/318

In Schiller, N.O. & Meyer, A.S. (eds.), Phonetics and Phonology in Language Comprehension and Production, pp. 159-207. Mouton de Gruyter. 2003

Articulatory Phonology: A phonology for public language use

Louis Goldstein and Carol A. Fowler[1]

1. Introduction

The goals of the theoretical work that we describe here are twofold. We intend first to develop a realistic understanding of language forms as language users know them, produce them and perceive them. Second we aim to understand how the forms might have emerged in the evolutionary history of humans and how they arise developmentally, as a child interacts with speakers in the environment.

A seminal idea is that language forms (that is, those entities of various grain sizes that theories of phonology characterize) are the means that languages provide to make between-person linguistic communication possible, and, as such, they are kinds of public action, not the exclusively mental categories of most theories of phonology, production and perception. A theory of phonology, then, should be a theory about the properties of those public actions, a theory of speech production should be about how those actions are achieved, and a theory of speech perception should be about how the actions are perceived.

A theory of the emergence of phonological structure in language, from this perspective, is about how particulate language forms emerged in the course of communicative exchanges between people. Therefore, it predicts that the forms will have properties that adapt them for public language use: for speaking and for being perceived largely from acoustic speech signals.

Articulatory Phonology provides the foundation on which we build these theoretical ideas.

2. Articulatory Phonology

2.1. Phonology as a combinatoric system

The most fundamental property of speech communication is its phonological structure: it allows a small (usually <100) inventory of primitive units to combine in different ways to form the vast array of words that constitute the vocabularies of human languages. It shares this combinatoric property with just a few other natural systems, such as chemical compounding and genetic recombination. The theoretical underpinnings of this class of natural systems have recently come under scrutiny (Abler 1989; Fontana and Buss 1996). In all such *self-diversifying* systems (Abler 1989; Studdert-Kennedy 1998), the atomic units are discretely distinct from one another, and they retain their discreteness when they combine to form new *objects*. This appears to be a necessary property of such systems. If, instead, the combination operation were to involve blending of units defined as points along some scale, the diversity of the combinations would tend to decrease as more and more atoms join – all combinations would tend toward the mean value of the scalar units. Other properties of these natural systems such as *recurring substructures* and *hierarchy* also have been shown to depend on the fact that combination involves creation of new objects in which atoms retain their discrete identities (Fontana and Buss 1996).

The combinatoric structure of speech appears to be at odds with measurements of speech that we can make in the laboratory. Early attempts to find discrete, re-combinable units in the acoustic record (Cooper et al. 1952; Harris 1953) yielded surprising failure, and such failures have been replicated ever since, and extended to the articulatory, electromyographic, aerodynamic, and auditory domains (but cf. Stevens' [1989, 1999] quantal theory which attempts to isolate some invariant acoustic properties). Continuous, context-dependent motion of a large number of degrees of freedom is what we find in physical records. As a response to this failure, phonological units (segments) have been removed from the domain of publicly observable phenomena, and have been hypothesized to be

fundamentally mental units that are destroyed or distorted in the act of production, only to be reconstructed in the mind of the perceiver (e.g., Hockett 1955; Ohala 1981).

Articulatory Phonology (Browman and Goldstein 1992a, 1995a) has proposed, following Fowler (1980), that the failure to find phonological units in the public record was due to looking at too shallow a description of the act of speech production, and that it is, in fact, possible to decompose *vocal tract action* during speech production into discrete, re-combinable units. The central idea is that while the articulatory and acoustic *products* of speech production actions are continuous and context-dependent, the actions themselves that engage the vocal tract and regulate the motions of its articulators are discrete and context-independent. In other words, phonological units are abstract with respect to the articulatory and acoustic variables that are typically measured, but not so abstract as to leave the realm of the vocal tract and recede into the mind. They are abstract in being coarse-grained (low dimensional) with respect to the specific motions of the articulators and to the acoustic structure that may specify the motions.

2.2. *Units of combination (atoms) are constriction actions of vocal organs*

Articulatory Phonology makes three key hypotheses about the nature of phonological units that allow these units to serve their dual roles as units of action and units of combination (and contrast). These are: that vocal tract activity can be analyzed into constriction actions of distinct vocal organs, that actions are organized into temporally overlapping structures, and that constriction formation is appropriately modeled by dynamical systems.

2.2.1. Constriction actions and the organs that produce them

It is possible to decompose the behavior of the vocal tract during speech into the formation and release of constrictions by six distinct

constricting devices or *organs*: lips, tongue tip, tongue body, tongue root, velum, and larynx. Although these constricting organs share mechanical degrees of freedom (articulators and muscles) with one another (for example, the jaw is part of the lips, tongue tip, and tongue body devices), they are intrinsically distinct and independent. They are distinct in the sense that the parts of the vocal anatomy that approach one another to form the constrictions are different, and they are independent in the sense that a constriction can be formed by one of these devices without necessarily producing a constriction in one of the others (a point made by Halle 1983, who referred to organs as *articulators*). Thus, constricting actions of distinct organs, actions known as *gestures*, can be taken as atoms of a combinatoric system – they satisfy the property of discrete differences (Browman and Goldstein 1989; Studdert-Kennedy 1998). Two combinations of gestures can be defined as potentially *contrasting* with one another if, for example, they include at least one distinct constriction gesture. The words *pack* and *tack* contrast with one another in that the former includes a lips gesture and the latter a tongue tip gesture.

The words *hill*, *sill*, and *pill* contrast with one another in the combination of constricting organs engaged at the onset of the word. *Hill* begins with a gesture of the larynx (glottal abduction); *sill* adds a gesture of the tongue tip organ to the laryngeal one and *pill* adds a lip gesture. These three gestures (larynx, tongue tip, and lips) all combine to create the contrasting molecule *spill*. The analysis of the onset of *spill* as composed of three gestures differs from an analysis of this form as composed of a sequence of two feature bundles (/s/ and /p/), in which each of those bundles has some specification for the larynx and a supralaryngeal gesture. Evidence in favor of the three-gesture specification has been discussed in Browman and Goldstein (1986).

Of course, not all phonological contrasts involve gestures of distinct organs. For example, *pin* and *fin* differ in the nature of the lip gestures at the beginning of the words. The discrete differentiation of the gestures involved in such contrasts critically depends on the public properties of a phonological system, and is discussed in the last section of this paper. However, it can also be argued that

between-organ contrasts are the primary ones within phonological systems. For example, while all languages contrast constriction gestures of the lips, tongue tip, and tongue body, within-organ contrasts (such as [p-f], or [t-T]) are not universal.

2.2.2. Coordination of gestures and overlap

While traditional theories of phonology hypothesize that the primitive units combine by forming linear sequences, Articulatory Phonology hypothesizes that gestures are coordinated into more elaborated "molecular" structures in which gestures can overlap in time. Such *coproduction* of gestures can account for much of the superficial context-dependence that is observed in speech. The reason for this can be found in the nature of the distinct constricting organs, which share articulators and muscles. When two gestures overlap, the activities of the individual mechanical degrees of freedom will depend on both (competing) gestures. For example, consider the coproduction of a tongue tip constriction gesture with the tongue body gesture for different vowels (as in /di/ and /du/). The same (context-independent) tongue tip gesture is hypothesized to be produced in both cases, but the contribution of the various articulatory degrees of freedom (tongue tip, tongue body) will differ, due to the differing demands of the vowel gestures. The theory of task dynamics (Saltzman 1986, 1995, as discussed below) provides a formal model of such context-dependent variability.

Gestures are hypothesized to combine into larger molecules by means of coordinating (bonding) individual gestures to one another. This coordination can be accomplished by specifying a phase relation between the pair of coupled dynamical systems that control the production of the gestures within a task dynamic model (Browman and Goldstein 1990; Saltzman and Byrd 2000). Figure 1 shows how the gestures composing the utterance "team," are arranged in time, using a display called a *gestural score*. The arrows connect individual gestures that are coordinated with respect to one another. Recent work within Articulatory Phonology (Browman and

Goldstein 2000; Byrd 1996) has shown that the pairs of coordinated of gestures vary in how tightly they are bonded. *Bonding strength* is represented in the figure by the thickness of the arrows. Note that while there is no explicit decomposition of this molecule into traditional segmental units, the gestures that constitute segments (tongue tip and larynx gesture for /t/, lip and velum gesture for /m/) are connected by the thicker lines. High bonding strength is also found for gestures that compose a syllable onset, even when they do not constitute a single segment (e.g., tongue tip and lip gestures in the onset cluster /sp/), and it is not known at present whether their bonding differs from the intrasegmental bonds. In principle, bonding strength can be used to used to define a hierarchy of unit types, including segments, onset and rimes, syllables, feet, and words. The more tightly bonded units are those that we would expect to cohere in speech production and planning, and therefore, in errors.

Bonding strength can also be seen to correlate with combinatoric structure of gestures and may, in fact, emerge from that combinatoric structure developmentally. Syllable onsets are tightly bonded, and the number and nature of consonant combinations that can occur as syllable onsets (in English and in languages generally) is highly restricted (usually in accordance with the sonority hierarchy). The coordination of the onset gestures with the vowel is much looser, and there are virtually no constraints on combinations of onsets with vowels in languages – it is here that systems typically exhibit free substitutability. Thus, each onset recurs frequently with different vowels, a distribution that could allow the coordination of the onsets to be highly stabilized (through frequency of experience), but would not afford such stabilization for particular onset-vowel combinations

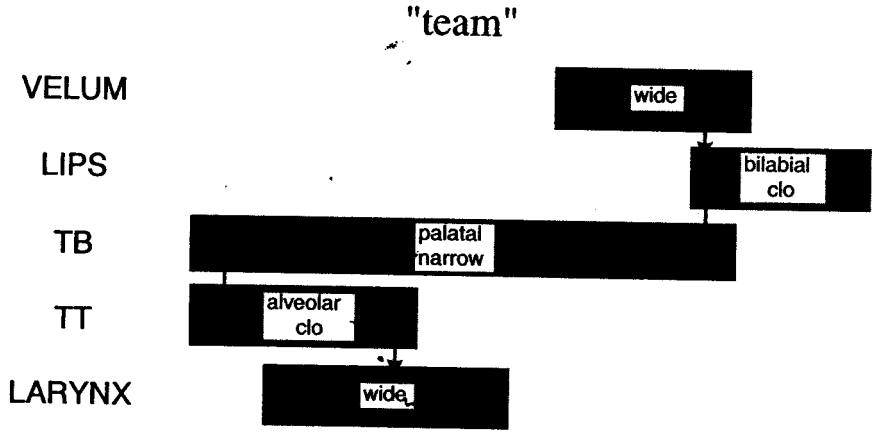

Figure 1. Gestural score for the utterance "team" as generated automatically by the model of Browman and Goldstein (1992a). The rows correspond to distinct organs (TB = "Tongue Body", TT = "Tongue Tip"). The labels in the boxes stand for gesture's goal specification for that organ. For example, "alveolar" stands for a *tongue tip constriction location* 56 degrees up from the horizontal and "clo" stands for a *tongue tip constriction degree* of –3.5mm (the tongue tip compresses against the palate). The arrows connect gestures that are critically coordinated, or phased, with respect to one another. The thicker arrows represent tighter *bonding strengths* between coordinated gestures.

not afford such stabilization for particular onset-vowel combinations (since they occur relatively infrequently). Thus, onsets (and in the limiting case, segments) can be considered *ions* of a combinatoric system: internally cohesive structures of atoms that recombine readily with other such structures.

2.2.3. Dynamical specification

One aspect of the physical description of speech seems particularly at odds with a description of discrete units: the fact that articulators move continuously over time. This continuous motion is modeled in most theories of phonology or speech production by assuming that

there are "targets" for phonological units, and that speech production involves "interpolation" between the targets (e.g., Keating 1990; Perrier, Loevenbruck and Payan 1996). In this view, the phonological information (the target) is not part of the production itself – it is hidden within the mind of the producer, and is only communicated though the resulting interpolation.

In Articulatory Phonology, gestural units are modeled as dynamical systems. A dynamical system is characterized by an equation (or set of equations) that expresses how the state of a system changes over time. Crucially, production of a gestural constriction can be modeled as a (mass-spring) dynamical system with a *fixed* set of parameter values (Browman and Goldstein 1995a; Saltzman 1995). During the time that a gesture is active, its equation is fixed and time-invariant, even though the articulators are moving continuously. Moreover, the *way* in which the articulators move over time is specific to the particular dynamical system involved, and therefore, the gestural unit is specified directly in the motion itself.

2.3. Evidence for gestures as units of speech production

How can we test the hypothesis that the act of speech production can be decomposed into gestures – actions that form local constrictions? There is little direct evidence in the products of speech production (acoustics or articulator movements) for discrete units, as has been long observed. One kind of indirect evidence involves analysis-by-synthesis. It is possible to examine articulator motions and to infer from them a plausible gestural structure (such as that in Figure 1). The resulting structure can then be used as input to a gestural production model in order to test whether motions matching those of the original data are generated. This strategy has, in limited contexts at least, been used successfully to support particular hypotheses about gestural decomposition (Browman 1994; Browman and Goldstein 1992b; Saltzman and Munhall 1989).

Recently, some more direct evidence for decomposing speech production and/or planning into gestural structures has been obtained

from speech production errors. Errors have long been used as evidence for linguistic units of various types (e.g., Fromkin 1973). The underlying assumption is that errors can result from systematic "misplacement" of linguistic units within a larger structure that is active during speech planning. For example, errors have been used to support the reality of abstract segments (internal, non-public units) as units of phonological encoding in speech production (e.g., Shattuck-Hufnagel 1983; Shattuck-Hufnagel and Klatt 1979). Transcriptions of errors suggest that segments are the most common units involved in errors (exhibiting changes in position, such as anticipations, perseverations, and exchanges). Evidence that the segments involved in errors are abstract (and not phonetic) is found in the fact that segments appear to be phonetically accommodated to the new contexts created for them by the errors. For example in the error *slumber party* → *lumber sparty* (Fromkin 1973), the /p/ is aspirated in *party* but unaspirated in *sparty*. In addition, errors involving segments have also been claimed to be (almost always) phonotactically well-formed strings of the language. These facts have been modeled (Shattuck-Hufnagel 1983) as arising during phonological planning by the insertion of a segmental unit into an incorrect *slot* in the phonological *frame*. Though incorrect, the slot is of the appropriate type for the segment in question, which accounts for the observed phonotactic well-formedness. Thus, utterances with errors are thought to be generally well-formed, both phonetically and phonotactically.

That this view of speech errors is complete was called into question by Mowrey and McKay (1990) who collected EMG data during the production of tongue twisters and found anomalous patterns of muscle activity that were not consistent with the notion that utterances with errors are phonetically and phonotactically well formed. Experiments in our laboratory have generated kinematic data that confirm and extend their results (Goldstein et al. submitted; Pouplier and Goldstein submitted;). These experiments involve repetition of simple two-word phrases such as *cop top*, in which the syllables are identical except for their initial consonants. Speakers who produce these phrases consistently produce errors of *gestural*

intrusion, in which a "copy" of the oral constriction gesture associated with one of the initial consonants is produced during the other initial consonant. For example, in *cop top* a copy of the tongue dorsum gesture for the initial /k/ in *cop* can be observed during the /t/ in top (and conversely, a copy of the tongue tip gesture for /t/ can be observed during the /k/). These intrusions are usually reduced in magnitude compared to an intended gesture (and therefore such errors are often not perceived), but they can become as large as intended gestures, in which case a speech error is perceived.

These gestural intrusion errors cannot result from moving an abstract segment to the wrong position within a phonotactically well-formed frame. First, an intruded gesture is typically partial in magnitude. If an abstract segment were to move to an incorrect position, there is no reason why it should result in a reduced gesture. Second, intruded gestures are produced concurrently with the gesture(s) of the intended consonant, which may not exhibit reduction (For example, when the tongue dorsum gesture of /k/ intrudes on the /t/ of *top*, the tongue tip gesture for /t/ is still produced). Every speaker exhibits more errors of gestural intrusion than reduction. Thus, the speaker appears to be producing the gestures of the two words at the same time, which is not a phonotactically well-formed structure. In contrast, gestural intrusion errors can be readily accounted for if the structures involved in speech production and/or planning are gestural. The intrusion errors can be viewed as spontaneous transitions to a more intrinsically stable mode of coordination in which all gestures are produced in a 1:1 frequency-locking mode (which is known to be the most stable coordinative state in motor activities generally, e.g., Haken et al. 1996). For example, in the correct production of *cop top*, the tongue dorsum gesture for /k/ is coordinated in a 1:2 pattern with the gestures for the rime /ɒp/ (*op*), as is the tongue tip gesture for /t/. In productions with gestural intrusion, the relation with the rime gestures becomes 1:1, and, if both initial consonants exhibit intrusion errors (which does occur), then all gestures are produced in a 1:1 pattern.

It is important to note that gestural structures remain discrete in space and time in gestural intrusion errors, even when they are partial. Spatially, the articulatory record shows the simultaneous production of two distinct gestures, not a blended intermediate articulation. Temporally, the erroneous gesture is synchronized with the other initial consonant gesture, and does not slide continuously between its "home" and its intrusion site.

In some cases, a reduction of an intended gesture is observed when a gestural intrusion occurs. Such errors can be modeled as resulting from the competition between the intrinsically stable 1:1 mode and the learned gestural coordination patterns (molecules) of the language. These learned coordination patterns (for English) do not allow for both gestures (e.g., tongue tip for /t/ and tongue dorsum for /k/) to occur together.

How does this approach to speech errors account for cases like *slumber party* → *lumber sparty* that have been used to argue for the role of abstract mental units in speech production? The oral constriction gesture for /s/ (and possibly its laryngeal gesture) could be assumed to intrude at the beginning of the word *party*, and to reduce at the beginning of the word *slumber*. Alternatively, the gestural ion for /s/ may be mis-coordinated or mis-selected in the production of the phrase (the possible existence of such errors has not been ruled out). Under either interpretation, the resulting pattern of coordination of the oral and laryngeal gestures for /s/ and /p/ could combine to produce an organization that would be identified as an unaspirated [p], though articulatory data would be required to determine exactly how this comes about (see Browman and Goldstein 1986; Munhall and Löfqvist 1992; Saltzman and Munhall 1989).

2.4. Phonological knowledge as (abstract) constraints on gestural coordination

The goal of a phonological grammar is to account for native speakers' (implicit) knowledge of phonological structure and regularities in a particular language, including an inventory of lexically contrastive units, constraints on phonological forms, and systematic alternations to lexical forms that result from morphological combination and embedding in a particular prosodic context. If phonological forms are structures of coordinated gestures, as hypothesized in Articulatory Phonology, then gestural analysis should reveal generalizations (part of speakers' knowledge) that are obscured when phonological form is analyzed in some other way (in terms of features, for example). The property that most saliently differentiates gestural analyses from featural analyses is that gestural primitives are intrinsically temporal and thus can be explicitly coordinated in time. We therefore expect to find phonological generalizations that refer to patterns or modes of coordination, abstracting away from the particular actions that are being coordinated. Several examples of abstract coordination constraints have, in fact, been uncovered.

The coordination relation specified between two closure gestures will determine the resulting aerodynamic and acoustic consequences – whether there is an audible release of trapped pressure between the closures or not. So in principle one could characterize a sequence of closures either in terms of their abstract coordination pattern or their superficial release characteristics. Which characterization do speakers employ in their phonological knowledge? Gafos (2002) has shown that a crucial test case can be found when the same abstract coordination pattern can give rise to different consequences depending on whether the closures employ the same or distinct organs. In such cases, he finds languages in which the generalization must be stated in terms of coordination pattern not in terms of the superficial release properties. In Sierra Popoluca (Clements 1985; Elson 1947), releases are found between sequences of heterorganic consonants, but lack of release is found for homorganic sequences. A

generalization can be stated in terms of an abstract coordination pattern (the onset of movement for the second closure beginning just before the release of the first closure), but not in terms of release. Note that an abstract decomposition of the articulatory record into gestures is required here. In the case of a homorganic sequence, it is not possible to observe the onset of movement of the second closure in the articulatory record – a single closure is observed. But the temporal properties of this closure fall out of a gestural analysis, under the assumption that there are two gestural actions in the homorganic case, coordinated just as they are in the heterorganic case (where they can be observed). In Sierra Popoluca, the generalization is a relatively superficial phonetic one (characterizing the form of the consonant sequence) that does not have consequences for the deeper (morpho)-phonology of the language. But Gafos then shows that in Moroccan Arabic, a similar coordination constraint interacts with other constraints (such as a gestural version of the obligatory contour principle) in determining the optimal set of (stem) consonants to fill a morphological template. In this case, constraints on the coordination of gestures contribute to an account of morphophonological alternations.

Another example of abstract coordination modes in phonology involves syllable structure. Browman and Goldstein (1995b) have proposed that there are distinct modes of gestural coordination for consonant gestures in an onset versus those in a coda, and that these modes are the public manifestations of syllable structure. In an onset, a synchronous mode of coordination dominates – consonant gestures tend to be synchronous with one another (to the extent allowed by an overall constraint that the gestures must be recoverable by listeners), while in a coda, a sequential mode dominates. Synchronous production is most compatible with recoverability when a narrow constriction is coproduced with a wider one (Mattingly 1981), and therefore is most clearly satisfied in the gestures constituting single segments, which typically involve combinations of a narrow constriction of the lips, tongue tip, or tongue body with a wider laryngeal or velic gesture (e.g., voiceless or nasal stops) or a wider supralaryngeal gesture (e.g., "secondary" palatalization, velarization,

or rounding). Indeed the compatibility of synchronous production with recoverability may be what leads to the emergence of such segmental ions in phonology. But there is evidence that even multi-segment gestural structures (e.g., /sp/) exhibit some consequences of a tendency to synchronize onsets (Browman and Goldstein 2000).

Browman and Goldstein show that there are several featurally distinct types of syllable position-induced allophony that can be modeled as lawful consequences of these different coordination styles of onsets and codas. For example, vowels in English are nasalized before coda nasal consonants (but not before onset nasals). Krakow (1993) has shown that this can be explained by the fact that the oral constriction and velum-lowering gestures are synchronous in onsets, but sequential in codas, with the velum gesture preceding. The lowered velum superimposed on an open vocal tract is what is identified as a nasalized vowel. A very similar pattern of intergestural coordination can account for the syllable position differences in [l] – "lighter" in onsets and represented featurally as [-back], "darker" in codas and represented as [+back] (Browman and Goldstein 1995b; Krakow 1999; Sproat and Fujimura 1993). In this case the two coordinated gestures are the tongue tip closure and tongue dorsum retraction. Thus, two processes that are quite distinct in terms of features are lawful consequences of a single generalization about gestural coordination – the distinct styles of coordination that characterize onsets and codas. Compatible differences in coordination have been found for [w] (Gick in press) and for [r] (Gick and Goldstein 2002).

3. Parity in public language use

Articulatory Phonology provides a foundation on which compatible theories of speech production and perception may be built. Because these compatible theories are about between-person communication, a core idea that they share is that there must be a common phonological currency among knowers of language forms, producers

of the forms and perceivers of them. That is, the language forms that language users know, produce and perceive must be the same.[2]

3.1. The need for a common currency in perceptually guided action, including speech

The common currency theme emerges from research and theories spanning domains that, in nature, are closely interleaved, but that may be traditionally studied independently. In speech, the need for a common currency so that transmitted and received messages may be the same is known as the parity requirement (e.g., Liberman and Whalen 2000).

The various domains in which the need for a common currency has been noted are those involving perception and action. An example is the study of imitation by infants. Meltzoff and Moore (1977, 1983, 1997, 1999) have found that newborns (the youngest 42 minutes after birth) are disposed to imitate the facial gestures of an adult. In the presence of an adult protruding his tongue, infants attempt tongue protrusion gestures; in the presence of an adult opening his mouth, infants attempt a mouth opening gesture.

It is instructive to ask how they can know what to do. As Meltzoff and Moore (e.g., 1997, 1999) point out, infants can see the adult's facial gesture, say, tongue protrusion, but they cannot see their own tongue or what it is doing. They can feel their tongue proprioceptively, but they cannot feel the adult's tongue. Meltzoff and Moore (1997, 1999) suggest that infants employ a "supramodal representation," that is, a representation that transcends any particular sensory system. To enable the infant to identify his or her facial or vocal tract "organs" with those of the model, the representation has to reflect what is common between the visually perceived action of the adult model and the proprioceptively perceived tongue and its action by the infant. The commonality is captured if the supramodal representation is of "distal" objects and events (tongues, protrusion gestures), not of the perceptual-system-specific proximal properties (reflected light patterns, proprioceptive feels) of the stimulation.

We will use the term "common currency" in place of "supramodal representation" to refer, not to properties of a representation necessarily, but merely to what is shared in information acquired cross-modally, and as we will see next, to what is shared in perceptual and action domains. To be shared, the information acquired cross-modally and shared between perception and action plans has to be distal; that is, it has to be about the perceived and acted-upon world. It cannot be about the proximal stimulation.

Imitation is perceptually guided action, and Meltzoff and Moore might have noted that a similar idea of a common currency is needed to explain how what infants perceive can have an impact on what they do. Hommel et al. (2001) remark that if perceptual information were coded in some sensory way, and action plans were coded in some motoric way, perceptual information could not guide action, because there is no common currency. They propose a "common coding" in which features of the perceived and acted upon world are coded both perceptually and in action plans. Here, as for Meltzoff and Moore, common currency is achieved by coding distal, not proximal properties in perception, and, in Hommel et al.'s account, by coding distal action goals, not, say, commands to muscles, in action plans.

In speech, two sets of ideas lead to a conclusion that speech production and perception require a common currency.

First, infants imitate speech (vowels) beginning as young as 12 weeks of age (Kuhl and Meltzoff 1996), and they integrate speech information cross-modally. That is, they look longer at a film of a person mouthing the vowel they are hearing than at a film of a model mouthing some other vowel (Kuhl and Meltzoff 1982). Both findings raise issues that suggest the need for a common currency, first between perceived infant and perceived adult speech, and second between speech perceived by eye and by ear.

Although imitation of vowels may involve information obtained intra- rather than cross-modally (as in tongue protrusion), how infants recognize a successful imitation remains a challenging puzzle. The infant has a tiny vocal tract, whereas the adult model has a large one. Infants cannot rely on acoustic similarity between their

vowels and the model's to verify that they are imitating successfully. They need some way to compare their vowels with the adult's. Likewise, they need a way to compare visible with acoustically specified vowels. How can the infant know what facial gesture goes with what acoustic signal? The problem is the same one confronted by infants attempting to imitate facial gestures.

A solution analogous to that proposed by Meltzoff and Moore (1997, 1999) is to propose that listeners extract distal properties and events from information that they acquire optically and acoustically.[3] In this case, the distal events are gestural. Infants can identify their vowel productions with those of the adult model, because they perceive actions of the vocal tract (for example, tongue raising and backing and lip protrusion). In the gestural domain, these actions are the same whether they are achieved by a small or a large vocal tract. (This is not to deny that a nonlinear warping may be required to map an infant's vocal tract onto an adult's. It is to say that infants have lips, alveolar ridges on their palates, soft palates, velums, etc. They can detect the correspondence between the organs of their vocal tract and regions of their vocal tract with those of an adult. They can detect the correspondence of their actions using their organs in those regions with the actions of an adult.) The actions are also the same whether they are perceived optically or acoustically.

A second route to a conclusion that speaking and listening require a common currency is different from the first, but not unrelated to it. It concerns the fact that language serves a between-person communicative function. Liberman and colleagues (e.g., Liberman and Whalen 2000) use the term "parity" to refer to three requirements of language if it is to serve its communicative function. The first two relate to between-person communication. (The third is that, within a language user, specializations for speaking and listening must have co-evolved, because neither specialization would be useful without the other.) The first is that what counts for the speaker as a language form also has to count for the listener. As Liberman and Whalen put it, /ba/ counts, a sniff does not. If speakers and listeners did not share recognition of possible language forms, listeners would not know which of the noises produced by a speaker

should be analyzed for its linguistic content. The second requirement is more local. It is that, for language to serve as a communication system, characteristically, listeners have to perceive accurately the language forms that talkers produce. There has to be "parity" or sufficient equivalence between phonological messages sent and received.

Articulatory Phonology provides a hypothesis about a common currency for speaking and listening. Like the other common currencies that we have considered, this one is distal.

3.2. Articulatory Phonology provides the common currency for speech

To communicate, language users have to engage in public, perceivable activity that counts as doing something linguistic for members of the language community. Listeners have to perceive that activity as linguistic and to perceive it accurately for communication to have a chance of taking place.

Language forms are the means that languages provide for making linguistic messages public. If language is well adapted to its public communicative function, then we should expect the forms to be (or to be isomorphic with) the public actions that count as talking. In particular, we should expect the forms to be such that they can be made immediately available to a listener – available, that is, without mediation by something other than the language forms. This is their nature in Articulatory Phonology.

However, in most phonologies, as we noted earlier (section 2.1), this is not their nature. In phonologies other than Articulatory Phonology, atomic language forms are mental categories. Moreover, in accounts of speech production and perception, the activities of the vocal tract that count as speaking are not isomorphic with those language forms, due in part, to the ostensibly destructive or distorting effects of coarticulation. This means that elements of phonological competence are only hinted at by vocal tract actions. Further, because vocal tract actions cause the acoustic signals that constitute the

listener's major source of information about what was said, the acoustic signals likewise do not provide information that directly specifies the elements of phonological competence. Perception has to be reconstructive. In this system, talkers' phonological messages remain private. If communication succeeds, listeners represent the speaker's message in their head. However, transmission of the message is quite indirect. Language forms go out of existence as the message is transformed into public action and come back into existence only in the mind of the listener. This is not a parity-fostering system.

In contrast, Articulatory Phonology coupled with compatible theories of speech production and perception represents a parity-fostering system. Elements of the phonological system are the public actions of the vocal tract that count as speaking. What language users may have in their heads is knowledge about those phonological elements, not the elements themselves. If phonological atoms are public actions, then they directly cause the structure in acoustic speech signals, which, then, provides information directly about the phonological atoms. In this theoretical approach, language forms are preserved throughout a successful communicative exchange; they are not lost in the vocal tract and reconstituted by the perceiver.

Note that this hypothesis is not incompatible with the possibility that there are certain types of phonological processes that depend on the coarser topology of the gestural structures involved (e.g., syllable count, syllabification, foot structure and other prosodic domains) and not on the detailed specification of the actions involved. The hypothesis would claim that these coarser properties are ultimately derivable from the more global organization of public actions and do not represent purely mental categories that exist independently of any actions at all.

3.3. An integrative theory of phonological knowing, acting and perceiving

The gestures of Articulatory Phonology are dynamical systems that, as phonological entities, are units of contrast; as physical entities, they are systems that create and release constrictions in the vocal tract. A compatible theory of speech production is one that spells out how those systems generate constrictions and releases and how the gestures that form larger language forms are sequenced. To maintain the claim that atomic language forms are preserved from language planning to language perception, the account has to explain how coarticulated and coarticulating gestures nonetheless maintain their essential properties.

One such account is provided by task dynamics theory.

3.3.1. Task dynamics

Speech production is coordinated action. Like other coordinated actions, it has some properties for which any theory of speech production needs to provide an account. One property is equifinality. This occurs in systems in which actions are goal-directed and in which the goal is abstract in relation to the movements that achieve it. In Articulatory Phonology, the gesture for /b/ is lip closure. A gesture counts as a lip closure whether it is achieved by a lot of jaw closing and a little lip closing or by jaw opening accompanied by enough lip closing to get the lips closed.

Talkers exhibit equifinality in experiments in which an articulator is perturbed on line (e.g., Gracco and Abbs 1982; Kelso et al. 1984; Shaiman 1989). They also exhibit it in their coarticulatory behavior. In the perturbation research of Kelso et al., for example, a speaker repeatedly produced /baeb/ or /baez/ in a carrier sentence. On a low proportion of trials, unpredictably, the jaw was mechanically pulled down during closing for the final consonant in the test syllable. If the consonant was /b/, extra downward movement of the upper lip achieved lip closure on perturbed as compared to unperturbed trials.

If the consonant was /z/, extra activity of a muscle of the tongue allowed the tongue to compensate for the low position of the jaw. Equifinality suggests that gestures are achieved by systems of articulators that are coordinated to achieve the gestures' goals. These systems are known as synergies or coordinative structures. The synergy that achieves bilabial closure includes the jaw and the two lips, appropriately coordinated.

It is the equifinality property of the synergies that achieve phonetic gestures that prevents coarticulation from destroying or distorting essential properties of gestures. An open vowel coarticulating with /b/ may pull the jaw down more-or-less as the mechanical jaw puller did in the research of Kelso et al. However, lip closure, the essential property of the labial stop gesture is nonetheless achieved.[4]

The task dynamics model (e.g., Saltzman 1991, 1995; Saltzman and Kelso 1987; Saltzman and Munhall 1989) exhibits equifinality and therefore the nondestructive consequences of coarticulation needed to ensure that gestures are achieved in public language performance.

In the task dynamics model, the synergies that achieve gestures are modeled as dynamical systems. The systems are characterized by attractors to the goal state of a phonetic gesture. These goal states are specified in terms of "tract variables." Because, in Articulatory Phonology, gestures create and release constrictions, tract variables define the constriction space: they are specific to the organs of the vocal tract that can achieve constrictions, and they specify a particular constriction degree and location for that constricting organ. For example, /d/ is produced with a constriction of the tongue tip organ; the relevant tract variables are "TTCL" and "TTCD" (tongue tip constriction location and constriction degree). They are parameterized for /d/ so that the constriction location is at the alveolar ridge of the palate and constriction degree is a complete closure. In task dynamics, gestural dynamics are those of a damped mass-spring, and the dynamics are such that gesture-specific tract variable values emerge as point attractors. The dynamical systems exhibit the equifinality property required to understand how

coarticulation can perturb production of a gesture without preventing achievement of its essential properties.

In one version of task dynamics, gestures are sequenced by the gestural scores generated by Articulatory Phonology from a specification of coordination (phase) relation among the gestures. Each gesture is associated with an activation wave that determines the temporal interval over which the gesture exerts an influence on the dynamics of the vocal tract. Due to coarticulation, activation waves overlap.

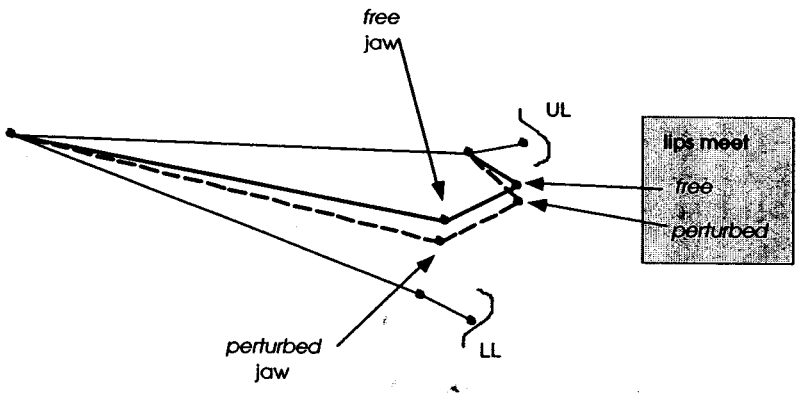

Figure 2. Simulation of lip and jaw behavior under task dynamic control. The dashed line shows the position of the jaw and lips when the jaw is *perturbed* (by being mechanically restrained in a low position), and the solid line shows the positions when the jaw is *free* to move. The lips close in both cases, but increased movement of both upper and lower lips contributes to the closure goal in the *perturbed* condition (after Saltzman 1986).

Figure 2 shows a model vocal tract (lips and jaw only) under task dynamic control achieving lip closure under perturbed and unperturbed (free) conditions. When the jaw is prevented from closing (the dashed line in Figure 2), the lips compensate by moving farther than on unperturbed productions (the solid line). Figure 3 shows equifinal production of /d/ during coarticulation with different vowels. The top of the figure shows the shape of the front part of the

tongue during the /d/ closure in the utterances /idi/, /ada/, /udu/ (as measured by x-ray microbeam data). The bottom of the figure shows the front part of the tongue during closure for /idi/, /ada/, /udu/ as generated by the computational gestural model (Browman and Goldstein 1992a), incorporating the task-dynamics model. The constriction of the tip of the tongue against the palate is the same (in location and degree) across vowel contexts, even though the overall shape of the tongue varies considerably as a function of vowel.

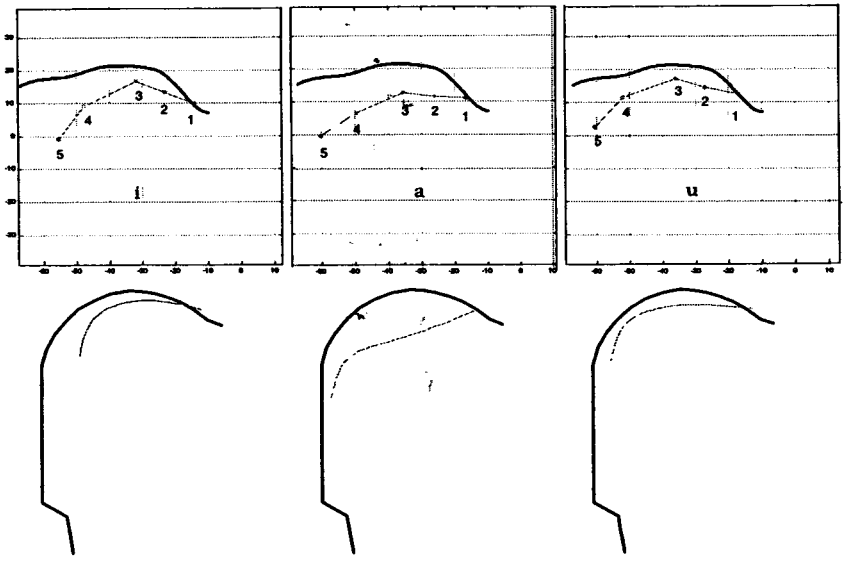

Figure 3. Equifinality of tongue tip closure: data and simulation. **Top**: Spatial positions of X-ray pellets during medial consonants of /pi'tip/, /pa'tap/, and /pu'tup/. Pellets on the surface of the tongue are connected by line segments are numbered from 1 (anterior) to 5 (posterior). Curve is an estimate of the position of speaker's palate. Units are millimeters. **Bottom**: Simulation of /idi/, /ada/, /udu/ using computational gestural model (Browman and Goldstein 1992a), incorporating task dynamics. Front part of tongue during /d/ is shown superimposed on overall outline of vocal tract. (Simulation was not an attempt to model the specific data shown, but employed general principles of gestural score constriction for English.)

Equifinality (or motor equivalence) in the task space is also a key property of other models of speech production, for example that developed by Guenther and his colleagues (e.g., Guenther 1995; this volume). A salient difference between Guenther's model and Saltzman's task dynamics model is that Guenther's model employs acoustic states in addition to orosensory states (which include constriction parameters) to specify speech goals, while the task dynamics model is completely constriction based. Summarizing the arguments for or against acoustic goals for various types of phonological units goes well beyond the scope of this paper. Suffice it to say that the final experiments on this question have not yet been performed.

3.3.2. The direct realist theory of speech perception

In the task dynamics account, gestures, the atoms of the phonology of languages, directly structure the acoustic speech signal. This allows, but does not guarantee, that listeners receive acoustic structure having sufficient information to specify the atoms. If this happens, then, in the context of the theories of Articulatory Phonology and task dynamics, language forms are preserved throughout a communicative exchange. Gestures are the common currency.

In James Gibson's direct realist theory of perception (e.g., Gibson 1966, 1979), informational media such as light for seeing, and acoustic energy for hearing do specify their distal sources. This is because there is a causal relation between properties of distal objects and events and structure in those media, and because distinctive properties of objects and events tend to structure the media distinctively. In this way, structure in media imparted to sensory systems serves as information for its distal source. In the account, whereas perceivers detect patterned energy distributions that stimulate their sense organs (the "proximal stimulus"), they perceive the properties and events in the world ("distal" events) that the energy distributions provide information about.

In the direct realist theory of speech perception (Best 1995; Fowler 1986, 1996), the gestures (distal events) that causally structure acoustic speech signals (proximal stimulation) are, in turn, specified by them. When the same gestures causally structure reflected light, they may also be specified by reflected light structure. This gives rise to the common currency that allows infants as well as adults to integrate speech information cross-modally.

Readers unfamiliar with direct realism sometimes ask how infants learn to connect the patterns of stimulation at their sense organs with the properties that stimulation specifies in the world. This question, however, has to be ill-posed. The only way that perceivers can know the world is via the information in stimulation. Therefore, the infant cannot be supposed to have two things that need to be connected: properties in the world and stimulation at sense organs. They can only have stimulation at the sense organs that, from the start, gives them properties in the world. This is not to deny that perceptual learning occurs (e.g., E. Gibson and Pick 2000; Reed 1996). It is only to remark that learning to link properties of the world with stimulation at the sense organs is impossible if infants begin life with no such link enabled.

The claim that listeners to speech perceive gestures has not been accepted in the speech community, but there is evidence for it. Summaries of this evidence are available elsewhere. Here we provide just a few examples.

3.3.2.1. Perception tracks articulation, I: Findings that led to Liberman's motor theory

Two findings were pivotal in Liberman's development of his motor theory of speech perception (e.g., Liberman 1957, 1996; Liberman et al. 1967). Both findings reflect how listeners perceive stop consonants. Liberman and colleagues had found that identifiable stop consonants in consonant-vowel syllables could be synthesized by preserving either of two salient cues for it, the second formant transition into the vowel or the stop burst.

In a study in which consonants were specified by their second formant transitions, Liberman and colleagues (Liberman et al. 1954) found that the transitions for the /d/ in /dV/ syllables were quite different depending on the vowel. /di/ and /du/ provided a striking comparison in this respect. The second formant transition of /di/ is high and rising in frequency to the level of the high second formant (F2) for /i/. The transition of /du/ is low and falling in frequency down to the level of the low F2 for /u/. Isolated from the rest of the syllable, the transitions sound different from one another in just the ways that their acoustic properties suggest; that is, they sound like high and low pitch glides. In context, however, they sound alike. Liberman (1957) recognized that the property that the /d/s in /di/ and /du/ share is articulatory. They are produced as constriction gestures of the tongue tip against the alveolar ridge of the palate. Because of coarticulation with the vowel, the transitions, which are generated after the constriction is released, are determined not only by the consonantal gesture, but by the vowel gesture or gestures as well.

Liberman and colleagues (Liberman, Delattre, and Cooper 1952) reported a complementary finding. /di/ and /du/ provide instances in which very different acoustic cues lead to identical perceptions of consonants that were produced in the same way. The complementary finding was that identical acoustic cues can signal very different percepts in different vocalic contexts. In this study, voiceless stops were signaled by the bursts that are produced just as stop constriction gestures are released. Liberman et al. found that the same stop burst, centered at 1440 Hz was identified as /p/ before vowels identified as /i/ and /u/, but as /k/ before /a/. Because of coarticulation with the following vowel, a stop burst at 1440 Hz had to be produced with different constriction gestures in the context of /i/ and /u/ versus /a/. These two findings led Liberman (1957) to conclude that when articulation and acoustics "go their separate ways," perception tracks articulation.

3.3.2.2. Perception tracks articulation, II: Parsing

That perception tracks articulation has been shown in a different way. Coarticulation results in an acoustic signal in which the information for a given consonant or vowel is distributed often over a substantial interval, and in that interval, production of other gestures also shapes the acoustic signal. Listeners parse the signal along gestural lines, extracting information for more than one gesture from intervals in which the acoustic signal was shaped by more than one gesture. Perception of fundamental frequency (F0) provides a striking example. The fundamental frequency of a speech signal is the result of converging effects of many gestures. Most notably, the fundamental frequency contour marks an intonational phrase. In addition, however, it can be raised or lowered locally by production of high or low vowels, which have, respectively, high and low intrinsic F0s (e.g., Whalen and Levitt 1995; Whalen et al. 1995). (That is, when talkers produce /i/ and /a/, attempting to match them in pitch, /i/ has a higher F0 than /a/.) The F0 contour may also be raised on a vowel that follows a voiceless obstruent (e.g., Silverman 1986, 1987). Remarkably, listeners do not perceive the F0 contour as the intonation contour; they perceive it as the intonation contour after parsing from F0 the effects of intrinsic F0 of vowels and of consonant devoicing. That is, two equal F0 peaks (marking pitch accents in the intonation contour) are heard as different in pitch height if one is produced on an /i/ and one on an /a/ vowel (Silverman 1987) or if one occurs on a vowel following an unvoiced and one a voiced obstruent (Pardo and Fowler 1997). The parsed F0 information is not discarded by perceivers. Parsed information about vowel intrinsic F0 is used by listeners as information for vowel height (Reinholt Peterson 1986); parsed information about consonant voicing is used as such (Pardo and Fowler 1997).

Nor are findings of gestural parsing restricted to perception of information provided by F0. Listeners use coarticulatory information for a forthcoming vowel as such (e.g., Fowler and Smith 1986; Martin and Bunnell 1981, 1982), and that coarticulatory information does not contribute to the perceived quality of the vowel with which

it overlaps temporally (e.g., Fowler 1981; Fowler and Smith 1986). That is, information in schwa for a forthcoming /i/ in an /ˊbi/ disyllable does not make the schwa vowel sound high; rather, it sounds just like the schwa vowel in /ˊba/, despite substantial acoustic differences between the schwas. It sounds quite different from itself cross-spliced into a /ba/ context, where parsing will pull the wrong acoustic information from it.

3.3.2.3. The common currency underlying audiovisual speech perception

In the McGurk effect (McGurk and MacDonald 1976), an appropriately selected acoustic syllable or word is dubbed onto a face mouthing something else. For example, acoustic /ma/ may be dubbed onto video /da/. With eyes open, looking at the face, listeners report hearing /na/; with eyes closed, they report hearing /ma/. That is, visible information about consonantal place of articulation is integrated with acoustic information about voicing and manner. How can the information integrate?

Two possibilities have been proposed. One invokes associations between the sights and sounds of speech in memory (e.g., Diehl and Kluender 1989; Massaro 1998). The other invokes common currency. Visual perceivers are well-understood to perceive distal events (e.g., lips closing or not during speech). If, as we propose, auditory perceivers do too, then integration is understandable in the same way that, for example, Meltzoff and Moore (1997) propose that cross-person imitation of facial gestures is possible. There is common currency between events seen and heard.

As for the association account, it is disconfirmed by findings that replacing a video of a speaker with a spelled syllable can eliminate any cross-modal integration, whereas replacing the video with the haptic feel of a gesture retains the effect (Fowler and Dekle 1991). There are associations in memory between spellings and pronunciations (e.g., Stone, Vanhoy, and Van Orden 1997; Tanenhaus, Flanigan, and Seidenberg 1980). Any associations

between the manual feel of a speech gesture and the acoustic consequence should be considerably weaker than those between spelling and pronunciation. However, the magnitudes of cross-modal integration worked the other way.

Returning to the claim of the previous sections that listeners track articulations very closely, we note that they do so cross-modally as well. Listeners' parsing of coarticulated speech leads to "compensation for coarticulation" (e.g., Mann 1980). For example, in the context of a preceding /l/; ambiguous members of a /da/ to /ga/ continuum are identified as /ga/ more often than in the context of /r/. This may occur, because listeners parse the fronting coarticulatory effects of /l/ on the stop consonant, ascribing the fronting to /l/, not to the stop. Likewise, they may parse the backing effects of /r/ on the stop (but see Lotto and Kluender 1998, for a different account). Fowler, Brown and Mann (2000) found qualitatively the same result when the only information distinguishing /r/ from /l/ was optical (because an ambiguous acoustic syllable was dubbed onto a face mouthing /r/ or /l/), and the only information distinguishing /da/ from /ga/ was acoustic. Listeners track articulation cross-modally as well as unimodally.

3.3.2.4. Infants perceive gestures audiovisually

We have already cited findings by Kuhl and Meltzoff (1982) showing that four to five month old infants exhibit cross-modal matching. Given two films of a speaker mouthing a vowel, infants look longer at the film in which the speaker is mouthing the vowel they hear emanating from a loud speaker situated between the film screens. Converging with this evidence from cross-modal matching that infants extract supramodal, that is, distal world properties both when they see and hear speech is evidence that five month olds exhibit a McGurk effect (Rosenblum, Schmuckler, and Johnson 1997). Thus, infants extract gestural information from speech events before they can appreciate the linguistic significance of the gestures.

Development of that appreciation will accompany acquisition of a lexicon (e.g., Beckman and Edwards 2000).

3.3.3. A short detour: Mirror neurons

Our chapter is about the character of public language use. However, by request, we turn from that topic to consider a mechanism that, by some accounts, might plausibly underlie maintenance of a common currency between production and perception of speech. That is the mirror neuron uncovered in the research of Rizzolatti and colleagues (e.g., Gallese et al. 1996; Rizzolatti et al. 1996).

Mirror neurons, found in area F5 of the monkey cortex, fire both when the monkey observes an action, for example, a particular kind of reaching and grasping action, and when the monkey performs the same action. This is, indeed, a part of a mechanism that connects action to corresponding perceptions, and so part of a mechanism that arguably might underlie speech perception and production if those skills have the character that we propose. Indeed, it is notable that mirror neurons are found in the monkey homologue of Broca's area in the human brain, an area that is active when language is used.

However, we hasten to point out that mirror neurons in no way *explain* how production and perception can be linked or can be recognized to correspond. Indeed, mirror neurons pose essentially the same theoretical problem that we are addressing in this chapter, but now at the level of neurons rather than of whole organisms. How can a neuron "recognize" the correspondences that mirror neurons do? They fire in the brain of a monkey when a *human* (or other monkey) performs an appropriate action or when the same action is performed by the monkey itself. But how can the same action be recognized as such? Presumably the theory of mirror neuron performance will have to include some concept of common currency of the sort we have proposed for human speaker/listeners. In short, mirror neurons must constitute the culmination of the operations of a very smart neural mechanism the workings of which we understand no better than we understand the achievements of perceiver/actors.

3.3.4. Articulatory Phonology, task dynamics and direct realism

Articulatory Phonology, task dynamics and direct realism together constitute a mutually consistent account of public communication of language forms. Articulatory Phonology provides gestures that are public actions as well as being phonological forms. Task dynamics shows how these forms can be implemented nondestructively despite coarticulation. Direct realism suggests that gestures are perceptual objects.

Together, these accounts of phonological competence, speech production and speech perception constitute a perspective on language performance as parity-fostering.

4. Public language and the emergence of phonological structure

We have argued that the most basic phonological units must be discrete and re-combinable, and also that phonological units should provide a currency common to speech production and perception. Putting these two desiderata together means that we should find the same categorical units common to both speech perception and production that can serve as phonological primitives. In this section, we review some preliminary work that investigates the basis for decomposing production and perception into discrete units. This leads, in turn, to some predictions about the emergence of phonological categories in the course of phonological development and these appear to be borne out.

4.1. Distinct organs

We argued that one basis for discrete units in speech production could be found in the constricting organs of the vocal tract (lips, tongue tip, tongue body, tongue root, velum, larynx). They are anatomically distinct from one another and capable of performing

constricting actions independently of one another. Note, too, that there is generally no continuum on which the organs lie, so that the notion of a constriction intermediate between two organs is not defined, e.g., between a velic constriction and a laryngeal constriction. (Although one could argue that the boundaries between the tongue organs are less sharp in this particular sense, they still exhibit distinctness and independence.) Studdert-Kennedy has proposed that this "particulation" of the vocal tract into organs could have provided the starting point for the evolution of (arbitrary) discrete categories in language (Studdert-Kennedy 1998; Studdert-Kennedy and Goldstein in press).

But what of discrete organs in speech perception? Evidence for perception of discrete organs can be found in the experiments on facial mimicry in infants described above (Meltzoff and Moore 1997). The imitations by infants are specific to the organ used by the adult model, even though they are not always correct in detail (for example, an adult gesture of protruding the tongue to the side might be mimicked by tongue protrusion without the lateral movement). Meltzoff and Moore report that when a facial display is presented to infants, all of their facial organs stop moving, except the one involved in the adult gesture.

So infants are able to distinguish facial organs (lips, tongue, eyes) as distinct objects, capable of performing distinct classes of actions. The set of organs involved in speech intersects with those on the face (lips and tongue common to both, though the tongue is further decomposed into tongue tip and tongue body), and also includes the glottis and the velum. As gestures of the speech organs are usually specified by structure in the acoustic medium (in addition to, or instead of) the optic medium, we would predict that acoustic medium could also be used to trigger some kind of imitative response on the part of infants to gestures of the speech-related organs (not necessarily speech gestures). So, for example, we would expect that infants would move their lips if presented with an auditory lip smack, or with [ba] syllables. Such experiments are being pursued in our laboratory.[5] Prima facie evidence that infants can, in fact, use acoustic structure in this way can be found in their categorical

response to speech sounds produced with distinct organs, e.g., "place" distinctions – oral gestures of lips vs. tongue tip vs. tongue body – in newborns (Bertoncini et al. 1987). While it had been claimed (Jusczyk 1997) that young infants exhibit adult-like perception of *all* phonological contrasts tested (not just between-organ contrasts) some recent reports suggest that certain contrasts may not be so well perceived by young infants, or may show decreased discriminability at ages 10-12 months, even when they are present in the ambient language. These more poorly discriminated contrasts are all within-organ contrasts. For example, Polka, Colantonio, and Sundara (2001) found that English-learning infants aged 6-8 months showed poor discrimination of a /d – ɖ/ contrast, a within-organ distinction that is contrastive in English. Best and McRoberts (in press) report decreased discriminability for a variety of within-organ contrasts at ages 10-12 months, *regardless* of whether they are contrastive in the language environment in which the child is being raised, but good discrimination of between-organ contrasts, even when the segments in question are not contrastive in the learning environment (e.g. labial vs. coronal ejective stops for English-learning infants).

The remarkable ability of infants to distinguish organs leads naturally to a view that distinct organs should play a key role in the emergence of phonology in the child. Recent analyses of early child phonology provide preliminary support for such a view. Ferguson and Farwell (1975), for example, showed that narrow phonetic transcriptions of a child's earliest words (first 50) are quite variable. The initial consonant of a given word (or set of words) was transcribed as different phonetic units on different occasions, and Ferguson and Farwell (1975) argue that the variability is too extreme for the child to have a coherent phonemic system, with allophones neatly grouped into phonemes, and that the basic unit of the child's production must therefore be the whole word, not the segment. However, if the child's word productions are re-analyzed in terms of the organs involved, it turns out that children are remarkably consistent in the organ they move at the beginning of a given word (Studdert-Kennedy 2000, in press; Studdert-Kennedy and Goldstein,

in press), particularly the oral constriction organ (lips, tongue tip, tongue body). Thus, children appear to be acquiring a relation between actions of distinct organs and lexical units very early in the process of developing language. Organ identity is common to production and perception and very early on is used for lexical contrast.

4.2. Mutual attunement and the emergence of within-organ contrasts

As discussed earlier, distinct actions of a given organ can also function as contrastive gestures. Such contrastive gestures typically differ in the attractor states of the tract variables that control the particular organ. For example, *bet*, *vet*, and *wet* all begin with gestures of the lips organ, but the gestures contrast in the state (value) of the Lip Aperture (LA) tract variable (degree of lip constriction). Lips are most constricted for "bet," less so for *vet*, and least constricted for *wet*. Tongue tip gestures at the beginning of the words *thick* and *sick* differ the value of the Tongue Tip Constriction Location (TTCL) tract variable (position of the tongue tip along upper teeth and/or palate). The contrasting attractor values along LA or TTCL are in principle points along a continuum. How are these continua partitioned into contrasting states?

One hypothesis is that the categories emerge as a consequence of satisfying the requirement that phonological actions be shared by members of a speech community. In order to satisfy that requirement, members must attune their actions to one another. Such attunement can be seen in the spread of certain types of sound changes in progress (e.g. Labov 1994), in the *gestural drift* found when a speaker changes speech communities (Sancier and Fowler 1997), and in the babbling behavior of infants as young as 6 months old (de Boysson-Bardies et al. 1992; Whalen, Levitt, and Wang 1991). Mutual attunement must be accomplished primarily through the acoustic medium. Because the relation between constriction parameters and their acoustic properties is nonlinear (Stevens 1989),

certain regions of a tract variable continuum will afford attunement, while others will not. Thus, the categories we observe could represent just those values (or regions) of the tract variable parameters that afford attunement. They are an example of self-organization through the public interaction of multiple speakers.

It is possible to test this hypothesis through computational simulation of a population of agents that acts randomly under a set of constraints or conditions. (For examples of self-organizing simulations in phonology, see Browman and Goldstein 2000; de Boer 2000; Zuraw 2000). In a preliminary simulation designed to investigate the partitioning of a tract variable constriction continuum into discrete regions, agents interacted randomly under the following three conditions: (a) Agents attempt to attune their actions to one another. (b) Agents recover the constriction parameters used by their partners from the acoustic signal, and that recovery is assumed to be noisy. (c) The relation between constriction and acoustics is nonlinear.

The simulation investigated an idealized constriction degree (CD) continuum and how it is partitioned into three categories (corresponding to stops, fricatives, and glides). Figure 4 shows the function used to map constriction degree to a hypothetical acoustical property, which could represent something like the overall amplitude of acoustic energy that emerges from the vocal tract during the constriction. The crucial point is that the form of the nonlinear function follows that hypothesized by Stevens (1989) for constriction degree and several other articulatory-acoustic mappings. Regions of relative stability (associated with stops, fricatives, and glides) are separated by regions of rapid change. The constriction degree continuum was divided into 80 equal intervals.

Two agents were employed. Each agent produced one of the 80 intervals at random, with some *a priori* probability associated with each interval. At the outset, all probabilities were set to be equal. The simulation then proceeded as follows. On each trial, the two agents produced one of the 80 intervals at random. Each agent then recovered the CD produced by its partner from its acoustic property, and compared that CD value to the one it produced itself. If they

matched, within some criterion, the agent incremented the probability of producing that value of CD again. The recovery process works like this. The function in Figure 4 is a true function, so an acoustic value can be mapped uniquely onto the CD value. However, since we assume the acoustics to be noisy in the real world, a range of CD values is actually recovered, within +/- 3 acoustic units of that actually produced.

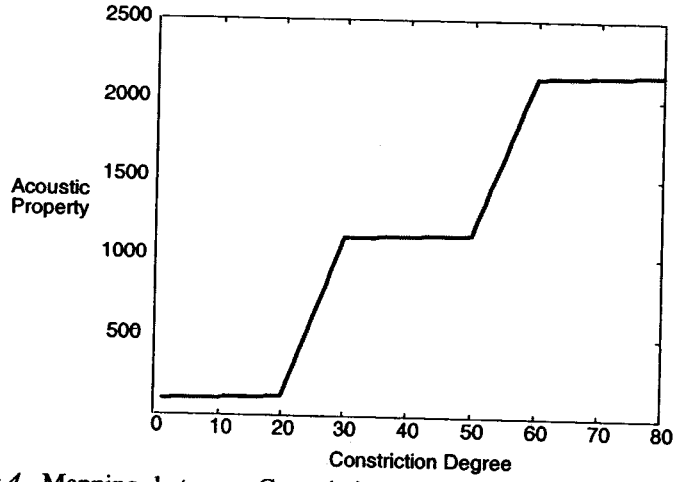

Figure 4. Mapping between Constriction Degree and a hypothetical Acoustic Property used in agent-based computational experiment. Shape of function after Stevens (1989).

Typical results of such a simulation (60,000 trials) are shown in Figure 5. For both agents (T1 and T2), the CD productions are now partitioned into 3 modes corresponding to the stable states of Figure 4. These values of CD in these regions are relatively frequently matched because of their acoustic similarity: Several values of CD fall into the +/-3 acoustic unit noise range in these regions, while only a few fall within the +/- 3 range in the unstable regions. Thus, mutual attunement, a concept available only in public language use, gives rise to discrete modes of constriction degree, under the influence of a nonlinear articulatory-acoustic map.

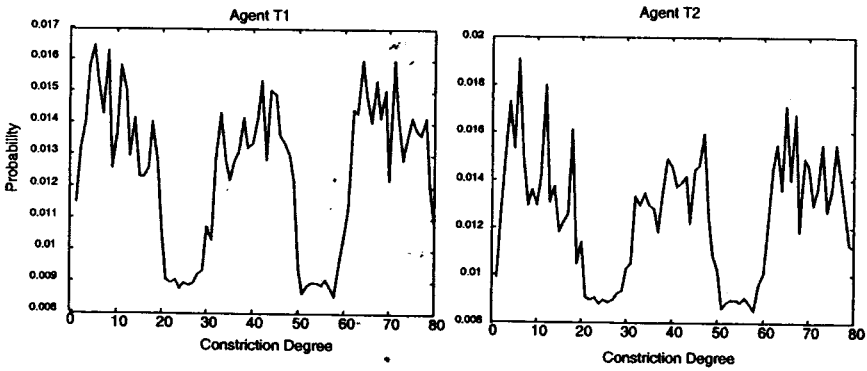

Figure 5. Results of agent-based experiment. Probability distributions of the two agents (T1 and T2) emitting particular values of constriction degree (60,000 iterations).

The role of attunement in partitioning within-organ contrasts is further supported by the developmental data described above. While children's early words are consistent in the oral constriction organ employed, and match the adult models in this regard, they are quite variable in within-organ properties, such as constriction degree (or constriction location). The simulations suggest that within-organ categories emerge only from attunement, which presumably takes some time. This conclusion is further bolstered by the recent perceptual findings with infants 10-12 months of age (Best and McRoberts in press), showing reduced discrimination for within-organ contrasts, even when the contrasts can be found in the language the child is about to acquire. At this age, infants have only begun to attune their vocal behavior to the language environment (de Boyssson-Bardies et al. 1992), and therefore partitioning of within-organ categories is expected to be incomplete.

5. Conclusions

In the fields of linguistics and psycholinguistics, there is an almost exclusive focus on the individual speaker/hearer and on the covert

knowledge (linguistics) or the covert mental categories and processes (psycholinguistics) that support language use. In these approaches, there has appeared to be no scientific harm in studying phonological competence, speech production and speech perception independently, and that is how research in these domains has proceeded for the most part. Independent investigation of these domains means that theories of phonology, speech production and speech perception are each largely unconstrained by the others. Phonological forms are not constrained to be producible in a vocal tract, descriptions of vocal tract activities need not be, and are not, descriptions of phonological forms, and neither phonological forms, nor vocal tract activities need to be perceivable.

There *is* scientific harm in this approach, however, because language use is almost entirely a between-person activity, and it matters whether or not listeners perceive the language forms that speakers intend to convey. In our view, the prevalent views that language forms are mental categories, that coarticulation ensures that vocal tract activity is not isomorphic with the forms, and that listeners perceive acoustic cues are the erroneous consequences of the exclusive focus on the individual language user.

We start from the premise that languages evolved to be spoken and heard and, therefore, that language forms – the means that languages provide for making language use public – are likely to be public events. We have shown that a phonological system can be composed of public vocal events – that is, gestures. The gestures, not acoustic cues, can be supposed to be perceived. And public language forms can be shown to emerge; with the properties of ready produceability and perceivability that language forms do have, from imitative communicative exchanges between people. This, we argue, is real language.

Notes

1. Preparation of the manuscript was supported by NICHD Grant HD-01994, and NIDCD grants DC-03782 and DC-02717 to Haskins Laboratories.
2. Obviously, this sameness has to be within some tolerance. For example, not everyone who speaks the same language speaks the same dialect.
3. This is not the solution adopted by Meltzoff and Kuhl (1994) to explain audiovisual integration of speech. They propose that, during cooing and babbling, as infants produce speech-like acoustic signals, they learn a mapping from articulations to acoustic signals. Articulations, like other facial or vocal tract actions, can be perceived supramodally. Accordingly, the articulation-acoustic mapping can underlie audiovisual integration of speech.
4. It is, of course, possible that the essential properties of a gesture may fail to be completely achieved in some prosodic, stylistic, or informational contexts. For example, a closure gesture may be reduced and fail to result in complete closure. In such cases, the reduction can serve as information about the context. If the case of coarticulation, however, the essential properties of a gesture would be *systematically* obscured (and never achieved), if it were not for equifinality.
5. MacNeilage (1998) has argued that speech emerges from oscillatory movements of the jaw without specific controls for lips, tongue tip, and tongue body. It is possible that infants' early production of utterances with the global rhythmical properties of speech have the properties he proposes. However, infants may have some control of individual movements of the separate organs. It is just their integration into a more global structure that occurs only after that global structure is established through mandibular oscillation.

References

Abler, William
1989 On the particulate principle of self-diversifying systems. *Journal of Social and Biological Structures* 12: 1-13.

Beckman, Mary and Jan Edwards
2000 The ontogeny of phonological categories and the primacy of lexical learning in linguistic development. *Child Development* 71: 240-249.

Bertoncini, Josiane, Ranka Bijeljac-Babic, Sheila E. Blumstein and Jacques Mehler
1987 Discrimination in neonates of very short CV's. *Journal of the Acoustical Society of America* 82: 31-37.

Best, Catherine T.
1995 A direct realist perspective on cross-language speech perception. In: Winifred Strange and James J. Jenkins (eds.), *Cross-Language Speech Perception*, 171-204. Timonium, MD: York Press.

Best, Catherine T. and Gerald W. McRoberts
In press Infant perception of nonnative contrasts that adults assimilate in different ways. *Language and Speech.*

de Boer, Bart
2000 Self-organization in vowel systems. *Journal of Phonetics* 28: 441-465.

de Boysson-Bardies, Benedicte, Marilyn Vihman, Liselotte Roug-Hellichius,Catherine Durand, Ingrid Landberg and Fumiko Arao
1992 Material evidence of infant selection from the target language: a cross-linguistic study. In: Charles Ferguson, Lise Menn, and Carol Stoel-Gammon (eds.), *Phonological Development: Models, Resesarch, Implications*, 369-391. Timonium, MD: York Press.

Browman, Catherine P.
1994 Lip aperture and consonant releases. In: Patricia Keating (ed.), *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*, 331-353. Cambridge: Cambridge University Press.

Browman, Catherine P. and Louis Goldstein
1986 Towards an articulatory phonology. *Phonology Yearbook* 3: 219-252.

Browman, Catherine P. and Louis Goldstein
1989 Articulatory gestures as phonological units. *Phonology* 6: 151-206.

Browman, Catherine P. and Louis Goldstein
1990 Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics* 18: 299-320.

Browman, Catherine P. and Louis Goldstein
1992a Articulatory phonology: An overview. *Phonetica* 49: 155-180.

Browman, Catherine P. and Louis Goldstein
1992b Targetless schwa: An articulatory analysis. In: Gerard J. Docherty and D. Robert Ladd (eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody,* 26-56. Cambridge: Cambridge University Press.

Browman, Catherine P. and Louis Goldstein
1995a Dynamics and articulatory phonology. In: Robert Port and Timothy van Gelder (eds.), *Mind as Motion: Explorations in the Dynamics of Cognition,* 175-193. Cambridge, MA: MIT Press.

Browman, Catherine P. and Louis Goldstein
1995b Gestural syllable position effects in American English. In: Fredericka Bell-Berti and Lawrence Raphael (eds.), *Producing Speech: Contemporary Issues,* 19-33. New York: American Institute of Physics.

Browman, Catherine P. and Louis Goldstein
2000 Competing constraints on intergestural coordination and self-organization of phonological structures. *Bulletin de la Communication Parlée* 5: 25-34.

Byrd, Dani
1996 Influences on articulatory timing in consonant sequences. *Journal of Phonetics* 24: 209-244.

Clements, G. N.
1985 The geometry of phonological features. *Phonology Yearbook* 2: 225-252.

Cooper, Franklin S., Pierre C. Delattre, Alvin M. Liberman, John M. Borst and Louis J. Gerstman
1952 Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America* 24: 597-606.

Diehl, Randy L. and Keith R. Kluender
1989 On the objects of speech perception. *Ecological Psychology* 1: 121-144.

Elson, Ben
1947 Sierra Popoluca syllable structure. *International Journal of American Linguistics* 13: 13-17.

Ferguson, Charles A. and Carol B. Farwell
1975 Words and sounds in early language acquisition. *Language* 51: 419-439.

Fontana, Walter and Leo Buss
1996 The barrier of objects: From dynamical systems to bounded organizations. In: John Casti and Anders Karlquist (eds.), *Boundaries and Barriers,* 56-116. Addison-Wesley, Reading, MA.

Fowler, Carol A.
1980 Coarticulation and theories of extrinsic timing. *Journal of Phonetics* 8: 113-133.

Fowler, Carol A.
1981 Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech and Hearing Research* 46: 127-139.

Fowler, Carol A.
1986 An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics* 14: 3-28.

Fowler, Carol A.
1996 Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America* 99: 1730-1741.

Fowler, Carol A., Julie Brown and Virginia Mann
2000 Contrast effects do not underlie effects of preceding liquid consonants on stop identification in humans. *Journal of Experimental Psychology: Human Perception and Performance* 26: 877-888.

Fowler, Carol A. and Dawn J. Dekle
1991 Listening with eye and hand: Crossmodal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance* 17: 816-828.

Fowler, Carol A. and Mary Smith
1986 Speech perception as "vector analysis": An approach to the problems of invariance and segmentation. In: Joseph S. Perkell and Dennis H. Klatt (eds.), *Invariance and Variability in Speech Processes*, 123-136. Hillsdale, NJ: Lawrence Erlbaum Associates.

Fromkin, Victoria
1973 *Speech Errors as Linguistic Evidence.* The Hague: Mouton.

Gafos, Adamantios
2002 A grammar of gestural coordination. *Natural Language and Linguistic Theory* 20: 269-337.

Gallese, Vittorio, Luciano Fadiga, Leonardo Fogassi and Giacomo Rizzolatti
1996 Action recognition in the premotor cortex. *Brain* 119: 593-609.

Gibson, Eleanor and Anne Pick
2000 *An Ecological Approach to Perceptual Learning and Development.* Oxford: Oxford University Press.

Gibson, James J.
1966 *The Senses Considered as Perceptual Systems.* Boston, MA: Houghton-Mifflin.

Gibson, James J.
1979 *The Ecological Approach to Visual Perception.* Boston, MA: Houghton Mifflin.

Gick, Bryan
In press Articulatory correlates of ambisyllabicity in English glides and liquids. In: John Local, Richard Ogden and Rosalind Temple (eds.), *Laboratory Phonology VI: Constraints on Phonetic Interpretation*. Cambridge: Cambridge University Press.

Gick, Bryan and Louis Goldstein
2002 Relative timing of the gestures of North American English /r/. *Journal of the Acoustical Society of America* 111: 2481.

Goldstein, Louis, Marianne Pouplier, Larissa Chen and Dani Byrd
Submitted Dynamic action units slip in speech production errors. *Nature*.

Gracco, Vincent L. and James H. Abbs
1982 Compensatory response capabilities of the labial system to variation in onset of unanticipated loads. *Journal of the Acoustical Society of America* 71: S34.

Guenther, Frank
1995 Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review* 102, 594-621.

Haken, Hermann, Lieke Peper, Peter J. Beek and Andreas Dafferthofer
1996 A model for phase transitions. *Physica D* 90: 179-196.

Halle, Morris
1983 On distinctive features and their articulatory implementation. *Natural Language and Linguistic Theory* 1: 91-105.

Harris, Cyril M.
1953 A study of the building blocks in speech. *Journal of the Acoustical Society of America* 25: 962-969.

Hockett, Charles
1955 *A Manual of Phonetics*. Bloomington, Indiana: Indiana University Press.

Hommel, Bernhard, Jochen Müsseler, Gisa Aschersleben and Wolfgang Prinz
2001 The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences* 24: 849-937.

Jusczyk, Peter
1997 *The Discovery of Spoken Language*. Cambridge, MA: MIT Press.

Keating, Patricia A.
1990 The window model of coarticulation: Articulatory evidence. In: John Kingston and Mary E. Beckman (eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, 451-470. Cambridge: Cambridge University Press.

Kelso, J. A. Scott, Betty Tuller, Eric Vatikiotis-Bateson and Carol A. Fowler
1984 Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance* 10: 812-832.

Krakow, Rena
1993 Nonsegmental influences on velum movement patterns: Syllables, segments, stress and speaking rate. In: Marie Huffman and Rena Krakow (eds.), *Phonetics and Phonology, 5: Nasals, Nasalization and the Velum,* 87-116. New York: Academic Press.

Krakow, Rena
1999 Physiological organization of syllables: a review. *Journal of Phonetics* 27: 23-54.

Kuhl, Patricia and Andrew Meltzoff
1982 The bimodal perception of speech in infancy. *Science* 218: 1138-1141.

Kuhl, Patricia and Andrew Meltzoff
1996 Infant vocalizations in response to speech: Vocal imitation and developmental change. *Journal of the Acoustical Society of America* 100: 2425-2438.

Labov, William
1994 *Principles of Linguistic Change. Volume 1: Internal Factors.* Oxford: Basil Blackwell.

Liberman, Alvin M.
1957 Some results of research on speech perception. *Journal of the Acoustical Society of America* 29: 117-123.

Liberman, Alvin M.
1996 *Speech: A Special Code.* Cambridge, MA: Bradford Books.

Liberman, Alvin, Franklin S. Cooper, Donald Shankweiler and Michael Studdert-Kennedy
1967 Perception of the speech code. *Psychological Review* 74: 431-461.

Liberman, Alvin, Pierre Delattre and Franklin S. Cooper
1952 The role of selected stimulus variables in the perception of the unvoiced-stop consonants. *American Journal of Psychology* 65: 497-516.

Liberman, Alvin M., Pierre Delattre, Franklin S. Cooper and Louis Gerstman
1954 The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied* 68: 1-13.

Liberman, Alvin M. and Douglas H. Whalen
2000 On the relation of speech to language. *Trends in Cognitive Sciences* 4: 187-196.
Lotto, Andrew and Keith Kluender
1998 General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics* 60: 602-619.
MacNeilage, Peter F.
1998 The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences* 21: 499-511.
Mann, Virginia
1980 Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics* 28: 407-412.
Martin, James G. and Timothy Bunnell
1981 Perception of anticipatory coarticulation effects in /stri, stru/ sequences. *Journal of the Acoustical Society of America* 69: S92.
Martin, James G. and Timothy Bunnell
1982 Perception of anticipatory coarticulation effects in vowel-stop consonant-vowel syllables. *Journal of Experimental Psychology: Human Perception and Performance* 8: 473-488.
Massaro, Dominic
1998 *Perceiving Talking Faces*. Cambridge, MA: MIT Press.
Mattingly, Ignatius
1981 Phonetic representation and speech synthesis by rule. In: Terry Myers, John Laver and John Anderson (eds.), *The Cognitive Representation of Speech*, 415-420. Amsterdam: North Holland Publishing Company.
McGurk, Harry and John MacDonald
1976 Hearing lips and seeing voices. *Nature* 264: 746-748.
Meltzoff, Andrew N. and Patricia Kuhl
1994 Faces and speech: Intermodal processing of biologically relevant signals in infants and adults. In: David Lewkowicz and Robert Likliter (eds.), *The Development of Intersensory Perception: Comparative Perspectives*, 335-369. Hillsdale, NJ: Lawrence Erlbaum.
Meltzoff, Andrew N. and M. Keith Moore
1977 Imitation of facial and manual gestures by human infants. *Science* 198: 75-78.
Meltzoff, Andrew N. and M. Keith Moore
1983 Newborn infants imitate adults' facial gestures. *Child Development* 54: 702-709.

Meltzoff, Andrew N. and M. Keith Moore
1997 Explaining facial imitation: a theoretical model. *Early Development and Parenting* 6: 179-192.

Meltzoff, Andrew N. and M. Keith Moore
1999 Persons and representation: Why infant imitation is important for theories of human development. In: Jacqueline Nadel and George Butterworth (eds.), *Imitation in Infancy*, 9-35. Cambridge: Cambridge University Press.

Mowrey, Richard and Ian MacKay
1990 Phonological primitives: Electromyographic speech error evidence. *Journal of the Acoustical Society of America* 88: 1299-1312.

Munhall, Kevin G. and Anders Löfqvist
1992 Gestural aggregation in speech: Laryngeal gestures. *Journal of Phonetics* 20: 111-126.

Ohala, John
1981 The listener as a source of sound change. In: Carrie Masek, Roberta Hendrick and Mary Frances Miller (eds.), *Papers from the Parasession on Language and Behavior,* 178-203. Chicago: Chicago Linguistics Society.

Pardo, Jennifer and Carol A. Fowler
1997 Perceiving the causes of coarticulatory acoustic variation: Consonant voicing and vowel pitch. *Perception & Psychophysics* 59: 1141-1152.

Perrier, Pascal, Hélène Loevenbruck and Yohan Payan
1996 Control of tongue movements in speech: the Equilibrium Point Hypothesis perspective. *Journal of Phonetics* 24: 53-75.

Polka, Linda, Connie Colantonio and Megha Sundara
2001 A cross-language comparison of d-/ð/ perception: Evidence for a new developmental pattern. *Journal of the Acoustical Society of America* 109: 2190-2201.

Pouplier, Marianne and Louis Goldstein
Submitted Asymmetries in speech errors: Production, perception and the question of underspecification. *Journal of Phonetics.*

Reed, Edward
1996 *Encountering the World: Toward an Ecological Psychology.* Oxford: Oxford University Press.

Reinholt Peterson, Niels
1986 Perceptual compensation for segmentally-conditioned fundamental-frequency perturbations. *Phonetica* 43: 31-42.

Rizzolatti, Giacomo, Luciano Fadiga, Vittorio Gallese and Leonardo Fogassi
1996 Premotor cortex and the recognition of motor actions. *Cognitive Brain Research 3*: 131-141.

Rosenblum, Lawrence, Mark Schmuckler and Jennifer Johnson
1997 The McGurk effect in infants. *Perception & Psychophysics* 59: 347-357.

Saltzman, Elliot L.
1986 Task dynamic coordination of the speech articulators: A preliminary model. Generation and modulation of action patterns. In: Herbert Heuer and Christoph Fromm (eds.), *Experimental Brain Research, Series 15*, 129-144. New York: Springer-Verlag.

Saltzman, Elliot L.
1991 The task dynamic model in speech production. In: Herman F. M. Peters, Wouter Hulstijn and C. Woodruff Starkweather (eds.), *Speech Motor Control and Stuttering*, 37-52. Amsterdam: Elsevier Science Publishers.

Saltzman, Elliot L.
1995 Dynamics and coordinate systems in skilled sensorimotor activity. In: Robert Port and Timothy van Gelder (eds.), *Mind as Motion: Explorations in the Dynamics of Cognition*, 150-173. Cambridge, MA: MIT Press.

Saltzman, Elliot L. and Dani Byrd
2000 Task-dynamics of gestural timing: Phase windows and multi-frequency rhythms. *Human Movement Science* 19: 499-526.

Saltzman, Elliot L. and J. A. Scott Kelso
1987 Skilled action: A task-dynamic approach. *Psychological Review* 94: 84-106.

Saltzman, Elliot L. and Kevin G. Munhall
1989 A dynamical approach to gestural patterning in speech production. *Ecological Psychology* 1: 333-382.

Sancier, Michele and Carol A. Fowler
1997 Gestural drift in a bilingual speaker of Brazilian Portuguese. *Journal of Phonetics* 25: 421-436.

Shaiman, Susan
1989 Kinematic and electromyographic responses to perturbation of the jaw. *Journal of the Acoustical Society of America* 86: 78-88.

Shattuck-Hufnagel, Stefanie
1983 Sublexical units and suprasegmental structure in speech production. In: Peter MacNeilage (ed.), *The Production of Speech*, 109-136. New York: Springer-Verlag.

Shattuck-Hufnagel, Stefanie and Dennis Klatt
1979 Minimal uses of features and markedness in speech production: Evidence from speech errors. *Journal of Verbal Learning and Verbal Behavior* 18: 41-55.

Silverman, Kim
1986 F_0 cues depend on intonation: The case of the rise after voiced stops. *Phonetica* 43: 76-92.

Silverman, Kim
1987 *The Structure and Processing of Fundamental Frequency Contours.* PhD Dissertation, Department of Psychology, Cambridge University.

Sproat, Richard and Osamu Fujimura
1993 Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics* 21: 291-311.

Stevens, Kenneth N.
1989 On the quantal nature of speech. *Journal of Phonetics* 17: 3-45.

Stevens, Kenneth N.
1999 *Acoustic Phonetics.* Cambridge, MA: MIT Press.

Stone, Gregory, Mickie Vanhoy and Guy Van Orden
1997 Perception is a two-way street: Feedforward and feedback in visual word recognition. *Journal of Memory and Language* 36: 337-359.

Studdert-Kennedy, Michael
1998 The particulate origins of language generativity. In: James Hurford, Michael Studdert-Kennedy and Christopher Knight (eds.), *Approaches to the Evolution of Language*, 202-221. Cambridge: Cambridge University Press.

Studdert-Kennedy, Michael
2000 Imitation and the emergence of segments. *Phonetica* 57: 275-283.

Studdert-Kennedy, Michael
In press Mirror neurons, vocal imitation, and the evolution of particulate speech. In: Vittorio Gallese and Maxim Stamenov (eds.), *Mirror Neurons and the Evolution of the Brain and Language.* Amsterdam: John Benjamins.

Studdert-Kennedy, Michael and Louis Goldstein
In press Launching language: The gestural origin of discrete infinity. In: Morten Christiansen and Simon Kirby (eds.), *Language Evolution: The States of the Art.* Oxford: Oxford University Press.

Tanenhaus, Michael, Helen Flanigan and Mark Seidenberg
1980 Orthographic and phonological activation in auditory and visual word recognition. *Memory & Cognition* 8: 513-520.

Whalen, Douglas H., Andrea G. Levitt and Qi Wang
1991 Intonational differences between the reduplicative babbling of French- and English-learning infants. *Journal of Child Language* 18: 501-516.

Whalen, Douglas H. and Andrea G. Levitt
1995 The universality of intrinsic F0 of vowels. *Journal of Phonetics* 23: 349-366.

Whalen, Douglas H., Andrea G. Levitt, Pai Ling Hsaio and Iris Smorodinsky
1995 Intrinsic F0 of vowels in the babbling of 6-, 9-, and 12-month old French- and English-learning infants. *Journal of the Acoustical Society of America* 97: 2533-2539.

Zuraw, Kie
2000 *Patterned Exceptions in Phonology*. PhD Dissertation, Department of Linguistics, University of California, Los Angeles.