A dynamic neural field model of phonetic trace effects in speech errors

Michael C. Stern* **(michael.stern@yale.edu)**

Manasvi Chaturvedi* **(manasvi.chaturvedi@yale.edu)**

Jason A. Shaw (jason.shaw@yale.edu)

Department of Linguistics, Yale University
New Haven, CT 06511 USA

Abstract

Speech errors are often perceived as categorical substitutions of one sound for another, but phonetic analyses have consistently revealed that errorful productions retain a phonetic trace of the target category. These trace effects have been taken as evidence for the simultaneous activation of multiple categories, both exerting influence on speech production. We develop a dynamic neural field model of voice onset time (VOT) planning, showing how multiple activated categories can be resolved in the field to show trace effects. We evaluate model predictions against measurements of VOT for voiced and voiceless stops in speech error experiments and naturalistic corpora.

Keywords: Dynamic Field Theory; speech production; speech errors; cascading activation; voice onset time

Introduction

Early studies of speech errors, based on impressionistic transcription, argued that speech errors involve categorical substitutions which, despite being errors, nevertheless follow the rules of the grammar (Fromkin, 1971). It is now widely acknowledged that this is often not the case—various studies, looking at speech errors through laboratory tongue twister experiments or naturalistic corpora, have found that speech errors do not simply reflect a change in category. Errorful productions of a sound are systematically different from their canonical, non-errorful counterparts (Alderete et al., 2021; Frisch & Wright, 2002; Goldrick & Blumstein, 2006; Goldstein et al., 2007; McMillan & Corley, 2010; Mowrey & MacKay, 1990; Pouplier & Goldstein, 2010). For example, Goldrick & Blumstein (2006) reported a "trace effect" in errors elicited through a tongue twister experiment—they found that errorful productions of initial stops in tongue twisters showed a trace of the intended sound, which can be seen through voice onset time (VOT) measurements. VOT is the primary phonetic cue differentiating voiced and voiceless stops in English (Lisker & Abramson, 1964); voiced stops, /b/, /d/, /g/, have short VOT (~10 ms), while voiceless stops, /p/, /t/, /k/, have long VOT (~60 ms) (Chodroff & Wilson, 2017). In the tongue twister *keff geff geff keff,* an errorful production of the second syllable as *keff* resulted in a /k/ with a shorter VOT than a canonical production—i.e., the /k/ contained a *trace* of the intended /g/ (Goldrick & Blumstein, 2006).

Goldrick & Blumstein (2006) argue against the hypothesis that trace effects arise solely from articulatory factors, concluding that their findings support a *cascading activation* account of speech production, whereby the activation of competitor word forms cascades down to lower levels of planning, even if the competitor form is less activated than the target form. In an error, the competitor receives *more* activation than the target. However, the target still has some effect on articulation because of its non-zero activation.

In an analysis of podcasts, Alderete et al. (2021) found a similar trace effect in VOT measurements of naturalistic (non-experimentally induced) errors. For example, an errorful production of /b/ when the intended sound was /p/ (e.g., producing *bath* when the intended word was *path*) had a higher VOT than a canonical production of /b/ (when the intended sound was /b/). As in Goldrick & Blumstein (2006), the VOT of errorful productions was shifted towards the *target* sound.

According to the cascading activation account, this is because even when an unintended sound is produced, the *intended* sound still influences its production (Alderete et al., 2021). Thus, trace effects in speech errors arise from the activation of competing categories. In a computational implementation of this proposal, categorical representations of lexical items and phonemes vary continuously in their activation during speech planning, and multiple active representations simultaneously influence continuous phonetic levels of speech planning (Goldrick & Chu, 2014; Smolensky et al., 2014). By assuming that the activation of symbolic representations maps continuously to the temporal duration of articulatory gestures, Goldrick & Chu (2014) are able to derive some aspects of phonetic trace effects, like reduced VOT of errorful voiceless stops. Since VOT is a temporal dimension, it can appropriately be modelled by mapping gradient symbol activations to articulatory duration. On this approach, spatial reduction of articulation can only be achieved indirectly by limiting articulation time (cf. target undershoot: Lindblom, 1963), which has been argued to be insufficient (Pouplier & Goldstein, 2014).

We develop a neural-computational model of VOT planning which generates trace effects in speech errors from gradient activation of competitors during speech planning. In

* Equal contribution

this way, our model is similar to the model discussed above (Goldrick & Chu, 2014). However, unlike the previous model, our model does not adopt the assumption that activation of categories maps solely to articulation time. Rather, our model is generalizable to any dimension relevant to speech planning, spatial or temporal.

Dynamic Field Theory (DFT)

Dynamic Field Theory (DFT: Erlhagen & Schöner, 2002; Schöner et al., 2016) offers a promising framework to account for phonetic trace effects because it allows for categories to interact in a continuous feature space. This contrasts with many symbol-oriented speech production models in which category representations are discrete, even when the *activations* of representations are modeled as continuous (e.g., Dell, 1986; Dell et al., 2021; Levelt et al., 1999). In DFT, features relevant to perception, behavior and cognition are modeled as continuous parameters. Each parameter is represented by a functionally (not necessarily topographically) unified population of neurons. The neurons in a population can be arranged on an axis representing the parameter to which they are sensitive, with the position of each neuron on the axis representing the parameter value that maximizes the spike rate or activation level of that neuron (the peak of the neuron's tuning curve). The *distribution* of activation across the neurons in a population is modeled as a *dynamic neural field* (DNF). DNFs evolve over time under the influence of input (e.g., from sensory surfaces), lateral interactions, and noise. The dynamics of DNFs—described in greater detail in the following section—give rise to qualitative shifts in activity, e.g. from stable resting states to stable peaks of activation corresponding to percepts or movement plans.

An important characteristic of DNFs is their ability to make decisions in the presence of multiple inputs. For instance, in a task requiring a subject to reach in one of two directions, a DNF representing reaching direction might receive two inputs representing the two different reaching directions (the target and distractor). Initially, these inputs will increase activation in two different field locations. Via lateral inhibition, however, only one activation peak will ultimately form, driving the subject to reach in the corresponding direction. Importantly, the activation peak that ultimately forms may exhibit *traces* of the inhibited input. For instance, the direction in which the subject reaches may be pulled slightly in the direction of the distractor, compared to a trial with no distractor (Erlhagen & Schöner, 2002). These trace effects have also been observed and modeled using DNFs in other cognitive domains like motion perception (Giese, 1999) and eye saccade planning (Kopecz & Schöner, 1995). The neural activation distributions corresponding to trace effects in the planning of reaching movements have been observed in rhesus monkey motor cortex (Georgopoulos et al., 1986). We propose that phonetic trace effects in speech errors are similarly the result of the interaction of multiple inputs to DNFs. In this case, the relevant DNFs represent parameters of speech planning (e.g., Gafos & Kirov, 2009; Roon & Gafos, 2016; Tilsen, 2019). In the following sections, we describe a DNF model of VOT planning, and present simulations from the model that exhibit key aspects of phonetic trace effects in voicing errors.

Model structure

In this section we present a DNF model of VOT planning. The state of this DNF is assumed to govern the implementation of VOT (i.e. the temporal coordination of laryngeal and supralaryngeal gestures) with stable peaks of activation in the DNF driving behavioral dynamics. The model is summarized in Eq. 1:

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

The key component of the model is the activation field u defined over the VOT dimension x at each moment in time t. We set the field size, x, to 150, representing a 150 ms range of VOT targets. The rate of change of activation $\dot{u}(x,t)$ is inversely related to current activation $u(x,t)$, so Eq. 1 represents a dynamical system with an attractor at $h + s(x,t) + \int k(x - x')g(u(x',t))dx' + q\xi(x,t)$. τ is a time constant, with higher values corresponding to slower rates of field evolution. The resting level h is assumed to be below zero for all field locations (neurons), by convention at –5. Field input $s(x,t)$ is represented as a Gaussian distribution of the form

$$s(x,t) = a \exp\left[-\frac{(x-p)^2}{2w^2}\right] \quad (2)$$

where a controls the amplitude or strength of the input, p controls the position of the input in the field, and w controls the width of the input distribution. This treatment of speech intentions as *distributions* in feature space is similar to previous conceptualizations of phonetic goals as "ranges" (Byrd & Saltzman, 2003), "windows" (Keating, 1990) or "convex regions" (Guenther, 1995). Each neuron x' which exceeds an activation threshold contributes activation to other neurons x via the interaction kernel $k(x - x')$ given by

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

The effects of both excitatory and inhibitory interaction are modeled as Gaussian distributions centered on each neuron x'. c_{exc} and c_{inh} control the magnitude of excitatory and inhibitory interaction, respectively, and σ_{exc} and σ_{inh} control the width of each interaction distribution. c_{glob} contributes additional across-the-board inhibition from each above-

threshold neuron. In our model, as in most DNF models, $c_{exc}>c_{inh}>c_{glob}$ and $\sigma_{exc}<\sigma_{inh}$, so interaction is excitatory (positive effect on activation) for nearby neurons and inhibitory (negative effect on activation) for more distant neurons. Lateral excitation allows the formation of self-sustained above-threshold activation peaks which drive articulatory movement, while lateral inhibition prevents runaway expansion of activation peaks. The activation threshold for interaction is given by a sigmoidal function $g(u)$:

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

By convention, the threshold is $u = 0$. Finally, noise is simulated by adding normally distributed random values $\xi(x,t)$ weighted by a parameter q. This model, which we use to generate the simulations in the following section, was built using the MATLAB-based software COSIVINA (Schneegans, 2021).

Simulation results

Conditions for activation peak attraction

Using the DNF model of VOT planning described above, we can simulate the effects of simultaneous input from both a voiced category $s_{voiced}(x,t)$ and a voiceless category $s_{voiceless}(x,t)$. In this case, both inputs will contribute to field evolution according to Eq. 1. Since a single consonant production can only have a single VOT value, then lateral inhibition must be strong enough to prevent the formation of multiple stable activation peaks; ultimately, only one activation peak will form, even in the presence of two inputs. If the two inputs differ in their amplitude a, then the input with stronger amplitude will tend to dominate field evolution. However, if the two input distributions *overlap* in feature space, then the (ultimately inhibited) input with smaller amplitude can still contribute to the *field location* at which the activation peak forms. Crucially, this effect is sensitive to the width w of the Gaussian inputs.

To demonstrate this, we varied w from 1 to 50. For each value of w, we ran one simulation of field evolution under the influence of two inputs: s_{voiced} and $s_{voiceless}$. In each simulation, $p_{voiced} = 10$, $p_{voiceless} = 60$, $a_{voiced} = 4$, and $a_{voiceless} = 6$. Since $a_{voiceless} > a_{voiced}$, we expect $s_{voiceless}$ to dominate field evolution. The variation in w applied to both input distributions: s_{voiced} and $s_{voiceless}$ had the same value of w. Parameters of the field are listed in Table 1. Each simulation ran for 50 time steps, which was consistently found to be enough time for a stable activation peak to form.

Figure 1 displays the activation level at each VOT field location at the end of each simulation. For very small values of w (< 10), interaction between the two inputs was minimal, so that some below-threshold but above resting-level activation was maintained on the voiced side of the field without being inhibited by the above-threshold voiceless peak. In these cases, sub-threshold input from the voiced category had no impact on the VOT field location of the above-threshold peak, which was centered on 60 ms, the center of the $s_{voiceless}$ input distribution. For intermediate values of w (10 to 20), lateral inhibition from the voiceless peak reached further in the field, inhibiting activation on the voiced side of the field so that only the voiceless peak was apparent; this peak was still centered on 60 ms, the canonical field location for the voiceless category. Interestingly, at large values of w (> 20), overlap between the two input distributions was substantial enough that the voiceless activation peak was pulled in the direction of the voiced input. This is because—for large enough values of w—neurons on the more voiced (central) side of the voiceless peak received activation from *both* inputs s_{voiced} and $s_{voiceless}$, as well as lateral excitation, while neurons on the more voiceless (peripheral) side of the voiceless peak received less activation from s_{voiced}. Overlap in feature space between target distributions of opposing phonetic categories has previously been proposed to account for dissimilation effects between English vowels (Tilsen, 2009) and Mandarin tones (Tilsen, 2013).

Table 1: Field parameter values.

Parameter	Value
τ	20
h	-5
β	4
c_{exc}	15
c_{inh}	5
c_{glob}	0.9
σ_{exc}	5
σ_{inh}	12.5
q	1

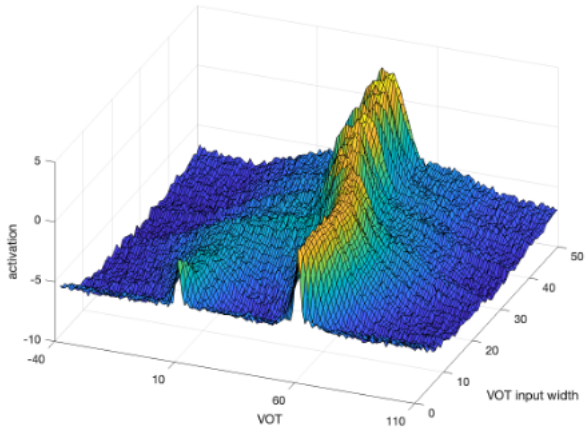

Figure 1: Effect of input width on overall field activation.

Figure 2 displays the field output as a function of w. The VOT target was calculated as the activation-weighted average of the field locations of above-threshold ($u > 0$) neurons. Figure 2 demonstrates that for values of w between about 20 and 35,

the VOT target was approximately a linear function of w, with higher values of w causing a larger trace effect (i.e. a smaller VOT target for a voiceless production). At very large values of w (> 35), field noise exerted a large influence on the precise value of the VOT target. At $w \approx 30$, the VOT target was reduced by about 10-15 ms, which approximates the magnitude of trace effects observed in speech errors elicited from tongue twister tasks (e.g., Goldrick et al., 2016; Goldrick & Blumstein, 2006). Interestingly, this value of w also approximates the standard deviation of VOT of voiceless plosives in American English (~30 ms) (Chodroff & Wilson, 2017).

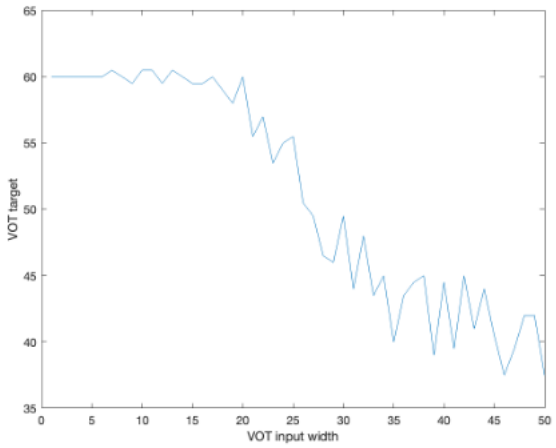

Figure 2: Effect of input width on VOT target.

Trace effects as activation peak attraction

We demonstrated in the previous section that an activation peak corresponding to a voiceless target can be pulled towards the voiced side of the feature space when a voiced input is simultaneously influencing the DNF, and there is overlap in feature space between the two inputs. In this section, we show more directly how this hypothesized neural phenomenon can give rise to phonetic trace effects in voicing errors. We assume, following previous work on speech errors (Goldrick et al., 2016; Goldrick & Blumstein, 2006; Goldrick & Chu, 2014), that some voicing errors are caused by the mental representation of the non-target voicing category becoming more active than the target category during planning. Trace effects reflect a non-zero activation of the target category in error production. We test this hypothesis with simulations of both voiceless and voiced stops as targets. For each case, we simulated 500 VOT productions in each of two conditions: (1) a non-error or canonical condition in which only an intended input influences the VOT planning DNF, and (2) an error condition in which both an unintended input and an intended, weaker input influences the DNF. In the canonical condition, $a_{target} = 6$ and $a_{competitor} = 0$. In the error condition, $a_{competitor} = 6$ and $a_{target} = 4$. w was set to 30 for all simulations. Otherwise, model parameters were identical to the simulations presented in the previous section. The simulation results are displayed in Figure 3 for voiceless stops and in Figure 4 for voiced stops. Both voiced and voiceless stops show phonetic trace effects consistent with those reported in the literature. For voiceless stops, mean VOT was reduced by ~13 ms in the error condition relative to the canonical condition. For voiced stops the phonetic trace effect was ~12 ms.

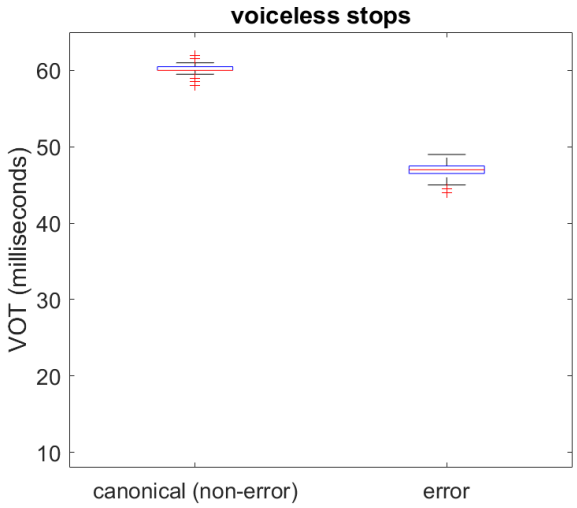

Figure 3: VOT of voiceless stops by condition.

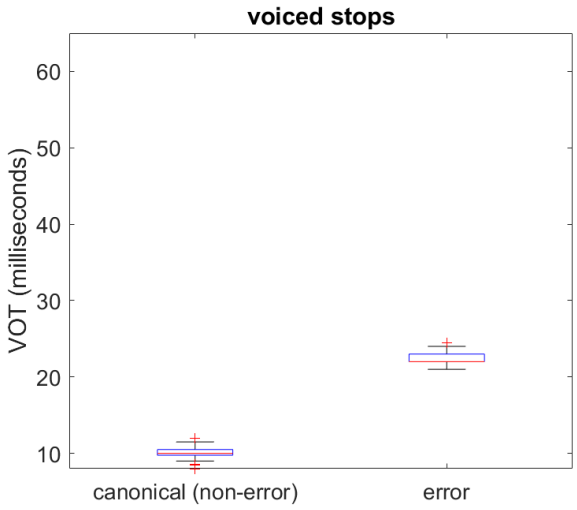

Figure 4: VOT of voiced stops by condition.

Errors arising from noise within the DNF

A major focus of computational models of speech production has been deriving the source of speech errors. So far, we have focused on errors which derive from cascading activation, i.e., multiple inputs into a phonetic planning field. In this type of speech error, the non-target category receives greater activation, and therefore exerts a greater influence over field evolution, but the target category (with less activation) still exerts some influence, deriving the trace effect. The error

thus originates at a level of planning preceding VOT target selection in the DNF. This is just one of many possible sources of speech errors, which also include, for example, articulatory factors (Goldstein et al., 2007; Mowrey & MacKay, 1990; Stemberger, 1983). Our account of trace effects in the framework of DFT opens up possibilities to understand and model other sources of errors and how they might interact with cascading activation. For example, the model presented here raises the additional possibility that some errors might originate from noise *within* the DNF. That is, even if the target input to the DNF is stronger than the non-target input, within-field noise might still lead to production of a VOT that would be classified as belonging to the non-target category. This possibility reduces the burden of higher level lexical and sub-lexical planning levels to account for all errors.

To investigate this possibility, we varied the noise amplitude in the DNF (q in Eq. 1) from 1 to 15 in steps of 0.5. For each value of q, we ran 500 simulations of field evolution. This allows us to observe the distribution of VOT values at each noise level. In each simulation, there were two inputs to the field: a voiceless input with amplitude $a_{voiceless} = 6$ and a voiced input with amplitude $a_{voiced} = 4$. This reflects a situation in which the intended voiceless category receives the most activation, but the unintended voiced category is still partially activated. A Bayesian classifier trained on the underlying category distributions $s_{voiceless}$ and s_{voiced} categorized each VOT target output by the DNF as either voiceless (no error) or voiced (error). Figure 5 displays the error rate (proportion of tokens categorized as voiced) as a function of noise amplitude q. At small values of q, the more active voiceless input overwhelmingly dominated field evolution, so errors were exceedingly rare. However, at large values of q (> 10), errors were observed at a non-negligible rate, sometimes exceeding 5% of productions.

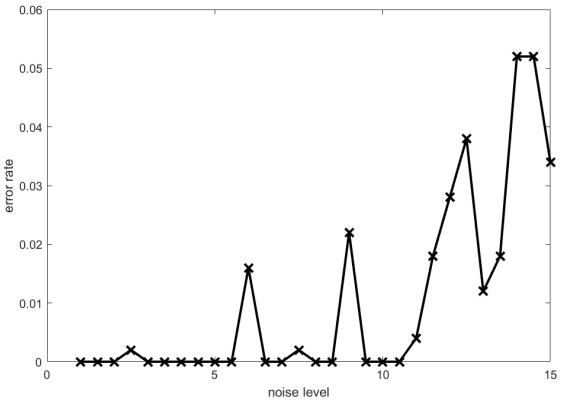

Figure 5: Error rate by noise level.

As seen in Figure 6, q did *not* have a consistent effect on median VOT. Rather, the effect of q on error rate appears to be driven by an increase in the number of outliers. These simulations support the hypothesis that, in addition to errors driven by greater activation of a competitor category compared to the target category, some errors might additionally be driven by noise within the process of VOT target selection, even when the target category representation exerts a greater overall influence on this process than the competitor.

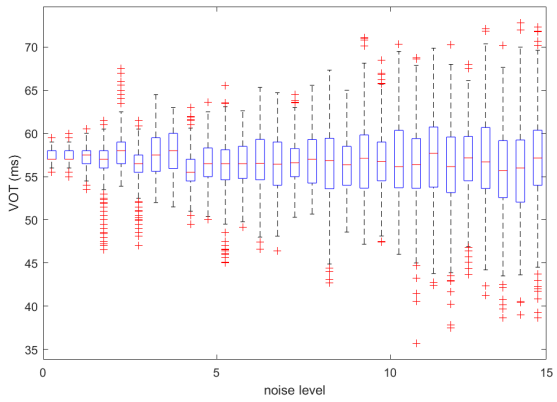

Figure 6: VOT by noise level.

Discussion

We demonstrated that trace effects in voicing errors can arise from the interaction of multiple inputs from higher levels of planning into a continuous phonetic planning space, consistent with a cascading activation account of speech planning and production. We obtained this result using a framework that can generalize to any phonetic dimension of speech, addressing the critique raised by Pouplier & Goldstein (2014). As discussed in the Introduction, the model in Goldrick & Chu (2014) restricts the effects of gradient symbolic activation on phonetic planning to articulation time alone, limiting the model's generality. Although in this paper we focused on VOT (a temporal dimension), our model can derive trace effects in both spatial and temporal dimensions. Moreover, our model provides wider empirical coverage of even the VOT facts, since it can derive trace effects in voiced as well as voiceless errors.

Moving beyond trace effects, we also explored the possibility that some errors arise from noise *within* the DNF governing VOT planning, rather than solely from noise in the process of lexical selection. In particular, we found that when within-field noise reached a particularly high level (q > 10), VOT targets became so variable that some targets were classified as voiced even when the intended voiceless input was stronger than the unintended voiced input. This relationship between response variability and error likelihood is consistent with the finding that VOT is more variable among errors than among canonical productions (Goldrick et al., 2016). In fact, even when errors arise from the process of lexical selection rather than within-field noise, our model is consistent with this finding. This is because the above-threshold activation peak tends to be wider in feature space when it reflects the contributions of two inputs (error case)

compared to one (canonical case). Previous work has related wider distributions of above-threshold activation to greater variability in behavior (Erlhagen & Schöner, 2002).

At low levels of field noise, the particular method we used to simulate VOT target selection, i.e. activation-weighted averaging of above-threshold neurons, yielded low VOT variance for both conditions (Figure 3, 4). That is, the small differences in the widths of the above-threshold activation peaks (for errors and for canonical productions) did not translate into increased VOT variability for errors. As we now discuss, whether the width of an above-threshold activation peak influences VOT variability depends both on the level of field noise and the particular method of target selection.

There are in principle different methods of VOT target selection from field activation. In addition to the method described above, we explored three others: (1) sampling from an activation-weighted distribution of above-threshold neurons, (2) selecting the single neuron with the highest activation, and (3) sampling from a uniform distribution of above-threshold neurons. At low levels of within-field noise, q, these methods showed very little difference in the resulting variability of VOT targets. However, we saw a clearer separation at higher levels of noise—while each method resulted in more VOT target variability as noise level was increased, some methods led to greater variability than others. In particular, we saw that target variability was greatest for method (3), followed by method (2), and lastly, method (1).

Activation-weighted averaging of above-threshold neurons, the method we used, was least sensitive to both noise scaling and the difference in above-threshold activation peak width between errorful and canonical productions. Importantly, at high levels of noise, all methods resulted in greater target variability when there were two inputs in the field as compared to one—consistent with the aforementioned finding of more VOT variability among errorful productions (Goldrick et al., 2016). Notably, methods that were more sensitive to within-field noise also showed a greater difference in variability between errorful and canonical simulations, with (3) showing the greatest difference, followed by (2), and then (1). We conclude from this brief exploration that noise can be useful for revealing the dynamics of the model. It is only in the presence of sufficient noise that we expose important differences in the underlying dynamics found across conditions.

As a final point, our DNF, defined in terms of a differential equation, is inherently temporal, capturing the stabilization of the field over time. The explicit incorporation of both temporal and feature gradience allows the generation of quantitative predictions regarding response time, response parameters, and the relationship between them. For example, Roon and Gafos (2016) used a similar approach within the DFT framework to model the influence of phonetic similarity between response and distractor on response times in a speech production task. Regarding speech errors, previous work has shown a relationship between speech rate and error rate, such that errors are more likely at faster speech rates (Goldstein et al., 2007). Future work could probe the temporal characteristics of the present model (for example, by varying τ, or by forcing target selection to occur at a particular time step) in order to generate additional predictions regarding the relationship between speech rate and the *phonetic properties* of errors.

Conclusion

We presented a dynamic neural field (DNF) model of voice onset time (VOT) planning. In the model, VOT targets are derived during speech planning from category inputs represented as distributions in feature space, lateral interactions (excitatory and inhibitory) between neurons in the field, and noise. The model allows simultaneous inputs from voiced and voiceless categories to explicitly interact in a continuous feature space. Through simulations, we demonstrated that the model generates trace effects in voicing errors consistent with those observed in experiments and naturalistic speech. We thus offered a neural-computational implementation of the proposal that trace effects arise from the interaction of multiple active categories during speech planning, consistent with the cascading activation account. We also discussed possible extensions of the model beyond phonetic trace effects.

Acknowledgments

Many thanks to Marisa Norzagaray and Irene Yi for discussion of ideas related to this paper.

References

Alderete, J., Baese-Berk, M., Leung, K., & Goldrick, M. (2021). Cascading activation in phonological planning and articulation: Evidence from spontaneous speech errors. *Cognition*, *210*. https://doi.org/10.1016/j.cognition.2020.104577

Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, *31*(2), 149–180. https://doi.org/10.1016/S0095-4470(02)00085-2

Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics*, *61*, 30–47. https://doi.org/10.1016/j.wocn.2017.01.001

Dell, G. S. (1986). A Spreading-Activation Theory of Retrieval in Sentence Production. *Psychological Review*, *93*(3), 283–321. https://doi.org/10.1037/0033-295X.93.3.283

Dell, G. S., Kelley, A. C., Hwang, S., & Bian, Y. (2021). The adaptable speaker: A theory of implicit learning in language production. *Psychological Review*, *128*(3), 446–487. https://doi.org/10.1037/rev0000275

Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review*, *109*(3), 545–572. https://doi.org/10.1037/0033-295X.109.3.545

Frisch, S. A., & Wright, R. (2002). The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue. *Journal of Phonetics*, *30*(2), 139–162. https://doi.org/10.1006/jpho.2002.0176

Fromkin, V. A. (1971). The Non-Anomalous Nature of Anomalous Utterances. *Language*, *47*(1), 27–52.

Gafos, A., & Kirov, C. (2009). A dynamical model of change in phonological representations: The case of lenition. *Approaches to Phonological Complexity*, 219–240. https://doi.org/10.1515/9783110223958.219

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, *233*, 1416–1419.

Giese, M. A. (1999). *Dynamic Neural Field Theory for Motion Perception*. Kluwer Academic Publishers.

Goldrick, M., & Blumstein, S. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes*, *21*(6), 649–683. https://doi.org/10.1080/01690960500181332

Goldrick, M., & Chu, K. (2014). Gradient co-activation and speech error articulation: Comment on Pouplier and Goldstein (2010). *Language, Cognition and Neuroscience*, *29*(4), 452–458. https://doi.org/10.1080/01690965.2013.807347

Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., & Byrd, D. (2007). Dynamic action units slip in speech production errors. *Cognition*, *103*(3), 386–412. https://doi.org/10.1016/j.cognition.2006.05.010

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, *102*(3), 594–621. https://doi.org/10.1037/0033-295X.102.3.594

Keating, P. A. (1990). The window model of coarticulation: articulatory evidence. In M. E. Beckman & J. Kingston (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech* (pp. 451–470). Cambridge University Press.

Kopecz, K., & Schöner, G. (1995). Saccadic motor planning by integrating visual information and pre-information on neural dynamic fields. *Biological Cybernetics*, *73*(1), 49–60. https://doi.org/10.1007/BF00199055

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–75. https://doi.org/10.1017/S0140525X99001776

Lindblom, B. (1963). Spectrographic Study of Vowel Reduction. *The Journal of the Acoustical Society of America*, *35*(5), 783–783. https://doi.org/10.1121/1.2142410

Lisker, L., & Abramson, A. S. (1964). A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements. WORD, *20*(3), 384–422. https://doi.org/10.1080/00437956.1964.11659830

McMillan, C. T., & Corley, M. (2010). Cascading influences on the production of speech: Evidence from articulation. *Cognition*, *117*(3), 243–260. https://doi.org/10.1016/j.cognition.2010.08.019

Mowrey, R. A., & MacKay, I. R. A. (1990). Phonological primitives: Electromyographic speech error evidence. *The Journal of the Acoustical Society of America*, *88*(3), 1299–1312. https://doi.org/10.1121/1.399706

Pouplier, M., & Goldstein, L. (2010). Intention in articulation: Articulatory timing in alternating consonant sequences and its implications for models of speech production. *Language and Cognitive Processes*, *25*(5), 616–649. https://doi.org/10.1080/01690960903395380

Pouplier, M., & Goldstein, L. (2014). The relationship between planning and execution is more than duration: Response to Goldrick & Chu. *Language, Cognition and Neuroscience*, *29*(9), 1097–1099. https://doi.org/10.1080/01690965.2013.834063

Roon, K. D., & Gafos, A. I. (2016). Perceiving while producing: Modeling the dynamics of phonological planning. *Journal of Memory and Language*, *89*, 222–243. https://doi.org/10.1016/j.jml.2016.01.005

Schneegans, S. (2021). *COSIVINA: A Matlab Toolbox to Compose, Simulate, and Visualize Neurodynamic Architectures* (Version 1.4). https://github.com/cosivina/cosivina

Schöner, G., Spencer, J., & Group, D. R. (2016). *Dynamic Thinking: A Primer on Dynamic Field Theory*. Oxford University Press.

Smolensky, P., Goldrick, M., & Mathis, D. (2014). Optimization and quantization in gradient symbol systems: A framework for integrating the continuous and the discrete in cognition. *Cognitive Science*, *38*(6), 1102–1138. https://doi.org/10.1111/cogs.12047

Stemberger, J. P. (1983). The nature of /r/ and /l/ in English: evidence from speech errors. *Journal of Phonetics*, *11*(2), 139–147. https://doi.org/10.1016/s0095-4470(19)30812-5

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. https://doi.org/10.1016/j.wocn.2009.03.004

Tilsen, S. (2013). Inhibitory mechanisms in speech planning maintain and maximize contrast. In A. Yu (Ed.), *Origins of Sound Change: Approaches to Phonologization* (pp. 112–127). Oxford University Press.

Tilsen, S. (2019). Motoric mechanisms for the emergence of non-local phonological patterns. *Frontiers in Psychology*, *10*(SEP), 1–25. https://doi.org/10.3389/fpsyg.2019.02143