In Language Perception and Production
Edited by: Allport, A., MacKay, D.,
Prinz, W. and Scheerer, E.
London: Academic Press, 1987

4 The phoneme as a perceptuomotor structure

Michael Studdert-Kennedy

Abstract

Studies of speech and writing face a paradox: the discrete units of the written representation of an utterance cannot be isolated in its quasi-continuous articulatory and acoustic structure. We may resolve the paradox by positing that units of writing (ideographs, syllabic signs, alphabetic letters) are symbols for discrete, perceptuomotor, neural control structures, normally engaged in speaking and listening. Focussing on the phoneme, for which an alphabetic letter is a symbol, the paper traces its emergence in a child's speech through several stages: hemispheric specialization for speech perception at birth, early discriminative capacity followed by gradual loss of the capacity to discriminate among speech sounds not used in the surrounding language, babbling, and first words. The word, a unit of meaning that mediates the child's entry into language, is viewed as an articulatory routine, a sequence of a few variable gestures of lips, jaw, tongue, velum and larynx, and their acoustic correlates. Under pressure from an increasing vocabulary, recurrent patterns of sound and gesture crystallize into encapsulated phonemic control units. Once a full repertoire of phonemes has emerged, usually around the middle of the third year, an explosive growth of vocabulary begins, and the child is soon ready, at least in principle, for the metalinguistic task of learning to read.

Ever since I . . . started to read . . . there has never been a line that I didn't *hear*. As my eyes followed the sentence, a voice was saying it silently to me. It isn't my mother's voice, or the voice of any person I can identify, certainly not my own. It is human, and it is inwardly that I listen to it.

Eudora Welty (1983, p. 12).

LANGUAGE PERCEPTION AND PRODUCTION
ISBN 0-12-052750-2

Introduction

Any discussion of the relation between speech and writing faces a paradox: the most widespread and efficient system of writing, the alphabet, exploits a unit of speech, the phoneme, for the physical reality of which we have no evidence. To be sure, we have evidence of its psychological reality. But, ironically, that evidence depends on the alphabet itself. How are we to escape from this circle?

First, let me elaborate the terms of the paradox. Since the earliest spectrographic, cineradiographic and electromyographic studies, we have known that neither the articulatory nor the acoustic flow of speech can be divided into a sequence of segments corresponding to the invariant segments of linguistic description. Whether the segments are words, morphs, syllables, phones or features, the case is the same. The reason for this is simply that we do not normally speak phoneme by phoneme, syllable by syllable, or even word by word. At any instant, our articulators are executing a complex interleaved pattern of movements of which the spatio-temporal coordinates reflect the influence of several neighboring segments. The typical result is that any isolable articulatory or acoustic segment arises as a vector of forces from more than one linguistic segment, while any particular linguistic segment distributes its forces over several articulatory and acoustic segments. This lack of isomorphism between articulatory-acoustic and linguistic structure is the central unsolved problem of speech research (Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1967; Pisoni, 1985). Its continued recalcitrance is reflected in the fact that (apart from a variety of technologically ingenious, but limited and brute force solutions) we are little closer to automatic speech recognition today than we were thirty years ago (Levinson and Liberman, 1981).

What then is the evidence for the psychological reality of linguistic segments? (I confine my discussion to the phoneme, although most of what follows would apply *mutatis mutandis* to all other levels of description.) First and foremost is the alphabet itself. Superficially, we might take the alphabet (or any other writing system) to be a system of movement notation analogous to those used by ethologists to describe, say, the mating behavior of Tasmanian devils (Golani, 1981). The difference lies in their modes of validation. The ethologist's units may or may not correspond to motor control structures in the devil's behavior; the units are sufficiently validated, if they lend order and insight to the ethologist's understanding of that behavior. By contrast, the alphabet (like music and dance notation) is validated by the fact that it serves not only to notate, but to control behavior: we both write and read. Surely, we could not do so with such ease, if alphabetic symbols did not correspond to units of perceptuomotor control. A writing system constructed from arbitrary units — phonemes and a half, quarter words — would be of limited utility. We infer then that lexical items (words, morphemes) are stored as sequences of abstract perceptuomotor units (phonemes) for which letters of an alphabet are symbols.

If this is so, those who finger the phoneme as a fictitious unit imposed on speech by linguists because they know the alphabet (e.g., Warren, 1976) have it backwards. Historically, the possibility of the alphabet was discovered, not invented. Just as the bicycle was a discovery of locomotor possibilities implicit in the cyclical motions of walking and running, so the alphabet was a discovery of linguistic possibilities implicit in patterns of speaking.

Of course, we do have other important sources of evidence that confirm the psychological reality of the phoneme: errors of perception (e.g., Browman, 1980) and production[1] (e.g., MacKay (Chap 18); Shattuck-Hufnagel, 1983), backward talking (Cowan, Leavitt, Massaro and Kent, 1982), aphasic deficit (e.g., Blumstein, 1981). But we can only collect such data because we have the metalinguistic awareness and notational system to record them. Illiterates may make speech errors (MacKay, 1970), and oral cultures certainly practise alliteration and rhyme in their poetry. But, like the illiterate child who relishes 'Hickory dickory dock', they probably do not know what they are doing (cf. Morais, Cary, Alegria and Bertelson, 1981). Thus, the data that confirm our inferences from the alphabet rest squarely on the alphabet itself.

The paradox I have outlined might be resolved, if we could conceptualize the relation between a letter of the alphabet (or a word) and the behavior that it symbolizes. Just how difficult this will be becomes apparent, if we compare the information conveyed by a spoken word with the information conveyed by its written counterpart. An experimenter may ask a willing subject either to repeat a spoken word or to read aloud its written form. The subject's utterances in the two cases will be indiscriminable, but the information that controlled the utterances will have been quite different. The distinction, due to Carello, Turvey, Kugler and Shaw (1984) (see also Turvey and Kugler, 1984) is between information that *specifies* and information that *indicates* or *instructs*. The information in a spoken word is not arbitrary: its acoustic structure is a lawful consequence of the articulatory gestures that shape it. In other words, its acoustic structure is *specific* to those gestures, so that the prepared listener can follow the specifications to organize his own articulation and reproduce the utterance. Of course, we do not need the full specification of an utterance, in all its phonetic detail, in order to perceive it correctly, as those who know a foreign language, yet speak it with an accent, demonstrate: capturing all the details calls for a subtle process of perceptuomotor attunement. But it is evident that the specification does suffice for accurate reproduction, given adequate perceptuomotor skill, both in the child who slowly comes to master a surrounding dialect and in the trained phonetician who precisely mimics that dialect.

By contrast, the form of a written word is an arbitrary convention, a string of symbols that *indicate* to a reader what he is to do, but do not tell him how to do it. What is important here is that indicational information cannot control action in the absence of information specific to the act to be performed. That is why we may find it easier to imitate the stroke of a tennis coach than to implement his verbal instructions. Similarly, we can only pronounce a written word, if we have information specifying the correspondences between the symbol string and the motor control structures that must be engaged for speaking. These are the correspondences that an illiterate has not discovered.

The question now is simply this: what is the relation between a discrete symbol and the continuous motor behavior that it controls? If a written symbol does indeed stand for a motor control structure, as argued above, we may put the question in a slightly more concrete form: What is the relation between a discrete motor control structure and the complex pattern of movements that it generates? The answer will certainly not come in short order. But perhaps we can clarify the question, and gain insight into possible lines of answer by examining how units of perceptuomotor control emerge, as a child begins to speak its first language.

Basic to this development is the child's capacity to imitate, that is, to reproduce utterances functionally equivalent to those of the adults around it. We have claimed above that an utterance specifies the articulation necessary to reproduce it. But until we spell out what specification entails, the claim amounts to little more than the observation that people can repeat the words they hear. At least three questions must be answered, if we are to put flesh on the bones.

First is the question of how a listener (or, in lipreading, a viewer) transduces a pattern of sound (or light) into a matching pattern of muscular controls, sufficient to reproduce the modeled event. We can say very little here other than that the acoustic/optic pattern must induce a neural structure isomorphic with itself. The pattern must be abstract in the sense that it no longer carries the marks of its sensory channel, but concrete in that it specifies (perhaps quite loosely, as we shall see below) the muscular systems to be engaged: no one attempts to reproduce a spoken utterance with his feet. The perceptuomotor structure is therefore specific to the speech system. Perhaps it is worth remarking that, in the matter of transduction, the puzzle of imitation seems to be a special case of the general puzzle of how an animal modulates its actions to fit the world it perceives.

The second question raised by imitation concerns the units into which the modeled action is parsed. Research in speech perception has been preoccupied with units of linguistic analysis: features and phonemes. These, as normally defined, are abstract units, unsuited to an account of imitation, because, whatever their ultimate function in the adult speaker, they do not correspond to primitives of motor control that a child might engage to imitate an utterance in a language that it does not yet know. The human vocal apparatus comprises several discrete, partially independent articulators (lips, jaw, tongue, velum, larynx) by which energy from the respiratory system is modulated. The perceptual units of imitation must therefore be structures that specify functional units of motor control, corresponding to actions of the articulators. Isolation of these units is a central task for future research. We will come back to this matter below.

A third issue for the study of speech imitation is the notorious many-to-one relation between articulation and the acoustic signal (Porter, Chap 5). Speakers who normally raise and then lower their jaws in producing, say, the word, 'Be!', may execute acoustically identical utterances with pipes clenched between their teeth. The rounded English vowel of, say, *coot* may be produced either with protruded lips and the tongue humped just in front of the velum, or with spread lips, the tongue further backed and the larynx lowered. Even more bizarre articulations are discovered by children, born without tongue blade and tip, who none the less achieve a surprisingly normal phonetic repertoire (MacKain, 1983). Thus, the claim that an utterance specifies its articulation cannot mean that it specifies precisely which articulators are to be engaged, and when. Rather, it must mean that the utterance specifies a range of functionally equivalent articulatory actions. Of course, functional (or motor) equivalence is not peculiar to speech and may be observed in animals as lowly as the mouse (Fentress, 1981, 1983; Golani, 1981). Solution of the problem is a pressing issue in general research on motor control. For speech (and for other forms of vocal imitation, in songbirds and marine mammals) we have an added twist: the arbiter of equivalence is not some effect on the external world — seizing prey, peeling fruit, closing a door — but a listener's judgment.

Early perceptual development

With all this in mind, let us turn to the infant. Perceptually, speech already has a unique status for the infant within a few hours or days of birth. Neonates discriminate speech from non-speech (Alegria and Noirot, 1982), and, perhaps as a result of intrauterine stimulation, prefer their mothers' voices to strangers' (DeCasper and Fifer, 1980). Studies of infants from one to six months of age, using a variety of habituation and conditioning techniques, have shown that infants can discriminate virtually any speech sound contrast on which they are tested, including contrasts not used in the surrounding language (see Eimas (1985) for review). However, similar results from lower animals (chinchillas, macaques) indicate that infants are here drawing on capacities of the general mammalian auditory system (see Kuhl (1986) for review).

Dissociation of left and right sides of the brain for speech and nonspeech sounds respectively, measured by relative amplitude of auditory evoked response over left and right temporal lobes, may be detected within days of birth (Molfese, 1977). Left and right hemisphere short term memories for syllables and musical chords, respectively, measured by habituation and dishabituation of the cardiac orienting response to change, or lack of change, in dichotic stimulation, are developing by the third month (Best, Hoffman and Glanville, 1982). These and other similar results (see Best, *et al.* (1982) and Studdert-Kennedy (1986) for review) are important, because many descriptive and experimental studies have established that speech perceptuomotor capacity is vested in the left cerebral hemisphere of more than 90% of normal adults.

At the same time, we should not read these results as evidence of 'hard wiring'. At this stage of development not even the modality of language is fixed. If an infant is born deaf, it will learn to sign no less readily than its hearing peers learn to speak. Recent studies of 'aphasia' in native American Sign Language signers show striking parallels in forms of breakdown between signers and speakers with similar left hemisphere lesions (Bellugi, Poizner and Klima, 1983). Thus, the neural substrate is shaped by environmental contingencies, and the left hemisphere, despite its predisposition for speech, may be usurped by sign (Neville, 1980, 1985; Neville, Kutas and Schmidt, 1982). Given the diversity of human languages to which an infant may become attuned, such a process of epigenetic development is hardly surprising.

Early motor development

The development of motor capacity over the first year of life may be divided into a period before babbling (roughly, 0–6 months) and a period of babbling (7–12 months) (Oller, 1980). At birth, the larynx is set relatively high in the vocal tract, so that the tongue fills most of the oral cavity, limiting tongue movement and therefore both the possible points of intraoral constriction, or closure, and the spectral range of possible vocalic sounds. Accordingly, early sounds tend to be neutral, vowel-like phonations, often nasalized (produced with lowered velum), with little variation in degree or placement of oral constriction. As the larynx lowers, the variety of nonreflexive, nondistress sounds increases. By the second trimester, sounds include labial trills ('raspberries'), squeals and primitive syllabic patterns, formed by a consonant-like closure followed by a vowel-like resonance. These syllabic patterns lack the precise timing of closure, release and opening characteristic of mature syllables.

In fact, the onset of true or canonical babbling (often a quite sudden event around the seventh month) is marked by the emergence of syllables with the timing pattern (including closing to opening ratio), typical of natural languages (Oller, 1986). In the early months, syllables tend to be reduplicated (e.g., [bababa], [mamama], [dadada]); these give way in later months to sequences in which both consonant and vowel vary. Phonetic descriptions of babbled consonants (e.g., predominance of stops, glides, nasals, scarcity of fricatives, liquids, consonant clusters) tend to be similar across many language environments, including that of the deaf infant (Locke, 1983). We may therefore view these preferences as largely determined by universal anatomical, physiological and aerodynamic constraints on vocal action. At the same time, as we might expect in a behavior geared for environmental shaping, the repertoire is not rigid: individual infants vary widely both in how much they babble and in the relative frequency of their babbled sounds (MacNeilage, Hutchinson and Lasater, 1981).

We should emphasize that segmental phonetic descriptors are simply a convenient, approximate notation of what a child seems to be doing with its articulators — the only descriptors we have, in the absence of cineradiographic or other quantitative data. We should not infer that the child has independent, articulatory control over consonantal and vocalic portions of a syllable. The syllable, formed by rhythmically opening and closing the mouth, is a natural, cohesive unit of speech, with temporal properties that may be determined, in part, by the resonant frequency of the jaw. Its articulatory structure is perhaps related — at least by analogy, if not by homology — to the soft, tongue- or lip-modulated patterns of sound observed in the intimate interactions of Japanese macaque monkeys (Green, 1975; MacNeilage, personal communication).

Early perceptuomotor development

Imitation, long thought to be the outcome of a lengthy course of cognitive development (e.g., Piaget, 1962), is now known to be an innate capacity of the human infant. Meltzoff and Moore (1977, 1983) have shown, in a pair of meticulously controlled studies, that infants, within 72 hours of birth, can imitate arbitrary facial gestures (mouth opening, lip protrusion) and within 12–21 days (perhaps earlier, but we have no data) can also imitate tongue protrusion and sequential closing of the fingers (of particular interest for sign language acquisition). Of course, these are relatively crude gestures, far from the subtly interleaved patterns of movement, coordinated across several articulators, that are necessary for adult speech. The importance of the work lies in its implication that optically conveyed, facial gestures, already at birth, induce a neural structure isomorphic with the movements that produce them.

We should not expect speech sounds to induce an analogous neuromotor control structure at birth, not only because the sounds are complex, but because, as language diversity attests, speech is learned. Nonetheless, we might reasonably predict an early, amodal, *perceptual* representation of speech, since this must be the ground on which imitation is based. At present, we have to wait until 4–5 months for this, perhaps because appropriate studies have not yet been done on younger infants. Kuhl and Meltzoff (1982) showed that infants of this age looked longer at the videotaped face of a woman repeatedly articulating the vowel they were hearing (either [i] or [a]) than at the same face articulating the other vowel *in synchrony*. The preference disappeared when the signals were pure tones, matched in amplitude and duration to the vowels, so that infant preference was evidently for a match between a mouth shape and a particular spectral structure. Since spectral structure is directly determined by the resonant cavities of the vocal tract, and since the shape and volume of these cavities are determined by articulation (including pattern of mouth opening for [i] and [a]), the correspondence between mouth shape (optic) and spectral structure (acoustic) reflects their common source in articulation. Evidently, infants of 4–5 months, like adults in recent studies of lip-reading (e.g., McGurk and MacDonald, 1976; Summerfield, 1979, in press; Campbell, Chap 7) already have an amodal representation of speech, closely related to the articulatory structures that determine phonetic form.

Just how close this relation is we may judge from a second study similar to that of Kuhl and Meltzoff (1982). MacKain, Studdert-Kennedy, Spieker and Stern (1983) showed that 5–6 month old infants preferred to look at the videotaped 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]). However, the two faces were presented to left and right of an infant's central gaze, and the preference for an acoustic-optic match was only significant when infants were looking at the right side display. We may interpret this result in light of studies by Kinsbourne and his colleagues (e.g., Kinsbourne, 1972; Lempert and Kinsbourne, 1982), demonstrating that attention to one side of the body facilitates processes for which the contralateral hemisphere is specialized. Infants might then be more sensitive to acoustic-optic correspondences in speech presented on their right sides than on their left. Thus, infants of 5–6 months may already have an amodal representation of speech in the hemisphere that will later coordinate the activity of their bilaterally innervated speech apparatus.

Signally absent from all of the foregoing is any indication that the infant is affected by the surrounding language. In fact, it has often been proposed (e.g., Brown, 1958) that the infant's phonetic repertoire drifts towards that of its native language during the babbling of the second half year, but, despite several studies, no firm evidence of babbling drift has been found (Locke, 1983). We do, however, have evidence of perceptual effects. Werker and her colleagues (Werker, 1982; Werker, Gilbert, Humphrey and Tees, 1981; Werker and Tees, 1984) have shown, in several cross-sectional and longitudinal studies, that, during the second half year, infants may gradually lose their capacity to distinguish sound contrasts not used in their native language. This is perhaps just the period when an infant is first attending to individual words and the situations in which they occur (Jusczyk, 1982; MacKain, 1982).

The general picture of perceptuomotor development over the first year, then, is of two parallel, independent processes, with production trailing perception. Doubtless, physiological changes in the left hemisphere are taking place, laying down neural networks that will later make contact. These processes may resemble those in songbirds, such as the marsh wren, in which the perceptual template of its species' song is laid down during a narrow sensitive period many weeks before it begins to sing (Kroodsma, 1981). The first behavioral evidence of a perceptuomotor link then appears with the bird's first song and, in the infant, with its first imitation of an adult sound.

First words and their component gestures

Up to this point we have talked easily of perceptual, or perceptuomotor, 'representations' without asking what is represented. During the 1970s, when intensive work on child phonology began, researchers quite reasonably assumed that units of acquisition would be those that linguists had found useful in describing adult language: features and phonemes. Little attention was paid to the fact that these units, as defined by linguists, were abstract descriptors that could not be specified either articulatorily or acoustically, and were therefore of dubious utility to the child striving to talk like its companions. The oversight was perhaps encouraged by division of labor between students of perception and students of production whose mutual isolation absolved them from confronting what the child confronts: the puzzle of the relation between listening and speaking.

Over the past decade, child phonologists have come to recognize the fact, borne in also by pragmatic studies (e.g., Bates, 1979), that a child's entry into language is mediated by meaning; and meaning cannot be conveyed by isolated features or phonemes. The child's earliest unit of meaning is probably the prosodic contour: the rising pitch of question and surprise, the falling pitch of declaration, and so on, often observed in stretches of 'jargon' or intonated babble (Menn, 1978). The earliest *segmental* unit of meaning is the word (or formulaic phrase).

Evidence for the word as the basic unit of contrast in early language is rich and subtle (Moskowitz, 1973; Ferguson and Farwell, 1975; Ferguson, 1978; Macken, 1979; Menn, 1983a). Here I simply note three points. First is the observation that phonetic forms mastered in one word are not necessarily mastered in another. For example, a 15-month old child may execute [n] correctly in *no*, but substitute [m] for [n] in *night*, and [b] for [m] in *moo* (Ferguson and Farwell, 1975). Thus, the child does not contrast [b], [m] and [n], as in the adult language, but the three words with their insecurely grasped onsets.

A second point is that early speech is replete with instances of consonant harmony, that is, words in which one consonant assimilates the place or manner of articulation of another — even though the child may execute the assimilated consonant correctly in other words. Thus, a child may produce *daddy* and *egg* correctly, but offer [gɔg] for *dog* and [dʌt] for *duck* and *truck*: the child seems unable to switch place of articulation within a syllable. Such 'assimilation at a distance' suggests that the word is 'assembled before it is spoken' as a single prosodic unit (Menn, 1983a, p. 16).

The third point is that individual words may vary widely in their phonetic form from one occasion to another. A striking example comes from Ferguson and Farwell (1975). They report ten radically different attempts by a 15-month old girl, K, to say *pen* within one half-hour session: [mãᵊ, ˇʌ̄, deᵈⁿ̣, hin, ᵐbõ, pʰɪn, tʰn̩tʰn̩tʰn̩, baʰ, d̥ʰauᴺ, buã].[2] On the surface, these attempts seem almost incomprehensibly diverse one from another and from their model. But the authors shrewdly remark that 'K seems to be trying to sort out the features of nasality, bilabial closure, alveolar closure, and voicelessness' (Ferguson and Farwell, 1975, p. 14). An alternative description (to be preferred, in my view, for reasons that will appear shortly) would be to say that all the *gestures* of the model (lip closure, tongue raising and fronting, alveolar closure, velum lowering/raising, glottal opening/closing) are to be found in one or other of these utterances, but that the gestures are incorrectly phased with respect to one another. For example, lip closure for the initial [p], properly executed with an open glottis and raised velum, will yield [^{m}b], as in [^{m}bõ], if glottal closure for [ɛn] and velum lowering for [n] are initiated at the same time as lip closure, tens of milliseconds earlier than in the correct utterance, [pɛn]. Thus, the adult model evidently specified for the child the required gestures, but not their relative timing. (Notice, incidentally, that the only gestures present in the child's attempts, but absent from the model, are tongue backing and tongue lowering for the sounds transcribed as [o], [a], and [u]. Four of these five 'errors' occur when the child has successfully executed initial lip closure, as though attention to the initial gesture had exhausted the child's capacity to assemble later gestures.)

One reason for preferring a gestural to a featural description of a child's — or for that matter of an adult's — speech is that it lends the description observable, physical content (Browman and Goldstein, 1986). We are then dealing with patterns of movement in space and time, accessible to treatment according to general principles of motor control (e.g., Kelso, Tuller and Harris, 1983; Saltzman and Kelso, 1987). For example, the problem of motor equivalence may become more tractable, because the gesture is a *functional* unit, an equivalence class of coordinated movements that achieve some end (closing the lips, raising the tongue, etc.) (Kelso, Saltzman and Tuller, 1986). Moreover, a gestural description may help us to explore the claim (based on the facts of imitation) that the speech percept is an amodal structure isomorphic with the speaker's articulation. Glottal, velic and labial gestures can already be isolated by standard techniques; tongue movements are more problematic, because they are often vectors of two or more concurrent gestures. Nonetheless, positing a concrete, observable event as the fundamental unit of production may help researchers to analyze articulatory vectors into their component forces, and to isolate the acoustic marks of those vectors in the signal.

Finally, to forestall misunderstanding, a gestural description is not simply a change of terminology. Gestures do not usually correspond one-to-one with either phonemes or features. The phoneme /m/, for example, comprises the precisely timed and coordinated gestures of bilabial closure, velum lowering and glottal closing. The gesture of bilabial closure corresponds to several features [− continuant], [+ anterior], [+ consonantal], etc. A gestural account of speech — that is, an account grounded in the anatomy and physiology of the speaker — will require extensive revision of standard featural or segmental descriptions (Browman and Goldstein, 1986).

From words to phonemes

To summarize the previous section, we have argued that: (1) an element of meaning, the word, is the initial segmental unit of contrast in early speech; (2) a word is a coordinated pattern of gestures; (3) an adult spoken word specifies for the child learning to speak, at least some of its component gestures, but often not their detailed temporal organization. (The third point does not imply that the child's perceptual representation is necessarily incomplete: the representation may be exact, and the child's difficulty solely in coordinating its articulators. The difference is not without theoretical interest, but, in the present context, our focus is on how a child comes to reorganize a holistic pattern of gestures into a sequence of phonetic segments, or phonemes. Whether the reorganization is perceptual, articulatory, or both need not concern us.)

What follows, then, is a sketch of the process by which phonemes seem to emerge as units of perceptuomotor control in a child's speech. I should emphasize that details of the process vary widely from child to child, but the general outline is becoming clear (Menn, 1983a, b).

We can illustrate the process by tracing how a child escapes from consonant harmony, that is, how it comes to execute a word (or syllable) with two different places (or manners) or articulation. Children vary in their initial attack on such words: Some children omit, others harmonize one or other of the discrepant consonants. For example, faced with the word *fish*, which calls for a shift from a labiodental to a palatal constriction, one child may offer [fɪʼ], another [ɪʃ]; faced with *duck*, one child may try [gʌk], another [dʌt]. Menn (1983b) proposes a perspicuous account of such attempts: the child has '. . . learned an articulatory program of opening and closing her mouth that allows her to specify two things: the vowel and one point of oral closure' (p. 5). Reframing this in terms of gestures (an exercise that we need not repeat in later examples), we may say that the child has learned to coordinate glottal closing/opening and tongue positioning (back, front, up, down, in various degrees) with raising/lowering the jaw, in order to approximate an adult word. This description of a word as an articulatory program, or routine, composed of a few variable gestures, is a key to the child's phonological development.

Consider here a Spanish child, Si, studied by Macken (1979) from 1 year 7 months to 2 years 1 month of age. At a certain point, Si seemed only able to escape from consonant place harmony by producing a labial-vowel-dental-vowel disyllable, deleting any extra syllable in the adult model. Thus, *manzana* ('apple') became [manna]; *Fernando* became [mannə] or [wanno], with the initial [f] transduced as [m] or [w]; *pelota* ('ball') became [patda]. In some words, where the labial and dental were in the 'wrong' order, Si metathesized. Thus, *sopa* ('soup') became [pwæta], replacing [s] with [t], and *teléfono* became [fəntonno]. As Si's mastery increased, the class of words, subject to the labial–dental routine, narrowed: *manzana* became [tʃænna], *Fernando* became [tçɪnalto], and so on.

These examples make two important points: (1) the child brings adult words with similar patterns (e.g., *manzana*, *Fernando*, *pelota*) within the domain of a single articulatory routine, demonstrating use of the word as a unit; (2) at the same time, the child selects as models adult words that share certain gestural patterns, demonstrating an incipient grasp of their segmental structure.

We may view the developmental process as driven by the conflicting demands of articulatory 'ease' and lexical accumulation. As long as the child has only a few words, it needs only one or two articulatory routines. Initially, it exploits these routines by adding to its repertoire only words composed of gestural patterns similar to those it has already 'solved', and by avoiding words with markedly different patterns. (For evidence and discussion of avoidance and exploitation in early child phonology, see Menn, 1983a.) Once the initial routines have been consolidated, new routines begin to emerge under pressure from the child's accumulating vocabulary. New routines emerge either to handle a new class of adult words, not previously attempted, or to break up and redistribute the increasing cohort of words covered by an old articulatory routine.

Phonological development seems then to be a process of: (1) diversifying articulatory routines to encompass more and more different classes of adult model; (2) gradually narrowing the domain within a word to which an articulatory routine applies. The logical end of the process (usually reached during the third year of life, when the child has accumulated some 50–100 words) is a single articulatory routine for each phonetic segment. Development is far from complete at this point: there must ensue, at least, the systematic grouping of phonetic variants (allophones) into phoneme classes, and the discovery of language-specific regularities in their sequencing ('phonotactic rules'). But the emergence of the phonetic segment as a perceptuomotor unit brings the entire adult lexicon, insofar as it is cognitively available, within the child's phonetic reach. This signals the onset of the explosive vocabulary growth, at an average rate of some 5–7 words a day, that permits an average 6-year old American child to recognize an estimated 7000–11,000 root words, depending on family background (Templin, 1957; cf. Miller, 1977).

Conclusion

We began with a paradox: the apparent incommensurability of the quasi-continuous articulatory and acoustic structure of speech with the discrete units of its written representation. To resolve the paradox, we proposed that an alphabetic letter (or an element in a syllabary, or an ideograph) is a symbol for a discrete, perceptuomotor control structure. We then traced the emergence of such structures as encapsulated patterns of gesture in a child's speech. Implicit in their derivation is that a child, once possessed of them, is, at least in principle, ready for the metalinguistic task of learning to write and read (cf. Asbell, 1984).

What we have left unresolved is the relation between discrete motor control structures (word, syllable, phoneme) and the coordinated patterns of gesture that they generate. Perhaps we should regard the postulated structures as conceptual place-holders. Their functional analysis must await advances in neurophysiology and in the general theory of motor control.

Acknowledgement

My thanks to Björn Lindblom and Peter MacNeilage for critical comments, and discussion. The paper was written while the author was on sabbatical leave as a Fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford, California. The financial support of the City University of New York and of the Spencer Foundation is gratefully acknowledged.

Notes

[1] Speech errors display a number of well-known biases. For example, word-initial errors are more common than word-medial and word-final errors; metathesis occurs only between segments that occupy the same position in the syllabic frame; and the phonetic form of an error is adjusted to the context in which the erroneous segment occurs, not to the context from which it was drawn. Thus, speech errors often reflect phonetic processes that follow access of a phonemically specified lexical item. An adequate account of speech errors must therefore accommodate not only the phoneme as the fundamental phonological unit of all spoken languages, but the processes of lexical access and phonetic execution that give rise to biases in speech error types and frequencies. For a model of speech errors within these constraints, see Shattuck-Hufnagel (1983); for fuller discussion of the issue, see Lindblom, MacNeilage and Studdert-Kennedy (forthcoming).

[2] The first two items listed were immediate imitations of an adult utterance. Later items were identified by their '. . . consistency in reference or accompanying

action' (Ferguson and Farwell, 1975, p. 9). Interobserver agreement in the study from which these transcriptions were drawn was over 90%. The validity of the assumed target, *pen*, is further attested by the many featural (or gestural) properties common to the target and each of the child's attempts. The attempts did not include, for example, [gog], in which only glottal closure would be shared with the presumed target.

References

Alegria, J., and Noirot, E. (1982). Oriented mouthing activity in neonates: Early development of differences related to feeding experience. In J. Mehler, E. C. T. Walker and M. Garrett (Eds) *Perspectives on Mental Representation*. Hillsdale, NJ: Erlbaum, pp. 389–97.

Asbell, B. (1984). Writer's workshop at age 5. *New York Times Magazine*, February 26th.

Bates, E. (1979). *The Emergence of Symbols*. New York: Academic Press.

Bellugi, U., Poizner, H. and Klima, E. S. (1983). Brain organization for language: clues from sign aphasia. *Human Neurobiology*, 2, 155–70.

Best, C. T., Hoffman, H. and Glanville, B. R. (1982). Development of infant ear asymmetries for speech and music. *Perception and Psychophysics*, 31, 75–85.

Blumstein, S. E. (1981). Phonological aspects of aphasia. In M. T. Sarno (Ed.) *Acquired Aphasia*. New York: Academic Press.

Browman, C. P. (1980). Perceptual processing: slips of the ear. In V. A. Fromkin (Ed.) *Errors in Linguistic Performance*. New York: Academic Press.

Browman, C. and Goldstein, L. (1986). *Towards an articulatory phonology. Phonology Yearbook*, 3, 219–252.

Brown, R. (1958). *Words and Things*. Glencoe, IL: Free Press.

Carello, C., Turvey, M. T., Kugler, P. N. and Shaw, R. E. (1984). Inadequacies of the computer metaphor. In M. J. Gazzaniga (Ed.) *Handbook of Cognitive Neuroscience*. New York: Plenum.

Cowan, N., Leavitt, L. A., Massaro, D. W. and Kent, R. D. (1982). A fluent backward talker. *Journal of Speech and Hearing Research*, 25, 48–53.

DeCasper, A. J. and Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science*. 28, 1174–6.

Eimas, P. D. (1985). The perception of speech in early infancy. *Scientific American*, 252, 46–52.

Fentress, J. C. (1981). Order in ontogeny: relational dynamics. In K. Immelmann, G. W. Barlow, L. Petrinovich and M. Main (Ed.) *Behavioral Development*. New York: Cambridge University Press, pp. 338–71.

Fentress, J. C. (1983). Hierarchical motor control. In M. Studdert-Kennedy (Ed.) *Psychobiology of Language*. Cambridge, MA: MIT Press, pp. 40–61.

Ferguson, C. A. and Farwell, C. B. (1975). Words and sounds in early language acquisition: English initial consonants in the first fifty words. *Language*, 51, 419–30.

Ferguson, C. A. (1978). Learning to pronounce: The earliest stages of phonological development in the child. In F. D. Minifie and L. L. Lloyd (Eds) *Communicative and Cognitive Abilities — Early Behavioral Assessment*. Baltimore, MD: University Park Press, pp. 273–97.

Golani, I. (1981). The search for invariants in motor behavior In K. Immelmann, G. W. Barlow, L. Petrinovich and M. Main (Eds) *Behavioral Development*. New York: Cambridge University Press, pp. 372–90.

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*, Volume 4. New York: Academic Press, pp. 1–102.

Jusczyk, P. W. (1982). Auditory versus phonetic coding of speech signals during infancy. In J. Mehler, E. C. T. Walker, and M. Garrett (Eds) *Perspectives on Mental Representation*. Hillsdale, NJ: Erlbaum, pp. 361–87.

Kelso, J. A. S., Tuller, B. and Harris, K. S. (1983). A 'dynamic pattern' perspective on the control and coordination of movement. In P. F. MacNeilage (Ed.) *The Production of Speech*. New York: Springer, pp. 137–73.

Kelso, J. A. S., Saltzman, E. L. and Tuller, B. (1986). The dynamical perspective on speech production: Data and theory. *Journal of Phonetics*, **14**, 29–59.

Kinsbourne, M. (1972). Eye and head turning indicates cerebral lateralization. *Science*, **176**, 539–41.

Kroodsma, D. E. (1981). Ontogeny of bird song. In K. Immelmann, G. B. Barlow, L. Petrinovich and M. Main (Eds) *Behavioral Development*. New York: Cambridge University Press, pp. 518–32.

Kuhl, P. K. (1986). Infants' perception of speech: constraints on the characterizations of the initial state. In B. Lindblom and R. Zetterström (Eds) *Precursors of Early Speech*. Basingstoke, UK: MacMillan, pp. 219–44.

Kuhl, P. K. and Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, **218**, 1138–44.

Lempert, H. and Kinsbourne, M. (1982). Effect of laterality of orientation on verbal memory. *Neuropsychologia*, **20**, 211–14.

Levinson, S. E. and Liberman, M. Y. (1981). Speech recognition by computer. *Scientific American*, **244**, 64–86.

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

Lindblom, B., MacNeilage, P. F., and Studdert-Kennedy, M. (forthcoming). *Evolution of Spoken Language*. Orlando, FL: Academic Press.

Locke, J. (1983). *Phonological acquisition and change*. New York: Academic.

MacKain, K. S. (1982). Assessing the role of experience in infant speech discrimination. *Journal of Child Language*, **9**, 527–42.

MacKain, K. S. (1983). Speaking without a tongue. *Journal of the National Student Speech Language Hearing Association*, **11**, 46–71.

MacKain, K. S., Studdert-Kennedy, M., Spieker, S. and Stern, D. (1983). Infant intermodal speech perception is a left hemisphere function. *Science*, **219**, 1347–9.

MacKay, D. G. (1970). Spoonerisms: The structure of errors in the serial order of speech. *Neuropsychologia*, **8**, 323–50.

Macken, M. A. (1979). Developmental reorganization of phonology: A hierarchy of basic units of acquisition. *Lingua*, **49**, 11–49.

MacNeilage, P. F., Hutchinson, J. and Lasater, S. (1981). The production of speech: Development and dissolution of motoric and premotoric processes. In J. Long and A. Baddeley (Eds) *Attention and Performance IX*. Hillsdale, NJ: Erlbaum.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746–8.

Meltzoff, A. N. and Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, 198, 175–8.

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

Menn, L. (1978). *Pattern, control, and contrast in beginning speech: A case study in the development of word form and word function*. Bloomington, IN: Indiana University Linguistics Club.

Menn, L. (1983a). Development of articulatory, phonetic, and phonological capabilities. In B. Butterworth (Ed.) *Language Production*, Vol 11. London: Academic Press.

Menn, L. (1983b). Language acquisition, Aphasia, and Phonotactic Universals. Paper presented at 12th Annual University of Wisconsin — Milwaukee Linguistics Symposium.

Miller, G. A. (1977). *Spontaneous Apprentices*. New York: The Seabury Press.

Molfese, D. L. (1977). Infant cerebral asymmetry. In S. J. Segalowitz and F. A. Gruber (Eds) *Language Development and Neurological Theory*. New York: Academic Press.

Morais, J., Cary, L., Alegria, J. and Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition*, 7, 323–31.

Moskowitz, A. I. (1973). The acquisition of phonology and syntax. In K. K. J. Hintikka, J. M. E. Moravsik and P. Suppes (Eds) *Approaches to Natural Language*. Dordrecht, Netherlands: Reidel.

Neville, H. J. (1980). Event-related potentials in neuropsychological studies of language. *Brain and Language*, 11, 300–18.

Neville, H. J. (1985). Effects of early sensory and language experience on the development of the human brain. In J. Mehler and R. Fox (Eds) *Neonate Cognition*. Hillsdale, NJ: Erlbaum, pp. 349–63.

Oller, D. K. (1980). The emergence of the sounds of speech in infancy. In G. H. Yeni-Komshian, J. F. Kavanagh and C. A. Ferguson (Eds) *Child Phonology, Vol 1: Production*, New York: Academic Press, pp. 93–112.

Oller, D. K. (1986). Metaphonology and infant vocalizations. In B. Lindblom and R. Zetterström (Eds) *Precursors of Early Speech*. Basingstoke, UK: MacMillan, pp. 21–35.

Piaget, J. (1962). *Plays, Dreams and Imitation in Childhood*. New York: Norton.

Pisoni, D. B. (1985). Speech perception: Some new directions in research and theory. *Journal of the Acoustical Society of America*, 78, 381–8.

Saltzman, E. L. and Kelso, J. A. S. (1987). Skilled actions: a task dynamic approach. *Psychological Review*, 94, 84–105.

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

Summerfield, Q. (in press). Preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell (Eds) *Hearing by Eye*. Hillsdale, NJ: Erlbaum.

Shattuck-Hufnagel, S. (1983). Sublexical units and suprasegmental structure in speech production planning. In P. F. MacNeilage (Ed.) *The Production of Speech*. New York: Springer, pp. 109–36.

Studdert-Kennedy, M. (1986). Sources of variability in early speech development. In J. S. Perkell and D. H. Klatt (Eds) *Invariance and Variability in Speech Processes*. Hillsdale, NJ: Erlbaum, pp. 58–76.

Templin, M. (1957). *Certain Language Skills of Children*. Minneapolis: University of Minnesota Press.

Turney, M. T. and Kugler, P. N. (1984). A comment on equating information with symbol strings. *American Journal of Physiology*, **246** (*Regulatory, Integrative, Comparative Physiology*, **15**) R925–7.

Warren, R. M. (1976). Auditory illusions and phonetic processes. In J. J. Lass (Ed.) *Contemporary Issues in Experimental Phonetics*. New York: Academic Press.

Welty, E. (1983). *One Writer's Beginnings*. New York: Warner Books.

Werker, J. F. (1982). *The development of cross-language speech perception: The effect of age, experience and context on perceptual organization*. Unpublished doctoral dissertation, University of British Columbia, Vancouver, BC.

Werker, J. and Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, **7**, 49–63.