Journal of Memory and Language 89 (2016) 222–243

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Perceiving while producing: Modeling the dynamics of phonological planning

Kevin D. Roon [a,b,*], Adamantios I. Gafos [b,c]

[a] The CUNY Graduate Center, 365 Fifth Avenue, Suite 7107, New York, NY 10016, USA
[b] Haskins Laboratories, 300 George Street, Suite 900, New Haven, CT 06511, USA
[c] Universität Potsdam, Department Linguistik, Haus 14, Karl-Liebknecht-Straße 24-25, 14476 Potsdam, Germany

ARTICLE INFO

Article history:
Received 3 November 2014
revision received 26 January 2016
Available online 9 March 2016

Keywords:
Speech perception
Speech production
Computational modeling
Response time modulation
Phonological planning

ABSTRACT

We offer a dynamical model of phonological planning that provides a formal instantiation of how the speech production and perception systems interact during online processing. The model is developed on the basis of evidence from an experimental task that requires concurrent use of both systems, the so-called response–distractor task in which speakers hear distractor syllables while they are preparing to produce required responses. The model formalizes how ongoing response planning is affected by perception and accounts for a range of results reported across previous studies. It does so by explicitly addressing the setting of parameter values in representations. The key unit of the model is that of the dynamic field, a distribution of activation over the range of values associated with each representational parameter. The setting of parameter values takes place by the attainment of a stable distribution of activation over the entire field, stable in the sense that it persists even after the response cue in the above experiments has been removed. This and other properties of representations that have been taken as axiomatic in previous work are derived by the dynamics of the proposed model.

Introduction

Discussion about the links between speech perception and production has traditionally been concerned with whether the objects of speech perception are acoustic or articulatory (see Diehl, Lotto, & Holt, 2004; Fowler, 1996; Galantucci, Fowler, & Turvey, 2006; Liberman & Mattingly, 1985; Ohala, 1996, among many others). Despite disagreement on answers to that theoretical question, the assertion that speech perception and production are tightly linked is not contentious (see, e.g., Diehl et al., 2004; Hickok & Poeppel, 2000; Moulin-Frier, Laurent, Bessière, Schwartz, & Diard, 2012). More attention is now being paid to understanding better how perception and production are related, and to what representations are involved in the link between the two. Nevertheless, very little to no attention has been paid to developing explicit computational models of the online interaction between speech perception and production. We present a dynamical, computationally explicit model of the process by which phonological production parameters are set. The model focuses on a specific task that requires the concurrent use of both speech perception and production, and thereby sheds light on the nature of the representations involved in the perception–production link.

There is good evidence for facilitation in speech production response times (RTs) when perceived stimuli share phonemes with intended productions in a variety of experimental paradigms (Forster & Davis, 1991; Galantucci,

* Corresponding author at: The CUNY Graduate Center, 365 Fifth Avenue, Suite 7107, New York, NY 10016, USA.
E-mail addresses: kroon@gc.cuny.edu (K.D. Roon), gafos@uni-potsdam.de (A.I. Gafos).

http://dx.doi.org/10.1016/j.jml.2016.01.005
0749-596X/

Fowler, & Goldstein, 2009; Kerzel & Bekkering, 2000; Schriefers, Meyer, & Levelt, 1990). Beyond shared phonemes, studies have attempted to further probe the specificity of representations involved in the perception–production link by also seeking to uncover effects on RTs attributable to linguistic properties corresponding to distinctive features. The results have been mixed. For instance, several studies have sought (without success) a feature-level effect for the feature of place, corresponding to sharing of primary oral articulator between the perceived stimulus and the required response (Galantucci et al., 2009; Gordon & Meyer, 1984; Mitterer & Ernestus, 2008; Roelofs, 1999). The lack of an effect in these studies is particularly surprising given the undisputed status of the oral articulator in the description of linguistic contrasts (Chomsky & Halle, 1968; Ladefoged & Maddieson, 1996). Another set of studies has uncovered feature-level effects for primary oral articulator as well as for voicing, evidenced both by modulations of production RTs (Gordon & Meyer, 1984, for voicing; Klein, Roon, & Gafos, 2015, for articulator; Mousikou, Roon, & Rastle, 2015, for voicing; Roon & Gafos, 2015, for both) and by modulations of the phonetic output of speakers (Goldinger, 1998; Nielsen, 2007; Tilsen, 2009; Yuen, Brysbaert, Davis, & Rastle, 2010) driven by (in)compatibility between recently perceived stimuli and utterances produced. It can be reasonably argued that the inconsistency in finding feature-level effects is due to the variety in the experimental tasks across the various studies, which included responding to an auditory cue based on learned cue–response pairs (Gordon & Meyer, 1984; Roelofs, 1999), responding to a visual cue in the presence of various distractors (Galantucci et al., 2009; Kerzel & Bekkering, 2000; Roon & Gafos, 2015), reading aloud with masked primes (Mousikou et al., 2015), and shadowing spoken stimuli (Mitterer & Ernestus, 2008). However, if we focus on results from a series of studies that use the same experimental task, the response–distractor task, it turns out that the results for feature-level effects are reliably consistent. These latter results offer a rich and sufficiently coherent dataset that makes possible the formalization of the link between perception and production. In the present study, therefore, we provide a computationally explicit model of these feature-level effects in the response–distractor task. The model will be shown to account for the range of results from studies using this task by proposing a link between speech perception and production that is situated in the process of phonological planning.

In a response–distractor task, participants learn pairs of visual cues and spoken syllables (e.g., "if you see && say *ba*, if you see ## say *da*"). Participants are instructed that they will repeatedly see these cues and that they should say the corresponding syllable that they have learned as quickly as possible, but not so quickly that they make a lot of mistakes. They are also told that they will hear various things over headphones while they are performing the task, and that they should ignore what they hear. As shown in Fig. 1, participants first see a fixation box alerting them to the beginning of the trial. After 500 ms, participants see a cue instructing them which syllable to say. Shortly after the presentation of the cue, participants hear one of

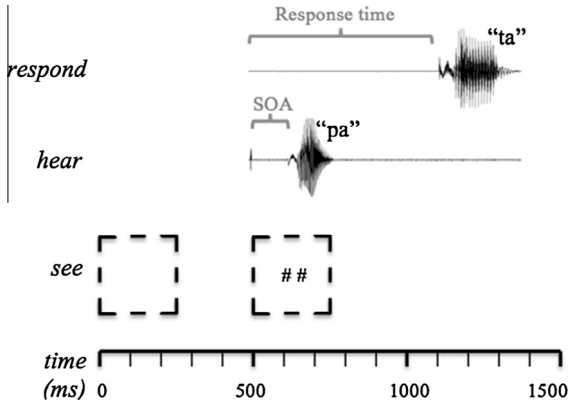

Fig. 1. Time line of one trial from the response–distractor task. The participant's task is to produce *ta* upon seeing the visual cue ##. At an SOA of 100 ms, the participant hears an auditory distractor, which is the syllable *pa*.

a number of various linguistic distractors, a non-speech sinusoidal tone equal in length to the linguistic distractors, or no distractor. The timing of the distractor relative to the cue is such that the distractor always follows the cue by a set duration, that is, a positive Stimulus Onset Asynchrony (SOA) is used. The response time on the trial is measured as the time from the onset of the visual cue to the acoustic onset of the produced response. The crucial experimental manipulation consists of systematically varying the (in)-compatibility between the distractor and response along various phonological parameters.

The design of the response–distractor task is well suited to provide evidence of effects attributable to the interaction of the speech production and perception systems. Any results from experimental tasks that present some priming or distractor stimulus at any time before the participant has decided on a response (e.g., a shadowing task) are open to being interpreted as reflecting "selection" effects (or, "stimulus-response" compatibility effects, see Galantucci et al., 2009; Kerzel & Bekkering, 2000; Kornblum, 1994, for discussion). A prime or distractor stimulus may bias the participant toward (or away from) a particular response, thereby speeding up (or slowing down) RTs, but the nature of the bias is highly task-dependent. That is, depending on the task, the bias may be driven by congruency along any number or combination of parameters—acoustic, articulatory, visual, orthographic—and not be driven solely by sharing speech-specific (acoustic or articulatory) properties. In contrast, in the response–distractor task, the distractor stimulus is presented so close in time to the beginning of the utterance that any influence of the distractor stimulus must reflect involvement of the production system in perception since it is simply too late for any other representations to be involved. Effects on RTs that are attributable to the interaction of the speech production and perception systems have therefore been dubbed "perceptuo-motor" effects (also referred to as "stimulus-stimulus" compatibility effects, Kornblum, 1994).

Kerzel and Bekkering (2000) and Galantucci et al. (2009) use this response–distractor task to show that

phonemic identity between response and distractor yields facilitative effects on RTs. The main difference between the two studies is that Kerzel and Bekkering (2000) used silent videos of speakers producing the distractor stimuli, while Galantucci et al. (2009) used auditory stimuli only. Fig. 2 illustrates the basic findings from Galantucci et al. (2009), which were consistent with the results found by Kerzel and Bekkering (2000) despite the different modalities of the distractor stimuli. In order to understand the effects of the linguistic distractors on RTs, the conditions that did not involve linguistic distractors must be examined first. There are two consistent non-linguistic influences on RTs: the presence of any distractor vs. no distractor, and SOA. The presence of a distractor increased RTs, regardless of whether the distractor was a speech syllable or a tone, as RTs were fastest when there was no distractor (bar with vertical pattern). RTs increased monotonically as SOAs increased. These non-linguistic influences presumably arise from some other cognitive process (or processes) involved in this task that do not involve the (specifically) speech perception–production link. The RT slowdown of the Tone condition at various SOAs compared to the No Distractor condition therefore can be treated as a neutral baseline RT reference indicating the influence of these other processing demands, but not reflecting any influence of the process that generates perceptuo-motor effects. The main perceptuo-motor effect from Galantucci et al. (2009) is the dependence of RTs on phonemic identity. Specifically, RTs were shorter than the neutral tone distractor within a given SOA if the distractor was the same syllable as the response (e.g., *ba*–*ba*, white bars in Fig. 2) and longer if the distractor had a different onset consonant from the response, which in the case of the Galantucci et al. (2009) experiment meant that they differed in articulator (e.g., *ba*–*da*, bars with horizontal shading in Fig. 2).

Roon and Gafos (2015) used the same task to reveal perceptuo-motor effects beyond phonemic identity and found effects both of articulator and voicing (Fig. 3). The key difference in experimental design between Roon and Gafos (2015) and Galantucci et al. (2009) was that the response and distractor were never identical in the former, which allowed for teasing apart individual feature-level effects. Specifically, in Roon and Gafos (2015)'s articulator experiment (Fig. 3A), distractors never matched responses in voicing, but had an articulator that was either congruent with the response (e.g., response *pa*–distractor *ba*) or incongruent (e.g., *pa*–*da*). In their voicing experiment (Fig. 3B), distractors never matched responses in articulator, but had voicing that was either congruent (e.g., *ta*–*pa*) or incongruent (e.g., *ta*–*ba*) with the response. In both experiments, RTs were slower in the incongruent case than in the congruent case. These results were the first to provide clear evidence for independent effects of articulator and voicing in this task.

Taken together, the results from the above experiments, which all employed the response–distractor task to isolate perceptuo-motor effects, securely establish perceptuo-motor interactions beyond cases of complete identity between required responses and distractors, and provide design characteristics for a model of the perception–production link. We next present such a model and demonstrate its efficacy in capturing these results and others, as well as in making novel predictions.

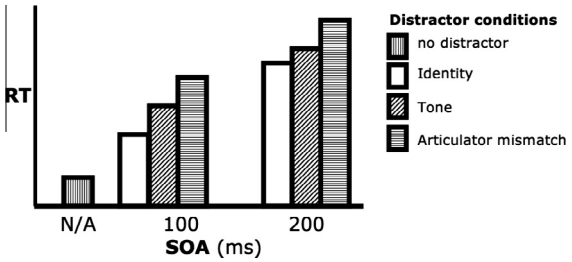

Fig. 2. Schematic representation of the results from Galantucci, Fowler, and Goldstein (2009). RTs were faster when distractors were identical to responses (white bars), and slower when they mismatched in articulator (bars with horizontal shading).

Model of phonological planning

Consider a syllable *ta*, beginning with a tongue tip constriction as required for a /t/ followed by tongue back vowel and glottal gestures as required for an /a/. Upon presentation of a visual cue indicating that the required response is the syllable *ta*, a speaker must assemble a set of parameter values that specify the required vocal tract actions. These include (but are not limited to) articulator-specific parameters referring to the constriction location and constriction degree of the articulator forming the constriction required for the initial consonant (Browman & Goldstein, 1990; Guenther, 1995; Saltzman & Munhall, 1989), as well as the parameter specifying the voicing for that consonant to be voiceless. For instance, in *ta*, the speaker must set a constriction location value for the tongue tip articulator (and not the tongue back, as would be the case for *ka*), and a degree of constriction (for a stop like /t/, that is "full closure" as opposed to "critical", as would be the case for the fricative in *sa*) to be effected by this articulator. For voicing, the speaker must set the oral–laryngeal timing needed for properly coordinating the consonant's release with the onset of modal voicing for the vowel, known as the Voice Onset Time parameter (VOT, Lisker & Abramson, 1964). In our model, each such parameter corresponds to a planning field. Fig. 4 shows the components of the model for the response–distractor task. It includes three planning fields for each potential speech articulator (limited to those relevant to the data considered here: Tongue Tip, Tongue Back, and Lower Lip; shown in orange shaded rectangles), another planning field for Voicing (shown in the blue shaded rectangle), inputs to the fields (shown in ovals), and a Monitor function. Inputs to the fields as well as interaction within and across fields determine in a mathematically explicit way described below the actual parameter values to be produced. The Monitor function decides when all of the required parameter values have been determined. At that point, those parameter values are sent to implementation. Implementation is separate from our model, and is a system that controls the online movements of articulators, such as the

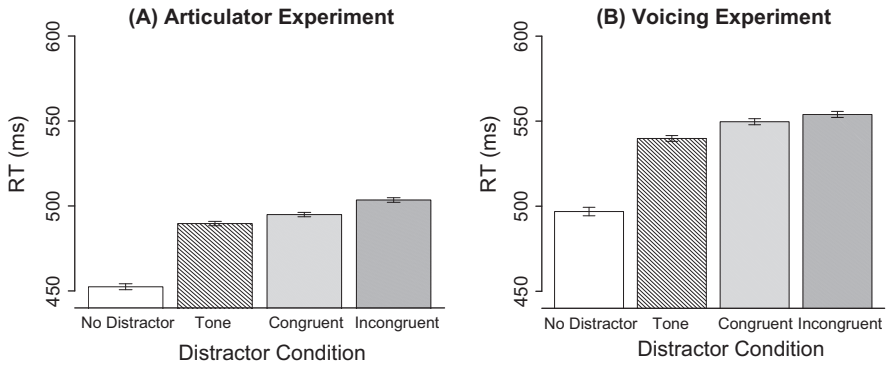

Fig. 3. Results from Roon and Gafos (2015). (A) Articulator experiment. (B) Voicing experiment.

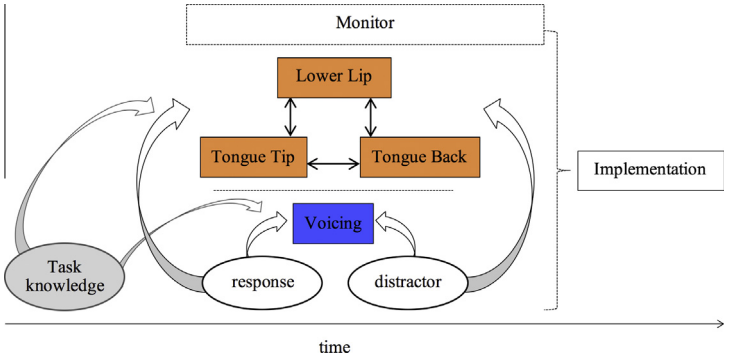

Fig. 4. Model of phonological planning. Shaded rectangles represent planning fields: orange for articulator planning fields and blue for voicing. Double-pointed arrows represent cross-field inhibition. Ovals represent three sources of input. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Task Dynamics Model (Saltzman & Munhall, 1989) or DIVA (Guenther, 1995, et seq.).

The planning fields in Fig. 4 evolve over time and determine the specific parameter settings for the phonological parameters in an intended utterance. The evolution of the fields is specified by a dynamical system. A dynamical system is a formal system whose internal state changes over time in a mathematically explicit way. The workings of the proposed model are based on Dynamic Field Theory ("DFT"), a theoretical framework originally developed in the context of movement planning (see Erlhagen & Schöner, 2002, for a general formulation of the theory; see Kopecz & Schöner, 1995, for an earlier formulation in the context of oculomotor tasks; see Schöner, Spencer, & DFT Research Group, 2016, for a comprehensive survey of the current state of the theory), and by now extended to domains as wide-ranging as motion preparation (Hock, Schöner, & Giese, 2003), behavioral choice in the A-not-B infant perseverative-reaching paradigm (Thelen, Schöner, Scheier, & Smith, 2001), and turn-taking in dyadic communication (Sandamirskaya & Schöner, 2008). The mathematical foundations of DFT derive from the landmark analytical treatment of neural field dynamics by Amari (1977). Amari's key equation for field dynamics is given in (1). In this equation, A is the field (a function of the continuous variables x, t), h is the field's resting activation, $dA(x,t)$ is the change in activation at x at time t, τ is a constant corresponding to the rate of decay of the field, $Input(x,t)$ is time-dependent input to the system (i.e., a cue specifying a required response or a perceived distractor) in the form of a localized activation spike, $S(x,t)$ is a term expressing interactions among different field sites, and *noise* contributes stochastic random noise to the activation evolution.

Main stochastic differential equation for field evolution :

$$\tau dA(x,t) = -A(x,t) + h + Input(x,t) + S(x,t) + noise \quad (1)$$

Eq. (1) can be broken down into simpler components to better understand how it functions. The core component $\tau dA(x,t) = -A(x,t) + h$ states that the rate of activation change $dA(x,t)$ is a linear function of current activation $A(x,t)$ and specifically that it is inversely related to the current activation $A(x,t)$ plus some constant h. This relation is an instance of exponential decay dynamics. To see this, let us arbitrarily select a single location for x, which we call x^i, and plot its activation $A(x^i,t)$ over time. As shown in the top left panel of Fig. 5, in the absence of any input or interaction, activation $A(x^i,t)$ converges exponentially to the resting level h and stays there once h is reached (at this level the right hand side of the equation becomes zero, which

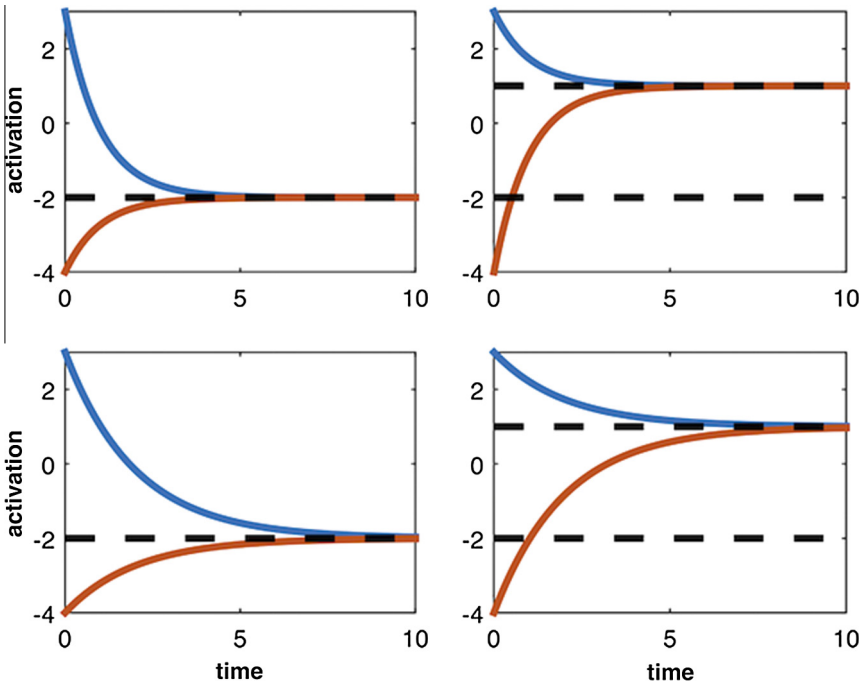

Fig. 5. Trajectories of the simplified linear dynamics $\tau dA(x,t) = -A(x,t) + h$. Top left: In the absence of input, field activation at a particular point converges to the resting level, the "off" state. Here $h = -2$ (dashed line) and the time scale is specified by $\tau = 1$. Top right: With added input $Input(x,t) = 3$, activation converges to resting level h plus input (top dashed line) with $\tau = 1$. Bottom left: In the absence of input, activation converges to resting level $h = -2$ with a slower time scale specified by $\tau = 2$. Bottom right: With added input $Input(x,t) = 3$, activation converges to resting level h plus input with $\tau = 2$.

means that the rate of change on the left hand side becomes zero and thus no further change is due). Let us call this resting activation level the "off" state for the parameter represented by this equation governing the evolution of x^i. In the terminology of dynamical systems, the starting activation of x^i is known as an *initial condition*, and the activation it converges to, in this case the resting activation, is known as a stable fixed point or an *attractor*. If the input term, $Input(x,t)$, is non-zero, then the system will move toward a new attractor equal to the resting activation plus the input term, as shown in the top right panel. The speed of the process is modulated by the τ term, which defines the time scale of the planning process, with the top two panels in Fig. 5 showing faster convergence than the bottom two.

Fig. 6 depicts the same dynamics as above in a way that fully captures the system's behavior regardless of initial conditions and without solving the equation $\tau dA(x,t) = -A(x,t) + h$ as was done to obtain the trajectories in Fig. 5. Let us denote the right hand side of $\tau dA(x,t) = -A(x,t) + h$ by $f(A)$. Without solving $\tau dA(x,t) = f(A)$, one can fully describe the behavior of $A(x,t)$ by considering just three cases. If $f(A)$ is positive, the rate of change $dA(x,t)$ must be positive and thus $A(x,t)$ will increase by an amount given by $dA(x,t)$. If $f(A)$ is negative, $A(x,t)$ will decrease. If $f(A)$ is zero, $A(x,t)$ stays the same. The values of A for which the latter is true are called fixed points—these are the points where the line representing $f(A)$ intersects the A axis. Thus, $f(A)$ can be seen to specify a vector which indicates the direction of change for A and also the magnitude of the change and for this reason it is known as a *vector field* of the dynamical system. The arrows on the A axis of Fig. 6 show the vector field by taking representative values of A and drawing on top of each of these values an arrow pointing in the direction of change, that is, to the right/left for positive/negative $f(A)$. The stability of the fixed point is indicated by the arrows (both to its left and to its right) pointing toward it. This much background is sufficient to illustrate one essential point, which is that the dynamics controlling the change of activation are self-stabilizing: when the system finds itself below or above the resting level, due to setting its initial conditions of activation at this level or due to perturbations that may be applied to it (e.g., noise introduced by stochastic forces) during the course of its evolution, the system converges back to that level of activation. This property of dynamical systems, which derives formally from the state dependence of the dynamics and specifically from the rate of activation change $dA(x,t)$ being inversely related to the current activation $A(x,t)$, plays a key role in formalizing the concept of representation and in setting and maintaining parameter values in our model.

In moving from the single parameter linear dynamics to fields, the parameter x turns to a continuum of locations representing the range of possible parameter values (e.g., constriction locations) as opposed to a single location x^i above. This continuum is shown by an axis in our field representations with each point along that axis associated with an activation value (hence we can still speak of activation at x). Single location activation now turns to a

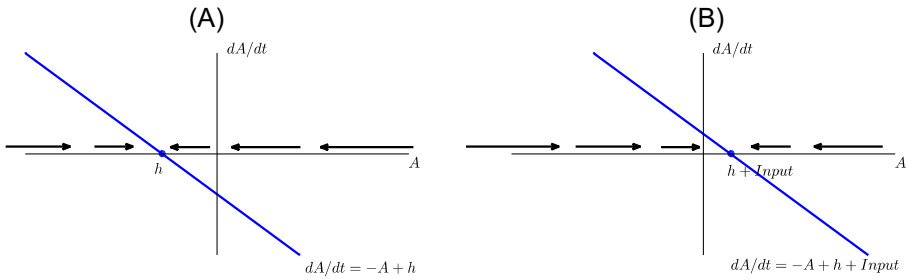

Fig. 6. Linear dynamics with corresponding vector fields. (A) In the absence of input, the fixed point is the resting level h, which represents the "off" state of the system. (B) With added input, the fixed point is lifted higher to an input-determined value, namely, that of the resting level h plus input. When input is removed, the system returns back to the "off" state. In this linear system, there is no qualitative change in the dynamics as input strength is scaled. Specifically, there is a single fixed point throughout.

distribution of activation over that continuum of locations represented by the parameter axis. Issues of stability in the field dynamics correspondingly translate to the existence and specification of regions of locations over which an activation distribution stabilizes. We first illustrate graphically the different stabilization scenarios with field dynamics in Fig. 7 and then turn to a discussion of how Eq. (1) prescribes these scenarios.

When input to the field is weak, as in the case of a small spike introduced and removed shortly thereafter, the field relaxes back to its resting level. This is shown in Fig. 7A, which illustrates what is involved as we move from the single parameter exponential growth/decay dynamics to fields. It is now the evolution of activation along the entire field represented by the constriction location axis that is depicted. A small spike raises activation values in a region between the anterior and posterior ends of the constriction location continuum. Eventually, the spike wanes as the retraction of input results in the field relaxing back to its resting level (this is analogous to what happens with the no-input case of the simple dynamics illustrated above). At each time step, evolution is noisy, as shown by the small random perturbations throughout the field.

In contrast, inputs of sufficient strength and duration lead to stabilization, i.e., to a state of activation distribution where a peak formed above the resting level can be maintained.[1] This is illustrated in Fig. 7B. The figure shows input to the Tongue Tip Constriction Location field introduced at time step 200 and evolving over time to an eventually stable peak with higher activation at some intermediate value of Tongue Tip Constriction Location on the anterior–posterior axis (note that the noise in the field is still present but less visible than in Fig. 7A due to the larger range of activation values displayed). This peak is stable in the sense that, once achieved, it persists indefinitely, even in the absence of further input. Indeed, in the example shown in Fig. 7B, there is no input to the field after time step 500 but the single-peak distribution of activation in the field remains. This is the "on" state of a planning field. It is when the dynamics have reached this stable, non-resting activation state that we say a parameter value in a representation has been set. In our model, the phonetic parameter value of the peak in this second stable "on" state is what is sent to implementation.

Whereas Fig. 7A and B illustrate cases of monomodal or single-peak input, Fig. 7C turns to a case where the input is bimodal. Specifically, Fig. 7C shows that the buildup of activation for a posterior constriction location can be suppressed by the introduction of an incongruent input. The activation buildup for the posterior constriction starts off similar to the buildup in Fig. 7B, but 300 time steps into the evolution of the posterior constriction peak, a second, incongruent anterior constriction location input is introduced into the field. As that peak rises, it inhibits the buildup of the posterior activation peak, resulting in a brief dip in its maximum value around time step 1000. The incongruent anterior constriction location activation is ultimately not sufficient to prevent the field from stabilizing with the posterior constriction location peak, and it soon dies out due to the inhibition introduced by the posterior peak. Nevertheless, the introduction of the incongruent peak does result in achievement of the stable "on" state being delayed compared to the field evolution depicted in Fig. 7B. Indeed, a crucial function of dynamic fields in DFT in general and in our model specifically is to provide a mechanism to resolve multiple—and potentially conflicting—inputs to the planning process. In Eq. (1), this mechanism corresponds to the interaction term $S(x,t)$, which crucially endows fields with this capacity of decision. As we formally explicate below, how close the peaks of the two inputs are to each other, as well as their relative strength, width, and timing all affect the field's achievement of a single stable state.

We now characterize formally the $Input(x,t)$ term. Inputs to the model take the form of activation distributions. The key idea is that each phonological parameter is not specified by a single numerical value, but rather by an activation distribution depicting the continuity of its phonetic detail. These distributions in the model are defined by (2), and examples are illustrated in Fig. 8.

$$input = e^{-(x-val+noise)^2}/2\sigma^2 \tag{2}$$

In this equation, *val* indicates the mean of the distribution, and includes a small noise term. The standard deviation of the distribution (σ) defines the width of an input. For

[1] The reason why the strength of input leads to stabilization in field dynamics requires considering the effects of the interaction term $S(x,t)$ in the right hand side of Eq. (1). We therefore return to this reason below after we describe the interaction term.

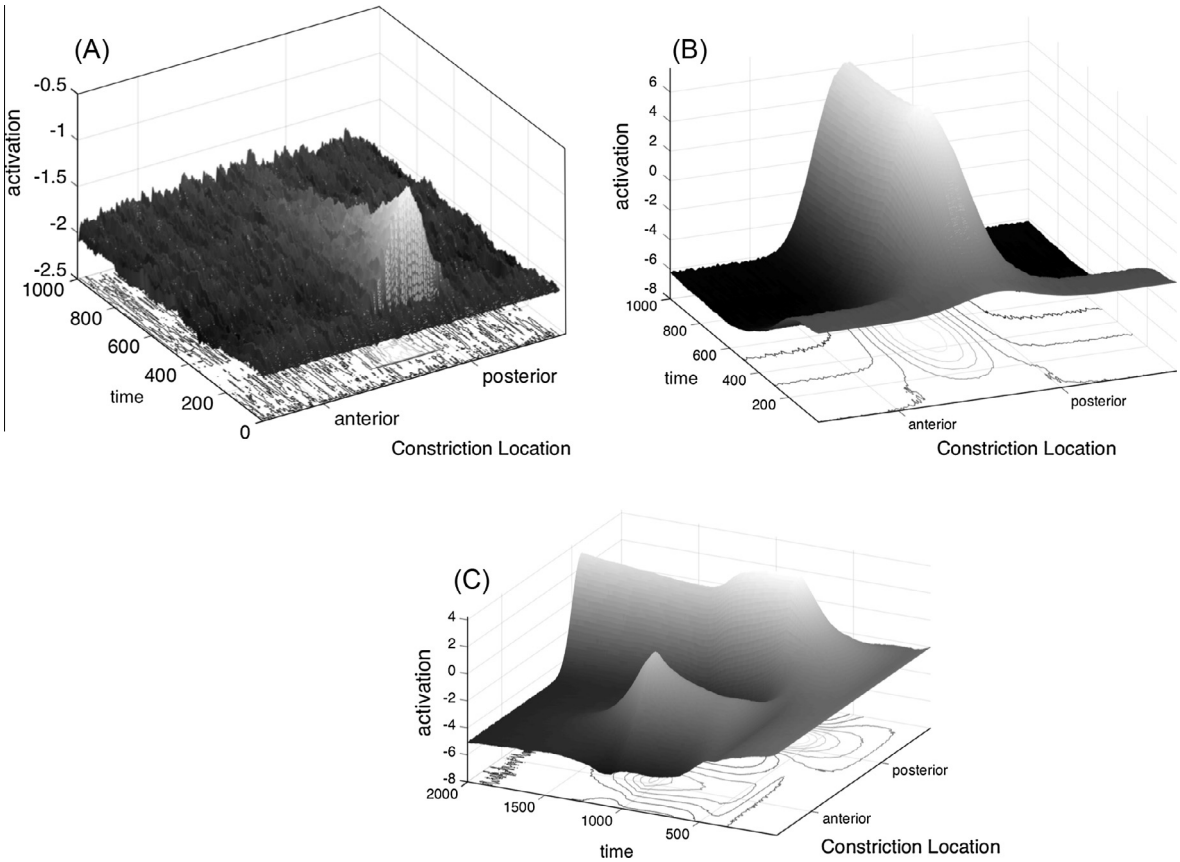

Fig. 7. Planning field for the Tongue Tip constriction location. (A) Insufficient input to the field results in activation levels returning to rest, i.e., the "off" state. (B) Sufficient input to the field results in a self-sustaining peak of activation, i.e., the "on" state. (C) Stable peak temporarily inhibited by incongruent input.

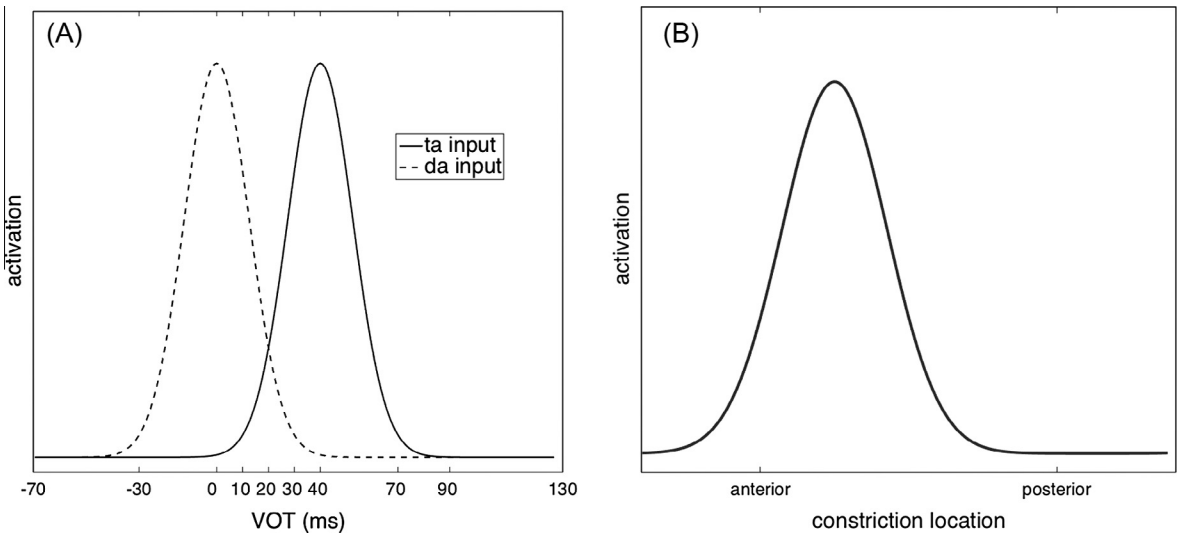

Fig. 8. (A) Representations of Voice Onset Times for syllable-initial stops differing in voicing: voiced (e.g., *da*, dashed line), and voiceless (e.g., *ta*, solid line). (B) Representation of an alveolar Tongue tip constriction location typical of English *ta* or *da*.

instance, a speaker producing VOTs around 45 ms for *ta* has an activation distribution for voicing with a localized peak around that value (versus say, at 70 ms for another speaker or context), as shown by the solid line in Fig. 8A. In contrast, a voiced syllable such as *da* will have an activation peak around 0 ms VOT, shown by the dotted line in

Fig. 8A. The same applies to articulator-specific parameters. Thus, the parameter relevant to the constriction location of the tongue tip is represented by a continuum of constriction locations from dental (most anterior) to post-alveolar (most posterior). An example of a tongue tip constriction location for a typical English *ta* is shown in Fig. 8B. Localized peaks in this axis reflect (language-, lexical item-, and) participant-specific modes for constriction location values, e.g., constriction locations for American English /t/ are more posterior than those of French (Dart, 1998). Overall, $Input(x,t)$ represents three kinds of input, defined in (3).

$$Input(x,t) = r^* \, input_{RESPONSE}(x,t) + d^* \, input_{DISTRACTOR}(x,t) + p^* \, input_{TASK}(x,t) \quad (3)$$

The inputs are added to the field by the terms $input_{RESPONSE}(x, t)$, $input_{DISTRACTOR}(x, t)$ and $input_{TASK}(x, t)$, as required by the particular trial modeled (i.e., for some trials, there is no linguistic distractor and hence the corresponding input would not be present). The first input term, $input_{RESPONSE}(x, t)$, reflects input to the fields specified by the required response, e.g., assuming participants should produce *da* when they see ##, presentation of the visual cue ## introduces a peak of activation in the Tongue Tip field and not the Tongue Back field, as would be the case if the visual cue was associated with *ga* instead. The second input term, $input_{DISTRACTOR}(x, t)$, reflects input corresponding to the perceived distractor, e.g., presentation of an auditory distractor introduces a local peak of activation in its corresponding fields. The other input term, $input_{TASK}(x, t)$, reflects task knowledge and specifies contributions to activation fields based on the participant's expectation of possible responses. For example, in simulating a trial from the articulator experiment where the potential responses within the experimental block are either *ta* or *ka*, small amounts of input are introduced for an alveolar constriction in the Tongue Tip planning field, for a velar constriction in the Tongue Back planning field, and for a voiceless VOT value (e.g., 45 ms) in the Voicing field. For a trial from the voicing experiment where the potential responses are *ta* or *da*, small amounts of input are introduced for an alveolar constriction in the Tongue Tip planning field, and two inputs are introduced to the Voicing field, one for a voiced VOT value (e.g., 5 ms) and another for a voiceless VOT value (e.g., 45 ms). In their trial-initial states, fields are in states of preparedness reflecting the possible responses of the task at hand. The scaling factors r, d, and p, scale the response, distractor, and task inputs, respectively. The response input is scaled such that it is sufficient on its own to generate the necessary peaks of activation to produce the response. The weight of the distractor is strong enough to affect the evolution of the fields without having the participant produce the distractor instead of the required response, which did not happen. The activation strength of the task-knowledge input was the maximum that could be added to the fields without triggering a self-stabilizing peak in any field.

We now turn to the formal component of the dynamics that enables the buildup and stabilization of activation distributions, as opposed to single activation values, over an entire field. Understanding how this is achieved requires considering the interaction component of the dynamics, the $S(x,t)$ term in Eq. (1). Interaction means that the evolution of activation of any given parameter value x depends on its own activation, exactly as with the single parameter exponential growth/decay dynamics, but also on the activation levels of the other parameter values x' within the same field. In other words, field sites are connected and influence the activation of other sites, as in the so-called recurrent networks of connectionist models. In dynamic fields, parameter values excite each other when they are local (nearby one another) and inhibit each other when they are not local (global inhibition). To appreciate what "excite or inhibit" means in the context of an evolving field, recall our main field evolution equation in (1). The interaction term $S(x,t)$ contributes to the rate of activation change denoted by the left hand side of the equation $dA(x,t)$. To say that a field parameter value excites (inhibits) another nearby (far away) parameter value is to say that the former raises (lowers) the rate of activation change of the latter. Locally excitatory and non-locally inhibitory interaction is achieved by Eq. (4) for the within-field interaction (we turn to the cross-field interaction below). This equation represents a convolution operation where the convolution kernel $w(x)$ is applied to a nonlinear transformation of the field expressed by the function f.

$$Interaction_{WITHIN\text{-}FIELD} = \int w(x - x') f[A(x', t)] dx' \quad (4)$$

We first consider the term $f[A(x,t)]$. Not all values of x participate equally in the interaction. Specifically, only sufficiently activated values of x can participate in changes to the field. This is achieved by transforming the activation $A(x, t)$ using some "threshold" function f (for antecedent notions of this by now widely accepted property of neural activation propagation, see Grossberg, 1973). This function admits different implementations. It can be a "hard" threshold implemented by a step function so that $f(A) = 0$ when $A(x, t)$ is less than θ, thus zeroing the transformed activation so that this value of x has no participation in the interaction, and 1 otherwise. Alternatively, it can be a "soft" threshold as specified by the sigmoid in (5), where the term β controls the steepness of the threshold (see Fig. 9). In the neighborhood of θ, the greater the activation, the greater its interactive influence, i.e., the bigger the transformed $f(A)$. As activation gets farther away from θ (farther higher or lower), then $f(A)$ becomes less sensitive to differences in activation and thus such differences have relatively little effect on the strength of their interactive influence. In sum, thresholding ensures that only sufficiently activated (near θ) values of x are instigators of activation change elsewhere in the field and that the strength of their effect on other field locations depends nonlinearly on their activation.

$$f(A) = \frac{1}{1 + \exp[-\beta(A - \theta)]} \quad (5)$$

Given this transformed $f[A(x,t)]$, the interaction induces changes in the field as some value(s) of x approaches the soft threshold (θ). These changes can be either excitatory or inhibitory. Whether it is one or the other and the degree

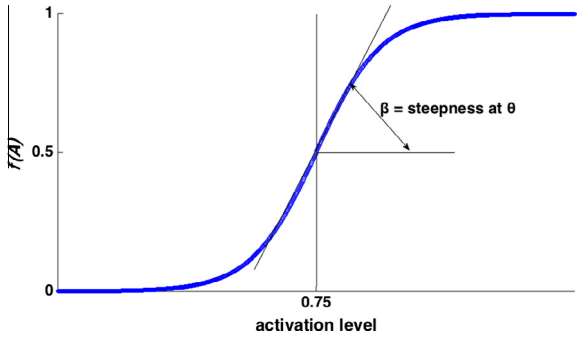

Fig. 9. Sigmoid threshold function defined in (5). The sigmoid function is most sensitive to activation values around θ, which in the model is 0.75. Activation values much lower than θ have no effect on the interaction, while activation values much greater than θ have a uniform, positive effect on the interaction.

of the corresponding activation change is determined by the interaction kernel defined in (6) below, and illustrated in Fig. 10. This kernel consists of two components, an excitatory component expressed by the first positive Gaussian term containing w_{excite} and a second inhibitory component expressed by the $w_{inhibit}$ term.

$$w(x) = w_{excite}e^{-(x^2/2\sigma_w^2)} - w_{inhibit} \quad (6)$$

Whether activation change is excitatory or inhibitory depends on the distance between the values of x partaking in the interaction. Specifically, the convolution kernel's σ_w term defines the width of the excitatory region. For values within a local range defined by σ_w, the kernel is positive and thus excitatory, with w_{excite} being the degree of excitation. Outside of that range, it is inhibitory (the $w_{inhibit}$ term overtakes the positive first term). This is how within-field interaction encompasses both local excitation and lateral inhibition, two properties crucial to the buildup and maintenance of stable local peaks of activation over an entire field. We can now, in particular, understand why input-contributed localized peaks of activation sometimes wane out, with the field relaxing back to its resting level, and other times lead to the generation of a stable peak maintained even after the input has been retracted (in the words of Amari, 1977: 77, 'a fixed size of localized excitation, once evoked by stimulation, can be retained in the field persistently even after the stimulus vanishes'). The two scenarios were illustrated in Fig. 7A and B, respectively. The difference is due to the effects of interaction. Unlike in Fig. 7A, in Fig. 7B the input-contributed activation values were sufficiently high to be above the threshold of the functional term $f[A(x,t)]$ in Eq. (4). This engages the interaction term. Interaction in turn sets up a wave of change throughout the field where local excitation in the neighborhood of a peak sustains local activation levels above values contributed by the input and suppresses activation levels in field locations non-local to that peak. Eventually, the field reaches a stable state which persists even after input has been removed. To appreciate how this happens, consider in Fig. 11A a nonlinear system with two stable fixed points shown by the two filled circles, separated by an unstable fixed point shown by an open circle (for the unstable fixed point, the arrows of the vector field point away from it). This system can be in two possible states given by the two stable fixed points, the lower stable fixed point being the "off" state and the higher stable fixed point being an "on" state. When input of sufficient strength is introduced, it results under appropriate parameter conditions in a change from Fig. 11A to B. The bistable attractor landscape in the vector field has changed qualitatively to one where only an "on" remains at activation values higher than those of the input-contributed activation. This "on" state formally expresses the notion of setting parameter values in our model. It is notable that this qualitative change is caused by a quantitative increase in input strength. Such a change is not possible with the linear dynamics described in Figs. 5 and 6. In those systems, input does not result in changing the number of fixed points. Input only shifts the location of the single fixed point. In sum, the nonlinearity in the dynamics of Eq. (1) endows fields with behaviors not accessible to the linear dynamics reviewed above.

Interaction furthermore endows fields with the capacity to reach stable activation distributions even in the face of input with multiple competing or ambiguous peaks. The case of a single-peaked input was illustrated in Fig. 6B above. As we have seen, given sufficient input strength,

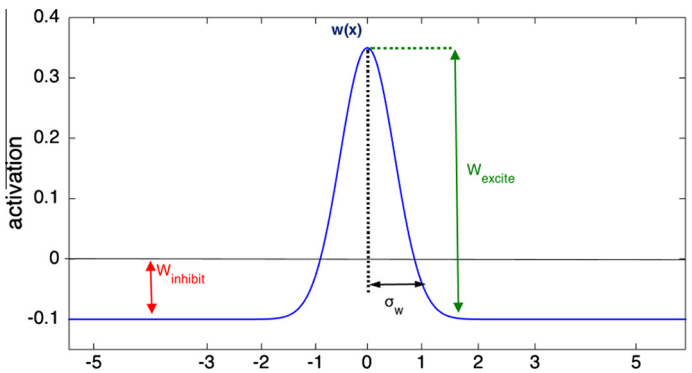

Fig. 10. The interaction term $w(x)$, showing the values for (6) used in the model (see Appendix A). The x axis is defined along arbitrary units of constriction location.

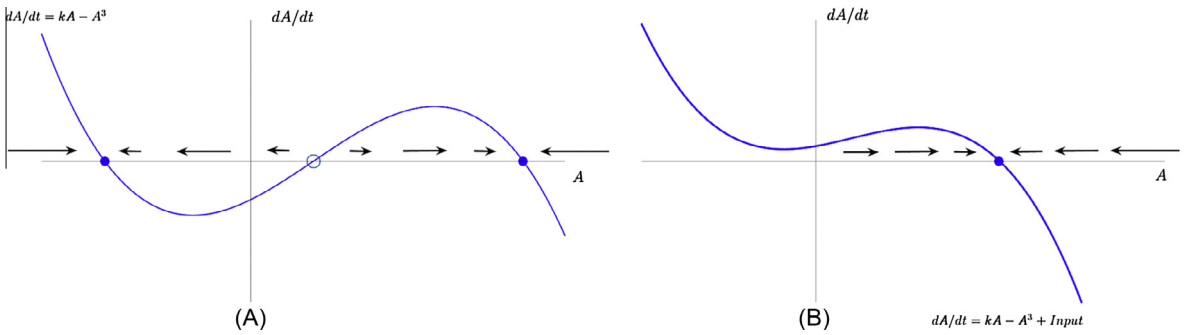

Fig. 11. Nonlinear dynamics with corresponding vector fields. (A) The system $dA/dt = f(A) = kA - A^3$, which is nonlinear due to the cubic term, describes a bistable regime. There are two stable fixed points (filled circles) separated by an unstable fixed point (open circle). (B) With added input, the nonlinear system moves to a regime where only one stable fixed point exists. Unlike in Fig. 6, where added input resulted in no qualitative change to the dynamics, input strength in the nonlinear system results in a bifurcation where the system has changed from two stable fixed points to one stable fixed point. This change corresponds to a form of decision-making and provides a formal expression of the notion of setting a parameter value for the required response in our model.

activation builds up locally as nearby values of x excite each other (local excitation), eventually reaching a stable activation distribution over the entire field. Second, lateral inhibition suppresses activation levels in the field in locations other than those near the activation peak, effectively disallowing two or more self-stabilizing peaks to coexist within a field (formally, this is due to the $w_{inhibit}$ term). Thus, when two inputs to a field are sufficiently distant, for example, as in Fig. 7C with a posterior and an anterior constriction location (or as in the case of one voiced and another voiceless input to the Voicing field, which will be illustrated in Fig. 13B below), both peaks inhibit each other due to lateral inhibition (as illustrated in Fig. 7C). In terms of deciding between the two peaks, it is the relative strengths, widths, and timing of the two competing inputs plus the noisy evolution of activation that determine which peak wins. In terms of RTs, this means that whichever peak ultimately stabilizes takes longer to do so than it would have without the other, incongruous input.

Our interaction term also involves a component introducing interactions among different fields, in the form of cross-field inhibition. This is necessitated by two considerations. First, unlike the basic model of Dynamic Field Theory with one field (Erlhagen & Schöner, 2002), in the case of speech we have multiple fields representing the multiple organs or articulators. Second, specific task demands of the task we model impose the specific constraint that required responses involve at most one supra-glottal articulator, e.g., Tongue Tip, Tongue Back, or Lower Lip. Cross-field inhibition is indicated in Fig. 4 by the bidirectional arrows between articulator fields. That the cross-field interaction in our model takes the form of inhibition (and not both excitation and inhibition as with the within-field interaction) is because of this constraint. Cross-field inhibition, that is, is necessary to effect this exclusivity condition among the different articulators. Thus, each articulator field inhibits the activation level of the other two articulator fields when the inhibiting field's activation level rises above a cross-field threshold χ. Unlike the soft, field-internal threshold θ, the cross-field threshold χ is a hard threshold, meaning that no cross-field inhibition is introduced until some activation value of some articulator planning field passes χ. Thus, $interaction_{CROSS\text{-}FIELD}$ was defined such that at each time step t in the evolution of the field, if the maximum activation value is greater than or equal to χ in a given field, the activation levels for all values of x in the other two articulator planning fields are reduced by a set amount. In sum, the fully expanded form of the interaction term $S(x,t)$ from (1) reads as in (7).[2]

$$S(x,t) = interaction_{WITHIN\text{-}FIELD} + interaction_{CROSS\text{-}FIELD}(x,t) \quad (7)$$

Functionally, the model sends production values to implementation at the point when the Voicing planning field and one articulator planning field achieve a stable "on" state. This is determined in the model by a Monitor, which waits until the activation level for some x value in both the Voicing field and one articulator field (one of the Lower Lip, Tongue Tip, or Tongue Back) reach a criterion value κ. The numerical value of κ in the model serves as a computational convenience for indicating that once some activation level of an x value has achieved κ, the field will inevitably stabilize with an "on"-state peak. At that point the Monitor chooses the parameter values x with the highest activation level from those two fields (voicing and the constriction location of one of the articulator fields) to be sent to implementation. The time step in the evolution of the model at which the Monitor make this choice serves as the RT on that trial. In other words, the intention to produce a particular combination of constriction and voicing values reaches a stable state, which drives the implementation of that constriction and voicing combination. Given the behavior of the Monitor, whichever field evolves more slowly determines the RT on the trial.

[2] Note that the equation that defines the evolution of the Voicing field differs from the one that defines the evolution of the articulator fields (7) only in that its interaction term $S(x,t)$ does not contain a term for cross-field inhibition. The Voicing field neither inhibits nor is inhibited by any other planning field, since it is not an articulator and functions independently of which primary oral articulator is involved in the utterance. This design reflects the fact that voicing and articulator are cross-classifying parameters for English consonants (Chomsky & Halle, 1968; Ladefoged, 1999).

As will become clear below, sometimes it is the Voicing field and sometimes it is an articulator field that evolves more slowly.

To sum up, the proposed model of phonological planning provides a formal and computationally explicit instantiation of how perception affects the online buildup of phonological plans in the response–distractor task. We highlight here the essential properties of our model and how those properties set it apart from other models in the speech motor control and phonological literature. The most important property of our model is time dependence. In our model and using the example of a lexical item containing a syllable *ta*, the tongue tip constriction location, constriction degree, and voicing parameters for this /t/ are not statically assigned to their canonical values. Rather, assigning values to these parameters is a time-dependent process, captured as the evolution of a dynamical system. This system governs how the activation distributions in the planning fields representing parameters change in time. Thus, activation distributions like those shown in Fig. 7 are not static but evolve in explicit ways. Our model's time dependence in setting parameter values stands in contrast to other formally explicit models with components devoted to the control and execution of speech movements (Browman & Goldstein, 1990; Guenther, 1995; Saltzman & Munhall, 1989). In these models, assignment of values to parameters is instantaneous. Of course, movement execution in these models does unfold in time, but with parameters such as target location and stiffness set from the start and kept fixed during the lifetime of the movement. That is, in these models the targets arrive fully specified from some preceding sources, usually taken from the phonological inventory of the language (a notable exception is Nam & Saltzman, 2003, on setting the parameters of temporal coordination of gestures). However, models of phonological representation that could produce such targets (e.g., Browman & Goldstein, 1989; Chomsky & Halle, 1968) have no formal notion of the time course by which those representations are assembled.

Finally, the key representational unit in our model is that of the dynamic field. Fields are continuous (in the parameter space they represent), self-stabilizing, interactive, and noisy, in ways explicitly captured by the dynamics we have described in this section. Using fields is a generalization of a similar idea put forth in Byrd and Saltzman (2003), where gestural parameters are stored as ranges of possible values. In our model, each range is approximated by an activation field in memory; hence, there is a range of values but also activations associated with those values and of course dynamics governing the evolution of activation values on top of that range. Representing targets by activation fields is also a generalization of two well-known proposals about the nature of speech targets, Keating's "windows" (Keating, 1990) and Guenther's "convex regions" (Guenther, 1995). In Guenther's model of speech production, speech targets take the form of convex regions over orosensory dimensions. Unlike other properties of targets in Guenther's model, the convexity property does not fall out from the learning dynamics of the model. Rather, it is an enforced assumption. No such assumption about the nature of the distributions underlying target specification needs be made in our model.

Simulations

We now turn to illustrating the model at work. In doing so, we simulated the results from the articulator and voicing experiments of Roon and Gafos (2015), as well as those reported by Galantucci et al. (2009). For the purpose of illustration and without loss of generalization, we take the required response on all simulated trials to be *ta*. Therefore in the simulations of these experiments, the distractor in the Identity condition was *ta*, the Tone condition represented the case of a non-speech distractor, and the Incongruent condition distractor was *ga*.[3] In the simulated experiments, the distractor was introduced 250 time steps after the start of the trial and 150 time steps after the presentation of the visual cue, reflecting its timing relative to the presentation of the visual cue in the actual experiments (i.e., a positive SOA). The only differences between the two simulations were that the Congruent distractor was *ka* in the simulated voicing experiment and *da* in the articulator experiment, and that the task-knowledge inputs reflected the possible responses of *ta* or *da* for the voicing experiment but *ta* or *ka* for the articulator experiment. The values for all of the model parameters used in the simulations are found in Appendix A. A link to the MATLAB scripts (MATLAB 2014, The MathWorks Inc., Natick, MA) can be found in Appendix B.

Trial simulations by condition

Fig. 12 illustrates evolutions of the planning fields during a single trial in each of four experimental conditions: the Tone, Congruent, and Incongruent conditions from the voicing experiment of Roon and Gafos (2015), plus the Identity condition from Galantucci et al. (2009). Each panel in Fig. 12 shows how the maximum activation level for the four planning fields unfolds as a function of time steps in the model. The black line shows the evolution of the Tongue Tip field, the light gray line shows the Lower Lip field, the dark gray line shows the Tongue Back field, and the blue line shows the Voicing field. Differences in the rate of rise of the maximum activation level of the fields predict differences in experimental RTs.

We begin with Fig. 12A, which shows the evolution of the four planning fields in the Tone condition. Since the tone distractor is not a speech syllable, the behavior of the fields in this tone condition serves as a baseline reference to how the planning fields evolve in the other conditions with a speech distractor. On all trials simulated in Fig. 12, the response always involves the tongue tip, but a voiceless (*ta*) or voiced (*da*) response is equally likely. Therefore, at the start of the trial, the Tongue Tip and Voicing fields have higher activation levels than the Lower Lip and Tongue Back fields due to the task input, since no possible response will involve the lower lip or the tongue back

[3] The response in the Galantucci et al. (2009) experiment was *da*, not *ta*, and the Incongruent distractor common to both experiments in Roon and Gafos (2015) was *ba*, but these differences are immaterial in the model.

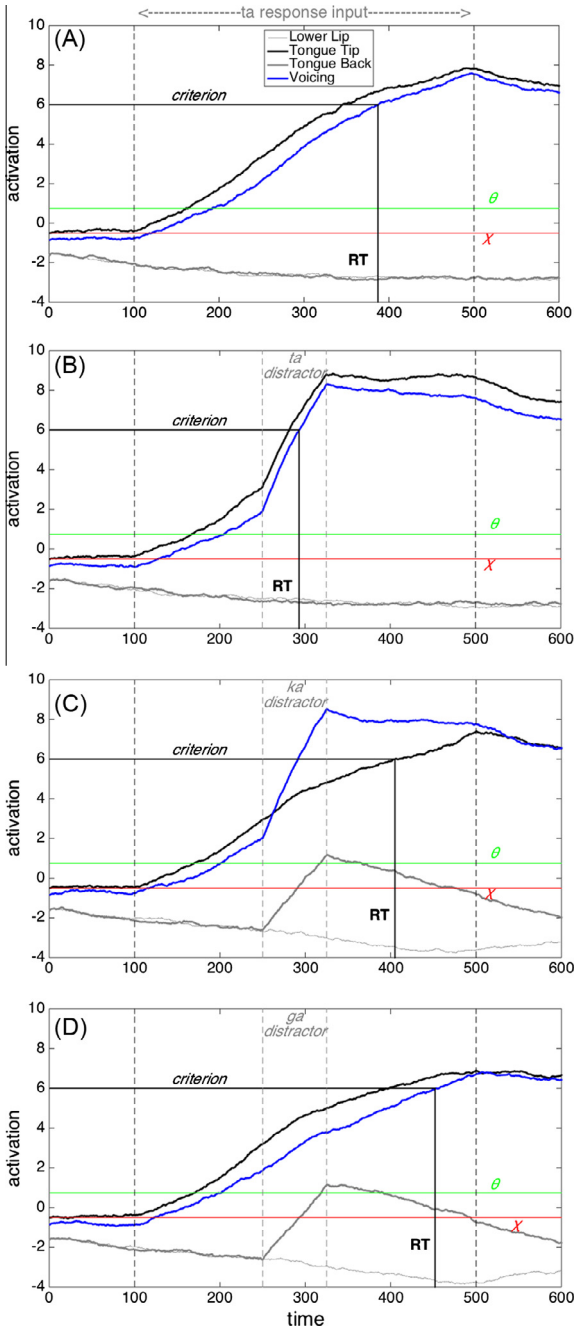

Fig. 12. Evolution of planning fields in individual simulated trials from four experimental conditions. (A) The non-speech Tone condition. (B) The Identity condition. (C) The Congruent condition (matched voicing, mismatched articulator). (D) The Incongruent condition. Vertical black dashed lines at time steps 100 and 500 indicate the duration of the response input. Vertical gray dashed lines (B–D) at time steps 250 and 325 indicated duration of the distractor input. A vertical black solid line indicates the response time (RT) on each simulated trial. The within-field threshold (θ) is indicated by the horizontal green line. The cross-field threshold (χ) is indicated by the horizontal red line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(the Voicing field has a slightly lower activation level for reasons we explain in the section "*Unknown voicing vs. unknown articulator*" below). The activation levels of the Tongue Tip and Voicing fields start to rise at time step 100, the point at which the participant begins planning the required utterance based on the appearance of the visual cue on that trial (here ## instructing the participant to say *ta*, indicated by the vertical black dashed lines at time steps 100 and 500, indicating the duration of the response input resulting from this visual cue). The horizontal green line drawn at activation level 0.75 indicates the value of the soft threshold (θ) that determines the engagement of the within-field interaction term. The Tongue Back and Lower Lip fields receive no input, apart from random fluctuations due to stochastic noise. The cross-field inhibition threshold (χ) is indicated by the horizontal red line drawn at activation level -0.5. As the activation level of the Tongue Tip field increases and continues past χ, it takes away activation from the Tongue Back and Lower Lip fields, as a result of this cross-field inhibition. The Tongue Tip and Voicing activation levels continue to rise until they both have passed the criterion value (κ), indicated by the black line drawn at activation level 6. The time step at which the second field passes κ is marked as the RT on that trial (the vertical line at about time step 390). At that time step, the Monitor takes the maximum parameter values from the Voicing and Tongue Tip fields and passes them to implementation.

Fig. 12B shows the evolution of the fields in the Identity case from the experiment of Galantucci et al. (2009). In this case, participants are required to respond with *ta* and the distractor is also *ta*. From time step 0 to 250, all fields evolve in the same way as in the Tone condition. The distractor is presented at time step 250, thus the vertical gray dashed lines at time steps 250 and 325 indicate the duration of the input from the distractor. In this condition the distractor inputs are the same as those for the response. Therefore, the activation level for the Tongue Tip and Voicing fields rises at a much faster rate than in the Tone condition because both inputs add activation to the same range of parameter values, in addition to the local excitation being generated by the interaction term. Both fields therefore cross κ earlier than in the Tone condition, and the simulated RT is shorter, around time step 290.

Fig. 12C shows the evolution of the fields in the Congruent case (from the voicing experiment of Roon & Gafos, 2015) on a trial with a *ta* response and *ka* distractor. Since the response and distractor share the same voicing, the evolution of the Voicing field in this condition is qualitatively the same as in the Identity case. The evolution of the Tongue Tip field is different, however. When the distractor input starts at time step 250, the activation level of the Tongue Back field begins to rise, and eventually crosses χ, introducing cross-field inhibition to the Tongue Tip field. The distractor input ends at time step 325, but by that time the Tongue Back field maximum is above θ, so it maintains a somewhat elevated activation level for some time due to the interaction term, and the cross-field inhibition of the Tongue Tip field by the Tongue Back

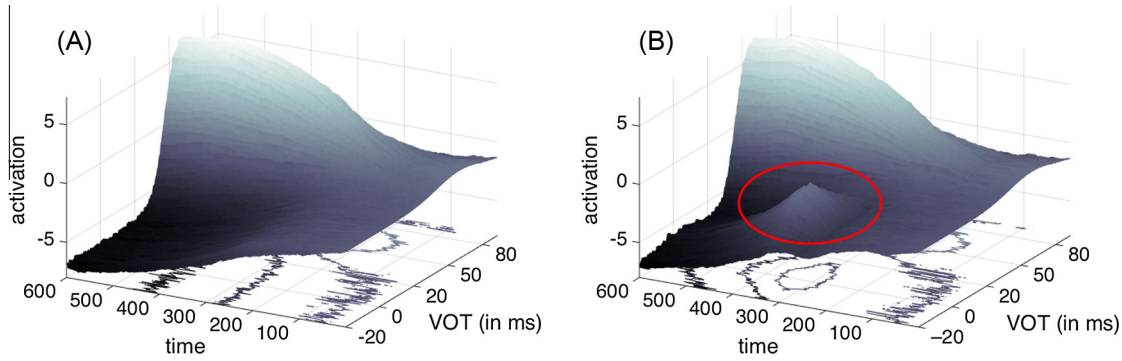

Fig. 13. Evolution of the Voicing field for a *ta* response in two conditions from the voicing experiment: (A) the Tone condition, in which there is no linguistic distractor, corresponding to Fig. 12A, and (B) the Incongruent condition, where the voicing of the *ta* response and the *ga* mismatch in voicing, corresponding to Fig. 12D. The red circle indicates the incongruent voicing introduced by the *ga* distractor during the ongoing planning of *ta*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

field therefore persists. As a result, the rate of rise of the Tongue Tip field activation level slows down compared to its rise in the Tone condition. Due to the cross-field inhibition introduced to the Tongue Tip field by the Tongue Back field, the Monitor has to wait longer for the Tongue Tip field to cross κ, and thus the RT on this trial is longer than in the Tone condition, in this case at about time step 405.

Lastly, the evolution of the fields in a trial from the Incongruent condition is shown in Fig. 12D, with a *ta* response and a *ga* distractor. The evolution of the Tongue Tip field is effectively the same as in the Congruent condition, due to the cross-field inhibition introduced from the mismatching articulator of the distractor. The RT on this simulated trial is determined by the relatively slow rate of evolution of the Voicing field, which is due to the incompatible Voicing input from the *ga* distractor.

Whereas Fig. 12 shows the evolution of the maximum activation level for each of the four planning fields, Fig. 13 illustrates the effects of incompatible inputs introduced to the same field. Fig. 13A shows the evolution of the Voicing field for the Tone condition. Activation as a function of time is now shown throughout the entire range of VOT values. The single input corresponding to a voiceless response contributes a peak of activation whose mean VOT value is near 50 ms. Given the within-field dynamics and the lack of any other input from a speech distractor in this condition, the field rises quickly to a self-sustained maximum activation around that VOT value. However, fields do not simply reproduce input. The Voicing field evolution in this single input case is contrasted in Fig. 13B with its evolution in a condition where competition leads to decision of one versus another peak when within-field lateral inhibition is engaged. Fig. 13B shows that the introduction of distractor input with incongruent voicing (*ga*) results in two peaks of activation forming in the Voicing field, a large peak in the voiceless end of the VOT continuum for the required response (*ta*) and a second, smaller peak at the voiced end of the continuum for the distractor (*ga*). These peaks inhibit each other due to lateral inhibition (as seen in Fig. 7C). The rate of rise for the voiceless response required for *ta* therefore is lower than in the neutral Tone condition (as can be seen by comparing the rise in activation of the Voicing field Fig. 12A and D). As a result, the Monitor has to wait longer for the Voicing field to reach κ, which it does at about time step 450.

Simulation results

The RTs predicted in the dynamical model of phonological planning are determined by the totality of deterministic relations and interactions between the model components shown in the box diagram of Fig. 4 and explained above, but they are also affected by non-deterministic or stochastic forces in the model dynamics. Hence, the model's efficacy in capturing the range of past experimental results can be determined by sampling across many repetitions of actuating or simulating the individual trial conditions. The relative arrangement of RTs across the different simulations are then compared to those obtained in experimental data.

The results of the model simulations of both the voicing and articulator experiments from Roon and Gafos (2015) and the experiment from Galantucci et al. (2009) are shown in Fig. 14. Each experiment included 150 simulated trials for each of four conditions: Identity, Tone, Congruent, and Incongruent, yielding 600 trials per simulated experiment. On each trial, the RT was calculated as the time step at which both the Voicing field plus one articulator field reached criterion, minus 100, since that is the time step at which the cue is presented. The Identity condition (i.e., response *ta*–distractor *ta*) yielded the fastest RTs, which were shorter than a neutral Tone. The Congruent condition (i.e., *ta*–*ka* or *ta*–*da*, respectively) had RTs slower than in the Tone condition, but faster than in the Incongruent condition (i.e., *ta*–*ga*). This is the same relative arrangement of RTs found in the experimental results.

The Identity condition had the fastest simulated RTs because only in that condition were all inputs to the fields mutually reinforcing. This resulted in RTs faster than in the Tone condition, in which there was neither inhibiting nor reinforcing inputs. The slow-down in the Congruent condition relative to the Tone condition has its source in different model components. Specifically, in the articulator

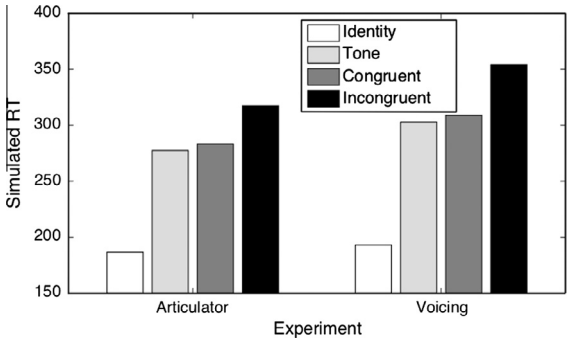

Fig. 14. Results of model response time simulations of the articulator (left) and voicing (right) experiments. Distractor conditions: Identity (white bars) was when the distractor was the same as the response (e.g., *ta–ta*); Tone (light gray bars) was when there was no linguistic distractor; Congruent (dark gray bars) was when the distractor either mismatched the response in voicing but matched in articulator for the articulator experiment (e.g., *ta–da*) or when the distractor mismatched the response in articulator but matched in voicing for the voicing experiment (e.g., *ta–ka*); Incongruent (black bars) was when the distractor and response mismatched in both voicing and articulator, (e.g., *ta–ga*).

experiment simulation, the slow-down was the result of the within-field inhibition introduced by the mismatched voicing between the response and distractor (*ta–da*). In the voicing experiment simulation, the slow-down is due to the cross-field inhibition introduced by the mismatched articulator between the response and distractor (*ta–ka*). In both experiments, the slow-down of RTs for the Incongruent condition was due to the combination of cross- and within-field inhibition introduced by the mismatch in both articulator and voicing (*ta–ga*).

Unknown voicing vs. unknown articulator

Independent from the effects of articulator and voicing congruency, the experiments in Roon and Gafos (2015) revealed an unexpected result. Speakers responded slower when they did not know the voicing of the planned response than when they did not know the articulator. Specifically, in the voicing experiment of Roon and Gafos (2015), participants always knew the primary articulator in their response but the probability of the voicing for a given response was 50–50% (e.g., *ta* or *da*). In the articulator experiment the reverse was true. The voicing parameter for the response was known, but participants could not predict which of two articulators would be needed in their response (e.g., *ta* or *ka*). RTs in the experiment where voicing was unknown were 52 ms slower on average than in the experiment where the articulator was unknown (compare Fig. 3A and B), independent of distractor condition. This result was new. It was also not predicted by any model or theory of speech production.

This cross-experiment difference was replicated by our model, as can be seen in Fig. 14. RTs for the voicing experiment simulations were longer across the board than those for the articulator experiment. Note that this was not an effect of distractor—it applied across distractor conditions, including the Tone condition, just as in the experimental results. The source of the cross-experiment RT differences in the model lay in the difference between the trial-initial states of the planning fields due to differences in task-knowledge inputs. Fig. 15 illustrates the differences between the trial-initial states of the simulations of the two experiments. The top panel of Fig. 15A shows the trial-initial state of the VOT planning field in the articulator experiment, in which the voicing of the response was known and the possible responses were *ta* or *ka*. A single peak of activation was introduced in each of the Voicing, Tongue Tip and Tongue Back fields. The bottom panel of Fig. 15A shows that the maximum activation level of each of those three fields was higher than the resting activation level shown for the Lower Lip field, which received no trial-initial input, since no possible lower lip response was anticipated. The top panel of Fig. 15B shows that the trial-initial state of the Voicing field was different in the voicing experiment, in which the voicing of the response was not known and the possible responses were *ta* or *da*. The critical difference was that in the voicing experiment there were two small peaks introduced into one field, due to the equal probability of a voiced or voiceless response on each trial, whereas in the articulator experiment no one field received two incompatible trial-initial inputs. The introduction of two inputs of trial-initial, incompatible activation to the Voicing field resulted in peaks, albeit small ones, that were sufficiently close to the threshold θ to introduce some lateral inhibition in the field. This lateral inhibition entails two small activation peaks inhibiting each other and lowering the overall level of activation in the Voicing field at the start of the trial, as can be seen by comparing the trial-initial maximum activation of the Voicing field (represented by the blue lines) in the bottom panels of Fig. 15A and B. In contrast, the dynamics of the cross-field inhibition are different, and did not depress the trial-initial state of activation in any field. Since the Monitor requires a Voicing value before sending parameter values to implementation, it had to wait longer for the Voicing field to stabilize in all conditions because the trial-initial state of the Voicing field was lower in the voicing experiment (Fig. 15B) than in the articulator experiment (Fig. 15A).

The different natures of the within- and cross-field inhibition in the model were designed to meet different functional and theoretical requirements. In the unknown-articulator case, speakers must be prepared for one or the other response on each trial and produce the required response as quickly as possible upon seeing the cue. In the trial-initial state, higher activation levels introduced by the task input reflect this state of preparedness (Kornhuber & Deecke, 1965). Crucially, concurrently higher activation levels in multiple articulator planning fields do not run afoul of any fundamental representational principles. In other types of tasks or utterances, articulator planning fields do not inherently inhibit each other. For example, many speech sounds require concurrent constrictions of multiple articulators, e.g., concurrent lip rounding along with tongue tip and tongue back constrictions for English /ɹ/ (Campbell, Gick, Wilson, & Vatikiotis-Bateson, 2010). In such a case, concurrent activation is desirable, and cross-field inhibition would be detrimental. The cross-field inhibition of our model is therefore specific to

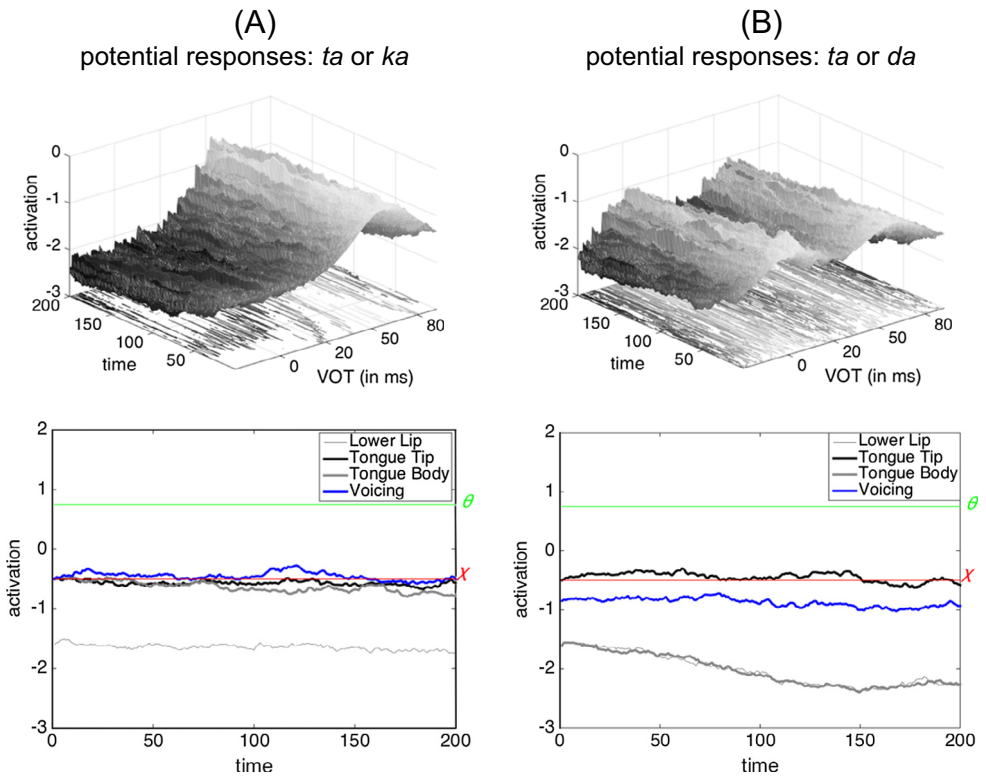

Fig. 15. Effects of task knowledge on the trial-initial state of the Voicing planning field. All fields reflect the activation levels in the absence of input other than task knowledge, i.e., no response or distractor input. (A) The top panel shows the trial-initial state of the Voicing planning field in the articulator experiment. The bottom panel shows the maximum activation levels of three articulator fields (gray scale) and the Voicing field (blue). (B) The top panel shows the trial-initial state of the Voicing planning field in the voicing experiment. The bottom panel shows the maximum activation levels of the fields. In the bottom panels of A and B, the within-field threshold (θ) is shown by the green line at 0.75, and the cross-field threshold (χ) is shown by the red line at −0.5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

this task and the utterances involved, and serves to suppress potential but not cued articulators as quickly as possible once sufficient evidence for the cued articulator has built up.

The unknown-voicing case is different. No utterance can be both voiced and voiceless. Preparing two inherently conflicting responses by introducing incongruous inputs to one field (here, the Voicing field), violates this basic principle of phonological representation. Planning fields serve the purpose of determining a single production value based on one or more potentially conflicting—and mutually exclusive—inputs. Although in the model the within-field threshold θ is numerically higher than the cross-field threshold χ, the interaction term in fact influences the field when activation levels are lower than θ, with the result that the within-field inhibition affects the fields at lower activation levels than the cross-field inhibition does, as illustrated above. The design of the planning fields, including the fact that within-field inhibition engages early on in planning as part of the inherent field dynamics, therefore reflects this basic representational principle. This representational constraint cannot be superseded or modified by task demands.

Thus in the model, slower RTs in the experiment with unknown voicing are not caused by unknown voicing *per se*, but rather by incompatible inputs to one planning field—that is, two inputs that are inherently mutually exclusive—at the beginning of the trial. For example, just as no segment can simultaneously be voiced and voiceless, the tongue tip cannot simultaneously make dental and post-alveolar constrictions. The model predicts that RTs should be similarly modulated regardless of the field that receives such conflicting inputs, and that conflicting inputs to separate fields should not slow down RTs to the same degree.

Some support for this model prediction can be found in the results reported in Roon, Klein, and Gafos (2014), which used fricative-initial responses and distractors with the same participants as in the voicing experiment from Roon and Gafos (2015). Specifically, in this experiment, responses were either *fa* or *ʃa* (“*sha*”), and distractors were either *ha* or *sa*, in addition to the neutral tone and no distractor conditions. The two potential responses involved two different articulator planning fields (Lower Lip for *fa*, Tongue Tip for *ʃa*) with fixed voicing throughout the experimental session. The model predicts that RTs should be faster in this experiment than in the voicing experiment from Roon and Gafos (2015), where the articulator was known but the voicing was not. The reason for this prediction can be traced to differences in the trial-initial states across

the two experiments and the implications of these differences in terms of field evolution, as discussed above. The trial-initial state of the Voicing planning field of the voicing experiment from Roon and Gafos (2015) is as shown in Fig. 15B, since there was conflicting task-knowledge input given the uncertain voicing of the response. The trial-initial state of the Voicing field in the model of the fricative-initial experiment is the same as in Fig. 15A, since the voicing of the two possible responses was the same. Therefore, no conflicting inputs were introduced to the trial-initial state of any field.

While these data are useful because they allow for within-subject comparisons, we note that for phonetic reasons it is not possible to definitively compare RTs between stop- and fricative-initial utterances with acoustic data only. The onset of a fricative has an acoustic consequence of measurable aperiodic noise in a spectrogram. The acoustic consequence of a stop onset is silence. Since the utterances in these data were not preceded by any other sound, it is impossible to determine when the oral closure for stops took place. This issue can be illustrated by results from a study by Rastle, Croot, Harrington, and Coltheart (2005), which also illustrates some important facts that bear on interpreting our data. That study used a delayed-naming task in which English speakers saw a cue indicating the syllable they were about to say, and then waited for a go signal before speaking. Rastle et al. (2005) measured the latencies of two acoustic landmarks from that go signal for each onset: the onset of acoustic energy of any kind, and the acoustic start of the vowel. Latencies for the onsets relevant to the experiments discussed here (before /əː/ and /a/ only) are shown in Fig. 16. For the stops /g, k, d, t/, the acoustic onset, indicated by the number at the left edge of each gray box, indicates the release of the oral closure. For fricatives /f, ʃ/ the acoustic onset indicates the beginning of frication. The right edge of each gray box indicates the onset of phonation for the vowel regardless of manner. Therefore, the number inside each gray box indicates VOT for the stops and frication duration for the fricatives. Differences in latencies are indicative of inherent properties of producing those onsets, since all planning was presumably complete at the time of the go signal. While the onset of aperiodic energy indicates the achievement of a constriction for fricative-initial utterances (represented by the gray bars in the bottom two rows of Fig. 16), stop-initial utterances begin with silence, reflecting the closure of the vocal tract (represented by the cloud in the top of Fig. 16). As noted above, the achievement of that closure cannot be determined from the acoustics.

There are two ways to explore whether these two sets of RTs from Roon et al. (2014) and Roon and Gafos (2015) are consistent with the prediction of the model. The first way to compare the two experiments is to estimate the closure duration of the stop-initial utterances (i.e., the size of the "cloud" in Fig. 16 for each stop) and subtract those estimates from the RTs reported in the Roon and Gafos (2015) voicing experiment, which were calculated from the release of the oral closure. These adjusted RTs can then be compared to the onset of aperiodic energy of the fricative-initial responses from Roon et al. (2014). A study of the acoustic closure durations for

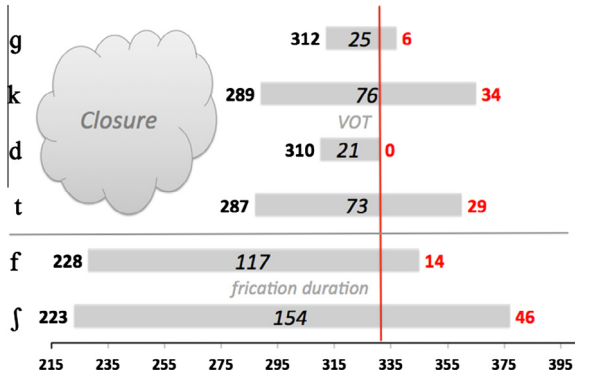

Fig. 16. Naming latencies (ms) for English CV syllables by initial consonant, as reported by Rastle et al. (2005). Boxes start at the average acoustic onset for each consonant (leftmost number) and end at phonation onset. Numbers inside boxes show the average frication duration of fricatives and VOT of stops. Rightmost number indicates phonation onset lag vs. /d/.

American English stops from an extremely large spoken corpus by Byrd (1993) reported the following closure durations: /t/ = 53 ms, /d/ = 52 ms, /k/ = 60 ms, and /g/ = 54 ms. Fig. 17A shows the comparison across the experiments, broken down by response. Based on this estimation, the RTs in the fricative-initial experiment were 39 ms shorter than the stop-initial experiment, per the prediction of the model.

A second way to estimate an appropriate comparison between the two sets of RTs is to measure RTs from the onset of phonation for the vowel, which is an acoustic landmark common to each experiment (represented by the end of all of the gray bars in Fig. 16). However, Rastle et al. (2005) showed that there are inherent differences in naming latencies based on the initial consonant that would need to be taken into consideration in such a comparison. The shortest time to phonation in the Rastle et al. (2005) data was 331 ms for /d/-initial responses (marked by the vertical red line). The rightmost (red[4]) numbers in Fig. 16 indicate the difference between phonation onset times of the vowel in that consonant context compared to the baseline of /d/, i.e., the "phonation onset lag". RTs for all trials for both the stop- and fricative-initial experiments were then recalculated by subtracting the corresponding phonation onset lag from each trial. Fig. 17B shows the mean RTs adjusted for phonation onset lag. By this measure, RTs in the stop-initial experiment (gray bars) were still longer than in the fricative-initial experiment (white bars), here by 25 ms.

While these analyses should be interpreted very cautiously and statistical assessment would not be appropriate, both comparisons provide tentative support for the prediction of the model that RTs should be shorter when the potential responses for a given trial do not involve parameters that are inherently mutually exclusive, compared to trials that force a choice between mutually exclusive parameters. There are many ways to test this

[4] For interpretation of color in Fig. 16, the reader is referred to the web version of this article.

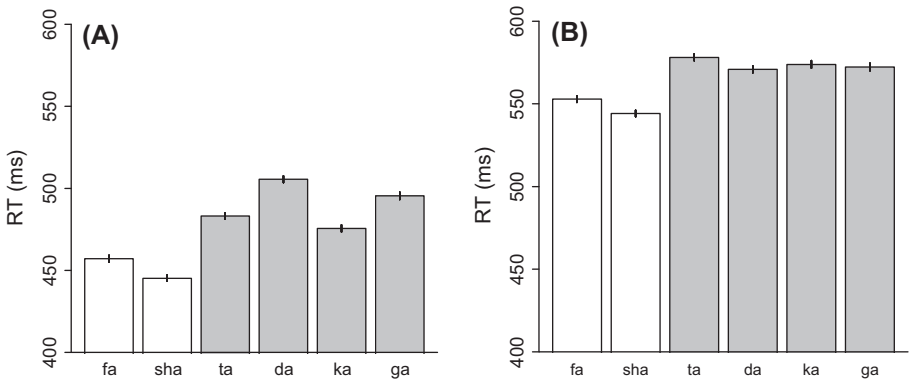

Fig. 17. Comparisons of mean RTs of Roon et al. (2014, white bars) with the Voicing experiment from Roon and Gafos (2015, gray bars). (A) RTs measured from frication onset compared with estimated RTs for the voicing experiment adjusted for closure duration. (B) All RTs measured from phonation onset and adjusted for reported intrinsic RT differences.

prediction using a within-subject design. One would be to combine tasks from the two different experiments of Roon and Gafos (2015). The same participants in one experiment would produce in one task *ta* or *da* and in another task *ta* or *ka*. Another way would be to compare RTs from a task in which potential responses are *θa* or *fa* with RTs from a task in which the potential responses are *θa* or *sa*. RTs in the former should be faster than in the latter, since *θa* and *sa* require different constriction locations of the same articulator (the tongue tip), while *θa* and *fa* require different articulators (the tongue tip and lower lip, respectively).

Accounting for additional experimental data

The response–distractor task and the results from the studies that adopt it offer a rich but sufficiently coherent dataset that makes model development possible. The model we have developed on the basis of this dataset formally instantiates, for the first time, how ongoing response planning is affected by perception and accounts for a range of results reported across several previous studies. It is specifically the time course dimension in setting phonological parameters for production while listening to speech that our quantitative model simulations above have focused on. However, the basic principles of the model we have developed, especially, time-dependence and local excitation/lateral inhibition, are not bound to a specific task. In this section, we show that these principles can be used to develop accounts or derive new predictions for a variety of other experimental settings. What follows serves to demonstrate further the nature of the model's principles as well as the model's promise in elucidating other aspects of the link between perception and production in speech.

Effects of within- and across-category variation

In the experimental results we have discussed so far, the distractor stimuli had fixed VOT values. It is a prediction of the model that distractor and response VOTs do not need to be identical in order to excite each other. Speed-up in RTs for congruent response–distractor pairs should be observed even in the presence of variability in the phonetic detail of the distractor stimuli.

The continuous representations used in our dynamic fields provide a formal way of simultaneously accommodating both the categorical nature of phonological contrasts, e.g., voiced /d/ vs. voiceless /t/, and the variation in phonetic detail within a given category, e.g., VOT. Thus, within any given category, say, the voiceless, the continuous difference in VOT of $/t/^{vot\,=\,60\ ms}$ and $/t/^{vot\,=\,80\ ms}$ are close enough that activation of one value increases neighboring voiceless activation levels, via local excitation. Across the two categories, an exemplar of a voiced /da/ and a voiceless /ta/ occupy two regions in the VOT continuum that are sufficiently distant from each other so that activation of one results in suppression of the other, via lateral inhibition. "Close enough" in our description of local excitation above is elaborated in the model by the kernel term of the interaction, which is parameterized for distance within the relevant phonetic space (here, VOT) and also for the slope of excitation as a function of distance (thus effecting more or less excitation, depending on distance). Local excitation and lateral inhibition predict specific effects of distractors on responses. Hearing a distractor with a mismatched voicing category (e.g., *da–ka*) should result in slower RTs than in matched distractor–response pairs (e.g., *ta–ka*), due to lateral inhibition between the distractor and response VOTs. The same applies when the mismatch is in terms of articulator. These are the results of Galantucci et al. (2009) and Roon and Gafos (2015) that we have focused on so far. In these experiments, the stimulus for a given distractor always used the same sound file, and thus had the same phonetic properties. As we have seen, the model predicts that phonetic variability in VOT within voicing category should not affect the inhibition effects introduced by another parameter, e.g., articulator. That is, RTs should be longer for response–distractor pairs like *ka–ta* or *ta–ka* than with pairs like *ka–ka* and *ta–ta*, even if the voiceless distractors vary in their specific VOT within the voiceless range.

An experiment by Klein et al. (2015) tested this prediction in a response–distractor task with German speakers. Distractor stimuli were *ta* and *ka*. In contrast to other

response–distractor experiments, the VOT of the distractors was not kept fixed. Specifically, for each distractor type, *ta* and *ka*, six stimuli were generated with VOTs ranging from 45 to 120 ms in 15-ms steps. Participants always responded with *ta* or *ka*. The predictions of the model were borne out. RTs were slower when the distractor and response mismatched in articulator than when they matched in articulator. This replicates the articulator effect reported by Galantucci et al. (2009) and Roon and Gafos (2015). The Klein et al. (2015) results further extend that finding, showing that this effect of articulator congruency is obtained despite within-category variation in VOT. Specifically, VOT step did not interact with articulator congruency regardless of whether the distractor matched (e.g., *ta–ta*) or mismatched (e.g., *ta–ka*) the articulator of the response. In sum, as predicted by our model, robust congruency effects of articulator are obtained regardless of the within-category variation in VOT.

Accounting for multiple (mis)articulations

We noted above that some of the properties of the model accounting for the RT results in the core datasets from the response–distractor experiments we have reviewed above are task-specific. These properties include the variable values of the cross-field inhibition and the functioning of the Monitor. We first describe the way in which these properties reflect task-specific constraints and then turn to how lifting these constraints or imposing different constraints offers a handle to accounting for data from other experimental tasks.

In the experimental datasets modeled above, all of the responses involved syllable-initial stops that have only one oral articulator. No response consonant required multiple oral constrictions, as would be the case for consonants such as /w, l, or r/ in English or doubly-articulated / $\widehat{kp}, \widehat{gb}$ / in, e.g., Yoruba (Ladefoged & Maddieson, 1996). The cross-field inhibition for stops with one primary oral articulator may not be the same as for stops involving multiple oral articulators. In addition, the stimuli in these experiments were designed such that when participants realized they had to produce a stop with one articulator, it was also clear that the other articulators would not be needed. Therefore, the specifics of the task in the response–distractor studies we have considered so far implicate a stricter form of cross-field inhibition than in normal speech production, though this was not tested explicitly.

The function attributed to the Monitor in the model may well be task-influenced if not task-specific. In response–distractor experiments, participants were instructed to reply as quickly as they could after the display of the cue indicating the response on that trial. The Monitor criterion variable (κ) was set to an activation value where it could be safely assumed that an articulator field and the Voicing field would stabilize once that value was passed. It seems reasonable to expect that in a different task, the read-out of field values could be externally imposed (as in the timed movement paradigm of Ghez et al., 1997; Schouten & Bekker, 1967), as opposed to being left to the inherent dynamics of field evolution. In such a case, the chosen production values could reflect the influence of multiple evolving articulator fields. For us, this would mean that the Monitor could be forced to choose production values at a particular point in time, thus revealing the gradual nature of parameter setting.

Yuen et al. (2010) present a task where this may plausibly have been the case. Their participants had to produce nonsense response utterances (e.g., "*kab*") based on a visual cue, which was presented immediately following an auditory distractor. Crucially, the timing of distractors, cues, and responses was tightly controlled. Participants heard three tones each timed to be 500 ms apart. The first indicated the start of the trial and the second was simultaneous with the presentation of the cue indicating the required response. The third tone indicated the target onset time of the response, i.e., participants had to respond in sync with a beep that followed 500 ms after the presentation of the cue. The distractor stimulus was presented between the first and second tones. There was also a phoneme-monitoring component to the task, in which participants were occasionally asked after their response whether the distractor contained a particular phoneme. This component was included to ensure attention to the distractors. Data were collected using electropalatography, which registers regions of tongue–palate contact as some part of the tongue raises to form a constriction on the palate. The results of interest were that /k/ responses ("*kab*") preceded by a /t/-initial distractor ("*tab*") showed increased alveolar contact compared to the same responses with /k/-initial distractors ("*kab*").

The results from these conditions can be simulated using our model with minimal assumptions and changes to the simulations reported in the previous section. In terms of the model, there are only a few material differences between this experimental task and the response–distractor task (see Appendix A for specific differences in the parameter value settings, and Appendix B for a link to the MATLAB scripts that were used to simulate this experiment). First, in this task the distractor preceded the response cue. This requires implementing negative SOAs in the model as opposed to positive ones. Second, the participants needed to attend to the distractor in this task, whereas they were told to ignore it in the response–distractor task. Therefore, the distractor input needs to be weighted such that the activation level of its planning field remains sufficiently high, but not so much that the participants respond with the distractor, which they were not reported to have done. Third and lastly, since the participant's response time was fixed, this task puts different constraints on the way that production values are sent to implementation in the model. We make the minimal assumption that the values sent to implementation reflect directly (i.e., linearly) the activation values of the corresponding fields. That is, for fields with "on" states, higher activation results in stronger constriction.

Fig. 18A illustrates the evolution of the fields in the simulation of a single trial in the congruent case, where the distractor and response are the same, i.e., both "*kab*". Time step 0 corresponds to the time of the tone that preceded the cue presentation. Shortly after that tone, the distractor

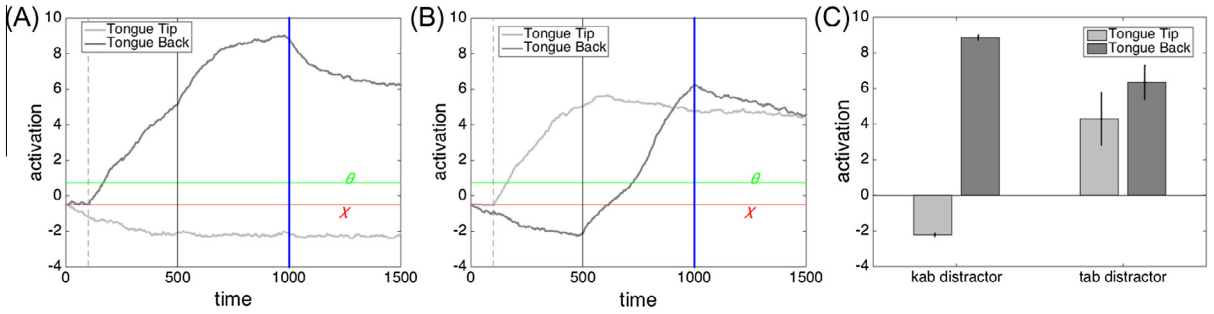

Fig. 18. Simulation of the result from Yuen et al. (2010). (A) Evolution of the Tongue Tip and Tongue Back fields on a trial with the congruent (identical) distractor–response pair, *kab–kab*. (B) Evolution of the same fields on a trial with the incongruent distractor–response pair, *tab–kab*. (C) Mean activation levels of the Tongue Tip and Tongue Back fields at time step 1000 across 50 simulated trials for each distractor–response combination (error bars indicate one standard deviation).

stimulus input begins, indicated by the vertical dashed line at time step 100 in Fig. 18A. The activation level of the Tongue Back planning field begins to rise, and the activation level of the Tongue Tip planning field begins to fall due to cross-field inhibition. The input for the required response begins at time step 500, resulting in the continued rise of the activation level of the Tongue Back planning field. At time step 1000, the response input stops, and the Monitor specific to this task needs to send values from all planning fields that have stabilized in the "on" state to implementation. In this case, only the Tongue Back field has such an activation level. Therefore, the production value of only the Tongue Back with the maximum activation level is sent to implementation. Fig. 18B shows a simulated trial of the incongruent condition, where the distractor is "*tab*". In this case, the distractor input raises the activation level of the Tongue Tip planning field, which behaves much the same as the Tongue Back planning field does in the congruent condition, rising toward a self-sustaining "on" peak and inhibiting the Tongue Back field. However, at time step 500 in this condition, the response input results in the activation level of the Tongue Back field rising, overcoming the inhibition from the Tongue Tip field and eventually inhibiting the activation level of the Tongue Tip field around time step 600, so that at time step 1000 the Tongue Back field has achieved a higher activation level than the Tongue Tip field. Cross-field inhibition lowers the levels of both articulator fields, but does not prevent either of them from achieving and maintaining an "on" state. Therefore, in this condition, both the Tongue Tip and Tongue Back fields have achieved an "on" state, and their weighted production values are sent to implementation. On this trial, the model therefore predicts both dorsal (tongue back) and alveolar (tongue tip) constrictions, but with the dorsal contact being greater than the alveolar, since the Tongue Back field has higher activation than the Tongue Tip. This is what was found by Yuen et al. (2010). Fig. 18C shows the mean maximum activation levels of the Tongue Tip and Tongue Back planning fields at time step 1000 in the two different distractor conditions across 100 simulated trials (50 for each distractor–response pair). On the left, the mean activation of Tongue Tip planning remains below resting level since there is no input to it and it is inhibited by the Tongue Back field. On the right, the activation level of the Tongue Tip field is roughly equal to the "on" activation level, meaning that on average, the planning field corresponding to the distractor stimulus achieves a stable "on" state and a tongue tip constriction is therefore sent to Implementation. The activation level of the Tongue Tip planning field is lower than the Tongue Back planning field, so that even though tongue tip constrictions are sent to Implementation, they are weaker than the tongue back constrictions. In summary, these simulations show that the model of phonological planning that accounts for RT differences in the response–distractor task can also provide an account of modulations in articulation in another task where response times are externally imposed.

VOT modulation

Finally, the model also makes predictions about the nature and phonetic detail of the other main phonological parameter of the actual responses, i.e., voicing. The combination of inputs to the planning process can also result in modulations of the implemented values of the utterance being planned. We specifically focus here on the effects of within-category gradient differences in input values and the consequences of such differences for the value chosen for implementation.

The dynamics of DFT are such that, on the one hand, when two inputs are sufficiently close to each other, even if they are not the same, they excite each other. This mutual excitation results in a faster buildup of activation for parameter values in the region of the two inputs than if there were no re-enforcing input, thus the increased rate of activation buildup. That is, local excitation introduced by parameter values sufficiently close to each other increases not just the activation levels of these values but also the activation level of neighboring parameter values. Therefore, given two inputs that are sufficiently close to each other, one having peak a maximum activation at parameter value x_1 and the other having a maximum at x_2, all parameter values between x_1 and x_2 are excited by both inputs. Assuming that the combined inputs are of sufficient strength for the field to stabilize with a single peak

of activation, the parameter value with the maximum activation level when the field stabilizes will be a value between x_1 and x_2, determined by the combined influence of the relative activation levels of the peaks, the width of those peaks, and noise.

On the other hand, when there are two incompatible inputs to the same field, they do not mutually excite parameter values that lie between them; they only mutually inhibit each other. This means that in the case of two compatible (i.e., close) inputs, the field reaches a stable state with a peak faster than when there is no reinforcing input, but the actual parameter value chosen for output will be an intermediary value between the maxima of the inputs. It also means that in the case of two incompatible (i.e., distant) inputs, the field stabilizes more slowly than when there is only one input, but there is no influence of one input on the other in terms of the parameter value that gets sent to implementation. This behavior is qualitatively the same as seen in the model of saccade planning developed by Kopecz and Schöner (1995).

Therefore, the model predicts that the VOT of a response should be modulated by the VOT of a distractor. Consider a scenario where the intended VOT of the response /t/$^{\text{vot = 60 ms}}$ and a distractor comes in with a different VOT value, e.g., /t/$^{\text{vot = 105 ms}}$. Perception of the distractor influences the on-going planning of the response. Specifically, the distractor's VOT contributes a localized increase in activation to the VOT activation field of the planned response, shifting (in our example to a more extreme value) the locus of maximum activation toward the distractor's VOT value. Thus, it is predicted that the VOT of the response should accommodate to that of the distractor. This prediction will be tested in a future study. Specifically, during a baseline block, participants will be prompted to produce 50 tokens of *ta* and 50 tokens of *ka*, without auditory distractors, in order to obtain a baseline VOT profile for each participant. The VOTs of each syllable in this block will be measured automatically, using software developed in our lab. This will permit us to use the participants' baseline VOTs to generate proximal and non-proximal VOTs for distractors, and thereby assess the extent of modulation in phonetic details.

Conclusions

Perceptuo-motor effects obtained using the response–distractor paradigm offer insights on the nature of the perception–production link and help to identify design requirements that any account of this link must satisfy. We have argued that the source of at least one class of perceptuo-motor effects observed in response–distractor tasks is found in the process of phonological planning, that perceived stimuli affect this process, and that the principles of excitation and inhibition embedded in an explicit computational framework are crucial in the planning process. A range of response time results concerning both complete identity and partial identity between planned responses and perceived inputs can be explained by the proposed model. The proposed model and the experimental results from the response–distractor paradigm add coherence to an otherwise confusing set of previous psycholinguistic results by showing that fundamental properties involved in phonetic description of linguistic contrast do play a role in the interaction between speech production and perception. Finally, the model serves as a tool for deriving new predictions that can be used to guide further experimental work on the relation between speech perception and speech production.

Acknowledgments

KDR gratefully acknowledges support from NSF Grant BCS-0951831 and from NIH Grant DC-002717 to Haskins Laboratories and the City University of New York for preparation of this manuscript. AIG gratefully acknowledges support by ERC AdG 249440.

Appendix A. Model parameter values

The constriction location input distribution for all articulator fields had a mean (*val*) of 0 and standard deviation of 2, defined on an arbitrary scale of constriction locations that ranged from −10 to 10. For the Voicing parameter, distributions for all voiced stimuli input had a mean of 5 ms VOT and 45 ms for voiceless stimuli, both with a standard deviation of 45 ms. The variable values used were: $\tau = 150$ and $h = -2.1$. Noise was added across all x values in each field at every time step in the evolution of the field. It was implemented by a discretized Wiener process with time step dt using a normal distribution with zero mean and unit variance, i.e., $dW = \sqrt{dt}\ N(0,1)$. The time step was set to 1/150. The response input weight (r) was 2.7, and was the same for inputs to both the articulator and Voicing field of the required response. The response input lasted for 400 time steps. The weight of the task input (p) was 0.7. There were two different distractor input weights, one for the articulator parameter (d_{artic}), which was 9.5 for all articulator fields, and one for the voicing parameter (d_{voice}), which was 11. This difference is due to the fact that the dynamics that give rise to the within-field and cross-field inhibition are markedly different (as we discuss in the main text). The distractor input lasted for 75 time steps. The cross-field inhibition threshold (χ) was −0.5. The amount of cross-field inhibition subtracted on each step from other fields when an articulator field was above (χ) was 1.25. The values for the parameters of the interaction kernel term (Eq. (6)) were the same in all four activation fields: $\theta = 0.75$, $w_{excite} = 0.45$, $w_{inhibit} = 0.1$, $\sigma = 1$. For the sigmoid threshold function (Eq. (5)), β was always 1.5. The criterion value (κ) was 6. A small amount of noise was included in the $input_{RESPONSE}(x, t)$, but not for $input_{DISTRACTOR}(x, t)$, since the distractor stimulus was the same across trials. The settings of the parameter values in the simulations of the task used by Yuen et al. (2010) were the same as in the response–distractor task, except the following changes: SOA was −500, the response input duration was 500 time steps and its weight (r) was 3.5, the distractor input duration was 100 time steps, and the weights of the distractor were 2.5 for the articulator parameter (d_{artic}) and 2 for the voicing parameter (d_{voice}).

The specific values of the variables in the above equations are not meaningful in and of themselves. The parameters are interrelated so as to implement specific concepts, e.g., settling to a state corresponding to a localized distribution of activation over an entire field (a continuum) of phonetic values and maintaining that distribution even in the absence of input (stability). We note that while logically it should be the case that there are other (potentially unlimited) sets of parameter values that could qualitatively match our data, the broad generalizations or predictions from the model do not depend on the specific parameter values and rather follow from the general principles of Dynamic Field Theory. For example, that congruency is faster than incongruency holds true for a wide class of parameter values and implementations of the interaction term in the dynamics. Their values relative to each other are more informative.

Appendix B. Supplementary material

Supplementary materials associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jml.2016.01.005.

References

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

Browman, C. P., & Goldstein, L. M. (1989). Articulatory gestures as phonological units. *Phonology, 6*, 201–251.

Browman, C. P., & Goldstein, L. M. (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics, 18*, 299–320.

Byrd, D. (1993). 54,000 American stops. *UCLA working papers in phonetics* (Vol. 83, pp. 97–116).

Byrd, D., & Saltzman, E. L. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics, 31*(2), 149–470.

Campbell, F., Gick, B., Wilson, I., & Vatikiotis-Bateson, E. (2010). Spatial and temporal properties of gestures in North American English /r/. *Language and Speech, 53*, 49–69.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*.New York: Harper & Row.

Dart, S. N. (1998). Comparing French and English coronal consonant articulation. *Journal of Phonetics, 26*, 71–94.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology, 55*, 149–179.

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

Forster, K. I., & Davis, C. (1991). The density constraint on form-priming in the naming task: Interference effects from a masked prime. *Journal of Memory and Language, 30*, 1–25.

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America, 99*(3), 1730–1741.

Galantucci, B., Fowler, C. A., & Goldstein, L. M. (2009). Perceptuomotor compatibility effects in speech. *Attention, Perception, & Psychophysics, 71*(5), 1138–1149.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review, 13*(3), 361–377.

Ghez, C., Favilla, M., Ghilardi, M. F., Gordon, J., Bermejo, R., & Pullman, S. (1997). Discrete and continuous planning of hand movements and isometric force trajectories. *Experimental Brain Research, 115*, 217–233.

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

Gordon, P. C., & Meyer, D. E. (1984). Perceptual-motor processing of phonetic features in speech. *Journal of Experimental Psychology: Human Perception and Performance, 10*(2), 153–178.

Grossberg, S. (1973). Contour enhancement, short-term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics, 52*, 213–257.

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.

Hickok, G. S., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences, 4*(4), 131–138.

Hock, H. S., Schöner, G., & Giese, M. (2003). The dynamical foundations of motion pattern formation: Stability, selective adaptation, and perceptual continuity. *Perception & Psychophysics, 65*(3), 429–457.

Keating, P. A. (1990). The window model of coarticulation: Articulatory evidence. In J. Kingston & J. Beckman (Eds.), *Papers in laboratory phonology I* (pp. 451–470). Cambridge: Cambridge University Press.

Kerzel, D., & Bekkering, H. (2000). Motor activation from visible speech: Evidence from stimulus response compatibility. *Journal of Experimental Psychology: Human Perception and Performance, 26*(2), 634–647.

Klein, E., Roon, K. D., & Gafos, A. I. (2015). Perceptuo-motor interactions across and within phonemic categories. *Paper presented at the 18th international congress of phonetic sciences, Glasgow, UK*.

Kopecz, K., & Schöner, G. (1995). Saccadic motor planning by integrating visual information and pre-information on neural dynamic fields. *Biological Cybernetics, 73*, 49–60.

Kornblum, S. (1994). The way irrelevant dimensions are processed depends on what they overlap with: The case of Stroop- and Simon-like stimuli. *Psychological Research Psychologische Forschung, 56*(3), 130–135.

Kornhuber, H. H., & Deecke, L. (1965). Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale (Changes in brain potentials with willful and passive movements in humans: The readiness potential and reafferent potentials). *Pflüger's Archiv für die gesamte Physiologie des Menschen und der Tiere, 284*, 1–17.

Ladefoged, P. (1999). *American English handbook of the International Phonetic Association* (pp. 41–44). Cambridge: Cambridge University Press.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Malden, MA: Blackwell Publishing.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1–36.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*, 384–422.

Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition, 109*(1), 168–173.

Moulin-Frier, C., Laurent, R., Bessière, P., Schwartz, J.-L., & Diard, J. (2012). Adverse conditions improve distinguishability of auditory, motor, and perceptuo-motor theories of speech perception: An exploratory Bayesian modelling study. *Language and Cognitive Processes, 27*, 1240–1263.

Mousikou, P., Roon, K. D., & Rastle, K. (2015). Masked primes activate feature representations in reading aloud. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 636–649.

Nam, H., & Saltzman, E. L. (2003). A competitive, coupled oscillator model of syllable structure. *Paper presented at the proceedings of the 15th international congress of phonetic sciences, Barcelona*.

Nielsen, K. Y. (2007). Implicit phonetic imitation is constrained by phonemic contrast. *Paper presented at the 16th international congress of phonetic sciences (ICPhS XVI), Saarbrücken, Germany*.

Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America, 99*(3), 1718–1725.

Rastle, K., Croot, K. P., Harrington, J. M., & Coltheart, M. (2005). Characterizing the motor execution stage of speech production: Consonantal effects on delayed naming latency and onset duration. *Journal of Experimental Psychology: Human Perception and Performance, 31*(5), 1083–1095.

Roelofs, A. (1999). Phonological segments and features as planning units in speech production. *Language and Cognitive Processes, 14*(2), 173–200.

Roon, K. D., Klein, E., & Gafos, A. I. (2014). Distractor effects on response times in fricative production. *Paper presented at the 10th international seminar on speech production (ISSP), Cologne, Germany*.

Roon, K. D., & Gafos, A. I. (2015). Perceptuo-motor effects of response-distractor compatibility in speech: Beyond phonemic identity. *Psychonomic Bulletin & Review, 22*(1), 242–250.

Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1*(4), 333–382.

Sandamirskaya, Y., & Schöner, G. (2008). Dynamic field theory and embodied communication. In I. Wachsmuth & G. Knoblich (Eds.). *Modeling communication with robots and virtual humans* (Vol. 4930, pp. 260–278). Berlin Heidelberg: Springer.

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

Schouten, J. F., & Bekker, J. A. M. (1967). Reaction time and accuracy. *Acta Psychologica, 27*, 143–153.

Schriefers, H. J., Meyer, A. S., & Levelt, W. J. M. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language, 29*, 86–102.

Thelen, E., Schöner, G., Scheier, C., & Smith, L. B. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching. *Behavioral and Brain Sciences, 24*, 1–86.

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

Yuen, I., Brysbaert, M., Davis, M. H., & Rastle, K. (2010). Activation of articulatory information in speech perception. *Proceedings of the National Academy of Sciences (Social Sciences), 107*(2), 592–597.