ISSP 2024 - 13th International Seminar on Speech Production
13-17 May 2024, Autrans, France

A dynamic neural field model of vowel diphthongisation

Sam Kirkham[1], Patrycja Strycharczuk[2]

[1]*Lancaster University, UK*
[2]*University of Manchester, UK*
s.kirkham@lancaster.ac.uk, patrycja.strycharczuk@manchester.ac.uk

Abstract

We advance a computational model of vowel diphthongisation that situates phonological representations in dynamic neural fields (DNFs), which represent the time-varying activation of neural populations that are sensitive to a given phonetic parameter range. We model all long vowels as two separate inputs to the DNF, with input timing governed by a coupled oscillator model that generates an anti-phase relationship between inputs. The location of time-varying maximum activation in the DNF forms a noisy dynamic target, which is used as input to a task dynamic model of gestural coordination. We find that spatial characteristics of long vowels are well captured by the model, which exhibits gradient variation between monophthongs and diphthongs. We also show that a simplified model of production/perception can simulate changes in a speaker's phonological planning representations, which could represent a mechanism behind sound change if transmitted across a community.

Keywords: Articulatory Phonology, Task Dynamics, Dynamic Field Theory, computational modelling, vowels

1. Introduction

The variable diphthongisation of vowels in English is a widely attested form of synchronic variation, such as the monothongisation of GOAT and PRICE in the dialects of Northern England, as well as diphthongisation of tense monophthongs, such as FLEECE and GOOSE (Hughes, Trudgill, and Watt 2012). Some speakers even alternate between such variants, such as producing variably diphthongal or monophthongal vowels. The issue of variable diphthongisation also underpins accounts of diachronic change, such as the diphthongisation of Middle English /i/ and /u/ into present-day /aɪ/ and /aʊ/, as a consequence of the English Great Vowel Shift (Jespersen 1909).

In Strycharczuk et al. (submitted), we account for the gradient nature of diphthongisation by proposing a compositional two-target model for all long vowels (following precedents in Labov, Ash, and Boberg 2006; Popescu and Chitoran 2022). In this view, a short monophthong is short because it has a single target, while a long monophthong is long because it is comprised of two sequentially-timed gestures, each of which has identical targets. A diphthong has the same underlying structure as a long monophthong (two targets), but has different parameters for each of the targets, thus yielding movement from the first target to the second. However, this model does not contain appropriate mechanisms that would help to explain observed variability in vowels, such as the role of perceptually-driven change and the mechanisms behind variability in an individual speaker. One possibility is that each speaker has a single target for the component gestures, but that over time a community drifts towards a new set of targets. This hypothesis is untenable, as we know that speakers can also be highly variable. An alternative is that an individual speaker has a distribution of targets, which would facilitate an account of observed variability. But where do these distributions originate and how do they undergo change?

We outline a solution by grounding phonological representations in a dynamical planning field. Specifically, we use the mathematical and conceptual insights of dynamic field theory (DFT) (Schöner, Spencer, and The DFT Research Group 2016), which have proven to be a versatile tool for dynamical models of phonological planning (Kirov and Gafos 2007; Tilsen 2007; Roon and Gafos 2016; Tilsen 2019; Harper 2021; Shaw and Tang 2023; Stern and Shaw 2023). A dynamic neural field (DNF) model situates phonological planning in an activation field over a phonetic parameter range. A dynamical equation specifies the evolution of field activation until some value reaches a threshold, which is selected as the parameter value for speech production. We then model production and perception as inputs to the field, allowing us track how the field develops over real-time speech planning, as well as over longer timescales. The following model is inspired by integrative dynamical models of timing, planning and execution (Tilsen 2018; Tilsen 2019), as well as by the proof-of-concept DFT model of sound change in Kirov and Gafos (2007).

2. Model architecture

2.1. Dynamic neural field model

A phonological planning representation is modelled as a dynamic neural field, which evolves according to (1) (Schöner, Spencer, and The DFT Research Group 2016). τ dictates the rate of field evolution, $-u(x,t)$ is time-dependent activation at each field site x, h is the resting level of the neural field, $s(x,t)$ represents an input to the field, and $\xi(x,t)$ is Gaussian noise scaled by a factor q.

$$\tau \dot{u}(x,t) = -u(x,t) + h + s(x,t) + \int k(x-x')g(u(x',t))dx' + q\xi(x,t) \quad (1)$$

An input $s(x,t)$ represents any task-specific input, such as phonological planning units or perceptual input, and is modelled in (2) as a Gaussian distribution over a parameter x with amplitude a, centroid p and width w. A model can have multiple inputs, which are summed as $s_1(x,t) + s_2(x,t) + s_n(x,t)$.

$$s(x,t) = \sum_i a_i \exp\left[-\frac{(x-p_i)^2}{2w_i^2}\right] \quad (2)$$

10.21437/issp.2024-52

The interaction kernel $k(x - x')$ in (3) defines excitatory and inhibitory forces across the DNF. Each field location only contributes to above-threshold activation when it exceeds a threshold of $u = 0$. Interaction is excitatory for nearby locations and inhibitory for distal locations. c_{exc}, σ_{exc} are the mean and standard deviation of the excitatory component, while c_{inh}, σ_{inh} are the mean and standard deviation of the inhibitory component. c_{glob} is a global inhibition constant.

$$k(x - x') = \frac{c_{exc}}{\sqrt{2}\pi\sigma_{exc}} \exp\left[-\frac{(x - x')^2}{2\sigma_{exc}^2}\right] - \frac{c_{inh}}{\sqrt{2}\pi\sigma_{inh}} \exp\left[-\frac{(x - x')^2}{2\sigma_{inh}^2}\right] - c_{glob} \quad (3)$$

The interaction kernel is gated by a sigmoidal function $g(u)$, where β is the slope of the sigmoid and α is a threshold, typically set to $\alpha = 0$, whereby only activation values above zero contribute to supra-threshold activation.

$$g(u) = \frac{1}{1 + \exp(-\beta(u - \alpha))} \quad (4)$$

2.2. Coupled oscillator model of gestural timing

We model phonological planning as separate planning inputs $s_{nuc}(x, t)$, $s_{glide}(x, t)$ for the nucleus and offglide. The relative timing of these inputs is determined via the coupled oscillator model in (5) (Tilsen 2018). Φ_{ij} is the relative phase between oscillators i, j, such that $\Phi_{ij} = \theta_i - \theta_j$. C_{ij} is a matrix of coupling strengths between oscillators i, j, where $C_{ij} > 0$ is in-phase and $C_{ij} < 0$ is anti-phase.

$$\dot{\theta}_i = 2\pi f_i + \Sigma_j C_{ij} \sin(\Phi_{ij}) \quad (5)$$

We model all planning units with the same oscillator frequency f = 4 Hz and each unit lasts for 200 ms. If two vowel planning units of 200 ms are coupled anti-phase then the offglide will begin 100 ms after the nucleus. This does not mean, however, that the period of activation will be 300 ms, as there is a time lag between an input to the DNF and activation reaching the threshold. Above-threshold activation can also persist after an input is removed, due to stability-promoting mechanisms in the model. We ensure realistic vowel durations by setting input amplitudes such that activation relaxes to resting level shortly after an input is removed. While we believe that the timing of gestural onsets via coupled oscillators is neurally plausible, the notion of fixed input durations is likely not, so this represents a simplifying heuristic in lieu of a more realistic mechanism, such as feedback-induced gestural suppression (Tilsen 2019).

2.3. Task dynamic model

The DNF governs gestural selection, activation durations, and time-varying gestural targets. We model gestural dynamics using the model in (6) from Saltzman and Munhall (1989), where m is mass, b is a damping coefficient, k is a stiffness coefficient. The task dynamic literature conventionally defines $m = 1$ and $b = 2\sqrt{m}k$, which makes (6) a critically damped oscillator (see Iskarous 2017 for an accessible overview of this model).

$$m\ddot{x} + b\dot{x} + k(x - T(t)) = 0 \quad (6)$$

Gestures are commonly represented by a single target T, but the DNF produces time-varying activations across a parameter range, which represent a dynamic target $T(t)$. Tilsen (2019) proposes a DNF model with dynamic targets, whereby an activation-weighted target supplants the gestural blending mechanism of Saltzman and Munhall (1989). In our study, the target simply tracks the location of peak activation. This enforces stricter selection dynamics, as sudden changes in the location of peak activation results in sudden changes in the target.

The presence of neural noise in the DNF means that the location of peak activation is often a noisy function of time, so how do we avoid overly noisy gestural trajectories? The key concept is that the time-varying location of peak activation is a dynamic input $T(t)$ to the model in (6), not the actual articulatory movement trajectory. The stiffness term k acts as a restoring force that governs the acceleration of the system. Lower values of k constrain movement between dynamic target values, essentially acting as a low-pass filter that forces smoothness on trajectories. Importantly, this is not a form of ad-hoc smoothing, but inherent to the dynamics of the system, allowing smooth gestural trajectories to emerge from noisy neural outputs.

2.4. Computational implementation

All computational models in this paper were implemented in Python 3.9.13, with numerical integration computed using `scipy.integrate.solve_ivp`. Numerical parameters are as follows: DNF [$x = [-10, 10]$, $\tau = 50$, $h = -2$, $\xi = \mathcal{N}(0, 1)$, $q = 3$, $\Delta t = 0.001$, $\Delta x = 0.1$]; kernel: [$c_{exc} = 1, c_{inh} = 0.5, c_{glob} = 0.1, \sigma_{exc} = 1, \sigma_{inh} = 3$], sigmoid: [$\alpha = 0, \beta = 1.5$]; coupled oscillator: [$\Delta t = 0.001, f = 4$]; task dynamics: [$\Delta t = 0.001$, $m = 1$, $k = 2/\Delta t$]. Parameter values encode relative relationships between elements and the specific values are not integral to the model.

3. Simulation results

3.1. Long monophthongs vs. diphthongs

Figure 1 shows example DNFs for three cases: (1) a long monophthong with two identical targets, p = [0,0]; (2) a diphthong with two different targets, p = [3,0]; (3) a diphthong with a bigger distance between two targets, p = [5,0]. The first target has amplitude $a = 3$ and the second $a = 6$ to represent the difference in relative blending weight in favour of the offglide in traditional task dynamic models. The parameter range represents an abstraction for Tongue Body Constriction Location (TBCL), where 0 is a palatal vowel and 5 is a pharyngeal vowel (the parameter range is purely heuristic for the purposes of illustration). In the top row, the left panel shows a single peak: the second input is at the same location as the first, thereby boosting the peak to a higher activation level. The middle panel shows the emerge of two peaks, which briefly overlap, causing a sudden change in the location of maximum activation. The right panel shows a similar dynamic, but the first input sits at a higher value on the parameter range, resulting in a larger difference in the location of peak activation between onset and offset.

The second row of Figure 1 shows the location of peak activation on the parameter axis (the noisy field means there are minor jumps in this value between time-steps). The third and fourth rows show the output of task dynamic simulations, with the dynamic target as time-varying input (initial position $x = -1$; initial velocity = 0 for all examples). Note that the position and velocity trajectories are moderately smooth, with some minor perturbations in the velocity signal. This demonstrates that the distinction between a long monophthong (single velocity minimum) and a diphthong (two velocity minima) can be captured by the model.

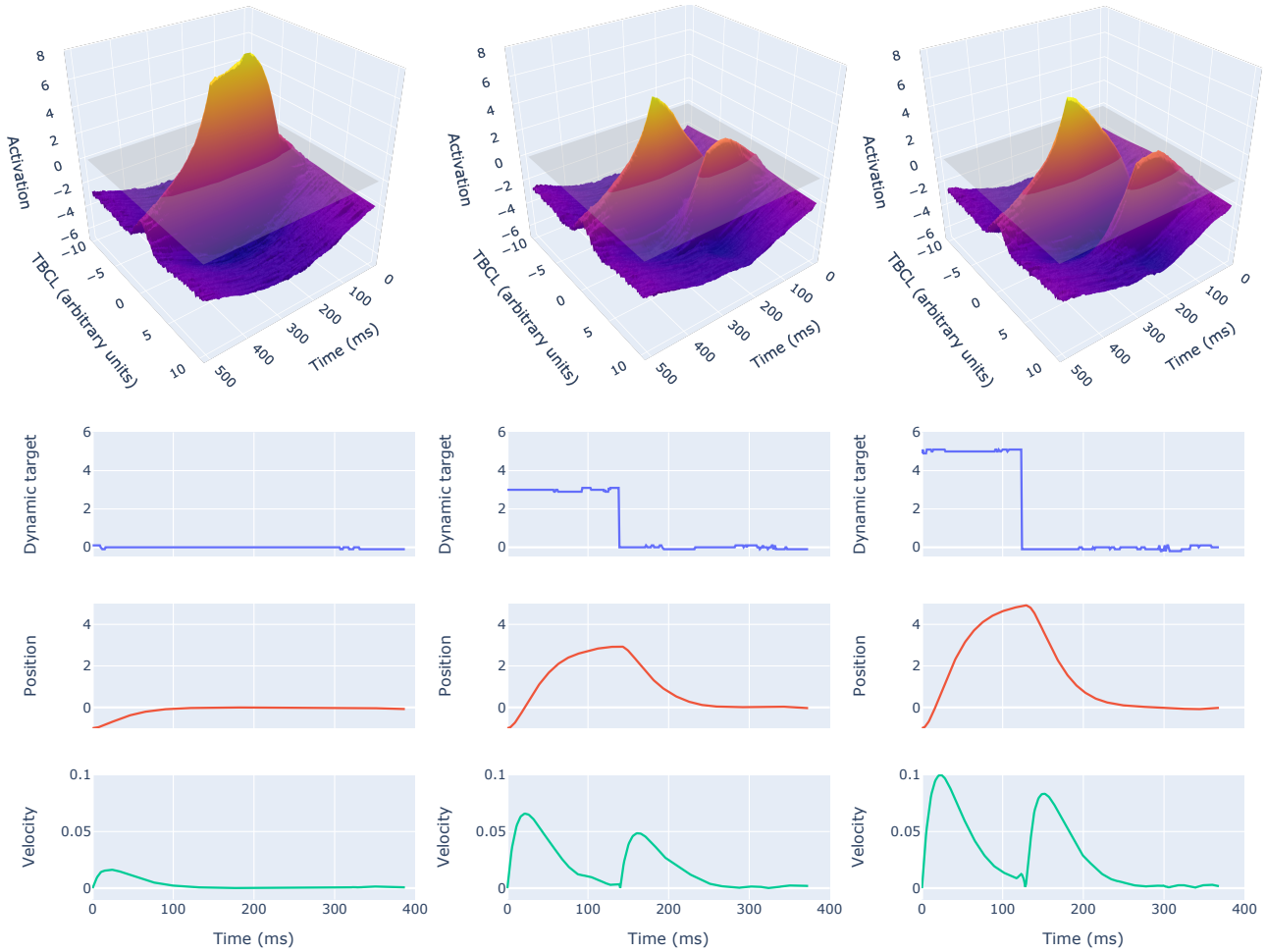

Figure 1: *ROW 1: DNFs for three vowels with increasing degrees of diphthongisation from left-to-right (grey plane = threshold). ROW 2: time-varying location of peak activation in each DNF. ROWS 3 & 4: Task dynamic simulation based on dynamic targets from each DNF. Time in rows 2–4 corresponds to the interval between the onset and offset of supra-threshold activation in each field.*

3.2. Production-perception model

We now present a model of how a speaker's phonological planning representation could undergo change from a long monophthong to a diphthong. This is a highly simplified model of production-perception inspired by Kirov and Gafos (2007) in which two speakers (A and B) interact. Specifically, speaker A produces a long monophthong with two identical targets. They then perceive speaker B producing the same vowel, but with a slightly different phonetic target for the nucleus. This represents perceptual input to speaker A's DNF, which changes their memory trace for the next production to a minor degree. This process repeats, with speaker A producing a vowel, perceiving speaker B's vowel, and so on. This is obviously a highly idealised model of interaction, as the influence is unidirectional (speaker B influences speaker A, but speaker A does not influence speaker B) and the only variation in speaker B's production is due to the addition of random noise added to their target value.

A long vowel is comprised of two inputs: $s_{nuc}(x,t)$ and $s_{glide}(x,t)$. We keep $s_{glide}(x,t)$ constant across production-perception loops, but vary $s_{nuc}(x,t)$ according to (7), where α and γ are weights for the respective task and perceptual inputs. Nucleus and glide both begin with $p = 0, w = 0.7$, with input amplitudes of $a = 3$ (nucleus) and $a = 6$ (offglide). Across production-perception loops, the current $s^i_{nuc}(x,t)$ is:

$$s^i_{nuc}(x,t) = \alpha s^{i-1}_{nuc}(x,t) + \gamma s_{perception}(x,t) \quad (7)$$

$s_{perception}(x,t)$ is defined as in equation (2) for $s(x,t)$, with $a = 0.3$, $w = 0.7$, except p is calculated as:

$$p = \arg\max_x u(x,t) + bias + q\xi \quad (8)$$

where $\arg\max_x u(x,t)$ is the TBCL parameter corresponding to the location of maximum activation (sampled at t = 100), $bias$ is a numerical value representing the difference between speaker A's target and the perceived phonetic target from speaker B (here $bias$ = 1.5), and q is a weighting factor that scales Gaussian noise ξ in the range [0,1]. The task input $s_{nuc}(x,t)$ is weighted by $\alpha = 0.99$, representing very slow memory decay, and the $s_{perception}(x,t)$ input is weighted by $\gamma = 0.2$. Higher values of γ increase the influence of the perceptual input, resulting in faster change over repeated loops.

The production-perception loop was run for 150 iterations and the resulting activation distributions at several iteration

steps are shown in Figure 2. After a number of interactions with this 'biased' speaker B, speaker A's activation field for the nucleus shifts away from the initial state towards a new peak. Notably, the offglide peak does not change very much at all, showing that this target remains stable. The nucleus, however, undergoes change, with the resulting vowel being gradually more diphthongal because the centroid of the nucleus distribution increasingly diverges from the offglide as the iterations increase.

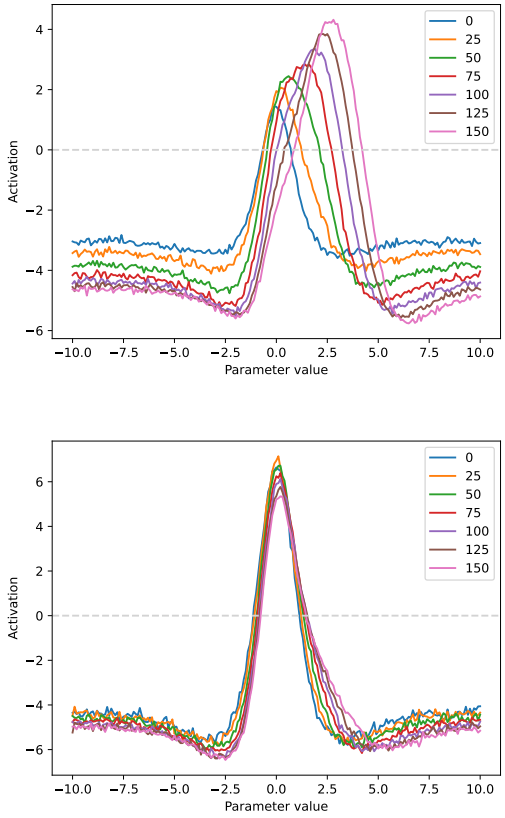

Figure 2: *Activation distributions at selected steps of the production-perception loops for nucleus target sampled at* $t = 100$ *(top) and offglide target at* $t = 300$ *(bottom).*

4. Discussion and conclusion

In summary, we model gestural selection, activation and articulatory dynamics using a combination of dynamic field theory, coupled oscillators and task dynamic models. This allows us to pose specific mechanistic connections between different components of the model, which yields behaviourally-realistic articulatory trajectories for long monophthongs and diphthongs, grounded in neurally-plausible dynamical mechanisms. We also use the same mathematical and conceptual language to propose a mechanism for variation and change in the phonological representations of individual speakers, thereby identifying a clear link between short-term synchronic variation and medium-term change in the diphthongisation of long vowels.

In terms of future research, the model assumes that gestural parameters, such as TBCL, are directly retrievable in perception. While speakers can undoubtably infer articulatory gestures from acoustics, the mapping is unlikely to be linear or perfect and a more realistic model requires a perceptual-acoustic field that projects to a tract variable field. Second, our model of between-speaker interactions is highly idealised and our future research aims to develop more complex models of interaction between small groups of speakers. Finally, while our DNF claims to be a neural model, we make no claims about cortical or subcortical localisation. Instead, the DNF is an abstraction that models the functional behaviour of a neural population, which may actually be distributed over different areas of the brain (Schöner, Spencer, and The DFT Research Group 2016).

5. Acknowledgements

This research was supported by Arts and Humanities Research Council grants AH/S011900/1 and AH/Y002822/1.

6. References

Harper, Sarah Kolin (2021). "Individual differences in phonetic variability and phonological representation". PhD thesis. Los Angeles, CA: University of Southern California.

Hughes, Arthur, Peter Trudgill, and Dominic Watt, eds. (2012). *English Accents and Dialects: An Introduction to Social and Regional Varieties of English in the British Isles*. Fifth. London: Hodder.

Iskarous, Khalil (2017). "The relation between the continuous and the discrete: A note on the first principles of speech dynamics". In: *Journal of Phonetics* 64, pp. 8–20.

Jespersen, Otto (1909). *A Modern English Grammar on Historical Principles*. London: George Allen & Unwin Ltd.

Kirov, Christo and Adamantios I. Gafos (2007). "Dynamic phonetic detail in lexical representations". In: *Proceedings of the 16th International Congress of Phonetic Sciences*, pp. 637–640.

Labov, William, Sharon Ash, and Charles Boberg (2006). *The Atlas of North American English*. Berlin: Mouton de Gruyter.

Popescu, Anisia and Ioana Chitoran (2022). "Linking gestural representations to syllable count judgements: A cross language test". In: *Laboratory Phonology* 13.1, pp. 1–48.

Roon, Kevin D. and Adamantios I. Gafos (2016). "Perceiving while producing: Modeling the dynamics of phonological planning". In: *Journal of Memory and Language* 89.2, pp. 222–243.

Saltzman, Elliot and Kevin G. Munhall (1989). "A dynamical approach to gestural patterning in speech production". In: *Ecological Psychology* 1.4, pp. 333–382.

Schöner, Gregor, John P. Spencer, and The DFT Research Group (2016). *Dynamic Thinking: A Primer on Dynamic Field Theory*. Oxford: Oxford University Press.

Shaw, Jason A. and Kevin Tang (2023). "A dynamic neural field model of leaky prosody: proof of concept". In: *Proceedings of the Annual Meeting on Phonology 2022*, pp. 1–12.

Stern, Michael C. and Jason A. Shaw (2023). "Neural inhibition during speech planning contributes to contrastive hyperarticulation". In: *Journal of Memory and Language* 132.104443, pp. 1–16.

Strycharczuk, Patrycja, Sam Kirkham, Emily Gorman, and Takayuki Nagamine (submitted). "Towards a dynamical model of English vowels: Evidence from diphthongisation". In.

Tilsen, Sam (2007). "Vowel-to-vowel coarticulation and dissimilation in phonemic-response priming". In: *UC Berkeley Phonology Lab Annual Report* 3.1, pp. 416–458.

— (2018). "Three mechanisms for modeling articulation: selection, coordination, and intention". In: *Cornell Working Papers in Phonetics and Phonology*, pp. 1–49.

— (2019). "Motoric mechanisms for the emergence of non-local phonological patterns". In: *Frontiers in Psychology* 10, pp. 1–25.