A Dynamic Neural Model of Tonal Downstep*

Manasvi Chaturvedi & Jason A. Shaw
Yale University

1 Introduction

Tonal downstep has been defined as a contrastive drop in pitch from a preceding tone (Leben, 2014). A slightly more inclusive definition describes downstep as a pitch drop that cannot be explained by pitch declination (Connell, 2011). Our focus in this paper is specifically on a pitch drop in a high tone (H) that is triggered by a preceding low tone (L). This type of pattern has variably been treated as either phonetic in nature or as phonological in nature. For example, downstep of this type has been modelled with phonetic implementation rules (e.g., Liberman & Pierrehumbert, 1984); it has also been modelled as feature spreading, a phonological process (e.g., Snider, 1998). In this paper, we propose an account that incorporates insights from both phonological and phonetic treatments of downstep. We arrive at this account by considering the general neural mechanisms that underlie action selection and control, and considering downstep through this lens. Situated within the broader perspective of action selection and control, we see downstep as a natural consequence of rapid transitions between goals on the same metric dimension (here, pitch).

We formalize our neural-based account of downstep in Dynamic Field Theory (DFT) (Schöner & Spencer, 2016). DFT describes changes in the state of activation over time within a dynamic neural field (DNF). DNFs are sensitive to continuous metric dimensions, such as pitch. Activation peaks within a field give rise to discrete percepts and actions. DFT has been applied to both non-categorical language behavior – such as trace effects in speech errors (Stern et al., 2022), phonetic accommodation (Kirkham et al., 2025), hyperarticulation due to lexical competitors (Stern & Shaw, 2023), gradient sound change (Shaw & Tang, 2023; Gafos & Kirov, 2009) – as well as categorical alternations (Shaw, 2025). Here we consider tonal downstep, a phenomenon that has aspects of both.

Although formalized using a general approach to cognition grounded in neural processing, our model does not abandon key insights from previous accounts. We make crucial use of the proposal, from phonological theory, that register spreading is important for downstep, while at the same time, similarly to phonetic accounts, derive continuous changes in pitch over time.

There are three crucial ingredients to our account of downstep in an HLH sequence: (1) the insight that downstep involves the combination of low (l) register with H tone (as in Snider (1998)), 2) the representation of tone targets within the *continuous* dimension of pitch, which draws on the representational structure of a DNF 3) an assumption that activation peaks in a DNF are inhibited upon successful completion of an action (pitch target). Such inhibition of activation following successful target achievement aids rapid transitions between targets. This mechanism may be crucial for languages that specify sequences of distinct pitch targets within a short temporal span, such as on consecutive tone bearing units, i.e., a typical lexical tone language.

A primary contribution of this paper is the illustration of how phonological patterns can be derived from the interaction between the cognitive mechanisms involved in the selection of action plans, and the control of action over time. This approach is original in that–rather than abstracting away from time–it relies on a neural process model with states that change over time on the appropriate (millisecond) timescale for phonological cognition.

* **Thank you to the audience at AMP 2024 (Rutgers University), and the members of the DYNAMICS group (Linguistics department, Yale University), for their comments and suggestions.**

Proceedings of *AMP 2023/2024*

The rest of the paper is structured as follows: first, we will go over some data from Snider (1998) and discuss his register-based analysis of downstep in Bimomba; second, we will provide an overview of DFT, the framewwork for our model architecture, and present some simulations. We end with a short discussion of the results and directions for future work.

2 Background

2.1 *Downstep in Bimboa (Snider 1998)* Although there are language-specific variations in downstep, we focus here on the pattern in Bimoba, a Gur language of the Niger-Congo family spoken in Northern Ghana. There are some phonetic data published for this pattern. Figure 1 is based on the measurements reported in Snider (1998: pp. 87-89). The f_0 values correspond to the Bimoba examples in 1 and 2 (Snider, 1998: p. 86). The reported measurements are single values of f0 for each syllable.

The comparison in pitch between an utterance with all H tones vs. one with an intervening L shows us that the L causes a lowering of subsequent Hs that is far beyond that caused by normal declination over the utterance (Snider, 1998). Note that this lowering results in a tone that is still *phonologically* an H, even though it is *phonetically* closer to an L.

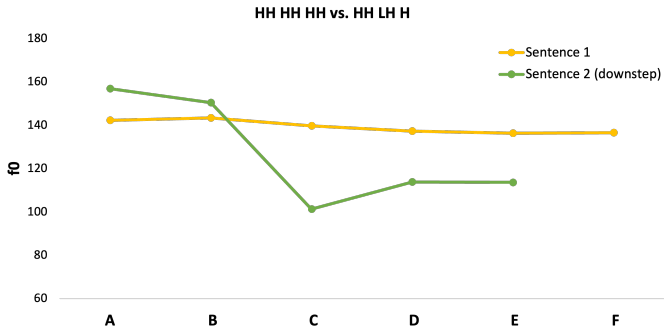


Figure 1: f_0 contour of HH HH HH vs. HH LH H utterances.

(1) *gbátúk* *mírí* *gbátúk*
bushbaby cut.PRES bushbaby
'bushbaby is cutting a bushbaby'

(2) *gbátúk* *gòt* *↓gbátúk*
bushbaby look.at.PAST bushbaby
'bushbaby looked at a bushbaby'

Snider (1998) accounts for downstep via a model that proposes a 'register tier' in the phonological representation. This register tier combines with the 'tone tier' to give the final tonal specification for each mora. Following Snider's notation, we use "l" and "h" for low and high register and L and H for low and high tone. Downstep, then, is the combination of a low (l) register feature with a H tone (as compared to a H tone with high (h) register) – giving us a tone output that is phonologically still an H, due to the tonal tier, but phonetically lower than the non-downstepped counterparts, due to the register tier. This is illustrated by the schema in Figure 2, reproduced from Snider (1998: p. 95). The Figure shows the low register (l) is spread across two tone nodes, while the second high register (h) is delinked from the final tone node.

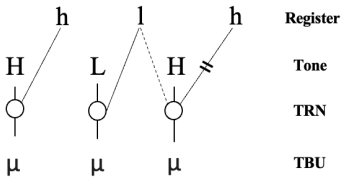


Figure 2: Snider (1998)'s characterization of downstep (p. 95).

Snider (1998)'s phonological analysis receives some converging evidence from facts about the *Mid* tones in Bimoba: 1) The observation that downstepped H tones are phonetically close to Mid tones in the language, and 2) The observation that while Mid tones downstep High tones, Low tones do not downstep Mids (p. 96). To account for these facts, Snider proposes that both the Mid and the High tones in Bimoba have the characterization in Figure 2. Since Mids already have a low register feature in their representation, they are ineligible to be downstepped by preceding Low tones. For the same reason, however, they can downstep subsequent Highs.

As we will outline in §3, we use Snider's insight that it is the low register feature, which spreads from Low/Mid tones, that is responsible for downstep, while at the same time, connecting this intuition to a continuous representation of pitch.

2.2 *Dynamics of action control* As mentioned in §1, discrete cognitive events emerge in DFT via the evolution of activation within neural fields (DNF) sensitive to particular continuous dimensions.

Activation in a DNF evolves over time according to equation 3 (Schöner & Spencer, 2016). $\dot{u}(x,t)$ represents the change in activation over a particular location in the field (particular value of the dimension the field is defined over; x) at a particular moment in time, t. Change in activation is related to the *current* activation, u, defining a point attractor system at each field location. Under the influence of inputs (s, the term in green, modeled as Gaussian distributions), interactions between the neurons in the field (the term in purple), and the history of activation (i.e., the state of the field before the inputs are added), activation across the field can stabilize, with some field locations above threshold ($>$ zero), representing a cognitive decision or detection of percept. In our case, the above threshold activation peak represents selection (a "decision") of a pitch target.

$$\tau\dot{u}(x,t) = -u(x,t) + h + s(x,t) + \int k(x-x')g(u(x',t))dx' + q\xi(x,t) \quad (3)$$

Pitch targets selected by the DNF are fed to the articulatory dynamics, which we model via the linear second order dynamical system in equation 4. This dynamical system precisely characterizes a physical mass on a spring, and has been applied to the dynamics of a wide range of speech movements (Saltzman & Munhall, 1989), including pitch control (Gao, 2008). In this model, T is the target, x represents the current state (pitch value), and k is the stiffness parameter, which dictates the speed at which the system evolves towards the target state, m is mass and b is damping. In our simulations, we fixed the values of the control parameters to ensure critical damping: m is 1; b is 4; k is 4. The only remaining parameter is T, which is set to the location of above threshold activation peak in the DNF. While the articulatory dynamics in 4 is sufficient for our purposes–it specifies a point attractor, driving pitch to the selected pitch target over time–the precise details of the pitch contours, which we do not have access to for Bimoba, may be better captured by other dynamical systems. See, for example, Iskarous et al. (2024), who model pitch trajectories in English pitch accents, Stern & Shaw (2024), who propose an articulatory dynamics with fewer parameters than the mass spring, and Flemming & Cho (2017) who optimize over competing pitch targets in their model of Mandarin contour tones. For our purposes, what matters is that the equation evolves towards the target, T, which comes from the location in the DNF where a stable, above threshold, activation peak has formed. Thus, our system incorporates a dynamics for pitch target selection (equation 3) and a dynamics for pitch control (equation 4) over time.

$$m\ddot{x} + b\dot{x} + k(x-T) = 0 \quad (4)$$

The two dynamical systems interact to produce phonological patterns, in our case, tonal downstep. At each timestep, if there is a stable, above threshold peak in the pitch field (activation > 0), the target of the

gesture dynamics, T, is set to the location of that peak, i.e., the field location with maximum activation. If there is no above threshold peak, the dynamics evolve towards a neutral value, $T = 150Hz$. Over time, the rise and fall of activation peaks over different locations in the DNF condition transitions between different pitch targets. Our system makes crucial use of the interaction between the current state of pitch production, at a moment in time, and the dynamics of pitch target selection. In our model, this is achieved via a condition of satisfaction mechanism (CoS) (Schöner, 2020), which triggers the neural dynamics to progress to the next target only when the current target has been successfully achieved. This allows the successful sequencing of pitch targets.

In the next section, we detail our CoS mechanism, after a discussion of approaches to serial order sequencing via CoS within DFT more generally.

2.3 *Serial order sequencing in DFT* The basic picture that we are pursuing for downstep is one in which the neural dynamics selects a target for the gesture dynamics and senses through perception when the target has been achieved, triggering a transition in the neural dynamics that leads to the next target. Within the DFT literature, this type of sensing mechanisms is referred to as a Condition of Satisfaction (CoS). CoS in DFT has been implemented via an architecture that includes a combination of DNFs and dynamics nodes representing three crucial aspects of action: 1) Ordinal position in a sequence, 2) Target (often referred to as 'Intention') selection, 3) Perceptual input (Sandamirskaya, 2016; Sandamirskaya & Schoner, 2008). The target (or 'intention') is an internal representation of the action to be completed. When the target matches the actual perceptual input (for example, when the intention to look for a color matches visual input of that color (Sandamirskaya & Schoner, 2008)), it satisfies the CoS, through activation of a CoS node/field, indicating that the action has been successfully completed. This CoS can then inhibit the current action, leading to a transition to the next action.[1]

Like this general approach in DFT, we rely on a CoS system in our model of downstep to achieve sequential pitch targets. We implement the inhibition of completed actions via an inhibitory **input** into the field. Once the field stabilizes at a particular value, the gesture control dynamics evolves towards this value, as described previously. When the current pitch value (x) in equation 4 is close to the value in the field (within 3 Hz), the CoS system is activated. After an interval of time (80ms in the simulations below), chosen to represent the self-perception of pitch, an inhibitory input is added at the location of the stabilized peak in the pitch field, along with the input to the DNF for the next tone target. This achieves rapid transition from the current tone target to the next one.

In the next section, we report model parameters and three simulations. The simulations demonstrate how the components of our model lead to downstep, under certain parameter values.

3 DFT Model of downstep

Our model is based on three crucial ingredients: independent representations for tone and register, the representation of tone and register as distributions over continuous values, and the inhibition of completed pitch targets. The model parameters are listed in Table 1. The first four parameters are related to inputs to the DNF associated with phonological categories (first three) and inhibition following CoS (fourth). These inputs to the DNF take a Gaussian shape, with a center, p, and width, w. There are H & L tone inputs and a low (l) register input, all excitatory (with positive values). The last input (fourth row in the table) drives the CoS system described previously. The center of this input, p, is always at the location of the current above-threshold activation peak (the target). It functions to flatten an existing activation peak (overcoming hysteresis) enabling rapid transition to the next pitch target (activation peak in another location in the DNF). The remaining parameters specify properties of the DNF itself, including, the rate of activation change over time, τ, resting level (the point attractor when there are no inputs), h, DNF noise, q, and within-field interactions. The within-field parameters, local excitation (strength: c_{exc}, width: σ_{exc}), local inhibition (strength: c_{inh}, width: σ_{inh}), and global inhibition (strength: c_{glob}), are set so that together they enable selection dynamics. That is, when one part of the field crosses the activation threshold, an activation peak

[1] The full mechanism of inhibition leading to a transition to the next state in DFT involves an interaction between ordinal nodes, encoding sequence position, and memory nodes, which excite the next action in the sequence. *Nodes* in DFT have the same dynamics as fields, but are not defined over a continuous dimension, and are equivalent to a single location within a field. For a tutorial on serial order sequencing in DFT see Sandamirskaya (2016).

will be stabilized within the field.

Parameter	Description	Value
$p_{Htone}(w)$	Center of input (width)	180(5)
$p_{Ltone}(w)$	Center of input (width)	120(5)
$p_{lregister}(w)$	Center of input (width)	140(10)
$p_{inhibition}(w)$	Center of input (width)	At stabilized peaks(5)
τ	Evolution rate	20
h	Resting level	-5
β	Sigmoid slope	4
c_{exc}	Excitation strength	20
c_{inh}	Inhibition strength	5
c_{glob}	Global inhibition	0.9
σ_{exc}	Excitation width	5
σ_{inh}	Inhibition width	25
q	Noise	1.5

Table 1: Model parameters.

Using the parameters described above, we simulate downstep in an HLH sequence. Note, we assume that there is low (l) register input, which quantitatively is higher and wider than the L tone input. Given the dynamics of the DNF, in equation 3, the field will settle to the resting activation level, h, in the absence of inputs. The H tone input, the first in the HLH sequence, will cause an activation peak to form, providing a target to the gesture dynamics and raising pitch in the direction of the target. When the CoS is met, the location of the activation peak will be inhibited, the H tone input removed, and the L tone and low register (l) inputs introduced. The same cycle of events occurs when an activation peak forms at the low pitch end of the DNF, except that the low register (l) input is not removed once the final H input is added. That is, in this system, CoS removes tone inputs (L, H) but not register inputs. A register input, introduced by the L tone + l register combination, remains in the field even after CoS for the L tone. This mechanisms effectively captures the intuition of low register spreading (Snider, 1998). Note that our implementation does not make use of a high register (h), which may not be necessary for Bimoba, given the three tone (L,M,H) system.

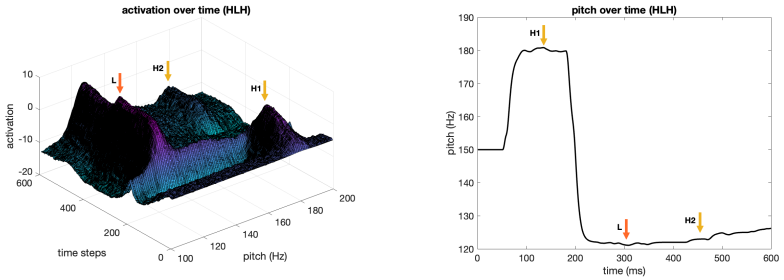
The next section will be divided into 3 different parts. In order to show that both inhibition of completed tones and l register perseverance (into the second H in an HLH sequence) is needed, we will simulate an HLH tone sequence under 3 conditions: 1. Without inhibition after target achievement. 2. Without low (l) register perseverance, and finally, 3. **With** both of the above.

4 Model Simulations

For each of the simulations below, the figure on the left describes the state of the neural field over the course of the HLH sequence. The pitch dimension is on the x-axis, the y-axis is time (each timestep can be interpreted as 1 ms), and the z-axis is the activation. The figure on the right is the result of the articulatory dynamics, driven by the varying pitch targets from the DNF.

4.1 *No inhibition of L target.* Figure 3 demonstrates that if there is no inhibition of the completed targets, we do not see downstep. In the neural dynamics (left), we see an activation peak at the high end of the field, corresponding to the first H tone. In the pitch contour (right), we see that the pitch starts at 150 Hz and then rises, once an activation peak forms at the high pitch end of the DNF. Following this first high peak, an activation peak forms at the low end of the DNF, driven by the combination of the L tone and l register inputs. This causes a drop in pitch (right; around 200ms). Even after the L tone input is removed, via CoS, and the second H input is added to the field, we do not see an above-threshold activation peak at the high end of the field. The arrow labeled H2 in the neural dynamics (left) shows a below-threshold increase in activation. The strength of the input is the same as for the first H, but the state of the field at the time of the second H input is different, owing to global inhibition (a component of the interaction kernal), so the consequence is only a below-threshold increase in activation. The high end of the field is inhibited (due to global inhibition during the low pitch activation peak) and the low pitch end of the field remains active, due

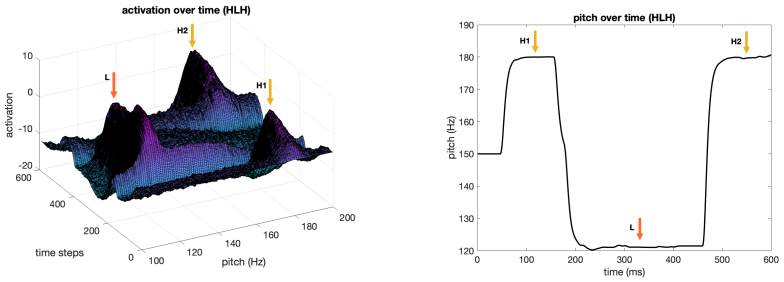
to the persistence of the l register (and the absence of inhibition). Under these conditions, without inhibiting the field following target achievement (CoS), to work against hysteresis, we see that pitch slowly moves to higher values. This is due to a transition from the L tone input dominating the location of the activation peak to the l register (once the L tone input is removed). Note that the second H tone input does not influence the location of the activation peak under these conditions.



(a) Activation over time. **(b)** Pitch over time.

Figure 3: Activation and pitch over time for an HLH sequence simulated without inhibition of L target.

4.1.1 *No perseverance of L register.* Figure 4 demonstrates that if l register is turned off along with L tone (after the HL in HLH), the last H surfaces at exactly the same value as the first one – i.e., there is no downstep. Note that, here, we incorporate inhibition of completed targets, unlike the previous simulation.

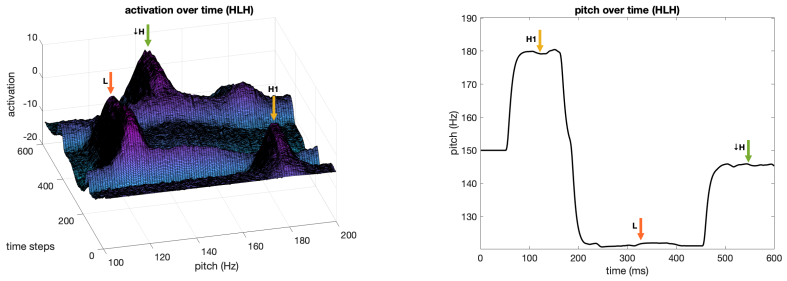


(a) Activation over time. **(b)** Pitch over time.

Figure 4: Activation and pitch over time for an HLH sequence simulated without perseverance of l register.

With these two simulations, we have demonstrated that without either l register perseverance or L target inhibition, the second H is not downstepped. The next simulation incorporates both of these aspects.

4.1.2 *Downstepped H: L register perseverance and L tone target inhibition.* With both the persistence of l register past the L tone and the inhibition of the L target following CoS, we see that the field stabilizes at a downstepped location. While there is still an H input at the non-downstepped location, the overall activation at lower ends of the fields results in a lower H realization (↓ H).



(a) Activation over time. **(b)** Pitch over time.

Figure 5: Activation and pitch over time for an HLH sequence with l register, and target inhibition.

With these 3 simulations taken together, we conclude that l register perseverance into the second H, and pitch target inhibition are both crucial ingredients to achieving a downstepped H.

5 Discussion

In this paper, we presented a dynamic neural field model of downstep, based on data from the Gur language, Bimoba, analyzed by Snider (1998) in the framework of Register Tier Theory. We demonstrated that downstep can be derived via the evolution of activation in a continuous neural field defined over pitch values, and that such a model highlights three necessary components for achieving downstep: (1) The perseverance of low (l) register into the second H of an HLH sequence (see simulation 2), (2) The representation of tone and register as distributions over continuous values, and (3) The inhibition of the DNF after successful achievement of a tone target in order to facilitate a rapid transition to the next target (see negative evidence in simulation 1). Here, *rapid* is important. Given enough time, we may get to a downstepped value even without inhibition–but, in language production, speakers do not have the luxury of 'enough time'. Compare the pitch track in Figure 3 to the one in Figure 6 below. Figure 6 shows the pitch contour simulated from the exact same parameters as simulation 2 (§4.1), except that we have doubled the time of the simulation, allowing 1200ms instead of 600ms. Given the additional time, we can see that pitch does eventually reach the value of a downstepped H even without inhibition. This slow timecourse of pitch evolution is not tenable for a typical lexical tone language, which specifies contrastive tones on consecutive tone-bearing units (although it may be sufficient for languages with only sparse specification of tone).

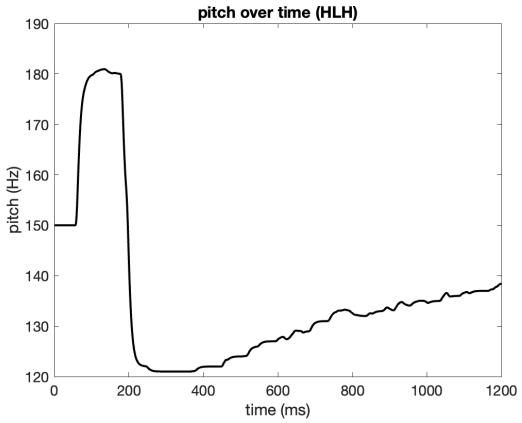


Figure 6: Pitch over time of an HLH sequence without target inhibition, and with extra time.

Our model also makes an important empirical prediction–since downstep relies on the presence of l register and the inhibition of the second tone target (the L target), we do not need the first H in the HLH in order for the second to be downstepped. Whether this is indeed the case in Bimoba is not evident in the data presented in Snider (1998) and thus requires new empirical work for confirmation. Additionally, we

departed from Snider (1998) in not proposing a high (h) register. Our system derives downstep with the tonal inventory of L tone, H tone, l register, which may have some independent cross-linguistic motivation (c.f. Lionnet, 2022).

Our model demonstrates the potential to derive phonological patterns from cognitive mechanisms that interact over the time scale appropriate for speech production. While we see this as a promising direction for future research integrating phonology with cognitive neuroscience, there are some limitations of the current model. Generally, in downstep, each L in a sequence is described as "setting a new pitch ceiling" such that subsequent Hs are realized at lower and lower pitch values–leading to pitch stepping (Connell, 2011). We have not extended the model to capture this phenomenon. In its current form, our model of HLH predicts that in a longer sequence, with multiple Ls (e.g., HLHLH), the second and third Hs will be produced at a similar pitch value. Thus, while we are optimistic about the general approach, future work is needed to extend the model, including exploration of a wider range of possibilities for integrating insights from atemporal phonological theory and dynamical approaches to phonological cognition.

References

Connell, Bruce (2011). Downstep. Oostendorp, Marc Van, Colin J Ewen, Elizabeth Hume & Keren Rice (eds.), *The Blackwell companion to phonology: Suprasegmental and Prosodic Phonology*, John Wiley & Sons, Ltd, vol. 2, 1–24.

Flemming, Edward & Hyesun Cho (2017). The phonetic specification of contour tones: Evidence from the mandarin rising tone. *Phonology* 34, 1–40.

Gafos, Adamantios & Christo Kirov (2009). A dynamical model of change in phonological representations: The case of lenition. *Approaches to Phonological Complexity*, De Gruyter Mouton, 219–240.

Gao, Man (2008). Tonal alignment in mandarin chinese: An articulatory phonology account. (Unpublished doctoral dissertation).

Iskarous, Khalil, Jennifer Cole & Jeremy Steffman (2024). A minimal dynamical model of intonation: Tone contrast, alignment, and scaling of american english pitch accents as emergent properties. *Journal of Phonetics* 104, p. 101309.

Kirkham, Sam, Patrycja Strycharczuk, Rob Davies & Danielle Welburn (2025). Modelling change in neural dynamics during phonetic accommodation. *arXiv preprint arXiv:2502.01210* .

Leben, William R (2014). The nature(s) of downstep. *SLAO/1er Colloque International, Humboldt Kolleg Abidjan*.

Liberman, Mark & Janet Pierrehumbert (1984). Intonational invariance under changes in pitch range and length. Aronoff, Mark & Richard Oehrle (eds.), *Language and Sound Structure*, Cambridge, MA: MIT Press, 157–233.

Lionnet, Florian (2022). Tone and downstep in paicî (oceanic, new caledonia). *Phonological Data and Analysis* 4, 1–47.

Saltzman, Elliot L & Kevin G Munhall (1989). A dynamical approach to gestural patterning in speech production. *Ecological psychology* 1:4, 333–382.

Sandamirskaya, Yulia (2016). Autonomous sequnece generation in dynamic field theory. Schöner, Gregor & John P Spencer (eds.), *Dynamic thinking: A primer on dynamic field theory*, Oxford University Press, chap. 14, 353–368.

Sandamirskaya, Yulia & Gregor Schoner (2008). Dynamic field theory of sequential action: A model and its implementation on an embodied agent. *2008 7th IEEE international conference on development and learning*, IEEE, 133–138.

Schöner, Gregor (2020). The dynamics of neural populations capture the laws of the mind. *Topics in Cognitive Science* 12:4, 1257–1271.

Schöner, Gregor & John P Spencer (2016). *Dynamic thinking: A primer on dynamic field theory*. Oxford University Press.

Shaw, Jason A (2025). Unifying phonological and phonetic aspects of speech in dynamic neural fields: the case of laryngeal patterns in japanese. *Phonological Studies* 28, 1–12.

Shaw, Jason A & Kevin Tang (2023). A dynamic neural field model of leaky prosody: proof of concept. *Proceedings of the Annual Meetings on Phonology*.

Snider, Keith L (1998). Phonetic realisation of downstep in bimoba. *Phonology* 15, 77–101.

Stern, Michael C & Jason A Shaw (2023). Neural inhibition during speech planning contributes to contrastive hyperarticulation. *Journal of Memory and Language* 132, p. 104443.

Stern, Michael C & Jason A Shaw (2024). Towards a minimal dynamics for gestures: a law relating velocity and position. *Proc. issp 2024*, 185–188.

Stern, Michael C, Manasvi Chaturvedi & Jason A Shaw (2022). A dynamic neural field model of phonetic trace effects in speech errors. *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 44.