Toward a grammar of timing in speech production

Christopher Geissler
Department of Linguistics
Carleton College

January 25, 2024
Slides available on cageissler.github.io/resources

Roadmap

- **“Discrete phonology, continuous phonetics”**
- Coupled oscillators: timing in phonology
- Problems
  - Effects of tone
  - Surface timing goals
- Next steps: typology of coordination
- Conclusion

Discrete phonology

Discrete behavior

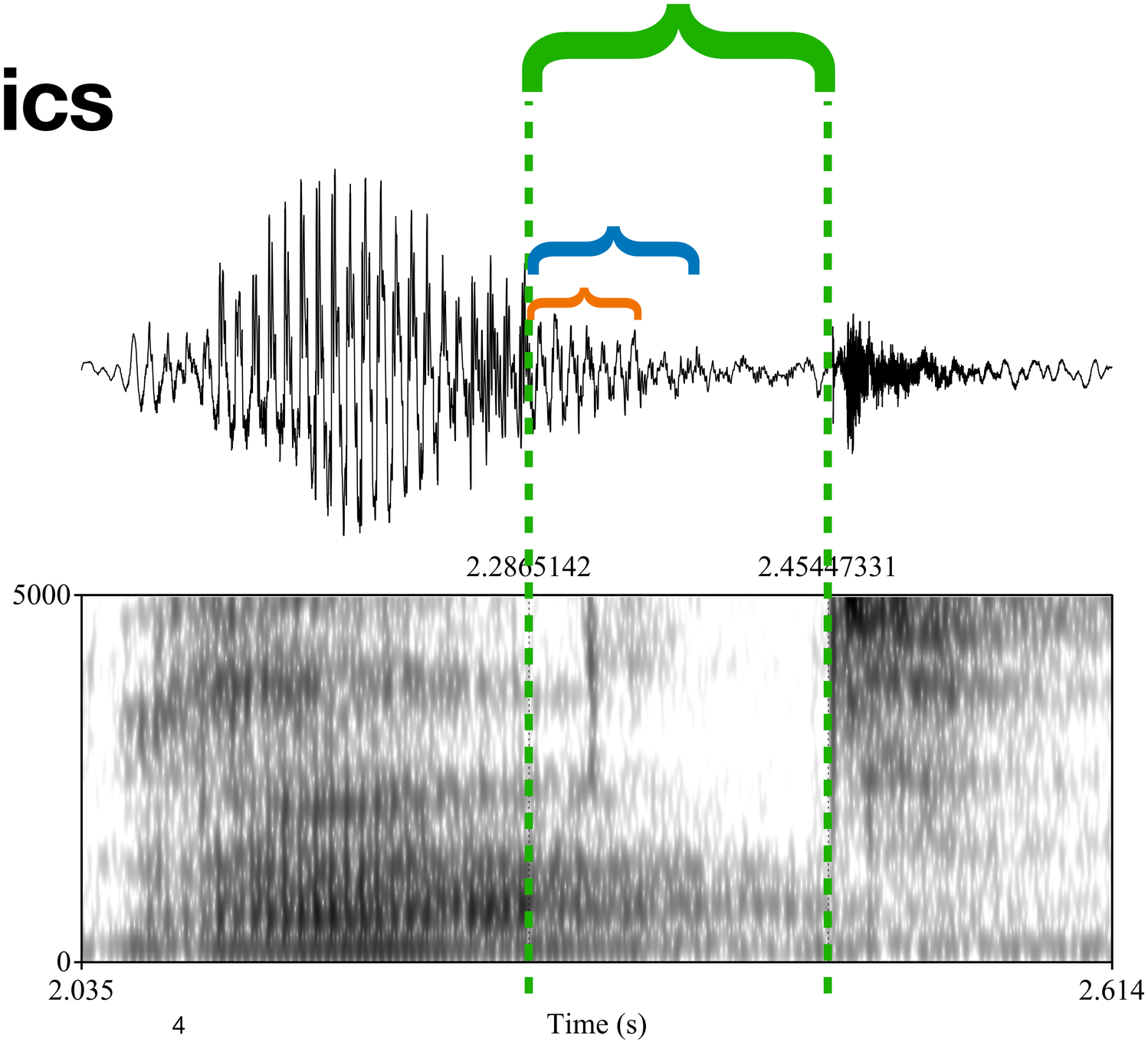
- In German, voiced consonants are voiceless when they occur at the end of words (but not elsewhere):

  - *Ra̲d̲* ‘wheel’ [ʀɑ̲t̲], but plural *Räd̲er* [ʀɛ̲d̲ɐ]
  - compare:
    Ra̲t̲ ‘council’ [ʀɑ̲t̲], but plural *Rät̲e* [ʀɛ̲t̲ə]

Intro-level phonetics

Continuous behavior

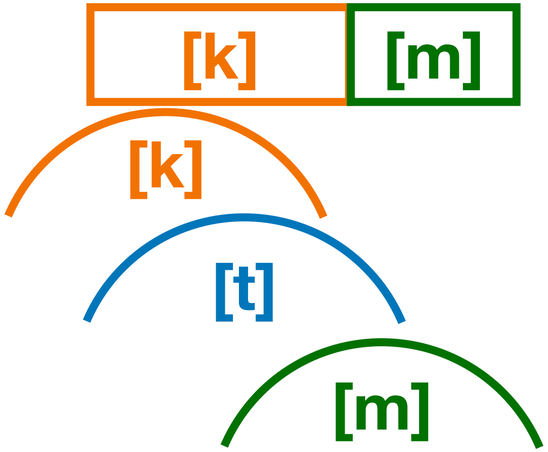
- *Rat̲/Rad̲* ‘wheel’ [ʀɑt̲]
- Where does the voicing end?
  - The whole closure?
  - Periodic sound?
  - Regular periodicity?



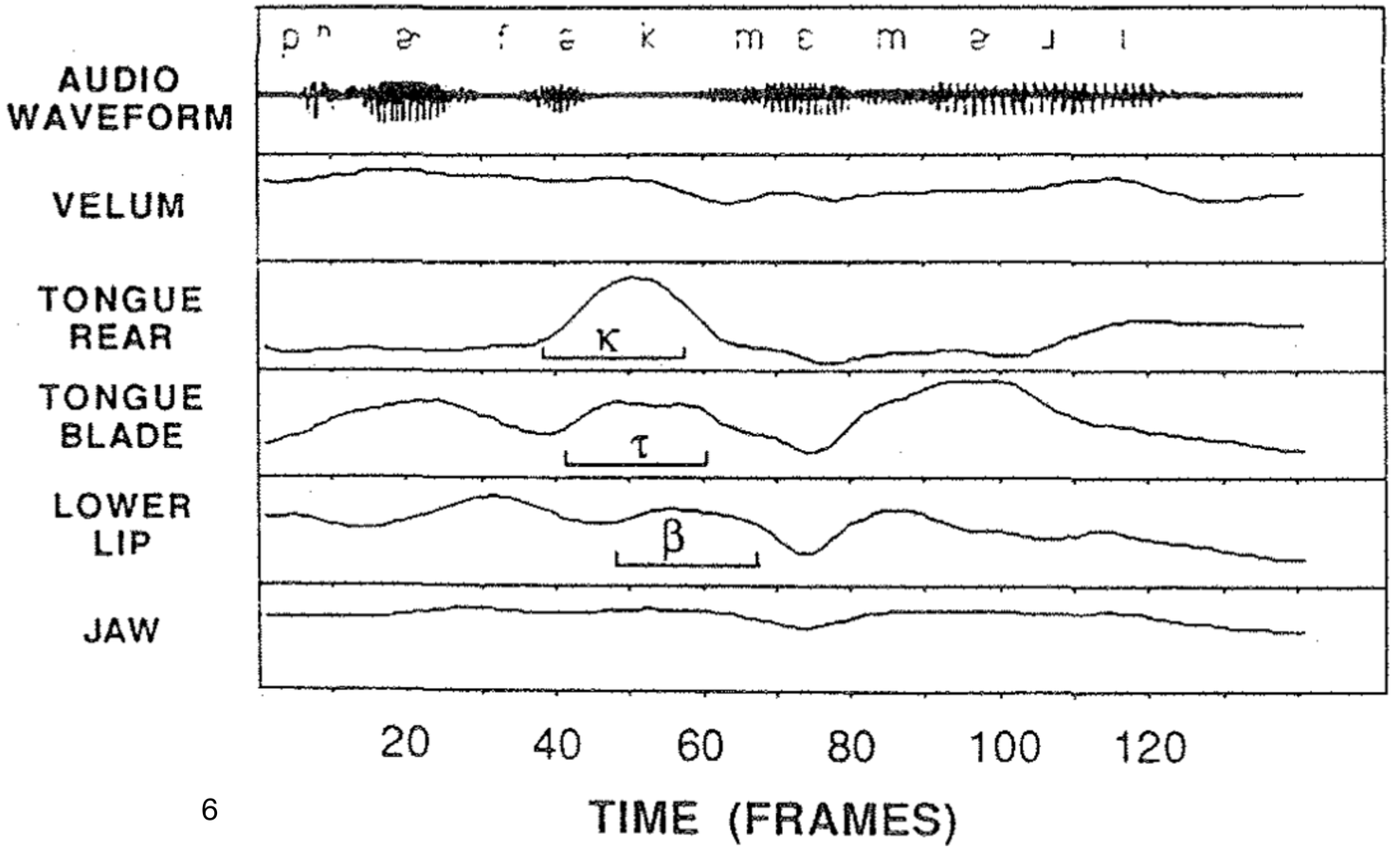
Probabilistic discrete phonology

- In English, t/d at the end of a word sometimes isn't there
  - *rift* = [ɹɪft̚] or [ɹɪf_]; *build* = [bɹɬd] or [bɹɬ]
  - More likely among some groups
  - More likely in some social contexts
  - More likely around some sounds
  - More likely in *mist* than in *missed*

Articulatory complications



- *Perfect memory*
- At least some “deleted” t’s/d’s are visible in articulation, but not in acoustics
  - (Actually it’s most)



Midsagittal sections

(Browman & Goldstein 1988, Purse 2019)

Gestures

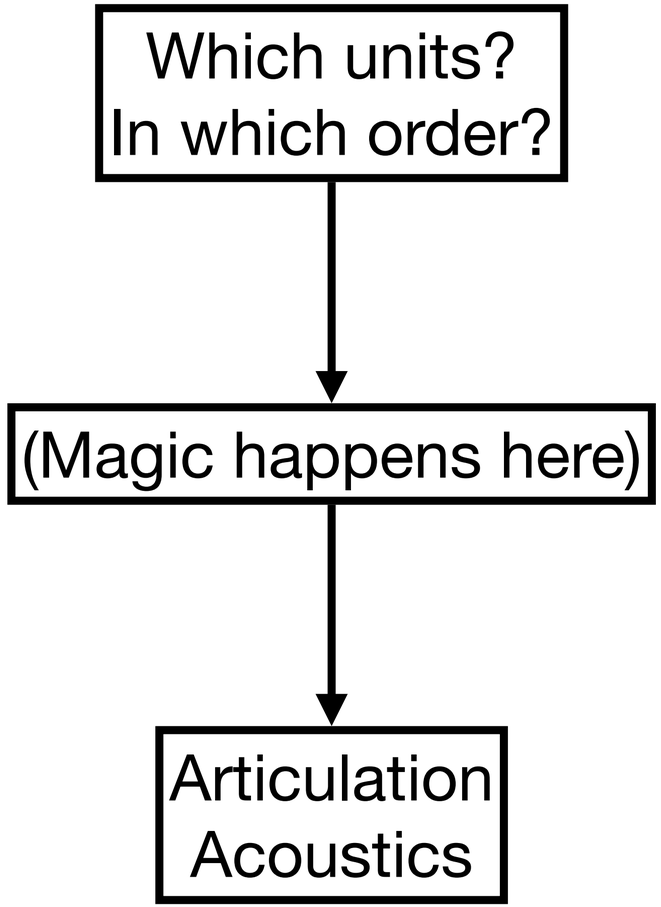
In Articulatory Phonology

- Abstract, hierarchical control unit for linguistically-defined goal-directed movement *(Pouplier 2020)*
  - Motor equivalence
  - Equifinality

A Theory of the Interface

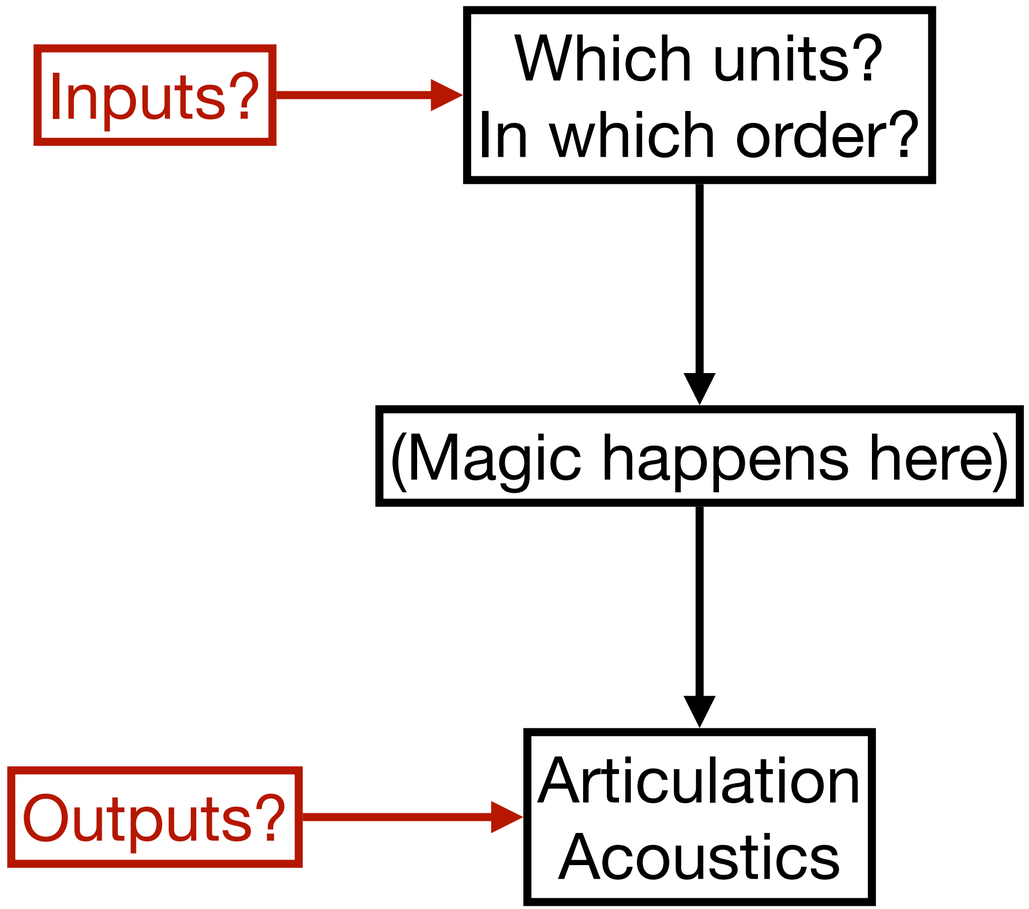
- “Phonology”

- Phonetic observables



A Theory of the Interface

- "Phonology"
- Phonetic observables



Roadmap

- **“Discrete phonology, continuous phonetics”**
- Coupled oscillators: timing in phonology
- Problems
  - Effects of tone
  - Surface timing goals
- Next steps: typology of coordination
- Conclusion

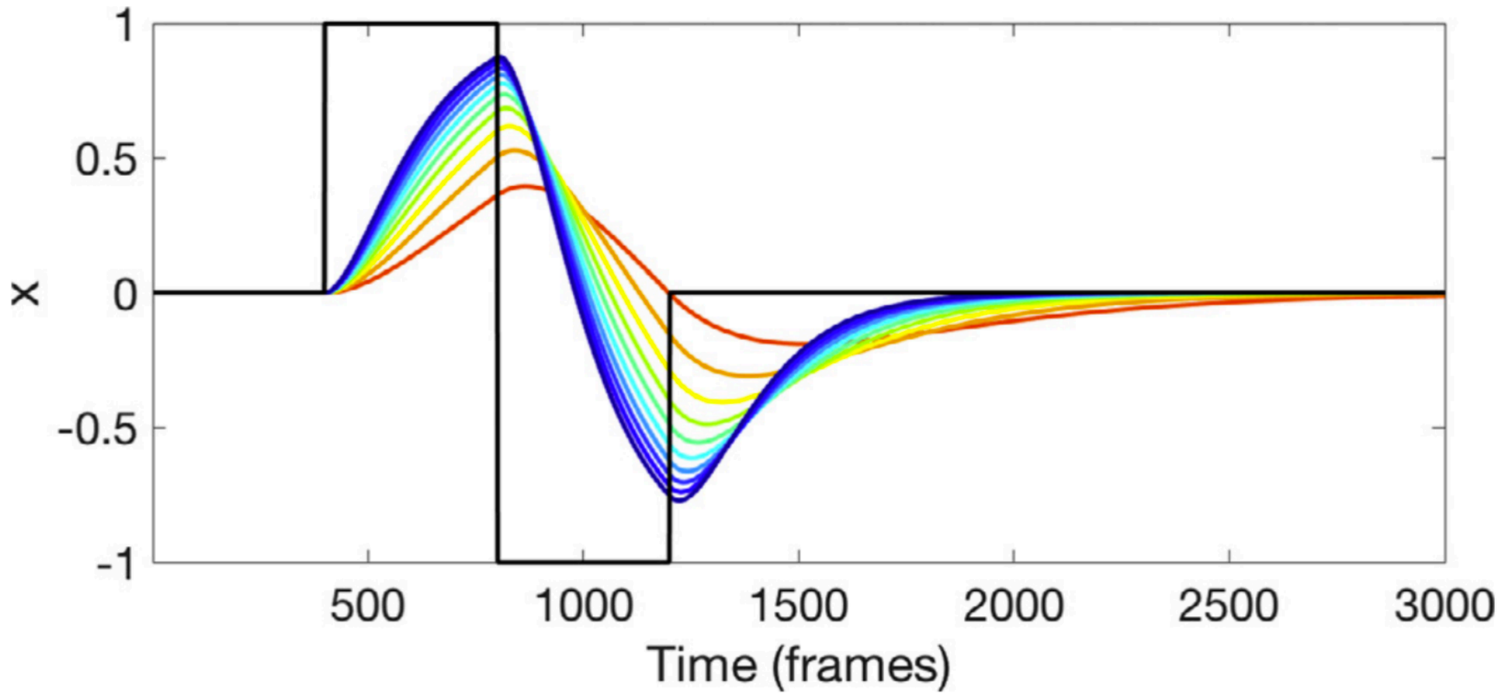
Oscillator model

(Haken et al. 1985, Saltzman & Munhall 1989, Nam & Saltzman 2003)

- Model kinematics as critically-damped mass-spring oscillator
- Asymptotically approaches target (equilibrium position) as fast as possible

$$ma + bv + k(x - C) = 0$$

acceleration, velocity, position, stiffness, target

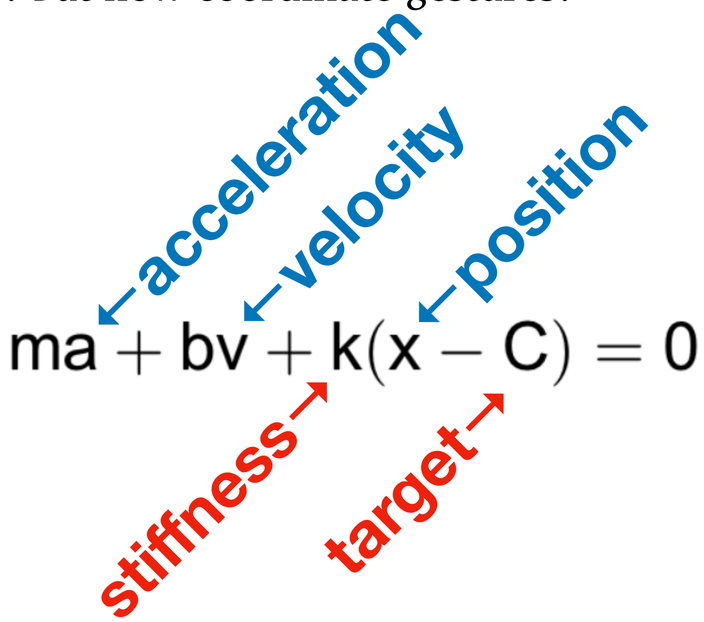


(Iskarous 2013)

Oscillator model

(Haken et al. 1985, Saltzman & Munhall 1989, Nam & Saltzman 2003)

- Requires: position(/velocity/acceleration) at a point, and target
- Does not reference *time* of target attainment
- A position is a function of another position/vel/acc and stiffness
- Stiffness is a function of natural frequency and mass
- Has equifinality, consistent with motor equivalency
- … but how coordinate gestures?



Oscillator model

(Haken et al. 1985, Saltzman & Munhall 1989, Nam & Saltzman 2003)

Note absence of target time

$$ma + bv + k(x - C) = 0$$

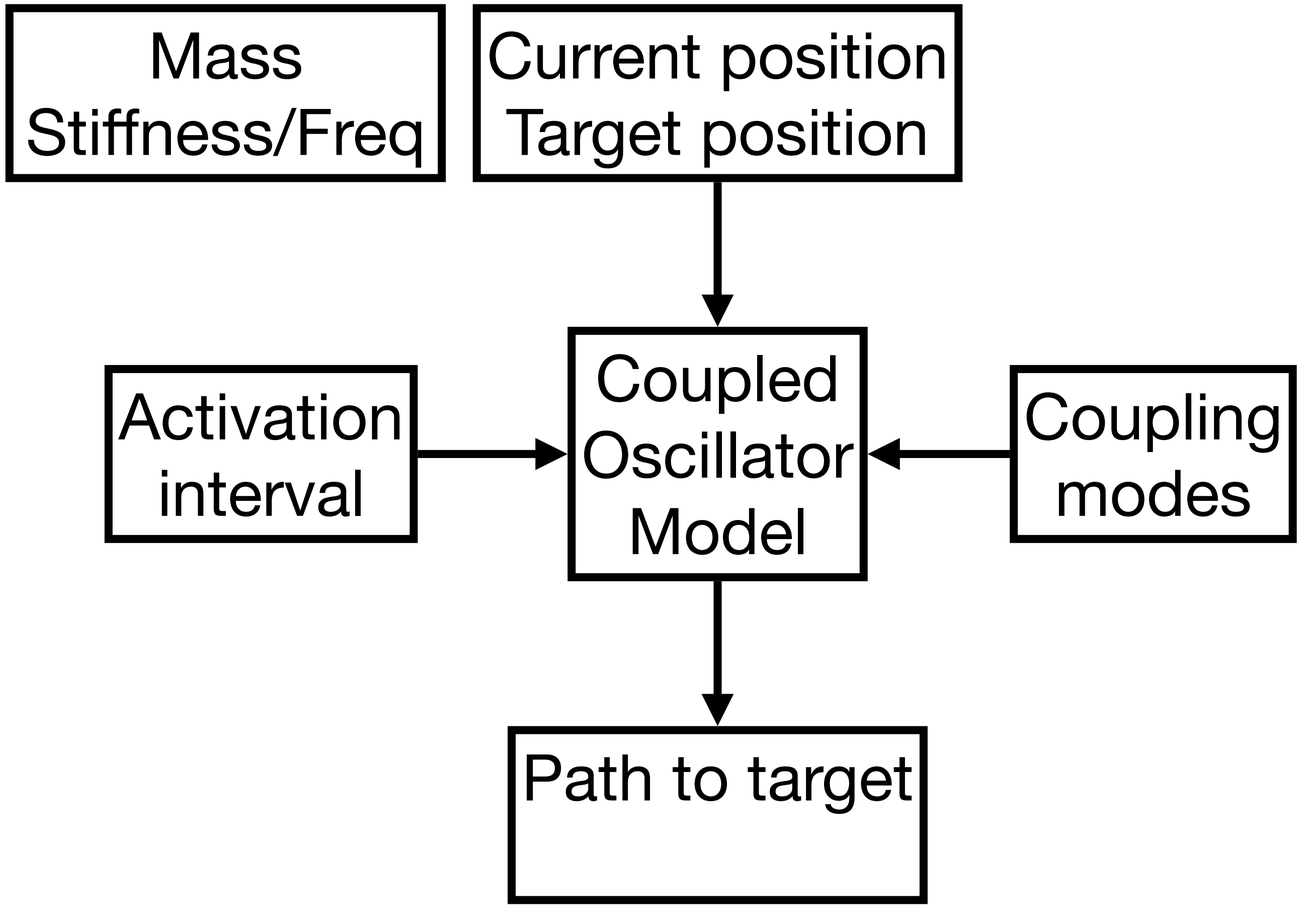
acceleration

velocity

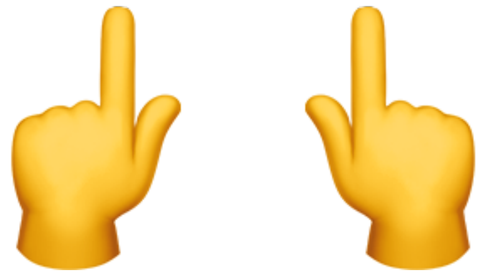
position

stiffness

target



Bimanual tapping interlude



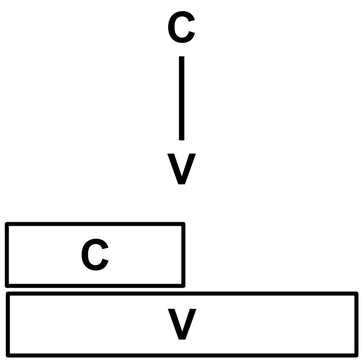
Oscillators

- Synchronization in non-speech and speech movements:
  - “pa… pa… pa… pa.pa[…]pa.pa.pa.pa”
  - “ap… ap… ap… ap.ap.[…]pa.pa.pa.pa”
- Tapping: “in-phase” more stable than “anti-phase”
  (both more stable than any other phasing)
  … in speech too?

CV vs. VC syllables

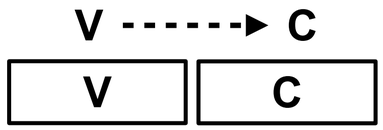
in-phase

[pa]	
LIPS	Labial closure
TONGUE TIP	
TONGUE BODY	pharyngeal wide



anti-phase

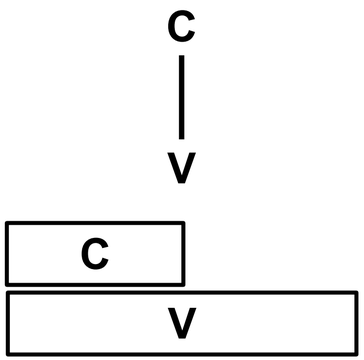
[ap]	
LIPS	labial closure
TONGUE TIP	
TONGUE BODY	pharyngeal wide



CV vs. VC syllables

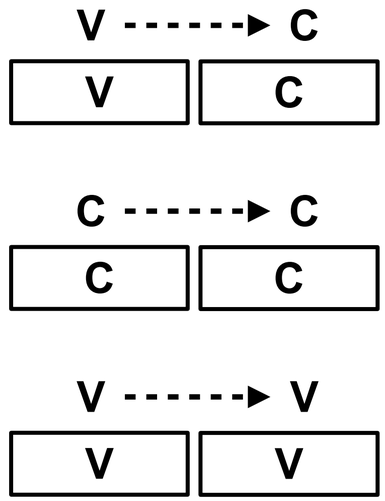
in-phase

[pa]	
LIPS	Labial closure
TONGUE TIP	
TONGUE BODY	pharyngeal wide



anti-phase

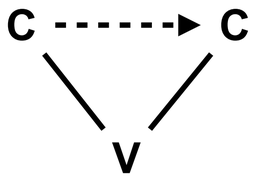
[ap]	
LIPS	labial closure
TONGUE TIP	
TONGUE BODY	pharyngeal wide

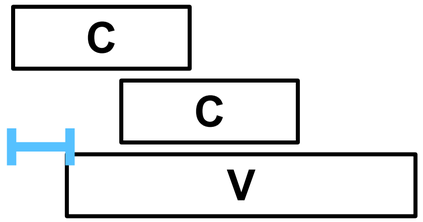


What about clusters?

- Empirically, onset clusters overlap

/spa/ ‘spa’	
LIPS	labial closure
TONGUE TIP	alveolar critical
TONGUE BODY	pharyngeal wide





What about tone?

- Empirically, V lags following C
  - (In *lexical tone* languages only)

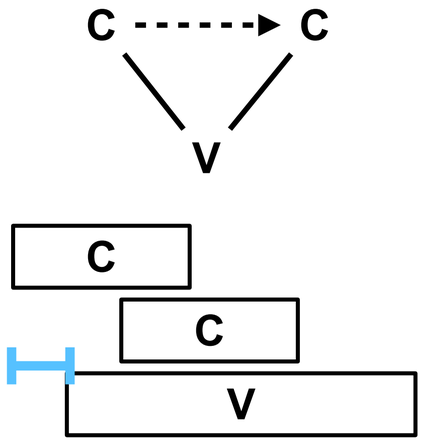
/pá/	
LIPS	labial closure
TONGUE TIP	
TONGUE BODY	pharyngeal wide
pitch (?)	high

(Gao 2008, Niemann et al. 2011, Karlin 2014)

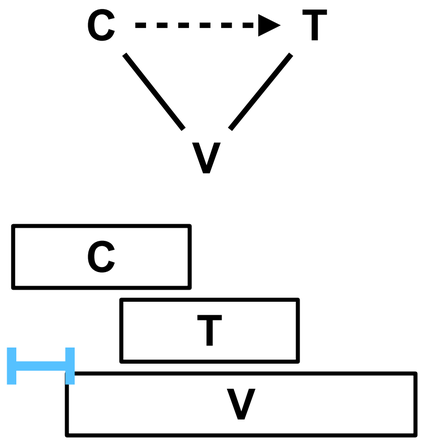
Competitive coupling account 🎉

- Unifies clusters and tone (neat for typology)
- Unifies syllables (and up?), contrast, and planning

/spa/ ‘spa’	
LIPS	labial closure
TONGUE TIP	alveolar critical
TONGUE BODY	pharyngeal wide



/pá/	
LIPS	labial closure
TONGUE TIP	
TONGUE BODY	pharyngeal wide
pitch (?)	high



Roadmap

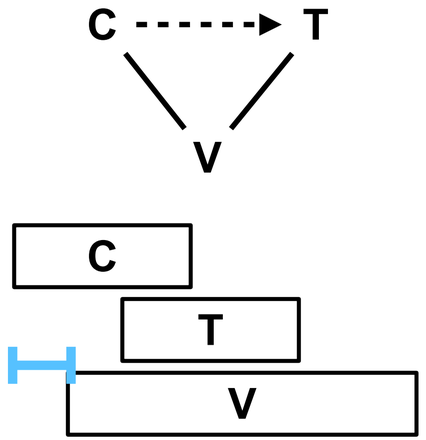
- Phonology and articulatory gestures
- Coordinating gestures: the Coupled Oscillator Model
- Problems
  - **Tibetan tone study**
- Toward solutions: Analysis-by-synthesis
- Conclusion

Doubts

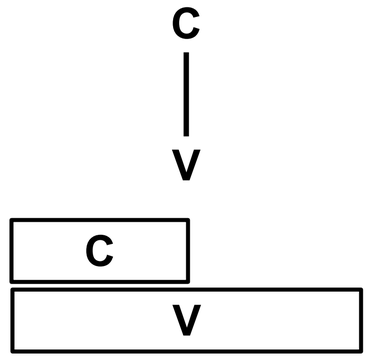
- “Competitive coupling” … is weird
  - Is it much better than just stipulating phasing?
  - Doesn’t work for clusters of 3+
- Can we really generalize from ‘papapapa’ to regular speech?
- Should we rely on *start* of a gesture or the *end* of a gesture?

Predictions of Coupled Oscillator Model

- If there is a tone gesture in a syllable:
  - C-V timing like in clusters:
    C-V lag positive, ~50ms



- If there is no tone in that syllable:
  - Simultaneous C & V:
    C-V lag ~0ms



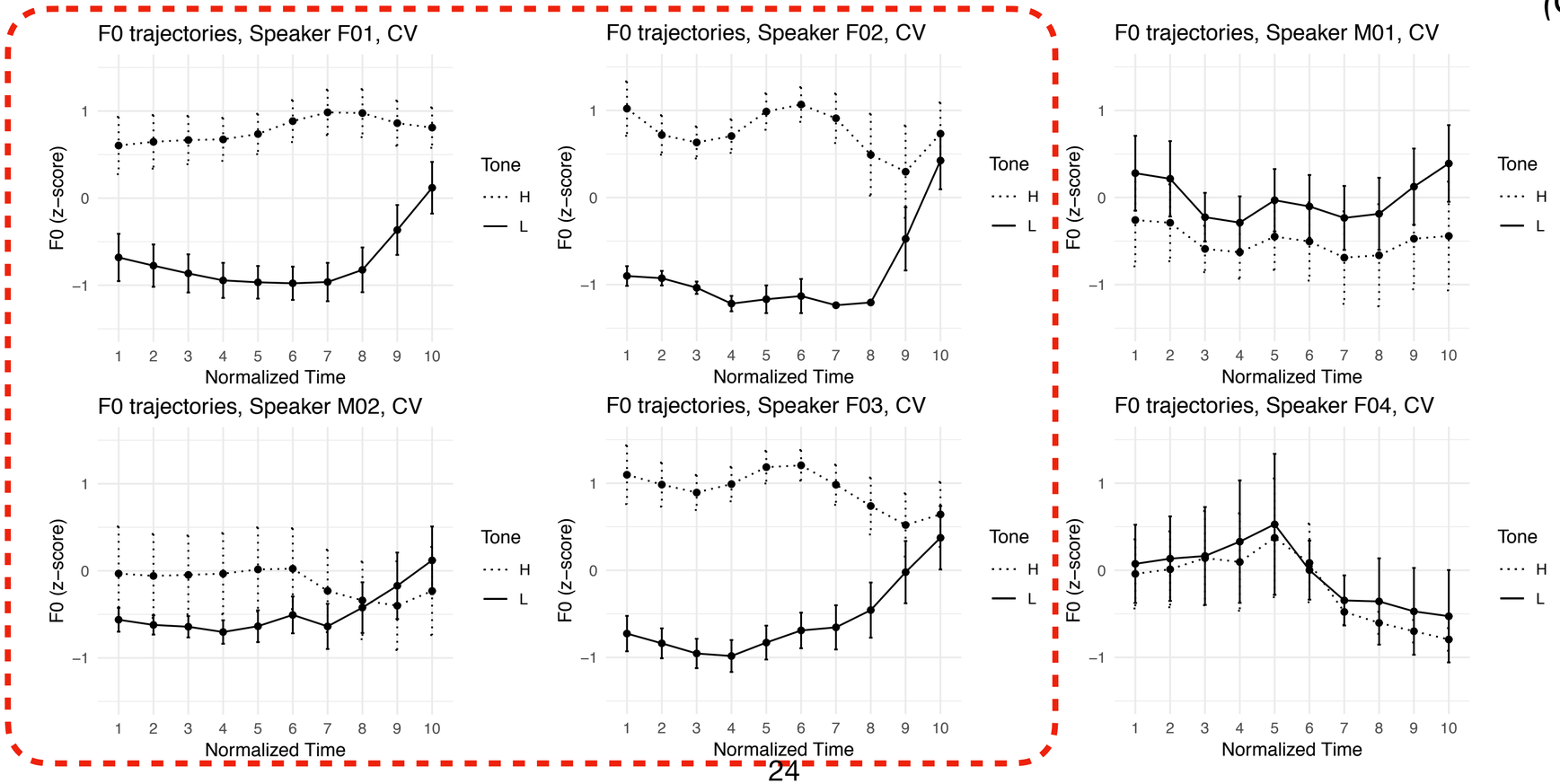
The perfect test case?

A language where some speakers produce tone and others don't

(Geissler 2019, 2021)

- 4 speakers produce a tone contrast, two do not (images: /mV/)

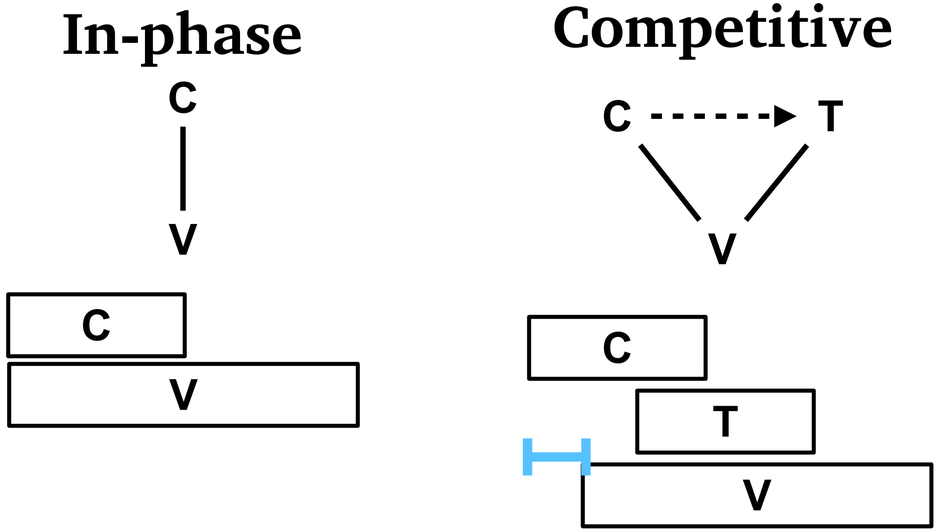
(Geissler et al. 2021)

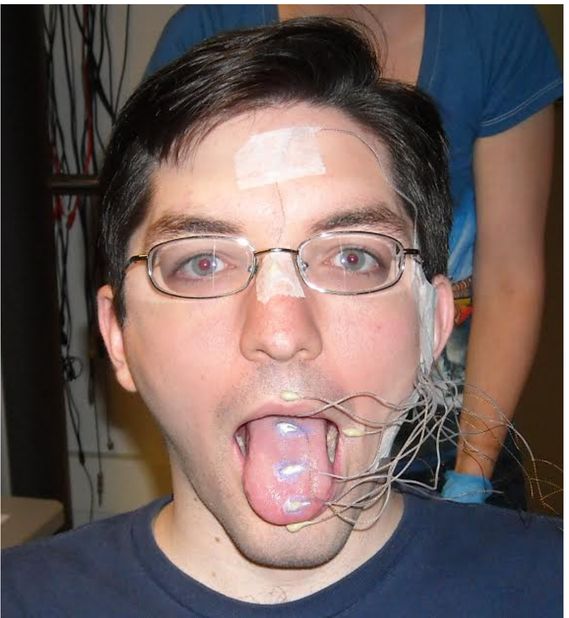


EMA study

articulatory trajectories

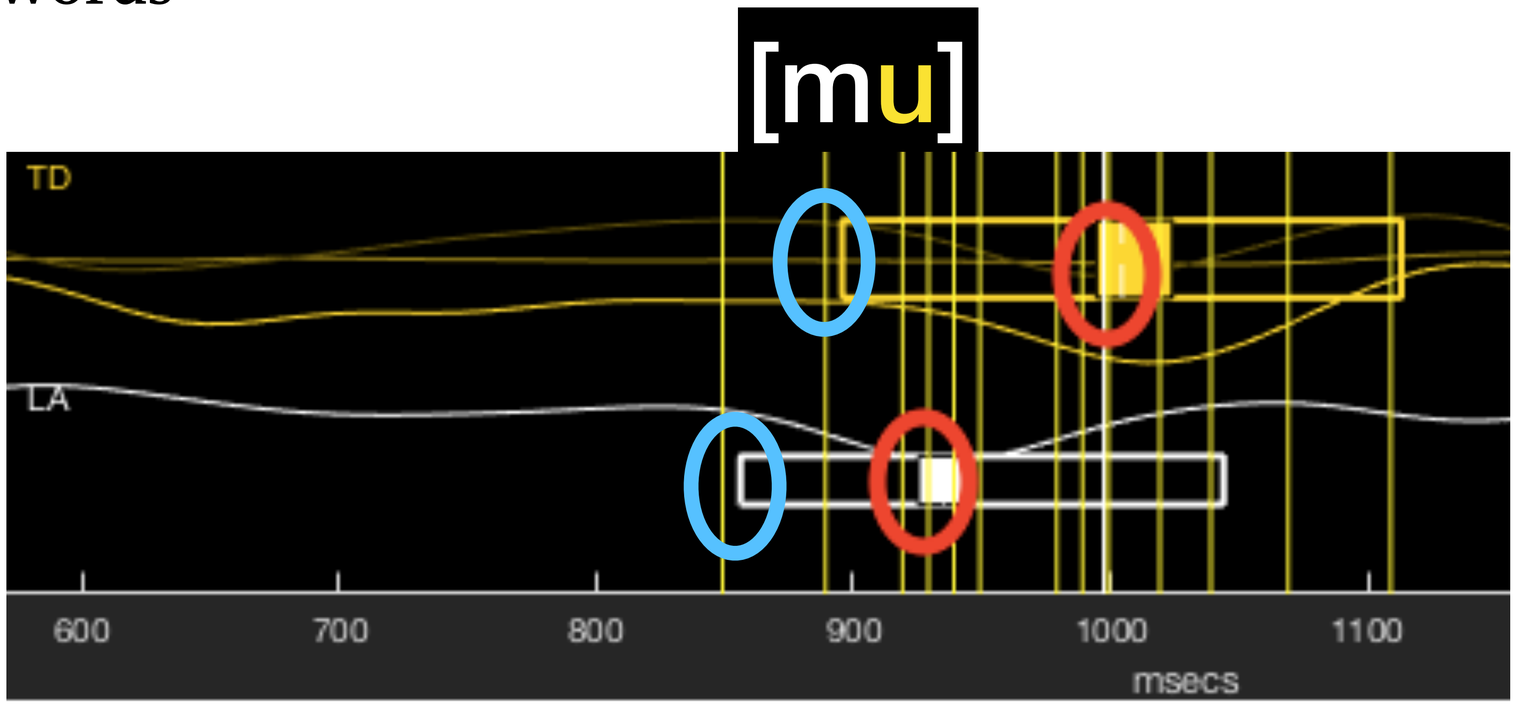
- [p p^h m]: distance between lip sensors
- [i]→[u o a]: tongue dorsum retraction
- H, L tones; 1- and 2-syllable words
- C-V lag as diagnostic of tone





Tongue Dorsum: front ↓ back

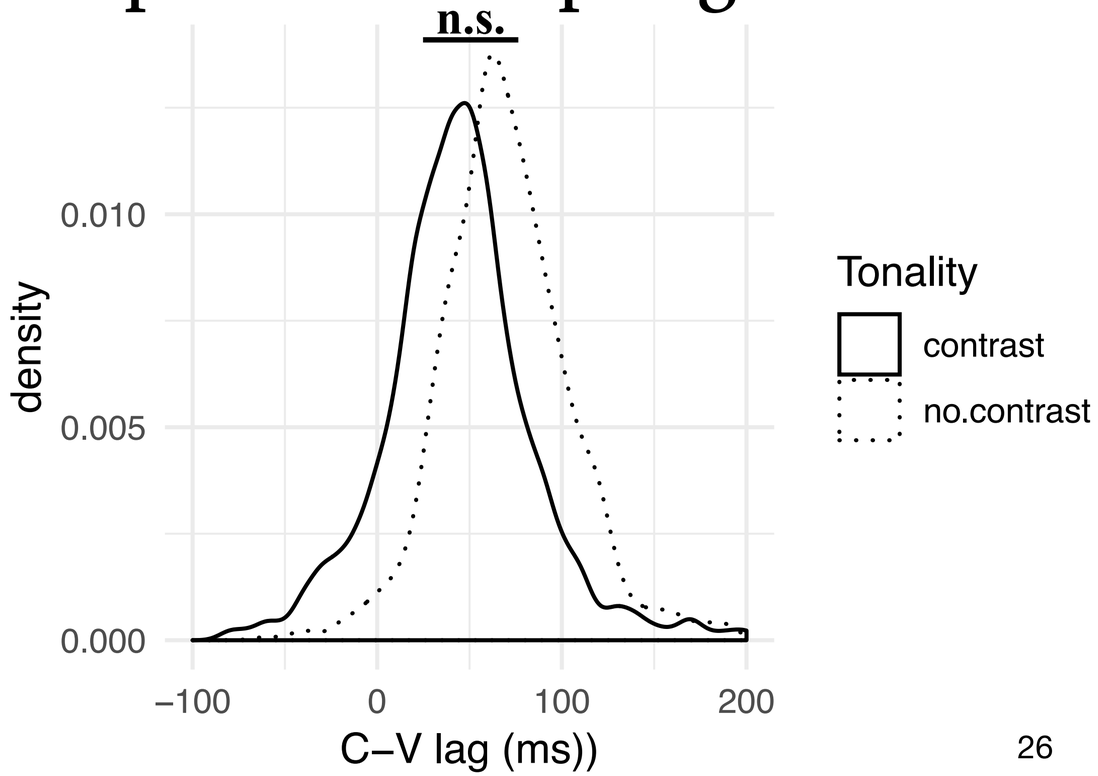
Lip Aperture: open ↓ closed

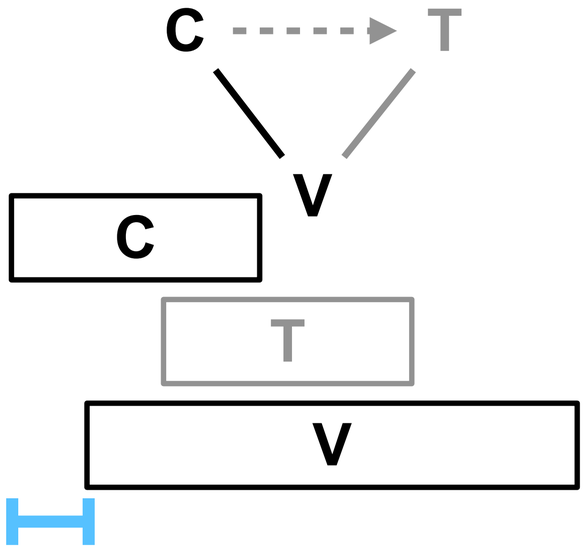


(Data: Zhang, Geissler, & Shaw 2019)
(Mview software: Tiede 2005)

Results: C-V lag in all speakers

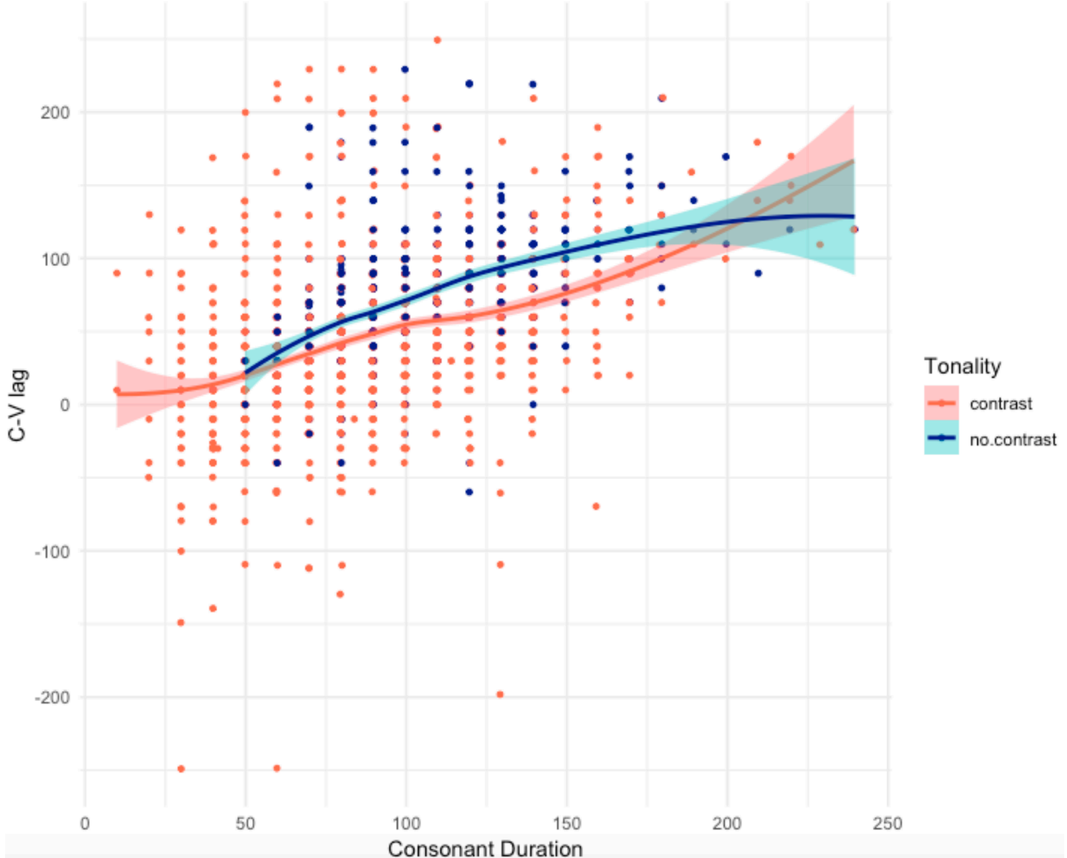
- There is a positive C-V lag… for speakers with ***and*** without the tone contrast (and in both tones)
- Competitive Coupling has no explanation for the 50ms lag

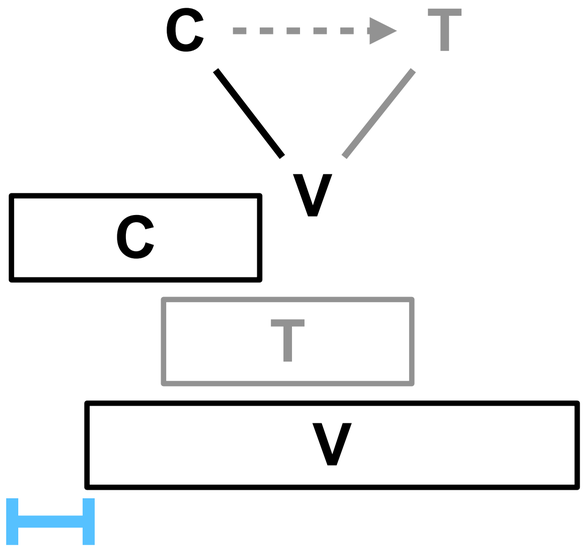




Results: C-V lag ~ C duration

- C-V lag increases with C duration—not just onset-onset timing
- Again—for both tonal and non-tonal speakers

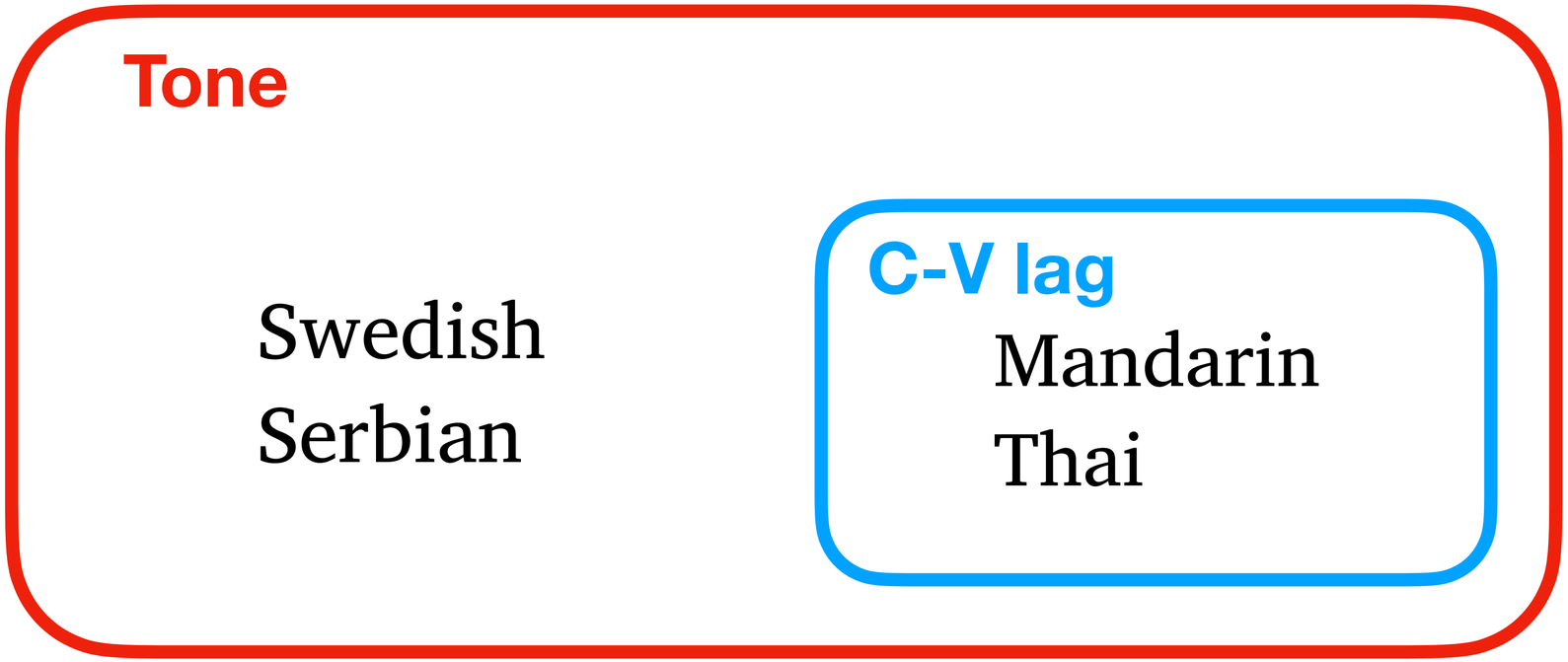




Cross-linguistic evidence (before)

No tone,
no C-V lag

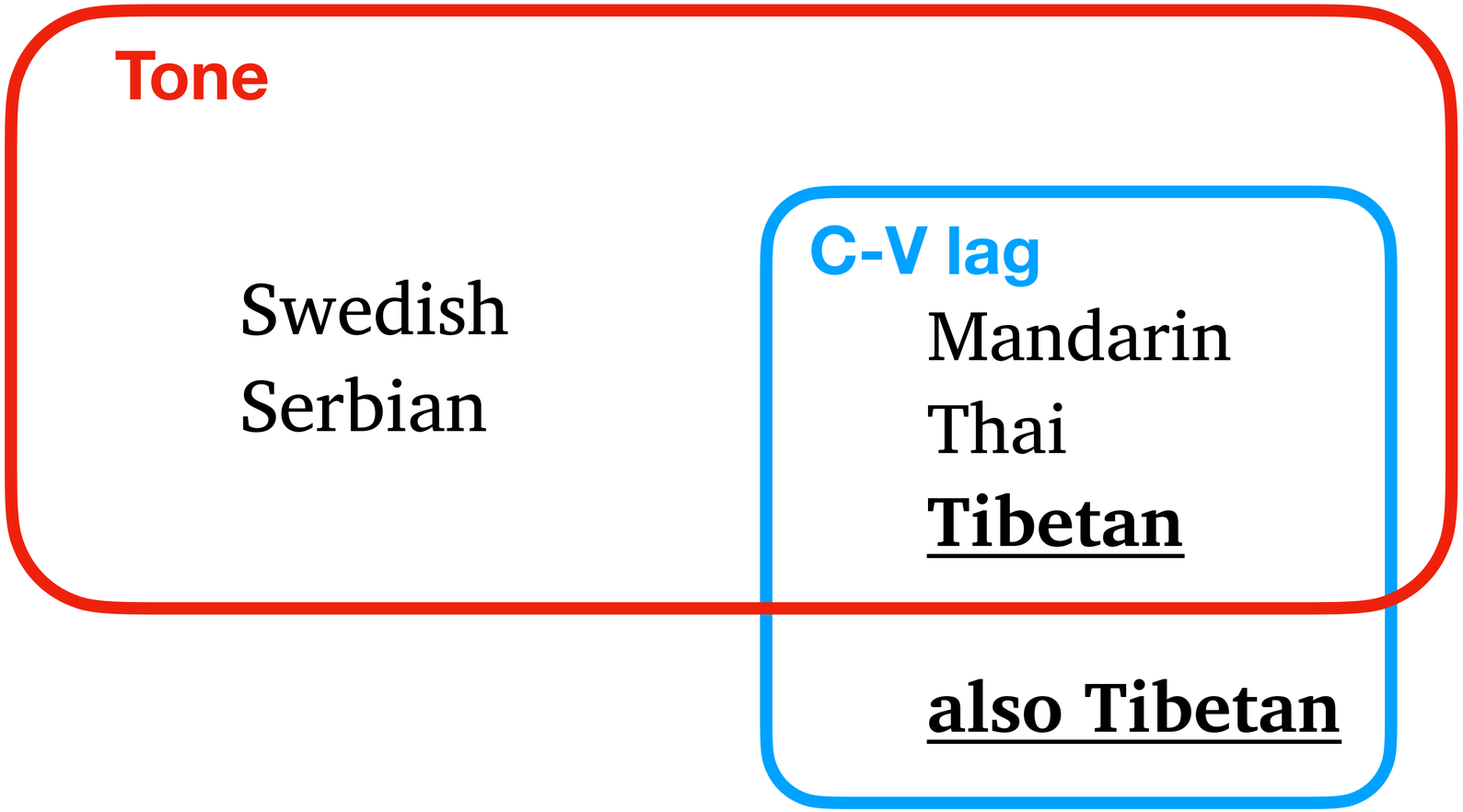
Arabic
Catalan
English
German
Georgian
Italian
Romanian



Cross-linguistic evidence (after)

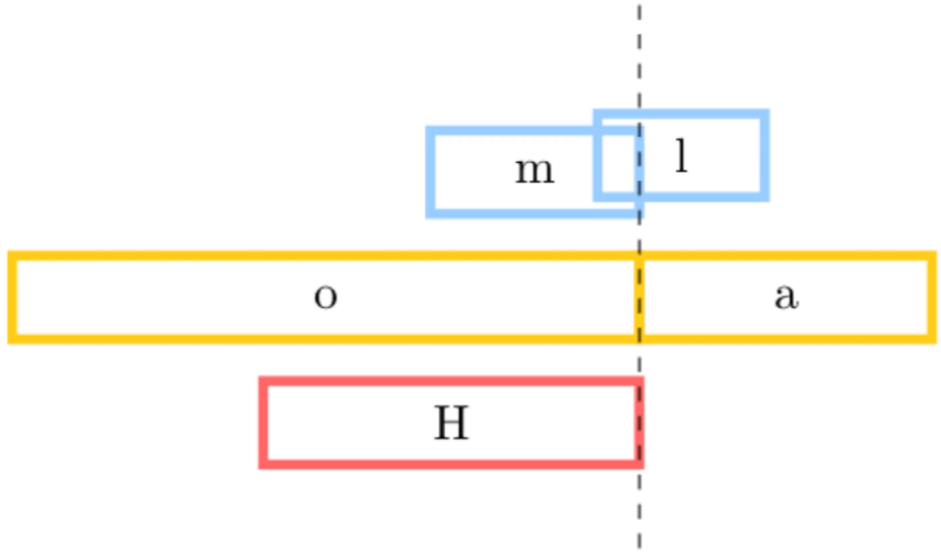
No tone,
no C-V lag

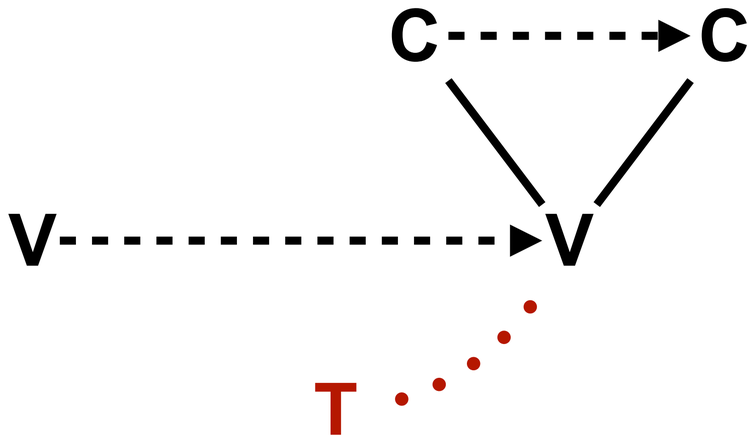
Arabic
Catalan
English
German
Georgian
Italian
Romanian



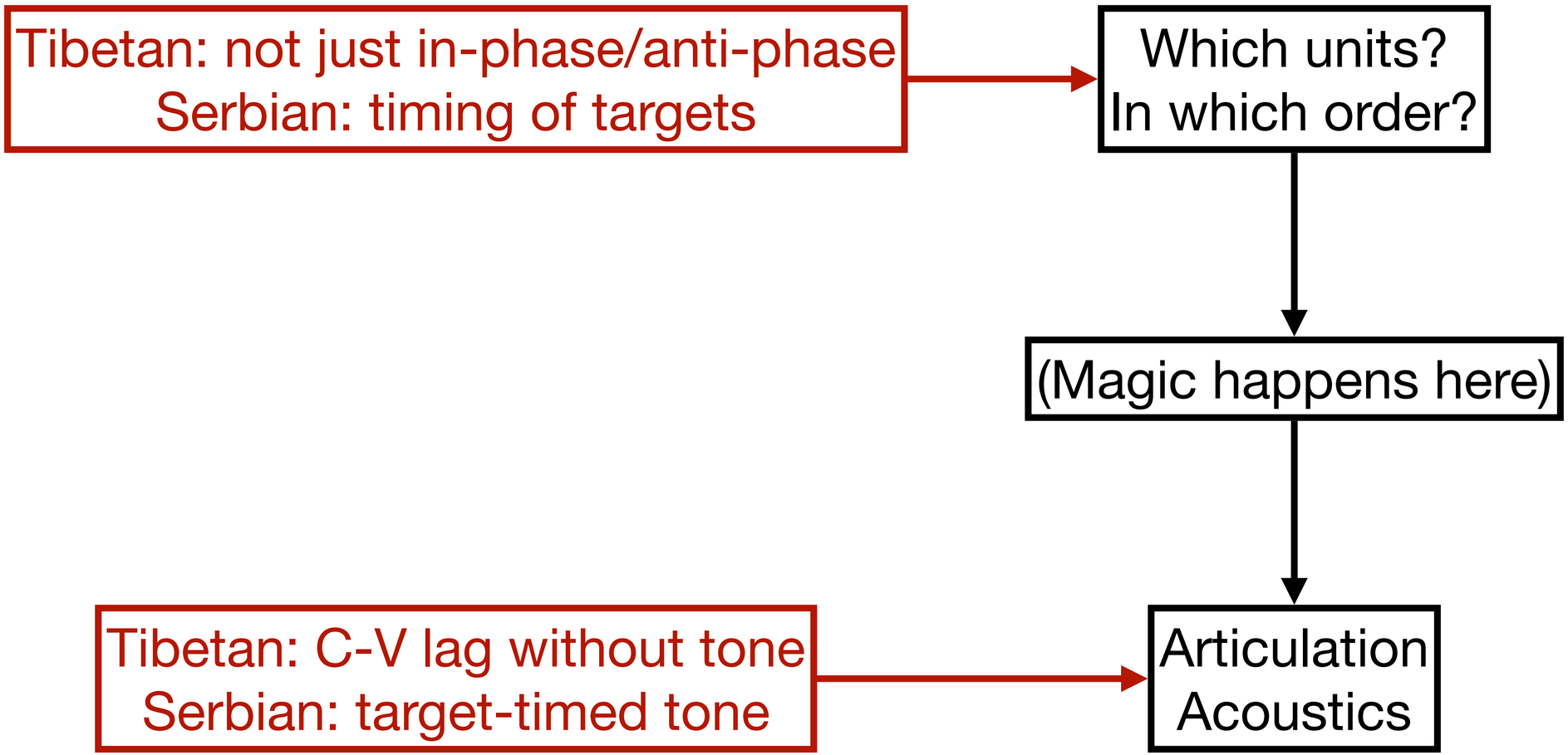
Karlin (2022)

- Coordination of tones in two BCS dialects: Belgrade and Valjevo Serbian
- Valjevo rising accent: [ő.mla]
  target of H timed to
  start of V2

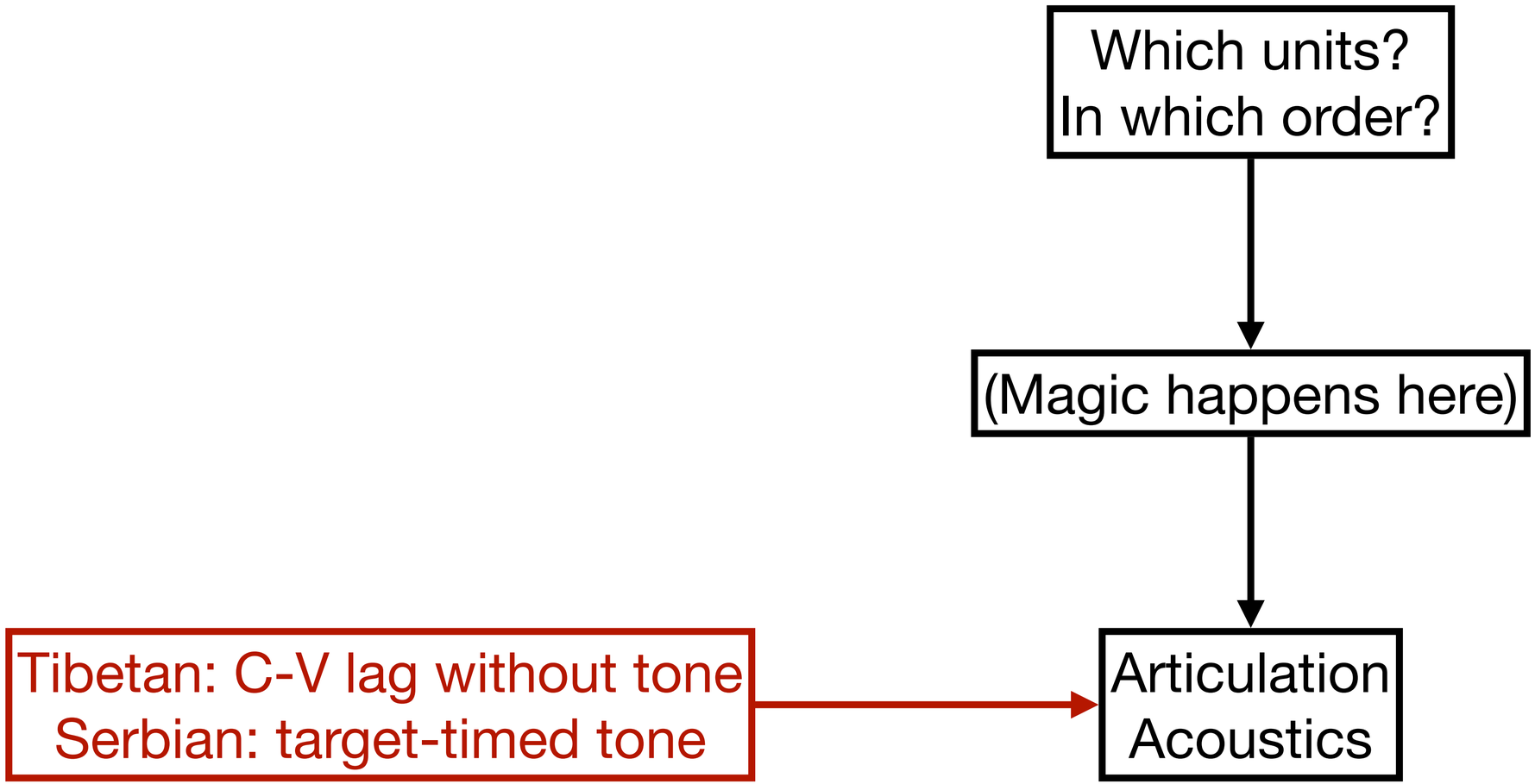




A Theory of the Interface



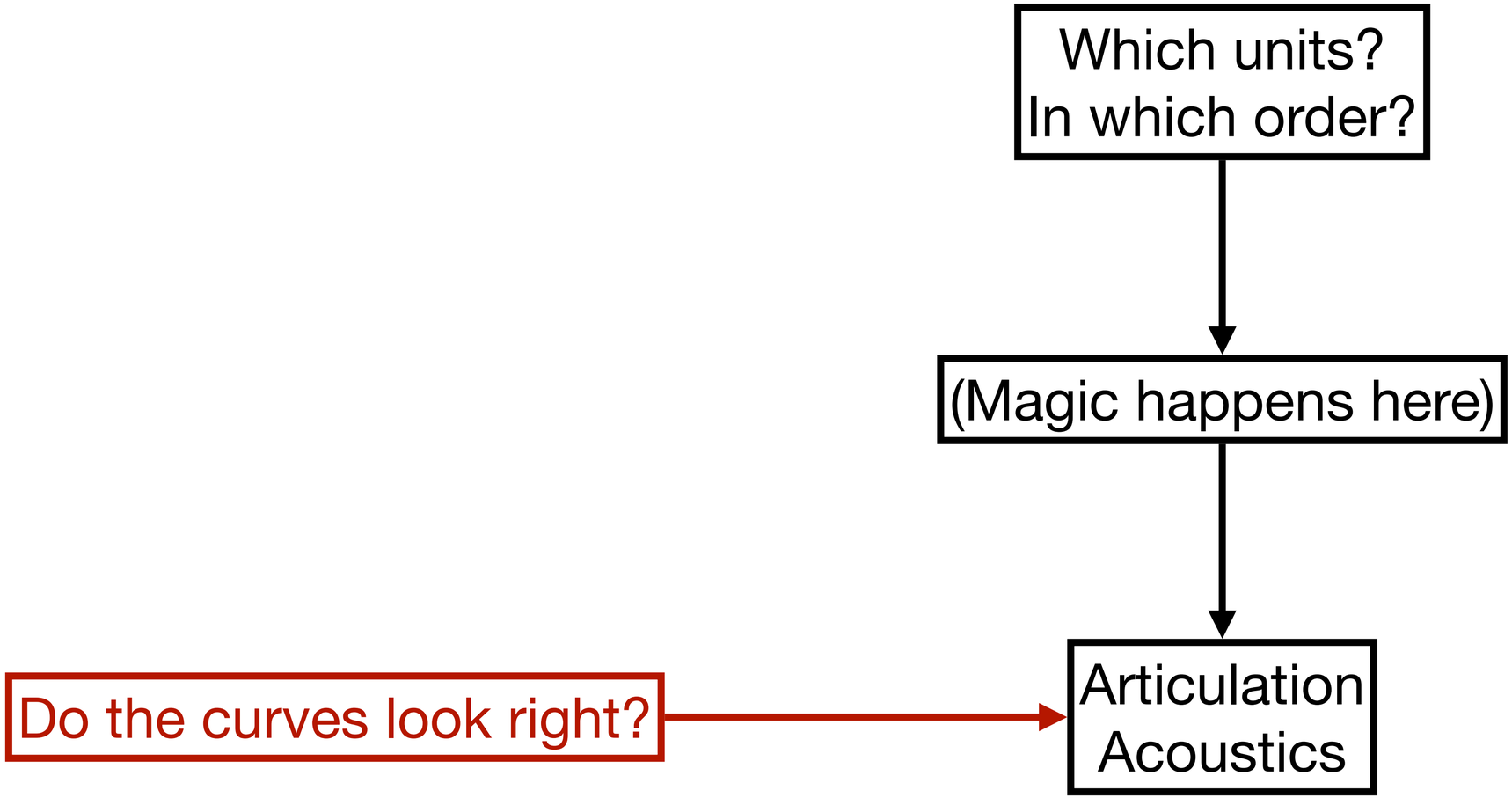
A Theory of the Interface



Roadmap

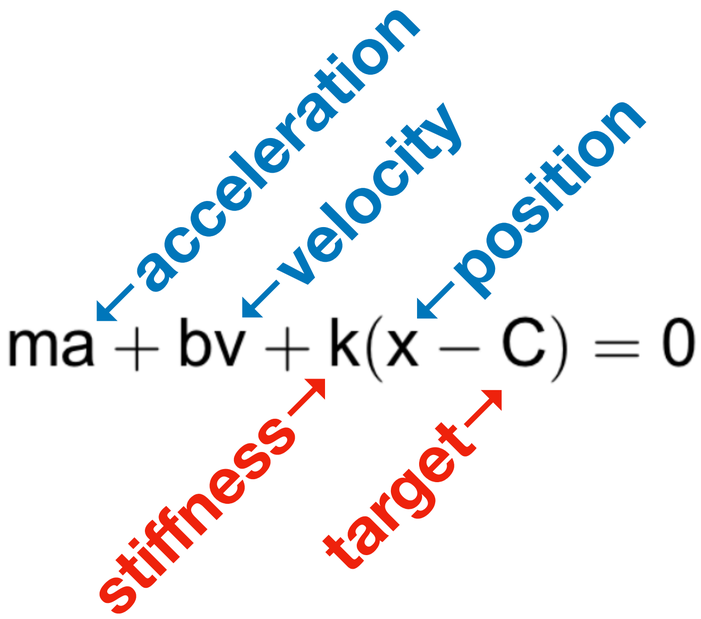
- Phonology and articulatory gestures
- Coordinating gestures: the Coupled Oscillator Model
- Problems
  - Tibetan tone study
- **Toward solutions: Analysis-by-synthesis**
- Conclusion

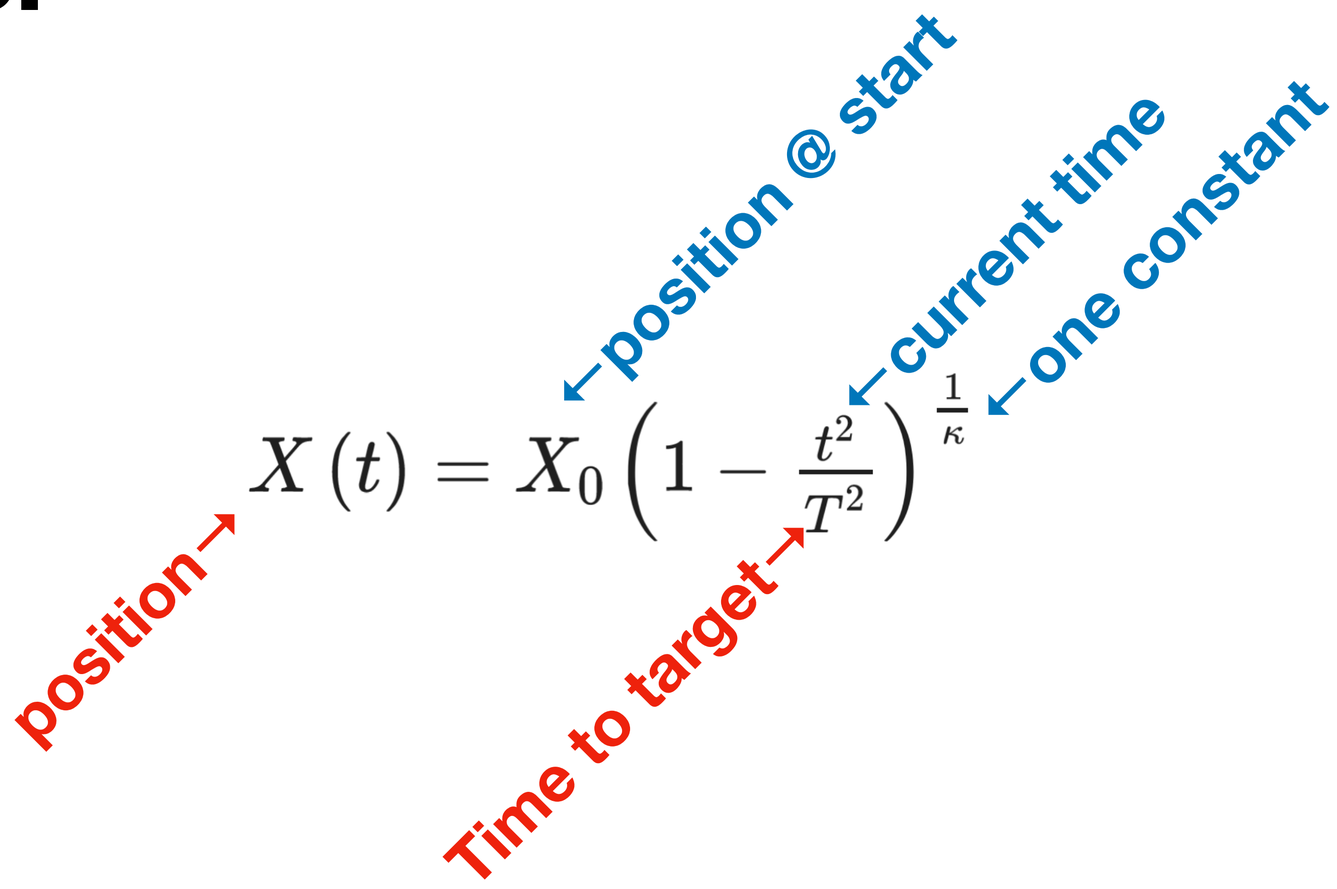
A Theory of the Interface



General Tau model

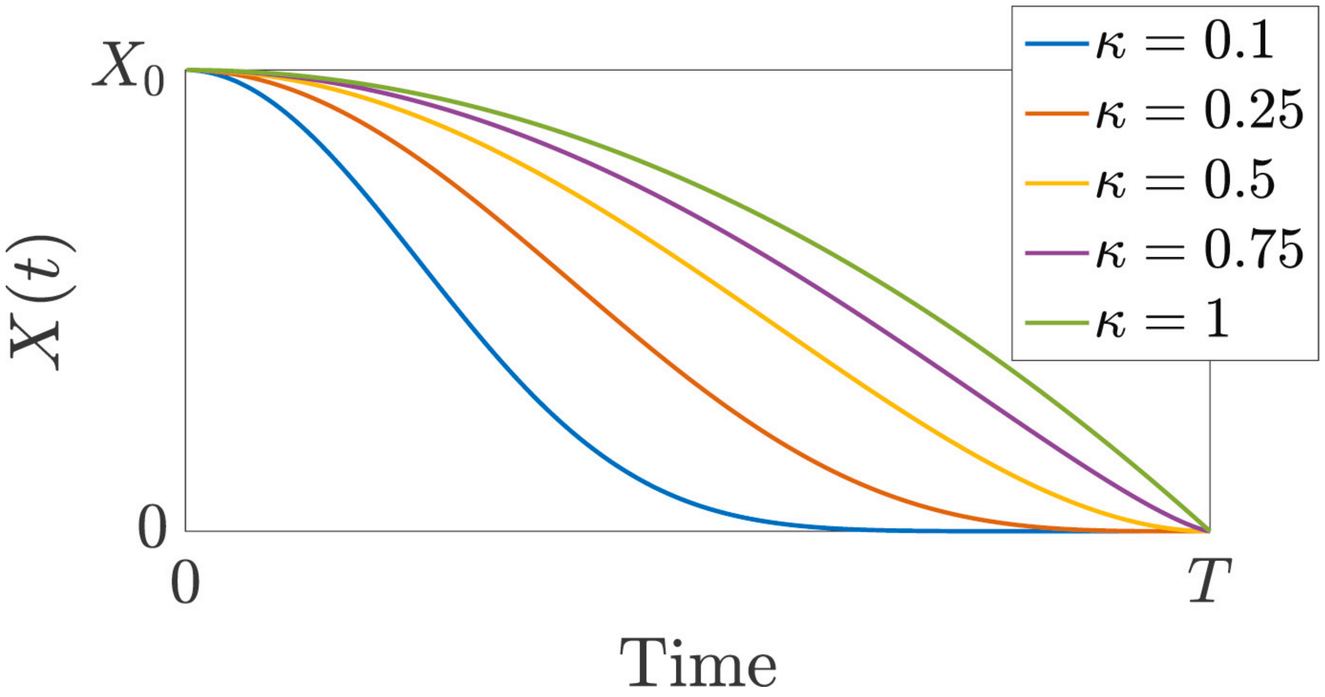
(Lee 1998, Elie et al. 2023)





General Tau model

(Lee 1998, Elie et al. 2023)



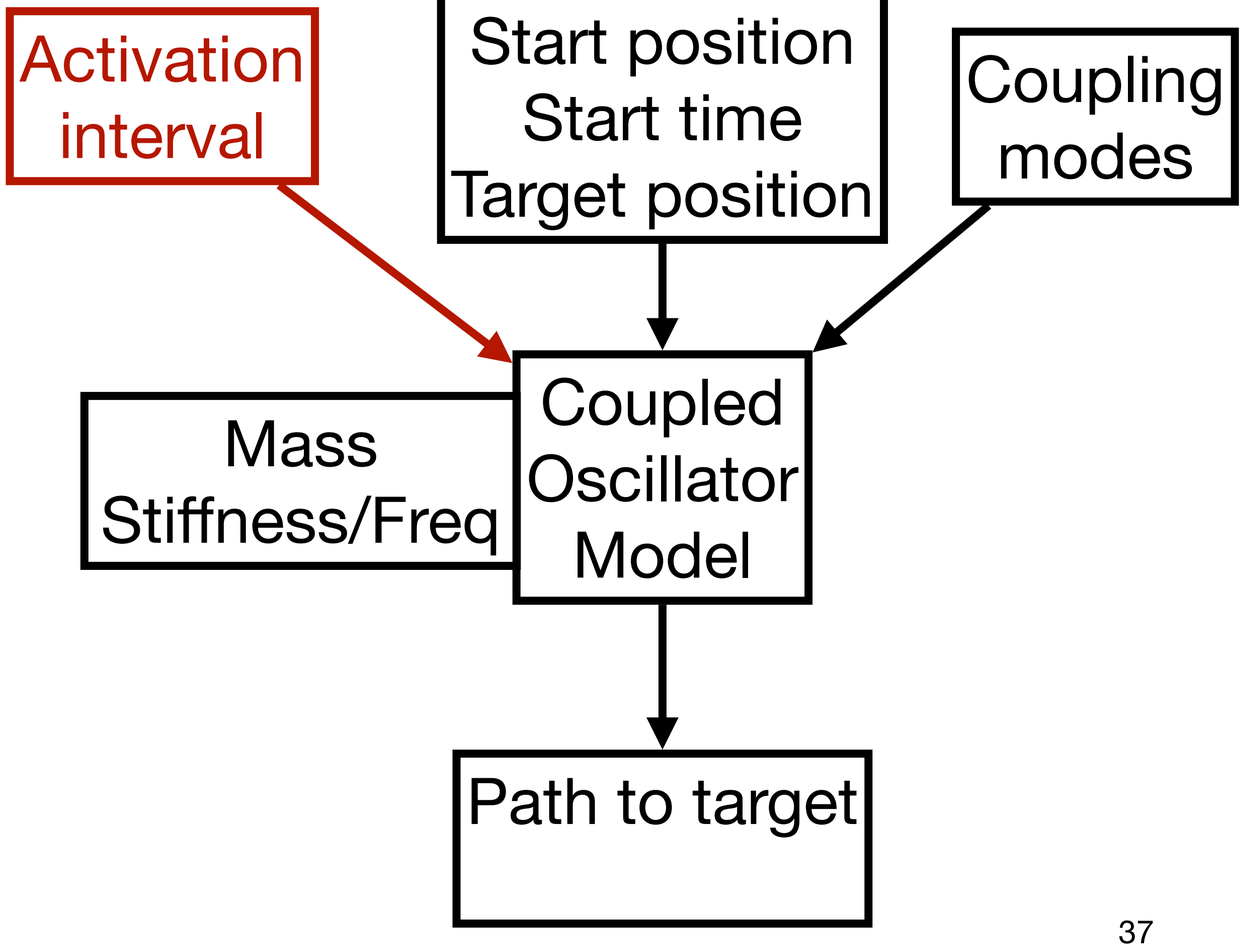
$$X(t) = X_0 \left(1 - \frac{t^2}{T^2}\right)^{\frac{1}{\kappa}}$$

← position @ start
← current time
← one constant
Time to target →

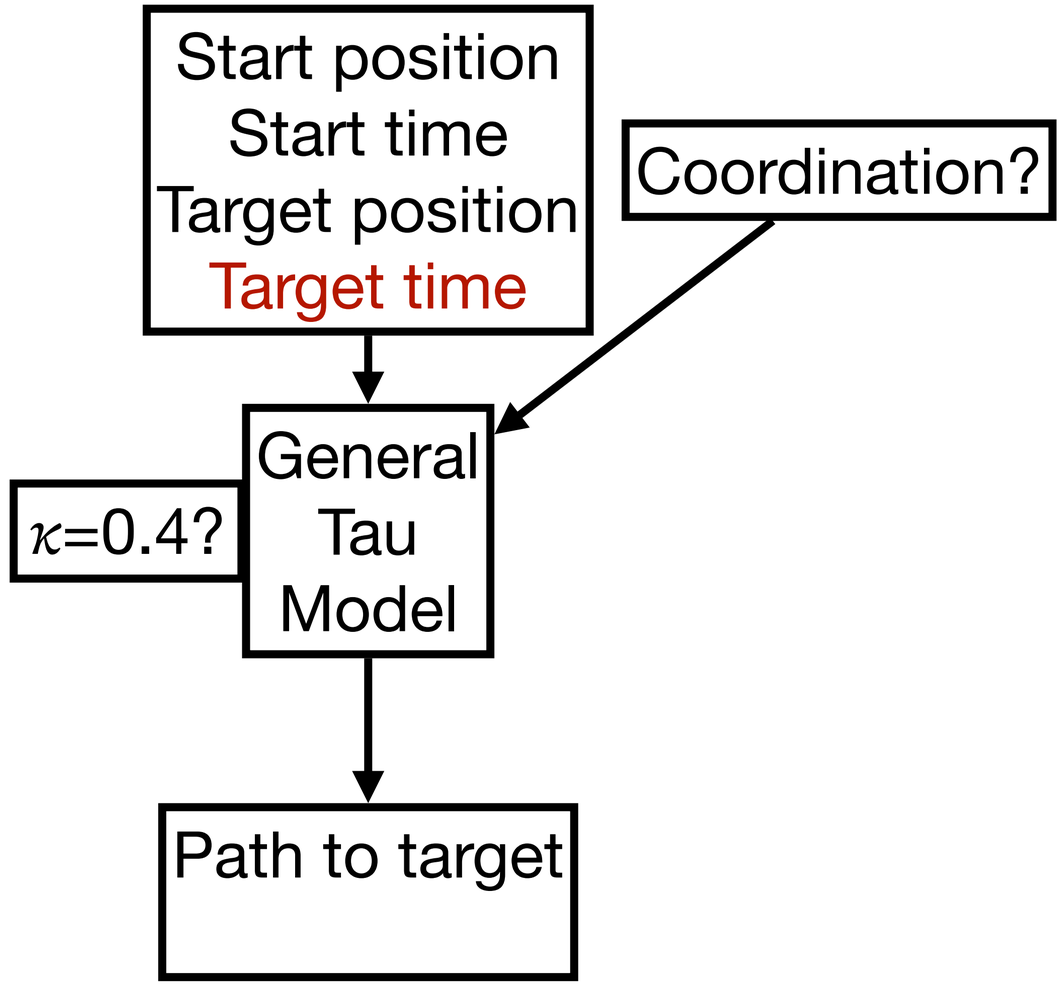
Symmetrical when κ = 0.4

Oscillator vs. Tau Models

$$ma + bv + k(x - C) = 0$$



$$X(t) = X_0\left(1 - \frac{t^2}{T^2}\right)^{\frac{1}{\kappa}}$$



There’s another problem

There’s another problem

WHEN DOES A GESTURE START

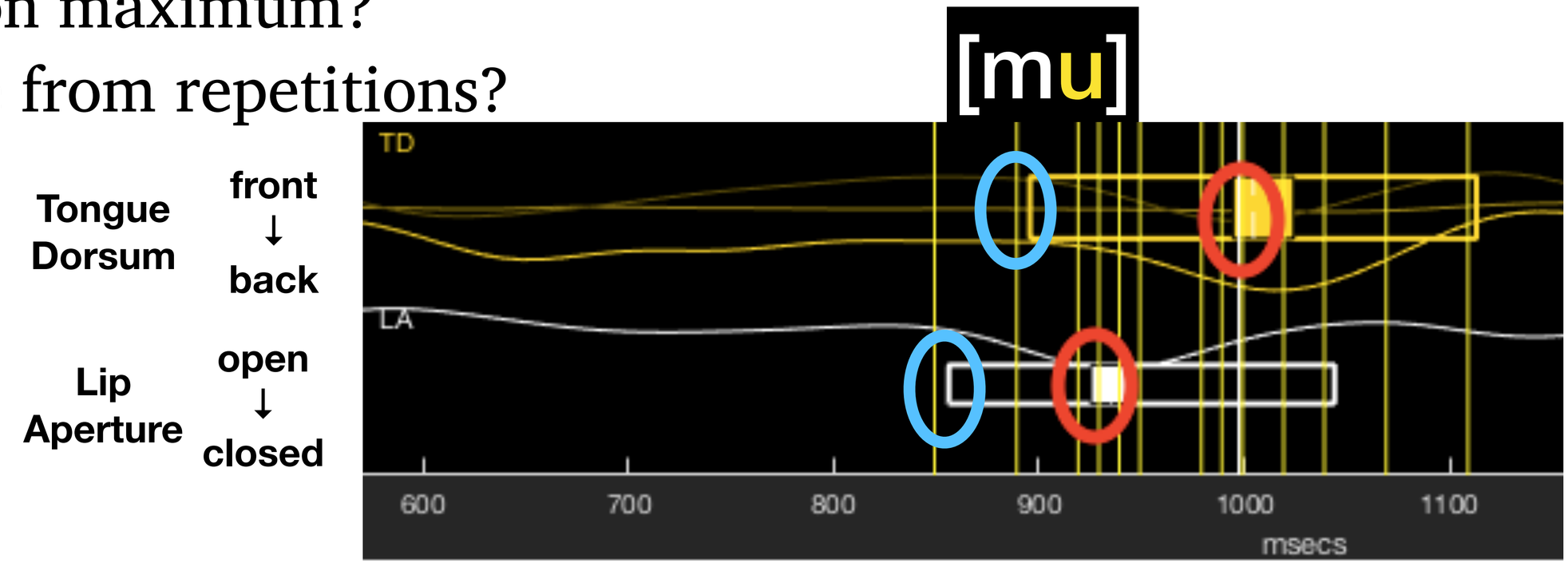
Velocity zero-crossing?

Velocity 20% of peak?

Acceleration maximum?

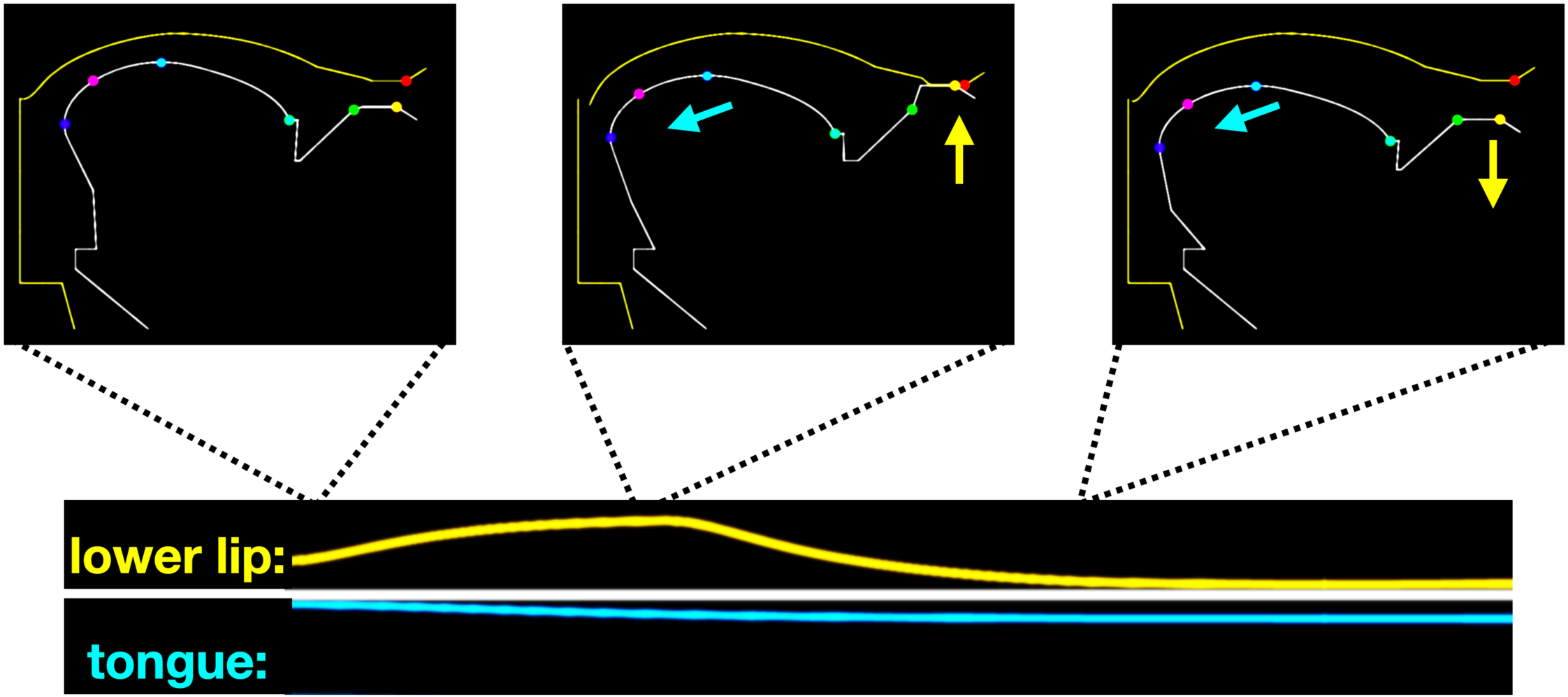
Divergence from repetitions?

…



Articulatory simulation

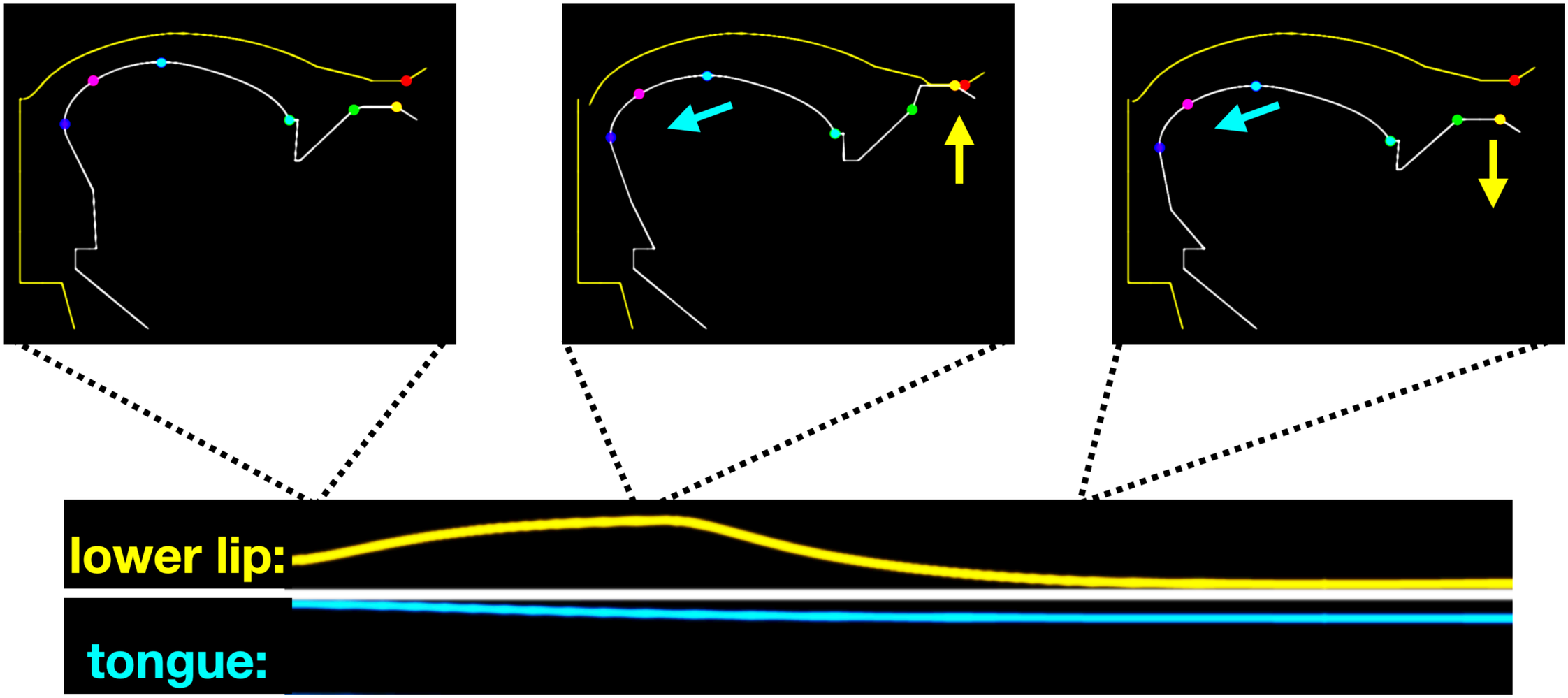
TADA: Task Dynamics Application *(Nam et al. 2004)*



Articulatory simulation

TADA: Task Dynamics Application *(Nam et al. 2004)*

Images from a different study sanity-checking the Tibetan experiment results *(Geissler 2022)*



Analysis-by-synthesis: <five>

- Diphthong targets can’t be separated with kinematic data
- Make a simulation, then tweak it, → 34,000 simulations
  Compare to 525 tokens from X-ray Microbeam Database

Bad fit **Good fit**

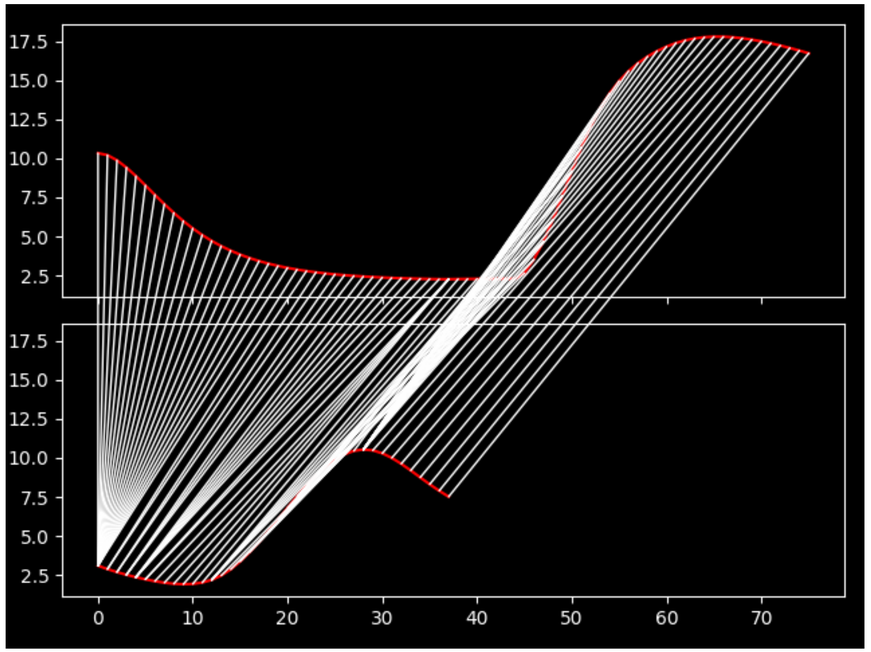
Simulated

Real

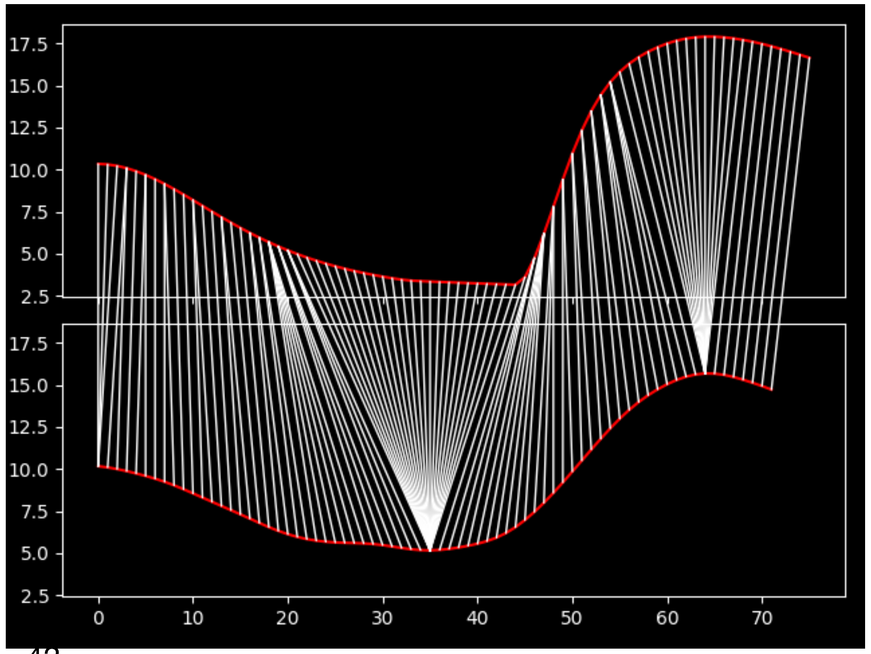
Analysis-by-synthesis: <five>

- Diphthong targets can’t be separated with kinematic data
- Make a simulation, then tweak it, → 34,000 simulations
  Compare to 525 tokens from X-ray Microbeam Database

Bad fit



Good fit



Simulated

Real

Interim findings

Analysis-by-Synthesis of <five>

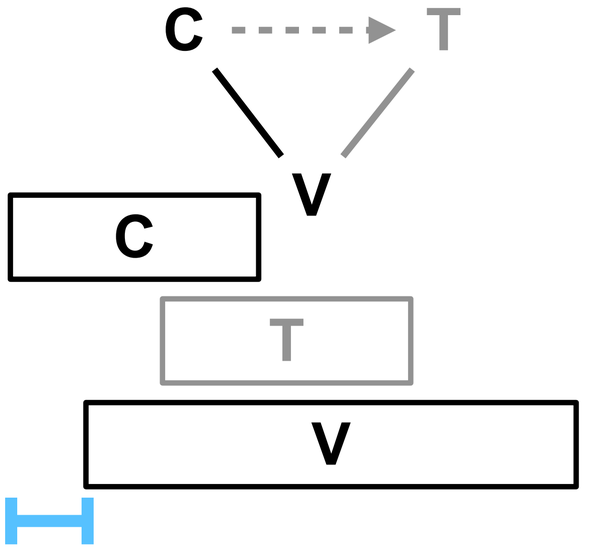
- We got some results!
  - [a] portion of dipthong timed to rest of word
  - [ɪ] portion more free to vary across tokens
- Still a lot to do
  - Extremely computationally-intensive
  - Which dimensions of variation? How much to vary?
  - What's the best way to compare curves?

Roadmap

- Phonology and articulatory gestures
- Coordinating gestures: the Coupled Oscillator Model
- Problems
  - Tibetan tone study
- Toward solutions: Analysis-by-synthesis
- **Conclusion**

What have we learned

- Tibetan: some speakers have tone, others don't
  BUT they all have the same C-V lag
  → problem for Competitive Coupling of Tone
- Studying coordination requires consistent, reliable, practical ways to identify gestures
  → Analysis-by-Synthesis might help



Theory ↔ Data

- Observation: Gestures!
  - Theory: Oscillators!
- Observation: overlap in clusters
  - Theory: Coupled Oscillators!
    What else can this do? Tone!
- Observations: or can it?
  - Theory: …

Theory ↔ Data

- Observation: Gestures!
  - Theory: Oscillators!
- Observation: overlap in clusters
  - Theory: Coupled Oscillators! What else can this do? Tone!
- Observations: or can it?
  - Theory: …

For now:

- Gather new observations & reevaluate old ones
  → descriptive generalizations
- Gather pieces that might help us with the next iteration
  → tools (e.g. simulators)
  → insights from other fields

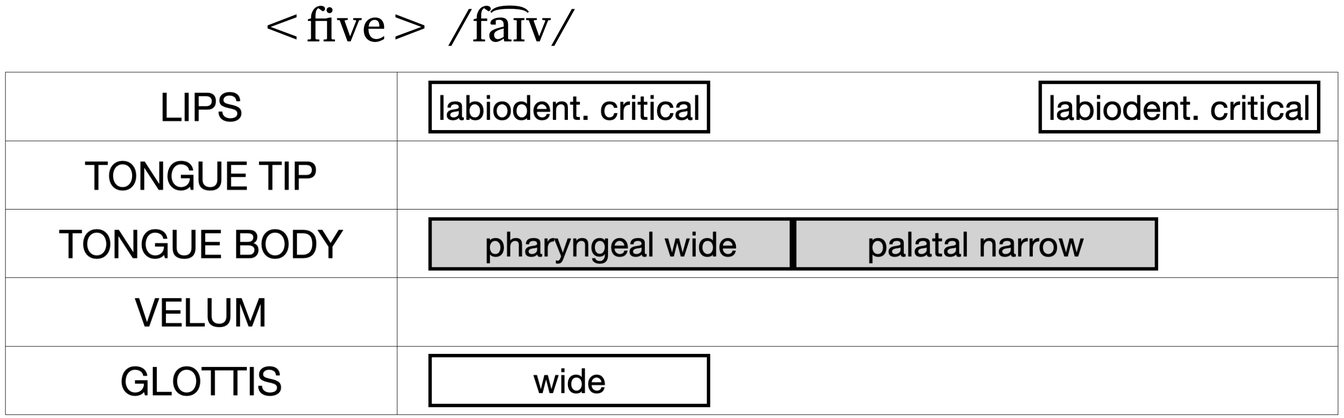
ཐུགས་རྗེ་ཆེ།

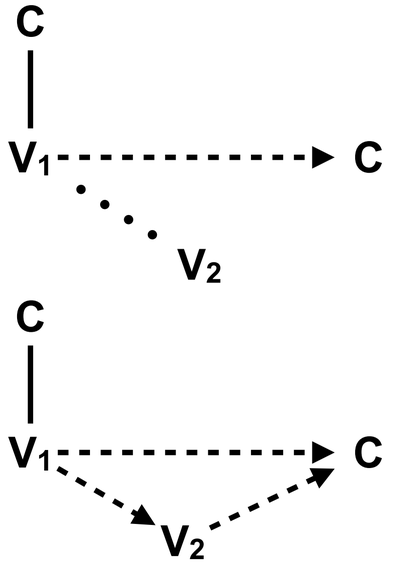
Thank you!

Pocket slides

What about diphthongs?

- Can approximately describe with in-phase/anti-phase
- How do diphthongs change when they get shorter?





(Goldstein et al. 2000)

Articulatory study

Geissler et al. (2021), Geissler (2021ch4)

- H1: variation in timing conditioned by presence/absence of lexical tone
  - speakers with tone contrast will have competitive coupling (pos. C-V lag)
  - speakers without tone contrast will have in-phase C-V timing (no C-V lag)
- H2: timing convergence:
  - all speakers will have similar coordination patterns despite interspeaker variation in presence/absence of tone
- What kind of tone contrast is there?
  - If H-∅, then difference will be visible in high vs. low tone words
  - If H-L, then no difference in timing by tone.

EMA Study conclusions

- H1: variation in timing conditioned by presence/absence of lexical tone
  - speakers with tone contrast will have competitive coupling (pos. C-V lag)
  - speakers without tone contrast will have in-phase C-V timing (no C-V lag)
- ✔ **H2: timing convergence:**
  - all speakers have similar coordination patterns despite interspeaker variation in presence/absence of tone
- What kind of tone contrast is there?
  - If H-∅, then difference will be visible in high vs. low tone words
  - ✔ If **H-L**, then no difference in timing by tone.

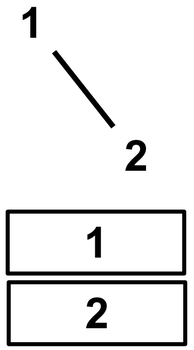
The temporal basis of complex segments

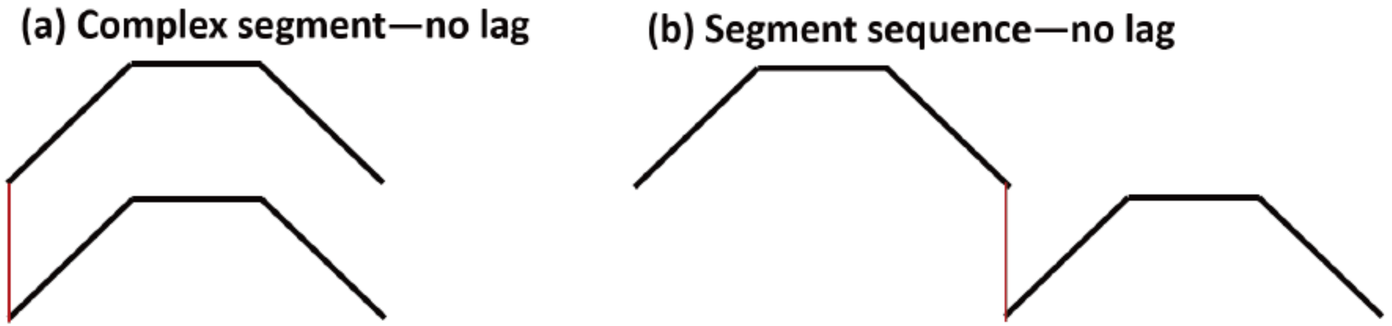
Shaw et al. 2019

The temporal basis of complex segments

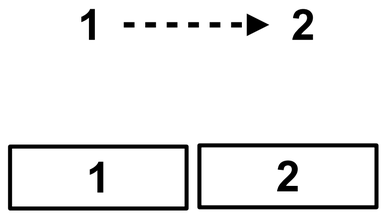
Shaw (2019): predictions

In-phase

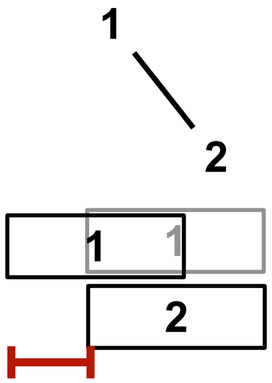


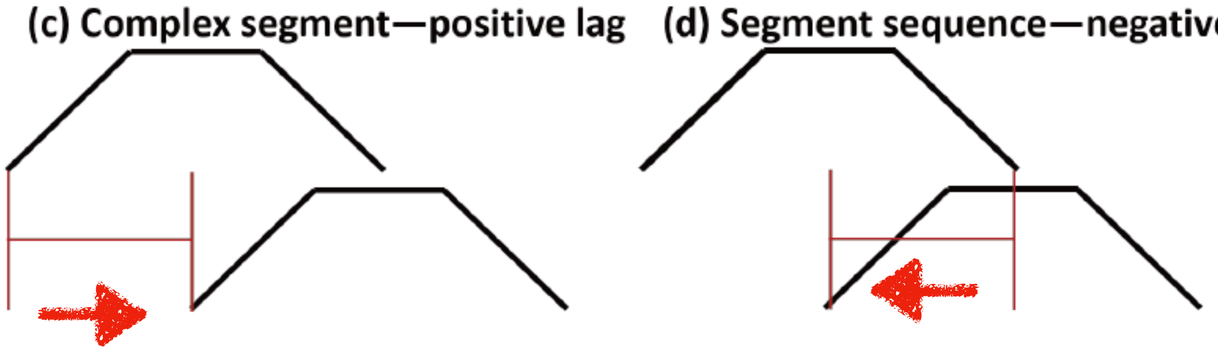


Anti-Phase



In-phase + lag (offset)





Anti-Phase - lag (offset)

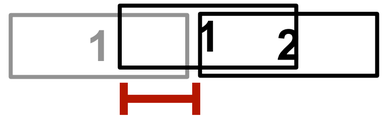


Figure 1: Hypothesized gestural coordination patterns for complex segments (a), (c) and segment sequences (b), (d)

The temporal basis of complex segments

Shaw (2019): results

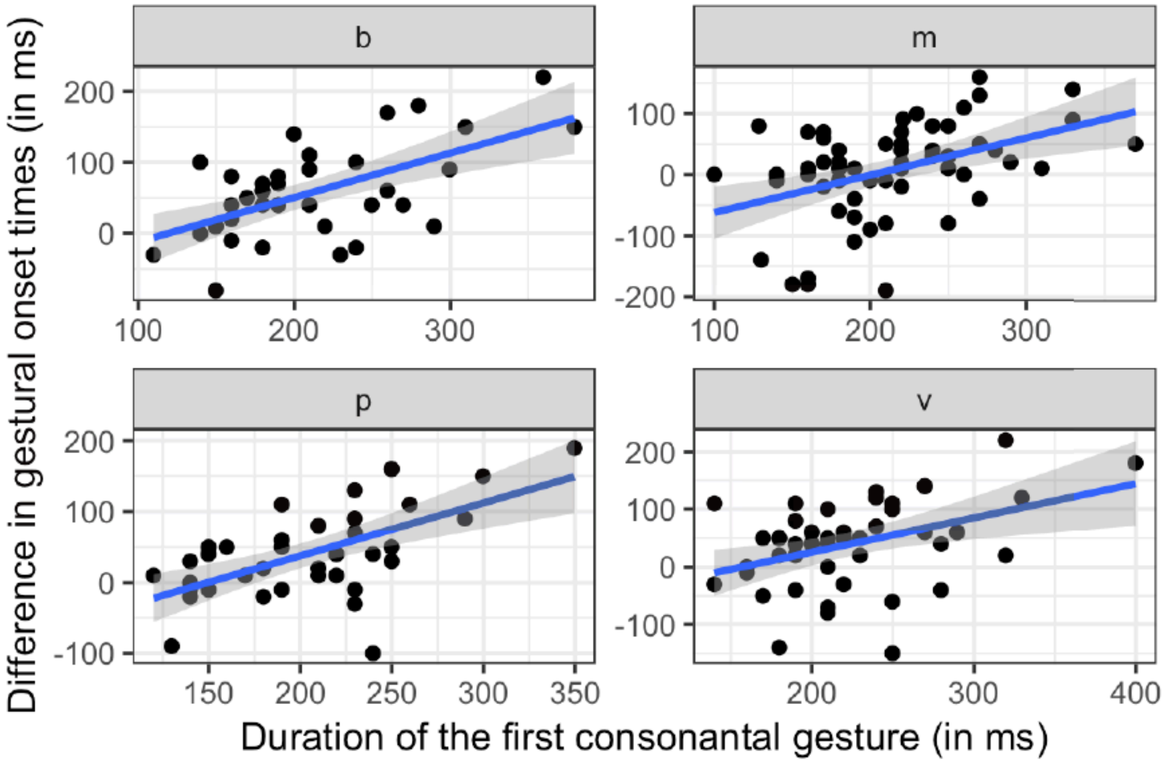


Figure 4: Correlations for the data from the English experiment

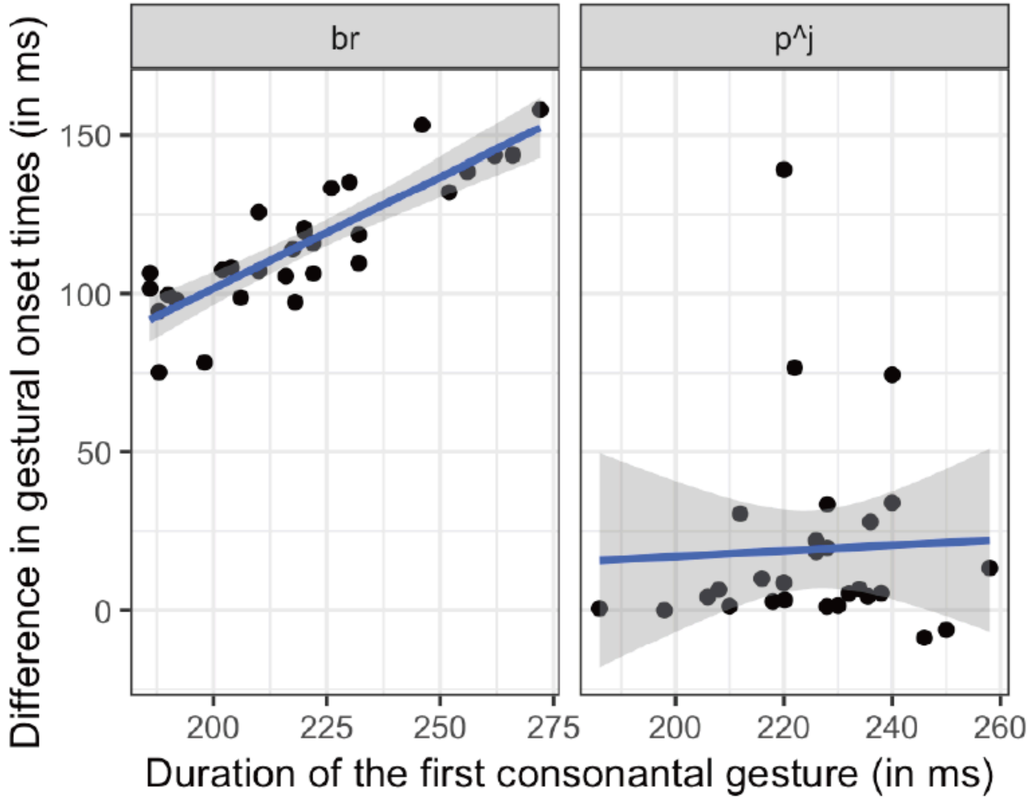


Figure 2: Correlations for the Russian data

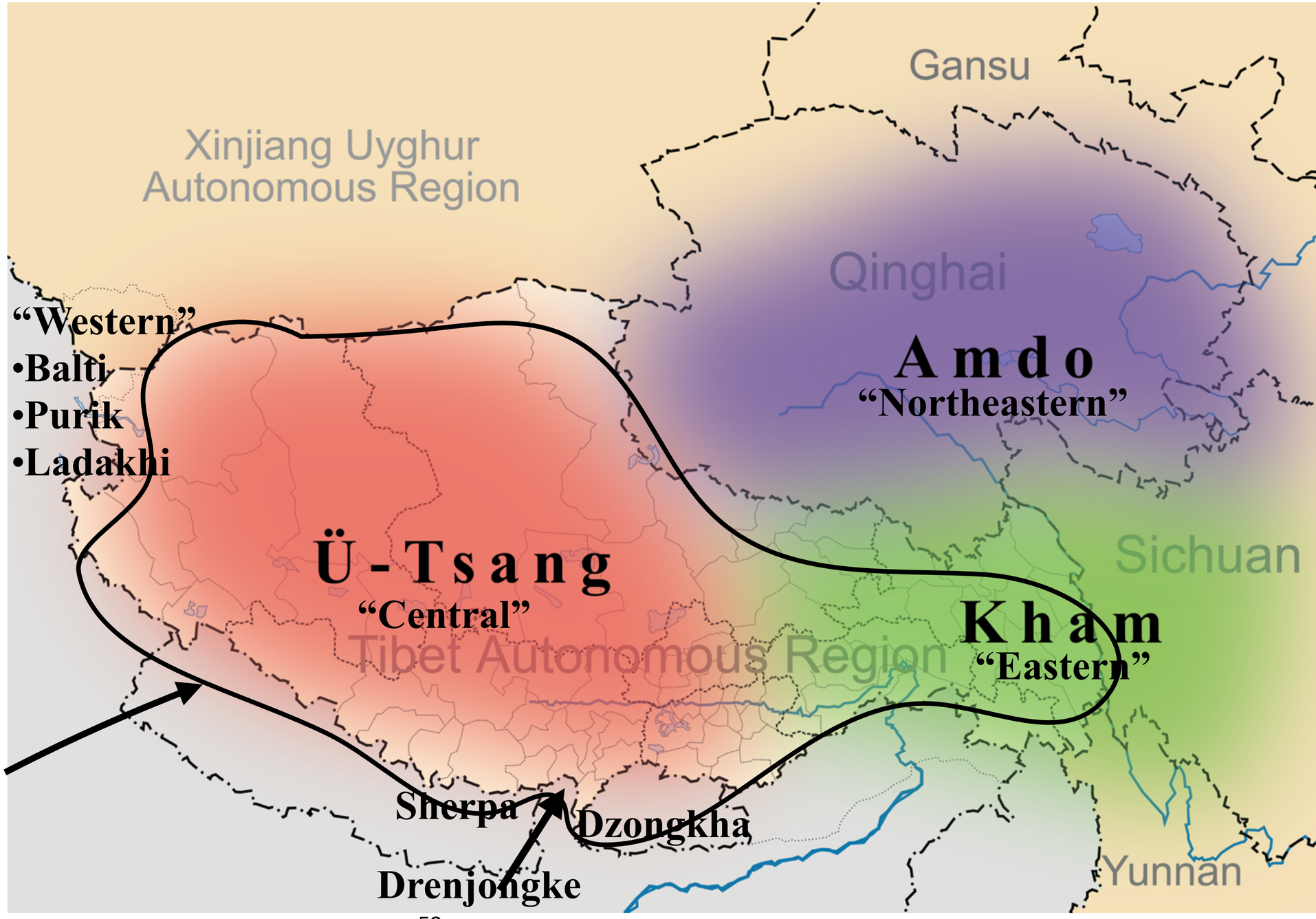
Tibetan dialects

Tibetan

བོད་སྐད་

- “archaic”/ “cluster”
- “innovative”/ “non-cluster”
- dialect continuum
- post-1959 diaspora

Approx. extent of tone



Dialects: Natural laboratory

- tonogenesis
- laryngeal variation
- cluster simplification
- vowel shifts, spirantization, retroflexion, palatalization
- evidential, honorifics, modality, etc.

Written (Classical) Tibetan	Balti (Western)	Rebkong (Northeastern)	Tokpe Gola (Central)	Gloss
khrag	[kʂʌk]	[t͡ɕɤɣ]	[t^{h}ʌ́k] ([t^{h}ák])	‘blood’
rtswa	[xst͡soa]	[xt͡sa]	[t͡sá]	‘grass’
spyang ki	[spjaŋ.ˈku]	[xt͡ɕaŋ.ˈk^{h}ɤ]	[t͡ʃáŋ.gú]	‘wolf’
bcu bdun	[t͡ɕub.ˈdʊn]	[t͡ɕɤb.ˈdɤn]	[t͡ʃúp.tő] ([t͡ʃúp.tỹ])	‘seventeen’

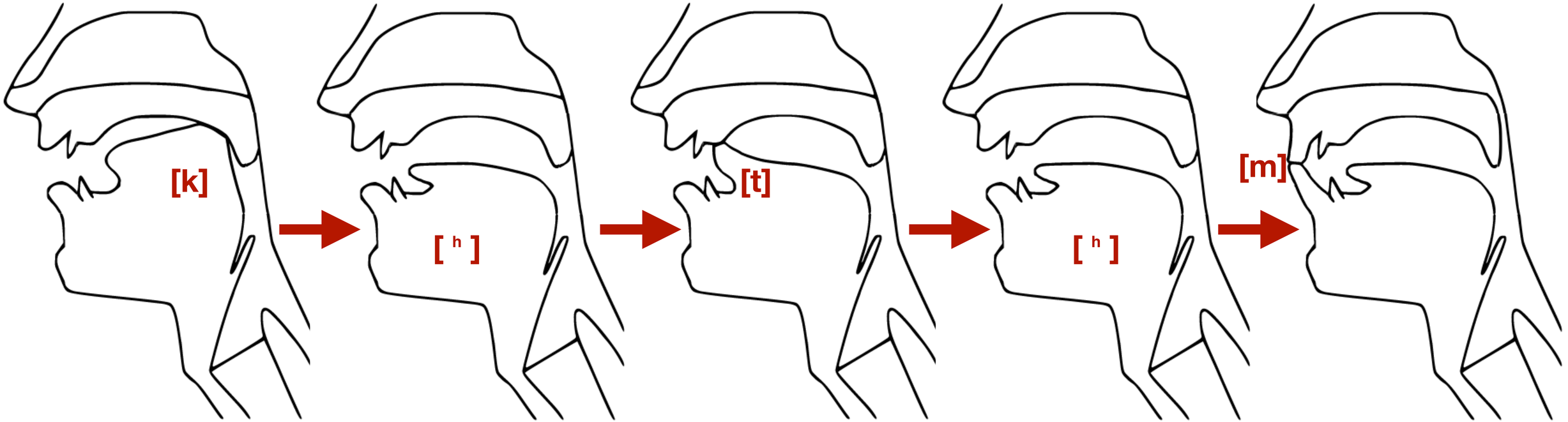
(Adapted from Caplow 2013)

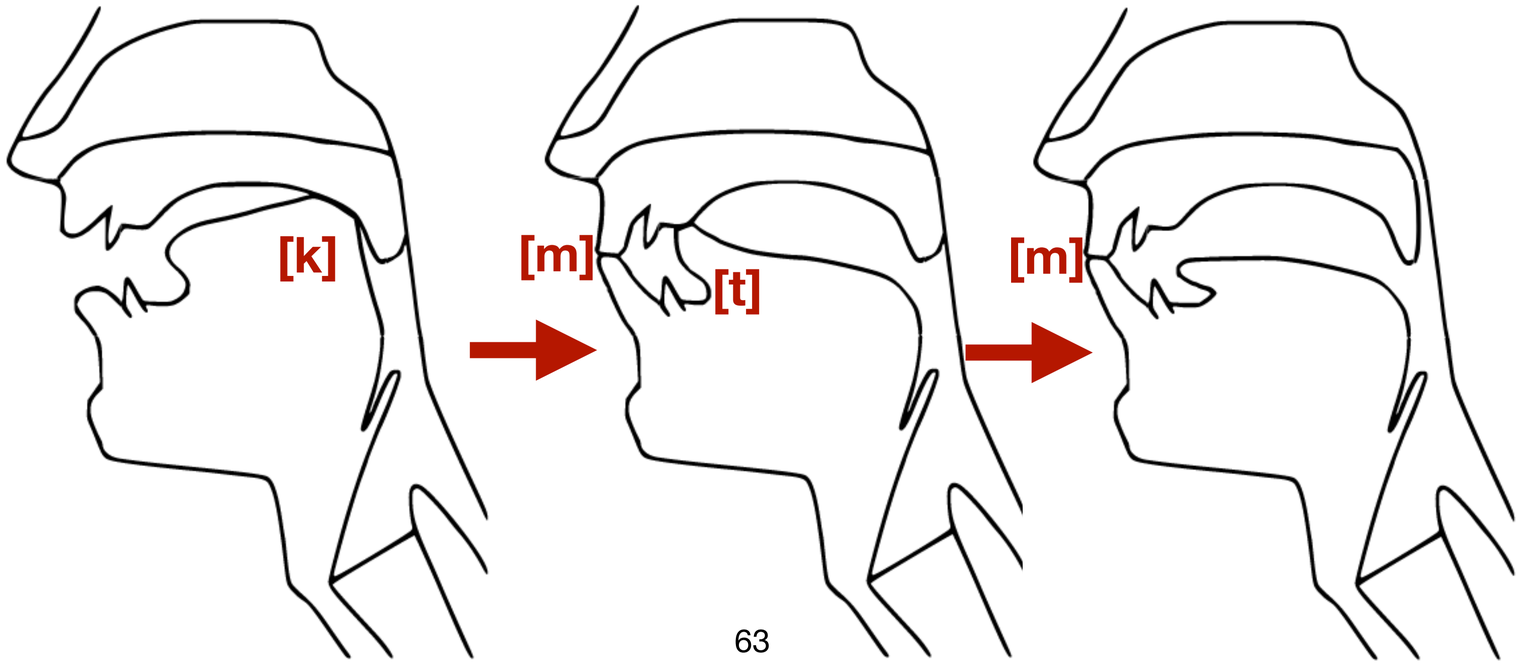
Tonogenesis
(tonal dialects only)

- Voiceless onsets > high tone
  - *p^{h}ar ‘over there’ > H
    *sa ‘earth’ > H
- Voiced onsets > low tone
  - *bar ‘between’ > L
    *za ‘eat’ > L
    *mar ‘butter’ > L
- Sonorants with pre-initial > high tone
  - *sman ‘medicine’ > H

Laryngeal contrasts

	Etymological onsets							Innovative features
Orthography	སྤ་	ཕ་	བ་	སྦ་	ས་	ཟ་	བཟ་	
Old Tibetan	sᵊpa	pʰa	ba	sᵊba	sa	za	bᵊza	aspiration allphonic
Northeastern and Western dialects	spa	pʰa	ba ~ wa	ʁba	sa	za	za	cluster simplification aspirated/unaspirated contrast
Eastern dialects	pá	pʰá	pà	bà	sá	zà	zà	tonogenesis cluster simplification
Central dialects (Lhasa)	pá	pʰá	pʰà	pà	sá	sà	sà	voiced clusters > voiceless voiced simplex > aspirated



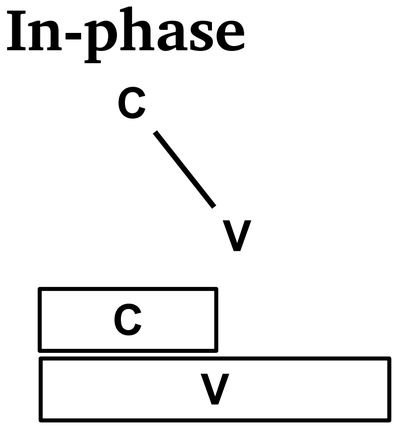


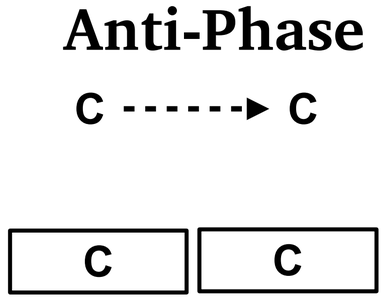
back to slide

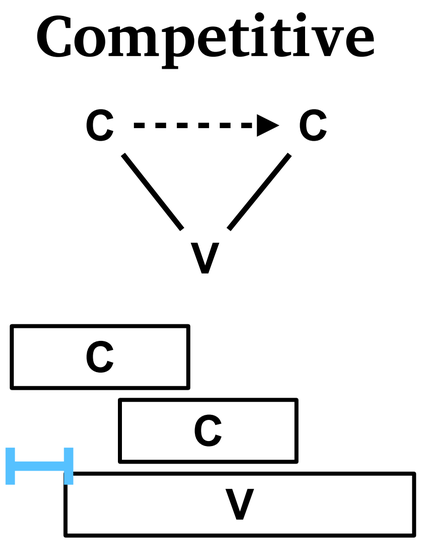
Interactive Sagittal Section

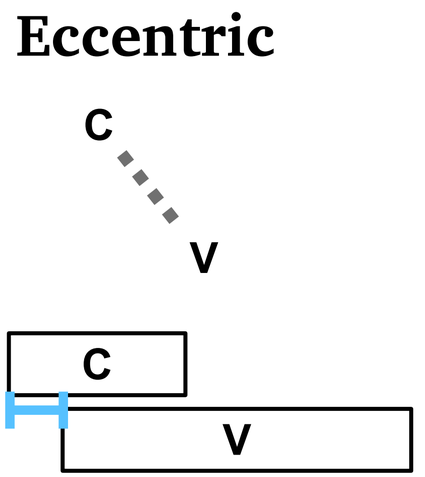
Coordinating gestures in time

- Gestural coupling modes:
  - *In-phase coupling:* (synchronous) and *Anti-phase coupling* (sequential) are most stable
  - *Competitive coupling:* combination of in-phase and anti-phase coupling relations
  - *Eccentric coupling:* one coupling relation, just not intrinsically stable







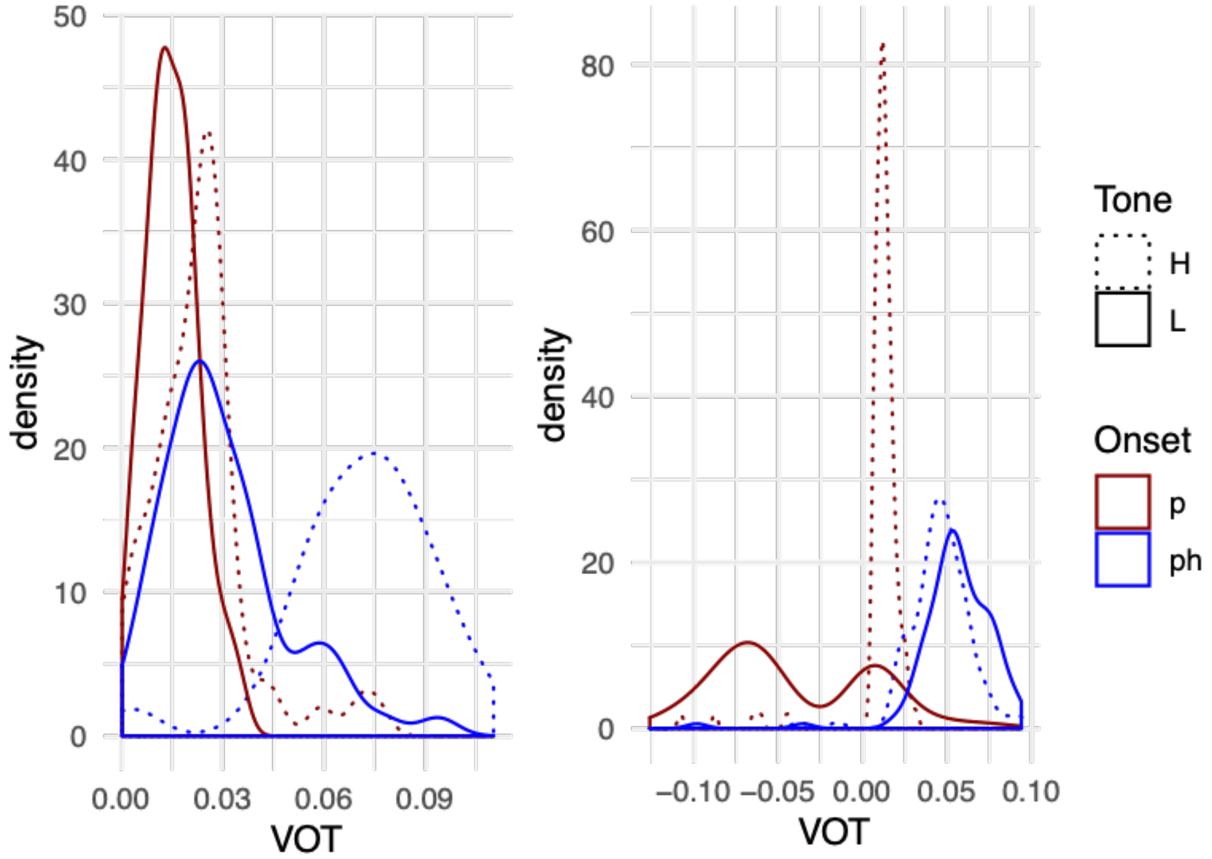


(Nam & Saltzman 2003, Nam et al. 2009, Goldstein 2011)

Two systems of laryngeal contrasts

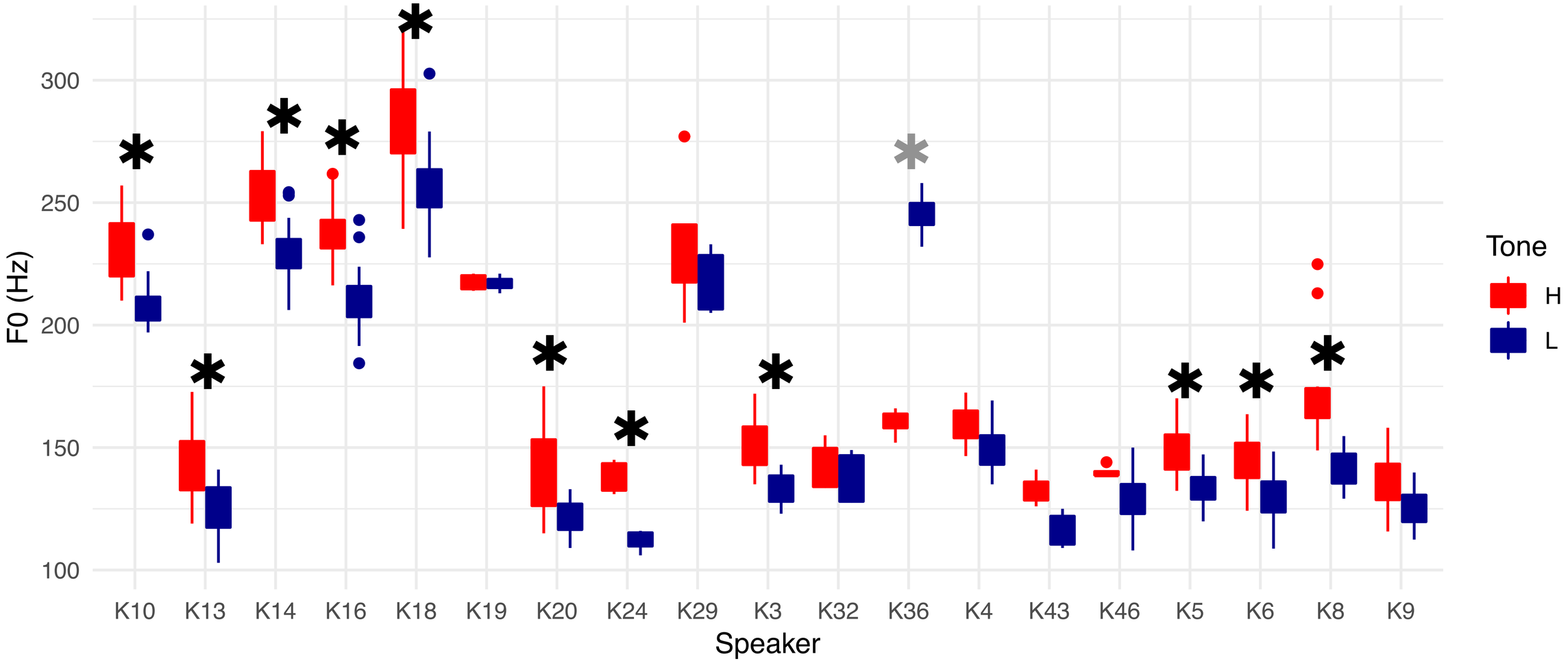
even in speakers with no F0 contrast (!!!)

- Both conditioned by etymological tone category:
- Left speaker
  - no prevoicing
  - long VOT only with H tone
- Right speaker:
  - prevoicing with L tone
  - long VOT with both tones

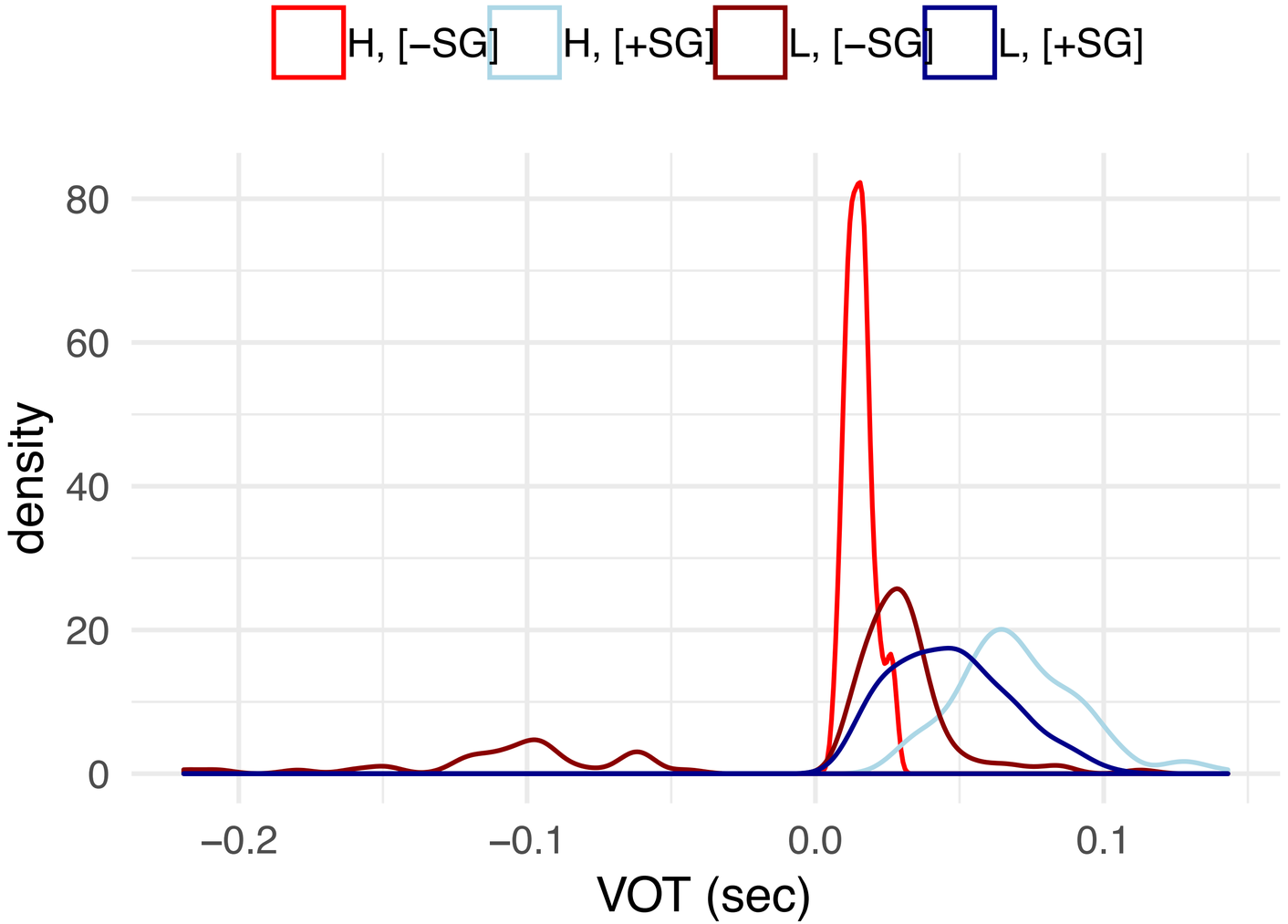


Does H have higher pitch than L?

Yes for 11/19, no for 7/19



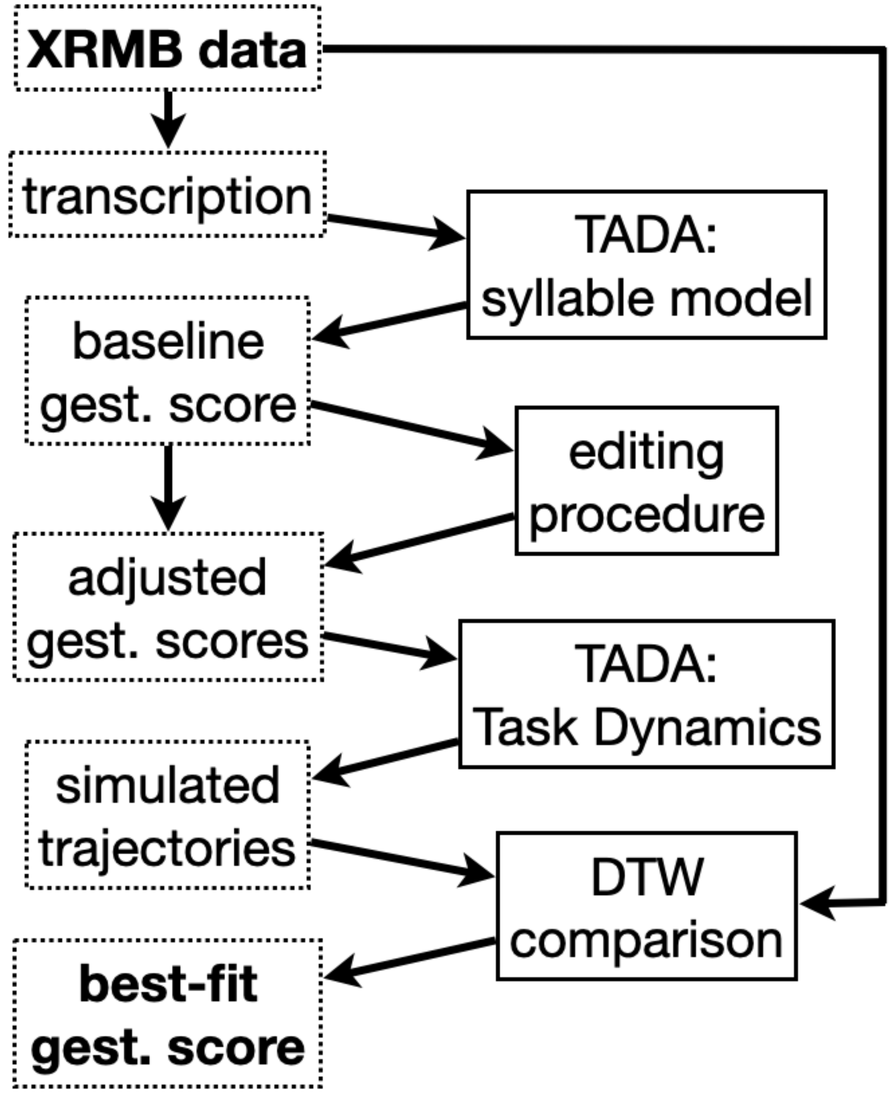
Consonant and tone categories



<five> study: methods

O'Reilly, Geissler, & Tang (2023)

- Ideal test case?
  - diphthongs: all four modes
  - C's with lips, V's with tongue
  - available data



Timing in phonology and/or phonetics?

- "Discrete Phonology" vs. "Gradient Phonetics"
- Speech timing as phonology
  - Is timing *intrinsic* or *extrinsic* to phonology?
  - Are gestures coordinated at *beginning* or *end*?
  - *Symbolic* vs. *phonetically-enriched* representations?

(Turk & Shattuck-Hufnagel 2021)