Phonetic accommodation and inhibition in a dynamic neural field model

Sam Kirkham (s.kirkham@lancaster.ac.uk)
Linguistics and English Language, Lancaster University, UK

Patrycja Strycharczuk (patrycja.strycharczuk@manchester.ac.uk)
Linguistics and English Language, University of Manchester, UK

Rob Davies (r.l.davies@lancaster.ac.uk)
Linguistics and English Language, Lancaster University, UK

Danielle Welburn (dani_welburn@hotmail.com)
Linguistics and English Language, Lancaster University, UK

Abstract

Short-term phonetic accommodation is a fundamental driver behind accent change, but how does real-time input from another speaker's voice shape the speech planning representations of an interlocutor? We advance a computational model of change in speech planning representations during phonetic accommodation, grounded in dynamic neural field equations for movement planning and memory dynamics. A dual-layer planning/memory field predicts that convergence to a model talker on one trial can trigger divergence on subsequent trials, due to a delayed inhibitory effect in the more slowly evolving memory field. The model's predictions are compared with empirical patterns of accommodation from an experimental pilot study. We show that observed empirical phenomena may correspond to variation in the magnitude of inhibitory memory dynamics, which could reflect resistance to accommodation due to phonological and/or sociolinguistic pressures. We discuss the implications of these results for the relations between short-term phonetic accommodation and sound change.

Keywords: phonetic accommodation; shadowing task; neural dynamics; computational modelling; dynamical systems

Introduction

Spoken language is rarely static; when two people converse, they subtly (and sometimes not so subtly) modulate their voices in response to one another. This *phonetic accommodation* is a fundamental characteristic of spoken interaction, representing the ebb and flow of human communication (Giles, 1973; Pardo, 2006). Experimental evidence for phonetic accommodation comes from the shadowing task paradigm, which involves a speaker reading a series of words, followed by imitating a (usually pre-recorded) model talker producing the same words (Goldinger, 1998). This allows for the assessment of how much change occurs as a result of shadowing the model talker, while a post-test recording of the same words can be used to establish persistence of any adaptations. There is considerable evidence that speakers converge towards a model talker in the shadowing task and that listeners are also sensitive to these short-term changes (Goldinger, 1998; Namy et al., 2002; Shockey et al., 2004; Tilsen, 2009). While accommodation is often cast as an automatic process (Goldinger, 1998), Babel (2012) finds that accommodation is socially-mediated, with the degree of adaptation based on the perceived social characteristics of the model talker.

Phonetic accommodation is short-term, but its accumulation over time is a key driver of accent change. Accent change generally slows significantly during adulthood, but change nonetheless can occur, especially with increased exposure to different accents (Evans & Iverson, 2007; Harrington et al., 2019). Exemplar models of speech processing hypothesize a mechanism behind these changes, such that the phonetic representations used in speech production are influenced by stored instances of language from speech perception (Gubian et al., 2023; Johnson, 2007; Pierrehumbert, 2002). In this sense, hearing another talker creates an episodic memory trace, which exerts a small bias on subsequent speech production. It stands to reason that over many such instances, small accent changes could occur, and this process of change would become accelerated if spread across a community.

In this study we advance a computational model of phonetic imitation, where imitation-based shadowing is used as an experimentally-constrained proxy for phonetic accommodation. Previous approaches are largely exemplar-based; e.g. Goldinger (1998) models data from a shadowing task using Minerva2 (Hintzman, 1984). We take inspiration from this work, but advance an alternative dynamic neural field model (Schöner et al., 2016), which incorporates exemplar-like dynamics using a biophysically-inspired account of perception, action and memory (Gafos, 2006; Tilsen, 2009). The motivation for this approach is that complex motor synergies are hypothesized to represent the locus of speech planning (Fowler, 1980; Kelso et al., 1986), and dynamic field models are well-developed for auditory-motor dynamics underlying speech production and perception (Gafos, 2006; Harper, 2021; Roon & Gafos, 2016; Stern & Shaw, 2023; Tilsen, 2019).

Our aim in this study is to examine excitatory and inhibitory dynamics in phonetic accommodation, inspired by previous work on response priming (Tilsen, 2007; Roon & Gafos, 2016). We investigate how shadowing a model talker can lead to convergence on one trial, but propose a novel mechanism for observed divergence on subsequent trials. Rather than arising from online selective inhibition (Houghton & Tipper, 1996), divergence may reflect a delayed inhibitory effect in a coupled memory field, which we implement as a dual-layer model. As such, convergence during one trial can induce suppression in memory, leading to repulsion on subsequent trials even in the absence of an external input.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

Neural field model of phonetic accommodation

Our candidate model comes from the class of dynamic field models of movement planning (Erlhagen & Schöner, 2002), which are inspired by a long history of research in synergetics, self-organization and neural information processing (e.g. Amari, 1977; Grossberg, 1980; Haken, 1977; Kelso, 1995). A dynamic neural field (DNF) functionally represents a neural population that is sensitive to a perceptual or movement parameter dimension. A DNF's evolution is shaped by inputs to the field, such as perceptual and task-related input, as well as memory dynamics and field interactions, such as self-excitation. DNFs are relatively well developed as models of the neural dynamics underpinning speech planning, execution and perception (e.g. Gafos, 2006; Kirkham & Strycharczuk, 2024; Roon & Gafos, 2016; Stern & Shaw, 2023; Tilsen, 2007, 2019), and see Schöner et al. (2016) for an introduction and different applications across the cognitive sciences.

Model architecture

We now outline a minimal model architecture for phonetic planning and memory during short-term phonetic accommodation. A dynamic neural field (DNF) evolves according to the Amari (1977) model that underpins Equation (1):

$$\begin{aligned}\tau\dot{u}(x,t) = &-u(x,t)+h+c_{\text{memory}}u_{\text{memory}}(x,t)\\ &+c_{\text{auditory}}s_{\text{auditory}}(x,t)+c_{\text{response}}s_{\text{response}}(x,t)\\ &+\int k(x-x')g(u(x',t))dx'+q\xi(x,t)\end{aligned} \quad (1)$$

where 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 coefficient q (Schöner et al., 2016). As we are dealing with acoustic measurements, we assume that x represents a one-dimensional reduction of the $F1 \sim F2$ acoustic feature space, but a more realistic model could capture the coupling between acoustic, perceptual and articulatory representations using a multi-layer model.

Inputs $s(x,t)$ are Gaussian distributions over a parameter x with amplitude a, centroid p and width w,

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

Response input $s_{\text{response}}(x,t)$ represents retrieval of a speech planning representation in response to the experiment's visual prompt and is weighted by c_{response}. We model this as retrieval of the appropriate representation from long-term memory (Roon & Gafos, 2016). Auditory input $s_{\text{auditory}}(x,t)$ is an auditory-perceptual input that couples the model talker's production to activation dynamics with strength c_{auditory}. We here treat c_{auditory} as capturing the degree of attention to the incoming speech, which we expect is very high during a shadowing task, but lower in normal conversational interaction. In the shadowing block, auditory input cues the response input, whereas in the non-shadowing block the response is cued by a visual prompt (although we do not explicitly model the visual cue in this study).

The interaction kernel $k(x-x')$ in (3) specifies excitatory and inhibitory forces across the activation field.

$$\begin{aligned}k(x-x') = &\frac{c_{\text{excite}}}{\sqrt{2}\pi\sigma_{\text{excite}}}\exp\left[-\frac{(x-x')^2}{2\sigma^2_{\text{excite}}}\right]\\ &-\frac{c_{\text{inhibit}}}{\sqrt{2}\pi\sigma_{\text{inhibit}}}\exp\left[-\frac{(x-x')^2}{2\sigma^2_{\text{inhibit}}}\right]-c_{\text{global}}\end{aligned} \quad (3)$$

The interaction kernel is gated by a sigmoidal function

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

where β is the slope of the sigmoid and α is a threshold value of u. Each field location only contributes to above-threshold activation when it exceeds a threshold of $u = \alpha$, where typically $\alpha = 0$. Interaction generates local excitation and lateral inhibition, meaning that activation close to an input's centroid will be excited, whereas more distal activation will be inhibited. c_{excite}, c_{inhibit} and σ_{excite}, σ_{inhibit} are the mean and standard deviation of the excitatory/inhibitory components, and c_{global} is a global inhibition constant.

The interaction kernel is a key part of our model, because any new inputs that are very close to the current activation peak will excite that location, causing activation to drift towards the input location, resulting in a compromise value between the speaker's planned target and the perceived target from the model talker (Erlhagen & Schöner, 2002). However, inhibitory forces mean that some inputs may cause dissimilation, causing activation to drift *away* from the planned targets in a direction opposite the model talker's input (Tilsen, 2007).

Short-term memory dynamics are achieved by a Hebbian layer (Samuelson et al., 2011). This is represented by the memory field $u_{\text{memory}}(x,t)$ in (5) coupled to $u(x,t)$ with strength c_{memory} and is subject to local interactions $w(x-x')$.

$$\dot{u}_{\text{mem}}(x,t) = \begin{cases}\frac{1}{\tau_{\text{mem}}}[-u_{\text{mem}}(x,t) \\ +\int w(x-x')g(u(x',t))dx'], & g(u) > \alpha\\ \\ \frac{1}{\tau_{\text{decay}}}[-u_{\text{mem}}(x,t)], & g(u) \le \alpha\end{cases} \quad (5)$$

When activation in the online planning field $u(x,t)$ is greater than threshold α the sigmoid $g(u)$ gates activation into the memory field at a rate determined by τ_{mem}. When $g(u) \le \alpha$ memory at those field locations undergoes decay at a rate of τ_{decay}. The memory field evolves on a slower timescale than the field dynamics, while memory decay evolves the slowest, such that $\tau_{\text{decay}} > \tau_{\text{mem}} > \tau$. This reflects the fact that memory formation happens faster than memory decay. This suggests that even if a speaker's current production is shifted in the direction of the model talker, the concomitant effects on

the memory field will be relatively small. As such, we predict that any post-shadowing convergence or divergence will be minimal over a small number of trials.

Note that in our simulations we use an interaction kernel for both the main parameter field $k(x-x')$ and the memory field $w(x-x')$. The memory kernel is specified for local inhibition but not global inhibition, and we hypothesize that the memory field may have different interaction dynamics due to learning patterns and longer-term attentional dynamics. For example, previous research suggests that phonetic accommodation varies between different vowels (Evans & Iverson, 2007; Babel, 2012). While in some cases this can be partly explained due to the distance between a field's current activation pattern and the model talker's input, there are cases where speakers converge during shadowing and diverge post-shadowing, which varies between vowels and speakers. We hypothesize that this could represent different patterns of lateral inhibition in *memory* for different vowels, which may be a consequence of phonological, perceptual or sociolinguistic pressures. Note that we locate any such differences in the memory field and not in the online planning field. This predicts that speakers will typically converge towards the model talker, but may vary in the post-shadowing response depending on the current state of the memory field.

Simulating experimental trials

A simulated interaction proceeds as follows in three blocks.

1. **Baseline**. The visual prompt cues $s_{\text{response}}(x,t)$ input, which raises activation above threshold and triggers production at the parameter value corresponding to peak activation. We assume $s_{\text{auditory}}(x,t) = 0$ during the baseline block, meaning it has no influence on activation. The field dynamics leave an activation trace in $u_{\text{memory}}(x,t)$.

2. **Shadowing**. Auditory input from the model talker $s_{\text{auditory}}(x,t) > 0$ raises sub-threshold activation in $u(x,t)$ and the response input $s_{\text{response}}(x,t)$ subsequently raises activation above threshold and production occurs. High attention to the model talker reflected in a large c_{auditory} value means that input amplitude is high, which causes its effects to persist over time. These dynamics leave an activation trace in the updated $u_{\text{memory}}(x,t)$ field.

3. **Post-shadowing**. The visual prompt cues $s_{\text{response}}(x,t)$, while $s_{\text{auditory}}(x,t) = 0$, which raises activation above threshold, cues production, and leaves a memory trace.

Note that the memory field $u_{\text{memory}}(x,t)$ is active on each trial and shapes the evolution of the planning field based on its current state, which is also updated during each trial.

All simulations were implemented in Python using NumPy (Harris et al., 2020) and SciPy (Virtanen et al., 2020), with visualizations made using Matplotlib (Hunter, 2007). Numerical solutions were calculated using an Explicit Runge-Kutta method of order 5(4) via SciPy's integrate.solve_ivp function.

A pilot experiment on phonetic accommodation

Experiment design

We also report an experimental pilot study of phonetic accommodation in a shadowing task, where speakers of Northern Anglo British English shadow a model talker with an accent different from their own. Due to the small sample size, our experiment is not an empirical assessment of the facts behind phonetic accommodation, but is instead used to generate plausible empirical scenarios. We subsequently attempt to model these empirical scenarios in order to generate hypothesized mechanisms behind patterns of accommodation.

The experiment featured three blocks: pre-test, shadowing, post-test (Babel, 2010; Goldinger, 1998). In the pre-test and post-test blocks, speakers read aloud single hVd words along with a set of target words. The shadowing block required speakers to identify and repeat the same target words produced by a model talker, who was a male Standard Southern British English speaker aged 21. We specifically focus on two vowels that differ substantially between the participants and model talker: BATH and STRUT. Standard Southern British English realizes these vowels as [ɑ] and [ʌ] respectively, while in almost all varieties of Northern Anglo English these vowels are produced as [a] and [ʊ]. These vowels represent the most characteristic difference between northern and southern varieties in England (Wells, 1982); they are highly salient to listeners and can also undergo change as a consequence of long-term different-accent exposure (Evans & Iverson, 2007). The BATH target words were *bath, chance, fast, mast, staff*, while the STRUT words were *strut, bust, chuck, fun, mud*. Another 10 non-BATH/STRUT words were also included in all experimental blocks as distractor stimuli.

Participants and recording

All participants were first-language speakers of Northern Anglo English, aged between 19–22 years old. 18 speakers completed the experiment, with 13 speakers (11 female, 2 male) included in the analysis (two speakers were removed as they recognized the model talker, and three speakers were removed due to significant distortion in the audio recordings). The experiment was adminstered using PsychoPy (Peirce et al., 2019), with audio recorded in a sound-attenuated booth at 44.1 kHz using a Beyerdynamic Opus 55 headset microphone (5cm from the mouth), pre-amplified and digitized using a Sound Devices USBPre 2 audio interface. Audio stimuli were delivered using Beyerdynamic DT 770 headphones.

Data processing

Recordings were force-aligned using Montreal Forced Aligner (McAuliffe et al., 2017) and formant estimation was optimized using the FastTrack algorithm (Barreda, 2021; Fruehwald & Barreda, 2023), with a 20-step search window of 4000–7000 Hz, 25 ms window length, 2 ms step size, 5th-order DCT smoothing. Formant values were extracted from vowel midpoints and by-speaker z-scored across hVd and target words. The degree of accommodation was quantified by

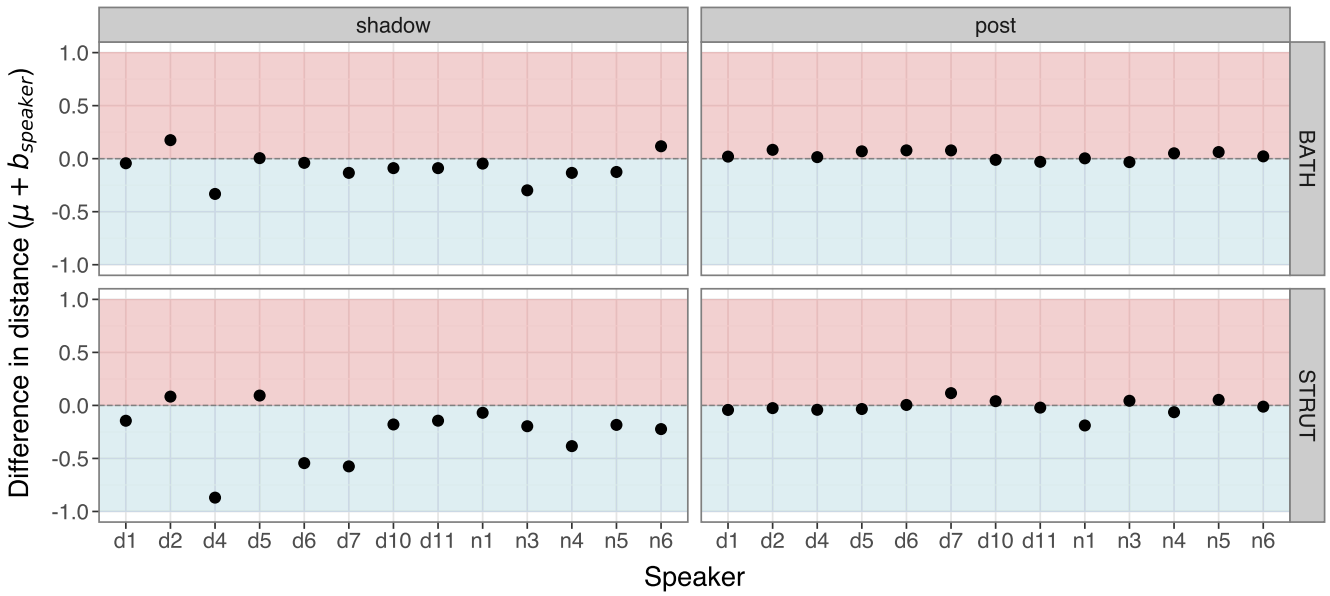

Figure 1: By-speaker difference in distance values for vowel and block. Values are from the Bayesian model and represent each speaker's random intercept coefficient added to the model's grand intercept. Negative values indicate convergence towards the model talker (blue shading); positive values indicate divergence (red shading); zero values indicate no accommodation. Speaker labels (d/n) refer to data collected by different experimenters and do not reflect any differences in speaker characteristics.

calculating the Euclidean distance d between each speaker's production and the corresponding model talker production.

We subtract the Euclidean distance value for the baseline block from the shadowing and post-test blocks in order to calculate a 'difference in distance' metric that captures the degree of accommodation (Babel, 2012). Values of zero indicate no accommodation, negative values indicate convergence (reduced distance from model talker), and positive values indicate divergence (increased distance from model talker). All data analysis was carried out in the Python programming language, using the packages NumPy (Harris et al., 2020), Pandas (The pandas development team, 2020) and plotnine (The plotnine development team, 2025).

Results

We fit four Bayesian linear mixed models to the difference in distance (d) values for each combination of task (shadowing/post) and vowel (BATH/STRUT). Each model contains a grand intercept μ and random intercepts b_{word} and b_{speaker}. Models were written in Stan and run using cmdstanpy (Stan Development Team, 2024) with weakly informative priors $\mu \sim \mathcal{N}(0,2)$ and $b \sim \mathcal{N}(0,1)$, using 4 MCMC chains, 500 warmup iterations, and 2000 sampling iterations. Speakers accommodate to the model talker during shadowing to a greater extent for STRUT ($\bar{d} = -0.25$, 95% CI $[-0.60, 0.13]$) than BATH ($\bar{d} = -0.08$, 95% CI $[-0.30, 0.15]$). The vowels differ only minimally post-shadowing, with BATH slightly diverging from the model talker post-shadowing ($\bar{d} = 0.03$, 95% CI $[-0.22, 0.25]$), while STRUT is closer to the pre-shadowing baseline ($\bar{d} = 0.01$, 95% CI $[-0.19, 0.16]$). Note that in all cases the wide credible intervals point towards extensive between-speaker variation and do not support vowel-specific differences on a group level.

Figure 1 shows speaker-specific difference in distance values. These values represent the fitted by-speaker random intercepts, added to the model's grand intercept, which provides an estimate of each speaker's difference in distance value. The plot reflects some convergence in BATH, with the majority of speakers below the zero line, followed by the majority of speakers showing a small amount of divergence post-shadowing. However, some speakers do show small divergence during shadowing, such as d2 and n6, but all speakers are very close to baseline or above post-shadowing. The STRUT data is more variable during shadowing, with some speakers showing substantial convergence to the model talker (e.g. d4, d6, d7). Two speakers diverge during shadowing for STRUT (d2, d5), one of whom also diverged for BATH (d2). The post-shadowing STRUT data shows strong clustering around the zero line, indicating a return to the baseline production. Speaker n1 is the only one who is slightly closer to the model talker post-shadowing than during shadowing, but these differences remain small.

Summary and next steps

The overall picture from this sample is considerable variability in accommodation. Both vowels show minor amounts of divergence in post-shadowing, with BATH showing slightly greater divergence on a speaker-specific level. While our results are insufficient to claim a robust vowel-specific effect, Evans & Iverson (2007) report a longitudinal study in which

northern speakers converge to SSBE STRUT to a greater extent than BATH over a period of two years, which the authors suggest is due to the greater salience of BATH in northern English. While our data do not support a group-level vowel effect, there are certainly individual speakers who follow this pattern. This points to a more generic observation that convergence to a model talker can potentially result in a subsequent return to baseline *or* subsequent divergence. In the following section, we use our computational model to explore the mechanisms that could generate these two phenomena.

Simulating interactional scenarios

Motivations and approach

We now use our model to replicate two potential observations: (1) convergence followed by return to baseline; (2) convergence followed by divergence. Our primary interest is identifying which mechanisms in our model are required to capture these patterns. We refer to these cases as STRUT (return to baseline) and BATH (divergence) as this is the trend in the literature, but these should be taken as more general examples that are within the model's scope. To provide some empirical validity to the simulations, we focus on modelling the small vowel-specific effects from the empirical data. While these are very small, we view this as preferable to modelling potentially larger effects that are not observed in our data and may therefore be unrealistic.

For all simulations we centre the idealized speaker's memory trace at zero across a field of $x \in [-10, +10]$ (the majority of the field either side of zero is subject to significant inhibition, so it is unlikely that such areas receive any significant activation). The inputs $s_{\text{auditory}}(x,t)$ are based on the average z-scored distance from the model talker, with STRUT $= -1.4$ and BATH $= -1.2$. These input values represent differences from the idealized speaker's existing representation. The planning field interaction kernel $k(x-x')$ is defined as $c_{\text{excite}} = 2, \sigma_{\text{excite}} = 0.2, c_{\text{inhibit}} = 1, \sigma_{\text{inhibit}} = 2, c_{\text{global}} = 0.5$. The default memory kernel $w(x-x')$ is identical to the field kernel, except $c_{\text{global}} = 0$ and $\sigma_{\text{excite}} = 0.1$. Temporal parameters are $\tau = 25$, $\tau_{\text{memory}} = 150$, $\tau_{\text{decay}} = 500$, while $c_{\text{memory}} = 10$, $c_{\text{auditory}} = 10$, $c_{\text{response}} = 1$, $h = -2$, $q = 3$, $\beta = 1.5$. All inputs $s(x,t)$ have $a = 10$ and $w = 0.5$.

All simulations lasted for a duration of 300 ms. Inputs are constant over time because we make the assumption that monophthongs are one-target vowels (Strycharczuk et al., 2024). To represent the acoustic parameter selected for speech production, we sample at the time-step corresponding to peak activation, with the x-location of peak activation representing the selected parameter value for production.

We first initialize short-term memory $u_{\text{memory}}(x,t)$ as a zero-valued flat field and then run simulations with a single input $a = 100, p = 0, w = 0.5$, which when coupled to the memory field serves to update $u_{\text{memory}}(x,t)$ based on the resulting field activation. This represents the existing short-term memory for each vowel, while the response input $s_{\text{response}}(x,t)$ is drawn from longer-term phonological memory. For each simulation, the initial memory state is the memory state at the end of the previous simulation.

Phonetic convergence and return to baseline

The model straightforwardly captures the dynamics of phonetic convergence followed by return to near-baseline. In the STRUT vowel simulations, peak activation is at $x = -0.22$ during shadowing and $x = 0.02$ during post-shadowing. This is close to the empirical means of $\bar{d} = -0.25$ and $\bar{d} = 0.01$, with a relative difference between shadowing and post-shadowing of $\bar{d}_{\text{diff}} = 0.26$ and $x_{\text{diff}} = 0.24$, which represents good agreement between model and data.

The occurrence of accommodation during shadowing versus return to (near) baseline post-shadowing is a consequence of the auditory input and inhibitory dynamics. This is shown in Figure 2 (top), where the small bump at the left of the activation field during shadowing (orange line) represents the effects of $s_{\text{auditory}}(x,t)$. While this input does not reach the activation threshold of 0, it does slightly pull the activation centroid leftwards, resulting in the small degree of observed accommodation towards the model talker. The degree of accommodation is attenuated as the input occurs in a region of parameter space that is subject to considerable inhibition (i.e. the negative values near the base of the primary activation peak). This small amount of accommodation has a very minimal effect on the memory field in Figure 2 (bottom), with an almost undetectable rightwards shift of the memory peak as a consequence of greater inhibition in short-term memory. The memory field evolves more slowly than the parameter activation field, meaning that any production effects are very gradual. Note that while these effects are very small, they reflect the average empirical changes in speech production as a consequence of minimal short-term exposure to the model talker.

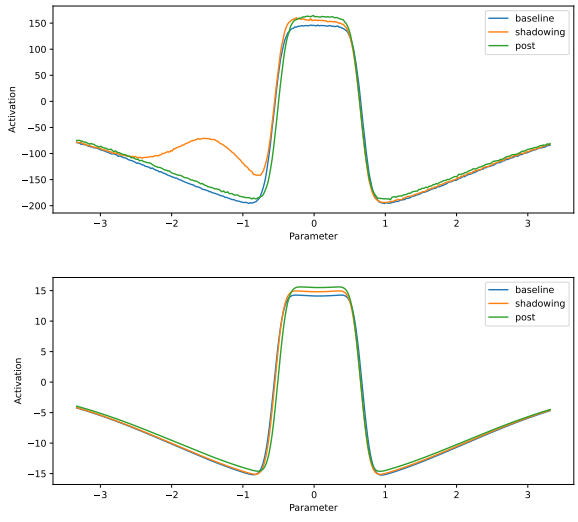

Figure 2: Activation field (top) and memory field (bottom) for STRUT simulations. The x parameter range has been truncated to highlight small differences in the activation peaks.

Phonetic convergence followed by divergence

We now examine how our model can reproduce the general effect of convergence (shadowing) followed by divergence (post-shadowing). We specifically model the small average effect for BATH, but note that some speakers diverge to a much greater extent than this. In doing so, we turn to the memory kernel, which situates the differences between BATH and STRUT in vowel-specific memories, rather than purely metric parameter differences. We run the same simulation as for STRUT but with three changes. First, $s_{\text{auditory}}(x,t)$ has $p = -1.2$ rather than $p = -1.4$ to reflect the empirical baseline distance from the model talker for BATH. Second, the memory kernel has higher local inhibition, with $c_{\text{inhibit}} = 1.8$ (compared to $c_{\text{inhibit}} = 1$ for STRUT). Third, the memory kernel also has higher $\sigma_{\text{inhibit}} = 3$ (compared with $\sigma_{\text{inhibit}} = 2$ for STRUT). This specifies the memory kernel for BATH as having stronger/wider local inhibition, which corresponds to the memory field's increased resistance in this region. This is in line with previous literature showing that BATH is more resistant to accommodation than STRUT for northern speakers.

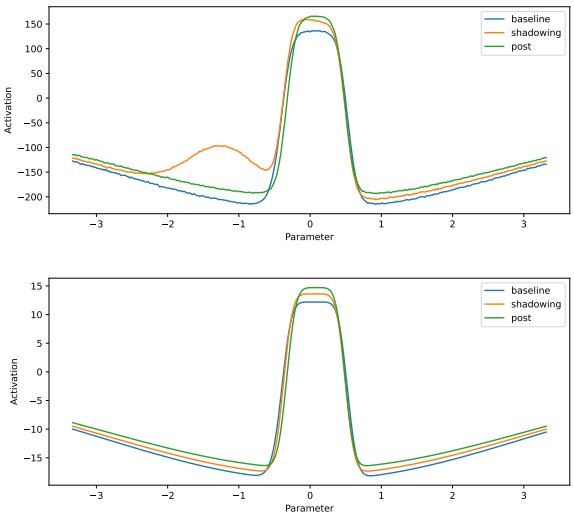

Figure 3: Activation field (top) and memory field (bottom) for BATH simulations. The x parameter range has been truncated to highlight small differences in the activation peaks.

The resulting simulations for BATH show peak activation at $x = -0.1$ for shadowing and $x = 0.02$ for post-shadowing, which is close to the mean empirical magnitude of accommodation ($\bar{d} = -0.08$) and divergence ($\bar{d} = 0.03$). If we take the relative difference between shadowing and post-shadowing then $\bar{d}_{\text{diff}} = 0.11$ and $x_{\text{diff}} = 0.12$, showing good agreement between model and data. This can be seen in Figure 3, where sub-threshold auditory input peak on the left-hand side (orange) pulls the activation peak slightly leftwards, representing accommodation, but inhibitory memory dynamics repel this slightly and the subsequent post-shadowing production is shifted slightly rightwards, representing divergence. Note that the differences between conditions in the memory trace are slightly larger than those in Figure 2, as a consequence of stronger inhibitory dynamics in the memory field for BATH.

Discussion

We proposed a dynamic neural field model of phonetic accommodation that qualitatively captures some observed empirical phenomena. While the majority of speakers show some convergence and a return to baseline, some speakers converge during shadowing and then diverge post-shadowing, resulting in a greater distance from the model talker *after* shadowing. This cannot be straightforwardly modelled using a single-layer field, which leads us to propose a delayed inhibitory effect in memory, due to the different temporal scales of online planning and short-term memory. Indeed, previous resarch suggests that inhibition may be part of the long-term memory of a gesture (Tilsen, 2007) and we show that modelling these vowel differences in the memory field, rather than the planning field, exposes a potential mechanism.

Previous research finds vowel-specific differences in accommodation; while this can be explained by inhibitory differences, why should inhibitory dynamics vary between vowels? One rationale is that changes in BATH have structural implications for northern speakers, whose vowel in the PALM/START lexical sets is phonetically similar to SSBE BATH. Convergence would lead to potential merger between vowel categories, so greater inhibition may prevent category merger. Additionally, the BATH vowel is a strong shibboleth of the north/south divide (Wells, 1982) and northern speakers tend to resist change in this vowel (Evans & Iverson, 2007). Our empirical data shows that inhibitory dynamics are not sufficient to completely block accommodation, suggesting an automatic dimension to accommodation (Goldinger, 1998). However, inhibitory dynamics attenuate the magnitude of accommodation and trigger divergence in short-term memory, which bolsters the maintenance of this salient accent feature.

In the present study, we have only modelled the very small average effects from the statistical model, but varying degrees of accommodation and persistence can be modelled through variation in the input weighting (i.e. reflecting attention to the model talker) and in excitatory/inhibitory forces. Variation in these parameters is likely to be a potential locus of speaker-specific variation, which could explain differences in adaptation. For example, reducing memory inhibition for BATH produces a return closer to baseline, rather than dissimilation.

In summary, vowel-specific phonetic accommodation can be modelled as differences in short-term inhibitory memory dynamics, which we hypothesize is motivated by phonological contrast and socially-motivated resistance to change. Vowels with weaker inhibitory dynamics are predicted to undergo greater accommodation, which if repeated over many interactions could lead to sound change. In future work we plan to conduct a more comprehensive experimental study, as well as integrate acoustic-perceptual representations with nonlinear gestural models of articulatory control (Kirkham, 2025a,b; Sorensen & Gafos, 2016; Stern & Shaw, 2024).

Acknowledgments

This research was funded by UKRI/AHRC grant AH/Y002822/1 awarded to SK.

References

Amari, S.-i. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, *27*(2), 77–87.

Babel, M. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society*, *39*(4), 437–456.

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*(1), 177-189.

Barreda, S. (2021). Fast Track: Fast (nearly) automatic formant-tracking using Praat. *Linguistics Vanguard*, *7*(1), 20200051.

Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review*, *109*(3), 545–572.

Evans, B. G., & Iverson, P. (2007). Plasticity in vowel perception and production: A study of accent change in young adults. *Journal of the Acoustical Society of America*, *121*(6), 3814–3826.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, *8*(1), 113–133.

Fruehwald, J., & Barreda, S. (2023). *fasttrackpy.* Zenodo. Retrieved from https://doi.org/10.5281/ZENODO.10212099

Gafos, A. I. (2006). Dynamics in grammar. In L. Goldstein, D. Whalen, & C. T. Best (Eds.), *Laboratory phonology 8: Varieties of phonological competence* (pp. 51–79). Berlin: Mouton de Gruyter.

Giles, H. (1973). Accent mobility: A model and some data. *Anthropological Linguistics*, *15*(2), 87–105.

Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Grossberg, S. (1980). Biological competition: Decision rules, pattern formation, and oscillations. *Proceedings of the National Academy of Sciences*, *77*(4), 2338–2342.

Gubian, M., Cronenberg, J., & Harrington, J. (2023). Phonetic and phonological sound changes in an agent-based model. *Speech Communication*, *147*, 93–115.

Haken, H. (1977). *Synergetics: An introduction*. Berlin: Springer-Verlag.

Harper, S. K. (2021). *Individual differences in phonetic variability and phonological representation*. Unpublished doctoral dissertation, University of Southern California, Los Angeles, CA.

Harrington, J., Gubian, M., Stevens, M., & Schiel, F. (2019). Phonetic change in an Antarctic winter. *Journal of the Acoustical Society of America*, *146*(5), 3327–3332.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362.

Hintzman, D. L. (1984). Minerva 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(1), 96–101.

Houghton, G., & Tipper, S. P. (1996). Inhibitory mechanisms of neural and cognitive control: Applications to selective attention and sequential action. *Brain and Cognition*, *30*(1), 20–43.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.

Johnson, K. (2007). Decisions and mechanisms in exemplar-based phonology. In M.-J. Solé, P. S. Beddor, & M. Ohala (Eds.), *Experimental approaches to phonology* (pp. 25–40). Oxford: Oxford University Press.

Kelso, J. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. Cambridge, MA: MIT Press.

Kelso, J. S., Saltzman, E. L., & Tuller, B. (1986). The dynamical perspective on speech production: data and theory. *Journal of Phonetics*, *14*(1), 29–59.

Kirkham, S. (2025a). Discovering dynamical laws for speech gestures. *Cognitive Science*, *49*(5), e70064.

Kirkham, S. (2025b). Scaling laws for nonlinear dynamical models of articulatory control. *JASA Express Letters*, *5*(2), 1–7.

Kirkham, S., & Strycharczuk, P. (2024). A dynamic neural field model of vowel diphthongisation. *Proc. ISSP 2024 – 13th International Seminar on Speech Production*, 193–196.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proc. Interspeech 2017* (pp. 498–502).

Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology*, *21*(4), 422–432.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, *119*(4), 2382–2393.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203.

Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology 7* (pp. 101–139). Berlin: Mouton de Gruyter.

Roon, K. D., & Gafos, A. I. (2016). Perceiving while producing: Modeling the dynamics of phonological planning. *Journal of Memory and Language*, *89*(2), 222–243.

Samuelson, L. K., Smith, L. B., Perry, L. K., & Spencer, J. P. (2011). Grounding word learning in space. *PLoS ONE*, *6*(12), e28095.

Schöner, G., Spencer, J. P., & The DFT Research Group. (2016). *Dynamic thinking: A primer on dynamic field theory*. Oxford: Oxford University Press.

Shockey, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics*, *66*(3), 422–429.

Sorensen, T., & Gafos, A. I. (2016). The gesture as an autonomous nonlinear dynamical system. *Ecological Psychology*, *28*(4), 188–215.

Stan Development Team. (2024). *Stan Reference Manual, v2.36.0.* https://mc-stan.org.

Stern, M. C., & Shaw, J. A. (2023). Neural inhibition during speech planning contributes to contrastive hyperarticulation. *Journal of Memory and Language*, *132*(104443), 1–16.

Stern, M. C., & Shaw, J. A. (2024). Towards a minimal dynamics for gestures: A law relating velocity and position. *Proc. ISSP 2024 – 13th International Seminar on Speech Production*, 262–265.

Strycharczuk, P., Kirkham, S., Gorman, E., & Nagamine, T. (2024). Towards a dynamical model of English vowels: Evidence from diphthongisation. *Journal of Phonetics*, *107*, 1–26.

The pandas development team. (2020). *pandas-dev/pandas: Pandas.* https://doi.org/10.5281/zenodo.3509134.

The plotnine development team. (2025). *plotnine: A grammar of graphics for Python.* https://doi.org/10.5281/zenodo.1325308.

Tilsen, S. (2007). Vowel-to-vowel coarticulation and dissimilation in phonemic-response priming. *UC Berkeley Phonology Lab Annual Report*, *3*(1), 416–458.

Tilsen, S. (2009). Subphonemic and cross-phonemic priming in vowel shadowing: Evidence for the involvement of exemplars in production. *Journal of Phonetics*, *37*(3), 276–296.

Tilsen, S. (2019). Motoric mechanisms for the emergence of non-local phonological patterns. *Frontiers in Psychology*, *10*(2143), 1–25.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272.

Wells, J. C. (1982). *Accents of English: Volumes 1–3*. Cambridge: Cambridge University Press.