Gestural coordination in the living lexicon of spoken words

UCL – Speech Science Forum
February 11, 2021

Jason A. Shaw
Yale University

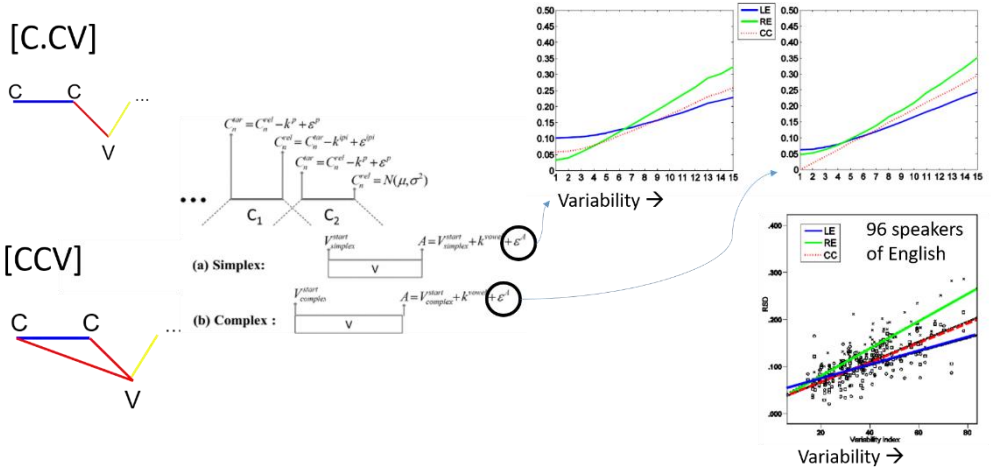
Yale

Last time (July 16th, 2012)

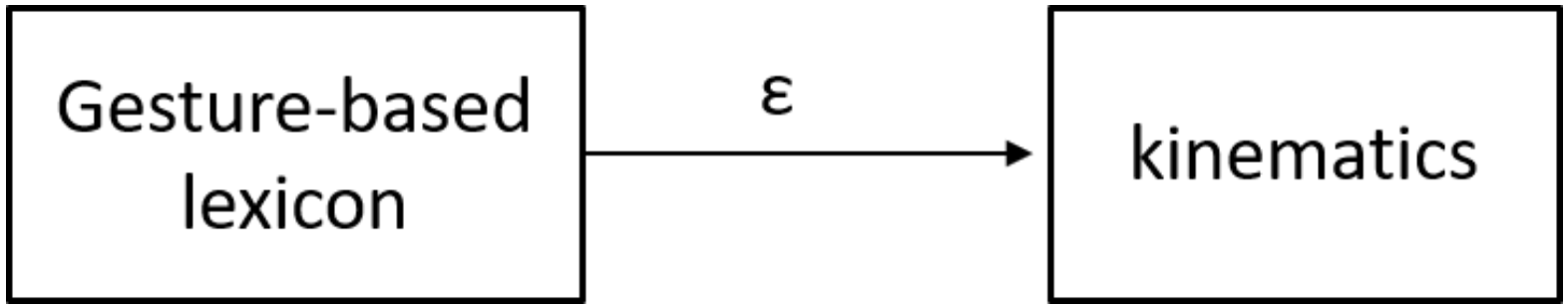
- **Dynamic invariance** in the phonetic expression of syllable structure

Shaw, J. A., Gafos, A. I., Hoole, P., & Zeroual, C. (2011). Dynamic invariance in the phonetic expression of syllable structure: a case study of Moroccan Arabic consonant clusters. *Phonology*, 455-490.

Shaw, J. A., & Gafos, A. I. (2015). Stochastic Time Models of Syllable Structure. *PLoS One, 10*(5), e0124714 01247 11-0124736.



- Abstract phonological structure conditions non-arbitrary variation in the kinematics.



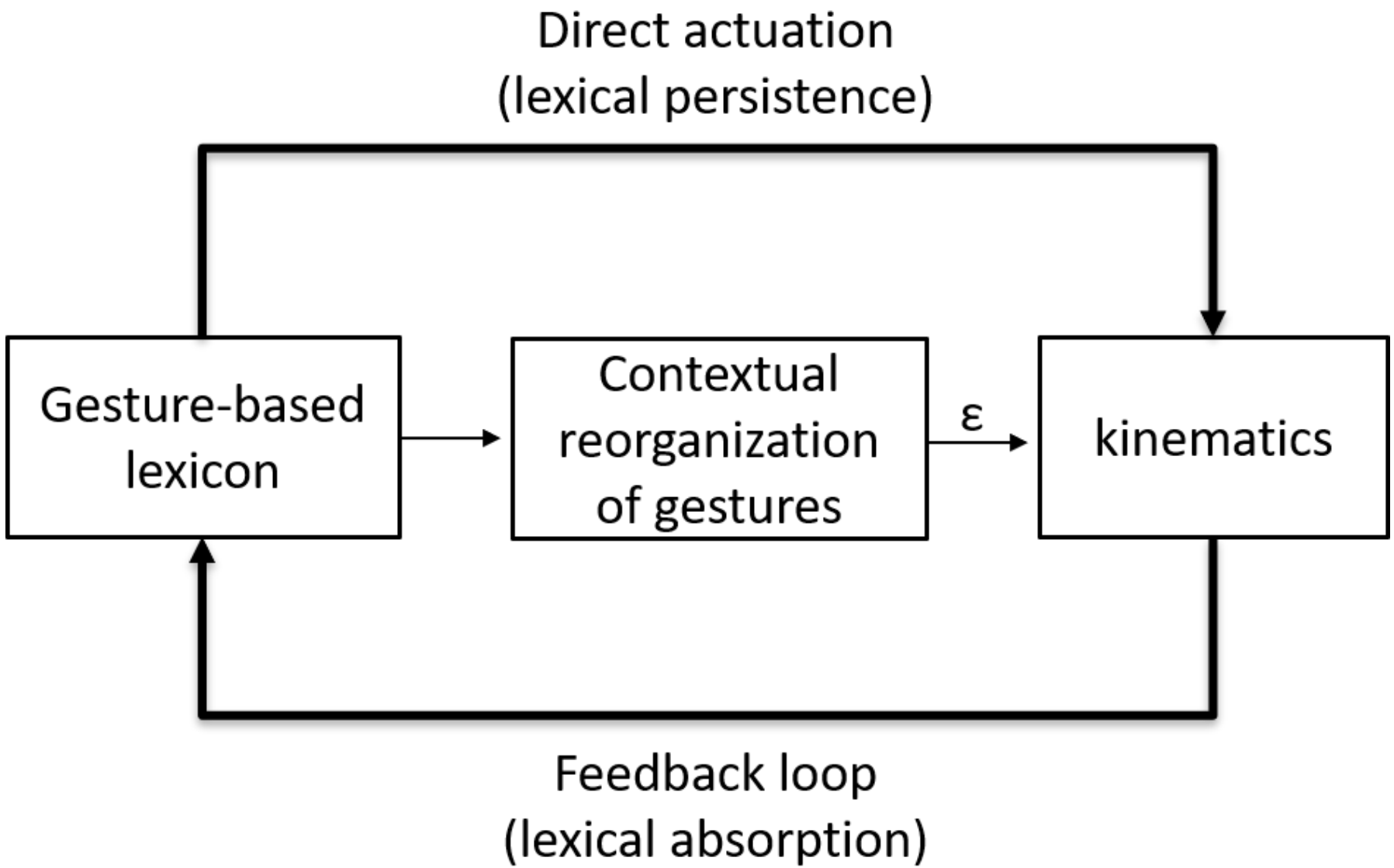
Dynamic invariance: variation in the kinematics follows from noisy actuation of coordinated gestures

This talk

1) **Dynamic invariance:** still a good idea!
   - Gestural basis for complex segments (Russian, English)

2) Gestural reorganization **conditioned by linguistic context** (language-specific)
   - Gesture deletion triggers re-organization of gestural coordination (Japanese)
   - Morpho-syntax conditions re-organization of gestural coordination (Mandarin)
   - Tone exogenesis with (Mandarin) and without (diaspora Tibetan) re-organization of segmental gestures

3) Living lexicon: word-specific phonetics
   - **Lexical absorption**: words take on the phonetic detail of the prosodic environments in which they are typically produced (Mandarin)
   - **Lexical persistence**: phonetic resistance to structurally-conditioned pitch accent reduction (Japanese).


Direct actuation
(lexical persistence)
Gesture-based lexicon
Contextual reorganization of gestures
ε
kinematics
Feedback loop
(lexical absorption)

This talk

1) **Dynamic invariance:** still a good idea!
   - Gestural basis for complex segments (Russian, English)

2) Gestural coordination is **conditioned by linguistic context**
   - Gesture deletion triggers re-organization of gestural coordination (Japanese)
   - Re-organization of gestural coordination precipitates tone loss (Mandarin)
   - Tone loss proceeds without gestural re-organization (diaspora Tibetan)

3) Living lexicon: word-specific phonetics
   - **Lexical absorption**: words take on the phonetic detail of the prosodic environments in which they are typically produced (Mandarin)
   - **Lexical persistance**: phonetic resistance to structurally-conditioned pitch accent reduction (Japanese).

Gestural basis for complex segments

collaborators

THE GRADUATE CENTER
CITY UNIVERSITY OF NEW YORK

Sejin Oh, PhD Candidate
CUNY/Haskins

Haskins Laboratories

Karthik Durvasula,
Michigan State University

MICHIGAN STATE UNIVERSITY

Alexei Kochetov,
University of Toronto

UNIVERSITY OF TORONTO

Segment sequences vs. Complex segments

- Descriptively, we recognize **segment sequences** as **distinct from complex segments:**
  - segment sequences: /pj/, /kw/, /kp/, /ps/
  - complex segments: /p^{j}/, /k^{w}/, /k͡p/, /p͡s/

- What is the basis for this structural distinction?

Phonological diagnostics for complex segmenthood

- **Contrast**: in rare cases, languages contrast complex segments and segment sequences:

  e.g., Russian C^j vs. Cj (near) minimal pairs

  a) /l^{j}ut/ ‘fierce’ /ljut/ ‘pour (3p pl)’
  b) /d^{j}at^{j}el/ 'woodpecker' /djakon/ 'deacon'
  c) /p^{j}ok/ ‘bake (3ps pst)’ /pjot/ ‘drink (3ps)’
  d) /b^{j}ust/ ‘bust’ /bjut/ ‘beat (3p pl)’

- Morpho-phonological patterns, segment distribution, language games

Phonetic diagnostics for complex segmenthood

- At least in cases of contrast, there must be phonetic differences, but…
- Complex segments are **not** systematically **shorter** in phonetic duration than gesturally matched segment sequences (Gouskova & Stanton 2019, c.f., Trubetzkoy 1939)

- We pursue the hypothesis that there is a **gestural basis** to the distinction with kinematic consequences:

HYPOTHESIS (Shaw, Durvasula, Kochetov, 2019)

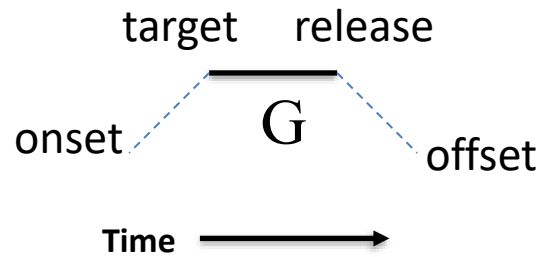
ℍ: **complex segments** involve gestures coordinated according to **onset** landmarks

Key assumptions ($\mathbb{A}_1$, $\mathbb{A}_2$)

$\mathbb{A}_1$: **Gestures** are forces that drive articulators to task goals over time (e.g., Browman & Goldstein 1986)

$\mathbb{A}_2$: A gesture can be decomposed into a series of states or **landmarks** (Gafos 2002)

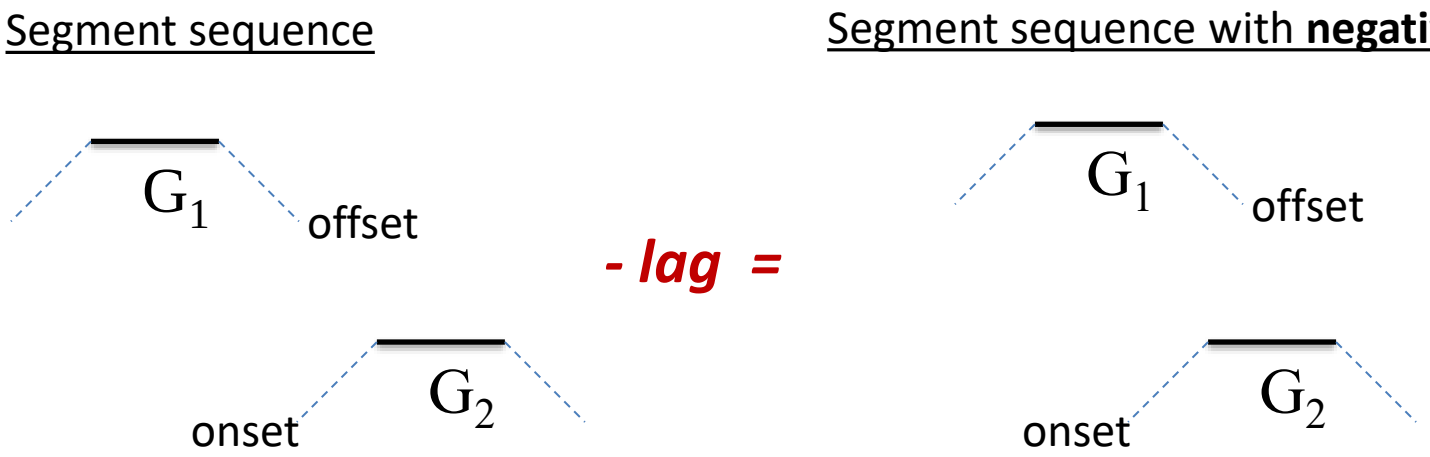
Gesture landmarks:



Key assumptions ($\mathbb{A}_3$, $\mathbb{A}_4$)

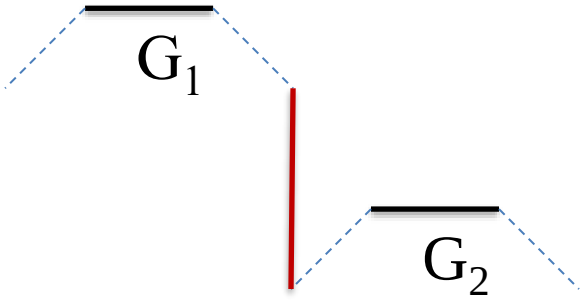
$\mathbb{A}_3$: **Coordination** relations between gestures make reference to **gesture landmarks**: e.g., the *onset* of G_2 is coordinated with the *offset* of G_1 (e.g., Gafos 2002)

$\mathbb{A}_4$: There may be a **consistent +/- lag** between coordinated landmarks (e.g., Shaw & Gafos 2015)

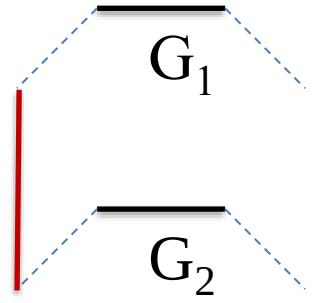


Lag can cause surface ambiguity

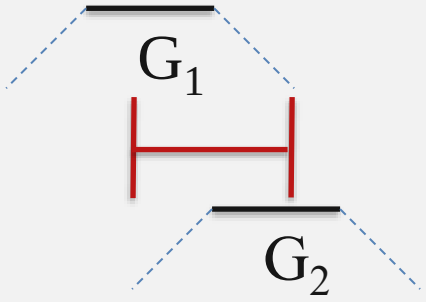
Segment sequence – no lag



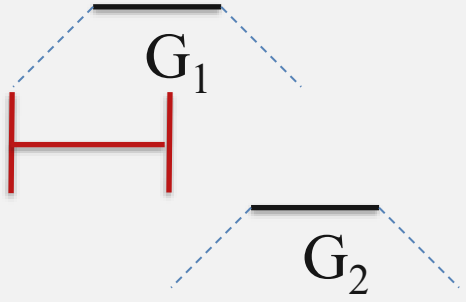
Complex segment– no lag



Segment sequence – negative lag



Complex segment – positive lag



similar patterns of gesture overlap can derive from different coordination relations

Stochastic models of coordination: approach

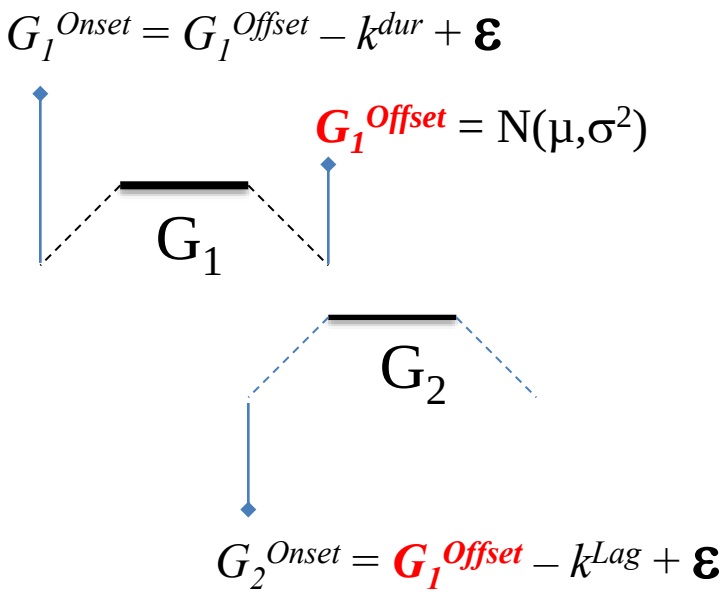
(following Shaw & Gafos 2010, 2015; Gafos et al. 2014; Shaw et al. 2011)

Guiding principle: phonetic variation derives from noisy actuation of discrete gestures and coordination relations between them.

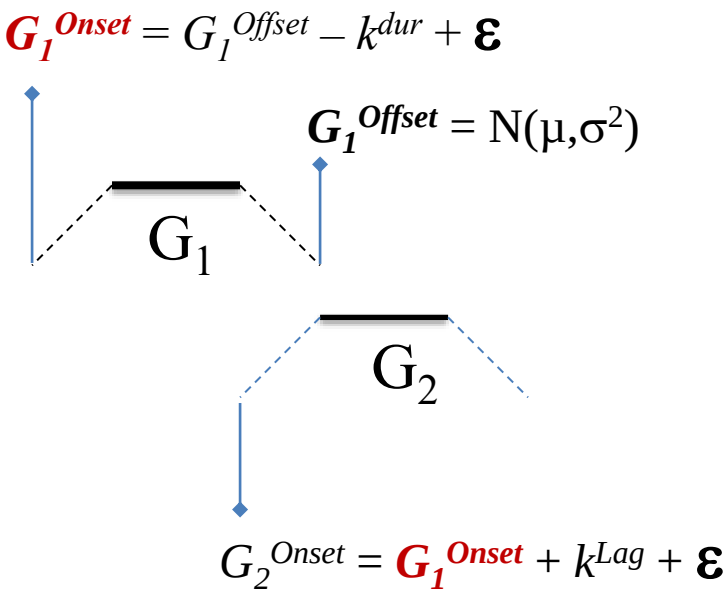
- Define coordination relations as statistical dependencies between gesture landmarks
- Simulations:
  - **Random variation**: kinematics as noisy actuation of dynamics
  - **Controlled variation**: introduce systematic variation in one phonetic parameter to observe how other phonetic parameters vary.
- Identify differences in **structure-specific covariation** across competing hypotheses.

Random variation: each landmark simulated with noise

Segment sequence, e.g., [pj]

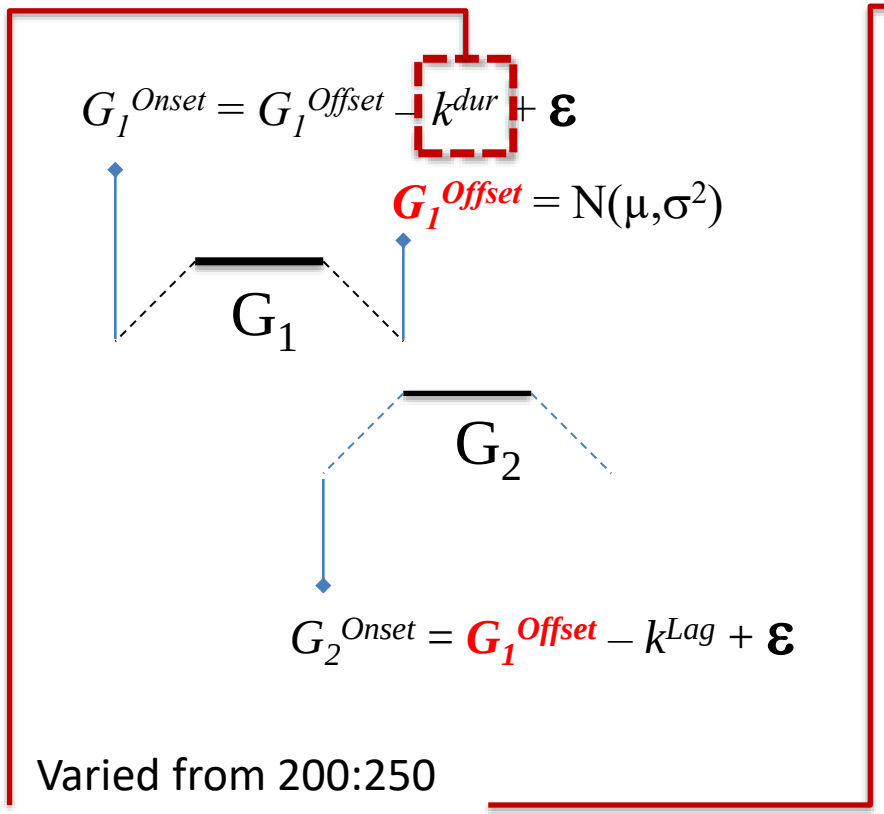


Complex segment , e.g., [p^j]



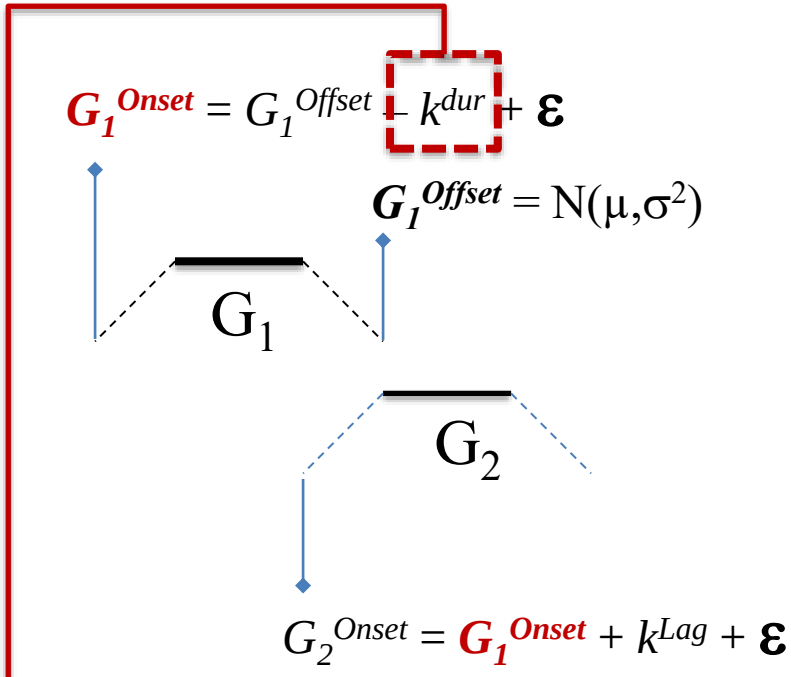
Controlled variation: G_1 duration varied systematically

Segment sequence, e.g., [pj]



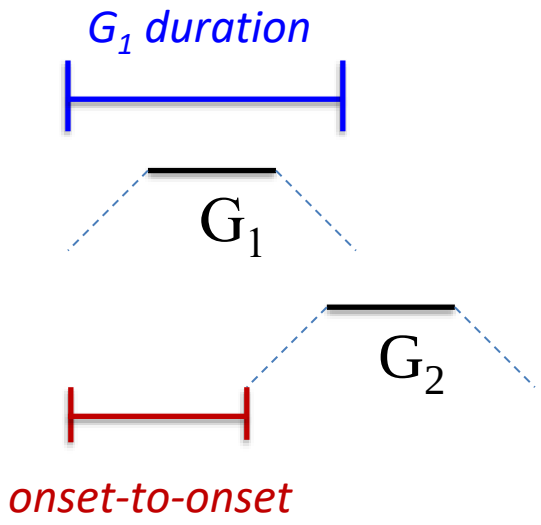
Varied from 200:250 ms in 1 ms steps

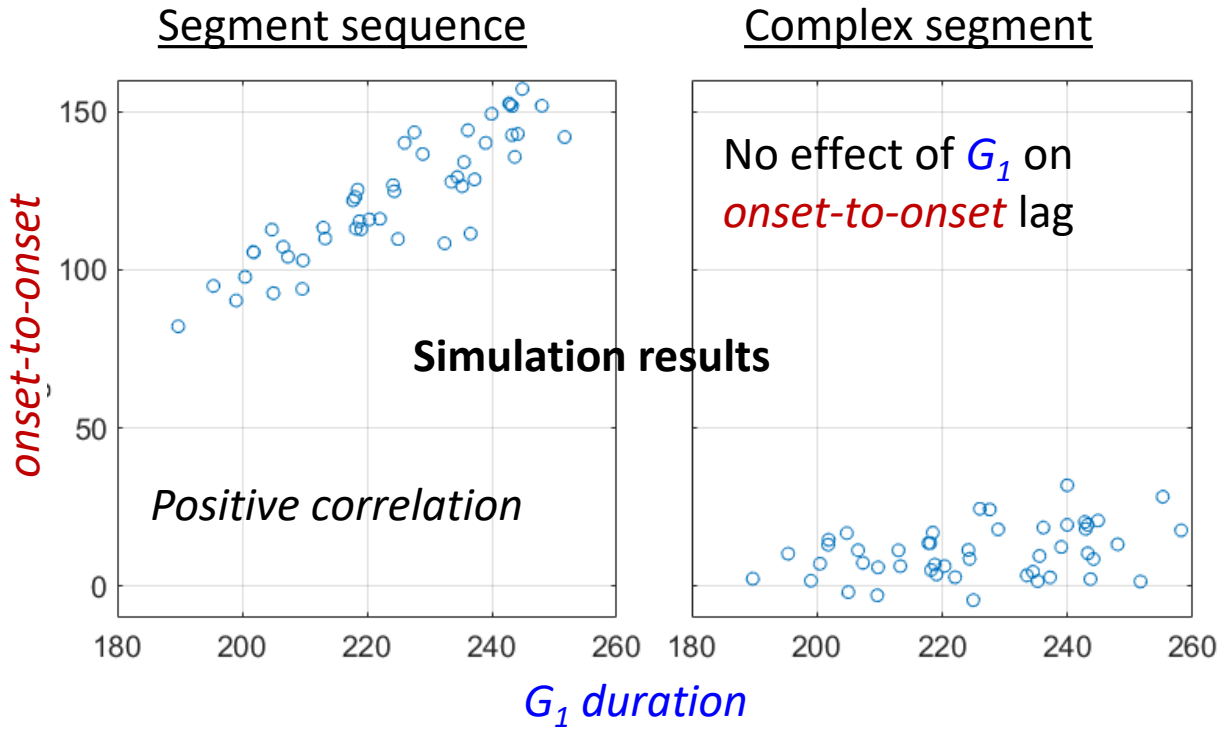
Complex segment , e.g., [p^j]



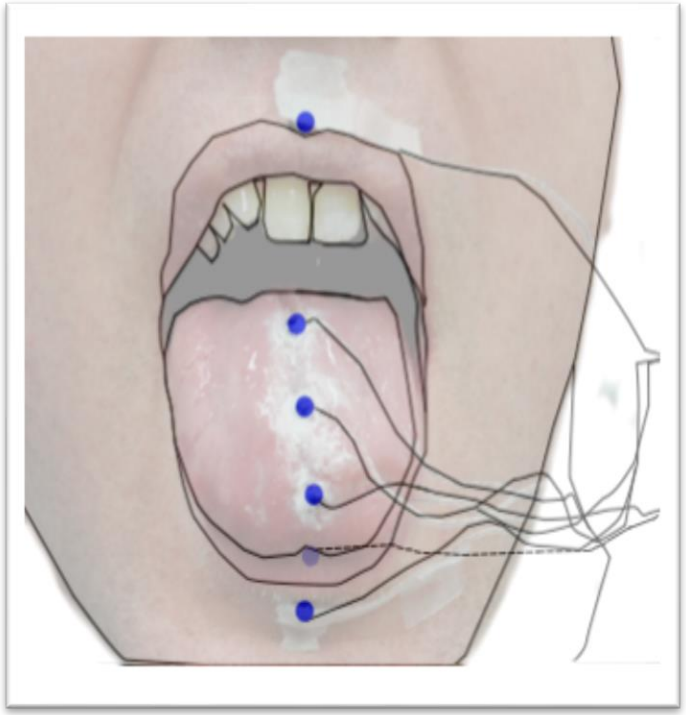
Coordination relations constrain phonetic covariation

Key simulation result: How *variation in G_1 duration* influences *onset-to-onset lag* **depends** on coordination relations.





Empirical tests



Sensor placement
for NDI Wave data

Fleshpoint tracking using **EMA & X-Ray Microbeam**

(1) Russian palatalized labial vs. control sequence

- Subset of EMMA data from Kochetov (2006)
- 3 female speakers
- /p^{j}/ & /br/ sequences
- 2 items per sequence; 4-5 reps

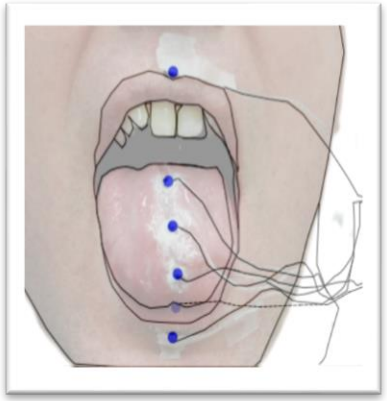
(2) English /bj/ sequences

- Wisconsin X-Ray Microbeam (Westbury 1994)
- 20 speakers, 1 rep per speaker
- Task 33: “beautiful” in word list

(3) Russian vs. English

- New NDI Wave 3D EMA data
- 8 speakers (4 Russian), 20-30 repetitions per item
- Russian: /b^{j}/, /p^{j}/, /m^{j}/, /v^{j}/ items in carrier phrase
- English: /bj/, /pj/, /mj/, /vj/ items in carrier phrase

Data measurement

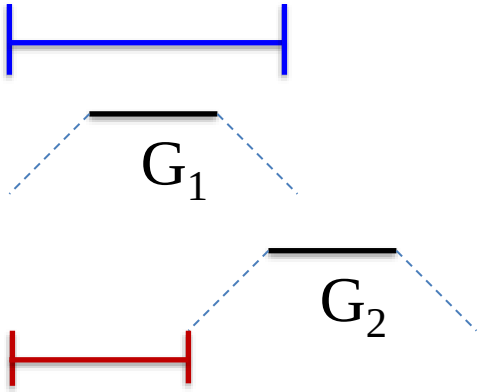


Gestures parsed according to primary articulator: tongue blade for [j]; tongue tip for [r] (rhotic trill); lip aperture for [m], [p], [b], [v]

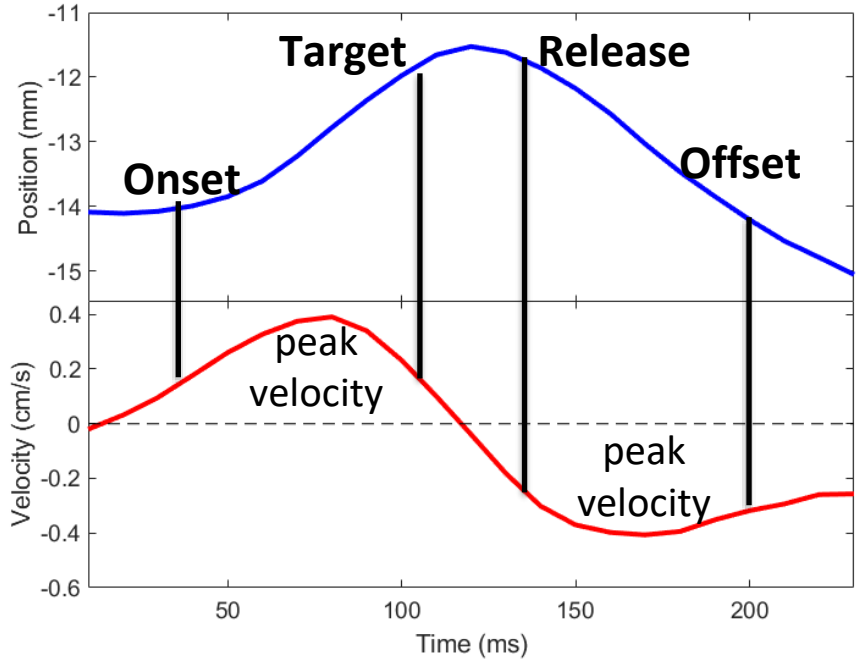
Landmarks: **Onset, Target, Release, Offset** determined by 20% threshold of peak velocity in Mview (Tiede 2005)

Dependent measures

G_1 duration = **Offset (G_1) – Onset (G_1)**

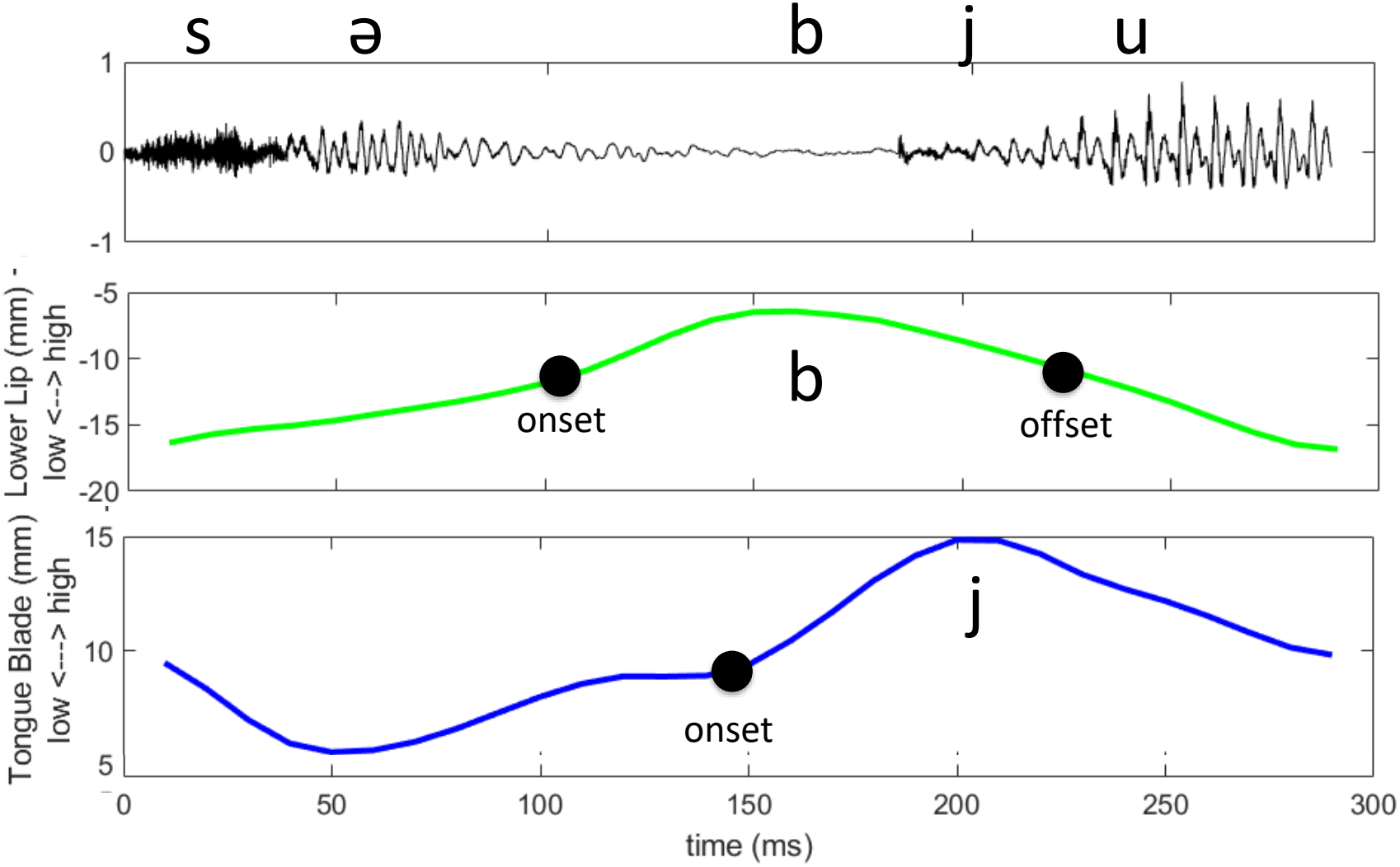


onset-to-onset = **Onset (G_2) – Onset (G_1)**



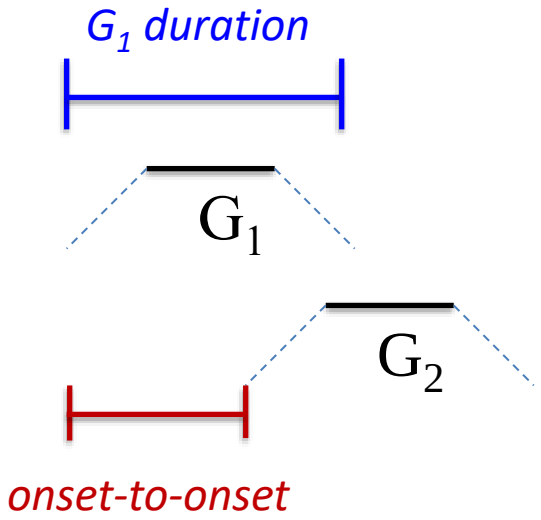
Representative token (English data)

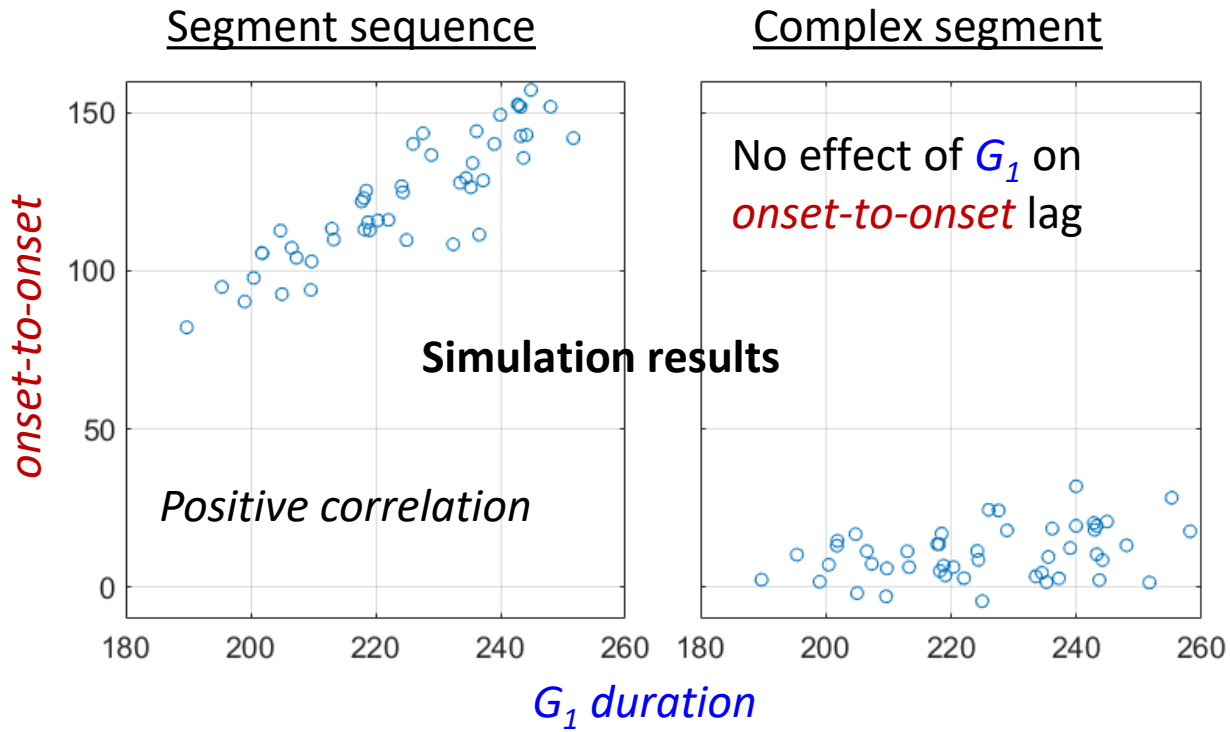
"It'**s a bu**<u>tte</u> perhaps"



Predictions

1. **Segment sequence timing (all English data and Russian /br/)**: the lag between the onsets of gestures increases with the duration of the first gesture.
2. **Complex segment timing (Russian /p^{j}/)**: the lag between the onsets of gestures is not affected by the duration of the gestures.



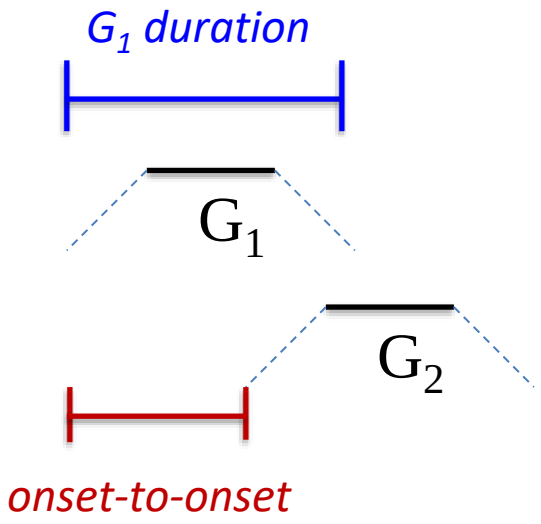


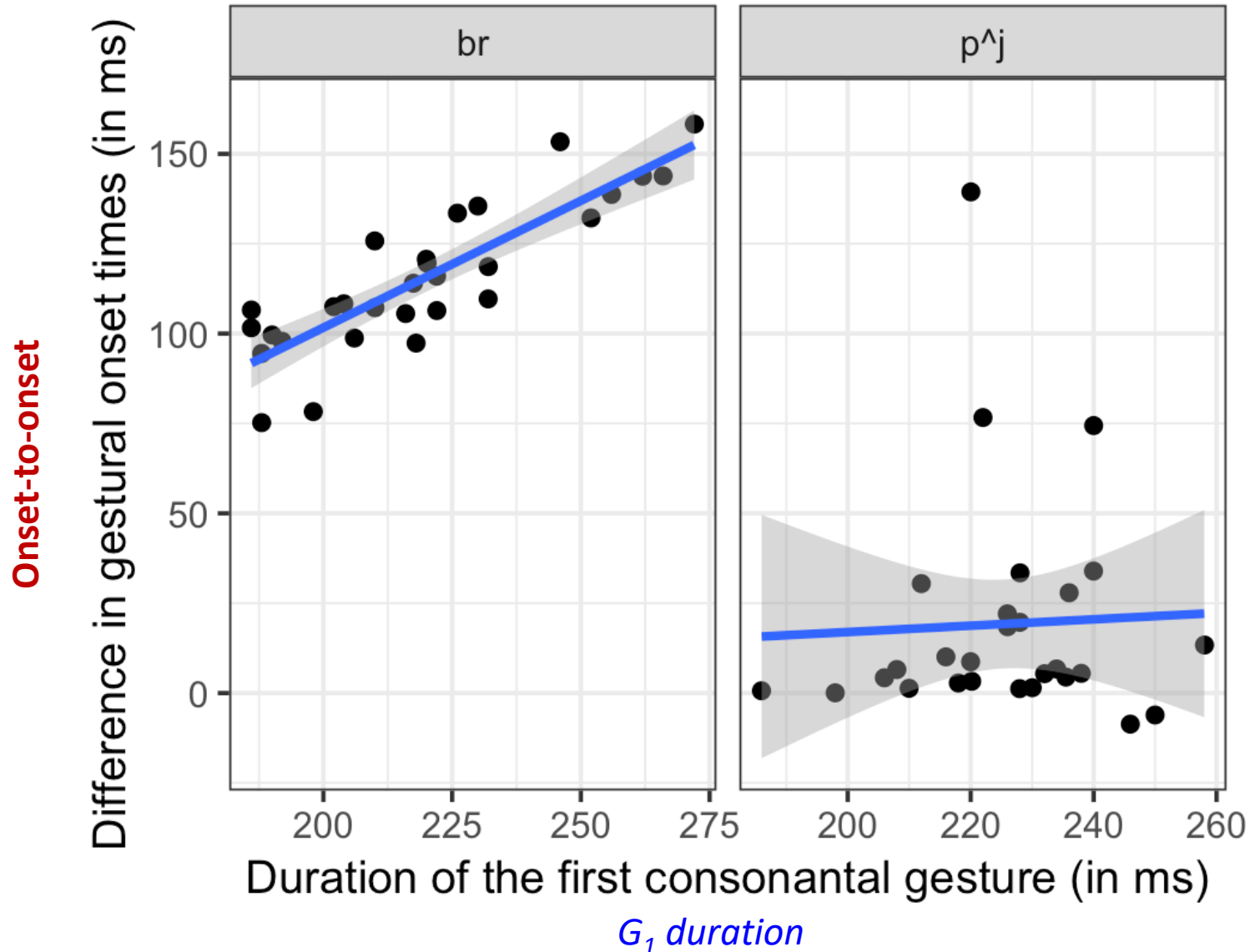
/**br**at#p^{j}atava/ ‘брат пятого’ /**br**at#padaja/ ‘брат падая’

Russian data

(Kochetov 2006)

/tat#**p^{j}**apɨ/ ‘тат пяпы’ /ta#**p^{j}**apɨ/ ‘та пяпы’



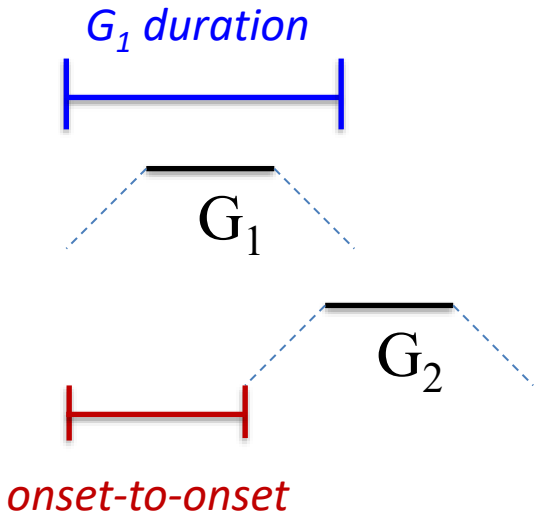


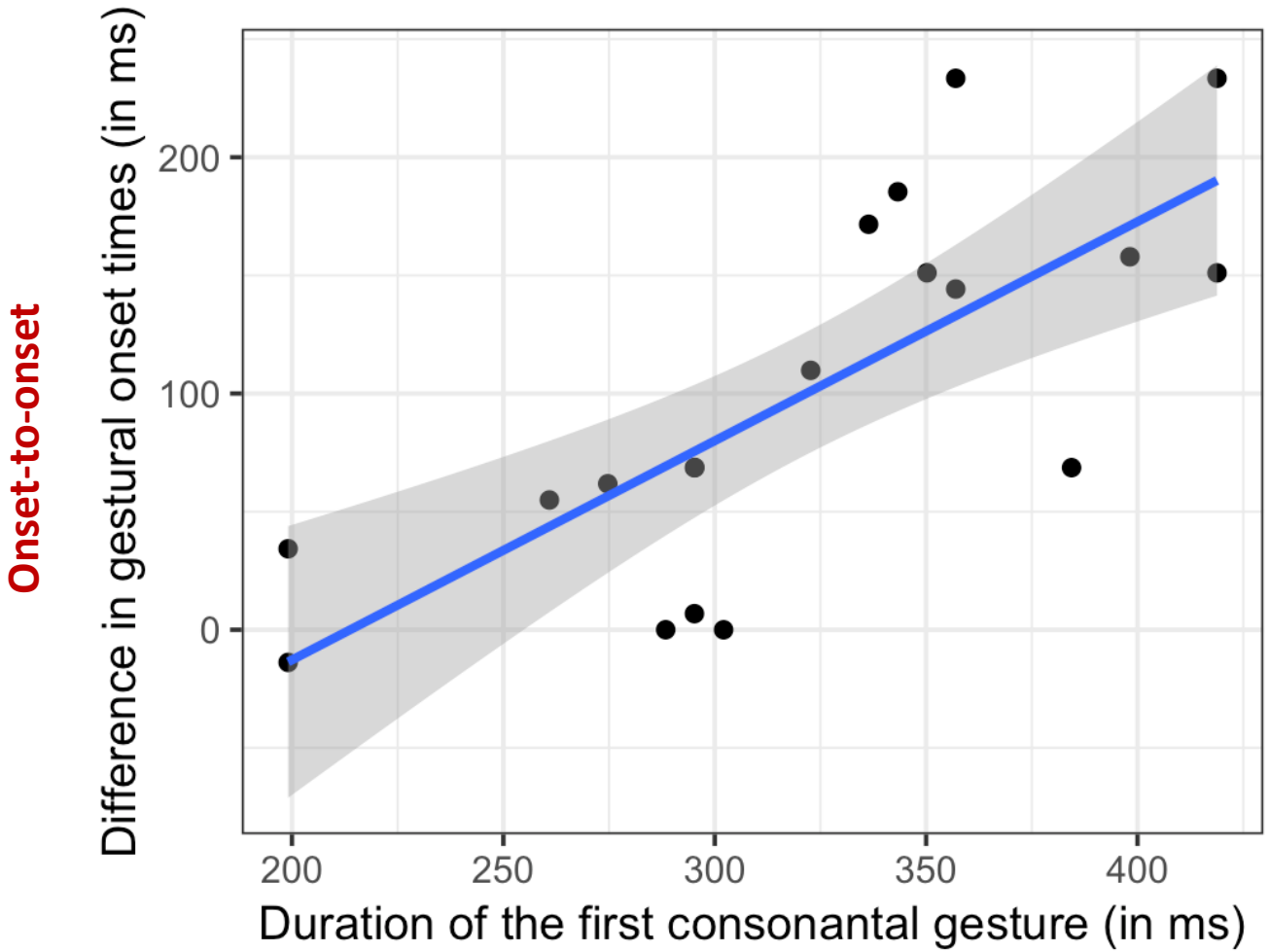
3 speakers; 4-5 reps per item

English control data

(X-ray Microbeam)

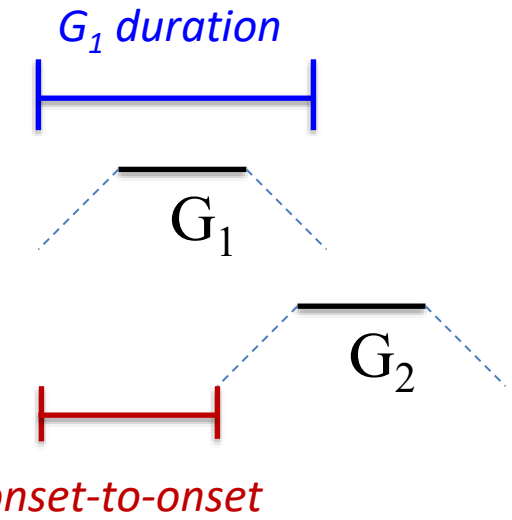
20 speakers
(1 rep)
"beautiful"





New EMA experiment

- No main effect of language on onset-to-onset lag.
- Strong interaction between *G_1 duration* and language



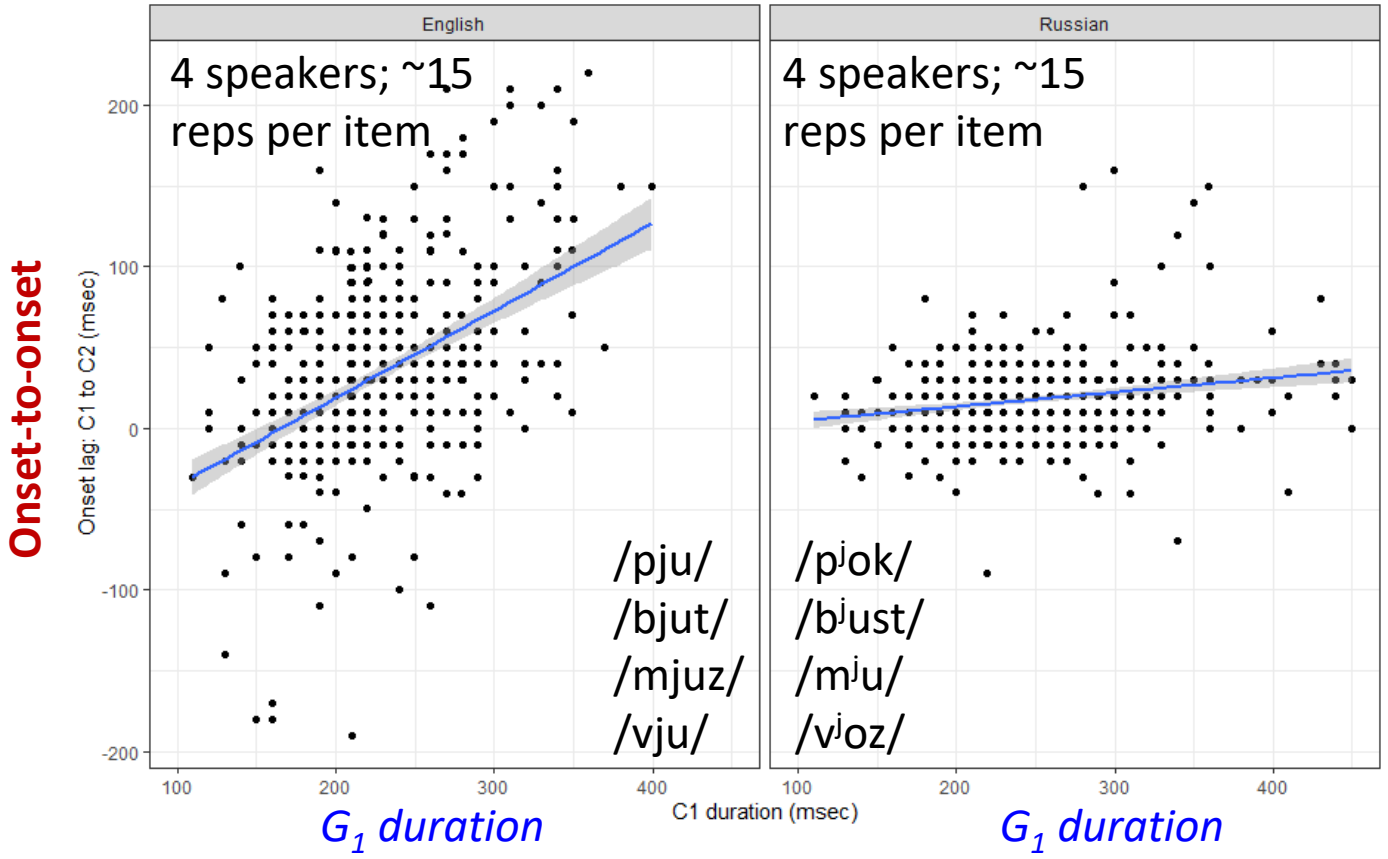


Table 1: LME Model comparison (*onset-to-onset* lag~)	Df	AIC	logLik	Chisq	Pr(>Chisq)
1 + (1|subject)+(1|item)	4	10270.7	-5131.35	NA	NA
1 + G_1 duration + (1|subject)+(1|item)	5	10159.6	-5074.8	113.1	<0.00001
1 + G_1 duration + language + (1|subject)+(1|item)	6	10161.6	-5074.8	0.002	0.96
1 + G_1 duration * language + (1|subject)+(1|item)	7	10076.9	-5031.5	86.6	<0.00001

Discussion: gestural basis of complex segments

- Predictions borne out:
  - **English** labial-palatal gestures timed as **segment sequences**
  - **Russian** labial-palatal gestures timed as **complex segments**

- Phonologically relevant **dynamics** can be diagnosed in the **kinematics** because of how **coordination relations** structure variability (Shaw et al. 2011; see also Oh 2020 on Korean coda nasals)

- Consistent with view of the lexicon as consisting of discrete gestures and coordination relations between them.

Future directions

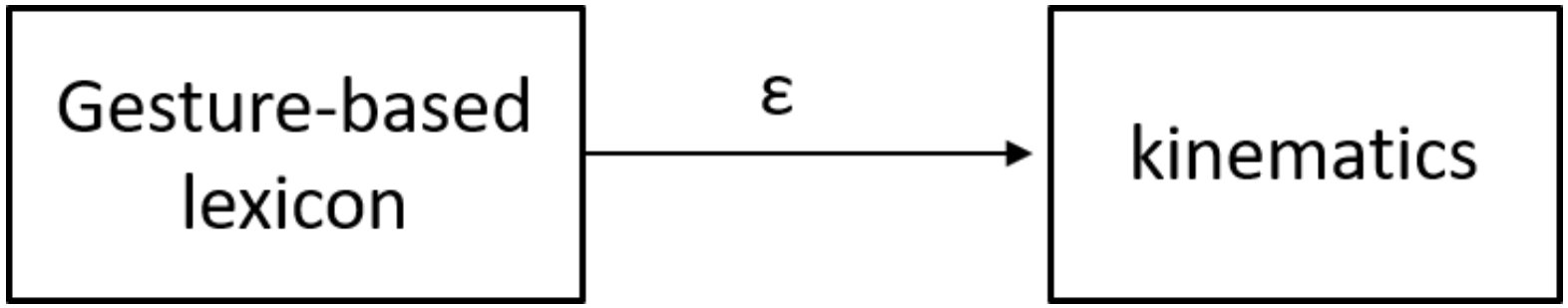
- We focused here on underlyingly palatalized consonants of Russian, but consonant-glide sequences are also described as “palatal” (Timberlake 1984) while “plain” consonants are velarized/uvularized (Roon et al. 2019), e.g.:

  /**p**$^{\text{ɣ}}$jot/ ‘drink (3ps)’ → [**p**$^{\text{j}}$jot]

- Do underlying plain (velarized/uvularized) consonants also show gestural timing characteristic of complex segments? (Oh et al., 2020, in prep)

- Gestural basis of complex segments may generalize to other cases, including those not traditionally thought of as “complex”:

  → pre-nasalized stops, etc., but also aspirated stops, nasals, ….



Dynamic invariance: variation in the kinematics follows from noisy actuation of gestures

This talk

1) **Dynamic invariance:** still a good idea!
   - Gestural basis for complex segments (Russian, English)

2) Gestural reorganization **conditioned by linguistic context** (language-specific)
   - Gesture deletion triggers re-organization of gestural coordination (Japanese)
   - Morpho-syntax conditions re-organization of gestural coordination (Mandarin)
   - Tone exogenesis with (Mandarin) and without (diaspora Tibetan) re-organization of segmental gestures

3) Living lexicon: word-specific phonetics
   - **Lexical absorption**: words take on the phonetic detail of the prosodic environments in which they are typically produced (Mandarin)
   - **Lexical persistence**: phonetic resistance to structurally-conditioned pitch accent reduction (Japanese).

Gesture coordination is sensitive to linguistic context: **collaborators**

Japanese

CVC → CC

Mandarin

Tibetan

CV̀ → CV

Shigeto Kawahara
Keio University

慶應義塾大学 言語文化研究所
The Keio Institute of Cultural and Linguistic Studies

Muye (Andy) Zhang
Yale, PhD Candidate

Chris Gesissler
Yale, PhD Candidate

NSF Yale

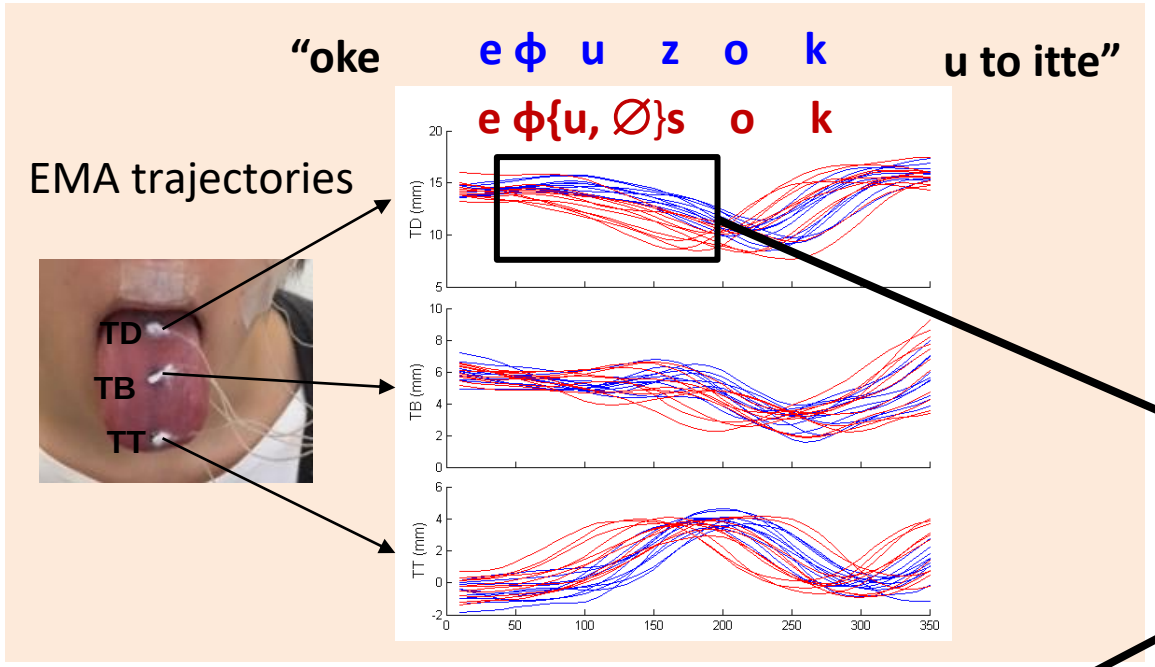
High vowel devoicing in Tokyo Japanese

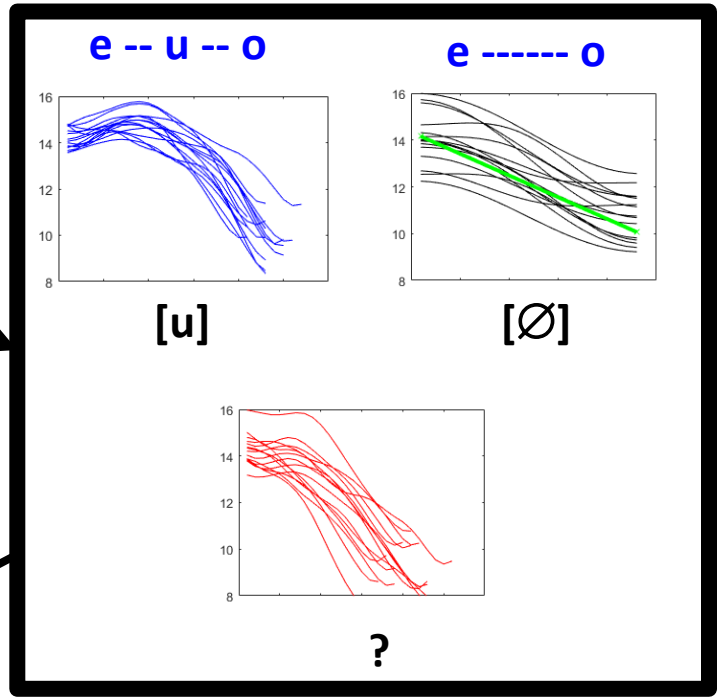
high vowels are devoiced between two voiceless consonants and between a voiceless consonant and a pause

ʃ̥utaise:	‘individuality’	ʃudaika	‘theme song’
ʃ̥isen	‘eye gaze’	ʃizen	‘nature’
ɸ̥usoku	‘shortage’	ɸuzoku	‘affiliated’
tʃ̥ikai	‘pledge’	tʃigai	‘difference’
katsu̥toki	‘win time’	katsudo:	‘life activities’
aʃ̥ika	‘sea lion’	saʃiga	‘inserted picture’

Fujimoto, M. (2015). Chapter 4: Vowel devoicing. In H. Kubozono (Ed.), *The handbook of Japanese phonetics and phonology.* Berlin: Mouton de Gruyter.

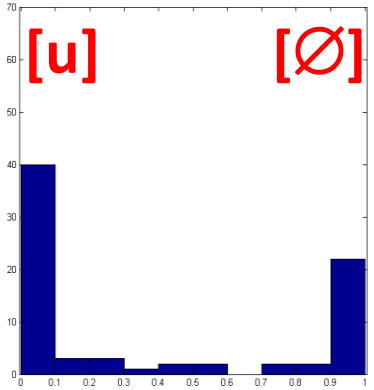
The lingual gesture of devoiced vowels is optionally deleted



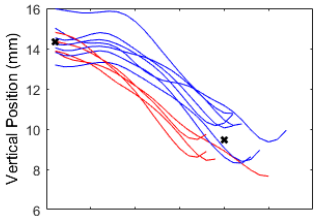


Bayesian classification of devoiced trajectories

(six speakers)



Largely categorical results—most tokens either [u] or [∅]



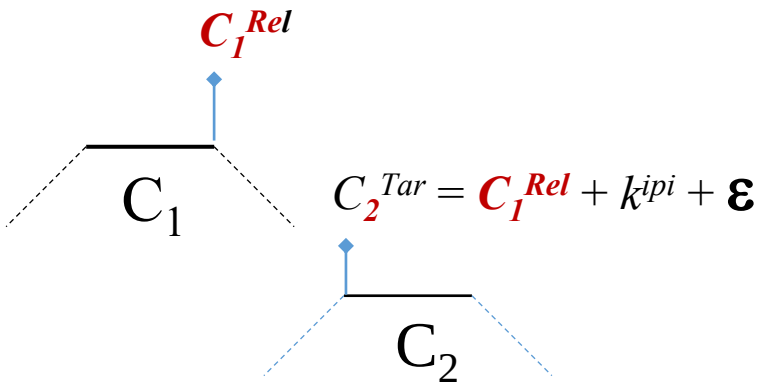
Shaw, J. A., & Kawahara, S. (2018). Assessing surface phonological specification through simulation and classification of phonetic trajectories. *Phonology*, *35*(3), 481-522.

CC vs. CVC

What happens to gestural coordination when the vowel height target for /u/ is deleted?

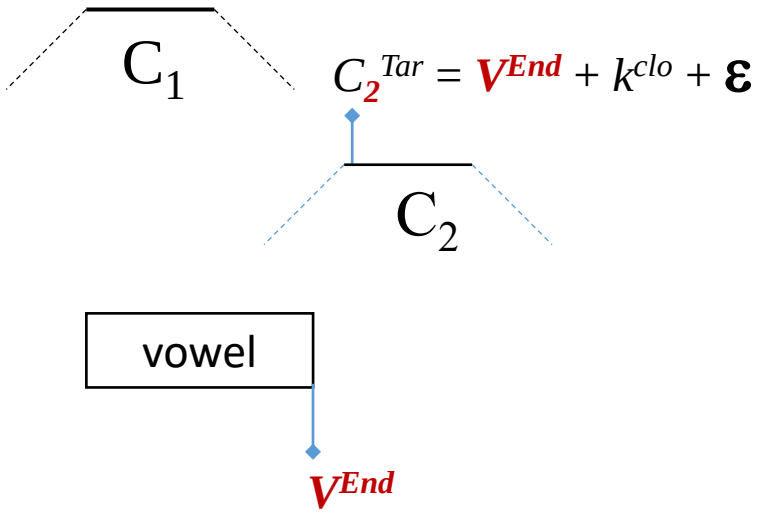
CC sequence

C_2 is timed to the release of C_1

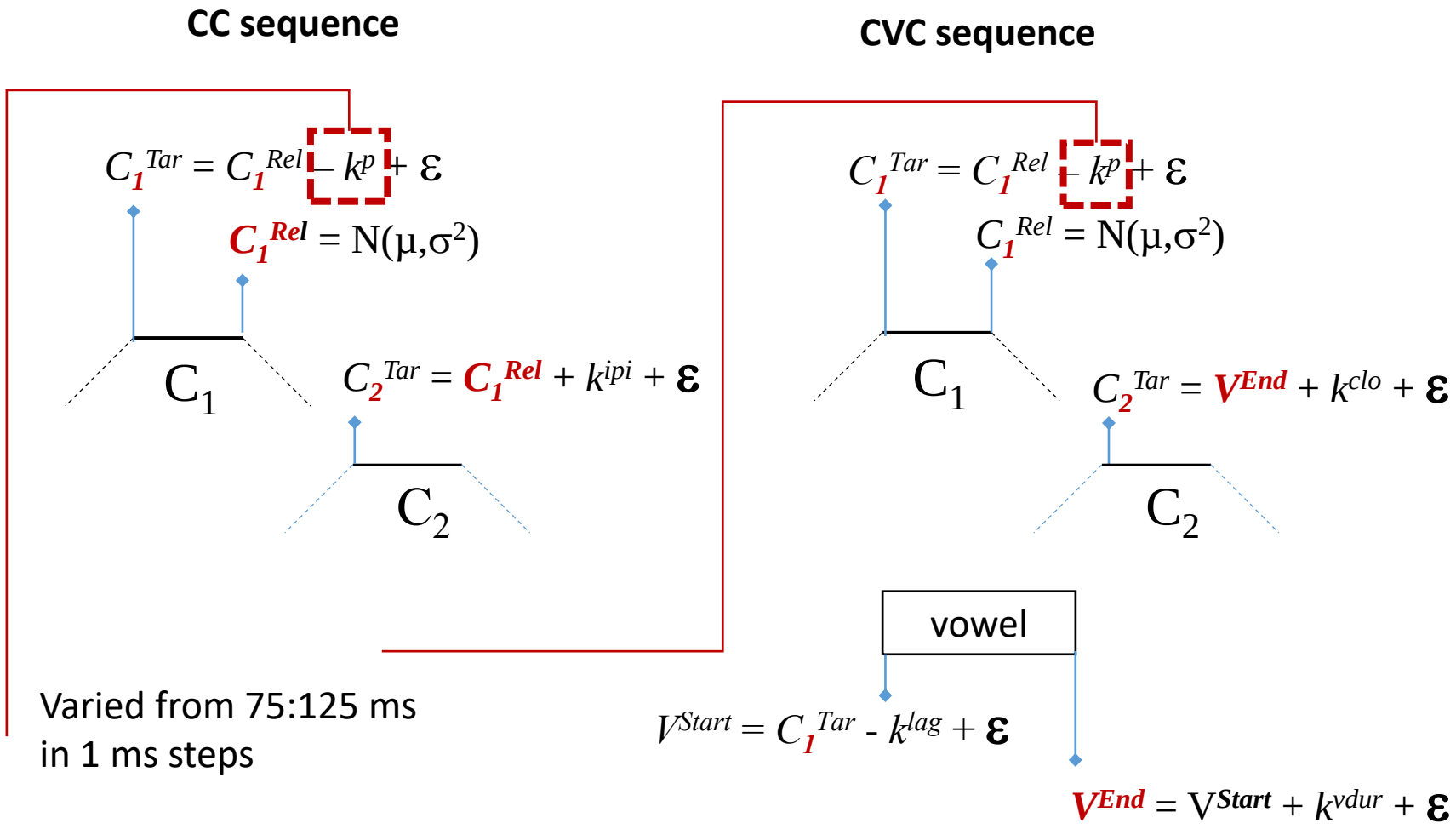


CVC sequence

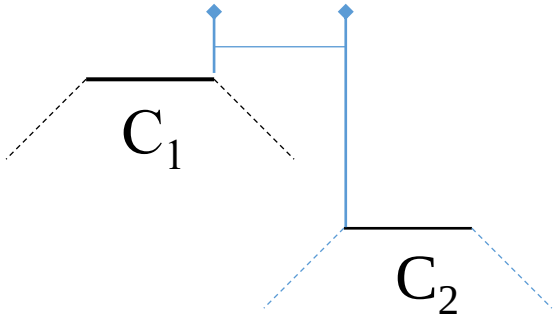
C_2 is timed to the end of V

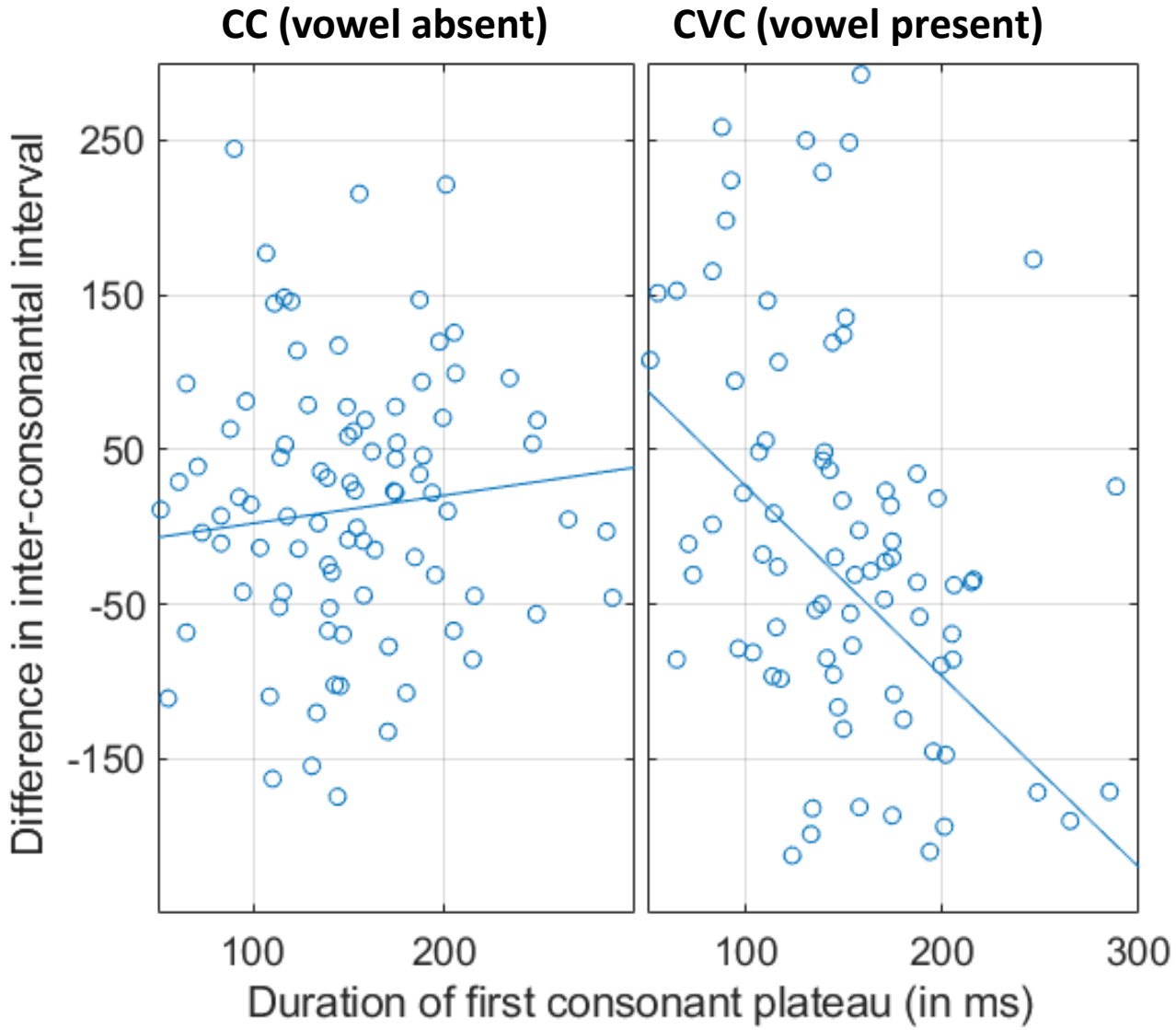


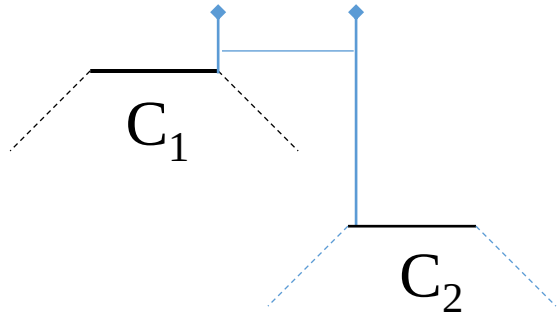
Simulation algorithm (effect of C_1 plateau duration on IPI)



Simulation results







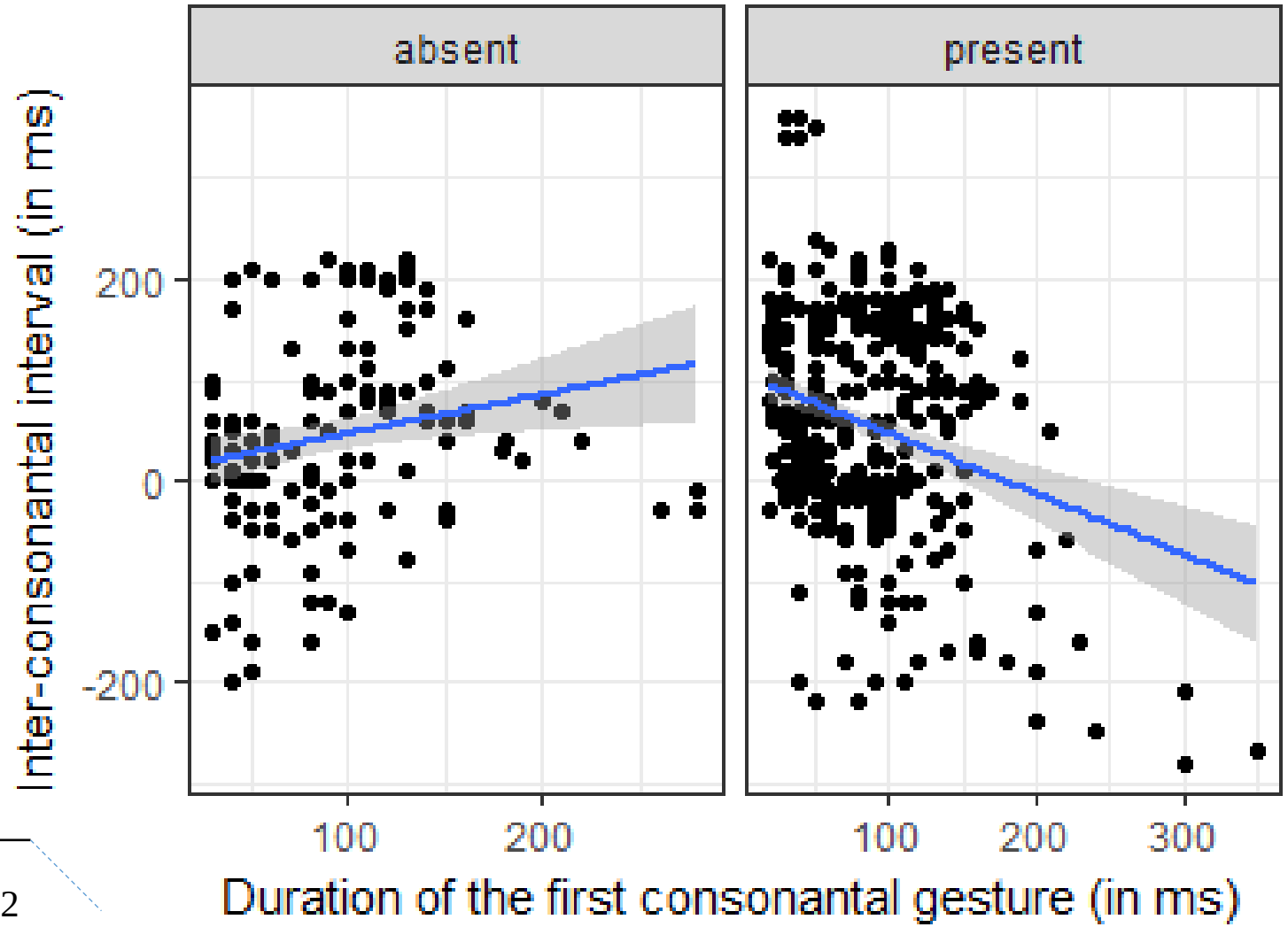
vowel

Japanese data

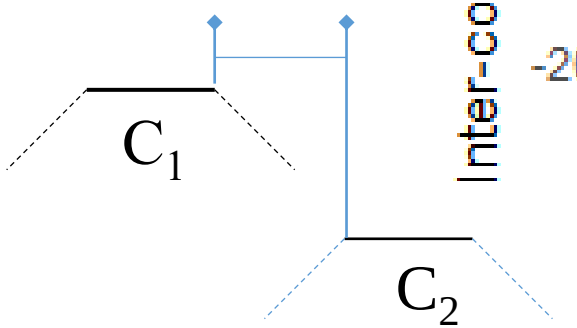
6 speakers;
10-15 reps
per item

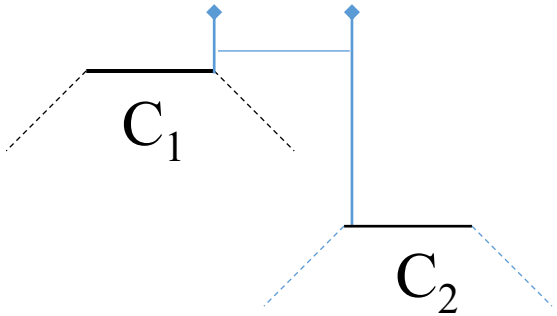
Shaw, J. A., & Kawahara, S. (2018). The lingual articulation of devoiced /u/ in Tokyo Japanese. *Journal of Phonetics, 66*, 100-119.

[ma**st**a:] [**ɸs**oku]
[**ʃt**aise:] [ka**tst**oki]



[ma**sʉt**a:] [**ɸʉs**oku]
[**ʃʉt**aise:] [ka**tsʉt**oki]

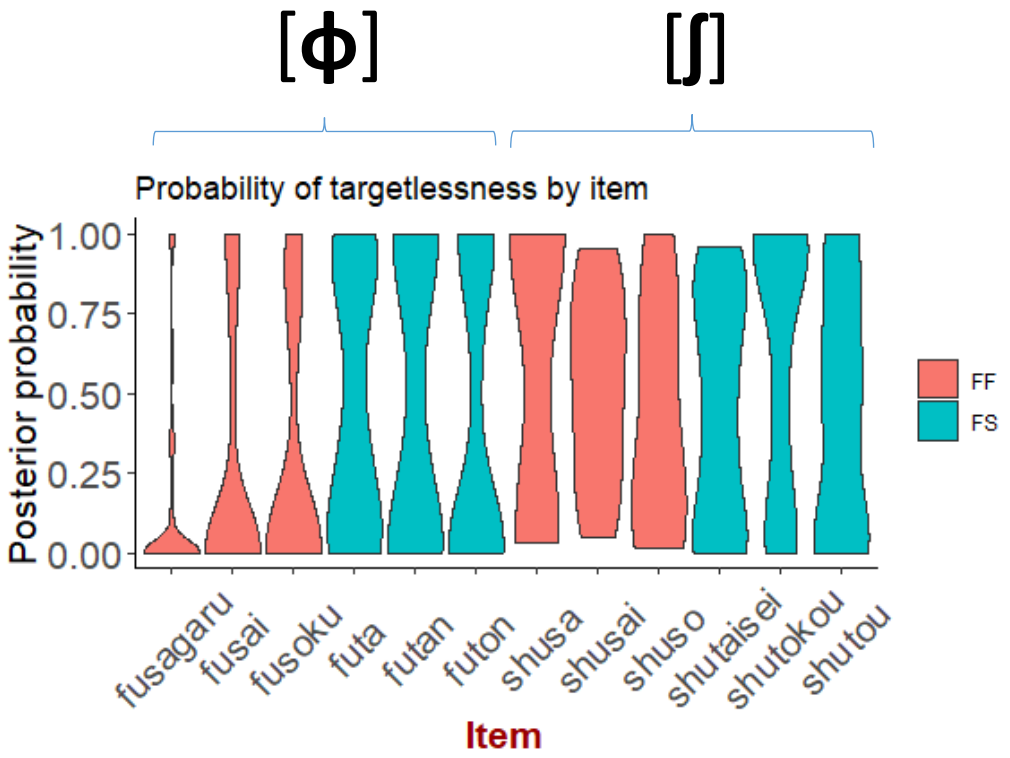


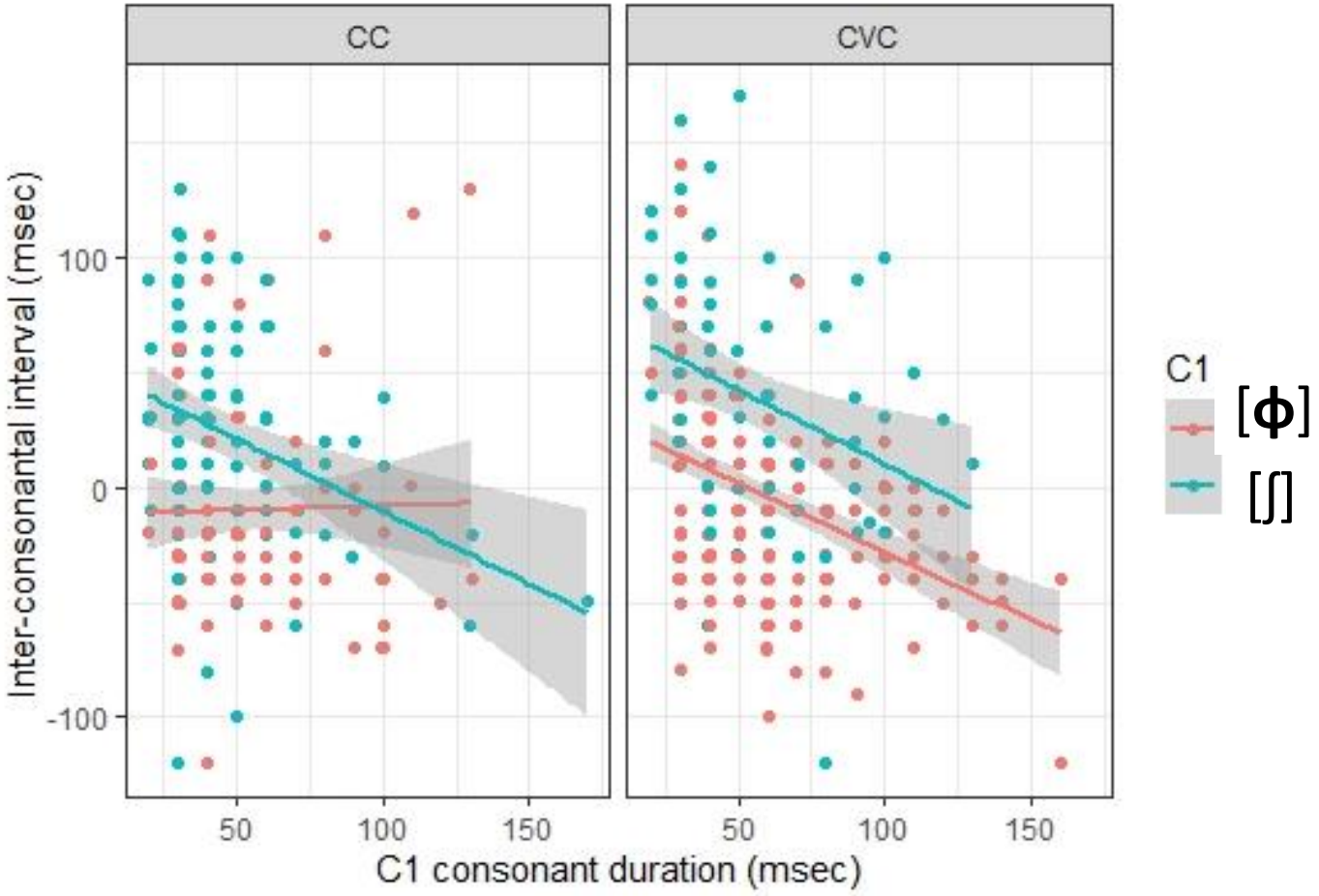


vowel

Replication

6 new speakers; more items; 10-15 reps





Only coronal-initial cluster showed gestural reorganization.

Shaw, J. A., & Kawahara, S. (2021). More on the articulation of devoiced /u/ in Tokyo Japanese: effects of surrounding consonants. manuscript, Yale University and Keio University. 47 pgs.

Discussion: discontinuous variation

- In Tokyo Japanese, **devoicing triggers variable deletion** (categorical) of a vowel height target in [u]

- Deletion of vowel height target triggers gestural re-organization (categorical) for [**ɸ**]-initial words but not for [**ʃ**]-initial words.

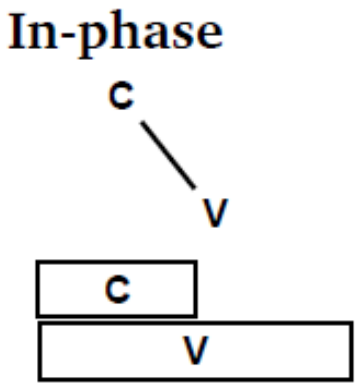
- Possibly related to lexical gap:

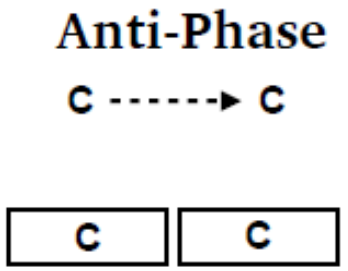
	/u/	**/i/**
ʃ	**ʃu**	**ʃi**
ɸ	**ɸu**	-- ***ɸi**

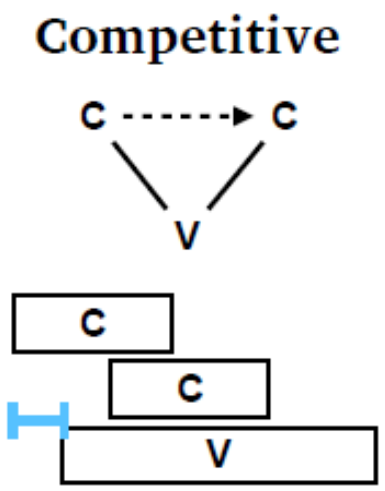
- Gestural reorganization except when contrast is at stake?

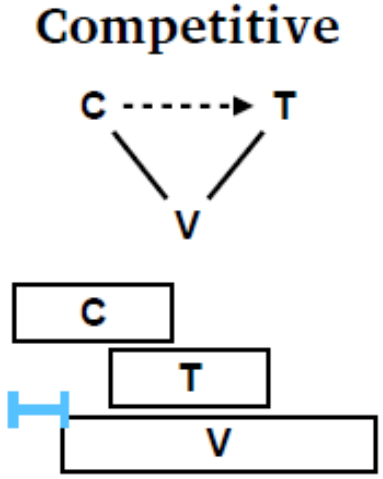
Gesture complexity and tone

- So far, we've looked at rather simple distinctions:
- Combinations of coordination relations can apply competing forces, resolved by compromise.
- In this respect tone has been observed to behave similarly to segmental gestures.

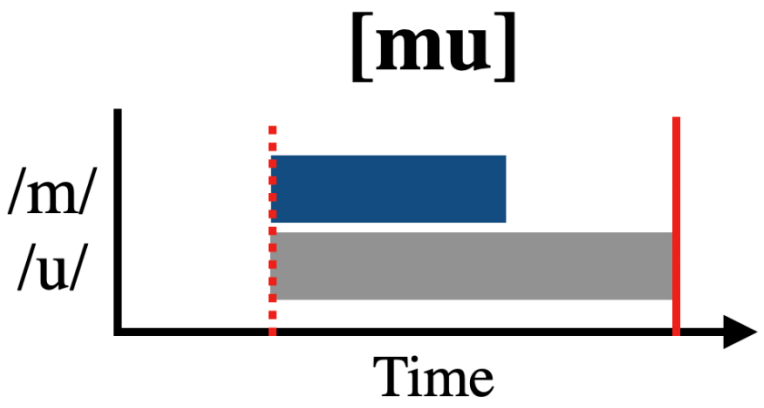






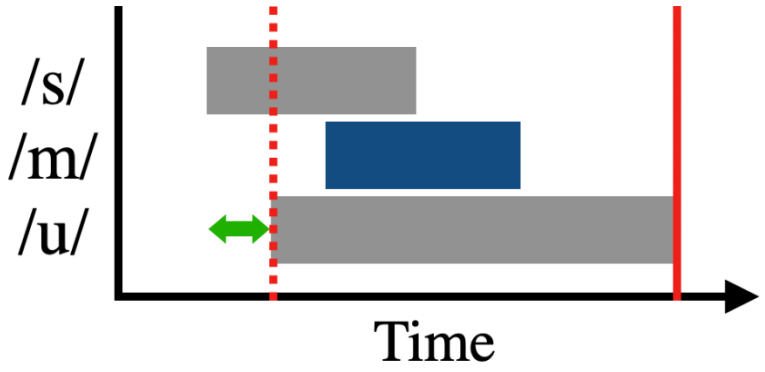


Tone as gesture



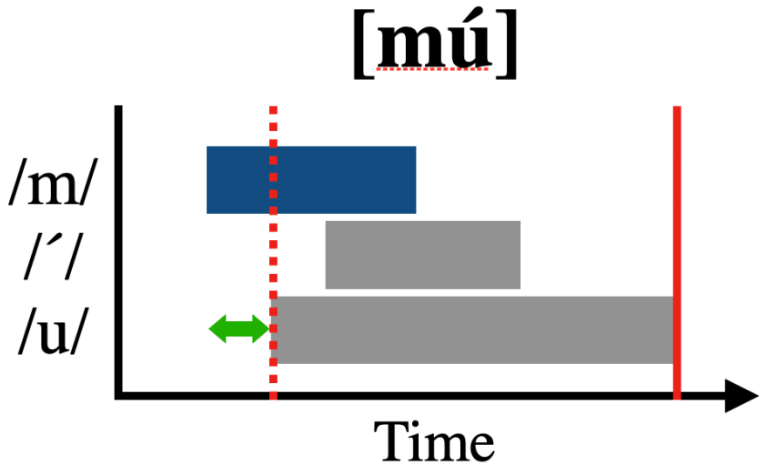
Russian (Kozhevkikov & Chistovich, 1965)
American English (Lofqvist & Gracco, 1999)

+ C



Romanian (Marin, 2013),
American English (Marin & Pouplier, 2010),
Polish (Hermes et al., 2017)

+ T



Mandarin Chinese (Gao, 2008; Shaw & Chen, 2019)
Thai (Karlin & Tilsen, 2015)
Lhasa Tibetan (Hu, 2016)

Mandarin Chinese tones

Lexical tones (1-high, 2-rising, 3-low, 4-falling)

tang: 1: 汤 soup 2: 糖 candy 3: 躺 lie-down 4: 烫 scalding-hot

Toneless syllables or “neutral tone” (Chen & Xu, 2006)

- lexically toneless, e.g. *ma* QUES, *le* PERF → “absent”
- disyllabic words,
  e.g. /bō.lí/ ‘glass’ > [bō.**li**], /yún.cǎi/ ‘cloud’ > [yún.cai] → “reduced”
  especially in compounds, e.g. [bō.li.bēi] glass cup

Muye (Andy) Zhang

Mandarin materials – 7 sets

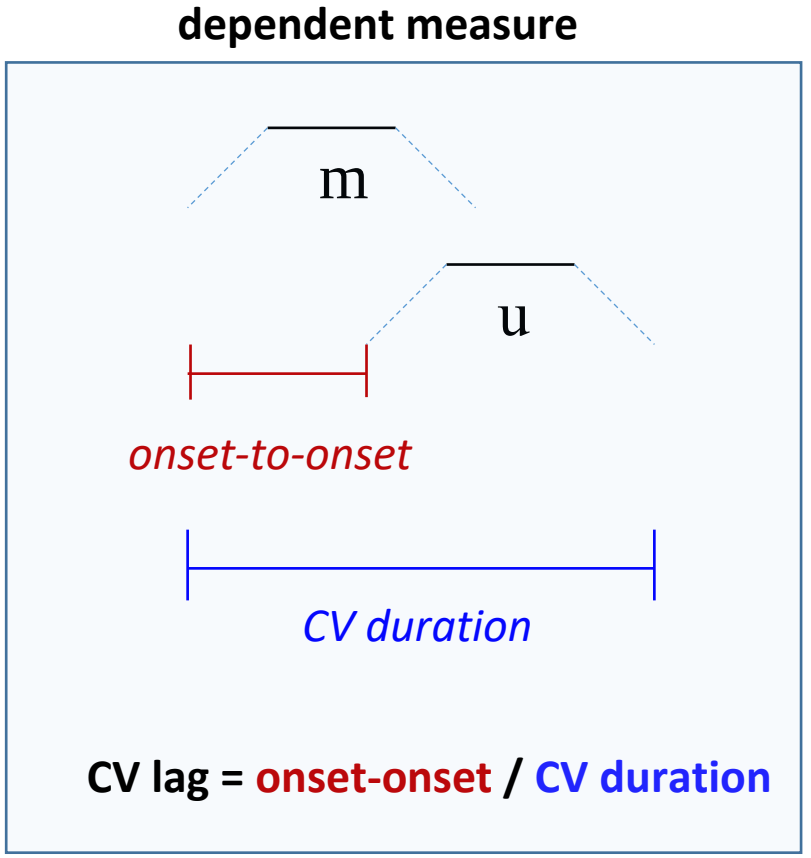
full	这一类兔子长大没有妈妈 this type of rabbit grows up without a mother	我们给它起名叫 we call it a	自 母 兔 zi4 mu3 tu4 self mother rabbit
reduced	这一类兔子喜欢看英文书 this type of rabbit enjoys reading english books	我们给它起名叫 we call it a	字母 兔 zi4mu3 tu4 letter rabbit
absent	这一类兔子喜欢看英文书 this type of rabbit enjoys reading English books	那我不知道：会写英文 but i'm not sure: does it know how to write English letters?	字 吗 他 zi4 ma ta1 letter QUES 3.SG.M

- Participants read context silently and then read aloud target words in sentences and in isolation

Methods

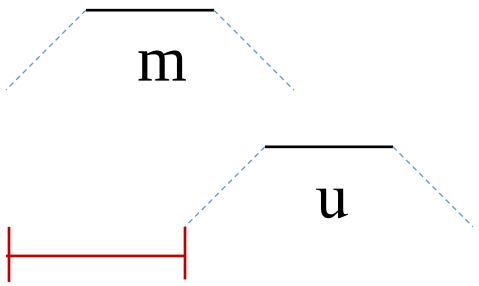
- Electromagnetic Articulography (EMA)
  - upper/lower lip sensors => closure in /m/ gesture
  - tongue dorsum sensor => retraction in /u/ gesture
- 11 participants
  - 6 female
  - ages 19-37 (mean 22;4)
  - native speakers of mandarin
- 6,798 tokens
  (2 pronunciations (sentence/isolation) x 3 conditions (full/reduced/absent) x 7 sets = 42 tokens per block; 12-19 blocks per participant)

- **Tone presence/absence** determined by **Bayesian classification** (Shaw & Kawahara 2018).



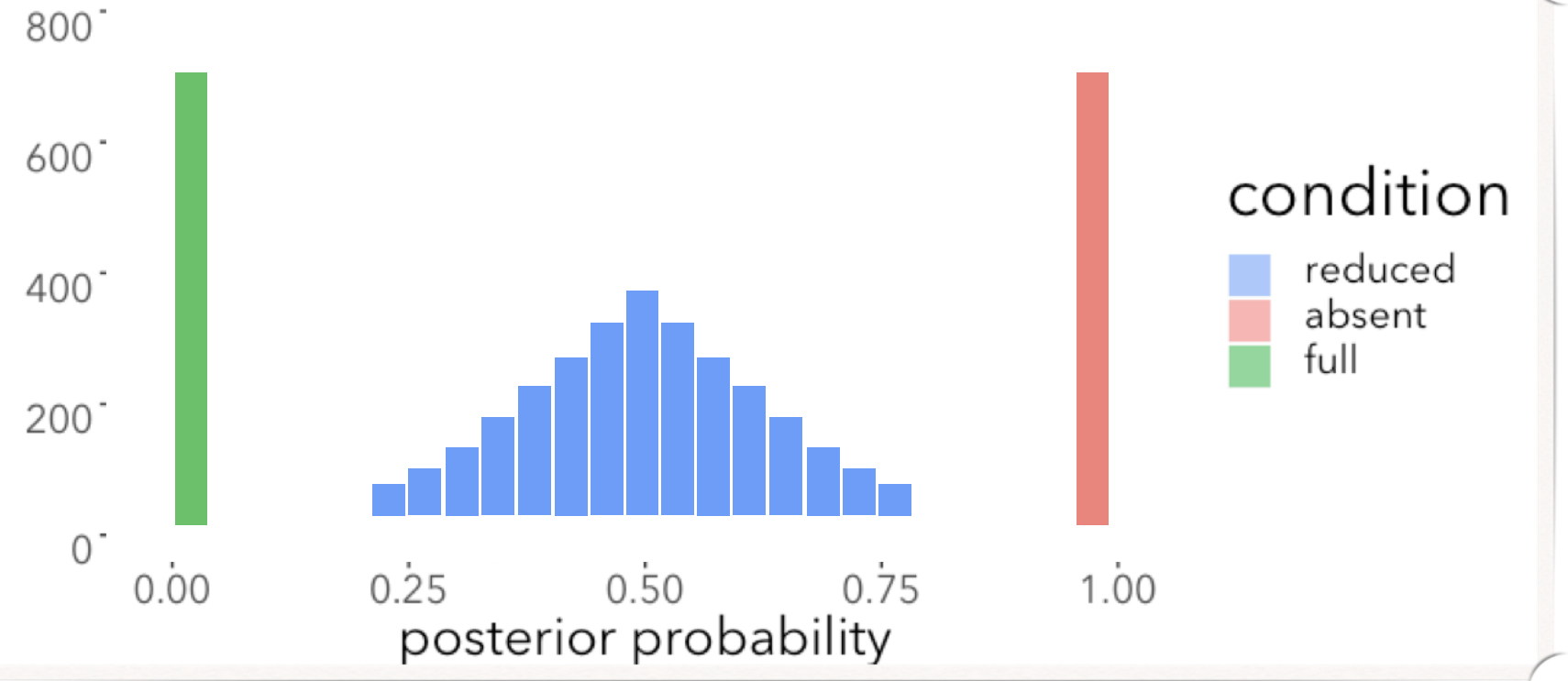
Results

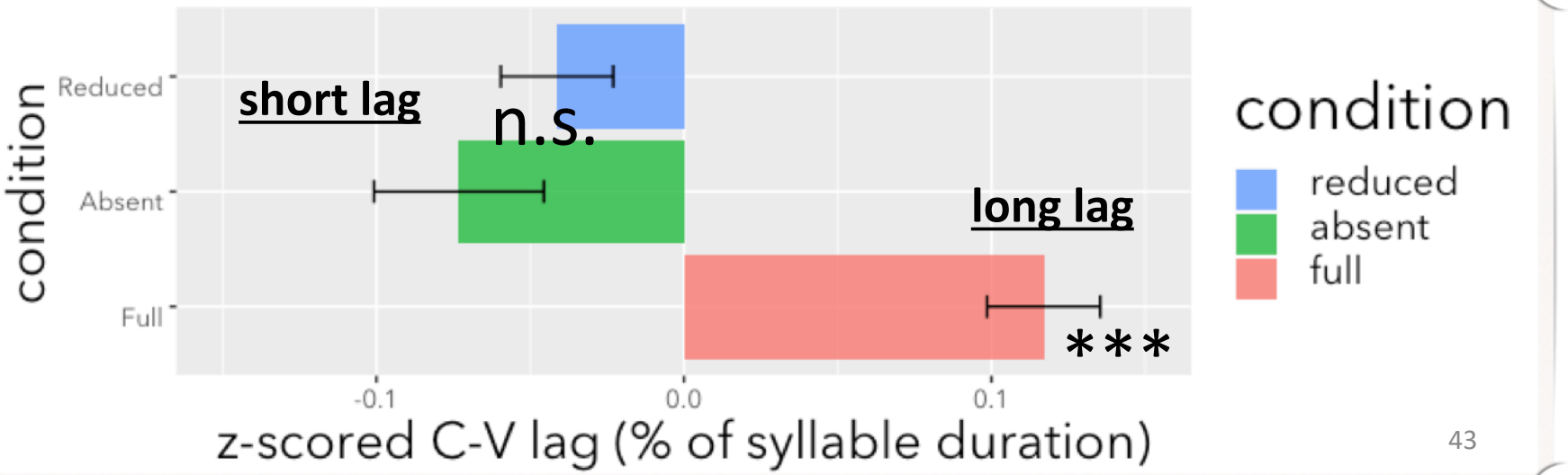
Tone presence/absence based on f0



full-tone

absent-tone (phonetic interpolation)



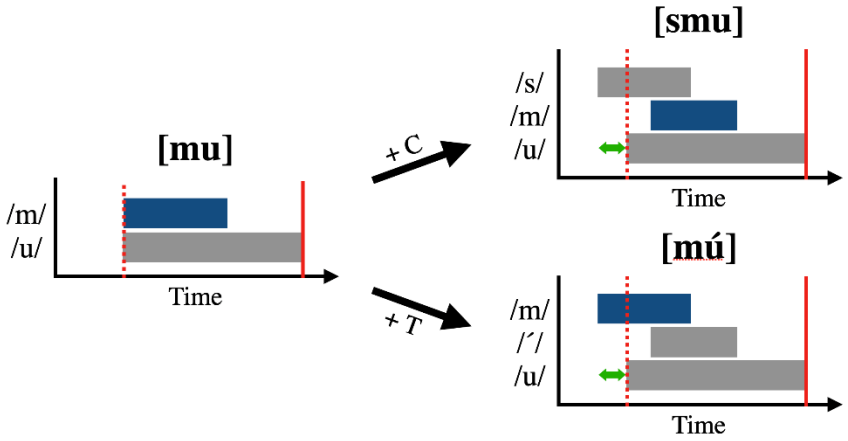


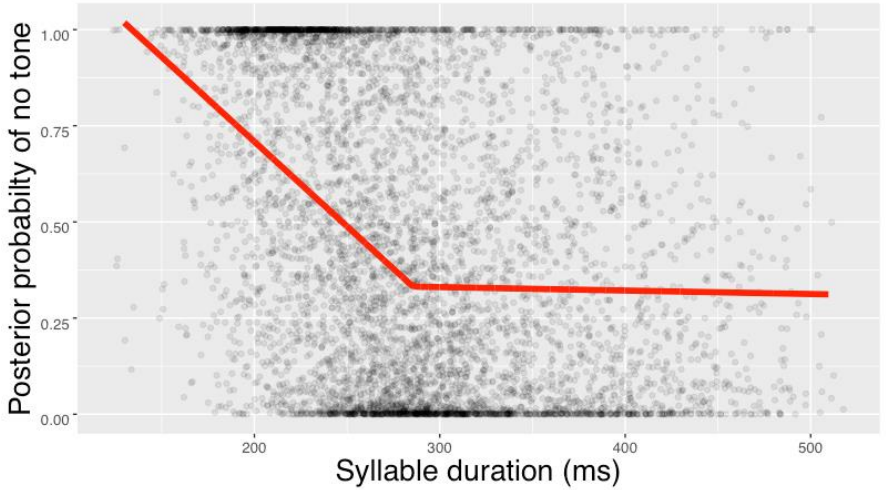
Discussion

- As expected,
  long CV lag for full tone syllables; short lag for neutral tone

- Surprisingly,
  "Reduced" syllables showed **full tone pitch trajectory but short CV lag**

- **Morpho-syntactic context triggers shift in gestures**; tone undershoot/loss follows



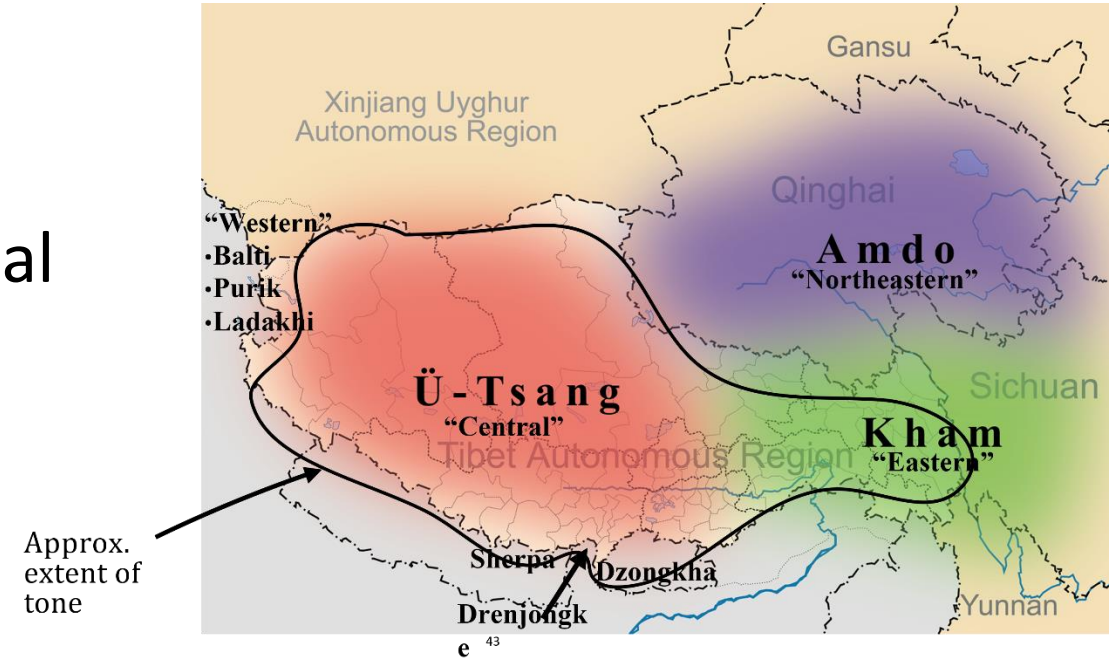


diaspora Tibetan

- Derives from a mix of Tibetan varieties some of which have lexical tone and some of which do not.

- Tonal dialects have two-way tonal contrast:
  - High tone (H)
  - Rising tone (LH)

Chris Gesissler
Yale, PhD Candidate



Methods

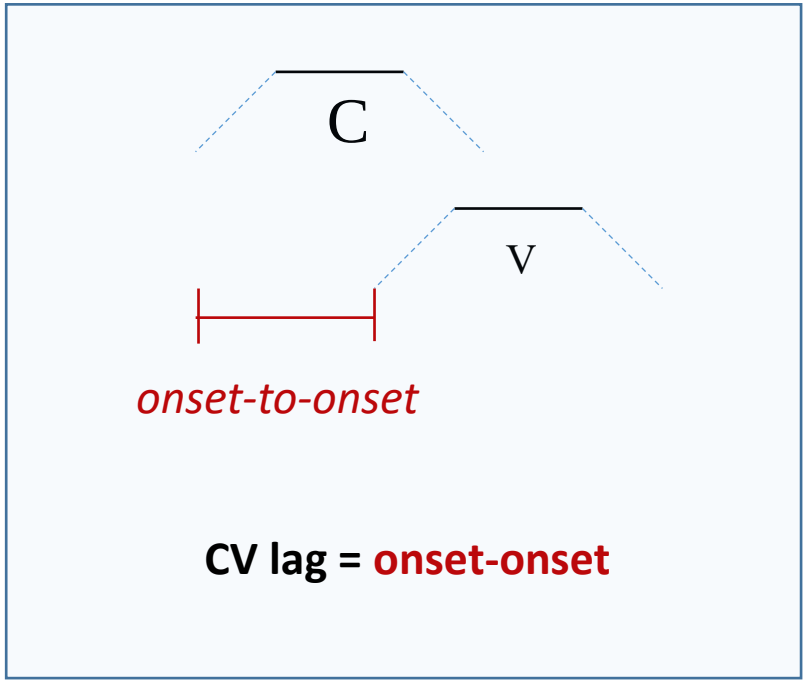
- Electromagnetic Articulography (EMA)
  - upper/lower lip sensors => closure in [p p^h m]
  - tongue dorsum sensor => retraction in [u o a]
- 6 participants
  - 4 female
  - ages 19-37
  - native speakers of diaspora Tibetan
- 3,862 tokens for analysis
  - 72 items read in carrier phrase:
  - 5-10 reps per item

ཚོག་འདི་ ___ འདུག

$\widetilde{ts^h}$ík ʈĭ ___ ʈŭk

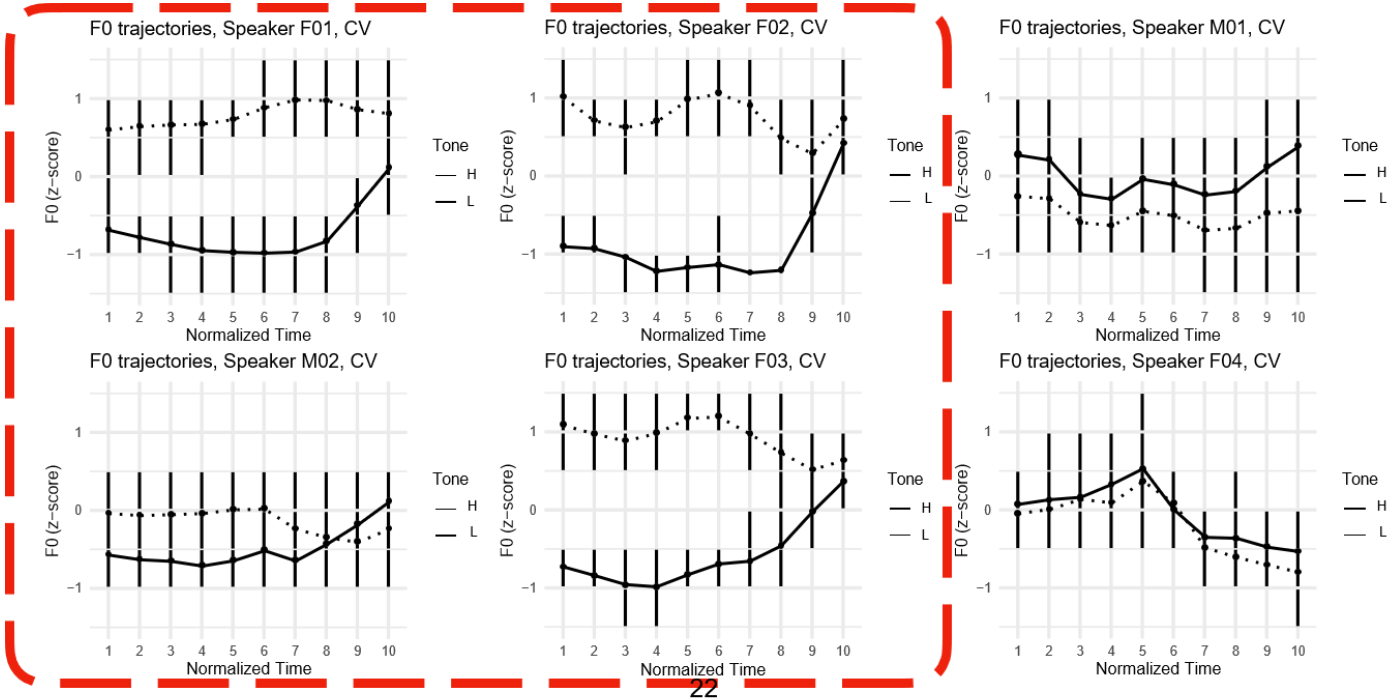
‘This word is ___.’

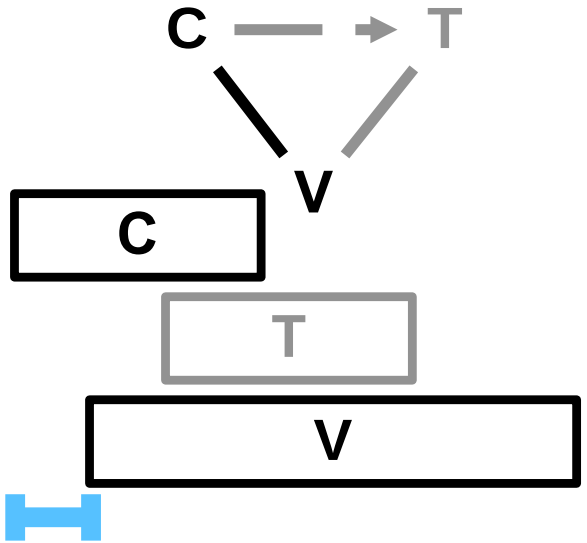
dependent measure

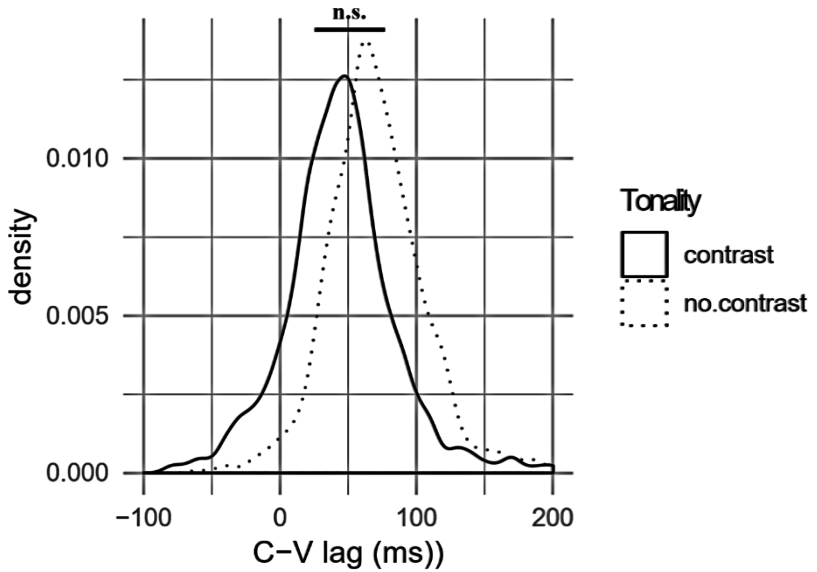


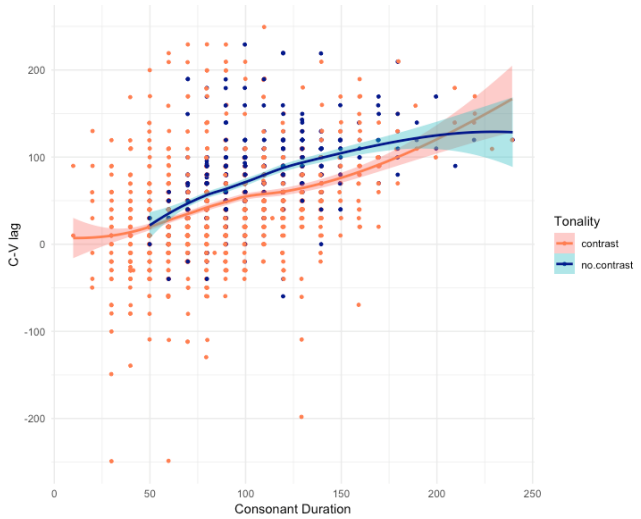
Results

- 4 speakers produce tone contrast; 2 do not
- **All 6 speakers show long CV lag**

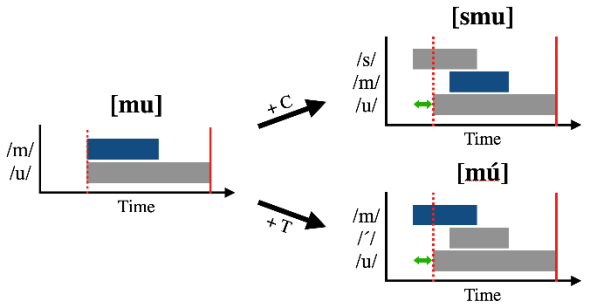




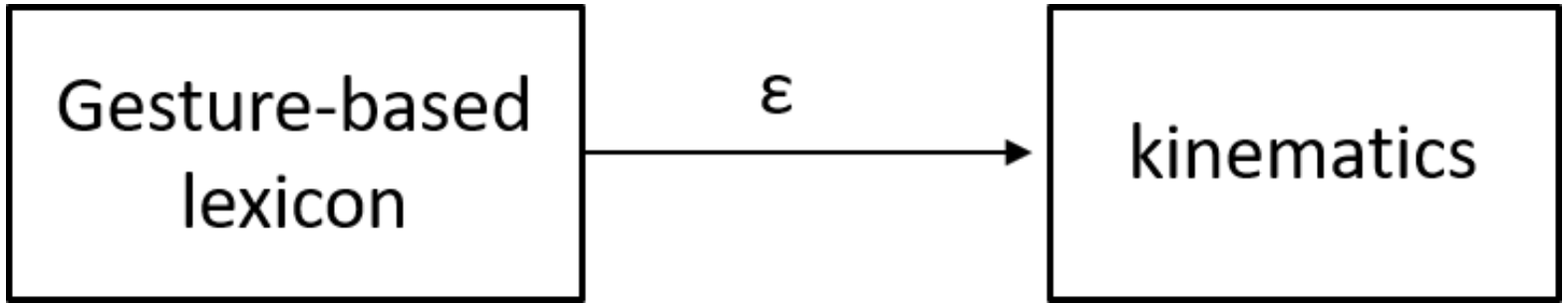


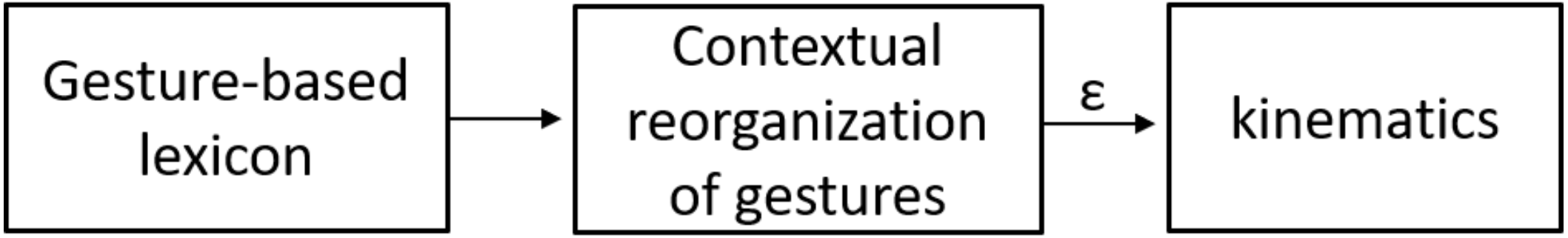


Discussion



- "Lexical tone languages" tend to have long lag C-V timing
- Even when they've lost tone (Tibetan) the timing pattern can persist in the community, indicating that it is not the presence of the tone *per se* that conditions long lag (synchronically)
- Likewise in Mandarin it is not the loss of tone that triggers synchronous timing (in "reduced" condition), but rather the synchronous timing that causes tone undershoot (and ultimate loss)





Japanese

Devoicing triggers deletion of vowel height target which triggers change in coordination

Mandarin

Context (morpho-syntactic or maybe prosodic) triggers change in gesture coordination which leads to tone undershoot

This talk

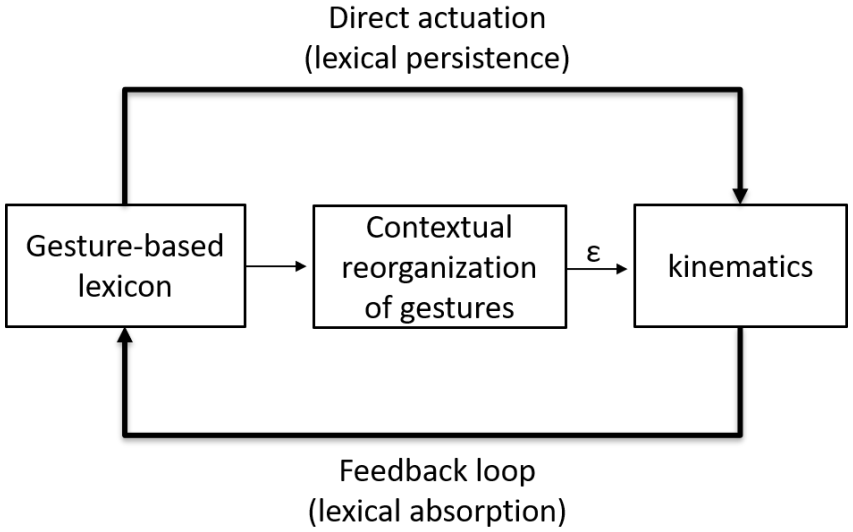
1) **Dynamic invariance:** still a good idea!
   - Gestural basis for complex segments (Russian, English)

2) Gestural coordination is **conditioned by linguistic context**
   - Gesture deletion triggers re-organization of gestural coordination (Japanese)
   - Re-organization of gestural coordination precipitates tone loss (Mandarin)
   - Tone loss proceeds without gestural re-organization (diaspora Tibetan)

3) Living lexicon: word-specific phonetics
   - **Lexical absorption**: words take on the phonetic detail of the prosodic environments in which they are typically produced (Mandarin)
   - **Lexical persistance**: phonetic resistance to structurally-conditioned pitch accent reduction (Japanese).

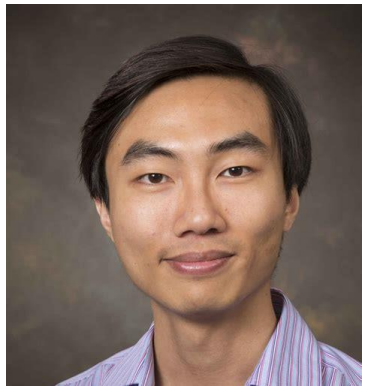
Living lexicon

- **Lexical persistence**: resistance to structurally conditioned reduction (Kawahara, Shaw, Ishihara, 2021)



- **Lexical absorption**: lexical items take on the phonetic detail of the prosodic environments in which they are typically produced (Tang & Shaw, 2021)

Lexical absorption: Prosody leaks in the lexicon



Kevin Tang
U. of Florida

Tang, K., & Shaw, J. A. (2021). Prosody leaks into the memories of words. *Cognition*, *210*, 104601.

Contents lists available at ScienceDirect

Cognition

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

Prosody leaks into the memories of words

Kevin Tang [a,*], Jason A. Shaw [b]

[a] Department of Linguistics, University of Florida, Gainesville, FL 32611-5454, USA
[b] Department of Linguistics, Yale University, New Haven, CT 06520, USA

ARTICLE INFO

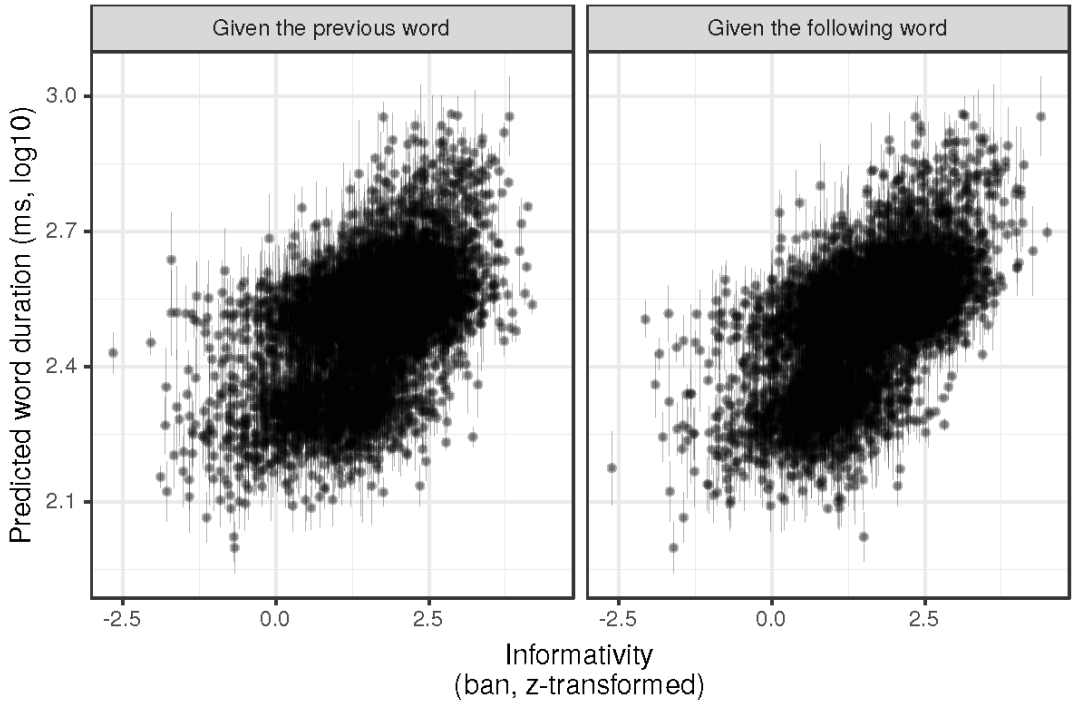
Keywords:
Probabilistic reduction
Prosodic prominence
Lexical representation
Speech production
Predictability
Informativity

ABSTRACT

The average predictability (aka informativity) of a word in context has been shown to condition word duration (Seyfarth, 2014). All else being equal, words that tend to occur in more predictable environments are shorter than words that tend to occur in less predictable environments. One account of the informativity effect on duration is that the acoustic details of probabilistic reduction are stored as part of a word's mental representation. Other research has argued that predictability effects are tied to prosodic structure in integral ways. With the aim of assessing a potential prosodic basis for informativity effects in speech production, this study extends past work in two directions; it investigated informativity effects in another large language, Mandarin Chinese, and broadened the study beyond word duration to additional acoustic dimensions, pitch and intensity, known to index prosodic prominence. The acoustic information of content words was extracted from a large telephone conversation speech corpus with over 400,000 tokens and 6000 word types spoken by 1655 individuals and analyzed for the effect of informativity using frequency statistics estimated from a 431 million word subtitle corpus. Results indicated that words with low informativity have shorter durations, replicating the effect found in English. In addition, informativity had significant effects on maximum pitch and intensity, two phonetic dimensions related to prosodic prominence. Extending this interpretation, these results suggest that predictability is closely linked to prosodic prominence, and that the lexical representation of a word includes phonetic details associated with its average prosodic prominence in discourse. In other words, the lexicon absorbs prosodic influences on speech production.

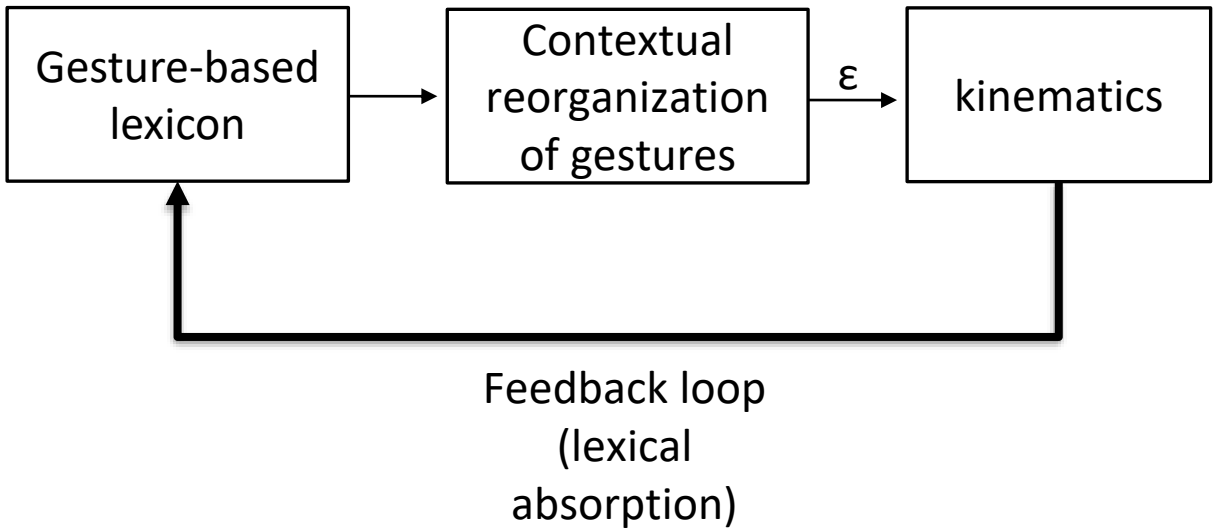
The argument

- **Prosodic prominence**, as dictated by contextual predictability, influences *word duration*, *pitch*, and *intensity* in Mandarin Chinese.
- A word's **informativity** (average contextual predictability) reflects aggregate influences of prosodic prominence on lexical representations.
- **Informative words** have longer duration, higher pitch, greater intensity, *even in prosodically weak positions*, because they tend to occur in prosodically prominent positions.
- Hence, prosodic prominence leaks into the lexicon.



Dependent variable (either Duration, Maximum intensity, Intensity range, Maximum pitch or Pitch range) ~ Frequency + Forward predictability + Backward predictability + ***Forward informativity*** *+* ***Backward informativity*** *+ Word length + Preceding disfluency + Following disfluency + Preceding pause duration + Following pause duration + Preceding speech rate + Following speech rate + Previous self-mention + Previous cross-speaker mention + Age + Gender + Syntactic category + (1 | Word type) + (1 | Tone sequence) + (1 + Forward informativity + Backward informativity | Speaker)*

Living lexicon: feedback loop



Lexical persistence: Failure of prosodic reduction

Shigeto Kawahara
Keio University

Shin Ishihara
Lund University

Kawahara, S., Shaw, J. A., & Ishihara, S. (2021). Assessing the prosodic licensing of wh-in-situ in Japanese. *Natural Language & Linguistic Theory*, 1-20.

Nat Lang Linguist Theory
https://doi.org/10.1007/s11049-021-09504-3

Check for updates

Assessing the prosodic licensing of wh-in-situ in Japanese

A computational-experimental approach

Shigeto Kawahara[1] · Jason A. Shaw[2] · Shinichiro Ishihara[3]

Received: 8 May 2020 / Accepted: 22 January 2021

Abstract The relationship between syntactic structure and prosodic structure has received increased theoretical attention in recent years. Richards (2010) proposes that Japanese allows wh-elements to stay in situ because of a certain aspect of its prosodic system. Specifically, in contrast to some other languages like English, Japanese can prosodically group wh-elements together with their licensers. This prosodic grouping is phonetically signaled by eradication or reduction of the lexical pitch accents of intervening words. In this theory, a question still remains as to whether each syntactic derivation is checked against its phonetic realization, or what allows Japanese wh-elements to stay in situ is more abstract phonological prosodic structure, whose phonetic manifestations can potentially be variable. This paper reports an experiment which addressed this question, by testing whether there is eradication or reduction of lexical pitch accents based on the detailed analysis of F0 contours. Our analysis makes use of a computational toolkit that allows us to assess the presence of tonal targets on a token-by-token basis. The results demonstrate that almost all speakers produce some wh-sentences which show reduction or eradication of the lexical pitch accents, as well as some that do not. Those tokens that show reduction or eradication directly support the prediction of Richards' (2010) theory. The variability observed in the results suggests that the property of Japanese that allows their wh-elements to stay in situ must be abstract, phonological prosodic structure, whose phonetic realiza-

A puzzle in syntactic theory

- Some languages (e.g. Tagalog) show **overt wh-movement**; some languages allow their **wh-elements to stay in situ** (e.g. Japanese).

- **Minimalist Syntax**: those that move overtly have a strong (uninterpretable) feature that needs to be checked. Japanese wh-elements on the other hand have a weak (interpretable) feature.

- **Richards (2010)** attempts to derive this difference from an independently observable difference. (Further developed in **Richards 2016**).

Richards, N. (2010). *Uttering trees* (Vol. 56). MIT Press.

Richards, N. (2016). *Contiguity theory* (Vol. 73). MIT Press.

Richards' (2010) proposal in a nutshell

- All languages attempt: "to create a prosodic structure for *wh*-questions in which the *wh*-phrase and corresponding complementizer are separated by as few prosodic boundaries as possible" (p. 145).

- Japanese has a prosodic means to group the wh-phrase and its complementizer, and hence does not need to resort to overt wh-movement.

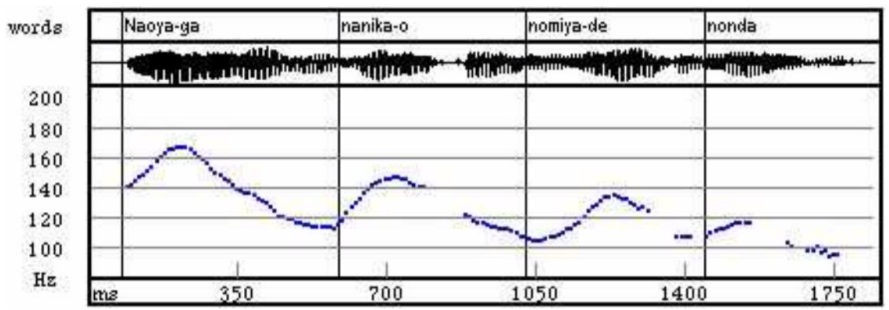
- Tagalog on the other hand does not have that prosodic strategy, so its wh-elements needs to move overtly.

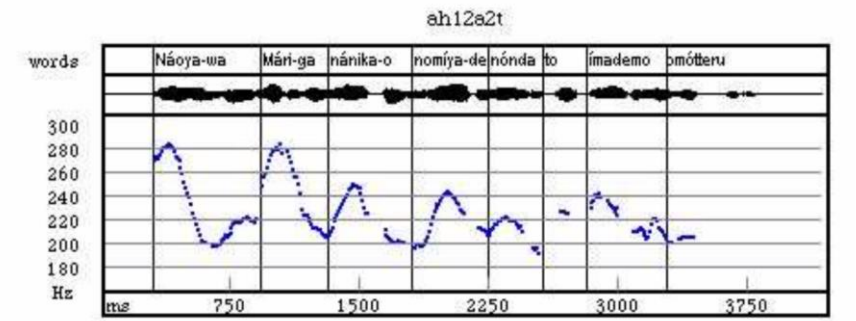
One source of inspiration for Richards (2010)

- Post-wh accent in Japanese is *eradicated* (Deguchi & Kitagawa 2002). Sample pitch tracks from Ishihara (2001).

declarative

(28a): Non-interrogative sentence

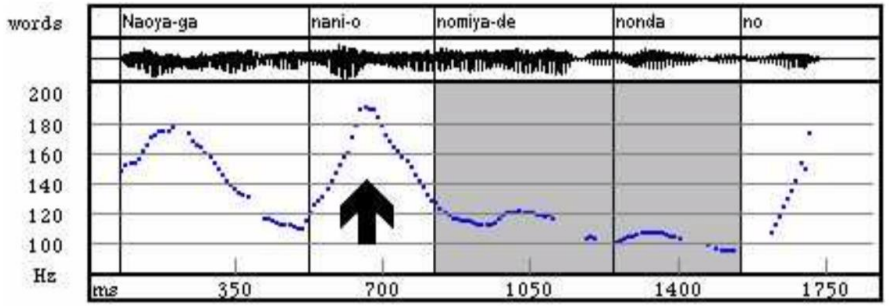




wh-sentence

(28b): *Wh*-question

wh eradication?



(29b): *Wh*-question

wh eradication?

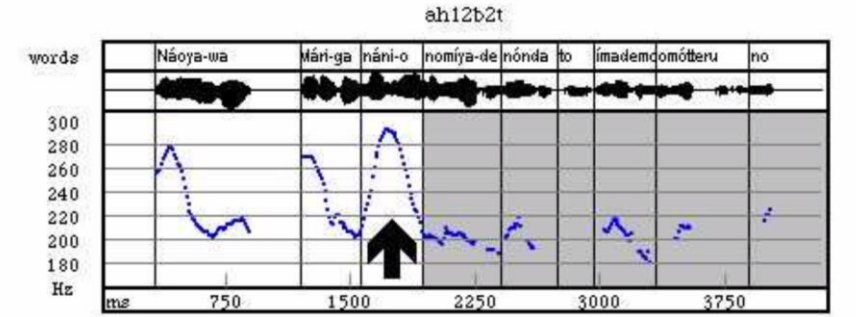
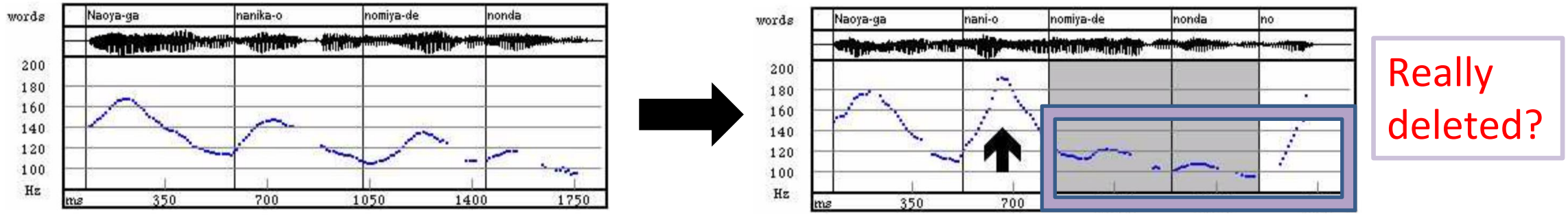


Figure 3-1: Single *wh*-question

Deguchi & Kitagawa (p.74)

“Another important prosodic effect of focus pointed out by Ishihara (2000) (extending the original observation by Ladd (1996)) is that an emphatic accent is accompanied by what we label as "**eradication**" of lexical accents. That is, when one or more of lexical accents follow an emphatic accent, their H tones (H*) are all suppressed. As a result, the lowest pitch induced by the emphatic accent is inherited and prolonged with further gradual declination up to the right boundary of some clausal structure”



Method

- **Tone presence/absence** determined by **Bayesian classification** (Shaw & Kawahara 2018).
- Nine Tokyo Japanese speakers (4 female)
- 6 items per condition; 2 repetitions each (24 tokens per subject)

(1) Control sentences: $Word_1$ $Word_{2[-wh]}$ $Word_3$ $Word_4$ Verb

(2) Test sentences: $Word_1$ $Word_{2[+wh]}$ $Word_3$ $Word_4$ Verb

(1) 丸山は$_1$ エルメスの$_{[-wh]2}$ 襟巻きに$_3$ 飲み物を$_4$ こぼしました。
Maruyama-TOP Hermes-GEN scarf-DAT drink-ACC spilled

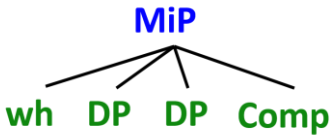
(2)丸山は$_1$ どの人の$_{[+wh]\ 2}$ 襟巻きに$_3$ 飲み物を$_4$ こぼしましたか？
Maruyama-TOP Who scarf-DAT drink-ACC spilled-Q

Tone eradication?

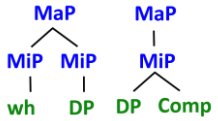
Ishihara, Shinichiro. 2011. "Focus Prosody in Tokyo Japanese Wh-Questions with Lexical Unaccented Wh-Phrases." *Proceedings of ICPhS* XVII: 946–49.

All speakers

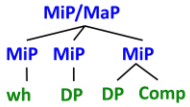
Clear tokens of tone eradication

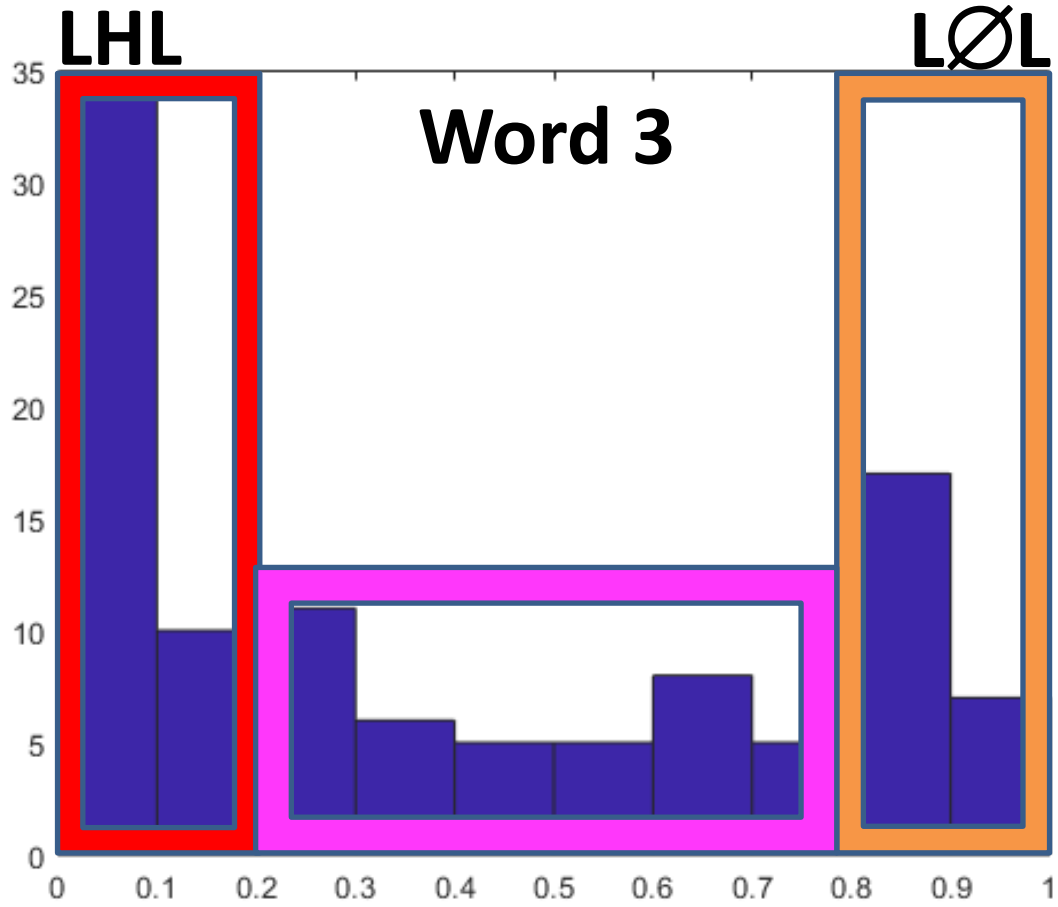


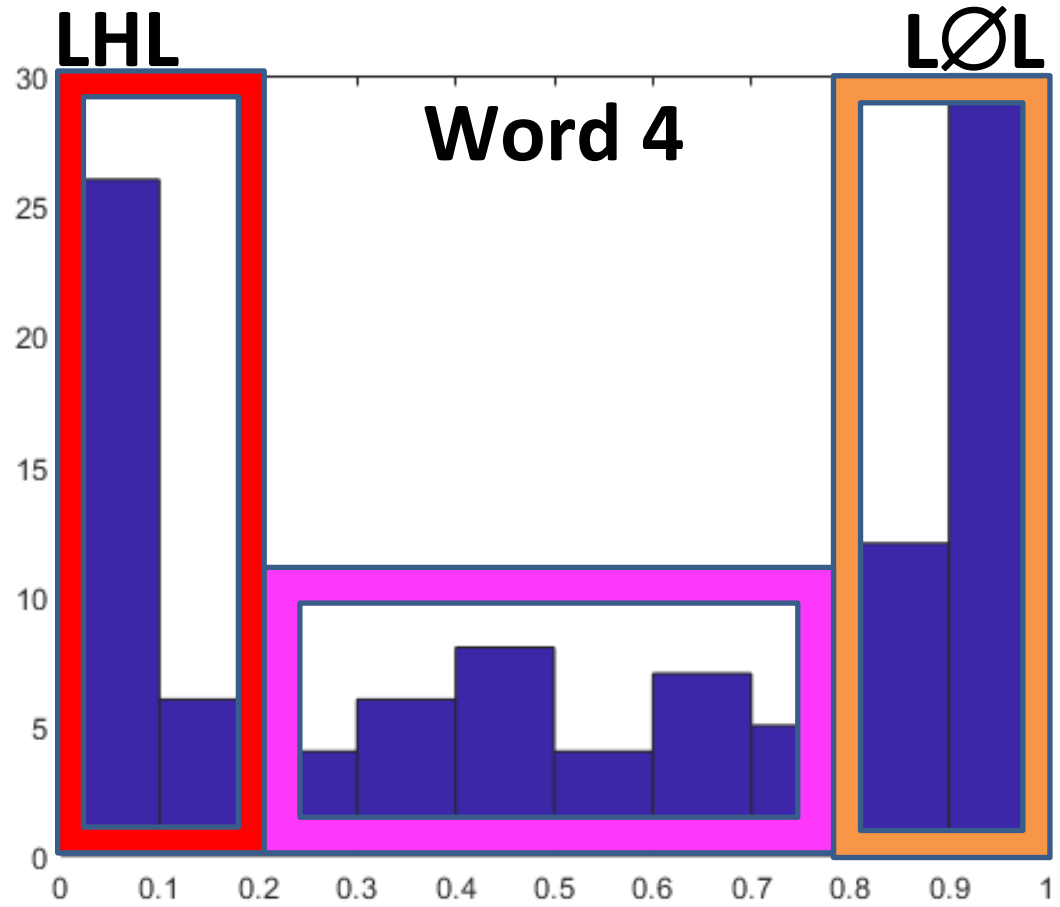
Full tone (same as [-wh])



Reduced tone



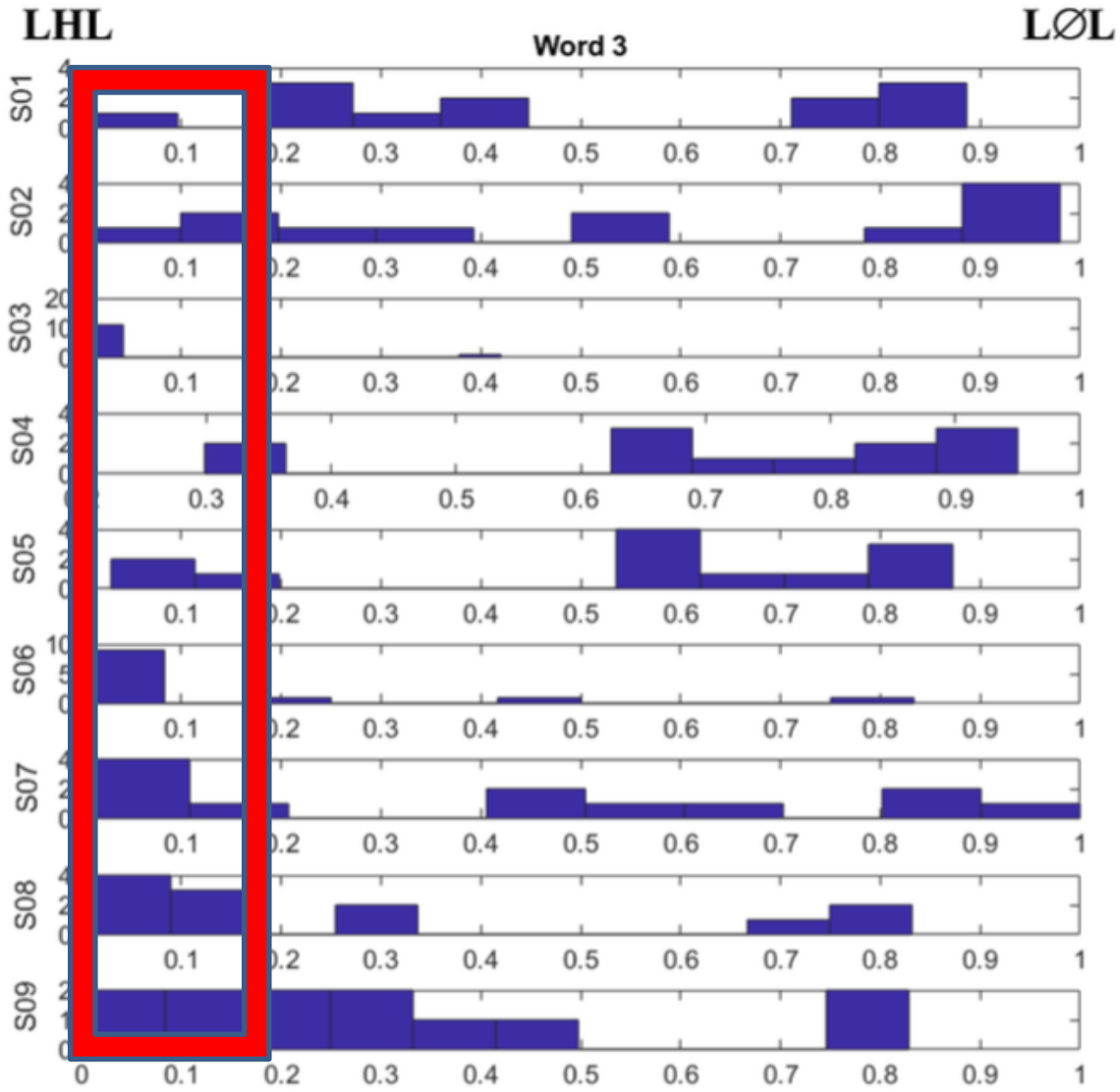


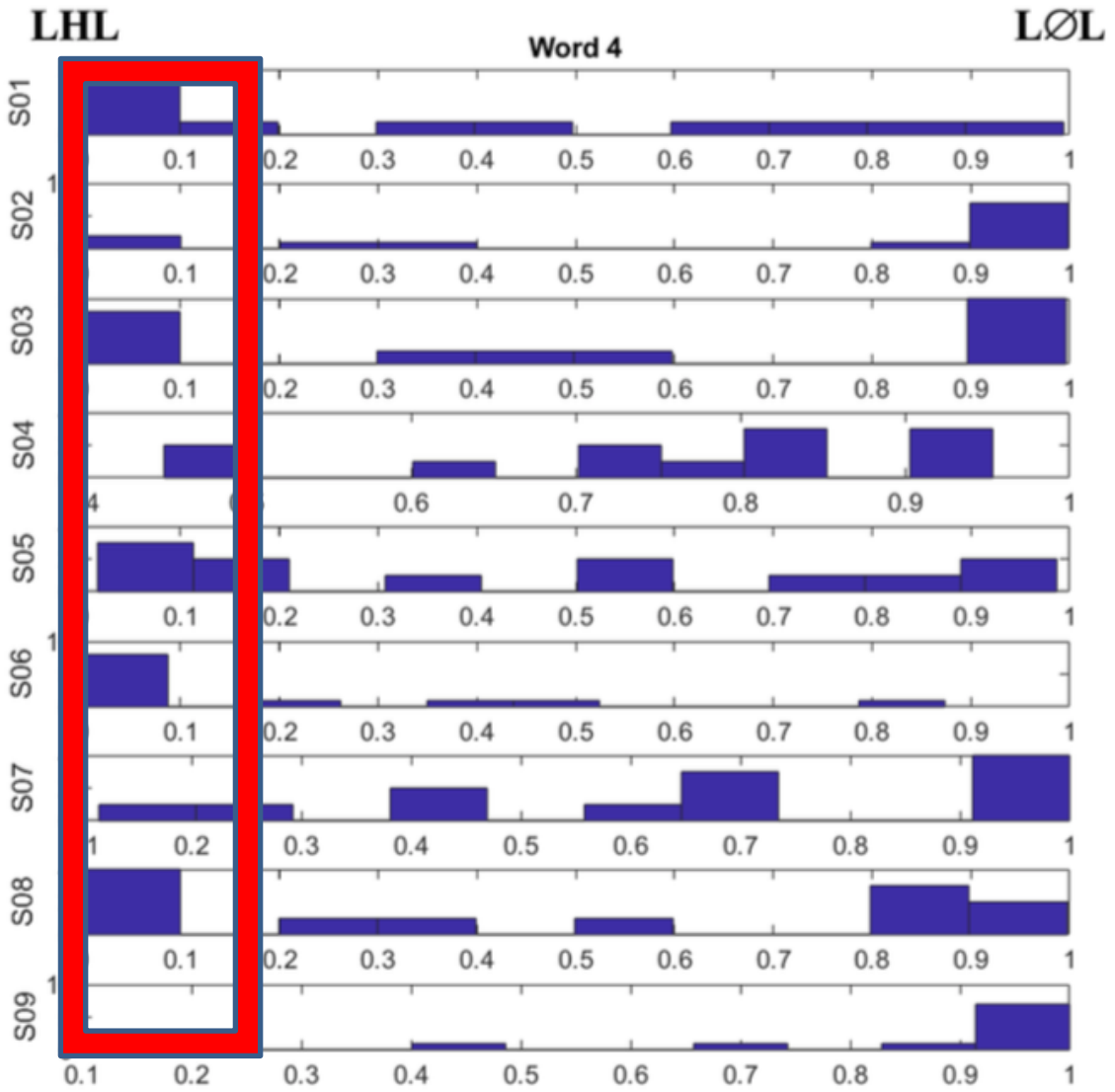


Posterior probability of targetlessness

Results by speaker

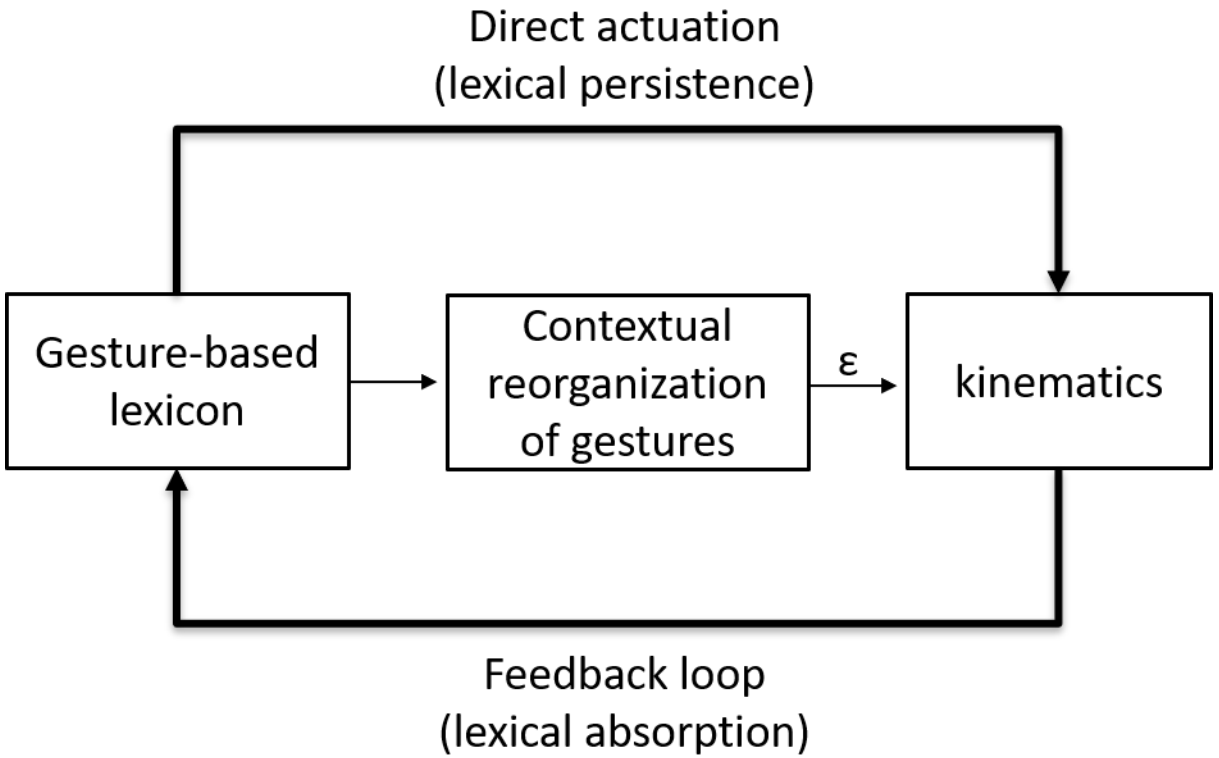
Every speaker produces some tokens without reduction





Living lexicon: direct actuation

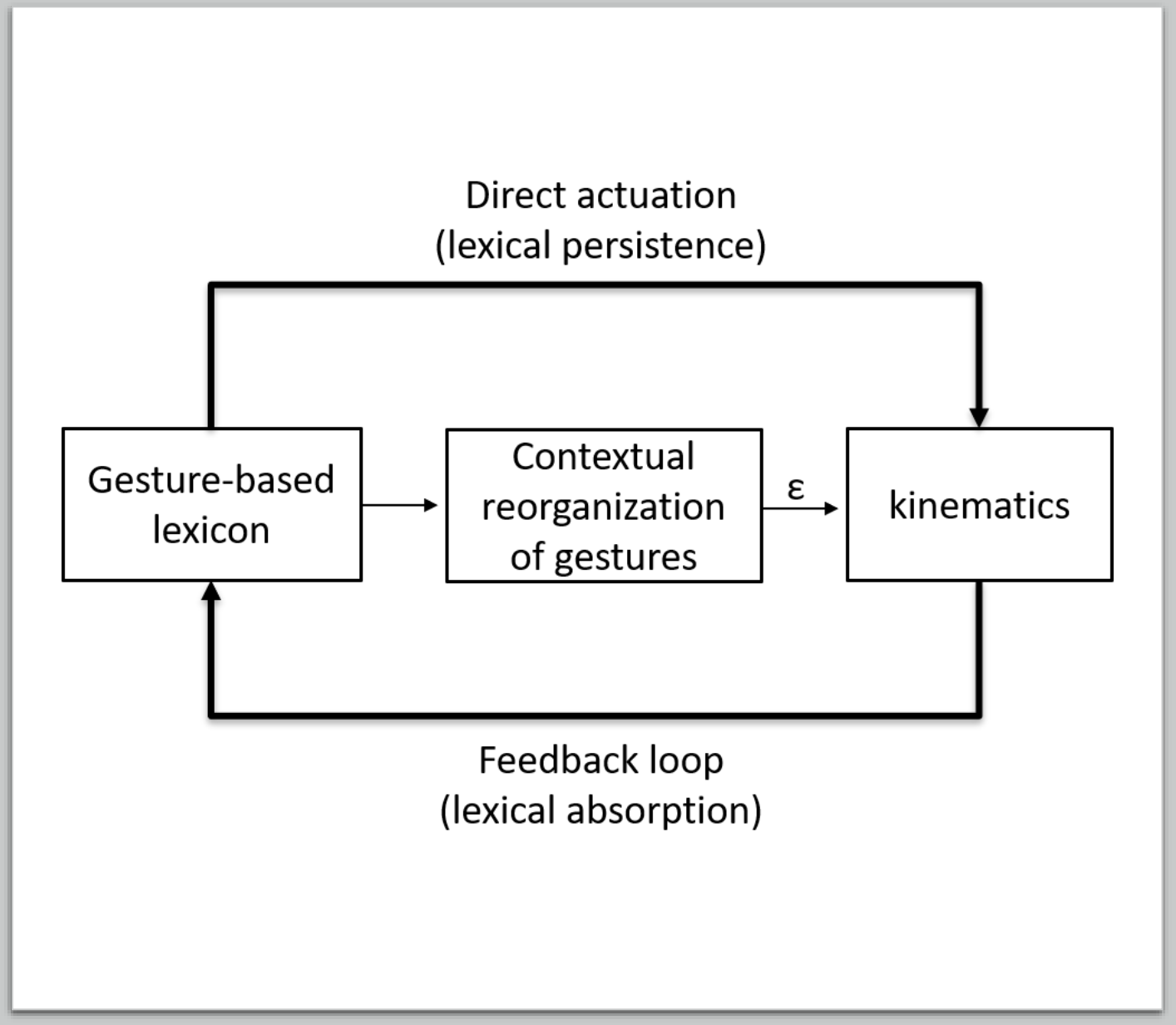
Japanese
Prosodic context conditions tone eradication, but lexically specified pitch accents can still get through.



General discussion

- Different gestural coordination patterns can be distinguished in the kinematics because they structure variation in specific ways (**dynamic invariance**), but there's more…
- **Context**, including prosodic context, conditions gestural reorganization and can **feedback** into the lexicon.
- Contextual factors can also be by-passed, c.f. motor program reuse.

Gestural coordination in the living lexicon of spoken words



Thank you!

EXTRA SLIDES

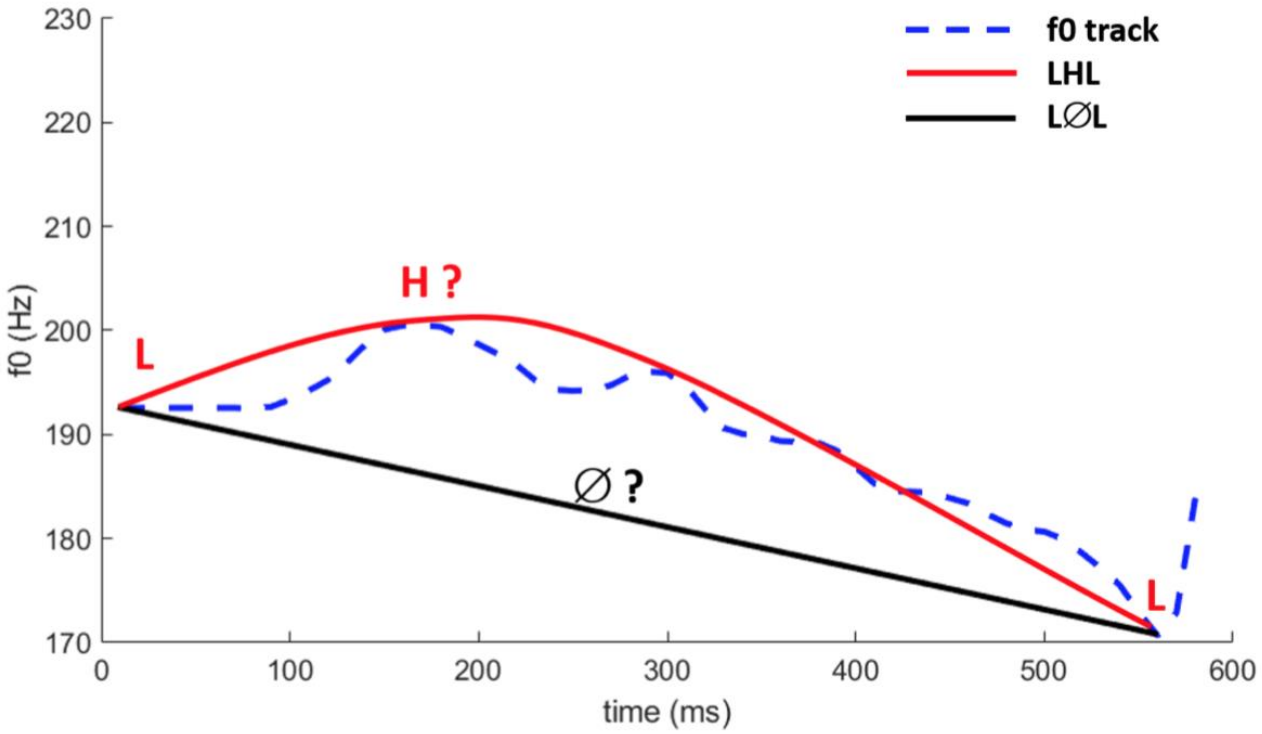
Approach

Which phonological structure is responsible for the phonetic data?

- Setup stochastic generators of f0 based on competing phonological hypotheses:
  H_1: LHL
  H_2: L∅L

- Use stochastic generative model to assign probabilities of phonological hypotheses to phonetic data.

- Allows for token-by-token analysis of f0 contours



Step 1: Discrete Cosine Transform (DCT)

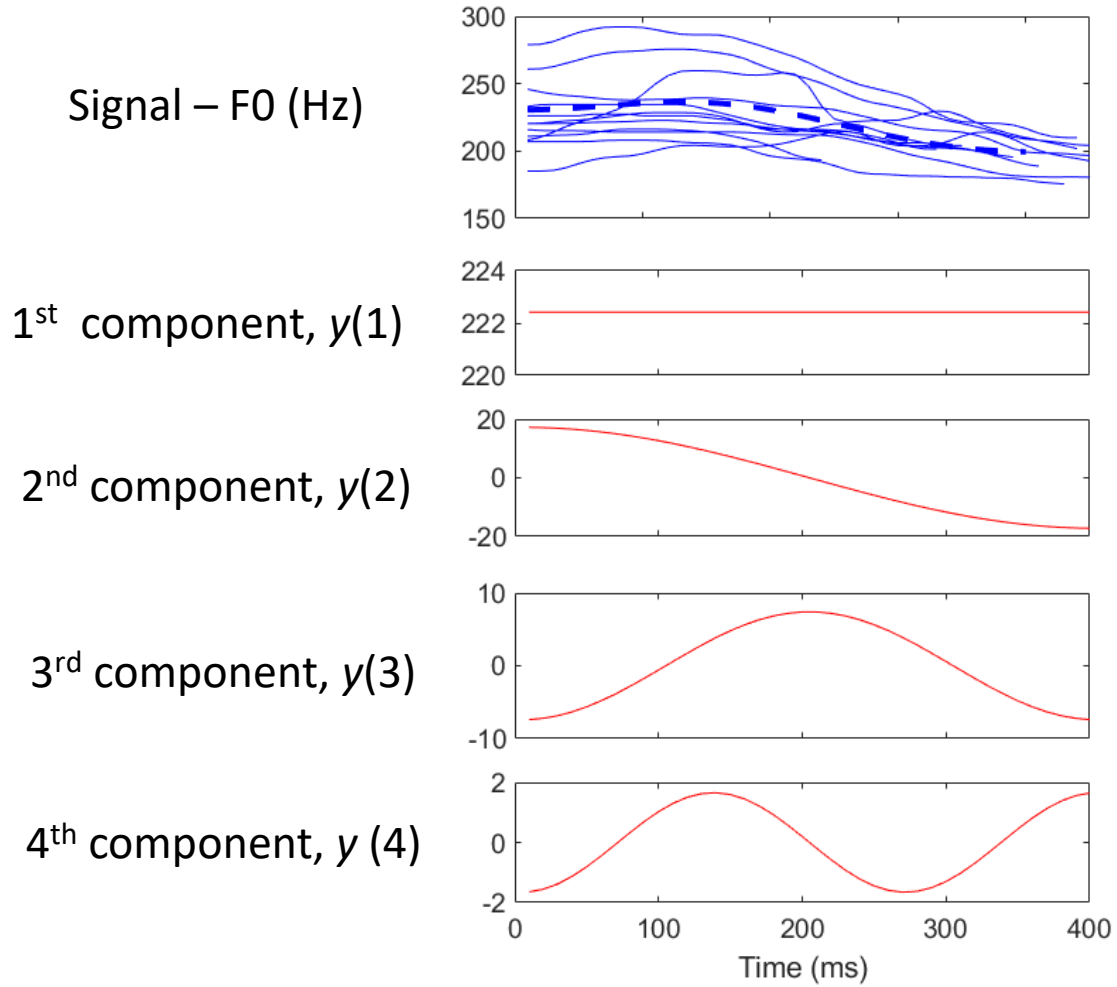
Represent f0 trajectory as the sum of Cosines:

$$y(k) = w(k) \sum_{n=1}^{L} x(n) \cos\left(\frac{\pi(2n-1)(k-1)}{2L}\right)$$

$$k = 1,2, \dots L$$

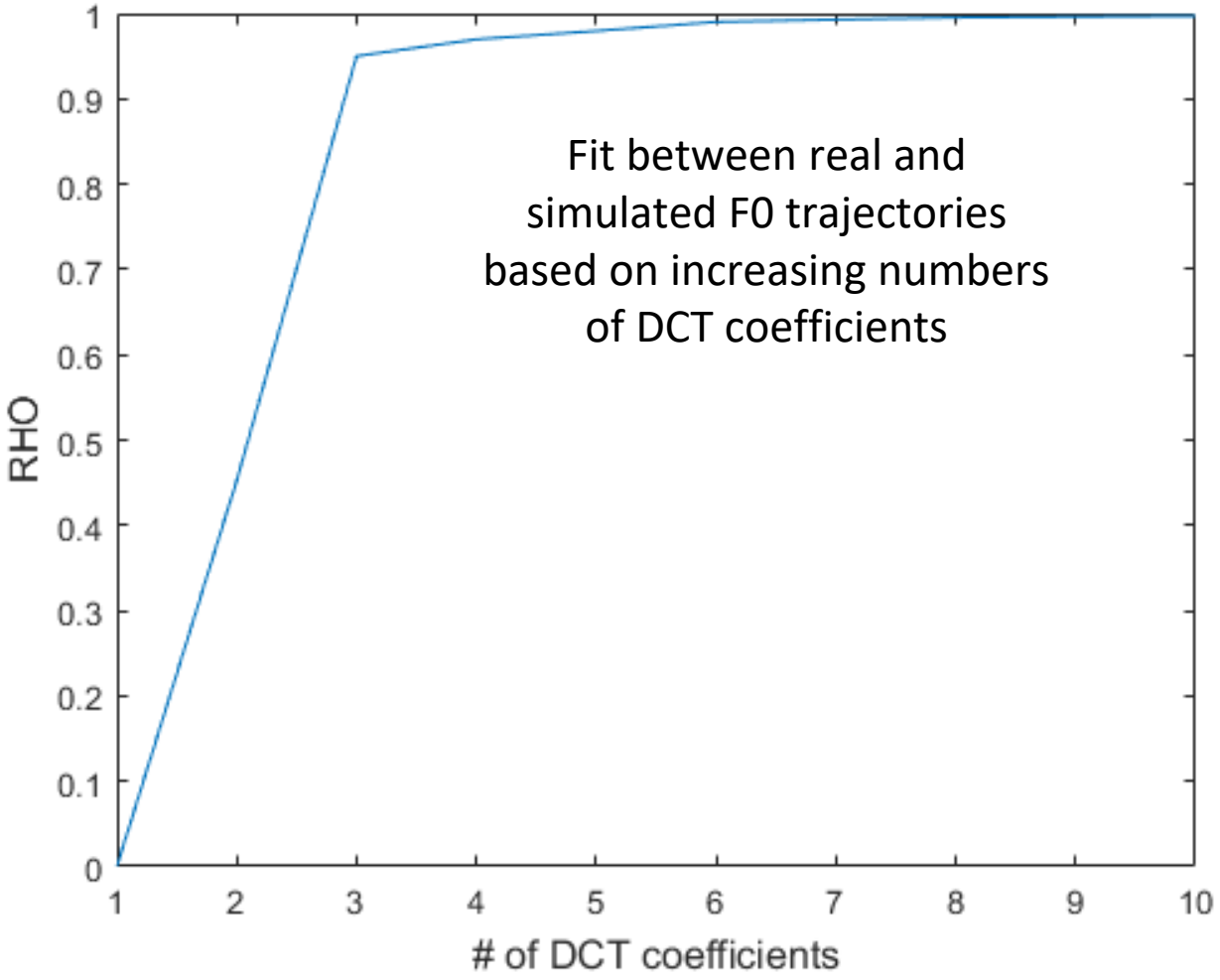
Where *L* is the number of data samples and ***x(n)*** **is the trajectory to be modelled** and:

$$w(k) = \begin{cases} \frac{1}{\sqrt{L}} & k = 1 \\ \sqrt{\frac{2}{L}} & 2 \le k \le L \end{cases}$$

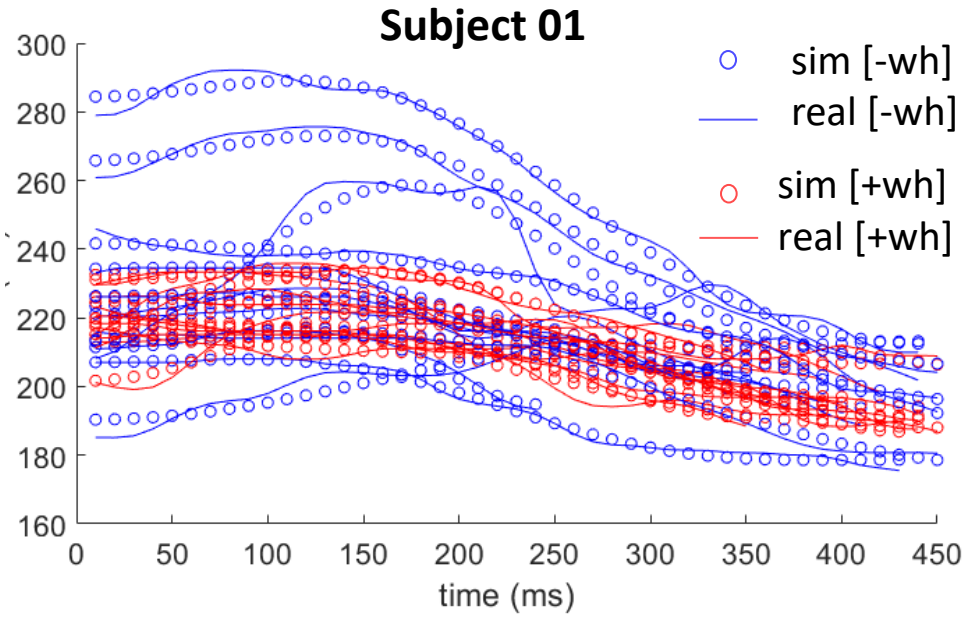


Shaw, J. A., & Kawahara, S. (2018). Assessing surface phonological specification through simulation and classification of phonetic trajectories. *Phonology, 35*(3), 481-522. doi:10.1017/S0952675718000131

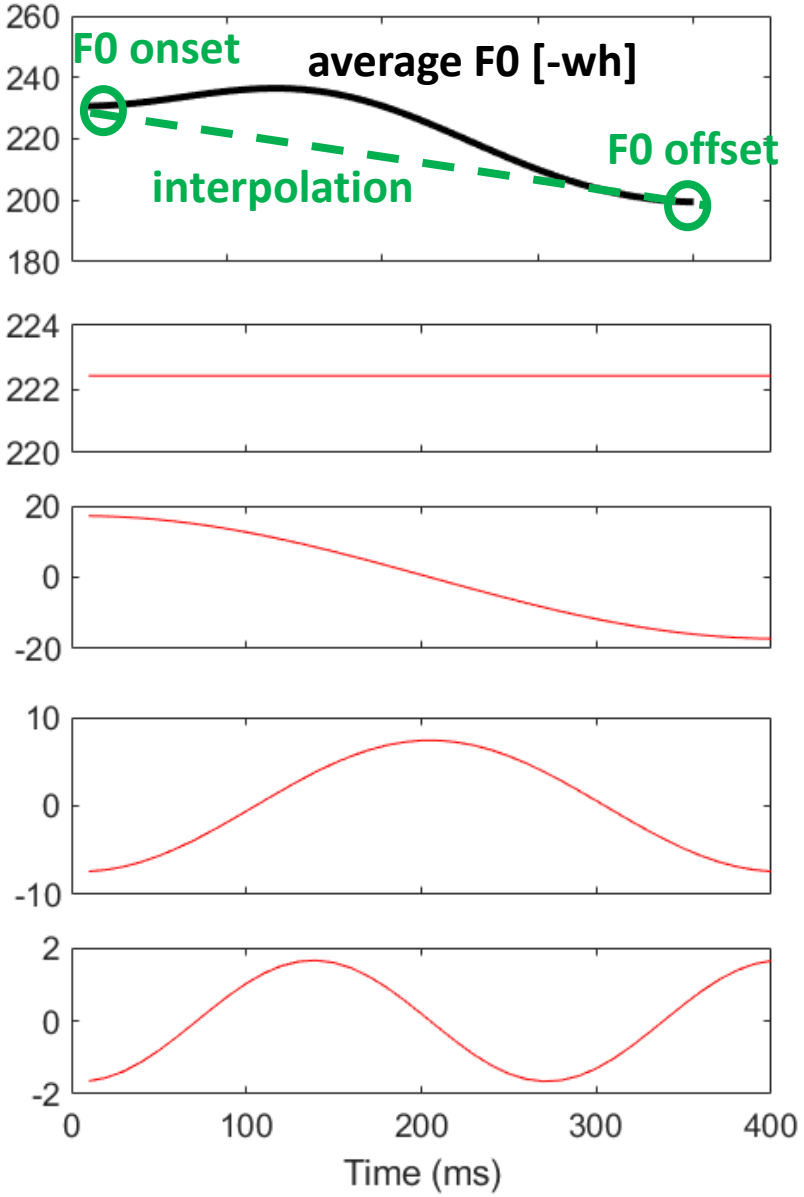
Fit between real and simulated F0 using iDCT



- Simulations from 4 DCT components explain > 90% of variance for all 9 speakers



Step 2: F0 of L∅L (the noisy null)



Simulate F0 trajectories from DCT components:

Interpolation trajectory $y(k) \sim \mathrm{N}(\mu(k), \sigma(k))$

Target present [-Wh] $y(k) \sim \mathrm{N}(\mu(k), \sigma(k))$

$$x(n) = \sum_{n=1}^{L} w(k)y(k)\cos\left(\frac{\pi(2n-1)(k-1)}{2L}\right)$$

$$n = 1,2,\ldots L$$

Where *L* is the number of data samples and *x(n)* the trajectory to be simulated and:

$$w(k) = \begin{cases} \frac{1}{\sqrt{L}} & k = 1 \\ \sqrt{\frac{2}{L}} & 2 \le k \le L \end{cases}$$

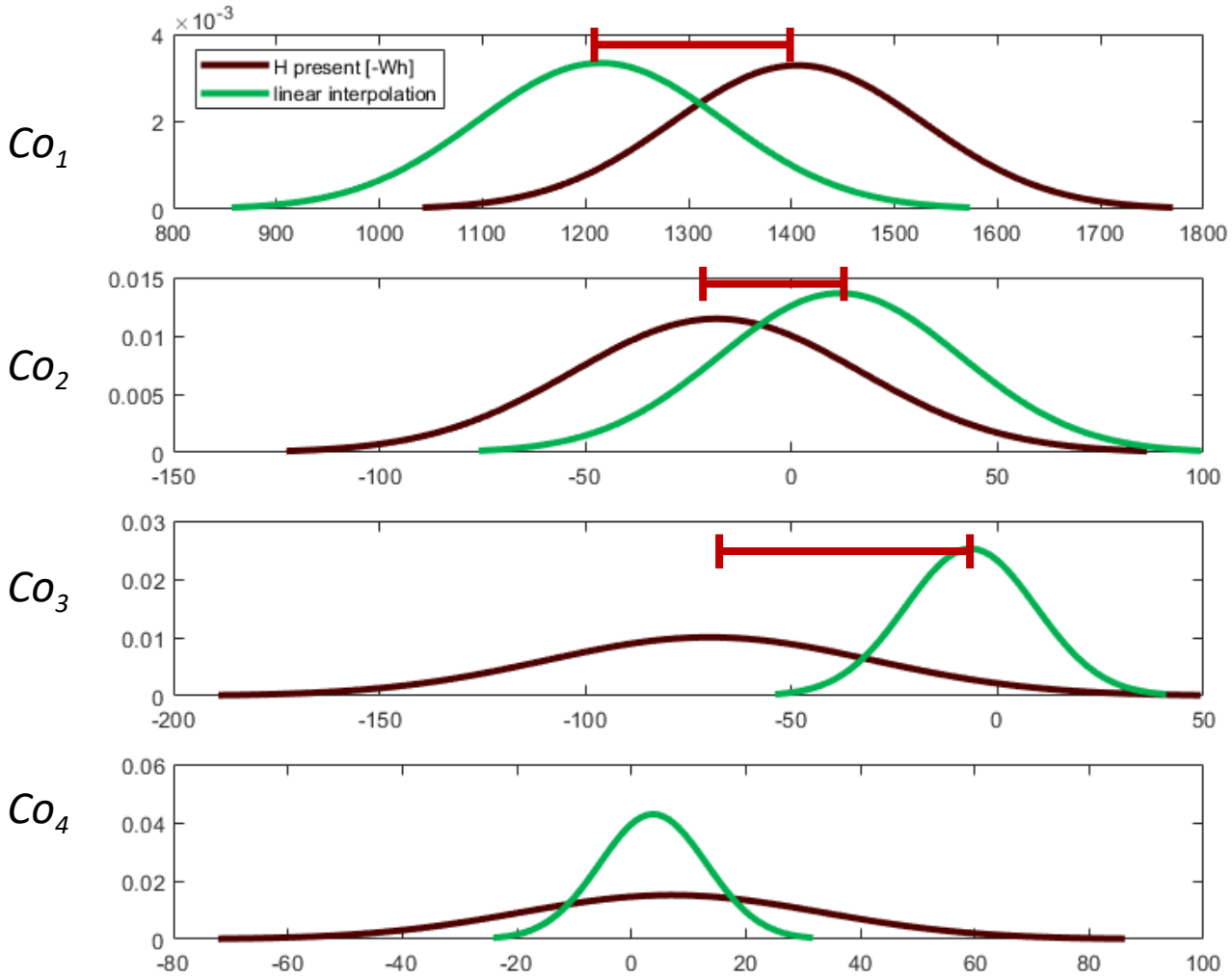
Step 3: Bayesian classifier

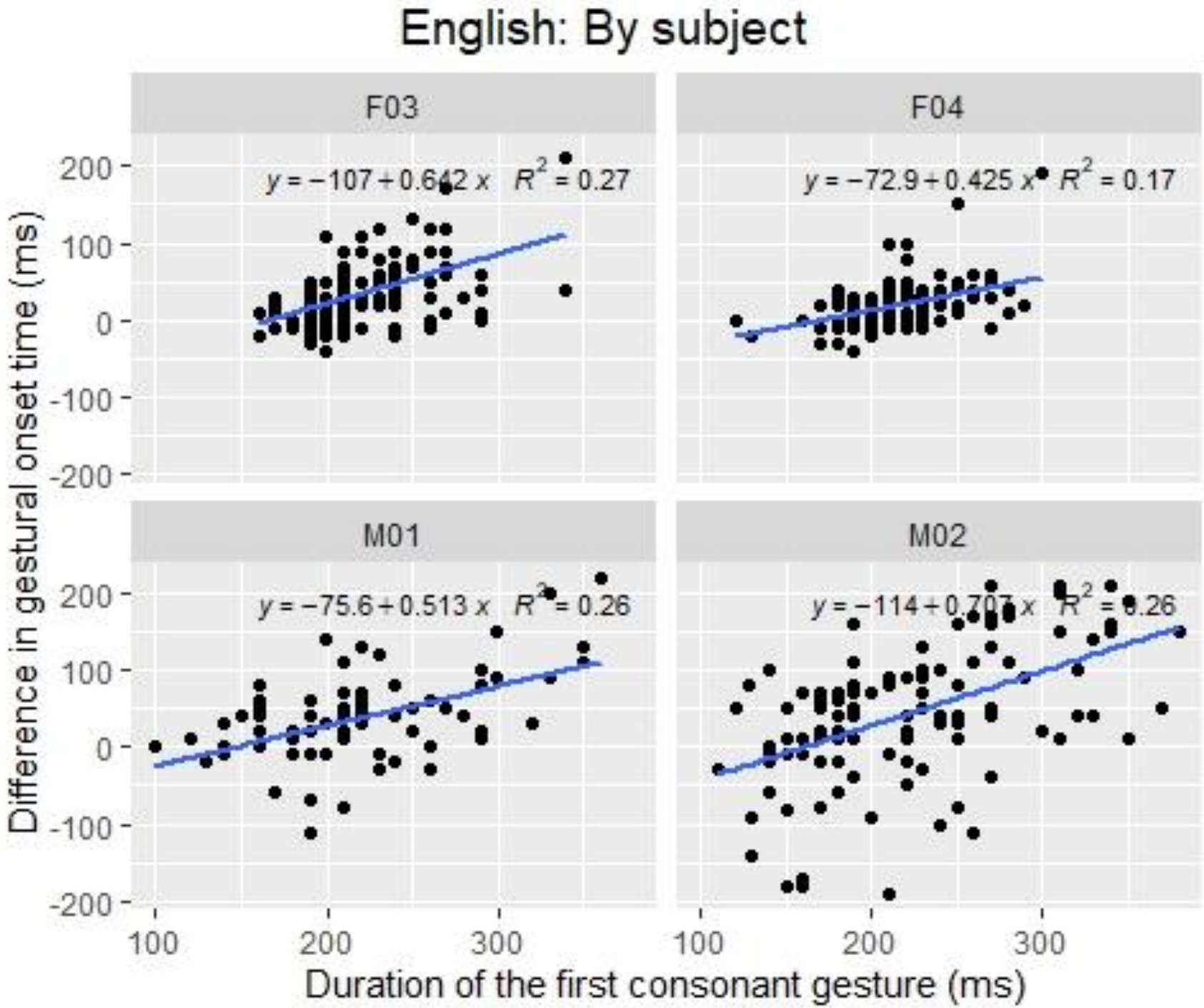
- Training data
  - ❖**[-Wh] Word3 Word4**
  - ❖**Linear interpolation**
- Test data
  - ❖[+Wh] Word3 and Word4

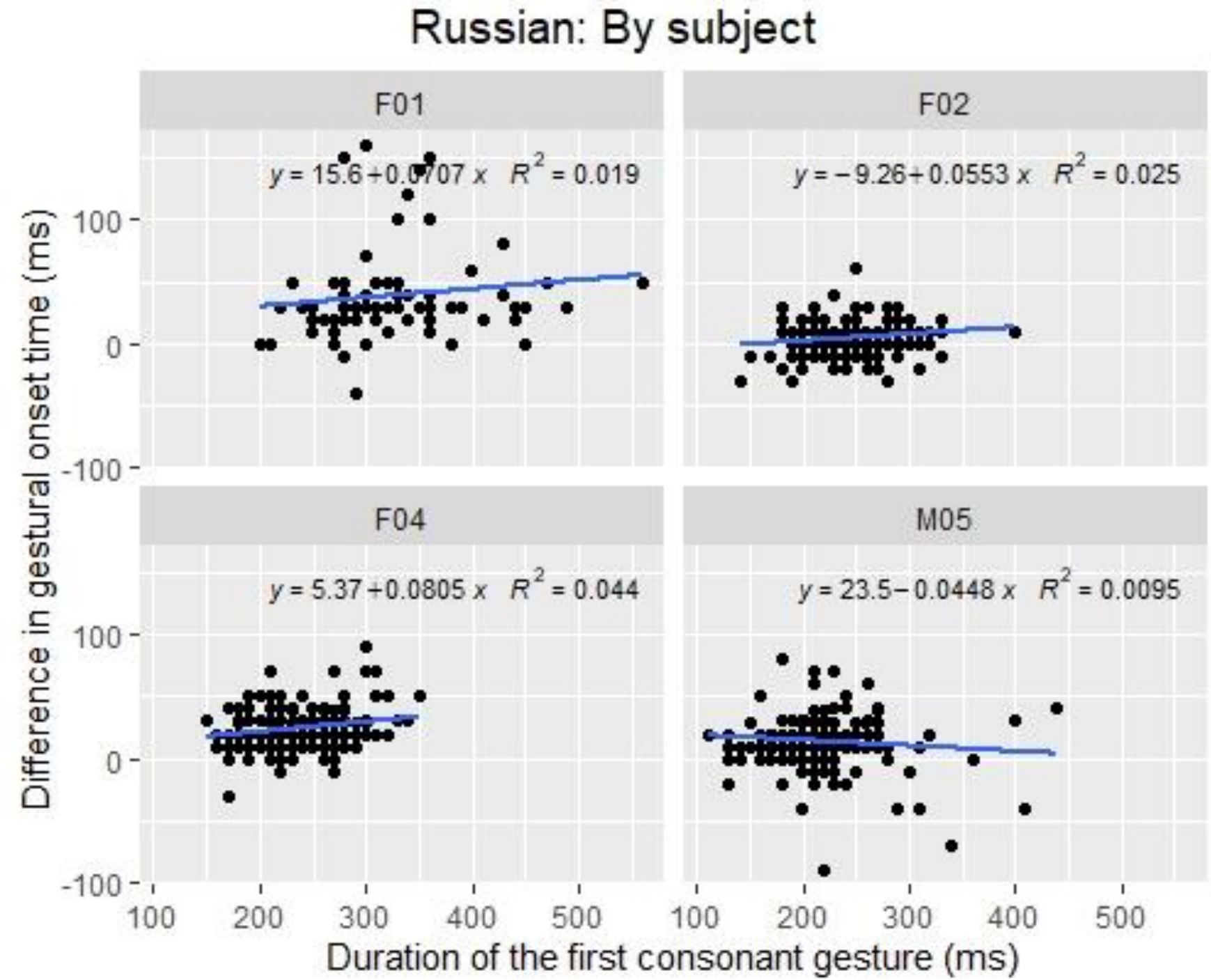
$$p(T|Co_i, \ldots, Co_n) = \frac{p(T) \times \Pi_{i=1}^{n} p(Co_i|T)}{\Pi_{i=1}^{n} p(Co_i)}$$

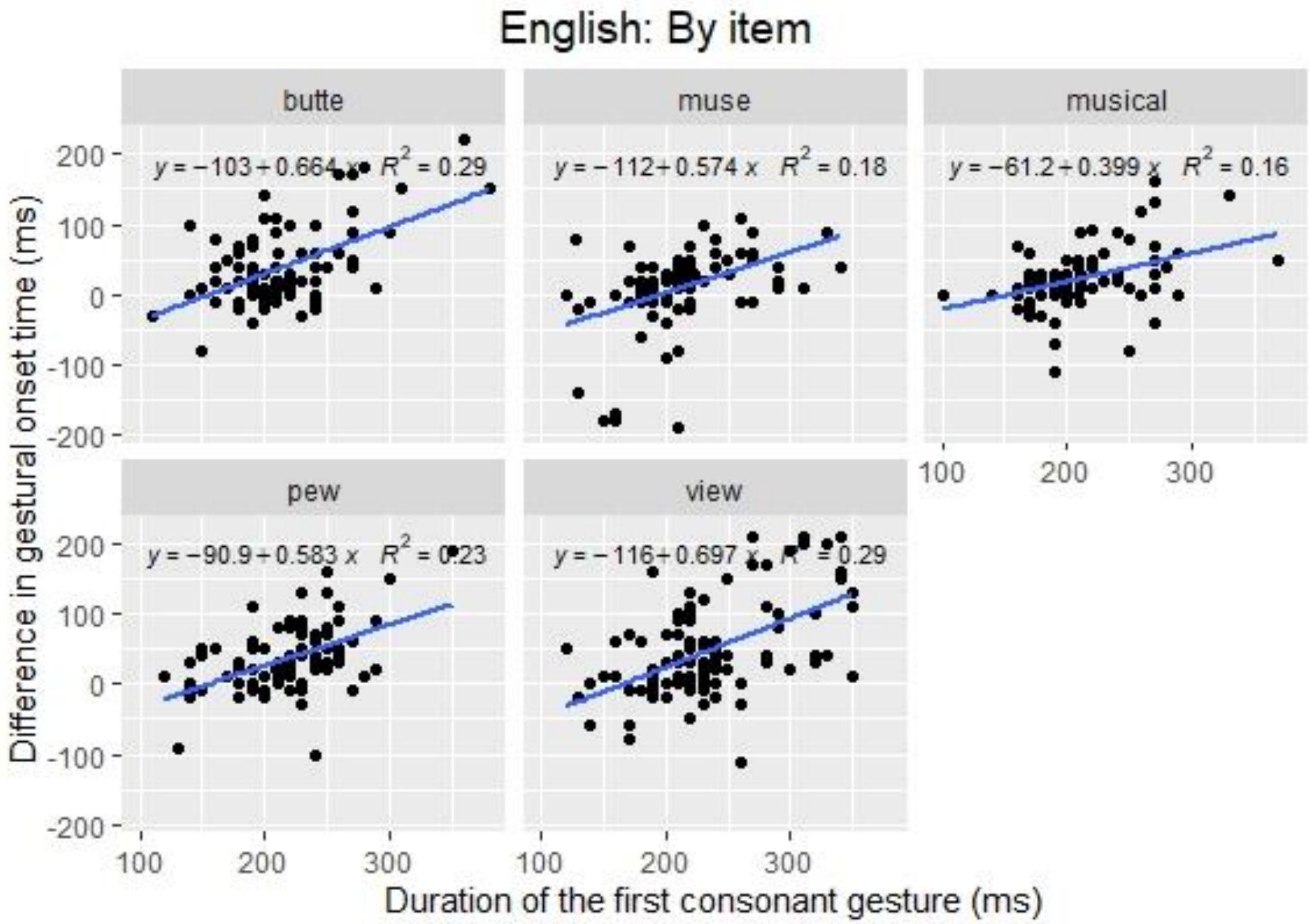
where Co_i is the ith DCT coefficient

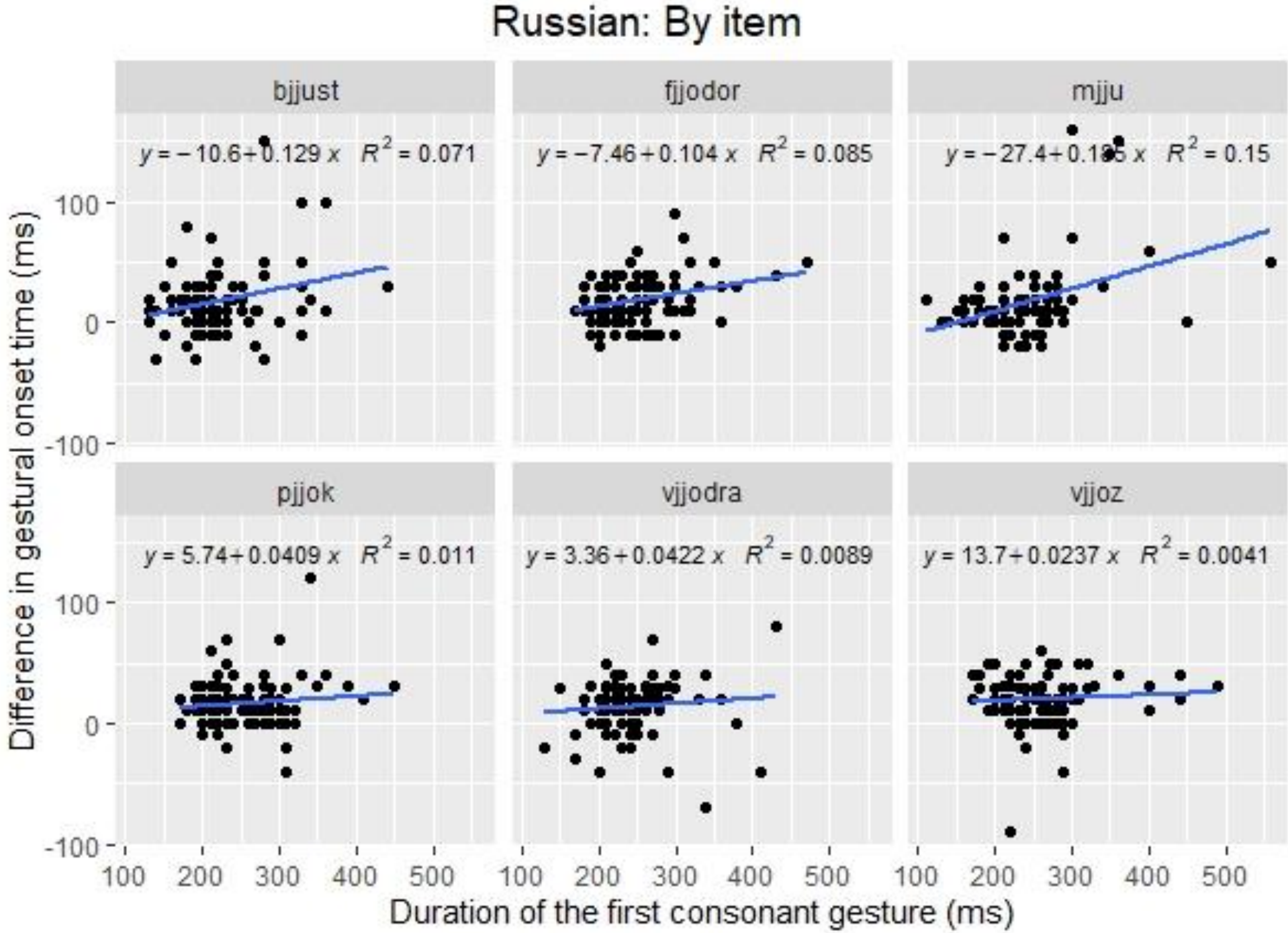
Parameters (4 DCT Coefficients)




English: By subject
F03
F04
M01
M02
y = −107 + 0.642 x R² = 0.27
y = −72.9 + 0.425 x R² = 0.17
y = −75.6 + 0.513 x R² = 0.26
y = −114 + 0.707 x R² = 0.26
Difference in gestural onset time (ms)
200
100
0
-100
-200
100
200
300
Duration of the first consonant gesture (ms)


Russian: By subject
F01
F02
F04
M05
y = 15.6 + 0.0707 x R² = 0.019
y = −9.26 + 0.0553 x R² = 0.025
y = 5.37 + 0.0805 x R² = 0.044
y = 23.5 − 0.0448 x R² = 0.0095
Difference in gestural onset time (ms)
100
0
-100
100
200
300
400
500
Duration of the first consonant gesture (ms)


English: By item
butte
muse
musical
pew
view
y = −103 + 0.664 x R² = 0.29
y = −112 + 0.574 x R² = 0.18
y = −61.2 + 0.399 x R² = 0.16
y = −90.9 + 0.583 x R² = 0.23
y = −116 + 0.697 x R² = 0.29
Difference in gestural onset time (ms)
200
100
0
-100
-200
100
200
300
Duration of the first consonant gesture (ms)


Russian: By item
bjjust
fjjodor
mjju
pjjok
vjjodra
vjjoz
y = −10.6+0.129 x R² = 0.071
y = −7.46+0.104 x R² = 0.085
y = −27.4+0.185 x R² = 0.15
y = 5.74+0.0409 x R² = 0.011
y = 3.36+0.0422 x R² = 0.0089
y = 13.7+0.0237 x R² = 0.0041
Difference in gestural onset time (ms)
100
0
-100
100
200
300
400
500
Duration of the first consonant gesture (ms)

Methods: Russian

Speakers

Four native speakers of Russian (3 female and 1 male)

Stimuli

Target

C^j	Cj
[p^{j}ok]	[pjot]
[b^{j}ust]	[bjut]
[m^{j}u]	[mju]
[f^{j}odor]	[fjord]
[v^{j}oz]	[vjoş]
[v^{j}odra]	[vjotsa]

Fillers

sonority		
plateau	[ptaşka]	[tkat^j]
falling	[lgat^j]	[rvat^j]
rising	[blat]	[brak]

Carrier phrase: [ʌˈna ____ pəftʌˈr^{j}ilʌ]. ‘She ____ repeated.’

Methods: English

Speakers

Four native speakers of English (2 female and 2 male)

Stimuli

Target	Fillers
butte	frap
pew	
view	
mew	
musical	

- Carrier phrase: ‘It’s a ____ perhaps.’