Timing in talking: What is it used for, and how is it controlled?

Timing in talking: What is it used for, and how is it controlled?

Alice Turk[a] & Stefanie Shattuck-Hufnagel[b]

[a]Linguistics & English Language, U. of Edinburgh, turk@ling.ed.ac.uk; [b]Speech Communication Group, Research Lab of Electronics, Massachusetts Institute of Technology

Abstract In the first part of the paper, we summarize the linguistic factors which shape speech timing patterns, including the prosodic structures which govern them, and suggest that speech timing patterns are used to aid utterance recognition. In the spirit of Optimal Control Theory, we propose that recognition requirements are balanced against requirements such as rate of speech and style, as well as movement costs, to yield (near-)optimal planned surface timing patterns; additional factors may influence the implementation of that plan. In the second part of the paper, we discuss theories of timing control in models of speech production and motor control. We present three types of evidence that support models of speech production that involve extrinsic timing. These include 1) increasing variability with increases in interval duration, 2) evidence that speakers refer to and plan surface durations, and 3) independent timing of movement onsets and offsets.

Key index words or phrases: extrinsic speech timing, prosodic structure, speech production, smooth signal redundancy, optimal control theory, phonetic implementation.

1. Introduction

Timing is an integral part of every aspect of speech production: individual movements of the rib cage, oral articulators and laryngeal structures; their coordinated motor activity; and the speech sounds they produce. Understanding speech production therefore requires understanding timing: what it is used for, and how it is controlled. In this paper, we first review our current understanding of what speakers use timing for, and how this understanding was acquired by researchers, and then focus on two different views of how timing is controlled: with and without an extrinsic timekeeping mechanism. We then present evidence that seems to require an extrinsic timekeeping mechanism. Space prevents us from detailing the methods involved in measuring timing, but see [1] for measurement methods based on acoustic landmarks [2], and [3] for a method based on landmarks in movement traces.

2. What is speech timing used for?

The traditional way of determining what speakers use timing for is to conduct controlled experiments in which a factor of interest is systematically varied, keeping other factors constant. For example, in experiments testing whether vowel type has a systematic effect on duration, different vowels can be embedded in a constant carrier phrase, e.g. *Say dad again*, vs. *Say did again*. Such experiments have shown systematic differences between different speech sounds, e.g. [4], which are therefore hypothesized to have a characteristic

'intrinsic' duration [5]. Analogously, experiments that vary higher-level prosodic structure have shown systematic effects of prominence and constituent boundaries on duration. For example, a comparison of *dad* in *Say DAD again* vs. in *SAY dad again* shows that *DAD* is systematically longer when phrasally prominent (see [6] for a review). Moreover, depending on how the speaker chooses to produce a syntactic string, words before major constituent boundaries are often systematically longer than constituent-medial words, e.g. *cousin* is longer in *Mary GEORGE's cousin] [baked the cake*, where it is at the end of a phrase, as compared to *cousin* in *Mary's cousin GEORGE] [baked the cake*, where it is medial.

Experiments conducted from the 1950s through the 1980s established a long list of factors that affect speech timing. These include:

- Vowel and consonant type
- Contextual factors, e.g.
  - Prominence (word-stress, phrasal stress)
  - Syntax
  - Predictability
  - Adjacent segment type
- Global factors, e.g.
  - Speech rate
  - Speech style (e.g. clear vs. relaxed)

([7-10] inter alia; see [4, 11, and 6] for reviews). In addition, there are many other possible factors not yet integrated into current models, that may also influence speech timing, e.g. speaking to a periodic beat.

However, since the late 70's and 80's, e.g. [12,13], it has become clear that the view that each factor has a separate, direct effect on timing is problematic. Syntax is problematic because it has only an indirect influence on phonetic form, and predictability is problematic because many of its effects appear to be shared with other factors. In the first part of this paper we show how these factors relate to prosodic structure, which we see as a central part of the interface between language and speech. On our view, prosodic structure, segmental identity and segmental context are the factors that have a direct effect on the speaker's surface phonetic plan, including speech timing. When planning speech production, speakers balance these factors against non-grammatical factors such as speech rate and other stylistic requirements, clarity requirements, and movement costs (e.g. energy, time) to yield a specification of the desired temporal patterns for a spoken utterance.

2.1 The problem with syntax

Although it is clear that some syntactic manipulations have a measurable effect on duration (and other phonetic parameters), not all do. Consider for example:

- Mary George's cousin]? ate a piece of cake
- Her cousin]? ate a piece of cake
- She]? ate a piece of cake.

In these examples, where]? Is used to indicate a possible site of boundary related cues, the likelihood of these cues decreases for shorter subject noun phrases. That is, the longer subject noun phrase (*Mary George's cousin*) is more likely than the shorter ones (*Her cousin* and *She*) to show boundary-related phonetic cues such as pre-boundary lengthening and pause, even though they all share the same syntactic structure [14-17]. There are also some phonetic indicators of constituent boundaries that occur where syntax would not predict them, as in *Sesame street is brought to you by]…the Children's Television Workshop*, where a break occurs within a prepositional phrase [18]. Finally, levels of embedding found in syntax are often absent in speech [19]: For example, the utterance below has a right branching syntactic structure (Figure 1, top), whereas its spoken phrasing is flatter (Figure 1, bottom):

Figure 1 about here

2.2 Prosodic constituent structure

Along with other findings from segmental phonology (e.g. [12]) and intonational phonology [20], these findings suggest that a structure that is influenced by syntax, but not isomorphic to it, directly defines the groupings observed in speech. This structure, called prosodic constituent structure, is hierarchical, and includes constituents such as words and perhaps feet or syllables at lower levels, and phrases of various sizes at higher levels. Although there are debates about many aspects of the prosodic hierarchy, e.g. about the number of levels in the hierarchy, and about the name and definition of each constituent type, there is general agreement about its hierarchical nature, and about the fact that it is flatter and more symmetric than syntactic structure [21,12]. An example of prosodic structure is shown in Figure 2.

Figure 2 about here

Prosodic constituent structure is a likely linguistic universal, although different languages may elect different sets of levels from the universal hierarchy [22].
It has measurable effects on durational phenomena such as initial lengthening, final (or pre-boundary) lengthening, polysyllabic shortening (the shortening of syllables when more occur in a constituent), polysegmental shortening (the shortening of segments when more occur in a constituent) and pause (see [24] and [5] for reviews). Support for the universality of prosodic structure comes from the ubiquitous occurrence of final and initial lengthening patterns that reflect a structural hierarchy in languages of the world [23], [25].

Phrasally-related initial and final lengthening affect specific parts of initial and final words, respectively. Initial lengthening appears to be primarily localized on the initial C in phrase-initial CV and CCV sequences [26, 27]. In final position, most of the lengthening occurs on the rhyme of the final word. Smaller, but significant amounts of lengthening have also been observed on lexically stressed syllable rhymes when the lexically stressed syllable is pre-final, as in *Michigan* or *Trinidad* ([28] for Dutch, [29] for American English). Lengthening at other sites, e.g. the onset consonant of the phrase-final syllable rhyme, has also been observed, but these effects are sporadic in the sense that they appear to be

study- or material-dependent, and may possibly be speaker-dependent. For both initial and final lengthening, the magnitude of the durational effects varies with boundary strength: stronger boundaries (e.g. phrases) are generally associated with greater degrees of lengthening [30, 31] but interestingly not with a longer string in the domain of lengthening ([28]). See [32-34] for discussions of polysyllabic shortening.

Prosodic constituent structure also affects non-durational phonetic parameters, such as constituent-initial and final voice quality modifications (e.g. [35-39]), supralaryngeal articulatory modifications, (e.g. phrase-initial strengthening, syllable-final lenition [40-42], the use of word- or phrasal-prominence near the beginnings or ends of constituents (e.g. [43, 16]), as well as intonational phenomena, e.g. phrase-final lowering, phrase-initial reset (cf. [20, 44] among others).

2.3. Prosodic prominence structure

Prosodic structure also includes prosodic prominence structure, which describes different degrees of stress/accent found in words and phrases. For example, in the neutral prosodification of the phrase *Mary's cousin George*, *George* is the most prominent word in the phrase, and is said to bear phrasal stress (also called sentence stress, or accent). In the words *Mary* and *cousin,* the word-initial syllables *Ma(r)-* and *cou-* are more prominent than the second syllables in these words, and are said to bear word- or lexical stress*.* Figure 3 shows a grid-like representation of prominence structure [45-47], illustrated for this phrase.

Figure 3 about here

Like prosodic constituent structure, prosodic prominence structure is hierarchical, with word-stress near the bottom of the hierarchy, and phrasal stress at higher levels [48]. It too has measurable effects on duration, but the effects of prominence on duration appear to be different from those related to prosodic constituent boundaries [32, 49-51]. For example, monosyllabic words show different effects of phrasal prominence vs. final lengthening: prominence increases the nucleus duration most, followed by the syllable onset, then optionally the coda, whereas final lengthening increases the nucleus duration most, followed by the coda, then (optionally) the onset. Prosodic prominence structure affects duration but also affects other articulatory parameters such as articulatory distinctiveness and voice quality, and their acoustic consequences (e.g. formant structure, and spectral balance) [52,53].

2.4 Prosodic structure as the interface between language and speech

The indirect effects of factors such as syntax, utterance length, focus etc. on surface phonetics, via prosody, are illustrated in Figure 4 (based on a similar figure in [54]). We propose that prosodic structure serves as an interface between language on the one hand, and speech on the other (see also [55]), so that it is prosodic structure that exercises direct influence on the phonetic plan. During speech planning, prosodic effects on phonetic parameters such as duration are balanced against the effects of segmental identity and context, as

well as non-grammatical factors (e.g. rate and style of speech, clarity requirements, movement costs, etc., on those same parameters.

Figure 4 about here

Several aspects of Figure 4 are worthy of comment. First, we assume that the non-grammatical factors have a direct influence on the plans for surface phonetic form, rather than influencing the phonological plan. Although factors like rate and style of speech have been described as directly affecting aspects of prosody (e.g. fewer "breaks" at faster rates of speech, cf. [56], our view is that a speaker plans the same prosodic structure (i.e. same relative prominence and relative boundary strength structure) for a given utterance at different rates of speech, but that the planned phonetic manifestation of this structure would be different at different rates. This is because the rate-of-speech requirement would be balanced against the prosodic structure requirement in determining optimum surface phonetic characteristics that meet the competing demands. Second, the factors mentioned in this figure are intended to be a preliminary indicator of factors that might be at work, and may not be exhaustive. Related to this, there are other factors that are known to influence phonetics that remain to be investigated, for example, the adjustments that might be made in response to an interlocutor (possibly including non-speech input), a noisy environment, or intense emotion. These adjustments might relate to phonological planning, e.g. choices of prosodic structure, or might be non-grammatical, e.g. reflected in specifications of rate or clarity, and would therefore be balanced against prosodic structure requirements in influencing the phonetic plan. And there are other candidate factors such as cognitive processing costs and constraints, whose effects are not yet well-understood. This figure is therefore intended as a tool for identifying and thinking about factors that influence phonetic planning and as a proposal for how they might interact.

2.5. The problem with predictability

If we accept that prosodic structure has a measurable effect on duration, then another factor in the list becomes problematic: predictability. What we refer to as 'predictability' is the likelihood of a word given its context (linguistic and pragmatic/real world) and frequency of use; i.e. the likelihood that a word can be guessed from its context. It has long been observed that more-predictable words are produced with shorter durations than less-predictable words [57-60] For example, Lieberman [57] observed that more predictable words are shorter and less acoustically salient; he found that the word *nine* in *A stitch in time saves nine* (highly predictable context) was shorter than the word *nine* in *The number that you will hear is nine,*.

The problem with predictability as a factor affecting duration is that it is unclear whether prosodic structure and predictability are both motivated as separate factors affecting duration. This is because prosodic structure and predictability are not independent; When predictability is low, syllables are more likely to be prosodically prominent, and words are more likely to be demarcated using prosodic boundary correlates such as initial- and final-lengthening and pause. For example, the word *operas* in the phrase *health operas* is more likely to bear phrasal stress than the word *issues* in the phrase *health issues*, possibly because

issues in this context is more predictable [61,62]. In addition, the word *nine* may be longer in the phrase *The number that you will hear is nine* than in the phrase *A stitch in time saves nine*, because the *nine* in the former sentence is less predictable, and therefore the word boundary will be more saliently marked by lengthening on the word-initial /n/.

2.6 Prosodic structure as the interface between predictability and acoustic salience

[61-63] proposed that prosodic structure is the interface between predictability and acoustic salience, that is, prosodic structure is used to control acoustic salience in order to signal relative predictability. Aylett [61] proposed that by doing this, prosodic structure makes all words in an utterance equally easy to recognize. This proposal was termed the SMOOTH SIGNAL REDUNDANCY HYPOTHESIS (Figure 5, based on a similar figure in [63]).

Figure 5 about here

In the sentence *Who's the author?*, "*Who's* in its context (___*the author?*) is more predictable than *author* in its (full) context (*Who's the ___?); the* is even more predictable (context: *Who's__author?);* and furthermore, the word-initial syllable *, au(th*)- is relatively unpredictable compared to the second syllable *–(th)or.* The SMOOTH SIGNAL REDUNDANCY HYPOTHESIS states that an utterance's predictability profile (also called language redundancy) is inversely reflected in the prosodic structure of the elements (e.g. syllables and words) in the utterance. Prosodic structure is used to control the acoustic salience of surface phonetics (through prosodic prominence and boundary strength) so that the recognition likelihood of each element in the utterance is approximately equal, i.e. signal redundancy is smooth. As discussed in [62], the smooth signal redundancy profile is advantageous because it increases the likelihood of recognizing all of the elements in the utterance. The p(recognition) [1] of the entire sequence corresponds to the product of the p(recognition) of each element in the sequence, and will therefore be greater if p(recognition) of each element is equal, than if p(recognition) of different elements is different.

As discussed in [63], the idea that prosodic structure reflects predictability provides an explanation for the effect of utterance length on the likelihood of boundary occurrence and on boundary strength. This is because words are harder to guess (less predictable) in longer utterances. To understand why this is, consider a two-syllable utterance. All things being equal, there are two possible ways to parse such an utterance. As a sequence of two monosyllabic words, or as a single disyllabic word.

Parsing option 1: [syl]$_{word}$ [syl]$_{word}$

Parsing option 2: [syl syl]$_{word}$

[1] Note that by definition, probability values never exceed 1.

For a three-syllable utterance, the number of possible parsings increases to four:

Parsing option 1: [syl]$_{word}$ [syl]$_{word}$ [syl]$_{word}$

Parsing option 2: [syl]$_{word}$ [syl syl]$_{word}$

Parsing option 3: [syl syl]$_{word}$ [syl]$_{word}$

Parsing option 4: [syl syl syl]$_{word}$

And for a four-syllable utterance, the number of possible parsings is even larger, i.e. 8. However, when a phrase boundary is inserted anywhere in the utterance, the number of possible parsings is halved. As this example illustrates, when predictability is relatively low because an utterance is long, prosodic structure can be used to increase recognition likelihood by signaling constituent boundaries.

[61,62] proposed that predictability is a composite factor which directly influences prosodic structure, and thereby indirectly controls acoustic salience. That is, all of the factors shown at the top of Figure 6 contribute to the predictability of elements in an utterance. For example, a word's lexical frequency, together with its syntactic and semantic context, its real-world context (pragmatics), and utterance length, combine to predict how likely a particular word would be (i.e. how easily a word could be guessed) in that particular context. Aylett [61] refers to this predictability as 'language redundancy'. Our current hypothesis is that the predictability of each element in an utterance relates to its predictability on the basis of both preceding and following elements (i.e. the full context), as well as its frequency of use and likelihood on the basis of real world context, but note that it is an important research question to determine exactly what contributes to an element's predictability/language redundancy. As discussed in [63], the speaker can compute predictability (language redundancy) on the basis of his/her own language and real-world experience. Information about the listeners' knowledge or experience can be incorporated in the computation, but need not be.

As noted above, our hypothesis is that language redundancy is used to plan prosodic structure in order to make the recognition likelihood of each element equal. This goal of even recognition likelihood (or SMOOTH SIGNAL REDUNDANCY) is balanced against other goals, such as speaking clearly, quickly or in rhythm, as well as movement costs (e.g. time, energy) when speakers plan the surface phonetic properties of a spoken utterance..

Figure 6 about here

[61,62] provide supporting evidence for the view that prosodic prominence

structure reflects predictability: Both prosodic prominence structure and predictability (word frequency, syllable transitional probability and first vs. second mention of a word) largely accounted for the same variance in syllable duration in a large corpus study of spontaneous speech. Further supporting evidence includes findings that word durations are longer, and pauses and intonational boundaries more likely, in less predictable sequences [64, 15], discussed in [63].

2.7. Summary of Part I

What is speech timing used for? We propose that one of its main purposes is to make utterances easier to recognize, by signaling the identity of individual speech sounds (e.g. *did* vs. *dad*), and also signaling (and compensating for) the relative predictability of syllables and words in larger utterances. Because timing effects are implemented on very specific stretches of speech that relate to prosodic constituents (e.g. final lengthening occurs primarily on the rhyme of the final syllable; prominence-related lengthening occurs primarily on the stressed syllable nucleus and onset), it appears that predictability does not have a direct effect on surface phonetics (including timing), but rather its effects are mediated by prosodic structure. See other arguments in [63]. We propose that the goal of making speech easier to recognize by smoothing signal redundancy is balanced against other goals and costs when planning surface sound durations in speech.

3. Part II. How is speech timing controlled?

Here we address two different views of speech timing control: With and without an extrinsic timekeeper. Both approaches assume that surface timing patterns result from processes available for general non-speech motor control, but propose very different mechanisms to generate those surface phenomena. Extrinsic timing approaches involve the use of a system-extrinsic timekeeper, which tracks, represents, and specifies time in units that are not defined within the system (in the case of speech, the system would be the speech motor control system). In contrast, intrinsic timing systems do not involve system-extrinsic timekeepers. In such systems, all aspects of surface timing emerge from within-system characteristics. Any within-system timing specification is made in terms of within-system units, e.g. within-system oscillator periods or phasing. We note that we will call extrinsic any system that involves at least some timing computation by an extrinsic timing mechanism. However, we suspect that in many, if not all, extrinsic timing systems there may be aspects of surface timing that are emergent and don't need to be specified by the extrinsic timekeeper.

We first present the two approaches, and then three types of timing phenomena that suggest extrinsic timekeeper control.

3.1. Timing with an extrinsic timekeeper

Extrinsic timekeepers can be used in motor control for a variety of functions including tracking the passage of time, measuring time, representing time, as well as specifying time as a parameter of movement. Theories of speech and non-speech motor control that assume an extrinsic timekeeper include DIVA [65,66, based on 67], and many Optimal Control Theory models, e.g. [68]. These

models assume that desired movement durations can be specified as part of the plan for an utterance, and that the passage of time (and/or the time remaining) within a movement can be continuously tracked during implementation of that plan. Within these models, state (e.g. spatial) information is also tracked continuously, and timing information is integrated with state information to generate appropriate movement velocities at each time point. For example, DIVA [65] and VITE [67] assume that at each point in time, a temporal GO signal is multiplied by the difference vector (distance remaining to the target assuming a straight line path) to give instantaneous movement speed. In [67], GO is a function of time that is proportional to 1 divided by the time-to-target-attainment at the current instantaneous movement speed, cf. Lee's tau, [69]. Because the GO signal in [67] is an increasing function of time, and the distance to the target decreases as a function of time, multiplying GO by the distance remaining until the target at each point in time yields a bell-shaped velocity profile [70]. The same GO for different movement distances leads to equal movement durations for both, with higher peak velocities for the movement involving a greater distance. A larger GO for a given movement distance will yield a faster speed and therefore a shorter movement duration.

Optimal Control Theory models assume that we generate movements that are optimal in the sense that they meet task requirements at minimum cost. Many models of motor control in the Optimal Control Theory framework are like DIVA in that they assume that we continuously monitor the states of our effectors (e.g. their position and velocity) in relation to the task goals, continuously updating our motor commands on the basis of state information to accomplish goals in a near-optimal way (but see [71,72] for an exception). In these models, movements are generated via a control policy that determines the optimal movement from any current state given the task goals and costs of movement. The control policy (which can be a solution to a set of equations) is determined by minimizing a cost function defining the task goals, costs of movement, and their relative weightings in the current situation. Cost function minimization leads to the specification of values for all of the parameters in the model.

Although Optimal Control Theory models do not necessarily require the use of an extrinsic timekeeper, many models developed within this framework use time as a parameter of movement and/or as a cost, and therefore assume one, see e.g. [68,70,71]. In many Optimal Control Theory models that use extrinsic timekeepers, cost function minimization leads to the specification of movement parameters, including movement duration, where the optimal movement duration is the one that best satisfies the task requirements and movement costs. This movement duration results from several aspects of the cost function, including the specification of time as a task requirement, the cost of time, the cost of temporal inaccuracy, and the temporal consequences of other movement costs, e.g. spatial inaccuracy at the movement target, or endpoint [70,73-75]. The goals of a movement will determine whether all of these aspects are included in the cost function. For example, if a movement must be produced within a certain time (as in tasks with a periodic rhythm), time would be an explicit task requirement, and spatial inaccuracy would be included in the costs.

In contrast to tasks that require a specified duration as a task goal, purely spatial tasks might not involve an explicit goal for movement duration, but there would be temporal consequences of other task requirements, e.g. of spatial accuracy at target achievement, since faster movements can be produced when there are less stringent spatial accuracy requirements. In addition, empirical findings show that movements are usually produced in the minimum time consistent with other task requirements, suggesting that time itself is a cost [74,75].

Why should time be a cost? One possibility is that longer movements have more temporal variability [76]. This could be explained by the view that the mechanism that meters out time is variable, and hence more variability is expected to accumulate for longer duration intervals. However, this would not explain minimized durations observed in tasks where temporal accuracy is not an issue. [70,73], following [74], offer an explanation that relates movement speed to reward. That is, moving fast is desirable because we get to a rewarding state quickly; moving slowly is sub-optimal because it delays the next desirable state. Evidence in the literature supports the view that getting to a rewarding state more quickly is preferred; For example, [77] found that thirsty undergraduates preferred to receive a small amount of water now, rather than more later. See [70] for additional evidence.

The Optimal Control Theory framework is particularly attractive for speech timing, which appears to involve the influence of many different prioritized factors. It has been used successfully to model simple movements, and to model aspects of speech timing [71,72]. We note however, that although many if not most Optimal Control Theory models of motor control assume an extrinsic timekeeper, this theoretical framework is a theory of parameter value optimization, and can also be used in intrinsic timing models that don't use extrinsic timekeepers.

Simko & Cummins' Embodied Task Dynamics model [71,72] is an interesting case: an example of a theory of speech motor control in which time is used only as a cost (where surface utterance duration is penalized), but not a parameter of movement. In avoiding the use of time as a parameter of movement, [71,72] is similar to the Articulatory Phonology/Task Dynamics approach, discussed in more detail below. However, even though time is not a parameter of movement in this model, an extrinsic timekeeping mechanism is nevertheless required to specify and represent the utterance duration quantity that it penalizes. On the definition that we presented above in Section 2.2., we would therefore classify it as an extrinsic timing model, even though it makes less extensive use of an extrinsic timekeeper than other types of extrinsic models.

In sum, many models of motor control use extrinsic timekeepers and many of these are Optimal Control Theory models. In the following section, we discuss a different approach, that is, timing without an extrinsic timekeeper in Articulatory Phonology/Task Dynamics. Although this model currently provides the most comprehensive account of timing effects in speech production, we believe extrinsic models should be considered, for reasons laid out in Section 3.3.

3.2. Timing without an extrinsic timekeeper in Articulatory Phonology/Task Dynamics

The main theory of speech production that assumes that surface timing phenomena can be produced without an extrinsic timekeeper is Articulatory Phonology/Task Dynamics (AP/TD) [78-84]. This theory is particularly important because it currently provides the most comprehensive account of timing phenomena observed in speech, and has led to a number of significant insights into the nature of speech production, such as the understanding that coarticulation between adjacent sounds is often a matter of articulatory overlap rather than of feature changes in the phonemic features that define the words. The model is based on oscillators; this key feature enables it to produce surface timing patterns without an extrinsic timekeeper.

AP/TD is unlike traditional phonological theories which assume that units of phonological contrast are symbolic, i.e. do not contain quantitative specifications for how articulatory movement should unfold. In AP/TD, units of phonological contrast are gestures, defined as equations of motion that determine how constrictions will be formed in the vocal tract; constriction releases are modeled as movement back to a neutral vocal tract position. In this framework, each dimension of gestural movement towards a constriction goal is modelled as movement towards an equilibrium position in a damped, mass-spring system (analogous to the movement of a mass attached to a spring). The gesture's starting position is analogous to the position to which the mass attached to the spring is stretched, and the equilibrium position is the target position that is approached by the mass after releasing the spring. Because the system is critically damped, the mass doesn't oscillate, but rather asymptotes towards (approaches, but never quite reaches) the equilibrium position. It can thus be described as having point attractor dynamics. The time required to approximate a constriction target (gestural settling time) is intrinsic to the system because it is dictated by the parameters of the mass-spring oscillator, i.e. its stiffness, mass, and damping coefficients.

Other aspects of timing within AP/TD are also determined by oscillators. As we explain below, point-attractor oscillators are additionally used to adjust the timing of gestures at positions defined by prosodic structure, i.e. for final lengthening and prominence-related lengthening [82,83]. AP/TD also uses two types of freely-oscillating oscillators (i.e. oscillators with limit cycle rather than point attractor dynamics): 1) gestural planning oscillators, and 2) a hierarchy of coupled supra-segmental planning oscillators (syllable, foot, and phrase oscillators). These oscillators are used during utterance planning to determine 1) relative timing among gestures (inter-gestural coordination) 2) the amount of time that each gesture shapes the vocal tract (gestural activation), and 3) some aspects of timing attributed to supra-segmental (i.e. prosodic) structure.

In this framework, inter-gestural timing is determined by the relative phasing among gestural planning oscillators assigned to each gesture, and does not need to be specified by an extrinsic timekeeper. For example, if two gestural planning oscillators entrain in-phase during utterance planning, then the physical gestures that correspond to each planning oscillator will begin at the same time.

Other phasing relationships are also possible, but the most stable entrainment patterns are predicted to be the most common, i.e. in-phase and anti-phase. See [84] for a more complete discussion of inter-gestural timing.

The amount of time that each gesture is active (i.e. its activation interval) is derived from other parameters within the system and does not need to be specified extrinsically. Gestural activation intervals specify the amount of time that a gesture actively shapes the vocal tract. Gestures whose activation intervals are as long as their settling times, will have enough time to approximate their targets. On the other hand, if gestural activation intervals are shorter than gestural settling times, targets will not be approximated and undershoot will occur. If gestural activation intervals are longer than gestural settling times, then gestures will continue to asymptote towards their targets for the length of the activation interval (and will thus appear to be in a quasi-steady state for the duration of the activation interval).

Gestural activation interval timing is intrinsic because activation intervals are specified within the model as a fixed proportion of each planning oscillator's cycle. Because gestural planning oscillations are coupled to the oscillations of the supra-segmental hierarchy of syllable-, foot- and phrase-oscillations, the physical duration of activation will depend on the frequency of oscillation of this whole planning oscillator ensemble, i.e. on overall speech rate. When speech rate (i.e. planning oscillator ensemble frequency) increases, activation intervals will be physically shorter, and undershoot will be more likely, although gestural activation intervals will still correspond to the same gestural planning oscillator proportion. Likewise, when speech rate decreases, activation intervals will be longer, and more time will be spent asymptoting (getting closer and closer) to the gesture's target.

Temporal aspects of prosodic structure are also intrinsic to the system and do not need to be extrinsically specified. There are two aspects of prosodic timing in this framework: First, interactions among higher level organizing oscillators (e.g. syllable, foot, phrase) specify the rates of syllable, foot, and phrase production. These oscillation rates in turn affect the rates of planning oscillators for individual gestures, which determine gestural activation intervals because each activation interval corresponds to a proportion of a planning oscillator cycle. The second aspect of prosodic timing has to do with adjustments that are made to all gestures that are concurrently active within a specified interval (mentioned briefly above). For example, the lengthenings that commonly occur at prosodically-privileged positions in an utterance, e.g. boundary-related and prominence-related lengthening, are generated by proportionally stretching the activation intervals of boundary-adjacent or prominent gestures [82, 83].

Global timing, i.e. overall speech rate, is specified by the utterance-specific oscillation rate of the ensemble of supra-segmental and gestural planning oscillators, but again does not involve the specification of surface duration [83].

In the current form of AP/TD, surface timing characteristics cannot be specified, nor is there a mechanism that can keep track of the output durations while they are being produced, or measure them after they are produced. These features are not required in the model, because once speakers have chosen a rate of

speech and have imposed prosodic boundaries and prominences on an ordered sequence of gesturally-specified words, surface timing patterns emerge from the interacting mechanisms of the system.

Simko & Cummins' Embodied Task Dynamic model [71,72] is similar to AP/TD in that it uses mass-spring oscillator systems for gestures. In this model, some aspects of surface timing are emergent, i.e. they result from the stiffness specification of the mass-spring system, and other aspects result from the coordination of these oscillators in terms of their phasing. However, as discussed above, Simko & Cummins' model can't be considered a strictly intrinsic timing model because it uses an extrinsic timekeeping mechanism to represent an utterance duration cost and therefore the surface duration of each utterance.

Although the use of intrinsic timing has the advantage of minimizing the planning required for each utterance, the three kinds of evidence we present in the next section are difficult to reconcile with intrinsic timing approach adopted in AP/TD, and are suggestive of extrinsic timekeeper control.

3.3. Evidence for extrinsic timing

In sections 3.3.1 and 3.3.2 which follow, we provide evidence that challenges the intrinsic timing aspect of the AP/TD model, because it supports the use of an extrinsic timekeeping mechanism in speech and non-speech motor control. In Section 3.3.3, we present evidence which, although it is not as conclusively challenging, is difficult to account for in mass-spring systems. These lines of evidence motivate us to consider extrinsic timing models that include time as a parameter of movement.

3.3.1. Increasing variability with increases in interval duration: Evidence for an extrinsic timekeeper

Patterns of variability in the timing of intervals support an extrinsic timing mechanism. Many studies show more variability in interval duration for longer intervals defined by movement [85-92], and as explained in [86, p. 422], these findings are expected in extrinsic timing models: "the mechanism that *meters out* intervals of time … is variable, and the amount of variability is directly proportional to the length of the interval of time to be metered out." (This is because time is metered out in smaller units than the total interval, and the variability in each inter-tick interval adds up.) The relationship of variability to mean duration follows Weber's law, with an approximately constant coefficient of variation (standard deviation/mean) across a range of intervals (from tens of milliseconds to seconds and possibly longer), for both humans and animals, consistent with an extrinsic timing mechanism [84, 85, 90-94]. Support for the view that the same timekeeping mechanisms are used in perception comes from a Weber relationship between difference threshold and interval duration in perceptual discrimination tasks, e.g.. [88].

The Weber relationship between standard deviation and interval duration, suggestive of noise in a timing process and therefore of an extrinsic timing mechanism, is observed in many production tasks, including:

1) Single timed interval production, where participants reproduce a single interval to match the duration of a model, using e.g. taps [87, 88, 91, 96], for intervals ranging from 0 ms to 1050 ms.

2) Movements made to a metronome: [86], for moving a stylus to and from a target, with inter-beat intervals from 200-500 ms.

3) Movements made to an internally-generated rhythm in a continuation paradigm [89, 91, 91, 96] among others: participants first produce a movement (e.g. tapping) in synchrony with a metronome (pacing phase), and continue the rhythm after the metronome is turned off (continuation phase). Typically, interval duration measurements are made from the continuation phase; standard deviations and mean interval durations are computed over a series of trials. [91] found patterns of increased variability for longer tapping interval duration for intervals ranging from 325-500 ms. [92] observed increased variability for longer tapping and continuous circle drawing intervals, as well as back-and-forth line drawing intervals, for intervals ranging from 300-500 ms (see also [96]).

4) Speech movements and intervals. Byrd & Saltzman [99] found that variability increased with movement duration, for measured durations of lip aperture closings associated with a trans-boundary /m/-schwa-/m/ sequence. Movements of different durations were elicited in conditions designed to systematically vary the prosodic boundary strength before the second /m/. For example, the target sequence –***mam***- in *mom**mam**ia* was described as having no word boundary before its second /m/, whereas in *mom**ma m**imi* the second /m/ was separated by a word boundary from the preceding vowel. In other cases, the second /m/ was separated from the preceding vowel by a stronger boundary, and was either phrase- or utterance-initial. Movement durations were generally longer for stronger boundaries, because of constituent-final lengthening, whose magnitude increases with boundary strength (cf. [30] for acoustic measures). Data from [29] show a similar pattern for phrase-final vs. phrase-medial word-final syllable rhyme measures, based on landmarks in the acoustic signal. Rhyme duration means and standard deviations were considerably higher for phrase-final words as compared to phrase-medial words; That is, monosyllabic words (e.g. *Tom)* had phrase-final mean durations of 346 ms (82 ms s.d.) vs. phrase-medial mean durations of 193 ms (47 ms s.d.).

In Articulatory Phonology/Task Dynamics, longer movement durations at phrase boundaries arise by stretching the activation intervals in the vicinity of the boundary, that is, by **decreasing** the oscillation rate of a planning oscillator ensemble in a specified interval, while leaving the number of oscillations the same. Within this framework, therefore, there are no additional "ticks" of an utterance-specific clock that could be used to explain the source of the additional temporal variability. Thus, the substantial body of evidence supporting increased variability with longer-duration movements is inconsistent with the AP/TD model of motor timing.

3.3.2. Surface timing constraints and goal specifications: Evidence that surface durations are part of the phonetic plan for an utterance

Within Articulatory Phonology/Task Dynamics, desired surface durations can't be specified as part of the utterance plan. For example, gesture durations in phrase-final position reflect the settling-time of their mass-spring system, their gestural activation interval, and an adjustment which lengthens the gestural activation intervals at the boundary [82]. But in AP/TD, the surface duration emerges from these mechanisms alone, and cannot be specified in the original utterance plan.

However, [100] suggest that a constraint on surface durations of phonemically short vowels in phrase-final position may be required to preserve the short vs. long phonemic contrast in Northern Finnish. In Northern Finnish disyllabic words with a phonemically short vowel in the word-final syllable (CVCV(C)), the final-syllable vowel is described as half-long because its duration is intermediate between that of the short vowel in other contexts and that of the contrasting long vowel (VV). The authors observed that the magnitude of final, accentual, and combined lengthening on the half-long vowel was restricted compared to lengthening on the other vowel types (e.g. 17% combined accentual + final lengthening on the half-long vowel vs. 68% on the long vowel in the same context). Support for a surface duration constraint also comes from observations that lengthening magnitudes were smaller for half-long vowels with longer phrase-medial durations; [100] found a negative correlation between phrase-medial half-long vowel durations and the magnitude of phrase-final lengthening. These results are consistent with the view that the surface durations of the (phonemically short) half-long vowel are restricted in order to avoid endangering the phonemic short vs. long vowel quantity contrast in this language. Although it is possible *implement* this type of effect in AP/TD, the effect is difficult to *explain* within the theory, since surface durations can't be measured, represented, or referred to as motivating factors.

Additional support for the representation of surface durations can be found in studies of rate of speech effects and durational correlates of prosodic structure and quantity [101-103]. These studies find that there is considerable variability in the strategies that different speakers use to implement these factors, but that nevertheless speakers all achieve a common surface duration pattern of relatively long surface durations e.g. in phrase-final position, at slow speech rates, and for phonemically long vowels. These findings challenge intrinsic timing in AP/TD because they suggest the equivalence of different strategies that result in similar surface duration patterns, and therefore support the specification of surface duration goals.

Summary: The two types of evidence we presented in sections 3.3.1 and 3.3.2 strongly support the use of extrinsic timekeepers to measure, represent and specify surface movement and/or interval durations in speech. This evidence therefore supports models like DIVA/VITE in which duration is a planned parameter of movement. Although duration is not a parameter of movement in Simko & Cummins' [71,72] model, this model could probably be modified to account for these data, since Simko & Cummins use an extrinsic timekeeper to

specify an utterance duration cost. However, it might need to be amended to measure, represent and specify durations of constituents smaller than the utterance. Currently in their model, although whole-utterance duration cost specification requires an extrinsic timer, surface timing of constituents smaller than the utterance (e.g. syllables, individual gestures) arise from phasing relations among gestures and from gestural stiffness, and is not specified directly. In Section 3.3.3. we present evidence which challenges this approach as well as AP/TD's approach because it is difficult to account for in mass-spring systems. This evidence therefore motivates the consideration of extrinsic timing models of speech production which include time as a parameter of movement (and not simply as a cost).

3.3.3. Independent planning of the timing of movement onset vs. target attainment: Evidence difficult to account for in mass-spring models

In [69], Lee commented "it is frequently not critical when a movement starts—just so long as it does not start too late. For example, an experienced driver who knows the car and road conditions can start braking safely for an obstacle a bit later than an inexperienced driver…" This type of example suggests that timing variability may be different at target attainment vs. movement onset, difficult to account for in mass-spring models such as AP/TD, but relatively straightforward to account for in extrinsic timing models that allow separate timing specification and prioritization for target attainment vs. other parts of movement [104].

Several studies have confirmed the finding of differential variability in the timing of target attainment, compared to the timing of other movement events such as movement onset ([92, 105-108], for non-speech motor activity; [109] for speech). For example, [106] showed that the timing of initiating forehand drives in table tennis was more than twice as variable as the timing of paddle contact with the ball. Forehand drives in this experiment had average movement times that ranged between 92 and 178 ms. Timing accuracy at paddle-ball contact was estimated on the basis of the ratio of standard deviation of the direction of travel of the paddle and its mean rate of change, and was calculated to be within 2-5 ms. In contrast, movement time standard deviations ranged from 5-21 ms, depending on the player, showing that movement initiation times were much more variable.

[109] showed a similar pattern of timing variability for upper lip protrusion movements during spoken /i_u/ sequences, where the number of intervocalic consonants varied systematically. They observed lower variability in the timing of target attainment (maximum protrusion) relative to voicing onset for /u/, as compared to the timing of a point shortly after movement onset (maximum acceleration), relative to voicing onset for the same vowel. This pattern suggests a tighter temporal coordination of maximum lip protrusion with voicing onset than of lip protrusion movement onset with voicing onset. These findings suggest that target attainment timing is controlled independently of movement onset timing, and that target attainment timing takes higher priority. These findings are not predicted by mass-spring models in which the timing of movement onset is not independent from the timing of target achievement. That is, while AP/TD does provide a mechanism for separately adjusting the timing of the beginning and the end of an activation interval (by applying its prosodic

"stretching" mechanism to a proportion of the interval), it doesn't provide a mechanism by which these timings could be differently variable.

In contrast, an extrinsic timing mechanism can in principle 1) plan the timing of movement onset independently of the timing of target attainment, and 2) account for the possibility of different degrees of variability in these two time points, as would be the case if the timing of target attainment has a higher priority than the timing of the movement onset, resulting in online adjustments to achieve high priority goals.

The separate control of different parts of a movement is also supported by evidence from spatial variability at target achievement vs. other parts of movement. The first line of evidence for differential degrees of variability at different points in a movement trajectory comes from work by Todorov & Jordan [68], who found lower spatial variability at target achievement compared to elsewhere in movements, for a task in which participants moved a pointer through a series of circular targets on a flat table[2]. When analyzing their results, they sampled each movement trajectory at 100 equally spaced points along the path. They computed the average movement path, and determined spatial deviations from the average path at each of the 100 points. Results showed that spatial deviations from the average path were lowest at the circular targets, and higher in between. [110] report similar results for shorter-than-a-second reaching movements (variability greater for 1st half of reaching movement, compared to the second half as the hand approached the target), as do [111] for two reaching tasks. [111] found that spatial variability was lowest at the beginning and end of each movement, and highest in between. Presumably the variability was low at the beginning of movement because the movements started from a fixed point, and was low at the end of movement because the end was the target.

These results suggest that actors are able to identify parts of a trajectory that relate most closely to task performance, and are able to prioritize spatial accuracy in these parts of the trajectory. The results are also consistent with the view that actors make use of a feedback-based error correction system to implement error correction in the parts of the trajectory whose accuracy has been prioritized. On this view, errors in planned movement trajectories (as evidenced by deviations from the mean) can be left uncorrected if they do not interfere with task performance. These findings suggest that models like AP/TD might need to be modified to incorporate a feedback-based correction system. In addition, these data suggest that separate parts of movement are identified so that spatial accuracy can be prioritized, something that would be straightforward if these same points were also identified for differential timing prioritization in an extrinsic timing model. Different degrees of spatial (as well as timing) variability at different parts of movement may be difficult, though perhaps not impossible, to implement in mass-spring systems.

[2] Target-to-target movement durations were comparable to those observed in speech (i.e. approximately 100-400 ms).

3.4. Part II: Summary

In this section, we reviewed two types of timing control theory: 1) without an extrinsic timekeeper, exemplified by Articulatory Phonology/Task Dynamics, and 2) with an extrinsic timekeeper, exemplified by DIVA and many types of Optimal Control Theory models. Several findings challenge models such as AP/TD that don't make use of an extrinsic timekeeper. These findings include greater timing variability for longer duration intervals as compared to shorter duration intervals, the apparent use of a durational constraint in Northern Finnish, as well as the use of different strategies to achieve the same duration patterns as a speech planning goal. Additionally, we presented evidence of differential timing variability at movement end as compared to movement onset. This evidence is does not provide direct support for the use of an extrinsic timekeeper, but is more straightforward to account for in models which include one. Models of speech timing control that involve an extrinsic timekeeper are therefore worth investigating, although they will require extensive development to account for the range of phenomena currently captured by Articulatory Phonology/Task Dynamics.

4. Conclusion

Understanding speech timing requires an understanding of both what timing is used for, and how it is controlled. We propose that one goal of speech timing is to make speech understandable, and that this goal is balanced against other goals, such as speaking quickly, to give the surface timing properties of speech. This view is based on findings from controlled experiments, as well as from analyses of relationships among factors proposed to account for surface timing patterns. We also presented two alternative ways of modelling surface timing patterns 1) as an emergent property of motor control, without the involvement of an extrinsic timekeeper (as in Articulatory Phonology/Task Dynamics) and 2) as the result of desired durational specifications made possible by an extrinsic timekeeper (as in DIVA/VITE and many Optimal Control Theory models, where desired durational specifications are balanced against other task requirements and costs to generate (near-)optimal movements). Although the Articulatory Phonology/Task Dynamics framework currently exceeds other models in its ability to account for speech timing phenomena, several findings present challenges for this framework, and raise the possibility that models of motor control that involve an extrinsic timekeeper may ultimately provide a simpler and more comprehensive account of speech timing behaviour.

While there are aspects of what timing is used for, and of the structures that govern it, that still remain to be discovered, our current understanding of these two aspects of speech timing is more advanced than our understanding of the mechanisms that are used to control it. It is hoped that advances in experimentation, modelling, and neuroscience will eventually lead to a better match between our understanding of speech timing patterns and our models of how these patterns arise.

5. Acknowledgements

We thank Jelena Krivokapic and two anonymous reviewers for useful comments on previous versions of this manuscript, Elliot Saltzman and Louis Goldstein for tutorial discussions on Articulatory Phonology/Task Dynamics, and Dave Lee for discussions of General Tau theory. Any errors are ours. This work was supported by an Arts and Humanities Research Council fellowship (AH/1002758/1) to the first author, and NIH R01-DC008780 to the second author.

6. References

[1] Turk, A., Nakai, S., & Sugahara, M. (2006). Acoustic segment durations in prosodic research: a practical guide. In S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter & J. Schliesser (Eds.), *Methods in Empirical Prosody Research* (pp. 1-28). Berlin, New York: De Gruyter.

[2] 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*(4), 1873-1891.

[3] Perkell, J. S., Zandipour, M., Matthies, M. L., & Lane, H. 2002 Economy of effort in different speaking conditions. I. A preliminary study of intersubject differences and modeling issues. *Journal of the Acoustical Society of America, 112*(4), 1627-1641.

[4] Peterson, G., & Lehiste, I. 1960 Duration of syllable nuclei in English. *Journal of the Acoustical Society of America 32*(6), 693-703.

[5] Klatt, D. H. 1976 Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America, 59*(5), 1208-1221.

[6] Fletcher, J. 2010 The Prosody of Speech: Timing and Rhythm. In W. J. Hardcastle, J. Laver & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (2nd edition ed., pp. 521-602): Blackwell.

[7] Delattre, P. C. 1966 A comparison of syllable length conditioning among languages. *International Review of Applied Linguistics, 4*(183-198).

[8] Lindblom, B. 1968 Temporal organization of syllable production *Reports of the 6th International Congress of Acoustics* (pp. B29-B30). Tokyo.

[9] Lehiste, I. 1970*Suprasegmentals*. Cambridge, MA: The MIT Press.

[10] Nooteboom, S. 1972 *Production and perception of vowel duration; a study of durational properties of vowels in Dutch.* Unpublished PhD dissertation. University of Utrecht.

[11] van Santen, J. P. H. 1992 Contextual effects on vowel duration. *Speech Communication, 11*, 513-546.

[12] Selkirk, E. O. 1978 On prosodic structure and its relation to syntactic structure. In T. Fretheim (Ed.), *Nordic Prosody II* (pp. 111-140). Trondheim: TAPIR.

[13] Nespor, M., & Vogel, I. 1986 *Prosodic Phonology*. Dordrecht: Foris Publications.

[14] Watson, D., & Gibson, E. 2004The relationship between intonational phrasing and syntactic structure in language production. *Language and Cognitive Processes, 19*(6), 713–755.

[15] Watson, D., Breen, M., & Gibson, E. 2006 The role of syntactic obligatoriness in the production of intonational boundaries. *Journal of Experimental Psychology: Learning, Memory and Cognition, 32*, 1045–1056.

[16] Astésano, C., Bard, E. G., & Turk, A. 2007 Structural influences on initial accent placement in French. Language and Speech 50(3), 423-446.
[17] Kainada, E. 2010 *Phonetic and phonological nature of prosodic boundaries: evidence from Modern Greek.* Unpublished PhD dissertation. The University of Edinburgh.
[18] Jackendoff, R. personal communication
[19] Chomsky, N., & Halle, M. 1968 *The Sound Pattern of English*. New York: Harper & Row.
[20] Beckman, M. E., & Pierrehumbert, J. B. 1986 Intonational structure in Japanese and English. *Phonology Yearbook 3*, 255-309.
[21] Gee, J. P., & Grosjean, F. 1983 Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology, 15*, 411-458.
[22] Jun, S.-A. (Ed.) 2005 *Prosodic Typology: The Phonology of Intonation and Phrasing* Oxford University Press.
[23] Vaissière, J. 1983 Language Independent Prosodic Features. In A. Cutler & D. R. Ladd (Eds.), *Prosody: Models and Measurements* (pp. 53-65): Springer Verlag.
[24] White, L. 2002 *English speech timing: a domain and locus approach.* Unpublished Ph.D. dissertation, U. of Edinburgh, Edinburgh, UK.
[25] Keating, P., Cho, T. , Fougeron, C. , & Hsu, C-S. 2003 Domain-initial strengthening in four languages. In J. Local, R. Ogden & R. Temple (Eds.), *Phonetic interpretation: Papers in laboratory phonology VI* (pp. 143–161). Cambridge: Cambridge University Press.
[26] Cho, T., & Keating, P. 2009 Effects of initial position versus prominence in English. *Journal of Phonetics, 37*, 466-485.
[27] Bombien, L., Mooshammer, C., Hoole, P., & Kühnert, B. 2010 Prosodic and segmental effects on EPG contact patterns of word-initial German clusters. *Journal of Phonetics, 38*, 388-403.
[28] Cambier-Langeveld, T. 1997 The domain of final lengthening in the production of Dutch. In J. Coerts & H. d. Hoop (Eds.), *Linguistics in the Netherlands* (pp. 13-24).
[29] Turk, A. E., & Shattuck-Hufnagel, S. 2007 Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics, 35*(4), 445-472.
[30] Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. 1992 Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America, 91*(3), 1707-1717.
[31] Keating, P. 2006 Phonetic encoding of prosodic structure. In J. Harrington & M. Tabain (Eds.), *Speech production: Models, phonetic processes, and techniques* (pp. 167-186). New York and Hove: Psychology Press.
[32] Turk, A. E., & Shattuck-Hufnagel, S. 2000 Word-boundary-related duration patterns in English. *Journal of Phonetics, 28*, 397-440.
[33] Turk, A. 2012 The temporal implementation of prosodic structure. In A. Cohn, C. Fougeron & M. Huffman (Eds.), *The Oxford Handbook of Laboratory Phonology* (pp. 242-253). Oxford: Oxford University Press.
[34] Turk, A., & Shattuck-Hufnagel, S. 2013 What is speech rhythm? A commentary inspired by Arvaniti & Rodriquez, Krivokapić, and Goswami & Leong. *Laboratory Phonology,* 4(1), 93-118.
[35] Pierrehumbert, J., & Talkin, D. 1992 Lenition of /h/ and glottal stop. In G. J. Docherty & D. R. Ladd (Eds.), *Papers in Laboratory Phonology II: Gesture, Segment Prosody*: Cambridge University Press.

[36] Ogden, R. 2004 Non-modal voice quality and turn-taking in Finnish. In E. Couper-Kuhlen & C. Ford (Eds.), *Sound Patterns in Interaction: Cross-Linguistic Studies from Conversation* (Vol. 29-62). Amsterdam: John Benjamins.
[37] Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. 1996 Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics, 24*(4), 423-444.
[38] Redi, L., & Shattuck-Hufnagel, S. 2001 Variation in the realization of glottalization in normal speakers. *Journal of Phonetics, 29*(4), 407-429.
[39] Tanaka, H. 2004 Prosody for marking transition-relevance places in Japanese conversation: The case of turns unmarked by utterance-final objects. In E. Couper-Kuhlen & C. Ford (Eds.), *Sound Patterns in Interaction: Cross-linguistic Studies From Conversation* (pp. 63–96). Amsterdam: John Benjamins.
[40] Fougeron, C., & Keating, P. 1997 Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America, 101*, 3728-3740.
[41] Keating, P., Cho, T., Fougeron, C., & Hsu, C.-S. 2003 Domain-initial strengthening in four languages. In J. Local, R. Ogden & R. Temple (Eds.), *Phonetic interpretation: Papers in laboratory phonology VI* (pp. 143–161). Cambridge: Cambridge University Press.
[42] Lavoie, L. 2001 *Consonant Strength: Phonological Patterns and Phonetic Manifestations*. London: Routledge.
[43] Shattuck-Hufnagel, S., Ostendorf, M., & Ross, K. 1994 Stress shift and early pitch accent placement in lexical items in American English. *Journal of Phonetics, 22*, 357-388.
[44] Ladd, D. R. 2008 *Intonational Phonology.* (2nd ed.): Cambridge University Press.
[45] Hayes, B. 1983 A grid-based theory of English meter. *Linguistic Inquiry, 14*(3), 357-393.
[46] Selkirk, E.O. 1984 Phonology and Syntax: the relation between sound and structure. Cambridge, MA: MIT Press.
[47] Halle, M., & Vergnaud, J.-R. 1987 *An Essay on Stress*. Cambridge: MIT Press.
[48] Beckman, M. E., & Edwards, J. 1994 Articulatory evidence for differentiating stress categories. In P. Keating (Ed.), *Papers in Laboratory Phonology III: Phonological Structure and Phonetic Form*. Cambridge: Cambridge University Press.
[49] Beckman, M. E., & Edwards, J. 1992 Intonational categories and the articulatory control of duration. In Y. Tohkura, E. Vatikiotis-Bateson & Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure*. Tokyo: OHM Publishing Co., Ltd.
[50] Turk, A. E., & Sawusch, J. R. 1997 The domain of accentual lengthening in American English. *Journal of Phonetics, 25*, 25-41.
[51] Mo, Y., Cole, J., & Hasegawa-Johnson, M. 2010 *Prosodic effects on temporal structure of monosyllabic CVC words in American English.* Paper presented at the 5th Speech Prosody conference, Chicago, IL.
[52] Heuven, V. J. J. P. v., & Sluijter, A. M. C. 1996 Notes on the phonetics of word prosody. In R. Goedemans, H. v. d. Hulst & E. Visch (Eds.), *Stress patterns of the world, part 1: background (HIL Publications)* (pp. 233-269). The Hague: Holland Academic Graphics.
[53] Cho, T. 2006 Manifestation of prosodic structure in articulation: Evidence from lip kinematics in English. In L. Goldstein, D. H. Whalen & C. T. Best (Eds.), *Laboratory Phonology* (Vol. 8, pp. 519-548). Mouton de Gruyter.
[54] Shattuck-Hufnagel, S., & Turk, A. 1996 A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research, 25*(2),

193-247.
[55] Keating, P., & Shattuck-Hufnagel, S. 2002 A prosodic view of word form encoding for speech production. *UCLA Working Papers in Phonetics, 101*, 112-156.
[56] Caspers, J. 1994 *Pitch movements under time pressure: Effects of speech rate on the melodic marking of accents and boundaries in Dutch*. The Hague: Holland Academic Graphics
[57] Lieberman, P. 1963 Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech, 6*(3), 172-187.
[58] Fowler, C., & Housum, J. 1987 Talkers' signaling of "New" and "Old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language, 26*, 489–504.
[59] Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. 2001 Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (Eds.), *Frequency and the Emergence of Linguistic Structure* (pp. 229–254). Amsterdam: John Benjamins.
[60] Bell, A., Brenier, J., 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.
[61] Aylett, M. P. 2000 *Stochastic suprasegmentals: Relationships between redundancy, prosodic structure and care of articulation in spontaneous speech.*, Unpublished PhD dissertation, The University of Edinburgh.
[62] Aylett, M., & Turk, A. 2004 The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration on spontaneous speech. Language and Speech, 47, 31-56.
[63] Turk, A. 2010 Does prosodic constituency signal relative predictability? A Smooth Signal Redundancy hypothesis. *Journal of Laboratory Phonology, 1*, 227-262.
[64] Gahl, S., & Garnsey, S. 2004 Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language, 80*, 748–775.
[65] Guenther, F. H. 1995 Speech Sound Acquisition, Coarticulation, and Rate Effects in a Neural-Network Model of Speech Production. *Psychological Review, 102*(3), 594-621.
[66] Guenther, F. 2006 Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders 39* 350–365.
[67] Bullock, D., & Grossberg, S. 1988 Neural Dynamics of Planned Arm Movements - Emergent Invariants and Speed Accuracy Properties during Trajectory Formation. *Psychological Review, 95*(1), 49-90.
[68] Todorov, E., & Jordan, M. I. 2002 Optimal feedback control as a theory of motor coordination. *Nature Neuroscience, 5*(11), 1226-1235.
[69] Lee, D. N. 1998 Guiding movement by coupling taus. *Ecological Psychology, 10*(3-4), 221-250.
[70] Shadmehr, R., & Mussa-Ivaldi, S. 2012 *Biological learning and control: How the brain builds representations, predicts events, and makes decisions.* Cambridge, MA: The MIT Press.
[71] Simko, J., & Cummins, F. 2010 Embodied Task Dynamics. *Psychological Review, 117*(4), 1229-1246.
[72] Simko, J., & Cummins, F. (2011). Sequencing and optimization within an embodied Task Dynamic model. *Cognitive Science, 35*, 527–562.
[73] Shadmehr, R., Orban de Xivry, J. J., Xu-Wilson, M., & Shih, T.-Y. 2010

Temporal discounting of reward and the cost of time in motor control. *The Journal of Neuroscience, 30*(31), 10507–10516.

[74] Harris, C. M., & Wolpert, D. M. 2006 The main sequence of saccades optimizes speed-accuracy trade-off. *Biological Cybernetics, 95*(1), 21-29.

[75] Tanaka, H., Krakauer, J. W., & Qian, N. 2006 An optimization principle for determining movement duration. *Journal of Neurophysiology, 95*, 3875-3886.

[76] Hancock, P. A., & Newell, K. M. 1985 The movement speed-accuracy relationship in space-time. In H. Heuer, U. Kleinbeck & K.-H. Schmidt (Eds.), *Motor Behavior: Programming, Control, and Acquisition* (pp. 153-185). Berlin: Springer-Verlag.

[77] Jimura, K., Myerson, J., Hilgard, J., Braver, T. S., & Green, L. 2009 Are people really more patient than other animals? Evidence from human discounting of real liquid rewards. *Psychonomic Bulletin & Review, 16*(6), 1071-1075.

[78] Browman, C. P., & Goldstein, L. 1985 Dynamic modeling of phonetic structure. In V. A. Fromkin (Ed.), *Phonetic Linguistics* (pp. 35-53). New York: Academic Press.

[79] Browman, C. P., & Goldstein, L. 1992 Articulatory phonology: an overview. *Phonetica, 49*(3-4), 155-180.

[80] Saltzman, E., & Kelso, J. A. S. 1987 Skilled Actions: A Task-Dynamic Approach. *Psychological Review, 94*(1), 84-106.

[81] Saltzman, E. L., & Munhall, K. 1989 A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1*(4), 333-382.

[82] Byrd, D., & Saltzman, E. 2003 The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics, 31*(2), 149-180.

[83] Saltzman, E., Nam, H., Krivokapic, J., & Goldstein, L. 2008 *A task-dynamic toolkit for modeling the effects of prosodic structure on articulation.* Paper presented at the Speech Prosody 2008, Campinas, Brazil.

[84] Nam, H., Goldstein, L., & Saltzman, E. 2010 Self-organization of syllable structure: a coupled oscillator model. In F. Pellegrino, E. Marisco & I. Chitoran (Eds.), *Approaches to phonological complexity*. Berlin, New York: Mouton de Gruyter.

[85] Treisman, M. 1963 Temporal discrimination and the indifference interval - Implications for a model of the internal clock. *Psychological Monographs, 77*(13), 1-31.

[86] Schmidt, R. A., Zelaznik, H., Hawkins, B., Frank, J. S., & Quinn, J. T. 1979 Motor-output variability: A theory for the accuracy of rapid motor acts. *Psychological Review, 86*(5), 415-451.

[87] Rosenbaum, D. A., & Patashnik, O. 1980 A mental clock setting process revealed by reaction times. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in Motor Behavior* (pp. 487-499): North-Holland Publishing Company.

[88] Rosenbaum, D. A., & Patashnik, O. 1980 Time to time in the human motor system. In R. S. Nickerson (Ed.), *Attention and performance* (Vol. VIII). Hillsdale, New Jersey: Erlbaum.

[89] Wing, A. M. 1980 The long and short of timing in response sequences. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in Motor Behavior* (pp. 469-484): North-Holland Publishing Company.

[90] Ivry, R., & Corcos, D. M. 1993 Slicing the variability pie - component analysis of coordination and motor dysfunction. *Variability and Motor Control*, 415-447.

[91] Ivry, R. B., & Hazeltine, R. E. 1995 Perception and production of temporal intervals across a range of durations - Evidence for a common timing mechanism. *Journal of Experimental Psychology-Human Perception and Performance, 21*(1), 3-18.

[92] Spencer, R. M. C., & Zelaznik, H. N. 2003 Weber (slope) analyses of timing

variability in tapping and drawing tasks. *Journal of Motor Behavior, 35*(4), 371-381.
[93] Gibbon, J. 1977 Scalar Expectancy Theory and Weber's law in animal timing. *Psychological Review, 84*(3), 279-325.
[94] Gibbon, J., Malapani, C., Dale, C. L., & Gallistel, C. R. 1997 Toward a neurobiology of temporal cognition: Advances and challenges. *Current Opinion in Neurobiology, 7*(2), 170-184.
[95] Buhusi, C., & Meck, W. H. 2005 What makes us tick? Functional and neural mechanisms of interval timing. *Nature Reviews Neuroscience, 6*, 755-765.
[96] Merchant, H., Zarco, W., Bartolo, R., & Prado, L. 2008 The context of temporal processing is represented in the multidimensional relationships between timing tasks. *PLoS ONE, 3*(9), e3169.
[97] Lewis, P. A., & Miall, R. C. 2009 The Experience of Time: Neural Mechanisms and the Interplay of Emotion, Cognition and Embodiment. *Philosophical Transactions: Biological Sciences, 364*(1525), 1897-1905.
[98] Robertson, S. D., Zelaznik, H. N., Lantero, D. A., Bojczyk, K. G., Spencer, R. M., Doffin, J. G., et al. 1999 Correlations for timing consistency among tapping and drawing tasks: Evidence against a single timing process for motor control. *Journal of Experimental Psychology-Human Perception and Performance, 25*(5), 1316-1330.
[99] Byrd, D., & Saltzman, E. 1998 Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics, 26*(2), 173-199.
[100] Nakai, S., Turk, A., Suomi, K., Granlund, S., Ylitalo, R., & Kunnari, S. 2012 Quantity and constraints on the temporal implementation of phrasal prosody in Northern Finnish.
[101] Berry, J. 2011 Speaking rate effects on normal aspects of articulation: Outcomes and issues. *Perspectives on Speech Science and Orofacial Disorders, 21*, 15-26.
[102] Edwards, J., Beckman, M.E., & Fletcher, J. 1991 The articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America, 89*(1), 369-382.
[103] Hertrich, I., & Ackermann, H. 1997 Articulatory control of phonological vowel length contrasts: Kinematic analysis of labial gestures. *Journal of the Acoustical Society of America, 102*(1), 523-536.
[104] Shaffer, L. H. 1982 Rhythm and timing in skill. *Psychological Review, 89*(2), 109-122.
[105] Billon, M., Semjen, A., & Stelmach, G. E. 1996 The timing effects of accent production in periodic finger-tapping sequences. *Journal of Motor Behavior, 28*(3), 198-210.
[106] Bootsma, R., & van Wieringen, P. C. 1990 Timing an attacking forehand drive in table tennis. *Journal of Experimental Psychology: Human Perception and Performance, 16*(1), 21-29.
[107] Craig, C., Pepping, G. J., & Grealy, M. 2005 Intercepting beats in predesignated target zones. *Experimental Brain Research, 165*(4), 490-504.
[108] Zelaznik, H. N., & Rosenbaum, D. A. 2010 Timing Processes Are Correlated When Tasks Share a Salient Event. *Journal of Experimental Psychology-Human Perception and Performance, 36*(6), 1565-1575.
[109] Perkell, J. S., & Matthies, M. L. 1992 Temporal measures of anticipatory labial coarticulation for the vowel /u/ - within-subject and cross-subject variability. *Journal of the Acoustical Society of America, 91*(5), 2911-2925.
[110] Paulignan, Y., MacKenzie, C., Marteniuk, R., & Jeannerod, M. 1991 Selective perturbation of visual input during prehension: 1. The effects of changing object

position. *Experimental Brain Research, 83*, 502-512.
[111] Liu, D., & Todorov, E. 2007 Evidence for the flexible sensorimotor strategies predicted by Optimal Feedback Control. *The Journal of Neuroscience, 27*(35), 9354 –9368.

Figure Captions

Figure 1. Schematic diagram of the syntactic structure for "This is the cat that ate the rat that ate the cheese" (top), and a possible prosodic structure for the same utterance (bottom).

Figure 2. An example prosodic structure for *Mary's cousin George baked the cake.*

Figure 3. A grid-like representation of prominence structure for *Mary's cousin George.*

Figure 4. Prosodic structure as the interface between language and speech. Based on a similar figure in [54], illustrating some of the factors that influence phonetic planning. This diagram is intended as a tool for identifying and thinking about factors that influence phonetic planning, and is a proposal for how they interact.

Figure 5. The complementary relationship between predictability (language redundancy) and acoustic salience yields smooth signal redundancy (equal recognition likelihood throughout an utterance). Based on a similar figure in [63].

Figure 6. Factors that shape surface phonetics and their relationship to predictability, acoustic salience, and recognition likelihood. Based on similar figures in [62,63]

Short Title for Page Headings: Timing in Talking