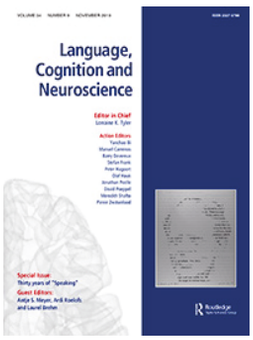

Routledge
Taylor & Francis Group

Language, Cognition and Neuroscience

ISSN: 2327-3798 (Print) 2327-3801 (Online) Journal homepage: https://www.tandfonline.com/loi/plcp21

Toward an (even) more comprehensive model of speech production planning

Stefanie Shattuck-Hufnagel

To cite this article: Stefanie Shattuck-Hufnagel (2019) Toward an (even) more comprehensive model of speech production planning, Language, Cognition and Neuroscience, 34:9, 1202-1213, DOI: 10.1080/23273798.2019.1650944

To link to this article: https://doi.org/10.1080/23273798.2019.1650944

LANGUAGE, COGNITION AND NEUROSCIENCE
2019, VOL. 34, NO. 9, 1202–1213
https://doi.org/10.1080/23273798.2019.1650944

Routledge
Taylor & Francis Group

REVIEW ARTICLE

Toward an (even) more comprehensive model of speech production planning

Stefanie Shattuck-Hufnagel

Research Laboratory of Electronics, MIT, Cambridge, MA, USA

ABSTRACT

Since the publication of *Speaking* in 1989, with its extraordinary goal of modelling the entire process of human speech generation from message conceptualisation to articulation, encompassing results from a wide range of empirical studies, much new information has emerged about three aspects of speech production that were not clearly in focus at that time. This evidence has revealed 1) the systematic patterns of context-governed surface phonetic variation, and the active control of these patterns exercised by speakers and listeners, 2) the depth and pervasiveness of prosodic influences on those patterns, and 3) the close alignment of co-speech gestures with the prosodic structure of an utterance. This paper reviews some of that evidence, and suggests how its implications may constrain models of speech production planning, as those models become more comprehensive in their treatment of higher-level structures, and of aspects of the communicative act beyond the articulation of lexical and syntactic elements.

ARTICLE HISTORY
Received 16 October 2018
Accepted 25 June 2019

KEYWORDS
Speech production planning; feature-cue-based processing; prosodic control of phonetics; co-speech gesture; prosody-gesture timing

1. *Speaking* as a comprehensive model/ treatment of the human speech production process

One of most striking accomplishments in Levelt's (1989) volume *Speaking* was the comprehensive nature of its treatment of the human speech production process. Starting with the formulation of an intended message in the mind of the speaker, and ending with the provision of instructions to the articulatory apparatus, the book knitted together what was currently known about that process from a wide variety of sources. Levelt saw both the need for and the value of formulating a comprehensive model, to ensure that the various components do not violate each others' assumptions. Soon many researchers were adopting this approach, testing hypotheses about individual steps in the process within the framework of the overall model presented in *Speaking* and elaborated/amended in Levelt, Roelofs, and Meyer (1999), henceforth LRM99. Other researchers were inspired by the explicitness of the model to carry out analyses from different points of view (Caramazza, 1997; Caramazza & Miozzo, 1997; Miozzo & Caramazza, 1997; Dell, Burger, & Svec, 1997), resulting in some lively interchanges (e.g. Roelofs, Meyer, & Levelt, 1998; Schiller & Caramazza, 2002; see also articles in this volume by many of these authors).

As has often been noted, some of the topics most actively investigated in this followup activity have been the process by which individual words are selected in the lexicon, and the process of producing an isolated word as an utterance. In fact, Levelt et al.'s (1999) paper in Behavioral and Brain Science was explicitly focused on the part of the model concerned with the retrieval and encoding of single words. This focus grew out of the observation that, as of 1989, there was little information available about this part of the speaking process, and the resulting series of experiments by the Nijmegen group were designed to explore it empirically (Levelt, 2002a). In addition, investigators have been making good use of techniques for brain imaging that have emerged over the intervening years, to enrich our understanding of what parts of the brain are active in various stages of the planning and production of a spoken utterance, and to develop models of brain function during those processes. (See papers in this issue for surveys of these topics.)

At the same time, there have been extensive developments in the linguistic and cognitive/psycholinguistic literature that bear on the question of how speakers plan and produce an entire utterance. Three major developments of this kind are of particular concern here. The first relates to context-governed surface phonetic variation: the accumulation of evidence that speakers (and listeners) attend to and control detailed aspects of the acoustic speech signal that are not in themselves contrastive (in the sense of distinguishing one word or morpheme from another), but are nevertheless informative about other aspects of the spoken utterance or of the

CONTACT Stefanie Shattuck-Hufnagel sshuf@mit.edu

speaking situation. The pervasive and sometimes-extreme nature of these patterns of variation, and their dependence on context, raise the question of how they are to be dealt with in models of the utterance planning process, and whether retrieval of canonical articulatory patterns for individual syllables, adopted in the Nijmegen approach, is the most appropriate mechanism for formulating utterance-specific instructions for the articulatory control system. The second development is the emergence of a modern theory of higher-level prosodic constituents and prominence patterns, along with ample behavioural evidence for the role of hierarchical prosodic structure in governing systematic patterns of surface phonetic variation. These developments have added urgency to the question of how phrase-level prosodic contours are planned, how meaning-appropriate intonational targets are selected, and how to model the effect of prosodic constituent structure and prominence on surface phonetic detail, in the form of e.g. boundary- and accent-related lengthening, and various other aspects of lenition and fortition. The third development concerns co-speech gesture: a set of observations that have revealed the central role of co-speech movements of the hands, face and other parts of the body in signalling aspects of a speaker's intended message, along with data indicating how those gestures intersect with the prosody of the spoken signal. The aim of this paper is to summarise some of the most pertinent findings in each of these three areas, i.e. extreme yet context-systematic surface phonetic variation, prosodic governance of surface form, and gesture-speech interaction, and to explore their implications for further extending both the conceptual version of the Nijmegen approach described in Levelt (1989), and in the implemented LRM99 model of word-form encoding of single words. An important aspect of these desireable future developments is that they be consistent with the model's unique spirit of addressing all, or at least most, of the known facts about human speech production planning. As we shall see, the various presentations of the Nijmegen approach, taken together, have presaged most of these questions, even if (like most production models) it has not developed explicit answers to them.

The body of this paper is structured into three sections which address relevant findings in these three areas. Section 2 addresses systematic patterns of context-governed surface phonetic variation, and their implications for the nature of speech planning. Section 3 summarises findings about how phrase-level prosodic structure governs surface phonetics, and their implications for the role of higher-level prosody in the planning process. Finally, Section 4 presents findings about the connections between spoken utterances and the gestures of hands and other body parts that often accompany those utterances, suggesting the need to integrate these two streams of information in production.

2. The feature-cue-based nature of context-governed surface phonetic variation and its implications for speech planning models

One of the most important discoveries in speech science over the past few decades has been the degree to which speakers and listeners attend to, control and make use of detailed acoustic aspects of the speech signal—a level of detail that in previous decades had been largely viewed as noise. This misapprehension of variability as noise came about because of a fundamental flaw in our understanding of the relationship between the contrastive phonemic categories that define a word, and how those categories are implemented in the speech signal. This flaw lay in the assumption that if words are represented in the lexicon in terms of discrete serially-ordered bundles of distinctive features, then these phonemic segments should necessarily be observable as discrete, identifiable units in the speech wave form. Hockett (1955) may have been one of the first to point out that the implementation process that maps phonological representations to their phonetic realisation in the speech signal does not result in a simple 1:1 relationship. He illustrated this observation in a famous (but partially misleading) analogy, describing this relationship as similar to the process by which a sequence of separate Easter eggs with specific designs might be run through an old-fashioned washer-wringer roll. He noted that what would emerge from such a process would be an unsegmented mass of material, with little trace of its original structure as discrete serially-ordered elements. This image vividly captures the degree to which the surface phonetic form of the words in an utterance often fails to directly reflect the segmental character of their phonological representations in the lexicon, but it is misleading in that it fails to capture the highly-systematic relationship that nevertheless can be detected between that segmental representation and the resulting surface phonetic outcome (see Pierrehumbert, 1990; Pierrehumbert, 2016; Kazanina, Bowers, & Idsardi, 2017 for discussion). As a growing number of research reports have documented, this relationship is far from random. And because it is systematic, even a highly-reduced utterance conveys information both about the features and phonemes of the speaker's intended words (see Kazanina et al., 2017 for discussion), and about the context which influenced the specific pattern of reduction or variation. Moreover, the facts

about surface phonetic variation suggest the need for an important change in our understanding of the nature of the units that must be accounted for in a comprehensive model of the speech production planning process. That is, these findings suggest that speakers explicitly represent and manipulate individual acoustic cues to the distinctive features of contrastive speech sounds.

The smearing or overlapping of information about successive sound segments when they are realised in the speech signal, as well as the fact that this process leaves critical cues behind (termed the phonetic "residue" by Kohler (1999)), is illustrated by many well-known observations. One particularly familiar example is the pattern of vowel formant transitions as the speaker forms or releases an oral constriction for an adjacent consonant. In such regions, information about the vowel and information about both the following (or preceding) consonant *and* the vowel is present during the same time period, but these ovelapping cues are parsed out appropriately by a listener. Equally familiar is the finding that, in American English, the vowel before a coda consonant is longer when the coda is voiced and shorter when the coda is unvoiced, at least in utterance-final position (called "prepausal" in Crystal & House, 1988 region), where the vowel is generally long enough to reliably signal this difference. Thus in such cases, the region in the signal that is usually thought of as most closely related to the vowel contains cues to the distinctive features of the following consonant, and listeners have little difficulty in parsing the duration cue in this region as evidence for the phonological category of the coda consonant.

Another illustration of the cue-overlap phenomenon arises when one segment appears to have taken on the characteristics of an adjacent segment, as when e.g. the coda /n/ in *one guy* appears to be produced as an /ng/, apparently adopting the place feature of the following /g/. One of the most important insights contributed to our understanding of such phenomena by the developers of Articulatory Phonology (Browman & Goldstein, 1986 et seq.) is that many such cases do not involve a change in the feature specification at all; instead, both of the appropriate constriction gestures occur, but there is extensive temporal overlap between the two gestures, so that their acoustic consequences overlap in time as well. However, interestingly, it appears that speakers can differ in whether they employ gestural overlap or feature change in such cases (Ellis & Hardcastle, 2002), and there is evidence that some instances of utterance-specific context adjustments are accomplished by one of these mechanisms vs the other. For example, Zsiga (1997) showed that in Igbo, the process of vowel harmony involves a binary feature change, while the coarticulation of adjacent vowels is process that involves gradient articulatory overlap. Moreover, whether overlap or feature change occurs, it appears that listeners can quite easily parse the overlapping cues into evidence for the successive underlying phonemes. This was demonstrated by Manuel (1995), who showed that listeners interpreted the interdental nasal in *win those* in English as evidence for an alveolar nasal /n/ followed by a dental /dh/, and by Gow (2001), who showed (using a cross-modal lexical decision task) that listeners correctly perceived the /-t + b-/sequence in a heavily coarticulated utterance of e.g. *right berries*, instead of *ripe berries*, even though they reported hearing *ripe berries* in an explicit naming task.) Hawkins (2011) provides evidence for similar listener sensitivity in the case of phonetic signalling of morphemic contrasts, in eye movement studies showing that the non-morphemic *mis-* in e.g. *mistake* is perceived differently from the morphemic *mis-* in e.g. *mis-type*.

The understanding that co-articulation involves gestural overlap, at least in many cases, illustrates clearly that the acoustic cues in a speech signal to two successive phonological elements of the words of an utterance may not always be separate in time. But this is far from the only way in which the tidy arrangement of successive phonological segments of an utterance can lose their beads-on-a-string structure when produced in a spoken utterance. Continuous speech, particularly in casual speaking contexts, can undergo even more radical changes in the cues to the speaker's intended words and their phonemic segments. A range of studies have illustrated the severity of what Johnson (2004) has called "massive reductions", and have demonstrated their pervasive occurrence in typical communicative speech, as well as the systematic nature of their relationship both to the phonemic representations of the words of an utterance and to the contexts in which they occur. Familiar examples of massive reductions in English include both single words (e.g. *probably* produced as ~*prah-y*) and word sequences (e.g. *why don't you* produced as something like ~*wyncha*), usually involving words or phrases that have a high frequency of occurrence or are highly predictable in context (Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Turk and White 1999; Bybee, 2009).

Even when such reduced productions reflect little of the original syllable structure or organisation into a sequence of discrete phonemic segments, they nevertheless often include individual cues to some of the features of the word's (or words', in the case of reduced word sequences or phrases) defining phonemes. Moreover, these few cues can be enough to allow the listener to recognise the word in context, as demonstrated by

Niebuhr and Kohler (2011). Their study showed that the German adverb *eigentlich* can sometimes be so extremely reduced that it closely resembles the word *eine*; nevertheless, German listeners can use the few residual acoustic cues (available in the duration, and the distribution of palatality and nasality in the signal) to distinguish between the two candidates.

These observations have profound implications for models of speech production planning, because they make it clear that an adequate model must provide a mechanism by which extreme smearing and reduction of information about sequences of phonemic elements can occur. For example, a model such as LRM99, in which pre-compiled articulatory plans for syllables are retrieved, will require considerable elaboration of those syllable-sized articulatory plans to account for e.g. highly-reduced (or sometimes, in contrast, greatly-strengthened) productions of words like *probably* in American English. Compare, for example, the highly reduced form of this word mentioned above (something like ~prah-y), with the exaggerated form produced when the word carries two pitch accents, a low on the first syllable and a high on the last, followed by a low boundary tone (where both /b/ segments and the /l/ are clearly articulated.)

Levelt (2002b) points to this issue himself, noting that, after retrieval of the stored syllable-sized motor programmes (proposed in Articulatory Phonology), "In addition, phonetic encoding involves the further coarticulatory integration of successive articulatory syllables" (p.5). Cases of within-word and across-word-boundary overlap, as well as massive reductions, suggest that this further integration may be so extreme as to completely transform the syllable-sized motor programmes. In fact, the pervasive and often extreme nature of such reductions (and strengthenings) raises the question of whether it would be desireable to consider an alternative to the retrieval of stored syllable-sized articulatory patterns, avoiding the problem that such stored patterns must subsequently be altered in complex ways for particular utterances.

Such an alternative is suggested by the individual-feature-cue-based processing representations proposed for speech perception by Stevens (2002). This model draws on Halle's (1992) proposal for partitioning features into two classes: *articulator-free features* (such as [consonantal], [vocalic] and [continuant]) that do not specify an articulator, and *articulator-bound features* (such as [labial] and [nasal]) that do specify an articulator. Stevens proposed that an early step in the speech recognition process is to detect the abrupt acoustic events that signal articulator-free features, (i.e. "acoustic edges" in the spectrogram, produced by consonantal closures and releases, as well as vowel maxima and glide minima), which he termed Landmarks. Stevens proposed further that these Landmark cues in the signal are interpreted as evidence for the distinctive manner features of the sounds intended by the speaker. These Landmark events also identify regions in the signal that are rich in additional cues to the second type of features, the articulator-bound features. These include cues to place (such as formant transitions into a consonant closure or out of a consonant release) and to voicing. In Stevens' model of the perception process, later steps involve integrating information about cues to both of these types of features, articulator-free and articulator-bound, into sequences of feature-bundle representations, i.e. phonemes. The key idea in this proposal for our purposes here is the proposal that identifying individual feature cues is a critical step in the process; if individual feature cues are significant and discrete aspects of the representations used in the perception of spoken utterances, then these acoustic cues are also candidates for representations that are actively used in the production planning process.

Such a feature-cue-based approach has been taken by Turk & Shattuck-Hufnagel (forthcoming) as a part of their three-component model of word-form encoding during the process of utterance planning. This model critically invokes symbol-based lexical representations and a planning process based on individual acoustic cues to features. It proposes that individual context-appropriate feature cues are selected for a particular utterance during the operation of a Phonological Planning Component, while the quantitative values of those acoustic cues (and the articulatory movements required to generate them) are subsequently computed during the operation of a separate Phonetic Planning Component. The output of the Phonetic Planning Component provides input to the third component, for Motor-Sensory Implementation, which tracks the movements of the articulators and adjusts them to ensure that the targets are reached on time. Turk and Shattuck-Hufnagel propose this structure in part because its reliance on the selection of individual context-appropriate acoustic cues to features, and subsequent computation of the quantitative values for those acoustic cues, provides an account of context-specific adjustments in surface phonetic form, including massive reductions or lenitions (and of what might be called massive fortitions, in especially clear speech) which reflect differences in the choice of cues and gradient adjustment of their quantitative values for particular contexts. It also provides a mechanism for the translation from contrastive symbolic phonemic representations in the lexicon to gradient quantitative representations that can provide

instructions to the articulatory tract; this translation process is more difficult to model by the traditional means of selecting context-specific symbolic allophones, due to the gradient nature of many reduction processes and the fact that acoustic cues from several contextual allophones may occur together, as when a coda /t/ is both glottalized and released in American English. While this feature-cue-based model provides a promising approach to accounting for a range of observations about human speech processing in both perception and production, such as the gradient phonetic entrainment of co-interlocutors in conversations (Babel, 2012; Nielson, 2011; Pardo, 2006) and covert contrast in early child speech (Kornfeld, 1971; Macken & Barton, 1980; McAllister Byun, Richtsmeier, & Maas, 2013; Richtsmeier, 2010), it remains for the moment a conceptual framework that awaits implementation for rigorous testing.

Other approaches to modelling the occurrence of highly reduced (or, in contrast, strengthened) productions of words and word sequences include that of Articulatory Phonology, noted above, which relies on gestural overlap in time and/or gestural reduction in space to account for acoustic reduction, and that of Exemplar Theory (Bybee, 2009; also explored in Pierrehumbert, 2001), which relies on the retrieval of stored representations of earlier-produced tokens of the appropriate degree of strength. Although each of these approaches has its attractions, and usefully accounts for aspects of observed surface variation, each also faces challenges. The gesture-based approach does not provide a natural account of observations like the use of alternative articulators to produce cues to the same phonemic category (as when the same speaker produces /r/ with either a tongue tip or tongue dorsum constriction on different occasions (Tiede, Boyce, Espy-Wilson, & Gracco, 2010), or when the same speaker produces /t/ with either a tongue tip closure or a glottal constriction on different occasions (Heyward, Turk, & Geng, 2014)). The exemplar-based approach raises questions about how highly variable exemplars could be produced in the first place, particularly in systematic patterns related to context, in order to be stored in the cloud of related exemplars for later retrieval and re-use. It also leaves unanswered (although perhaps not without the possibility of an answer) the question of how language users represent phonological equivalence. The Feature Underspecified Lexicon approach developed by Lahiri and Reetz (2002, 2010) for speech perception also offers some insights into how the lexical representations that are accessed in production can support surface variation, by opening the door to speaker choice of how to signal a lexically underspecified alveolar consonant in different contexts. But whatever the approach, it appears obligatory that an adequate model of speech production must provide an account of systematic context-governed massive reductions (and fortitions) that involve changes at the level of individual cues to distinctive features. This requirement lays out an interesting path for future extension of the Nijmegen approach, bringing it closer to the ultimate goal of an implemented model that produces not only articulatory movements but also acoustic wave forms.

The next section samples the growing body of evidence that prosodic structure above the level of the word is one of the most powerful factors influencing systematic context-governed surface phonetic variation, and thus must play a significant role in the representations that support speech production planning.

3. The influence of phrase-level prosody on surface phonetic forms and words, and its implications for models of production planning

Systematic patterns of acoustic-phonetic variation in words and their sounds in different contexts have long been observed simply by listening, and by the 1950s the advent of convenient tools for displaying and analyzing speech wave forms made it possible to put quantitative meat on the bones of those perceptual observations. These tools enabled careful measurement of phenomena that had been observed perceptually, such as the effects of adjacent sounds on the realisation of a target sound, and the variation of sounds across positions in a larger constituent (e.g. in American English, aspiration of word-onset voiceless stops even before reduced vowels, as in *tomorrow* or *potato*, but not word-internally, as in *butter* or *iota*, A. Cooper, 1991, and differences in the realisation of /t/ in *sin tax* vs *syntax*, Gow & Gordon, 1995). Studies also began to reveal the quantitative effects of constituent structure on timing, e.g. the lengthening of utterance-final elements (Klatt, 1976) and the shortening of vowel nuclei in polysyllabic words (Lehiste, 1972). But these patterns of surface variation were initially viewed as systematic with respect to lexical structure (words) and syntactic structure (clauses and sentences). A further development enabled discovery of an additional powerful influence on surface phonetics: the emergence of modern prosodic theory, in the form of hierarchical structures of both constituents (Beckman & Pierrehumbert, 1986; Hayes, 1984, 1989; Liberman & Prince, 1977; Nespor & Vogel, 1986; Pierrehumbert, 1980; Pierrehumbert & Beckman, 1988; Selkirk, 1984; see Shattuck-Hufnagel & Turk, 1996 for a review), and prominences (Beckman & Edwards, 1994). While decades later there

is still active discussion of the nature of the constituents in the prosodic hierarchy, and of the degree to which the levels are fixed and discrete (see e.g. Wagner, 2010; Wagner & Watson, 2010), as well as its relation to syntax (Steedman, 2000), few practitioners now doubt that there are levels of prosodic constituents, and that they are correlated with a range of acoustic cues (Brugos, Breen, Veilleux, Barnes, & Shattuck-Hufnagel, 2018; Cole & Shattuck-Hufnagel, 2018).

With the theoretical advances in the understanding of spoken prosody, a wealth of acoustic phonetic analyses of surface phonetic variation emerged, focused on speech samples which were elicited specifically to contrast the effects of different levels of prosodic structure on surface phonetic form. These studies soon revealed that prosodic structure, rather than morpho-phonological structure, was the factor that governed a large proportion of this variation. For example, Ferreira (1993) showed that the duration lengthening at the ends of spoken constituents was better predicted by prosodic constituent structure than by syntactic structure. In the articulatory domain, Fougeron and Keating (1997) showed that articulatory strengthening of a constituent-onset consonant reflected the levels in the prosodic hierarchy. Investigators working in the domain of Articulatory Phonology modelled the gestural mechanisms by which hierarchical degrees of boundary-related lengthening (Byrd & Saltzman, 2003) might occurs, as well as how Lehiste's (1972) finding of polysyllabic shortening of a root syllable (as in *sleep*) as more syllables were added to the word (as in *sleepy, sleepiness)* might arise) (Krivokapić, 2012).

Another example of systematic prosodically-governed surface phonetic variation is found in the distribution of a voice quality variation called Irregular Pitch Periods (IPPs) or glottalisation. This phenomenon occurs quite often in typical speech, but its distribution was not seen as predictable on the basis of then-current theories of surface structure. For example, Umeda's (1978) analysis of the occurrence of IPPs found no systematic relation to syntactic structure, aside from its tendency to occur utterance-finally. In contrast, Pierrehumbert and Talkin (1991) analysed the occurrence of this phenomenon through the lens of prosodic structure, and found that in their sample of read laboratory speech, IPPs occur preferentially at the onset of a higher-level prosodic constituent, the Intonational Phrase, as well as at the onset of syllables that bear prominence. Similarly, Dilley, Shattuck-Hufnagel, and Ostendorf (1996) examined Intonational-Phrase-initial vowels in speech read by FM radio newscasters, finding similar probabilistic distributions of IPPs, and moreover reported that the distribution reflects the hierarchical distinction between Full and Intermediate Intonational Phrases. More recently, Garellek (2014) has argued that phrase-initial and accented-syllable episodes of IPPs arise by different mechanisms, with only the episodes occurring at the onsets of accented syllables attributable to articulatory strengthening; Shattuck-Hufnagel (2017) has reported data from a different dialect of American English showing that, for some speakers, even non-prominent Intonational-Phrase-initial vowels can be marked with IPPs.

The implication of such findings is that prosodic structure at the phrase or utterance level is a critical influence on the surface phonetic shapes of words in a spoken utterance. This observation is consistent with the function of the Prosody Generator in the 1989 version of the Nijmegen model (Levelt 1989). The Prosody Generator is designed to deal with suprasegmental aspects of an entire intonational phrase, including phrase-level syllabification (across word boundaries) with insertion of appropriate phonetic segments into that syllabic structure, and the specification of an intonational contour. (The discussion of the Prosody Generator in the 1999 presentation of the model focuses on its role in constructing syllables within PWds, more than on its role in generating higher-level structure, because the 1999 presentation, although more recent, deals largely with the encoding of single PWds rather than with entire phrases or utterances.)

Although the Prosody Generator is designed to deal with the prosodic structure of an entire phrase or utterance, it does so incrementally, as the outputs of previous processing modules become available. One of its major tasks is to construct Prosodic Words (which can include more than one lexical word, as in e.g. the now-famous example *escort + us*) and syllabify them for the Phonetic Encoder; the Phonetic Encoder will use the syllabified string to access the motor programmes for each syllable (as in *e + scor + tus*), and it will do this one PWd at a time. As discussed in Chapter 10 of *Speaking*, an interesting question that arises as a result of the commitment to incremental processing concerns how segmental interaction errors between the sounds of two different PWds can occur, particularly in cases when the two PWds are separated by one or more lexical words, as in *wagging their tails → tagging their wails*, *caught me totally by surprise → taught me coatally by surprise*, or *How can you say anything with your thumb in your mouth → thay anything with your sumb in your mouth*[1], which is sometimes the case. This is challenging to account for in a model where PWds are phonologically and phonetically encoded one at a time.[2] The incremental approach is consistent with many aspects of the experimental chronometric literature related to the

timing of speech initiation, as is tellingly argued in the 1999 presentation. However, there is evidence that speakers represent at least some prosodic aspects of an entire phrase or utterance before beginning to speak, and in future iterations of the model it will be important to specify more explicitly the ways in which larger prosodic constituents play a role in the planning process.

In addition to the occurrence of sound-level interaction errors between PWds, evidence for the representation of at least some aspects of larger prosodic constituents before the articulation of the utterance is begun includes 1) reports of higher initial F0 and deeper inhalations before longer utterances (Rochet-Capellan & Fuchs, 2013), suggesting some knowledge of the overall length of the utterance; 2) results of tongue twister experiments that show effects of phrase-level prosodic constituents and prominences on segmental interaction errors (Bierne & Croot, 2018; Croot, Au, & Harper, 2010); and 3) findings by Wheeldon and Lahiri (1997, 2002) that, under certain conditions (i.e. delayed production), initiation times for utterances in Dutch reflect the number of Prosodic Words in the utterance, even when the number of syllables is held constant (though under other conditions (immediate production), initiation times reflect the complexity of the first PWd of the utterance). Wynne, Wheeldon, and Lahiri (2018) have extended the latter findings to provide evidence that compound words are treated as single PWds, in Dutch. As Levelt (2002a) notes, the view that the phonetic stage of word-form encoding for articulation takes place one PWd at a time does not preclude the possibility that larger constituents play a role in speech planning as well, and it is even possible that speakers do not employ the same fixed level of representation on all occasions, i.e. that the processing representation repertoire is flexible.

However, there is an alternative model of the role of prosody in speech production which does not adopt an incremental approach: the Prosody First model of Keating and Shattuck-Hufnagel (2002). In that model, the speaker begins with a minimal generic prosodic representation for the planned phrase or utterance, and enriches it appropriately as more detailed information about the morpho-syntactic shape of the utterance is generated and transferred into the prosodic planning frame. In this approach, as in Shattuck-Hufnagel's scan-copier model (Shattuck-Hufnagel, 1992), the candidate words for the phrase or utterance are activated together, but they are serially ordered (i.e. selected for association with the structural slots of the planned utterance) one at a time, in early-to-late order. In such an approach, the availability of segmental content for at least some of the upcoming words of a planned utterance provides an account of how interaction errors can involve sounds from two different PWds, while the existence of a planning frame with segment-sized slots that is independent of its (eventual) contents provides an account of the fact that, in an exchange like the K*nicks and the Celtics* → the *Sicks and the Neltics*, a segment displaced by an earlier error (here, the /n/) appears in the precise serial location where the displacing segment (here, the /s/), should have occurred.

We note that the Nijmegen model already bends the incremental principle a bit, in order to account for the prevalence of inter-PWd sound-level interaction errors. That is, it postulates that in some cases, the sound segments for a second PWd are available during the encoding of the previous PWd. It would be of considerable interest to determine how often the two interacting PWds are adjacent, and how often they are separated by an intervening PWd. If the latter is rare, then perhaps the Nijmegen approach of allowing at least occasional departures from incrementality might be enough to account for between-PWd errors.

Additional functions of the Prosody Generator which have not yet been fully developed include the selection of Pitch Accent and Boundary Tone types (suggested as under the control of a component called "Intonational Meaning" in the 1989 version), the hierarchical nature of constituent-final lengthening related to the level of the prosodic constituent (Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992), and the tendency to divide utterances into equal-length prosodic constituents rather than adhering strictly to the syntactic structure (Gee & Grosjean, 1983). Levelt noted in 1989, that "The ways of the Prosody Generator are still quite enigmatic" (p. 398)—and they continue to be. For a model of the utterance-form-encoding process which deals with these issues, see Turk and Shattuck-Hufnagel (forthcoming).

This section has presented a sampling of the wide range of evidence that phrase-level prosody is one of the important factors governing systematic context-governed surface phonetic variation, particularly (but not exclusively) in timing and duration. It has also argued that these facts support the view that prosodic planning involves higher-level constituents from its inception. But surface form variation is not the only aspect of an utterance that appears to be under the influence of speech prosody; an emerging body of evidence demonstrates the close relationship between the prosody of a spoken utterance and the co-speech gestures of the hands and other non-oral "articulators", and these observations also argue for planning at the level of the

intended message. The next section describes some of this evidence, raising the question of how a speech production planning model might encompass co-speech movements as well as movements of the speech articulatory tract.

4. The close alignment between speech prosody and co-speech gestures, and its implications for production planning models

While the phrase-level prosodic structure of spoken utterances was coming to the fore as a major factor influencing the timing and other aspects of the surface phonetic form of a spoken utterance, considerable progress was being made in a different area of speech production: understanding the contribution of co-speech gesture to the communicative function of an utterance. In the work of Kendon (1972, 1980, 2004) and McNeill (1996, 2005) and their colleagues, substantial theoretical formulations emerged about the way in which certain movements of the hands, face and other body parts form part of an utterance. These theoretical proposals were supported by exquisitely detailed analyses of audio-video recordings of both experimentally-induced utterances and spontaneous conversations, building on early video analyses by e.g. Condon and Ogston (1966, 1967) and others. For example, Kendon (1972, 1980) proposed that speech and gesture are two aspects of an utterance, neither of them primary, and are planned together to convey the totality of the speaker's intended message. He also proposed a complex hierarchy of gestural organisation, with a number of optional movement phases that can accompany each gestural "stroke" (such as preparation, pre-stroke hold and post-stroke hold), and with successive gestures grouped together into larger constituents signalled by kinematic characteristics, such as consistent use of the right or left hand, consistent posture or consistent location in space, and by kinematic events such as constituent-final recovery (i.e. relaxation to rest position of the hands). He also proposed that these two streams of information are aligned in both their timing and their contribution to the meaning of an utterance, as when the gesture illustrates the manner in which an action that is described more abstractly in the utterance was carried out. McNeill (2018) proposed a related idea, which he called a "growth point", i.e. the seed of an idea to be communicated in a multi-modal message, and the concept of a "catchment", the repeated use of a gestural feature across multiple (not necessarily immediately adjacent) co-speech gestures, which reveals the underlying structure of the idea. McNeill also noted the tight alignment between the speech of an utterance and its gestures. Melinger and Levelt (2004) directly tested the hypothesis that co-speech gestures contribute to the communicative power of the utterance they accompany; their experimental results demonstrated that speakers describing a visual display included fewer spatial terms in their spoken description, when they also included gestural "tracing" movement in the air that conveyed that information. Such results support the view that speakers plan a communicative act to combine the information in their speech and in their gestures.

Kendon's work is of particular interest to us here because of the link he proposed between the hierarchy of prosodic constituents and the organisation of the accompanying gestures. His observations suggested that speakers organise successive gestures into units which are revealed by their shared characteristics, with a change in one or more of these characteristics when a new unit begins—and that these often-multi-gestural constituents align in time with the prosodic constituents. In his 1980 text, Kendon lays out a hierarchy of spoken prosodic constituents of his own devising, described in more detail in Kendon (1972). That hierarchy takes the constituent called the Intonational Phrase in Autosegmental Metrical Theory (here called a Tone Unit) as its starting point, and groups these phrases into successively higher level structures, so that Tone Units group into Locutions (roughly, sentences), which group into Locution Groups, which group into Locution Clusters (roughly, paragraphs), and finally into the Discourse. He then describes, for a 2-minute sample taken from a film made in a London pub (from a single speaker who was addressing a group of people), the relationship between these prosodic structures in the spoken part of the Discourse and the organisational structures in the gestural part. He notes that "each level of organization was matched by a distinctive pattern of bodily movement." For example, the speaker adopted a particular body posture when he began the discourse, and abandoned it at the end. For each of the next-lower constituents in the Discourse, i.e. the Locution Clusters, the speaker used his arms differently. Further, within each Locution Cluster, each Locution was characterised by a particular type of head movement. Finally, for individual Tone Units, distinctive movement patterns were again observed. (pp. 210–211).

As can be seen in this discussion, Kendon observed a striking degree of co-organisation between the prosodic structures in the speech and the kinematic structures in the gestures in this sample. In this paper and in subsequent publications, including his 2004 book *Gesture*, many additional examples illustrating the same phenomena are presented. Shattuck-Hufnagel and Ren (2018) have also presented data, from an exhaustive analysis

of the co-speech gestures in a half-hour sample of academic-lecture-style speech in American English, which provides additional support for the hypothesis that these gestures are organised into groups. Their results provide preliminary evidence that these gestural groupings are aligned with higher-level prosodic constituents, signalled by longer inter-constituent pauses. Further evidence for the close relationship between spoken prosody and gesture comes from a number of studies showing that the stroke of a gesture has a high probability of occurring in conjunction with a spoken syllable that bears a prosodic prominence (pitch accent); see for example Loehr, 2004; Renwick, Shattuck-Hufnagel, & Yasinnik, 2004; Shattuck-Hufnagel, Yasinnik, Veilleux, & Renwick, 2007; Loehr, 2012 and Jannedy & Mendoza-Denton, 2005, inter alia.

Although it is far from established that all speakers co-organise their gestures and their speech in these ways at all times, it is quite clear that at least some speakers do this at least some of the time. And so it is worth asking what are the implications for a model of speech production planning. Different gesture researchers have proposed alternative points in the production process at which gestures might be planned, including pre-linguistically, i.e. without access to the linguistic formulation process (Krauss, Chen, & Chawla, 1996; Krauss, Chen, & Gottesman, 2000, and in a different vein, de Ruiter, 2000), or under the simultaneous influence of the linguistic plan and spatio-motor properties of the objects or actions being spoken about (Kita et al., 2007; Kita & Özyürek, 2003; Özyürek, Kita, Allen, Furman, & Brown, 2005). In any case, it appears that a fully comprehensive model of the speech planning process must include a mechanism that can provide an account of speech-gesture alignment in time and the relationship of both speech and gesture to structure and meaning. This is because in at least some cases, some aspects of the message are conveyed by the gestures, rather than by the speech, and this allocation must be determined before further detailed specifications of the speech and the gestures can be generated. For example, Kita et al. (2007) tested the likelihood that speakers of English who were asked to describe the movements of objects that involved both a path (change of location in space) and a manner (e.g. rolling) sometimes used a gesture that demonstrated both aspects of the movement and at other times gestures that demonstrated only one aspect. While the primary interest for that study was the question of whether the syntactic choice of the speaker influenced the gestural choice, the results also showed that speakers have options for which aspect of an utterance they will convey with the spoken words of an utterance and which with a supportive gesture, and thus that these decisions must be planned. The question of where in the planning process such decisions are made will need to be addressed by a fully comprehensive model.

5. Conclusion

When the volume *Speaking* appeared in 1989, it integrated what was currently known about the processes involved in speech production processing into an elegant and comprehensive model, starting with the formulation of a message and ending with the actions of the articulatory system that produced the acoustic wave form of a spoken utterance. Since that time, a substantial body of knowledge has accumulated concerning the systematicity of context-governed non-contrastive surface phonetic variability, the role of higher-level prosody in governing that variability, and the connection of co-speech gesture to spoken prosody. Each of these sets of findings has implications for the nature of the speech production planning process. The blueprint provided by the production model in *Speaking* and partially implemented in LRM99 has inspired a wide range of investigations into aspects of that process that were not envisioned at the beginning, and the goal of developing a comprehensive model of the speech planning and implementation process which was laid out by that work is very much alive. Perhaps someday a volume will appear called *Speaking Prosodically-Governed Acoustic Phonetics with Appropriate Gestures and Occasional Errors.* That will indeed continue the spirit of comprehensiveness so beautifully embodied in the 1989 volume that started it all.

Notes

1. Examples cited here are drawn from the MIT Speech Error Corpus, a collection of 11,000+ errors collected by the author and her colleagues, by listening to everyday speech over the past decades; individual errors have been largely labelled for error type, error unit, direction of error and error ambiguity (which is pervasive, Shattuck-Hufnagel, 1987).
2. Levelt et al. (1999) describe a mechanism by which such cross-PWd errors can occur, postulating for example that in the encoding of *red socks*, the phonemic segments for both syllables are available during the process of Phonetic Encoding, by which syllable-position-specific elements like syllable onsets are associated with their appropriate slots. the result that the onset /s/ can be mis-selected for the first onset position and the onset /r/ for the second onset position, and if their binding-by-checking mechanism (Roelofs, 1997) fails to detect the mis-associations, the error *sed rocks* will be produced. This account opens the door to the possibility that longer

stretches of the phrase or utterance are planned phonologically.

Disclosure statement

No potential conflict of interest was reported by the author.

References

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*, 177–189.

Beckman, M., & Edwards, J. (1994). Articulatory evidence for differentiating stress categories. In P. A. Keating (Ed.), *Phonological structure and phonetic form: Papers in laboratory Phonology III* (pp. 7–33). Cambridge, UK: Cambridge University Press.

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, *3*, 255–309.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*, 92–111.

Bierne, M., & Croot, K. (2018). The prosodic domain of phonological encoding: Evidence from speech errors. *Cognition*, *2018*, 177. (e-publication ahead of print). doi:10.1016/j.cognition.2018.03.004

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

Brugos, A., Breen, M., Veilleux, N., Barnes, J., & Shattuck-Hufnagel, S. (2018). Cue-based annotation and analysis of prosodic boundary events. *Proceedings of Speech Prosody IX*, Poznan, Poland. 245–249.

Bybee, J. (2009). *Phonology and language use*. Cambridge: Cambridge University Press.

Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, *31*, 149–180.

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, *14*(1), 177–208. doi:10.1080/026432997381664

Caramazza, A., & Miozzo, M. (1997). The relation between syntactic and phonological knowledge in lexical access: Evidence from the 'tip-of-the-tongue' phenomenon. *Cognition*, *64*, 309–343.

Cole, J., & Shattuck-Hufnagel, S. (2018). Quantifying phonetic variation: Landmark labelling of imitated utterances. In F. Cangemi, M. Clayards, O. Niebuhr, B. Schuppler, & M. Zellers (Eds.), *Rethinking reduction* (pp. 164–204). Berlin: Mouton de Gruyter.

Condon, W. S., & Ogston, R. D. (1966). Sound film analysis of normal and pathological behavior patterns. *Journal of Nervous and Mental Disease*, *143*, 338–347. doi:10.1097/00005053-196610000-00005

Condon, W. S., & Ogston, R. D. (1967). A segmentation of behavior. *Journal of Psychiatric Research*, *5*, 221–235. doi:10.1016/0022-3956(67)90004-0

Cooper, A. M. (1991). *An articulatory account of aspiration in English*. PhD dissertation. Yale University.

Croot, K., Au, C., & Harper, A. (2010). Prosodic structure and tongue twister errors. In C. Fougeron, B. Kuhnert, M. D'Imperio, & N. Valee (Eds.), *Laboratory Phonology 10* (pp. 433–461). Berlin & New York: Mouiton de Gruyter.

Crystal, T. H., & House, A. S. (1988). Segmental durations in connected-speech signals: Current results. *Journal of the Acoustical Society of America*, *83*, 1553–1573.

Dell, G. S., Burger, L. K., & Svec, W. R. (1997). Language production and serial order: A functional analysis and a model. *Psychological Review*, *104*(1), 123–147.

de Ruiter, J. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and gesture* (pp. 248–311). Cambridge: Cambridge University Press.

Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, *24*(4), 425–444.

Ellis, L., & Hardcastle, W. J. (2002). Categorical and gradient properties of assimilation in alveolar to velar sequences: Evidence from EPG and EMA data. *Journal of Phonetics*, *30*, 373–396.

Ferreira, F. (1993). The creation of prosody during sentence processing. *Psychological Review*, *100*, 233–253.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, *101*, 3728–3740.

Garellek, M. (2014). Voice quality strengthening and glottalization. *Journal of Phonetics*, *45*, 106–113.

Gee, J. P., & Grosjean, F. (1983). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, *15*, 411–458.

Gow, D. (2001). Assimilation and anticipation in continuous spoken word recognition. *Journal of Memory and Language*, *45*, 133–159.

Gow, D. W., & Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology, Human Perception and Performance*, *21*, 344–359.

Halle, M. (1992). Phonological features. In W. Bright (Ed.), *International encyclopedia of linguistics, vol. 3* (pp. 207–212). Oxford: Oxford University Press.

Hawkins, S. (2011). Does phonetic detail guide situation-specific speech recognition? *Proceedings of the International Congress of Phonetic Sciences XVII*, Saarbrueken, 9–18.

Hayes, B. (1984). The phonology of rhythm in English. *Linguistic Inquiry*, *15*, 33–74.

Hayes, B. (1989). The prosodic hierarchy in meter. In P. Kiparsky & G. Youmans (Eds.), *Phonetics and Phonology I: Rhythm and Meter* (pp. 201–260). New York: Academic Press.

Heyward, J., Turk, A., & Geng, C. (2014). Does /t/ produced as [ʔ] involve tongue tip raising? Articulatory evidence for the nature of phonological representations. Poster presented at the *14th Conference on Laboratory Phonology*, Tokyo.

Hockett, C. (1955). *A manual of phonology*. Indiana University Publications in Anthropology and Linguistics 11.

Jannedy, S., & Mendoza-Denton, N. (2005). Structuring information through gesture and Intonation. In S. Ishihara, M. Schmitz, & A. Schwarz (Eds.), *Interdisciplinary studies on information structure 03* (pp. 199–244). Potsdam: Universitätsverlag Potsdam.

Johnson, K. (2004). Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (Eds.), *Spontaneous speech: Data and analysis. Proceedings of the 1st Session of the 10th International Symposium* (pp. 29–54).

Tokyo, Japan: The National International Institute for Japanese Language.
Kazanina, N., Bowers, J. S., & Idsardi, W. (2017). Phonemes: Lexical access and beyond. *Psychonomic Bulletin and Review*, *24*, 1–27.
Keating, P., & Shattuck-Hufnagel, S. (2002). A prosodic view of word form encoding for speech production. *UCLA Working Papers in Phonetics*, *101*, 112–156.
Kendon, A. (1972). Some relationships between body motion and speech. In A. Seigman & B. Pope (Eds.), *Studies in dyadic communication. elmsford* (pp. 177–216). New York: Pergamon Press.
Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key (Ed.), *Nonverbal communication and language* (pp. 207–227). The Hague: Mouton.
Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.
Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal:: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, *48*, 16–32.
Kita, S., Özyürek, A., Allen, S., Brown, A., Furman, R., & Ishizuka, T. (2007). Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. *Language and Cognitive Processes*, *22*(8), 1212–1236. doi:10.1080/01690960701461426
Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, *59*, 1208–1221.
Kohler, K. (1999). Articulatory prosodies in German reduced speech. In: *Proceedings of the XIVth International Congress of Phonetic Sciences*, San Francisco, Volume 1, 89–92.
Kornfeld, J. R. (1971). What initial clusters tell us about a child's speech code. Quarterly Progress Report 101, Research Laboratory of Electronics, Massachusetts institute of Technology, 218–221.
Krauss, R. M., Chen, Y., & Chawla, P. (1996). Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In M. Zanna (Ed.), *Advances in experimental social psychology* (pp. 389–450). San Diego, CA: Academic Press.
Krauss, R. M., Chen, Y., & Gottesman, R. F. (2000). Lexical gestures and lexical access: A process model. In D. McNeill (Ed.), *Language and gesture* (pp. 261–283). New York: Cambridge University Press.
Krivokapić, J. (2012). Prosodic planning in speech production. In S. Fuchs, M. Weihrich, D. Pape, & P. Perrier (Eds.), *Speech planning and dynamics* (pp. 157–190). Bern: Peter Lang.
Lahiri, A., & Reetz, H. (2002). Underspecified recognition. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology VII* (pp. 637–677). Berlin: Mouton de Gruyter.
Lahiri, A., & Reetz, H. (2010). Distinctive features: Phonological underspecification in representation and processing. *Journal of Phonetics*, *38*, 44–59.
Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America*, *51*(6B), 2008–2024.
Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
Levelt, W. J. M. (2002a). Picture naming and word frequency: Comments on Alario, Costa and Caramazza. *Language and Cognitive Processes*, *17*(3), 299–319.
Levelt, W. J. M. (2002b). Phonological encoding in speech production: Comments on Jurafsky et al., Schiller et al., and van Heuven & Haan. In C. Gussenhover & N. Warner (Eds.), *Laboratory Phonology VII* (pp. 87–99). Berlin: Mouton de Gruyter.
Levelt, W. J. M., Roelofs, A., & Meyer, A. (1999). A theory of lexical access in speech production. *Behavioral and Brain Science*, *22*, 1–75.
Liberman, M., & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, *8*(2), 249–336.
Loehr, D. (2004). *Gesture and intonation*. PhD Thesis, Georgetown University.
Loehr, D. P. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology*, *3*(1), 71–89.
Macken, M. A., & Barton, D. (1980). A longitudinal study of the acquisition of the voicing contrast in American English word-initial stops, as measured by voice onset time. *Journal of Child Language*, *7*, 433–458.
Manuel, S. Y. (1995). Speakers nasalize /dh/ after /n/, but listeners still hear /dh/. *Journal of Phonetics*, *23*(4), 453–476.
McAllister Byun, T., Richtsmeier, P., & Maas, E. (2013). Covert contrast in child phonology is not necessarily extragrammatical. *LSA Annual Meeting Extended Abstracts*, *4*(28), 1–5. doi:10.3765/exabs.v0i0.786
McNeill, D. (1996). *Hand and mind: What gestures reveal about thought*. Chicago, Illinois: University Of Chicago Press.
McNeill, D. (2005). *Gesture and thought*. Chicago, Illinois: University Of Chicago Press.
McNeill, D. (2018). Growth Points. Retrieved from http://mcneilllab.uchicago.edu/writing/growth_points.html
Melinger, A., & Levelt, W. J. M. (2004). Gesture and the communicative intention of the speaker. *Gesture*, *4*(2), 119–141.
Miozzo, M., & Caramazza, A. (1997). On knowing the auxiliary of a verb that cannot be named: Evidence for the independence of grammatical and phonological aspects of lexical knowledge. *Journal of Cognitive Neuropsychology*, *9*, 160–166.
Nespor, M., & Vogel, I. (1986). *Prosodic phonology*. Berlin: De Gruyter.
Niebuhr, O., & Kohler, K. (2011). Perception of phonetic detail in the identification of highly reduced words. *Journal of Phonetics*, *39*, 319–329.
Nielson, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*, 132–142.
Özyürek, A., Kita, S., Allen, S., Furman, R., & Brown, A. (2005). How does linguistic framing influence co-speech gestures? Insights from crosslinguistic differences and similarities. *Gesture*, *5*(1–2), 219–240.
Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, *119*, 2382–2393.
Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation*. PhD thesis, Massachusetts Institute of Technology.
Pierrehumbert, J. B. (1990). Phonological and phonetic representation. *Journal of Phonetics*, *18*, 375–394.
Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency effects and the emergence of lexical structure* (pp. 137–157). Amsterdam: John Benjamins.

Pierrehumbert, J. B. (2016). Beyond abstract vs. episodic. *Annual Review of Linguistics*, *2*, 33–52.
Pierrehumbert, J. B., & Beckman, M. (1988). *Japanese tone structure*. Cambridge MA: MIT Press.
Pierrehumbert, J., & Talkin, D. (1991). Lenition of /h/ and glottal stop. *Papers in Laboratory Phonology II*, Cambridge University Press, Cambridge UK. 90–117.
Renwick, M., Shattuck-Hufnagel, S., & Yasinnik, Y. (2004). The timing of speech-accompanying gestures with respect to prosody (Abstract). *Journal of the Acoustical Society of America*, *115*, 2397.
Richtsmeier, P. T. (2010). Child phoneme errors are not substitutions. *Toronto Working Papers in Linguistics 33*. Retrieved from http://twpl.library.utoronto.ca/index.php/twpl/article/view/6889
Rochet-Capellan, A., & Fuchs, S. (2013). The interplay of linguistic structure and breathing in German spontaneous speech. *Proceedings of Interspeech*, *2013*, 1128–1132.
Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition*, *64*, 249–284.
Roelofs, A., Meyer, A., & Levelt, W. J. M. (1998). A case for the lemma/lexeme distinction in models of speaking: Comment on Caramazza and Miozzo (1997). *Cognition* , *69*, 219–230.
Schiller, N., & Caramazza, A. (2002). The selection of grammatical features in word production: The case of plural nouns in German. *Brain & Language*, *81*(1–3), 342–357.
Selkirk, E. O. (1984). *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: MIT Press.
Shattuck-Hufnagel, S. (1987). The role of word onset consonants in speech production planning: New evidence form speech error patterns. In E. Keller & M. Gopnik (Eds.), *Motor and sensory processing in language* (pp. 17–51). Hillsdale, NJ: Erlbaum.
Shattuck-Hufnagel, S. (1992). The role of word structure in segmental serial ordering. *Cognition*, *42*, 213–259.
Shattuck-Hufnagel, S. (2017). Individual differences in the signalling of prosodic structure by changes in voice quality. *The Journal of the Acoustical Society of America*, *142*, 2521. doi:10.1121/1.5014213.
Shattuck-Hufnagel, S., & Ren, A. (2018). The prosodic characteristics of non-referential co-speech gestures in a sample of academic lecture-style speech. *Frontiers of Psychology*, *09*, 1514.
Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, *25*(2), 193–247.
Shattuck-Hufnagel, S., Yasinnik, Y., Veilleux, N., & Renwick, M. (2007). A method for studying the time alignment of gestures and prosody in American English: 'Hits' and pitch accents in academic-lecture-style speech. In A. Esposito, M. Bratanic, E. Keller, & M. Marinaro (Eds.), *Fundamentals of Verbal and Nonverbal Communication and the Biometric issue* (pp. 34–44). Brussels: NATO.
Steedman, M. (2000). Information structure and the syntax-phonology interface. *Linguistic Inquiry*, *31*(4), 649–689.
Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, *111*, 1872–1891.
Tiede, M. K., Boyce, S. E., Espy-Wilson, C., & Gracco, V. (2010). Variability of North American English /r/ production in response to palatal perturbation. In B. Maassen & P. van Lieshout (Eds.), *Speech motor control: New developments in Basic and Applied research* (pp. 53–67). Oxford: Oxford University Press.
Turk, A. E., & Shattuck-Hufnagel, S. (forthcoming). *Speech timing*. Oxford: Oxford University Press.
Turk, A., & White, L. (1999). Structural influences on accentual lengthening in English. *Journal of Phonetics*, *27*, 171–206.
Umeda, N. (1978). Occurrence of glottal stops in fluent speech. *Journal of the Acoustical Society of America*, *64*(1), 88–94.
Wagner, M. (2010). Prosody and recursion in coordinate structures and beyond. *Natural Language and Linguistic Theory*, *28*, 183–237.
Wagner, M., & Watson, D. (2010). Experimental and theoretical advances in prosody: A review. *Introduction to Special Issue of Language and Cognitive Processes*, *25*(7), 905–945.
Wheeldon, L., & Lahiri, A. (1997). Prosodic units in speech production. *Journal of Memory and Language*, *37*, 356–381.
Wheeldon, L., & Lahiri, A. (2002). The minimal unit of phonological encoding: Prosodic or lexical word. *Cognition*, *85*(2), B31–B41.
Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, *91*(3), 1707–1717.
Wynne, H., Wheeldon, L., & Lahiri, A. (2018). Compounds, phrases and clitics in connected speech. *Journal of Memory and Language*, *98*, 45–58.
Zsiga, E. C. (1997). Features, gestures and Igbo vowels: An approach to the phonology-phonetics Interface. *Language*, *73*(2), 227–274.