Phonology **38** (2021) 165–171. © The Author(s), 2021. Published by Cambridge University Press
doi:10.1017/S0952675721000099

Alice Turk and Stefanie Shattuck-Hufnagel (2020). *Speech timing: implications for theories of phonology, phonetics, and speech motor control*. (Oxford Studies in Phonology and Phonetics 5.) Oxford: Oxford University Press. Pp. xv + 370.

Jason A. Shaw *
Yale University

1 Overview and structure of the volume

There is increasing awareness that the temporal dimension of speech, in particular the relative timing of speech movements, contains rich information about phonological structure. Relating abstract phonological structure to the temporal unfolding of realistically variable speech data remains a major interdisciplinary challenge. It is this challenge that is taken up in *Speech timing: implications for theories of phonology, phonetics, and speech motor control*, henceforth *Speech timing*.

The book has eleven chapters, including a short introduction and a conclusion. The main proposal – a sketch of a model mapping phonological representations to continuous movements of articulators – comes in the second half of the book, particularly in Chapters 7 and 10. The second half also includes chapters on optimisation (Ch. 8: 'Optimization') and general mechanisms for timing (Ch. 9: 'How do timing mechanisms work?'). These provide a unique synthesis of speech and non-speech literature, which is highly accessible for linguists, and serves to motivate aspects of the main proposal.

The first half of the book provides a description and critique of the theory of Articulatory Phonology, developed in the Task Dynamics framework (AP/TD) (e.g. Browman & Goldstein 1986, Saltzman & Munhall 1989). On the view of the authors:

> AP/TD currently provides the most comprehensive account of systematic spatiotemporal variability in speech … it represents the standard which any alternative theory must match or surpass, and provides a clear advantage over traditional phonological theories as a model of the speech production process (pp. 8–9).

The review of AP/TD in Chapter 2 of *Speech timing* highlights some key characteristics of the theory, particularly those related to prosodic modulation of timing and those implemented in the Task Dynamics Application, which simulates articulatory trajectories and the resulting acoustics from specifications of phonological representations. The AP/TD overview sets the stage for an exposition of empirical phenomena in Chapters 3–6 that the authors interpret as a challenge to AP/TD and as motivation for an alternative approach.

Thus the basic rhetorical strategy is to first problematise an existing theory, AP/TD, and then to present an alternative approach.

* E-mail: JASON.SHAW@YALE.EDU.

2 Major arguments in the volume

Speech timing describes the 'Phonology-Extrinsic-Timing-Based Three-Component approach', abbreviated as 'XT/3C'. Although, as described above, this approach is presented as an alternative to AP/TD, it has many features that may be more familiar to readers of *Phonology* than AP/TD. AP/TD represents phonological contrasts in terms of gestures, dynamically defined speech production tasks. Gestures have temporal extent and are coordinated in time. The dynamical approach of AP/TD defines at once discrete phonological tasks and the changes in articulator position over time that achieve them. From this standpoint, phonology and phonetics are not clearly distinct – rather they are different levels of description of the same system. In contrast, XT/3C is characterised as having three components: (i) phonological planning, (ii) phonetic planning and (iii) motor-sensory implementation.

The phonology-extrinsic timing aspect of the framework, the 'XT' of 'XT/3C', highlights a key contrast with AP/TD – phonological representations in XT/3C do not specify time beyond the linear order of segments. Moreover, in XT/3C, phonological representations do not commit to the particular phonetic dimensions that will serve as targets for the speech production process or how progress towards those goals will unfold in time. Such specifications come only at later stages in the XT/3C speech production process, when the phonological representation is input into the phonetic planning component. Representations in the phonological planning component of XT/3C consist of phonemes, hierarchical prosodic structure, metrical prominence and abstract characterisations of speech rate and style. Fully specified utterances in the phonological planning component project linearly ordered acoustic cues, which are given quantitative targets in the phonetic planning module. Also in the phonetic planning module, articulatory trajectories are computed to achieve the acoustic targets in the linear order specified by the phonological component. The cost of deviating from acoustic targets is balanced against the cost of movement (movement effort and movement time), both of which could be sensitive to any aspect of the phonological representation. Movement control is proposed to be guided by instantaneous awareness of the location of articulators and the distance that they would have to move to achieve acoustic targets, i.e. the movement endpoints; it is suggested that this idea can be appropriately formalised in tau theory, following Lee (1998). Finally, the motor-sensory implementation component issues the motor commands at appropriate times to achieve the planned acoustic cues to phonological structure.

Some of the key contrasts between AP/TD and XT/3C discussed at length in *Speech timing* are: (i) spatio-temporal (AP/TD) *vs*. symbolic (XT/3C) phonological representations; (ii) articulatory (AP/TD) *vs*. acoustic (XT/3C) phonetic targets; (iii) phonology-intrinsic (AP/TD) *vs*. phonology-extrinsic (XT/3C) timing; (iv) onset-triggered (AP/TD) *vs*. endpoint-triggered (XT/3C) movements. The XT/3C assumption of symbolic phonological representations, (i), appears to dictate much of how the rest of the XT/3C architecture, i.e. (ii)–(iv), differs from AP/TD. In AP/TD, the mapping from phonological representations to observable speech behaviour is compositional. Since gestures are specifications of how phonological contrasts shape speech production over time, any contextual deviation must be attributed to another overlapping gesture. This includes the π- and μ-gestures involved in prosodic modulation (Byrd & Saltzman 2003, Katsika *et al*. 2014). Positing symbolic phonological representations liberates XT/3C from having to account for contextual variation, including prosodic

effects, compositionally, as spatio-temporal modifications. Rather, in XT/3C, since phonological representations lack spatio-temporal content, there is no 'default' to modify. Any aspect of the phonological representation, including combinations of segmental contrast and phonological position can condition unique, context-specific phonetic targets. Since phonological representations lack spatio-temporal content, this needs to be provided in other components, which relate to (ii)–(iv).

The above aspects of XT/3C that differ from AP/TD are each motivated with some discussion of empirical data. Most of the data that is key to the discussion is already published. One exception is a study presenting some proof of concept of the applicability of Lee's theory to speech, cited as Lee & Turk (in preparation) and presented briefly in §9.2.1 (pp. 259–261). Although the relevant data is generally not new, it is uniquely synthesised as motivation for XT/3C.

The empirical motivation for loosening the strict compositionality of AP/TD comes in part from the observation that phrase-final lengthening in Finnish is attenuated when a vowel-length contrast is at stake (Nakai *et al*. 2009, Nakai *et al*. 2012). The Finnish data is interpreted as an argument against phonology-intrinsic timing, because the lengthening effect of the prosodic boundary is not uniform across short and long vowels, an apparent violation of the compositionality inherent in AP/TD. The XT/3C alternative is that surface durations result from an optimisation of cues to both phonological contrast and prosodic structure. Since the symbolic representations of XT/3C are phonetically unconstrained, they can reflect sensitivity to contrast maintenance that flouts the cues to prosodic boundaries just when contrast is at stake.

Key evidence for acoustic phonetic targets include well-known cases in which it appears that different combinations of articulatory constrictions vary in order to maintain relatively stable acoustic targets, such as the trade-off between tongue-dorsum retraction and lip rounding to maintain a relatively stable F2 for /u/ in American English. In addition to examples from speech, *Speech timing* also reviews empirical studies documenting other motor behaviours, such as typing and tapping, which tend to show greater temporal variability at movement onsets than at movement endpoints. One speech production study is cited as evidence for this claim (Perkell & Matties 1992). Another possible line of evidence for endpoint-triggered movements comes from spatially conditioned gestural timing (for discussion, see Shaw & Chen 2019), although there may be alternative explanations for these observations that are also consistent with onset-triggered movement.

3 Critical discussion of data and claims

Although *Speech timing* raises some important issues that will guide future research on the topic, to really adjudicate between approaches it will be necessary to consider a wider range of data and to work out XT/3C in greater detail, ideally to the extent that it can make quantitative predictions. I elaborate on these points in this section.

Speech timing does not review cases in which articulation maintains stability across contexts even at the expense of acoustic cues to phonemic contrast, or cases where articulatory differences corresponding to distinct phonemic contrasts are masked in the acoustics. Extreme examples include gestural hiding and covert contrasts. In gestural hiding, gestures are produced that have little or no acoustic consequences, because of how the gestures overlap in time, such as the tongue-tip

gesture for /t/ movement after lip closure for /m/ in the English sequence *perfec*/t/*memory* (Browman & Goldstein 1989). Covert contrasts are phonemic contrasts that are produced distinctly in articulation but still sound indistinct to listeners, and are thought by some researchers to be widespread in both typical and atypical language development. These observations appear to me to present a challenge for XT/3C or any theory in which articulatory trajectories are computed as optimal means to achieve acoustic targets. Quite specifically, the challenge situated within the XT/3C framework is to identify the combinations of costs such that optimisation will result in a large articulatory movement with no acoustic consequences.

There are also cases in which the same articulatory posture appears to be reused across segments. For example, the articulatory posture observed for fricative vowels in Suzhou Chinese, including both alveolar fricative vowels, which are restricted in their distribution to follow alveolar fricatives, and postalveolar fricative vowels, which are not subject to such restriction, have the same articulatory posture as corresponding fricative consonants in the language (Faytak 2018). Faytak argues that this and other observations of articulatory uniformity, i.e. reusing articulatory postures across different contrasting segments, may have its basis in learning – speakers may reuse already practiced motor routines if they accomplish 'merely good enough' rather than optimal acoustic outcomes (Faytak 2018: 29). This reasoning appears to me to be incompatible with XT/3C, although, again, the precise predictions depend much on how the interacting costs function in the optimisation process.

In addition to such cases of spatial uniformity, there are also cases of articulatory timing uniformity in the literature. For example, Shaw & Davidson (2011) pursued the hypothesis that native English speakers would produce Russian consonant clusters, e.g. word-initial /zb dg/, etc., in ways that minimise the perceptual distance between what they hear and what they produce – this hypothesis is consistent with XT/3C, but it was not supported by the data. Rather, computational simulations deriving the acoustic data from the articulatory timing of consonant gestures suggested that speakers imposed uniformity in timing – producing fricative–stop clusters like /zb/ with the same pattern of relative timing as stop–stop clusters like /dg/, even though this results in different acoustic deviations from Russian: [zb] *vs*. [dᵊg]. Several studies have reported systematicity in the timing between articulatory movements such that different consonants enter into the same temporal relations (despite different spatial targets). Key empirical observations include: (i) timing is language-specific, in that similar sequences of segments can show different patterns of articulatory timing across languages (e.g. Hermes *et al*. 2017), and (ii) within languages, different sequences of segments, such as rising *vs*. falling sonority syllable onsets, can sometimes show similar patterns of relative timing (e.g. Shaw *et al*. 2011). These patterns can be derived from AP/TD straightforwardly, but it remains to be seen if they might also emerge from surface optimisation of acoustic targets and articulatory costs, as in XT/3C.

Of course, the timing between consonant gestures is not always consistent across consonants of different identities. For example, initial /kn/ clusters in German have longer temporal lag than initial /kl/ clusters (Bombien *et al*. 2013) and Italian /sC/ clusters pattern together in their timing to the exclusion of rising sonority clusters (Hermes *et al*. 2013). In some cases, the differences in gestural timing correspond to differences in syllable structure, providing some cross-linguistic support for the observation that higher levels of linguistic structure, such as syllables, may be found in characteristic patterns of temporal

organisation between articulatory gestures (Browman & Goldstein 1988). This points to a broader issue – how exactly phonological structure conditions variation in articulatory coordination. Within the XT/3C framework, any such correspondence is indirect, in that it is mediated by the linearisation of acoustic cues to phonological structure, although the precise consequences of this architecture depend on the cost function at play in optimisation.

The last ten years have produced a wide empirical base of studies reporting high temporal resolution articulatory data that can constrain theorising about principles underlying speech timing, within and across languages. Establishing whether XT/3C can account for this range of patterns in a principled way requires further development of the model. In my view, two further developments are required.

One step is to converge on a precise characterisation of how cost enters into the optimisation of surface timing, including the integration of costs for time, energy/effort and spatial variation at the endpoint of movement, since, within the XT/3C framework, optimisation is required to generate quantitative predictions. *Speech timing* provides an extensive discussion of the issues involved in treating speech production as a complex optimisation problem, including relevant non-speech motor literature, and why it is difficult to identify with precision what the relevant costs are for speech and how these costs might interact in the optimisation process. This discussion, mostly localised in Chapter 8, includes insightful critiques of optimisation approaches that posit durational targets for linguistic units with penalties for deviating from targets. *Speech timing* argues that the durations of linguistic units should not be specified in underlying representations, because they can emerge instead from optimisation. In XT/3C, phonemes, syllables, etc., gain temporal extent only through the joint optimisation of surface durations between acoustic landmarks and the articulatory movements that give rise to them. One major challenge for XT/3C is therefore to derive the range of temporal patterns, including cases of apparent systematicity in articulatory timing, from the proposed optimisation procedure.

A second step for XT/3C, in my view, is to develop the theory of how phonological representations project linearly ordered acoustic/auditory cues. This may be as simple as adopting existing phonological proposals, or aspects of them. The mapping from abstract phonemic representations to context-specific acoustic/auditory cues in XT/3C resembles the mapping from underlying to surface representations in generative phonology, particularly versions of generative phonology that offer an optimisation-based unification of phonetics and phonology (e.g. Boersma 1998, Flemming 2001). *Speech timing* makes it clear that the phonological component of XT/3C is not intended to be isomorphic with the phonological grammar, even though they make use of similar abstract linguistic units (p. 268). However, it seems to me that there is potential to do a lot of what phonological grammars are designed to do within the XT/3C framework, including mappings that can be described in terms of phonological rules. This is because, unless further constrained, the framework leaves open the possibility that the same abstract phoneme could be mapped to radically different acoustic realisations in different phonological contexts. One constraint on phoneme to acoustic landmark projection proposed in XT/3C is that this mapping balances language redundancy and acoustic/auditory information so as to maintain consistent recognition probability over time. It would be interesting to consider how language redundancy relates to grammatical rules, particularly in cases in which they make potentially conflicting predictions. Working out the

mapping from phonological representations to acoustic/auditory cues would reveal the degree to which XT/3C, as a model of speech production, can be integrated with existing models of phonological grammar.

4 Conclusion

There are a number of properties which make *Speech timing* essential reading for students and researchers interested in relating abstract phonological structure to time-dependent articulatory and acoustic properties. From start to finish, the book offers a balanced review of a significant amount of relevant research, some of which is not succinctly reviewed elsewhere. Perhaps even more valuable is that *Speech timing* keeps consistent tabs on the empirical facts that motivate each component of the model, maintaining contact between data and theory. The book can also be read with an eye to even broader conceptual issues. These include how speech motor control relates to other aspects of motor control, such as limb movement – 'is speech special?' – and whether speech is better conceptualised as a complex computation, as in the optimisation problem characterised by XT/3C, or as the interplay of forces tending toward equilibrium, as in the dynamical systems of AP/TD.

The general style of the book, reasoning from theoretical predictions (of AP/TD) to empirical data and back to theoretical alternatives, encourages situating speech data as evidence for competing theoretical hypotheses. Although the theoretical dichotomies in (i)–(iv) are possibly too sharp – it seems unlikely that speech production targets are entirely articulatory or entirely acoustic/auditory, or that articulatory timing refers only to movement onsets or only to movement endpoints – they serve to highlight important and unresolved issues, which can be clarified by further experimental work. Data seemingly supporting one approach or the other could ultimately be specific to particular language varieties, phonological contexts, experimental tasks, individual speakers or stages of language development. For example, in AP/TD it is generally assumed that the production goals of tone gestures are acoustic in nature (e.g. Gao 2008) and some gesture-based models consider both movement onsets and offsets (endpoints), as well as other landmarks, to be available for coordination (Gafos 2002). Nevertheless, the theoretical issues raised in *Speech timing* can focus much future research on computational and experimental aspects of relating phonological structure to the speech signal, including approaches that hybridise aspects of AP/TD and aspects of XT/3C (see e.g. Parrell & Lammert 2019).

REFERENCES

Boersma, Paul (1998). *Functional phonology : formalizing the interactions between articulatory and perceptual drives*. PhD dissertation, University of Amsterdam.

Bombien, Lasse, Christine Mooshammer & Philip Hoole (2013). Articulatory coordination in word-initial clusters of German. *JPh* **41**. 546–561.

Browman, Catherine P. & Louis Goldstein (1986). Towards an articulatory phonology. *Phonology Yearbook* **3**. 219–252.

Browman, Catherine P. & Louis Goldstein (1988). Some notes on syllable structure in articulatory phonology. *Phonetica* **45**. 140–155.

Browman, Catherine P. & Louis Goldstein (1989). Articulatory gestures as phonological units. *Phonology* **6**. 201–251.

Byrd, Dani & Elliot Saltzman (2003). The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. *JPh* **31**. 149–180.
Faytak, Matthew D. (2018). *Articulatory uniformity through articulatory reuse: insights from an ultrasound study of Sūzhōu Chinese*. PhD dissertation, University of California, Berkeley.
Flemming, Edward (2001). Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology* **18**. 7–44.
Gafos, Adamantios I. (2002). A grammar of gestural coordination. *NLLT* **20**. 269–337.
Gao, Man (2008). *Mandarin tones: an Articulatory Phonology account*. PhD dissertation, Yale University.
Hermes, Anne, Doris Mücke & Bastian Auris (2017). The variability of syllable patterns in Tashlhiyt Berber and Polish. *JPh* **64**. 127–144.
Hermes, Anne, Doris Mücke & Martine Grice (2013). Gestural coordination of Italian word-initial clusters: the case of 'impure s'. *Phonology* **30**. 1–25.
Katsika, Argyro, Jelena Krivokapić, Christine Mooshammer, Mark Tiede & Louis Goldstein (2014). The coordination of boundary tones and its interaction with prominence. *JPh* **44**. 62–82.
Lee, David N. (1998). Guiding movement by coupling taus. *Ecological Psychology* **10**. 221–250.
Lee, David N. & Alice Turk (in preparation). *Vocalizing by tauG-guiding articulators*.
Nakai, Satsuki, Sari Kunnari, Alice Turk, Kari Suomi & Riikka Ylitalo (2009). Utterance-final lengthening and quantity in Northern Finnish. *JPh* **37**. 29–45.
Nakai, Satsuki, Alice Turk, Kari Suomi, Sonia Granlund, Riikka Ylitalo & Sari Kunnari (2012). Quantity constraints on the temporal implementation of phrasal prosody in Northern Finnish. *JPh* **40**. 796–807.
Parrell, Benjamin & Adam C. Lammert (2019). Bridging dynamical systems and optimal trajectory approaches to speech motor control with dynamic movement primitives. *Frontiers in Psychology* **10:2251**. https://doi.org/10.3389/fpsyg.2019.02251.
Perkell, Joseph S. & Melanie L. Matties (1992). Temporal measures of anticipatory labial coarticulation for the vowel /u/: within- and cross-subject variability. *JASA* **91**. 2911–2925.
Saltzman, Elliot & Kevin G. Munhall (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology* **1**. 333–382.
Shaw, Jason A. & Wei-rong Chen (2019). Spatially conditioned speech timing: evidence and implications. *Frontiers in Psychology* **10:2726**. https://doi.org/10.3389/fpsyg.2019.02726.
Shaw, Jason A. & Lisa Davidson (2011). Perceptual similarity in input–output mappings: a computational/experimental study of non-native speech production. *Lingua* **121**. 1344–1358.
Shaw, Jason A., Adamantios I. Gafos, Philip Hoole & Chakir Zeroual (2011). Dynamic invariance in the phonetic expression of syllable structure: a case study of Moroccan Arabic consonant clusters. *Phonology* **28**. 455–490.