

10.2  
Lieberman

296 .

10.1

Marlee

Reprinted from M. von Cranach, K. Foppa, W. Lepenies and D.  
Ploog (eds.), *Human ethology: claims and limits of a new discipline*  
© Maison des Sciences de l'Homme and Cambridge University  
Press 1979 Printed in Great Britain



## 10. Ontogeny of auditory perception<sup>1</sup>

---

### 10.1. Development of auditory perception in relation to vocal behavior<sup>2</sup>

PETER MARLER

Of the contributions of classical ethology to behavioral biology, none has had more far reaching consequences than the demonstration of innate factors in behavioral development, and particularly their influence on the ontogeny of responsiveness to environmental stimuli. Among vertebrates the emphasis on innate responsiveness has been especially strong in birds, as embodied in such notions as the 'releaser' and the 'innate release mechanism' – bringing to mind, for example, the now classical studies over a twenty-year period on feeding responses of gulls to visual stimulation (Tinbergen 1951, Tinbergen and Perdeck 1950, Hailman 1967, 1970). In theory, if not always in practice, ethologists have placed equal emphasis on the intercalation of innate factors with learning in the development of bird behavior. Those in neighboring disciplines are nevertheless prone to infer that the responsiveness of birds to stimuli is largely genetically programmed, except in such special situations as imprinting (e.g. Gibson 1977). This is contrasted with the human situation, where innate influences are thought to play a minor, even vestigial, role. According to this view, non-human primates fall somewhere between these extremes. In fact little work has been done on the development of responsiveness in monkeys and apes to stimuli that control their

<sup>1</sup> The comments on this section (pp. 705–10) take account of the fact that auditory perception may be a subject less familiar to readers of this volume, and provide a brief overview of the experimental techniques.

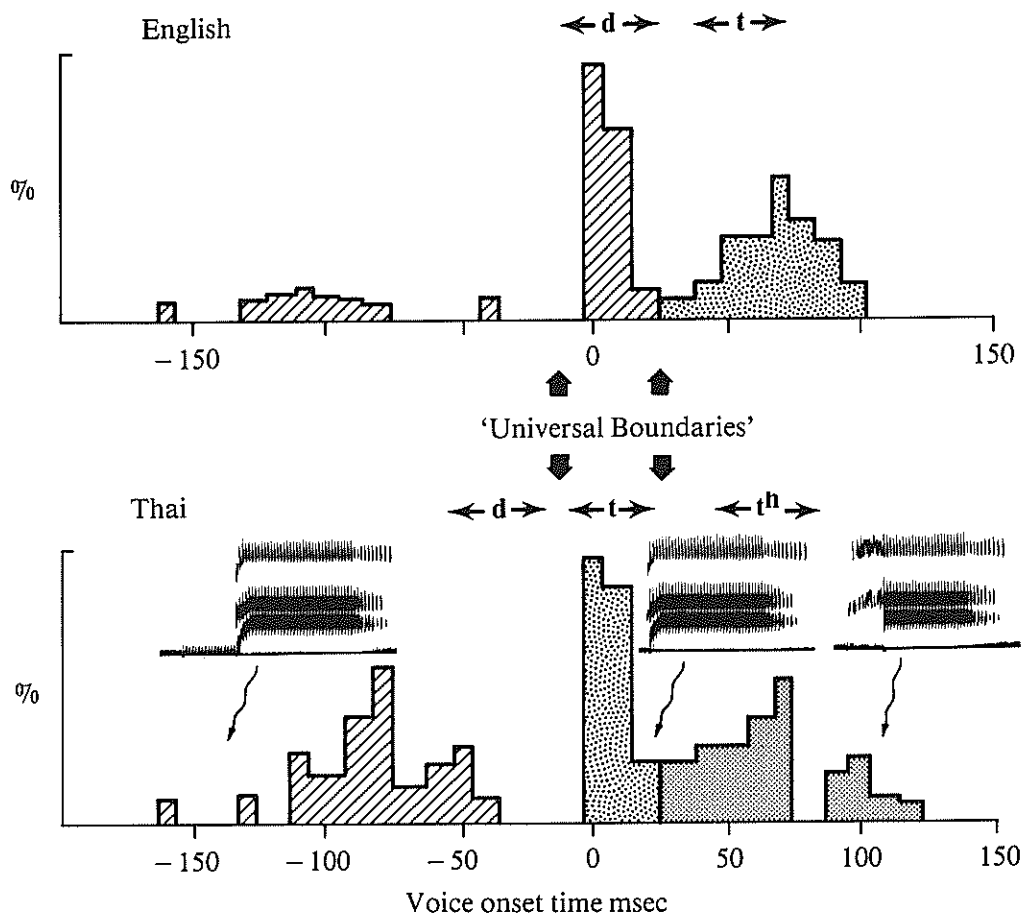
<sup>2</sup> Research for this paper was supported by grants from NIMH (MH14651) and NSF (BNS7519431). The author is indebted to conference participants, especially to Dr Alvin Liberman and Dr Detlev Ploog for discussion and criticism, and to Dr Stephen Zoloth for reviewing the manuscript and adding new material.

behavior in nature, and the emphasis has been on the learnability of responsiveness. Actually there are hints of a significant degree of innate responsiveness, as for example in Sackett's (1966) studies of responses of infant monkeys to two-dimensional representations of facial expressions.

With such exceptions, there has been little attempt to explore the possibility of innate responsiveness of non-human primates to biologically significant stimuli in infancy. Nor has the possibility been explored of any special adult facility in learned perceptual processing of natural sign stimuli, irrespective of the richness or paucity of prior experience of them. The aim of this paper is to present new information on a study in progress on responsiveness of the Japanese macaque both to natural vocal stimuli and to synthetic stimuli molded on the natural vocabulary, and on a study of the intercalation of genetic and environmental influences in the development of responsiveness to vocal stimuli in birds.

The methodology of the animal studies I shall describe was stimulated by that used in recent investigations of the perceptual processing of speech stimuli by human adults and infants. As components in the revolution in our understanding of the perceptual capacities of human infants in the last fifteen years, these investigations have done much to establish a sensible balance between nativistic and empiricist views of the development of human perception. Just as ethology played its part in inspiring new approaches to the study of human behavior, I believe that the logic and sophistication of experimentation on human perceptual development has outpaced progress in ethological studies of animal perception. Ethologists can now learn much of benefit from human studies. I will illustrate this point with a brief review of an impressive, thought-provoking body of data on the development of speech perception that merits close attention from ethologists working on analogous problems with animals.

Approaching recent research on the structure of speech sounds as a novice, I was astonished to discover that cross-cultural descriptions of certain physical features of speech patterns reveal the existence of universals in the properties that define boundaries between some functionally distinct patterns of sounds. I can best illustrate the results from these comparative vocal 'ethograms' by reference to the distinction in many unrelated languages between critical pairs of voiced and unvoiced consonants. I have in mind the property known as 'voice-onset-time' (VOT), a focus of special study since it is one of the few characteristics of speech that can be reliably measured from the frequency-time sound spectrograms on which so many bioacoustical studies are based.



1. Measurements of speech sounds: histograms of voice-onset-times in stop consonants in English and Thai. The large arrows indicate the position of the perceived ‘universal’ boundaries. Inserts show three examples of synthetic speech with VOTs of -150 msec., +10 msec. and +100 msec. (After Lisker and Abramson 1964, Cutting and Eimas 1975.)

An example from English is shown in figure 1. The cross-cultural studies of Lisker and Abramson (1964) have shown that all languages studied employ voice-onset-time as one criterion for differentiating speech sounds, and that, when employed, the boundaries always fall in approximately the same places, at one of two locations (figure 1). Similar universals in speech sounds have been found in the patterns of formant onset that differentiate sounds produced at different points of articulation, labial, alveolar, and velar (e.g. [ba]-[da]-[ga]). There is a long list of

other universals (e.g. Greenberg 1966, 1969, Studdert-Kennedy 1977), but the features of consonants I have mentioned have the advantage that they are specific and lend themselves to precise analysis and experimental control.

When such universals are discovered in ethograms of animal behavior, recurring in separate populations of the same species, and contrasting strikingly with the distribution of vocal dialects, local feeding traditions and the many other divergent traits of local populations, an ethologist is likely to entertain the possibility of genetic developmental controls.

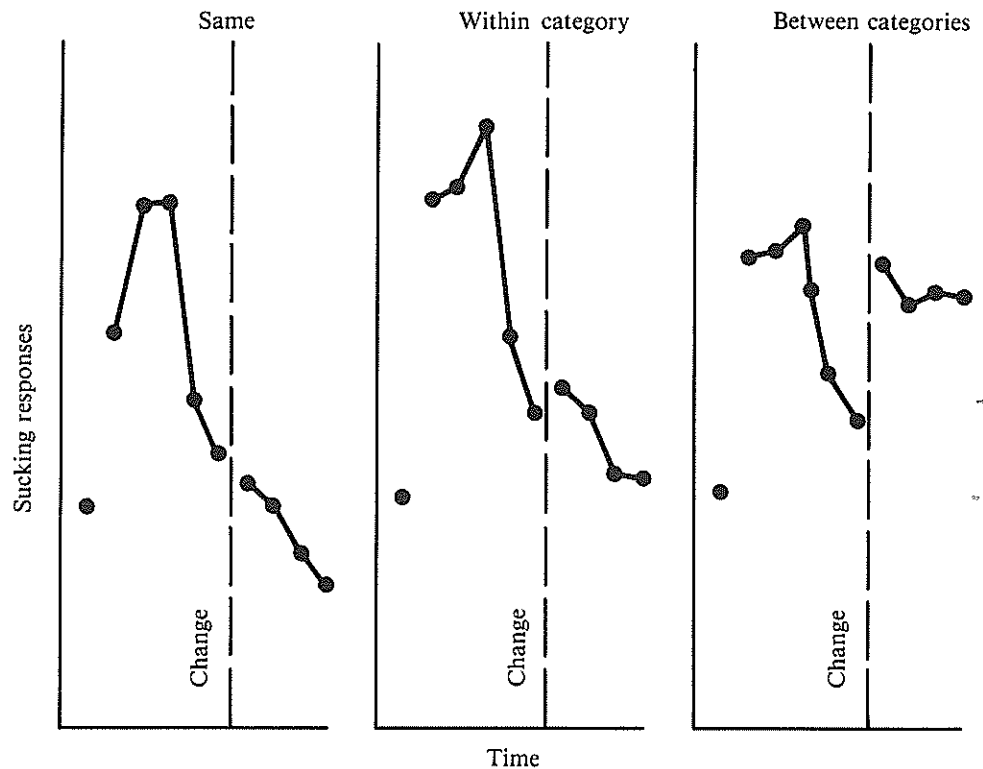
A further revelation to me was that although we hear speech sounds as discretely distinct from one another, they are often distributed in actual speech in 'graded' continua rather than in discretely separate categories. For example, histograms of voice-onset-times used in speech reveal that although the values across a given boundary tend to be grouped separately, such as that on the VOT dimension between [pa] and [ba], there are nevertheless intermediate values. These occur frequently enough to invite us to ask why we are not more often confused as to precisely which consonant a speaker intends (figure 1). This implication was not lost on experimental psychologists, and led to a series of studies on responsiveness of adult subjects to such graded speech sound continua drawn first from natural speech patterns, and subsequently created by computer synthesis, with all characteristics under experimental control (Liberman 1957, Liberman *et al.* 1961, 1967).

Such series as the voice-onset-time continuum, in ten or twenty millisecond steps from [pa] to [ba] to [mba], created by the speech synthesis facilities of the Haskins Laboratories in New Haven, have provided the basis for many insights into the mysteries of speech perception. In particular, they have led to evidence for a distinctive mode of perceptual processing that has become known as 'categorical perception'. Although not unique to the perception of speech sounds (e.g. Cutting and Rosner 1974), nor restricted to the auditory modality (Pastore 1976), it is especially well exemplified by studies of responsiveness of human subjects to complex acoustic continua such as a voice-onset-time series. Asked to label sounds on such a continuum, an English-speaking subject divides the continuum into two parts, labelling one side [pa], the other [ba]. The sharp boundary between them coincides with the trough in voice-onset-time productions. This boundary recurs in different languages, though with subtle details that vary consistently from one to another. In some languages, such as Thai, there is a second boundary, around -20 msec. VOT, shared by the speech pattern of many other cultures.

There is another characteristic of so-called 'categorical perception' of speech sound continua. Adult subjects, asked to discriminate between sound pairs differing by small increments on the VOT continuum, display greater sensitivity to variations in the zone of the boundary than to within-category variations under certain testing conditions (Studdert-Kennedy *et al.* 1970). They behave as though they were desensitized to within-category variations in this particular property of speech sounds, while being acutely sensitive to small changes at the boundary. This contrasts with the 'continuous' perception of other kinds of sound properties such as pitch or loudness. Categorical processing has the consequence of grouping stimuli in classes, imposing a particular kind of order on varying patterns of stimulation by a process of quantization. Although virtually unexplored by ethologists, we should seriously entertain the possibility that animals exhibit analogous perceptual phenomena. It is easy to imagine circumstances in which they could be of value, hence the genesis of some of the animal studies to be described shortly.

Much of what I have described about adult perception of speech could be thought of as a consequence of the rich perceptual and motor experience of speech that any adult brings to bear on a given task of speech-labelling or discrimination. Another set of recent findings suggests that special perceptual predispositions are also involved, irrespective of prior experience with speech. A variety of human infant response measures, including habituation of a sucking response, heart-rate changes and evoked brain potentials, indicate responsiveness to similar boundary values between functionally distinct speech sounds in subjects as young as one month of age (e.g. Dorman 1974, Eimas *et al.* 1971, Eimas 1975, Morse 1972, Wood, Goff and Day 1971). Figure 2 illustrates the kinds of results leading to this interpretation, derived from the work of Eimas. The early age at which these results are obtained led to the speculation that responsiveness to some of these boundary properties may be innate.

Evidence for an innate component was obtained by Lasky, Syrdal-Lasky and Klein (1975) in studies of speech perception of four-to-six-month-old infants living in a Spanish-speaking environment. There are slight but consistent differences in voice-onset-time boundaries in adult production and perception in English and Spanish. These led to the prediction that infants would demonstrate boundary limits different from those obtained by Eimas with children living in English-speaking environments, if these were acquired through infantile experience of speech patterns. The infants proved to be responsive to boundaries in both



2. A typical result of an experiment by Eimas and his colleagues on perception by a four-month-old infant of synthetic speech sounds varying in voice-onset-time. With repeated playbacks of a given sound triggered by the sucking response, habituation occurs. A new sound is then substituted. If the new sound is the same as the old one there is no change (same). If the new one is different, it evokes small or large response increases depending on whether it is on the same side (within category) or the opposite side (between categories) of the 'universal' VOT boundary at about 30 msec. (After Eimas *et al.* 1971.)

regions of the VOT continuum that are universals, the so-called 'English' and the 'Thai' boundaries, with no sign that experience of the distributions used in Spanish had affected their speech perception.

An innate component is implied by a study of Streeter (1975), although with evidence of acquired components as well. Infant perception of boundaries along the VOT continuum was studied in children exposed to Kikuyu in infancy. This language has only one labial stop consonant, with a VOT of about -60 msec. Perhaps as a consequence of exposure to the pattern of usage, two-month-old-infants were responsive to one bound-



ary along the VOT continuum somewhere between 0 and -30 msec. They seemed more responsive to this boundary than the subjects Eimas had studied in an English-speaking environment. However, the Kikuyu-exposed infants, although lacking experience of anything equivalent to a [p], proved responsive to a boundary somewhere between +10 and +40 msec., thus resembling infants exposed to English and various other languages. Streeter concluded that there is evidence of interaction between nature and nurture, and that some phonetic or acoustic discriminations may be universal whereas others seem to require or are reinforced by previous relevant exposure.

Other studies have demonstrated responsiveness in infants between one and six months of age to the variations in second- and third-formant transitions in synthetic speech patterns that establish boundaries between the different articulation points distinguishing labial, alveolar and velar stop consonants. Infants also seem responsive to differences between vowel sounds.

The potential lability of predispositions that human infants may bring to segmentation of speech sound continua is clear. The [ra]-[la] distinction that Japanese adults find so difficult, unemphatic in Japanese, is probably easier for infants, though only American subjects have been tested thus far (Eimas 1974). However, even though the stimulus patterns on which learned responsiveness in adulthood is based are likely to be more complex than those of infants, with more redundancy, perhaps involving configurational features, and sometimes so changed that the effective stimulus set no longer contains those that match the original predisposition, the latter must surely play an ontogenetic role in setting the trajectory for learning to respond to a more elaborate array of abstracted features.

Such possibilities are indicated in a study by Kuhl and Miller (1975). The formant patterns that distinguish different vowel sounds are complicated by variations in the fundamental frequency of different voices, likely to be a serious distraction for an infant learning to respond to speech. Given the importance of vowel coding in speech, we might expect a predisposition to focus more strongly on formant patterns than on pitch in early responses. By independently varying the two features in sounds presented to infants, Kuhl and Miller (1975) were able to show that variations in vowel pattern are indeed more salient or arresting for human infants than variations in pitch. This is not to say that they are unresponsive to pitch variations. However, the salience of pitch is lower than that of variations in vowel patterns, thus imposing some order in the

process of learning to extract different features from the complex array of stimuli that speech sounds present.

Below I have summarized some of these findings about speech patterns, and speech perception in adults and infants, that seem to me of particular interest to biologists. They show that the human organism brings some well-defined perceptual predispositions to the task of developing responsiveness to the complex of sound stimuli that speech represents. Some are manifest in initial encounters, and are thus developed without prior experience of the stimuli involved.

- (1) There are cross-cultural universals in acoustic properties defining boundaries between functionally distinct speech sounds.
- (2) Some functionally distinct speech sounds are not discretely separated but connected by a continuous series of graded intermediates.
- (3) Adults process graded speech sound continua 'categorically', by reference to boundaries, rather than 'continuously'.
- (4) Pre-speech infants are sensitive to some of these same 'universal boundaries'.

However early in human development such predispositions are manifest, we are hardly likely to view them as developmental instructions for designing infants as automata. Instead it seems natural to think of them as initial instructions to set the trajectory for development of learned responsiveness to a more elaborate array of abstracted features. Eventually these are embodied in the centrally generated 'schemata' invoked by many psychologists in conceptualizing the development of human perceptions of complex stimuli (Marler 1977). I now want to present animal data from current experiments by Stephen Zoloth and myself and collaborators at the University of Michigan's Kresge Hearing Institute, suggesting that there are parallels in the perception of conspecific vocal sounds by both monkeys and birds.

One reason I was so intrigued to learn of the graded nature of speech sounds is that grading has proved to be an interesting characteristic of sounds of several higher primates. In addition to the rhesus monkey (Rowell 1962, Rowell and Hinde 1962), it has now been described in several other monkey species, as well as the chimpanzee (Marler 1976). Even some species originally thought to have discretely organized vocal repertoires, such as the squirrel monkey, are now known to exhibit more grading than had been originally thought (Winter, Ploog and Latta 1966, Schott 1975). It is virtually impossible to assay the communicative function of such sounds until we have some understanding of how they are

processed during perception. The Japanese macaque has proved to be an ideal subject for further pursuit of this problem.








In a thorough study of the usage of sounds in wild Japanese macaques in relation to the circumstances of the vocalizer, Green (1975) has subjected the entire vocal repertoire to exhaustive analysis. Sound patterns intergrade freely in many parts of the repertoire. By subjecting the sounds to an arbitrary acoustical taxonomy, Green was able to show that even subtle variations in fine structure correlate well with varying circumstances of production, thus potentially encoding information of value to companions. One subsystem consists of a variety of coos, data for which are shown in figure 3.

A feature identified as significant by Green is the temporal position of a frequency rise, which may occur at any point in the coo. Early and late positions correlate with different circumstances of production. So-called 'smooth early highs' (SE) are contact coos given by isolated animals, by individuals in subgroups separated from the main group, or by young animals separated from regular companions within the group. Vocalizers are usually relatively calm, and smooth early highs seem to function mainly to maintain group cohesion.

Animals producing smooth late highs (SL) are more highly aroused. Though again the mood is affiliative, here the vocalizer is actively soliciting contact, as for example in the sexual solicitation of oestrous females in early stages of consortship. The call is typically given by a subordinate towards a dominant (see figure 3). A careful analysis of the position of 'highs' in natural usage reveals a distribution reminiscent of that for voice-onset-times, on the [pa]-[ba] continuum, for example (figure 4). As such it lends itself to similar kinds of questions about the perceptual processing by a species of its own vocal signals (Zoloth and Green, in press).

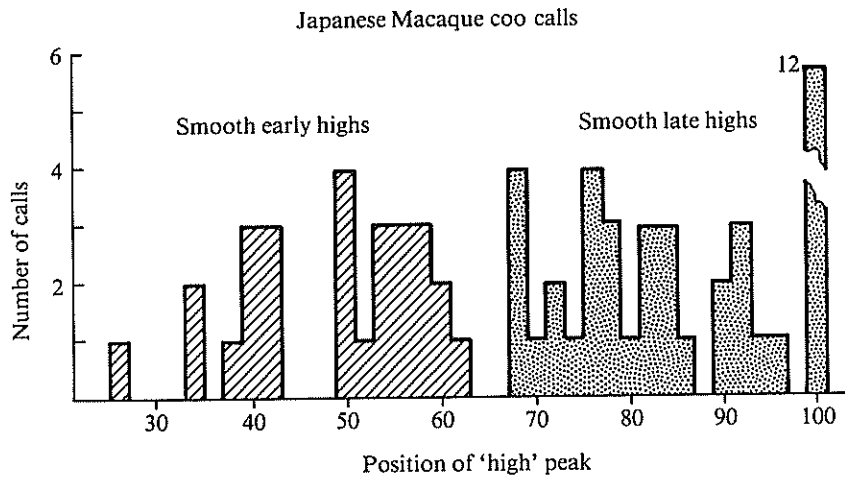
In the first publication from what is planned as a series of studies, Beecher *et al.* (1976) have trained both Japanese macaques and other monkeys in the laboratory to respond to playback of different classes of coo calls as cues, with one class as positive, the other negative. Eventually the monkeys will be exposed to intermediate forms, both natural and synthetic. Species differences in rates of generalization to new members of the classes of smooth early and smooth late highs bear directly on the theme of perceptual predispositions that are involved in learning to respond to biologically significant stimuli.

As in speech so in these monkey sounds certain physical features such as pitch and spectral composition vary in different renditions and from

Coo type		Distinguishing criteria			
	Name	Midpoint pitch	Position of highest peak	Duration	Other features
	Double	$\leq 510$ Hz	N.A.	N.A.	Two overlapping harmonic series
	Long low	$\leq 510$ Hz	N.A.	$\geq 0.20$ sec	N.A.
	Short low	$\leq 590$ Hz	$\neq 1$	$\leq 0.19$ sec	N.A.
	Smooth early high	$\geq 520$ Hz	$< 2/3$	N.A.	No dip
	Dip early high	$\geq 520$ Hz	$< 2/3$	N.A.	Dip
	Dip late high	$\geq 520$ Hz	$\geq 2/3$	N.A.	Dip
	Smooth late high	$\geq 520$ Hz	$\geq 2/3$	N.A.	No dip

		Type of coo vocalization								
		Low			Early High		Late High			
		Double	Long	Short	Smooth	Dip	Dip	Smooth		
Situation	Separated male	xxxx xxxxxx xxxxxx		xx	xx		xx	xx	26	
	Female minus infant	xxxx	xxx						7	
	Non-consorting female		xxxxxx	x					9	
	Female at young		xxxxxx	x					11	
	Dominant at subordinate			xxx xxxxxx			xx		12	
	Young alone				x xxxxxx	xxxxxx	x		23	
	Dispersal				x xxxxxx	xx	xxxxxx	xx	17	
	Young to mother				xxxx	xxxx	xxxx	x xxxxxx	26	
	Subordinate to dominant				xxx	xxxx	xxxx	xxxxxx	43	
	Oestrous female					x	xxxx	xxxxxx	52	
		24	19	14	32	31	56	50		

3. A taxonomy of different types of 'coo' calls of the Japanese macaque, together with a table of frequency of usage of each type in a variety of social situations. (After Green 1975.)

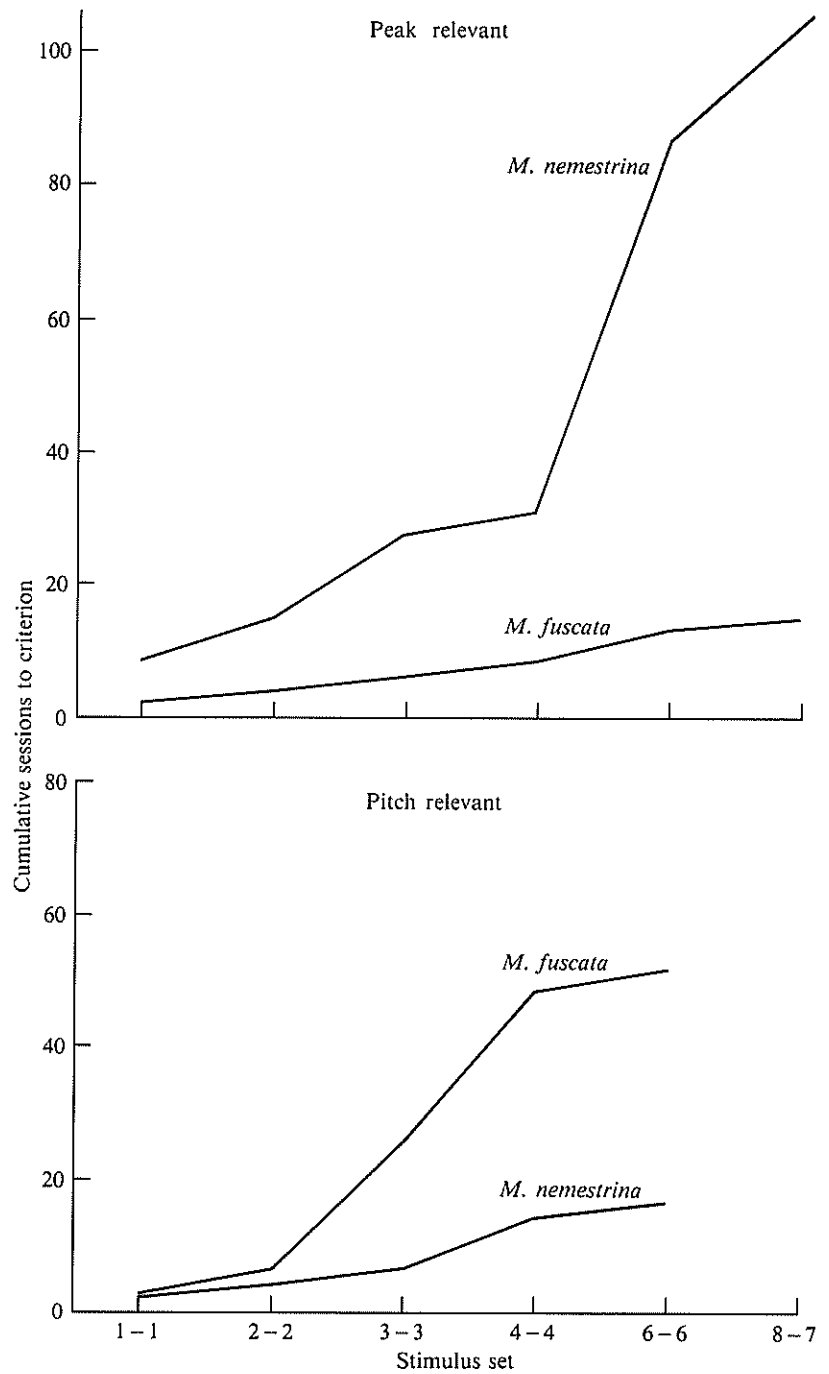


4. Distribution of frequency-peak positions in field recordings of Japanese macaque 'coo' calls. (After Zoloth and Green, in press.)

individual to individual. This variation seems to slow down the rate with which generalization occurs to new members of the two classes, when these are characterized by high position. Nevertheless, Japanese macaques proceed to achieve a high criterion of performance quite rapidly. The next step was a comparison with performance of two other species of monkeys with the same sounds. One, the vervet monkey, does not use coos. The other, a pig-tailed macaque, does have coo-type calls, but the details of its usage are unknown.

Members of these two control species had enormous difficulty in generalizing to new tokens of the two coo classes – smooth early highs and smooth late highs – in this situation (figure 5). This result is consistent with the notion that the SE-SL distinction is a conceptually relevant one for the Japanese macaque, and thus easy to demonstrate, but an alien one for the other species. Nevertheless, by exhaustive training the control species were eventually able to reach a similar level of performance to the Japanese macaques, showing that the task is not impossible for them, just harder.

In subsequent experiments the ability of Japanese macaques and the control species to group coo calls on the basis of either high position or pitch was compared. First two groups of coos were selected, smooth early



5. Generalization rates of two monkey species while being trained to discriminate between sets of natural calls that differ in frequency–peak position. With the ‘peak relevant’ task they had to discriminate on this basis while ignoring variations in starting pitch. The ‘pitch relevant’ task imposed the opposite requirement, obviously more difficult for *M. fuscata*. (By courtesy of Stephen Zoloth, after Beecher *et al.* 1976).

and smooth late highs, carefully counterbalanced for other acoustic variables such as pitch and duration. In this experiment, which replicated our previous work, the three Japanese monkeys acquired the discrimination faster than the three controls.

Next, we used the *same* coos but this time sorted on the basis of pitch. The task for the animals was to distinguish high and low pitched calls, and ignore high position. Since each group contained both smooth early and smooth late highs, the position of the peak was not relevant to the task. The result was in complete contrast to the previous experiment. The Japanese macaques acquired the discrimination more slowly than the other species. The pig-tailed macaques, for example, had no trouble at all in classifying the Japanese macaque calls on the basis of starting pitch, a relatively simple cue. Thus it appears that Japanese macaques are better able to classify groups of coos when they are sorted by peak position than by pitch, while the opposite was true for control species.

We assume that learning proceeds fastest when the discriminative stimuli differ consistently along dimensions that are meaningful to the subjects, and that for Japanese macaques classification according to the position of the 'high' is an easier conceptual task. This reflects a predisposition to process the coos in a way that parallels their apparent meanings. Of course it is too early to tell whether innate perceptual mechanisms are involved. All subjects were wild-caught, with a history of experience of reception and production of such sounds, experience which the control species lack. Whatever the developmental basis, the stimuli are clearly not equivalent to conspecific and alien adults, as classical learning theory would have led us to expect. Further research will tell if there is proneness to categorical divisions of these acoustic continua, and whether some way can be found to ask similar questions of infant monkeys.

Biological approaches to animal learning have in fact called several assumptions of learning theory into question in recent years, especially the principle of equipotentiality (Seligman and Hager 1972, Shettleworth 1972, Hinde and Stevenson-Hinde 1973). A feature of the song learning process in which many male oscine birds engage in youth is its selectivity, such that a male presented with a natural choice of songs of different species to copy will selectively learn conspecific models. I present now an example which parallels the results of human infant studies in some respects, providing a more viable comparison than any non-human primate studies yet available on the ontogeny of auditorily controlled behavior.

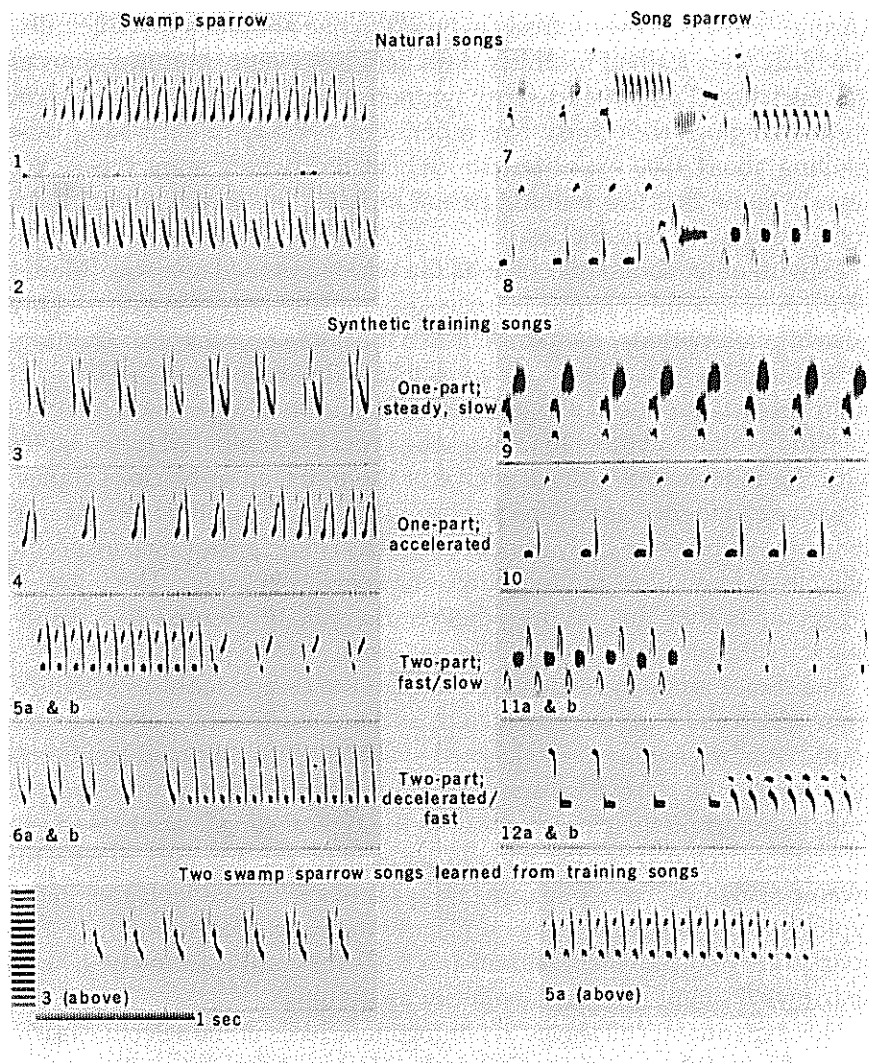
Song and swamp sparrows are closely related congeneric species whose male songs, although similar in duration, are very different in their temporal organization or 'syntax'. The simple song of the swamp sparrow consists of a slow trill of similar slurred liquid notes. That of the song sparrow is much more complex, with several distinct parts, consisting of many short diverse notes and a trill near the end (figure 6). Within these different, relatively stable, species-specific, syntactical patterns, both exhibit much individual variability in the acoustic structure of the so-called 'syllables' from which the songs are constructed.

Although the preferred micro-habitats of the two species differ, they are very often within earshot. Both engage in song learning, and the singing behavior of males reared in social isolation is significantly abnormal. Yet there is no evidence that the two species learn one another's song under field conditions, thus setting the stage for this experiment on selective learning (Marler and Peters 1977).

Our aim was to present male swamp sparrows in youth with both swamp and song sparrow songs to see if selective learning occurred. If so, we sought also to specify some of the acoustic parameters on which the selectivity is based. For this purpose, series of artificial songs were created by editing out distinctly different 'syllables' from tape recordings of normal local song of the two species and then splicing them together in a variety of simple but artificial syntactical patterns. These were chosen to explore the possible significance of some of the organizational features by which normal songs of the two species differ. Thus 'swamp-sparrow-like' patterns included sequences of identical syllables at various steady rates. 'Song-sparrow-like' features included variable rates of delivery of syllable sequences (accelerating, decelerating) and a multipartite structure (two parts), all features present in song sparrow song and lacking in that of the swamp sparrow. Ten such patterns were created, using sixteen different song sparrow syllables. Then an equivalent set was created from swamp sparrow syllables, again sixteen in all. The syllable types were sufficiently distinct that if imitation occurred we would be able to determine the temporal pattern from which each had been selected. Although there were more details to the design of stimuli, this outline will serve to illustrate the result which was striking and, to us, unexpected.

In the first experiment eight male swamp sparrows were taken from wild nests between three and ten days of age and reared by hand, together with female age mates, in small groups in acoustically shielded chambers. Song sparrows of a similar age were also present during





6. Sound spectrograms of natural song sparrow and swamp sparrow songs and artificial training songs. Natural songs are shown at the top. Syllables from these and others were assembled in synthetic songs, some created from swamp sparrow syllables (e.g. 3-6), some from song sparrow syllables (e.g. 9-12), some in 'swamp-sparrow-like' patterns (e.g. 3, 4, 9, 10), some in 'song-sparrow-like' patterns (e.g. 5, 6, 11, 12); syllables from songs 1 and 2 can be seen in songs 3 and 4, syllables from song 8 in songs 10 and 11. Only swamp sparrow syllables were learned. At the bottom are two songs of male swamp sparrows copied from training songs 3 and 5. A 1 sec. time marker is given at the bottom left, with a 500 Hz interval frequency scale.

rearing, so that they were freely exposed to one another's juvenile calls. The birds were trained for thirty days between twenty and fifty days of age. Each heard a set of the twenty synthetic songs arranged in bouts as in normal singing, with thirty-two repetitions per day of each song type totalling about 1000 exposures to each of the twenty song types. We already knew that this training period includes the sensitive period for vocal learning in this species.

As with other songbirds, such as the white-crowned sparrow (Marler 1970), male swamp sparrows learn to sing from memory. Kept on roughly normal photoperiods, they came into song some months after training. After songs had crystallized we were able to determine that the group of eight subjects produced altogether nineteen syllable types. As in nature each individual had more than one song type. We compared the syllables with the models and judged twelve of the nineteen to be close copies. Every one of them was a swamp sparrow syllable. Thus the male swamp sparrow exhibits extremely selective vocal learning, accepting conspecific syllables for imitation and rejecting song sparrow syllables. The interesting point is that this occurs whether they are presented in swamp-sparrow-like or song-sparrow-like patterns.

The choice is clearly made at the level of the components or 'syllables' from which the song is constructed, and not the overall pattern of the song. Thus while four of the learned syllables were extracted from one-part songs, the normal swamp sparrow pattern, eight were extracted from two-part models, much closer to the typical song sparrow pattern. Five of the accepted models were in series with a steady rate, normal for swamp sparrows, but seven came from a series with a variable rate, more typical of song sparrow patterns. Clearly the song syllables of these two species are not equivalent stimuli as a basis for vocal learning of swamp sparrows brought into the laboratory as nestlings.

This experiment still leaves the ontogenetic basis of the selective learning in doubt. What of the possibility that the few days of life in the nest in the wild before capture could provide a basis for selectivity? To test this possibility eggs were removed from the nests of wild swamp sparrows early in incubation and, with some difficulty, hatched and reared under canaries in the laboratory. These cross-fostered subjects were then trained from thirty to fifty days of age with synthetic songs like those used in the previous experiment. All subjects behaved similarly, selecting only swamp sparrow syllables for imitation, showing that an innate predisposition is involved. Obviously there is much in common with speech perception in human infants. In fact, many parallels can be struck between

avian song learning and the development of the perception and the production of speech in human infants, as summarized below.

- (1) Motor learning has dominant role in developing patterns of sound production
- (2) Learning results in local vocal variants (dialects).
- (3) All species members share some species-specific vocal characteristics (universals).
- (4) Selective responsiveness to species-specific sounds during vocal learning (templates).
- (5) Learning occurs most readily in certain life stages (sensitive periods).
- (6) Extrinsic reinforcement (e.g. social or food) not prerequisite for vocal learning.
- (7) Early deafness affects vocal development more than late deafness.
- (8) Hearing important for access to external models and to monitor own vocalization (template matching).
- (9) Progression from highly variable to more stereotyped sounds during vocal development (subsong and babbling).
- (10) Lateralization of neural control (hypoglossal and hemispheric dominance).

Just as young of our own species are predisposed to respond selectively to particular aspects of speech sounds before themselves speaking, so some young songbirds are responsive to species-specific features of song before they themselves begin to sing. In both cases initial responsiveness is manifest to relatively simple, elementary properties with full appreciation of more complex aspects of adult sounds remaining to be shaped through learning. Such perceptual predispositions are valuable as biological constraints on the vocal learning process, serving to focus the young organism's attention on an appropriate set of complex sounds, and on particular properties that they exhibit. In the birds' case, they are guided to a set of conspecific models, focussing attention on a particular subset of properties that they exhibit, sufficient to reduce the potential hazard of learning the wrong song. This is achieved without sacrificing the ability to learn more complex features of natural song. Human infants stand to benefit not only from being encouraged to attend closely to sounds of speech, but also from guidance in embarking on its perceptual analysis. Speech sounds are enormously complex, and there is still controversy about which of the multitude of acoustic features exhibited are in fact the best purveyors of meaning. It could only benefit the infant to have guidance in the extraction of particularly meaningful features from the multitudes of varying reliability that speech presents.

I see great promise in unifying such classical ethological concepts as the 'releaser' and the 'innate release mechanism' with psychological concepts concerning perceptual development such as that of learned 'schemata'. To postulate innate responsiveness to certain stimuli in the young organism is by no means to commit it to a life of behavioral automaticity. On the contrary, it is likely that in animals, as in man, innate responsiveness in infancy will often become so heavily overlain and transformed by learning in the transition from youth to adulthood that its consequences will be difficult to detect. Nevertheless I believe that the consequences for behavioral ontogeny are likely to be considerable, tending to guide the young organism along certain species-specific developmental trajectories without necessarily sacrificing the many advantages that accrue from behavioral plasticity.

Such guidelines are likely to be especially important in the development of communicative behavior. While solitary behavioral innovations may be of immediate value in certain domains, such as feeding behavior, in communication there are special conditions to be satisfied before innovations can become effective. Participants must share some common rules in their behavior. I believe that, in our own species, innate constraints on development of the perception of stimuli generated by signaling behavior must aid in achieving this end, while still allowing the extraordinary diversity of culturally-determined behaviors that is diagnostic of the human condition.

### Summary

A brief 'ethologist's' review has been presented of data on the development of speech perception in human infants, adult perception of natural and synthetic speech sounds, and descriptive 'ethograms' of the acoustic structure of speech. These demonstrate that there are universals in the placing of boundaries on the acoustic continua that occur in natural speech and that infants show innate predispositions to observe similar boundaries. The theme of perceptual predispositions brought to bear on learning tasks was then extended to animal studies. Japanese macaques confronted with discrimination tasks in which the cues are natural or synthetic calls of their own species learn to generalize much more quickly than other monkey species trained with the same Japanese monkey calls. Japanese monkeys seem predisposed to process their calls in the laboratory in ways that parallel their apparent natural meanings. Some songbirds show an innate selectivity in accepting acoustic models for song

learning. In human and animal studies initial responsiveness seems to be focussed on elementary sound properties, with full appreciation of more complex aspects of sound signals of the species remaining to be shaped through learning. Innate perceptual predispositions guide the young organism in learning to extract meaningful features from complex sound stimuli. They will thus encourage the sharing of simple perceptual rules by species members, facilitating communication without sacrificing the advantages that accrue from the developmental plasticity of signalling behaviors and their perception.

## 10.2. An ethological approach to language through the study of speech perception<sup>1</sup>

A. M. LIBERMAN

### Introduction

It is, I hope, appropriate to the purposes of this volume that my approach be the reverse of that taken by Peter Marler (chapter 10.1). Where he begins with the biology of communication in animals and looks toward man, I would begin with the biology of language in man and look toward the animals. I should emphasize that my aim is to complement what he has said, not to contradict it. Indeed, there is nothing I would want to contradict, for I find in his contribution the best hope we have for understanding certain aspects of human communication. I think especially in this connection of the seminal research on the learning of song by certain birds. That work has greatly enlightened us about the acquisition of language by children; more so, by a striking irony, than most of those vastly more numerous studies of language learning in humans that investigated the memorization of lists of unconnected (or unnaturally connected) words. Perhaps there is a lesson here for us human ethologists, which is that we can learn about language from birdsong if only because both are systems with biological function and biological integrity, whereas the rote learning of lists of words is not. But I will say little more about the work Peter Marler has described. I will only express my admiration, acknowledge my debt, and then take my own stance, which is, as I said, 180 degrees away.

<sup>1</sup> The preparation of this paper was aided by a grant to Haskins Laboratories from the National Institute of Child Health and Human Development.

### **Requirements for an ethology of language**

To study language from an ethological point of view, we should meet at least two requirements. The first is to establish that language does have an ethology worth studying, for if we adopt certain views about language, we should conclude that it does not. Thus, we might suppose that language is an invention, as far removed from its biological base as the kinds of things people do when they build and use automobiles and sewing machines. That view is uncommon, perhaps, but we should understand nevertheless that there are aspects of language for which it is exactly right. Written language, for example, *is* an invention of sorts and, accordingly, not very interesting from an ethological point of view. In spoken language, on the other hand, an ethologist will surely find phenomena that are close to their biological roots, but he will just as surely encounter others that, like written language, represent cultural artifacts. At all events, the ethologist must make his way carefully, seeking out the former and avoiding the latter.

There is another, far more common way of looking at language that would also make us hesitate to study its ethology. In this view, language is seen, not as unnatural, but as secondary, the epiphenomenal result of more basic processes. Language has been so regarded by many psychologists, including some who are of very different, even opposite, theoretical persuasions. Thus, from the view of a 'cognitive' psychologist like Piaget, language is an aspect of those same processes that underlie cognitive activities in general (Piaget 1968). At the other extreme of psychological theory, a behaviorist like Skinner treats language as another set, albeit a large one, of conditioned responses (Skinner 1957). If we find reason to agree with either, then we should not want to investigate language, whether from an ethological point of view or from some other, but rather those more basic processes of which it is presumably a reflection. As for the communicative behavior of animals, we should then suppose that it differs from ours for reasons that have nothing to do with a faculty of language as such. We might, for example, even suppose that animals do not talk because they have nothing to say. In any case, we should want to study language from an ethological point of view only after we have, by appropriate research, found characteristics that distinguish it from nonlinguistic processes in human beings and, perhaps, from all processes in nonlinguistic animals.

A second, and even more obvious, requirement for an ethology of

language is that its distinctive characteristics – or, at least, those we choose to study – be accessible to scientific investigation. It hardly suits our ethological purposes to have identified formal properties of syntax, for example, if we cannot determine how their underlying processes compare with those that result in the many other things that human creatures do. In the ideal case, indeed, we should want to determine in what form, if any, these same processes exist in nonhuman animals; and, in order to gain further insight into such biological predispositions to language as there may be, we should also want to be able to study these processes in human infants, including especially those who are too young to talk.

My aim is to suggest that both requirements can rather easily be met by putting our attention on speech perception. I use the term 'speech perception' in its narrowest sense to refer to just those events that occur when, on being presented with the sounds of speech, a listener perceives a string of consonants and vowels. There is nothing here of syntax or meaning, only the relation between acoustic signal and the phonetic message it conveys. In that relation we can, I think, find phenomena that imply the existence of biological specializations for language. These can be studied, not only in adult human beings, but also in presumably nonlinguistic animals and in patently prelinguistic (human) infants. For us ethologists, then, speech perception can be a window on language. Not perfectly transparent, to be sure, but likely nevertheless to afford a better view – at least for some purposes – than we can get by looking in at the more abstract levels of syntax or semantics. But we will best see what we are looking for if the biologically interesting characteristics of speech perception represent the special characteristics of language. Bear with me, then, while I say how language is special, at least in my view, and then how that is exemplified in speech. My colleagues and I have written on this matter at greater length in several papers (Liberman *et al.* 1967, Liberman 1970, Liberman and Studdert-Kennedy, in press); however, I cannot presume that these have been widely read, so I will offer a brief review.

### **A special characteristic of language and how speech partakes of it**

Surely, the special characteristic of language is grammar, if by grammar we mean the peculiar codes that make sense of the relation between sound and meaning. So I will speak of grammatical codes, but instead of dealing with their form, which is what students of language most com-



monly worry about, I will ask rather about their function. For the moment, then, our concern is not with what those grammars are, but with what they do.

To appreciate the function of grammatical codes, it is helpful to see the nature and limits of agrammatic communication. In an agrammatic mode, which is common among animals and in man's nonlinguistic communication, the relation of message to signal is straightforward. Each message is directly linked to a signal, and each signal differs holistically from all others. There is no grammatical structure, only a list of all possible messages and their corresponding signals.

Notice, now, that if all human communication were of that kind, we should not have to wonder about distinctive linguistic processes. At the one end, the signals would have to be discriminated and identified, but that is what auditory perception is all about. At the other end, the messages to which those signals are so directly connected would have to be comprehended and stored, but that is the business of processes that lie squarely in the cognitive domain. So, if we knew all about auditory perception and all about cognition, we should understand perfectly the perception of agrammatic communication.

Such agrammatic communication would, of course, be quite limited, so much so that most of what we might want to say would be unsayable. For agrammatic communication would work well only if there were agreement in number between the messages we are capable of composing and the holistically different signals we can produce and perceive. But the number of messages we can generate and comprehend is uncountably large, or so we might immodestly assume, while, in contrast, our vocal tracts and ears can cope efficiently with only a relatively small number of signals. It is precisely there, in that incompatibility, that we see the function of grammar; for the need is to match the potentialities of the message-generating intellect to the limitations of the sound-transmitting vocal tract and sound-perceiving ear. In fact, that is what grammatical processes do. They restructure the message, often drastically, so as to make it differentially appropriate for the unmatched organs – those primarily associated with thinking, remembering, listening, and breathing–eating – that must deal with it, each in its own way.

To appreciate how very great that grammatical restructuring is, and what it does for you and me, consider what would happen if you were to try to recall what I have said thus far. Assuming even the best case – that is, that I have made sense – you could not possibly remember how I ordered phonetic segments into words, words into sentences, and sen-

tences into coherent discourse. But, by using the grammatical processes of phonology and syntax, you could extract from my utterances such ideas as they might have contained. Those ideas, represented now in some presumably nonlinguistic form, you could store in (long-term) memory. On the occasion of recall you could once again use grammatical processes to restructure the ideas into transmittable language. But your language would be a paraphrase of mine. Your language and mine would both be metaphors, as it were, related to each other and to their (more or less) common meaning only by grammatical codes.

Thus, grammar serves to match a message generator, at the one extreme, to a transmitter and receiver at the other. In so doing, it makes human communication vastly more various and efficient than it would otherwise be. But the gain is achieved at a price, since grammar entails a peculiar complication in the relation between message and signal, and a need for correspondingly peculiar processes to deal with it. It is, I should think, in connection with those processes that the biological specializations for language exist.

What can we say, now, about the shape that those grammatical processes might take? Looking at the matter from the standpoint of the perceiver, we see that all the grammatical complications he must cope with are just those that he produces when he assumes the role of speaker. This is to say that the complications are internal to himself. But the same is not true for all other forms of perception. The complications of shape constancy, for example, are external to the perceiver, though he may have an internal model to deal with them; they are expressed, not in the rules of grammar, but in those of projective geometry. At all events, the special grammatical processes that are necessary to perceive language might be expected to have something in common with those that produce it. If so, the key to grammatical codes would lie in the manner of their production.

So much for grammatical codes in general. Now what about speech? Has the need for grammatical restructuring ended with the production of the abstract representations of consonants and vowels? Given that the processes of syntax and phonology have produced, finally, a string of phonetic segments, can those segments be effectively represented at the acoustic level in an agrammatic way, one acoustic segment for each phonetic segment? Plainly, they cannot. Indeed, we can see the need for grammatical restructuring even more clearly in speech than at the other levels of language. The difficulties that agrammatic communication would encounter at this level have been described in other papers (Liberman, Mattingly, and Turvey 1972, Liberman and Studdert-

Kennedy, in press). For our purposes here it is enough to speak of only one of these, and that but briefly. Consider, then, that the phonetic message is often transmitted at rates of twenty-five, or even thirty, segments per second, at least for short stretches. Surely, it would be impossible to speak that fast if the message segments were produced agrammatically by a string of discrete gestures, one for each segment and each in its proper turn. More to the point, given our emphasis on speech perception, is the fact that it would be impossible to listen that fast if each segment were, in similar fashion, represented by a unit sound, since twenty-five or thirty such sound units per second would far overreach the temporal resolving power of the ear.

In fact, the phonetic segments are not transmitted agrammatically. There is a kind of grammar that links the phonetic message to the acoustic signal; and, like the proper grammars of syntax and phonology, the grammar of speech serves to match the requirements of the one level to the limitations of the other. That is done, roughly speaking, in the following way. First, the message segments, of which there are most commonly about two to three dozen, are broken down into a somewhat smaller number of features. Each feature is assigned, as it were, to a gesture that can be made more or less independently of the others. The gestures are organized into units larger than a segment – the coding unit may be as long as a syllable or, in some cases, even longer – and then co-articulated in such a way that gestures corresponding to features of successive message segments are produced at the same time or else greatly overlapped.

By this means, a speaker can produce phonetic segments at rates several times faster than the rates at which he must change the state of any muscle. Moreover, by encoding information about several successive message segments into the same segment of the signal, he significantly reduces the number of acoustic segments per second that the listener's ear must resolve. But, as in the other grammatical conversions, these gains have a cost: there is no direct correspondence in segmentation between units of the message and units of the signal; also, the shape of the signal that carries the information for a given segment of the message will vary, often in apparently peculiar ways, depending on the nature of the other message segments that are simultaneously encoded with it.

Thus, in perception of speech, as in all of language, there is a peculiar complication in the relation between message and signal and, presumably, a need for an equally peculiar perceiver. Moreover, as will have been obvious by this time, the complications the perceiver must cope with are

only those that were introduced by the speaker. Once again, then, the key to the code is in the manner of its production. We might expect, therefore, that the key would somehow be part of the perceptual process. If so, it would, I should think, be the biologically distinctive part.

### **Some comments on the claim that speech perception is biologically special**

To suppose that speech perception is special is, of course, to imply that it is not ordinary. The view that speech perception is only ordinary sees it as an overlaid function, carried out by auditory processes no different from those we use when we hear the roar of a lion, the pattern of rain, or the clang of a bell. That view is the narrow analogue of the broader one, referred to earlier, that regards the whole of language as a more or less incidental result of processes of cognition or conditioning that are not specifically linguistic. The contrary view, which I present here, is that speech perception – and perhaps, the larger language system of which it is a part – depends on biological specializations. In the case of speech perception, these specializations can be of at least two kinds.

The first, which is perhaps the less interesting from an ethological point of view, would be a specialization of the auditory system. Thus, there may be devices specialized to respond to just those aspects of the acoustic signal that are phonetically relevant. Such devices would be useful for the purpose of extracting from the signal those physically inconspicuous parts – e.g. rapid frequency modulations – that are nevertheless of great importance from a phonetic point of view. If devices of that kind exist, they would represent auditory specializations, not phonetic (or linguistic) ones, though they would have developed in connection with speech. The distinction between auditory and phonetic specialization, which is, I think, an important one, has two aspects. First, the kind of auditory specialization I have imagined could only sharpen and clarify the signal; it could not manage the grammatical peculiarities of the relation between that signal and the phonetic message it encodes. Second, the specialized auditory mechanisms would be called into play in the processing of all sounds; hence, their perceptual consequences would be characteristic of all auditory perception, not just of the perception of speech. That is surely an important consideration, for if the specialization were extreme, then perception of other biologically important sounds would be altered and, perhaps, impaired.

The assumption of auditory specializations for speech has, of course, a

counterpart for language in general – to wit, that the cognitive processes have undergone evolutionary changes that make them somehow better adapted for language. It is difficult even to imagine what such specializations would be like, but, simply to make the point, let us take account of the digital nature of language and suppose that the brain of a linguistic animal like man might be, in general, better adapted to digital processes. As in the auditory case, these cognitive specializations would apply in general and not just to those activities that are associated with language. Processes that are, perhaps, better carried out in an analogue mode – e.g. spatialization – might then be adversely affected.

In any case, there may be a class of specializations for speech perception that would properly be considered auditory. Unfortunately, we can find very little evidence that bears one way or the other on the matter (see Liberman and Studdert-Kennedy, *in press*), so, having remarked the possibility that special auditory devices might exist, we turn our attention to a second possible specialization for speech, one of potentially greater interest to ethologists.

This other specialization would serve to deal with the grammatical peculiarities in the relation between acoustic signal and phonetic message – in particular, the lack of correspondence in segmentation and the context-conditioned variation. It would presumably be called into play only in the perception of phonetic structures, leaving nonspeech perception entirely unaffected; and the result of its operation would be a distinctive mode of perception, the phonetic mode. Hence it would deserve to be called a phonetic specialization.

The analogous assumption for the other aspects of language is, of course, that they, too, depend on specialized grammatical processes and result in a distinct linguistic mode of perceiving or thinking. I should suppose, then, that, as I have already suggested, the phonetic device we are here considering would be an integral part of the larger specialization.

Following the argument of the earlier section, I will assume that the special characteristic of the phonetic device is a biologically based link between perception and production. Given such a link, speech perception is constrained as if by 'knowledge' of what a vocal tract does when it makes linguistically significant gestures. The development of that knowledge may depend in important ways on experience, much as the learning of song by some birds does, but we should suppose that, again as in the case of the birds, those effects of experience rest on a strong (and specialized) biological base. It is difficult, in the present state of our ignorance, to know how that biological base should be characterized. Putting the mat-

ter negatively, I should say that it seems hardly conceivable that, starting with a *tabula rasa*, the child could ever learn what he needs to know in order to make sense of the peculiarities of the speech code. For if he had only the speech signal, and no knowledge of vocal tracts, he could entertain an indefinitely large number of hypotheses about how the signal was produced, and a corresponding number of hypotheses about the nature of the coded relation to the message. To 'break' the code, the child would presumably need some biologically given limits on the kinds of hypotheses he should consider – that is, some biologically given 'knowledge' about what vocal tracts do. A somewhat similar argument has been made by Chomsky (1959) about the acquisition of syntax. The point of the argument is that there is no automatic 'discovery procedure' by which a child can infer the grammar of a language from the (mostly degenerate) examples he is offered. One supposes, then, that the child is biologically predisposed to entertain only certain kinds of hypotheses, presumably those that capture just the aspects of grammar that are universal. In the case of speech, I am tempted to assume, analogously, a special biological endowment that, given appropriate experience, enables the child to learn what he needs to know about the relation of sound to the manner of its production, and so to acquire the key to the code.

#### **A few examples of putatively special phenomena of speech perception**

The point of this paper, it will be recalled, is that there are phenomena of speech perception that can be shown to rest on specialized phonetic processes, and that these can be looked for and studied, not only in adult humans, but in human infants and nonhuman animals as well. I can offer only a few examples of such phenomena here, and even these I must deal with all too briefly. And, in order to keep technical phonetic and acoustic details to a minimum, I will limit the examples to those that deal with a single and simple acoustic cue: silence. The reader who may wish to find additional examples is referred to a recent review (Lieberman and Studdert-Kennedy, in press) and to the studies cited there; he may also wish to see a short paper (Lieberman and Pisoni 1977) written recently. Several of the examples I will use here have appeared in those papers, though I will also take advantage of some new data and examples.

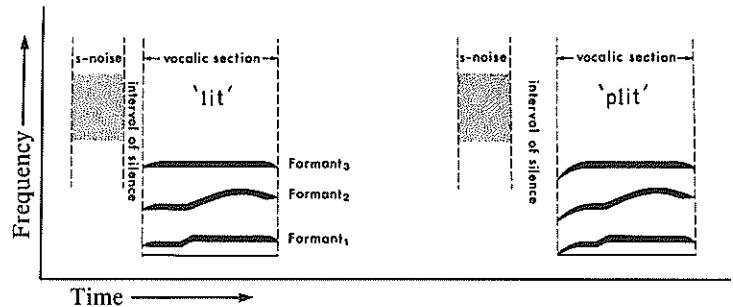
Before presenting the examples, which derive entirely from studies of adult humans, I should say a word about how easy or difficult it might be

to use them in research on animals and infants. To make it as easy as possible, I have chosen only those examples – generally very simple ones – for which it is possible to imagine straightforward behavioral tests. (Unfortunately, some of the most interesting phenomena of phonetic perception do not lend themselves to tests of that kind and can, therefore, only be investigated in adult humans. See Liberman *et al.* 1967, Liberman and Studdert-Kennedy, in press.) For some of the examples I will offer, it is quite likely that the appropriate tests can easily be made. Others may prove less feasible. Among the latter are those in which we may discover that the subject – in particular, the animal subject – lacks even the basic sensory capacity that is necessary (but presumably not sufficient) for the phenomenon we wish to study. In such cases we should hope to find animals that have the necessary sensory capabilities, or else develop other examples of the same general phenomenon that present fewer difficulties of a purely sensory sort. At all events, we ought, in one way or another, to be able to make the appropriate tests.

*When silence sounds like sound*

Articulatory gestures that produce linguistically significant contrasts (for example, 'rabid' versus 'rapid') typically have acoustic consequences that are numerous, diverse, and distributed over a considerable stretch of the acoustic signal. It is of interest from our point of view that these various acoustic consequences have an equivalence in phonetic perception. That equivalence is established by demonstrating that each such acoustic consequence – let us call it a 'cue' – is more or less sufficient (with all other cues held constant) to produce the perceived phonetic contrast. The rather considerable evidence bearing on that point is reviewed in Liberman and Studdert-Kennedy (in press). I will offer one example here.

Consider the contrast between the words 'slit' and 'split'. To articulate the stop consonant [p], which is the distinguishing segment in 'split', the speaker must close his vocal tract for 50 msec. or so after making the hissing noise associated with the fricative [s], and then open it as he undertakes the remaining (vocalic) section of the syllable. (Omitting the p in 'slit', the speaker does not completely close his vocal tract.) Among the acoustic consequences of the closing and opening are two – shown schematically in figure 1 – that will concern us here. One is the interval of silence between s-noise and the vocalic section that corresponds to the closure: relatively short silence for 'slit', relatively long silence for 'split'. The other is the effect on the acoustic spectrum at the beginning of the



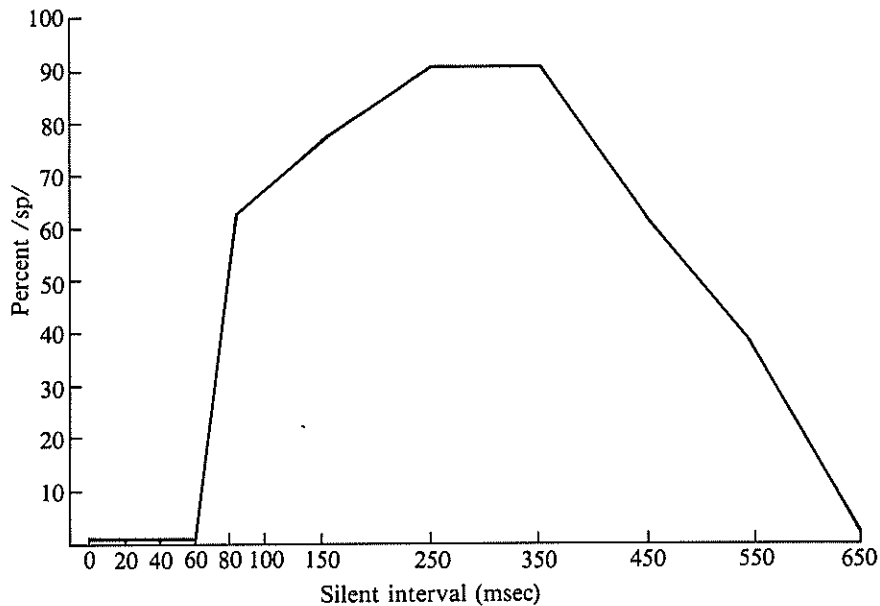
1. Schematic representations of the temporal (interval of silence) and spectral (initial formant transition) cues for the perceived distinction between 'slit' and 'split'. (From Erickson *et al.* 1977. A full account is in preparation.)

vocalic section that is a result of the subsequent opening of the vocal tract: the formants are initially level in 'slit', rapidly changing in 'split'.

Let us look first at the effect of the silence cue as shown in an experiment by Dorman, Raphael and Liberman (1976). They started with a (real speech) recording of the word 'lit'. To this they prefixed a brief patch of s-noise, separating it from 'lit' by intervals of silence that varied from 0 to 650 msec. These stimuli were randomized and presented to listeners for judgment as 'slit' or 'split'. The results are shown in figure 2, where we see that, with silent gaps from 0 to 60 msec., all listeners reported 'slit'; then, quite abruptly, they heard 'split'; finally, at 450 msec. of silence they began, though now rather slowly, to hear 'slit' once again. We will not concern ourselves here with the question: why 'split', not 'sklit' or 'stlit'? That is a separate issue, quite unrelated to our present interest, which is only in the presence or absence of a stop consonant. What we have seen in that connection is that appropriate variations in the amount of silence are sufficient to produce the perceived contrast between the absence of a stop in 'slit' and its presence in 'split'. If we suppose that the silence provides the phonetic information that the speaker did (or did not) close his vocal tract, and that the listener has a device specialized to make the appropriate interpretation, then silence proves to have just the sound we should expect it to have: it sounds like a stop consonant.

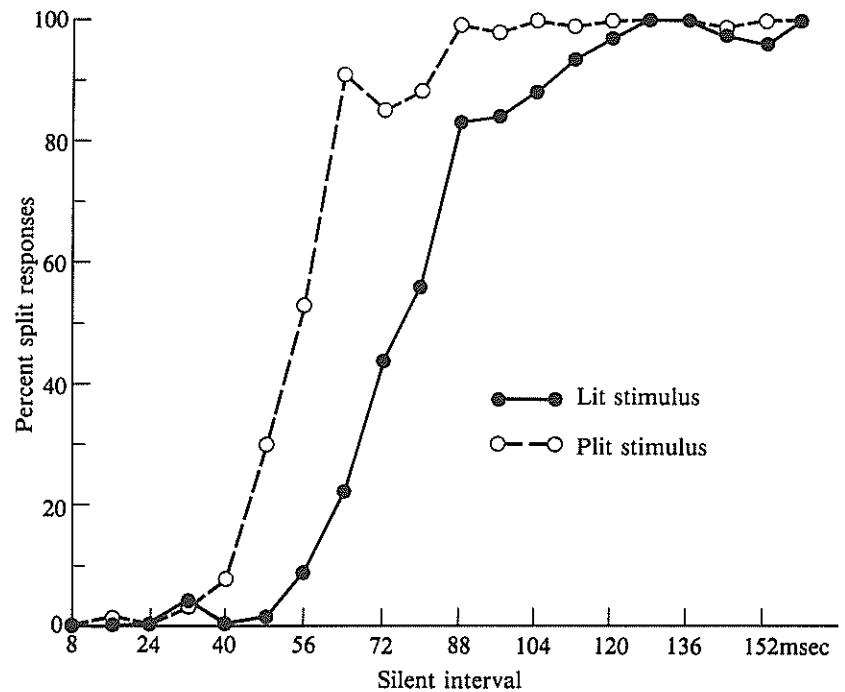
What, then, of the spectral cue? What does it sound like and how does it relate to the silence? For an answer, we turn to a recent experiment by Erickson *et al.* (1977), in which the effect of silence on the 'slit-split' contrast was investigated under each of two stimulus conditions. In one,





2. The effect of the interval of silence on the perceived distinction between 'slit' and 'split'. (From Dorman, Raphael and Liberman 1976: 203.)

illustrated by the example at the left of figure 1, there was a patch of (synthetic) s-noise, followed by a (silent) gap that varied in steps of 16 msec. from 8 to 152 msec., followed then by a (synthetic) vocalic section having at its onset straight formants appropriate for the syllable 'lit'. In the other set, illustrated by the example at the right of figure 1, all aspects of the stimuli were the same except that there were, at the onset of the vocalic section, formant transitions appropriate for the stop consonant [p] in the syllable 'plit'. These stimuli were randomized and presented to listeners for judgment. The results are shown in figure 3. There we see two almost parallel functions – one for s-noise plus 'lit', the other for s-noise plus 'plit' – that show the percentage of 'split' judgments plotted against the duration of the silent gap. The two functions are displaced with reference to each other in such a way as to indicate that, in order to convert 'slit' to 'split', 20 msec. less silence is necessary when the spectral cues appropriate for the stop [p] are present. In effect, then, there is, in this instance of phonetic perception, an equivalence between 20 msec. of silence, on the one hand, and, on the other, the presence or absence of certain transitions. Thus, a temporal cue sounds like a spectral cue.



3. A trading relation in the perception of 'slit' and 'split' between temporal and spectral cues. (From Erickson *et al.* 1977.)

One is naturally led to ask why two such different physical events – and, indeed, it is difficult to imagine any two that would be more different – should sound the same. The answer is presumably to be found in the fact that these different acoustic cues are the distributed consequences of the same linguistically significant gesture. They sound alike, then, because both signal to a biologically specialized phonetic perceiver that the speaker did or did not close and open his vocal tract in a way appropriate for the production of 'split' (or 'slit').

Would such an equivalence exist in animals? On the basis of what we know about auditory systems in general, I should think it unlikely. In any case, we can find out, and, as we will see in a moment, by fairly simple procedures. Moreover, we can also test for the equivalence in human infants at various ages. We should especially want to do that because, given the repeated association of spectral and temporal cues in speech, we are tempted to suppose that the equivalence is learned. But could the learning of such phenomenological equivalence possibly be arbitrary?

That is, could it depend on the fact of association and nothing else? Surely, two different sounds that are frequently associated in the world will come to be associated in a listener's mind; on hearing one, he will expect to hear the other. Or, if they always signify the same thing, then each may become a sufficient sign. But will they, with any amount of association, come actually to sound alike, as in the case of the speech cues they clearly do? The experimental literature on 'acquired similarity' – as that possibility used to be called – together with the normal experience of all of us suggest that they will not. I should think, therefore, that, while experience may be necessary for establishing the perceived equivalence of the speech cues, it is not sufficient. In any case, we may hope to learn from research on infants whether the development of the equivalence comes so early and so suddenly as to imply a strong biological predisposition to profit from the experience in the particular and, perhaps, particularly human way that produces the effect we have observed in our experiments.

But let us look now at how we can, by simple behavioral tests, determine whether animals and infants hear the silence and spectral cues as we do. One possibly interesting experimental plan, drawn from the results shown in figure 3 and previously discussed, is sketched in figure 4. There we see four pairs of stimuli, the corresponding percepts, and the relevant characterization of the cues. Pairs I and II illustrate that the perceived contrast between 'slit' and 'split' can be produced by either of two acous-

	Description of stimuli		Percept	Characterization of cues			
	Gap	Vocalic		Temporal	Spectral	Temporal	Spectral
Pair I	s-noise	short – – – <i>lit</i> short – – – <i>plit</i>	<i>slit</i> <i>split</i>	– p – p	– p + p	→ same	different
Pair II	s-noise	short – – – <i>lit</i> long – – – <i>lit</i>	<i>slit</i> <i>split</i>	– p + p	– p – p	→ different	same
Pair III	s-noise	short – – – <i>lit</i> long – – – <i>plit</i>	<i>slit</i> <i>split</i>	– p + p	– p + p	→ different	different
Pair IV	s-noise	short – – – <i>plit</i> long – – – <i>lit</i>	<i>split</i> <i>split</i>	– p + p	+ p – p	→ different	different

4. Diagrams illustrating the phonetically equivalent effects of spectral and temporal cues and, also, the phonetically different effects produced by the two ways in which those cues may be combined. (From Liberman and Studdert-Kennedy, in press; also in Liberman and Pisoni 1977.)

tic cues – one spectral, the other temporal. The patterns in each of these pairs differ by only one cue. Pairs III and IV illustrate the effects of combining the two cues, but in different ways and with different consequences. In Pair III, the spectral and temporal cues are combined in such a way as to 'add' to each other and so enhance the perceived difference between 'slit' and 'split'. In Pair IV, on the other hand, the two cues are combined so as to 'cancel' the perceived difference, with the result that listeners hear 'split' in both cases. Putting aside further discussion of what these patterns sound like, let us consider simply the relative difficulty that a human adult would have in discriminating them. That can only be inferred from the (phonetic) identification data of figure 3, but it has now been verified by direct measures of discriminability in an experiment by Fitch *et al.* (in preparation). In terms of increasing difficulty, the order is: Pair III (two-cue difference), Pairs I and II (one-cue difference), Pair IV (two-cue difference). The point to note is that the easiest and hardest discriminations have in common that the patterns differ by two cues. One of the pairs with two cues different (Pair IV) is harder to discriminate than either of the pairs that differ by only one cue (Pairs I and II). This is so, presumably, because the cues have, as it were, positive and negative signs, or vector-like directions, for the perceptions they induce; we should suppose that such signs, or such directions, exist only in the phonetic mode. If so, then animals should show a different order of relative difficulty. For them, Pairs III and IV ought to be of approximately equal difficulty, since the patterns in each differ by two cues; and both of these pairs should be easier than Pairs I and II, in which the patterns are distinguished by only one cue in each case. If the expected difference between adult humans and animals is found, we should want then to test infants at various ages. At all events, the test should be an easy one to make: the dependent variable is only ease of discrimination; the comparison to be made is only in the relative order of difficulty; and the expected result with animals (if obtained) cannot reasonably be attributed to inattention or lack of motivation.

*When the sound of silence depends on how many are talking: phonetic constraints of an ecological sort*

I have suggested that the biologically distinctive characteristic of phonetic perception is that it is governed as if by knowledge of what a vocal tract does when it makes linguistically significant gestures. We should ask now: whose vocal tract? A proper regard for the ecological realities sug-

gests it can hardly be that of the listener, nor yet of the speaker; for if the listener is to cope with the fact that he is commonly exposed to several talkers simultaneously, he must make his perceptual calculations in terms of some abstract conception of the behavior of vocal tracts in general.

To see how that is, consider another example of the sound of silence. Suppose we record the sentence, 'You will please say *shop* again.' Now we introduce silence between the end of 'say' and the beginning of 'shop', and find, not surprisingly, that, with just the right amount of silence, listeners hear, 'You will please say *chop* again.' I say 'not surprisingly' because, to produce the affricate – that is, the stop-initiated fricative – in 'chop' instead of the fricative in 'shop', a speaker must close his vocal tract for a brief period and, in the process, introduce a brief period of silence (Dorman, Raphael and Liberman 1976). To a listener, that silence provides the information that the speaker closed his vocal tract just long enough to have said 'chop'; hence, 'chop' is what the listener perceives. But if there were two speakers, one saying 'You will please say', the other 'shop again', then the period of silence between the words 'say' and 'shop' (or 'say' and 'chop') would not, in principle, supply useful phonetic information: given intentional collaboration, or the accidents of speech when two are talking, one person might say 'You will please say' and the other 'chop again' with zero interval of silence between 'say' and 'chop'. Our experiments reveal that listeners behave as if they knew that perfectly well.

One of those experiments (Dorman, Raphael, and Liberman 1976; for a similar experiment, see Dorman *et al.* 1975) was performed with the utterance I have here used as an example. In one condition, we recorded a male speaker saying 'You will please say shop again.' In the other, we recorded a female saying 'You will please say' and joined that to the 'shop again' as previously recorded by the male. In both conditions, the experimental variable was the duration of silence between the words 'say' and 'shop'. The result was quite straightforward. In the one-voice condition, all listeners heard 'say shop' or 'say chop' depending on the presence (or absence) of the appropriate amount of silence. In the two-voice condition, on the other hand, listeners heard 'say shop' at all intervals of silence. Thus, they behaved as if, knowing that two vocal tracts can do what one vocal tract cannot, their perception of speech was governed by some quite abstract conception of vocal tracts in general. It would, I should think, be interesting from an ethological point of view to find out when infants begin to behave that way.

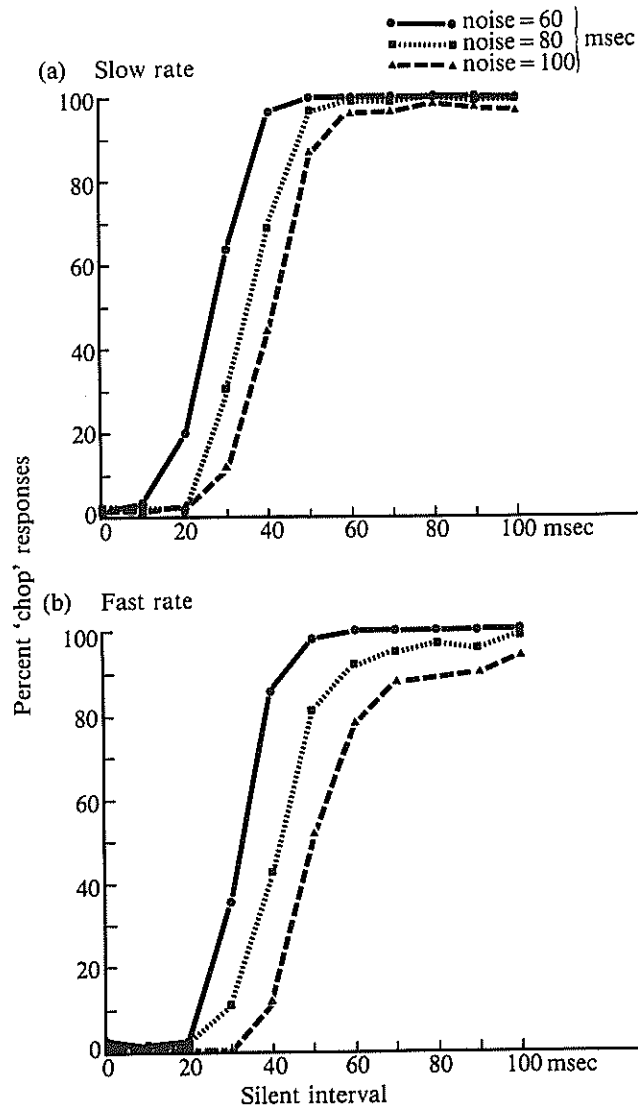
*When the sound of silence depends on how fast the speaker is talking*

One of the interesting problems that a speech perceiver must contend with is that which is created when speakers articulate at different rates. The problem is the more interesting because variations in rate do not affect all portions of the acoustic signal equally (Gaitenby 1965, Lehiste 1970, Huggins 1972): some portions are stretched (or compressed), others not, or not to the same degree. Presumably, then, the listener cannot make the necessary adjustment by, as it were, simply multiplying the acoustic signal by a constant factor. If he adjusts properly, it is as if he had some knowledge of the disproportionalities that are associated with rate variations, or, more generally, of the articulatory mechanisms that generate them. It would be particularly interesting, then, to know whether animals can make those adjustments, and at what age infants do so.

Unfortunately, the matter of adjustment for rate has been very little studied in adult human beings, so we do not have a large set of examples to choose from. There is, however, a recent study (Repp *et al.*, in press) that is particularly appropriate for our purposes if only because it deals with the same silence cue to which we have become accustomed.

In this study, we are concerned once more, then, with the role of silence in converting 'Please say shop again' to 'Please say chop again.' But this experiment adds two new variations: one is in the duration of the noise associated with the fricative [sh] in 'shop'; the other is in the rate at which the carrier sentence is articulated, more slowly in one condition and more rapidly in the other. (The several durations of the friction noise were the same in the two-rate conditions.) As I have implied, the experimental variable in all cases was the duration of silence between 'say' and 'shop'.

The results are shown in figure 5. There, the percentage of 'chop' judgments is plotted as a function of the duration of silence between 'say' and 'shop'. As usual, the silence cue is sufficient to produce the perceptual contrast between 'shop' and 'chop'. We also see that at both rates of articulation the amount of silence necessary for the affricate 'chop' is greater as the duration of the friction noise is longer. This is so because duration of friction noise is itself a cue to the fricative-affricate distinction – longer noise biases the perception toward the fricative [sh]. Thus, we have still another relation between different acoustic cues, similar in principle to the one between temporal and spectral cues described earlier. In this case, the relation establishes an equivalence in phonetic perception between durations of silence and durations of noise.



5. The effects on the perceived distinction between 'You will please say shop again' and 'You will please say chop again' of orthogonal variations in the duration of the silent interval (immediately preceding 'shop'), the duration of friction noise (in 'shop'), and the rate of articulation of the sentence frame. (From Repp *et al.*, in press; also in Liberman *et al.* 1977.)

More relevant to our present concern, however, is the variation in rate of articulation. There, we see an effect that is, apparently, paradoxical: when the rate of articulation of the carrier sentence is increased, while the duration of the noise is held constant, listeners require more silence to hear 'chop'. But perhaps that comparison is not altogether proper, since it assumes that the duration of noise should remain fixed as rate of articulation increases, whereas it would, presumably, be shorter as the rate is faster. So, perhaps the more appropriate way to view the results is to look at the different effects on the silence cue of the two ways of shortening the duration of the noise: within the same rate condition, which is roughly equivalent to what would have happened (to the noise) if the speaker had changed the articulation from 'shop' to 'chop'; or by moving across the rate conditions, which is roughly equivalent to what would have happened to the duration of the noise, if, while continuing to say 'shop', the speaker had simply speeded up. Consider, first, the effect of shortening the noise while holding the rate constant. At the slow rate, we see that the 'boundary' value for the silence cue – that is, the point on the silence continuum at which the listeners' judgments are 50% 'shop' and 50% 'chop' – goes down from about 41 msec. at 100 msec. of noise to 35 at 80 and then to 28 at 60; at the fast rate, the boundary for the silence cue moves from 50 at 100 msec. of noise to 42 at 80 and then to 37 at 60. But look now at the different effect of shortening the noise across rate conditions – that is, when the same reductions in noise durations are made by changing the rate of articulation. As the noise duration is reduced from 100 msec. at the slower rate to 80 at the faster rate, the boundary for the silence cue does not decrease at all; in fact, it increases very slightly from 41 msec. to 42 msec.; then, with further reduction in noise duration from 80 msec. (slower rate) to 60 (faster rate), the boundary value for the silence cue again increases slightly, this time from 35 to 37 msec.

I believe that the disproportionality in the perceptual results just described may reflect the listeners' sensitivity to a disproportionality in the acoustic effects of varying the rate of articulation: perhaps duration of the noise associated with the fricative generally changes more with rate of articulation than does the stop closure. If so, then we see here an instance of the rather complex perceptual adjustment to rate variation that we earlier wondered about and, in that connection, a demonstration of how very exact is the listener's knowledge of the particulars of vocal tract dynamics. But I hasten to say that, at this writing, we do not have definitive data about the acoustic effects of varying articulatory rate, so we can have no great confidence in that particular interpretation. Still, it is



of interest from our point of view that the two ways of varying the duration of noise – by changing the phonetic segment, as it were, or by changing the rate of articulation – have different perceptual consequences for our adult listeners. How special are the calculations that underlie that distinction?

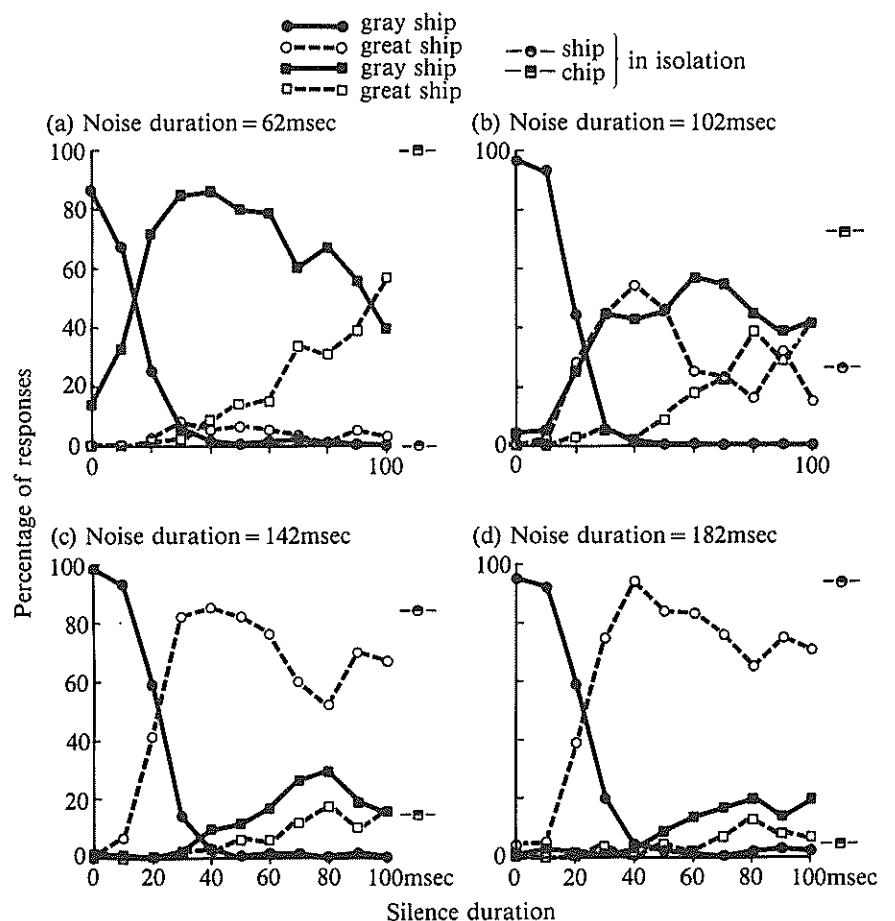
*When the sound of silence is heard (or not) in one syllable depending on an acoustic cue in the next syllable*

As I said in an earlier section of this paper, a very general and important characteristic of the speech code is that there is no direct correspondence in segmentation between segments of the phonetic message and segments of the acoustic signal. The rapid switches of sound source that occur during articulation often cause the information about a single phonetic segment to be spread through several acoustic segments; on the other hand, normal co-articulation often collapses information about several phonetic segments into a single segment of sound. A general consequence is that, at any given instant, the speech signal is likely to be carrying information about more than one phonetic segment. It is this characteristic that justifies our speaking of the relation between phonetic message and acoustic signal as a code rather than a cipher – or, indeed, as grammatical rather than agrammatical – and it is this same characteristic of speech that, perhaps more clearly than any other, would appear to be beyond the capacity of ordinary auditory devices. The task is not simply to respond to a complex acoustic pattern, but to recover a message from a signal in which it is peculiarly encoded. In that sense, perceiving speech is not so much a matter of complex pattern recognition as it is of crypt-analysis.

We should have no difficulty finding examples of the curious relation in segmentation between message and signal; their number is legion. The trick is rather to find some that lend themselves to relatively simple behavioral tests with animals and infants. A recent study (Liberman *et al.* 1977, and Repp *et al.* in press) presents a possibility, and happily for us, it deals yet again with the silence cue.

We begin with a recording of the sentence, 'He saw the gray ship.' The experimental variable is, as in all our other examples, the duration of silence – in this case, between the words 'gray' and 'ship'. The parameter of the experiment is the duration of the friction noise associated with the fricative [sh] in 'ship' (or the affricate [ch] in 'chip'). There were eleven durations of silence, ranging from 0 to 100 in steps of 10 msec.; the

durations of friction noise were set at 62, 102, 142, and 182 msec. The resulting patterns were randomized and presented to listeners for judgment as '(He saw the) *gray ship*, *gray chip*, *great ship*, or *great chip*'. The results are shown in figure 6, where the responses are plotted as a function of duration of silence for each duration of friction noise. We see, first, that at all noise durations the listeners reported hearing 'gray ship' when the duration of silence was less than approximately 30 msec. That is, when the silence was insufficiently long, listeners heard neither a stop



6. The relation between duration of silence and duration of friction noise in the perception of fricative, affricate, and stop consonant, demonstrating how the perception of a phonetic segment (the [t] of 'great') is determined by an acoustic cue (the duration of the friction noise of 'ship') that lies in the following syllable. (From Repp *et al.*, in press; also in Liberman *et al.* 1977.)

(as in 'great') nor a stop-initiated fricative (i.e. affricate, as in 'chip'), which is exactly what we should expect, given the results described earlier. If we look now at the results obtained with the shortest duration of noise (62 msec.), we see another familiar result: when the duration of the silence becomes long enough, listeners perceive a stop-like effect, reporting the affricate in 'gray chip' (instead of 'gray ship'). But consider now the result when the duration of fricative noise is long (182 msec.). To produce a stop-like effect in this condition, more silence is necessary, as it was indeed in the experiment described earlier. What is novel and particularly interesting, however, is that, under the particular conditions of this experiment, the stop-like effect is attached to the end of the 'gray' syllable: the listeners perceive, not 'gray chip', but 'great ship'. Thus, with all other aspects of the pattern held constant, it is possible to interconvert between the words 'gray' and 'great' – that is, to add or subtract the syllable-final stop consonant – by altering the duration of the noise associated with the fricative (affricate) in the next syllable. The effect is the more interesting, since adding the [t], which is commonly assumed to be a transient from an acoustic point of view, is accomplished by making the noise in the next syllable longer (that is, less transient).

It is not difficult to find a plausible explanation for the results just described. Consider, as we have before, that an appropriate interval of silence signals that the speaker closed his vocal tract, as he must to produce either a stop or an affricate. The listener will, as a consequence, tend to hear one or the other of those phones. But the duration of the friction noise is also an important cue for the affricate, distinguishing it from the corresponding fricative: relatively short noise for affricate, relatively long for fricative. Given a relatively short duration of noise at the onset of the second syllable, the listener takes the relatively long silence (hence closure) to mean an affricate: he hears 'gray chip'. But given a relatively long duration of noise at the onset of the second syllable, and the same relatively long silence as before, the listener perceives the fricative in 'ship', which accords with the long duration of the noise; then he adds a stop consonant [t] to the end of the previous syllable 'gray', which converts it to 'great' and allows him to take account of the vocal-tract closure that was signaled by the silence.

It is somewhat beside the point whether that account is exactly correct or not. For our purposes, the important fact is that a sufficient cue for the distinction 'gray' versus 'great' is in the duration of noise associated with the fricative (or affricate) at the beginning of the next syllable. Will animals decode the signal that way, and at what age will infants do it?

**Summary**

My aim has been to suggest that an ethologist might want to study the perception of speech because it is an integral and reasonably representative part of the language process, yet it can be investigated experimentally, not only in adult humans, but also in nonhuman animals and in human infants. Moreover, research on adult humans has already uncovered certain phenomena that strongly imply the existence of biological specializations for phonetic (as distinguished from auditory) perception. I have offered several examples of these, all taken from recent research on the role of silence in the perception of stop and affricate consonants.

---

## Comments on papers by Marler and Liberman

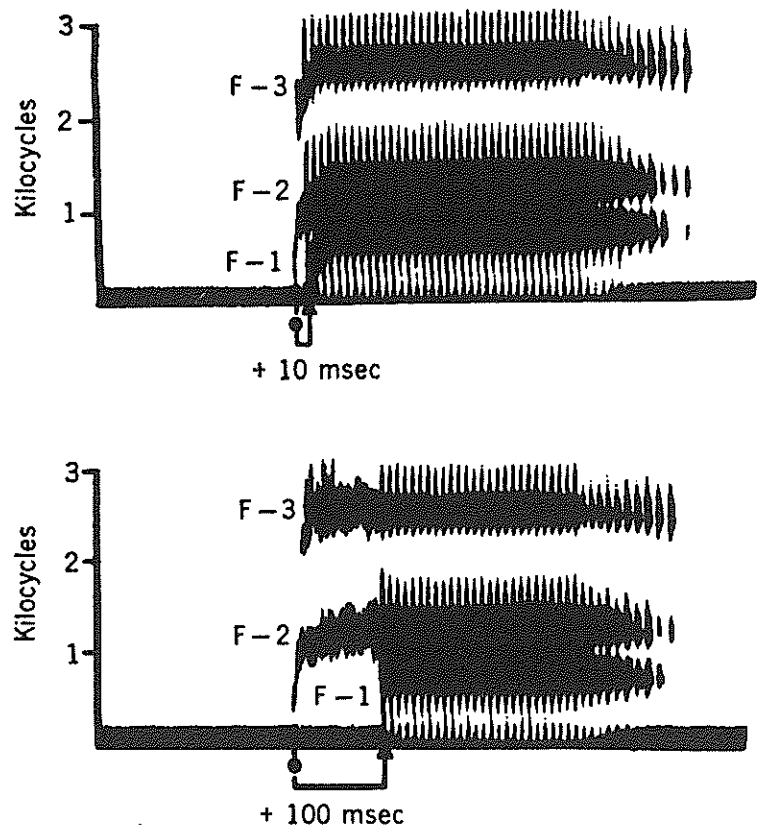
D. PLOOG

The two papers by Peter Marler and Alvin Liberman which I will now review are papers with complementary approaches. They refer to different sets of experimental data but share similar aims. Marler focuses on animal communication, especially in monkeys and birds, and looks toward man, while Liberman focuses on language, and looks towards the animals.

I have the feeling that facts and concepts in the field of auditory perception are less well-known to ethologists, anthropologists, and psychologists and perhaps technically more difficult to understand than some other aspects of ethology with which this volume is concerned. Therefore I shall take the liberty of going into some technicalities. In the course of my comments you will, I hope, develop a feeling for the beauty of these experiments because of the precision of measurement that can be achieved, because of the reproducibility of the results, and because of the straightforward strategies that can be employed to ask specific questions.

One of the chief concerns in ethological research is the observation and, if possible, the quantification of motor behaviour. In the studies of auditory perception and speech perception discussed by Marler and Liberman a very specialized kind of motor behaviour, namely vocal behaviour, is used as stimulus material. Please note that we are dealing here with natural stimuli chosen from the species-specific vocal repertoire of monkeys, birds, and man.

On a sound spectrogram as in figure 1, the frequency modulation of a sound is usually plotted over time. However, the spectrogram does not show what the perceiver of a sound perceives. This is a key issue for the understanding of the experiments: the distribution of acoustic energy as



1. Spectrograms of synthetic speech showing two conditions of voice-onset-time (VOT): slight voicing lag in the upper figure and long voicing lag in the lower figure. The symbols F-1, F-2, and F-3 represent the first three formants, that is, the relatively intense bands of energy in the speech spectrum (Eimas *et al.* 1971).

reflected in the spectrogram does not correspond to the percept of the stimulus receiver. This is definitely so in the perception of speech sounds, as you will see in a moment. Whether this is also true for monkeys, birds, and other animals we do not know as yet, but for reasons which I will explain later it would be of great interest to find out.

The next thing one should know for the understanding of the experiments is the way in which a given vocal gesture – our stimulus configuration – can be technically manipulated. Listen to the sound shift from [pa] to [ba]: this series, and many such sounds, can be produced by a machine. Speech sounds and animal calls can be synthesized. Each parameter which contributes to the sound, such as frequency transitions, pauses in

these transitions, and the duration of transitions, can be manipulated independently of the others, and the corresponding percept of the listener can be ascertained. In the [ba]–[pa] example, the one parameter that is systematically varied is the parameter called voice-onset-time (VOT). The VOT (see figure 1) is the time between the release of the consonant and the onset of voicing, or, to be more precise, the number of milliseconds by which the release of a consonant precedes voicing. Consonants with shorter VOTs are perceived as voiced (as in [ba]); consonants with longer VOTs are perceived as voiceless or, better, unvoiced.

This is, I think, all we need to know about the technicalities. Marler is very much intrigued by investigations of the perceptual processing of speech stimuli by human adults and infants. Such studies merit close attention from ethologists working on analogous problems with animals. Since the human data which Marler cites bear directly on Liberman's contribution – and are in fact either generated or incited by him – I will give you only one example for the human infant. Again I will use the [ba]–[pa] paradigm. One-month-old infants are exposed to the [ba]–[pa] sequence of speech sounds. The VOT is extended in 10- or 20-msec. steps from, say 10 to 60 msec. Neither these young infants nor adults who listen to such a VOT continuum perceive the continuum as such but rather as a sudden shift from [ba] to [pa]. The shift always occurs at a sharp boundary, about 40 msec. VOT. Before that time both infant and adult listeners hear [ba], thereafter they hear [pa]. I will not explain here how this sudden shift in the percepts can be precisely ascertained in young infants. I will merely state that this finding has been confirmed by several investigators with varying methods and different speech sounds. This evidence for a distinctive mode of perceptual processing has become known as categorical perception. Moreover, testing infants in very different linguistic environments has led to the same results, that speech sounds are perceived in the categorical mode. Obviously, categorical processing has the consequence of grouping stimuli into classes, imposing a particular kind of order on the varying acoustic patterns.

For the purpose of comparison let us now shift to monkeys. Although there are a number of studies on the vocalizations of various species of monkeys, very little is known about the functions of vocal signals. As Marler rightly states, it is virtually impossible to assay the communicative function of these signals until we have some understanding of how the signals are processed during perception.

Among the many graded calls of the Japanese macaque there is a variety of 'coos' on which I want to focus now (see figure 3, chapter 10.1

above). Steven Green has identified a significant feature within the variety of coos: the temporal position of a frequency rise. The rise may occur at any point of the coo (see 'Coo type', figure 3). Early and late frequency rises correlate with different circumstances of production. For example, the rises referred to as 'smooth early highs' are contact coos emitted by isolated animals; 'smooth late highs', on the other hand, are typically emitted by a subordinate monkey towards a dominant one. During one phase of experiments Japanese macaques were trained to respond conditionally to playback of different classes of coo calls. They achieved a high criterion of performance quite rapidly, whereas the two other species trained – the vervet monkey and the pig-tailed macaque – had enormous difficulty in generalizing to new tokens of the two classes. The smooth early–smooth late distinction appears to be a relevant distinction for the Japanese macaque but an alien one for the other species. One is tempted to argue that Japanese macaques exhibit a special predisposition to process coo sounds in a particular way that parallels their apparent meaning. The stimuli are clearly not equivalent to conspecific and alien adults, as classical learning theory would have led us to expect. That the principal of equipotentiality – one of the assumptions of learning theory – does not hold for a number of learning processes is beautifully demonstrated by Marler's experiments on song learning in swamp sparrows. A male of this species presented with a natural choice of songs of a congeneric species to copy will selectively learn conspecific models. There are surprising parallels between the song learning of birds and the development of the perception of speech in human infants. Just as our infants are predisposed to respond selectively to particular aspects of speech sounds before speaking themselves, so some young songbirds are responsive to species-specific features of song before they themselves begin to sing.

This might be a good point to turn back to the human case. In the studies on speech perception which I have discussed so far certain phenomena have come to light that strongly imply the existence of biological specializations for phonetic – as distinguished from auditory–perception. This is, in fact, the key issue of Liberman's paper, which is concerned with the relation between the acoustic signal and the phonetic message it conveys. There is a kind of grammar that links the phonetic message to the acoustic signal. How this transformation takes place will be exemplified in a few minutes. The transformational process involves the encoding of information about several successive message segments into the same signal, with the result that the speaker significantly reduces the number of acoustic segments per second that the listener's ear must



resolve. As we shall see, the key to the phonetic code is in the manner of its production. To make things relatively easy, Liberman has concentrated on one striking phenomenon, the role of silence in the perception of consonants.

There is space for only one example. Suppose, he says, we record someone saying the sentence 'You will please say *shop* again.' Now we introduce silence between the end of 'say' and the beginning of 'shop'. How this is done in the right sequence you can see in figure 1 of chapter 10.2 above, although it is not the same example I am reporting now. With just the right amount of silence we find that listeners hear, 'You will please say *chop* again.' To produce the affricate in 'chop' instead of the fricative in 'shop' a speaker must close his vocal tract briefly, thus introducing a period of silence. To a listener, *that* silence provides the information that the speaker closed his vocal tract just long enough to have said 'chop'; so 'chop' is what the listener perceives. But if there were two speakers, one saying, 'You will please say', the other 'shop again', then the period of silence between the words 'say' and 'shop' (or 'say' and 'chop') would not, in principle, supply useful phonetic information. In experiments of this sort with one or two speakers and varying amounts of silence between 'say' and 'shop' the results were quite straightforward. In the one-voice condition, all listeners heard 'say shop' or 'say chop' depending on the presence or absence of the appropriate amount of silence. In the two-voice condition, on the other hand, listeners heard 'say shop', which was presented to them, at *all* intervals of silence. Thus, they behaved as if they knew that two vocal tracts can do what one vocal tract cannot. As Liberman says, it would be interesting to find out when infants begin to behave that way.

In my opinion the most important notion Liberman advances is the idea that the special characteristic of the phonetic device is a link between perception and production. Given such a link, he says, speech perception is constrained as if by innate 'knowledge' of what a vocal tract does when it makes linguistically significant gestures. For the language acquisition process this means that the child would need some biologically given 'knowledge' about what vocal tracts do.

Summing up the message of both papers, I should like to repeat that speech perception depends on biological specialization and is considered to be an integral part of the larger specialization for language. We may think of the perceptual predisposition in human infants as initial instructions to set the trajectory for development of learned responsiveness in the language acquisition process.

## 710      Ontogeny of auditory perception

The comparative studies on monkeys and birds may remind us that the specialization in humans might be *one* special device among many for auditory perception. The study of speech perception seems to disclose the most pertinent ethological problems which we have been discussing here: the development of perception, a specialized stimulus-response relationship, selective learning through species-specific predisposition, and thereby the nature-nurture intercalation – all at a level of analysis which is very close to central nervous mechanisms, an aspect of speech perception which I cannot discuss here.

## References

- Beecher, M. D., Zoloth, S. R., Petersen, M., Moody, D. and Stebbins, W. 1976. Perception of conspecific communication signals by Japanese macaques (*Macaca fuscata*). *Journal of the Acoustical Society of America*, 60(1):89 (abstract).
- Chomsky, N. 1959. Review of *Verbal behavior* by B. F. Skinner. *Language*, 36(1):26–58.
- Cutting, J. E. and Eimas, P. D. 1975. Phonetic feature analyzers and the processing of speech in infants. In J. F. Kavanagh and J. E. Cutting (eds.), *The role of speech in language*. Cambridge, Mass.: MIT Press.
- Cutting, J. E. and Rosner, B. 1974. Categories and boundaries of speech and music. *Perception and Psychophysics*, 16:564–70.
- Dorman, M. F. 1974. Auditory evoked correlates of speech sound discrimination. *Perception and Psychophysics*, 15:215–20.
- Dorman, M. F., Raphael, L. J. and Liberman, A. M. 1976. Further observations on the role of silence in the perception of stop consonants. *Haskins Laboratories Status Report on Speech Research*, SR-48:199–207.
- Dorman, M. F., Raphael, L. J., Liberman, A. M. and Repp, B. 1975. Masking-like phenomena in speech perception. *Journal of the Acoustical Society of America*, 57 (Supplement 1: S48(A)) (full text in *Haskins Laboratories Status Report on Speech Research*, SR-42/43, 265–76).
- Eimas, P. D. 1974. Auditory and linguistic processing of the cues for speech: Discrimination of the r-l distinction by young infants. *Perception and Psychophysics*, 16:513–21.
1975. Infant perception. In L. B. Cohen and P. Salapatek (eds.), *Infant perception: from sensation to cognition*, vol. II. New York: Academic Press.
- Eimas, P. D., Siqueland, E. R., Jusczyk, P. and Vigorito, J. 1971. Speech perception in infants. *Science*, 171:303–6.
- Erickson, D. M., Fitch, H. L., Halwes, T. G. and Liberman, A. M. 1977. Trading relation in perception between silence and spectra. *Journal of the Acoustical Society of America*, 61 (Supplement 1): S46–S47(A).
- Fitch, H. L., Erickson, D. M., Halwes, T. G. and Liberman, A. M. In preparation.
- Gaitenby, J. H. 1965. The elastic word. *Haskins Laboratories Status Report on Speech Research*, SR-2:3.1–3.12.

- Gibson, E. A. 1977. The development of perception as an adaptive process. In I. L. Janis (ed.), *Current trends in psychology*. Los Altos, California: William Kaufman Inc.
- Green, S. 1975. Variation of vocal pattern with social situation in the Japanese monkey (*Macaca fuscata*): a field study. In L. A. Rosenblum (ed.), *Primate behavior*, vol. IV. New York: Academic Press.
- Greenberg, J. H. 1966. *Language universals, with special reference to feature hierarchies*. The Hague: Mouton.
1969. Language universals: a research frontier. *Science*, 166:473-8.
- Hailman, J. P. 1967. Ontogeny of an instinct. *Behaviour Supplement*, 15:1-59.
1970. Comments on the coding of releasing stimuli. In L. R. Aronson, E. Tobach, D. S. Lehrman and J. S. Rosenblatt (eds.), *Development and evolution of behavior*. San Francisco: Freeman.
- Hinde, R. A. and Stevenson-Hinde, J. (eds.) 1973. *Constraints on learning*. New York: Academic Press.
- Huggins, A. W. F. 1972. On the perception of temporal phenomena in speech. *Journal of the Acoustical Society of America*, 51:1279-90.
- Kuhl, P. K. and Miller, J. D. 1975. Speech perception in early infancy: discrimination of speech sound categories. *Journal of the Acoustical Society of America*, 58 (Supplement 1): S56.
- Lasky, R., Syrdal-Lasky, A. and Klein, R. 1975. VOT discrimination by four-to-six-month-old infants from Spanish environments. *Journal of Experimental Child Psychology*, 20:215-25.
- Lehiste, I. 1970. *Suprasegmentals*. Cambridge: MIT Press.
- Liberman, A. M. 1957. Some results of research on speech perception. *Journal of the Acoustical Society of America*, 29:117-23.
1970. The grammars of speech and language. *Cognitive Psychology*, 1:301-23.
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P. and Studdert-Kennedy, M. 1967. Perception of the speech code. *Psychological Review*, 74:431-61.
- Liberman, A. M., Harris, K. S., Kinney, J. and Lane, H. 1961. The discrimination of relative onset time of the components of certain speech and non-speech patterns. *Journal of Experimental Psychology*, 61:379-88.
- Liberman, A. M., Mattingly, I. G. and Turvey, M. T. 1972. Language codes and memory codes. In A. W. Melton and E. Martin (eds.), *Coding processes in human memory*. Washington, DC: V. H. Winston.
- Liberman, A. M. and Pisoni, D. B. 1977. Evidence for a special speech-perceiving subsystem in the human. In T. H. Bullock (ed.), *Recognition of complex acoustic signals*. Life Sciences Research Report 5. Berlin: Dahlem Konferenzen.
- Liberman, A. M., Repp, B. H., Eccardt, T. and Pesetsky, D. 1977. Some relations between duration of silence and duration of friction noise as joint cues for fricatives, affricatives, and stops. *Journal of the Acoustical Society of America*, 62 (Supplement 1: S78(A)).
- Liberman, A. M. and Studdert-Kennedy, M. In press. Phonetic perception. In R. Held, H. Leibowitz, and H.-L. Teuber (eds.), *Handbook of sensory physiology*, vol. VIII, *Perception*. Heidelberg: Springer-Verlag Inc.
- Lisker, L. and Abramson, A. S. 1964. A cross-language study of voicing of initial stops: acoustical measurements. *Word*, 20:384-422.
- Marler, P. 1970. A comparative approach to vocal learning: song development in

- white-crowned sparrows. *Journal of Comparative and Physiological Psychology*, 71:1-25.
1976. Social organization, communication and graded signals: Vocal behavior of the chimpanzee and the gorilla. In P. Bateson and R. A. Hinde (eds.), *Growing points in ethology*. Cambridge: Cambridge University Press.
1977. Development and learning of recognition systems. In T. H. Bullock (ed.), *Recognition of complex acoustic signals*. Berlin: Dahlem Konferenzen.
- Marler, P. and Peters, S. 1977. Selective vocal learning in sparrows. *Science*, 198:519-21.
- Morse, P. A. 1972. The discrimination of speech and non-speech stimuli in early infancy. *Journal of Experimental Child Psychology*, 14:477.
- Pastore, R. E. 1976. Categorical perception: a critical re-evaluation. In S. K. Hirsh et al. (eds.), *Hearing and Davis: Essays honoring Hallowell Davis*. St Louis, Missouri: Washington University Press.
- Piaget, J. 1968. *The language and thought of the child*. 3rd edn. London: Routledge and Kegan Paul.
- Raphael, L. J., Dorman, M. F. and Liberman, A. M. 1976. Some ecological constraints on the perception of stops and affricates. *Journal of the Acoustical Society of America*, 59 (Supplement 1): S25A.
- Repp, B. H., Liberman, A. M., Eccardt, T. and Pesetsky, D. 1978 (in press). Perceptual integration of cries for stop, fricative, and affricative manner. *Journal of Experimental Psychology: Human Perception and Performance*, 4.
- Rowell, T. E. 1962. Agonistic noises of the rhesus monkey. *Symposia of the Zoological Society of London*, 8:91-6.
- Rowell, T. E. and Hinde, R. A. 1962. Vocal communication by the rhesus monkey (*Macaca mulatta*). *Proceedings of the Zoological Society of London*, 138:279-94.
- Sackett, G. P. 1966. Monkeys reared in isolation with pictures as visual input: evidence for an innate releasing mechanism. *Science*, 154:1468-73.
- Schott, D. 1975. Quantitative analysis of the vocal repertoire of squirrel monkeys (*Saimiri sciureus*). *Zeitschrift für Tierpsychologie*, 38:225-50.
- Seligman, M. E. P. and Hager, P. L. 1972. *Biological boundaries of learning*. New York: Appleton-Century-Crofts.
- Shettleworth, S. J. 1972. Constraints on learning. In D. S. Lehrman, R. A. Hinde and E. Shaw (eds.), *Advances in the study of behavior*. New York: Academic Press.
- Skinner, B. F. 1957. *Verbal behavior*. New York: Appleton-Century-Crofts.
- Streeter, L. A. 1975. Language perception of 2-month old infants shows effects of both innate mechanisms and experience. *Nature*, 259:39-41.
- Studdert-Kennedy, M. G. 1977. Universals in phonetic structure and their role in linguistic communication. In T. H. Bullock (ed.), *Recognition of complex acoustic signals*. Berlin: Dahlem Konferenzen.
- Studdert-Kennedy, M., Liberman, A. M., Harris, K. S. and Cooper, F. S. 1970. The motor theory of speech perception: a reply to Lane's critical review. *Psychological Review*, 77:234-49.
- Tinbergen, N. 1951. *The study of instinct*. London: Oxford University Press.
- Tinbergen, N. and Perdeck, A. C. 1950. On the stimulus situation releasing the begging response in the newly-hatched herring gull chick (*Larus a. argentatus* Pont.). *Behaviour*, 3:1-38.

714      Ontogeny of auditory perception

- Winter, P., Ploog, D. and Latta, J. 1966. Vocal repertoire of the squirrel monkey (*Saimiri sciureus*), its analysis and significance. *Experimental Brain Research*, 1:359-84.
- Wood, C. C., Goff, W. P. and Day, R. S. 1971. Auditory evoked potentials during speech perception. *Science*, 173:1248-51.
- Zoloth, S. and Green, S. In press. Perception of intergraded vocal signals by Japanese macaques. *Brain Behavior and Evolution*.