The Perception of Speech in Early Infancy

2

by Peter D. Eimas
January 1985

In perceiving speech human beings detect discrete phonemic categories and ignore much of the acoustic variation in the speech signal. Research with infants suggests the underlying perceptual mechanisms are innate

How is it that a child swiftly and seemingly without much effort learns to speak and understand? The process of language acquisition begins well before the first birthday, and most children use language with considerable skill by their third year. In contrast to the learning of reading or arithmetic, a child masters language without formal teaching; indeed, much of the learning takes place within a fairly limited linguistic environment, which does not specify precisely the rules governing competent language use.

A possible explanation for the swift growth of a child's language skills is that language is not as complex as is generally thought, and consequently that such simple psychological principles as conditioning and generalization account for the speed with which it is learned. But research during the past several decades on the nature of language and the processes by which it is produced and understood has revealed not underlying simplicity but increasing complexity.

Experiments carried out by my colleagues and me at Brown University and by other investigators elsewhere have supported a different explanation, one derived from the view, of which the linguist Noam Chomsky is the most notable exponent, that inborn knowledge and capacities underlie the use of language. In studies of speech perception by infants we have found these young subjects are richly endowed with innate perceptual mechanisms, well adapted to the characteristics of human language, that prepare them for the linguistic world they will encounter.

The search for inborn mechanisms of speech perception developed from studies of the relation of the speech signal to phonemes, the perceptual units that correspond to the consonants and vowels of language. Phonemes are the smallest units of speech that affect meaning: only one phonemic difference sets apart the words *late* and *rate*, yet they are entirely distinct in meaning.

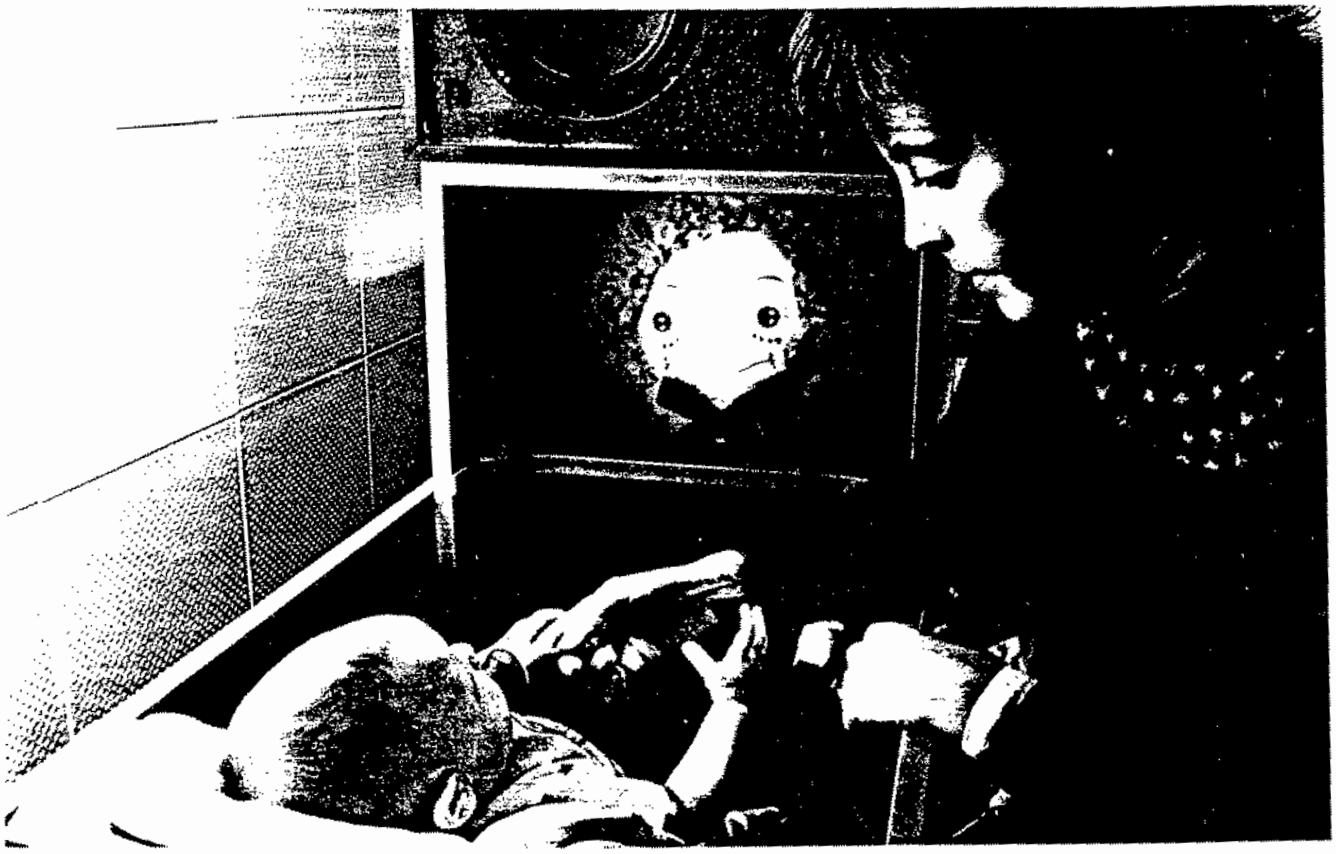
Workers at the Haskins Laboratories in New Haven, the Massachusetts Institute of Technology, Sweden's Royal Institute of Technology and elsewhere have shown that the speech signal is a complex of acoustic units: brief segments bounded by momentary pauses or peaks in intensity. The segments vary in duration and in the frequency, temporal relations and intensity of their constituent bands of concentrated acoustic energy, known as formants, and of noiselike acoustic components known as aspiration and frication. The variation in these acoustic parameters provides the information that is critical to the perception of phonemes.

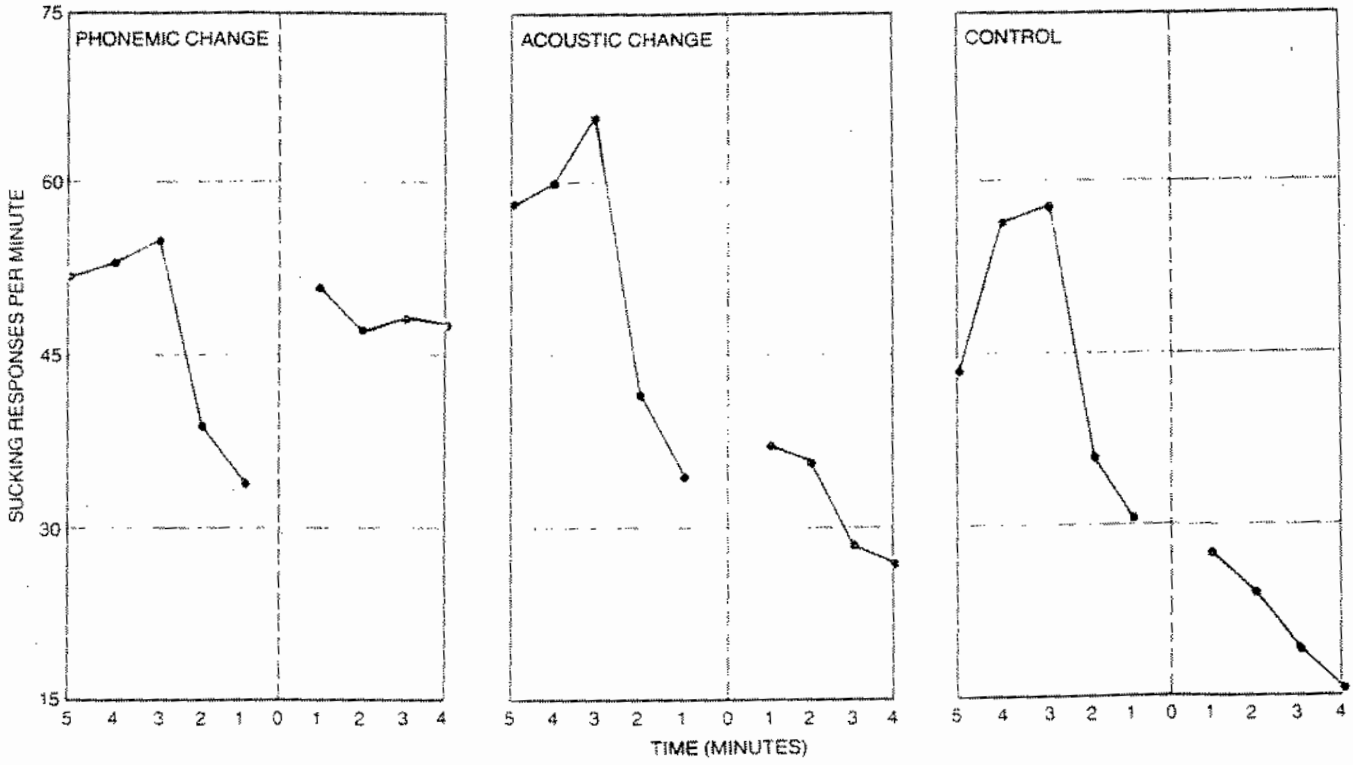
No direct, one-to-one correspondence holds, however, between individual acoustic segments and the phonemes we perceive. A single acoustic segment may encompass a consonant and a vowel; conversely, two distinct acoustic segments may contribute to a single consonantal sound. Furthermore, there is no direct relation between the segments' frequency and temporal characteristics and the phonemes we hear. A listener may recognize a range of stimuli, varying widely in a number of acoustic traits, as instances of the same phoneme. On the other hand, a small change in a single acoustic cue may in some situations change the phoneme that is perceived.

Consider the acoustic information that is sufficient to signal the distinction between the voiced stop consonant that begins the word *bin* and the voiceless stop consonant in *pin*. In both cases the speaker completely blocks the flow of air through the vocal tract just before the release of the utterance; in *bin*, however, the vocal cords begin to vibrate almost simultaneously with the release of air, whereas in *pin* vocal-cord vibration is delayed. The interval between the release of air and the onset of vocal-cord vibration, or voicing, is known as voice-onset time; it holds the crucial acoustic information that enables a hearer to distinguish *bin* from *pin*. No single value of voice-onset time defines each phoneme, however. Instead hearers typically perceive a range of values, reflecting different speakers, different instances of speech and differences in the surrounding phonemic environment, as examples of the same phoneme.

The acoustic variables that define other phonemes are similarly fluid. For example, many phonemes are differentiated by place of articulation, the site of the constriction of the vocal tract that occurs as the sound is formed; the initial sounds of *bin* and *din* are examples. Among the acoustic cues that correspond to place of articulation and enable a hearer to distinguish such phonemes are the initial frequencies of the second and third formants: the formants that fall second and third from the bottom on a scale of frequency. Again no single value of these acoustic parameters characterizes each phoneme; a range of onset frequencies can signal the same place of articulation. Yet in spite of the variation in the sounds corresponding to each phoneme we have little trouble deciding whether a speaker said *din* or *bin*. We are able in effect to listen through the variation in the signal and make categorical judgments of phonemic quality.

Experimental results confirm that in





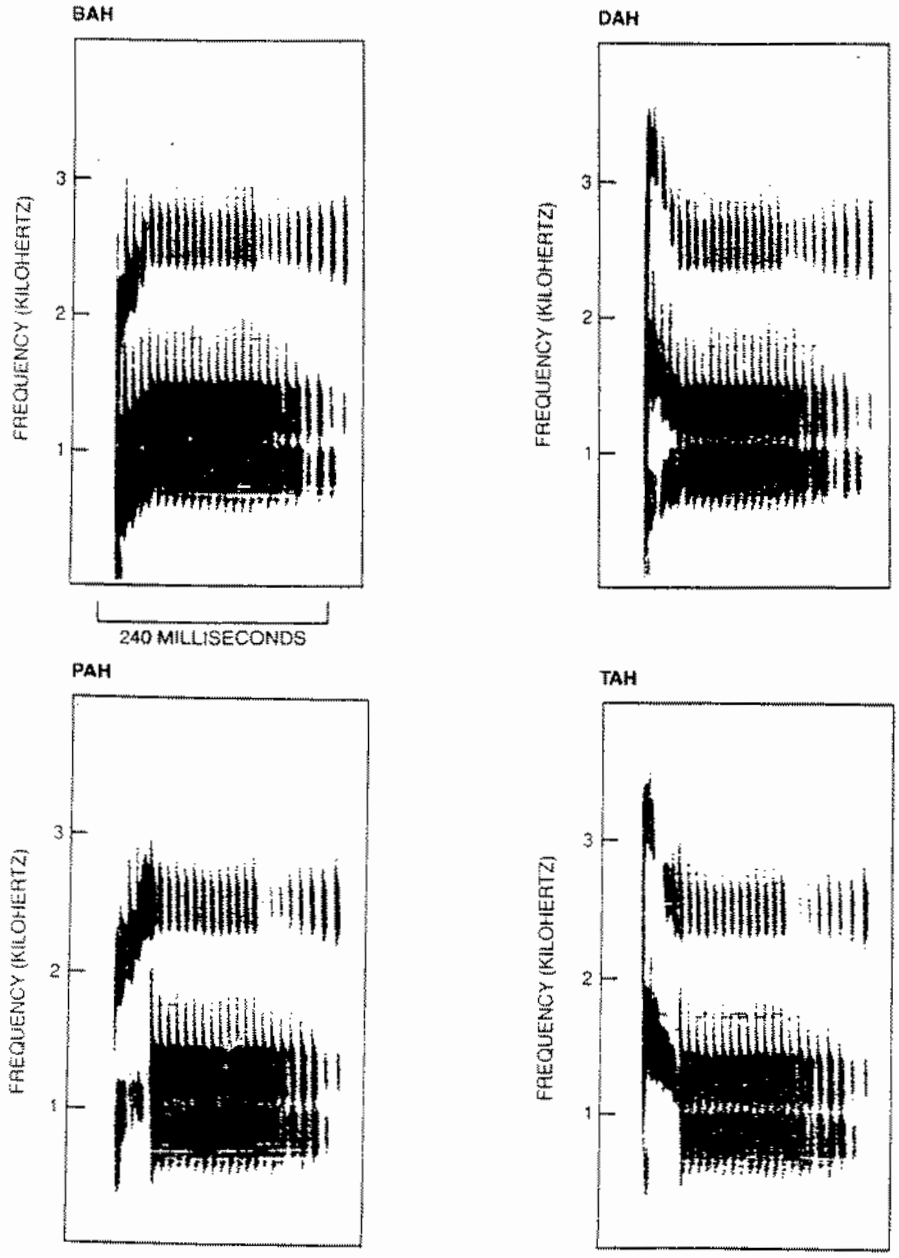
INFANTS' SUCKING RATE indicates their response to a series of speech sounds. In the author's experimental setup (*top*) syllables of synthetic speech were played through the loudspeaker above the screen display of Raggedy Ann while a four-month-old infant sucked on a pacifier connected to recording instruments. Graphs of mean sucking rate (*bottom*), recorded under various experimental conditions with a number of infants, show that when a syllable beginning with a particular consonant was repeated, sucking rate first increased and then decreased as the stimulus became familiar. In some cases the sound changed at a time indicated by the broken line. In one group (*bottom left*) the new sound represented a different consonant: sucking rate increased sharply, showing that the infants perceived a contrast. In a second group (*bottom middle*) the stimulus differed acoustically from the preceding sound but corresponded to the same consonant, and there was little change in sucking rate. A control group of infants (*bottom right*) experienced no change in stimulus.

the perception of speech we are ordinarily aware of discrete phonemic categories rather than of the continuous variation in each acoustic parameter: we perceive speech categorically. In experiments conducted by Leigh Lisker and Arthur S. Abramson of the Haskins Laboratories adults heard computer-generated speech sounds that embodied a range of different values of voice-onset time. In spite of the many variants of voice-onset time the subjects heard nearly all the stimuli either as a voiced phoneme such as the initial consonant of BAH or as a voiceless phoneme such as the consonant that begins PAH. The boundary—the voice-onset time at which listeners began to hear PAH instead of BAH—was situated at about 30 milliseconds following the initial release of air.

To confirm the categorical nature of speech perception the experimenters asked subjects to distinguish pairs of stimuli differing in voice-onset time. If both sounds represented voicing delays of less than 30 milliseconds, the listeners generally heard them as two identical instances of BAH; if the voice-onset times of both were longer than 30 milliseconds, the listeners tended to hear two PAH's, indistinguishable although acoustically different. Only when the stimuli straddled the 30-millisecond boundary could subjects distinguish them consistently. Catherine G. Wolf, then at Brown University, obtained similar evidence of categorical perception in school-age children.

How much of this mechanism of categorical perception, which enables us to perceive speech reliably in spite of the lack of precision of the speech signal, is innate? The fact that speakers of different languages are attuned to somewhat different phonemic distinctions suggests that the influence of the linguistic environment on speech perception is powerful. Japanese speakers fail to perceive the contrast between the phonemes /r/ and /l/, a standard distinction in English; English speakers do not notice a fundamental contrast in voicing that distinguishes certain phonemes in Thai. Yet certain phonemic distinctions are present in languages throughout the world. It seemed possible to my colleagues and me that strong biological determinants, modified by later linguistic experience, might underlie the categorical perception of speech. To find out whether this is the case we did experiments with infants not yet able to speak, in whom one would expect the influence of their parents' language to be minimal.

In 1971 Einar R. Siqueland, Peter W. Jusczyk, James Vigorito and I examined the perception of voice-onset time in one- and four-month-old infants. We exposed the infants to three different pairs of sounds. The voice-onset times of one pair were 20 and 40 milliseconds; thus the stimuli fell on opposite sides of the category boundary recognized by adult speakers of English and other languages. To adult ears the stimuli sounded like the syllables BAH and PAH. In each of the other pairs, with voice-onset times of zero and 20 milliseconds and 60 and 80 milliseconds, both stimuli fell on the same side of the voiced/voiceless boundary; both were instances of BAH or PAH.



SPECTROGRAMS of syllables beginning with different stop consonants, so called because they require an interruption in the flow of air through the vocal tract, show the underlying differences in acoustic characteristics. The four acoustic signatures differ in the frequency and timing of their component bands of acoustic energy, known as formants. The consonants paired horizontally are distinguished by the frequency at which the formants begin, a reflection of the point within the vocal tract at which constriction occurs. The frequency of the highest formant of the sound BAH, for example, begins at about two kilohertz and then rises, while that of the third formant of DAH begins at about three kilohertz and falls. Consonants paired vertically differ in voice-onset time, a measure of the delay between the release of air and the vibration of the vocal cords. In spectrograms for BAH and DAH a voice-onset time of zero is evident in the presence of periodicity, a series of spiky vertical striations that indicate vocal-cord vibration, at the beginning of all three formants. In spectrograms for PAH and TAH there is a gap before the lowest formant appears and periodicity begins in the two higher formants, reflecting a longer delay in onset of voicing.

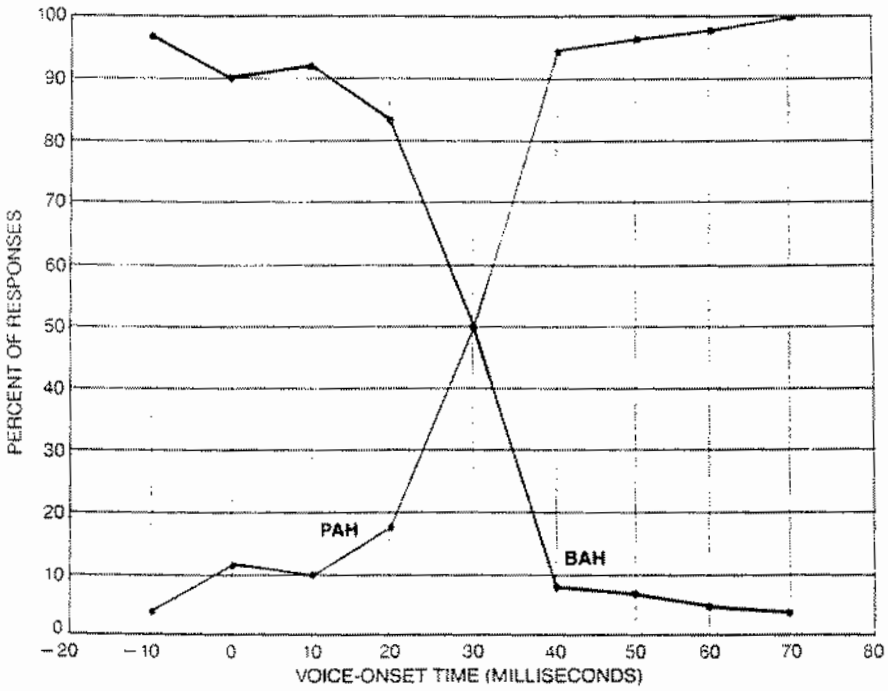
Infants a few months old cannot report their perceptions directly. In order to gauge the infants' responses to the stimuli we resorted to a methodology called the high-amplitude sucking procedure. Each infant sucked on a pacifier wired to a pressure transducer, which in turn was connected to recording instruments. We adjusted the setup's sensitivity separately for each infant so that in every case the apparatus recorded a base-line rate of sucking of 20 to 40 times a minute.

Once the experiment was under way, each time the apparatus recorded an instance of sucking one sound of a stimulus pair was played. When an infant encounters a new stimulus, its rate of sucking typically increases for several minutes, then gradually settles back to the base-line rate, presumably as a result of familiarization. When the sucking rates of our subjects fell to a preset level as they grew accustomed to the first sound, we shifted the stimulus to the other sound of the stimulus pair. If an infant grows familiar with one stimulus and then encounters a stimulus it perceives as different, its rate of sucking ordinarily increases.

The results showed that infants, like people who command a language, perceive differences in voice-onset time categorically. When both the sounds to which an infant subject was exposed lay on the same side of the 30-millisecond boundary, the shift from one sound to another evoked no increase in sucking rate. The infants appeared not to notice the change in voice-onset time. On the other hand, when the stimuli fell on opposite sides of the boundary, a sharp increase in sucking rate occurred at the shift, indicating that the infants perceived a change.

Other investigators and I have discovered further perceptual boundaries in infants' responses to the acoustic information in speech. Like adults, they respond categorically to changes in the onset frequency of the second and third formants, an acoustic cue that indicates differences in the place of articulation of a consonant. The same pattern holds in their responses to the acoustic cues that signal the distinctions between nasal and stop consonants, exemplified by the initial sounds of MAH and BAH, and between stop consonants and semivowels such as the initial sound of WAH.

It is difficult to see how learning could account for the mode of perception we have demonstrated in infants. What events during the first few weeks of life would train an infant to respond categorically to gradations of acoustic properties? A simpler view is that categorization occurs because a child is born with perceptual mechanisms that



CATEGORICAL PERCEPTION is reflected in curves showing the relative proportions of responses when children were asked to identify a synthetic speech sound with a particular voice-onset time as an instance of a voiced (BAH) or a voiceless (PAH) consonantal sound. Instead of a linear change in the percentages the curves show that at voice-onset times of less than 30 milliseconds the children almost always identified the stimulus as BAH; when voice-onset time exceeded 30 milliseconds, they tended to hear the sound as PAH. The perceptual tendency shifted abruptly at 30 milliseconds. The study, done by Catherine G. Wolf at Brown University, suggests that perceptual categories, rather than continuous gradations in the acoustic properties of the speech signal, shape the perception of speech.

are tuned to the properties of speech. These mechanisms yield the forerunners of the phonemic categories that later will enable the child unthinkingly to convert the variable signal of speech into a series of phonemes and thence into words and meanings.

If these perceptual mechanisms do represent a biological endowment, they should be universal. The same perceptual patterns should occur in infants of every linguistic background. In research reported in 1975 Robert E. Lasky, Ann Syrdal-Lasky and Robert E. Klein, then at the Institute of Nutrition in Panama, investigated the perception of voice-onset time by Guatemalan infants, born into a Spanish-speaking environment. The group's experimental methods differed from those used in our 1971 study: in place of changes in sucking rate they used changes in heart rate as the gauge of infants' response to the speech patterns. The study also tested sensitivity to a voicing category we had omitted, one found among stop consonants at the beginning of syllables in Thai and in a number of other languages although not in English. In this so-called prevoiced category the vocal cords begin to vibrate up to 100 milliseconds before the release of air, in a kind of prefatory hum.

Lasky and his co-workers exposed the infants to three pairs of stimuli. In the first pair the voice-onset times fell at 20 and 60 milliseconds after consonantal release: thus the two sounds lay on opposite sides of the voiced/voiceless boundary recognized by speakers of English and other languages, although as it happens not by Spanish speakers. The stimuli in the second pair had voice-onset times of 60 and 20 milliseconds prior to consonantal release and fell on opposite sides of the prevoiced/voiced boundary of Thai. In the sounds of the final pair voicing began 20 milliseconds before and 20 milliseconds after consonantal release. Spanish speakers, in contrast to speakers of many other languages, perceive the voiced/voiceless boundary as falling between those two values.

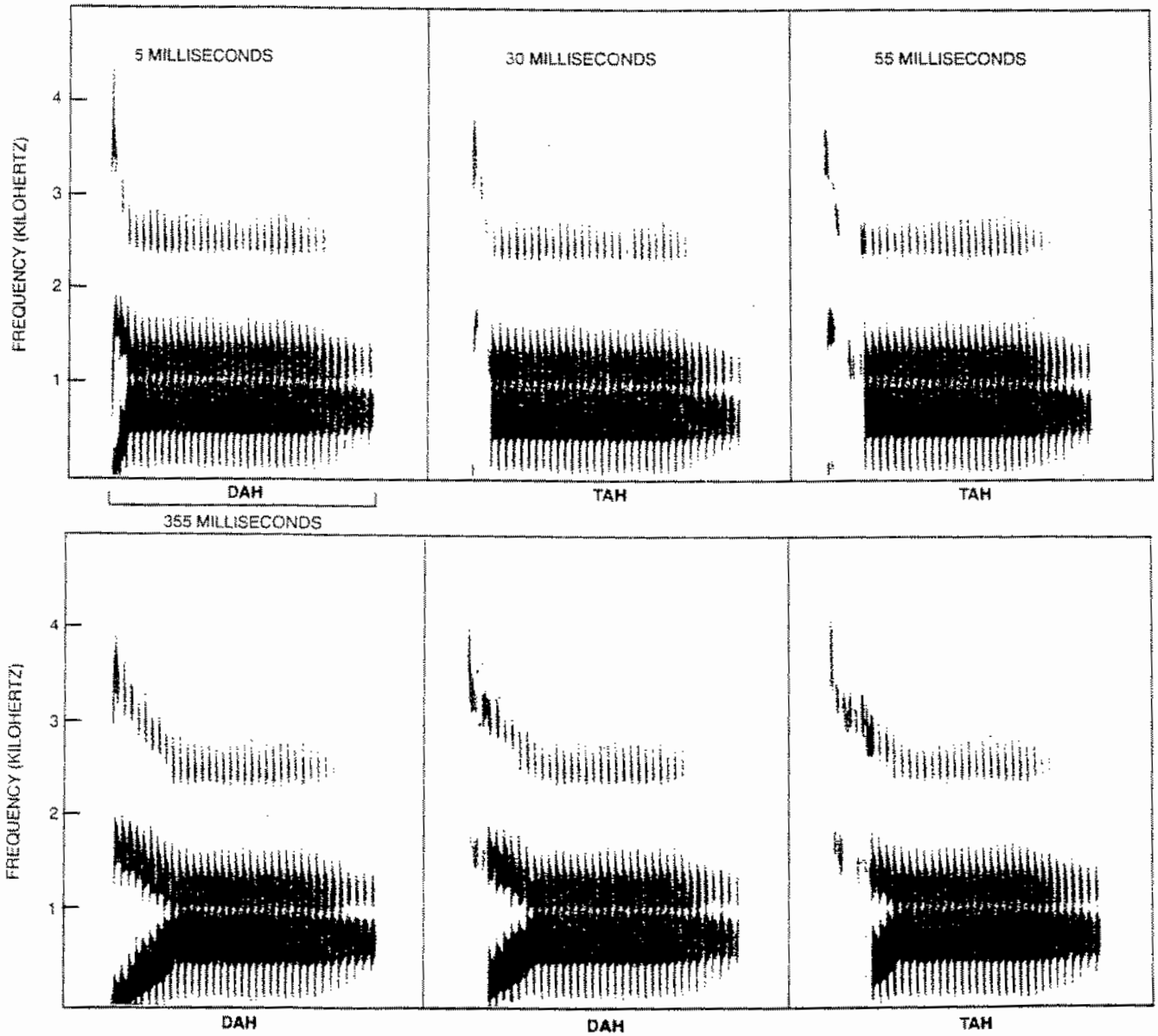
The tracings of heart rate recorded any increases that occurred when the infants, having grown accustomed to the first sound of a stimulus pair, heard the second sound. The data showed the young subjects responded to the prevoiced/voiced distinction, with a boundary between 60 and 20 milliseconds before consonantal release, and also to the voiced/voiceless distinction

with a boundary between 20 and 60 milliseconds following release. The voicing distinction peculiar to Spanish evoked no change in heart rate.

In 1976 Lynn A. Streeter, then at the Bell Laboratories, published evidence that infants born into a Kikuyu-speaking culture in Kenya display much the same perceptual pattern as the Guatemalan babies. Richard N. Aslin, David B. Pisoni, Beth L. Hennessy and Alan J. Perey of Indiana University recently completed the study of voice-onset-time sensitivity among infants in English-speaking communities by showing that they respond to the prevoiced/voiced contrast just as they do to the voiced/voiceless distinction. It appears that infants the world over are equipped with an inborn sensitivity to these three categories of voicing, whether or not the distinctions are important in their parent language.

The perception of speech is a complex and subtle process, which the studies of categorical perception described so far probe only in the simplest terms. The acoustic information that enables a listener to perceive distinctions in voicing illustrates the point. So far we have treated the essential information as a single continuum of time measuring the interval between consonantal release and the beginning of voicing. In ordinary speech, however, an interplay of temporal and spectral factors governs the perception of voicing distinctions. These acoustic properties interact in what might be called perceptual trading relations: a change in the value of one property alters the value of another at which the perceptual boundary falls.

For example, because of functional characteristics of the mechanisms of articulation, the frequency of the first, or lowest, formant rises as voice-onset time increases. Our perceptual system seems to be attuned to the relation: a frequency change can substitute for a



SHIFT IN A PERCEPTUAL BOUNDARY can occur when two acoustic cues are altered independently. The consonants beginning the six syllables shown in spectrographic form vary in voice-onset time and in the onset frequency of the lowest formant; to adult ears the sounds are the syllables DAH and TAH. At a high onset frequency (*top row, visible in second and third spectrograms*) infants detected the DAH/TAH contrast between sounds having voice-onset times of five and 30 milliseconds. When the onset frequency was low (*second and third spectrograms in bottom row*), voice-onset time had to increase to between 30 and 55 milliseconds before the infants reacted to the phonemic contrast. Such interactions between two acoustic variables are known as perceptual trading relations.

change in the temporal cue. When the first formant begins at a higher frequency, the effect is the same as if the voice-onset time had lengthened. As a result, at higher onset frequencies adults perceive the voiced/voiceless boundary earlier in the continuum of voice-onset times.

The same subtleties are apparent in the perceptual systems of infants. In 1983 Joanne L. Miller of Northeastern University and I showed that one perceptual trading relation found in adults also holds in infants' responses. We found that the voice-onset time at which three- and four-month-old infants recognize a shift from the voiced initial sound of the syllable DAH to the voiceless sound of TAH varies with the onset frequency of the first formant.

A second complication in the perceptual process arises from the fact that the category boundaries perceived by adults shift not only as a result of the interplay of multiple cues but also with variations in acoustic context. In this respect as well infants display the forerunners of more mature patterns of perception. Miller and I have shown that infants, like adults, distinguish the stop consonant of BAH and the semivowel of WAH differently depending on the duration of the vowel sound that follows. The acoustic basis of the distinction is the length of the formant transitions: the periods needed for the central frequencies of the formants to reach the values appropriate for the vowel that follows. In the case of BAH the formant transitions are swift; with WAH they are slower. The longer the vowel duration is, however, the slower the formant transitions must be before infants recognize a change in stimulus from BAH to WAH.

Other quite complex effects of context on the categorization of speech by infants have been demonstrated. Jusczyk and his associates at the University of Oregon found a shift in the formant-onset frequencies at which infants detect a distinction between phonemes differing in place of articulation. The boundary value varied depending on whether an additional band of noiselike acoustic energy was present, signaling a fricative rather than a stop consonant.

BABY RECOGNIZES A PHONEMIC CONTRAST in an experiment devised by Patricia K. Kuhl of the University of Washington to investigate infants' ability to distinguish contrasting phonemes from acoustically varied instances of the same phoneme. In this case the baby, its attention held by a toy, ignored variations in speaker and intonation among repetitions of the vowel sound /a/, as in POP (*top*). When instances of the vowel /i/, as in PEEP, interrupted the sequence, the infant turned away from the toy and toward the loudspeaker (*bottom*), indicating recognition of the linguistically important contrast. The sight of a mechanical stuffed rabbit, illuminated in its case on top of the loudspeaker when the contrasting phoneme was played, served to reward the infant's accurate response.

The complex mechanism of categorical perception enables an individual to recognize phonemes consistently in spite of great variation in crucial acoustic parameters. Other kinds of variability blur the definition of the speech signal even further. The length of syllables, along with other temporal characteristics of speech, changes with rate of speech and patterns of emphasis; wide variations in the fundamental frequency of voicing and in the spacing of resonant frequencies occur as a result of the speaker's sex, age and emotional state. Some mechanism must enable us to listen through the variation to hear the same phoneme each time it is spoken. This phenomenon of perceptual constancy cannot be investigated directly in infants. But studies of infants' ability to form equivalence classes—groups of stimuli that evoke the same response in spite of obvious differences—suggest infants possess at least the forerunners of perceptual constancy.

Patricia K. Kuhl and her colleagues at the University of Washington have investigated the formation of equivalence classes for the sounds of speech in six-month-old infants. In the first stage of each experiment the Kuhl group trained infants to turn their head 90 degrees toward a loudspeaker whenever a series of contrasting stimuli interrupted a background sound; the sight of a colorful, moving toy that appeared above the loudspeaker as the contrasting sequence was played rewarded successful responses. In one experiment identical instances of the vowel sound /a/, as in POP, served as the background stimulus; identical versions of /i/, as in PEEP, provided the contrast. Once the training was complete the stimuli were varied: the vowels, /a/ and /i/, remained the same but the infants now heard both vowels in a variety of voices and intonations. Sequences without contrasting stimuli, in which every sound was a variant of /a/, served as controls.

The infants' success in singling out

contrasting stimuli and ignoring within-category acoustic variations during the control trials was impressive. When both inappropriate head-turnings and missed contrasts were counted, they averaged about 80 percent correct: in seven out of eight cases the infants scored better than if their responses had reflected chance. When Kuhl and her colleagues repeated the experiment with the acoustically less distinctive vowels /a/ and /ɔ/ (as in PAW), the infants still could detect equivalent sounds, although less reliably; the proportion of correct responses fell to 67 percent and only four out of eight infants bettered the expected score for random responses.

When both the background and the contrasting sequences included arbitrarily chosen variants of both /a/ and /i/, however, the infants could not learn to differentiate members of the two sequences, in spite of the reward elicited by a correct response. They could not be trained to recognize an arbitrary grouping of sounds that had no linguistic property in common. They could respond correctly, indicating they had organized diverse stimuli into equivalence classes, only when the background and contrasting sequences represented different categories of speech. The finding is further evidence that long before infants can speak and understand they are particularly sensitive to the acoustic distinctions crucial to the comprehension of speech. It adds weight to the case for a set of inborn mechanisms that are specialized for speech perception.

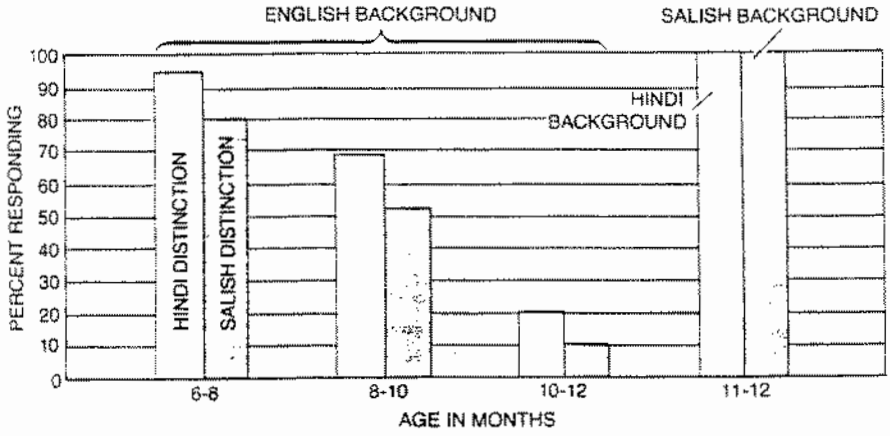
The diversity among the sound systems of human languages makes it clear that environmental factors affect the perceptual dispositions with which an infant is born. What happens as the linguistic environment created by a child's parents and companions interacts with inborn perceptual mechanisms? It appears that perceptual horizons narrow as a child learns his or her native language. The child retains and probably sharpens those perceptual capacities that correspond to phonemic distinctions in the parental language but loses the ability to detect distinctions that do not occur in the native language.

Studies of voice-onset-time perception testify to the decline in some discriminative powers as the infant develops. While infants from diverse linguistic backgrounds respond to contrasts in prevoiced, voiced and voiceless initial consonants, adult speakers of some languages, including English, recognize only the distinction between the voiced and the voiceless categories. Although native adult Japanese speakers are virtually unable to perceive the distinction between the sounds of /r/ and /l/ without special training, I have found that the distinction is among those to which American infants—and presumably Japanese infants—have an innate sensitivity. Similarly, research by Janet F. Werker of Dalhousie University in Nova Scotia and Richard C. Tees of the University of British Columbia showed that six- to eight-month-old infants from an English-speaking background readily distinguish phonemic contrasts in Hindi and Salish, a North American Indian language. When they were tested again at the age of 12 months, the same infants, like English-speaking adults, did not detect the contrasts to which they had earlier been sensitive.

The decline in perceptual abilities through exposure to a restricted environment is familiar. When kittens are raised wearing goggles that limit the visual input of one eye to a series of horizontal stripes and that of the other eye to vertical stripes, corresponding areas of the visual cortex lose their sensitivity to stripes running in other directions. Such losses seem to be irreversible, no matter how varied the animal's later surroundings. In contrast, we can recover at least some of our initial capacities to detect the acoustic information underlying phonemic contrasts. For instance, when the acoustic information critical to phonemic distinctions in Hindi and Salish is embodied in sounds that usually are not heard as speech, English speakers can detect differences to which they are ordinarily insensitive.

Apparently the restricted linguistic environment of one's native language does not inactivate unused perceptual mechanisms completely. We learn to listen primarily for the acoustic distinctions that correspond to phonemic contrasts in our own language. Given the right task or instructions, however, we can detect unfamiliar acoustic distinctions even though we do not perceive them as marking phonemic contrasts. Furthermore, with enough experience the perception of non-native distinctions begins to operate at the phonemic level: after considerable experience with spoken English, native speakers of Japanese can distinguish the phonemes /r/ and /l/ categorically and almost as accurately as native English speakers. The fact that perceptual mechanisms available to us as infants can still operate in adulthood, after long disuse, confounds hypotheses that early experience with language immutably alters some of the mechanisms of speech perception.

The most dramatic demonstration of the innate mechanisms of perception other workers and I have studied, however, takes place in infancy, as a child begins to learn its parents' language. It is now clear that an infant is born with many of the underpinnings of later speech perception and comprehension. It may be that like the specialized anatomy of the vocal tract and the speech centers in the brain these innate perceptual capacities evolved specifically for the perception and comprehension of speech. They are an evolutionary answer to the need for each infant to acquire its parents' language and culture as early in life as possible. The effectiveness of these mechanisms is reflected in the swiftness with which a child joins the community of language.



WANING OF UNUSED PERCEPTUAL POWERS is evident in the responses of infants from an English-speaking background to linguistic contrasts that are foreign to English. When Janet F. Werker of Dalhousie University in Nova Scotia and Richard C. Tees of the University of British Columbia simultaneously tested infants in different age groups, the proportion responding to consonantal contrasts from Hindi and Salish, a North American Indian language, fell rapidly with age. One-year-old Hindi and Salish infants, in contrast, retain the capacity to perceive the linguistic contrasts native to their respective languages.