Human Neurobiol (1983) 2:191–195

Human Neurobiology

On learning to speak

M. Studdert-Kennedy

Department of Communication Arts and Sciences, Queens College, City University of New York, NY 11367 and Haskins Laboratories, New Haven, CT 06510, USA

Summary. Every language, spoken or signed, deploys a large lexicon, made possible by permutation and combination of a small set of linguistic elements. In speech, rapid interleaving of the gestures that form these elements (consonants and vowels) leads to a complex acoustic signal in which the boundaries between elements are lost. However, for the child learning to speak, the initial task is not to recover these elements, but simply to imitate the sound pattern that it hears. Studies of "lipreading" in adults and infants suggest that imitation is mediated by an amodal representation, closely related to the dynamics of articulation, and that a left-hemisphere perceptuo-motor mechanism specialized to make use of this representation develops during the first six months of life. By drawing on this specialized mechanism, the infant learns the recurrent patterns of acoustic structure and articulatory gesture from which linguistic segments must be presumed to emerge.

Key words: Phonology – Motor requirements – Imitation – Articulation

As a system of animal communication, language has the distinctive property of being open, that is, fitted to carrying messages on an unlimited range of topics. Human cognitive capacity is, of course, greater than that of other animals, but this may be a consequence as much as a cause of linguistic range. Other primate communication systems have a limited referential scope – sources of food or danger, personal and group identity, sexual inclination, emotional state, and so on – and a limited set of no more than 10–40 signals (Wilson 1975, p 183). In fact, 10–40 holistically distinct signals may be close to the upper range of primate perceptual and motor capacity. The distinctive property of language is that it has finessed that upper limit, by developing a double structure, or dual pattern (Hockett 1958).

The two levels of patterning are phonology and syntax. The first permits us to develop a large lexicon, the second permits us to deploy the lexicon in predicating relations among objects and events. My present concern is entirely with the first level. A 6-year old middle-class American child already recognizes some 13,000 words (Templin 1957), while an adult's recognition vocabulary may be well over 100,000. Every language, however primitive the culture of its speakers by Western standards, deploys a large lexicon. This is possible because the phonology, or sound pattern, of a language draws on a small set (roughly between 20 and 100 elements) of meaningless units – consonants and vowels – to construct a very large set of meaningful units, words (or morphemes). These meaningless units may themselves be described in terms of a smaller set of recurrent, contrasting phonetic properties or features. Evidently, there emerged in our hominid ancestors a combinatorial principle (later, perhaps, extended into syntax) by which a finite set of articulatory gestures could be repeatedly permuted to produce a very large number of distinctively different patterns.

"Articulatory gesture" refers, at a gross level, to opening and closing the mouth. Repeated constriction of the vocal tract, somewhere between lips and glottis, to form consonants, and repeated opening of the tract by lowering the jaw, to form vowels, give rise to the basic consonant vowel syllable from which the sound patterns of all spoken languages are formed. The varying phonetic qualities of consonants and vowels are determined by the precise shape of the vocal tract through which sound – the buzz of vocal fold vibration or the hiss of air blown through a narrow constriction – is filtered. The shape of the resonating cavities of the vocal tract is determined by fine positioning of the articulators: raising, lowering, fronting or backing the tip, blade or body of the tongue, raising or lowering the velum, rounding or spreading the lips, and so on.

Thus, permutation and combination of some two dozen gestures provides ". . . a kind of impedance match between an open-ended set of meaningful symbols and a decidedly limited set of signaling devices" (Studdert-Kennedy and Lane 1980, p 35). Yet permutation and combination alone would not suffice for a flexible and open-ended system of communication, if the gestures were not executed rapidly enough to evade the limits of short-term memory and to match the natural rate of thought and action.

What this "natural rate" may be we do not know. But for English, at least, a typical rate of speech is of the order of 150 words/min. This reduces to roughly 10 to 15 phonemes (consonants and vowels)/s. As Cooper has remarked, such rates can be achieved ". . . only if separate parts of the articulatory machinery – muscles of the lips, tongue, velum, etc. – can be separately controlled, and if . . . a change of state for any one of these articulatory entities, taken together with the current state of others, is a change to . . . another phoneme . . . It is this kind of parallel processing

that makes it possible to get high-speed performance with low-speed machinery" (Liberman et al. 1967, p 446). Thus, repeated use of a small set of interleaved gestures may not only expand the potential lexicon, but also ensure rapid execution of its elements.

Let me conclude this brief introduction by noting that the dual motoric structure of spoken language has no known parallel in any other system of animal behavior, except manual-facial sign languages. Over the past 15–20 years we have learned that American Sign Language (ASL) (the first language of over 100,000 deaf persons, and the fourth most common language in the United States (Mayberry 1978) is a fully independent language with its own characteristic formational ("phonological") structure and syntax (Klima and Bellugi 1979). Whether signed language is a mere analog of spoken language or a true homolog, drawing on the same neural structures, we do not yet know, although studies of sign language deficits following left hemisphere lesion (e.g., Kimura et al. 1976; Poizner et al. in press) reveal remarkable parallels with aphasic deficits of spoken language users.

In any event, my point here is simply that each ASL sign is formed by combining four intrinsically meaningless components: a hand configuration, a palm orientation, a place in the body space where it is formed, and a movement. These four classes of component, like the two segmental classes of spoken language (consonants and vowels), may also be described in terms of a smaller set of recurrent, contrasting features (e.g., Klima and Bellugi 1979, Chapter 7). There are some fifty values, or "primes", distributed across the four dimensions and their combination in a sign follows "phonological rules", analogous to those that constrain the structure of a syllable in spoken languages. In short, both spoken and signed languages exploit combinatorial principles of lexical formation. Moreover, it would seem that short-term memory and cognitive capacity have constrained signed and spoken languages to similar rates of communication. For, although each ASL sign takes roughly three times as long to form as an English word, the proposition rates in the two languages are almost identical (Klima and Bellugi 1979). This is possible because, while the phonological and syntactic structures of a spoken language are largely implemented by sequential organization over time, a signed language can exploit simultaneous manual and facial gestures distributed in space. Thus, both types of language are grounded in a capacity for rapid, precise and precisely coordinated movements of a small set of articulators.

In what follows, I shall have little further to say about signed languages. Here, I simply note two points. First, we do not talk with our toes, and we may doubt whether any imaginable system of human articulators, other than those of the hand and mouth, would be capable of the motor speed and precision necessary to implement language, as we know it. Second, whatever the evolutionary sequence may have been, the well-established (albeit imperfect) correlation between hemispheric specializations for language and manual praxis is, I assume, not mere coincidence. In all likelihood, the two modes of language draw on closely related neural structures.

I have dwelt so far on motor requirements. But there are perceptual demands also. If spoken language is indeed constructed from rapid sequences of consonants and vowels, the listener must somehow extract these recurrent elements from the signal. Yet, from the earliest spectrographic studies (Joos 1948) it has been known that the acoustic flow of speech cannot be readily divided into an alphabetic sequence of invariant segments corresponding to the invariant segments of linguistic description. The reason for this is simply that we do not speak segment by segment, or even syllable by syllable. At any instant, the several articulators are executing a complex, interleaved pattern of movements, of which the spatio-temporal coordinates reflect the influence of several neighboring segments. (The reader may test this by slowly uttering, for example, the words *call* and *keel*. He will find that the position of his tongue on the palate during closure for the initial consonant, [k], is slightly further back for the first word than for the second.) The consequence of this imbricated pattern of movement is, of course, an imbricated pattern of sound, such that any particular acoustic segment typically specifies more than one linguistic segment, while any particular linguistic segment is specified by more than one acoustic segment (Fant 1962; Liberman et al. 1967). This lack of isomorphism between acoustic and linguistic structure is the central unsolved problem of speech perception. Its continued recalcitrance is reflected in the fact that we are little closer to automatic phonetic transcription of speech now than we were thirty years ago (Levinson and Liberman 1981).

Many different approaches to the problem have been proposed, but I will not review them here (see Studdert-Kennedy 1980 for fuller discussion). Instead, I will attempt to recast the problem by setting aside, for the moment, the discrepancy between acoustic signal and linguistic description, and simply asking what we know about how a child learns to speak. I shall assume that, whatever the process, it is sufficiently general to permit the deaf child to learn to sign with as much ease as a hearing child learns to speak. I note, further, that when a child learns to sign or speak, it learns a specific dialect. That is to say, it gradually discovers, in the detailed acoustic or optic patterns of its caretakers' signals, specifications for a no less detailed pattern of motor organization.

Stated in this way, the problem becomes a special case of the general problem of imitation. Relatively few species imitate. The higher primates imitate general bodily actions, but vocal imitation is peculiar to a few species of songbirds, certain marine mammals and humans. The capacity to imitate is evidently a rare, specialized capacity for discovering links between perceived movements and their corresponding motor controls.

We may gain insight into the bases of speech imitation from recent studies of "lip-reading" in adults and infants. That adults can learn to lip-read is, of course, a commonplace of aural rehabilitation, but the theoretical implications of this capacity have only recently begun to emerge. McGurk and MacDonald (1976) demonstrated that listeners' perceptions of a spoken syllable often change, if they simultaneously watch a video display of a speaker pronouncing a different syllable. For example, if listeners are presented with the acoustic syllable [ba] repeated four times, while watching a synchronized optic display of a speaker articulating [ba, va, ða, da], they will typically report the latter, optically specified sequence. That the effect is not simply a matter of visual dominance in a sensory hierarchy (Marks 1978) is evidenced by the fact that certain combinations (e.g., acoustic [ba] with

optic [ga] may be perceived as clusters ([bga] or [gba]), or even as syllables corresponding to neither display ([da]). Thus, listeners' percepts seem to arise from a process by which two distinct sources of information, acoustic and optic, are actively combined at an abstract level where each has already lost its distinctive sensory quality. (For fuller discussion, see Summerfield 1979).

Further evidence for an amodal representation of speech comes from a cross-modal study of the so-called suffix effect by Campbell and Dodd (1980). A standard finding of short-term memory studies is that listeners, recalling a list of auditorily presented words, recall those at the end better than those in the middle (recency effect). The effect is reduced if the list is presented graphically. Moreover, Crowder and Morton (1969) demonstrated that the effect could be abolished, or significantly reduced, if a spoken word was appended to the list, not for recall but simply as a signal to begin recall (suffix effect). Presumably, the suffix "interferes" in some way with the representation of recent items. That this representation is at some relatively "low", yet structured, level is argued by the facts that the effect (1) is unaffected by degree of semantic similarity between suffix and list, (2) is reduced if suffix and list are presented to opposite ears, (3) does not occur if the suffix is a tone or burst of noise.

Campbell and Dodd (1980) used this paradigm to test listeners' recall of digits, either lip-read (without sound) or presented graphically, with and without the spoken suffix, "ten" (heard, but not seen). They found significant recency and suffix effects for the lip-read, but not for the graphic, lists. In a complementary study, Spoehr and Corin (1978) demonstrated that a lip-read suffix reduced recall of auditorily presented lists. Evidently, speech heard, but not seen, and speech seen, but not heard, share a common representation. Moreover, the fact that Campbell and Dodd did not find a suffix effect for graphically presented lists suggests that this shared representation is not at some abstract, phonological level where spoken and written language converge. Rather, these studies, like that of McGurk and MacDonald (1976), hint at a representation in some form common to both the light reflected and the sound radiated from mouth and lips.

Consider, now, that infants are also sensitive to structural correspondences between the acoustic and optic specifications of an event. Spelke (1976) showed that 4-month-old infants preferred to watch the film (of a woman playing "peekaboo", or of a hand rhythmically striking a wood block and a tambourine with a baton) that matched the sound track they were hearing. Dodd (1978) showed that 4-month-old infants watched the face of a woman reading nursery rhymes more attentively, when her voice was synchronized with her facial movements than when it was delayed by 400 ms. If these preferences were merely for synchrony, we might expect infants to be satisfied with any acoustic-optic pattern in which moments of abrupt change are arbitrarily synchronized. Thus, in speech they might be no less attentive to an articulating face whose closed mouth was synchronized with syllable amplitude peaks and open mouth with amplitude troughs than to the (natural) reverse. However, Kuhl and Meltzoff (1982) showed that 4–5 month old infants looked longer at the face of a woman articulating the vowel they were hearing (either [i] or [a]) than at the same face articulating the other vowel *in synchrony*. Moreover, the preference disappeared, when the signals were pure tones, matched in amplitude and duration to the vowels, so that the infant preference was evidently for a match between a mouth shape and a particular spectral structure. Similarly, MacKain et al. (1983) showed that 5–6 month old infants preferred to look at the face of a woman repeating the disyllable they were hearing (e.g., [zuzi] than at the synchronized face of the same woman repeating another disyllable (e.g., [vava]).

In both these studies, the infants' preferences were for natural structural correspondences between acoustic and optic information. Both studies hint at infant sensitivity to intermodal correspondences that could play a role in learning to speak. However, I am not suggesting that optic information is necessary, since the blind infant also learns to speak[1]. My intent rather is to gain leverage on the puzzle of imitation. What we need therefore is to establish that the underlying metric of auditory-visual correspondence is related to that of the auditory-motor correspondence required for an individual to imitate the utterances of another.

To this end we may note, first, the visual-motor link evidenced in the capacity to imitate facial expression and, second, the association across many primate species between facial expression and pattern of vocalization (Marler 1975; Hooff 1976; Ohala in press). Recently, Field et al. (1982) reported that 36-h-old infants could imitate the "happy, sad and surprised" expressions of a model. However, these are relatively stereotyped emotional responses that might be evoked without recourse to the visual-motor link required for imitation of novel movements. More striking is the work of Meltzoff and Moore (1977) who showed that 12- to 21-day-old infants could imitate both arbitrary mouth movements, such as tongue protrusion and mouth opening, and (of particular interest for the acquisition of ASL) arbitrary hand movements, such as opening and closing the hand by serially moving the fingers. Here mouth opening was elicited without vocalization; but had vocalization occurred, its structure would, of course, have reflected the shape of the mouth. Kuhl and Meltzoff (1982) do, in fact, report as an incidental finding of their study that 10 of their 32 4- to 5-month-old infants ". . . produced sounds that resembled the adult female's vowels. They seemed to be imitating the female talker, 'taking turns' by alternating their vocalizations with hers" (p 1140). If we accept the evidence that the infants of this study were recognizing acoustic-optic correspondences, and add to it the results of the adult lip-reading studies, calling for a metric in which acoustic and optic information are combined, then we may conclude that the perceptual structure controlling the infants' imitations was specified in this common metric.

Evidently, the desired metric must be ". . . closely related to that of articulatory dynamics" (Summerfield 1979, p 329). Following Runeson and Frykholm (1981) (see also Summerfield 1980), we may suppose that in the visual perception of an event we perceive not simply the surface kinematics (displacement, velocity, acceleration), but also the underlying biophysical properties that define the structure being moved and the forces that move it (mass, force, momentum, elasticity,

[1] I have often heard it said that blind children develop language more slowly than their sighted peers, but I know of no systematic study on the topic.

and so on). Similarly, in perceiving speech, we perceive not only its "kinematics", that is, the changes and rates of change in spectral structure, but also the underlying dynamic forces that produce these changes. In other words, to perceive speech is to perceive movements of the articulators, specified by a pattern of radiated sound, just as we perceive movements of the hand, specified by a pattern of reflected light.

The close link, for the infant, between perceiving speech and producing it, is further suggested by a curious aspect of the study by MacKain et al. (1983), cited earlier. This is the fact that infants' preferences for a match between the facial movements they were watching and the speech sounds they were hearing was statistically significant only when they were looking to their right sides. Fourteen of the eighteen infants in the study preferred more matches on their right sides than on their left. Moreover, in a follow-up investigation of familial handedness, MacKain and her colleagues have learned that six of the infants have left-handed first or second order relatives. Of these six, four are the infants who displayed more left-side than right-side matches.

These results can be interpreted in the light of studies by Kinsbourne and his colleagues. Kinsbourne (1972) found that right-handed adults tended to shift their gaze to the right, while solving verbal problems, to the left, while visualizing spatial relations; left-handers tended to shift gaze in the same direction for both types of task, with each direction roughly equally represented across the subject group. Lempert and Kinsbourne (1982) showed that the effect was reversible for right-handed subjects on a verbal task: subjects who rehearsed sentences, with head and eyes turned right, recalled the sentences better than subjects who rehearsed, while turned left. Thus, attention to one side of the body may facilitate processes for which the contralateral hemisphere is specialized.

Extending this interpretation to the infants of MacKain et al. (1983), we may infer that infants with a preference for matches on the right side, rather than the left, were revealing a left hemisphere capacity for recognizing acoustic-optic correspondence in speech. If, further, the metric specifying these correspondences is the same as that specifying the auditory-motor correspondences necessary for imitation (as was argued above), we may conclude that 5- to 6-month-old infants already display a speech perceptuo-motor link in the left hemisphere.

How early this link may develop we do not yet know. However, Best et al. (1982), testing 2-, 3- and 4-month-old infants dichotically, in a cardiac habituation paradigm, found a right-ear advantage for speech and a left-ear advantage for music in the 3 and 4 month olds, but only a left-ear advantage for music in the 2 month olds. We may suspect, then, that the perceptual component of the speech link begins to develop between the second and third months of life. By 5–6 months, close to the onset of babbling, the motor component is beginning to emerge. By the end of the first year, as babbling fades, the infant would be equipped with the perceptuo-motor mechanisms necessary for imitating the sounds of the language it is going to learn.

In conclusion, let me recall the paradoxical discrepancy between the speech signal and its linguistic description with which I began. The approach to imitation I have sketched deliberately sidesteps this problem. Yet it may ultimately contribute to its solution by focusing on the infant for whom the discrepancy does not yet exist, for the simple reason that the infant has not yet learned the phonetic categories of its language. Tracing the process by which the recurrent patterns of infant articulation coalesce into categorical linguistic units, evidenced by spoonerisms and other adult speech errors (Shattuck-Hufnagel 1979) is a task for the future. However, the task may be easier, if we see it as a problem in the development of a unique mode of motor control, characteristic of human language.

Acknowledgements. Preparation of this paper was supported in part by NICHD Grant No. HD-01994 to Haskins Laboratories. The argument of the paper extends and elaborates one first put forward as part of a paper read at a conference at the University of Connecticut, Storrs, in June 1980, to be published in: Warren WH, Shaw RE (eds) (in press) Persistence and Change: Proceedings of the First International Conference on Event Perception. Hillsdale, New Jersey

References

Best CT, Hoffman H, Glanville BR (1982) Development of infant ear asymmetries for speech and music. Perc Psychophys 31:75–85

Campbell R, Dodd B (1980) Hearing by eye. QJ Exp Psych 32:85–99

Crowder RG, Morton J (1969) Precategorical acoustic storage. (PAS) Perc Psychophys 5:365–373

Dodd B (1979) Lip reading in infants: Attention to speech presented in- and out-of-synchrony. Cog Psych 11:478–484

Fant CGM (1962) Descriptive analysis of the acoustic aspects of speech. Logos 5:3–17

Field TM, Woodson R, Greenberg R, Cohen D (1982) Discrimination and imitation of facial expressions by neonates. Science 218: 179–181

Hockett C (1958) A course in modern linguistics. Macmillan, New York

Hooff JARAM van (1976) The comparison of facial expression in man and higher primates. In: Cranach M von (ed) Methods of inference from human to animal behavior. Aldine, Chicago, pp 165–196

Joos M (1948) Acoustic phonetics. Lang Monog No 23, Vol 24 [Suppl]

Kelso JAS, Cook E, Olson ME, Epstein W (1975) Allocation of attention and the locus of adaptation to displaced vision. J Exp Psychol Hum Percept 1:237–245

Kimura D, Battison R, Lubert B (1976) Impairment of nonlinguistic hand movements in a deaf aphasic. Brain Lang 3:566–571

Kinsbourne M (1972) Eye and head turning indicates cerebral lateralization. Science 176:539–541

Klima ES, Bellugi U (1979) The signs of language. Harvard UP, Cambridge, Massachusetts

Kuhl PK, Meltzoff AN (1982) The bimodal perception of speech in infancy. Science 218:1138–1144

Lempert H, Kinsbourne M (1982) Effect of laterality of orientation on verbal memory. Neuropsychologia 20:211–214

Levinson SE, Liberman MY (1981) Speech recognition by computer. Sci Am 244:64–87

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) Perception of the speech code. Psychol Rev 74:431–461

MacKain KS, Studdert-Kennedy M, Spieker S, Stern D (1983) Infant intermodal speech perception is a left hemisphere function. Science 219:1347–1349

Marks LE (1978) Multimodal perception. In: Carterette EC, Friedman MP (ed) Handbook of perception, Vol VIII. Academic Press, New York, pp 321–339

Marler P (1975) On the origin of speech from animal sounds. In: Kavanagh JF, Cutting JE (eds) The role of speech in language. MIT Press, Cambridge, Massachusetts, pp 11–37

Mayberry RI (1978) Manual communication. In: Davis H, Silverman SR (eds) Hearing and deafness. 4th edn. Holt, Rinehart & Winston, New York

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:746–748

Meltzoff AN, Moore MK (1977) Imitation of facial and manual gestures by human neonates. Science 198:175–178

Ohala J (in press) Cross-language uses of pitch. Phonetica

Poizner H, Bellugi U, Iragui V (in press) Apraxia and aphasia for a visual-gestural language. Am J Physiol

Runeson S, Frykholm G (1981) Visual perception of lifted weight. J Exp Psychol Hum Percept 7:733–740

Shattuck-Hufnagel S (1979) Speech errors as evidence for a serial ordering mechanism in speech production. In: Cooper WE, Walker ECT (eds) Sentence processing: Psycholinguistic studies presented to Merrill Garrett. Lawrence Erlbaum Associates, Hillsdale, New Jersey

Spelke E (1976) Infants' intermodal perception of events. Cog Psych 8:553–560

Spoehr KT, Corin WJ (1978) The stimulus suffix as a memory coding phenomenon. Mem Cog 6:583–589

Studdert-Kennedy M (1980) Speech perception. Lang Speech 23:45–66

Studdert-Kennedy M, Lane H (1980) Clues from the differences between signed and spoken language. In: Bellugi U, Studdert-Kennedy M (eds) Signed and spoken language: biological constraints on linguistic form. Verlag Chemie Deerfield Park, Florida, pp 29–39

Summerfield Q (1979) Use of visual information for phonetic perception. Phonetica 36:314–331

Templin M (1957) Certain language skills of children. Minneapolis: University of Minnesota Press, Minneapolis

Wilson EO (1975) Sociobiology. The Belknap Press, Cambridge, Massachusetts