Handbook of Psychology, Vol. 4, Experimental Psychology, 2003,
John Wiley & Sons, Inc., New Jersey.

1271

CHAPTER 9

Speech Production and Perception

CAROL A. FOWLER

In order to convey linguistic messages that are accessible to listeners, speakers have to engage in activities that count in their language community as encodings of the messages in the public domain. Accordingly, spoken languages consist of forms that express meanings; the forms are (or, by other accounts, give rise to) the actions that make messages public and perceivable. Psycholinguistic theories of speech are concerned with those forms and their roles in communicative events. The focus of attention in this chapter will be on the phonological forms that compose words and, more specifically, on consonants and vowels.

As for the roles of phonological forms in communicative events, four are central to the psycholinguistic study of speech. First, phonological forms may be the atoms of word forms as language users store them in the mental lexicon. To study this is to study phonological competence (that is, knowledge). Second, phonological forms retrieved from lexical entries may specify words in a mental plan for an utterance. This is phonological planning. Third, phonological forms are implemented as vocal tract activity, and to study this is to study speech production. Fourth, phonological forms may be the finest-grained linguistic forms that listeners extract from acoustic speech signals during speech perception. The main body of the chapter will constitute a review of research findings and theories in these four domains.

Before proceeding to those reviews, however, I provide a caveat and then a setting for the reviews. The caveat is about the psycholinguistic study of speech. Research and theorizing in the domains under review generally proceed independently and therefore are largely unconstrained by findings in the other domains (cf. Kent & Tjaden, 1997, and Browman & Goldstein, 1995a, who make a similar comment). As my review will reveal, many theorists have concluded that the relevant parts of a communicative exchange (phonological competence, planning, production, and perception) fit together poorly. For example, many believe that the forms of phonological competence have properties that cannot be implemented as vocal tract activity, so that the forms of language cannot literally be made public. My caveat is that this kind of conclusion may be premature; it may be a consequence of the independence of research conducted in the four domains. The stage-setting remarks just below will suggest why we should expect the fit to be good.

Preparation of this chapter was supported by NICHD grant HD-01994 and NIH grants DC-02717 and DC-03782 to Haskins Laboratories.

In the psycholinguistic study of speech, as in psycholinguistics generally (see chapter by Treiman, Clifton, Meyer, & Wurm in this volume), the focus of attention is almost solely on the individual speaker/hearer and specifically on the memory systems and mental processing that underlie speaking or listening. It is perhaps this sole focus of attention that has fostered the near autonomy of investigations into the various components of a communicative exchange just described. Outside of the laboratory, speaking almost always occurs in the context of social activity; indeed, it is, itself, prototypically a social activity. This observation matters, and it can help to shape our thinking about the psycholinguistics of speech.

Although speaker/hearers can act autonomously, and sometimes do, often they participate in cooperative activities with others; jointly the group constitutes a special purpose system organized to achieve certain goals. Cooperation requires coordination, and speaking helps to achieve the social coordinations that get conjoint goals accomplished (Clark, 1996).

How, at the phonological level of description, can speech serve this role? Speakers speak intending that their utterance communicate to relevant listeners. Listeners actively seek to identify what a talker said as a way to discover what the talker intended to achieve by saying what he or she said. Required for successful communication is achievement of a relation of sufficient equivalence between messages sent and received. I will refer to this relation, at the phonological level of description, as *parity* (Fowler & Levy, 1995; cf. Liberman & Whalen, 2000).

That parity achievement has to be a typical outcome of speech is one conclusion that emerges from a shift in perspective on language users, a shift from inside the mind or brain of an individual speaker/listener to the cooperative activities in which speech prototypically occurs. Humans would not use speech to communicate if it characteristically did not. This conclusion implies that the parts of a communicative exchange (competence, planning, production, perception) have to fit together pretty well.

A second observation suggests that languages should have parity-fostering properties. The observation is that language is an evolved, not an invented, capability of humans. This is true of speech as well as of the rest of language. There are adaptations of the brain and the vocal tract to speech (e.g., Lieberman, 1991), suggesting that selective pressures for efficacious use of speech shaped the evolutionary development of humans.

Following are two properties that, if they were characteristic of the phonological component of language, would be parity fostering. The first is that phonological forms, here consonants and vowels, should be able to be made public and therefore accessible to listeners. Languages have forms as well as meanings exactly because messages need to be made public to be communicated. The second parity-fostering characteristic is that the elements of a phonological message should be preserved throughout a communicative exchange. That is, the phonological elements of words that speakers know in their lexicons should be the phonological elements of words that they intend to communicate, they should be units of action in speech production, and they should be objects of speech perception. If the elements are not preserved—if, say, vocal tract actions are not phonological things and so acoustic signals cannot specify phonological things—then listeners have to reconstruct the talker's phonological message from whatever they can perceive. This would not foster achievement of parity.

The next four sections of the chapter review the literature on phonological competence, planning, production, and perception. The reviews will accurately reflect the near independence of the research and theorizing that goes on in each domain. However, I will suggest appropriate links between domains that reflect the foregoing considerations.

PHONOLOGICAL COMPETENCE

The focus here is on how language users know the spoken word forms of their language, concentrating on the phonological primitives, consonants and vowels (*phonetic* or *phonological* segments). Much of what we know about this has been worked out by linguists with expertise in phonetics or phonology. However, the reader will need to keep in mind that the goals of a phonetician or phonologist are not necessarily those of a psycholinguist. Psycholinguists want to know how language users store spoken words. Phoneticians seek realistic descriptions of the sound inventories of languages that permit insightful generalizations about universal tendencies and ranges of variation cross-linguistically. Phonologists seek informative descriptions of the phonological systematicities that languages evidence in their lexicons. These goals are not psychologically irrelevant, as we will see. However, for example, descriptions of phonological word forms that are most transparent to phonological regularities may or may not reflect the way that people store word forms. This contrast will become apparent below when theories of linguistic phonology are compared specifically to a recent hypothesis raised by some speech researchers that lexical memory is a memory of word tokens (exemplars), not of abstract word types.

Word forms have an internal structure, the component parts of which are meaningless. The consonants and vowels are also discrete and permutable. This is one of the ways

in which language makes "infinite use of finite means" (Von Humbolt, 1936/1972; see Studdert-Kennedy, 1998). There is no principled limit on the size of a lexicon having to do with the number of forms that can serve as words. And we do know a great many words; Pinker (1994) estimates about 60,000 in the lexicon of an average high school graduate. This is despite the fact that languages have quite limited numbers of consonants and vowels (between 11 and 141 in Maddieson's (1984) survey of 317 representative languages of the world).

In this regard, as Abler (1989) and Studdert-Kennedy (1998) observe, languages make use of a "particulate principle" also at work in biological inheritance and chemical compounding, two other domains in which infinite use is made of finite means. All three of these systems are self-diversifying in that, when the discrete particulate units of the domain (phonological segments, genes, chemicals) combine to form larger units, their effects do not blend but, rather, remain distinct. (Accordingly, words that are composed of the same phonological segments, such as "cat," "act," and "tack," remain distinct.). In language, this in part underlies the unboundedness of the lexicon and the unboundedness of what we can use language to achieve. Although some writing about speech production suggests that, when talkers coarticulate, that is, when they temporally overlap the production of consonants and vowels in words, the result is a blending of the properties of the consonants and vowels (as in Hockett's, 1955, famous metaphor of coarticulated consonants and vowels as smashed Easter eggs), this is a mistaken understanding of coarticulation. Certainly, the acoustic speech signal at any point in time is jointly caused by the production of more than one consonant or vowel. However, the information in its structure must be about discrete consonants and vowels for the particulate principle to survive at the level of lexical knowledge.

Phonetics

Feature Systems

From phonetics we learn that consonants and vowels can be described by their featural attributes, and, when they are, some interesting cross-linguistic tendencies are revealed. Feature systems may describe consonants and vowels largely in terms of their articulatory correlates, their acoustic correlates, or both. A feature system that focuses on articulation might distinguish consonants primarily by their place and manner of articulation and by whether they are voiced or unvoiced. Consider the stop consonants in English. *Stop* is a manner class that includes oral and nasal stops. Production of these consonants involves transiently stopping the flow of air through the oral cavity. The stops of English are configured as shown.

	Bilabial	Alveolar	Velar
oral stops: voiced	b	d	g
unvoiced	p	t	k
nasal stops: voiced	m	n	N

The oral and nasal voiced stops are produced with the vocal folds of the larynx approximated (adducted); the oral voiceless stops are produced with the vocal folds apart (abducted). When the vocal folds are adducted and speakers exhale as they speak, the vocal folds cyclically open and close releasing successive puffs of air into the oral cavity. We hear a voicing buzz in consonants produced this way. When the vocal folds are abducted, air flows more or less unchecked by the larynx into the oral cavity, and we hear such consonants as unvoiced.

Compatible descriptions of vowels are in terms of height, backing, and rounding. Height refers to the height of the tongue in the oral cavity, and backing refers to whether the tongue's point of closest contact with the palate is in the back of the mouth or the front. Rounding (and unroundedness) refers to whether the lips are protruded during production of the vowel as they are, for example, in the vowel in *shoe*.

Some feature systems focus more on the acoustic realizations of the features than on the articulatory realizations. One example of such a system is that of Jakobson, Fant, and Halle (1962), who, nonetheless, also provide articulatory correlates of the features they propose. An example of a feature contrast of theirs that is more obviously captured in acoustic than articulatory terms is the feature [±grave]. Segments denoted as [+grave] are described as having acoustic energy that predominates in the lower region of the spectrum. Examples of [+grave] consonants are bilabials with extreme front articulations and uvulars with extreme back places of articulation. Consonants with intermediate places of articulation are [−grave]. Despite the possible articulatory oddity of the feature contrast [±grave], Jakobson, Fant, and Halle had reason to identify it as a meaningful contrast (see Ohala, 1996, for some reasons).

Before turning to what one can learn by describing consonants and vowels in terms of their features, consider two additional points that relate back to the stage-setting discussion above. First, many different feature systems have been proposed. Generally they are successful in describing the range of consonants and vowels in the world's languages and in capturing the nature of phonological slips of the tongue that speakers make (see section titled "Speech Errors"). Both of these

observations are relevant to a determination of how language users know the phonological forms of words. Nonetheless, there are differences among the systems that may have psychological significance. One relates back to the earlier discussion of parity. I suggested there that a parity-fostering property of languages would be a common currency in which messages are stored, formulated, sent, and received so that the phonological form of a message is preserved throughout a communicative exchange. Within the context of that discussion, a proposal that the features of consonants and vowels as language users know them are articulatory implies that the common currency is articulatory. A proposal that featural correlates are acoustic suggests that the common currency is acoustic.

A second point is that there is a proposal in the literature that the properties of consonants and vowels on which language knowledge and use depends are not featural. Rather, the phonological forms of words as we know them consist of "gestures" (e.g., Browman & Goldstein, 1990). Gestures are linguistically significant actions of the vocal tract. An example is the bilabial closing gesture that occurs when speakers of English produce /b/, /p/, or /m/. Gestures do not map 1:1 onto either phonological segments or features. For example, /p/ is produced by appropriately phasing two gestures, a bilabial constriction gesture and a devoicing gesture. Because Browman and Goldstein (1986) propose that voicing is the default state of the vocal tract producing speech, /b/ is achieved by just one gesture, bilabial constriction. As for the sequences /sp/, /st/, and /sk/, they are produced by appropriately phasing a tongue tip (alveolar) constriction gesture for /s/ and another constriction gesture for /p/, /t/, or /k/ with a single devoicing gesture that, in a sense, applies to both consonants in the sequence.

Browman and Goldstein (e.g., 1986) have proposed that words in the lexicon are specified as sequences of appropriately phased gestures (that is, as *gestural scores*). In a parity-fostering system in which these are primitives, the common currency is gestural. This is a notable shift in perspective because the theory gives primacy to public phonological forms (gestures) rather than to mental representations (features) with articulatory or acoustic correlates.

Featural Descriptions and the Sound Inventories of Languages

Featural descriptions of the sound inventories of languages have proven quite illuminating about the psychological factors that shape sound inventories. Relevant to our theme of languages' developing parity-fostering characteristics, researchers have shown that two factors, perceptual distinctiveness and articulatory simplification (Lindblom, 1990), are major factors shaping the consonants and vowels that languages use to form words. Perceptual distinctiveness is particularly important in shaping vowel inventories. Consider two examples.

One is that, as noted earlier, vowels may be rounded (with protruded lips) or unrounded. In Maddieson's (1984) survey of languages, 6% of front vowels were rounded, whereas 93.5% of back vowels were rounded. The evident reason for the correlation between backing and rounding is perceptual distinctiveness. Back vowels are produced with the tongue's constriction location toward the back of the oral cavity. This makes the cavity in front of the constriction very long. Rounding the lips makes it even longer. Front vowels are produced with the tongue constriction toward the front of the oral cavity so that the cavity in front of the constriction is short. An acoustic consequence of backing/fronting is the frequency of the vowel's second formant (i.e., the resonance associated with the acoustic signal for the vowel that is second lowest in frequency [F2]). F2 is low for back vowels and high for front vowels. Rounding back vowels lowers their F2 even more, enhancing the acoustic distinction between front and back vowels (e.g., Diehl & Kluender, 1989; Kluender, 1994).

A second example also derives from the study of vowel inventories. The most frequently occurring vowels in Maddieson's (1984) survey were /i/ (a high front unrounded vowel as in *heat*), /a/ (a low central unrounded vowel as in *hot*) and /u/ (a high back rounded vowel as in *hoot*), occurring in 83.9% (/u/) to 91.5% (/i/) of the language sample. Moreover, of the 18 languages in the survey that have just three vowels, 10 have those three vowels. Remarkably, most of the remaining 8 languages have minor variations on the same theme. Notice that these vowels, sometimes called the point vowels, form a triangle in vowel space if the horizontal dimension represents front-to-back and the vertical dimension vowel height:

i u

a

Accordingly, they are as distinct as they can be articulatorily and acoustically. Lindblom (1986) has shown that a principle of perceptual distinctiveness accurately predicts the location of vowels in languages with more than three vowels. For example, it accurately predicts the position of the fourth and fifth vowels of five-vowel inventories, the modal vowel inventory size in Maddieson's survey.

Consonants do not directly reflect a principle of perceptual dispersion as the foregoing configuration of English stop consonants suggests. Very tidy patterns of consonants in voicing, manner, and place space are common, yet such patterns mean that phonetic space is being densely packed. An important consideration for consonants appears to be

articulatory complexity. Lindblom and Maddieson (1988) classified consonants of the languages of the world into basic, elaborated, and complex categories according to the complexity of the articulatory actions required to produce them. They found that languages with small consonant inventories tend to restrict themselves to basic consonants. Further, languages with elaborated consonants always have basic consonants as well. Likewise, languages with complex consonants (for example, the click consonants of some languages of Africa) always also have both basic and elaborated consonants as well. In short, language communities prefer consonants that are easy to produce.

Does the foregoing set of observations mean that language communities value perceptual distinctiveness in vowels but articulatory simplicity in consonants? This is not likely. Lindblom (1990) suggests that the proper concept for understanding popular inventories both of vowels and of consonants is that of "sufficient contrast." Sufficient contrast is the equilibrium point in a tug-of-war between goals of perceptual distinctiveness and articulatory simplicity. The balance shifts toward perceptual distinctiveness in the case of vowel systems, probably because vowels are generally fairly simple articulatorily. Consonants vary more in that dimension, and the balance point shifts accordingly.

The major global observation here, however, is that the requirements of efficacious public language use clearly shape the sound inventories of language. Achievement of parity matters.

Features and Contrast: Onward to Phonology

An important concept in discussions of feature systems is contrast. A given consonant or vowel can, in principle, be exhaustively described by its featural attributes. However, only some of those attributes are used by a language community to distinguish words. For example, in the English *till,* the first consonant is /t/, an unvoiced, alveolar stop. It is also "aspirated" in that there is a longish unvoiced and breathy interval from the time that the alveolar constriction by the tongue tip is released until voicing for the following vowel begins. The /t/ in *still* is also an unvoiced, alveolar stop, but it is unaspirated. This is because, in the sequence /st/, although both the /s/ and the /t/ are unvoiced, there is just one devoicing gesture for the two segments, and it is phased earlier with respect to the tongue constriction gesture for /t/ than it is phased in *till.* Whereas a change in any of the voicing, manner, or place features can create a new word of English (voicing: *dill;* manner: *sill;* place: *pill*), a change in aspiration does not. Indeed, aspiration will vary due to rate of speaking and emphasis, but the /t/ in *till* will remain a /t/.

Making a distinction between contrastive and noncontrastive features historically allowed a distinction to be made also in how consonants and vowels were characterized. Characterizing them as *phonological segments* (or phonemes) involved specifying only their contrastive features. Characterizing them as *phonetic segments* (or phones) involved specifying fairly exactly how they were to be pronounced. To a first approximation, the contrastive/noncontrastive distinction evolved into another relating to predictability that has had a significant impact on how modern phonologists have characterized lexical word forms. Minimally, lexical word forms have to specify unpredictable features of words. These are approximately contrastive features. That is, that the word meaning "medicine in a small rounded mass to be swallowed whole" (Mish, 1990) is *pill,* not, say, *till,* is just a fact about English language use. It is not predictable from any general phonological or phonetic properties of English. Language users have to know the sequence of phonological segments that compose the word. However, the fact that the /p/ is aspirated is predictable. Stressed-syllable initial unvoiced stops are aspirated in English. An issue for phonologists has been whether lexical word forms are abstract, specifying only unpredictable features (and so giving rise to differences between lexical and pronounced forms of words), or whether they are fully specified.

The mapping of contrastive/noncontrastive onto predictable/unpredictable is not exact. In context, some contrastive features of words can be predictable. For example, if a consonant of English is labiodental (i.e., produced with teeth against lower lip as in /f/ or /v/), it must be a fricative. And if a word begins /skr/, the next segment must be [+vocalic]. An issue in phonology has been to determine what should count as predictable and lexically unspecified. Deciding that determines how abstract in relation to their pronounced forms lexical entries are proposed to be.

Phonology

Most phonologists argue that lexical forms must be abstract with respect to their pronunciations. One reason that has loomed large in only one phonology (Browman & Goldstein's, e.g., 1986, Articulatory Phonology) is that we do not pronounce the same word the same way on all occasions. Particularly, variations in speaking style (e.g., from formal to casual) can affect how a word is pronounced. Lexical forms, it seems (but see section titled "Another Abstractness Issue"), have to be abstracted away from detail that distinguishes those different pronunciations. A second reason given for abstract word forms is, as noted above, that some properties of word forms are predictable. Some linguists have argued that

lexical entries should include just what is phonologically unpredictable about a word. Predictable properties can be filled in another way, by rule application, for example. A final reason that words in the lexicon may be phonologically abstract is that the same morpheme may be pronounced differently in different words. For example, the prefixes on *inelegant* and *imprecise* are etymologically the same prefix, but the alveolar /n/ becomes labial /m/ before labial /p/ in *imprecise*. To capture in the lexicon that the morpheme is the same in the two words, some phonologists have proposed that they be given a common form there.

An early theory of phonology that focused on the second and third reasons was Chomsky and Halle's (1968) generative phonology. An aim there was to provide in the lexicon only the unpredictable phonological properties of words and to generate surface pronunciations by applying rules that provided the predictable properties. In this phonology, the threshold was rather low for identifying properties as predictable, and underlying forms were highly abstract.

A recent theory of phonology that appears to have superseded generative phonology and its descendents is optimality theory, first developed by Prince and Smolensky (1993). This theory accepts the idea that lexical forms and spoken forms are different, but it differs markedly from generative phonology in how it gets from the one to the other.

In optimality theory, there are no rules mediating lexical and surface forms. Rather, from a lexical form, a large number of candidate surface forms are generated. These are evaluated relative to a set of universal constraints. The constraints are ranked in language-particular ways, and they are violable. The surface form that emerges from the competition is the one that violates the fewest and the lowest ranked constraints. One kind of constraint that limits the abstractness of underlying forms is called a *faithfulness constraint*. One of these specifies that lexical and surface forms must be the same. (More precisely, every segment or feature in the lexical entry must have an identical correspondent in the surface form, and vice versa.) This constraint is violated in *imprecise*, the lexical form of which will have an /n/ in place of the /m/. A second constraint (the identical cluster constraint in Pulleyblank, 1997) requires that consonant clusters share place of articulation. It is responsible for the surface /m/.

On the surface, this model is not plausible as a psychological one. That is, no one supposes that, given a word to say, the speaker generates lots of possible surface forms and then evaluates them and ends up saying the optimal one. But there are models that have this flavor and are considered to have psychological plausibility. These are network models. In those models (e.g., van Orden, Pennington, & Stone, 1990), something input to the network (say, a written word) activates far more in the phonological component of the model than just the word's pronunciation. Research suggests that this happens in humans as well (e.g., Stone, Vanhoy, & Van Orden, 1997). The activation then settles into a state reflecting the optimal output, that is, the word's actual pronunciation. From this perspective, optimality theory may be a candidate psychological model of the lexicon.

Another theory of phonology, articulatory phonology (Browman & Goldstein, 1986), is markedly different from both of those described above. It does not argue from predictability or from a need to preserve a common form for the same morpheme in the lexicon that lexical entries are abstract. Indeed, in the theory, they are not very abstract. As noted earlier, primitive phonological forms in the theory are gestures. Lexical entries specify gestural scores. The lexical entries are abstract only with respect to variation due to speaking style. An attractive feature of their theory, as Browman and Goldstein (1995a) comment, is that phonology and phonetics are respectively macroscopic and microscopic descriptions of the same system. In contrast to this, in most accounts, phonology is an abstract, cognitive representation, whereas phonetics is its physical implementation. In an account of language production incorporating articulatory phonology, therefore, there need be no (quite mysterious) translation from a mental to a physical domain (cf. Fowler, Rubin, Remez, & Turvey, 1980); rather, the same domain is at once physical and cognitive (cf. Ryle, 1949). Articulatory phonology is a candidate for a psychological model.

Another Abstractness Issue: Exemplar Theories of the Lexicon

Psychologists have recently focused on a different aspect of the abstractness issue. The assumption has been until recently that language users store word types, not word tokens, in the lexicon. That is, even though listeners may have heard the word *boy* a few million times, they have not stored memories of those few million occurrences. Rather, listeners have just one word *boy* in their lexicon.

In recent years, this idea has been questioned, and some evidence has accrued in favor of a token or exemplar memory (see chapter by Goldstone & Kersten in this volume). The idea comes from theories of memory in cognitive psychology. Clearly, not all of memory is a type memory. We can recall particular events in our lives. Some researchers have suggested that exemplar memory systems may be quite pervasive. An example theory that has drawn the attention of

speech researchers is Hintzman's (e.g., 1986) memory model, MINERVA. In the model, input is stored as a trace, which consists of feature values along an array of dimensions. When an input is presented to the model, it not only lays down its own trace, but it activates existing traces to the extent that they are featurally similar to it. The set of activated traces forms a composite, called the echo, which bears great resemblance to a type (often called a prototype in this literature). Accordingly, the model can behave as if it stores types when it does not.

In the speech literature, researchers have tested for an exemplar lexicon by asking whether listeners show evidence of retaining information idiosyncratic to particular occurrences of words, typically, the voice characteristics of the speaker. Goldinger (1996) provided an interesting test in which listeners identified words in noise. The words were spoken in 2, 6, or 10 different voices. In a second half of the test (after a delay that varied across subjects), he presented some words that had occurred in the first half of the test. The tokens in the second half were produced by the same speaker who produced them in the first half (and typically they were the same token) or were productions by a different speaker. The general finding was that performance identifying words was better if the words were repeated by the speaker who had produced them in the first half of the test. This across–test-half priming persisted across delays between test halves as long as one week. This study shows that listeners retain token-level memories of words (see also Goldinger, 1998). Does it show that these token-level memories constitute word forms in the mental lexicon? Not definitively. However, it is now incumbent on theorists who retain the claim that the lexicon is a type memory to provide distinctively positive evidence for it.

PHONOLOGICAL PLANNING

Speakers are creators of linguistic messages, and creation requires planning. This is in part because utterances are syntactically structured so that the meaning of a sentence is different from the summed meanings of its component words. Syntactic structure can link words that are distant in a sentence. Accordingly, producing a syntactically structured utterance that conveys an intended message requires planning units larger than a word. Planning may also be required to get the phonetic, including the prosodic, form of an utterance right.

For many years, the primary source of evidence about planning for language production was the occurrence of spontaneous errors of speech production. In approximately the last decade other, experimentally generated, behavioral evidence has augmented that information source.

Speech Errors

Speakers sometimes make mistakes that they recognize as errors and are capable of correcting. For example, intending to say *This seat has a spring in it,* a speaker said *This spring has a seat in it* (Garrett, 1980), exchanging two nouns in the intended utterance. Or intending to say *It's the jolly green giant,* a speaker said *It's the golly green giant* (Garrett, 1980), anticipating the /g/ from *green*. In error corpora that researchers have collected (e.g., Dell, 1986; Fromkin, 1973; Garrett, 1980; Shattuck-Hufnagel, 1979), errors are remarkably systematic and, apparently, informative about planning for speech production.

One kind of information provided by these error corpora concerns the nature of planning units. Happily, they appear to be units that linguists have identified as linguistically coherent elements of languages. However, they do not include every kind of unit identified as significant in linguistic theory. In the two examples above, errors occurred on whole words and on phonological segments. Errors involving these units are common, as are errors involving individual morphemes (e.g., *point outed;* Garrett, 1980). In contrast, syllable errors are rare and so are feature errors (as in Fromkin's, 1973, *glear plue sky*). Rime (that is, the vowel and any postvocalic consonants of a syllable) errors occur, but consonant-vowel (CV) errors are rare (Shattuck-Hufnagel, 1983). This is not to say that syllables and features are irrelevant in speech planning. They are relevant, but in a different way from words and phonemes.

Not only are the units that participate in errors tidy, but the kinds of errors that occur are systematic too. In the word error above, quite remarkably, two words exchanged places. Sometimes, instead, one word is anticipated, but it also occurs in its intended slot (*This spring has a spring in it*) or a word is perseverated (*This seat has a seat in it*). Sometimes, noncontextual substitutions occur in which a word appears that the speaker did not intend to say at all (*This sheep has a spring in it*). Additions and deletions occur as well. To a close approximation, the same kinds of errors occur on words and phonological segments.

Errors have properties that have allowed inferences to be drawn about planning for speech production. Words exchange, anticipate, and perseverate over longer distances than do phonological segments. Moreover, word substitutions appear to occur in two varieties: semantic (e.g., saying

summer when meaning to say *winter*) and form-based (saying *equivocal when* meaning to say *equivalent*). These observations suggested to Garrett (1980) that two broad phases of planning occur. At a functional level, lemmas (that is, words as semantic and syntactic entities) are slotted into a phrasal structure. When movement errors occur, lemmas might be put into the wrong phrasal slot, but because their syntactic form class determines the slots they are eligible for, when words anticipate, perseverate, or exchange, they are members of the same syntactic category. Semantic substitution errors occur when a semantic neighbor of an intended word is mistakenly selected. At a positional level, planning concerns word forms rather than their meanings. This is where sound-based word substitutions may occur.

For their part, phonological segment errors also have highly systematic properties. They are not sensitive, as word movement errors are, to the syntactic form class of the words involved in the errors. Rather, they are sensitive to phonological variables. Intended and erroneous segments in errors tend to be featurally similar, and their intended and actual slots are similar in two ways. They tend to have featurally similar segments surrounding them, and they come from the same syllable position. That is, onset (prevocalic) consonants move to other onset positions, and codas (postvocalic consonants) move to coda positions.

These observations led theorists (e.g., Dell, 1986; Shattuck-Hufnagel, 1979) to propose that, in phonological planning, the phonemes that compose words to be said are slotted into syllabic frames. Onsets exchange with onsets, because, when an onset position is to be filled, only onset consonants are candidates for that slot. There is something intuitively displeasing about this idea, but there is evidence for it, theorists have offered justifications for it, and there is at least one failed attempt to avoid proposing a frame (Dell, Juliano, & Govindjee, 1993). The idea of slotting the phones of a word into a structural frame is displeasing, because it provides the opportunity for speakers to make errors, but seems to accomplish little else. The phones of words must be serially ordered in the lexical entry. Why reselect and reorder them in the frame? One justification has to do with productivity (e.g., Dell, 1986; Dell, Burger, & Svec, 1997). The linguistic units that most frequently participate in movement errors are those that we use productively. That is, words move, and we create novel sentences by selecting words and ordering them in new ways. Morphemes move, and we coin some words (e.g., *videocassette*) by putting morphemes together into novel combinations. Phonemes move, and we coin words by selecting consonants and vowels and ordering them in new ways (e.g., *smurf*). The frames for sentences (that is, syntactic structure) and for syllables permit the coining of novel sentences and words that fit the language's constraints on possible sentences and possible words.

Dell et al. (1993; see also Dell & Juliano, 1996) developed a parallel-distributed network model that allowed accurate sequences of phones to be produced without a frame-content distinction. The model nonetheless produced errors hitherto identified as evidence for a frame. (For example, errors were phonotactically legal the vast majority of the time, and consonants substituted for consonants and vowels for vowels.) However, the model did not produce anticipations, perseverations, or exchanges, and, even with modifications that would give rise to anticipations and perseverations, it would not make exchange errors. So far, theories and models that make the frame-content distinction have the edge over any that lack it.

Dell (1986) more or less accepted Garrett's (1980) two-tiered system for speech planning. However, he proposed that the lexical system in which planning occurs has both feedforward (word to morpheme to syllable constituent to phone) links and feedback links, with activation of planned lexical units spreading bidirectionally. The basis for this idea was a set of findings in speech error corpora. One is that, although phonological errors do create nonwords, they create words at a greater than chance rate. Moreover, in experimental settings, meaning variables can affect phonological error rates (see, e.g., Motley, 1980). Accordingly, when planning occurs at the positional level, word meanings are not irrelevant, as Garrett had supposed. The feedforward links in Dell's network provide the basis for this influence. A second finding is that semantic substitutions (e.g., the *summer/winter* error above) tend to be phonologically more related than are randomly re-paired intended and error words. This implies activation that spreads along feedback links.

In the last decade, researchers developed new ways to study phonological planning. One reason for these developments is concern about the representativeness of error corpora. Error collectors can only transcribe errors that they hear. They may fail to hear errors or mistranscribe them for a variety of reasons. Some errors occur that are inaudible. This has been shown by Mowrey and MacKay (1990), who measured activity in muscles of the vocal tract as speakers produced tongue twisters (e.g., *Bob flew by Bligh Bay*). In some utterances, Mowrey and MacKay observed tongue muscle activity for /l/ during production of *Bay* even though the word sounded error free to listeners. The findings show that errors occur that transcribers will miss. Mowrey and MacKay also suggest that their data show that subphonemic errors occur, in particular, in activation of single muscles. This conclusion is not yet warranted by their data, because other, unmonitored

muscles for production of an intruding phoneme might also have been active. However, it is also possible that errors may appear to the listener tidier than they are.

We know, too, that listeners tend to "fluently restore" (Marslen-Wilson & Welsh, 1978) speech errors. They may not hear errors that are, in principle, audible, because they are focusing on the content of the speaker's utterance, not its form. These are not reasons to ignore the literature on speech errors; it has provided much very useful information. However, it is a reason to look for converging measures, and that is the next topic.

Experimental Evidence About Phonological Planning

Some of the experimental evidence on phonological planning has been obtained from procedures that induce speech errors (e.g., Baars, Motley, & MacKay, 1975; Dell, 1986). Here, however, the focus is on findings from other procedures in which production response latencies constitute the main dependent measure.

This research, pioneered by investigators at the Max Planck Institute for Psycholinguistics in the Netherlands, has led to a theory of lexical access in speech production (Levelt, Roelofs, & Meyer, 1999) that will serve to organize presentation of relevant research findings. The theory has been partially implemented as a computational model, WEAVER (e.g., Roelofs & Meyer, 1998). However, I will focus on the theory itself. It begins by representing the concepts that a speaker might choose to talk about, and it describes processes that achieve selection of relevant linguistic units and ultimately speech motor programs. Discussion here is restricted to events beginning with word form selection.

In the theory, selection of a word form provides access to the word's component phonological segments, which are abstract, featurally underspecified segments (see section titled "Features and Contrast: Onward to Phonology"). If the word does not have the default stress pattern (with stress on the syllable with the first full vowel for both Dutch and English speakers), planners also access a metrical frame, which specifies the word's number of syllables and its stress pattern. For words with the default pattern, the metrical frame is constructed online. In this theory, as in Dell's, the segments are types, not tokens, so that the /t/ in *touch* is the very /t/ in *tiny*. This allows for the possibility of form priming. That is, preparing to say a word that shares its initial consonant with a prime word can facilitate latency to produce the target word. In contrast to Dell's (1986) model, however, consonants are not exclusively designated either onset consonants or coda consonants. That is, the /t/ in *touch* is also the very /t/ in *date*.

Accessed phonological segments are spelled out into phonological word frames. This reflects an association of the phonological segments of a word with the metrical frame, if there is an explicit one in the lexical entry, or with a frame computed on line. This process, called prosodification, is proposed to be sequential; that is, segments are slotted into the frame in an early-to-late (left-to-right) order.

Meyer and Shriefers (1991) found evidence of form priming and a left-to-right process in a picture-naming task. In one experiment, at some stimulus onset asynchrony (SOA) before or after presentation of a picture, participants heard a monosyllabic word that overlapped with the monosyllabic picture name at the beginning (the initial CV), at the end (the VC), or not at all. On end-related trials, the SOA between word and picture was adjusted so that the VC's temporal relation to the picture was the same as that of the CV of begin-related words. On some trials no priming word was presented. The priming stimulus generally slowed responses to the picture, but, at some SOAs, it did so less if it was related to the target. For words that overlapped with the picture name in the initial CV, the response time advantage (over response times to pictures presented with unrelated primes) was significant when words were presented 150 ms before the pictures (but not 300 ms before) and continued through the longest lagging SOA tested, when words were presented 150 ms after the picture. For words overlapping with the picture name in the final VC, priming began to have an effect at 0 ms SOA and continued through the 150-ms lag condition. The investigators infer that priming occurs during phonological encoding, that is, as speakers access the phonological segments of the picture name. Perhaps at a 300-ms lead the activations of phonological segments shared between prime and picture name have decayed by the time the picture is processed. However, by a 150-ms lead, the prime facilitates naming the picture, because phonemes activated by its presentation are still active and appropriate to the picture. The finding that end-related primes begin facilitating later than begin-related items, even though the overlapping phonemes in the prime bore the same temporal relation to the picture's presentation as did the overlapping CVs or initial syllables, suggests an early-to-late process.

Using another procedure, Meyer (1990, 1991) also found form priming and evidence of a left-to-right process. Meyer (1990) had participants learn word pairs. Then, prompted by the first word of the pair, they produced the second. In homogeneous sets of word pairs, disyllabic response words of each pair shared either their first or their second syllable. In heterogeneous sets, response words were unrelated. The question was whether, across productions of response words in homogeneous sets, latencies would be faster than to response

words in heterogeneous sets, because segments in the overlapping syllables would remain prepared for production. Meyer found shorter response latencies only in the homogeneous sets in which the first syllable was shared across response words. In a follow-up study, Meyer (1991) showed savings when word onsets were shared but not when rimes were shared. On the one hand, these studies provide evidence converging with that of Meyer and Shriefers (1991) for form priming and left-to-right preparation. However, the evidence appears to conflict in that Meyer (1990, 1991) found no end-overlap priming, whereas Meyer and Shriefers did. Levelt et al. (1999) suggested, as a resolution, that the latter results occur as the segments of a lexical item are activated, whereas the results of Meyer reflect prosodification (that is, merging of those segments with the metrical frame).

The theory of Levelt et al. (1999) makes a variety of predictions about the prosodification process. First, the phonological segments and the metrical frame are retrieved as separate entities. Second, the metrical frame specifies only the number of syllables in the word and the word's stress pattern; it does not specify the CV pattern of the syllables. Third, for words with the default stress pattern, no metrical frame is retrieved; rather, it is computed online.

Roelofs and Meyer (1998) tested these predictions using the implicit priming procedure. In the first experiment, in homogeneous sets, response words were disyllables with second-syllable stress that shared their first syllables; heterogeneous sets had unrelated first syllables. Alternatively, homogeneous (same first syllables) and heterogeneous (unrelated first syllables) response words had a variable number of syllables (2–4) with second-syllable stress. None of the words in this and the following experiments had the default stress pattern, so that, according to the theory, a metrical frame had to be retrieved. Priming (that is, an advantage in response latency for the homogeneous as compared to the heterogeneous sets) occurred only if the number of syllables was the same across response words. This is consistent with the prediction that the metrical frame specifies the number of syllables. A second experiment confirmed that, with the number of syllables per response word held constant, the stress pattern had to be shared for priming to occur. A third experiment tested the prediction that shared CV structure did not increase priming. In this experiment, response words were monosyllables that, in homogeneous sets, shared their initial consonant clusters (e.g., *br*). In one kind of homogeneous set, the words shared their CV structure (e.g., all were CCVCs); in another kind of homogeneous set, they had different CV structures. The two homogeneous sets produced equivalent priming relative to latencies to produce heterogeneous responses. This is consistent with the claim of the theory that the metrical frame only specifies the number of syllables, but not the CV structure of each syllable. Subsequent experiments showed that shared number of syllables with no segmental overlap and shared stress pattern without segmental overlap give rise to no priming. Accordingly, it is the integration of the word's phonological segments with the metrical frame that underlies the priming effect.

Finally, in a study by Meyer, Roelofs, and Schiller, described by Levelt et al. (1999), Meyer et al. examined words with the default stress pattern for Dutch. In this case, no metrical frame should be retrieved and so none can be shared across response words. Meyer et al. found that for words that shared their initial CVs and that had the default stress pattern for Dutch, shared metrical structure did not increase priming.

The next process in the theory is phonetic encoding in which talkers establish a gestural score (see section titled "Feature Systems") for each phonological word. This phase of talking is not well worked out by Levelt et al. (1999), and it is the topic of the next major section ("Speech Production"). Accordingly, I will not consider it further here.

Disagreements Between the Theories of Dell, 1986, and Levelt et al., 1999

Two salient differences between the theory of Dell (1986), developed largely from speech error data, and that of Levelt et al. (1999), developed largely from speeded naming data, concern feedback and syllabification. Dell's model includes feedback. The theory of Levelt et al. and Roelof and Meyer's (1998) model WEAVER do not. In Dell's model, phones are slotted into a syllable frame, whereas in the theory of Levelt et al., they are slotted into a metrical frame that specifies the number of syllables, but not their internal structure.

As for the disagreement about feedback, the crucial error data supporting feedback consist of such errors as saying *winter* for *summer,* in which the target and the error word share both form and meaning. In Dell's (1986) model, form can affect activation of lexical items via feedback links in the network. Levelt et al. (1999) suggest that these errors are monitoring failures. Speakers monitor their speech, and they often correct their errors. Levelt et al. suggest that the more phonologically similar the target and error words are, the more likely the monitor is to fail to detect the error.

The second disagreement is about when during planning phonological segments are syllabified. In Dell's (1986) model, phones are identified with syllable positions in the lexicon, and they are slotted into abstract syllable frames in the course of planning for production. In the theory of Levelt et al. (1999), syllabification is a late process, as it has to be to allow resyllabification to occur. There is evidence favoring

both sides. As described earlier, Roelofs and Meyer (1998) reported that implicit priming occurs across response words that share stress pattern, number of syllables, and phones at the beginning of the word, but shared syllable structure does not increase priming further. Sevald, Dell, and Cole (1995) report apparently discrepant findings. Their task was to have speakers produce a pair of nonwords repeatedly as quickly as possible in a 4-s interval. They measured mean syllable production time and found a 30-ms savings if the nonwords shared the initial syllable. For example, the mean syllable production time for KIL KIL.PER (where the "." signals the syllable boundary) was shorter than for KILP KIL.PER or KIL KILP.NER. Remarkably, they also found shorter production times when only syllable structure was shared (e.g., KEM TIL.PER). These findings show that, at whatever stage of planning this effect occurs, syllable structure matters, and an abstract syllable frame is involved. This disagreement, like the first, remains unresolved (see also Santiago & MacKay, 1999).

SPEECH PRODUCTION

Communication by language use requires that speakers act in ways that count as linguistic. What are the public events that count as linguistic? There are two general points of view. The more common one is that speakers control their actions, their movements, or their muscle activity. This viewpoint is in common with most accounts of control over voluntary activity (see chapter by Heuer in this volume). A less common view, however, is that speakers control the acoustic signals that they produce. A special characteristic of public linguistic events is that they are communicative. Speech activity causes an acoustic signal that listeners use to determine a talker's message.

As the next major section ("Speech Perception") will reveal, there are also two general views about immediate objects of speech perception. Here the more common view is that they are acoustic. That is, after all, what stimulates the perceiver's auditory perceptual system. A less common view, however, is that they are articulatory or gestural.

An irony is that the most common type of theory of production and the most common type of theory of perception do not fit together. They have the joint members of communicative events producing actions, but perceiving acoustic structure. This is unlikely to be the case. Communication requires prototypical achievement of parity, and parity is more likely to be achieved if listeners perceive what talkers produce. In this section, I will present instances of both types of production theory, and in the next section, both types of perception theory. The reader should keep in mind that considerations of parity suggest that the theories should be linked. If talkers aim to produce particular acoustic patternings, then acoustic patterns should be immediate perceptual objects. However, if talkers aim to produce particular gestures, then that is what listeners should perceive.

How Acoustic Speech Signals Are Produced

Figure 9.1 shows the vocal tract, the larynx, and the respiratory system. Articulators of the vocal tract include the jaw, the tongue (with relatively independent control of the tip or blade and the tongue body), the lips, and the velum. Also involved in speech is the larynx, which houses the vocal folds, and the lungs. In prototypical production of speech, acoustic energy is generated at a source, in the larynx or oral cavity. In production of vowels and voiced consonants, the vocal folds are adducted. Air flow from the lungs builds up pressure beneath the folds, which are blown apart briefly and then close again. This cycling occurs at a rapid rate during voiced speech. The pulses of air that escape whenever the folds are blown apart are filtered by the oral cavity. Vowels are produced by particular configurations of the oral cavity achieved by positioning the tongue body toward the front (e.g., for /i/) or back (e.g., for /a/) of the oral cavity, close to the palate (e.g., /i/, /u/) or farther away (e.g., /a/), with lips rounded (/u/) or not. In production of stop consonants, there is a complete

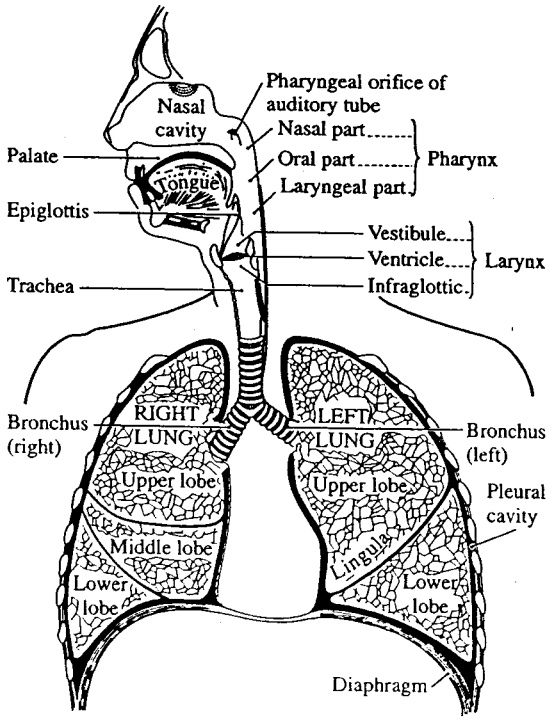

Figure 9.1 The speech sound producing system (from Borden, Harris, & Raphael, 1994). Reprinted with permission.

stoppage of airflow through the oral cavity for some time due to a constriction that, in English, occurs at the lips (/b/, /p/, /m/), with the tongue tip against the alveolar ridge of the palate (/d/, /t/, /n/) or with the tongue body against the velum (/g/, /k/, /ŋ/). For the nasal consonants, /m/, /n/, and /ŋ/, the velum is lowered, allowing airflow through the nose. For fricatives, the constriction is not complete, so that airflow is not stopped, but the constriction is sufficiently narrow to cause turbulent, noisy airflow. This occurs in English, for example, in /s/, /f/, and /θ/ (the initial consonant of, e.g., *theta*). Consonants of English can be voiced (vocal folds adducted) or unvoiced (vocal folds abducted).

The acoustic patterning caused by speech production bears a complex relation to the movements that generate it. In many instances the relation is nonlinear, so that, for example, a small movement may generate a marked change in the sound pattern (as, for example, when the narrow constriction for /s/ becomes the complete constriction for /t/). In other instances, a fairly large change in vocal tract configuration can change the acoustic signal rather little. Stevens (e.g., 1989) calls these "quantal regions," and he points out that language communities exploit them, for example, to reduce the requirement for extreme articulatory precision.

Some Properties of Speech That a Production Theory Needs to Explain

Like all intentional biological actions, speaking is coordinated action. Absent coordination, as Weiss (1941) noted, activity would consist of "unorganized convulsions." What is coordination? It is (cf. Turvey, 1990) a reduction in the degrees of freedom of an organism with a consequent reduction in its dimensionality. This reduces the outputs the system can produce, restricting them to the subset of outcomes consistent with the organism's intentions. Although it is not (wholly) biological, I like to illustrate this idea using the automobile. Cars have axles between the front wheels so that, when the driver turns the steering wheel, both front wheels are constrained to turn together. The axle reduces the degrees of freedom of movement of the car-human system, preventing movements in which the car's front wheels move independently, and it lowers the dimensionality of the system by linking the wheels. However, the reduction in power is just what the driver wants; that is, the driver only wants movements in which the wheels turn cooperatively.

The lowering of the dimensionality of the system creates macroscopic order consistent with an actor's intentions; that is, it creates a special purpose device. In the domain of action, these special purpose devices are sometimes called "coordinative structures" (Easton, 1972) or synergies. In the vocal tract, they are linkages among articulators that achieve coordinated action. An example is a transient linkage between the jaw and two lips that achieves lip closure for /b/, /p/, and /m/ in English.

An important characteristic of synergies is that they give rise to motor equivalence: that is, the ability to achieve the same goal (e.g., lip closure in the example above), in a variety of ways. Speakers with a bite block held between their teeth to immobilize the jaw (at a degree of opening too wide for normal production of /i/, for example, or too closed for normal production of /a/) produce vowels that are near normal from the first pitch pulse of the first vowel they produce (e.g., Lindblom, Lubker, & Gay, 1979). An even more striking finding is that speakers immediately compensate for on-line articulatory perturbations (e.g., Abbs & Gracco, 1984; Kelso, Tuller, Vatikiotis-Bateson, & Fowler, 1984; Shaiman, 1989). For example, in research by Kelso et al. (1984), on an unpredictable 20% of trials, a jaw puller pulled down the jaw of a speaker producing *It's a bab again* as the speaker was closing his lips for the final /b/ of *bab*. Within 20–30 ms of the perturbation, extra activity of an upper lip muscle (compared to its activity on unperturbed trials) occurred, and closure for /b/ was achieved. When the utterance was *It's a baz again,* jaw pulling caused extra activity in a muscle of the tongue, and the appropriate constriction was achieved. These responses to perturbation are fast and functional (cf. Löfquist, 1997).

These immediate and effective compensations contrast with others. When Savariaux, Perrier, and Orliaguet (1995) had talkers produce /u/ with a lip tube that prevented rounding, tongue backing could compensate for some acoustic consequences of the lip tube. Of 11 participants in the study, however, 4 showed no compensation at all (in about 20 attempts); 6 showed a little, but not enough to produce a normal acoustic signal for /u/; just 1 achieved full compensation. Similarly, in research by Hamlet and Stone (e.g., 1978; Hamlet, 1988), after one week's experience, speakers failed to compensate fully for an artificial palate that changed the morphology of their vocal tract. What is the difference between the two sets of studies that explains the differential success of compensation? Fowler and Saltzman (1993) suggest that the bite block and on-line perturbation studies may use perturbations that approximately occur in nature, whereas the lip tube and the artificial palate do not. That is, competing demands may be placed on the jaw because gestures overlap in time. For example, the lip-closing gesture for /b/ may overlap with the gestures for an open vowel. The vowel may pull down the jaw so that it occupies a more open position for /b/ than it does when /b/ gestures overlap with those for the high vowel /i/. Responses to the bite block and

to on-line perturbations of the jaw may be immediate and effective because talkers develop flexible synergies for producing vowels with a range of possible openings of the jaw and consonants with a range of jaw closings. However, nothing prevents lip protrusion in nature, and nothing changes the morphology of the vocal tract. Accordingly, synergies to compensate for those perturbations do not develop.

Indeed, gestural overlap (that is, coarticulation) is a pervasive characteristic of speech and therefore is a characteristic that speakers need to learn both to achieve and to compensate for. Coarticulation is a property of action that can only occur when discrete actions are sequenced. Coarticulation has been described in a variety of ways: as spreading of features from one segment to another (as when rounding of the lips from /u/ occurs from the beginning of a word such as *strew*) or as assimilation. However, most transparently, when articulatory activity is tracked, coarticulation is a temporal overlap of articulatory activity for neighboring consonants and vowels. Overlap occurs both in an anticipatory (right-to-left) and a carryover (perseveratory, left-to-right) direction. This characterization in terms of gestural overlap is sometimes called *coproduction*. Its span can be segmentally extensive as when vowel-to-vowel coarticulation occurs over intervening consonants (e.g., Fowler & Brancazio, 2000; Öhman, 1966; Recasens, 1984). However, it is not temporally very extensive, spanning perhaps no more than about 250 ms (cf. Fowler & Saltzman, 1993). According to the frame theory of coarticulation (e.g., Bell-Berti & Harris, 1981), in anticipatory coarticulation of such gestures as lip rounding for a rounded vowel (e.g., Boyce, Krakow, Bell-Berti, & Gelfer, 1990) or nasalization for a nasalized consonant (e.g., Bell-Berti & Krakow, 1991; Boyce et al., 1990) the anticipating gesture is not linked to the gestures for other segments with which it overlaps in time; rather, it remains tied to other gestures for the segment, which it anticipates by an invariant interval.

An interesting constraint on coarticulation is coarticulation resistance (Bladon & Al-Bamerni, 1976). This reflects the differential extent to which consonants or vowels resist coarticulatory encroachment by other segments. Recasens's research (e.g., 1984) suggests that resistance to vowels among consonants varies with the extent to which the consonants make use of the tongue body, also required for producing vowels. Accordingly, a consonant such as /b/ that is produced with the lips is less resistant than one such as /d/, which uses the tongue (cf. Fowler & Brancazio, 2000). An index of coarticulation resistance is the slope of the straight-line relation between F2 at vowel midpoint of a CV and F2 at syllable onset for CVs in which the vowel varies but the consonant is fixed (see many papers by Sussman, e.g., Sussman, Fruchter, Hilbert, & Sorish, 1999a). Figure 9.2 shows data from Fowler (1994).

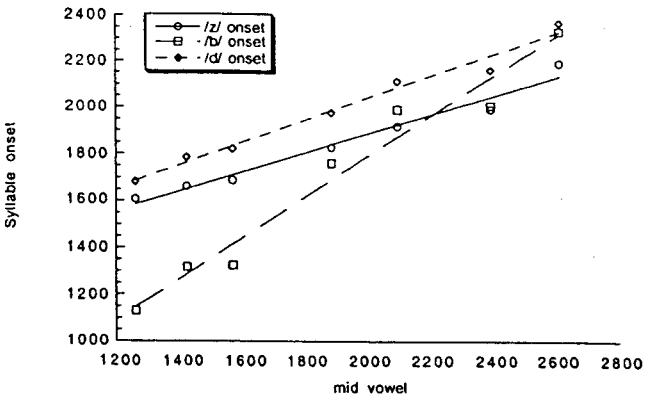

Figure 9.2 Data from Fowler (1994). Plots for /b/, /d/ and /z/ of F2 at vowel midpoint by F2 at syllable onset.

The low resistant consonant /b/ has a high slope, indicating considerable coarticulatory effect of the vowel on /b/'s acoustic manifestations at release; the slope for /d/ is much shallower; that for /z/ is slightly shallower than that for /d/. Fowler (1999) argues that the straight-line relation occurs because a given consonant resists coarticulation by different vowels to an approximately invariant extent; Sussman et al. (1999a; Sussman, Fruchter, Hilbert, & Sirosh, 1999b) argue that speakers produce the straight-line relation intentionally, because it fosters consonant identification and perhaps learning of consonantal place of articulation.

A final property of speech that will require an account by theories of speech production is the occurrence of *phase transitions* as rate is increased. This was first remarked on by Stetson (1951) and has been pursued by Tuller and Kelso (1990, 1991). If speakers begin producing /ip/, as rate increases, they shift to /pi/. Beginning with /pi/ does not lead to a shift to /ip/. Likewise, Gleason, Tuller, and Kelso (1996) found shifts from *opt* to *top*, but not vice versa, as rate increased. Phase transitions are seen in other action systems; for example, they underlie changes in gait from walk to trot to canter to gallop. They are considered hallmarks of nonlinear dynamical systems (e.g., Kelso, 1995). The asymmetry in direction of the transition suggests a difference in stability such that CVs are more stable than VCs (and CVCs than VCCs).

Acoustic Targets of Speech Production

I have described characteristics of speech production, but not its goals. Its goals are in contention. Theories that speakers control acoustic signals are less common than those that they control something motoric; however, there is a recent example in the work of Guenther and colleagues (Guenther, Hampson, & Johnson, 1998). Guenther et al. offer four reasons why

targets are likely to be acoustic (in fact, are likely to be the acoustic signal as they are transduced by the auditory system). Opposing a theory that speakers control gestural constrictions (see section titled "Gestural Targets of Speech Production") is that, in the authors' view, there is not very good sensory information about many vocal tract constrictions (e.g., constrictions for vowels where there is no tactile contact between the tongue and some surface). Moreover, although it is true that speakers achieve nearly invariant constrictions (e.g., they always close their lips to say /b/), this can be achieved by a model in which targets are auditory. Third, control over invariant constriction targets would limit the system's ability to compensate when perturbations require new targets. (This is quite right, but, in the literature, this is exactly where compensations to perturbation are not immediate or generally effective. See the studies by Hamlet & Stone, 1978; Hamlet, 1988; Savariaux et al., 1995; Perkell, Matthies, Svirsky, & Jordan, 1993.) Finally, whereas many studies have shown directly (Delattre & Freeman, 1968) or by suggestive acoustic evidence (Hagiwara, 1995) that American English /r/ is produced differently by different speakers and even differently by the same speaker in different phonetic contexts, all of the gestural manifestations produce a similar acoustic product.

In the DIVA model (Guenther et al., 1998), planning for production begins with choice of a phoneme string to produce. The phonemes are mapped one by one onto target regions in auditory-perceptual (speech-sound) space. The maps are to regions rather than to points in order to reflect the fact that the articulatory movements and acoustic signals are different for a given phoneme due to coarticulation and other perturbations. Information about the model's current location in auditory-perceptual space in relation to the target region generates a planning vector, still in auditory-perceptual space. This is mapped to a corresponding articulatory vector, which is used to update articulatory positions achieved over time.

The model uses mappings that are learned during a babbling phase. Infant humans babble on the way to learning to speak. That is, typically between the ages of 6 and 8 months, they produce meaningless sequences that sound as if they are composed of successive CVs. Guenther et al. propose that, during this phase of speech development, infants map information about their articulations onto corresponding configurations in auditory-perceptual space. The articulatory information is from orosensory feedback from their articulatory movements and from copies of the motor commands that the infant used to generate the movements. The auditory perceptual information is from hearing what they have produced. This mapping is called a *forward model;* inverted, it generates movement from auditory-perceptual targets. To this end, the babbling model learns two additional mappings, from speech-sound space, in which (see above) auditory-perceptual target regions corresponding to phonemes are represented as vectors through the space that will take the model from its current location to the target region, and from those trajectories to trajectories in articulatory space.

An important idea in the model is that targets are regions rather than points in acoustic-auditory space. This allows the model to exhibit coarticulation and, with target regions of appropriate ranges of sizes, coarticulation resistance. The model also shows compensation for perturbations, because if one target location in auditory-perceptual space is blocked, the model can reach another location within the target region. Successful phoneme production does not require achievement of an invariant configuration in either auditory-perceptual or articulatory space. This property of the model underlies its failure to distinguish responses to perturbation that are immediately effective from those that require some relearning. The model shows immediate compensations for both kinds of perturbation. It is silent on phase transitions.

Gestural Targets of Speech Production

Theories in which speakers control articulation rather than acoustic targets can address all or most of the reasons that underlay Guenther et al.'s (1998) conclusion that speakers control perceived acoustic consequences of production. For example, Guenther et al. suggest that if talkers controlled constrictions, it would unduly limit their ability to compensate for perturbations where compensation requires changing a constriction location, rather than achieving the same constriction in a different way. A response to this suggestion is that talkers do have more difficulty when they have to learn a new constriction. The response of gesture theorists to /r/ as a source of evidence that acoustics are controlled will be provided after a theory has been described.

Figure 9.3 depicts a model in which controlled primitives are the gestures of Browman and Goldstein's (e.g., 1986) articulatory phonology (see section titled "Feature Systems").

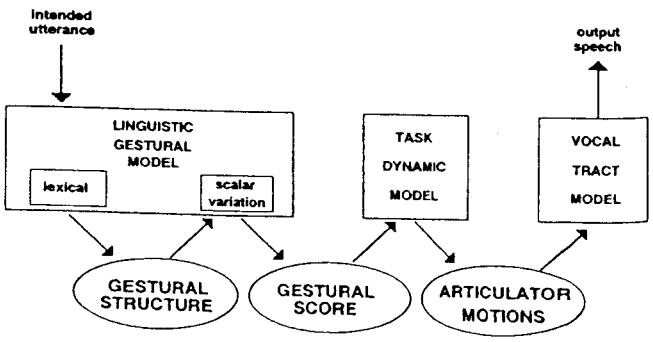

Figure 9.3 Haskins' Computational Gestural Model.

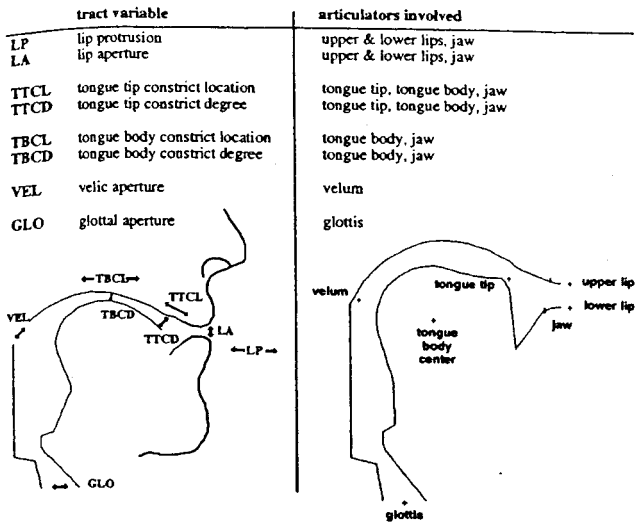

Figure 9.4 Tract Variables for gestures and the articulators comprising their coordinative structures.

Gestures create and release constrictions in the vocal tract. Figure 9.4 displays the tract variables that are controlled when gestures are produced and the gestures' associated articulators. In general, tract variables specify constriction locations (CLs) and constriction degrees (CD) in the vocal tract. For example, to produce a bilabial stop, the constriction location is a specified degree of lip protrusion and the constriction degree is maximal; the lips are closed. The articulators that achieve these values of the tract variables are the lips and the jaw.

The linguistic gestural model of Figure 9.3 generates *gestural scores* such as that in Figure 9.5. The scores specify the gestures that compose a word and their relative phasing. Gestural scores serve as input to the task dynamic model (e.g., Saltzman, 1991; but see Saltzman, 1995; Saltzman & Byrd, 1999). Gestures are implemented as two-tiered dynamical (mass-spring) systems. At an initial level the systems refer to tract variables, and the dynamics are of point attractors. These dynamics undergo a one-to-many transformation to articulator space. Because the transformation is one-many,

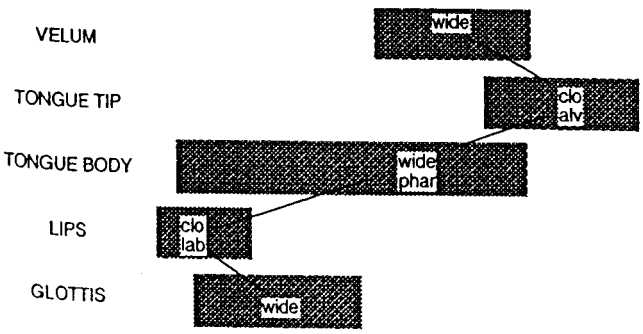

Figure 9.5 Gestural score for the word *pan*.

tract variable values can be achieved flexibly. Because the gestural scores specify overlap between gestures, the model coarticulates; moreover (e.g., Saltzman, 1991), it mimics some of the findings in the literature on coarticulation resistance. In particular, the high resistant consonant /d/ achieves its target constriction location regardless of the vowels with which it overlaps; the constriction location of the lower resistant /g/ moves with the location of the vowel gesture. The model also compensates for the kinds of perturbations to which human talkers compensate immediately (bite blocks and on-line jaw or lip perturbations in which invariant constrictions are achieved in novel ways). It does not show the kinds of compensations studied by Hamlet and Stone (1978), Savariaux et al. (1995), or Perkell et al. (1993), in which new constrictions are required. (The model, unlike that of Guenther et al., 1998, does not learn to speak; accordingly, it cannot show the learning that, for example, Hamlet and Stone find in their human talkers.) The model also fails to exhibit phase transitions although it is in the class of models (nonlinear dynamical systems) that can.

Evidence for Both Models: The Case of /r/

One of the strongest pieces of evidence convincing Guenther et al. (1998) that targets of production are acoustic is the highly variable way in which /r/ is produced. This is because of claims that acoustic variability in /r/ production is less than articulatory variability. Ironically, /r/ also ranks as strong evidence favoring gestural theory among gesture theorists. Indeed, in this domain, /r/ contributes to a rather beautiful recent set of investigations of composite phonetic segments.

The phoneme /r/ is in the class of multigestural (or *composite*) segments, a class that also includes /l/, /w/, and the nasal consonants. Krakow (1989, 1993, see also 1999) was the first to report that two salient gestures of /m/ (velum lowering and the oral constriction gesture) are phased differently in onset and coda positions in a syllable. In onset position, the velum reaches its maximal opening at about the same time as the oral constriction is achieved. In coda position, the velum reaches maximum opening as the oral articulators (the lips for /m/) begin their closing gesture. Similar findings have been reported for /l/. Browman and Goldstein (1995b), following earlier observations by Sproat and Fujimura (1993; see also Gick, 1999), report that in onset position, the terminations of tongue tip and tongue dorsum raising were simultaneous, whereas the tongue dorsum gesture led in coda position. Gick (1999) found a similar relation between lip and tongue body gestures for /w/.

As Browman and Goldstein (1997) remark, in multigestural consonants, in coda position, gestures with wider

constriction degrees (that is, more open gestures) are phased earlier with respect to gestures having more narrow constriction degrees; in onset position, the gestures are more synchronous. Sproat and Fujimura (1993) suggest that the component gestures of composite segments can be identified, indeed, as vocalic (V; more open) or consonantal (C). This is interesting in light of another property of syllables. They tend, universally, to obey a sonority gradation such that more vowel-like (sonorous) consonants tend to be closer to the syllable nucleus than less sonorous consonants. For example, if /t/ and /ɾ/ are going to occur before the vowel in a syllable of English, they are ordered /tɾ/. After the vowel, the order is /ɾt/. The more sonorous of /t/ and /ɾ/ is /ɾ/. Gestures with wider constriction degrees are more sonorous than those with narrow constriction degrees, and, in the coda position, they are phased so that they are closer to the vocalic gesture than are gestures with narrow constriction degrees. A reason for the sonority gradient has been suggested; it permits smooth opening and closing actions of the jaw in each syllable (Keating, 1983).

Goldstein (personal communication, October 19, 2000) suggests that the tendency for /r/ to become something like /ɔi/ in some dialects of American English (Brooklyn; New Orleans), so that *bird* (whose /r/-colored vowel is /ɚ/) is pronounced something like *boid,* may also be due to the phasing characteristics of coda C gestures. The phoneme /r/ may be produced with three constrictions: a pharyngeal constriction made by the tongue body, a palatal constriction made by the tongue blade, and a constriction at the lips. If the gestures of the tongue body and lips (with the widest constriction degrees) are phased earlier than the blade gesture in coda position, the tongue and lip gestures approximate those of /ɔ/, and the blade gesture against the palate is approximately that for /i/.

But what of the evidence of individual differences in /r/ production that convinced Guenther et al. (1998) that speech production targets are auditory-perceptual? One answer is that the production differences can look smaller than they have been portrayed in the literature if the gestural focus on vocal tract configurations is adopted. The striking differences that researchers have reported are in tongue shape. However, Delattre and Freeman (1968), characteristically cited to underscore the production variability of /r/, make this remark: "Different as their tongue shapes are, the six types of American /r/'s have one feature in common—they have two constrictions, one at the palate, another at the pharynx" (p. 41). That is, in terms of constriction location, a gestural parameter of articulatory phonology, there is one type of American English /r/, not six.

SPEECH PERCEPTION

The chapter began with the language knower. Then it explored how such an individual might formulate a linguistic message at the phonological level of description and implement the message as vocal tract activity that causes an acoustic speech signal. For an act of communication to be completed, a perceiver (another language knower) must intercept the acoustic signal and use it to recover the speaker's message. In this section, the focus is on how perception takes place.

Phonetic Perception

Preliminary Issues

I have suggested that a constraint on development of theories of phonological competence, planning, production, and perception should be an understanding that languages are likely to be parity fostering. Two parity-fostering characteristics are phonological forms that can be made public, and preservation of those forms throughout a communicative exchange. If theorists were to hew to expectations that languages have these properties, then we would expect to find perception theories in which perceptual objects are planned and produced phonological forms. We do not quite find that, because, as indicated in the introduction, research on perception, production, planning, and phonological description all have progressed fairly independently.

However, there is one respect in which perception theories intersect fairly neatly with production theories. They partition into two broad classes that divide according to the theorists' claims about immediate objects of speech perception. The majority view is that objects are acoustic. This is not an implausible view, given that acoustic signals are stimuli for speech perception. The minority view is that objects are gestural. Considerations of parity suggest a pairing of acoustic theories of speech perception with production theories like that of Guenther et al. (1998) in which speakers aim to produce acoustic signals with required properties. Gestural theories of speech perception are consistent with production theories, such as that of Saltzman and colleagues, in which speakers aim to produce gestures with particular properties.

Another issue that divides theorists is whether speech perception is special—that is, whether mental processes that underlie speech perception are unique to speech, perhaps taking place in a specialization of the brain for speech (a phonetic module, as Liberman & Mattingly, 1985, propose). There is reason to propose that speech processing is special. In speaking, talkers produce discrete, but temporally overlapping,

gestures that correspond in some way to the phonological forms that listeners must recover. Coarticulation ensures that there is no temporally discrete, phone-sized segmental structure in the acoustic signal corresponding to phonological forms and that the acoustic signal is everywhere context sensitive. If listeners do recover phonological forms when they listen, this poses a problem. Listeners have to use the continuous acoustic signal to recover the discrete context-invariant phonological forms of the talker's message. Because, in general, acoustic signals are not caused by sequences of discrete, coarticulated mechanical events, speech does appear to pose a unique problem for listeners.

However, there is also a point of view that the most conservative or parsimonious first guess should be that processing is not special. Until the data demand postulating a specialization, we should attempt to explain speech perception by invoking only processes that are required to explain other kinds of auditory perception. It happens that acoustic theorists generally take this latter view. Some gestural theorists take the former.

Acoustic Theories of Speech Perception

There are a great many different versions of acoustic theory (e.g., Diehl & Kluender, 1989; Kuhl, 1987; Massaro, 1987, 1998; Nearey, 1997; Stevens & Blumstein, 1981; Sussman et al., 1999a). Here, Diehl and Kluender's auditory enhancement theory will illustrate the class.

Acoustic theories are defined by their commitment to immediate perceptual objects that are acoustic (or auditory—that is, perceived acoustic) in nature. One common idea is that auditory processing renders an acoustic object that is then classified as a token of a particular phonological category. Auditory enhancement theory makes some special claims in addition (e.g., Diehl & Kluender, 1989; Kluender, 1994). One is that there is lots of covariation in production of speech and in the consequent acoustic signal. For example, as noted earlier, rounding in vowels tends to covary with tongue backness. The lips and the tongue are independent articulators; why do their gestures covary as they do? The answer from auditory enhancement theory is that both the rounding and the tongue backing gestures lower a vowel's second formant. Accordingly, having the gestures covary results in back vowels that are acoustically highly distinct from front (unrounded) vowels. In this and many other examples offered by Diehl and Kluender, pairs of gestures that, in principle, are independent conspire to make acoustic signals that maximally distinguish phonological form. This should benefit the perceiver of speech.

Another kind of covariation occurs as well. Characteristically, a given gesture has a constellation of distinct acoustic consequences. A well-known example is voicing in stop consonants. In intervocalic position (as in *rapid* vs. *rabid*), voiced and voiceless consonants can differ acoustically in 16 different ways or more (Lisker, 1978). Diehl and Kluender (1989) suggest that some of those ways, in phonological segments that are popular among languages of the world, are mutually enhancing. For example, voiced stops have shorter closure intervals than do voiceless stops. In addition, they tend to have voicing in the closure, whereas voiceless stops do not. Parker, Diehl, and Kluender (1986) have shown that low-amplitude noise in an otherwise silent gap between two square waves makes the gap sound shorter than it sounds in the absence of the noise (as it indeed is). This implies that, in speech, voicing in the closure reinforces the perception of a shorter closure for voiced than voiceless consonants. This is an interesting case, because, in contrast to rounding and backing of vowels where two gestures reinforce a common acoustic property (a low F2), in this case, a single gesture—approximation of the vocal folds during the constriction gesture for the consonant—has two or more enhancing acoustic consequences. Diehl and Kluender (1989; see also Kluender, 1994) suggest that language communities "select" gestures that have multiple, enhancing acoustic consequences.

A final claim of the theory is that speech perception is not special and that one can see the signature of auditory processing in speech perception. A recent example of such a claim is provided by Lotto and Kluender (1998). In 1980, Mann had reported a finding of "compensation for coarticulation." She synthesized an acoustic continuum of syllables that ranged from a clear /da/ to a clear /ga/ with many more ambiguous tokens in between. The syllables differed only in the direction of the third formant transition, which fell for /da/ and rose for /ga/. She asked listeners to identify members of the continuum when they were preceded by either of the two precursor syllables /al/ or /ar/. She predicted and found that listeners identified more ambiguous continuum members as /ga/ in the context of precursor /al/ than /ar/. The basis for Mann's prediction was the likely effect of coarticulation by /l/ and /r/ on /d/ and /g/. The phoneme /l/ has a tongue tip constriction that, coarticulated with /g/, a back consonant, is likely to pull /g/ forward; /r/ has a pharyngeal constriction that, coarticulated with /d/, is likely to pull /d/ back. When listeners reported more /g/s after /al/ and more /d/s after /ar/, they appeared to compensate for the fronting effects that /l/ should have on /g/ and the backing effects of /r/ on /d/.

Lotto and Kluender (1998) offered a different account. They noticed that, in Mann's stimulus set, /l/ had a very high

ending frequency of F3, higher than the starting F3s of any members of the /da/-to-/ga/ continuum. The phoneme /r/ had a very low ending frequency of F3, lower than the starting frequency of any members of the continuum. They proposed that the ending F3 frequencies of /al/ and /ar/ were exerting a contrast effect on the starting F3s of the continuum members. Contrast effects are pervasive in perception research across the sensory modalities (e.g., Warren, 1985, who, however, does not refer to them as contrast effects). For example, when individuals judge the heaviness of weights (Guilford & Park, 1931), they judge an intermediate weight lighter if they have just hefted a heavier weight than if they have just hefted a lighter weight. Lotto and Kluender suggested that the very high ending F3 of /l/ made following F3 onsets of continuum members effectively lower (and so more /g/-like) than they were; the very low F3 of /r/ made onset F3s effectively higher and more /d/-like.

They tested their hypothesis by substituting high and low sinewave tones for the precursor /al/ and /ar/ syllables of Mann (1980), and they found more /g/ judgments following the high than the low precursor tone. This cannot be compensation for coarticulation. It is, rather, according to Lotto and Kluender (1998), a signature of auditory processing showing up in speech perception judgments.

Comparisons like this between perception of speech and of nonspeech analogues has provided one way of testing claims of auditory theories. Parker et al. (1986) tested whether two acoustic properties were mutually enhancing. The test by Lotto and Kluender tested for evidence of auditory processing in speech perception. Generally, investigators have used speech/nonspeech comparisons as a way to test whether speech processing is specialized and distinct from auditory processing. Many tests have found closely similar response patterns to speech and closely similar nonspeech signals (e.g., Sawusch & Gagnon, 1995). As we will see, however, not all have.

Another test of auditory theories has been to compare responses by humans and nonhumans to speech signals. Clearly, nonhumans do not have specializations for human speech perception. If they show some of the markers of human speech perception, then it is not necessary to suppose that a specialization is responsible for the markers in humans. There are some striking findings here. Kuhl and Miller (1978) trained chinchillas in a go–no go procedure to move to a different compartment of a cage when they heard one endpoint of an acoustic voice onset time (VOT) continuum, but not when they heard a syllable at the other end. Following training, they were tested on all continuum members between the two endpoints as well as on the endpoints themselves. This allowed Kuhl and Miller to find a boundary along the continuum at which the chinchillas' behavior suggested that a voiced percept had replaced a voiceless one. Remarkably, the boundaries were close to those of humans, and there was an even more remarkable finding. In human speech, VOTs are longer for farther back places of articulation. That is, in English, /pa/ has a shorter VOT than /ta/, which has a shorter VOT than /ka/ (e.g., Zue, 1980). This may be because voicing cannot resume following a voiceless consonant until there is a sufficient drop in pressure across the larynx. With back places of constriction, the cavity above the larynx is quite small and the pressure correspondingly higher than for front constrictions. English listeners place VOT boundaries at shorter values for /pa/ than for /ta/ and for /ta/ than for /ka/, as do chinchillas (Kuhl & Miller, 1978). It is not known what stimulus property or auditory system property might underlie this outcome. However, most investigators are confident that chinchillas are not sensitive to transglottal pressure differences caused by back and front oral constrictions in human speech.

Another striking finding, now with quail, is that of Lotto, Kluender, and Holt (1997) that quail show "compensation for coarticulation" given stimuli like those used by Mann (1980).

Readers may be asking why anyone is a gesture theorist. However, gesture theories, like acoustic theories, derive from evidence and from theoretical considerations. Moreover, theorists argue that many of the claims and findings of acoustic theories are equally compatible with gesture theories. For example, findings that language communities gravitate toward phones that have mutually distinctive acoustic signals is not evidence that perceptual objects are acoustic. In gesture theories, the acoustic signal is processed; it is used as information for gestures. If the acoustic signals for distinct gestures are distinct, that is good for the gesture perceiver.

The most problematic findings for gesture theorists may be on the issue of whether speech perception is special. The negative evidence is provided by some of the speech/ nonspeech and human/nonhuman comparisons. Here, there are two lines of attack that gesture theorists can mount. One is to point out that not all such comparisons have resulted in similar response patterns (for speech/nonspeech, see below; for human/nonhuman, see, e.g., range effects in Waters & Wilson, 1976; see also Sinnott, 1974, cited in Waters & Wilson, 1976). If there are real differences, then the argument against a specialization weakens. A second line of attack is to point out that the logic of the research in the two domains is weak. It is true that if humans and nonhumans apply similar processes to acoustic speech signals (and if experiments are designed appropriately), the two subject groups should show similar response patterns to the stimuli. However, the logic required by the research is the reverse of that. It maintains that if humans

and nonhumans show similar response patterns, then the processes applied to the stimuli are the same. This need not hold (cf. Trout, 2001). The same can be said of the logic of speech/nonspeech comparisons.

Gesture Theories of Speech Perception

There are two gesture theories in the class, both largely associated with theorists at Haskins Laboratories. Gesture theories are defined by their commitment to the view that immediate objects of perception are gestural. One of these theories, the motor theory (e.g., Liberman & Mattingly, 1985; Liberman & Whalen, 2000), also proposes that speech perception is special. The other, direct realist theory (Best, 1995; Fowler, 1986, 1996), is agnostic on that issue.

The motor theory of speech perception was the first gesture theory. It was developed by Liberman (1957, see also 1996) when he obtained experimental findings that, in his view, could not be accommodated by an acoustic theory. He and his colleagues were using two complementary pieces of technology, the sound spectrograph and the pattern playback, to identify the acoustic cues for perception. They used the spectrograph to make speech visible in the informative ways that it does, identified possible cues for a given consonant or vowel, and reproduced those cues by painting them on an acetate strip that, input to the pattern playback, was transformed to speech. If the acoustic structure preserved on acetate was indeed important for identifying the phone, it could be identified as a cue.

One very striking finding in that research was that, due to coarticulation, acoustic cues for consonants especially were highly context sensitive. Figure 9.6 provides a schematic spectrographic display of the syllables /di/ and /du/. Although natural speech provides a much richer signal than that in Figure 9.6, the depicted signals are sufficient to be heard as

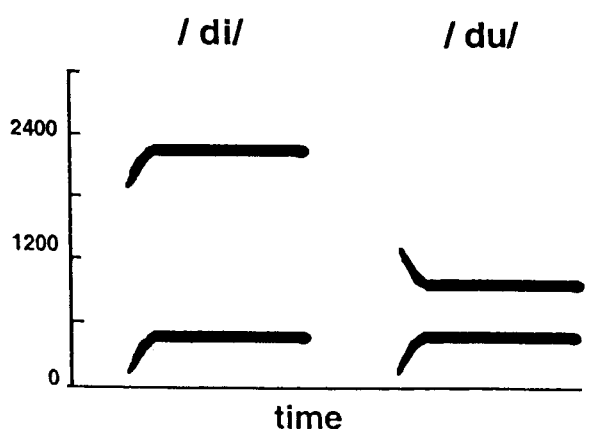

Figure 9.6 Schematic depiction of the synthetic syllables, /di/ and /du/.

/di/ and /du/. The striking finding was that the information critical to identification of these synthetic syllables was the transition of the second formant. However, that transition is high in frequency and rising in /di/, but low and falling in /du/. In the context of the rest of each syllable, the consonants sound alike to listeners. Separated from context, they sound different, and they sound the way they look like they should sound: two "chirps," one high in pitch and one lower.

Liberman (e.g., 1957) recognized that, despite the context sensitivity of the acoustic signals for /di/ and /du/, naturally produced syllables do have one thing in common. They are produced in the same way. In both syllables, the tongue tip makes a constriction behind the teeth. Listeners' percepts appeared to track the speaker's articulations.

A second striking finding was complementary. Stop consonants can be identified based on their formant transitions, as in the previous example, or based on a burst of energy that, in natural speech, precedes the transitions and occurs as the stop constriction is released. Liberman, Delattre, and Cooper (1952) found that a noise burst centered at 1440 Hz and placed in front of the vowels /i/ or /u/ was identified predominantly as /p/. However in front of /a/, it was identified as /k/. In this case, an invariant bit of acoustic structure led to different percepts. To produce that bit of acoustic structure before /i/ or /u/, a speaker has to make the constriction at the lips; to produce it before /a/, he or she has to make the constriction at the soft palate. These findings led Liberman to ask: "when articulation and the sound wave go their separate ways, which way does the perception go?" (Liberman, 1957, p. 121). His answer was: "The answer so far is clear. The perception always goes with articulation."

Although the motor theory was developed to explain unexpected research findings, Liberman and colleagues proposed a rationale for listeners' perception of gestures. Speakers have to coarticulate. Liberman and colleagues (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967) suggested that coarticulation is necessary to evade the limits of the temporal resolving power of the listener's ear. These limits were proposed to underlie the failure of Haskins researchers more than 50 years ago to train people to use an acoustic alphabet intended for use in a reading machine for the blind (see Liberman, 1996). Listeners could not perceive sequences of discrete sounds at anything close to the rates at which they perceive speech. Coarticulation provides a continuous signal evading the temporal resolving power limits of the ear, but it creates a new problem. The relation between phonological forms and acoustic speech structure is opaque. Liberman et al. (e.g., 1967) suggested that coarticulation required a specialization of the brain to achieve it. What system would be better suited to deal with the acoustic complexities

to which coarticulation gives rise than the system responsible for generating coarticulated speech? In later versions of the motor theory, this hypothesized specialization was identified as a phonetic module (cf. Fodor, 1983).

There is an independent route to a conclusion that speech perception yields gestures. Fowler's (e.g., 1986, 1996; see also Best, 1995; Rosenblum, 1987) direct realist theory derived that claim by developing a theory of speech perception in the context of a universal theory of perceptual function. That theory, developed by James Gibson (e.g., 1966, 1979), notes that perceptual systems constitute the only means that animals have to know their world. By hypothesis, they serve that function in just one general way. Stimulus structure at the sense organs is not perceived itself. Rather, it serves as information for its causal source in the environment, and the environment is thereby perceived. In vision, for example, light that reflects from objects in the environment is structured by the properties of the objects and takes on structure that is distinctive to those properties. Because the structure is distinctive to the properties, it can serve as information for them. Environmental events and objects, not the reflected light, are perceived. Fowler (1996) argued that, if even speech perception were wholly unspecial, listeners would perceive gestures, because gestures cause the structure in stimulation to the ear. And the auditory system (or the phonetic module), no less than the visual system, uses information in stimulation at the sense organ to reveal the world of objects and events to perceivers.

What does the experimental evidence show? An early finding that Liberman (1957) took to be compatible with his findings on /di/-/du/ and /pi/-/ka/-/pu/ was categorical perception. This was a pair of findings obtained when listeners made identification and discrimination judgments of stimuli along an acoustic continuum. Figure 9.7 displays schematic findings for a /ba/-to-/da/ continuum. Although the stimuli form a smooth continuum (in which the second formant transition is gradually shifted from a trajectory for /ba/ to one for /da/), the identification function is very sharp. Most stimuli along the continuum are heard either as a clear /ba/ or as a clear /da/. Only one or two syllables in the middle of the continuum are ambiguous. The second critical outcome was obtained when listeners were asked to discriminate pairs of syllables along the continuum. The finding was that discrimination was near chance among pairs of syllables both members of which listeners identified as /ba/ or both /da/, but it was good between pair members that were equally acoustically similar as the /ba/ pairs and the /da/ pairs, but in which listeners heard one as /ba/ and the other as /da/. In contrast, say, to colors, where perceivers can easily discriminate colors that they uniformly label as *blue*, to a first approximation,

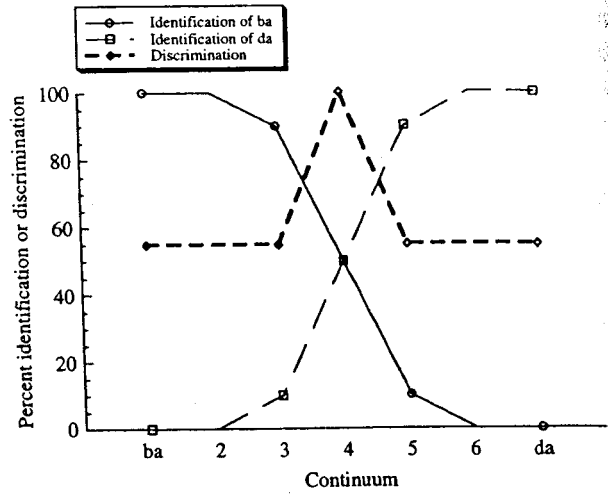

Figure 9.7 Schematic depiction of categorical identification and discrimination.

listeners could only discriminate what they labeled distinctively. The early interpretation of this finding was that it revealed perception of gestures, because the place of articulation difference between /ba/ and /da/, unlike the acoustic difference, is categorical.

This interpretation was challenged, for example, by Pisoni (e.g., Pisoni & Tash, 1974). In their study, Pisoni and Tash showed that *same* responses to pairs of syllables that were labeled the same but that differed acoustically were slower than to identical pairs of syllables. Accordingly, listeners have at least fleeting access to within-category differences. Despite this and other findings, the name *categorical perception* has endured, but now it is typically used only to refer to the data pattern of Figure 9.7, not to its original interpretation.

A set of findings that has a natural interpretation in gesture theories is the *McGurk effect* (named for one of its discoverers; McGurk and MacDonald, 1976). This effect is obtained when a videotape of a speaker mouthing a word or syllable (say, /da/) is dubbed with a different, appropriately selected, syllable (say, /ma/). With eyes open, listeners hear a syllable that integrates information from the two modalities. (In the example, they hear /na/, which takes its place of articulation from /da/ and its manner and voicing from /ma/.) The integration is expected in a theory in which gestures are perceived, because both modalities provide information about gestures. There is, of course, an alternative interpretation from acoustic theories. The effect may occur because of our vast experience both seeing and hearing speakers talk. This experience may be encoded as memories in which compatible sights and sounds are associated (but see Fowler and Dekle, 1991).

There are other findings that gesture theorists have taken to support their theory. For example, researchers have shown

that performance discriminating pairs of syllables can be better for stimuli that differ in one acoustic cue for a gesture than for stimuli that differ in that cue and one other for the same gesture (e.g., Fitch, Halwes, Erickson, & Liberman, 1980). This is unexpected on acoustic grounds. It occurs just when the two cues are selected so that the stimuli of a pair are identified as the same gesturally, whereas the pair differing in one cue are not always. Another finding is that people are remarkably rapid shadowers of speech under some conditions (e.g., Porter & Castellanos, 1980; Porter & Lubker, 1980). This has been interpreted as evidence that perceiving speech is perceiving gestures that constitute the instructions for the shadowing response. A third kind of finding has been research designed to show that listeners parse acoustic speech signals along gestural lines (e.g., Fowler & Smith, 1986; Pardo & Fowler, 1997). For example, when two gestures, say devoicing a preceding stop consonant and production of intonational accent, have convergent effects on the fundamental frequency (F0) pattern on a vowel, listeners do not hear the combined effects as the vowel's intonation or pitch. They hear the contribution to F0 made by the devoicing gesture as information for devoicing (Pardo & Fowler, 1997). Finally, Fowler, Brown, and Mann (2000) have recently disconfirmed the contrast account of compensation for coarticulation offered by Lotto and Kluender (1998). They used the McGurk effect to show that, when the only information distinguishing /al/ from /ar/ was optical, and the only information distinguishing /da/ from /ga/ was acoustic, participants provided more /ga/ responses in the context of precursor /al/ than /ar/. This cannot be a contrast effect. Fowler et al. concluded that the effect is literally compensation for coarticulation.

Motor theorists have also attempted to test their idea that speech perception is achieved by a phonetic module. Like acoustic theorists, they have compared listeners' responses to speech and to similar nonspeech signals, now with the expectation of finding differences. One of the most elegant demonstrations was provided by Mann and Liberman (1983). They took advantage of duplex perception, in which, in their version, components of a syllable were presented dichotically. The base, presented to one ear, included steady-state formants for /a/ preceded by F1 and F2 transitions consistent with either /d/ or /g/. An F3 transition, presented to the other ear, distinguished /da/ from /ga/. Perception is called *duplex* because the transitions are heard in two different ways at the same time. At the ear receiving the base, listeners hear a clear /da/ or a clear /ga/ depending on which transition was presented to the other ear. At the ear receiving the transition, listeners hear a nonspeech chirp. On the one hand, this can be interpreted as evidence for a speech module, because how

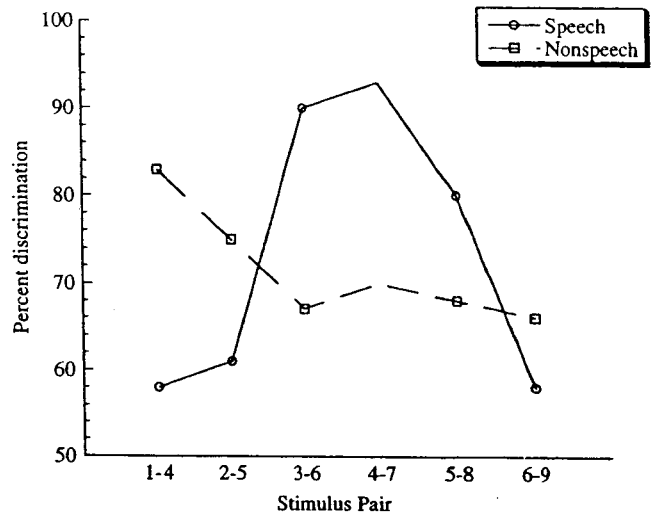

Figure 9.8 Results of speech and nonspeech discriminations of syllables and chirps (Mann & Liberman, 1983).

else, except with a separate perceptual system, can the same acoustic fragment be heard in two different ways at once? (However, see Fowler & Rosenblum, 1990, for a possible answer to the question.) On the other hand, it can provide the means of an elegant speech/nonspeech comparison, because listeners can be asked to attend to the syllable and make phonetic judgments that will vary as the critical formant transition varies, and they can be asked to attend to the chirps and make analogous judgments about them. Presented with a continuum of F3 transitions to one ear and the base to the other, and under instructions to discriminate syllable pairs or chirp pairs, listeners responded quite differently depending on the judgment, even though both judgments were based on the same acoustic pattern. Figure 9.8 shows that their speech discrimination judgments showed a sharply peaked pattern similar to that in Figure 9.7. Their chirp judgments showed a nearly monotonically decreasing pattern. This study, among others, shows that not all comparisons of speech and nonspeech perception have uncovered similarities.

Learning and Speech Perception

So far, it may appear as if speech perception is unaffected by a language user's experience talking and listening. It is affected, however. Experience with the language affects how listeners categorize consonants and vowels, and it affects the internal structure of native language phonological categories. It also provides language users with knowledge of the relative frequencies with which consonants and vowels follow one another in speech (e.g., Pitt & McQueen, 1998; Vitevitch & Luce, 1999) and with knowledge of the words of the

language. It is currently debated (e.g., Norris, McQueen, & Cutler, 1999; Samuel, 2000) whether particularly lexical knowledge, in fact, affects speech perception, but it is clear that it affects how listeners ultimately identify consonants and vowels.

Knowledge of Categories

The concept of category (see chapter by Goldstone & Kersten in this volume) remains rather fuzzy, although it is clear that it is required for understanding speech perception. Language users treat sets of physically distinct tokens of consonants and vowels as functionally equivalent. For example, English listeners treat tokens of /t/ as members of the same category when /t/s differ in aspiration due to variation in position in a syllable or stress, and when they differ due to speaking rate, coarticulatory context, dialect, foreign accent, and idiosyncratic speaker characteristics. (They treat them as functionally equivalent, for example, when they count physically distinct /t/s before /ap/ all as consonants of the word *top*.) The concept of category is meant to capture this behavior. Functional equivalence of physically distinct tokens may or may not imply that listeners represent consonants and vowels as abstract types. The section titled "Another Abstractness Issue: Exemplar Theories of the Lexicon" described exemplar theories of linguistic knowledge in which clusters of relevantly similar tokens underlie behaviors suggestive of type memories. Accordingly, the reader should interpret the following discussion of categories as neutral between the possibilities that abstract types are or are not components of linguistic competence.

From the earliest ages at which they are tested, infants show evidence of categorization. On the one hand, they exhibit something like categorical perception. Eimas, Siqueland, Jusczyk, and Vigorito (1971) pioneered the use of a high-amplitude sucking technique to test infants as young as one month of age. Infants sucked on a nonnutritive nipple. If they sucked with sufficient vigor they heard a speech syllable, for example, /ba/. Over time, infants increased their sucking rate under those conditions, but eventually they showed habituation: Their sucking rate declined. Following that, Eimas et al. presented different syllables to all infants except those in the control group. They presented a syllable that adult listeners heard as /pa/ or one that was acoustically as distant from the original /ba/ as the /pa/ syllable but that adults identified as /ba/. Infants dishabituated to the first syllable, showing that they heard the difference, but they did not dishabituate to the second.

Kuhl and colleagues (e.g., Kuhl & Miller, 1982) have shown that infants classify by phonetic type syllables that they readily discriminate. Kuhl and Miller trained 6-month-old infants to turn their head when they heard a phonetic change in a repeating background vowel (from /a/ to /i/). Then they increased the difficulty of the task by presenting as background /a/ vowels spoken by different speakers or with different pitch contours. These vowels are readily discriminated by infants, but adults would identify all of them as /a/. When a change occurred, it was to /i/ vowels spoken by the different speakers or produced with the different pitch contours. Infants' head turn responses demonstrated that they detected the phonetic identity of the variety of /a/ vowels and the phonetic difference between them and the /i/ vowels.

We know, then, that infants detect phonetic invariance over irrelevant variation. However, with additional experience with their native language, they begin to show differences in what they count as members of the same and different categories. For example, Werker and Tees (1984) showed that English-learning infants at 6–8 months of age distinguished Hindi dental and retroflex voiceless stops. However, at 10–12 months they did not. English- (non-Hindi-) speaking adults also had difficulty making the discrimination, whereas Hindi adults and three Hindi 10–12 month olds who were tested made the discrimination readily. One way to understand the English-learning infants' loss in sensitivity to the phonetic distinction is to observe that, in English, the distinction is not contrastive. English alveolar stops are most similar to the Hindi dental and retroflex stops. If an English speaker (perhaps due to coarticulation) were to produce a dental stop in place of an alveolar one, it would not change the word being produced from one word into another. With learning, categories change their structure to reflect the patterning of more and less important phonetic distinctions of the language to which the learner is exposed.

In recent years, investigators have found that categories have an internal structure. Whereas early findings from categorical perception implied that all category members, being indiscriminable, must be equally acceptable members of the category, that is not the case, as research by Kuhl (e.g., 1991) and by Miller has shown.

Kuhl (e.g., 1987) has suggested that categories are organized around best instances or prototypes. When Grieser and Kuhl (1989) created a grid of vowels, all identified as /i/ by listeners (ostensibly; but see Lively & Pisoni, 1995) but differing in their F1s and F2s, listeners gave higher goodness ratings to some tokens than to others. Kuhl (1991) showed, in addition, that listeners (adults and infants aged 6–7 months, but not monkeys) showed poorer discrimination of /i/ vowels close to the prototype (that is, the vowel given the highest goodness rating) than of vowels from a nonprototype (a vowel given a low goodness rating), an outcome she called

ıe "magnet effect." Kuhl, Williams, Lacerda, Stevens, and ,indblom (1992) showed that English and Swedish infants how magnet effects around different vowels, reflecting the ifferent vowel systems of their languages.

We should not think of phonological categories as having n invariant prototype organization, however. Listeners iden-ify different category members as best exemplars in different ontexts. This has been shown most clearly in the work of oanne Miller and colleagues. Miller and colleagues (e.g., Ailler & Volaitis, 1989) have generated acoustic continua anging, for example, from /bi/ to /pi/ and beyond to a very ong VOT /p/ designated */p/. Listeners make goodness judg-nents to the stimuli (in the example, they rate the goodness)f the consonants as /p/s), and Miller and colleagues get data ike those in Figure 9.9. (The VOT continuum is truncated at he long end in the figure.) The functions have a peak and raded sides. Miller and collaborators have shown that the ocation of the best rated consonant along an acoustic contin-ıum can vary markedly with rate of production, syllable ;tructure, and other variables. An effect of rate is shown in Figure 9.9 (where the legend's designations "125 ms" and '325 ms" are syllable durations for fast and slow produc-ions, respectively). Faber and Brown (1998) showed a :hange in the prototype with coarticulatory context. These findings suggest that the categories revealed by these studies have a dynamical character (cf. Tuller, Case, & Kelso, 1994).

How should the findings of Kuhl and colleagues and of Miller and colleagues be integrated? It is not yet clear. Kuhl (e.g., Kuhl & Iverson, 1995) acknowledges that her findings are as consistent with a theory in which there are actual prototypes in memory as with one in which prototypicality is an emergent property of an exemplar memory. It may be easier in an exemplar theory to understand how categories can change their structure dynamically.

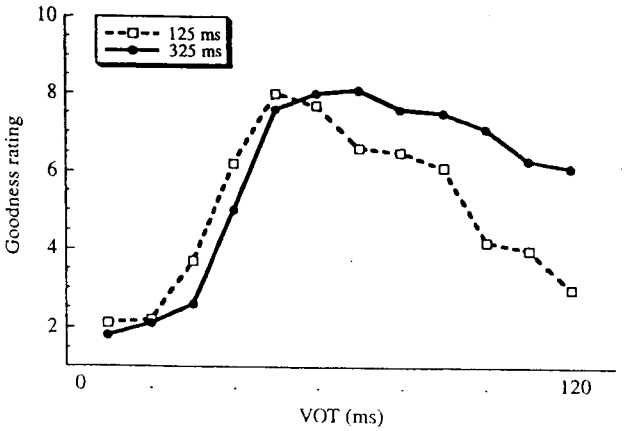

Figure 9.9 Goodness ratings along a /bi/-/pi/-*/pi/ continuum. Data similar to those of Miller and Volaitis (1989).

Possibly, Kuhl's magnet effect can also be understood from the framework of Miller's (e.g., Miller & Volaitis, 1989) findings if both sets of findings are related to Catherine Best's perceptual assimilation model (PAM; e.g., Best, 1994). PAM is a model that captures consequences of perceptual speech learning. In the model, experience with the language eventuates in the formation of language-specific categories. When listeners are given two nonnative consonants or two vowels to discriminate, and they fall into the same native category, discrimination is very poor if the phones are equally good exemplars of the category. Discrimination is better if one is judged a good and one a poor exemplar. This can be understood by looking at Figure 9.9. Tokens that fall near the peak of the goodness function sound very similar to listeners, and they sound like good members of the category. However, one token at the peak and one over to the left or right side of the function sound different in goodness and therefore presumably in phonetic quality. Functions with flat peaks and accelerating slopes to the sides of the function would give rise to a magnet effect. That is, tokens surrounding the peak would be difficult to discriminate, but equally acoustically similar tokens at the sides of the function (so a nonprototype and a token near to it) would differ considerably in goodness and be easily discriminable.

Lexical and Phonotactic Knowledge

Word knowledge can affect how phones are identified, as can knowledge of the frequencies with which phones follow one another in speech. Ganong (1980) showed that lexical knowledge can affect how a phone is identified. He created pairs of continua in which the phone sequence at one end was a word but the sequence at the other end was a nonword. For example, in one pair of continua, VOT was varied to produce a *gift*-to-*kift* continuum and a *giss*-to-*kiss* continuum. Ganong found that listeners provided more *g* responses in the *gift-kift* continuum than in the *giss-kiss* continuum. That is, they tended to give responses suggesting that they identified real words preferentially. This result has recently been replicated with audiovisual speech. Brancazio (submitted) has shown that participants exhibit more McGurk integrations if they turn acoustically specified nonwords into words (e.g., acoustic *besk* dubbed onto video *desk*, with the integrated McGurk response being *desk*) than if they turn acoustically specified words into nonwords (e.g., acoustic *bench* dubbed on to video *dench*).

Ganong's (1980) result has at least two interpretations. One is that lexical information feeds down and affects perceptual processing of consonants and vowels. An alternative is that perceptual processing of consonants and vowels is

encapsulated from such feedback; however, when the processor yields an ambiguous output, lexical knowledge is brought to bear to resolve the ambiguity. In the first account, the effect of the lexicon is on perceptual processing; in the second it is on processing that follows perception of phones. The Ganong paradigm has been used many times in creative attempts to distinguish these interpretations (e.g., Fox, 1984; Miller & Dexter, 1988; Newman, Sawusch, & Luce, 1997). However, it remains unresolved.

A second finding of lexical effects is phonemic restoration (e.g., Samuel, 1981, 1996; Warren, 1970). When the acoustic consequences of a phoneme are excised from a word (in Warren's classic example, the /s/ noise of *legislature*) and are replaced with noise that would mask the acoustic signal if it were present, listeners report hearing the missing phoneme and mislocate the noise. Samuel (1981) showed that when two versions of these words are created, one in which the acoustic consequences are present in the noise and one in which they are absent, listeners asked to make a judgment whether the phone is present or absent in the noise show lower perceptual sensitivity to phones in words than in nonwords. That the effect occurs on the measure of perceptual sensitivity (d′) suggests that, here, lexical knowledge is exerting its effect on phoneme perception itself. (However, that d′ can be so interpreted in word recognition experiments has been challenged; see Norris, 1995.)

A final lexical effect occurs in experiments on compensation for coarticulation. Mann and Repp (1981) found compensation for /s/ and /ʃ/ on members of a /ta/-to-/ka/ continuum such that the more front /s/ fostered /ka/ responses, and the more back /ʃ/ fostered /ta/ responses. Elman and McClelland (1988) used compensation for coarticulation in a study that seemingly demonstrated lexical feedback on perceptual processing of consonants. They generated continua ranging from /d/ to /g/ (e.g., *dates* to *gates*) and from /t/ to /k/ (e.g., *tapes* to *capes*). Continuum members followed words such as *Christmas* and *Spanish* in which the final fricatives of each word (or, in another experiment, the entire final syllables) were replaced with the same ambiguous sound. Accordingly, the only thing that made the final fricative of *Christmas* an /s/ was the listeners' knowledge that *Christmas* is a word and *Christmash* is not. Lexical knowledge, too, was all that made the final fricative of *Spanish* an /ʃ/. Listeners showed compensation for coarticulation appropriate for the lexically specified fricatives of the precursor words.

This result is ascribed to feedback effects on perception, because compensation for coarticulation is quite evidently an effect that occurs during perceptual processing of phones. However, Pitt and McQueen (1998) challenged the feedback interpretation with findings appearing to show that the effect is not really lexical. It is an effect of listeners' knowledge of the relative frequencies of phone sequences in the language, an effect that they identify as prelexical and at the same level of processing as that on which phonemes are perceived. Pitt and McQueen note that in English, /s/ is more likely to follow the final vowel of *Christmas* than is /ʃ/, and /ʃ/ is more common than /s/ following the final vowel of *Spanish*. (If readers find these vowels—ostensibly /ɔ/ and /ʌ/ according to Pitt and McQueen—rather subtly distinct, they are quite right.) These investigators directly pitted lexical identity against phone sequence frequency and found compensation for coarticulation fostered only by the transition probability variable. Lately, however, Samuel (2000) reports finding a true lexical effect on phoneme perception. The clear result is that lexical knowledge affects how we identify consonants and vowels. It is less clear where in processing the lexical effect comes in.

Pitt and McQueen's study introduces another knowledge variable that can affect phone identification: knowledge of the relative transition frequencies between phones. Although this logically could be another manifestation of our lexical knowledge, Pitt and McQueen's findings suggest that it is not, because lexical and transition-probability variables dissociate in their effects on compensation for coarticulation. A conclusion that transition probability effects arise prelexically is reinforced by recent findings of Vitevitch and Luce (1998, 1999).

There are many models of spoken-word recognition. They include the pioneering TRACE (McClelland & Elman, 1986), Marslen-Wilson's (e.g., 1987) cohort model, the neighborhood activation model (NAM; Luce, 1986; Luce & Pisoni, 1998), the fuzzy logical model of perception (FLMP; e.g., Massaro, 1987, 1998), and shortlist (e.g., Norris, 1994). (A more recent model of Norris et al., 1999, Merge, is currently a model of phoneme identification; it is not a full-fledged model of word recognition.)

I will describe just two models, TRACE and a nameless recurrent network model described by Norris (1993); these models represent extremes along the dimension of interactive versus feedforward only (autonomous) models.

In TRACE, acoustic signals are mapped onto phonetic features, features map to phonemes, and phonemes to words. Features activated by acoustic information feed activation forward to the phonemes to which they are linked. Phonemes activate words that include them. Activation also feeds back from the word level to the phoneme level and from the phoneme level to the feature level. It is this feedback that identifies TRACE as an interactive model. In the model, there is also lateral inhibition; forms at a given level inhibit forms at the same level with which they are incompatible. Lexical

effects on phoneme identification (e.g., the Ganong effect and phonemic restoration) arise from lexical feedback. Given an ambiguous member of a *gift*-to-*kift* continuum, the word *gift* will be activated at the lexical level and will feed activation back to its component phonemes, including /g/, thereby fostering identification of the ambiguous initial consonant as /g/. Lexical feedback also restores missing phonemes in the phonemic restoration effect.

In TRACE, knowledge of transition probabilities is the same as knowledge of words. That is, words and nonwords with high transition probabilities include phoneme sequences that occur frequently in words of the lexicon. TRACE cannot generate the dissociations between effects of lexical knowledge and transition probabilities that both Pitt and McQueen (1998) and Vitevitch and Luce (1998) report. A second shortcoming of TRACE is its way of dealing with the temporally extended character of speech. To permit TRACE to take in utterances over time, McClelland and Elman (1986) used the brute force method of replicating the entire network of feature, phone, and word nodes at many different points in modeled time.

Norris's (1993) recurrent network can handle temporally extended input without massive replication of nodes and links. The network has input nodes that receive as input sets of features for phonemes. The feature sets for successive phonemes are input over time. Input units link to hidden units, which link to output units. There is one set of output units for words and one for phonemes. The hidden units also link to one another over delay lines. It is this aspect of the network that allows it to learn the temporally extended phoneme sequences that constitute words. The network is trained to activate the appropriate output unit for a word when its component phonemes' feature sets are presented over time to the input units and to identify phonemes based on featural input. The network has the notable property that it is feedforward only; that is, in contrast to TRACE, there is no top-down feedback from a lexical to a prelexical level. Recurrent networks are good at learning sequences, and the learning resides in the hidden units. Accordingly, the hidden units have probabilistic phonotactic knowledge. Norris has shown that this model can exhibit the Ganong effect and compensation for coarticulation; before its time, it demonstrated findings like those of Pitt and McQueen (1998) in which apparently top-down lexical effects on compensation for coarticulation in fact arise prelexically and depend on knowledge of transition probabilities. This type of model (see also Norris et al., 1999) is remarkably successful in simulating findings that had previously been ascribed to top-down feedback. However, the debate about feedback is ongoing (e.g., Samuel, 2000).

SUMMARY

Intensive research on language forms within experimental psychology has only a 50-year history, beginning with the work by Liberman and colleagues at Haskins Laboratories. However, this chapter shows that much has been learned in that short time. Moreover, the scope of the research has broadened considerably, from an initial focus on speech perception only to current research spanning the domains of competence, planning production, and perception. Additionally, in each domain, the experimental methodolgies developed by investigators have expanded and include some remarkably useful ways of probing the psychology of phonology.

Theoretical developments have been considerable, too. Within each domain, competing theoretical views have grown that foster efforts to sharpen the theories and to distinguish them experimentally. Moreover, we now have theories in domains, such as planning, where earlier there were none. The scope and depth of our understanding of language forms and their role in language use has grown impressively. A relatively new development that is proving very useful is the use of models that implement theories. The models of Dell (1986) and Levelt et al. (1999) of phonological planning, of Guenther et al. (1998) and Saltzman (1991) on speech production, and of McClelland and Elman (1986) and Norris (1994), among others, of speech perception all help to make theoretical differences explicit and theoretical claims testable.

We have much more to learn, of course. My own view, made clear in this chapter, is that enduring advances depend on more cross-talk across the domains of competence, planning, production, and perception.

REFERENCES

Abbs, J., & Gracco, V. (1984). Control of complex gestures: Orofacial muscle responses to load perturbations of the lip during speech. *Journal of Neurophysiology, 51,* 705–723.

Abler, W. (1989). On the particulate principle of self-diversifying systems. *Journal of Social and Biological Structures, 12,* 1–13.

Baars, B., Motley, M., & MacKay, D. G. (1975). Output editing for lexical status from artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior, 14,* 382–391.

Bell-Berti, F., & Harris, K. (1981). A temporal model of speech production. *Phonetica, 38,* 9–20.

Bell-Berti, F., & Krakow, R. (1991). Anticipatory velar lowering: A coproduction account. *Journal of the Acoustical Society of America, 90,* 112–123.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. Goodman & H. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167–224). Cambridge, MA: MIT Press.

Best, C. T. (1995). A direct realist perspective on cross-language speech perception. In W. Strange & J. J. Jenkins (Eds.), *Cross-language speech perception* (pp. 171–204). Timonium, MD: York Press.

Bladon, A., & Al-Bamerni, A. (1976). Coarticulation resistance in English /l/. *Journal of Phonetics, 4,* 137–150.

Boyce, S., Krakow, R., Bell-Berti, F., & Gelfer, C. (1990). Converging sources of evidence for dissecting articulatory movements into gestures. *Journal of Phonetics, 18,* 173–188.

Brancazio, L. (2002). *Lexical influences in audiovisual speech perception.* Manuscript submitted for publication.

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

Browman, C., & Goldstein, L. (1990). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. Beckman (Eds.), *Papers in laboratory phonology: Vol. 1. Between the grammar and the physics of speech* (pp. 341–376). Cambridge, England: Cambridge University Press.

Browman, C., & Goldstein, L. (1995a). Dynamics and articulatory phonology. In R. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition* (pp. 175–193). Cambridge, MA: MIT Press.

Browman, C., & Goldstein, L. (1995b). Gestural syllable position effects in American English. In F. Bell-Berti & L. Raphael (Eds.), *Producing speech: Contemporary issues.* (pp. 19–33). New York: American Institute of Physics.

Browman, C., & Goldstein, L. (1997). The gestural phonology mode. In W. Hulstijn, H. F. M. Peters, & P. H. H. Van Lieshout (Eds.), *Speech production: Motor control, brain research and fluency disorders* (International Congress Series, No. 1146). Amsterdam: Elsevier.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* New York: Harper and Row.

Clark, H. (1996). *Using language.* Cambridge, England: Cambridge University Press.

Delattre, P., & Freeman, D. (1968). A dialect study of American r's by x-ray motion picture. *Linguistics, 44,* 29–68.

Dell, G. S. (1986). A spreading-activation theory of retrieval in speech production. *Psychological Review, 93,* 283–321.

Dell, G. S., Burger, L., & Svec, W. (1997). Language production and serial order: A functional analysis and a model. *Psychological Review, 104,* 127–147.

Dell, G. S., & Juliano, C. (1996). Computational models of phonological encoding. In T. Dijkstra & K. DeSmedt (Eds.), *Computational psycholinguistics: AI and connectionist models of language processing* (pp. 328–359). Philadelphia: Taylor & Francis.

Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints. *Cognitive Science, 17,* 149–195.

Diehl, R., & Kluender, K. (1989). On the objects of speech perception. *Ecological Psychology, 1,* 121–144.

Easton, T. (1972). On the normal use of reflexes. *American Scientist, 60,* 591–599.

Eimas, P., Siqueland, E., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 171,* 303–306.

Elman, J. & McClelland, J. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language, 27,* 143–165.

Faber, A., & Brown, J. (1998). The effect of consonant context on vowel goodness rating [Abstract]. *Journal of the Acoustical Society of America, 104,* 1759 (abstract).

Fitch, H., Halwes, T., Erickson, D., & Liberman, A. (1980). Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception and Psychophysics, 27,* 343–350.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology.* Cambridge, MA: MIT Press.

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

Fowler, C. A. (1994). Invariants, specifiers cues: An investigation of locus equations as information for place of articulation. *Perception & Psychophysics, 55,* 597–610.

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America, 99,* 1730–1741.

Fowler, C. A. (1999). The orderly output constraint is not wearing any clothes. *Behavioral and Brain Science, 21,* 265–266.

Fowler, C. A., & Brancazio, L. (2000). Coarticulation resistance of American English consonants and its effects on transconsonantal vowel-to-vowel coarticulation. *Language and Speech, 43,* 1–41.

Fowler, C. A., Brown, J., & Mann, V. (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, C. A., & Dekle, D. J. (1991). Listening with eye and hand: Crossmodal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 816–828.

Fowler, C. A., & Levy, E. (1995). Talker-listener attunements to speech events. *Journal of Contemporary Legal Studies, 6,* 305–328.

Fowler, C. A., & Rosenblum, L. D. (1990). Duplex perception: A comparison of monosyllables and slamming doors. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 742–754.

Fowler, C. A., Rubin, P., Remez, R., & Turvey, M. (1980). Implications for speech production of a general theory of action. In

B. Butterworth (Ed.), *Language production: Vol. 1. Speech and talk* (Vol. 1, pp. 373–420). London: Academic Press.

Fowler, C. A., & Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and Speech, 36,* 171–195.

Fowler, C. A., & Smith, M. (1986). Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In J. Perkell & D. Klatt (Eds.), *Invariance and variability of speech processes* (pp. 123–136). Hillsdale, NJ: Lawrence Erlbaum.

Fox, R. (1984). Effect of lexical status on phonetic categorization. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 526–540.

Fromkin, V. (1973). *Speech errors as linguistic evidence.* The Hague, the Netherlands: Mouton.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 110–125.

Garrett, M. (1980). Levels of processing in speech production. In B. Butterworth (Ed.), *Language production: Vol. 1. Speech and talk* (pp. 177–220). London: Academic Press.

Gibson, J. J. (1966). *The senses considered as perceptual systems.* Boston: Houghton Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

Gick, B. (1999). *The articulatory basis of syllable structure: A study of English glides and liquids.* Unpublished doctoral dissertation. Yale University, New Haven, CT.

Gleason, P., Tuller, B., & Kelso, J. A. S. (1996). Syllable affiliation of final consonant clusters undergoes a phase transition over speaking rates. *Proceedings of the International Conference on Speech and Language Processing* (pp. 276–278). Philadelphia, PA.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22,* 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105,* 251–279.

Grieser, D., & Kuhl, P. (1989). Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology, 25,* 577–588.

Guenther, F., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech. *Psychological Review, 105,* 611–633.

Guilford, J., & Park, D. (1931). The effect of interpolated weights upon comparative judgments. *American Journal of Psychology, 43,* 589–599.

Hagiwara, R. (1995). Acoustic realizations of American /r/ as produced by women and men. *UCLA Working Papers in Phonetics, 90.*

Hamlet, S. (1988). Speech compensation for prosthodontially created palatal asymmetries. *Journal of Speech and Hearing Research, 31,* 48–53.

Hamlet, S., & Stone, M. (1978). Compensatory alveolar consonant production induced by wearing a dental prosthesis. *Journal of Phonetics, 6,* 227–248.

Hintzman, D. (1986). "Schema abstraction" in a multiple trace memory model. *Psychological Review, 93,* 411–428.

Hockett, C. (1955). *A manual of phonetics.* Bloominton: Indiana University Press.

Humbolt, W. Von. (1972). *Linguistic variability and intellectual development.* (G. C. Buck & F. A. Raven, Trans.) Philadelphia: University of Pennsylvania Press. (Original work published 1836)

Jakobson, R., Fant, G., & Halle, M. (1962). *Preliminaries to speech analysis.* Cambridge, MA: MIT Press.

Keating, P. (1983). Comments on the jaw and syllable structure. *Journal of Phonetics, 11,* 401–406.

Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior.* Cambridge, MA: MIT Press.

Kelso, J. A. S., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C. A. (1984). Functionally-specific articulatory cooperation following jaw perturbation during speech: Evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 812–832.

Kent, R., & Tjaden, K. (1997). Brain functions underlying speech. In W. Hardcastle & J. Laver (Eds.), *The handbook of phonetic sciences* (pp. 221–255). Oxford, England: Blackwell Publishers.

Kluender, K. (1994). Speech perception as a tractable problem in cognitive science. In M. A. Gernsbacher (Ed.), *Handbook of Psycholinguistics* (pp. 173–217). San Diego, CA: Academic Press.

Krakow, R. (1989). *The articulatory organization of syllables: A kinematic analysis of labial and velar gestures.* Unpublished doctoral dissertation. Yale University, New Haven, CT.

Krakow, R. (1993). Nonsegmental influences on velum movement patterns: Syllables, segments, stress and speaking rate. In M. Huffman & R. Krakow (Eds.), *Phonetics and phonology: Vol. 5. Nasals, nasalization and the velum* (pp. 87–116). New York: Academic Press.

Krakow, R. (1999). Physiological organization of syllables: A review. *Journal of Phonetics, 27,* 23–54.

Kuhl, P. (1987). The special-mechanisms debate in speech research: Categorization tests on animals and infants. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 355–386). Cambridge, England: Cambridge University Press.

Kuhl, P. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics, 50,* 93–107.

Kuhl, P., & Iverson, P. (1995). Linguistic experience and the perceptual magnet effect. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 121–154). Baltimore: York Press.

Kuhl, P., & Miller, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America, 63,* 905–917.

Kuhl, P., & Miller, J. D. (1982). Discrimination of auditory target dimensions in the presence or absence of variation in a second dimension by infants. *Perception & Psychophysics, 31,* 279–292.

Kuhl, P., Williams, K., Lacerda, F., Stevens, K., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by six months of age. *Science, 255,* 606–608.

Levelt, W. J. M., Roelofs, A., & Meyer, A. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22,* 1–38.

Liberman, A. M. (1957). Some results of research on speech perception. *Journal of the Acoustical Society of America, 29,* 117–123.

Liberman, A. M. (1996). *Speech: A special code.* Cambridge, MA: Bradford Books.

Liberman, A. M., Cooper, F., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74,* 431–461.

Liberman, A. M., Delattre, P., & Cooper, F. (1952). The role of selected stimulus variables in the perception of the unvoiced-stop consonants. *American Journal of Psychology, 65,* 497–516.

Liberman, A. M., & Mattingly, I. (1985). The motor theory revised. *Cognition, 21,* 1–36.

Liberman, A. M., & Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences, 4,* 187–196.

Lieberman, P. (1991). *Uniquely human: Speech, thought and selfless behavior.* Cambridge, MA: Harvard University Press.

Lindblom, B. (1986). Phonetic universals in vowel systems. In J. Ohala & J. Jaeger (Eds.), *Experimental phonology* (pp. 13–44). Orlando, FL: Academic Press.

Lindblom, B. (1990). On the notion of "possible speech sound." *Journal of Phonetics, 18,* 135–142.

Lindblom, B., Lubker, J., & Gay, T. (1979). Formant frequencies of some fixed mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics, 7,* 147–161.

Lindblom, B., & Maddieson, I. (1988). Phonetic universals in consonant systems. In L. Hyman & C. N. Li (Eds.), *Language, speech and mind* (pp. 62–78). London: Routledge.

Lisker, L. (1978). Rapid vs rabid: A catalogue of acoustic features that may cue the distinction. *Haskins Laboratories Status Report on Speech Research, 54,* 127–132.

Lively, S. & Pisoni, D. (1995). On prototypes and phonetic categories: A critical assessment of the perceptual magnet effect in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 1665–1679.

Löfquist, A. (1997). Theories and models of speech production. In W. Hardcastle & J. Laver (Eds.), *The handbook of phonetic sciences* (pp. 405–426). Oxford, England: Blackwell.

Lotto, A., & Kluender, K. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics, 60,* 602–619.

Lotto, A., Kluender, K., & Holt, L. (1997). Perceptual compensation for coarticulation by Japanese quail (*coturnix coturnix japonica*). *Journal of the Acoustical Society of America, 102,* 1134–1140.

Luce, P. (1986). *Neighborhoods of words in the mental lexicon.* Unpublished doctoral dissertation. Indiana University, Bloomington, Indiana.

Luce, P., & Pisoni, D. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19,* 1–36.

Maddieson, I. (1984). *Patterns of sounds.* Cambridge, England: Cambridge University Press.

Mann, V. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics, 28,* 407–412.

Mann, V., & Liberman, A. (1983). Some differences between phonetic and auditory modes of perception. *Cognition, 14,* 211–235.

Mann, V., & Repp, B. (1981). Perceptual assessment of fricative-stop coarticulation. *Journal of the Acoustical Society of America, 69,* 1153–1169.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25,* 71–102.

Marslen-Wilson, W., & Welsh, A. (1978). Processing interaction and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10,* 29–63.

Massaro, D. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Lawrence Erlbaum.

Massaro, D. (1998). *Perceiving talking faces.* Cambridge, MA: MIT Press.

McClelland, J., & Elman, J. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86.

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

Meyer, A. (1990). The time course of phonological encoding in language production: The encoding of successive syllables of a word. *Journal of Memory and Language, 29,* 524–545.

Meyer, A. (1991). The time course of phonological encoding in language production: Phonological encoding inside a syllable. *Journal of Memory and Language, 30,* 69–89.

Meyer, A., & Schrieffers, H. (1991). Phonological facilitation in picture-word interference experiments: Effects of stimulus onset asynchrony and types of interfering stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition, 17,* 1146–1160.

Miller, J. L., & Dexter, E. R. (1988). Effects of speaking rate and lexical status on phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 369–378.

Miller, J. L., & Volaitis, L. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics, 46,* 505–512.

Mish, Frederick C. (1990). *Webster's Ninth New Collegiate Dictionary.* Springfield, MA: Merriam-Webster, Inc.

Motley, M. (1980). Verification of "Freudian slips" and semantically prearticulatory editing via laboratory-induced spoonerisms. In V. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand* (pp. 133–147). New York: Academic Press.

Mowrey, R., & MacKay, I. (1990). Phonological primitives: Electromyographic speech error evidence. *Journal of the Acoustical Society of America, 88,* 1299–1312.

Nearey, T. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America, 101,* 3241–3254.

Newman, R., Sawusch, J., & Luce, P. (1997). Lexical neighborhood effects in phonemic processing. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 873–889.

Norris, D. (1993). Bottom up connectionist models of "interaction." In G. Altmann & R. Shillcock (Eds.), *Cognitive models of speech processing: The second Sperlonga meeting* (pp. 211–234). Hillsdale, NJ: Lawrence Erlbaum.

Norris, D. (1994). Shortlist: A connectionist model of continuous word recognition. *Cognition, 52,* 189–234.

Norris, D. (1995). Signal detection theory and modularity: On being sensitive to the power of bias models of semantic priming. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 935–939.

Norris, D., McQueen, J., & Cutler, A. (1999). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23,* 299–370.

Ohala, J. (1996). Listeners hear sounds not tongues. *Journal of the Acoustical Society of America, 99,* 1718–1728.

Ohman, S. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *JASA, 39,* 151–168.

Orden, G. van, Pennington, B., & Stone, G. (1990). Word identification in reading and the promise of subsymbolic psycholinguistics. *Psychological Review, 97,* 488–522.

Pardo, J., & Fowler, C. A. (1997). Perceiving the causes of coarticulatory acoustic variation: Consonant voicing and vowel pitch. *Perception & Psychophysics, 59,* 1141–1152.

Parker, E. M., Diehl, R. L., & Kluender, K. R. (1986). Trading relations in speech and nonspeech. *Perception & Psychophysics, 39,* 129–142.

Perkell, J., Matthies, M., Svirsky, M., & Jordan, M. (1993). Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: A pilot "motor equivalence" study. *Journal of the Acoustical Society of America, 93,* 2948–2961.

Pinker, S. (1994). *The language instinct.* New York: William Morrow.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic boundaries. *Perception & Psychophysics, 15,* 285–290.

Pitt, M., & McQueen, J. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language, 39,* 347–370.

Porter, R., & Castellanos, F. X. (1980). Speech production measures of speech perception: Rapid shadowing of VCV syllables. *Journal of the Acoustical Society of America, 67,* 1349–1356.

Porter, R., & Lubker, J. (1980). Rapid reproduction of vowel-vowel sequences: Evidence for a fast and direct acoustic-motoric linkage. *Journal of Speech and Hearing Research, 23,* 593–602.

Prince, A., & Smolensky, P. (1993). *Optimality theory: Constraint interaction and satisfaction.* Unpublished manuscript, Rutgers University, New Brunswick, NJ, University of Colorado, Boulder.

Pulleyblank, D. (1997). Optimality theory and features. In D. Archangeli & D. T. Langendoen (Eds.), *Optimality theory: An overview* (pp. 59–101). Malden, MA: Blackwell.

Recasens, D. (1984). V-to-C coarticulation in Catalan VCV sequences: An articulatory and acoustical study. *Journal of Phonetics, 12,* 61–73.

Roelofs, A., & Meyer, A. (1998). Metrical structure in planning the production of spoken words. *Journal of Experimental Psychology: Learning, Memory and Cognition, 24,* 922–939.

Rosenblum, L. D. (1987). Towards an ecological alternative to the motor theory. *Perceiving-acting workshop, 2,* 25–28.

Ryle, G. (1949). *The concept of mind.* New York: Barnes and Noble.

Saltzman, E. (1991). The task dynamic model in speech production. In H. F. M. Peters, W. Hulstijn, & C. W. Starkweather (Eds.), *Speech motor control and stuttering* (pp. 37–52). Amsterdam: Elsevier Science.

Saltzman, E. (1995). Intergestural timing in speech production: Data and modeling. *Proceedings of the XIIIth International Congress of Phonetic Sciences, Stockholm, 2,* 84–88.

Saltzman, E., & Byrd, D. (1999). Dynamic simulations of a phase window model of relative timing. *Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco, 3,* 2275–2278.

Samuel, A. (1981). Phonemic restoration: Insights for a new methodology. *Journal of Experimental Psychology: General, 110,* 474–494.

Samuel, A. (1996). Does lexical information influence the perceptual restoration of phonemes. *Journal of Experimental Psychology: General, 125,* 28–51.

Samuel, A. (2000). Some empirical tests of Merge's architecture. In A. Cutler, J. McQueen, & R. Zondervan (Eds.), *Proceedings of the workshop on spoken word access processes* (pp. 51–54). Nijmegen: Max-Planck-Institute for Psycholinguistics.

Santiago, J., & MacKay, D. G. (1999). Constraining production theories: Principled motivation, consistency, homunculi, underspecification, failed predictions and contrary data. *Behavioral and Brain Sciences, 22,* 55–56.

Savariaux, C., Perrier, P., & Orliaguet, J. P. (1995). Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production. *Journal of the Acoustical Society of America, 98,* 2428–2442.

Sawusch, J., & Gagnon, D. (1995). Auditory coding, cues and coherence in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 635–652.

Sevald, C. A., Dell, G. S., & Cole, J. S. (1995). Syllable structure in speech production: Are syllables chunks or schemas. *Journal of Memory and Language, 34,* 807–820.

Shaiman, S. (1989). Kinematic and electromyographic responses to perturbation of the jaw. *Journal of the Acoustical Society of America, 86,* 78–88.

Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial-ordering mechanism in sentence production. In W. Cooper & E. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett* (pp. 295–342). Hillsdale, NJ: Lawrence Erlbaum.

Shattuck-Hufnagel, S. (1983). Sublexical units and suprasegmental structure in speech production planning. In P. MacNeilage (Ed.), *The production of speech* (pp. 109–136). New York: Springer-Verlag.

Sinnott, J. M. (1974). *A comparison of speech sound discrimination in humans and monkeys.* Unpublished doctoral dissertation, University of Michigan.

Sproat, R., & Fujimura, O. (1993). Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics, 21,* 291–311.

Stetson, R. H. (1951). *Motor phonetics.* Amsterdam: North Holland.

Stevens, K. (1989). On the quantal nature of speech. *Journal of Phonetics, 17,* 3–45.

Stevens, K., & Blumstein, S. (1981). The search for invariant correlates of phonetic features. In P. Eimas & J. Miller (Eds.), *Perspectives on the study of speech* (pp. 1–38). Hillsdale, NJ: Lawrence Erlbaum.

Stone, G., Vanhoy, M., & Van Orden, G. (1997). Perception is a two-way street: Feedforward and feedback in visual word recognition. *Journal of Memory and Language, 36,* 337–359.

Studdert-Kennedy, M. (1998). The particulate origins of language generativity: From syllable to gesture. In J. Hurford, M. Studdert-Kennedy, & C. Knight (Eds.), *Approaches to the evolution of language* (pp. 202–221). Cambridge, England: Cambridge University Press.

Sussman, H., Fruchter, D., Hilbert, J., & Sirosh, J. (1999a). Linear correlates in the speech signal: The orderly output constraint. *Behavioral and Brain Sciences, 21,* 287–299.

Sussman, H., Fruchter, D., Hilbert, J., & Sirosh, J. (1999b). Human speech: A tinkerer's delight. *Behavioral and Brain Sciences, 21,* 287–299.

Trout, J. D. (2001). The biological basis of speech: What to infer from talking to the animals. *Psychological Review, 108,* 523–549.

Tuller, B., Case, P., & Kelso, J. A. S. (1994). The nonlinear dynamics of speech categorization. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 3–16.

Tuller, B., & Kelso, J. A. S. (1990). Phase transitions in speech production and their perceptual consequences. In M. Jeannerod (Ed.), *Attention and performance: Vol. 13. Motor representation and control* (pp. 429–452). Hillsdale, NJ: Lawrence Erlbaum.

Tuller, B., & Kelso, J. A. S. (1991). The production and perception of syllable structure. *Journal of Speech and Hearing Research, 34,* 501–508.

Turvey, M. T. (1990). Coordination. *American Psychologist, 45,* 938–953.

Vitevitch, M., & Luce, P. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science, 9,* 325–329.

Vitevitch, M., & Luce, P. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language, 40,* 374–408.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167,* 392–393.

Warren, R. M. (1985). Criterion shift rule and perceptual homeostasis. *Psychological Review, 92,* 574–584.

Waters, R. S., & Wilson, W. A., (1976). Speech perception by rhesus monkeys: The voicing distinction in synthesized labial and velar stop consonants. *Perception & Psychophysics, 19,* 285–289.

Weiss, P. (1941). Self differentiation of the basic pattern of coordination. *Comparative Psychology Monograph, 17,* 21–96.

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

Zue, V. (1980). *Acoustic characteristics of stop consonants: A controlled study.* Bloomington: Indiana University Linguistics Club.