Fred Cummins (Dublin)*

Probing the Dynamics of Speech Production

1 Overview

Two specialized experimental methods are introduced which have proved useful in studying the dynamic properties of speech. They illustrate distinct ways in which we can consider experimental intervention into the speaking process, in order to bring to light characteristics of the dynamical control regime operative in the production of speech. Firstly, Speech Cycling (Section 4) illustrates how the external perturbation of the speech production system by a metronome can reveal innate patterns of rhythmic organization which are otherwise obscured. Then, in Synchronous Speech (Section 5), two speaker speak together, stripping their speech of superfluous para-linguistic sources of variability. The two speakers are mutually entrained, and the resulting speech exhibits considerably reduced inter-speaker variability for macroscopic prosodic parameters such as phrasing and pausing. Synchronous Speech represents a well-defined speaking style amenable to laboratory study. The two novel experimental approaches are argued to have much in common with other dynamical forms of intervention such as perturbation studies, reiterant speech, diadochokinetic tasks and the close shadowing of speech.

2 An apology for contrived experiments

> To state one argument is not necessarily to be deaf to all others, and that a man has written a book of travels in Montenegro, is no reason why he should never have been to Richmond.
>
> Robert Louis Stephenson, *An Apology for Idlers*

* Thanks are due to (Becky) Wang Bei and Elena Zvonik, who both contributed materially to the work described herein.

I feel compelled to begin this chapter with an apology[1] for the use of contrived methods in the elicitation of speech for investigative purposes. It is certainly a welcome fact that spontaneous speech, with all its intransigent messiness and vibrant colour, has become the principle object of study in the empirical analysis of the spoken form of language (Beckman, 1996). This is all the more appropriate as we realize that decades of rigid artificial laboratory tasks have delivered us data which oftentimes do not generalize well. The specific characteristics of speech generated in this or that experimental task frequently fail to reappear within the more pertinent corpus of speech as actually spoken. Similarly, casual everyday speech produces a host of effects which are difficult, if not impossible, to evoke in a laboratory setting.

But I firmly believe that artificial speaking conditions may, when dutifully employed, still furnish us with valuable information about the modus operandi of the speech production system, its capabilities, and especially its limitations. There are several well established elicitation paradigms which, though not without their detractors, have, I believe, sufficiently proved their mettle as to warrant their inclusion in the toolbox of any experimental speech scientist.

One such is the use of reiterant speech (Liberman and Streeter, 1978), in which each syllable in an utterance is replaced by a single form, such as /ma/. Thus, a model phrase such as “The linguist kicked the verb” is produced as “ma MA ma ma ma ma”. The reiterant version is, in a significant sense, simpler than the original, yet it preserves much that might still be of interest, including something akin to the original rhythmic and melodic form. Detractors have rightly objected that segmental detail and suprasegmental properties are not cleanly separable, and that the resulting prosody is not the same as that of the original utterance (Dauer, 1983). This can be admitted, without detracting from the interest which is directed to the reiterant production correctly interpreted. Thus, although we would be unwise to deduce details of the alignment of the nuclear accent with the segmental material from a reiterant production, there are commonalities between the intonation contours of both the reiterant and full versions which make a persuasive case for the existence of an intonation contour which is independent to a large degree of the specific segmental makeup of the relevant syllables. Used with caution, reiterant speech is a valuable tool for exploring speech.

Other avowedly artificial speaking tasks have proven themselves worthy of retention in the armoury. Among these is the shadowing of speech, in which a speaker attempts to speak along with a speech stream heard through headphones (Marslen-Wilson, 1973). The extreme temporal pressure which is thereby placed upon the speaking subject has the potential to reveal much about the real-time processing constraints that operate in both the perception and production of

[1] Apologia: a formal written defence of a cherished proposition.

speech. Perhaps the most valuable contribution of the shadowing technique is to allow us to rule out several *prima faciae* plausible hypotheses about the role of monitoring and feedback during production. The technique initially served to identify a peculiar skill set displayed by a small proportion of speakers: the ability to close-shadow a speaker with latencies as low as about 250 ms. Although few speakers display this ability (in the original study, 7 such speakers were selected from a screened sample of 65 individuals), the very fact that it can be done by some places strong limits on the role which active monitoring and processing can be assumed to play in a full account of speech communication (Bailly, 2001).

Tightly constrained experimental settings are of particular value in a clinical setting. 'Naturalness' is not a necessary quality, if the speech obtained is to be used primarily as an index of pathology. All that is required is that a given task be sufficiently well constrained that it is possible to identify with a high degree of confidence the characteristics of production by non-pathological speakers. For example, the production of syllables in the most rapid fashion possible, or oral diadochokinesia, is a useful performance indicator in assessing normal and pathological language development (Henry, 1990).

To argue for one approach is not to be blind to the advantages of others. For example, Swerts and Collier (1992) introduced a framework within which spontaneous speech could be collected which had a high probability of featuring specific adjective-noun combinations in a specific position. This is a thoroughly welcome approach to the study of spontaneous speech. But my purpose here is to point out some of the virtues which can accrue from the study of highly constrained, unnatural speech, as the speech production system is coerced into performing at or near its absolute limits. To justify this approach, it is useful to consider the rather abstract issue of how to obtain information about a complex dynamical system when ignorant about its inner organization and natural form.

3 On pushing and poking

Call it the Christmas Morning Syndrome. On that morning, the children get up, find presents under the tree, each of which (unless they are mere clothes) needs to be explored. The exploration of any mechanical system will involve pushing, poking, tweaking and otherwise mistreating it, and the result, sadly is inevitably at least one breakage. Despite this inevitable danger, the mode of exploration seems to me to be sound. Even when a system has an obvious and well defined function, it may be necessary to drive the system to some extremes if one wants to become truly acquainted with its capabilities and proclivities. If we fall short of driving the system to the limits of its abilities, we can often learn a lot from administering a well-timed jab or tweak, by shaking, bouncing, hefting and even throwing. This mode of exploration is as well known to speech researchers as it is to children (Sigurd, 1973; Xu, 2002)

There are essentially two ways in which we can administer a nudge or poke to a dynamic system in the service of exploration. In the first approach, we bring some force to bear from outside which influences the system in an asymmetric, one-way fashion. As a canonical example we might consider a parent pushing a child on a swing. The parent administers a well-timed push once per cycle, and the push serves to amplify the natural resonant cycle of the child/swing system. Although there is, strictly speaking, an effect on the parent just as there is on the child, the manifest effect is so asymmetrical that we can take this to be a central example of the unidirectional *forcing* of the child/swing system. The large-amplitude swinging motion which results is a latent behaviour of the system which is parameterized by the mass of the child/swing, the length of the chains, etc., but which only becomes evident when energy is injected into the system at just the right times. The external manipulation thus serves to reveal a natural, though latent, mode of operation of the system.

External forcing may reveal behaviours which are quite different from the unforced behaviour. This was graphically illustrated at 11:00 a.m. on November 7th, 1940, when an external forcing wind injected energy into the naturally swaying Tacoma Narrows Bridge, causing the small amplitude oscillations to be magnified many times over. The resulting violent shaking caused the bridge to be destroyed. The wind revealed a latent mode of behaviour which was ordinarily hidden.

A second manner in which latent behaviours of a complex dynamical system may be revealed is to entrain it to another system. Entrainment is most readily evident among periodic components. Thus when a group of male fiddler crabs wave their enlarged claws to attract the attention of a female, they exert a mutual influence on one another with the result that the group of suitors ultimately wave in synchrony (Backwell et al., 1998). Whether this results merely from watching one another, or from some competitive strategy, is not yet known. Certain firefly species of South East Asia likewise entrain their periodic flashing, so that a whole swarm will flash on and off in synchrony. These two examples of entrainment from the animal world are well known, and both are built from many strictly periodic components (Strogatz and Stewart, 1993). Other, more human, behaviours can illustrate entrainment too, without strict periodicity. Musicians in an ensemble are, of necessity, entrained, one to the other, as are dancers, and even synchronized swimmers. Entrainment is something which can emerge, and dissapear. We have all heard applause change quite naturally from a chaotic haphazard noise to a synchronous clap, sometimes across many hundreds or thousands of individuals.

In what follows, I will present some speech experiments which use both external forcing, and mutual entrainment to reveal some dynamic properties of speech. Speech Cycling is presented first, as a warmer-upper. It uses external forcing to reveal some surprisingly strong constraints on timing across and within phrases. We then present Synchronous Speech, in which one speaker is used to entrain another. In each case, we take our cue from examples such as the above, where a disturbance to the natural course of a dynamic system can be used to reveal latent, or natural, behaviours of the system, along with constraints and limitations on its motions.

4 External forcing: speech cycling

In moving our several limbs, we do not find stable, repeatable patterns which have arbitrary temporal relationships among the limbs. Rather, there are only a few stable cyclical relations. We call each such stable form of organization a gait, and we can observe that the same general principles of limb organization are to be found across all species which display symmetrical limb arrangements along their mid-line (Hildebrand, 1985). Although we bipeds have relatively few gaits (run, walk, sack-race-hop), quadrupeds have more, such as the pace, trot, gallop, canter, and Pepe le Pew's distinctive pronk. An experimental model of this general constraint on limb movement has been developed over many years by Scott Kelso and colleagues (Kelso, 1995). Typically, subjects were asked to wag their fingers in time with a metronome. Two fingers wagging periodically with the same frequency exhibit two and only two stable forms of organization: They can be exactly in phase, or in anti-phase. In the latter case, one finger is half way through its cycle just as the other is setting off. This model system exhibits a rich set of phenomena similar in many respect to limb ensembles. For example, the two stable patterns are not equally stable, and at high frequencies, the anti-phase pattern is seen to switch abruptly to an in-phase one. Just prior to the switch, critical fluctuations (an increase in the cycle-to-cycle variability) appear, only to disappear just after the switch. Many of these properties have been incorporated into a rigorous dynamical model in which each limb or finger is modelled as a self-sustaining oscillator which is coupled to the other (Haken et al., 1985). Extensions of the model have been made to accommodate more than two effectors, learning, discrete movement, etc.

If it were not quite obvious that the two-finger-wagging system has two and only two stable states (this can be readily demonstrated to oneself), it might have been discovered in an experiment done by Yamanishi et al. (1980) and later by Tuller and Kelso (1989). In these, subjects faced a pair of pacing light signals, and they tapped along with the flashing lights as the experimenters manipulated the relative phasing of the two lights. Tapping could be observed both while the pacing signal was flashing, and in a continuation phase thereafter. In this manner, it was possible to explore the relative stability of tapping at arbitrary phase relations. Although these two experiments had other purposes in mind, this procedure can serve as a scanning method to uncover stable repetitive behaviours. The Speech Cycling method, first introduced in Cummins and Port (1998) uses a variant of this set up to uncover stable repetitive behaviours of the speech production system.

In a Speech Cycling task, the subject repeats a short phrase along with a periodic auditory signal, or metronome. In its simplest form, there is a single, repeating tone which cues the phrase onset. Once subjects are familiar with the task, it is easy to manipulate the tempo of the metronome, e.g. by speeding it up, and observing the relative stability of the speech patterns which result. In this way, Tajima et al. (1999) were able to demonstrate clear rhythmic differences between Japanese, on the one hand, and Arabic and English on the other. English and Arabic, while similar, were also differentiated to a lesser degree by the way in which the medial stresses were bound to the overall repeating structure. The onsets of medial stresses in English were more tightly tied to specific phases within the overall repeating cycle than were Arabic stresses.

A somewhat more elaborate form of the phrase repetition task takes its cue from the phase-scanning methods employed in Yamanishi et al (1980) and Tuller and Kelso (1989). In this *targeted speech cycling*, subjects hear an alternating series of high and low tones. Phrase onsets are cued by the high tones, while the low tones specify a target time for the onset of a medial stress. Thus, in repeating the phrase “big for a duck”, the high tone cues “big”, and the low tone cues “duck” (Figure 1). The experimenter is free to specify an arbitrary phase relation for the medial stress onset within the overall Phrase Repetition Cycle (PRC). If the high and low tones form a regular isochronous series, subjects are being asked to produce stress-timed utterances[2].

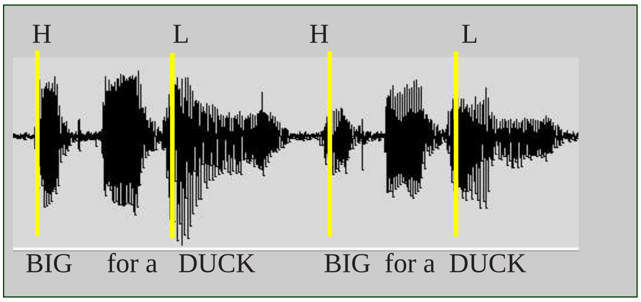

Figure 1: Targeted speech cycling set up. The tone sequence is indicated with ‘H’ and ‘L’. Two repetitions of one phrase are shown, and the stressed syllable onsets are marked by vertical lines.

In a targeted speech cycling experiment, it is thus possible to probe the con-

[2] The Abercrombian notion of a stress foot extending from one stressed syllable onset to the next, irrespective of intervening unstressed syllables, is used here (Abercrombie, 1967). Metrical phonologists may use other working definitions of the foot, but the present interpretation is conventionally used in typological studies of speech rhythm.

straints on macroscopic phrase timing, by specifying a target phase. In a typical trial, the target phase is selected from some range (say a uniform distribution bounded by 0.3 and 0.7), and the corresponding sequence of high and low tones is generated and played through headphones. The subject begins repeating the phrase, attempting to match her phrase onsets and medial stressed syllable onsets with the high and low tones, respectively. Subjects are trained to stop speaking and to skip a whole cycle when they need to draw breath, so that breathing requirements are cleanly separated from production data. Without this training, we found that some subjects would pant, drawing a little breath on each repetition cycle, and this produced manifest artifacts in the data. After a period of synchronization (typically about 15 phrase repetitions, with one or two breaths), the metronome stimulus ceases and the subject continues repeating the phrase, attempting to maintain the specified phase relationship. This methodology is essentially the same synchronization/continuation paradigm typically used in temporal interval production studies which employ metronome-influenced tapping.

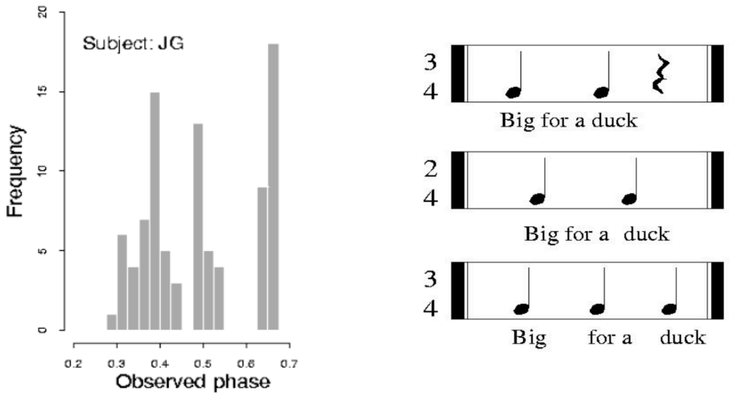

Figure 2: Left: Sample data from one subject in a targeted speech cycling experiment. On each trial, a target phase for the medial stress was drawn from a uniform distribution between 0.3 and 0.7. Measurements were made of the actually produced phase of the target onset within the repeating phrase cycle. The histogram shows the (trial median) phases produced in synchronization and continuation parts of the trial combined. Right: Musical notation depicting the three stable patterns which recur across subjects. From Cummins and Port (1998).

Using this procedure, we were able to demonstrate that there are quite strong constraints on stress foot timing when repeating an English phrase. Figure 2 (left) shows sample data from one subject, after 90 trials, in which a random phase between 0.3 and 0.7 was used on each trial. A strawman null hypothesis is that there are no rhythmic constraints on speech production. By this hypothesis, the phases produced should reflect the phases set as targets, with some additional noise. That is, they should reproduce a uniform distribution. In fact, as can be clearly seen, three patterns are produced much more frequently than any others. These three patterns are found reliably across subjects, and they correspond to an aliquot nesting of the stress foot within the PRC as illustrated in Figure 2 (right).

The use of alternating lights to cue finger taps was originally introduced by Yamanishi and colleagues (1980) to probe the relative stability of patterns. We adapted the procedure here in a speech context to probe for the existence of stable patterns. Although there are gross and manifest differences between the physical actuators involved in simultaneous limb wagging and in speech production, the experimental approach was able to demonstrate that some rather abstract organizational principles are common to both situations, and simple rhythmic constraints can be found to apply to each.

Speech cycling is just one form of external intervention which can be brought to bear on the functioning speech production system to unveil its internal constraints and abilities. Several other experimental set ups can be similarly interpreted. For example, several groups have studied movement after administration of an external perturbation to the jaw during syllable repetition (Abbs and Gracco, 1983; Saltzman and Munhall, 1989). These experiments have revealed that the smooth, rapid compensation which is observed after a perturbation is specific to the intended sound sequence; this finding in turn greatly constrains the space of plausible models of planning and control during production (Saltzman and Munhall, 1989).

Both speech cycling as described here, and Saltzman's mid-cycle perturbations constitute dynamic manipulations to the continuous flow of speech. These methods contrast with studies which involve the use of bite blocks (Lindblom et al., 1979), artificial palates (Flege et al., 1988), and lip tubes (Savariaux et al., 1995). These latter studies alter the context within which speech is produced, but do so statically.

5 Synchronous speech

The speech elicitation method to be described in the remainder of this chapter differs from the previous ones in that speech is reciprocally entrained by speech. At the heart of the method is the simple expedient of having two subjects read a prepared text together, attempting to remain in synchrony with one another. In order to satisfy the task demands, speakers must each adjust their speech so that it is by and large predictable for their co-speaker. Several characteristics of synchronous speech are immediately of interest: firstly, it is easy for speakers to do, and secondly, the resulting speech sounds very natural, despite the experimental manipulation.

5.1 Motivation

Before describing the methods we have employed and results obtained in detail, it is worthwhile motivating the procedure by recourse to an analogy. Much of what phoneticians and phonologists do on a day to day basis is concerned with the attempt to see beyond performance ‘noise’ in the speech signal to its ultimate linguistic constituents. It is a basic tenet of modern linguistics that language production involves the creative sequencing, in accordance with some strict rules, of context-free units. In the process of sequencing these units, they become profoundly affected by context and by the communicative situation in which they are employed. The Quixotic search for simple acoustic invariances which might signal the underlying units (Klatt, 1986; Lindblom, 1990) has served to highlight the considerable distance between the hypothetical context-free units and their much messier realization (Hockett, 1955).

In many respects, we are in a similar situation to a musicologist who has a recording of a performance and wishes to infer the score which gave rise to that performance (Cemgil et al., 2000). Unless the performer is a machine, the relationship between performance and score will not be simple. Along with a (slightly noisy) interpretation of specific note durations (or relative durations) there will always be layers of conventional and idiosyncratic timing variation overlaid on the signal. For a recording of the 14th violinist in the string section of an orchestra, the situation may be different than for a recording of a soloist. In the former case, the violin player is constrained by the requirement that she play in synchrony with the other strings. Variations from the noted durations must therefore be conventional if they are to be implemented similarly by all members of the string section. A clear example of such conventional timing modulation is the rallentendo at the end of a phrase. The soloist will overlay this conventional deviation from the score on her produced temporal intervals too, but will then typically add considerable idiosyncratic variability on top, greatly complicating the life of the musicologist. The freedom of the soloist to depart from the rigid timing expressed in the score is akin to the freedom of an individual speaker to layer expressive timing and idiosyncratic modulation on top of her spoken ’performance’.

Synchronous Speech is an attempt to constrain speakers in a manner analogous to the constraints on the ensemble player. By speaking in time with a co-speaker, the subject must, perforce, make her speech predictable for a co-speaker. She must therefore greatly limit the idiosyncratic and unpredictable elements to her speech timing if a reasonable degree of synchrony is to be attained.

Our initial expectations about the speech which might result were informed in part by familiarity with group speech modes such as prayer, chanting, and the repetition of familiar oaths or texts. A canonical example familiar to Americans is the repetition of the Pledge of Allegiance, while Irish school children will be more familiar with group recitation of specific prayers. The art of ‘Choral Speaking’ is practiced in some schools and even takes the form of a competitive activity in the *feis ceoil* of Ireland. However, all these situations are associated with an exaggeration of conventional prosody, and the use of more or less stylized patterns of timing and intonation which are quite distinct from conventional productions of a single individual. It was thus unclear to what extent the speech produced under synchronous speaking conditions might diverge from normal speech.

5.2 Methodology

The elicitation of Synchronous Speech is a simple matter. In its simplest form, subjects are seated opposite one another. Each wears a near-field head-mounted microphone, and recordings are done onto the right and left channels of a stereo sound file. For many purposes, the little bleed that occurs from one speaker into the channel of the other is of little account. If more stringent sound separation is required, subjects must be sound isolated, with headphones delivering sound from both participants. In this case, fine adjustment of the gain for each speaker is possible, which may help when one speaker has a significantly more resonant voice than the other.

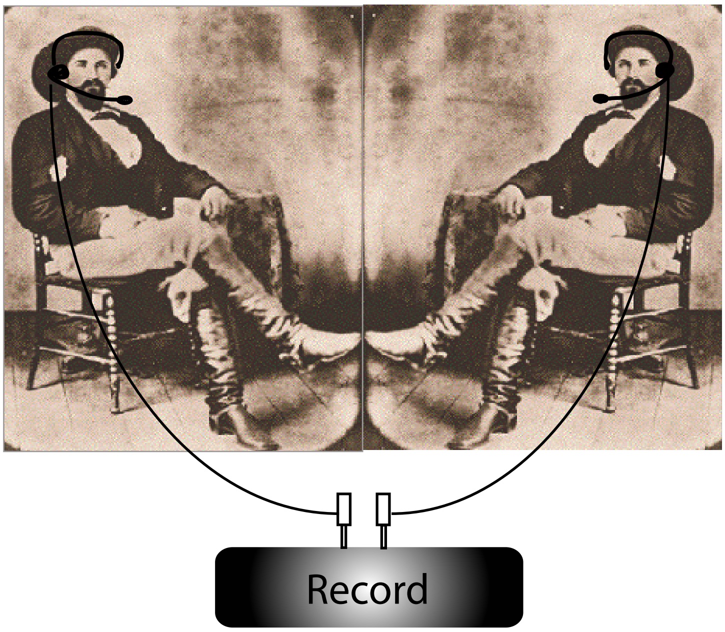

Figure 3: Recording synchronous speech. Speakers are recorded onto the right and left channels of a standard stereo file.

The text to be read should be no more than a single paragraph at a time.

Subjects can read this through silently first, and they are then asked to read in synchrony following a signal from the experimenter. A countdown (3–2–1–go) may be of some help to ensure a simultaneous start. No other preparation is typically required. As will be made clear below, no extensive practice is required. In our experiments, we have not controlled for subject familiarity, but it is to be suspected that dyads who are highly familiar with one another will find the synchronous reading task easier than strangers.

6 Characteristics of synchronous speech

One of the first things that became apparent when we started recording subjects, was that synchronous reading is not a particularly arduous task. Unlike some phonetic laboratory tasks, reading together has the hallmark of familiarity about it. Subjects did not require extensive training. In fact, we found that a single warm-up paragraph was typically all that was required to make subjects comfortable with reading in synchrony. Synchrony can be assessed by looking at the average lag between matched points in the parallel waveforms (Figure 4). Synchrony is typically very good. We found asynchronies of about 40 ms to be typical, with the proviso that values of about 60 ms are more typical at phrase onsets after pauses (Cummins, 2002).

In keeping with this informal appraisal, we studied the effect of practice by comparing mean asynchrony immediately after a warm-up paragraph, with that obtained after approximately 45 minutes of experimental participation, reading a variety of texts with the same co-speaker (Cummins, 2003a). To our surprise, there was no significant improvement over the whole session. This suggests that subjects are rapidly reaching asymptotic performance, and that practice with the task is of little or no value.

We also recorded synchronization performance under less variable conditions by having dyads read one and the same text in synchrony eight times in a row, and comparing synchronization performance at the beginning and end of the series (Cummins, 2003a). There was a slight improvement with this intensive practice which keeps both co-speaker and text constant, with mean asynchrony at phrase onset improving from 68 to 60 ms, and phrase medially from 40 to 27 ms.

Unlike the stylized prosody we expect when groups recite familiar prayers or oaths, synchronous speech elicited as above sounds like natural read speech. Durations are not manifestly different from those obtained from conventional readings. In one study, we compared relative durations (i.e. interval 1 expressed as a proportion of another interval 2) at a variety of time scales (Cummins, 2004). For intervals which spanned more than one syllable, we found a significant reduction in inter-speaker variability, without any attendant alteration to the mean value of the relative durations. This was true for the ratio of two phrases, and for the ratio of a pause over the following or the preceding phrase. For smaller

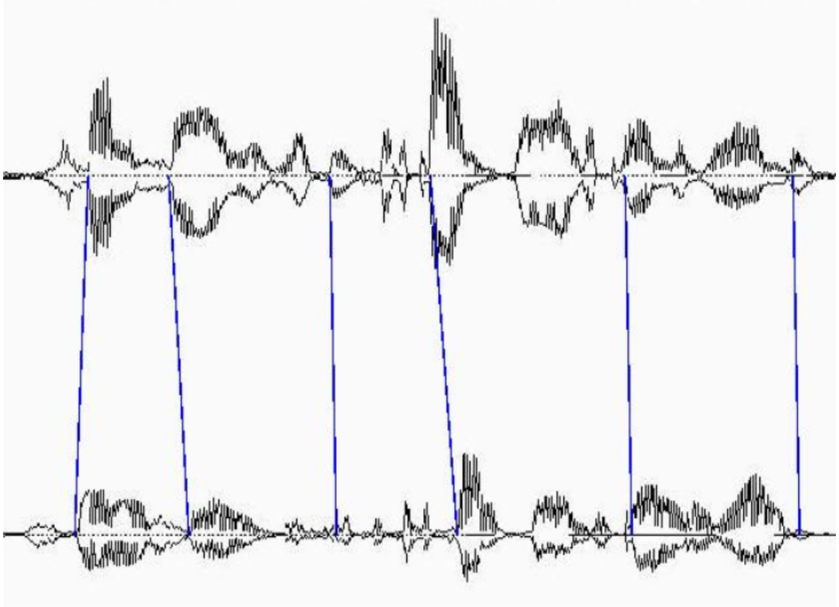

Figure 4: Measuring asynchrony. Corresponding points in the two waveforms are identified, and absolute lags computed.

intervals, there were no significant differences in either mean relative durations, or in inter-speaker variability. This was the case for the ratio of an unstressed to a stressed syllable, of a vowel to its containing word, and for the ratio of an onset consonant to the containing word.

Speaking rates in the synchronous condition are, on the whole, slow, but within the range of normal speaking rates. The pitch range used has been found to be reduced (Cummins, 2002) or unchanged (Wang and Cummins, 2003) compared to solo readings. There are no obvious artifacts to the speech, though a study to see whether listeners can distinguish synchronous speech from conventional readings has yet to be done.

We compared synchrony obtained when the two speakers were facing one another, with a condition in which they sat back to back, and so could not see one another. Despite the fact that they were reading from a sheet of paper, the ability to see the co-speaker, even peripherally, appeared to facilitate the synchronization process, as mean asynchrony in the no-vision condition increased from 63 to 80 ms at phrase onset, and from 42 to 51 ms phrase medially. The effect is small, however, and might plausibly be due in part to altered acoustics when speakers are turned back to back.

7 Applications

All our observations to date support the view that Synchronous Speech is not significantly different from speech obtained under more conventional “solo” speaking conditions, but that for macroscopic prosodic variables, there is a considerable reduction in inter-speaker variability. These results suggest two possible ways in which synchronous speech might be exploited: in the study of prosodic phenomena, and in the study of synchronization itself.

7.1 Studying prosody with synchronous speech

The first area of investigation in which we found Synchronous Speech to be of significant use was in the study of pauses in read speech. As reported above, we had first established that pauses exhibited less inter-speaker variation in the synchronous condition, where we express the pause as a proportion of either of the surrounding phrases. In a series of focussed studies of pauses in a variety of read texts, Zvonik and Cummins (2002, 2003) were able to identify a quantitative relationship between pause duration and the length of the surrounding phrases. In particular, they found that pauses of less than 300 ms duration have a restricted distribution which depends on boundary strength and on the length of the phrases on either side. For strong boundaries which occurred at the end of full sentences, the likelihood of a pause being short (i.e. less than 300 ms) increased if either of the flanking phrases was short (operationally defined as less than or equal to 10 syllables), but was greatly increased if both flanking phrases were short. The effect was superadditive. For weak boundaries, within sentences, only the preceding phrase length affected the chance of a short pause being produced. Critically, Zvonik found that data obtained using synchronous speech provided a substantially clearer picture than parallel data obtained from solo readings. The quantitative relationship could be demonstrated in both data sets, but the synchronous data provided more consistent, clearer results, with more reliable statistical differences between long and short pauses.

Similarly, in another study, I found Synchronous Speech to be invaluable in studying the inter-onset intervals resulting when subjects read sequences of trochaic words in an isochronous fashion (i.e. attempting to produce equally spaced word onsets) (Cummins, 2003b). The reasons for being interested in isochronous interval production are not relevant to the topic of this chapter, but the utility of Synchronous Speech is graphically illustrated in Figure 5. The top row shows intervals obtained for one sample list (left) and for all 8 lists (right) in the solo condition. Matching data from the synchronous condition is shown in the lower row, and the reduction in variability within the data set is quite clear.

Synchronous Speech thus clearly suggests itself as a powerful elicitation method for studying macroscopic temporal phenomena in read speech. Its principal advantage lies in drastically reducing inter-speaker variability without introducing manifest artifacts. It does so, it would seem, by tapping into speakers’ own

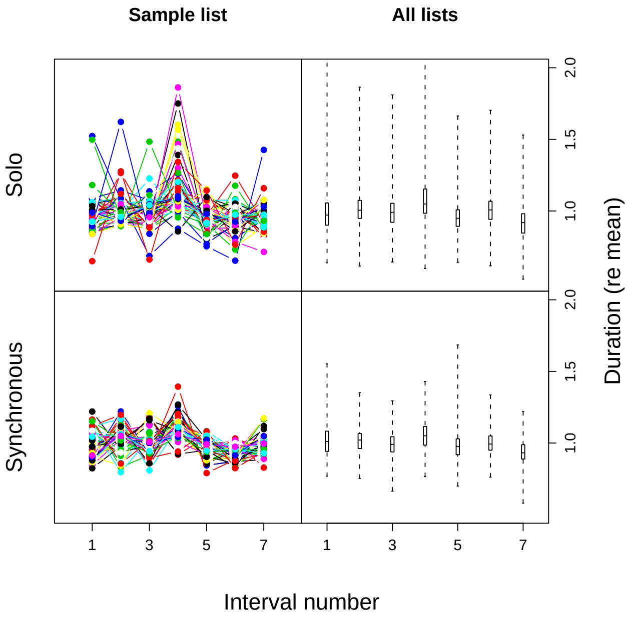

Figure 5: Comparison of inter-onset intervals obtained from readings of lists of eight trochees. Top: solo condition, Bottom: synchronous condition. Left: all intervals from one specific list. Right: intervals from all 8 lists combined. The whiskers span the entire data range. N=54. Reproduced from Cummins (2003b).

implicit knowledge of essential and inessential variation in speech, and requiring them to restrict variation to that which is predictable by a co-speaker.

7.2 Studying synchronization among speakers

I believe there is a second promising use of Synchronous Speech which remains largely untapped. There have been a few attempts in the past to provide objective criteria for delimiting distinct speech styles. For example, in Harnsberger and Pisoni (1999), an attempt was made to elicit speech which was consistently hypoarticulated by asking speakers to perform a concurrent digit retention task. The motivation was to develop a solid empirical laboratory paradigm for defining a specific speech style. The resulting speech did not turn out to convincingly represent a single style, unfortunately, as subjects differed greatly in the way in which they satisfied the task demands.

There is a considerable problem in the study of speech styles. When we think of a ’style’, we imagine speech produced within a rich network of social, cultural and communicative factors. Imagine, if you can, the speech of an elderly agitated librarian from Missouri, or speech of a wealthy drunken comedian from Utah. The style of speech which is indexed by these constraints is related in a very indirect way to the suite of variables we phoneticians measure. Although the notion of speaking style has an intuitive clarity and importance to us as speakers, it is notoriously difficult to pin down.

Herein lies one of the great outstanding problems of phonetics, and indeed of cognitive science as a whole. How do we bridge the gap between what we perceive as a relatively simple control space in which we intentionally change from speaking in one style to speaking in a different one and the much more complex space of observable effects which we can measure? The phenomenal space of stylistic control appears to be low-dimensional, in the sense that we effortlessly change style in a unitary manner, without attending to a vast array of individual indices. Nonetheless, the manifest effects of a style change are legion. Any variable we can think of may be affected as we change, for example, from a calm pedagogical tone of voice to an enraged diatribe. Only by having a range of well-defined speaking styles to study can we hope to relate the apparently simple space of volitional control to the rich and complex space of manifest effects. The task of relating dimensions of intentional control to the observed changes in a rich set of effectors is, of course, not confined to speech, but to the study of action and intention in the most general sense.

It is here that Synchronous Speech may offer a unique contribution, in that it represents a well-defined, easily replicable speaking style which subjects can adopt at will, and which exhibits relatively little variation across subjects. The set of prosodic parameters which have been found to be affected when speaking synchronously includes phrasal timing, pitch range and pause duration. These variables appear to be affected as a group under other conditions in which a wholesale change to speaking style is made. For example, several studies of mimicry have found that mimics imitating another speaker will alter just these variables, while making few, or unsystematic, changes to segmental variables (Eriksson and Wretling, 1997; Wretling and Eriksson, 1998; Zetterholm, 2002, 2003). This set is also collectively affected by many forms of dysarthria, such as the hypokinetic dysarthria characteristic of Parkinson’s Disease. The yoking together of these prosodic variables suggests that they may be modulated by a single control process in speaking, and that this higher level control process is effectively independent of the fine detail characterizing individual segments.

8 Discussion

As a tool for the study of macroscopic prosodic phenomena, Synchronous Speech has several desirable properties: It is easy to elicit from untrained subjects, requires minimal technical equipment, and produces speech with greatly reduced inter-speaker variability for precisely those prosodic variables which otherwise carry much idiosyncratic and expressive information. Pauses, foot and phrase-level timing, and intonation are examples of the kinds of variables which appear with improved clarity in studying Synchronous Speech.

As an object in its own right, however, the synchronization process has the potential to tell us much about the relationship between high and low level variables of control in speech production. Two synchronized speakers are mutually entrained. Throughout all our studies of Synchronous Speech, we have never yet found a dyad in which one speaker followed the other. Rather, all the evidence suggests that speakers are genuinely entraining, one to the other. This mutual entrainment produces a tightly locked compound system. Interestingly, we have observed that when one speaker makes a speech error, this often leads to a complete breakdown in the flow of speech for both speakers, and the reading process must be restarted from scratch. This is interestingly analogous to the disturbance observed with the use of Delayed Auditory Feedback (DAF) (Harrington, 1988). Under conditions of DAF, the normally tight coupling between production and auditory feedback is disrupted, and the effect on speech production is severe. The Synchronous Speaking condition seems to exhibit a similar level of coupling, but between individuals (cf also Schmidt et al. 1990).

The two novel techniques described here, Speech Cycling and Synchronous Speech, collectively illustrate some of the advantages to be gained by viewing the speech production system as a dynamical system with a rich set of behaviours. It suggests that natural or inherent behaviours of the system can be revealed either by external forcing, as with a metronome, or by mutual entrainment, as with Synchronous Speech.

References

Abbs, J. H. and V. L. Gracco (1983): Sensorimotor actions in the control of multimovement speech gestures. Trends in Neuroscience 6, 393–395.

Abercrombie, D. (1967): Elements of general phonetics. Chicago, IL: Aldine Pub. Co.

Backwell, P., M. Jennions, N. Passmore, and J. Christy (1998): Synchronized courtship in fiddler crabs. Nature 391, 31–32.

Bailly, G. (2001): Close shadowing natural vs synthetic speech. In: Proceedings of 4th ICSA ITR Workshop on Speech Synthesis, Perthshire, Scotland: ICSA.

Beckman, M. E. (1996): A typology of spontaneous speech. In: Sagisaka, Y., N. Campbell, and N. Higuchi (eds.), Computing Prosody: Computational Models for Processing Spontaneous Speech, New York: Springer Verlag, 7–26.

Cemgil, A. T., P. Desain, and B. Kappen (2000): Rhythm quantization for transcription. Computer Music Journal 24(2), 60–76.

Cummins, F. (2002): On synchronous speech. Acoustic Research Letters Online 3(1), 7–11.

Cummins, F. (2003a): Practice and Performance in Speech produced Synchronously. Journal of Phonetics 31(2), 139–148.

Cummins, F. (2003b): Rhythmic grouping in word lists: competing roles of syllables, words and stress feet. In: Proceedings of 15th ICPhS 2003, Barcelona, 325–328.

Cummins, F. (2004): Synchronization Among Speakers Reduces Macroscopic Temporal Variability. In: Proceedings of the 26th Annual Meeting of the Cognitive Science Society, 304–309.

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

Dauer, R. M. (1983): Stress-timing and syllable-timing reanalyzed. Journal of Phonetics 11, 51–62.

Eriksson, A. and P. Wretling (1997): How flexible is the human voice?–A case study of mimicry. In: Proceedings of EUROSPEECH, Rhodes, Greece, vol. 2, 1043–1046.

Flege, J. E., S. G. Fletcher, and A. Homiedan (1988): Compensating for a bite block in /s/ and /t/ production: palatographic, acoustic, and perceptual data. Journal of the Acoustical Society of America 83, 212–228.

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

Harnsberger, J. D. and D. B. Pisoni (1999): Eliciting speech reduction in the laboratory II: Calibrating cognitive loads for individual talkers. Tech. Rep. Progress Report No. 23, Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN 47405.

Harrington, J. (1988): Stuttering, delayed auditory feedback, and linguistic rhythm. Journal of Speech and Hearing Research 31, 36–47.

Henry, C. E. (1990): The development of oral diadochokinesia and non-linguistic rhythmic skills in normal and speech-disordered young children. Clinical Linguistics and Phonetics 4(2), 121–137.

Hildebrand, M. (1985): Walking and running. In: Hildebrand, M., D. M. Bramble, K. F. Liem, and D. B. Wake (eds.), Functional Vertibrate Morphology, Cambridge, MA: Harvard University Press, chap. 3, 38–57.

Hockett, C. (1955): A Manual of Phonology. Chicago: University of Chicago.

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

Klatt, D. (1986): The Problem of variability in speech recognition and in models of speech perception. In: Perkell, J. and D. Klatt (eds.), Invariance and Variability in the Speech Processes, Hillsdale, NJ: Lawrence Erlbaum Associates, 300–320.

Liberman, M. Y. and L. A. Streeter (1978): Use of nonsense-syllable mimicry in the study of prosodic phenomena. Journal of the Acoustical Society of America 63(1), 231–233.

Lindblom, B. (1990): Explaining phonetic variation: a sketch of the H&H theory. In: Hardcastle, W. J. and A. Marchal (eds.), Speech Production and Speech Modelling, Dordrecht: Kluwer Academic, 403–439.

Lindblom, B., J. Lubker, and T. Gay (1979): Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. Journal of Phonetics 7, 147–161.

Marslen-Wilson, W. (1973): Linguistic structure and speech shadowing at very short latencies. Nature 244, 522–523.

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

Savariaux, C., P. Perrier, and J. P. Orliaguet (1995): Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: as tudy of the control space in speech production. Journal of the Acoustical Society of America 98, 2428–2842.

Schmidt, R. C., C. Carello, and M. T. Turvey (1990): Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. Journal of Experimental Psychology: Human Perception and Performance 16(2), 227–247.

Sigurd, B. (1973): Maximum rate and minimal duration of repeated syllables. Language and Speech 16, 373–395.

Strogatz, S. H. and I. Stewart (1993): Coupled oscillators and biological synchronization. Scientific American , 102–109.

Swerts, M. and R. Collier (1992): On the controlled elicitation of spontaneous speech. Speech Communication 11, 463–468.

Tajima, K., B. A. Zawaydeh, and M. Kitahara (1999): A comparative study of speech rhythm in Arabic, English and Japanese. In: Proceedings of the 14th International Congress of the Phonetic Sciences.

Tuller, B. and J. A. S. Kelso (1989): Environmentally-specified patterns of movement coordination in normal and split-brain subjects. Experimental Brain Research 75, 306–316.

Wang, B. and F. Cummins (2003): Intonation Contour in Synchronous Speech. Journal of the Acoustical Society of America 114(4(2)), 2397.

Wretling, P. and A. Eriksson (1998): Is articulatory timing speaker specific? – Evidence from imitated voices. In: Proc. FONETIK 98, 48–52.

Xu, Y. (2002): Maximum speed of pitch change and how it may relate to speech. Journal of the Acoustical Society of America 111(3), 1399–1413.

Yamanishi, J., M. Kawato, and R. Suzuki (1980): Two coupled oscillators as a model for the coordinated finger tapping by both hands. Biological Cybernetics 37, 219–225.

Zetterholm, E. (2002): Intonation pattern and duration differences in imitated speech. In: Proc. Speech Prosody 2002, Aix-en-Provence, 731–734.

Zetterholm, E. (2003): The same but different: three impersonators imitate the same target voices. In: Proc of 15th ICPHS, Barcelona.

Zvonik, E. and F. Cummins (2002): Pause duration and variability in read texts. In: Proc. ICSLP, Denver, CO, 1109–1112.

Zvonik, E. and F. Cummins (2003): The effect of surrounding phrase lengths on pause duration. In: Proceedings of EUROSPEECH, Geneva, CH., 777–780.