Critically-damped oscillators and General Tau Theory exhibit similar error across speakers with different vocal tract dimensions

Rationale: Dynamical models are useful for bridging discrete and continuous aspects of speech. In this study, we compare the ability of two models, critically-damped oscillators (CDO) and General Tau Theory (GTT), to predict articulatory trajectories across speakers. Rather than applying an optimization method, we instead use kinematic measurements to determine model parameters. This provides a new kind of perspective on how well the models fit observed data. We also investigated whether differences in vocal tract dimensions affect model fit.

Methods: We use data from the X-Ray Microbeam Database (Westbury et al. 1994), specifically the non-word productions of consonants between schwa and a low-back vowel ("uh-ba", etc.). We selected labial and alveolar stops, nasals, and fricatives (/pbmfv/, /tdnsz/) in this environment, each of which was produced once per speaker. Closing movements of the lower lip and tongue tip were extracted using 20% velocity thresholds: trajectories were considered to start when velocity had reached 20% of peak, and to end when velocity had slowed to 20% of peak.

Specific measurements of these trajectories were used to set the parameters of each model. For the CDO model, we measured the starting position, maximum velocity, and position at the point of maximum velocity. For the GTT model, we measured the starting position, time at peak velocity, and end time of the movement. We then used these parameterized models to estimate the position of the vocal tract at each point in time. For each model trajectory, error was calculated as the summed euclidean distance between the estimated and actual positions.

Vocal tract dimensions were estimated from third and fourth formants at the schwa midpoint, following the assumption that this reflects resonant frequencies in a tube open at one end.

Results: Both models exhibited similar error. For CDO-estimated positions, the median distance was 13.5mm (IQR 9.6-21.1). For GTT-estimated positions, the median distance was 13.3mm (IQR 10.5-19.5). The two models' error was correlated with each other: ($r(29)=.95$, $p < .0001$).

Linear mixed-effects models were fit to the, and model comparison was used to find the significant predictors of error. Results were similar for CDO and GTT. All models included a random intercept for speaker (random-slope models failed to converge or had singular fits). Model fit was improved by including fixed effects of trajectory duration, trajectory magnitude (distance), and place of articulation (labial or alveolar). Vocal tract length and speaker gender did not affect model fit.

Discussion: CDO and GTT performed similarly in how they fit the data. Better fits are possible when optimization is performed over the full trajectory, but the current method, based on the point of peak velocity, led to substantial errors. Since the models performed similarly on speakers of different gender and vocal tract size, we find no bias according to vocal tract dimensions. However, both models exhibited large errors with our parametrization methods.