In M. Arbib (Ed.) The Handbook of Brain Theory and Neural Networks, 2nd Edition. Cambridge:MIT Press. 1072-1076, 2002.

Speech Production

Dani Byrd and Elliot Saltzman

Introduction

Understanding speech production requires a synthesis of perspectives found in physiology, motor control, cognitive science, and linguistics. This article presents work in the areas of motor control, dynamical systems and neural networks, and linguistics that is critical to understanding the functional architecture and characteristics of the speech production system.

Centuries of research in linguistics have provided considerable evidence that there are fundamental cognitive units that structure language. Spoken word forms are not unstructured wholes but rather are composed from a limited inventory of *phonological units* that have no independent meaning but can be (relatively freely) combined and organized in the construction of word forms. While languages differ in their selection of phonological units, within a given language there is a relatively small fixed set. Unlike certain other domains of human movement, in which the existence of component action units remains controversial (see MOTOR PRIMITIVES; ARM AND HAND MOVEMENT CONTROL), the production of speech by the lips, tongue, vocal folds, velum (the port to the nasal passages), and respiratory system can be understood as arising from choreographed linguistic action units.

A variety of micro- to macro-level units have been suggested as phonological units, among them features, gestures, phonemes (roughly, segments), moras, syllables, subsyllabic constituents (such as the syllable onset, nucleus, rime, and coda), gestural structures, and metrical feet (see Ladefoged, 2001). Some of these hypothesized units are mutually exclusive by definition (e.g., features and gestures); others have been assumed to coexist or to be hierarchically structured (e.g., feet and syllables and moras). For example, the word "phone" has three phonemes forming one syllable and could be transcribed /fon/, while the word "bone" (/bon/) contrasts in its initial unit and thereby in its meaning. (Two sounds are *contrastive* in a language if a change from one to the other can potentially change the meaning of a word.) Such pairs in a language are called minimal pairs and are appealed to as evidence for certain phonological units. Other types of evidence for phonological units can be gleaned from languages' word formation processes, language games, speech errors, diachronic language changes, and language acquisition. (And certain of these units, for example phonemes and syllables, form the basis of some orthographic systems.)

However, linguists and speech scientists have recognized that when phonological units are made manifest in word and sentence production, their spatiotemporal realization by the articulatory system, and consequent acoustic character presented to the auditory system, is highly variable and context dependent. In fact, the articulatory movements specific to adjacent units are not sequential in nature but are highly overlapped (i.g., co-articulated) and interactive. This has consequences that make the physical speech signal quite different from its familiar orthographic symbolic representation. In the acoustic domain, there is no invariant realization for a particular phonological unit across different contexts, an observation that has been termed *lack of invariance*. Additionally, the edges or boundaries between units are not implicit in the speech signal, a feature we can refer to as *lack of segmentability*. There are no pauses or "blank spaces" systematically demarcating phonological units—neither gestures, nor segments, nor words. (While there are amplitude and spectral discontinuities in the acoustic signal, they do not always indicate edges, although they may be indicative of certain important information regarding segment identity.) This *parallel transmission* of information in the acoustic signal due to co-articulated articulatory movements results in a highly efficient yet complex perceptual event that encodes and transmits information at high rates (see MOTOR THEORIES OF PERCEPTION).

Efforts to understand the relationship between phonological units that structure words and their variable physical realization in fluent speech are an important component of speech production research. A common view is that certain linguistic information seems to be lexically specified (i.e., encoded in our stable knowledge of a particular word), whereas other aspects of word and phrase production seem best understood as resulting from principled modulations of phonological units in the performance of speaking and by peripheral properties of the physical speech production system. For this reason, the speech production system is sometimes viewed as having two components, one, traditionally referred to as *phonology*, concerned with categorical and linguistically contrastive information, and the other, traditionally referred to as *phonetics*, concerned with gradient, noncontrastive information. However, current work in connectionist and dynamical systems models blurs this dichotomy. Speech scientists' views on the cognitive organization of the speech production process are shaped by their hypotheses regarding:

- the coordinate systems in which controlled variables are defined,
- the dynamic versus symbolic nature of phonological units (i.e., primitives),
- the higher-level organization of these units,
- the role of the speaker-listener relationship in shaping speech behavior,
- child language acquisition yielding adult phonology.

We discuss some of the hallmarks of this research in the following sections.

The Speech Production Apparatus

Speaking involves the orchestrated creation and release of constrictions in the supralaryngeal vocal tract (Figure 1). These constrictions are made by the lips and tongue (tip, body, and root) and serve to shape the resonance frequencies of the vocal tract tube. Many speech sounds are differentiated by the location and degree of these constrictions, and most speech sounds (though not all) are *pulmonic*; that is, they are generated using airflow from the lungs.

The sound excitation sources can be several but are primarily the vibratory airstream generated by rapid opening and closing of the vocal folds (*voicing*) and turbulence noise generated at narrow constrictions (*frication*). Vowels, which are voiced and produced with a relatively less constricted vocal tract than consonants, are differentiated in large measure by their first three resonance frequencies or *formants*, determined by the vocal tract shape. Consonants are differentiated largely by movement of the formant frequencies as constrictions at specific locations along the vocal tract are formed and released, by characteristic noise created at the constrictions themselves, and by the presence, absence, and timing of vocal fold vibration. Additionally, the vocal tract tube has a side branch to the nasal passageways that is opened for certain speech sounds by lowering the velum (soft palate). During speech, air may exit the vocal tract from the mouth (velum closed/raised) or from the nose (velum lowered and an anterior closure in the mouth), or from both (velum lowered and no oral closure). (Further information on the articulation and acoustics of speech can be found in

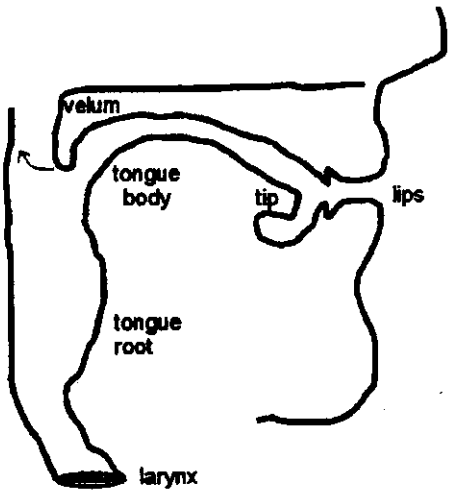

Figure 1. Schematic anatomy of the vocal tract showing the supralaryngeal constrictors (the upper and lower lips, the tongue tip, tongue body, and tongue root); the velum, which rises or lowers to prevent or allow airflow into the nasal passages; and the larynx, which houses the vocal folds that vibrate when adducted under appropriate aerodynamic conditions.

Ladefoged's [2001] *A Course in Phonetics*; for a more sophisticated account of the mapping between vocal tract shape and acoustics, see Stevens' [1998] *Acoustic Phonetics*.) A satisfactory account of human speech production abilities must encompass the great variety of speech sounds used contrastively in the world's languages, including those that contrast in aerodynamic mechanisms, tone, length, and phonation type (see Ladefoged, 2001, and citations therein, e.g., Ladefoged and Maddieson, 1996).

Modeling the Speech Production Process

Linguists have generally adopted a symbolic representation of spoken language. This has proved to be useful in investigating the structure of words and higher-level grammatical processes. Scientists whose primary interest is speech motor control, however, have generally adopted the nonsymbolic formulations provided by dynamical systems and connectionist approaches. The research community interested in spoken language has begun to synthesize these approaches.

Many issues faced in the control and coordination of the speech articulators are the same as those faced in understanding nonspeech skilled actions. In both cases, the multilevel geometry of the system must be specified in terms of a set of appropriate reference frames (coordinate systems) and the set of mappings that is defined among them (see GEOMETRICAL PRINCIPLES IN MOTOR CONTROL; LIMB GEOMETRY, NEURAL CONTROL). Additionally, the appropriate dynamics must be specified within this set of coordinate systems. For speech production, at least four types of coordinate systems and associated dynamics are posited generically in many current models (Figure 2). At the most concrete peripheral level, the *plant* is defined by the actual articulators (e.g., jaw, upper lip) with their neuromuscular (reflexive and muscle activation) and biomechanical dynamics, and may be represented in a coordinate space defined, for example, according to muscle forces and/or equilibrium lengths (see, e.g., Sanguineti, Laboissière, and Ostry, 1998). Commands to this most peripheral level can be shaped with reference to the motions of an internal model of the plant (*model articulators*) (see SENSORIMOTOR LEARNING), whose simulated neuromuscular and biomechanical behavior can provide significant constraints on the spatial patterning and relative timing (e.g., co-articulation) of movement commands. In turn, model articulator trajectories are shaped with reference to a set of task-space coordinates in which the goals of the language's phonological primitives are represented. Although there are differences in the exact nature of these coordinates among models—e.g., acoustic/auditory goals (Bailly, Laboissière, and Schwartz, 1991; Guenther, 1995) versus vocal tract constriction goals (Browman and Goldstein, 1995; Saltzman and Munhall, 1989)—models generally invoke static attractor dynamics to implement these goals. Thus, in all such models, when a particular phonological primitive is activated, the articulators will move in a coordinated manner to create a task-space *gesture* that attains the acoustic, auditory, or constriction targets and remains there until another primitive is activated or the current primitive is deactivated. In this sense, an *articulatory gesture* is a goal-directed movement of the vocal tract represented in a constriction task space and with an intrinsic duration reflecting the time constant of the attractor.

In order for this arrangement to work, a forward model must represent the mapping from current (model) articulator state to task-space state. A task-space dynamics must define the corresponding set of task-space state velocities ("forces"), and an inverse model must map these task-space state velocities to a corresponding set of (model) articulator state velocities. Finally, a set of activation (or "GO-signal"; e.g., Guenther, 1995) coordinates is required to orchestrate the patterning of these gestural units over time and vocal tract space.

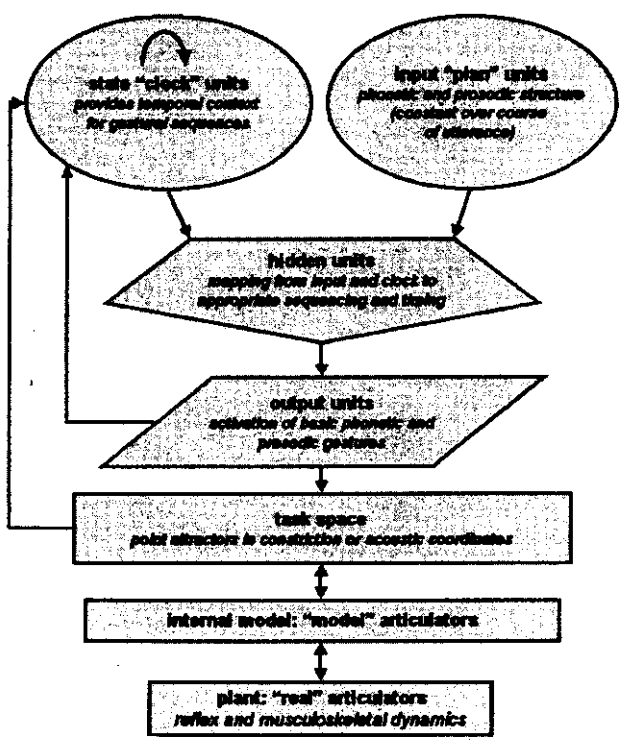

Figure 2. Schematic architecture of the speech production process. Gestural activations are viewed as outputs of a simple recurrent network that drive the articulatory plant through reference signal (model articulator) trajectories generated by task-space dynamics. Feedback connections from the task space and output units to the state unit "clock" modulate clock time flow, owing to the evolving state of the plant and gestural activations, respectively.

Although models generally have adopted connectionist dynamics to shape these activation trajectories, models can be distinguished on the basis of which of two approaches to intergestural timing is adopted. In *chain* models, gestural onsets are triggered whenever an associated preceding gesture either achieves near-zero velocity as it attains its target or passes through another kinematically defined critical point in its trajectory, such as peak tangential velocity (e.g., Guenther, 1995; see also Browman and Goldstein, 1995). In contrast, *clock* models have adopted architectures that are based on or similar to simple recurrent, sequential networks. In these models (e.g., Bailly et al., 1991; Saltzman and Munhall, 1989; see also SEQUENCE LEARNING and references therein), a network's *state unit* activity defines a dynamical flow with a time scale that is intrinsic to the intended sequence and that creates a temporal context within which gestural activations can be shaped by the network's output units. The resultant activation trajectories are determined by the manner in which a static input and the evolving state unit trajectories are mapped nonlinearly onto the output units (see Figure 2). Furthermore, it is noteworthy that many higher-level linguistic properties of both lexical and grammatical encoding can be captured using similar simple recurrent architectures (Dell, Chang, and Griffin, 1999).

Dynamical Units (Gestures) as Phonological Primitives

In our own work we have adopted *articulatory phonology* as a formal account of phonological units and their organization (Browman and Goldstein, 1995, and references to Browman and Goldstein's work therein), and *task dynamics* (Saltzman and Munhall, 1989) as a quantitative model that implements these phonological units in the speech production system in a multilevel dynamical system defined across activation, task-space, and model articulator coordinates. Articulatory phonology views lexical representations as being composed of gestural primitives that act as combinatorial units. Gestures have two functions: they function as units of information (i.e., linguistic contrast) and as units of action (i.e., speech production) (Browman and Goldstein, 1995). The activation waveforms of the gestures in a given utterance are coordinated, or phased, with respect to one another in a highly temporally overlapped pattern that yields the coarticulatory effects that are ubiquitous in speech (Browman and Goldstein, 1995).

To this point, we have mainly discussed the dynamic gesture as the primitive unit of organization in speech production. But gestures can additionally cohere in macro-level structures. Studdert-Kennedy and Goldstein (2002) describe this using the metaphor of gestures as atoms that combine in regular patterns of coordination with one another to form molecules corresponding to larger phonological units such as segments and syllables. Segment-sized units can be viewed as coherent ions—combinations of gestural atoms (one or several) that recur in many different molecules (Studdert-Kennedy and Goldstein, 2002; see also Saltzman and Munhall, 1989). In this way, macro-level phonological structure can be viewed as emerging from micro-level gestural primitives.

Although underlyingly invariant (or at least relatively stable) control units have been postulated in speech production, and certainly for lexical representation, the exact spatiotemporal realization of these units varies according to both local and prosodic context. By way of example, consider the three /o/'s in the sentence "He said *phone* not folk on the telephone": the [o] in "phone" will differ from the [o] in "folk" because of the different following consonantal context (e.g., there will be nasal airflow during the [o] in "phone," yielding different formant patterns). It will also differ from the [o] in "telephone" because of the emphatic stress placed on it in its first occurrence in the sentence. These are examples of co-articulatory variability and prosodic variability, respectively. Variation due to neighboring articulation (local context) is straightforwardly accounted for by the overlap of gestural units of action (e.g., Saltzman and Munhall, 1989; Browman and Goldstein, 1995, and references therein). However, variation due to prosody (e.g., phrasal position or informational prominence) requires the expression of the underlying primitives to be modulated for communicative ends in the production of a particular utterance. One approach we have pursued to implementing this modulation is via prosodic gestures that have no independent realization in vocal tract space but act vicariously to shape the time course of constriction (or auditory) gestures (Byrd et al., 2000). For example, a prosodic gesture at a phrase edge might slow the central clock (whose rate of time flow determines the local utterance rate), thereby time-stretching the gestural activations and inducing the articulatory slowing and acoustic lengthening that have been observed at phrase edges.

Although we and many in the speech production community adopt the general approach of articulatory phonology to linguistic representation, it is not without competition. A sense of this debate can be gleaned from a 1992 theme issue of the journal *Phonetica* on articulatory phonology (see Browman and Goldstein, 1992, cited in Browman and Goldstein, 1995). This view stands in contrast to a more traditional view among linguists that sees phonological primitives as symbolic (atemporal) elements such as features and segments rather than dynamic gestures. In the symbolic approach, the smallest units are typically binary features, such as [−continuant], [+aspirated], [+labial] (which could jointly describe the segment /p/ in /pat/), that are hierarchically incorporated into larger phonological units such as the syllable, (phonological) word, and phrase. These phonological units are then realized according to various and sundry phonetic implementation rules that mediate between lexical representation and physical realization.

The Role of the Speaker-Listener Relationship in Performance

Speech behavior is adaptive to communicative and situational demands (Lindblom, 1990), and for this reason, listeners can play an important role in shaping the production of the speech signal. The speaker-listener interaction might affect word forms diachronically in the form of language change (see, e.g., Ohala, 1993; see also LANGUAGE EVOLUTION AND CHANGE) or synchronically as a function of the listener's abilities, the opaqueness of the signal, and the environment in which the communication task is taking place. Lindblom (1990) credits the speaker with a "tacit awareness of the listener's access to sources of information independent of the signal and his judgement of the short-term demands for explicit signal information" (p. 403). In Lindblom's view, it is crucial to characterize what constitutes sufficient discriminability in the signal for the listener and how speakers operate to balance the benefits of providing this discriminability against the costs of articulatory precision and effort. This viewpoint diverges from theories of speech production in which the listener plays little or no role in shaping on-line speaker behavior.

The Organization of Speech: Adult Phonology

Whether the units of spoken language are symbolic or dynamic, they must be acquired in learning and utilized in behavior. The acquisition of the units of speech production and the organization of those units in the developing and adult lexicon into patterns appropriate to the language being acquired are broad topics that have just begun to be explored computationally.

While languages vary not only in their inventories of contrastive units but also in the structures (sequential, syllabic, rhythmic) into which those units are marshaled, they also show a large degree of agreement regarding the factors affecting the linguistic acceptabil-

ity of word forms. A theoretical account of the bases for these cross-linguistic differences and similarities is provided by Prince and Smolensky's Optimality Theory (see, e.g., Prince and Smolensky, 1997; see also OPTIMALITY THEORY IN LINGUISTICS). This approach to phonology (and, by extension, grammar in general) capitalizes on the idea that constraints determining well-formedness of linguistic structure are many but finite, universal (i.e., shared by all human languages), and in conflict within any particular language. The individual constraints are ranked differently from language to language and thereby yield observed cross-language typology. Constraints are thought to concern structural complexity (*markedness* constraints) (where complexity might be conceived, for example, in terms of production, perception [especially distinctiveness], and/or processing) and the relationship of a produced form to some underlying, analogous, or related form (*faithfulness* constraints). All constraints are violable (and indeed many constraints will be violated by any particular word form); however, the optimal candidate word form will be the one that avoids *any* violation of a higher-ranked constraint, regardless of the number of violations of lower-ranked constraints; that is, evaluation is via *strict domination* (Prince and Smolensky, 1997). Prince and Smolensky (1997) have described how the process of selecting optimal word forms might be modeled using connectionist networks in which (1) constraints are embodied in the network connection weights, and (2) optimality is defined according to Lyapunov function ("harmony") values corresponding to activation patterns of the network. Given a particular input, held fixed across a given set of network elements, the network will settle to a pattern that maximizes harmony and that corresponds to the most well-formed linguistic structure. The universality of constraints and the adequacy of particular evaluatory mechanisms for constraint satisfaction are topics of debate. For example, see OPTIMALITY THEORY IN LINGUISTICS for a discussion of probabilistic and variable constraint ranking.

The Organization of Speech: Child Language Acquisition

Finally, we wish to briefly mention the difficult problem of how phonological units emerge in child language production from a signal that clearly cannot be characterized as a sequence of discrete units but is best viewed as an intricately overlapping pattern of vocal tract or auditory goals. The child learner is thus faced with the challenges of *lack of invariance* and *lack of segmentability* in the continuous audiovisual signals with which she is confronted. (Of course, the adult perceiver is in a similar circumstance but brings a much richer semantic and syntactic knowledge, gained through years of experience, to the task.) Further, the child learner faces the additional difficulty of an immature vocal tract apparatus. The acoustics resulting from linguistically significant vocal tract actions of the child are, in certain respects, vastly different from the acoustic properties resulting from articulatorily parallel gestures of the adult (e.g., formant frequencies and fundamental [voicing] frequency are much higher), though their spectral and temporal *patterning* may, importantly, be similar. Further, the child's vocal tract is growing over time, resulting in ongoing changes in the relationship between the child's own articulatory gestures and their acoustic consequences, changes that must be reflected in adjustment to the production system's internal model during development. A recent special issue of the journal *Phonetica* on emergence and adaptation (2000) includes many illuminating articles relevant to the acquisition and development of linguistic systems. In one, Michael Studdert-Kennedy (2000) discusses the emergence of the gestural unit in the process of child language acquisition as occurring via an engagement of the child's own vocal apparatus and its behavioral consequences. He speculates that exemplar models of learning, facial imitation, and mirror neuron systems might have roles to play in elucidating how children form gestural units defined in terms of vocal tract constrictions in the language acquisition process (Studdert-Kennedy, 2000; see also Studdert-Kennedy and Goldstein, 2002).

Although computational modeling on the acquisition of speech production (i.e., articulation) is still in its infancy (but see Guenther, 1995, and several follow-up articles on his DIVA model for an illuminating treatment of the development of the production system), it is clear that there is an interdependent relationship between the codeveloping perceptual and production systems that relies on a perceptuomotor link (see MOTOR THEORIES OF PERCEPTION), whether learned or innate, and on experience with the phonological and lexical patterning within the child's language. Marilyn Vihman's 1996 book, *Phonological Development*, is an ideal starting point for exploring this relationship, and Beckman and Edwards (2000) specifically address the importance of experience and lexical patterning in the acquisition of spoken language.

Discussion

In order for an individual to articulate a language, she must know words of that language, know how to combine words into phrases, and be able to instantiate those phrases in the physical world through the use of body effectors. Furthermore, this act generally takes place socially in a communicative context involving a perceiver. All of these aspects of producing language shape the research agenda in the field of speech production. In this article we reviewed hypotheses regarding the nature of the units that serve to form words and the architecture of the production system that executes those units. We briefly touched on how the child learner might accomplish this, how the perceiver might play a role in shaping production, and how patterns of word forms within and among languages might be characterized. Although many challenges remain, not the least of which is connecting these insights to theories of brain, it is clear that an interdisciplinary approach involving motor control, cognitive and brain science, and linguistics is essential.

Road Map: Linguistics and Speech Processing
Related Reading: Motor Theories of Perception; Speech Processing: Psycholinguistics

References

Bailly, G., Laboissière, R. L., and Schwartz, J. L., 1991, Formant trajectories as audible gestures: An alternative for speech synthesis, *J. Phonet.*, 19:9–23.

Beckman, M., and Edwards, J., 2000, The ontogeny of phonological categories and the primacy of lexical learning in linguistic development, *Child Dev.*, 71:240–249.

Browman, C., and Goldstein, L., 1995, Dynamics and articulatory phonology, in *Mind as Motion* (R. F. Port and T. van Gelder, Eds.), Cambridge, MA: MIT Press, pp. 175–194. ◆

Byrd, D., Kaun, A., Narayanan, S., and Saltzman, E., 2000, Phrasal signatures in articulation, in *Papers in Laboratory Phonology V* (M. B. Broe and J. B. Pierrehumbert, Eds.), Cambridge, Engl.: Cambridge University Press, pp. 70–87.

Dell, G., Chang, S. F., and Griffin, Z. M., 1999, Connectionist models of language production: Lexical access and grammatical encoding, *Cognit. Sci.*, 23:517–542.

Guenther, F. H., 1995, Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production, *Psychol. Rev.*, 102:594–621.

Ladefoged, P., 2001, *A Course in Phonetics*, 4th ed., Orlando, FL: Harcourt. ◆

Lindblom, B., 1990, Explaining phonetic variation: A sketch of the H&H theory, in *Speech Production and Speech Modeling* (W. Hardcastle and A. Marchal, Eds.), Dordrecht: Kluwer, pp. 403–439.

Ohala, J. J., 1993, Sound change as nature's speech perception experiment, *Speech Commun.*, 13:155–161.
Prince, A., and Smolensky, P., 1997, Optimality: From neural networks to universal grammar, *Science*, 275:1604–1610. ◆
Saltzman, E. L., and Munhall, K. G., 1989, A dynamical approach to gestural patterning in speech production, *Ecol. Psychol.*, 1:333–382.
Sanguineti, V., Laboissière, R., and Ostry, D. J., 1998, A dynamic biomechanical model for neural control of speech production, *J. Acoust. Soc. Am.*, 103:1615–1627.
Stevens, K., 1998, *Acoustic Phonetics*, Cambridge, MA: MIT Press.
Studdert-Kennedy, M., 2000, Imitation and the emergence of segments, *Phonetica*, 57:275–283.
Studdert-Kennedy, M., and Goldstein, L., 2002, Launching language: The gestural origin of discrete infinity, in *Language Evolution: The States of the Art* (M. Christiansen and S. Kirby, Eds.), Oxford, Engl.: Oxford University Press. ◆
Vihman, M., 1996, *Phonological Development*, Cambridge, Engl.: Blackwell. ◆