A dynamic neural field model of leaky prosody: proof of concept

Jason A. Shaw¹ & Kevin Tang²
Yale University¹, Heinrich-Heine University²
Formal tools for sub-phonemic patterns

• Our formalisms tend to privilege discrete symbolic units, as many phonological patterns are insightfully described in these terms.

• Some phenomena are more challenging for this level of description (or just fall outside the scope):
  • Incomplete neutralization
  • Gradual sound change
  • Sub-phonemic change in representations over a lifetime

• In this talk, we explore the potential of Dynamic Neural Fields for capturing sub-phonemic patterns.
Empirical phenomenon: “Leaky Prosody”

• Lexical items come to take on the phonetic characteristics of the prosodic environments in which they are typically produced (e.g., Seyfarth 2014; Sóskuthy & Hay 2017; Tang & Shaw 2021).

• In Mandarin Chinese, words that tend to attract a high degree of prosodic prominence are produced with relatively high pitch, even in prosodically weak environments; thus, prosody from context leaks into the lexicon (Tang & Shaw 2021).

• Effects are lexically specific and sub-phonemic synchronically but may provide seeds for gradual diachronic change.
  • Frequency/informativity effect on segment count (Zipf 1949; Piantadosi et al. 2012) may derive from frequency/informativity effect on ms duration (Wright 1970; Seyfarth 2014).
  • lexical tone/stress emerging from higher level prosodic prominence/intonation
  • Lexical tone loss in predictable environments.
Architectural sketch (Tang & Shaw 2021)

• Leaky prosody facts suggest that phonetic outputs may feedback into the lexicon.

• Possibly imperfect (incomplete) compensation for effect of prosodic environment on phonetic realization.


c.f., Turk & Shattuck-Hufnagel (2020)
Today: alternative “flat model”

• Potential advantage in learning surface distributions (distributional learning) vs. transformational rule (multi-factor regression/highly parameterized generative model)
Framework: Dynamic Field Theory (Schöner & Spencer 2016)

- Cognitive representations are **continuous** parameters (here, pitch) governed by populations of neurons.

- The distribution of **activation** across a neural population is represented by a dynamic neural field (DNF).

- Activation at each field location **evolves over time** under the influence of inputs until the system **stabilizes**
DFT: key properties for a flat model of leaky prosody

- Multiple inputs to a field can exert influence on stabilization.

- Perception/production modelled as time varying processes, c.f., purely statistical agent-based models (c.f., Harrington & Schiel 2017).

- Nested time scales: Learning occurs token-by-token (slow time scale) in response to production & perception (fast time scale).
Model overview: pitch target

Pitch input from three sources:

- **lexicon**: lexical pitch target
- **tone**: phonological pitch target
- **prosody**: prosodic pitch target
- Lexicon updated to incorporate stable pitch

Feedback

Pitch planning field (selection dynamics)

\[ \int k(x - x')g(u(x', t))dx' \]
Formal expression of the model

\[ \tau \dot{u}(x, t) = -u(x, t) + h + s_{\text{lex}}(x, t) + s_{\text{phon}}(x, t) + s_{\text{pros}}(x, t) + \int k(x - x')g(u(x', t))dx' + q\xi(x, t) \]

Change of activation ➔ Resting activation ➔ Inputs to the field ➔ Interaction kernel (property of the field) ➔ Noise

DNF parameter

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resting activation</td>
<td>( h = -5 )</td>
</tr>
<tr>
<td>Field evolution speed</td>
<td>( \tau = 20 )</td>
</tr>
</tbody>
</table>

Models built with the COSIVINA Toolbox in Matlab:
Formal expression: gaussian inputs

\[ \tau u(x, t) = -u(x, t) + h + s_{\text{lex}}(x, t) + s_{\text{phon}}(x, t) + s_{\text{pros}}(x, t) + \int k(x - x')g(u(x', t))dx' + q\xi(x, t) \]

\[ s(x, t) = a \exp \left[ -\frac{(x - p)^2}{2w^2} \right] \]
Simulation inputs are surface distributions

Input parameters based on Tang & Shaw (2021) corpus of 1,655 Mandarin speakers.

• Starting Lexical input = sample of high tone distribution (1/500th)
• Phonological pitch target = high tone distribution (~41,000)
• Prosodic context = distribution of pitch values at two levels of bigram surprisal
  • Low predictability (~10,000)
  • High predictability (~10,000)

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

<table>
<thead>
<tr>
<th>Input parameters</th>
<th>( S_{\text{lex}} ) (1st run)</th>
<th>( S_{\text{phon}} )</th>
<th>( S_{\text{pros}} ) (low, high)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a )</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>( p )</td>
<td>241</td>
<td>238</td>
<td>233, 226</td>
</tr>
<tr>
<td>( w )</td>
<td>99</td>
<td>94</td>
<td>100, 92</td>
</tr>
</tbody>
</table>

Formal expression: interaction kernel

\[
\tau \dot{u}(x, t) = -u(x, t) + h + s_{lex}(x, t) + s_{phon}(x, t) + s_{pros}(x, t) + \int k(x - x') g(u(x', t))dx' + q \xi(x, t)
\]

Parameters chosen to ensure selection dynamics

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

Sigmoidal gate

DNF parameters

- \( c_{exc} = 15 \)
- \( c_{inh} = 5 \)
- \( c_{glob} = 0.9 \)
- \( \sigma_{exc} = 5 \)
- \( \sigma_{inh} = 12.5 \)
Simulations

1. **Speech production planning as a time varying process (fast time scale)**: establish effect of prosodic context on pitch target
   - Initialize two words with identical pitch targets
   - Simulate one in a high prominence environment; one in a low prominence environment.

2. **Lexical learning as a time varying process (slow time scale)**: derive leaky prosody from updating the lexicon
   - update lexical representations based on where the field stabilizes on each fast time scale simulation
Nested timescales

**Fast (e.g., msec):**
- Timescale of speech planning

**Slow (e.g. hours):**
- Learning with each exposure to a word

- **Pitch planning field**
  
  \[
  \int k(x - x')g(u(x', t))dx'
  \]

- **Stable pitch**

- **Feedback**

**Lexicon**
- \(s_{\text{lex}}(x, t)\)

**Tone**
- \(s_{\text{phon}}(x, t)\)

**Prosody**
- \(s_{\text{pros}}(x, t)\)
Fast time scale: single trial, high vs. low prominence

Time course of speech planning

---

**Figure:**

- **Graph 1:**
  - **Title:** final activation by field location
  - **X-axis:** field position (Hz)
  - **Y-axis:** activation
  - **Legend:**
    - high prominence (red)
    - low prominence (blue)

- **Graph 2:**
  - **Title:** high prominence
  - **X-axis:** time
  - **Y-axis:** activation
  - **Legend:**
    - peak@241Hz

- **Graph 3:**
  - **Title:** low prominence
  - **X-axis:** time
  - **Y-axis:** activation
  - **Legend:**
    - peak@234Hz
Updating the lexical input from single trial

Stable pitch
(single trial output)

Lexical input
Updated lexical input

Samples from lexical input distribution; one sample is randomly selected and replaced with the new pitch value.
Slow time scale: lexical drift over 500 trials

Stable pitch
(single trial outputs)

Lexical input

![Graph showing stable pitch over 500 trials with single trial outputs.](image)

![Graph showing lexical input with high and low prominence.](image)
Discussion: achievements

• Leaky prosody effect derived from simple assumptions
  • A1: production inputs come from surface distributions
    • Lexical target: sample of distribution of f0 for high tone category
    • Phonological tone: complete distribution of f0 for high tone category
    • Prosodic context: distribution of f0 at a given level of surprisal
  • A2: inputs jointly influence pitch target
  • A3: flat model → stabilization instead of transformations

• Trial-by-trial variability

• Small lexical differentiation emerges over time from learning
Discussion: limitations

• Just one tone (high)
• Just two lexical items
• Just one feature dimension (pitch)
• No talker normalization (flat model)
• No signal transformations (ERB, MEL)
Discussion: parameter space

- Only **lexical** inputs (not phonological/prosodic) updated
  - Stable phonological input works against lexical drift.
  - Should persist even if phonological representations are updated...
  - Unless enough words shift in the same direction

- Amplitude of inputs the same (> $h$ ‘rest level’) for lexical, phonological, prosodic targets
  - Having lexical, phonological, and prosodic inputs leads to faster stabilization.
  - Predicts we should be able to have a pitch target with just one input.
\[ \tau \dot{u}(x, t) = -u(x, t) + h + s_{\text{lex}}(x, t) + s_{\text{phon}}(x, t) + s_{\text{pros}}(x, t) + \int k(x - x') g(u(x', t)) dx' + q\xi(x, t) \]