

Deriving phonetic realizations from the interaction of target constraints and constraints on articulatory dynamics

Edward Flemming and Amy Li

Massachusetts Institute of Technology

June 25, 2026

LabPhon 20 workshop: Articulatory control

1. Motivation

2. Analysis

3. Results

4. References

5. Appendix

1. Motivation

2. Analysis

3. Results

4. References

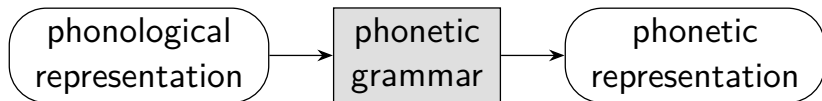
5. Appendix

Phonetic grammar

The phonetic component of a grammar maps phonological representations onto phonetic representations on a language-specific basis.

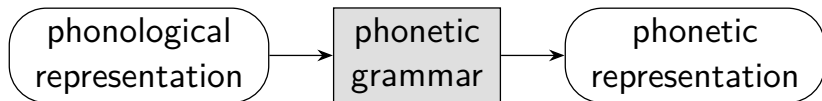
Phonetic grammar

The phonetic component of a grammar maps phonological representations onto phonetic representations on a language-specific basis.



Phonetic grammar

The phonetic component of a grammar maps phonological representations onto phonetic representations on a language-specific basis.

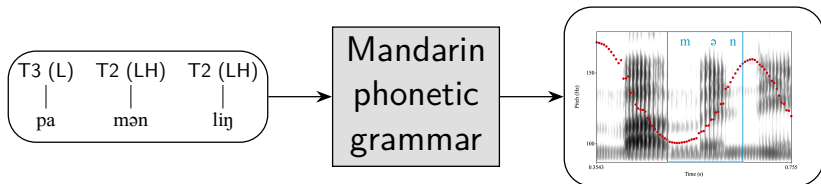


- ▶ High level goal: develop a theory of phonetic grammar.
- ▶ Starting point: phonetic grammar for tone and intonation.

Phonetic grammar: representations

Tone:

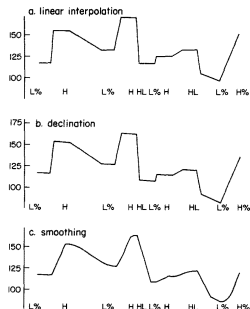
- ▶ Phonological representations are tone categories:
 - ▶ high, low, rise, etc.
- ▶ Phonetic representations are f_0 trajectories that are aligned to the segmental string (Pierrehumbert, 1980).
- ▶ A phonetic grammar for tone maps tone categories to f_0 trajectories.



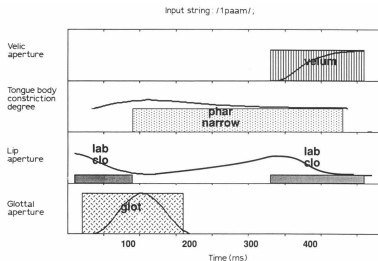
Control signals and trajectories

In most implemented models of phonetic realization:

- ▶ Phonological representations yield a control signal.
- ▶ The control signal is the input to a dynamical system or a smoothing filter that derives the output trajectory.
- ▶ There is no look-ahead to the output of the dynamical system when the control signal is generated.



(a) Beckman and Pierrehumbert (1988)



(b) Browman and Goldstein (1992)

An optimization-based model of phonetic grammar

Look-ahead is central to the generation of control signals.

- ▶ Generation = optimization over constraints.
- ▶ Constraints evaluate potential outputs.

Case study: variation in Mandarin rising tone

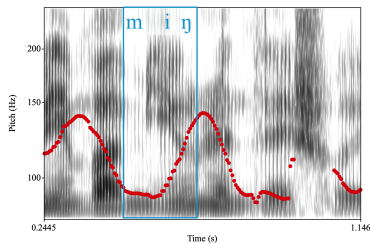
The phonetic realization of Mandarin rising tone varies with speech rate (Flemming and Cho, 2017).

- ▶ Mandarin speakers read sentences at three speech rates: normal, fast, slow.
- ▶ Each sentence: two rising tone syllables in a carrier phrase.

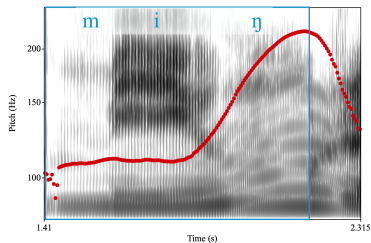
(1) tɕʰin̩ nǐn pà ˨˩˦ ˨˩˦ tsâi ʂó jǐ pjên
qǐng nín bă ˨˩˦ ˨˩˦ zài shuō yí biàn
'Please say ˨˩˦ again'

- ▶ Focus on examining the low-rise-rise context.
- ▶ Audio: normal example, fast example, slow example.

Case study: variation in Mandarin rising tone

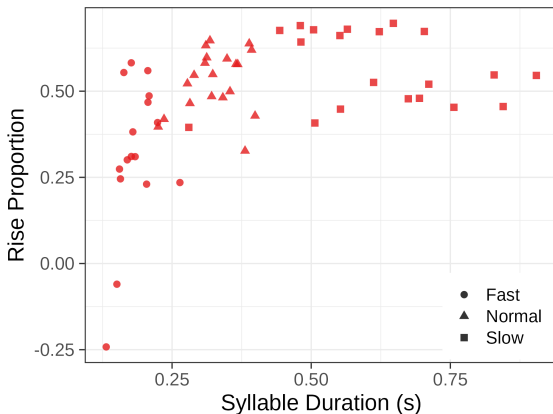


(a) rising tone, fast



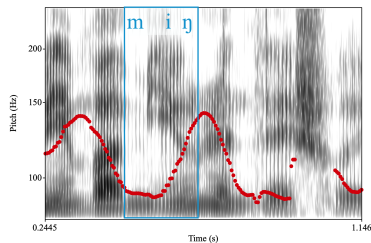
(b) rising tone, slow

Effect of syllable duration on Mandarin rising tone

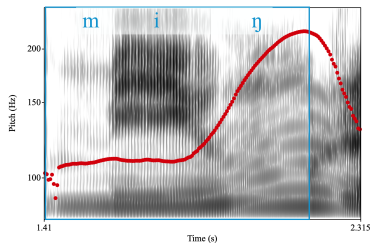


As syllable duration decreases (faster speech), the rise begins earlier in the syllable.

Case study: variation in Mandarin rising tone

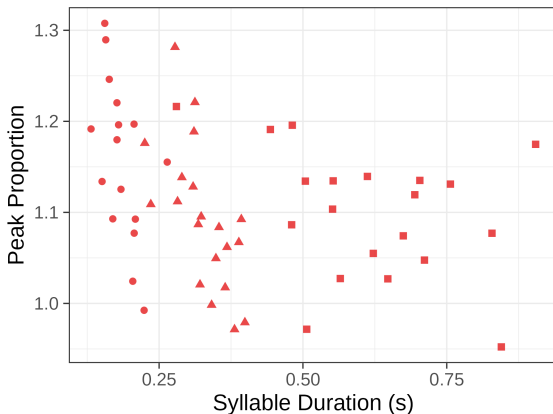


(a) rising tone, fast



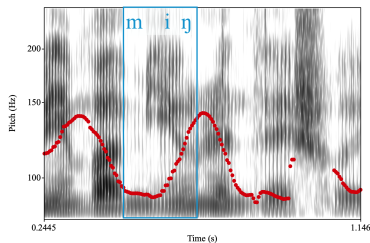
(b) rising tone, slow

Effect of syllable duration on Mandarin rising tone

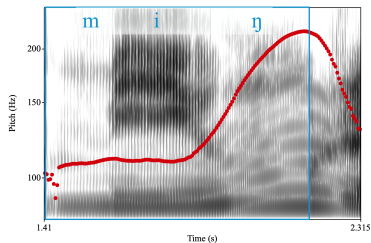


As syllable duration decreases (faster speech), the peak occurs later in the syllable.

Case study: variation in Mandarin rising tone

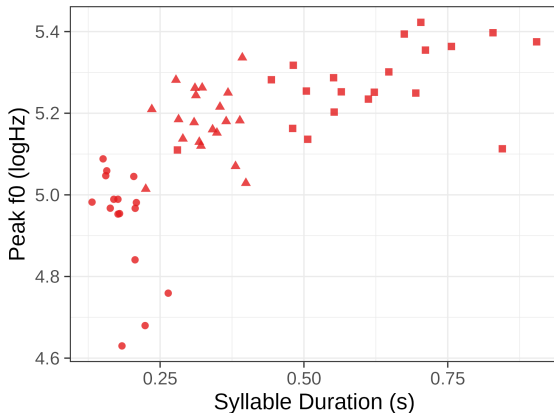


(a) rising tone, fast



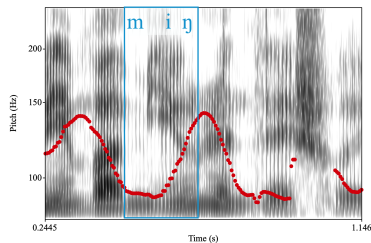
(b) rising tone, slow

Effect of syllable duration on Mandarin rising tone

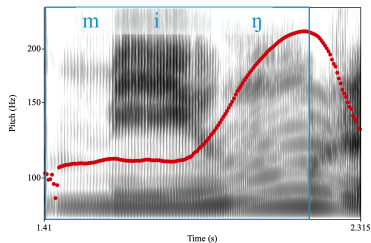


As syllable duration decreases (faster speech), the peak f0 gets lower.

Case study: variation in Mandarin rising tone

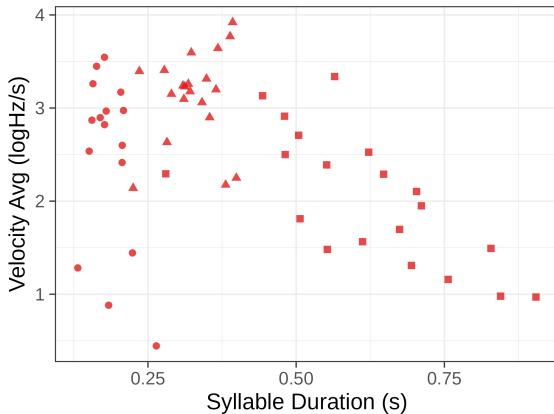


(a) rising tone, fast



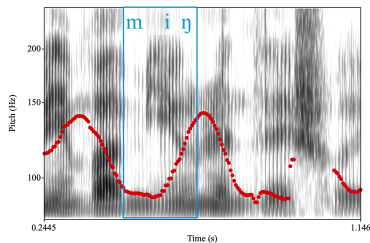
(b) rising tone, slow

Effect of syllable duration on Mandarin rising tone

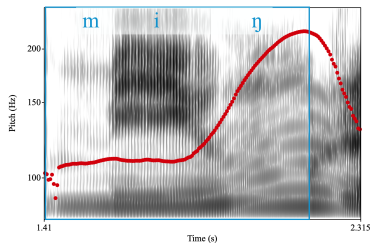


As syllable duration decreases (faster speech), the average velocity (between beginning of rise and peak) gets steeper.

Case study: variation in Mandarin rising tone



(a) rising tone, fast



(b) rising tone, slow

1. Motivation

2. Analysis

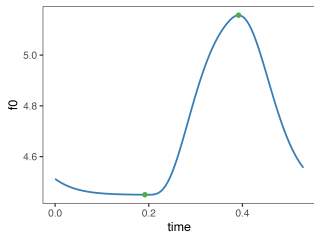
3. Results

4. References

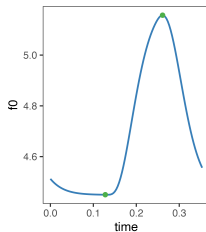
5. Appendix

Big idea of the analysis

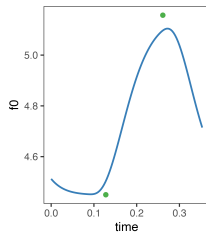
- ▶ Phonological representations impose targets on phonetic realization.
- ▶ Reaching the targets of a tone incurs increasing effort cost as syllable duration decreases.
- ▶ So tone trajectories are modified to reduce effort.
- ▶ The specific modifications are shaped by the target constraints.



(a) long syllable

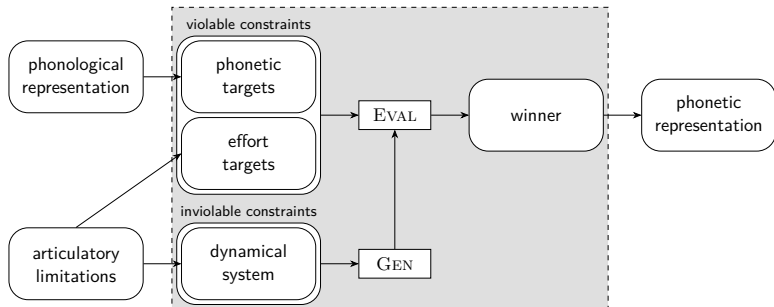


(b) short syllable

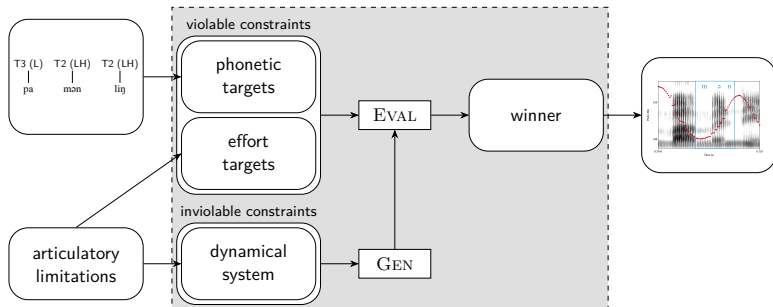


(c) undershoot

Overview of approach



Overview of approach to Mandarin rising tone



Targets of Mandarin rising tone

The phonetic targets (T_j) of Mandarin rising tone:

- ▶ **Rise time:** the rise should begin at a certain proportion of the syllable.
- ▶ **Peak time:** the peak should be reached at a certain proportion of the syllable.
- ▶ **Peak f0:** the peak should reach a certain f0 level.
- ▶ **Average velocity:** the average velocity between the rise beginning and the peak should occur at a certain speed.

Targets of Mandarin rising tone

The phonetic targets (T_j) of Mandarin rising tone:

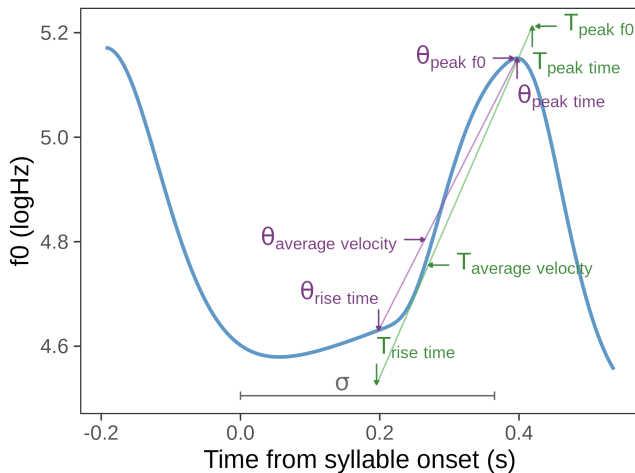
- ▶ **Rise time**: the rise should begin at a certain proportion of the syllable.
- ▶ **Peak time**: the peak should be reached at a certain proportion of the syllable.
- ▶ **Peak f0**: the peak should reach a certain f0 level.
- ▶ **Average velocity**: the average velocity between the rise beginning and the peak should occur at a certain speed.

Given a candidate trajectory, it is evaluated with respect to the weighted sum of constraint violations:

$$J(\theta) = \sum_j w_j (\theta_j - T_j)^2$$

$$j \in \{\text{rise time, peak time, peak f0, avg velocity}\}$$

Targets of Mandarin rising tone



$$J(\theta) = \sum_j w_j (\theta_j - T_j)^2$$

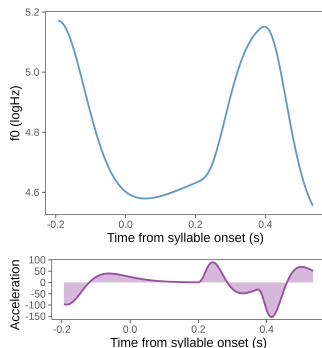
Violable constraints on articulatory movement

The speech production system imposes a violable effort minimization constraint.

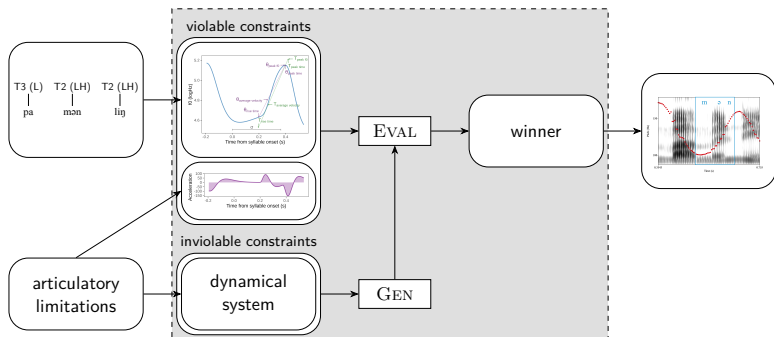
- ▶ Minimize acceleration over the trajectory (cf. Nelson (1983)):

$$\theta_{\text{RMS accl}} = \sqrt{\frac{1}{T} \int (f_0''(t))^2 dt}$$

- ▶ The target for RMS acceleration is 0 ($T_{\text{RMS accl}} = 0$).



Violable constraints on articulatory movement

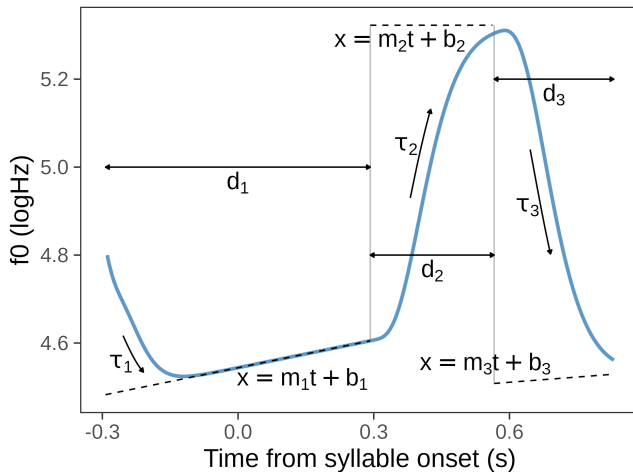


Inviolable constraints on articulatory movement

The speech production system imposes an inviolable constraint on f_0 trajectory shape.

- ▶ f_0 trajectories are the response of a dynamical system to a piecewise control signal.
 - ▶ General approach pioneered by Öhman and Lindqvist (1968) and Fujisaki and Nagashima (1969).
 - ▶ Specific model follows Birkholz et al. (2018).
- ▶ Dynamical system: cascade of 5 first-order linear systems.
- ▶ For the **duration** of each linear piece, f_0 asymptotes towards the **line** at a rate determined by a **time constant**, τ_i .
- ▶ Parameters of each control signal piece: $\{d_i, \tau_i, b_i, m_i\}$.

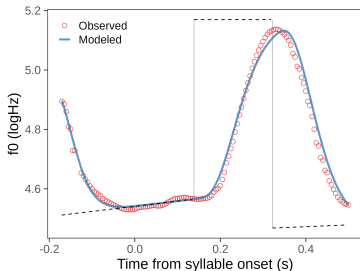
Inviolable constraints on articulatory movement



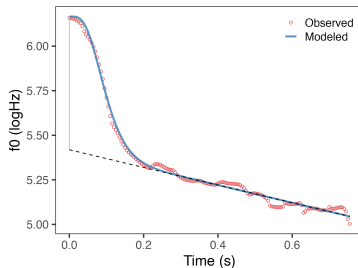
Inviolable constraints on articulatory movement

Parameters of each control signal piece: $\{d_i, \tau_i, b_i, m_i\}$.

- ▶ Rapid rises/falls are generated by steps between linear pieces.
- ▶ Sloped pieces are required to generate sloped linear plateaus observed in tone and intonation.

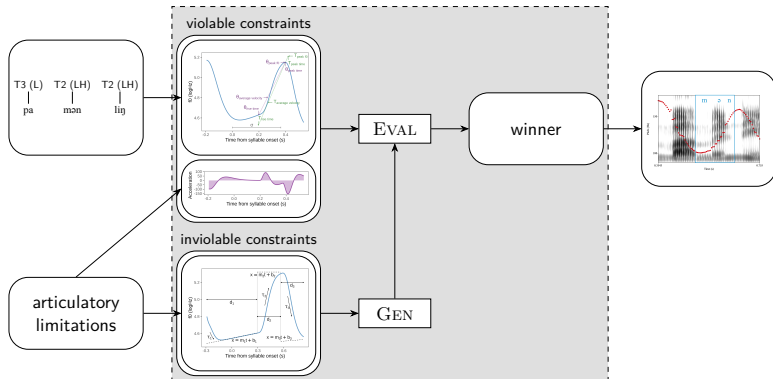


(a) Mandarin rising tone

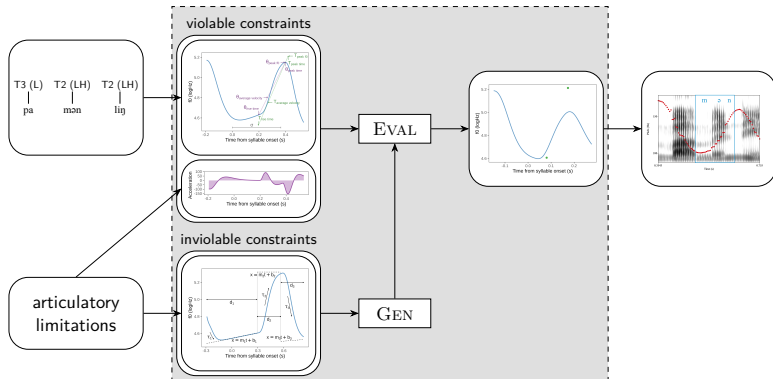


(b) English H* L-

Inviolable constraints on articulatory movement

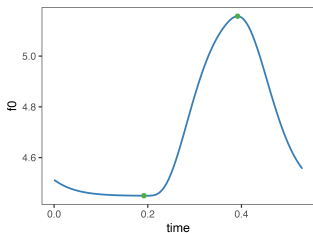


Full model in the Mandarin case study

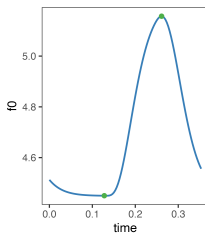


Deriving the observed patterns from constraint conflict

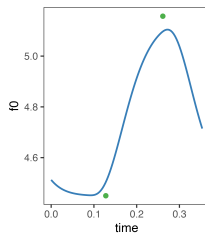
- ▶ When syllable duration is short, realizing T_{peak} time, T_{rise} time, $T_{\text{peak}} f_0 \Rightarrow$ high acceleration cost
- ▶ Constraint conflict is resolved by compromise: the best trajectory deviates from each conflicting target to minimize total cost.



(a) long syllable



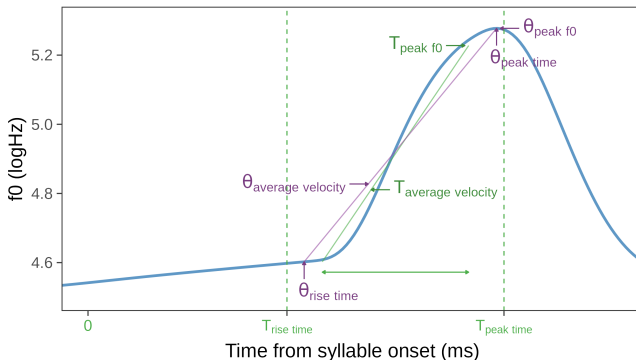
(b) short syllable



(c) undershoot

Deriving the observed patterns from constraint conflict

- ▶ When syllable duration is long, realizing $T_{\text{peak time}}$, $T_{\text{rise time}}$, $T_{\text{peak f0}} \Rightarrow \theta_{\text{avg velocity}}$ is low
- ▶ $T_{\text{avg velocity}}$ is high - a steep rise is preferred



1. Motivation

2. Analysis

3. Results

4. References

5. Appendix

Fitting the Mandarin data

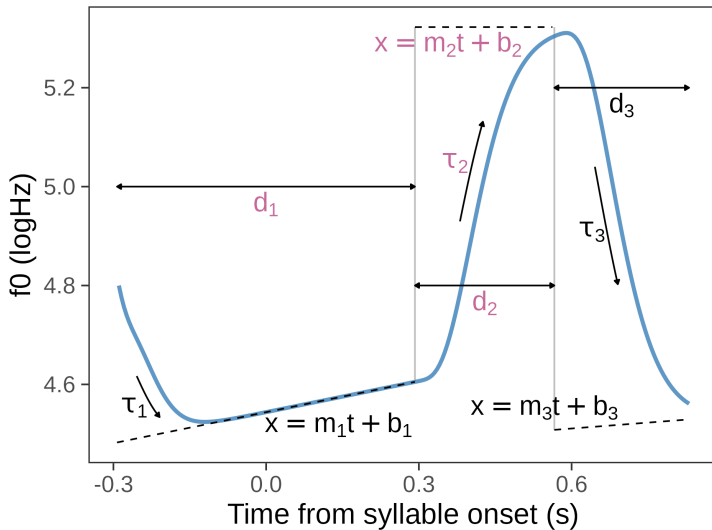
To evaluate the analysis, we fit the model to the data.

- ▶ Goal: Find the target values and constraint weights that yield the best fit to the data.

We are only modeling the rising tone, but evaluation of the constraints depends on the preceding and following context.

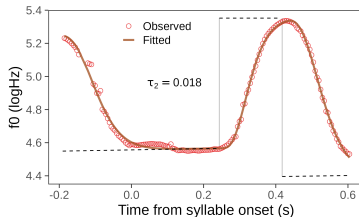
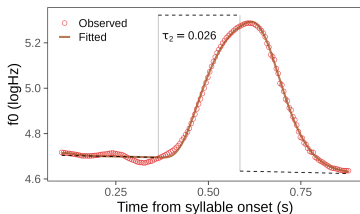
- ▶ Step 1: Fit the dynamical model to the observed f_0 contours to infer the control signal for the trajectory for the target tones and the preceding and following tones.
- ▶ Step 2: Take the fits from Step 1 for the preceding and following context as given, and use our grammar to derive the realization of the rising tone.

Fitting the Mandarin data

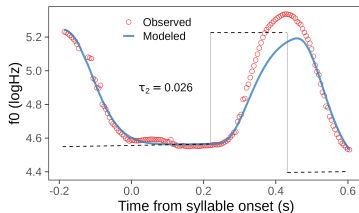
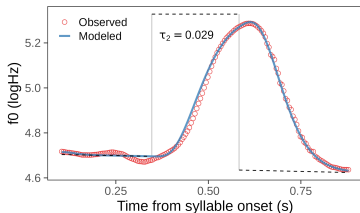


Fitting the Mandarin data

Step 1. Estimate the control signals that generate the f_0 trajectories.



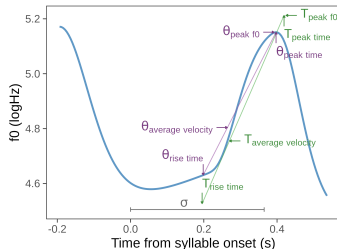
Step 2. Generate the rising tones, taking the context as given.



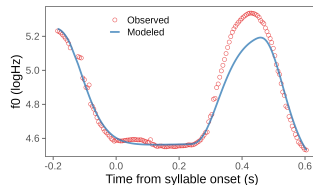
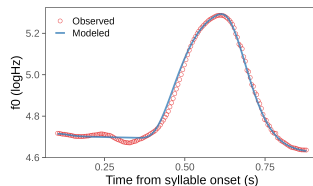
Fitting the Mandarin data

Targets and weights for one speaker:

j	T_j	w_j
rise time	0.54	1
peak time	1.15	1.97
peak f0	5.21 logHz	2.1e-01
avg velocity	3.07 logHz/s	1.1e-02
rms accel.	0 logHz/s ²	9.7e-06
final f0	from context	9.2



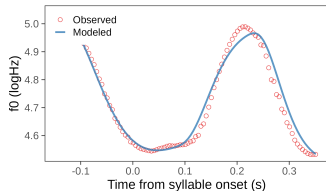
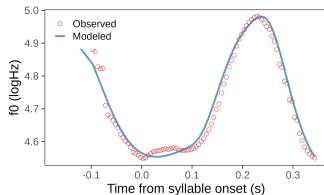
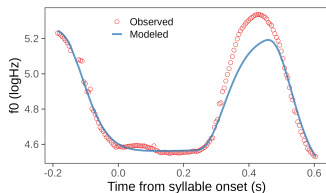
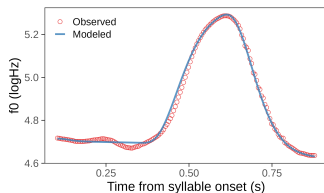
Given $\{T_j, w_j\}$, modeled trajectories are generated **per utterance**.



Find targets and weights that minimize MSE of the trajectories. ↻ 🔍 38/53

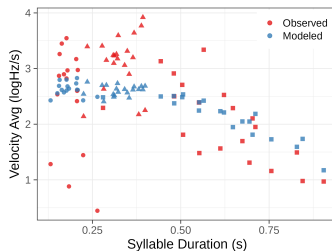
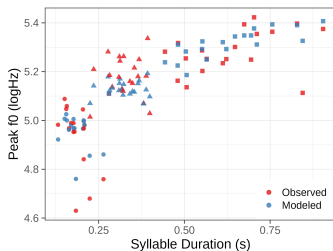
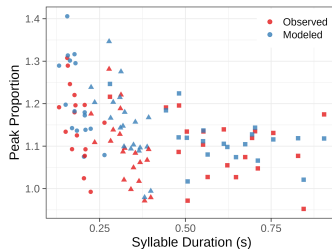
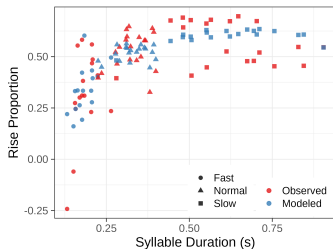
Results

Fit of the model to individual utterances varies, but the overall quantitative fit to the data is good.



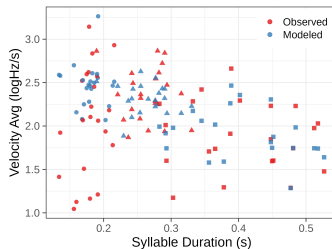
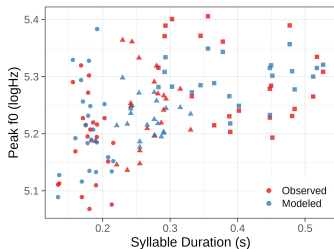
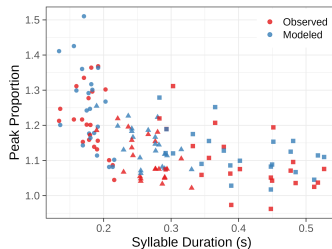
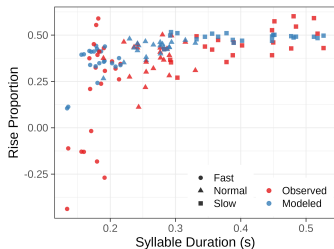
Results

The model derives the observed patterns of variation in realization of the rising tone as a function of duration.



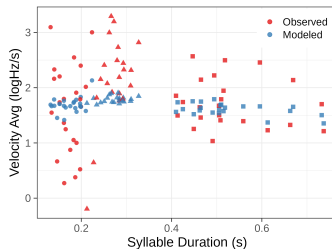
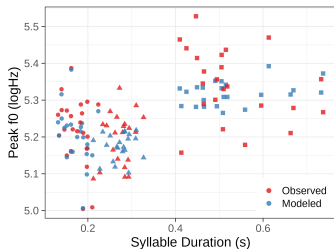
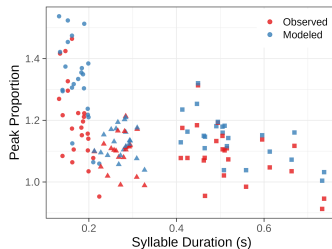
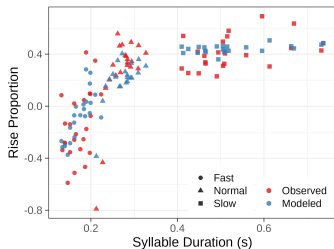
Results

The model derives the observed patterns of variation in realization of the rising tone as a function of duration (another speaker).



Results

The model derives the observed patterns of variation in realization of the rising tone as a function of duration (one more speaker).

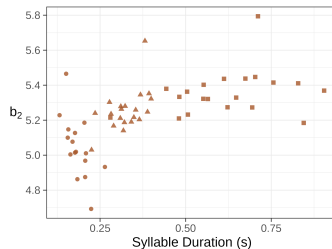
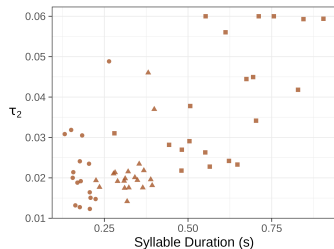
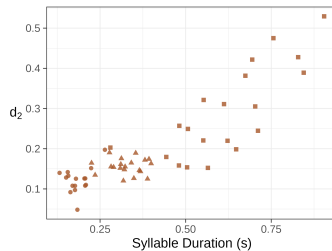
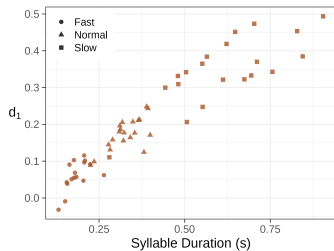


Look-ahead is central to the generation of control signals

As a function of speech rate, there are systematic trends in:

- ▶ Variation in the realization of the rising tone.
- ▶ $\Rightarrow$ Variation in the control signal.

Variation in control signal parameters



Look-ahead is central to the generation of control signals

As a function of speech rate, there are systematic trends in:

- ▶ Variation in the realization of the rising tone.
- ▶ $\Rightarrow$ Variation in the control signal.

Our analysis successfully derives the variation in the control signal.

- ▶ Via optimization with respect to **conflicting constraints on the f_0 trajectory**.
- ▶ Relies on look-ahead to the **output of the dynamical system** when generating the control signal.
- ▶ Cf. Simko and Cummins (2010), Elie et al. (2024), Shaw and Chen (2019).

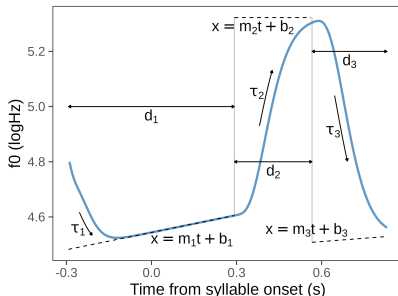
Bibliography I

- Beckman, M. E. and Pierrehumbert, J. B. (1988). Japanese Tone Structure. MIT Press, Cambridge, MA.
- Birkholz, P., Schmager, P., and Xu, Y. (2018). Estimation of pitch targets from speech signals by joint regularized optimization. In 2018 26th European Signal Processing Conference (EUSIPCO), pages 2075–2079.
- Browman, C. P. and Goldstein, L. (1992). Articulatory phonology: an overview. Phonetica, 49(3–4):155–180.
- Elie, B., Šimko, J., and Turk, A. (2024). Optimization-based planning of speech articulation using general tau theory. Speech Communication, 160.
- Flemming, E. and Cho, H. (2017). The phonetic specification of contour tones: The phonetic specification of contour tones: Evidence from the Mandarin rising tone. Phonology, 34(1):1–40.
- Fujisaki, H. and Nagashima, S. (1969). A model for the synthesis of pitch contours of connected speech. Annual Report of the Engineering Research Institute, 28:53–60.
- Nelson, W. (1983). Physical principles for economies of skilled movements. Biological Cybernetics, 46:135–147.
- Pierrehumbert, J. B. (1980). The Phonology and Phonetics of English Intonation. PhD thesis, Massachusetts Institute of Technology.
- Shaw, J. A. and Chen, W.-r. (2019). Spatially conditioned speech timing: Evidence and implications. Frontiers in Psychology, 10.

- Simko, J. and Cummins, F. (2010). Embodied task dynamics. Psychological Review, 117(4):1229–1246.
- Öhman, S. and Lindqvist, J. (1968). Analysis-by-synthesis of prosodic pitch contours. STUF - Language Typology and Universals, 21(1–6):164–170.

Step 1

- ▶ Utterances have the form: [tə^hɪŋ nɪn pà ˨ ˨ tsâi ʂó jǐ pjên].
- ▶ Extract f0 trajectory for the low-rise-rise sequence.
 - ▶ Start as early as possible in the fall towards the low tone.
 - ▶ End as close as possible to the f0 minimum for the second rising tone.
- ▶ Search for the control signal with three pieces that yields the best match to this f0 contour.



Step 2: generating trajectories

Generation is optimization:

- ▶ Given tonal context $\{\tau_1, b_1, m_1, \tau_3, b_3, m_3\}$, initial derivatives and the duration of the target syllable,
- ▶ Find the values of the control signal parameters $\{d_1, \tau_2, b_2, d_2\}$ that yield the f0 trajectory that best satisfies the grammatical constraints by minimizing the weighted sum of constraint violations.

$$\sum_j w_j (\theta_j - T_j)^2$$

$$j \in \{\text{rise time, peak time, peak f0, avg velocity, RMS accl, final f0}\}$$

$$T_{\text{RMS accl}} = 0, T_{\text{final f0}} \text{ from context.}$$

- ▶ Values for T_j and w_j are speaker-specific.

Dynamical model (Birkholz et al., 2018)

- ▶ Cascade of N identical first-order linear systems ($N = 5$).
- ▶ The input-output relation in the frequency domain is:

$$Y(s) \left[(1 + s\tau)^N \right] = X(s)$$

- ▶ The response to a control signal piece $x(t) = mt + b$ is:

$$y(t) = \left(c_0 + c_1 t + \dots + c_{N-1} t^{N-1} \right) e^{-t/\tau} + (mt + b)$$

- ▶ τ is the time constant.
- ▶ $c_0, \dots, c_{N-1}$ are constants that depend on the derivatives of $y(t)$ at the start of the piece ($t = 0$).

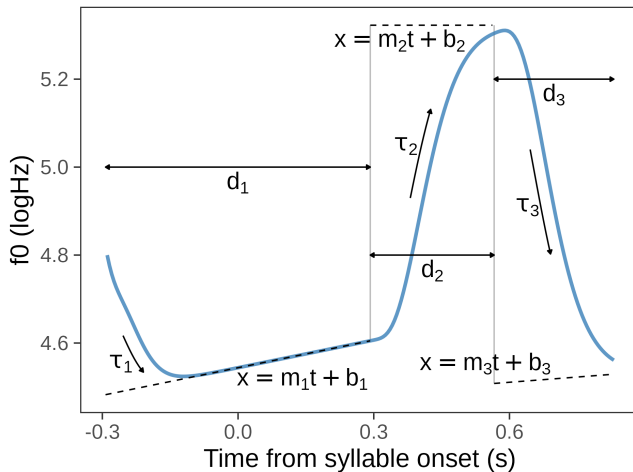
$$c_0 = y(0) - b$$

$$c_n = \left(y^{(n)}(0) - \frac{d^n}{dt^n}(mt + b) - \sum_{i=0}^{n-1} c_i \left(\frac{-1}{\tau} \right)^{n-i} \binom{n}{i} i! \right) \frac{1}{n!}$$

The derivatives of $y(t)$ are given by:

$$y^{(n)}(t) = \frac{d^n}{dt^n}(mt + b) + e^{-t/\tau} \cdot \sum_{i=0}^{n-1} t^i \left(\sum_{j=0}^{\min\{n-1-i, n\}} \left(\frac{-1}{\tau} \right)^{n-j} \binom{n}{j} c_{i+j} \cdot \frac{(i+j)!}{i!} \right)$$

Dynamical model (Birkholz et al., 2018)



Targets

