The speech production system is reconfigured to change speaking rate

Joe Rodd[1,2] Hans Rutger Bosker[1,3] Mirjam Ernestus[2,1] Antje S. Meyer[1,3] Louis ten Bosch[2,1]
[1]*Max Planck Institute for Psycholinguistics,* [2]*Radboud University, Centre for Language Studies,* [3]*Radboud University, Donders Institute for Brain, Cognition and Behaviour*

It is evident that speakers can freely vary stylistic features of their speech, such as speech rate, but how they accomplish this has hardly been studied, let alone implemented in a formal model of speech production. Much as in walking and running, where qualitatively different gaits are required cover the range of different speeds, we may predict there to be multiple qualitatively distinct configurations, or 'gaits', in the speech planning system that speakers must switch between to alter their speaking rate or style. Alternatively, control might rely on continuous modulation of a single 'gait'. This study investigates these possibilities through simulations of a novel connectionist computational model of the cognitive process of speech production, which mimics the temporal characteristics of observed speech.

Connectionist model Our model (Figure 1) is derived from Dell, Burger and Svec's [1] model of serial order in language production, and sequentially retrieves syllable-level motor plans in response to activation in a word level input node. A frame node mediates, encoding metrical structure and enforcing serial order. This model is the first of its type to predict the precise timing of motor plans and account for the ability to control rate in speech production.

The model has many parameters (connection weights, thresholds, etc.) that can be adjusted to fit a speaking rate. Different 'regimes' (combinations of parameter settings) can be engaged to achieve different speaking rates. We consider each parameter as a dimension of a high-dimensional 'regime space', in which the regimes occupy different locations.

Model training Our model approximated the distributions of observed syllable durations and syllable overlap durations in the PiNCeR corpus of Dutch disyllabic words produced at fast, medium and slow speaking rates. Syllable onset and offset were identified from the acoustic signal on the basis of spectral instability as an index of syllable overlap. Together, these duration distributions form a 'fingerprint' of the speech production system operating at a given rate. The model was trained separately for each speaking rate, by the evolutionary optimisation algorithm NSGA-III [2]. The training identified parameter values that resulted in the model to best approximate the duration distributions characteristic of each speaking rate. The fit of the model was assessed by calculating the Kullback-Leibler divergence between the model's predicted distributions and those taken from the corpus for each speaking rate.

Predictions In one gait system, where we 'speed-walk' to speak faster, the regimes used to achieve fast and slow speech are qualitatively similar, but quantitatively different. In regime space, they would be arranged along a straight line. Different points along this axis correspond to different speaking rates. In a multiple gait system, where we 'walk-speak' for slower speaking rates, but 'run-speak' to speak faster, this linearity would be missing. Instead, the arrangement of the regimes would be triangular, with no obvious relationship between the regions associated with each gait, and an abrupt shift in parameter values to move from speeds associated with 'walk-speaking' to 'run-speaking'.

Results Our model achieved good fits in all three speaking rates. In regime space, the arrangement of the parameter settings selected for the different speaking rates is clearly not triangular, suggesting that 'gaits' are present in the speech planning system (Figure 2). Further models fitted at intermediate points in regime space between the speaking rates revealed stark non-linearities between slow and medium and between slow and fast, but not between medium and fast (Figure 3). This leads us to conclude that one configuration is engaged for medium and fast speech, and a second qualitatively distinct configuration is engaged for slow speech. Thus, we provide the first computationally explicit connectionist account of the ability to modulate the speech production system to achieve different speaking styles.

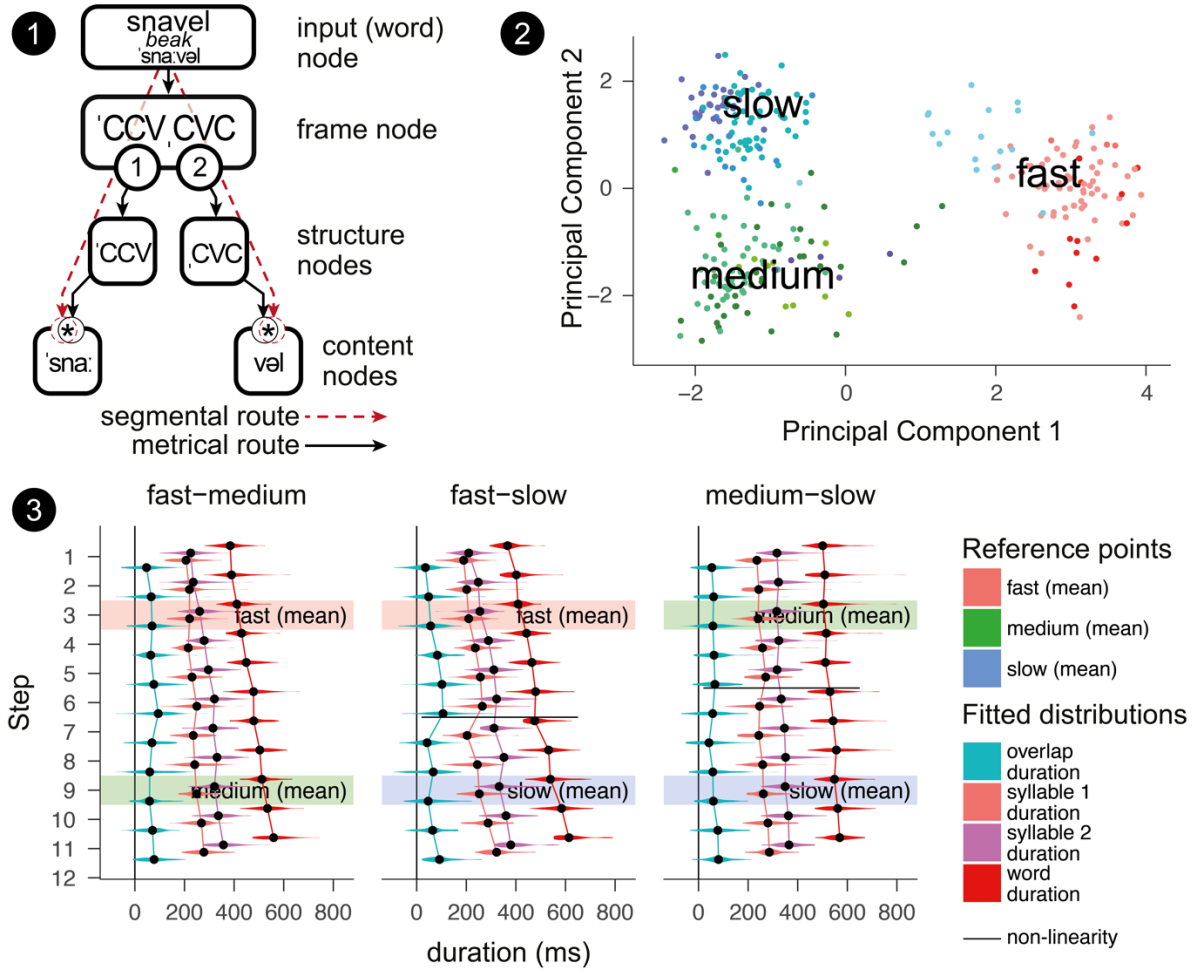

Figure 1. *The connectionist model, showing the nodes and connections required to produce the two syllables of the disyllabic Dutch word "snavel".*

Figure 2. *The best fitting parameter values for the three rate conditions, projected into PC1-PC2 space, showing the triangular arrangement of the parameters that best fit each rate.*

Figure 3: the predicted 'fingerprint' syllable duration, overlap duration and word duration distributions of interpolated points in regime space, illustrating the non-linearities (horizontal black lines) between fast and slow and medium and slow, and the absence of a non-linearity between fast and medium.

[1] G. S. Dell, L. K. Burger, and W. R. Svec, "Language production and serial order: A functional analysis and a model." *Psychological review*, vol. 104, no. 1, pp. 123–147, 1997.

[2] K. Deb and H. Jain, "An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 4, pp. 577–601, 2014.