Structured nonstationarity in articulatory timing

Sam Tilsen

Department of Linguistics, Cornell University
tilsen@cornell.edu

ABSTRACT

Many experiments investigating speech articulation examine the influence of linguistic or paralinguistic factors on articulatory timing. The current experiment differs from these by focusing on unconditioned variation—i.e. statistical noise—in articulatory timing. The aim is to assess whether distributions of intervals representing the relative timing of articulatory movements are well-modelled with a constant parameter distribution, such as a Gaussian with fixed mean and variance. Several hundred productions of the same target word were elicited from participants in a task where feedback encouraged identical productions on all trials. The results show that the distributions of articulatory timing intervals are not consistent with a constant-parameter model. Instead, distribution parameters show unconditioned drift over time, exhibit sporadic discontinuous transitions between modes, and have a complex, dynamic correlational structure. These results call into question typical assumptions of articulatory models.

Keywords: Articulatory timing, coordination, speech production, phonetics, motor control.

1. INTRODUCTION

Most models of articulatory control assume that some set of linguistically- and behaviourally-relevant control parameters accounts for the majority of variance in articulatory timing. Residual variance is treated as noise. The sources of this noise are rarely discussed, but a common presumption appears to be that the noise arises from measurement error and/or stochastic fluctuations at low levels of the nervous system. A reasonable null hypothesis is that residual variance is normally distributed and that articulatory interval distribution parameters (in the case of a Gaussian distribution, μ and σ) remain constant in the absence of conditioning factors. However, many studies of non-speech motor control have shown that temporal intervals rarely follow a constant-parameter distribution; moreover, recent studies have emphasized that understanding the processes responsible for generating noise is crucial for understanding motor behaviours [1,3].

The current study aimed to test the hypothesis that interval distributions conform to a constant parameter model by eliciting numerous repetitions of the same articulatory pattern from each of three speakers. The task included acoustically-based feedback in order to promote reduced variability in articulation. The results show that variance of articulatory intervals is not well characterized by a constant parameter model. Rather, articulatory intervals have nonstationary distributions and articulatory movements exhibit complex patterns of correlation which are dynamic on a range of timescales within an experimental session.

2. METHOD

2.1. Participants and task

Three native speakers of English with no speech or hearing problems participated in the experiment. Participants were seated in a quiet room in front of a computer monitor. Acoustic recordings were collected with a shotgun microphone located approximately 1.5 m from the participant. Articulator movements were recorded at 100 Hz with an NDI Wave electromagnetic articulograph. Sensors were located midsagittally on the upper lip (UL), lower lip (LL), tongue tip (TT: 2 cm from the apex), tongue body (TB: 4 cm posterior from the TT sensor, and lower gingiva (JAW) to capture jaw movement. Reference sensors located on the nasion and left/right mastoid processes were used to correct each frame for head movement in post-processing.

The experimental session was organized into blocks of 50 trials. Trial onsets were jittered 7.0s ± 3s to avoid rhythmic or list-reading effects. On each trial, a green box appeared on the monitor for 2.5 seconds, cueing participants to produce the target word: *demolish*. This word was chosen because it involves a word-medial stressed syllable with a bilabial closure in the context of a superior-to-inferior/anterior-to-posterior movement of the tongue body/root. This configuration of movements maximizes the extent to which the timing of the articulations for the bilabial closure are independent of the movements involved in producing the vowel [a]; the chosen configuration presumably minimizes the effect on measurements of timing attributable to mechanical coupling between the lower lip and tongue body via the jaw.

Participants were shown a printed version of the target word prior to the experiment and were instructed to try to say the word exactly the same way every time the green box appeared. Furthermore, after an initial set of trials, participants received a score (from 0-100) for each production, reflecting how well the production matched a target.

The target production was determined after the 7th trial of the experiment, using the acoustic signals from the preceding 5 trials (i.e. trials 3-7). Each of these trials was converted to a matrix consisting of 13 Mel frequency cepstral coefficients (MFCCs; freq. range: 75-4000 Hz; window size: 25 ms), computed at time steps of 10 ms. Each cepstral coefficient in the MFCC matrix was normalized independently across time steps. The contributing MFCC matrices were then time-aligned with an iterative algorithm based upon pair-wise cross-correlations. The resulting optimally-aligned MFCC matrices were averaged to create a target MFCC matrix.

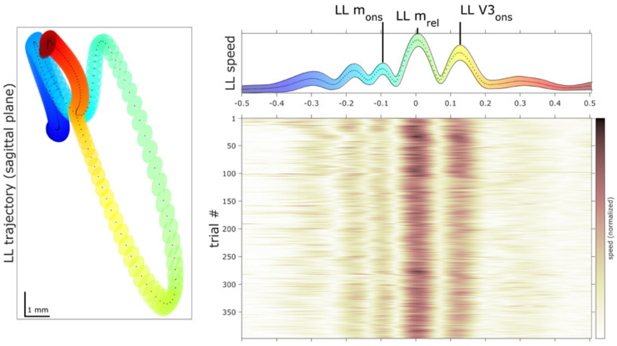
For each subsequent trial, the maximal cross-correlation (xc_{max}) of the target and trial MFCC matrices was used to calculate a score reflecting the degree of similarity between the production and the target. The score was obtained by linearly mapping the xc_{max} z-score to the range [0, 100] (z-scores were calculated relative to the distribution of all previous xc_{max} and truncated to the interval [-2, 2]).

2.2. Data processing and analysis

Participants S01, S02, and S03 performed 7, 8, and 9 blocks respectively (because of differing time constraints), resulting in 350, 400, and 450 productions of the target word. Articulatory landmarks (i.e. points in time with potential relevance to control, derived from sensor trajectories) were estimated with two goals in mind: (1) to avoid (as much as possible) a priori assumptions about which intervals of time are controlled, and (2) to ensure that only the most robustly identifiable temporal landmarks were used in the analysis. To these ends, consistently present maxima in the magnitudes of the sagittal-plane Euclidean velocities (i.e. speeds) of each of the five articulators/sensors (UL, LL, JAW, TT, TB) were located in each trial. To facilitate accurate landmark identification, sensor speeds were aligned by speaker and articulatory channel across trials using iterated cross-correlation with the mean.

Example speed maxima are shown in Fig. 1; these correspond to points in time when a sensor is moving relatively quickly, and are assumed to be relevant for assessing cognitive processes involved in control of articulatory timing. The rationale for using speed maxima rather than alternatives such as positional extrema or velocity threshold crossings is that speed extrema may be less subject to bias from anatomical or speaker-specific factors not directly relevant to control processes.

Figure 1: Example of aligned speed time series from participant S01, with LL trajectory (left), mean LL speed trajectory ±1 s.d. (upper right), and LL speed trajectories aligned across trials (lower right).



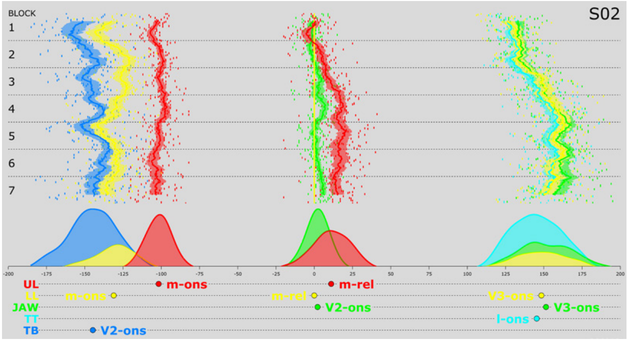
3. RESULTS

3.1 Non-stationarity of articulatory timing

Durations between articulatory landmarks generally did not have a constant mean across a recording session, and hence the hypothesis of constant distributional parameters can be rejected. Dynamic structure can be readily inferred by visual inspection of the moving average durations in Fig. 2. Note that landmarks are depicted relative to the LL [m] release speed maximum.

Many interesting dynamic features are evident in the relative timing of articulatory landmarks. For example, the UL [m] release speed maximum was closely synchronized with LL [m] release in the first two blocks, but subsequently the UL [m] release drifted so as to occur later.

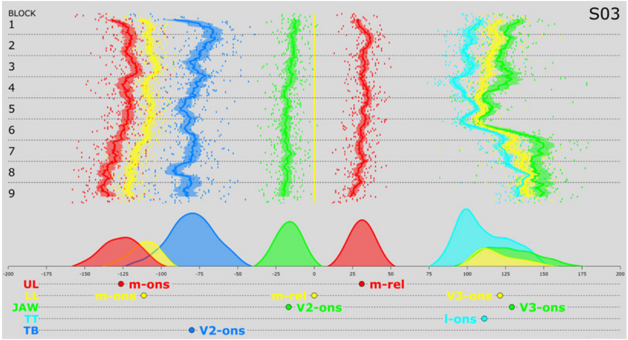
Figure 2: Articulatory landmarks and moving-averages with ±2 s.e. (50-trial window) for speaker S02. Landmarks are aligned to LL [m] release.



An even more extensive drift occurred for the LL and JAW movements associated with V3 ([ɪ]) and the TT movements associated with [l]. Another notable dynamic feature involves the timing of the V2 ([a]) TB lowering movement and [m] bilabial closure—the interval between these landmarks is approximately 25 ms in the first several blocks, but midway through the fourth block there appears to be a rapid shift toward nearly synchronous timing.

It is noteworthy that in some cases the timing of articulatory landmarks appears to change gradually, while in other cases there are more abrupt changes. An example of this can be seen in Fig. 3 for S03, where a gradual drift of V3/[l] onset movements is followed by an abrupt change in block 6. Some of the abrupt changes appear to be associated with the block structure (50 trials per block), but this is not so for all of them.

Figure 3: Articulatory landmarks and moving-averages with ±2 s.e. (50-trial window) for speaker S03. Landmarks are aligned to LL [m] release.



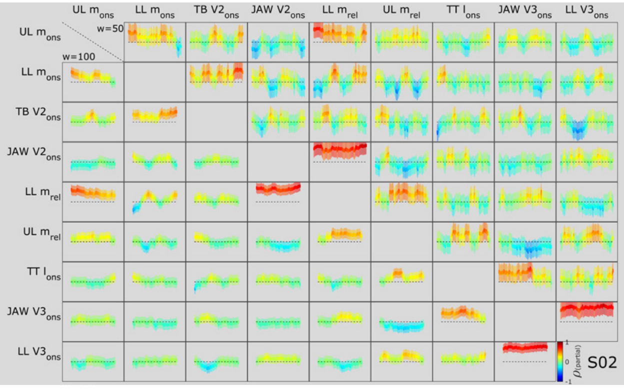
Inter-speaker differences in the relative timing of landmarks are also important factors. For example, the TB lowering associated with V2 onset occurs earlier for S02 and is more closely phased with LL [m] onset than UL [m] onset, whereas for S02 TB lowering for V2 onset occurs later and is less closely phased with LL [m] onset than UL [m] onset. S01 (not shown) exhibits a pattern in which TB lowering for V2 is more closely phased with LL/UL release landmarks as opposed to UL/LL onset landmarks. These observations are important because they implicate speaker-specific factors as an important source of variability in articulatory timing.

3.2 Dynamicity of correlations

One approach to investigating why articulatory timing distributions exhibit non-stationarity is to analyse the temporal correlations of landmarks. Analyses of these correlations show that like articulatory timing relations, they are dynamic over the course of a session and differ between speakers.

Figs. 4 and 5 show partial correlations between articulatory landmarks estimated with moving windows of 50 trials (upper triangle) and 100 trials (lower triangle). Note that the partial correlations estimated with a smaller window better capture the temporal dynamics of correlation structure but are associated with greater uncertainty (confidence intervals were estimated using a Monte Carlo procedure in combination with random permutation of trials). The partial correlation (i.e. the correlation between two variables which have been residualized by removing the effects of all remaining variables) is used here because it more directly represents the portion of the interaction between a pair of articulatory landmarks that is unique to that pair.

Figure 4: Moving-window partial correlations and 95% confidence intervals for speaker S02. Lower triangle window size is 100 trials, upper triangle window size is 50 trials.



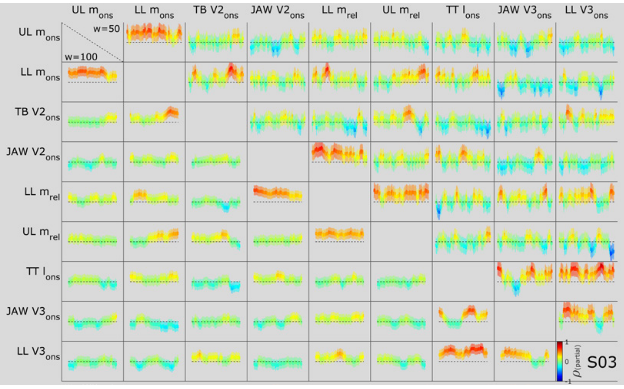
Inspection of the partial correlation time series in Figs. 4 and 5 reveals a number of striking patterns. Some landmarks are consistently correlated. For example, the most highly correlated landmark pairs are LL [m] release and JAW V2 onset, and LL V3 onset and JAW V3 onset. These correlations are not surprising giving the mechanical coupling between lower lip and jaw. Note however, that while S02 exhibits consistently high correlation between these pairs, for S03 the correlation diminishes over the course of the session.

Changes in partial correlation structure are indeed quite pervasive, and in some cases appear to involve multiple landmarks with complementary partial correlation patterns. For example, for S02 the LL and UL [m] release landmarks alternate between epochs of relatively high/low correlation with the LL [m] onset landmark. In other words, when the timing of the LL [m] release and LL [m] onset are positively correlated, the timing of the UL [m] release and LL [m] onset are more negatively correlated. This pattern may suggest that these three landmarks form a set of mutually coordinated movements.

In contrast to landmark pairs which exhibit either relatively consistent correlation or dynamic fluctuations in correlation, there are many landmark pairs which are mostly uncorrelated or only weakly correlated. These involve heterosyllabic landmark pairs, i.e. pairs in which one landmark is associated with [m]/V2 and the other is associated with [l]/V3. Yet there are some exceptions to this generalization. For S02 there is an epoch in the second half of the session in which the TT [l] onset is positively correlated with the UL [m] release. The correlation cannot be attributed to mechanical coupling and is not present in S03.

As with articulatory interval durations, the dynamics of partial correlation exhibit both relatively slow drift and more abrupt, discontinuous changes. For example, for S03 one can observe in the final block of the session the sudden emergence of a timing regime in which the UL [m] release and LL V3 onset are negatively correlated to a substantial degree (especially with the 50-trial window). Visual inspection of Figs. 4 and 5 indeed reveals that abrupt shifts in partial correlation are quite common.

Figure 5: Moving-window partial correlations and 95% confidence intervals for speaker S03.



4. DISCUSSION

Articulatory timing and partial correlations of movement landmarks were found to be highly non-stationary. The durations between articulatory landmarks exhibited changes on the timescale of an experimental session. These changes occurred in spite of a task design which encouraged identical productions across the session. Moreover, the changes took the form of relatively slow drifts and abrupt changes—these patterns are evidence of structure in the processes responsible for control of timing. These observations indicate that temporal intervals and their variance cannot be appropriately modelled by a distribution with a time-independent mean and variance. This raises the question of what processes are responsible for the observed non-stationarity of articulatory timing.

To some extent external factors, not directly related to control processes, may be partly implicated. For instance, session-scale dynamics of temporal intervals were not infrequently associated with the block-organization of the session (2 minute breaks intervened between blocks). A shorter break or no break at all would likely reduce this effect but risk fatiguing the speaker. Another factor which may be implicated is speakers' conscious attempts to achieve higher feedback scores. Participants likely attempt to identify parameters of their production that are associated with better scores, such as differences in vowel quality (d[i]molish vs. d[ə]molish) or subtle differences in vowel duration. So doing, they may mistakenly choose parameters which result in lower scores, and hence the interplay between the feedback score and conscious attempts to control the production process may be a source of non-stationarity.

In general the experimental results show that there is structure to the session-timescale dynamics of articulatory timing. Some of this structured non-stationarity may be attributable to the block-based organization of the session or feedback-design of the task, and future experiments should address these confounds. However, it is unlikely that all of the structure in the session-scale dynamics is a product of the experimental design, and this raises questions regarding what control parameters are changing and why they are changing. In the planning oscillators model of articulatory phonology [2,4], one possible interpretation is that non-stationarity arises from changes in phase-coupling forces between gestural planning oscillators which are selected together in a syllable [5,6,7]. This predicts that the patterns of variability observed between heterosyllabic gestures will differ from those observed between tauto-syllabic gestures. This prediction is consistent with the observations here but requires more data for confirmation.

One of the most important results of this study is that changes in relative timing occurred on multiple timescales, taking the form of both slow drifts and abrupt shifts. The origins of these changes remain an open question, and future studies of articulatory timing must investigate them in order to develop more accurate and realistic models of articulatory control.

5. REFERENCES

[1] Balasubramaniam, R., Torre, K. 2012. Complexity in Neurobiology: Perspectives from the study of noise in human motor systems. *Crit Rev Biomed Eng* 40.

[2] Browman, C., Goldstein, L. 1992. Articulatory phonology: An overview. *Phonetica* 49: 155–180.

[3] Riley, A., Turvey, T. 2002. Variability and determinism in motor behavior. *J Mot Behav* 34: 99–125.

[4] Saltzman, E., Nam, H., Krivokapic, J., Goldstein, L. 2008. A task-dynamic toolkit for modelling the effects of prosodic structure on articulation. *Proc. 4th Int. Conf. Speech Prosody*. Brazil: Campinas, 175–184.

[5] Tilsen, S. 2013. A dynamical model of hierarchical selection and coordination in speech planning. *PloS One* 8(4): e62800.

[6] Tilsen, S. 2014. Selection and coordination in temporally constrained production. *J. Phon*, 44: 26-46.

[7] Tilsen, S. 2014. Selection-coordination theory. *Cornell Working Papers in Phonetics and Phonology, 2014:*24-72.