Phase and Coordination in Speech Production

Fred Cummins

School of Computer Science and Informatics
University College Dublin
fred.cummins@ucd.ie
http://pworldrworld.com/fred

Abstract. When trying to understand behavioral systems, the measurement of time as phase offers many advantages over conventional clock time. We illustrate this with some experimental results in speech production, in which stable coordinations are evident using phase measurements. These stable coordinations may be related to the abstract constituents posited by linguists, but they are manifest only in the performance of an embodied system. Tying time measurement to the physical system also reveals a large role for individual difference in coordinative structures in speech.

Key words: speech production, phase, dynamical systems, embodiment

1 Introduction

Language is conventionally thought of as an abstract domain, relatively independent of its physical manifestation in speech, signing, writing, morse code or semaphore. Linguistic structure is thus typically thought of as removed from, and independent of, physical implementation. However, all language is produced using one physical system or another, and there are many reasons to privilege speech in this regard. Speech is the most common mode of language production. For the vast majority of language users, it is the first form of language experienced and learned. It also long predates all written or coded forms, and may be presumed to be the "natural condition" of a language producing subject[1].

While sequential order is of obvious centrality to all aspects of language, from sound sequencing, through morphology, syntax and semantics, the temporal dimension in speech is characterized by rich patterns of coordination among the speech articulators [14, 16]. The coordination of physically embodied articulators places constraints on speech production that serve to both delimit and define the space of possible speech events that can be reliably produced. In what follows, it will be argued that the embodied nature of speech production may allow the identification of hierarchical units in speech. These units are characterized by relatively stable temporal organization when time is measured as phase, rather than as clock time. They arise from well-learned coordination patterns, and

[1] The relations between sign, gesture and speech are the subject of much speculation, but little is known about their relative importance in the origins of language.

exhibit considerable variation across individuals. This empirical approach to the identification of units in language poses challenges and opportunities for theoretical linguistic accounts.

2 Measuring Time as Phase

The "blooming, buzzing, confusion" in which we are immersed is made interpretable in part because we parse the continuous flux into discrete events [13]. Two observable changes in the world may recur together, or at a relatively fixed offset, thus providing evidence that they ought to be considered as components of some larger whole. The identification of a fixed timing relation between two events depends in the first instance on their relation to one another, rather than their separate relations to a fixed temporal scale of reference (clock time). To give a concrete example, the right time for a goalkeeper to have his hands in a particular spot is not to be found on a clock, but is to be identified by using the trajectory of the ball as a referent. The timing of the goalkeeper's movements are intimately connected to those of the kicker and the ball, and are essentially and causally unrelated to the movements of uninvolved players, the referee, spectators, and passing birds. Thus, we will recognize the kick and subsequent save as an event, and distinguish it from a background of simultaneous but unrelated flux.

This insight underlies a long-standing discussion that pits embodied, dynamical models of motor coordination against other forms of computational modeling, such as the identification of putative motor programs. In its clearest form, the debate has been pitched as one of intrinsic versus extrinsic timing [14]. Extrinsic timing refers to models of temporal unfolding in which events are pegged with respect to a clock of some kind. Many models of temporal interval interval production, for example, assume an underlying clock that provides a stable sequence of periods, thus providing other processes with a temporal reference [19, 12]. Intrinsic timing, on the other hand. deals with the relative timing of events, where the events themselves serve as reciprocal temporal referents, as in the above example, where the trajectory of the ball provides the appropriate referent for timing the movement of the goalkeeper's hands. Intrinsic timing models thus need a way of expressing when one event component happens, in units that are provided by another event component. Phase measurement is one way this can be accomplished. If the event that is to serve as a referent has a fixed period, other events can be expressed as proportions of that period, thus providing a natural way of expressing temporal coordination that captures invariance across changes in absolute duration[2]. Phase is most readily expressed as the proportion of one period of a sinusoidal or periodic process, however the above arguments seek to emphasize that phase is best intuitively understood as relative timing

[2] Various conventions for describing phase exist, including ranges of 0 to 360 degrees, 0 to 2π radians, -π to π radians, or most simply, as proportions from 0 to 1. We will use the latter form here.

expressed in intrinsically meaningful units within a specific context. Phase is time made meaningful.

3 Phase Stability as the Hallmark of Meaningful Coordination

When you walk, one foot hits the ground half way through the cycle of the other foot. That is, there is a constant phase relation of 0.5 between the two legs. In fact, all gaits of all animals are characterized by constant phase relations (not necessarily 0.5) among the limbs [10], and different phase relations are the signature of different gaits. Phase relations remain invariant across rate changes within a single gait. This constancy of phase is a clear indicator that the limbs are meaningfully coordinated, one with the other. This can be contrasted with the temporal relationship obtaining between the elements involved in a sequence such as the making of a cup of tea. If we take the sequence to include boiling a fixed amount of water, infusing the tea, and the subsequent drinking, this sequence can also be done at a variety of tempi. However, not all parts of the sequence can be compressed with equal facility. Infusing may be shorter, drinking may be hurried, but boiling a fixed quantity of water will stubbornly resist temporal compression. In this case, if we define an overall cycle that lasts from filling the kettle and ends with finishing the cup of tea, then the phase at which drinking starts, for example, will change as the sequence is executed at different rates. There is a notable absence of temporal coordination between the diverse subparts to this action sequence.

Speech is complex sequential action, and it is an open question how that sequencing is achieved. Most linguistic descriptions emphasize serial order. Thus the sequence /pot/ contrasts systematically with the sequence /top/, even though the set of constituents are, at some abstract level of description, the same. Much of the structure of language as conventionally understood lies in the sequencing of elements, and in the grouping of sub-sequences into larger units within an ordered hierarchy. At the level of meaning, the smallest units, morphemes are conventionally assumed to group into larger units, words, which in turn partake in elaborate structural hierarchies such as phrases and sentences.

At the level of sounds, many accounts of speech structure posit atomic units at the level of the phone, with phone sequences organized within containing syllables. Above the syllable, theories of prosodic phonology typically posit several hierarchical layers that help to account for a wide variety of surface features of speech such as lengthening effects at the right edges of supposed constituents, or the blocking of processes such as vowel harmony by constituent boundaries. There has been a marked lack of agreement about the number and nature of levels required to accurately describe the prosodic structure of speech, and there are no effective procedures for the unambiguous identification of many proposed constituents [2, 8].

Linguistic theory typically regards the elements that are sequenced as disembodied symbol-like entities, and one of the principal distinctions between

phonology and phonetics is that the former is concerned with systematicity in the distribution of symbolic elements (phonemes, syllables, intonational phrases) while the latter is seen as the implementation thereof. The interface between the two thus becomes a challenge for a full account. Some have suggested that the implementation of such sequences within the constraints of the physical vocal tract makes the task of the recovery of the underlying symbol sequences difficult [11] while others have questioned the logic that deduces the presence of this presumed interface [15]. One influential theory that aspires to providing a full account of both linguistic and phonetic phenomena is Articulatory Phonology, in which the units of contrast are also simultaneously units of action, or gestures [3]. Within this approach, the question of how to appropriately coordinate the timing of gestures with respect to one another, or how to phase them, has long been a contentious issue [4]. It has recently been suggested that a suitably embodied instantiation of the theory may allow the discovery of physically optimal coordinative relations among gestures, but this work is still at an early stage [17]. Unfortunately, neither conventional articulatory phonology, nor the recent embodied task dynamic extension, can deal appropriately with the coordination of units much larger than individual gestures within syllables.

In the spirit of this embodied approach to the elements of phonology and phonetics, it is possible to ask if meaningful coordination of larger units than the syllable might be revealed by an appropriate experimental methodology that looked for evidence of phase stability across tempo variation. Rather than positing abstract underlying symbolic units that arise from the grammar of a language, and then seeking to uncover these units in the noisy signal that is the physical speech signal, one could adopt an alternative stance that starts with the physical signal, and looks for invariance across tempo change to uncover meaningful units of coordination. This experimental approach diverges from conventional strategies in several fundamental ways. Firstly, when we look at coordinated movement, we find that individuals differ, and these differences matter. We can readily identify an individual by their handwriting, their prosody, or their gait, because each individual has achieved a behavioral goal in an idiosyncratic manner. Skilled movement is the imposition of constraints upon a very high dimensional system, such that behavioral goals are fulfilled. This produces underspecified solutions, with the result that movement patterns are idiosyncratic [18]. We might not find a single grammar of movement for speakers of a language, but we might find individual structures that are demonstrably meaningful constituents in the speech of an individual. Secondly, coordinative units might be a function, not only of speaker, but also of speaking condition. Where linguistic theory tends to posit invariant underlying structures, a performative, embodied approach might uncover stable units of coordination in some speaking conditions that are simply not present in others, evan as word sequence is held constant.

One example of the identification of large units of prosodic structure that are specific to a speaking condition was provided by a series of speech cycling experiments [7]. In the canonical speech cycling experiment, speakers repeated a short phrase, such as “big for a duck” in time with a repeating sequence of

alternating high and low tones. They were instructed to attempt to align the onset of the phrase with the high tones, and the onset of the final stressed syllable ("duck") with the low tones. The experimental variable was the phase of the low tones within the repeating cycle of high tones. In one experiment, target phases drawn from a uniform distribution ranging from 0.3 to 0.7 were employed. On each trial, one phase was drawn randomly from this range, and subjects attempted to match it. The distribution of produced phases (i.e. the relative timing of the onset of "duck" with respect to the overall phrase repetition cycle) is illustrated in the left panel of Figure 1.

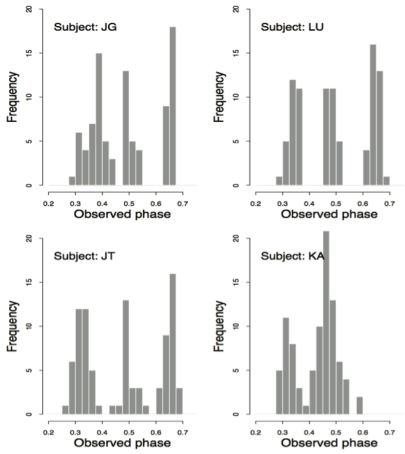

Fig. 1. Left: Distribution of phases of a medial stressed syllable onset within the overall phrase repetition cycle. Right: Schematic representation of the nesting of feet within the phrase cycle corresponding to the three phases that subject reliably produced.

It is immediately apparent that some phases are produced with greater frequency than others. In fact, three and only three phases are produced reliably, and each of these corresponds to the integral nesting of one unit within the overall phrase repetition cycle. The unit that is so nested is produced with a stable temporal relationship or phasing with respect to the containing cycle. This unit is, in fact, well known within phonology and corresponds to the stress foot as defined by Abercrombie [1] (the interval from the onset of one stressed syllable to the next). Within the strict constraints of the speech cycling task, the relative phasing of stress foot onsets within a containing repetition cycle is stable, and points to a meaningful unit of coordination.

In what follows, some new data from a repetition task are presented. The experimental goals are exploratory: we seek to ask whether units of coordination in speech production might be identified by phase stability. In contrast to the speech cycling experiment just described, we here vary articulation rate, and examine the relative constancy of selected phase relations across a range of

tempi. Many speech experiments make use of qualitative differences between normal speech and fast speech. In the present experiment, we treat articulation rate as a continuous variable and go to some lengths to ensure that data are obtained at a wide range of rates for each subject. Furthermore, we wish to inquire to what degree any stable coordinative patterns observed are specific to an individual speaker, or to a speech elicitation context, and to ask whether speakers are capable of varying phase as context varies.

4 Methods

Four subjects took part, two males and two females, all from the Eastern part of Ireland. Each read a short narrative text containing a target phrase. They were then instructed to repeat the target phrase again and again, and to vary their rate of speech as indicated by the experimenter's hand level. As they repeated the phrase, the experimenter raised or lowered his hand every four or five repetitions, encouraging the subjects to explore their range of potential articulatory rate variation. Every effort was made to ensure that the repetitions obtained spanned the range from the fastest to the slowest that the subjects could reliably achieve. Articulation rate was then indexed as the reciprocal of the interval duration from the first to the last stressed syllable onset. A similar procedure was then employed to obtain speech at a variety of amplitudes, but those data will not be reported here. The entire process was repeated for a second set phrase taken from a second text.

Each subject completed a second session which was structured as above, but this time all readings and repetition were done in synchrony with a matched subject (male with male, female with female) [6]. The synchronous repetition condition imposes strong temporal constraints upon subjects, and we wished to see to what degree any stable properties of phase variables were specific not only to an individual, but also to the conditions under which speech was elicited.

The two phrases employed were *Diving Deep Down in the Bay of BomBay* and *Big Dinosaurs and Bigger Daleks in Battle*. These were designed so as to provide a series of strong stresses that are separated by varying numbers of unstressed syllables. Vowel onsets for the capitalized syllables were measured by hand, and a variety of phase variables were explored by examining the variation in the proportion of one large interval occupied by some smaller interval, across a wide variety of articulation rates. For example, one could look at the phase of the onset of *Deep* within the containing interval delimited by the onsets of *Diving* and *Down*.

5 Results

Figure 2 shows the observed phase of *Deep* as defined above, collapsed across all rates for the two female speakers. In the central panel, it can be seen that the two speakers produce qualitatively different timing patterns for this small sub-phrase when they speak on their own. All phases observed were stable across

rate variation, as evidenced by the low R^2 values arising from correlation of the observed phase with tempo of articulation. The R^2 values obtained were 0.03 (F1, solo), 0.00 (F2, solo), 0.18 (F1, synchronous) and 0.14 (F2, synchronous). Although both are native speakers of closely matched dialects, their coordinative patterns in repeated speech belie highly individual solutions to the behavioral task of producing an acceptable utterance, just as their individual handwritings would also be found to differ. The right panel of Fig. 2 shows that when they are constrained to speak in synchrony, this particular difference can be overcome, as they produce similar coordinations, with a phase value lying intermediate between the two phases in the middle panel.

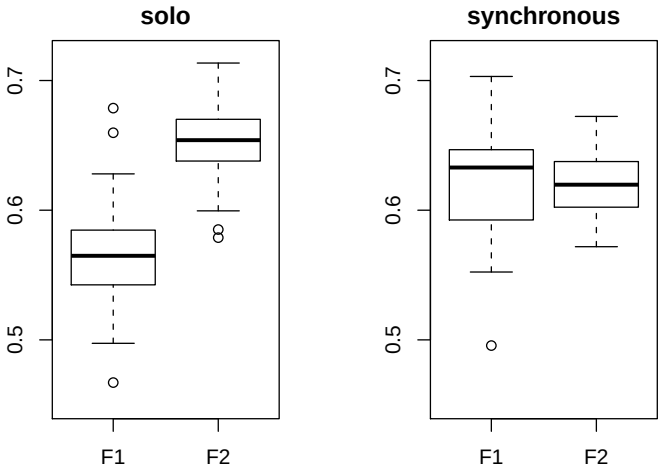

Fig. 2. Phase of the onset of *deep* within the containing interval bounded by the onsets of *diving* and *down* in the phrase *Diving deep down in the bay of bombay.* Data are from two subjects (F1, F2) speaking alone ("solo") or in synchrony with one another ("synchronous")

Phase stability is common, but not ubiquitous. In Fig. 3, two closely related phase variables are plotted as a function of articulation rate for speaker F1 speaking alone. As evident in the top half of the figure, the onset of *Bom* within the sequence *Bay of BomBay* is invariant across rate change. (Rate is indexed by the reciprocal of the period from the first to the last stressed syllable onset, with fast rates on the right of the figure.) This is in stark contrast to the related variable which indexes the relative timing of the onset of *Bay* within the subsequence *Down in the Bay of Bom.* The latter variable has a straightforward linear relationship to articulation rate. As the speaker speaks more rapidly, the unstressed syllables (and perhaps the initial stressed syllable) in the initial stress foot compress to a greater extent than those in the subsequence *Bay of.* These

data strongly suggest that the sequence *Bay of Bombay* is a meaningful, embodied, production unit in the speech of this person under these circumstances, while the syllables in *Down in the* do not form part of any such constituent.

However, these observations are not easy to square with conventional linguistic accounts, as the same phase variables, measured on the other female subject, yield R^2 values of 0.32 and 0.02 where subject F1 had 0.05 and 0.63, respectively. Thus phase stability here reveals a unit of coordination in the speech of an individual that is tied to that person, and quite probably also to the elicitation conditions. It is both embodied and performative. (For comparison, in the synchronous condition, subject F1 had corresponding R^2 values of 0.01 and 0.39, respectively, which are qualitatively similar to those seen in the solo condition, F2 had 0.38 and 0.05, again substantially the same as in the solo condition).

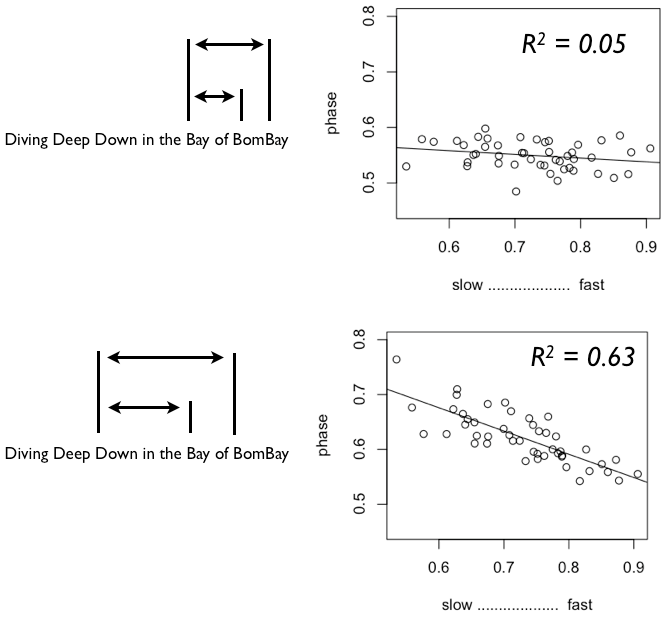

Fig. 3. Two different phase variables taken from the same subject (F1) speaking alone. The x-axis is tempo, with fast utterances on the right.

One further example will serve to further illustrate the character of coordinative structure as evidenced by phase stability. Fig. 4 shows the phase of the onset of *Daleks* in the subsequence *Big Daleks in Battle*, for two male subjects in both solo and synchronous conditions. Although the two subjects had no overt difficulty in synchronizing with one another, their phase data clearly reveal coordinative differences between the two subjects that are invariant across speaking conditions. Where subject M1 exhibits a strong linear relationship between this phase variable and rate, M2 produces almost constant phase values.

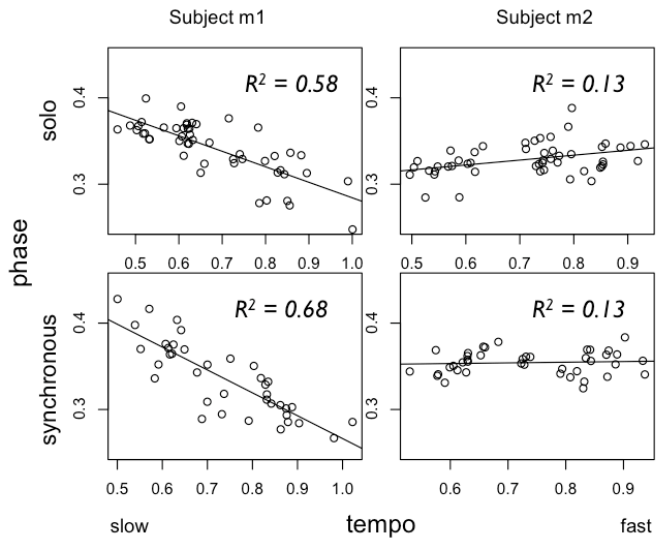

Fig. 4. Phase of the onset of *Daleks* in the phrase *Big Daleks in Battle*.

6 Discussion

The experimental methodology employed here reveals that phase measurements that index the proportional duration of one interval within a larger containing interval may reveal much about coordinative strategies in speech production. Speakers may differ in their coordinative strategies in uttering the same text (Fig 2). This seems to be akin to individual differences found in other forms of motor activity that are highly skilled and that satisfy behavioral goals within a system that can potentially achieve those goals in many ways. There are thus direct parallels to be drawn between coordinative patterns in speech production and individual characteristics of handwriting, gait, etc. Similar phase-based variables have previously been shown to reliably index individual speakers better than speech elicitation circumstances [5]. However in the present case, qualitative difference in phase values were not always invariable. The phase variable shown in Fig. 2 changed for both speakers in the synchronous speaking condition, whereas that observed in Fig. 4 remained invariant across speaking condition, despite behaving quite differently for the two male subjects.

The clear evidence of phase stability across a wide range of articulation rates serves to identify some units (e.g. *Bay of BomBay*) as meaningful wholes that are distinct from those subsequences that display variable phases across tempo change. The units so identified do not stand in any simple correspondence to prosodic units within any conventional linguistic theory. They are functions of embodied speech production in the performance of specific individuals under specific circumstances. It may not be the case that all syllables of any given utterance lie within such units. They thus pose a challenge to conventional ana-

lytical accounts. Some recent developments within phonology, such as optimality theory, have opened up room for consideration of individual phonologies that may differ even among speakers of matched dialects [9]. Perhaps there is room here then to bridge the gap between embodied accounts of behaviour and formal linguistic models.

References

1. David Abercrombie. *Elements of general phonetics.* Aldine Pub. Co., Chicago, IL, 1967.
2. Mary E. Beckman and Janet Pierrehumbert. Positions, probabilities and levels of categorization. In *Proceedings of 8th Austrlian International Conference on Speech Science and Technology*, pages 2–18, 2000.
3. Catherine P. Browman and Louis Goldstein. Articulatory phonology: An overview. *Phonetica*, 49:155–180, 1992.
4. Dani Byrd. A phase window framework for articulatory timing. *Phonology*, 13:139–169, 1996.
5. Fred Cummins. Speech rhythm and rhythmic taxonomy. In *Proceedings of Speech Prosody 2002*, pages 121–126, Aix en Provence, 2002.
6. Fred Cummins. Practice and performance in speech produced synchronously. *Journal of Phonetics*, 31(2):139–148, 2003.
7. Fred Cummins and Robert F. Port. Rhythmic constraints on stress timing in English. *Journal of Phonetics*, 26(2):145–171, 1998.
8. Fernanda Ferreira. Creation of prosody during sentence production. *Psychological Review*, 100(2):233–253, 1993.
9. Adamantios I. Gafos and Stefan Benus. Dynamics of phonological cognition. *Cognitive Science*, 30:905–943, 2006.
10. S. Grillner. Control of locomotion in bipeds, tetrapods, and fish. In V. B. Brooks, editor, *Handbook of Physiology, Motor Control.* Williams and Wilkins, Baltimore, Md, 1981.
11. Charles Hockett. *A Manual of Phonology.* University of Chicago, Chicago, 1955.
12. Richard B. Ivry and Thomas C. Richardson. Temporal control and coordination: the multiple timer model. *Brain and Cognition*, 48:117–132, 2002.
13. W. James. The Principles of Psychology, vols. 1 & 2. *New York: Holt*, 1890.
14. J. A. Scott Kelso, Elliott Saltzman, and Betty Tuller. The dynamical perspective in speech production: Data and theory. *Journal of Phonetics*, 14:29–60, 1986.
15. John J. Ohala. There is no interface between phonology and phonetics: a personal view. *Journal of Phonetics*, 18:153–171, 1990.
16. Robert F. Port, Fred Cummins, and J. Devin McAuley. Naive time, temporal patterns and human audition. In Robert F. Port and Timothy van Gelder, editors, *Mind as Motion*, pages 339–437. MIT Press, Cambridge, MA, 1995.
17. Juraj Simko and Fred Cummins. Sequencing of articulatory gestures using cost optimization. In *Proceedings of INTERSPEECH 2009*, Brighton, U.K., 2009.
18. Esther Thelen and Linda B. Smith, editors. *A Dynamic Systems Approach to the Development of Cognition and Action.* Bradford Books/MIT Press, Cambridge, MA, 1994.
19. Alan M. Wing and A. B. Kristofferson. Response delays and the timing of discrete motor responses. *Perception and Psychophysics*, 14(1):5–12, 1973.