音声研究　第13巻第３号
2009（平成21）年12月
●－●頁

Journal of the Phonetic Society
of Japan, Vol.13 No.3
December 2009, pp.00–00

Notes on Phase and Coordination, and Their Application to Rhythm and Timing in Speech

Fred Cummins and Juraj Simko*

音声のリズムとタイミングにおける位相と協調運動の役割について

フレッド・カミンズ・●●●

要旨：音声をはじめとする複雑な運動行動を理解するためには，その行動を構成する個々の運動の相対的なタイミングを把握する必要がある。そのような複雑な協調運動（coordination）の記述には，力学系（dynamical systems theory）が提供する概念や道具立てが適している。中でも「位相」（phase）の概念は，事象の相対的タイミングを表すのに中心的な役割を果たす。本稿では，位相が協調運動を捉えるのに有用であることを，音節や句の反復発話，手の運動，複数の話者が発話をシンクロナイズさせる同期発話，調音音声合成など，様々な具体例を用いて例証する。さらに，周期的・等時的な運動だけでなく，非周期的・非等時的な協調運動についても位相の概念を適用できることを示す。これにより，一話者内の運動の相対的タイミングだけでなく，複数の話者による運動の相対的タイミングも，協調運動に関する同一の枠組みで捉えることが可能となる。

Key words: Speech rhythm, Dynamic Systems, Phase, Timing, Coordination

1. Preliminaries

Timing is everything. Phoneticians care deeply about timing. In looking at speech sounds, or at articulatory movements, the raw data of the phonetician exhibits an intricate patterning in time, infused with the notion of precise change from one moment to the other. In producing speech, the movements we use are extraordinarily precise in their arrangement in both space and time. If the tongue misses its mark by a few millimeters, or a movement misses its goal by a few tens of milliseconds, speech can change into non-speech.

But linguists do not always share this concern with the continuous flow of time. For many approaches to language, simple serial order is seen as critical, but temporal details beyond that are considered irrelevant (Port et al. 1995). This seems to work fine for syntactic theory. It may even work for some approaches to phonology, in which the systematic patterning of symbolic elements is a consideration. But phonology must, ultimately, find expression in movement and usually in sound, and at that point, timing is of the essence, and must be given special consideration.

Dynamical Systems Theory provides a language for understanding change in time, and so it is a natural idiom for modeling speech patterning in time. In this chapter, I will present some examples of how some simple concepts from dynamical systems theory have proven useful in modeling temporal structure in speech, with a special focus on rhythm. I will not provide great detail about individual models and frameworks; for that, the reader should follow the citations herein to the primary sources. Instead, I hope that this overview will suffice to demonstrate just how useful the language of dynamics is for understanding patterns that live in time. A good generic introduction to dynamical systems theory is provided in Norton (1995).

2. Clocks and Time

When we use a clock, our time measurement is expressed with respect to several recurring periods, such as the hour, the minute and the second. These provide an independent scale against which instants and durations can be set. This is *clock time*, and it is essential if we are to relate moments and time intervals of independent processes. It is not the only way in which we can measure time.

Most complex events that merit our attention consist of sub-parts, and the event is recognizable as such only if the constituent parts work together to form an integrated whole. When that happens, we say that the con-

* University College Dublin, Ireland（アイルランド国立大学ダブリン校）

stituent events are *coordinated* with respect to each other. When we wish to speak of the temporal coordination of two events that are not independent, instead of using clock time, we may choose to use one event as a temporal reference for the other. For example, if a baseball is caught in mid-flight by a fielder, then we have two events (the motion of the ball, the motion of the fielder) that are not independent. The correct time for the fielder's hand to be in a certain position is most succinctly expressed, not in clock time, but by referring to the trajectory of the ball. Whenever the ball is in position to be caught is the right time for the hand to also be there. The essence of coordination lies in non-independence of events, and the essence of coordinative timing lies in using one event or process as a referent for the other.

The simplest case of coordinative timing arises when one of the coordinated events is, itself, periodic. In this case, it can serve much like a clock, and we have a straight-forward time scale which we call phase. There are several conventions for indexing phase. Hours, minutes and seconds can be interpreted as phase readings from the periodic motion of the hands of a clock. More usually, a point within a repeating cycle is identified using angular measurement (radians, degrees) or simple proportion. In what follows, we will adopt the latter convention, so that phase is a cyclic variable running from 0 to 1, and then repeating. The zero point at the start of one cycle is identical to the maximal phase value, 1, at the end of the previous cycle.

Phase measurement is useful because it allows us to express invariant relationships among coordinated events despite variation in the absolute rate at which these events occur. Every human walker with legs of equal length starts one stride midway through, or at phase 0.5 of, the other stride. Each leg cycle can act as a temporal referent for the other here. By using phase, we capture an invariant structural property of walking, irrespective of the fact that different people walk at different speeds. Phase is thus our principal basis for describing the temporal structure of coordinated events or processes. It is the natural idiom of the study of coordination.

3. Phase in Speech Timing

The only clearly periodic process routinely involved in speech production is the vibration of the vocal folds that gives rise to the F0 signal. Although not perfectly periodic, the laryngeal signal exhibits a relatively constant repeating structure, as seen in electroglottograph signals. The vibratory mode is sufficiently regular that phase is a natural way to express the relative alignment of the parts of the signal produced during phonation. This is no different from the kind of analysis brought to bear, for example, on ECG signals from the heart. Because phonation is essentially periodic, this use of phase is both familiar and thoroughly unsurprising.

But speech is interestingly structured in time in many different ways. The movements of the supra-glottal articulators are clearly highly coordinated, both with each other, and with the glottis itself. Beyond individual articulatory gestures, we find evidence of important temporal coordination within and among syllables, feet and phrases. It would be of great use to phoneticians if the temporal coordination that so infuses speech could be described using phase. There are many ways in which this can be done. In what follows, we shall look at phase as it relates to timing at a variety of levels, keeping a keen eye on the resulting consequences for our understanding of coordination and structure in speech.

One way of analyzing temporal structure using phase is to take a well circumscribed event and repeat it at a variety of rates. Repetition induces periodicity, which in turn allows the use of phase as a descriptor. This way of analyzing structure in speech has a long history, extending back at least to R. H. Stetson's "Motor Phonetics" of 1928. In that early work. Stetson had subjects repeat groups of one, two, three and four syllables at a range of rates, as he recorded airflow. From the recorded trace, he sought features that were invariant with respect to rate, and those that varied. One of his better known findings is the observation that syllables that are clearly distinct at slow rates, such as /at/ and /ta/ become perceptually identical at fast rate. Stetson believed this to arise from a conflation of the two forms into one canonical /ta/ form. This phenomenon was later studied by Tuller and Kelso (Tuller and Kelso 1990) who explicitly considered the relative phase of laryngeal and oral events in repeated series of the syllables /ip/ and /pi/. Because the entire syllable is repeated, one could take either the stream of glottal events, or the parallel stream of articulatory movement, as a referent. In this case, they selected the reference cycle to be the interval between successive lip aperture minima, and they analyzed the point within that cycle at which the peak glottal opening occurred. This phase variable successfully distinguished /ip/ from /pi/ at slow rates, and could be used to track the transition from /ip/ to /pi/ at fast rates. Within the coordination dynamics[1)] approach to movement (Kelso 1995), the phase measurements here are values of a collective variable that serves to index the relative stability of an entire coordinative structure. The

movement being described (a syllable) consists of several coordinated parts (glottal and supra-glottal gestures), which collectively make up a *coordinative structure*. Because phase here describes the relative timing of several gestures, it is called a *collective variable*.

Similar analysis by de Jong (2001) suggested that the apparent collapsing of two patterns into one at fast rates was not quite as clear as the perceptual reports would indicate, as some phase differences between CV and VC tokens remained even at the fastest rates. There are methodological differences between all these studies that prohibit any simplistic conclusions from these findings, but together they illustrate the relatively simple use of repetition in order to establish a periodic reference cycle, and hence allow phase measurement in the characterization of coordinated movement.

The utility of phase in identifying coordinative stability across rate change is further illustrated in Fig. 1. These data come from one subject in a small study in which speakers repeated a short, irregular phrase at a wide variety of rates (Cummins 2009a). Measurements were taken of the time of the vowel onset of the stressed syllables. From these, a variety of phase based measurements were possible, two of which are illustrated. In the top panel, the relative time of occurrence of the onset of 'Bom' within the interval demarcated by the onsets of the two instances of 'Bay' is shown, that is, phase is the ratio $\phi=a/b$. It is clear that this phase relation remains relatively invariant as rate changes substantially (rate is the reciprocal of the interval from the first to the last onset measurement in each phrasal token). The Pearson's r^2 coefficient of 0.05 illustrates that very little of the variability in measured phase is attributable to rate. Contrast this with the lower panel, based on exactly the same data, but with a phase variable defined as the relative time of occurrence of 'Bay' within the interval demarcated by the onsets of 'Down' and 'Bombay'. Here there is a very strong linear relation between rate and phase, whereby the onset occurs earlier within the interval as the subject speeds up. The Pearson's r^2 of 0.63 testifies to a very robust linear dependence of phase on rate. Empirical evidence such as this can be brought to bear in claims about constituent structure, as there is a *prima facie* case to be made that 'Bay of Bombay' exhibits a stability, and hence coordinative structure, not present in 'down in the Bay

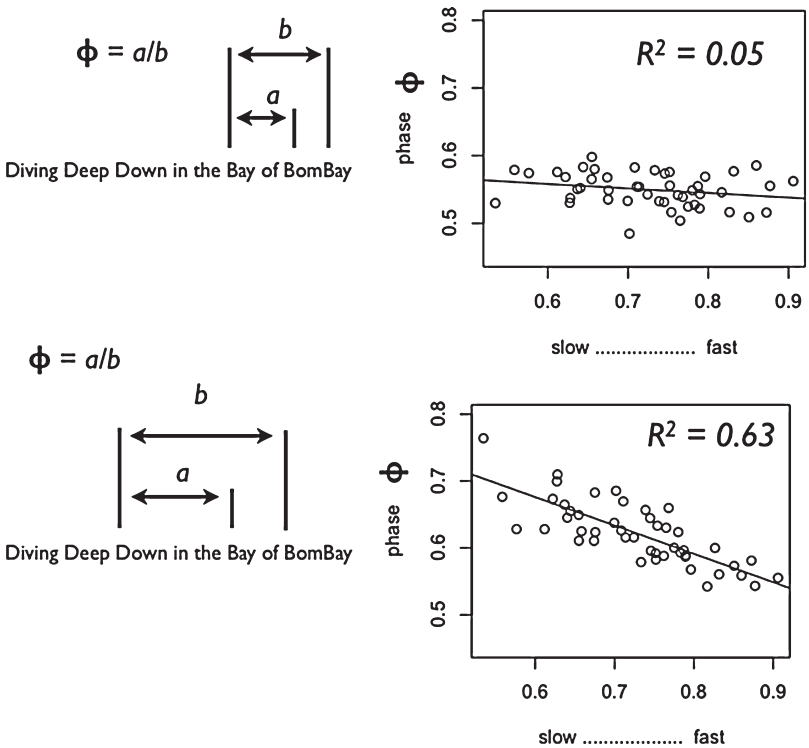

Fig. 1 Two different phase variables taken from the same subject repeating a short phrase. The x-axis is rate, with fast utterances on the right.

of Bom'. Care is required here though, as data from other subject show that phase stability varies considerably between subjects.

The use of phase-based variables in the characterization of temporal structure for repeated units of action is a feature of the venerable concept of a Generalized Motor Program (Gentner 1987). Within the GMP literature, invariant proportional timing (or phase) is taken as indicative of an underlying rate-independent representation of action. A more differentiated elaboration of the concept is found within coordination dynamics, where stability is understood to be an emergent property of coordinative structures that arise under specific boundary conditions (Kelso 1995), i.e. the global form of a pattern that is indexed by a collective variable is relatively resistant to change as long as certain constraints that dictate the nature of the task are held constant (e.g. the task instructions, the conditions under which the task is performed, etc.). Phase, within this approach, provides a window into pattern stability. Phase relations that are stable across changes in the boundary conditions that define a pattern serve to characterize coordinative structures. If the boundary conditions are changed beyond the limits of phase stability, as for example when a task is performed at a very fast or slow rate, we expect phase variability to increase, and ultimately to manifest qualitative change, as one pattern gives way to another (Kelso 1997).

The GMP notion of action representation places emphasis on the internal organization of the speech production system, strictly within the nervous system and bio-musculature of the speaker. It treats observed behaviour as the product of an internal system, divorced from its environment. This contrasts with the modeling approach within coordination dynamics, where relevant boundary conditions that constrain behavioural patterns may be identified within the organism, within the environment, and within the behavioural context of speech production. The resulting pattern is thus seen, not as the fully specified output of a complex, autonomous system, but rather as the emergent result of the interplay of many factors, only some of which are to be found within the speaker proper. The distinction is thus between a classical representationalist cognitive approach to skilled action (GMP) and an embodied and extended approach, in which behavioural patterns are not divorcable from the context and environment in which they occur (Clark 2008, Port and van Gelder 1995, Kelso 1995).

Speech is a rather idiosyncratic form of action. When we lift, kick, pull, throw or sit on something, our movement is intimately linked to the physical properties of the surrounding environment, and the resulting form of movement is very clearly constrained by the properties of both the movement system and the object of that movement. But when we speak, the articulators remain hidden, shielded from the world. The speech production system is capable of responding gracefully to physical interference from a pipe, bite block, or loose tooth, but by and large, the movement is not tightly *coupled* to the inertial properties of the surrounding environment of the speaker. 'Coupling' here means the mutual interaction and reciprocal influence of two systems, such as an organism and its immediate environment. This is not to say that the physical form of speech is divorced from context, however. There are many situations in which the temporal structure of speech produced is manifestly influenced by an external process or referent.

Consider first a simple call and response, as at prayer or roll call. Here there is a weak temporal association between the production of the caller and the responder. Such temporal links are typically weak, however. For example, there has been very little evidence to show that turn taking in conversation is very precisely timed, beyond the rather obvious constraint of serial ordering (Bull 1997). If a group response is required, as in a church setting or political rally, there will necessarily be a need for respondents to temporally align the start of their reply. Sometimes, groups speak together over extended periods. This is known as choral speaking, and is found in prayer, in the recitation of oaths, in group chants, etc. Typically the text repeated in these situations is heavily practiced and formulaic, and as a result, the prosody of the resulting speech is highly stylized and idiosyncratic. Ritual repetition of the rosary (a prayer sequence repeated many times in certain Roman Catholic rituals) provides a vivid example of this kind of idiosyncratic prosodic shaping.

But the constraints imposed by turn-taking or choral speaking are relatively weak compared to those imposed by speech set to music. The fact that speech can be produced in time with a musical beat illustrates the accommodative nature of speech timing. Anyone familiar with more elaborate forms of vocal music (scat singing, hip-hop, etc) will know that speech can accommodate a musical timing constraint in rich and varied ways. To be perceived as speech, certain temporal relations must be preserved, or else speech will not be accurately perceived. But to align appropriately with music, there must also be a great deal of flexibility and plasticity in the speech production system. This kind of context-driven temporal sculpting is difficult to account for in a

cognitivist, representationalist account, but fits nicely within a dynamical account of speech, in which the music or an underlying beat can serve as an additional constraint, contributing to the final pattern observed. Once an external signal has a marked influence on speech timing, dynamical models of coupling and entrainment provide a natural form of expressing the resulting blend of influences upon observed patterning in time. The terms 'coupling' and 'entrainment' are used more or less synonymously here to describe the mutual interaction of two systems, such that the patterning in time of each of them is manifestly and lawfully influenced by the patterning in time of the other. The relationship may be primarily one way, as when a boat is rocked by the ocean, and the reciprocal effect of the boat on the ocean is negligible, or it may be more nearly equal, as when two legs are entrained to each other in walking. In the latter case, a perturbation of one leg, will lead to a rapid response from both legs to preserve overall pattern stability.

4. Entrainment

Perhaps the best known model of coupling in overt human behavior is the Haken-Kelso-Bunz model of bimanual finger wagging (Haken et al. 1985). This model has been fully described in many other places (see e.g. Kelso 1995), so that a brief summary must suffice here. When subjects are instructed to wag two fingers periodically and at the same rate, there are two and only two stable ways to do this. In one pattern, each finger progresses through its cycle in lockstep with the other (the in-phase pattern), while in the other, each finger is half a cycle out of step with its partner (the anti-phase pattern). This model system comprises two coordinated hands, but the state of the system can be neatly captured with a single scalar variable, the phase difference, or relative phase, between the two hands (0 for in-phase, 0.5 or anti-phase).

This phase-based index serves to identify the two coexisting possible forms of coordination at most rates, that then become conflated into a single possible mode at fast rates. The transition from a bi-stable regime at intermediate rates, to a mono-stable regime at fast rates, has been intensively studied. In this instance, it is clear that the observation of a relatively fixed phase relation at most rates is not part of an invariant motoric representation stored centrally. Instead, phase plays the role of a collective variable that indexes the overall coordinative state of the two-hand system. It is emergent, rather than specified.

The modeling approach taken by Haken et al. (1985) is instructive. Firstly, the relative phase variable that indexes the entire coordinative pattern can be captured using a potential function, as illustrated in the right panel of Fig. 2. This function has two minima for most values of the parameter a/b, and a single deep minimum for fast rates. Importantly, the system can also be described as a composite system of two entrained subsystems: the two fingers or hands. For this, each finger is modelled as a self-sustaining oscillator, the exact form of which is selected to approximate the known dependence of amplitude on frequency. The two oscillators are linked by symmetrical and explicit coupling functions (Fig. 2, left, bottom). The original presentation of the model then showed how the simpler description of the composite system could be analytically derived from the more complex description of two combined systems (Haken et al. 1985). In this instance, the influence of one hand on the other is entirely matched by the reciprocal influence of the other hand on the first.

The finger wagging paradigm in the study of manual coordination inspired the development of the speech cycling paradigm, in which repetition is again used to study coordination (Cummins and Port 1998, Tajima 1998, Jankowski 2001). In its simplest form, a metronome provides a periodic signal, to which a speaker entrains, by attempting to align the onset of a small repeated phrase with the metronome tone (Tajima 1998). In the target speech cycling variation, two distinct alternating tones are provided, one of which is to be aligned with the phrase onset, and the other with the onset of a strong stressed syllable within the phrase (Cummins and Port 1998). By varying the phase of the overall repetition cycle at which the target for the medial stress occurs, it is possible to probe the ability of speakers to produce phrasal coordination with a variety of phase relations (Fig. 3). The striking experimental result is that speakers can only reliably produce a very restricted set of coordinative patterns, characterized by phases at 1/3, 1/2 and 2/3 of the overall repetition cycle (Fig. 4). Once more, phase serves to characterize the discrete behavioural patterns produced by subjects under sufficiently constrained experimental conditions. The identification of stable forms of coordination, characterized by stable and discrete relative phase values, allows us to see commonalities between the constraints and form of manual movement and of phrase repetition.

In the speech cycling example, we see the entrainment of speech-producing movement to an invariant, strictly periodic signal. The utility of phase measure-

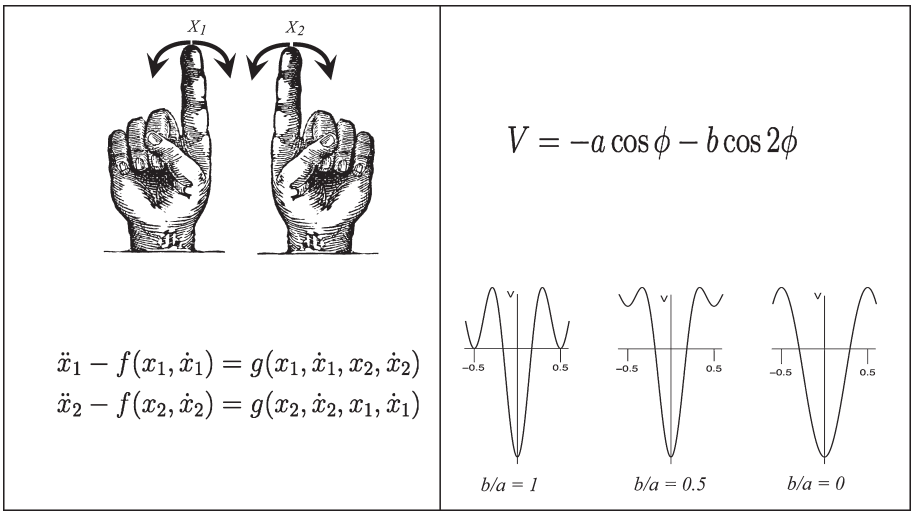

Fig. 2 Left: The behavioural task is to wag two fingers at the same frequency. Each finger is modeled as a self-sustaining oscillator, and a non-linear coupling term links the two systems. The state of each oscillator is a single number (x_1, x_2), and their time derivatives are given by $\dot{x}_1$, $\ddot{x}_1$ etc. Right: The same system can be described using only relative phase, ϕ. A potential function captures the distribution of observed relative phases. This function, V, defines a landscape over the system states (x-axis, ϕ), illustrating the relative stability of each phase. Different parameterizations of the potential function are obtained by varying the ration a/b.

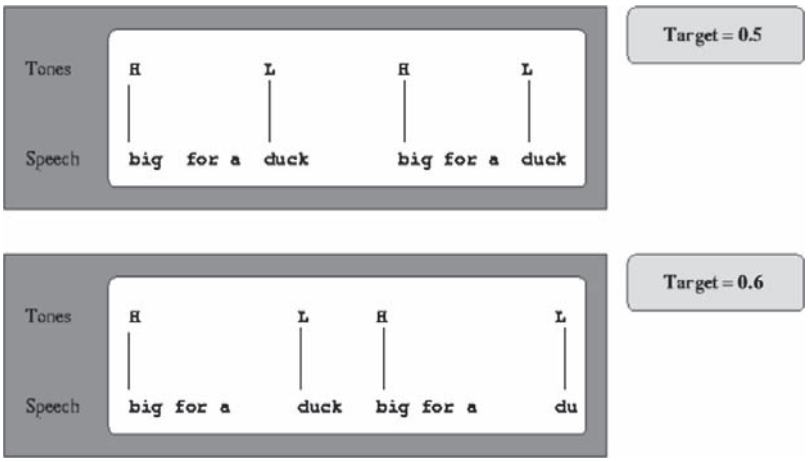

Fig. 3 In a Target Speech Cycling task, alternating high and low tones provide temporal targets for tone onsets for the onset of the phrase and a medial stressed syllable, respectively. The target phase is the point within the high-to-high cycle at which the low tone occurs.

ment in characterizing coordinative patterns arises through the use of repetition. Entrainment is found to other stimuli as well. We have already mentioned the entrainment of speech by music, and the mutual entrainment of a group of speakers in a choral speaking situation. An experimental analogue of the choral speaking situation has been developed, called Synchronous Speech (Cummins 2003, Kim and Nam 2008). Unlike conventional choral speaking, a synchronous speech task employs two people only, and makes us of a novel text, thereby avoiding the prosodic stylization so characteristic of oaths and prayers.

In a Synchronous Speech task, a pair of subjects first read a novel text to familiarize themselves with it. On a

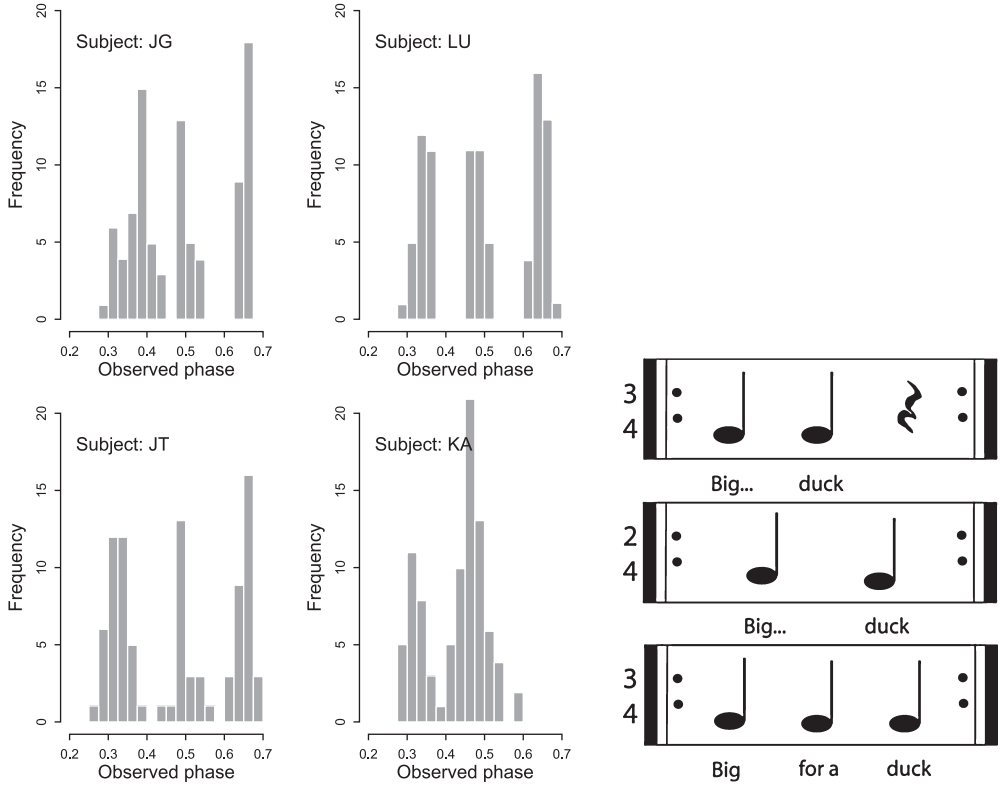

Fig. 4 Left: Histograms of observed phases in an experiment in which targets were drawn uniformly from the range [0.3, 0.7]. Right: Musical interpretation of the three dominant patterns observed.

starting signal, they then attempt to read the text in synchrony with one another. Remarkably, this turns out to be relatively straightforward for most subjects. Despite the known variability of speech timing found in general, subjects typically manage to coordinate their joint production such that lags observed between corresponding events in the parallel speech waveforms are of the order of 40 ms throughout a phrase, with a slight increase to about 60 ms at phrase onsets after a pause (Cummins 2003) (Fig. 5). Subjects typically do not need much practice at this task, and indeed practice does not, in general, improve their performance significantly (Cummins 2003). In listening to two speakers speaking in synchrony, it is hard not to think of them as locked to one another, each supporting and reinforcing the other. It is virtually never the case that one speaker leads consistently and the other follows. Rather, speakers enter into a protracted period of joint production, in which each is entrained to the other.

The synchronous speech paradigm presents an interesting challenge to the dynamical interpretation of speech production and movement. Although it looks as if two co-speakers are strongly entrained to each other, there is no periodic referent that would allow the simple computation of a phase variable, with which we might index the stability of the joint coordination. It seems incredible that speakers possess internal invariant representations that allow them to produce exquisitely

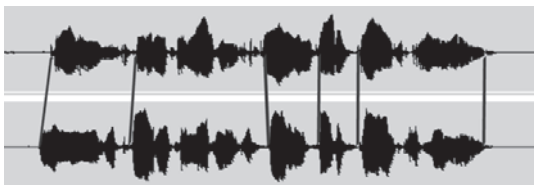

Fig. 5 Two speakers recorded speaking synchronously. A crude estimate of asynchrony can be obtained by comparing corresponding points in the two waveforms. A method for obtaining a more precise estimate of asynchrony is presented in (Cummins 2009c).

timed synchronous productions when speaking together. And yet there is no single temporal referent that can serve as a clock or *zeitgeber* in this instance. Apart from posing obvious methodological challenges for the dynamically-minded phonetics researcher, the larger question arises of just how it is that two speakers can maintain such precise synchrony, essentially without practice, in such a complex task, without a periodic signal with which they might calibrate their own production. We will return to the interpretation of apparently coupled, but aperiodic systems, in the next section.

5. Beyond Periodicity

Phase measurements based on a periodic reference signal are unproblematic. Even if, for example, the frequency of the reference slowly changes, phase is still meaningful and quantifiable. It allows us to locate one point in time with respect to the cycle of the referent. However, coordination does not demand periodicity. The example of the baseball fielder catching a ball illustrates this nicely. Both events, the ball flight and the hand movement of the catcher, are clearly coordinated, and timing in one is only comprehensible if it is understood in terms of the temporal unfolding of the other. But there is no periodic referent that would allow the definition of a phase variable.

An analogous situation arises when we try to describe the coordination of one discrete movement with another, as for example when the lowering of the velum is precisely timed with respect to lingual and glottal processes in the production of nasalized consonants and vowels. In general, it is clear that articulator timing demands a high degree of coordination among the many parts of the speech production apparatus, and that this must be achieved, even though at the level of articulatory gestures, there is no strict periodicity in speech.

This conundrum has been long recognized within the general area of the study of coordinated movement, and hearkens back to an old debate about so-called intrinsic versus extrinsic timing (Keller 1990, McAuley and Riess Jones 2003). If two processes are to be coordinated in time, then either one must act as the temporal referent for the other (possibly in symmetrical fashion), or they must both have access to some other source of timing information, often conceived of as an external clock. The former case is called intrinsic timing, the latter extrinsic. There are many models that explicitly assume external timekeepers (Wing and Kristofferson 1973, Howell 2001), but there is, as yet, no evidence for a dedicated clock in the brain (Ivry and Richardson 2002).

The issue of how best to describe the coordination of two overlapping articulatory gestures has arisen with a vengeance within the task dynamic implementation of articulatory phonology. Articulatory phonology (AP) is an extremely influential theory that provides a principled account for how units of action, gestures, can simultaneously serve as units of linguistic contrast. The theory has been influential in accounting for a variety of coarticulatory phenomena, and processes such as epenthesis, lenition, deletion, etc. (Browman and Goldstein 1990). Introductions may be found in (Browman and Goldstein 1992) and (Browman and Goldstein 1995). Fig. 6 shows a gestural score, which is a formal representation used to generate articulatory sequences within AP. The rectangular boxes pick out periods in which individual gestures are presumed active. This is thus akin to an underlying control structure for the given utterance.

Within the standard implementation of AP, this score (the boxes) is used to generate smooth movement of a set of model articulators, such as a jaw, tongue body, etc. This is done using the Task Dynamic model, first introduced in the context of limb movement, which provides a means for generating smooth movement in a system with many degrees of freedom, as it works towards the achievement of a sequence of behavioral goals (Saltzman and Kelso 1987, Saltzman and Munhall 1989). A full description would go beyond the bounds of the present article. In brief, a behavioral goal, e.g. a lip closure, is represented within the model as an abstract *task*, with an associated tract variable dynamical system which ensures smooth movement towards an abstract target. These tasks are mapped into a space

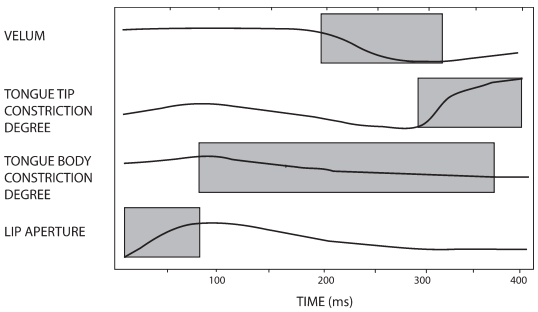

Fig. 6 Partial gestural score for the word /pan/. The score itself consists of the rectangular boxes which mark those periods of time during which a gesure is active and hence contributing directly to articulator movement. The solid lines illustrate the resulting motion of the associated vocal tract variables.

of model articulators, where they vie for control of multiple articulators at any given time. Their respective influences are combined, so that the resulting sequence of articulator movements is smooth, biologically plausible, and capable of driving, in turn, an articulatory synthesizer.

A major issue facing this approach has been the relative timing of the gestures: that is, how should gestures be timed with respect to each other. As with the baseball example, there is a need for tight temporal coordination, but no periodic referent. There are two distinct aspects to this problem: first we need a means to describe the relative coordination of two discrete, non-repeating gestures. Then we need a means for deriving appropriate relative timing relations among gestures, as described in the gestural score.

The standard approach to the first problem has been to make use of the fact that each gesture is represented within the model by an abstract dynamical system; specifically, each has the form of a simple mass-spring system[2], where the mass is initially displaced from its resting equilibrium, so that it moves smoothly towards the rest position. The dynamical system is set up so that smooth movement, without undershoot or oscillation results. This is done by incorporating critical damping into the equation for the mass spring system. However, one can also ignore the critical damping, in which case the solution to the mass-spring equation would be an undamped sinusoidal oscillation. This undamped cycle can play the role of a measurement scale, allowing us to describe events such as the on-and off-sets of other gestures, using a phase measurement based on this cycle. It must be stressed that this is a methodological convenience, allowing unambiguous measurement in an appropriate time scale. The undamped dynamical system does not generate movement within the model.

The second problem is more difficult, and a variety of approaches have been tried out. It has been suggested that two co-occurring gestures might exhibit invariant phase relations, but this turns out to be a poor way to describe empirically observed phase relations in real speech (Barry 1983, Kelso et al. 1986, Nittrouer et al. 1988). A more flexible approach suggested that the relative timing of two gestures might be multiply determined by factors such as speaking rate, focus, context, stress, etc, and that one might use these factors to derive phase windows that would predict phase relations probabilistically (Byrd 1996). This is a powerful approach, but somewhat unconstrained, and so not very predictive of actual phase values. Recent approaches have sought to constrain inter-gestural timing using a variety of oscillators (Saltzman et al. 2008).

A recent variant of the task dynamic model has suggested a novel way in which the sequencing problem among gestures might be approached (Simko and Cummins 2009b, Simko and Cummins 2009a, Simko 2009). The original task dynamic model is modified so that the dynamical systems associated with tasks are no longer abstract and context-free, but embodied in physical articulators and thus mutually coupled. This embodiment makes it possible to quantify the amount of effort required to make a specific gesture, or its articulatory cost. The articulatory cost, in turn is offset by the cost of being imprecise: a gesture that is not sufficiently effortful will result in imprecise or unintelligible speech. These two costs are combined with a third based on simple duration, so that any given sequence of gestures can be assigned a composite cost. It is then possible to search the space of possible relative timings (and system stiffnesses), to produce a sequence of gestures that minimizes this cost. The process is illustrated in Fig. 7, and an example of the smooth articulator trajectories that result is shown in Fig. 8.

The use of phase to describe the relative coordination of gestures in an utterance is possible only because the underlying dynamical model makes use of components, i.e. gestures, that are modeled *both* as single-shot movement patterns that terminate smoothly at a specific goal point, *and* as underlying cyclic systems. This is potentially useful for the further development of the theory of task dynamics and for articulatory phonology, as it may allow the identification of relatively invariant forms of coordination, or the quantitative characterization of coordinative relations that vary systematically with factors such as stress and rate. But the underlying cycles are purely abstract theoretical entities. They do service within the theory, but these cycles do not have clear correspondences in the physically instantiated realm of nervous system and muscle activity. In that more concrete realm in which we observe movement,

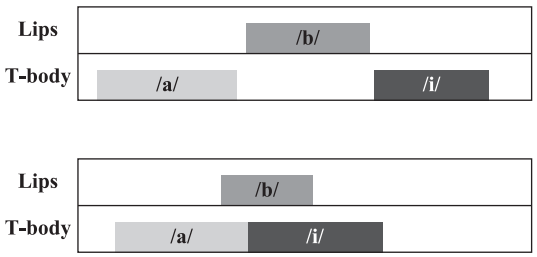

Fig. 7 Gestural score, before and after optimization. T-body = tongue body.

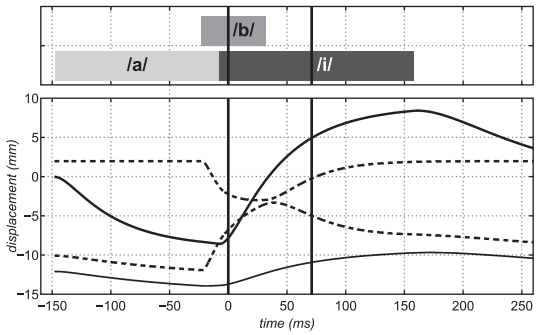

Fig. 8 Articulatory traces for an optimal utterance. Bold solid line = tongue body; dashed line = lips; thin solid line = jaw. The vertical lines are placed at the moment of complete consonant closure and release. Lip movement continues after closure, due to soft tissue compression.

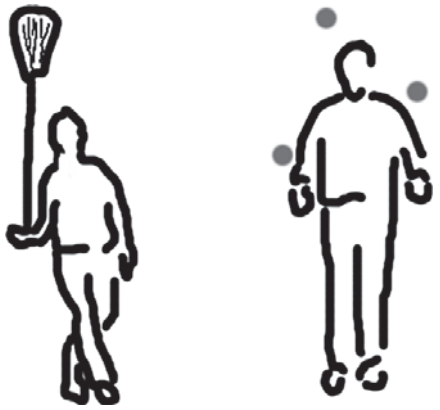

Fig. 9 Balancing and Juggling, illustrating the relationship between coupling degree and regularity in timing.

there is no convincing periodicity, and cyclic movement is the exception rather than the rule. Nonetheless we recognize that inter-gestural coordination is essential to the felicitous production of speech patterns. We are, as yet, missing a theoretical bridge that allows coordination and coupling to be expressed for both periodic and aperiodic systems.

One way to consider both periodic and aperiodic coordination is suggested by Fig. 9. In the right hand panel, a juggler constrains his movement so that it interacts smoothly and fluently with each of a set of balls. Two distinctive features of this situation are important to note. Firstly, the amount of time during which the juggler interacts with each ball is severely limited: there is haptic interaction during the dwell phase of the ball in his hand, and there is a limited amount of visual information, usually restricted to a portion of the ball's flight, near the apex of its trajectory. For most of the ball's trajectory, it is uncoupled with the juggler, flying freely through the air. The second important feature is that the resulting coordination pattern is built on strict periodicity. It seems entirely reasonable to suggest that these two features are non-independent, and that the limited amount of time in which one system (a ball) can interact with the other (the juggler) has as a consequence that stable coordination demands periodic structure. Periodicity provides predictiveness, which offsets the limited amount of mutual interaction between ball and man.

In the left panel, we also see two mutually coupled systems. Here, a man balances a broom on his hand. Movement of the broom arises from instability due to its position, and continuous compensation is required by the hand of the balancer. In this instance we have tight coupling of the two systems, but no periodicity in the final movement at all. A clock is here not necessary, as the two systems are in constant continuous contact, allowing a much tighter degree of coupling among the systems.

In speech production, events are being coordinated at a variety of levels. Gestures are coordinated with one another, producing syllables. Syllables are coordinated to produce larger prosodic units. Nowhere is there strict periodicity in the result. However, all these multiple levels of coordination are instantiated within the same physical system. There is thus a continuous opportunity for mutual entrainment of the processes responsible for timing the constituents. The fact that all constituents, from individual gestures to whole phrases, are produced within a single embodied system suggests that coordination patterns will arise in a manner more like the balancing of a broom than the juggling of a regular pattern.

If correct, this way of understanding coordination suggests something rather important about the synchronous speech situation as well. We noted that subjects seem to be able to entrain their speech with remarkable ease, without extensive practice, and to remain tightly time locked to a co-speaker for extended periods. All of this is achieved without any periodic scaffolding. The above consideration suggest that the two speakers should be regarded as being in continuous 'contact' with one another, and that this contact is mediated, not by touch, but by the acoustic pattern that is the publicly observable acoustic manifestation of speech. Having simultaneous access to their own acoustic production and the production of another person, seems to serve to

entrain the two production systems. This surely suggests that the production and perception of speech must be very deeply integrated with one another, as has been suggested on independent grounds (Liberman and Mattingly 1985, Liberman 1993).

6. End Notes

In this essay, I have attempted to bring together a number of problem domains within the broad field of coordination and speech, that share the characteristic of being usefully described using phase. I have sought to make the notion of phase somewhat more intuitive than a merely mathematical measurement of a periodic function. Phase, as presented here, is time made meaningful. It is the way in which we describe the co-dependence in time of two events that are non-independent, and it provides a window into the world of coordination and entrainment. Elsewhere, I have tried to construct a similar argument that much of what we discuss under the heading of 'rhythm' is best understood as entrainment among systems: entrainment of dancers to a musical beat, entrainment of musicians in an orchestra, entrainment of feet, hands and heads to a passionate speaker (Cummins 2009b). Rhythm is about linking and coupling systems so that they become coordinated.

If this approach is correct, then the distinctive characteristic of two temporally coordinated systems is not isochrony, regularity or periodicity. These are all attributes of a signal that may facilitate coordination, if the amount of coupling between the systems is limited. As that degree of coupling increases, for example by increasing the time in which the two signals are in extended physical contact (mediated through touch, or even sound), the need to rely on periodicity decreases, although the systems remain tightly linked. This may allow us to understand the coupling of gestures within an individual in the same way as we understand the coupling of movements across individuals.

From a practical point of view, an immediate consequence of this approach is that the description and modeling of complex movement is best done by studying how individual components of movement are related in time to one another, rather than to a single background time scale. This means placing a particular emphasis on those properties of the movement system that serve to constrain the task and to bring about the relative coupling of components. By the same token, it downplays the role of a central executive or controller, that might be supposed to directly specify or control the duration of the components of movement.

Acknowledgements

Much of this work was funded through a Science Foundation Ireland Principal Investigator grant, 04/IN3/I568, to the author. Thanks are due to Keiichi Tajima and Bob Port who co-developed the speech cycling paradigm.

Notes

1) Coordination dynamics refers to an approach to the study of movement most associated with the work of Scott Kelso. Within this approach, a complex movement pattern, such as the wagging of two limbs, is viewed both as the coordinated movement of several components (e.g. the two limbs), and as a whole in itself. One research task is to find the *collective variable* that indexes the state of the whole pattern, and relate it to the higher dimensional set of indices that capture the state of the individual components. In this example, each limb has an instantaneous phase, and the collective variable is the difference between the two phases. A good introduction to the field is Kelso (1995).
2) Norton (1995) provides a good concise introduction to basic dynamical modeling, including such standard systems as the mass-spring system. In any mass-spring system, the state of the system is given by the distance of the mass from its resting, or equilibrium, position. The movement towards the resting state is determined by two parameters: stiffness and damping, which influence movement speed, and deceleration, respectively.

References

Barry, W.J. (1983) "Some problems of interarticulator phasing as an index of temporal regularity in speech," *Journal of Experimental Psychology*: *Human Perception and Performance*, 9: 5, 826–828.

Browman, C. and Goldstein, L. (1990) "Tiers in articulatory phonology, with some implications for casual speech." In Kingston, J. and Beckman, M.E. (eds.) *Between the Grammar and Physics of Speech*: *Papers in Laboratory Phonology I*, chapter 19, pp. 341–376. CUP, Cambridge.

Browman, C.P. and Goldstein, L. (1992) "Articulatory phonology: An overview," *Phonetica* 49, 155–180.

Browman, C.P. and Goldstein, L. (1995) "Dynamics and articulatory phonology." In Port, R.F. and van Gelder, T. (eds.) *Mind as Motion*, chapter 7, pp. 175–193. MIT Press, Cambridge, MA.

Bull, M.C. (1997) *The Timing And Coordination of Turn-Taking*. PhD thesis, University of Edinburgh.

Byrd, D. (1996) "A phase window framework for articulatory timing," *Phonology*, 13: 139–169.

Clark, A. (2008) *Supersizing the mind: embodiment, action, and cognitive extension*. Oxford University Press, USA.

Cummins, F. (2003) "Practice and performance in speech produced synchronously," *Journal of Phonetics* 31: 2, 139–148.

Cummins, F. (2009a) "Phase and coordination in speech production." In Colye, L., Dunnion, J. and Freyne, J. (eds) *Proceedings of the 20th Irish Conference on Artificial Intelligence and Cognitive Science*, pp. 13–22.

Cummins, F. (2009b) "Rhythm as an affordance for the entrainment of movement," *Phonetica*, 66: 1–2, 15–28.

Cummins, F. (2009c) "Rhythm as entrainment: The case of synchronous speech," *Journal of Phonetics* 37: 1, 16–28.

Cummins, F. and Port, R.F. (1998) "Rhythmic constraints on stress timing in English," *Journal of Phonetics* 26: 2, 145–171.

de Jong, K. (2001) "Rate-induced resyllabification revisited," *Language and Speech* 44: 2, 197–216.

Gentner, D.R. (1987) "Timing of skilled motor performance: tests of the proportional duration model," *Psychological Review* 94: 2, 255–276.

Haken, H., Kelso, J.A.S. and Bunz, H. (1985) "A theoretical model of phase transitions in human hand movement," *Biological Cybernetics* 51, 347–356.

Howell, P. (in press (2001)) "The EXPLAN theory of fluency control applied to the treatment of stuttering by altered feedback and operant procedures." In Fava, E. (ed.) *Clinical Linguistics*: *Language Pathology*, *Speech Therapy*, *and Linguistic Theory*, Clinical Issues in Linguistic Theory. John Benjamins.

Ivry, R.B. and Richardson, T.C. (2002) "Temporal control and coordination: the multiple timer model," *Brain and Cognition* 48, 117–132.

Jankowski, L. (2001) "Replicating the speech cycling task paradigm with French material." In *Proceedings of the Conference ORAlity and GEstuality*, pp. 610–614, Aix-en-Provence, France.

Keller, E. (1990) "Speech motor timing." In Hardcastle, W.J. and Marchal, A. (eds.) *Speech Production and Speech Modelling*, pp. 343–364. Kluwer Academic, Dordrecht.

Kelso, J.A.S. (1995) *Dynamic Patterns*. MIT Press, Cambridge, MA.

Kelso, J.A.S. (1997) "Relative timing in brain and behavior: Some observations about the generalized motor program and self-organized coordination dynamics," *Human Movement Science* 16, 453–460.

Kelso, J.A.S., Saltzman, E. and Tuller, B. (1986) "The dynamical perspective in speech production: Data and theory," *Journal of Phonetics* 14, 29–60.

Kim, M. and Nam, H. (2008) "Synchronous speech and speech rate," *Journal of the Acoustical Society of America* 123: 5, 3736.

Liberman, A.M. (1993) "In speech perception, time is not what it seems." In Tallal, P., Galaburda, A.M., Llinás, R.R. and von Euler, C. (eds.) *Temporal Information Processing in the Nervous System*: *Special Reference to Dyslexia and Dysphasia*, volume 682. Annals of the New York Academy of Sciences, New York.

Liberman, A.M. and Mattingly, I.G. (1985) "The motor theory of speech perception revised," *Cognition* 21, 1–36.

McAuley, J.D. and Riess Jones, M. (2003) "Modeling effects of rhythmic context on perceived duration: a comparison of interval and entrainment approaches to short-interval timing," *Journal of Experimental Psychology: Human Perception and Performance* 29: 6, 1102–1125.

Nittrouer, S., Munhall, K., Kelso, J.A.S., Tuller, B. and Harris, K. (1988) "Patterns of interarticulator phasing and their relation to linguistic structure," *Journal of the Acoustical Society of America* 84: 5, 1653–1661.

Norton, A. (1995) "Dynamics: an introduction." In Port, R.F. and van Gelder, T. (eds.) *Mind as Motion: Explorations in the Dynamics of Cognition*, chapter 1, pp. 45–68. Bradford Books/MIT Press, Cambridge, MA.

Port, R. and van Gelder, T., (eds.) (1995) *Mind as Motion: Explorations in the Dynamics of Cognition*. Bradford Books/MIT Press, Cambridge, MA.

Port, R.F., Cummins, F. and McAuley, J.D. (1995) "Naive time, temporal patterns and human audition." In Port, R.F. and van Gelder, T. (eds.) *Mind as Motion*, pp. 339–437. MIT Press, Cambridge, MA.

Saltzman, E. and Kelso, J.A.S. (1987) "Skilled actions: A task dynamic approach," *Psychological Review*, 94, 84–106.

Saltzman, E. and Munhall, K. (1989) "A dynamical approach to gestural patterning in speech production," *Ecological Psychology* 1, 333–382.

Saltzman, E., Nam, H., Krivokapic, J. and Goldstein, L. (2008) "A task-dynamic toolkit for modeling the effects of prosodic structure on articulation," In *Proceedings of the speech prosody 2008 conference, Campinas, Brazil*.

Simko, J. (2009) *The Embodied Modelling of Gestural Sequencing in Speech*. PhD thesis, UCD School of Computer Science and Informatics, University College Dublin. Also released as Technical Report UCD-CSI-2009-07 available from http://www.csi.ucd.ie/biblio.

Simko, J. and Cummins, F. (2009a) "Embodied task dynamics," *Psychological Review*. Submitted.

Simko, J. and Cummins, F. (2009b) "Sequencing of articulatory gestures using cost optimization." In *Proceedings of INTERSPEECH 2009*, Brighton, U.K.

Stetson, R.H. (1951) *Motor Phonetics*. North-Holland, Amsterdam, 2nd (1st ed. 1928) edition.

Tajima, K. (1998) *Speech Rhythm in English and Japanese: Experiments in Speech Cycling*. PhD thesis, Indiana University, Bloomington, IN.

Tuller, B. and Kelso, J.A.S. (1990) "Phase transitions in speech production and their perceptual consequences." In Jeannerod, M. (ed.) *Attention and Performance XIII*, chapter 14, pp.429–452. Lawrence Erlbaum Associates, Hillsdale, NJ.

Wing, A.M. and Kristofferson, A.B. (1973) "Response delays and the timing of discrete motor responses," *Perception and Psychophysics* 14: 1, 5–12.

(Received Nov. 7, 2009, Accepted Feb. 26, 2010)