A sketch of an extrinsic timing model of speech production

Alice Turk [1], *Stefanie Shattuck-Hufnagel*[2]

[1] Department of Linguistics & English Language, University of Edinburgh, Edinburgh, Scotland
[2] Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA

turk@ling.ed.ac.uk, sshuf@mit.edu

Abstract

In this paper, we motivate and present a sketch of an extrinsic timing model of speech production. It is a three-stage model, involving 1) phonological planning, where symbolic segmental representations are sequenced and slotted into an appropriate prosodic structure, and where appropriate acoustic cues are selected for each segment in its context, and 2) phonetic planning, where cues are mapped onto sets of articulators, and appropriate values for spatial and temporal parameters of movement are computed, and 3) phonetic implementation, where articulator movements are generated, monitored, and updated. We cite model components from the literature that accomplish many of the required functions.

Index Terms: speech production, extrinsic timing, prosodic structure

1. Introduction

Articulatory Phonology/Task Dynamics (hereafter AP/TD [1,2]) is the model of speech production that currently provides the most comprehensive account of speech timing phenomena. Timing control in this model is intrinsic, that is, surface timing patterns emerge from properties of the system and do not need to be represented, specified, or tracked during an utterance using a system-extrinsic timekeeper. However, several lines of behavioral evidence challenge intrinsic timing as implemented in AP/TD, and support the view that timing control in speech production is extrinsic. In this paper, we first present three types of evidence that support extrinsic timing in speech production, and then discuss a preliminary sketch of an alternative model of speech production that involves symbolic phonological representations and extrinsic timing. We point out model components from the literature that can be used to implement the model.

2. Evidence for extrinsic timing

2.1. Increasing variability with increases in interval duration, as predicted by a "noisy timekeeper" model

Many studies show more variability in interval duration for longer intervals in a variety of motor tasks [3]; for speech production, see e.g. [4]. As explained in [4, 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) for a range of intervals in both humans and animals, consistent with an extrinsic timing mechanism [5,6].

2.2. Surface timing constraints and goal specifications suggest extrinsic timing

Within AP/TD, desired surface durations aren't specified as part of the utterance plan, but instead emerge from interacting components within the task dynamical system. 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 [7,8]. In AP/TD, the surface duration emerges from these mechanisms and is not explicitly specified in the original utterance plan. However, [9] 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. The authors observed that the magnitude of final, accentual, and combined lengthening on phonemically short vowels in word-final syllables was restricted compared to lengthening on phonemically long vowels (17% combined accentual + final lengthening on phonemically short vowels in a word-final syllables vs. 68% on long vowels in the same context). These results are consistent with the view that the surface durations of the phonemically short vowel are restricted in order to avoid endangering the phonemic short vs. long vowel quantity contrast in this language. Although it is possible to *implement* this type of effect in AP/TD, the effect is difficult to *explain* within the theory, since surface durations can't be referred to. Additional support for the representation of surface durations can be found in studies of speech rate effects and durational correlates of prosodic structure and quantity [10-12]; despite considerable variability in the strategies that different speakers use to implement these factors, 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.

2.3. Separate control of movement targets vs. onsets challenges intrinsic mass-spring models

In [13], Dave 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

10.21437/SpeechProsody.2014-36

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 [14].

Several studies have confirmed the differential variability in the timing of target attainment compared to the timing of other movement events such as movement onset ([15-18], for non-speech motor activity; [19] for speech). For example, [19] showed differences in of timing variability for onsets vs. target attainment for upper lip protrusion movements during spoken /i_u/ sequences. 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. These findings suggest that target attainment timing is controlled independently of movement onset timing, and that target attainment timing takes higher priority. Similar findings of differential variability at target attainment vs. elsewhere in movement have been observed for spatial characteristics of repeated non-speech movements, where spatial variability is lowest at a movement target and higher elsewhere, e.g. [20]. These findings add further support for the separate control of targets vs. other parts of movement.

3. Key features

The key feature of our proposed model sketch is that it involves extrinsic timing, with a way to assign different priorities to the timing of movement targets vs. other parts of movement such as onsets. Extrinsic timing implies a-temporal representations, and we therefore assume that representations are symbolic, because symbolic representations are a type of a-temporal representation. We favor symbolic representations because they offer a better account of phonological equivalence than alternative a-temporal representations such as spatial paths without timing. Mechanisms of phonetic implementation are required to map these symbolic representations onto their surface phonetic form. We therefore assume a three-stage model, involving 1) phonological planning, and 2) phonetic planning, and 3) phonetic implementation. We assume that timing specification is a part of phonetic planning that is separate from the specification of spectral/spatial information (see [21] for a similar view). Timing information is combined with spatial information to generate movements intended to get to their targets on time. We discuss the planning and implementation stages in more detail below.

4. Phonological planning

We assume that phonological planning involves sequencing symbolic segmental representations and slotting them into a prosodic frame that includes hierarchical constituent and prominence structure [22]. Following [23,24], we hypothesize that prosodic structure is planned with the goal of an even distribution of recognition likelihood by the listener throughout an utterance (called smooth signal redundancy). To this end, predictability information (from language and real-world context) is used to plan prosodic structure so that relatively unpredictable elements are highlighted, either by manipulating relative prosodic prominence, or by manipulating relative prosodic boundary strength (highlighting through edge demarcation). In the planning stage, other task requirements are identified, such as speaking quickly, or in a particular style (e.g. clear speech, periodic speech, etc.), as illustrated in Figure 1. These requirements are assigned relative priorities so that, in the Phonetic Implementation stage, they can be balanced against movement costs to yield optimal movements (see below and Figure 1).

Several aspects of Figure 1 are worthy of comment. First, the effects of predictability on planned phonetic form are assumed to be indirect, where predictability affects plnaned prosodic structure, and prosodic structure in turn affects planned surface phonetics. This view represents our current hypothesis, but we note that it is possible that predictability might have additional direct effects on phonetic form (in addition to those that are mediated by prosodic form). Second, we assume that the effects of non-grammatical factors, like rate and style of speech, on phonetic form have a direct effect on planned surface phonetics. Although these factors have been observed to affect aspects of prosody (e.g. fewer "breaks" at faster rates of speech, cf. [25]), our view is that a speaker would plan 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 correlates of this structure would be different at different rates because the rate of speech requirement would be balanced against the prosodic structure requirement in determining optimum phonetic characteristics that meet the competing demands. Third, the list of factors mentioned in the "Non-grammatical factors" box is intended to be a preliminary indicator of the many non-grammatical factors that might be at work, and may not be exhaustive.

At the planning stage, each symbolic representation in its context (prosodic, stylistic, etc.) is associated with a set of acoustic cues [26]. For example, in syllable initial position English /t/ might be associated with silence, then a relatively high frequency release burst + aspiration noise, but with a different set of cues in syllable final and ambisyllabic position.

5. Phonetic planning

Phonetic planning involves a) mapping cues onto sets of articulators, and b) assigning values to a set of movement parameters, including timing and spatial parameters (and perhaps accuracy goals).

Selected acoustic cues are mapped onto quantitative acoustic/constriction goals, which are achieved by sets of articulators, or synergies (see [27] for a plausible neural network model of the mapping between acoustics and articulatory goals). We assume that the constriction goals are very similar to the set defined by Saltzman & Munhall [2], and adopted by [27].

We assume that movement parameters include spatial aspects of the constriction target, e.g. lip aperture, tongue tip constriction with the alveolar ridge, etc., a default relative contribution of each articulator in a synergy, as well as a set of timing parameters.

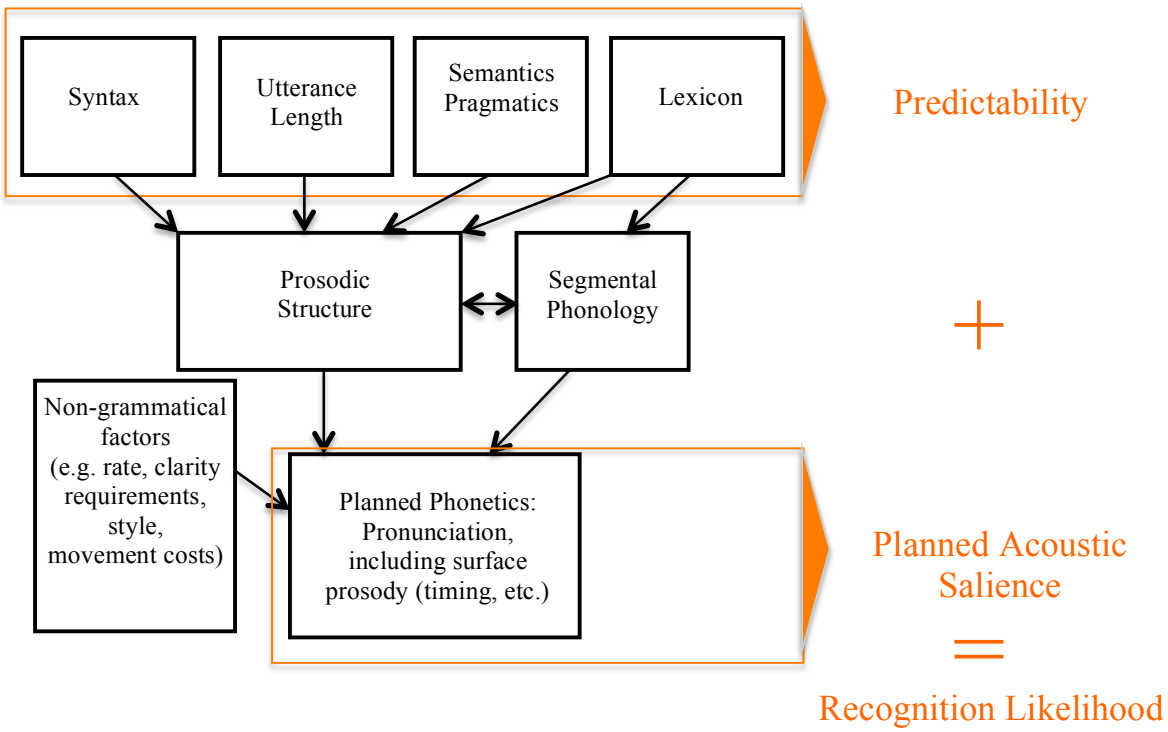

Figure 1: *Factors that shape planned surface phonetics and their relationship **to** predictability, acoustic salience, and recognition likelihood. Based on similar figures in [23,24]*

The timing parameters include:

1. Interval durations of various types, e.g. phrase-final rhyme durations for final lengthening, phrase-initial segment durations for initial lengthening, stressed syllable (or CV) durations for prominence-related lengthening; see [28] for more detail.
2. The timing of acoustic landmarks/constriction targets relative to the preceding one in a sequence.
3. The timing of movement onsets relative to the landmarks/targets
4. The time course of movement (tau, time-to-target achievement at the current movement rate, as a function of time [13]). When combined with spatial information, the tau function determines the velocity profile of movement. Following Lee [13], we assume that tau follows an *intrinsic tau guide*, represented by an equation that describes a family of finite movements (movements from rest that start with an acceleratory component and end after a finite duration): $tauG = k[½ (t-T^2/t)]$. The parameters of the equation are T, the duration of movement, and k, which describes the shape of the movement. The variable t is the elapsed time from the start of the movement. Tau-guided movements will have a single-peaked velocity profile if $k < 1$.

5.1. Determining parameter values

It is well-known that surface phonetic characteristics, including timing, vary systematically with prosodic and segmental context, as well as non-grammatical factors such as clarity requirements, rate, and style. Movement timing also varies with movement distance and accuracy requirements (Fitts' law [29]), where longer distance movements take longer in spite of increased movement speeds, and movements with higher spatial accuracy requirements take longer than movements with lower spatial accuracy requirements. All of these factors need to be taken into account in computing timing values. We assume that phonetic characteristics are also constrained by processing demands and movement costs, such as energy, time, and the cost of inaccuracy (see Figure 1). Following Optimal Control Theory [30-32], we propose that movement parameters are determined that represent the optimal balance between prioritized (or weighted) task requirements and movement costs (see [33] for an OCT interpretation of Fitts' law phenomena, and [34] for an example of the use of OCT in a model of speech production).

We acknowledge that computing all of the parameter values for movement is non-trivial, one reason being that parameter values are inter-dependent. For example, the timing between targets in a sequence will depend on timing requirements for supra-segmental intervals such as phrase-final syllable rhymes), and, as we mentioned earlier, movement timing parameter values depend on spatial parameters such as movement distance and accuracy. In the examples which follow, we illustrate the factors involved in determining the values for three of the types of timing parameter specifications defined above.

Example 1: For the timing of at least some intervals that are the sites of durational effects of prosodic context, we hypothesize that requirements for these intervals will have an influence on the timing between movement targets. For example, the timing between targets within the word-final

syllable rhymes will be longer if they are phrase-final than if they are phrase-medial. Because the phrase-final durational requirement is balanced against the cost of time, we expect phrase-final lengthening to be minimized where it can be. Evidence consistent with this view can be found in [37], who observed that it is not the case that e.g. every phrase-final segment has the same duration, rather, all segments of a particular type are longer in phrase-final position than medially, but the amount of absolute lengthening is segment-specific. Nevertheless, the relative durational rank ordering among segments is preserved.

Example 2: For the timing of an acoustic landmark/movement target with respect to a previous landmark, we assume that there is a cost for time that penalizes time between speech landmarks. We hypothesize that the time between targets will additionally depend on prosodic context and other factors, such as rate and style of speech. For example, if the speech rate is slow, the duration between targets will be longer than if the overall speech rate is fast. In cases where the two targets (X,Y) in a sequence involve the same articulators, the duration between targets will additionally depend on the distance between them, and on the target's spatial accuracy requirements, as well as on the energy cost for reaching the second target.

Example 3: For the timing of movement onset with respect to movement target achievement, we assume that costs for time and energy will constrain overall movement time, and that movement time will increase with the spatial accuracy requirement of the target, and will decrease with its timing accuracy requirement, because faster movements are more accurate in terms of their timing [35]. We hypothesize that the relative weighting of these two requirements might vary with speaker and style. Other factors may also affect movement time, such as prosodic position, where some syllable-final movements may be longer than syllable-initial movements (e.g. velum lowering for nasal stops, [36]).

5.2. Coordination for synchronized targets

Lee [13] presents a way of planning movement coordination within an extrinsic timing framework (General Tau theory in this case). On this theory, movements are coordinated through tau coupling, whereby movements whose tau functions are in constant proportion will end at the same time. Coordination can be achieved by coupling one or more movements onto the internal tau guide (mentioned above), or by coupling a movement onto a sensed movement tau. As explained in [40], when two movement tau functions are in constant proportion, e.g. tauA = ktauB, tauA reaches zero as the target is reached, and because tauB is in constant proportion to tauA, it reaches zero at the same time. On this theory, movement coordination involves movement *offset* coordination. It does not require the time course of the movements to be the same, nor is there a strict requirement for the movements to begin at the same time. What this means is that two coordinated movements might have velocity peaks that don't occur at the same time, but as long as their taus are in constant proportion, they will reach their targets at the same time. In addition, if one of two coordinated movements starts later than the other, it is assumed that the later onset movement is accelerated until $tau_{later} = ktau_{earlier}$, and then that the relation is maintained so that the two movements end at the same time [41].

6. Phonetic implementation

Phonetic implementation involves producing movement kinematics to meet planned phonetic goals. Following the VITE (Velocities-into-terminal-endpoints) model of [38], cited in early versions of DIVA [27], we assume that speakers constantly monitor positions relative to a) the planned target (either actual or predicted, depending on the type of predicted and/or sensory information available), as well as b) the time until planned target achievement at each time point (the tau function, [13]). This information is combined to generate appropriate movement velocities to get the articulators to the target on time. Based on evidence in e.g. [39], and following proposals in Optimal Feedback Control Theory [31-32], we further assume that movement corrections and updates can be made during a movement on the basis of sensory feedback and predicted states, and that corrections will be more likely for prioritized parts of movement, e.g. movement offsets, compared to other parts of movement.

7. Coarticulation in sequential movements

One of the key contributions of AP/TD is its account of coarticulation. In our model, coarticulation falls out of the relative timing of movement targets, and of the movement times required to produce the targets on time with required spatial accuracy.

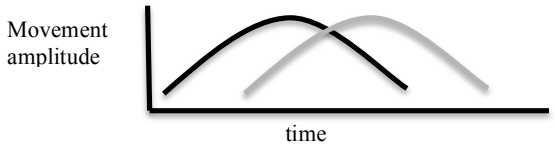

Figure 2. Schematic diagram of coarticulation, where the target of movement A (black) is timed to occur before the target of movement B (grey)

If movement targets are timed closely together, and are produced with different sets of articulators, then for a sequence of two targets AB, the movement onset of B will begin before the movement target A is reached. This is illustrated in Figure 2.

8. Conclusion

The main advantage of our proposal is that it is likely to provide a better fit to existing data in the literature than the AP/TD model, currently the best-worked-out model of speech production. In our experience, careful consideration of a well-motivated alternative to a dominant model can often result in improvements in both competing models. The major drawback to our proposal is that it is still only a model *sketch.* We have not implemented it, and as we note above, implementation will be non-trivial. Attempts to implement it will no doubt bring many deficiencies and oversights to light. However, we hope it will provide a framework for asking fruitful questions about how to model timing in speech production, and for interpreting timing data.

9. References

[1] C. P. Browman and L. Goldstein, "Dynamic modeling of phonetic structure," in *Phonetic Linguistics*, V. A. Fromkin, Ed., ed New York: Academic Press, 1985, pp. 35-53.

[2] E. L. Saltzman and K. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological Psychology,* vol. 1, pp. 333-382, 1989.

[3] R. B. Ivry and R. E. Hazeltine, "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,* vol. 21, pp. 3-18, Feb 1995.

[4] D. Byrd and E. Saltzman, "Intragestural dynamics of multiple prosodic boundaries," *Journal of Phonetics,* vol. 26, pp. 173-199, Apr 1998.

[5] M. Treisman, "Temporal Discrimination and the Indifference Interval - Implications for a Model of the Internal Clock," *Psychological Monographs,* vol. 77, pp. 1-31, 1963.

[6] J. Gibbon, "Scalar Expectancy Theory and Weber's law in animal timing," *Psychological Review,* vol. 84, pp. 279-325, 1977.

[7] D. Byrd and E. Saltzman, "The elastic phrase: modeling the dynamics of boundary-adjacent lengthening," *Journal of Phonetics,* vol. 31, pp. 149-180, Apr 2003.

[8] E. Saltzman, H. Nam, J. Krivokapic, and L. Goldstein, "A task-dynamic toolkit for modeling the effects of prosodic structure on articulation," in *Speech Prosody 2008*, Campinas, Brazil., 2008.

[9] S. Nakai, A. Turk, K. Suomi, S. Granlund, R. Ylitalo, and S. Kunnari, "Quantity constraints on the temporal implementation of phrasal prosody in Northern Finnish," *Journal of Phonetics,* vol. 40, pp. 796–807, 2012.

[10] J. Berry, "Speaking rate effects on normal aspects of articulation: Outcomes and issues," *Perspectives on Speech Science and Orofacial Disorders,* vol. 21, pp. 15-26, 2011.

[11] J. Edwards, M. E. Beckman, and J. Fletcher, "The articulatory kinematics of final lengthening," *Journal of the Acoustical Society of America,* vol. 89, pp. 369-382, 1991.

[12] I. Hertrich and H. Ackermann, "Articulatory control of phonological vowel length contrasts: Kinematic analysis of labial gestures," *Journal of the Acoustical Society of America,* vol. 102, pp. 523-536, 1997.

[13] D. N. Lee, "Guiding movement by coupling taus," *Ecological Psychology,* vol. 10, pp. 221-250, 1998.

[14] L. H. Shaffer, "Rhythm and timing in skill," *Psychological Review,* vol. 89, pp. 109-122, 1982.

[15] R. M. C. Spencer and H. N. Zelaznik, "Weber (slope) analyses of timing variability in tapping and drawing tasks," *Journal of Motor Behavior,* vol. 35, pp. 371-381, Dec 2003.

[16] M. Billon, A. Semjen, and G. E. Stelmach, "The timing effects of accent production in periodic finger-tapping sequences," *Journal of Motor Behavior,* vol. 28, pp. 198-210, Sep 1996.

[17] R. Bootsma and P. C. van Wieringen, "Timing an attacking forehand drive in table tennis," *Journal of Experimental Psychology: Human Perception and Performance,* vol. 16, pp. 21-29, 1990.

[18] C. Craig, G. J. Pepping, and M. Grealy, "Intercepting beats in predesignated target zones," *Experimental Brain Research,* vol. 165, pp. 490-504, Sep 2005.

[19] J. S. Perkell and M. L. Matthies, "Temporal measures of anticipatory labial coarticulation for the vowel /u/ - within-subject and cross-subject variability," *Journal of the Acoustical Society of America,* vol. 91, pp. 2911-2925, May 1992.

[20] Liu and E. Todorov, "Evidence for the flexible sensorimotor strategies predicted by Optimal Feedback Control," *The Journal of Neuroscience,* vol. 27, pp. 9354 –9368, 2007.

[21] A. Georgopoulos, "Cognitive motor control: spatial and temporal aspects," *Current Opinion in Neurobiology,* vol. 12, pp. 678–683, 2002.

[22] P. Keating and S. Shattuck-Hufnagel, "A prosodic view of word form encoding for speech production," *UCLA Working Papers in Phonetics,* vol. 101, pp. 112-156, 2002.

[23] M. Aylett and A. Turk, "The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration on spontaneous speech," *Language and Speech,* vol. 47, pp. 31-56, 2004.

[24] A. Turk, "Does prosodic constituency signal relative predictability? A Smooth Signal Redundancy hypothesis," *Journal of Laboratory Phonology,* vol. 1, pp. 227-262, 2010.

[25] J. Caspers, Pitch movements under time pressure: Effects of speech rate on the melodic marking of accents and boundaries in Dutch. The Hague: Holland Academic Graphics 1994.

[26] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *Journal of the Acoustical Society of America,* vol. 111, 2002.

[27] F. H. Guenther, "Speech Sound Acquisition, Coarticulation, and Rate Effects in a Neural-Network Model of Speech Production," *Psychological Review,* vol. 102, pp. 594-621, Jul 1995.

[28] A. Turk, "The temporal implementation of prosodic structure," in *The Oxford Handbook of Laboratory Phonology*, A. C. Cohn, C. Fougeron, and M. K. Huffman, Eds., ed: Oxford University Press, 2012, pp. 242-253.

[29] P. M. Fitts, "The information capacity of the human motor system in controlling the amplitude of movement," *Journal of Experimental Psychology,* vol. 47, pp. 381-391, 1954.

[30] R. E. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton University Press, 1957.

[31] E. Todorov and M. I. Jordan, "Optimal feedback control as a theory of motor coordination," *Nature Neuroscience,* vol. 5, pp. 1226-1235, Nov 2002.

[32] R. Shadmehr and S. Mussa-Ivaldi, *Biological learning and control: How the brain builds representations, predicts events, and makes decisions*. Cambridge, MA: The MIT Press, 2012.

[33] C. M. Harris and D. M. Wolpert, "Signal-dependent noise determines motor planning," *Nature,* vol. 394, pp. 780-784, Aug 20 1998.

[34] J. Simko and F. Cummins, "Sequencing and optimization within an embodied Task Dynamic model," *Cognitive Science,* vol. 35, pp. 527–562, 2011.

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

[36] R. A. Krakow, "Physiological organization of syllables: a review," *Journal of Phonetics,* vol. 27, pp. 23-54, 1999.

[37] J. van Santen and C. L. Shih, "Suprasegmental and segmental timing models in Mandarin Chinese and American English," *Journal of the Acoustical Society of America,* vol. 107, pp. 1012-1026, 2000.

[38] D. Bullock and S. Grossberg, "Neural Dynamics of Planned Arm Movements - Emergent Invariants and Speed Accuracy Properties during Trajectory Formation," *Psychological Review,* vol. 95, pp. 49-90, Jan 1988.

[39] D. Pélisson, C. Prablanc, M. A. Goodale, and M. Jeannerod, "Visual control of reaching movements without vision of the limb II. Evidence of fast unconscious processes correcting the trajectory of the hand to the final position of a double-step stimulus," *Experimental Brain Research,* vol. 62, pp. 303-311, 1986.

[40] D. N. Lee, A. P. Georgopoulos, M. J. O. Clark, C. M. Craig, and N. L. Port, "Guiding contact by coupling the taus of gaps," *Experimental Brain Research,* vol. 139, pp. 151–159, 2001.

[41] D. N. Lee, personal communication.