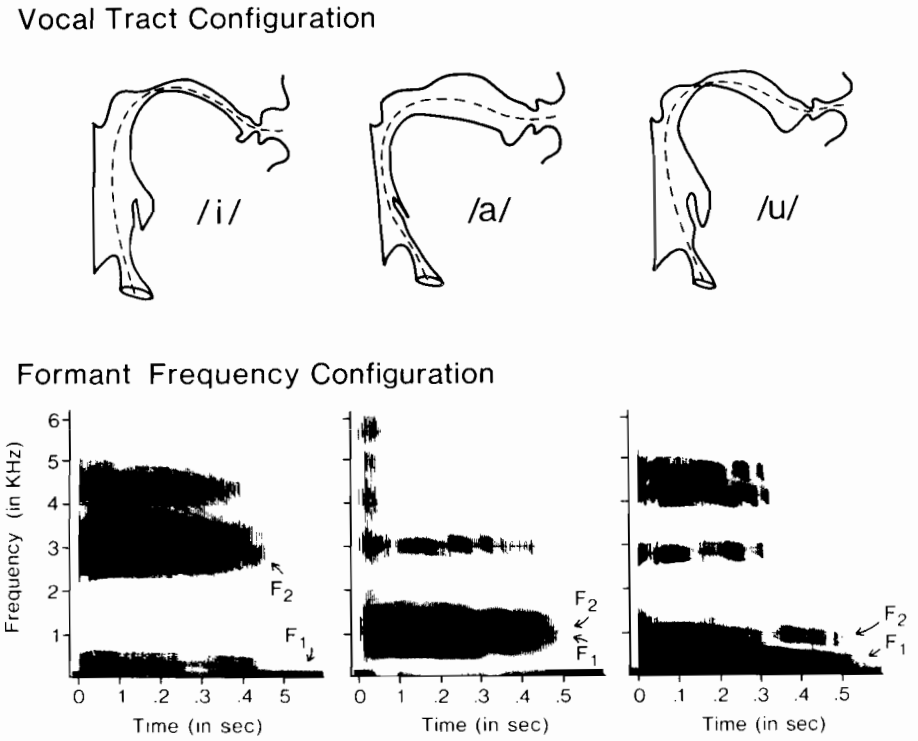

Figure 1 - Vocal tract configurations for the three «point» vowels /i/ as in «keep,» /a/ as in «cop,» and /u/ as in «coop» (top). Spectrographic displays mark the first two formant frequencies of each vowel (bottom).

artificially created by a computer, listeners will perceive the appropriate vowel (DELATTRE, LIBERMAN, COOPER & GERSTMAN, 1952).

A different representation of vowel formants is shown in *Figure 2*. This diagram plots the coordinate points in an acoustic space that correspond to the first (F_1) and second (F_2) formants for the vowels /i/, /a/, and /u/. The diagram demonstrates an important point regarding phonetic regularity. Notice that when plotted in this way the vowels define a space resembling a triangle on which /i/, /a/, and /u/ form the corners. This «vowel triangle» encompasses the acoustic coordinates of the vowels spoken in English. But more importantly, it contains those spoken in all other languages as well. For example, the vowels /i/, /a/, and /u/ occur in every language and always define the extremes of the vowel space. No languages contain vowels that fall outside of this triangular space. In fact, all of the vowels in every language of the world can be represented by only 19 points, all of which fall inside this space, and all of which fall in consistent spots in this space.

This raises questions about the nature of the constraints that restricted the selection of vowels in the world's languages: Was it the case that this space represented the vowels that could be most easily produced? Or were the contraints auditory in nature? That is, were the formant frequency patterns that fell within the triangular space somehow easier to hear than others? What guided the selection of such a restricted set of vowel categories for language?

The consonants also provide examples of phonetic regularity. Consonants are made

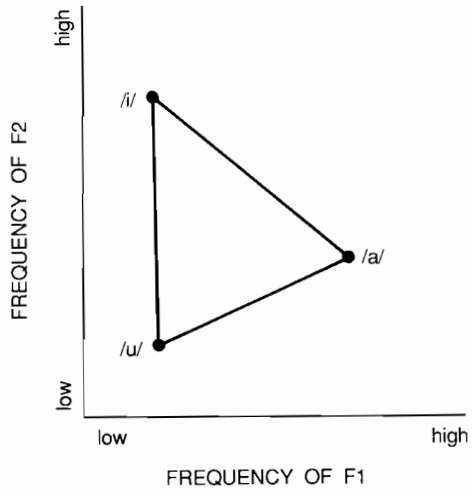

Figure 2 - Coordinate points representing the first formant (F_1) and the second formant (F_2) values of the three «point» vowels /i/ as in «keep,» /a/ as in «cop,» and /u/ as in «coop.» The three vowels form the corners of the «vowel triangle,» which encompasses the vowels spoken across all languages of the world.

up of a unique set of «distinctive features» (JAKOBSON, FANT & HALLE, 1969). For example, stop plosive consonants (/b, d, g, p, t, k/) are characterized by unique combinations of two features, one specifying a particular «place of articulation» and the other specifying a «manner of articulation». The place feature describes the location of the major constriction in the mouth that occurs when a particular speech sound is produced. The consonants /b/, /p/, and /m/ have the same place of articulation («bilabial») because they are produced by bringing the lips together. «Alveolars» such as /d/, /t/, and /n/ are created by touching the tongue tip to the alveolar ridge behind the teeth. «Velars» such as /k/ and /g/ are made when the back of the tongue touches the velum in the back of the mouth. Speech sounds that have the same place feature are distinguished by «manner» features.

The «manner» features indicate major distinctions in the way sounds are produced. For example, the consonants /b/, /d/, and /g/ are «voiced», while their counterparts /p/, /t/, and /k/ are «voiceless». For voiced sounds, the larynx starts vibrating at about the same time as the lips open; for voiceless sounds there is a delay between the opening of the lips and the onset of laryngeal vibration. Sounds such as /b/ and /p/, or /d/ and /t/ are identical but for this timing difference.

JAKOBSON, FANT & HALLE (1969) described the regularities that occur for the «voiced-voiceless» distinction. The distinction provides a particularly good example because it is phonemic in most of the world's languages in some form. The acoustic cues underlying the distinction can be examined in many different languages to see whether or not there are similarities across languages.

LISKER & ABRAMSON (1964) described a simple acoustic measure called «voice-onset time» (VOT) that separated voiced and voiceless sounds (like /b/ and /p/) based on the timing difference just described. *Figure 3* displays spectrograms of naturally produced versions of voiced and voiceless syllables in English plus a voiced syllable from an additional category used in languages other than English. For all three types, the onset of laryngeal vibration (voicing) and the burst of noise produced at the onset of the syllable are marked, and VOT is expressed in msec. For category 1 sounds, voicing precedes the onset of the burst, and VOT is a negative number. For category 2 sounds, voicing and the burst occur nearly simultaneously and VOT is nearly 0. For category 3 sounds, voicing lags the burst considerably, and VOT is a positive number. In most languages, only two of

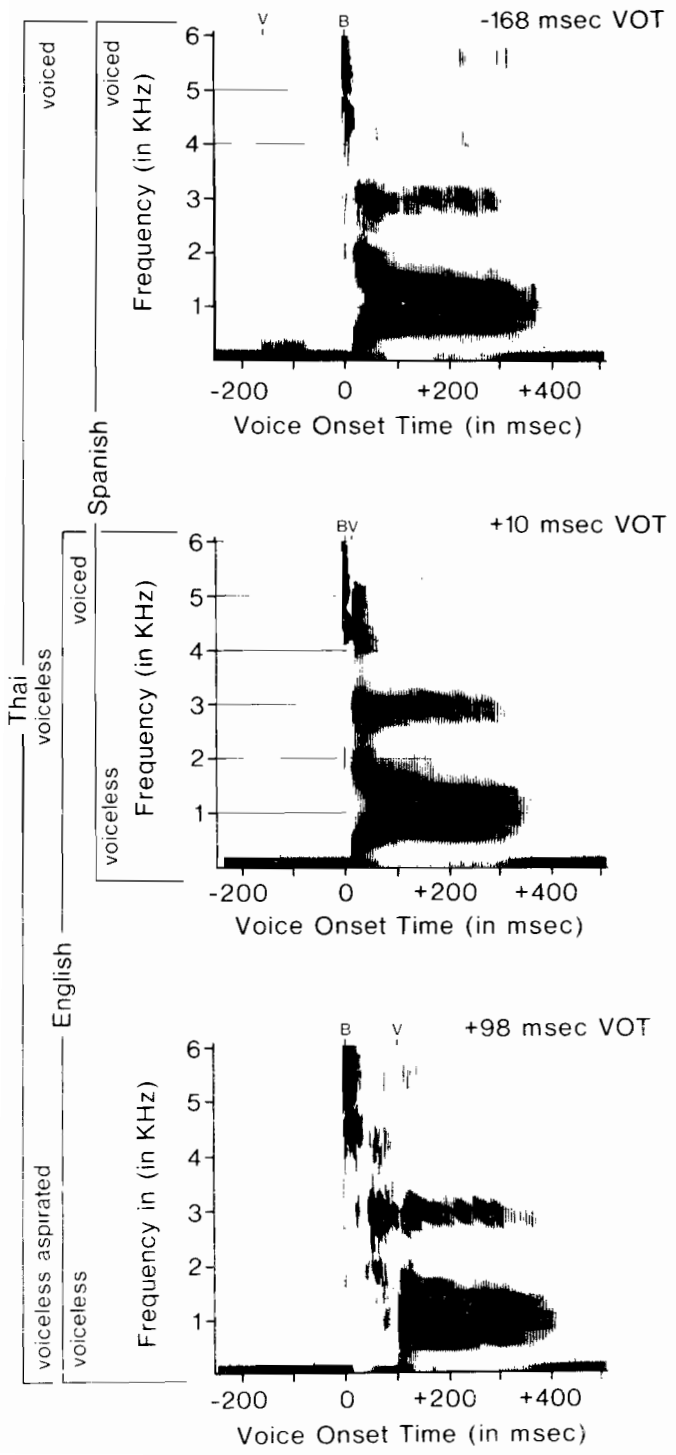

Figure 3 Spectrograms (frequency over time) of naturally produced syllables representing three voicing categories commonly used in the world's languages. Most languages use two of the three categories: Spanish and French use the first two, while English and Danish use the last two. Thai uses all three. The onset of voicing (V) and the onset of the burst (B) are marked.

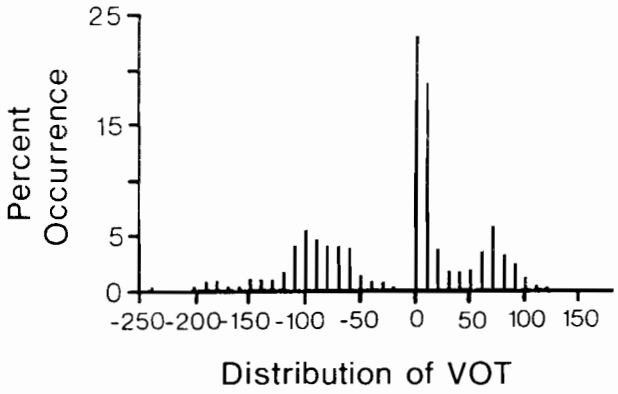

Figure 4 - Distribution of VOT values for bilabial consonants (/b/ and /p/) produced by the speakers of eleven different languages. VOT is trimodally distributed with values clustered at three different points along the continuum, showing the consistency in voicing categories used throughout the languages of the world. (From Lisker and Abramson, 1964).

these three categories are used. In English, categories 2 and 3 are phonemically contrastive, but in French and Spanish categories 1 and 2 are contrastive. In some languages (Thai, for example), all three categories are contrastive.

The remarkable point here is that the three categories shown in *Figure 3* are the only ones used for the voiced-voiceless distinction across all the world's languages. Of all the values of VOT that could have been chosen to represent these phonetic features, potentially any location on the continuum, only three locations have been picked for use in human languages. This point is illustrated in *Figure 4*, which displays the VOT values produced by talkers across eleven different languages. As shown, VOT is trimodally distributed (LISKER & ABRAMSON, 1964). There are three peaks in the VOT distribution, one for each of the types shown in *Figure 3*. The important point here is that speakers across eleven languages do not produce a random distribution of VOT's. Rather, the VOT values are clustered around points. Thus, the voiced-voiceless distinction is achieved similarly across many different languages.

In summary, the acoustic events that underlie the perception of speech features are not random across language. A rather small set of phonetic features makes up the phonetic inventory in the languages of the world. Moreover, the features consist of a restricted set of acoustic values among those that potentially could have been used. Such «phonetic universals» are common across languages. These restrictions on the articulatory/acoustic events underlying consonants led to similar arguments about the nature and origins of the constraints on phonetic features. Why were features restricted in this way? What was the nature of the pressure that guided the selection of the phonetic inventory? Were there limitations in the kinds of movements that could be made by the articulators? Or were the constraints auditory in nature? If both played a role, what guided the interactive process? What came first? Questions such as these led to studies on the perception of these sounds that will be described in the next sections.

Auditory Perception of Speech Features by Human Adults and Infants

Once the acoustic features of speech had been identified, research was aimed at determining how people perceived these acoustic events. The results of the earliest studies on the perception of speech, conducted in the early 1950's, verified that the acoustic features that had been identified using acoustic analysis techniques were in fact the ones

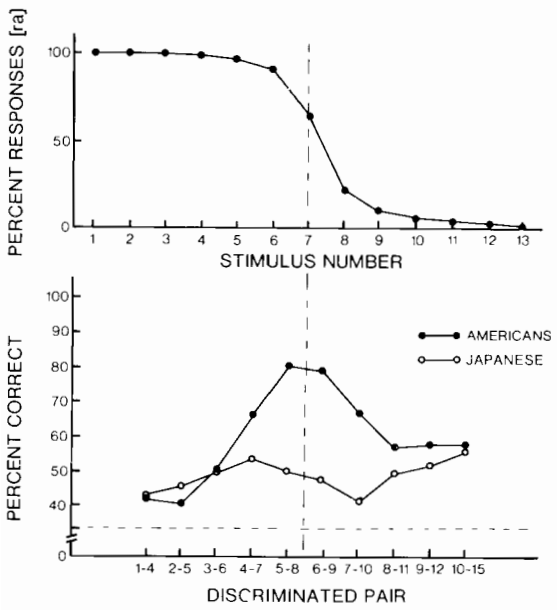

Figure 5 - A test of «categorical perception» of the syllables /ra/ and /la/ in American and Japanese adults. American adults demonstrate an increase in discriminibility at the phonetic boundary between /r/ and /l/ (dotted vertical line), while Japanese adults do not due to the fact that the /ra-la/ contrast is not phonemic in Japanese. (From Miyawaki *et al.*, 1975).

that were critical perceptually. These studies showed, for example, that the acoustic VOT «cue» reliably distinguished between sounds such as /b/ and /p/, or /d/ and /t/. But more importantly, in studying acoustic cues such as VOT researches uncovered a phenomenon that is the canonical demonstration of the phonetic relevance of human auditory perception. It has come to receive wide attention throughout psychology and the cognitive sciences (HARNAD, 1987). The phenomenon is «categorical perception», hereafter CP, discovered by a team of researchers from the Haskins Laboratories (LBERMAN, *et al.*, 1967).

The phenomenon became one of the hallmark questions that theories of speech perception needed to explain. It suggested that adults perceived the acoustic properties underlying phonetic features in an absolute «all or none» fashion. For example, when variants of /b/ were artificially synthesized with different VOT values, all appropriate for /b/, listeners reported an inability to hear the differences among them — in other words, all /b/'s were perceived to be identical. But further along the VOT continuum a kind of threshold in VOT was crossed, and listeners suddenly heard /p/. Additional variations in VOT to make a variety of /p/'s once again resulted in the perception of no differences. All /p/'s were perceived to be identical. Perception was thus «categorical»; stimuli from *different* categories were highly discriminable, but stimuli from the *same* category, while physically different, were virtually indiscriminable.

The CP phenomenon was formally demonstrated for several phonetic contrasts. The work on the /ra-la/ distinction is particularly rich. Results showed that even though the acoustic dimension changed continuously along the continuum in a stepwise fashion, perception of the syllables was discontinuous. The sounds were heard as a series of /ra/'s that changed abruptly to a series of /la/'s (*Figure 5*, top). The «boundaries» between categories did not occur at what might have appeared to be a logical division based on simple psychophysics. Second, the ability to discriminate between sounds taken from the series was tested. Pairs of syllables equally distant from one another on the continuum

were labeled as «same» or «different». The results showed that discrimination was constrained in a curious way. Adults appeared to be capable of discriminating only between sounds that fell in different categories (*Figure 5*, bottom) such that peaks in the discrimination function appeared at the boundaries between phonetic categories. Two /ra/'s on the continuum were very difficult to discriminate, but a /ra/ and /la/ stimulus, no further apart than the two /ra/'s, were quite discriminable (MIYAWAKI *et al.*, 1975). Even though the stimulus pairs were equally different from a physical standpoint, that is, even though an equal physical difference separated them, discrimination was not equal.

An additional point is illustrated by this example. CP occurred only for contrasts that were «phonemic» (changed the meaning of a word) in an adult speaker's language. Japanese adults, for whom the /ra-la/ distinction is not phonemic («fried rice» = «flied lice» to a Japanese speaker), did not produce the characteristic peak in the discrimination function for /ra-la/ stimuli (*Figure 5*, bottom) (MIYAWAKI *et al.*, 1975). Their ability to discriminate the stimuli hovered near chance. This result provided potent evidence that, in adults, CP was a phenomenon that was strongly tied to one's phonological categories — it appeared to be a linguistic phenomenon rather than one attributable to general auditory mechanisms.

Given these results, investigators became interested in posing the question to infants. Was CP learned? Or did infants display the tendency to partition acoustic continua so as to create phonetic categories right from the start? Two clearly different outcomes could be expected. On the one hand, since the stimuli are arranged on the continuum in physically equal steps, discrimination in the absence of any categorization provided by language should be equal for all pairs that were equally distant on the continuum. In other words, if infants were truly «prelinguistic», living up to the derivation of the word «infants» (from Latin, meaning «incapable of speech»), there ought not to be any categories at all — discriminability should be equal along the continuum, either equivalently good, or equivalently poor. Alternatively, the evidence on adults suggested that there were perceptual divisions of the continuum appropriate to language. For adults, the continuum was not continuous; it was divided into categories. If infants were endowed with an innate ability to perceive linguistic (in this case, phonetic) categories, then they might provide evidence of enhanced discrimination for pairs of sounds that straddled the adult-defined phonetic boundaries.

In 1971, Eimas and his colleagues at Brown answered the question. One-month-old infants demonstrated CP: they evidenced discrimination only for sounds that straddled the phonetic boundary on a voiced-voiceless (/ba-pa/) continuum, and failed to discriminate within-category stimuli (EIMAS *et al.*, 1971). The phenomenon was later demonstrated for stimuli along a place (/bæ-dæ-gæ/) continuum (EIMAS, 1974) and for stimuli along a manner (/ra-la/) continuum (EIMAS, 1975). Infants' ability to produce CP before a protracted period of listening to speech and before speech was produced suggested that the phenomenon was not learned. Infants appeared to divide the stimulus continuum just as adults did, right from the start. Instead of hearing a continuous stream of ever-changing sounds, they seemed to hear discrete categories of /ba/'s and /pa/'s, as we do.

Further evidence that the phenomenon was not learned came from studies showing that infants demonstrated CP even for phonetic contrasts that they had never heard, that is, ones from foreign languages (STREETER, 1976; LASKY, SYRDAL-LASKY & KLEIN 1976). Taken together, the results suggested that infants' auditory abilities were structured in such a way as to be perfectly suited to the task of language acquisition, and that these abilities were in place at birth.

Later studies demonstrated that infants' perception of speech evidenced even more

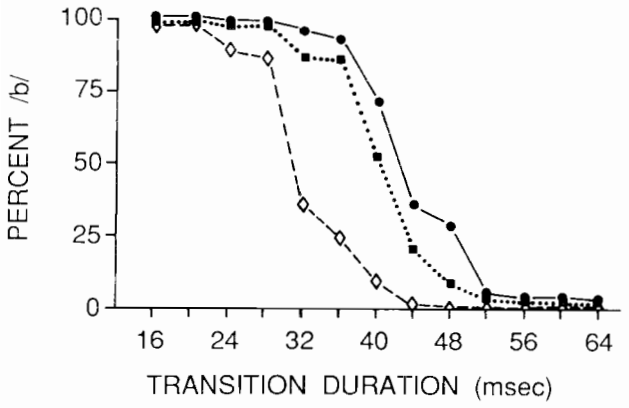

Figure 6 - Identification functions for syllables on various /ba-wa/ continua where the acoustic cue under manipulation is the duration of the first formant transition. The duration of the syllables varies from short (far left) to long (far right). The functions show that the phonetic boundaries (the 50% points on the psychometric functions) are altered by changes in the duration of the syllable such that longer syllables require longer formant transitions before the percept changes from /b/ to /w/. (From MILLER, 1981).

of the complexities found in adults' perception of speech. In adults, the locations of phonetic boundaries move with changes in the rest of the syllable. These effects are called «context effects». For example, when the rate of speech is decreased, the duration of the first-formant transition, which signals the difference between /b/ and /w/, must be made longer in order to maintain the /b/ percept (MILLER & LIBERMAN, 1979). In other words, the perceptual boundary on the /b-d/ continuum moves in accordance with syllabic rate (*Figure 6*). Now studies have been done which suggest that infants demonstrate these same effects. The studies involving infants were discrimination tests in which pairs of stimuli that straddle two of the boundaries shown in *Figure 6* (one with long syllables and one with short syllables) were tested, as were within-category pairs on the two continua. Data supporting the idea that infants demonstrate context effects are obtained if infants discriminate only those stimuli that straddle the two boundaries. EIMAS and MILLER's (1980) data on two-month-old infants demonstrate that this is the case.

To summarize: (a) Infants divide speech continua perceptually, just as adults do, prior to extensive experience in listening to speech and before speech is produced, thus suggesting the CP phenomenon is not learned. (b) Infants do this for all languages; that is, they produce CP for phonetic units they have never heard prior to the test. This supports the idea that the CP phenomenon is not learned. (c) Infants' peceptual boundaries are not «fixed» in a single location, but move with context, just as adults' boundaries do. All these data provide evidence in favor of the notion that infants' auditory abilities are linguistically relevant at birth, and this in turn suggests that infants are innately predisposed to detect the set of universal phonetic categories.

Theoretical Accounts of Infants' Phonetically Relevant Perception

There are two accounts that have been offered to explain infants' speech perception abilities. The original one argued that infants were born equipped with mechanisms specially evolved for the perception of speech — that they were endowed with a specification of the set of universal phonetic categories. By this account, infants entered the world with a «speech module» (FODOR, 1983), a device specially designed to detect all potential speech sounds. Importantly, the «special mechanism» account was based on listeners' knowledge of speech production; it was a «motor theory» (LIBERMAN *et al.*, 1967).

Motor theorists were pessimistic about the possibility that the speech mechanism was based on the detection of auditory invariants: they believed that the auditory events underlying speech were too complex and too context dependent to provide reliable cues to phonetic categories (LIBERMAN *et al.*, 1967). Instead the theory held that it was the underlying articulatory gesture that was recovered by the listener during perception (LIBERMAN & MATTINGLY, 1985). Thus, motor theory made two strong claims about the nature and specificity of the mechanisms involved in the perception of speech. It argued (a) that the perception of speech was accomplished by speech-specific mechanisms — that is, ones that had evolved specifically for the perception of speech, and (b) that these mechanisms were based on an articulatory representation of speech.

It was in this milieu that the first study of infants' abilities was undertaken. Young infants were not yet producing speech so it could be argued that they had no direct knowledge of articulation. The question was: Would infants' auditory abilities demonstrate phonetic relevance, even before they could speak? We now know the answer to this question. Infants' auditory abilities are phonetically relevant prior to the production of speech. But even though results on infants demonstrated that the CP phenomenon existed in the absence of an ability to produce speech, motor theorists argued that it was an articulatory representation of phonetic units that led to their perception; such articulatory representations were presumed to be innate in humans (LIBERMAN & MATTINGLY, 1985).

It was at this point that a second account was offered. KUHL & MILLER (1975 1978) argued that the motor theory was difficult to test in humans because it had developed the position that prearticulate infants, prior to experience in producing speech, were nevertheless in possession of the requisite knowledge of articulation that allowed perception to occur. KUHL & MILLER pointed out that there was, in fact, no way to design a test on infants that could support or refute the idea that infants, prior to their being able to speak, nevertheless perceived speech using knowledge of articulation. The only way the theory could have been tested was to identify a state during development in which the proposed requisite ability was absent; but infants were argued to be born with the requisite motor knowledge so there was no point during development in which the requisite knowledge was absent. Thus, tests on human infants were not critical tests of the motor theory that had been developed (KUHL, 1978, 1979b).

KUHL & MILLER believed that the CP phenomenon involved something more basic. They believed that infants were relying on auditory abilities that while matched to speech were not caused by a speech-specific mechanism. They presented an alternative hypothesis and proposed a test of it:

> .. infants 1 to 4 months old discriminate synthetic stimuli that fall on different sides of the English /b-p/ phonetic boundary (VOT of +22 msec) whether they are reared in an English-speaking environment where this boundary is phonemically relevant or in a Kikuyu-speaking environment where it is not. These facts lend themselves to at least two interpretations: either young infants demonstrate this perceptual boundary because their «speech processor» responds to its potential phonemic relevance or because the boundary is a natural psychophysical one that could be demonstrated by a nonhuman as well... (KUHL & MILLER, 1975, p. 69).

Thus, KUHL & MILLER proposed that infants' initial responsiveness to speech might be attributed to their more general auditory abilities and to the existence of «natural auditory boundaries», ones inherent to the auditory system. Speech, it was argued, had affixed phonetic significance to these auditory boundaries in the course of language evolution (KUHL & MILLER, 1975, 1978; KUHL, 1979b, 1981, 1986, 1987b).

The two accounts, «motor theory» and «general auditory abilites», differ greatly. The first one argues that the mechanisms underlying speech perception are ones that evolved

for speech with mechanisms based on articulatory representations of all of the speech units that could be phonemic in any language. The alternative view argues that speech perception in infants is not based on specialized mechanisms; on this view, phonetic units are not initially represented at all. The behaviors produced by infants listening to speech are due to their general auditory abilities, which happen to be well matched to the task of perceiving speech. While this theory made fewer assumptions about the baby and in that sense was more parsimonious, it lacked empirical support. What was needed was a test of the motion that CP could exist in the absence of special mechanisms, that it could be accounted for by auditory-level mechanisms alone.

The Perception of Speech by Animals: A Test of the Alternative Hypothesis

KUHL & MILLER (1975) devised a test of the «general auditory ability» hypothesis. The problem was posed by examining an animal's perception of speech sounds. The argument for using an animal model was straightforward. Certain animal species (KUHL, 1979b) provide good models of man's auditory level of processing in the absence of any ability to produce the sounds of speech. Animals have no recourse to special speech mechanisms of any kind. The animal reflects what is natural for the auditory-processing system when all linguistic influences have been stripped away and only auditory-level influences remain.

An important implication follows this reasoning. Theory holds that CP requires a specialized phonetic mechanism and that more general auditory mechanisms cannot account for it. Infrahuman species are not equipped with the specialized phonetic machinery; if they succeed, it is due to auditory-level processing. Stated simply, then, the hypothesis under test in animal experiments involving speech is this: Is auditory processing sufficient to reproduce the CP phenomenon? If animals and humans behave similarly we can then conclude that the phenomenon can exist in the absence of specialized mechanisms.

While research from this laboratory on animals' perception of speech is now quite extensive (KUHL & MILLER, 1975, 1978; KUHL,1978, 1979b, 1981, 1986b; KUHL & PADDEN, 1982, 1983; KUHL, STEVENS & PADDEN, In preparation), three examples will be cited here to illustrate the finding. The first is from the original study (KUHL & MILLER, 1975), which examined an animal's ability to categorize sounds from a speech-sound continuum. This test focused on the characteristic «labeling» functions obtained for speech. The question was whether animals' perceptual boundaires on speech continua coincided with humans' phonetic ones, or appeared someplace else. The second is from work which focused directly on tests of discrimination (KUHL & PADDEN, 1983). The question here was whether or not, in the absence of any experience in labeling the stimuli on the continuum, animals would demonstrate enhanced discriminability at the boundaries between phonetic categories, as human infants do. The third is an experiment just completed which tests animals on the context effect involving rate of speech that is shown by infants (KUHL *et al.* In preparation).

The first study (KUHL & MILLER, 1975) resembled the identification test in a typical adult CP experiment, only with animals. Recall that in these tasks adults are asked to identify stimuli drawn randomly from a continuum as either an example of stimulus category «A» or stimulus category «B» when the two categories are syllables that differ phonetically, such as /da/ and /ta/. The test stimuli are computer-generated and form a continuum in which the value of a particular dimension is altered in a stepwise fashion

from one end of the continuum to the other. The resulting data produce a psychometric function showing the percentage of time each stimulus was categorized as /da/ as a function of changes in the underlying acoustic dimension. The 50% point on the psychometric function is termed the «phonetic boundary» between the two categories.

We asked where animals would place the boundary on a phonetic continuum. In our first experiment we used chinchillas, mammals whose basic auditory abilities are similar ro man's (KUHL, 1979b). They were trained to distinguish computer-synthesized versions of the syllables /da/ and /ta/. The two stimuli were the endpoints on a /da-ta/ continuum. The test continuum ranged from 0 msec VOT (perceived by humans as a good instance of /da/) to +80 msec VOT (perceived by humans as a good instance of /ta/). To one of the endpoint stimuli animals were trained to cross a midline barrier in the 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, a generalization paradigm was used to test the stimuli between /da/ and /ta/ on the continuum. During this generalization test, half of the trials involved the endpoint stimuli. On these trials, all of the appropriate feedback was given, just as it had been during the training phase. This was done to ensure that the animal remained trained.

The critical trials were those in which new stimuli were tested, stimuli for which no previous training had been given. These new stimuli were located between the endpoints (+10 msec VOT to +to msec VOT, in 10-msec steps). These were the ones of greatest importance for theory because there were no clues telling the animal how to respond to them, and therefore how to divide up the continuum. During these trials, the feedback was arranged to indicate that the animal was always correct, no matter what the response. The feedback served to reinforce whatever the animal did naturally.

The experiment allowed us to ask where the perceptual boundary between /da/ and /ta/ was located for the animal. The data are shown in Figure 7 (top). The mean percentage of /da/ responses to each stimulus on the continuum are plotted for four chinchillas and human adults. The resulting phonetic boundaries, located at 35.2 msec VOT for humans and 33.3 msec VOT for animals, did not differ significantly.

The two functions were very similar and suggested that animals heard an abrupt change at precisely the location where human adults separate the /da/ and /ta/ categories. On the basis of these findings, we speculated that the boundaries for other phonetic categories might coincide with animals' natural perceptual boundaries and conducted additional experiments to test this.

Our tests were extended to continua involving other voiced-voiceless pairs, namely bilabial (/ba-pa/) and velar (/ga-ka/) contrasts (KUHL & MILLER, 1978). These stimuli were of interest because human listeners' voicing boundaries move with the place of articulation specified by the particular pair. There is no ready explanation for this movement in the location of the boundary. Motor theorists argue that the movement in the perceptual boundary is the result of a mechanism that is based on articulation (SUMMERFIELD & HAGGARD, 1977). Auditory theorists argue that it could be due to interactions between the acoustic events signaling these distinctions and thus explain it by recourse to auditory perception (KUHL, 1987b).

The new tests on the bilabial and velar stimuli were run exactly as the previous ones. The endpoint VOT values were 0 and +80 msec VOT, and animals were first trained to respond differentially to them. Again, training on the endpoint stimuli only provided the animal with an appropriate response. It did not, in any way, instruct the animal where to divide the test continuum. The important test stimuli were those between the endpoints (+10 to +70 msec VOT, in 10-msec steps), for which no training was given. They were

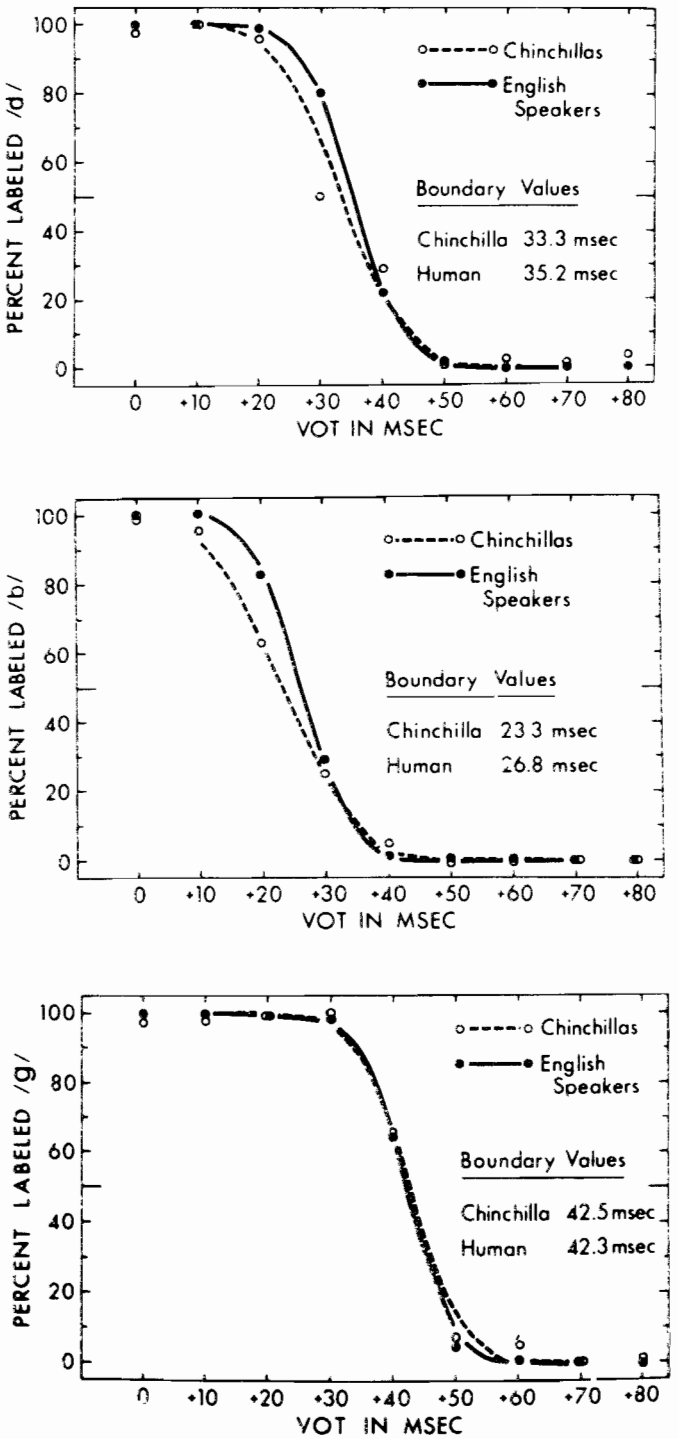

Figure 7 Identification functions derived from responses by animals (chinchillas) and human adults tested on stimuli from three voiced-voiceless continua, /d-t/ (top), /b-p/ (mid), and /g-k/ (bottom). In each case animals were trained to respond differently to the two endpoint stimuli (0 msec VOT and +80 msec VOT). Then, the intermediate stimuli were presented in the absence of any training. The locations of the boundaries for the two groups did not differ significantly. (From KUHL & MILLER, 1978).

presented on half the trials during the generalization phase.

The results again demonstrated excellent agreement between the human and animal data. The boundary values for the bilabial stimuli were 26.8 msec VOT for humans and 23.3 msec VOT for animals (*Figure 7*, mid), which were not significantly different. The boundary values for the velar stimuli were 42.3 msec VOT for humans and 42.5 msec VOT for animals (*Figure 7*, bottom). These values also did not differ significantly.

Having completed the three experiments, we examined individual categorization functions for the three places of articulation and verified that each human and animal subject had ordered their boundary values similarly. For every subject, the lowest boundary value occurred when listening to the bilabial series, and the highest boundary value occurred for the velar stimuli, with the alveolar boundary between these two. We had thus replicated in animals an unexplained perceptual interaction between two speech features, voicing and place, that had previously been observed in adults and thought to be due to an articulatory representation of speech (ABRAMSON & LISKER, 1970).

At this point we had examined animals' categorization functions, but had not examined animals' discrimination of specific pairs of stimuli from the continuum. Since it is the enhanced discriminability at the locations of phonetic boundaries and poor discriminability within categories — the «phoneme boundary effect» — that sets speech apart from other phenomena in psychophysics and in cognitive psychology (KUHL, 1987b), and since infants demonstrated this effect without learning or experience, we were motivated to test the effect directly.

The discrimination tests involved monkeys rather than chinchillas (KUHL & PADDEN, 1982, 1983). The technique involved training a monkey to initiate a trial by depressing a telegraph key. The animal was taught to lift the telegraph key when the two stimuli were Different, and was rewarded with a squirt of applesauce for doing so. If the stimuli were the Same, the monkey was rewarded with applesauce for holding the key down until the end of the trial. During training, we used stimuli that were easily discriminable, like a tone versus a noise.

This Same-Different procedure had several advantages. Once trained, animals could be tested on a variety of speech stimuli. Since it was a Same-Different procedure, the animals had no previous experience at all in categorizing the stimuli from the continuum. That made these tests more comparable to those run on human infants.

The design of the experiment was to choose pairs from a speech sound continuum, just as is done in tests on infants, and examine their discriminability. The CP effect predicts differential discriminability across the continuum, with pairs straddling the phonetic boundary being most discriminable. Two studies were conducted, one using stimuli from three different voicing continua (KUHL & PADDEN, 1982), the other using stimuli from a place continuum (KUHL & PADDEN, 1983). The results of the two experiments were identical — in both cases animals demonstrated the discrimination effect typical of CP.

These findings are here illustrated using the example involving the continuum varying in place (/b -d -g/) (KUHL & PADDEN, 1983). In experiments on human listeners using computer synthesized syllables, it has been shown that a change in the place of articulation from bilabial (/b/) to alveolar (/d/) to velar (/g/) can be governed by a change in the starting frequency of the second formant transition (MATTINGLY *et al.*, 1971). What we wanted to test with animals was whether there were any locations along a two-formant /b -d -g/ continuum where discriminability was enhanced, and if so, whether those particular locations coincided with the locations of human phonetic boundaries.

Animals' discrimination of these sounds was tested for seven pairs of stimuli, each

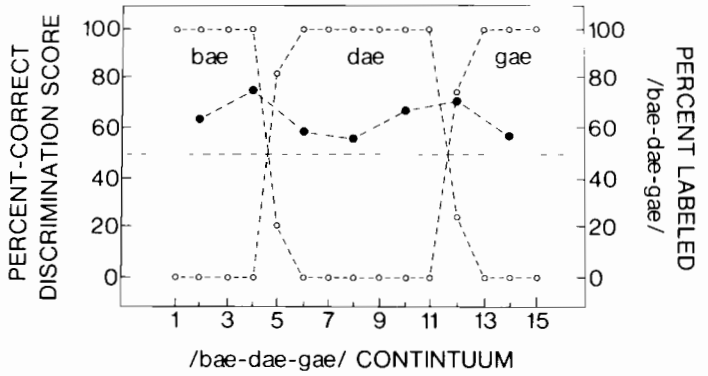

Figure 8 - Discrimination performance (filled circles) for animals tested on pairs of stimuli from a two-formant place of articulation (/b-d-g/) continuum. Performance is significantly enhanced at the /b-d/ and /d-g/ boundaries which are derived from humans identification functions (open circles).

separated by two steps on the continuum (pairs 1 vs. 3, 3 vs. 5, 5 vs. 7, and so on). The results are shown in *Figure 8*. The average per cent correct discrimination score is given for each pair. The data points are plotted at the stimulus located midway between the two stimuli forming the pair (so the data point for the 1 vs. 3 stimulus pair is plotted at stimulus number 2). The human identification functions are shown for comparison same.

As shown, the best discrimination performance for animals occurred on stimulus pairs 3 vs. 5, 9 vs. 11, and 11 vs. 13. These are exactly the pairs that involve stimuli from different phonetic categories for humans. Thus, discriminability was poor when the two stimuli were taken from the same phonetic category, like pair 1 vs. 3 (both /b/'s), pair 5 vs. 7 (both /d/'s), or pair 13 vs. 15 (both /g/'s). But discriminability was very good when the stimuli involved pairs taken from different phonetic categories, even though stimulus pairs were always separated by an equal physical distance on the continuum. These results suggested that even though the stimulus differences were equivalent from a *physical* standpoint, they were not equivalent auditorially.

The most recent finding with animals shows that their boundaries reflect some of the same complexities regarding changes with context that occur in human adults and infants (*Figure 6*). We have recently found, somewhat to our own surprise, that «context effects» exist in animals (KUHL *et al.*, in preparation). Our tests were conducted on the /ba-wa/ distinction. Recall that the perceptual boundary between /ba/ and /wa/ depends on the rate of speech. When the two syllables are spoken at «fast» versus «slow» rates of speech, the boundary between them is located at two different places. On the fast continuum, faster formant transitions are required to hear /b/ so the boundary between /b/ and /w/ is located at a faster transition rate. On the slow continuum, slower formant transitions are required to hear /b/ so the boundary is located at a slower transition rate. The problem for a theory of speech perception is to explain how the perceptual mechanism work — a «fixed» boundary cannot be specified because different boundaries are required for different rates of speech.

Our tests on macaques used the same stimuli used to test infants by EIMAS & MILLER (1980). The stimulus pairs had been chosen by these authors so that they included ones that straddled the adult-defined boundaries on both the «fast» and «slow» /ba-wa/ continua, as well as pairs that fell within each category on both continua. These same stimuli were used in tests on macaques. Our results replicated those obtained with infants. Macaques discriminated only the pairs that straddled adult human boundaries, while failing to show discrimination of pairs of stimuli that fell within a single phonetic category (KUHL *et al.*, in preparation). Thus it appears that macaques also show «context effects»

for speech; their boundaries on speech continua also move when the rate of speech is changed. Apparently, auditory-level mechanisms are sophisticated enough to adjust to rate information.

With the completion of those experiments we felt confident that animals' perception of the speech sounds we had tested was very similar to human infants' perception of those sounds. We had demonstrated that animals' «auditory» boundaries coincided with humans «phonetic» ones, and this raised, for the first time, the possibility that infants were relying on general auditory mechanisms, rather than phonetic ones, in producing speech phenomena. The data also raised the issue of speech evolution as KUHL & MILLER (1975) stated:

> The fact that the chinchillas respond to the synthetic speech as though an abrupt qualitative change occurs in the short voicing-lag region of the VOT continuum at precisely the place where many languages separate two phonemic categories lends support to the idea that speech-sound oppositions were selected to be highly distinctive to the auditory system (p. 72).

The Role of Audition in the Ontogenetic and Phylogenetic Development of Speech

With results showing that chinchillas and monkeys performed similarly to babies in speech perception experiments, theorists had no reason to impute «special mechanisms» to infants to explain phenomena such as CP (ASLIN, PISONI & JUSCZYK, 1983; EIMAS & TARTTER 1979; JUSCZYK, 1985; KUHL & MILLER, 1975). We now know that even perceptual phenomena as complex as context effects can be replicated in animals. One might have thought that the demonstration in infants of perceptual effects as difficult to understand as these would have provided strong and unambiguous support in favor of «special mechanisms». But they do not. Thus, the results on animals had changed the face of theorizing about the developmental course of infant speech perception. While the facts about what infants could do did not change, the mechanisms we attributed to those behaviours did change. On this newly emerging view, the infant's initial state could be characterized by a lack of speech-specific mechanisms; more general auditory and/or cognitive mechanisms might explain infants' capabilities (JUSCZYK, 1986; KUHL, 1978, 1987a; 1988).

But this conclusion about infants' abilities led to a new puzzle. Why were infants' general abilities so perfectly matched to a specific subsystem (i.e., language)? Why were the infant's *auditory* abilities phonetically relevant? Was it sheer accident that infants' general auditory abilities were perfectly suited to the acquisition of speech? It did not seem possible to argue that the match between the boundaries imposed by auditory perception and those imposed by linguistic categorization were due to chance. A new hypothesis was needed, one that addressed how infants' general auditory abilities could result in phonetically relevant perception.

The hypothesis we offered turned the original «special mechanisms» argument on its head. «Motor Theory» argued that a special perceptual mechanism evolved in man to detect articulatory invariants (LIBERMAN *et al.*, 1967). The «Auditory» hypothesis turns this argument around. In essence, the Auditory hypothesis holds that speech production evolved to match the properties of the ear rather than vice versa. On this view, the acoustics of speech were chosen to capitalize on invariant properties detected by the auditory system. This hypothesis explained why the acoustics of speech were so consistent across languages of the world (*Figs. 2 and 4*).

This suggestion was first raised by KUHL & MILLER (1975, 1978) and developed further by KUHL (1978, 1979b). Data and theorizing by LIEBERMAN (1984) and STEVENS

(1982, 1983) have added considerably this position. As originally proposed the theory emphasized the role that audition played in the evolution of language and described hypotheses that could be tested empirically. The findings supported the view that the evolution of language was continuous — that language grew out of, and capitalized on, existing systems (LIEBERMAN, 1984). Then, as now, the question for the theory is just how thoroughly auditory it should be. Did audition *per se* direct the selection of a phonetic inventory and the acoustic values of features? Or did auditory perception and motor skills interact in the process? If the two interacted, which came first?

KUHL & PADDEN (1983) described three alternatives regarding the influence that could have been exerted by audition in the evolution of speech. The three alternatives are: (1) audition did not provide a strong influence independent of articulation, on the evolution of speech categories; articulatory constraints guided evolution; (2) audition provided a strong and independent influence, but one that served to *structure initially* rather than *determine solely* speech categories; or (3) audition *per se* guided evolution; auditory considerations directed the selection and formation of speech categories.

The first alternative is, in essence, a motor theory. The second is an account that takes both auditory and articulatory constraints into account. The third argues that audition guided the evolution of speech categories in the absence of articulatory influences. Given the data now available, we believe that the second account is the most plausible of the three.

The first account argues that motor considerations, rather than auditory ones, played a singular role in the evolution of the phonetic categories of language. The most powerful data suggesting that this account is incorrect are those presented here and in other studies in which animals have been shown to categorize and discriminate speech sounds (BURDICK & MILLER, 1975; BARU, 1975; KUHL & MILLER, 1975, 1978; MORSE & SNOWDON, 1975; WATERS & WILSON, 1976; KUHL, 1981; KUHL & PADDEN, 1982, 1983; KUHL *et al.*, In preparation). Taken together, the data show that animals perceptually partition speech sounds into categories. Articulatory abilities are not a prerequisite for this; experience with speech is not a prerequisite for it, nor is the uniquely human destiny to acquire language. We have to argue that the structure imposed on the stimuli is due to auditory perception. This lends strong support to the claim that audition played an important role in shaping the acoustics of language, and in shaping the mechanisms that process it.

The second account holds that the auditory system provided a set of broad guidelines that initially structured the selection of phonetic candidates and their acoustic features. These guidelines might have taken the form of «natural auditory boundaries» (KUHL & MILLER, 1975), that determined the basic cuts locations of phonetic boundaries. Certain properties of sounds including temporal features, such as the relative timing of two events, and spectral features, involving distinctions like diffuse versus compact (STEVENS, 1982, 1983 for further candidates), and the way these properties are categorically processed by the auditory system would have produced a functional separation of sounds into categories. These categories would have been formed by a set of perceptual boundaries whose characteristics produced poor discriminability for stimuli falling on either side of boundary and good discriminability for stimuli straddling a boundary.

On this view, audition played a strong role, but it set the boundaries between categories rather crudely. The account does not go so far as to suggest that audition dictated more than the locations of boundaries for speech categories. It stipulates that speech capitalized on a set of preexisting boundaries, but then elaborated on them. Thus, this account suggests that the boundary effects typical of CP can be attributed to auditory-level analysis, but effects such as the «prototypicality» effects recently observed

in infants (described below), cannot be solely explained by processing at an auditory level. These effects may reflect analysis at a higher level. By this account, audition would have dictated the general locations of category boundaries, but not the centers of speech categories — audition would have *initially structured, but not solely determined* speech categories.

Nevertheless, on this view audition came first; it was the initial force that guided the selection of sounds that were used to communicate meaning. But the account holds that there is something more — something on top of these basic auditory sensitivities. The data to date on the perception of speech by animals do not contradict this general explanation. This account predicts that animals' and humans' perceptual results will coincide for many stimuli, but not all. At some point, animals' performance will break down, and the phenomena that cause this breakdown are likely to be the ones requiring higher-level processing by a more specialized mechanism.

The third alternative is the «solely auditory» account. It is the most thoroughly auditory of all, arguing for a deterministic role for audition in the evolution of speech. On this view, speech sounds form «natural classes». The natural class theory argues that even complex phenomena such as prototypicality are due to an auditory level of processing. The third account holds that it was audition that guided the formation of phonetic categories by inherently imposing natural auditory categories on the formation of phonetic categories. The articulatory mechanism evolved to achieve these auditory categories. By this account, audition guided both perception and production in the evolution of speech. Perception was guided by the existence of auditory boundaries and the existence of auditory prototypes. Production was guided in that articulatory maneuvers were adjusted as need be to produce sounds that resembled auditory prototypes.

The three alternatives can be tested. Continued comparisons — between human adults, human infants, and nonhuman animals — constitute a powerful method for evaluating the three alternatives. These tests should focus on more complex categorization tasks, particularly those that infants are good at. These could include tests of «trading relations» (BEST, MORRONGIELLO and ROBSON, 1981), and speech sound prototypes. Recent data from our lab suggest that human infants may have perceptual prototypes that match the centers of vowel categories (GRIESER & KUHL, 1983) and this ought to be tested in animals.

The prototype work is particularly relevant because the issue addressed in these studies was whether there were certain places in vowel space that were ideal «centers» of vowel categories from a perceptual standpoint. We argued that an ideal category center would exhibit certain properties. First, it would be judged to be a «good» instance of the vowel. Second, it would tend to represent a large number of the vowel stimuli surrounding it, providing a good «idealization» of the exemplars which resembled it but were not identical to it. Third, it should be easily learned and remembered. To test this in infants we examined whether there was an effect of stimulus «goodness» on infants' recognition of vowel categories.

To do this, a variety of points in a two-formant coordinate vowel space corresponding to different locations in the /i/ vowel category were selected. Adults rated the «goodness» of these vowels on a scale of 1 to 7. One stimulus that was rated as a «good» /i/ and another rated as a «poor» /i/ were chosen. While they varied in «goodness», both were perceived as /i/ rather than some other vowel. We then created variants that formed «rings» around the good and poor stimuli (*Figure 9*). On each of the four rings, eight stimuli were synthesized, for a total of 32 variants. Adults then rated the goodness of the variants around the «good» and «poor» /i/'s.

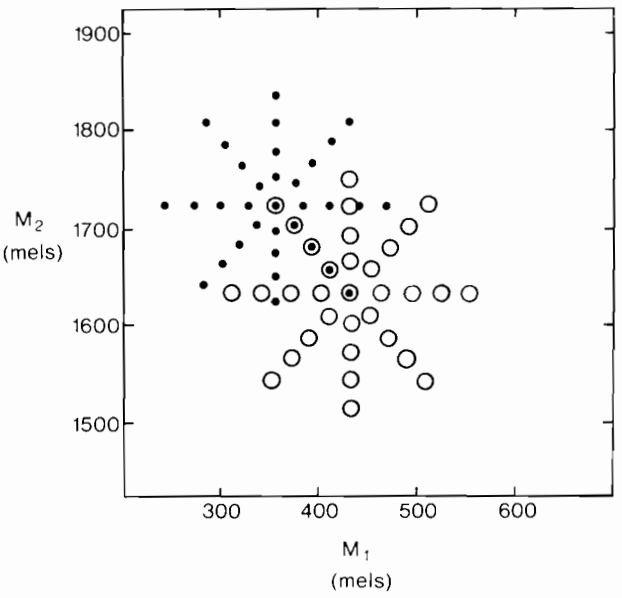

Figure 9 - Stimuli surronding the «good» and «poor» /i/ vowels. The «good» stimulus and its variants are shown as dots, the «poor» stimulus and its variants as circles. Results show that generalization is greater for stimuli surrounding the «good» vowel stimulus. In this figure the formants have been converted to mels. The «mel» scale adjusts the frequency of a stimulus to equate for perceived changes in its pitch; stimuli equally distant in mels are equally different in pitch, at least for pure tones.

The tests on infants were designed to examine whether generalization around a «good» stimulus differed from generalization around a «poor» stimulus. A head-turn conditioning technique (KUHL, 1985b) was used to examine infants' generalizations to variants around the two points. Infants were trained to produce a head-turn response to any stimulus other than the «target» stimulus. The target stimulus was either the «good» or the «poor» /i/. All of the variants around the target were randomly presented. The results showed that generalization around the «good» vowel stimulus was significantly broader than generalization around the «poor» vowel stimulus. In other words, infants produced fewer head-turns to the variants surrounding the «good» /i/ than to variants surrounding the «poor» /i/. Apparently, many more of the variants around the «good» stimulus resembled the «good» stimulus; a significantly smaller number of variants around the «poor» /i/ resembled the «poor» stimulus.

These studies suggest that some points in vowel space are better candidates for category centers than others, since they are associated with perceptual stability over a broad array of category variants. Other points in vowel space are poor candidates, since perception is not stable and generalization to novel exemplars is weak. These data support the notion expressed by STEVENS (1972), who argued that vowel categories were organized so as to take advantage of the quantal nature of perception.

An effect of stimulus goodness in 6-month-olds that mimics that shown by adults could indicate that infants organize vowel categories around a perceptually good vowel stimulus. This would be consistent with prototype theory (ROSCH, 1975). It would be interesting to test whether the prototype effect requires experience in listening to or producing sounds from a specific language. We also intend to test these effects with animals. If the perceptual grouping of stimuli in vowel space is due to effects that are purely auditory in nature, then the effects ought to be replicable in animals. But if it is due to speech processing mechanisms that take articulatory dynamics into account, or to specific language experience, then animals should not replicate these effects.

The major point to be made here is that continued study will either show that animals eventually fail on perceptual tasks such as trading relations and vowel prototypes that human infants succeed on, or alternatively, will show that both infants and animals demonstrate these phenomena. If the former pattern of results occur, then there is a definite dissociation between human infants and animals, which would support the view that speech sound evolution was strongly guided by audition. However we can now identify something additional to these basic auditory sensitivities which are in place in the human infant. Few would be surprized to find a dissociation somewhere between the speech processing mechanisms of human and nonhuman primates and these comparative studies will help establish the very point(s) of dissociation.

Beyond Audition: Infants' Intermodal Speech Perception Skills

Research in our laboratory on human infants has provided other clues to the workings of the speech processing system, and these findings also have implications for the evolution of speech. Our data show that developing infants' speech processing skills go beyond those involving a single sensory modality. Infants appear to process speech information in such a way as to make the information available to other sensory modalities, and to the motor system. In other words, they appear to process speech information intermodally. Evidence of the intermodal processing of speech in infants comes from two interrelated findings, «lipreading» and vocal imitation (KUHL & MELTZOFF, 1982, 1988).

The first one involves a demonstration that infants' perceive cross-modal equivalents for speech (KUHL & MELTZOFF, 1982, 1984). We reported a study showing that 18- to 20-week-old infants relate the auditory and visual concomitants of speech, something akin to what we as adults do when we «lip-read». *Figure 9* illustrates the technique. 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 the vowel /i/. A single sound, either /a/ or /i/, was auditorially presented from a loudspeaker located midway between the two facial images. The two facial images articulating the sounds moved in perfect synchrony with one another.

Infants' visual fixation to the two faces were recorded. 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. The effect was strong — of the total looking time, 73% was spent on the matched face ($p < .001$) and 24 of the 32 infants demonstrated the effect ($p < .01$). There were no other significant 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 the infant's left. (See KUHL & MELTZOFF, 1984 for full detail).

Thus, 4-month-old infants perceive auditory-visual concomitants for speech. They appear to know that /a/ sound go with faces displaying wide-open mouths and /i/ sounds with faces displaying retracted lips. We later extended this work and demonstrated this same effect for the vowels /i/ and /u/ (KUHL & MELTZOFF, 1988). These results indicate that infants also recognize that /u/ sounds go with pursed lips. Thus, infants can relate speech information they hear to speech information they see.

A second related finding is that infants in this experiment have a tendency to imitate the vowel sounds they hear (KUHL & MELTZOFF, 1988). Infants who heard the vowel /a/ produced vowel-like sounds whose formant frequencies were closer to adults' /a/'s than to

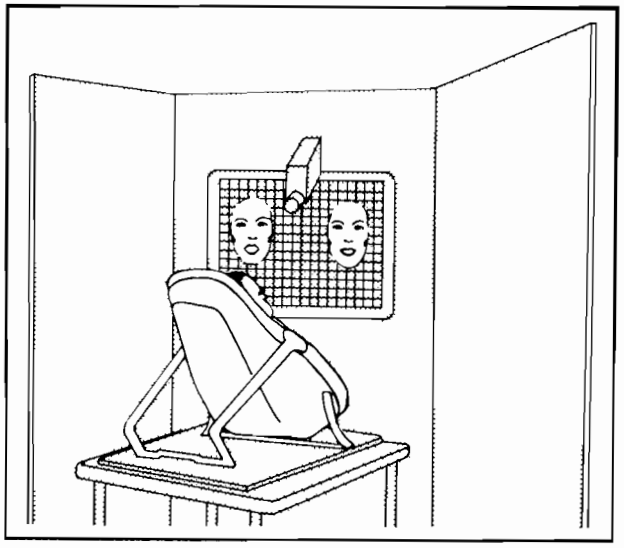

Figure 10 - Experimental set-up used to test the cross-modal perception of speech in infants. Infants see two faces producing the vowels /a/ and /i/ while a single sound (either /a/ or /i/) is presented from a loudspeaker located midway between the facial images. (From KUHL & MELTZOFF, 1982).

adults' /i/'s. Similarly, infants who heard /i/ produced vowel-like sounds whose formant frequencies were closer to adults' /i/'s than to adults' /a/'s. Moreover, infants copied the pitch contours of the vowels they heard (KUHL & MELTZOFF, 1982). LIEBERMAN (1984) has observed similar pitch and vowel matching in mother-infant interactions.

These results suggest that at least by four months of age, audition is linked to two other systems. An auditory signal drives infants' exploratory looking behaviour, causing them to seek out a visual signal that portrays to the eye an event that is equivalent to the one that they hear. The auditory signal also drives infants' motor behavior, prompting them to produce an articulatory maneuver that will result in an event that sounds equivalent to the one that they hear.

Thus, by four months of age, intermodal hook-ups are in place for speech. We have not investigated the developmental time course of these abilities, but have suggested a number of alternatives (KUHL & MELTZOFF, 1984). It will be very interesting to examine these skills in nonhuman primates. Will an auditory stimulus produce a visual search, and if so, will the search result in the detection of matches between faces and the sounds caused by their movements? Will an auditory stimulus induce sound production in the animal? These studies are yet to be done, but they will provide valuable information about the extent to which intermodal perceptual-motor connections are uniquely human.

Summary and Conclusion

One of the remarkable things about infants is the extent to which infant auditory perception is ideally suited to the perception of the phonetic categories of languages. Infant auditory perception is phonetically relevant. Theory holds that this is due to a «special mechanism» that evolved for speech. These mechanisms are argued to depend on an articulatory representation of speech. This «motor theory» assumes that in the absence of knowledge of speech production the ability to process speech perceptually cannot exist. Data reviewed here favor an alternative. The results of studies on two populations, human

infants and nonhuman animals, neither of whom are capable of producing articulate speech, nonetheless show that they are capable of perceiving speech. In fact, their perception of speech provides evidence of «natural auditory categories» that conform to phonetic ones. The implications of these findings for the evolution of speech are that while early hominids were incapable of producing articulate speech, given their vocal tracts, it is likely that they could nonetheless perceive speech. This in turn suggests that it was audition, specifically the existence of «natural auditory categories» that initially structured the formation of speech categories. This hypothesis explains the acoustic regularity observed for the phonetic features of language across cultures. Constraints on the formation of speech categories imposed by the fact that speech is a thoroughly intermodal event in humans may have been added «on top of» those provided by audition. Three alternative accounts of the evolution of speech categories were proposed and experiments that distinguish among them were described.

ACKNOWLEDGEMENTS — The author and her research are supported by grants from the National Science Foundation (BNS-8316318) and the National Institutes of Health (HD 18286 and HD 22514). The author thanks Andy Meltzoff and Phil Lieberman for comments on an earlier draft.

References

ABRAMSON A., & LISKER L., 1970. *Discriminability along the voicing continuum: Cross-language tests.* Proceedings of the Sixth International Congress of Phonetic Sciences 1967 (pp. 569-573). Prague: Academia.

ASLIN R. N., PISONI D. B. & JUSCZYK P. W., 1983. *Auditory development and speech perception in infancy.* In: M. M. Haith and J. J. Campos (Eds.), Infancy and the biology of development (v. 2, pp. 573-687) Carmichael's manual of child psychology (45th ed.). New York: Wiley.

ATLAS L., 1987. *A neural network model for vowel classification.* Paper presented at the International Conference on Acoustics, Speech, and Signal Processing, Dallas, 1987.

BARU A. V., 1975. *Discrimination of synthesized vowels [a] and [i] with varying parameters (fundamental frequency, intensity, duration, and number of formants) in dog.* In: G. Fant, & M.A.A. Tatham (Eds.), Auditory analysis and perception of speech (pp. 91-101). New York: Academic Press.

BEST C. T., MORRONGIELLO B. & ROBSON R., 1981. *Perceptual equivalence of acoustic cues in speech and nonspeech perception.* Perception and Psychophysics, 29: 191-211.

BURDICK C. K. & MILLER J. D., 1975. *Speech perception by the chinchilla: Discrimination of sustained [a] and [i].* Journal of the Acoustical Society of America, 58: 415-427.

DELATTRE P., LIBERMAN A. M., COOPER F. S. & GERSTMAN L. J., 1952. *An experimental study of the acoustic determinants of vowel color; observations on one- and two-formant vowels synthetized from spectrographic patterns.* Word, 8: 195-210.

EIMAS P. D., 1974. *Auditory and linguistic processing of cues for place of articulation by infants.* Perception and Psychophysics, 16: 513-521.

EIMAS P. D., 1975. *Auditory and phonetic coding of the cues for speech: Discrimination of the /r-l/ distinction by young infants.* Perception and Psychophysics, 18: 341-347.

EIMAS P. D. & MILLER J. L., 1980. *Contextual effects in infant speech perception.* Science, 209: 1140-1141.

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

EIMAS P. D. & TARTTER V. C., 1979. *On the development of speech perception: Mechanisms and analogies.* In: H. W. Reese & L.P. Lipsitt (Eds.), Advances in child development and behavior (Vol. 13, pp. 155-193). New York: Academic Press.

FANTZ R. L. & FAGAN J. S., 1975. *Visual attention to size and numbers of the pattern details by term and preterm infants during the first six months.* Child Development, 16: 3-18.

FERNALD A., 1985. *Four-month-old infants prefer to listen to Motherese.* Infant Behavior and Development, 8: 181-195.

FERNALD A., & KUHL P. K., 1987. *Acoustic determinants of infant preference for Motherese.* Infant Behavior and Development, 10: 279-293.

FODOR G., 1983. *The modularity of mind.* Cambridge, Mass.: MIT Press.

GRIESER D. & KUHL P. K., 1983. *Internal structure of vowel categories in infants: Effects of stimulus «goodness».* Journal of the Acoustical Society of America, 74: S102-103 (A).

GRIESER D. & KUHL P. K., 1987. *Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese.* Developmental Psychology, 24: 14-20.

HARNAD S., 1987. *Categorical perception: The groundwork of cognition.* Cambridge: Cambridge University Press.

JAKOBSON R., FANT C.G.M. & HALLE M., 1969. *Preliminaries to speech analysis: The distinctive features and their correlates.* Cambridge, Mass.: MIT Press.

JUSCZYK P. W., 1985. *On characterizing the development of speech perception.* In: J. Mehler and R. Fox (Eds.), Neonate cognition: Beyond the blooming, buzzing confusion (pp. 199-209). Hillsdale, N.J.: Erlbaum.

KUHL P. K., 1978. *Predispositions for the perception of speech-sound categories: A species-specific phenomenon?* In: F. D. Minifie & L. L. Lloyd (Eds.), Communicative and cognitive abilities - Early behavioral assessment (pp. 229-255). Baltimore: University Park Press.

KUHL P. K., 1979a. *Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories.* Journal of the Acoustical Society of America, 66: 1668-1679.

KUHL P. K., 1979b. *Models and mechanisms in speech perception: Species comparisons provide further contributions.* Brain, Behavior and Evolution, 16: 374-408.

KUHL P. K., 1981. *Discrimination of speech by nonhuman animals: Basic auditory sensitivities conducive to the perception of speech-sound categories.* Journal of the Acoustical of America, 70: 340-349.

KUHL P. K., 1983. *Perception of auditory equivalence classes for speech in early infancy.* Infant Behavior and Development, 6: 263-285.

KUHL P. K., 1985a. *Categorization of speech by infants.* In: J. Mehler & R. Fox (Eds.), Neonate cognition: Beyond the blooming, buzzing confusion (pp. 231-262). Hillsdale, N.J.: Erlbaum.

KUHL P. K., 1985b. *Methods in the study of infant speech perception.* In: G. Gottlieb & N.A. Krasnegor (Eds.), Measurement of audition and vision in the first year of postnatal life: A methodological overview (pp. 223-251). Norwood, N.J.: Ablex.

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

KUHL P. K., 1987a. *Perception of speech and sound in early infancy.* In: P. Salapatek & L. Cohen (Eds.), Handbook of infant perception: From perception to cognition (vol. 2, pp. 275-382). New York: Academic Press.

KUHL P. K., 1987b. *The special-mechanisms debate in speech: Categorization tests on animals and infants.* In: S. Harnad (Ed.), Categorial perception: The groundwork of cognition (pp. 355-386). Cambridge, Mass.: Cambridge University Press.

KUHL P. K., 1988. *On babies, birds, modules and mechanisms: A comparative approach to the acquisition of vocal communication.* In: R. J. Dooling & S. Hulse (Eds.), The comparative psychology of complex acoustic perception. Hillsdale, N.J.: Erlbaum.

KUHL P. K. & MELTZOFF A. N., 1982. *The bimodal perception of speech in infancy,* Science, 218: 1138-1141.

KUHL P. K. & MELTZOFF A. N., 1984. *The intermodal representation of speech in infants.* Infant Behavior and Development, 7: 361-381.

KUHL P. K. & MELTZOFF A. N., 1988. *Speech as an intermodal object of perception.* In: A. Yonas (Ed.), Perceptual development in infancy: Minnesota symposia on child psychology (v. 2, pp. 235-266). Hillsdale, N.J.: Erlbaum.

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

KUHL P. K. & MILLER J. D., 1978. *Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli.* Journal of the Acoustical Society of America, 63: 905-917.

KUHL P. K. & PADDEN D. M., 1982. *Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques.* Perception and Psychophysics, 32: 542-550.

KUHL P. K. & PADDEN D. M., 1983. *Enhanced discriminability at the phonetic boundaries for the place feature in macaques.* Journal of the Acoustical Society of America, 73: 1003-1010.

KUHL P. K., STEVENS E. & PADDEN D. M., In preparation. *Context effects in speech are demonstrated by macaques.*

KUHL P. K., WOLAK K. A. & GREEN K. P., In preparation. *Infants' perception of vowel categories.*

LASKY R.E., SYRDAL-LASKY, A. & 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: 215-225.

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

LIBERMAN A. M. & MATTINGLY I., 1985. *The Motor Theory of speech perception revised.* Cognition, 21: 1-36.

LIEBERMAN P., 1984. *The biology and evolution of language.* Cambridge, Mass.: Harvard University Press.

LINDBLOM, B., 1986. *Phonetic universals in vowel systems.* In: J.J. Ohala (Ed.), Experimental phonology (pp. 13-44). New York: Academic Press.

LISKER L. & ABRAMSON A. S., 1964. *A cross-language study of voicing in initial stops: Acoustical measurements.* Word, 20: 384-422.

MATTINGLY I. G., 1972. *Speech cues and sign stimuli.* American Scientist, 60: 327-337.

MATTINGLY I. G. & LIBERMAN A. M., 1986. *Specialized perceiving systems for speech and other biologically significant sounds.* Haskins Status Report on Speech Research, SR-86/87, 25-43.

MELTZOFF A. N. & MOORE M. K., 1977. *Imitation of facial and manual gestures by human neonates.* Science, 198: 75-78.

MELTZOFF A. N. & MOORE M. K., 1983. *Newborn infants imitate adult facial gestures.* Child Development, 54: 702-709.

MILLER J.L., 1981. *Some effects of speaking rate on phonetic perception.* Phonetica, 38: 159-180.

MILLER J. L. & LIBERMAN A. M., 1979. *Some effects of later-occurring information on the perception of stop consonant and semivowel.* Perception and Psychophysics, 25: 457-465.

MIYAWAKI K., STRANGE W., VERBRUGGE R., LIBERMAN A. M., JENKINS J. J. & 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: 331-340.

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

ROSCH E. H., 1975. *Cognitive reference points.* Cognitive Psychology, 7: 532-547.

STEVENS K. N., 1972. *The quantal nature of speech: Evidence from articulatory-acoustic data.* In: E. E. David, Jr., & P. D. Denes (eds.), Human communication: A unified view (pp. 51-66). New York: McGraw-Hill.

STEVENS K. N., 1982. *Constraints imposed by the auditory system on the properties used to classify speech sounds: Evidence from phonology, acoustics, and psychoacoustics.* In: T. Myers, J. Laver, & J. Anderson (Eds.), Advances in psychology: The cognitive representation of speech. Amsterdam: North-Holland.

STEVENS K. N., 1983. *Design features of speech sound systems.* In: P. F. MacNeilage (Ed.), The production of speech. New York: Springer-Verlag.

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

SUMMERFIELD O. & HAGGARD M., 1977. *On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants.* Journal of the Acoustical Society of America, 62: 435-448.

WATERS R. S. & WILSON W. A., JR., 1976. *Speech perception by rhesus monkeys: The voicing distinction in synthesized labial and velar stop consonants.* Perception and Psychophysics, 19: 285-289.

Received: 25 April 1987. Accepted 25 July 1987.