Chapter 21

Temporal patterning in speech and birdsong

Louis Goldstein
Department of Linguistics, University of Southern California

Speech and birdsong are complex motor behaviors in which patterning over time is itself informational. This is obvious in the case of speech, but in birdsong, too, the sequencing (and possibly timing) of syllables determines in part the well-formedness of the song. Despite gross differences in function, in the physical substrate (method of sound production), in brain structure, and in the scale of the animals, recent work has revealed a surprising degree of similarity in their solutions to the problem of controlling temporal patterning. There are differences, too, of course, and when we find them, it deepens our understanding about the (unique) structure of speech. Because Steve has had a lasting interest in birds and birdsong (Anderson 2006), this seemed to be an appropriate context to review these similarities. Two of them will be the focus of discussion here: decomposition of the behavior into a *sequence of discrete motor units* and the role of an *internal clock system*, partly independent of the units themselves.

1 Discrete decomposition

One of the foundational motivations for Articulatory Phonology (Browman & Goldstein 1992; 1995) is to address the apparent incompatibility between the discrete phonetic and phonological structure of speech, on the one hand, and the observation that the vocal tract articulators move in a continuous fashion, producing continuous modulation of the acoustics, on the other. AP hypothesizes that it is possible to model the continuous motion of the articulators as arising from discrete, context-independent dynamical control systems, called *gestures*, that govern the formation of phonologically-relevant constrictions within the vocal tract (for example bringing the tongue tip to the palate, with a particular degree of constriction). The control parameters of these dynamical systems (*target*, e.g., the phonologically-specified degree of constriction, and *stiffness*, the time constant that determines the amount of time required for the system to settle at its target value) remain fixed during the duration of the constriction action (roughly a consonant or a vowel), even though the articulators are moving. The decomposition of

Louis Goldstein. 2017. Temporal patterning in speech and birdsong. In Claire Bowern, Laurence Horn & Raffaella Zanuttini (eds.), *On looking into words (and beyond)*, 457–471. Berlin: Language Science Press. DOI:10.5281/zenodo.495460

speech into a pattern of gesture activations over time (or *gestural score*), is an abstract analysis (as Steve argued in his earliest work (Anderson 1974) must be the case for a phonetic representation), and can only be discovered by use of explicit dynamical and acoustical models. We cannot observe it directly. Recently, Nam et al (2012) showed that with the use of the TaDA gestural production model, it is possible to parse an acoustic signal into the maximum likelihood gestural score that could have produced it.

A similar strategy is employed by Amador et al. (2013), to decompose zebra finch song into discrete gestures. The authors first developed a dynamical model of sound production in the syrinx, the sound production organ in songbirds. The syrinx is a vibratory system located at the base of the trachea. The trachea divides into two tubes at that point, each of which hosts a pair of vibrating membranes called labia. The two pairs of labia can vibrate together or separately, or can be sequenced (Riede & Goller 2010). The model developed in Amador et al. (2013) (for a single pair of labia) allowed them to generate sound from two dynamical control parameters: the average tension in the labia, and tracheal pressure. Then, using a table-lookup scheme, they were able to estimate the time functions of these two parameters from audio recordings of the sound. This representation was then validated by generating audio from those time functions and playing those sounds to zebra finches. Neural responses from the synthetic song were highly similar (nearly identical) to the responses obtained by playing the original song (BOS, Bird's Own Song). Next, they showed that the derived time functions are essentially discrete: they exhibit sequences of intervals of time during which the values of the control parameters remain essentially fixed (just as with speech gestures), and refer to these intervals as elemental gestures of the song. The key similarity to speech gestures is that the continuous song can be decomposed into discrete intervals of time, longer by an order of magnitude than the periodicity of the song, during which the dynamical parameters are essentially fixed.

There are also some salient differences between speech gestures and zebra finch song gestures. Most superficial is that the control parameters for the zebra finch gestures are different from those of speech gestures that control the constrictions of the suprlaryngeal structures (as expected, because forming constrictions is not generally thought to be part of the bird's song behavior), but they are similar to those control parameters involved in controlling tone and intonation in speech (McGowan & Saltzman 1995). A somewhat deeper difference is that the zebra finch gestures (as analyzed in Amador et al. 2013) are strictly sequential, while speech gestures exhibit various types of overlap in time. At first blush, this makes sense, as the mechanism of sound production in birds is generally thought to be limited to a single device, the syrinx, while the distinct vocal organs of the human vocal tract can each make their own contribution to the filtering action of sound generated at the larynx. There are, however, as noted above, two sets of labia comprising the syrinx (Trevisan et al. 2007), and some avian species employ primarily one set, while others (like zebra finches) use both sets, simultaneously or sequentially. Trevisian et al. (2007) have shown how such symmetry-breaking (different functioning of the two sides at the same time) arises in the species that employ either one side only or both sides, but not symmetrically. So there is the possibility that distinct patterns of gesture overlap may yet be uncovered.

Another obvious potential difference between speech gestures and birdsong is compositionality. A small set of discrete speech gestures are employed in different combinations to create the set of segments and syllables in a language. It is unknown whether gestures in birdsong are compositional in this sense. Zebra finch songs (on which a large bulk of the research on song birds has been done) have been described as having a hierarchical structure. Syllables are the most immediately identifiable units, as they consist of the vocalization intervals produced by the bird that are bounded by (silent) inspirations. The total number of distinct syllables in a given bird's inventory is relatively small, on the order of 20 or so. As described in Yu & Margoliash (1996), syllables can be composed of distinct notes, and in turn, sequences of syllables form *motifs*, that can be repeated as part of a song. It is unclear to what extent the elemental gestures identified in Amador et al. (2013) can form parts of more than one syllable in an individual's inventory. Yu and Margoliash (1996) do report instances of distinct syllables in a bird's inventory that begin with the same note (and end with different notes). To the extent that those shared notes are produced with the same gestures, this would be evidence for some limited compositionality.

2 Clocks

2.1 Clocks in speech?

A lively debate in the 1980's sparked by the work of Carol Fowler (Fowler 1980) considered whether speech units had their own intrinsic timing as dynamical systems (as argued by Fowler), or whether the timing is imposed externally by some kind of central clock. Gestures in Articulatory Phonology are units with intrinsic timing; the time required for a gesture to reach its goal state is determined by its dynamical stiffness parameter. But what of the time between gestures? For a sequence of two gestures, x_1 and x_2, we can ask how the system controls when to trigger x_2 with respect to x_1. A natural answer to this is that x_2 is triggered when some reference state value of x_1 is achieved $\{x_{1,1}\}$. Sequencing in motor systems is often modeled by a mechanism of "competitive selection" of the sequenced items (Bullock & Rhodes 2002; Grossberg 1978), where feedback from the completion of element x_1 (achievement of its target state) allows it to be deactivated and element x_2 to be selected and triggered. In the case of speech, this feedback could be kinesthetic, orosensory and/or acoustic. However, there is an argument that this cannot be the complete story for speech. Consider the gestural score for the word "spot" in the left panel of Figure 1. The boxes represent intervals in which the the supralaryngeal gestures would be active, in some token of this word. The onset of the lip closure for /p/ initiates at a moment when the tongue tip fricative gesture (for /s/) is at a particular state (close to its target and not moving much), as marked by the vertical line in the figure. A possible mechanism for sequencing the these gestures would be to learn that in producing the word "spot," the lip closure gesture "waits" until the system has feedback that the tongue tip is near the alveolar ridge and is moving with little velocity, and at that point it is triggered. But now consider the phrase "toss spot"

shown on the right side of Figure 1. Here, the state that the lip closure is looking for, in order to trigger, occurs too early, because the tongue tip is already in position for /s/. And there is a interval of time (shaded in white) during which the state of the tongue tip does not change much, so there is no information in its state that can inform when the /p/ should be triggered. Nonetheless, its timing must show some regularity, as the [s] in "toss spot" systematically differs in duration from, for example "pa spot." Figure 2 shows the kinematics of the tongue tip in a sequence of identical supralaryngeal gestures ("had tied"). It is clear that there is indeed a considerable stretch (70 ms or so) marked with a yellow box where its state does not change, and so information about when to trigger a next gesture is lacking.

This case appears to argue that a simple state-based chain triggering will not work for speech, in the general case, and relative timing must be specified in some way separately from the actual gestural content of the units. It would also be possible to argue, in this case, that the lip gesture is triggered when the tongue tip gesture begins to release, but this just pushes the problem back onto the tongue tip release gesture. How does it know know when to trigger, without access to information about time? It can't just use the position and velocity of the tongue tip. Tilsen (2016) has recently proposed a theory that gestural sequencing based on feedback from the preceding gestures does indeed characterize the system at early stages of the child's development, gradually shifting to a different, coordination-based scheme as described below (for at least some syllable-contexts). It could also be countered that the cases presented here involve timing across words, and perhaps word-sequencing is controlled by a separate mechanism from within-word gesture sequencing. However, the same issue would arise within words in the case of geminate consonants (consonants that are maintained for a long temporal interval). And of course, the existence of geminates at all is itself *prima facie* evidence for some independence of timing and gestural content, as the same constrictions can be maintained for different durations, under linguistic control.

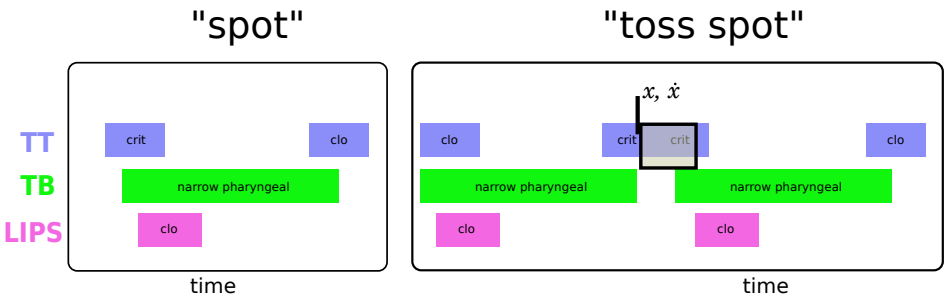

Figure 1: Gestural scores. On the left for the word "spot" and on the right for the phrase "toss spot." Rows represent (from the) top, gestures of the Tongue Tip, Tongue Body and Lips. See text. Shaded area represents interval of time during which state of tongue tip is not changing.

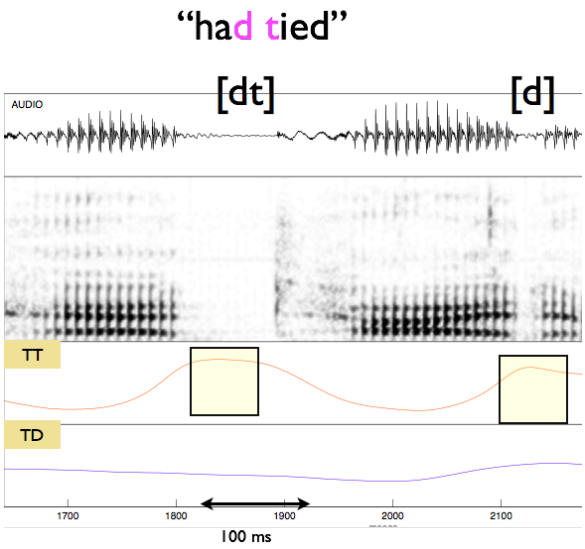

Figure 2: Time functions of the vertical position of the Tongue Tip (TT) and Tongue Dorsum (TD) in the phase "had tied." See text.

The coupled oscillator model of syllable structure (Goldstein, Byrd & Saltzman 2006) proposes a specific alternative to gesture sequencing, in which the clock machinery is separate from the particular gestures that form the syllable. In this model, the gestures composing a syllable are triggered by a system of planning oscillators (clocks) that are coupled to one another in distinct modes. Each planning oscillator triggers activation of a gesture. Specifically, clocks that trigger gestures comprising onset consonants (consonants preceding the vowel in a syllable) are coupled in-phase (the most stable mode) to the vowel gesture and clocks that trigger gestures comprising coda consonants (consonants following the vowel in a syllable) are coupled in anti-phase mode to the vowel. If every gesture is triggered at phase 0 degrees of its planning oscillator, then two gestures that are coupled in phase will be triggered synchronously. This synchronous triggering explains data that show that the onset of articulatory movement for an onset consonant and for the following vowel begin at roughly the same time (Goldstein, Byrd & Saltzman 2006). When two gestures are coupled in anti-phase mode, however, they will be triggered a half-period apart in time, which would be consistent with the observed time lag between the onset of the vowel gesture and the onset of a coda consonant gesture (Goldstein, Byrd & Saltzman 2006). The ensemble of oscillators can be formally represented as a (coupling) graph, and Figure 3 shows the coupling graph for the word "tab." Green edges represent in-phase coupling and the dashed red edge represents anti-phase coupling. Note that the same graph topology would underlie the timing of gestures in any CVC syllable, and in that sense, the clock is separate from, and independent of, the particular gestures that are deployed. The model has been used to explain patterns of syl-

lable typology, acquisition (Nam, Goldstein & Saltzman 2009), asymmetric coordination patterns in onset vs coda (Marin & Pouplier 2010), and weight, and it has been used as a diagnostic for the syllable structure of complex pre-vocalic clusters (Hermes, Muecke & Grice 2013; Shaw et al. 2009).

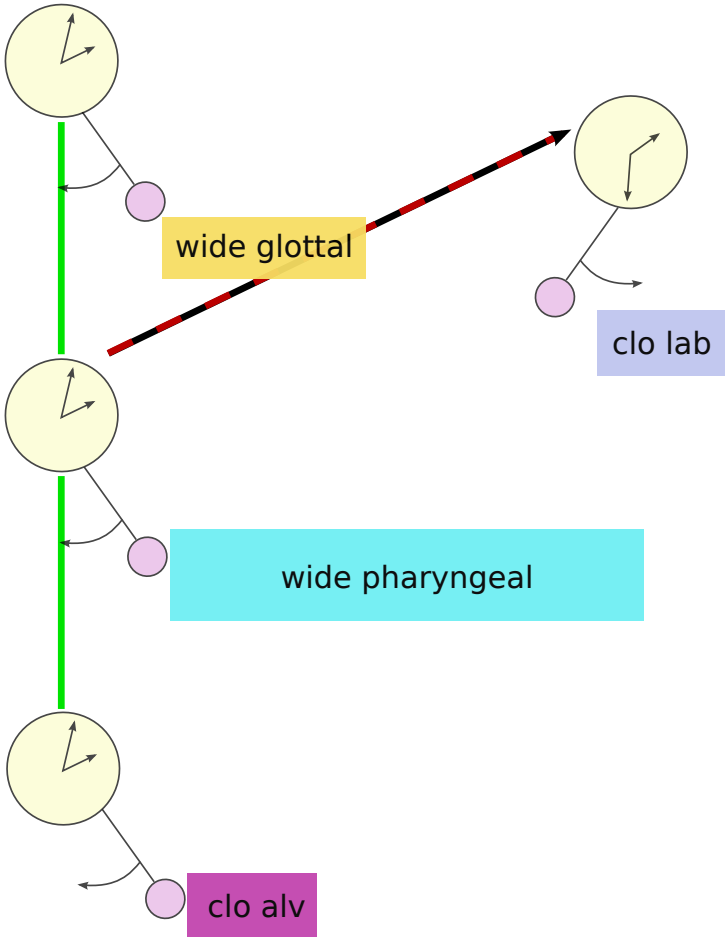

Figure 3: Coupling graph for the word "tab." Each clock represents one of the gestures in the word and they are the nodes of the coupling graph. From the top down the left, these are glottal abduction for the initial /t/, pharyngeal constriction (for the vowel), and tongue tip closure for the /t/. At the right is the lip closure for /b/. Green lines are graph edges that represent in-phase coupling and the red dashed line with arrowhead is the edge that represents anti-phase coupling. Boxes represent the gestural score for the the word (gesture activation over time) that results from running the coupled oscillator model.

The frequencies of the planning oscillator clocks are all defined with reference to the ticks of an overall speech rate clock. Prosodically-induced lengthening can thus be modeled as slowing of the rate of this overall clock, as has been proposed in the π-gesture model (Byrd & Saltzman 2003). Phrase edges are associated with local π-gestures, which function to slow the movements of all gestures that fall under the scope of the π-gesture. This model has been shown to account for the acoustic and kinematic correlates of boundary lengthening in a variety of languages (e.g. Greek; Katsika et al. 2014).

2.2 Clocks in birdsong?

Work on birdsong over the last 15 years has also revealed, within limits, separate control of timing and vocal organ activation patterns. Two areas of the avian cortex have been identified as significant for the production of the song: HVC, a pre-motor nucleus, and RA (robust nucleus of the arcopallium). HVC projects to RA, which in turn projects to the vocal motor neurons (and to midbrain vocal control areas). HVC was suspected to be a major site of timing control, and this was tested in a seminal study by Long and Fee (2008). Reasoning that cooling a brain region would result in a slowing of neural patterns, they used a miniature Peltier device to locally cool either HVC or RA. They found that cooling HVC resulted in slowing of the song, with the amount of slowing being proportional to the degree of cooling. Further, the slowing was fairly linear throughout the song. Syllable durations, onset lags, gap durations between motifs were all slowed to roughly the same degree, indicating that something like an overall clock (like the proposed speech clock) was being slowed. Consistent with the independence of timing account, there was very little change at all in the actual acoustics of the song, indicating that the control of the activation patterns at the level of the motor neurons remained intact, just spread out in time. (In other words, the rate at which the motor commands were issued was slowed down, but the commands were not changed, so the frequencies of the song were not altered by slowing). Conversely, even though spiking was decreased by cooling RA, the ability of (uncooled) HVC to drive RA and produce typical song speeds was not impaired, thus providing evidence for localizing timing control in HVC.

A more recent study of cooling from a different lab using canaries (Goldin et al. 2013) found that with more extreme cooling of HVC, the song begins to break down, exhibiting period-doubling of respiratory patterns, causing the emergence of additional syllables. The authors provide a formal model that predicts these transitions from the nonlinear interaction between the (hypothesized) neural pulse train (from HVC) and the dynamics of the respiratory cycle. Interestingly, this kind of period-doubling can also be observed in "gestural intrusions" human speakers produce when repeating phrases like "top cop", and a similar dynamical account has been proposed, less formally (Goldstein et al. 2007). That study found that when speakers produce such phrases repeatedly, they will begin to produce an "extra" copy of the tongue tip gesture of /t/ concurrently with the initial tongue dorsum gesture of "cop" (resulting in a co-produced /k͡t/) and conversely an extra tongue dorsum gesture during the initial tongue tip gesture of "top". These extra cycles of repeated tongue tip or tongue dorsum movement can be analyzed as a period doubling

– 2:1 to 1:1 transitions in frequency mode locking between the tongue tip (or tongue dorsum) oscillators and the lip oscillator of the syllable-final lip gesture (there is a lip gesture in every syllable, but a tongue tip or dorsum gesture only every other syllable; Goldstein et al. 2007). Since such period-doubling transitions in birdsong are analyzed by Goldin et al. (2013) as resulting from a presumed slowing of a clocking pulse in HVC, the results do not contradict the main finding and conclusion of the earlier work of Long and Fee. However, there is disagreement between the two research groups as to the nature of the temporal code in HVC and how it interacts (or not) with the rest of the system, as will be fleshed out a bit in the last section.

2.3 Brain-cooling in speech

The technique of focal brain cooling was recently employed with humans for the first time (Long et al. 2016) with patients undergoing brain surgery for either epilepsy or tumor resection. Cooling was applied in up to 4 locations in each subject, two in Broca's area within the left inferior frontal gyrus (IFG) and the others in the precentral gyrus (speech motor cortex). The hypothesis was that there would be a double dissociation with cooling in Broca's area causing changes in speech timing but not in articulatory quality, and that cooling the speech motor cortex would disrupt articulation, but not timing. Patients were recorded producing the digits from 1 to 20 or the names of the days of the week (one sequence per trial, with breaks between trials) while respective sites were being cooled, and also during control trials with no cooling. The utterances were judged for quality via crowd-sourcing on a scale from 0 (extremely degraded) to 1 (typical/normal). Timing was determined through durational measurements. Results supported the double dissociation. Cooling Broca's area resulted in changes to speech timing. Typically utterances were slowed down (both the actual articulation of words and the gaps between them were stretched), but some cases speeding up occurred. No effect was found on judged quality. When the speech motor cortex was cooled, ratings shifted to more degraded, with no effect on timing.

To examine the slowing more carefully, the authors generously made available the data from two of their subjects, one of which is analyzed here. Figure 4(a-b) shows boxplots for durations of the names of the days of the week (excluding pauses between names); on the left, control utterances are displayed and on the right, the trials with cooling of Broca's area. Results show fairly uniform slowing across the names of the days of week, except for "Friday," which shows less slowing. Somewhat surprising is that Friday is the shortest of all the words (even in the controls); there is certainly no tendency for list-final prosodic lengthening here. In terms of intonation, M-W generally appear to be produced with an extended High tone. Falling begins on "Thursday" and "Friday" is generally produced on a Low tone. So it is possible that the durations follow the prominence profile of the utterance. By itself, however, this does not explain the reduced percentage of slowing on "Friday." Another possibility is that cooling in Broca's area has a bigger effect on more complex syllable types (for example with coda consonants or clusters). The initial syllables of the days of the week all have closed syllables (with coda

consonants) except "Friday". To test this, the durations of the initial syllables and final syllable ("day") were analyzed separately, and the magnitude of lengthening of initial (dark blue bars) vs. final (light yellow bars) syllables are shown in Figure 5. Magnitude of lengthening is calculated as the ratio of the median duration of that syllable when cooled divided by the median duration of controls. The lengthening of the (open) syllable "day" is approximately the same across all the days' names. The lengthening of the first syllable in "Friday" (an open syllable) is about the same as for "day", while the other first syllables (that are closed) lengthen more. The most lengthening is observed on the first syllable of "Wednesday", which is also the most complex syllable, closed with a coda cluster. This is consistent with the hypothesis that more complex syllables exhibit more slowing due to cooling in Broca's area.

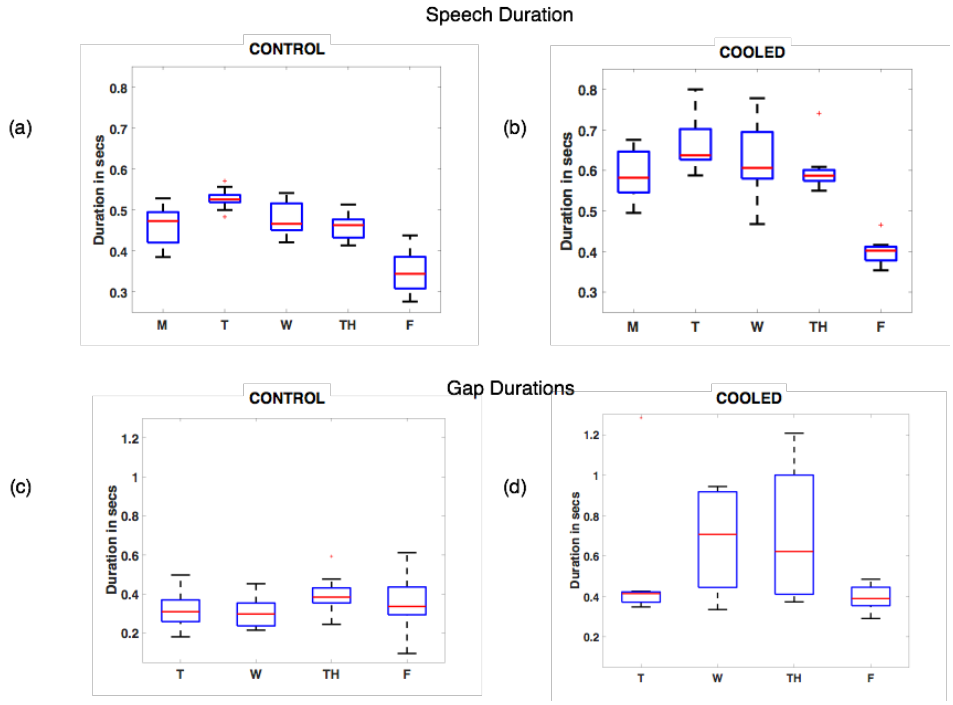

Figure 4: (a-b) Boxplots of duration of the names of days of the week from one patient. Control condition is shown in (a), cooling Broca's area is shown in (b). (c-d) Boxplots of silent gap durations before the production of the names of the week Tuesday-Friday. Control condition is shown in (c), cooling Broca's area is shown in (d).

Figure 4(c-d) shows the durations of the silent period before initiation of the words "Tuesday" to "Friday" from the time of completion of the preceding word. This shows a strikingly different pattern from that exhibited by the word durations. The silent gaps

before "Tuesday" and "Friday" show very little effect, while the gaps before "Wednesday" and "Thursday" show almost 3:1 lengthening. The pattern is reminiscent of the classic pattern of recall in short-term memory (Deese & Kaufman 1957; Ebbinghaus 1885; Brown, Neath & Chater 2007). The items in the middle of the list have more competitors on either side and therefore more interference, although many other models have been proposed. Such results could be modeled by a competitive queuing model of sequence selection (Bullock & Rhodes 2002), depending on exactly how the parameters are set. In any case, it is clear that more is going on than just clock slowing when Broca's area is cooled, unlike what is observed in zebra finch, though clock slowing is also going on.

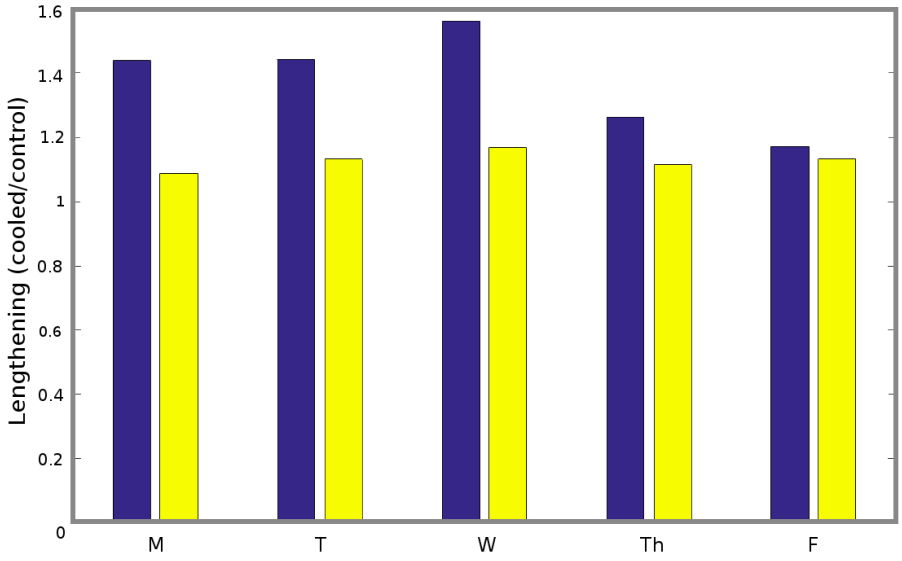

Figure 5: Magnitude of lengthening of initial (dark blue bars) vs. final (light yellow bars) syllables of the names of the days of the week. Magnitude of lengthening is calculated as the ratio of the median duration of that syllable when cooled divided by the median duration of that syllable in the uncooled control condition.

In summary, the results with humans generally confirm the dissociation of the timing clock from the articulatory gestural patterning that it paces. Differently from birds, however, where gaps between syllables and motifs were slowed in a roughly similar way to the actual song syllables, the gaps between the days of the week showed marked differences in response to cooling, depending on position in the list (for the one patient examined). This suggests that even for an over-learned list, mechanisms of selection of discrete individual words must still be in play, while for the bird, the entire song may just "run off" at different rates. This may be related to the relative lack of flexibility in the zebra finch song. Also the possible effect of syllable complexity on the magnitude

of slowing also suggests that more is going on for humans when Broca's area is cooled than just uniform slowing of the clock.

2.4 Continuous vs discrete temporal representations

The last point brings up the nature of temporal coding that characterizes the representation in HVC of the bird (compared for example to the clock model proposed for speech discussed above). The work of Fee and collaborators has consistently supported the view that the representation is a continuous-time representation of the song (in 10 ms or so slices). This is based on the earlier discovery (Hahnloser, Kozhevnikov & Fee 2002) of individual cells that burst sparsely in the song at a fixed lag from song onset. Theoretically then, there could be such cells for each 10ms sample in the sound, and they jointly produce a continuous representation. Further, their hypothesis is that the continuous-time representation completely drives (or enslaves) the downstream activity in RA and the vocal muscles to reproduce the song (Long & Fee 2008), which is why the slowing does not result in distortions to the song (but cf. the results discussed earlier with extreme values of cooling). An alternative discrete view was proposed by the Margoliash and Mindlin group (Amador et al. 2013). After discovering that it was possible to decompose the dynamical parameters governing song production into discrete gestures, as discussed above, they found that burst times of HVC neurons projecting to RA tended to be synchronized with the gestural extrema, for 14 of the 15 sites they examined with recordings of single neurons. This is exactly what would be predicted by coupled oscillator model of syllable structure described above: the clock mechanism generates a sequence of bursts that trigger their corresponding gestures. However, attempts to replicate this finding with a substantially larger population of cells, in both Long's lab (Picardo et al. 2016) and Fee's (Lynch et al. 2016), failed to replicate this finding. It is unclear why this is, apart from possible differences in sites examined and the types of electrodes used. It would not be surprising to find that both continuous and discrete representations co-exist in different subpopulations of neurons. The discrete representation would be useful during learning to produce individual "pieces" of the song on the way to mastery (assuming a continuous representation of the target song in auditory areas is any case available to the system). Consistent with this, Lynch et al. (2016) did find evidence of 10-Hz rhythmicity locked to song syllables, which was significant for HVC projections to Area-X (basal ganglia loop employed in learning) but not for HVC projections to RA. Given the stereotypy of zebra finch song, it is not surprising that a continuous-time representation could work. Obviously in the case of speech, we are capable of producing novel forms, and for that a discrete representation like the coupled oscillator model is really the only viable candidate (or compatibly, models like that of Bohland and Guenther, e.g., Bohland, Bullock & Guenther 2010).

3 Summary

Speech and birdsong share the property that their production can be decomposed into a sequence of discrete motor actions. In addition, the control of those actions is governed by a separate timing representation. The nature of the timing representation appears to be substantially different however, possibly due to the essential combinatoriality and productivity of human speech, though there is a lot still unknown about both speech and birdsong in this regard. It is interesting to consider why they should be as similar they are. One functional similarity is that while they are both species-specific capabilities, in both cases the specific behavioral forms must be learned by individuals (in the bird species in which the song is learned from experience). There are other odd similarities as well, such as the compatible frequency of their syllable rates. This flies in the face of hypotheses that the duration of the syllables in speech is related to the natural frequency of the jaw (e.g., Davis & MacNeilage 2004). A more likely cause may be the similarity of their auditory systems. In any case, the existence of a model system that can be probed in ways that speech cannot provides the opportunity of deepening our understanding of speech, particularly when we observe the particular places in which the systems diverge.

Acknowledgements

Many thanks to Michael Long and Kalman Katlowitz for making their data available, and to Jelena Krivokapic, Mairym Llorens, Sarah Bottjer, and two anonymous reviewers for their comments on an earlier draft.

References

Amador, Ana, Yonatan S. Perl, Gabriel B. Mindlin & Daniel Margoliash. 2013. Elemental gesture dynamics are encoded by song premotor cortical neurons. *Nature* 495(7439). 59–64.

Anderson, Stephen R. 1974. *The organization of phonology*. New York: Academic Press.

Anderson, Stephen R. 2006. *Doctor Dolittle's delusion: Animals and the uniqueness of human language*. New Haven: Yale University Press.

Bohland, Jason W., Daniel Bullock & Frank H. Guenther. 2010. Neural representations and mechanisms for the performance of simple speech sequences. *Journal of Cognitive Neuroscience* 22(7). 1504–1529.

Browman, Catherine & Louis Goldstein. 1992. Articulatory phonology: An overview. *Phonetica* 49(3–4). 155–180.

Browman, Catherine & Louis Goldstein. 1995. Dynamics and articulatory phonology. In Timothy van Gelder & Robert F. Port (eds.), *Mind as motion*, 175–193. Cambridge, MA: MIT Press.

Brown, Gordon, Ian Neath & Nick Chater. 2007. A temporal ratio model of memory. *Psychological Review* 114(3). 539–576.

Bullock, Daniel & Bradley J. Rhodes. 2002. Competitive queuing for planning and serial performance. *CAS/CNS Technical Report Series* 3(3). 1–9.

Byrd, Dani & Elliot Saltzman. 2003. The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics* 31. 149–180.

Davis, Barbara L. & Peter F. MacNeilage. 2004. The frame/content theory of speech evolution: From lip smacks to syllables. *Primatologie* 6. 305–328.

Deese, James & Roger A. Kaufman. 1957. Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of Experimental Psychology* 54(3). 180–187.

Ebbinghaus, Hermann. 1885. *Memory: A contribution to experimental psychology.* (Originally published 1885 in German. Translated by Henry A. Ruger and Clara E. Bussenius). New York: Teachers College, Columbia University.

Fowler, Carol A. 1980. Coarticulation and theories of extrinsic timing. *Journal of Phonetics* 8. 113–133.

Goldin, Matías A., Leandro M. Alonso, Jorge A. Alliende, Franz Goller & Gabriel B. Mindlin. 2013. Temperature induced syllable breaking unveils nonlinearly interacting timescales in birdsong motor pathway. *PLoS One* 8(6). e67814.

Goldstein, Louis, Dani Byrd & Elliot Saltzman. 2006. The role of vocal tract gestural action units in understanding the evolution of phonology. In Michael A. Arbib (ed.), *From action to language via the mirror neuron system*, 215–249. Cambridge: Cambridge University Press.

Goldstein, Louis, Marianne Pouplier, Larissa Chen, Elliot Saltzman & Dani Byrd. 2007. Dynamic action units slip in speech production errors. *Cognition* 103(3). 386–412.

Grossberg, Stephen. 1978. A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. *Progress in Theoretical Biology* 5. 233–374.

Hahnloser, Richard H. R., Alexay A. Kozhevnikov & Michale S. Fee. 2002. An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature* 419(6902). 65–70.

Hermes, Anne, Doris Muecke & Martine Grice. 2013. Gestural coordination of Italian word initial clusters – The case of 'impure s.' *Phonology* 30(1). 1–25.

Katsika, Argyro, Jelena Krivokapić, Christine Mooshammer, Mark Tiede & Louis Goldstein. 2014. The coordination of boundary tones and their interaction with prominence. *Journal of Phonetics* 44. 62–82.

Long, Michael A. & Michale S. Fee. 2008. Using temperature to analyse temporal dynamics in the songbird motor pathway. *Nature* 456(7219). 189–194.

Long, Michael A., Kalman A. Katlowitz, Mario A. Svirsky, Rachel C. Clary, Tara McAllister Byun, Najib Majaj, Hiroyuki Oya, Matthew A. Howard & Jeremy D.W. Greenlee. 2016. Functional segregation of cortical regions underlying speech timing and articulation. *Neuron* 89(6). 1187–1193.

Lynch, Galen F., Tatsuo S. Okubo, Alexander Hanuschkin, Richard H. R. Hahnloser & Michale S. Fee. 2016. Rhythmic continuous-time coding in the songbird analog of vocal motor cortex. *Neuron* 90(4). 877–892.

Marin, Stefania & Marianne Pouplier. 2010. Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model. *Motor Control* 14(3). 380–407.

McGowan, Richard S. & Elliot L. Saltzman. 1995. Incorporating aerodynamic and laryngeal components into task dynamics. *Journal of Phonetics* 23. 255–269.

Nam, Hosung, Louis Goldstein & Elliot Saltzman. 2009. Self-organization of syllable structure: A coupled oscillator model. In François Pellegrino, Egidio Marisco & Ioana Chitoran (eds.), *Approaches to phonological complexity*, 299–328. Berlin/New York: Mouton de Gruyter.

Nam, Hosung, Vikramjit Mitra, Mark Tiede, Mark Hasegawa-Johnson, Carol Espy-Wilson, Elliot Saltzman & Louis Goldstein. 2012. A procedure for estimating gestural scores from speech acoustics. *Journal of the Acoustical Society of America* 132(6). 3980–3989.

Picardo, Michel A., Josh Merel, Kalman A. Katlowitz, Daniela Vallentin, Daniel E. Okobi, Sam E. Benezra, Rachel C. Clary, Eftychios A. Pnevmatikakis, Liam Paninski & Michael A. Long. 2016. Population-level representation of a temporal sequence underlying song production in the zebra finch. *Neuron* 90(4). 866–876.

Riede, Tobias & Franz Goller. 2010. Functional morphology of the sound-generating labia in the syrinx of two songbird species. *Journal of Anatomy* 216. 23–36. DOI:doi:10.1111/j.1469-7580.2009.01161.x

Shaw, Jason, Adamantios I. Gafos, Philip Hoole & Chakir Zeroual. 2009. Syllabification in Moroccan Arabic: Evidence from temporal stability in articulation. *Phonology* 26. 187–215.

Tilsen, Sam. 2016. Selection and coordination: The articulatory basis for the emergence of phonological structure. *Journal of Phonetics* 55. 53–77. DOI:doi:10.1016/j.wocn.2015.11.005

Trevisan, Marcos A., Brenton Cooper, Franz Goller & Gabriel B. Mindlin. 2007. Lateralization as a symmetry breaking process in birdsong. *Physical Review E* 75(3 Pt 1). 031908.

Yu, Albert C. & Daniel Margoliash. 1996. Temporal hierarchical control of singing in birds. *Science* 273(5283). 1871–1875.