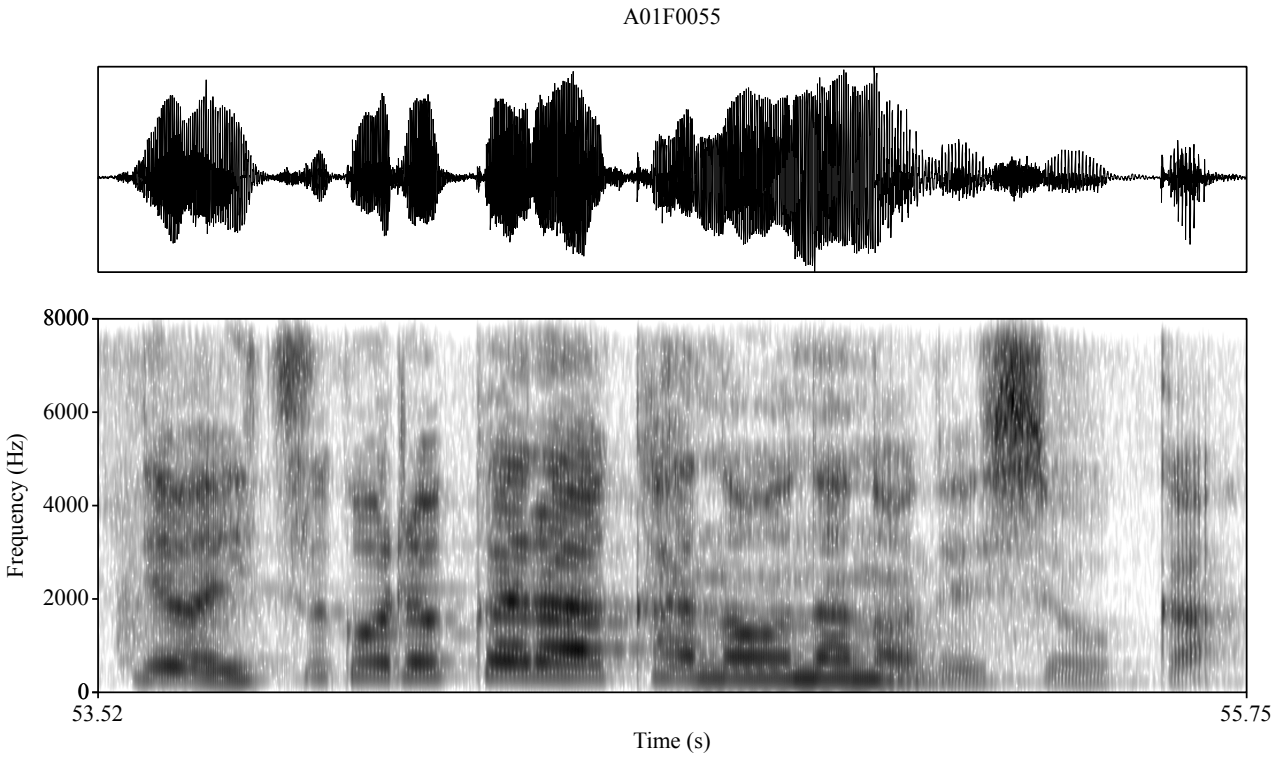
Speech rate effects on the Japanese stop voicing contrast

James Tanner, 五十嵐陽介, 前川喜久雄

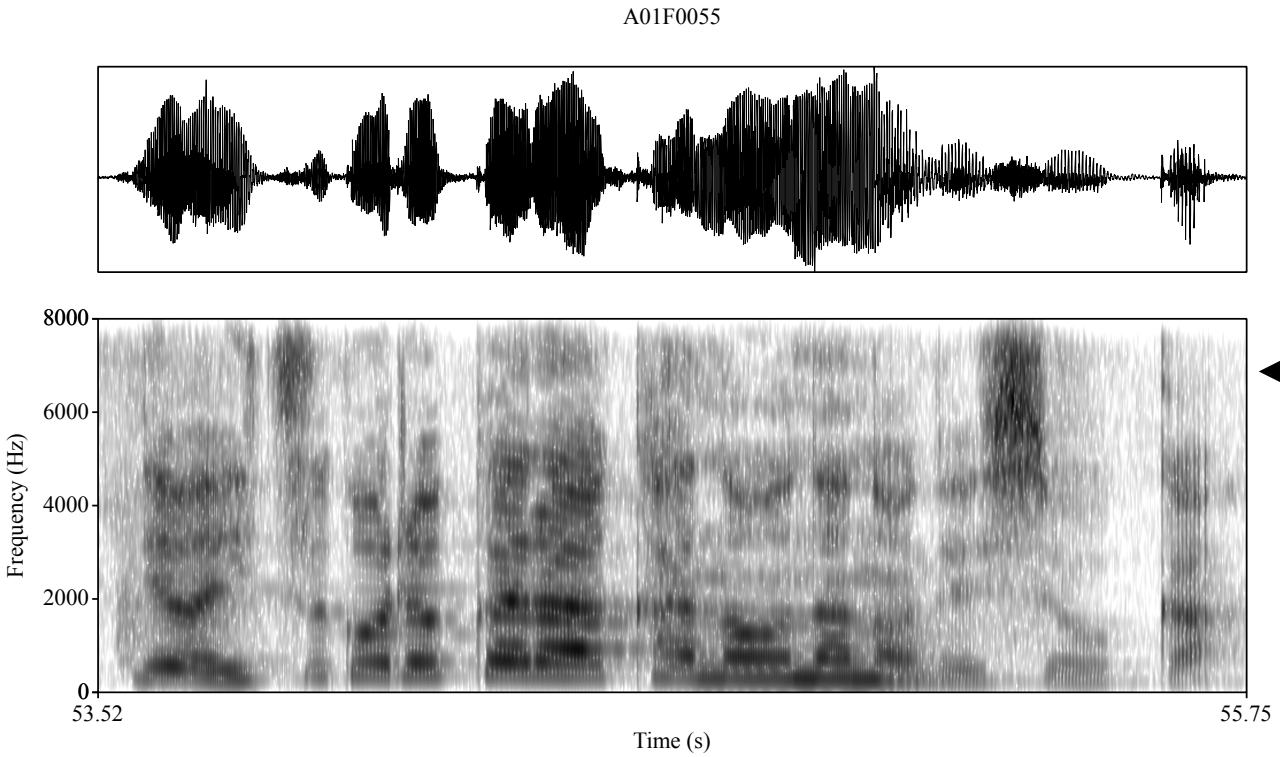
277th NINJAL Salon

Introduction



「ではいつ頃から可能なのでしょうか」

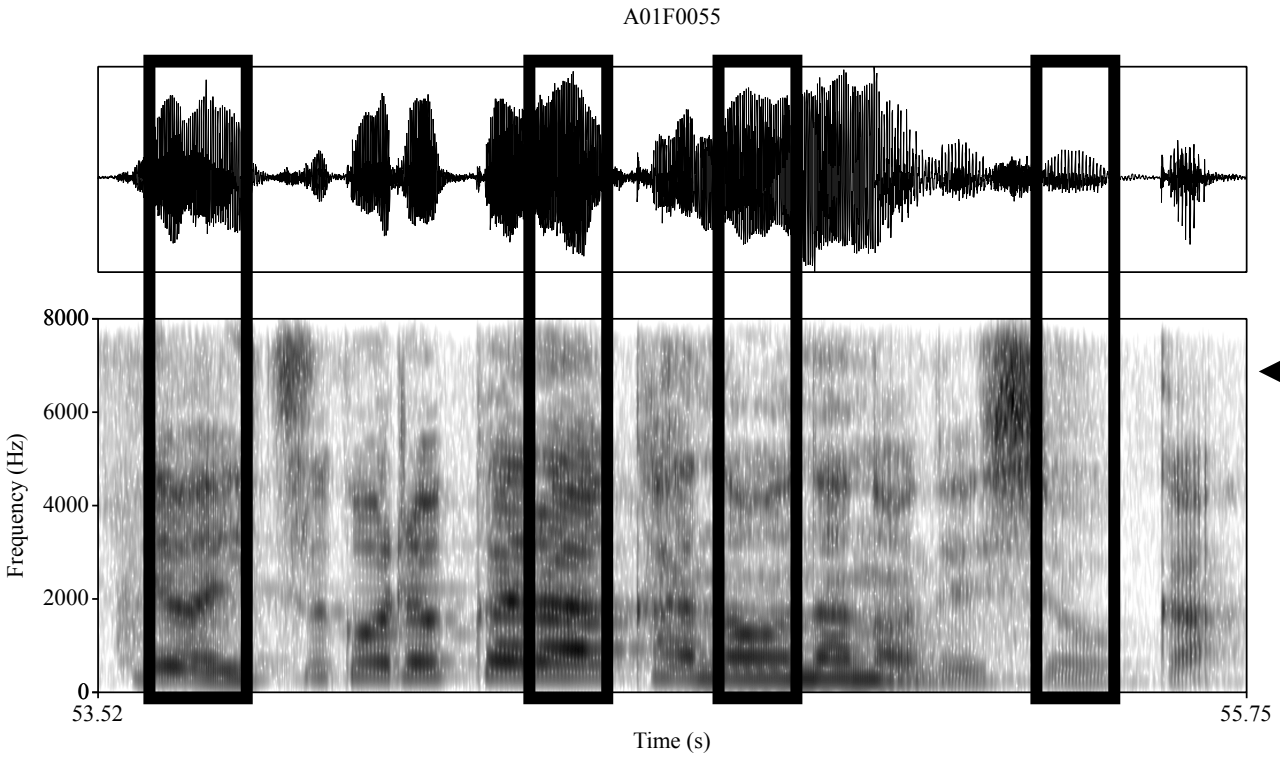
Introduction



「ではいつ頃から可能なのでしょうか」

- Speech production and perception: managing the mapping between these representations

Introduction



「ではいつ頃から可能なのでしょうか」

- Listeners must extract segmental acoustic information from the speech signal for later lexical access and syntactic/semantic processing

Introduction

- However speech itself is *highly variable*
  - Speech style, demographics (age, gender, region), language proficiency, emotional state, etc
- How do listeners manage this variability in perception?
- This variability must be somehow *structured* in production
- One type of variability: **speech tempo**

Liberman et al. (1967), Lisker (1986), Allen et al. (2003)

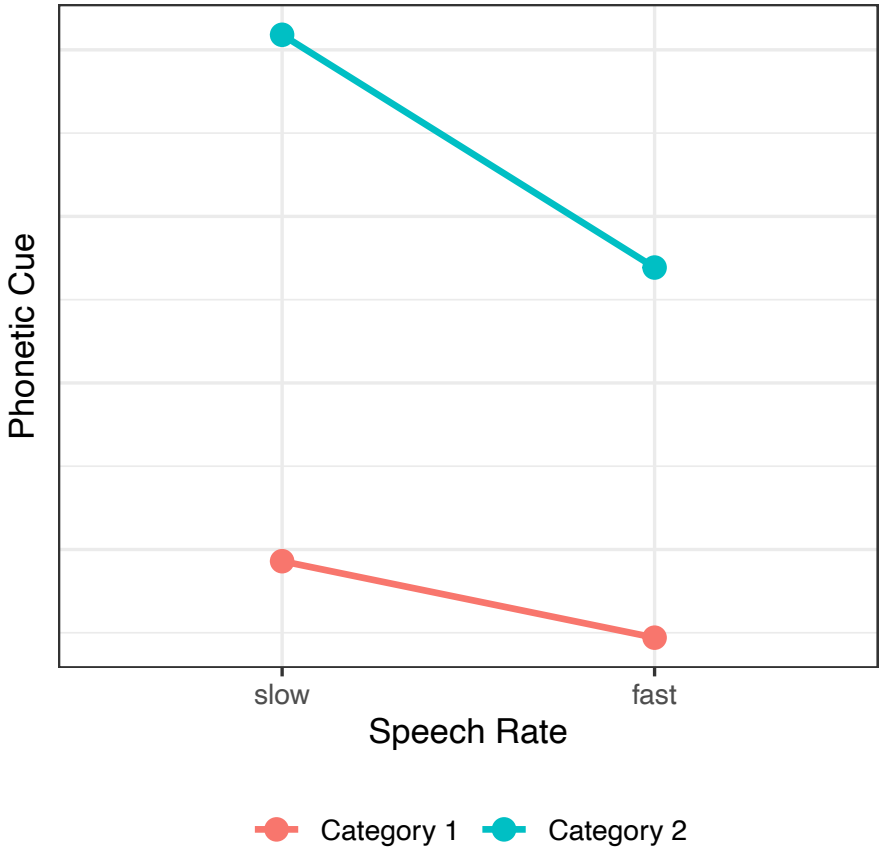
Introduction

- Speech **tempo**: the speed at which speech unfolds
- Speech **rate**: measure of speech tempo
  - e.g. syllables per second
- Changes the realisation of speech segments
  - Segments may become shorter
  - Changes in articulation plan

Miller et al. (1984), Shockey (1987), Davidson (2006), Kendall (2013)

Introduction

- Many phonological contrasts use temporal differences between segments
  - Japanese vowel length (/a/ vs /a:/)
  - Japanese singleton-geminate contrast (/t/ vs /t:/)
  - **This Study:** stop voicing contrasts (/t/ vs /d/)
- Contrasts *enhanced* in slower/more careful speech styles
- Both overall cue values and contrast sizes are compressed in faster speech



Port (1977), Miller et al. (1986), Kessinger & Blumstein (1997), Beckman et al. (2011)

Introduction

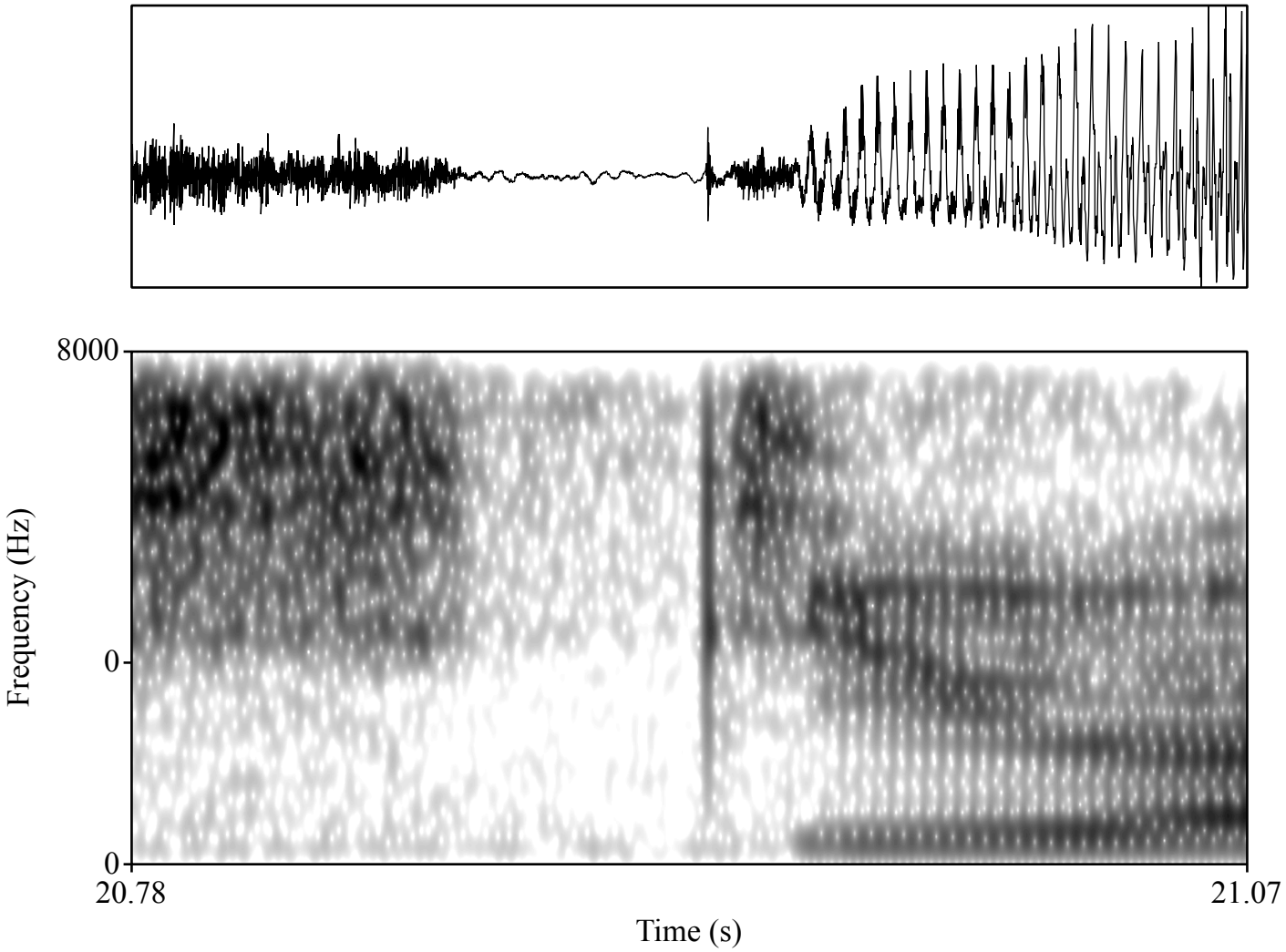
- How do listeners deal with speech rate variation in perceiving phonological contrasts?
  - Studies suggest listeners may ‘normalise’ rate differences by adjusting the perceptual boundary between categories
  - However listeners may maintain a single ‘optimal’ boundary which is robust to rate differences

Miller et al. (1986), Newman & Sawusch (1996), Kessinger & Blumstein (1997), Sawusch & Newman (2000), Nakai & Scobbie (2016), Ting & Kang (2023)

Introduction

- Phonological contrasts are phonetically *multidimensional*
- Stop contrasts are signalled by both *temporal* and *non-temporal* cues:

Scobbie (1998), Abramson & Whalen (2017), Clayards (2018)


8000
0
0
Frequency (Hz)
20.78
21.07
Time (s)

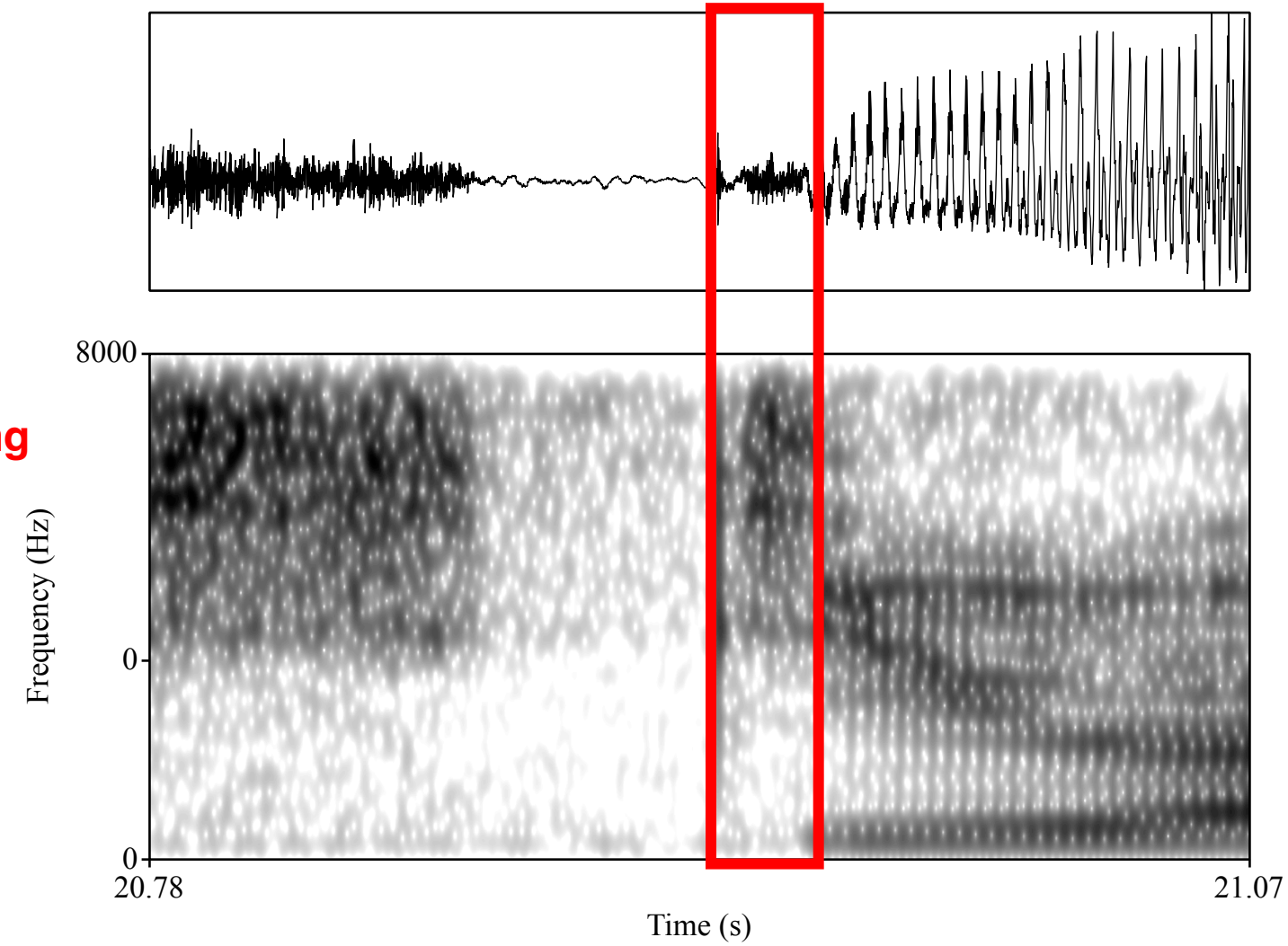
Voice Onset Time (VOT)

Period between burst and onset of voicing

Voiceless > Voiced

Primary cue to voicing in many languages (incl English)

Japanese contrasts ‘short’ and ‘intermediate’ VOTs



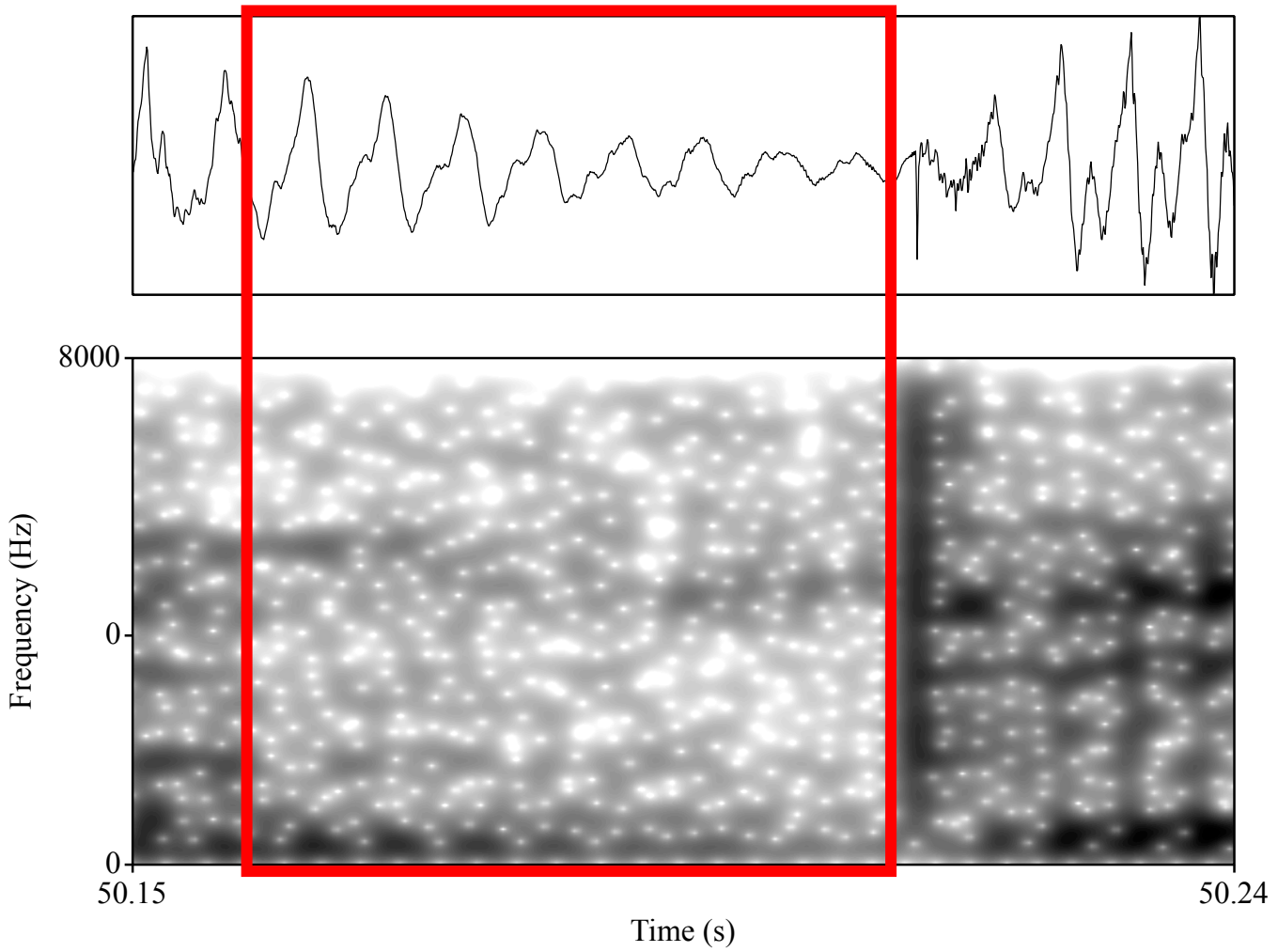
Lisker & Abramson (1964), Shimizu (1996), Riney et al. (2007), Takada (2011)

Voicing During Closure (VDC)

Sometimes called ‘negative VOT’ or ‘prevoicing’

May carry from preceding segment or start during closure

Prominent cue for voicing in Japanese

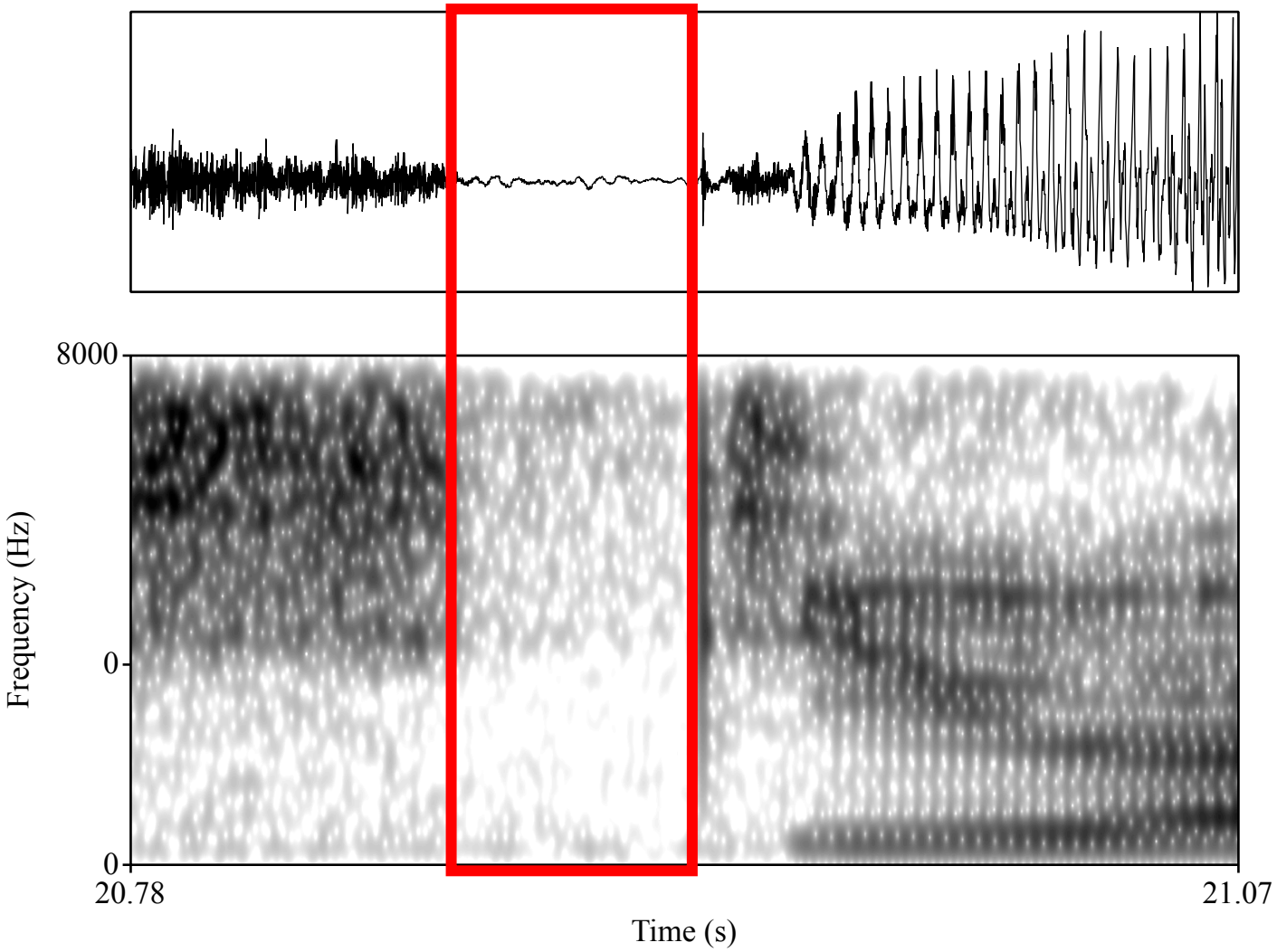


Ohala (1983), Westbury & Keating (1984), Davidson (2016, 2017), Gao & Arai (2019), Sonderegger et al. (2020), Tanner et al. (2020)

Closure Duration (CD)

Also cue to stop voicing across multiple languages

Voiced > Voiceless



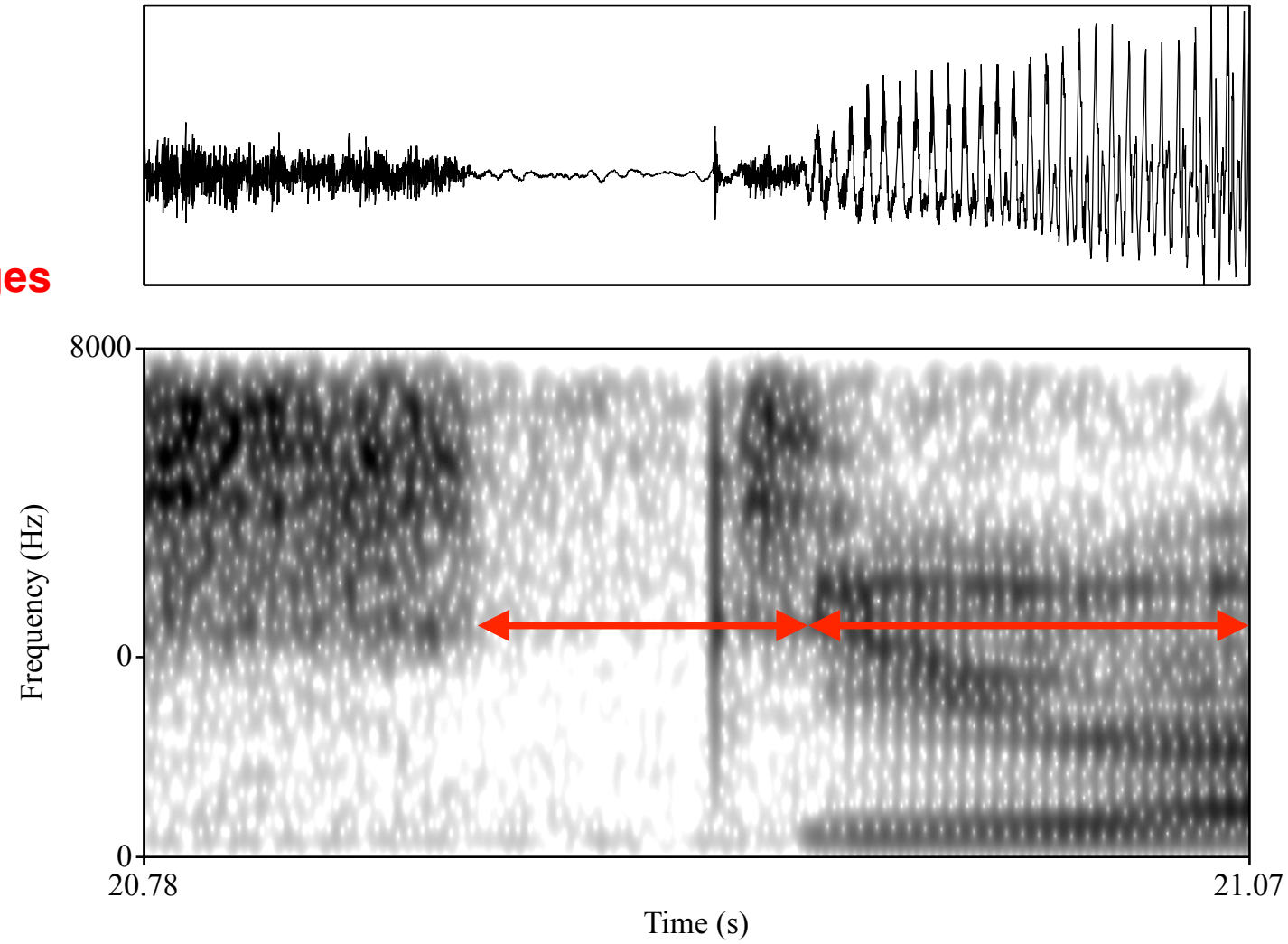
Lisker (1957), Han (1962), Cho & Keating (2001), Idemaru & Guion (2008)

Stop-Following-Vowel Ratio (SVFR)

Following vowel compensates for changes in stop duration

Shorter voiced stop -> longer vowel

Results in a 'stable' mora duration

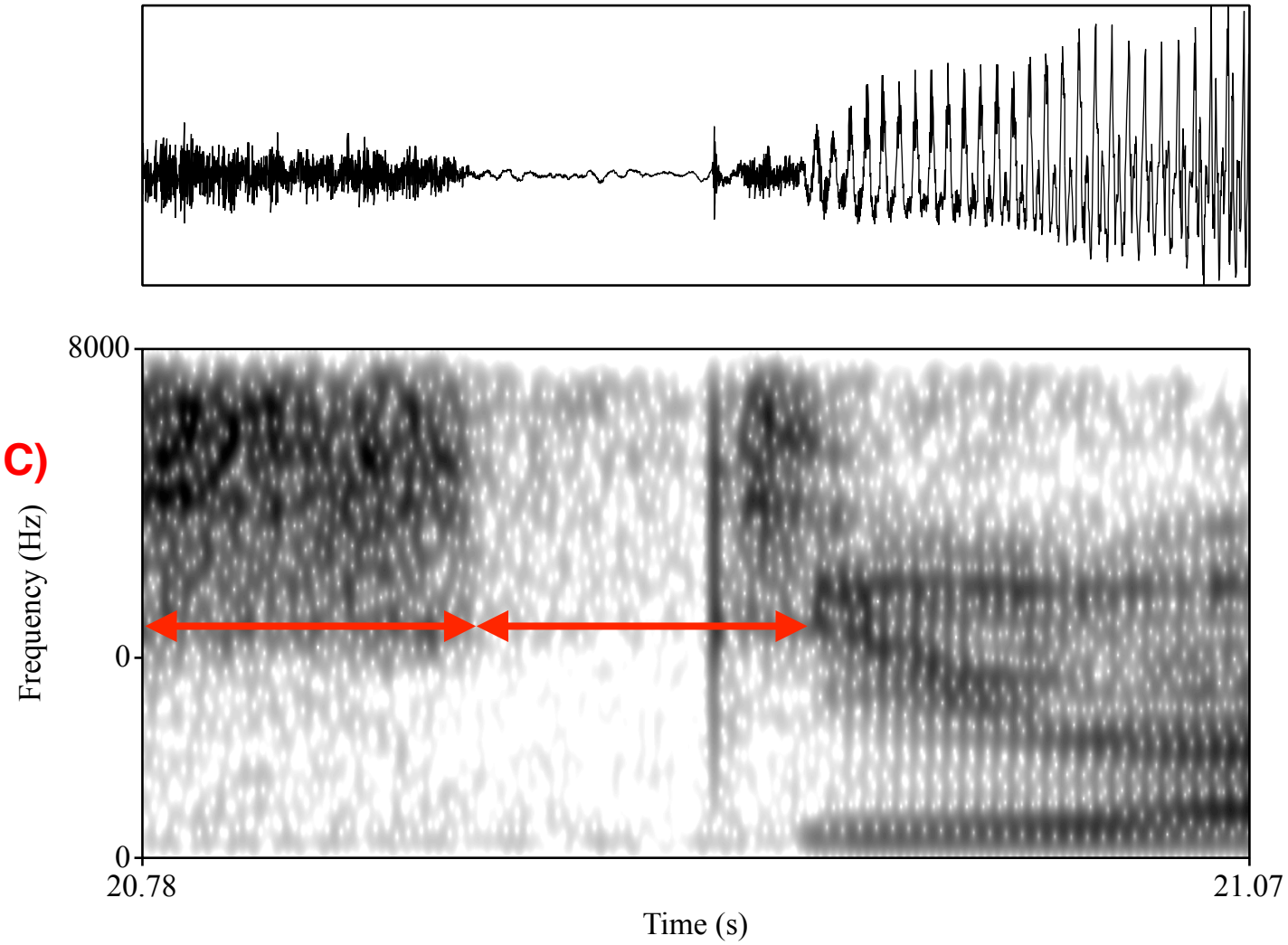


Klatt (1975), Port & Rotunno (1979), Minagawa-Kawai (1999), Warner & Arai (2001), Toscano & McMurray (2012)

Stop-Previous-Vowel Ratio (SPFR)

Cue to stop voicing in syllabic codas

Ratio below 1 (i.e. V > C) before voiced consonants



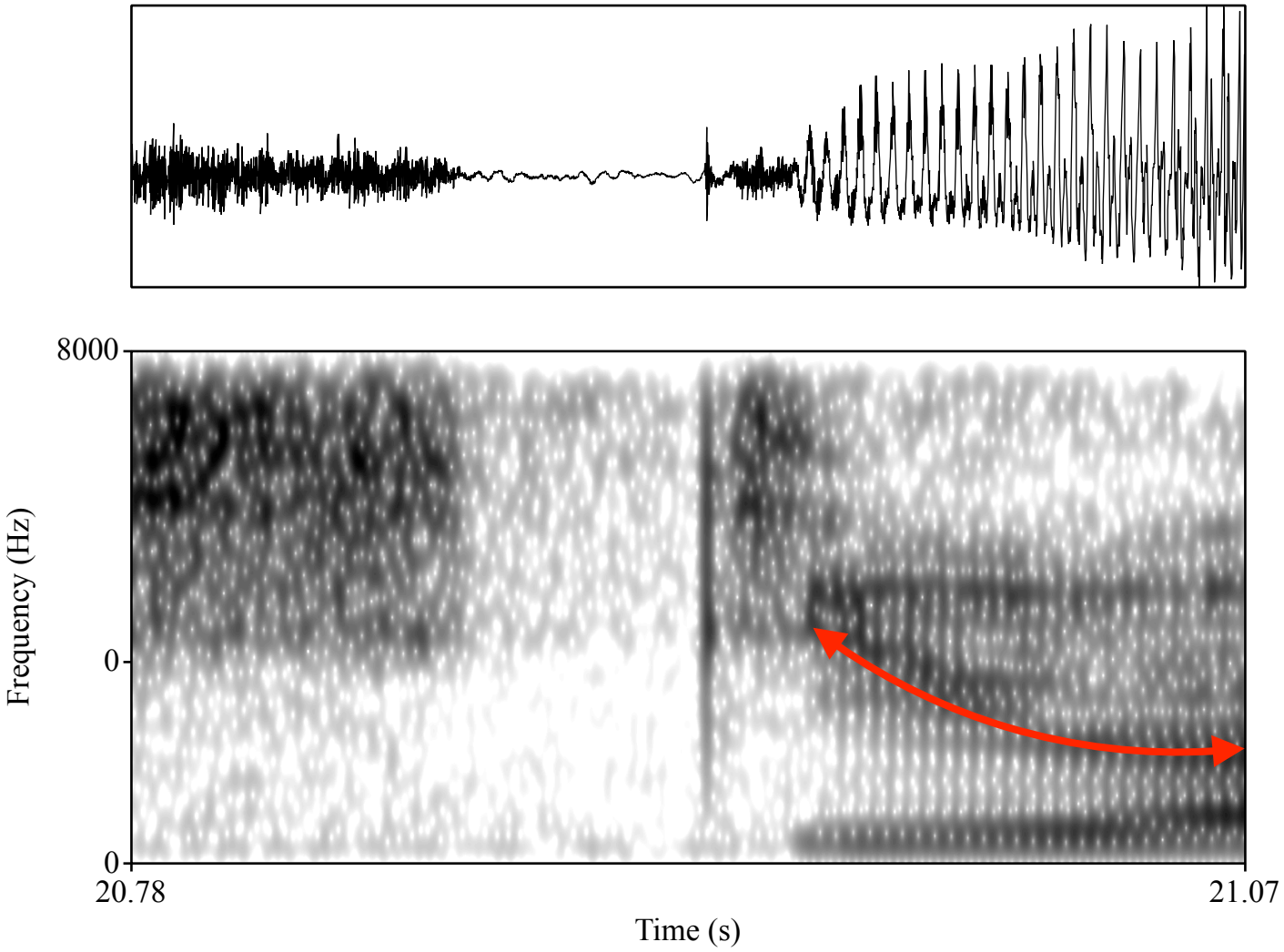
House (1961), Ohala (1970), Chen (1970), Port & Rotunno (1979)

Consonant F0 (CF0)

Cross-linguistic cue to stop voicing

CF0 higher for voiceless than voiced stops

Prominent in distinguishing utterance-initial stops in Japanese



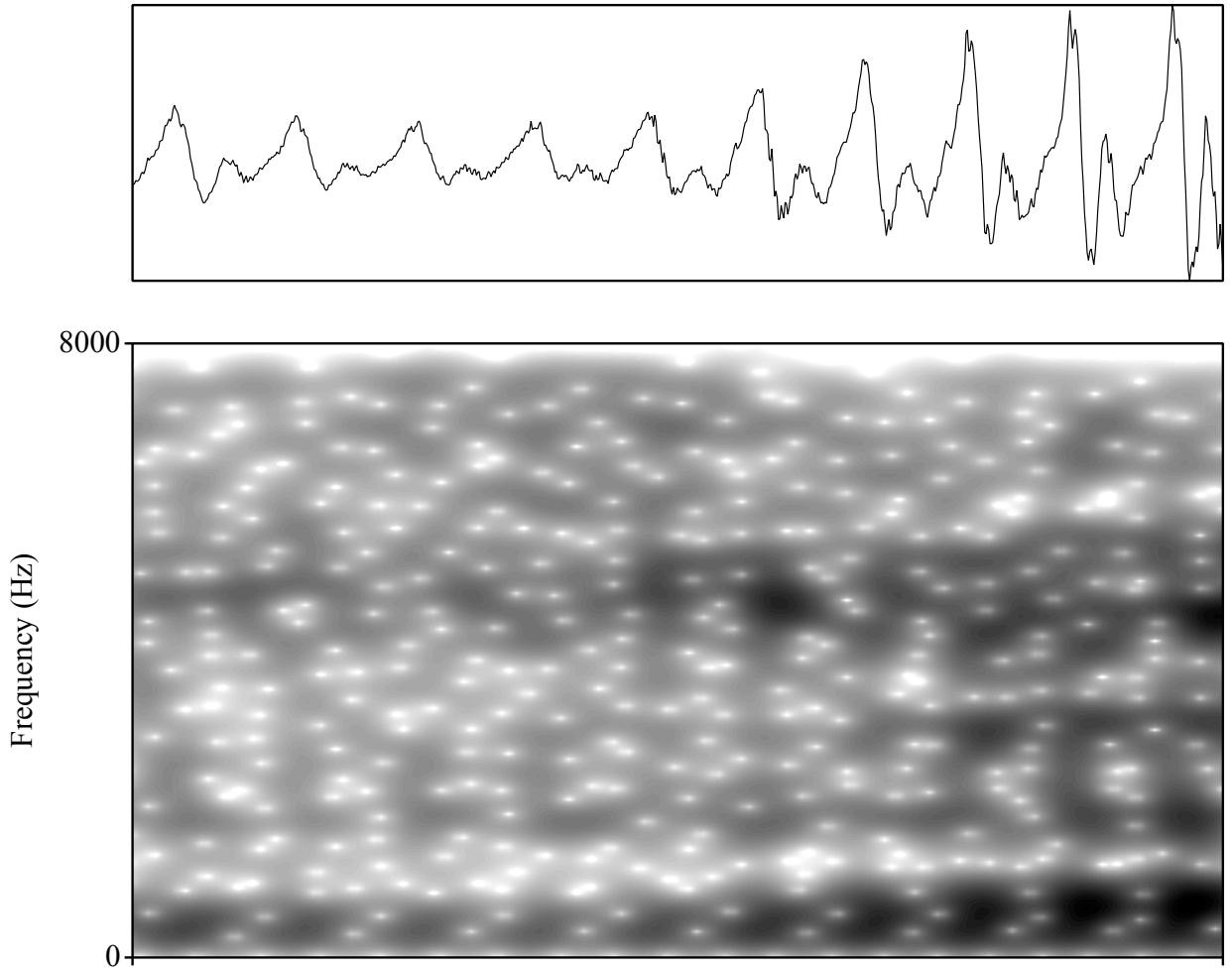
Kingston (2007), Gao & Arai (2019), Byun (2021, 2023), Ting et al. (2023)

Stop Lenition

Reduction from stops to fricatives or approximants

No clear distinction between closure & burst

More common for voiced stops to lenite



Kawakami (1977), Lavoie (2001), DiCanio (2012), Bouavichith & Davidson (2013), Maekawa (2018), DiCanio et al. (2021)

Introduction

- Previous work on speech rate effects on stop voicing cues
  - (Mostly) English/Germanic languages
  - (Mostly) Focus on single cue (VOT)
  - (Mostly) controlled speech

Kessinger & Blumstein (1997), Pickett et al. (1999), Beckman et al. (2011), Nakai & Scobbie (2016), Reinisch (2016), Ting & Kang (2023)

Introduction

- Previous work on speech rate effects on stop voicing cues
  - (Mostly) English/Germanic languages
  - (Mostly) Focus on single cue (VOT)
  - (Mostly) controlled speech

1. What about languages where VOT is not the primary cue?
2. How does speech rate affect *multiple* cues*?*
3. What are the effects in spontaneous speech?

Kessinger & Blumstein (1997), Pickett et al. (1999), Beckman et al. (2011), Nakai & Scobbie (2016), Reinisch (2016), Ting & Kang (2023)

Introduction

- *How do differences in speech rate modulate multiple cues to the stop voicing contrast?* (RQ1)
- *Does the importance of each cue change at different speech rates?* (RQ2)

Introduction

- *How do differences in speech rate modulate multiple cues to the stop voicing contrast?* (RQ1)
- *Does the importance of each cue change at different speech rates?* (RQ2)

- This study: examine speech rate effects on stop voicing cues across a large corpus of spontaneous Japanese
  - Japanese has different phonetic implementation of voicing
  - Uses other cues (VDC, CD, CF0) alongside VOT

Iverson & Salmons (1995), Shimizu (1996), Nasukawa (2005), Gao & Arai (2019), Tanner et al. (2020)

Methods

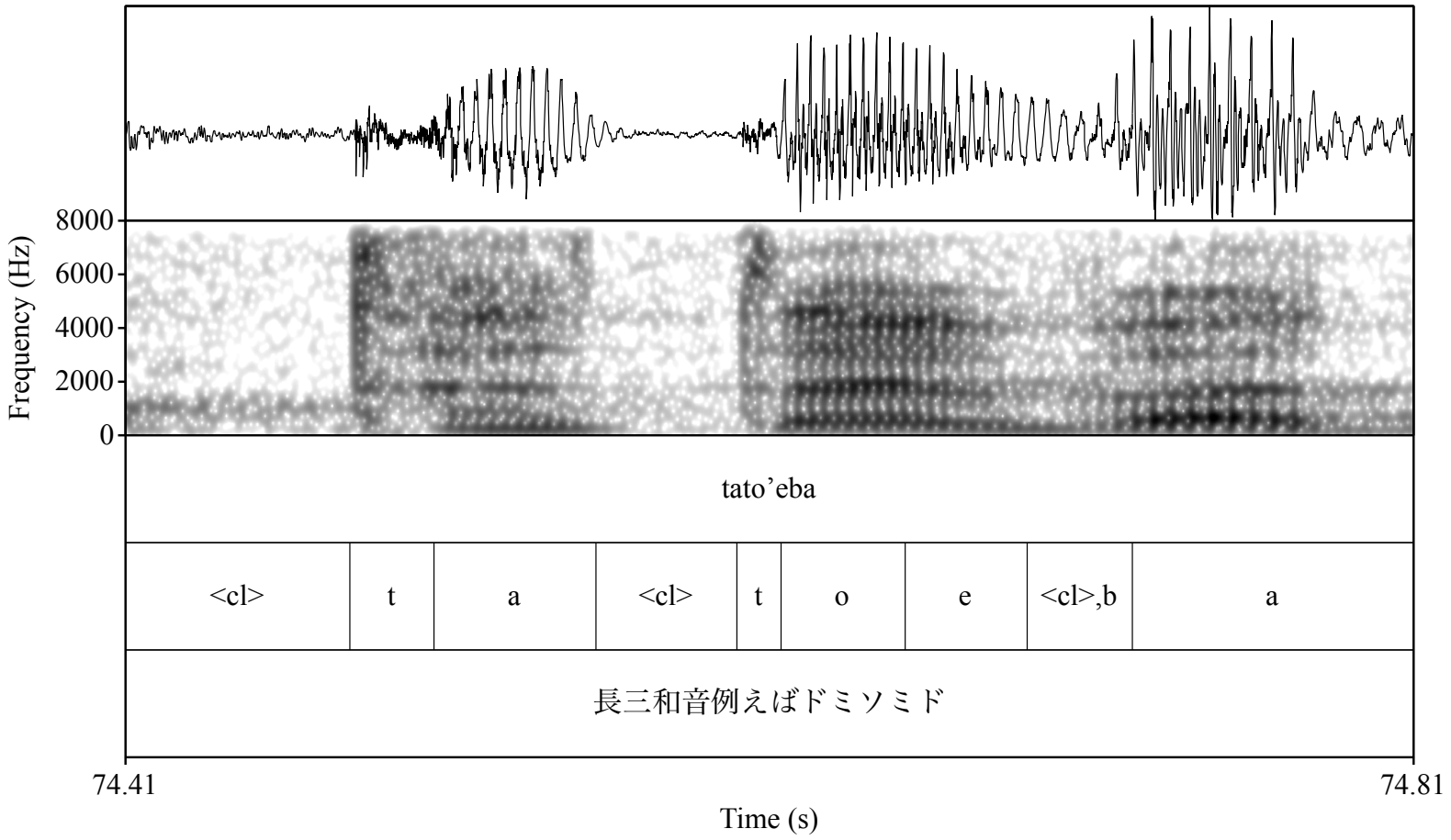
Data

大学共同利用機関法人 人間文化研究機構
国立国語研究所
National Institute for Japanese Language and Linguistics
NINJAL

- *Corpus of Spontaneous Japanese-Core* (日本語話し言葉コーパス, CSJ-Core)
  - ~45 hours (recorded 1999-2001)
  - 137 speakers (58 female), born 1939-1970
  - Monologues (lectures & public speaking), some conversations
- Substantial phonetic annotation
  - Vowel devoicing, voice quality, etc
  - For our purposes: **stop closure & bursts**

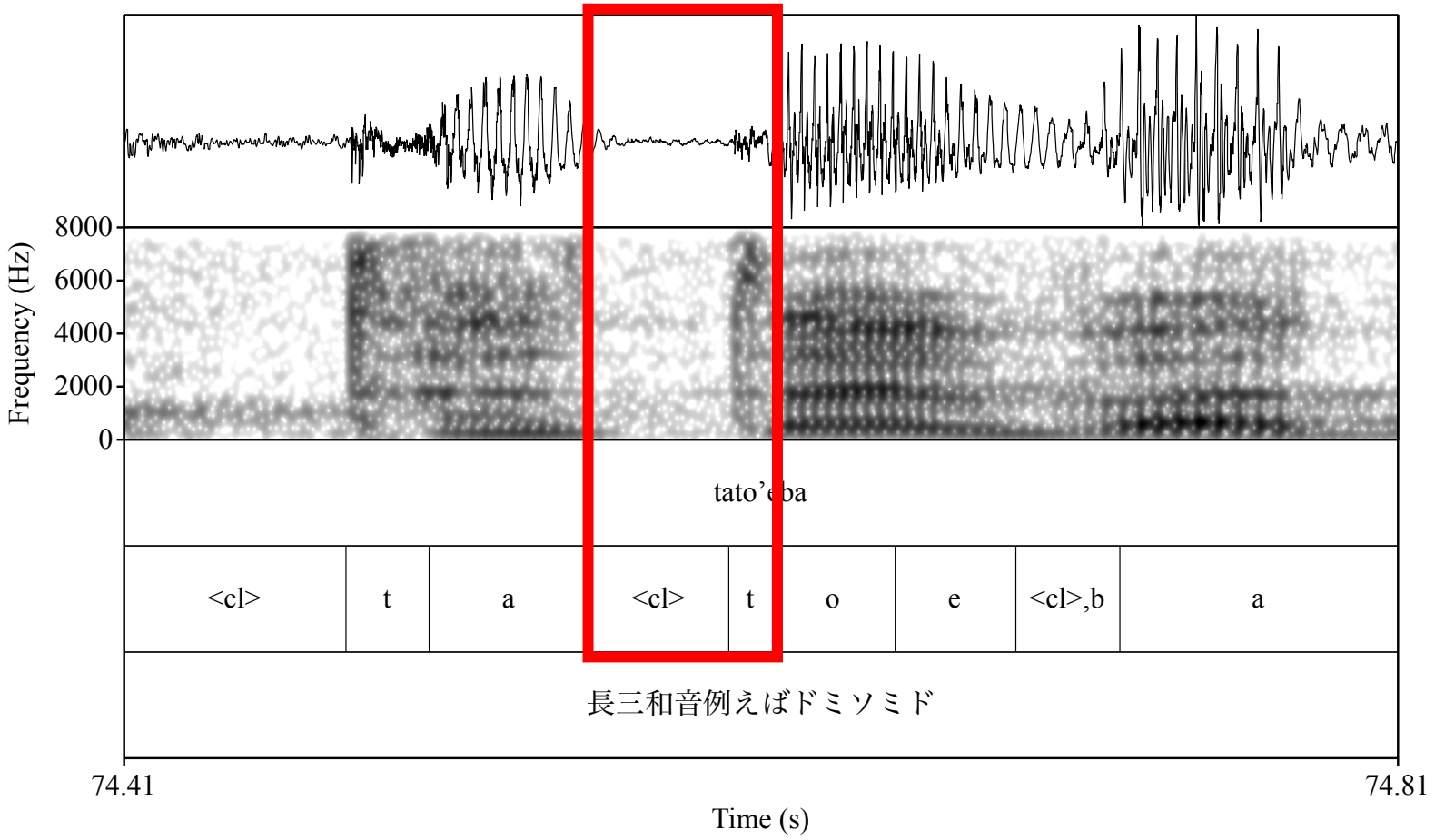
Maekawa et al. (2000), Maekawa (2003), Kikuchi & Maekawa (2003), Koiso et al. (2014)

Data


8000
6000
4000
2000
0
Frequency (Hz)
tato’eba
<cl>
t
a
<cl>
t
o
e
<cl>,b
a
長三和音例えばドミソミド
74.41
74.81
Time (s)

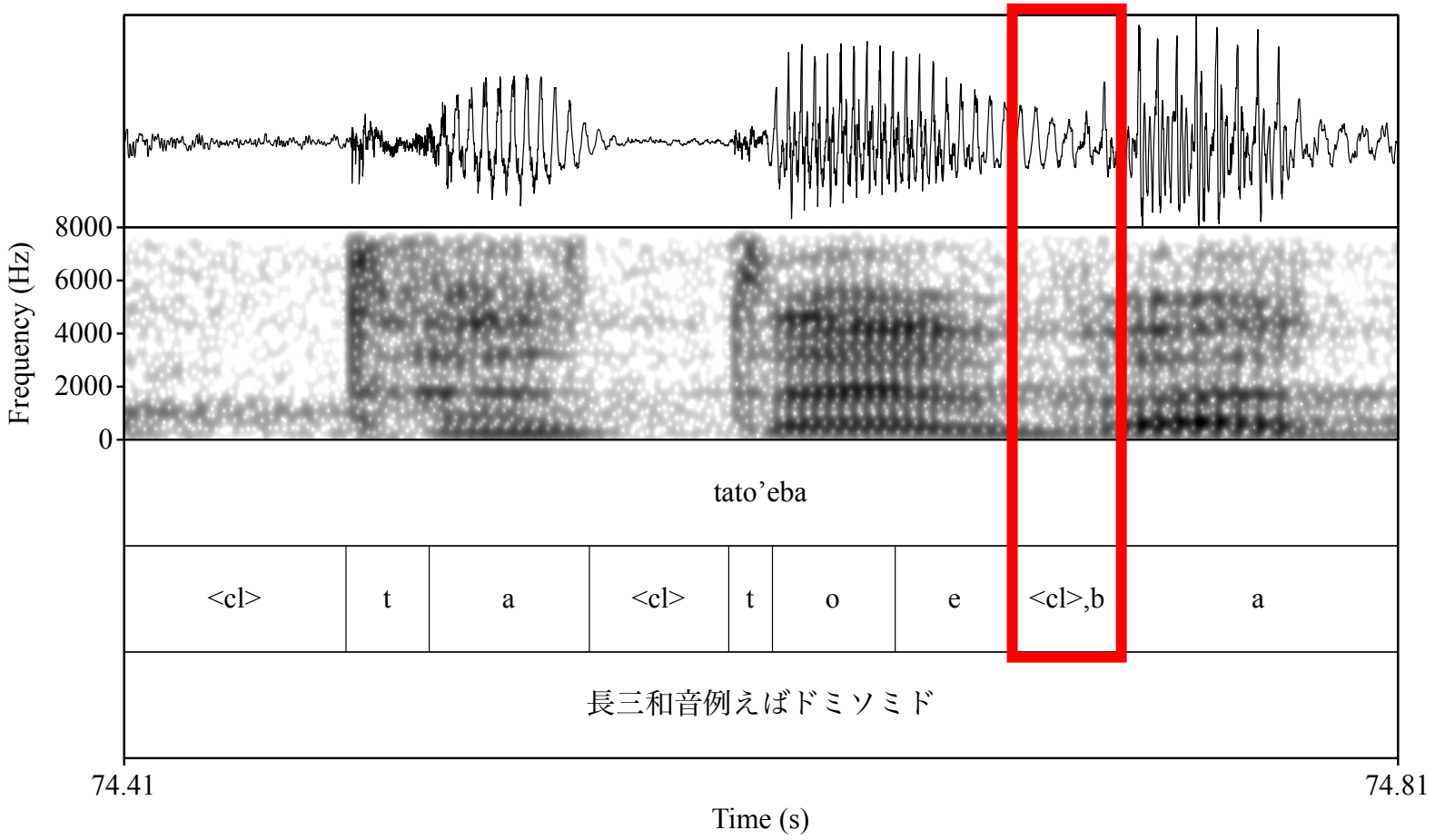
Data

Stops labelled as separate closure (<cl>) and burst ('VOT') sectio ns



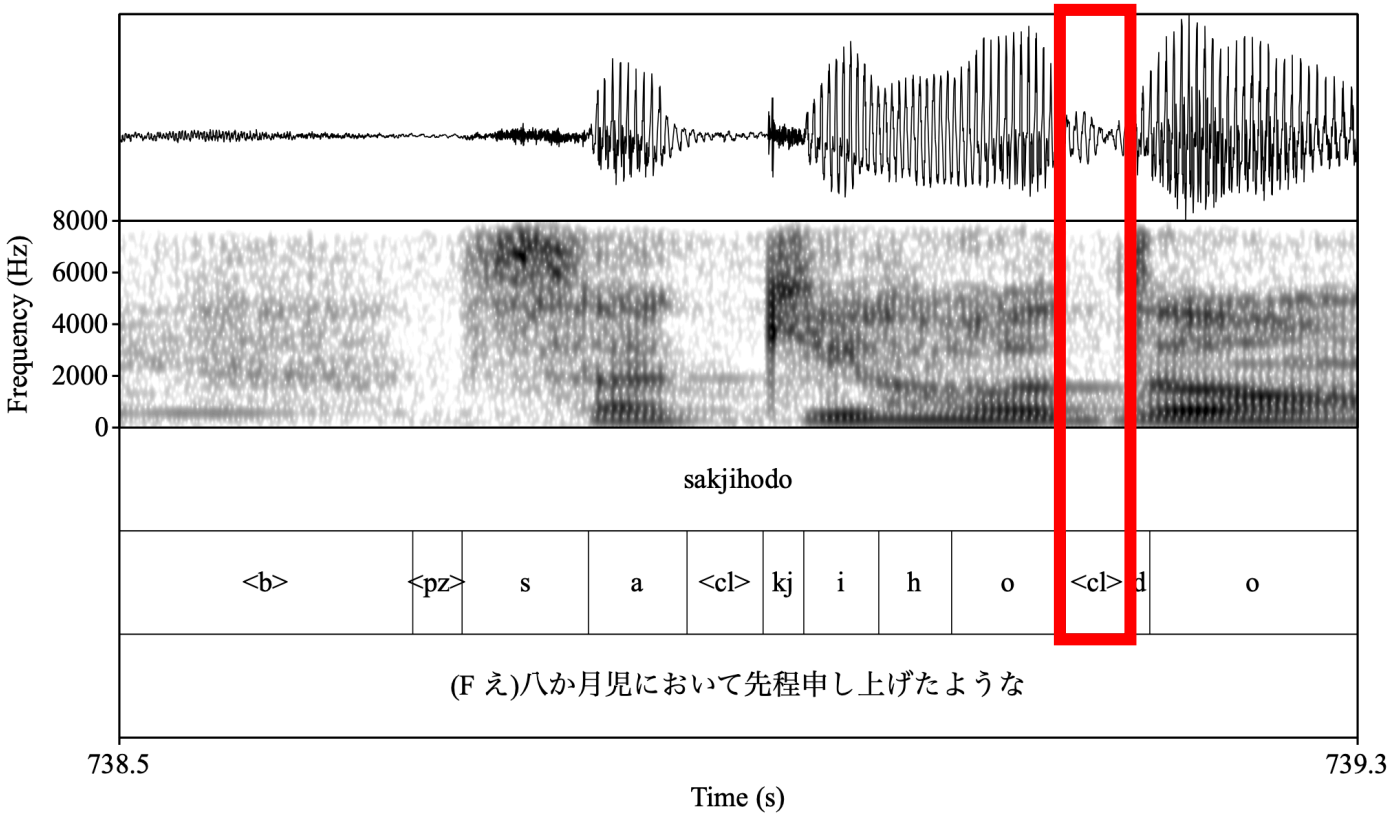
Data

Lenited stops labelled as single <cl>,phone label



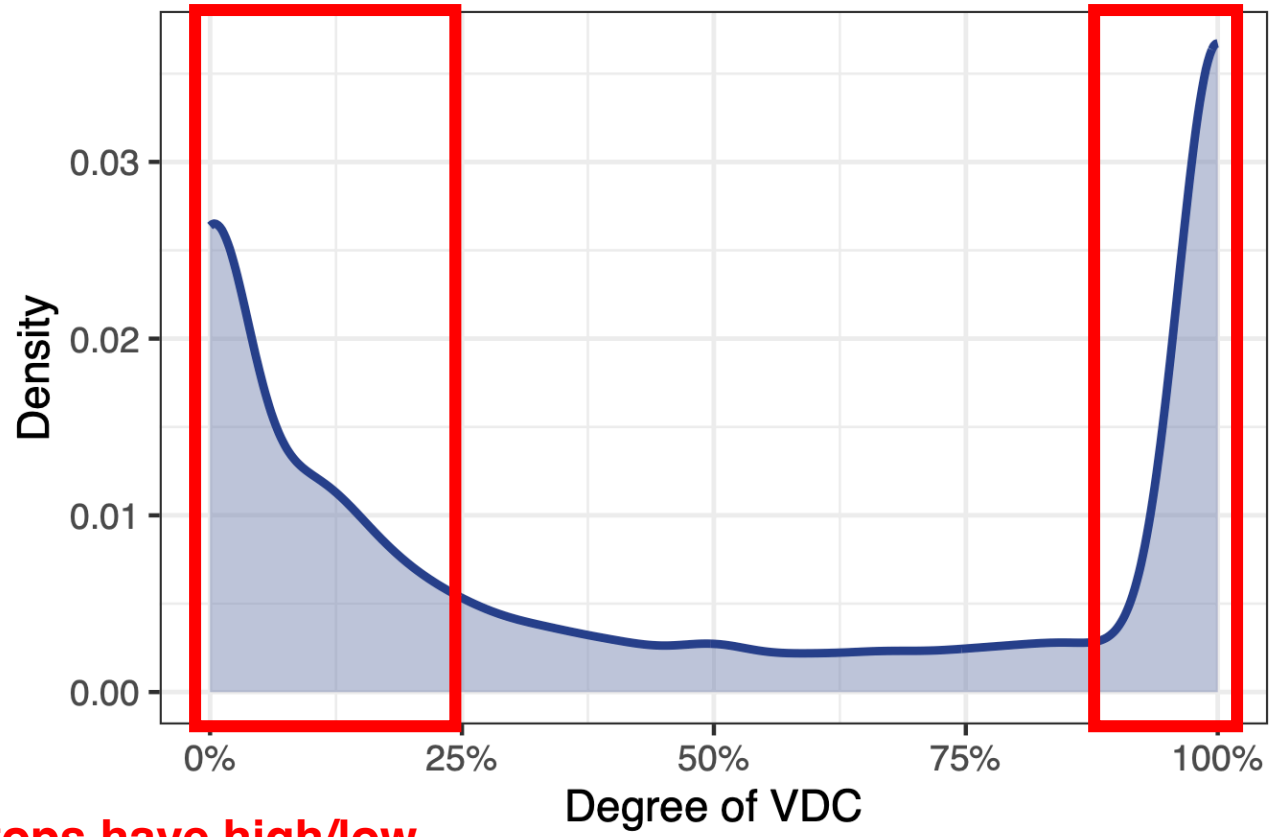
Data

- VDC measured as % of voicing within closure
  - Lenited stops = full '<cl>,stop'
- Extract '% of unvoiced frames' from Praat VoiceReport



Data

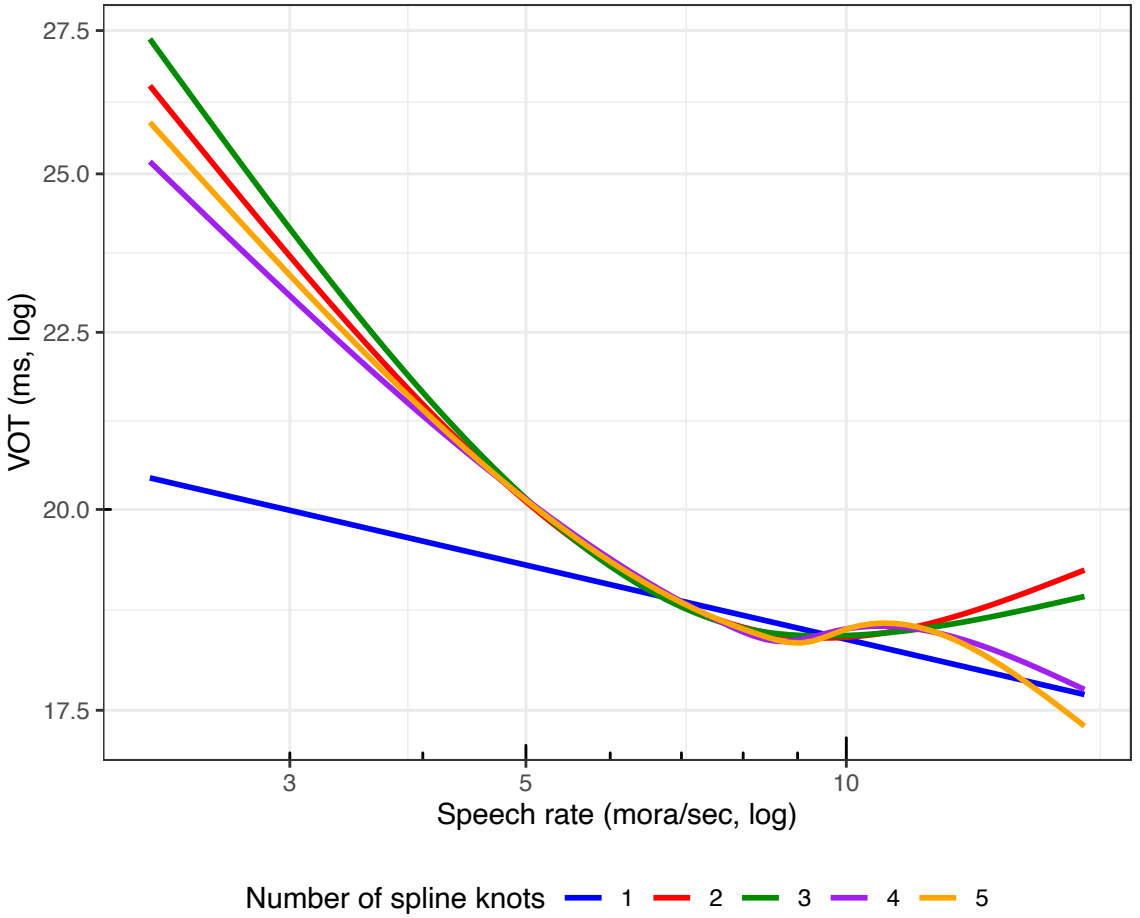
- VDC measured as % of voicing within closure
  - Lenited stops = full '<cl>,stop'
- Extract '% of unvoiced frames' from Praat VoiceReport
- CF0: measured at first 10% of vowel



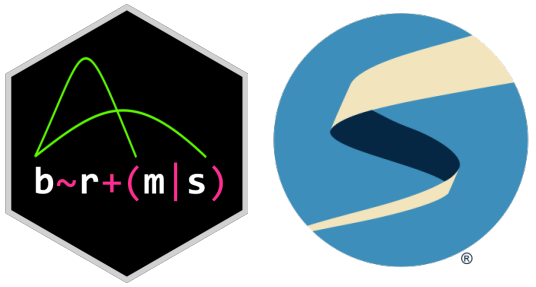
Most stops have high/low VDC%, with many stops with intermediate values

Data

- Speech Rate = mora per second (mora/sec) within utterance
- Allow speech rate to have **non-linear** effect
- Cubic splines = piecewise set of polynomial functions
  - More knots = more non-linearity
- Choose number of knots from visual examination



Models



- 94,879 tokens (67,735 non-lenited) utterance-medial tokens
- Bayesian multilevel ('mixed-effects') regressions in Stan/brms
  - VOT, CD, SPVR, SFVR = log-transformed linear model
  - CF0 = linear model
  - Lenition likelihood = logistic/binomial model
  - VDC = zero-one-inflated beta model
    - Models values within a bounded [0,1] range (e.g. percentages)

Models

- Effects of interest
  - Phonological voicing contrast (voiced vs voiceless)
  - Speech rate — two separate effects
    - Speech rate **mean**: effect of speaker's average speech rate
    - Speech rate **deviation**: effect of token-level speech rate, relative to speaker's average
    - Effect of 'Fast/slow speaker' vs 'Fast/slow speech’
  - Interaction *between* voicing and speech rates
- Controls
  - POA, gender, age, frequency, prosodic boundaries, etc
- Full random structure for speakers, word-level intercepts
  - Account for by-speaker/by-word variability

Stuart-Smith et al. (2015), Sonderegger et al. (2017), Tanner et al. (2020)

Models

RQ1: ***How do differences in speech rate modulate multiple cues to the stop voicing contrast?***

- Effects of interest
  - Phonological voicing contrast (voiced vs voiceless)
  - Speech rate — two separate effects
    - Speech rate **mean**: effect of speaker's average speech rate
    - Speech rate **deviation**: effect of token-level speech rate, relative to speaker's average
    - Effect of 'Fast/slow speaker' vs 'Fast/slow speech'
  - Interaction *between* voicing and speech rates
- Controls
  - POA, gender, age, frequency, prosodic boundaries, etc
- Full random structure for speakers, word-level intercepts
  - Account for by-speaker/by-word variability

Stuart-Smith et al. (2015), Sonderegger et al. (2017), Tanner et al. (2020)

Models

- Effects of interest
  - Phonological voicing contrast (voiced vs voiceless)
  - Speech rate — two separate effects
    - Speech rate **mean**: effect of speaker's average speech rate
    - Speech rate **deviation**: effect of token-level speech rate, relative to speaker's average
    - Effect of 'Fast/slow speaker' vs 'Fast/slow speech'
  - Interaction *between* voicing and speech rates
- Controls
  - POA, gender, age, frequency, prosodic boundaries, etc
- Full random structure for speakers, word-level intercepts
  - Account for by-speaker/by-word variability

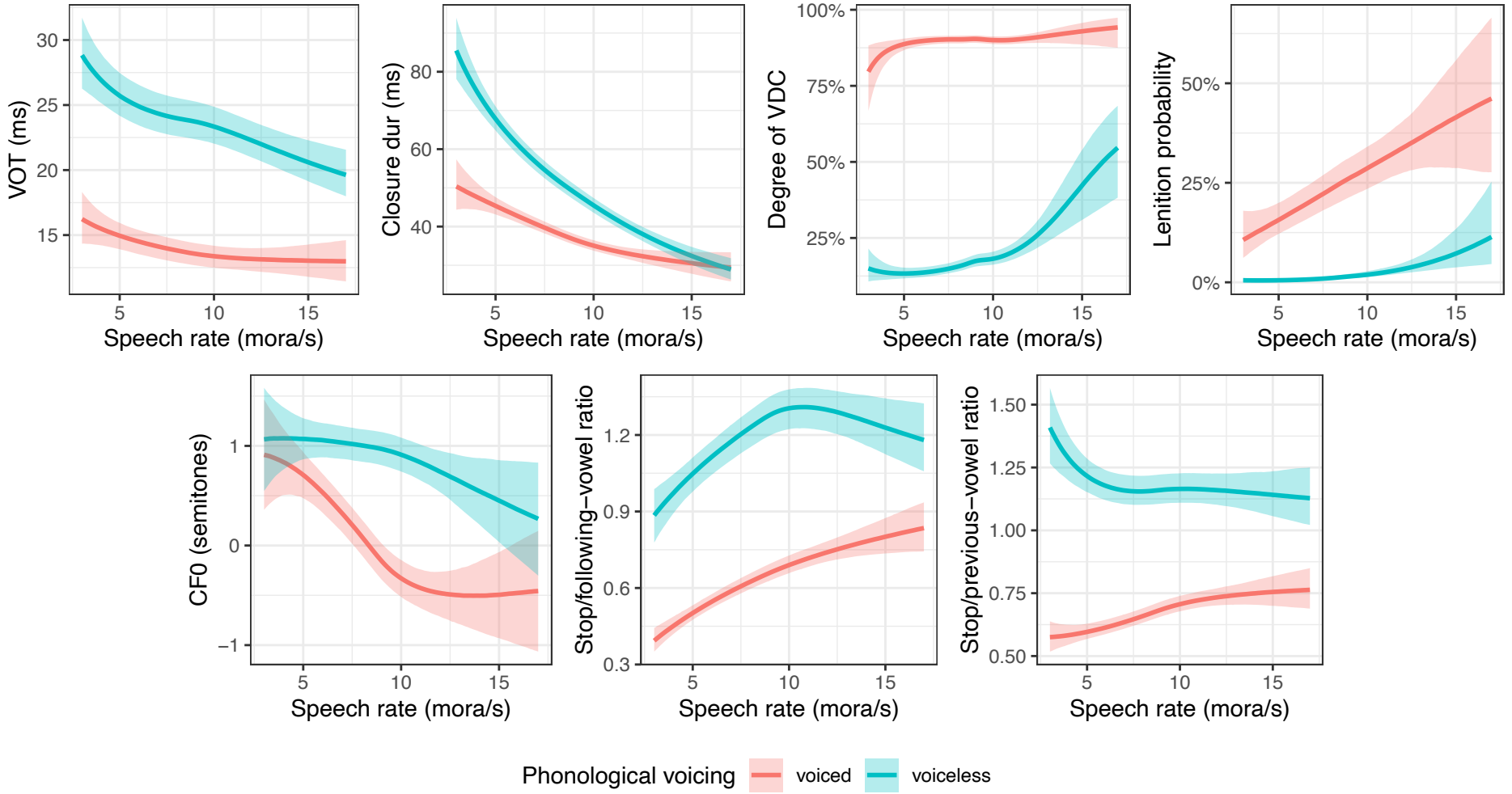
Speaker-level estimates used to address RQ2

Stuart-Smith et al. (2015), Sonderegger et al. (2017), Tanner et al. (2020)

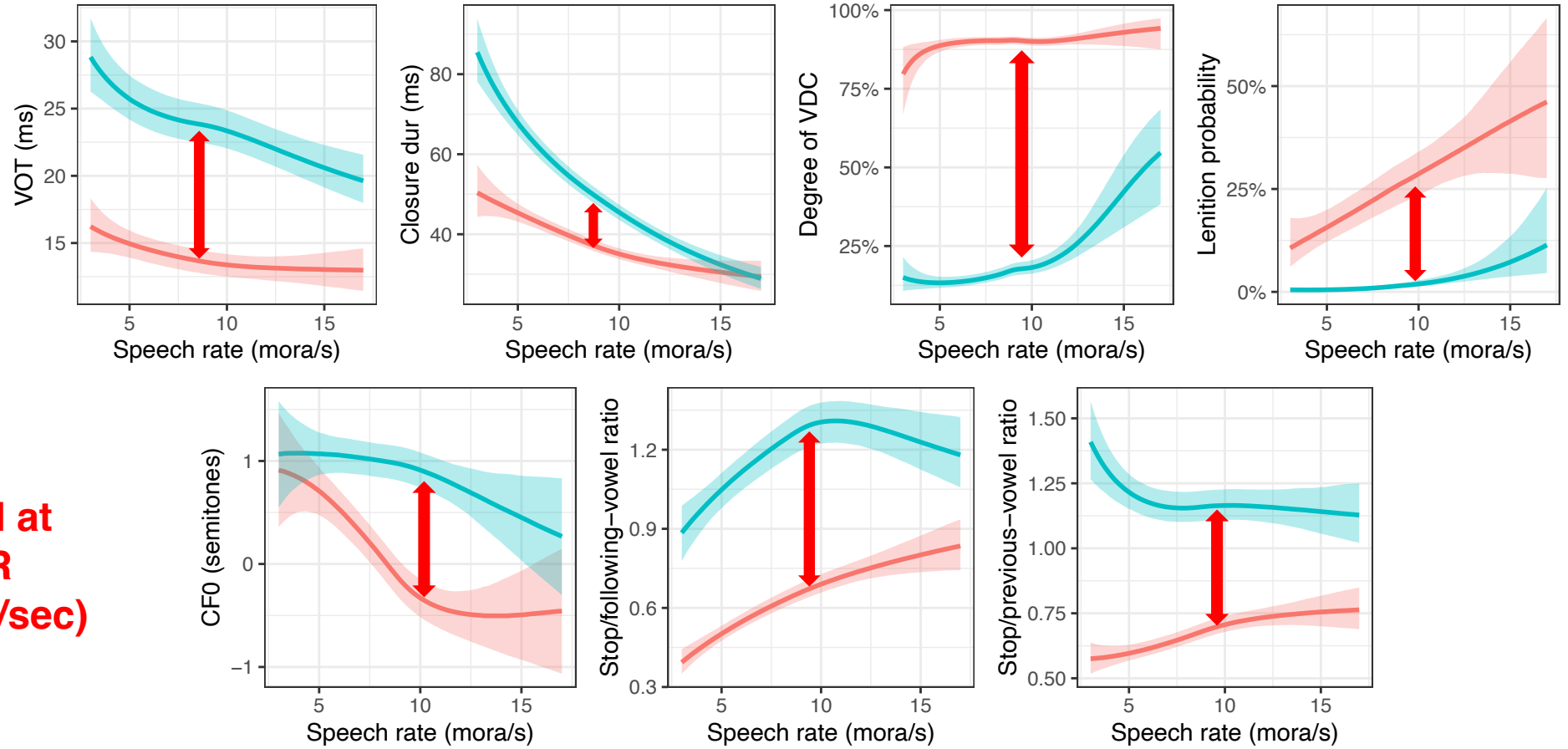
RQ1:

How do differences in speech rate modulate multiple cues to the stop voicing contrast?

RQ1: Speech rate effects on voicing cues



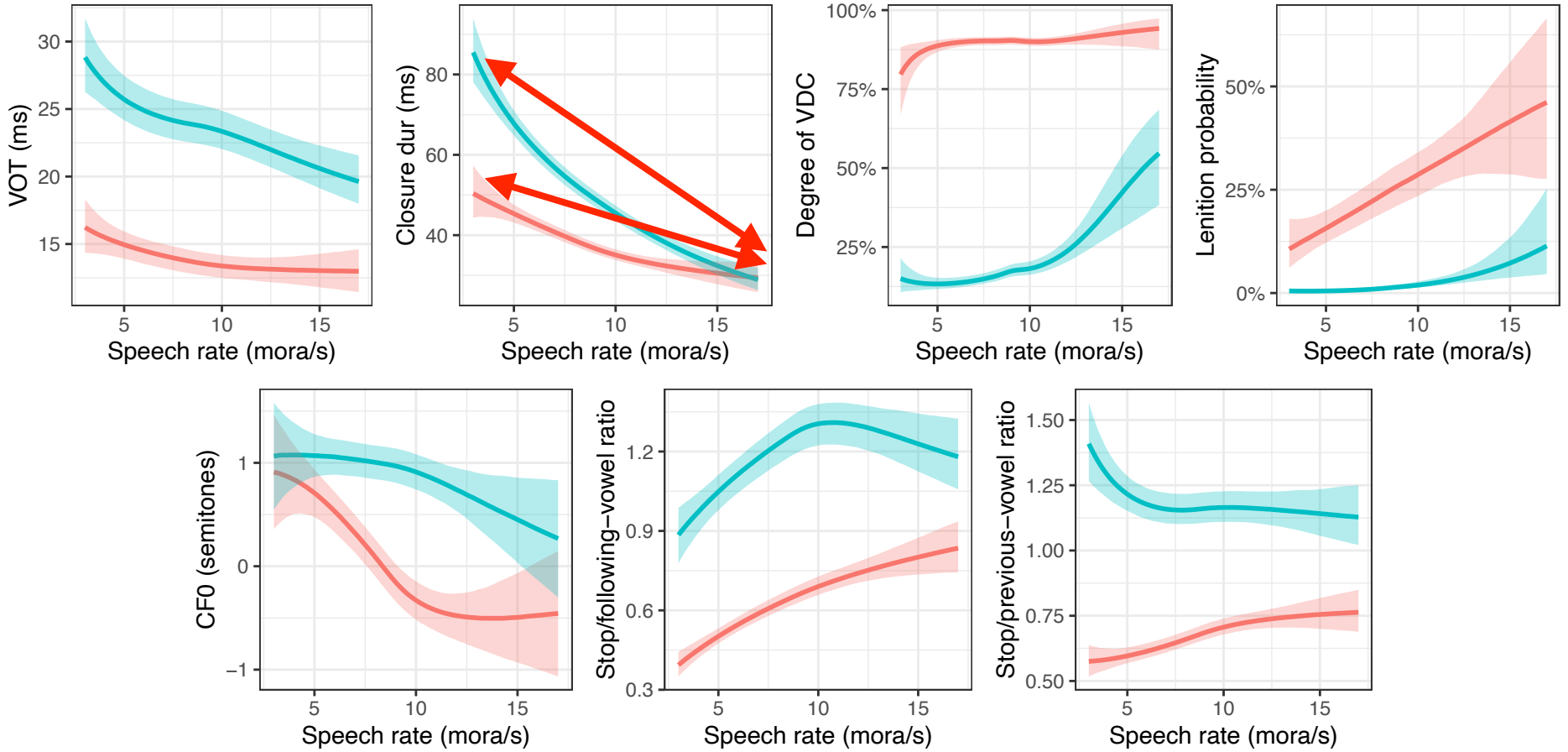
RQ1: Speech rate effects on voicing cues



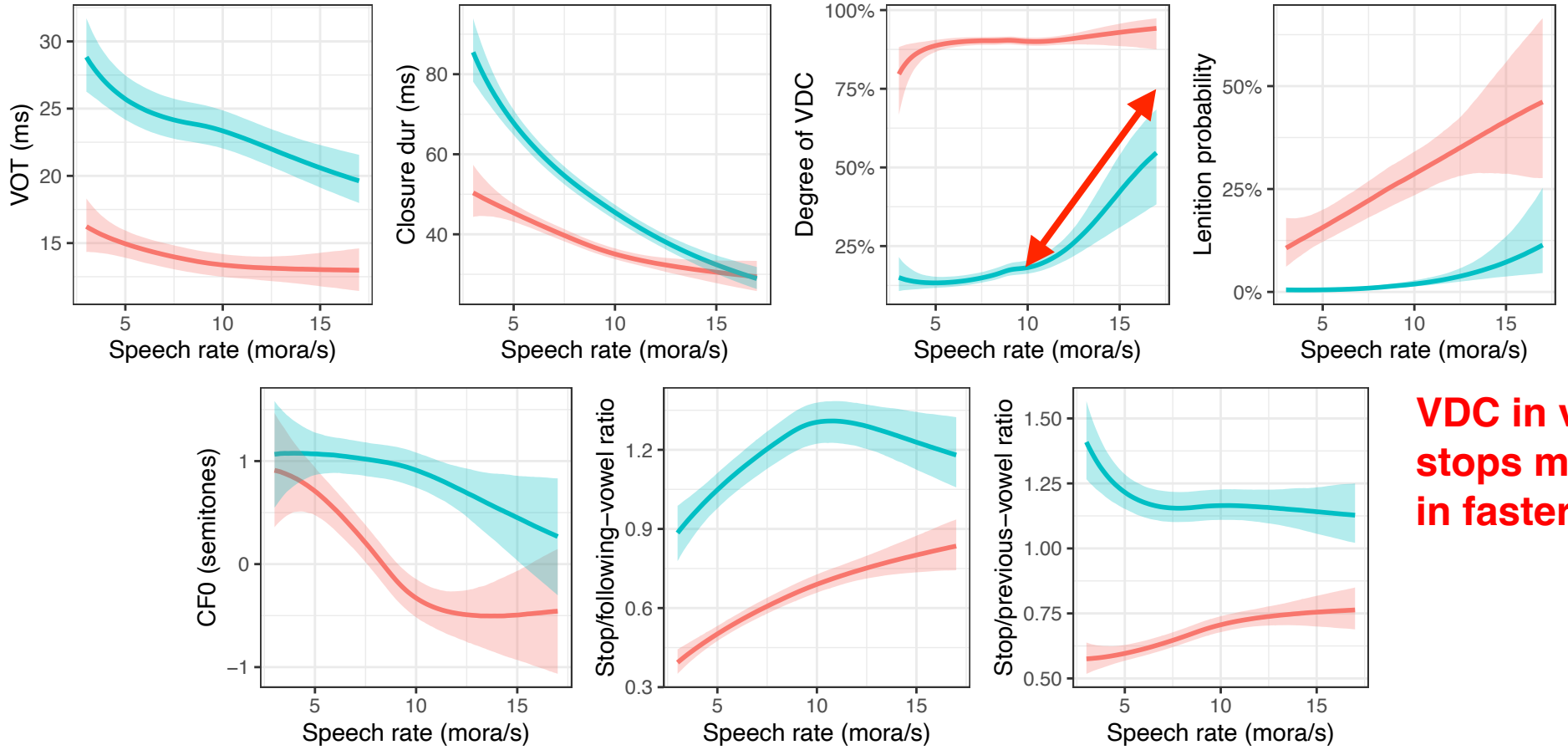
Voicing contrast maintained at average SR (8-10 mora/sec)

RQ1: Speech rate effects on voicing cues

Smaller increase for voiced CD at slower rates



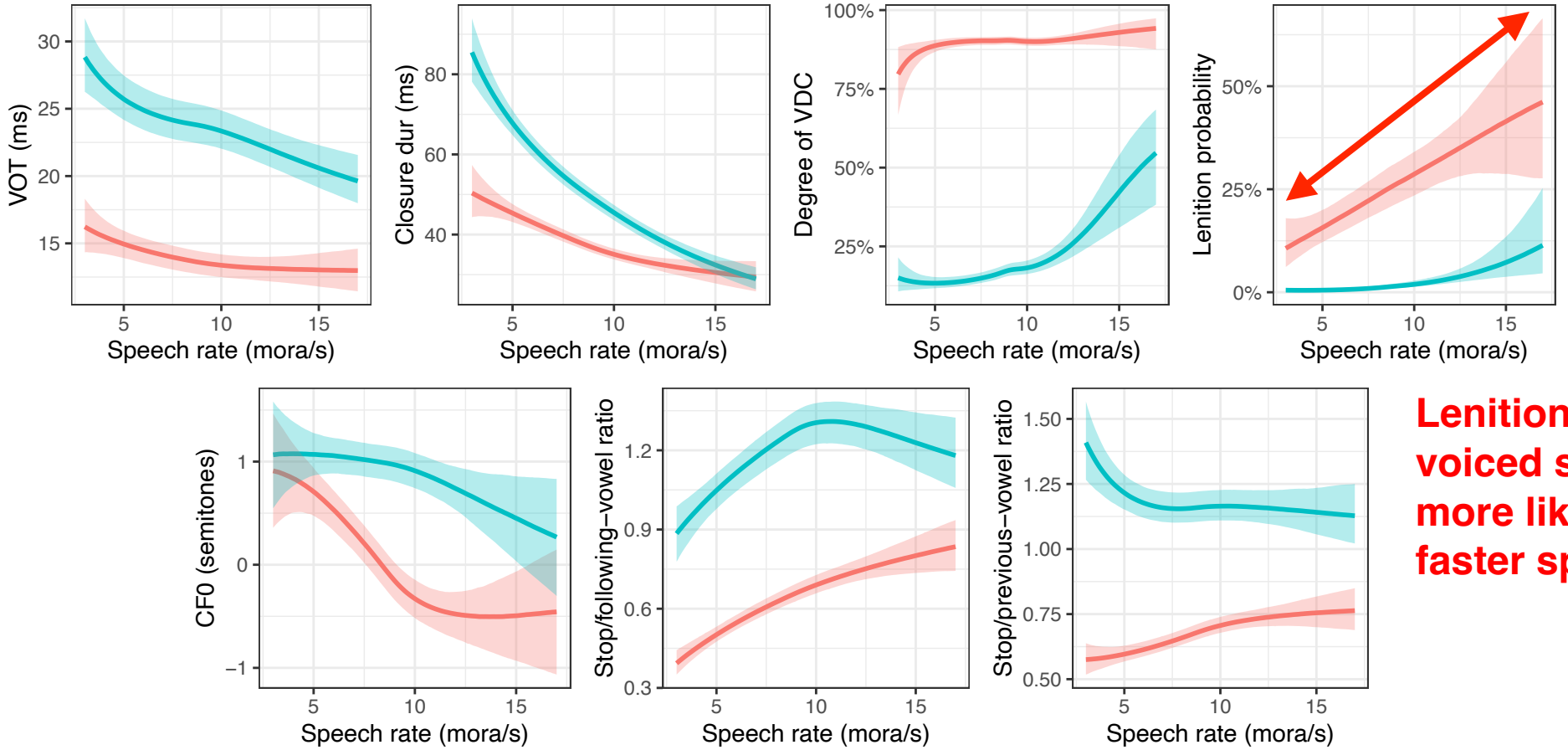
RQ1: Speech rate effects on voicing cues



VDC in voiceless stops more likely in faster speech

(Model estimated values)

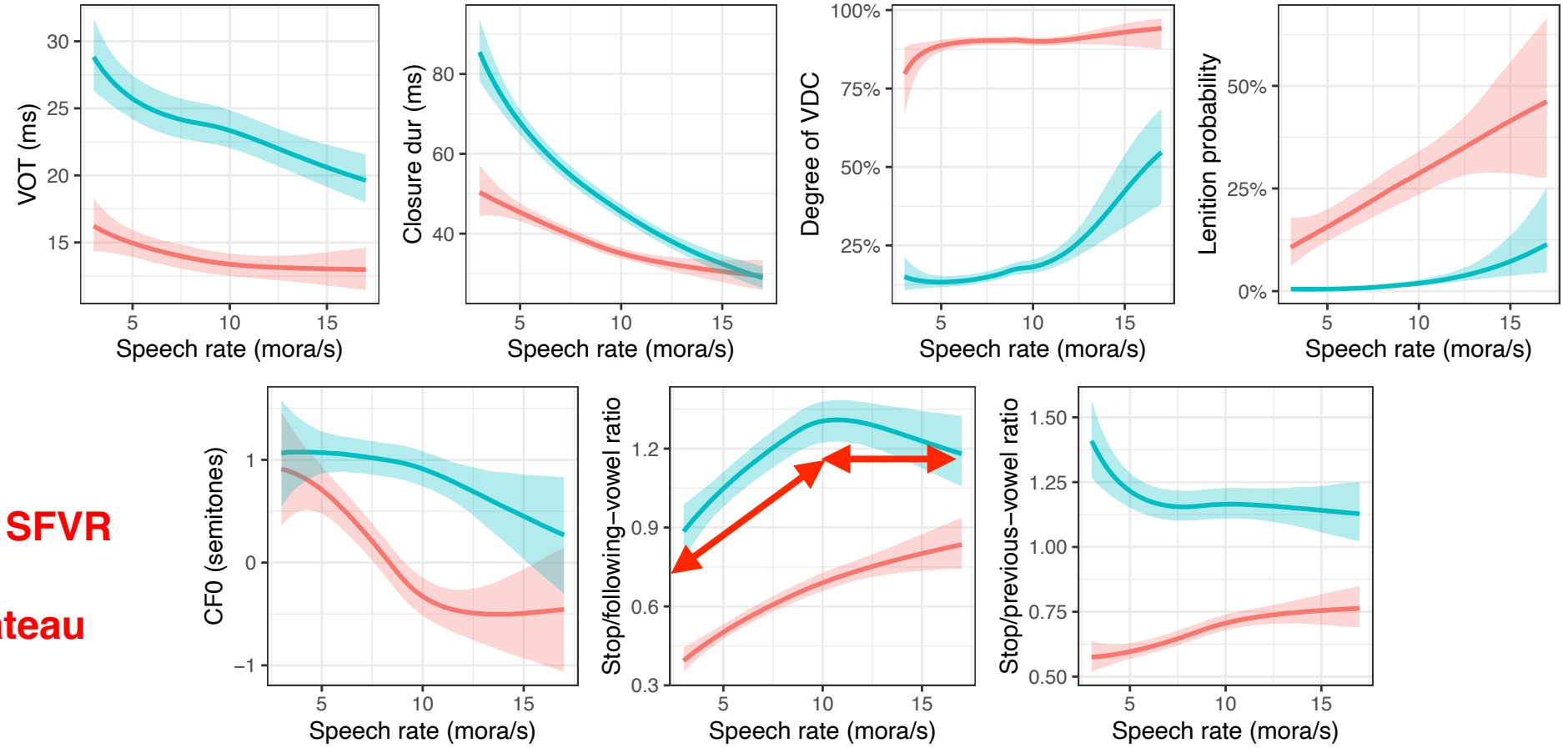
RQ1: Speech rate effects on voicing cues



Lenition of voiced stops more likely in faster speech

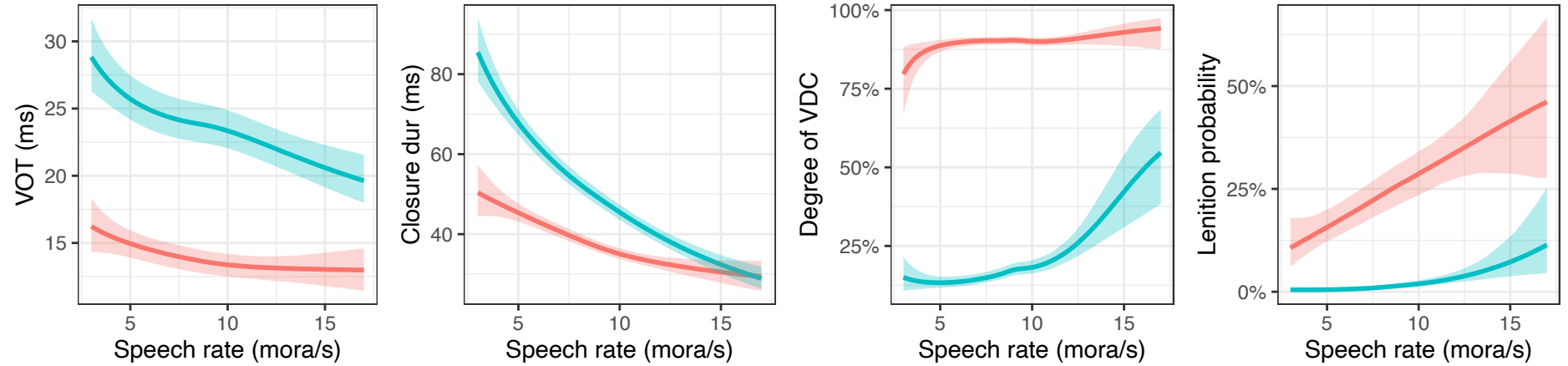
(Model estimated values)

RQ1: Speech rate effects on voicing cues



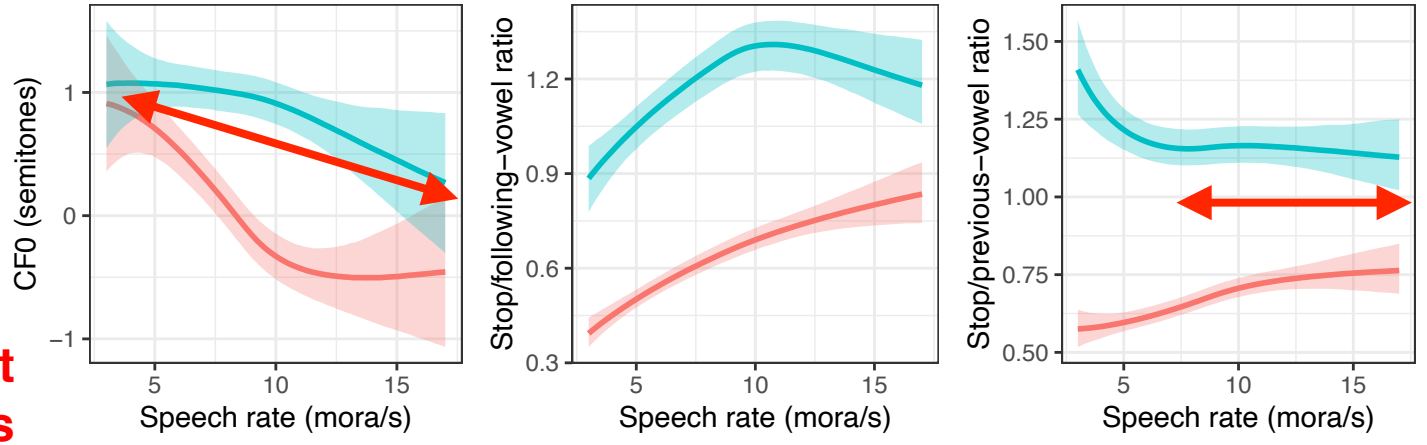
Voiceless SFVR increases before plateau

RQ1: Speech rate effects on voicing cues



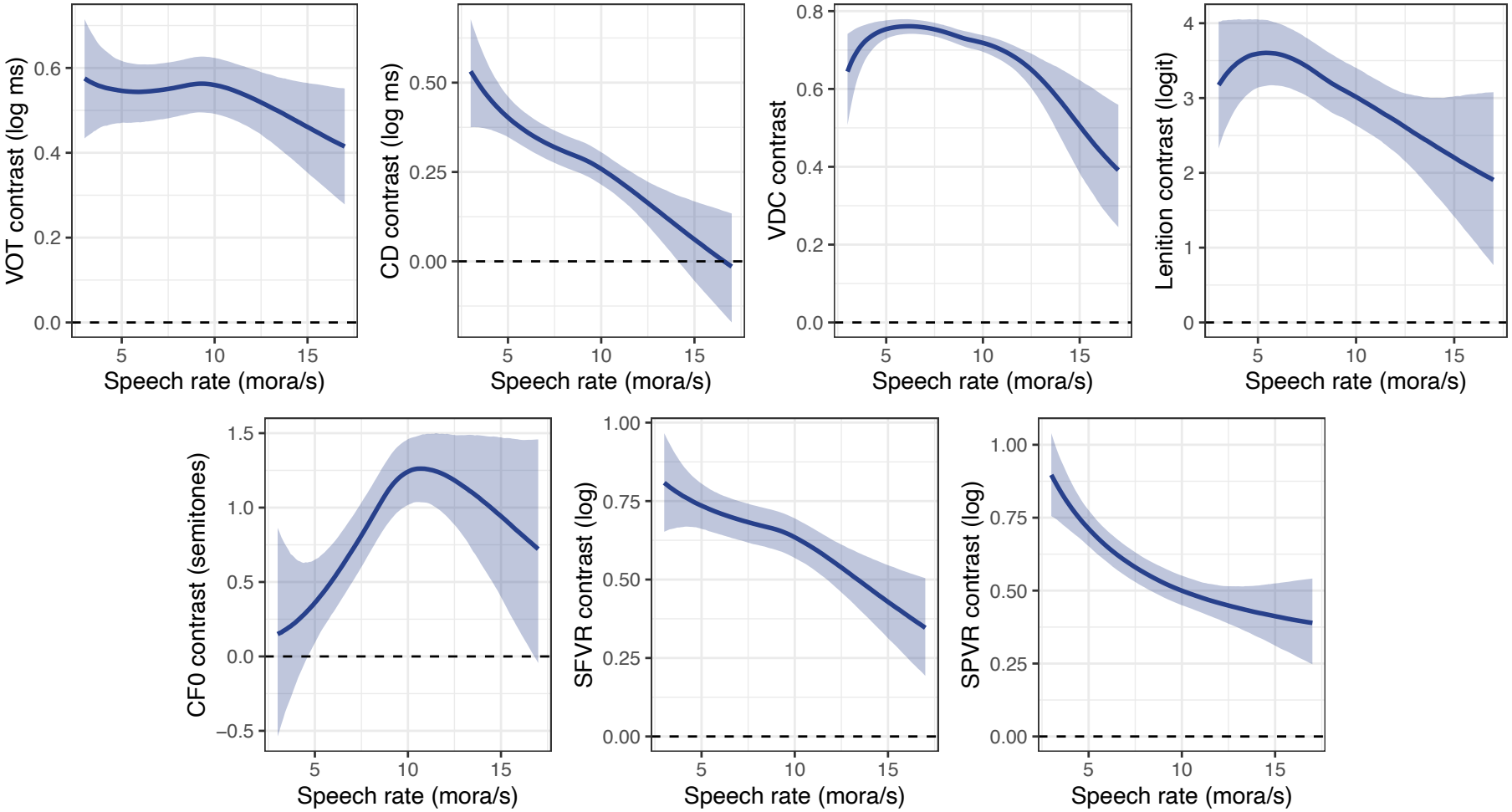
Little evidence for change in voiced CF0

CF0 contrast neutralised at extreme rates



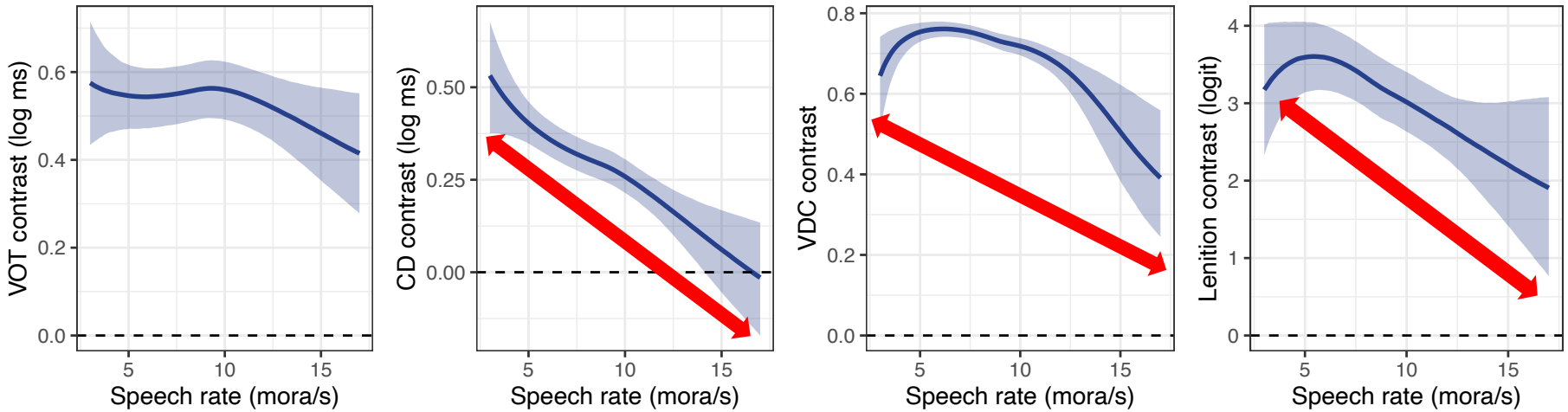
SPVR stable at medium-to-fast rates

RQ1: Speech rate effects on voicing cues

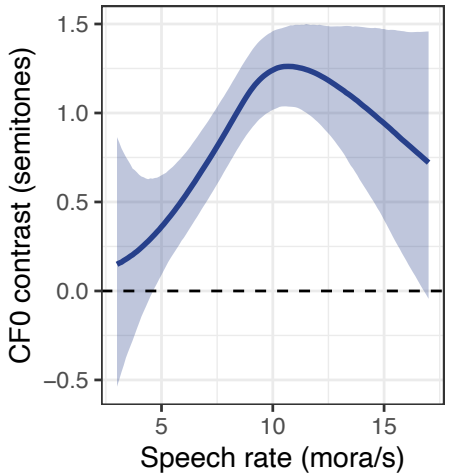


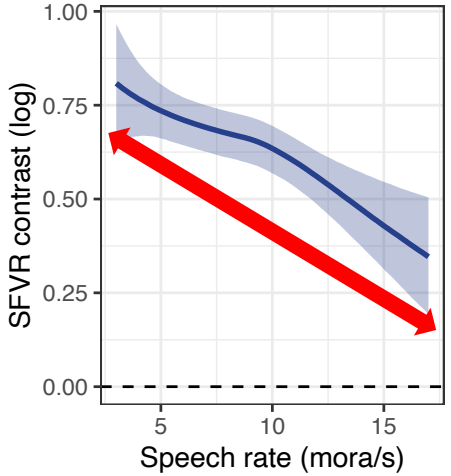
(Model estimated values)

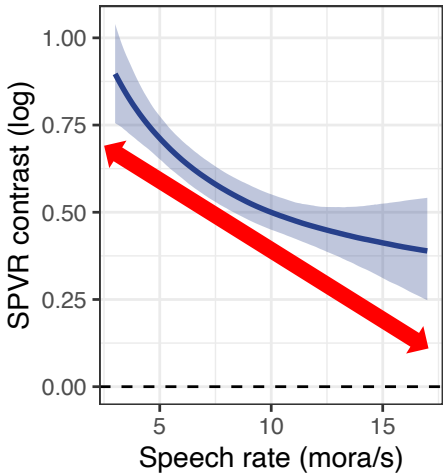
RQ1: Speech rate effects on voicing cues



Faster speech = smaller contrasts



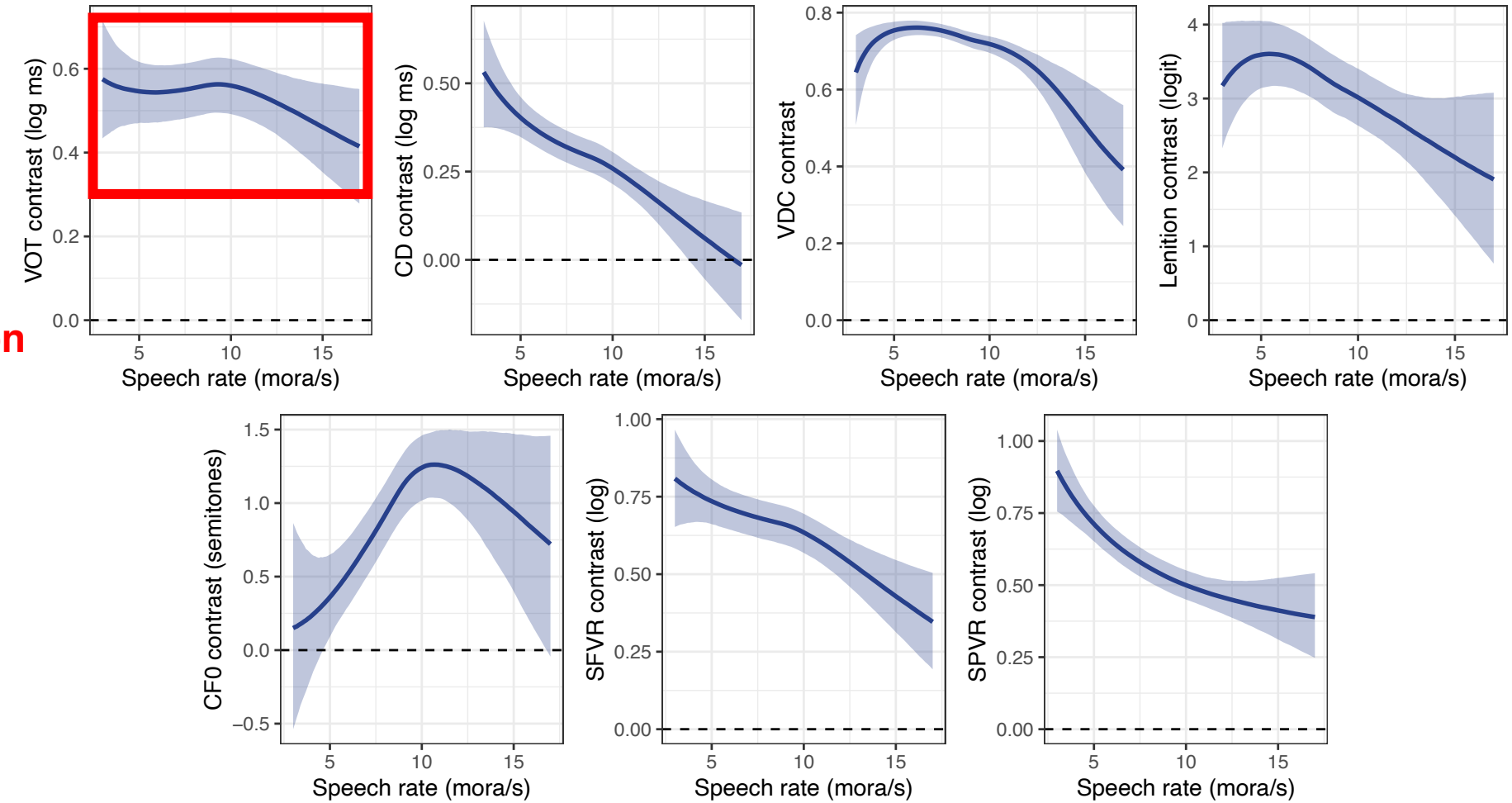




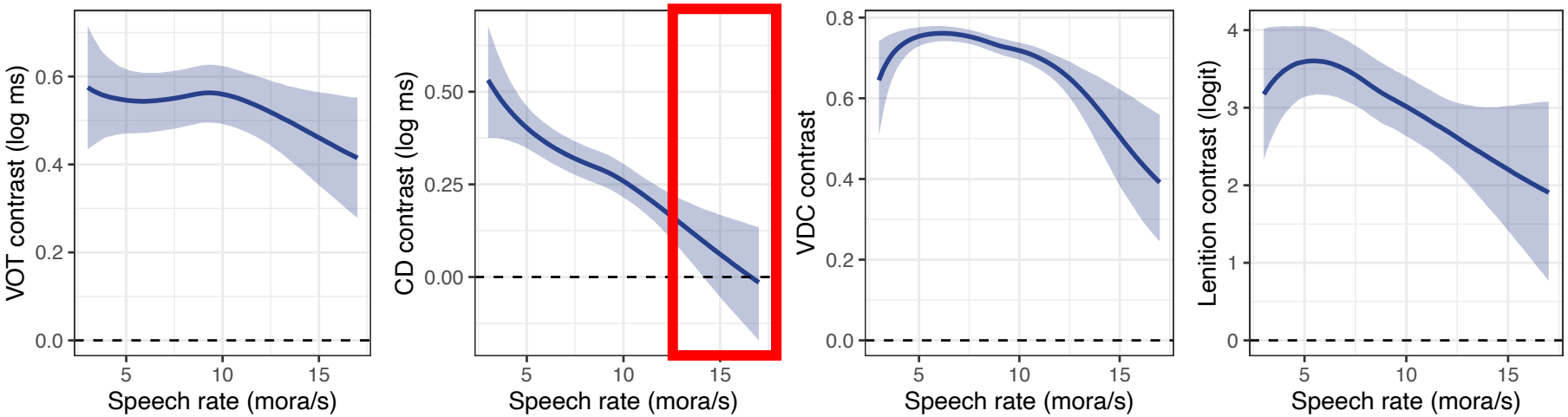
(Model estimated values)

RQ1: Speech rate effects on voicing cues

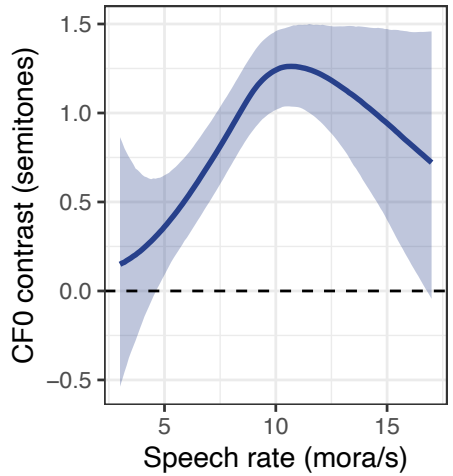
Little reduction of the VOT contrast

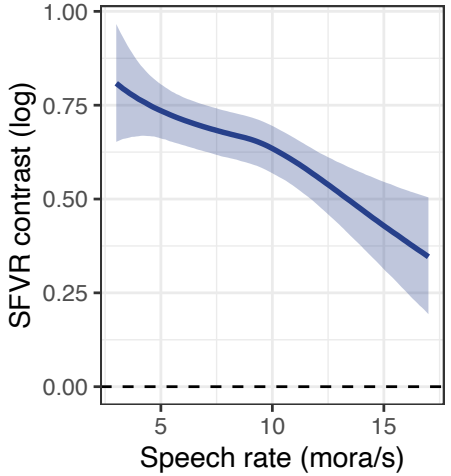


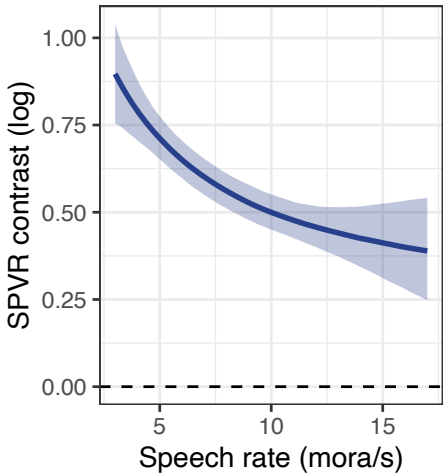
RQ1: Speech rate effects on voicing cues



CD contrast neutralised at ~15 mora/sec

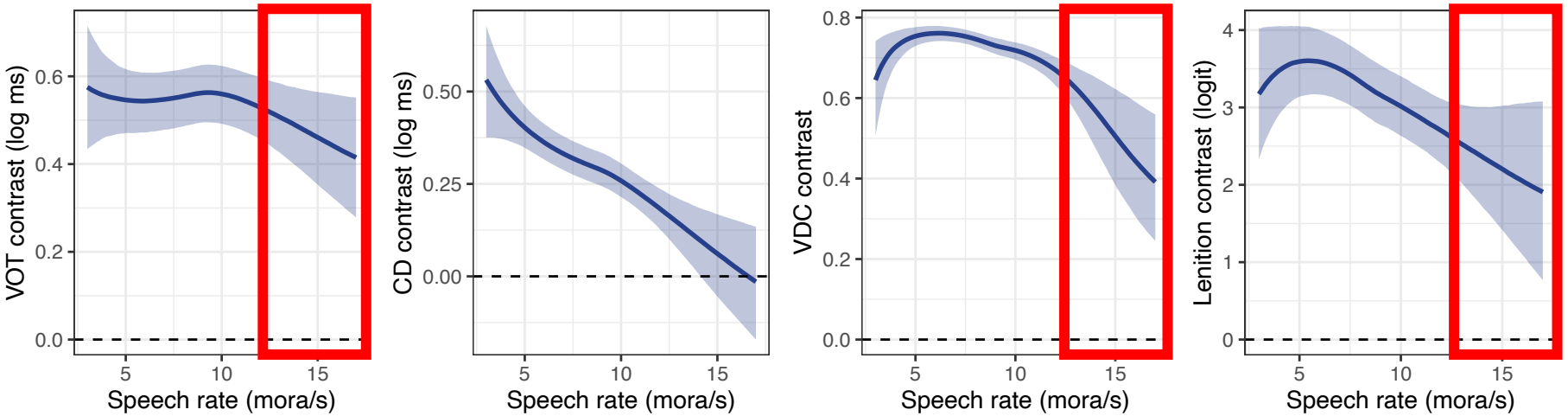




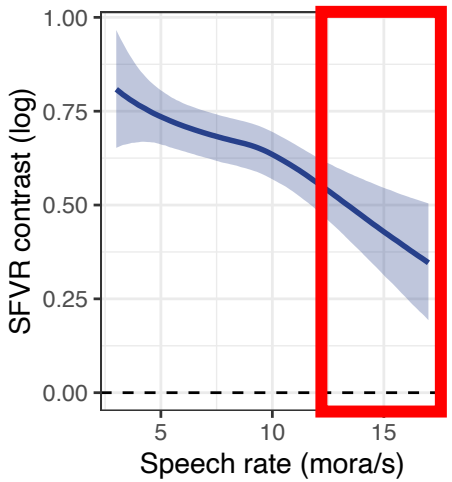


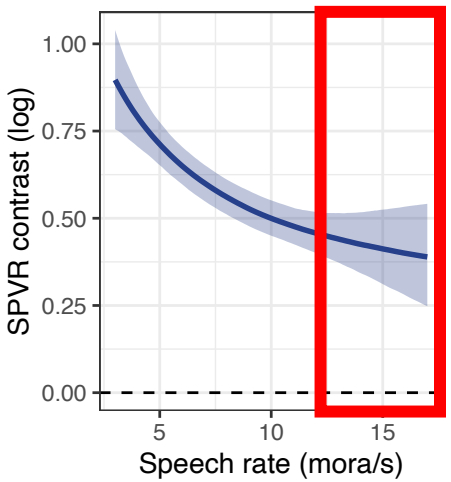
(Model estimated values)

RQ1: Speech rate effects on voicing cues



Contrast still present at fastest rates





RQ1: Summary

- All cues maintain voicing contrast at average speech rates
- Contrast becomes smaller at faster speech rates
  - Voiced CD short at slowest rates — constraint on elongating voiced closures
- Contrast size for VOT not affected by speech rate
- However contrasts are not neutralised at the fastest rates (except for CD)

RQ2:

Does the importance of each cue change at different speech rates?

RQ2: Methods

- *Does the importance of each cue change at different speech rates?*
- Classification experiment
  - Train classifier models to predict voicing labels
  - Compare the performance of classifiers:
    - Trained on different acoustic cues
    - At different speech rates

RQ2: Methods

- Random Forests (RFs)
  - Trained to predict a label for an observation (e.g. voiced vs voiceless) based on set of input variables (e.g. VOT, VDC, etc)
- Useful for comparing *importance of variables*

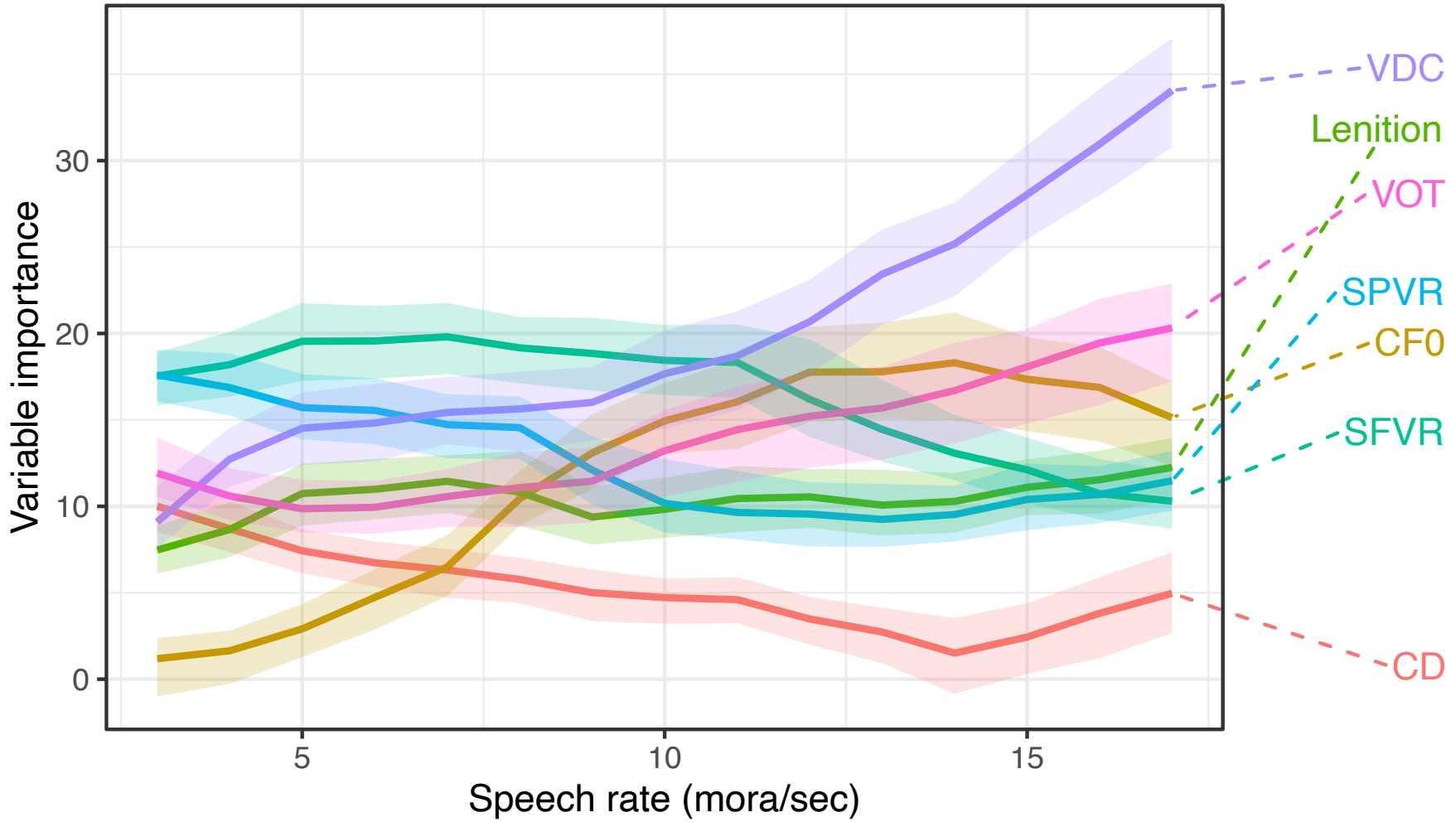
Breiman (2001), Strycharczuk et al. (2020), Sonderegger & Sóskuthy (2024)

RQ2: Methods

- For each 1 mora/sec speech rate interval (3-17 mora/sec):
  1. Train RF on 80% of speaker-predicted cue values (110 speakers)
  2. Predict voicing label from remaining 20% speaker values (27 speakers)
  3. Calculate *variable importance* for each variable
     - Reduction in accuracy if a variable/cue is shuffled

Breiman (2001), Strycharczuk et al. (2020), Sonderegger & Sóskuthy (2024)

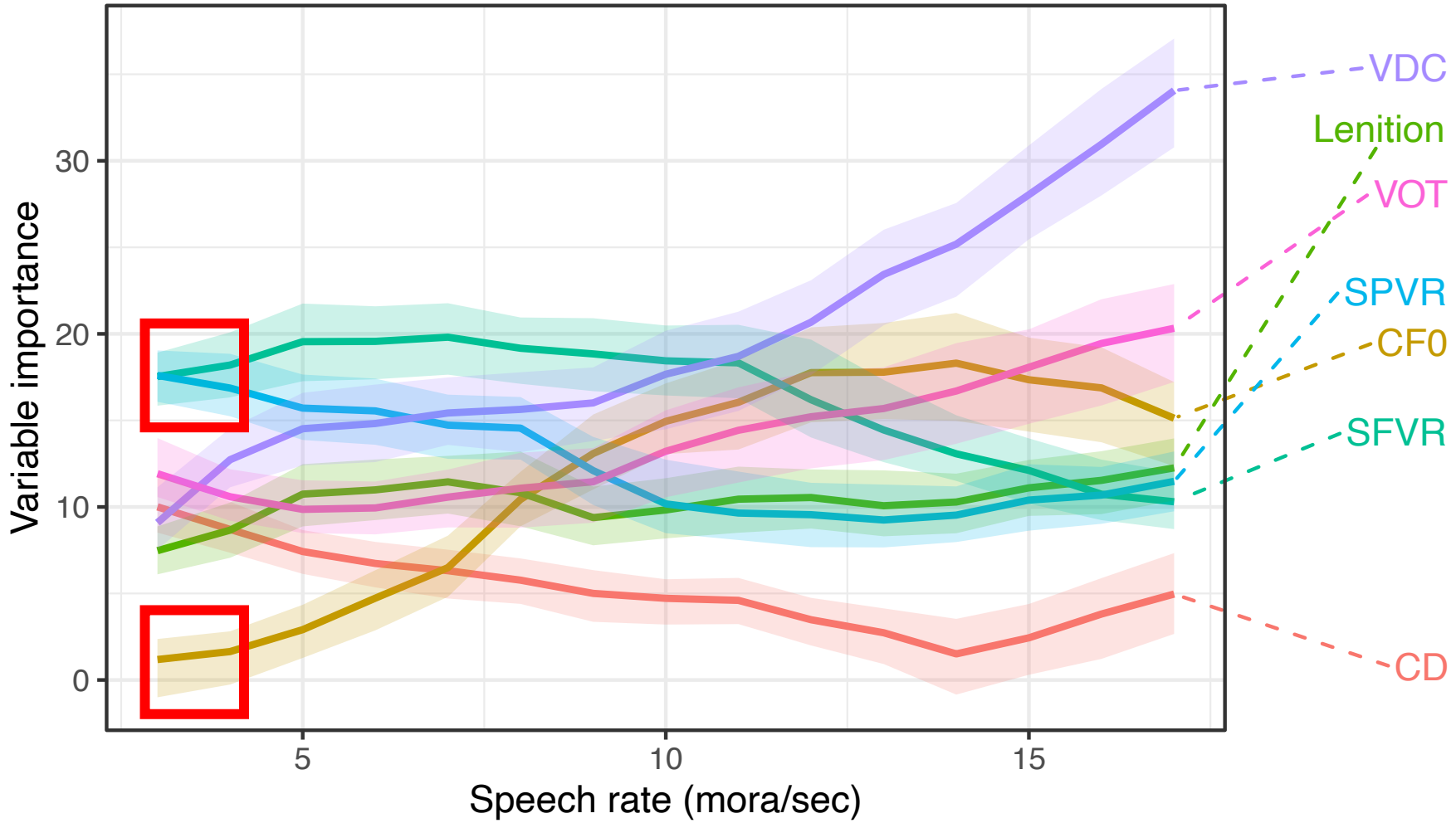
RQ2: Role of cues at different speech rates



RQ2: Role of cues at different speech rates

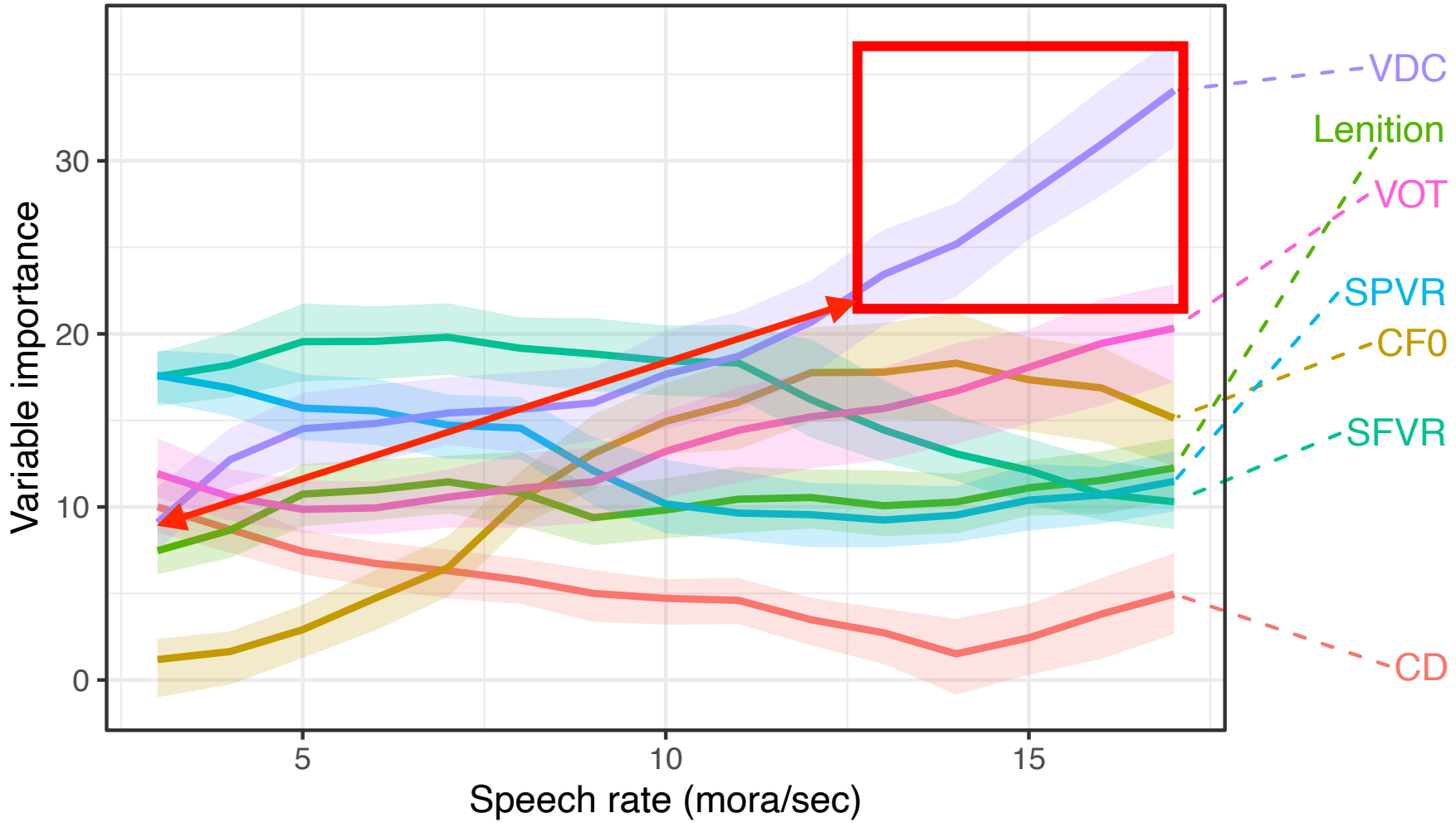
Relational cues most important at slowest rates

CF0 least important

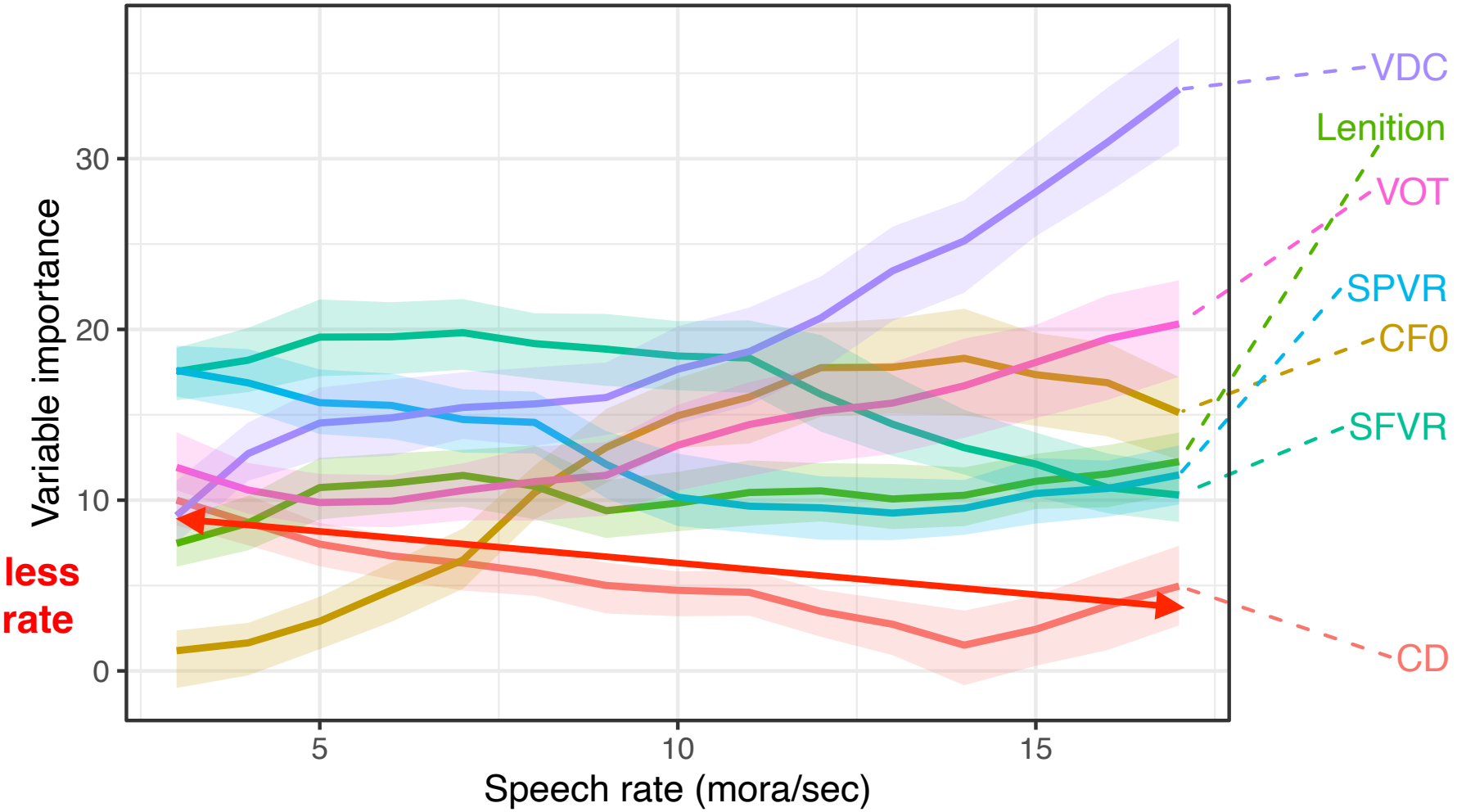


RQ2: Role of cues at different speech rates

VDC most important at fastest rates



RQ2: Role of cues at different speech rates



CD becomes less important as rate increases

Discussion

Summary

- RQ1: *How do differences in speech rate modulate multiple cues to the stop voicing contrast?*
  - Faster speech rate reduces both *absolute* cue values and the size of the voicing contrast
  - VOT contrast not affected by speech rate
    - Different from e.g. English
  - However only the contrast for Closure Duration is neutralised at the fastest speech rates

Summary

- RQ2: *Does the importance of each cue change at different speech rates?*
  - At slow speech rates, the voicing contrast is mainly cued by *relative* temporal cues — ratios between the stop and the following/previous vowel
  - VDC is the most prominent acoustic cue in fast speech

Discussion

- Supports finding that relative durational cues are informative at slow speech rates
- Importance of an acoustic cue is *contextual*
  - Some cues may be more prominent in extremely fast/slow speech
- Speech rate normalisation is also multidimensional
  - Listeners may adjust perception of multiple acoustic cues, but also which cues are attended to

Discussion

- Would we expect this rate normalisation to be language-specific?
  - Would a language with a predominantly VOT-based contrast (e.g. English) also exhibit rate-based changes in other cues?
- How do listeners track speech rate differences?
- How would this rate-specific cue prominence interact with other effects on segmental realisation (prosodic structure, lexical frequency, etc)?

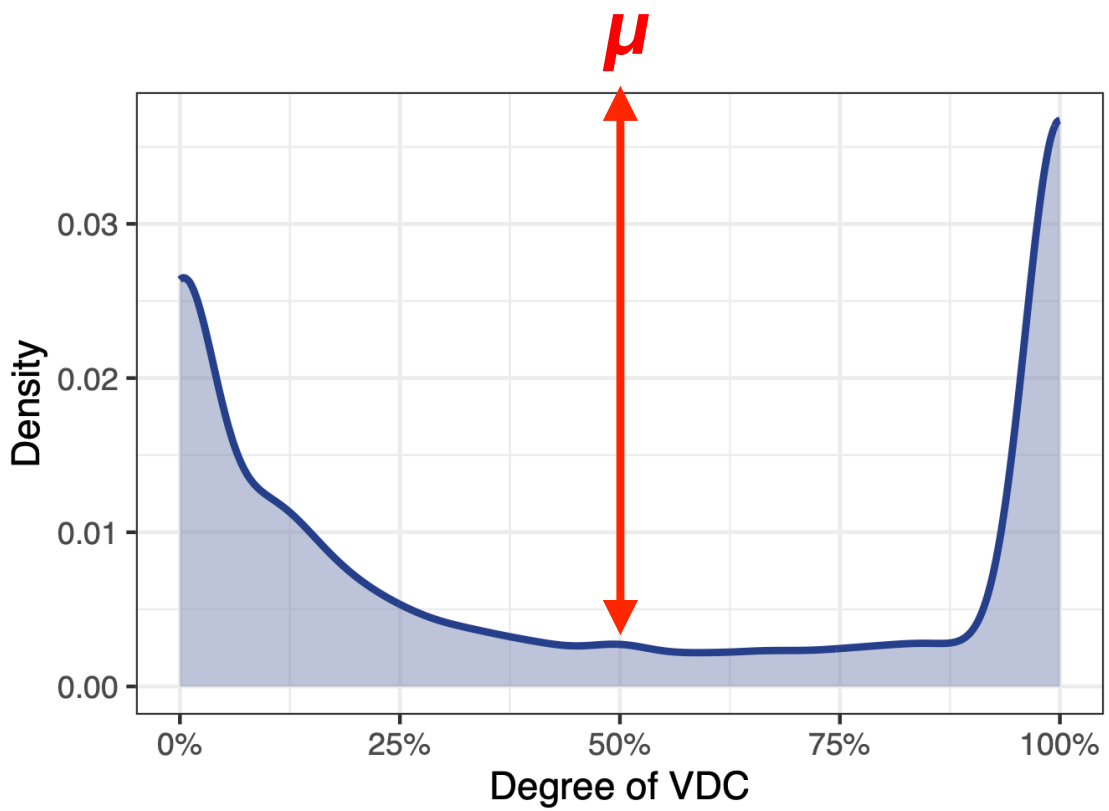
Thank you!

- Morgan Sonderegger
- Jane Stuart-Smith
- NINJAL Collaborative Research Project:
  - “Evidence-based Study on the Intonational Diversity of Japanese and Ryukyuan”
- British Academy

Extra slides

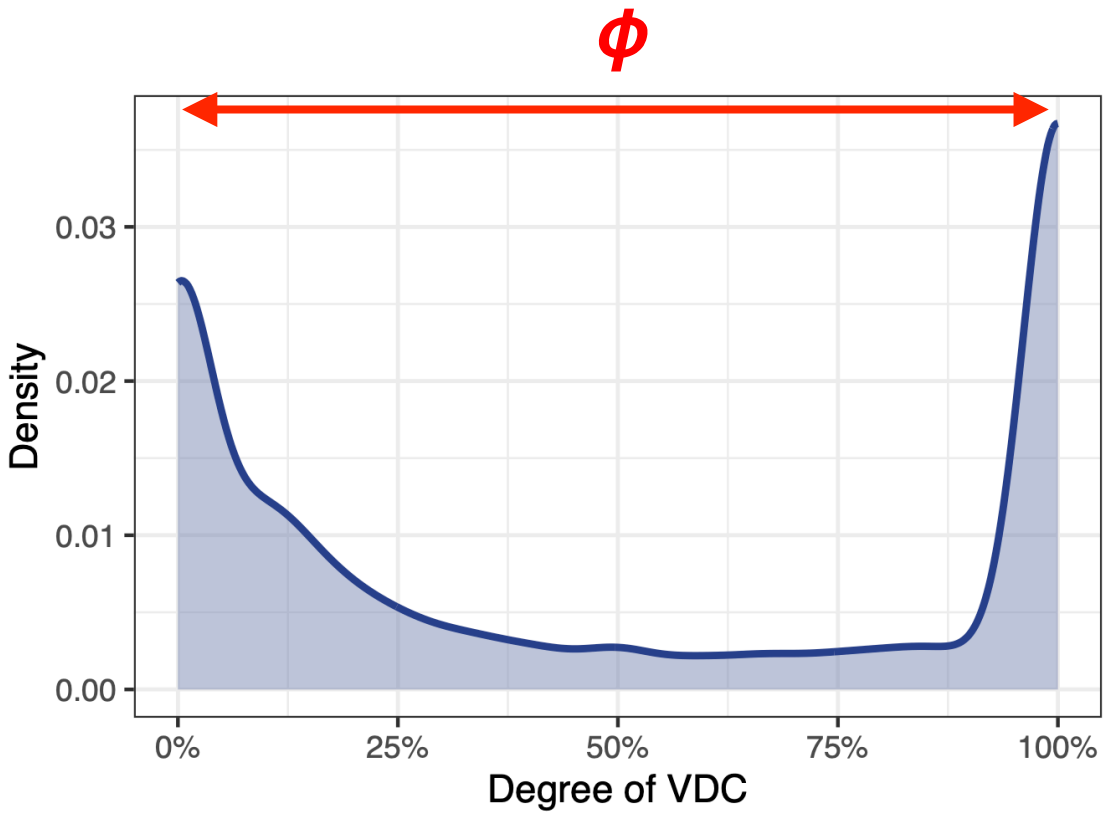
VDC: Zero-One-Inflated-Beta

- Model *distribution* of bounded variables (e.g. percentages)
- μ = central tendency



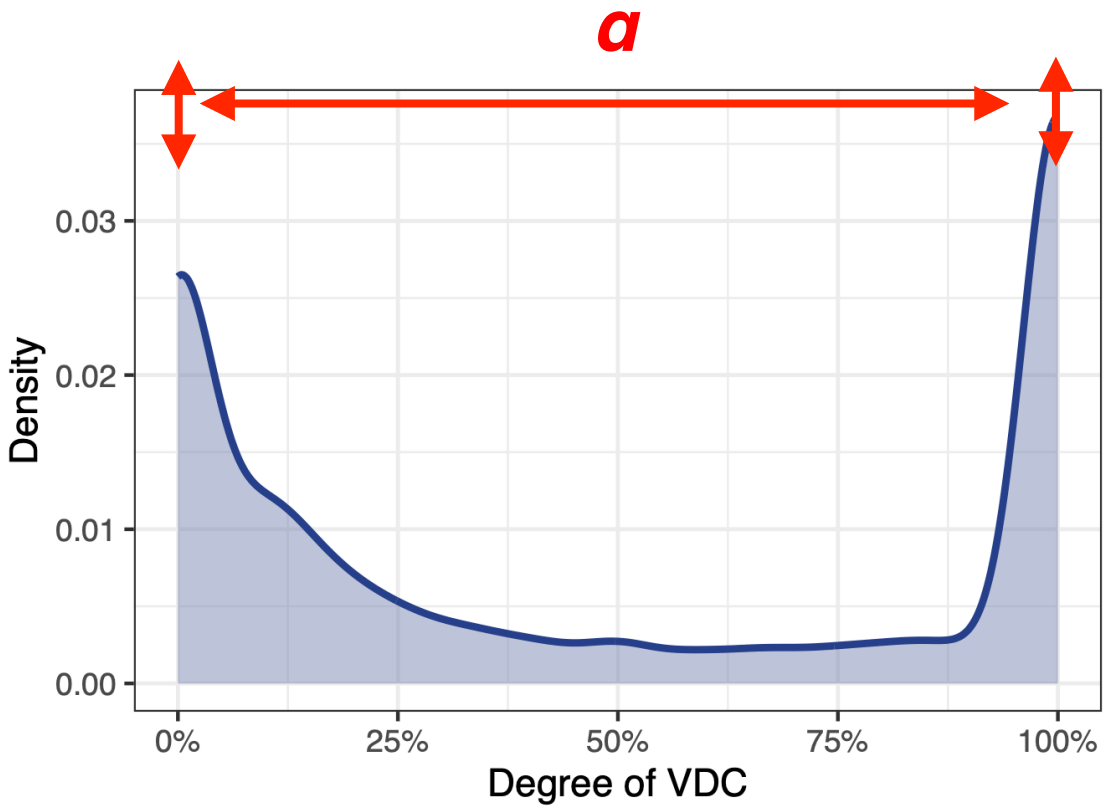
VDC: Zero-One-Inflated-Beta

- Model *distribution* of bounded variables (e.g. percentages)
- μ = central tendency
- ϕ = spread



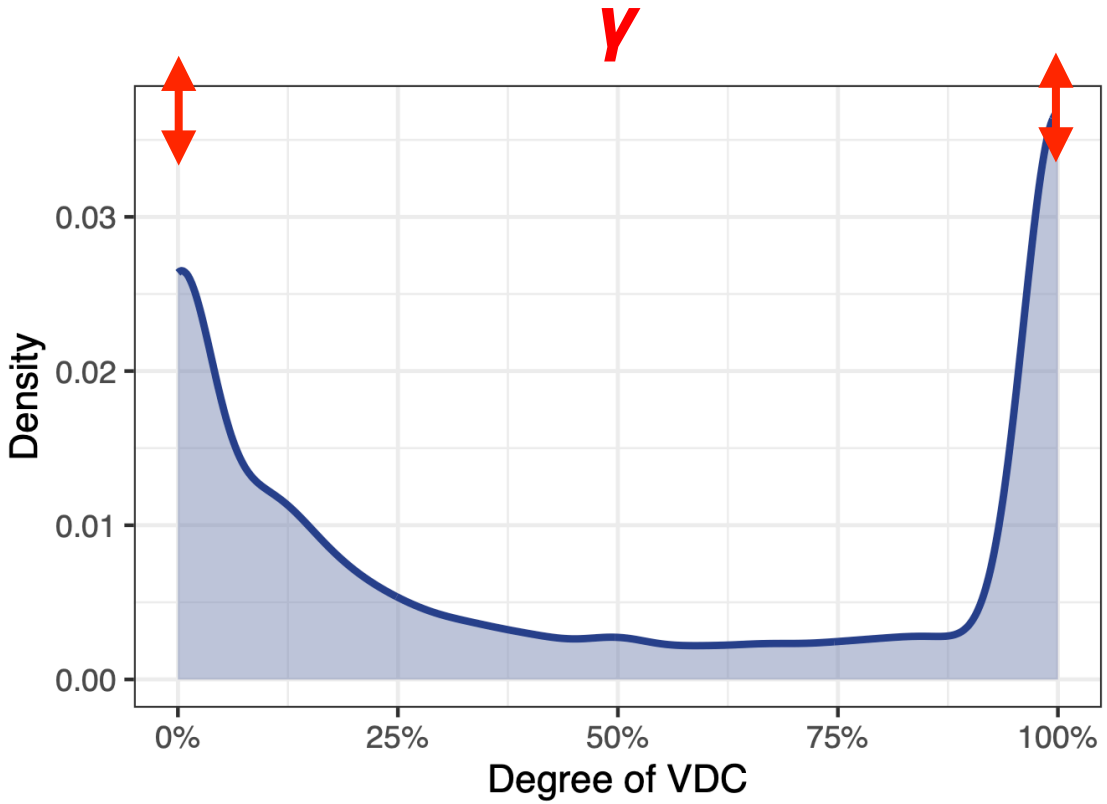
VDC: Zero-One-Inflated-Beta

- Model *distribution* of bounded variables (e.g. percentages)
- μ = central tendency
- ϕ = spread
- α = probability of {0,1} or [0,1]



VDC: Zero-One-Inflated-Beta

- Model *distribution* of bounded variables (e.g. percentages)
- μ = central tendency
- ϕ = spread
- α = probability of {0,1} or [0,1]
- γ = probability of 0 or 1 (i.e. logistic)



Priors

- Intercept ~ Normal(0,2)
- β ~ Normal(0,2)
- σ ~ Exponential(1)
- Corr ~ LKJ(1)

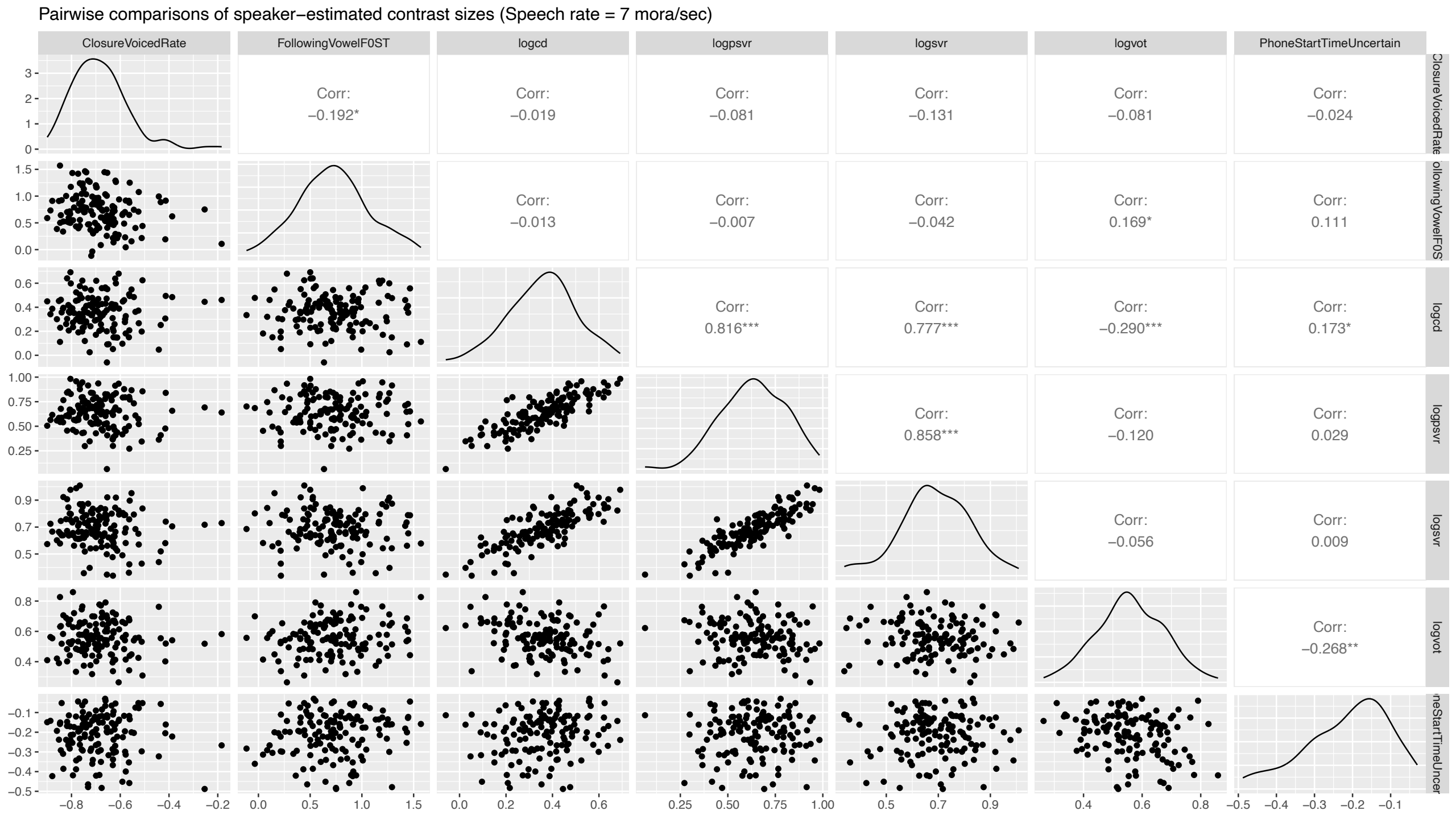
RQ1: Speech rate effects on voicing cues

- Evaluating evidence for speech rate effects
  1. Fit subset models *without* each effect of interest
     - Voicing, speech rate & speech rate-voicing interaction terms
  2. Compare the predictive accuracy between full and subset models with estimated LOO cross-validation (PSIS-LOO)

Vehtari et al. (2017), Paape & Vasishth (2022), Vehtari et al. (2023)

Table 1: Estimated log predictive densities ($\widehat{elpd}$) and standard error (SE) for the "Full" model (containing all specified model terms) and each nested model with respective terms removed. The model with the highest $\widehat{elpd}$ is shown in bold.

Cue	Model	$\widehat{elpd}$	SE
VOT	Full	−20635.29	266
	No voicing	−22829.36	262
	No speech rate	−20851.51	266
	No voicing × SR	**−20634.38**	266
CD	Full	**−29138.78**	505
	No voicing	−31805.45	473
	No speech rate	−31007.64	482
	No voicing × SR	−29176.30	505
VDC	Full	**−50881.73**	277
	No voicing	−88230.24	171
	No speech rate	−51299.38	277
	No voicing × SR	−51027.41	277
Lenition	Full	**−30928.73**	170
	No voicing	−33841.75	173
	No speech rate	−31216.29	171
	No voicing × SR	−30959.76	170
CF0	Full	**−232300.98**	214
	No voicing	−232835.92	212
	No speech rate	−232538.39	213
	No voicing × SR	−232366.34	214
SFVR	Full	**−64256.78**	280
	No voicing	−68256.61	269
	No speech rate	−64947.62	282
	No voicing × SR	−64320.87	280
SPVR	Full	**−54944.49**	290
	No voicing	−58902.29	280
	No speech rate	−55224.91	292
	No voicing × SR	−55097.43	291


Pairwise comparisons of speaker−estimated contrast sizes (Speech rate = 7 mora/sec)
ClosureVoicedRate
FollowingVowelF0ST
logcd
logpsvr
logsvr
logvot
PhoneStartTimeUncertain
Corr: −0.192*
Corr: −0.019
Corr: −0.081
Corr: −0.131
Corr: −0.081
Corr: −0.024
Corr: −0.013
Corr: −0.007
Corr: −0.042
Corr: 0.169*
Corr: 0.111
Corr: 0.816***
Corr: 0.777***
Corr: −0.290***
Corr: 0.173*
Corr: 0.858***
Corr: −0.120
Corr: 0.029
Corr: −0.056
Corr: 0.009
Corr: −0.268**