The SPADE project:
large-scale analysis of a spoken language across space and time

Morgan Sonderegger,
The SPADE Consortium

LVC Group, University of Edinburgh
12 July, 2019

SPADE

SPeech Across Dialects of English

August 2017 – July 2020

http://spade.glasgow.ac.uk/

Trans-Atlantic Platform Social Sciences and Humanities

DIGGING INTO DATA CHALLENGE

THE UNIVERSITY of EDINBURGH

University of Glasgow

McGill

NC STATE UNIVERSITY

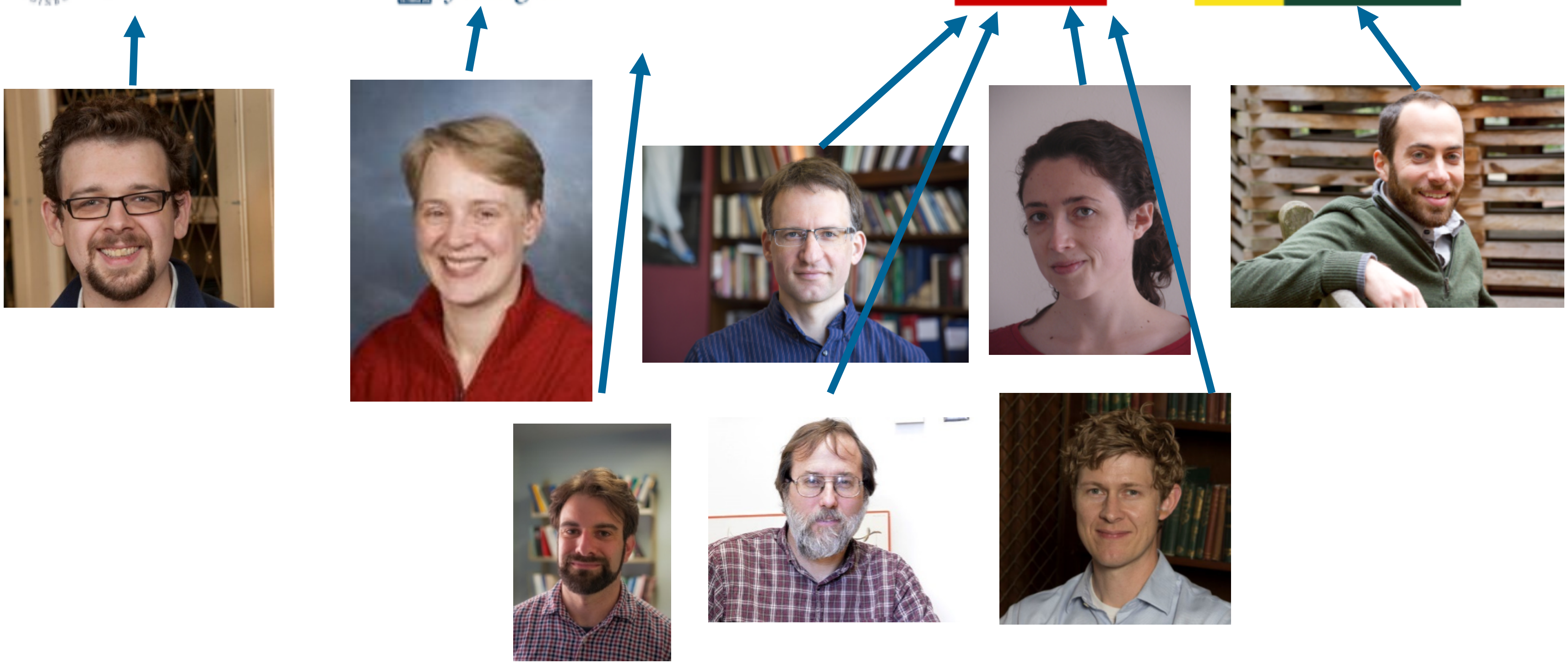
UNIVERSITY OF OREGON

SPADE

SPeech Across Dialects of English

investigators

THE UNIVERSITY of EDINBURGH | University of Glasgow | McGill | NC STATE UNIVERSITY | UNIVERSITY OF OREGON



http://spade.glasgow.ac.uk/

SPADE

SPeech Across Dialects of English

Postdocs

Project manager + data

Software development

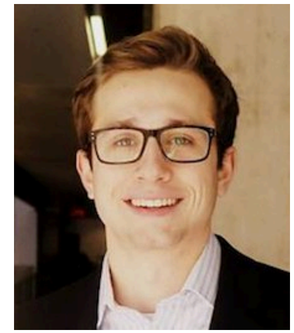
Graduates + undergrads: many

http://spade.glasgow.ac.uk/

- Developers

Michael McAuliffe

Arlie Coles (U. de Montréal)



Elias Stengel-Eskin (Johns Hopkins)

Michael Goodale, Sarah Mihuc (McGill)

- Docs, QA

James Tanner, Vanna Willerton (McGIll)

SPADE

SPeech Across Dialects of English

Project goals

- Software for large-scale automated analysis of speech datasets
- Data from ~40 dialects
  - public & private
  - Focus: sociolinguistic data
- Case studies: investigate how 'English' varies in time and space

Motivation

- Huge amount of annotated speech data exists
  - Corpora,
    academic labs,
    fieldwork…
- Across
  - Languages/dialects
  - Speech styles
  - Time
- + ever-better (semi)-automatic speech measurement tools

Motivation

- Great potential for speech analysis for different purposes

- Requires software for unified corpus analysis
  - Integrating speech datasets
  - Querying across them

- SPADE focus: sociolinguistic, phonetic datasets

Barriers

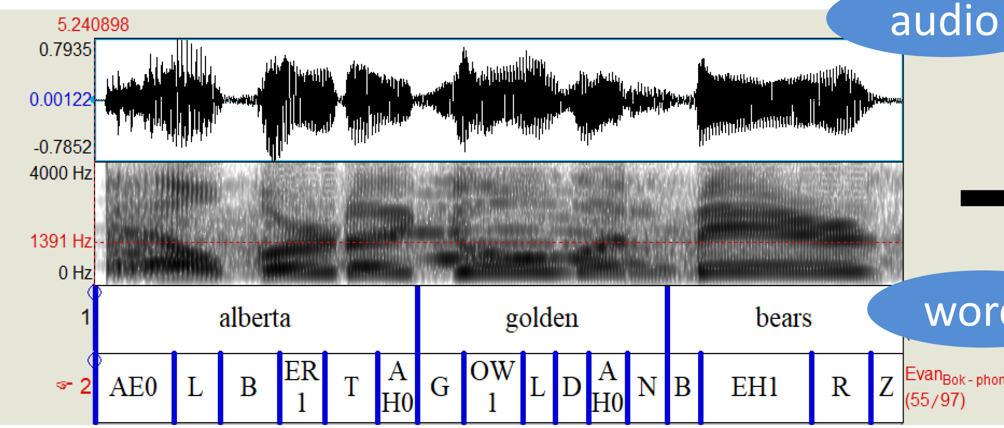
- Speech datasets:
  - Large
  - Complex
  - Diverse formats
- Access to many speech datasets
  - Costly or ethically restricted

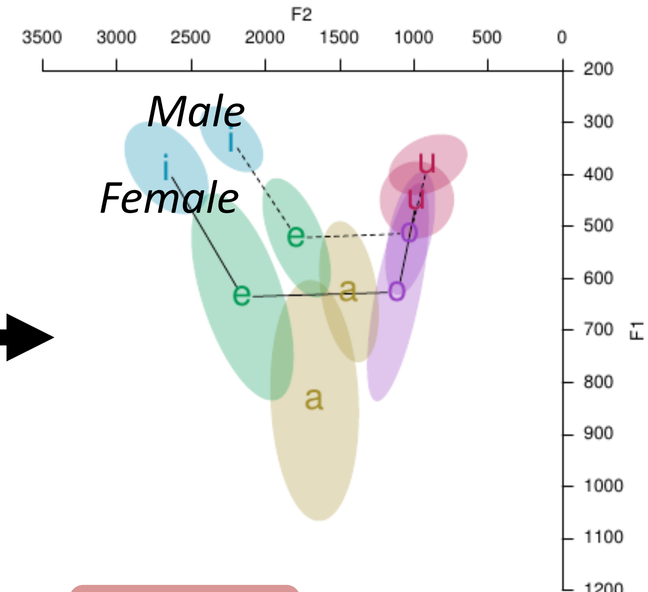
Most sociolinguistic, laboratory data

- Result: requires lots of specialized code, $$, effort, computational power

Raw data

Analysis





audio

words

phones

utterances

<u>Speaker M01</u>
Gender: M
Age: 35
…

speaker questionnaire

Gender

formants

Age

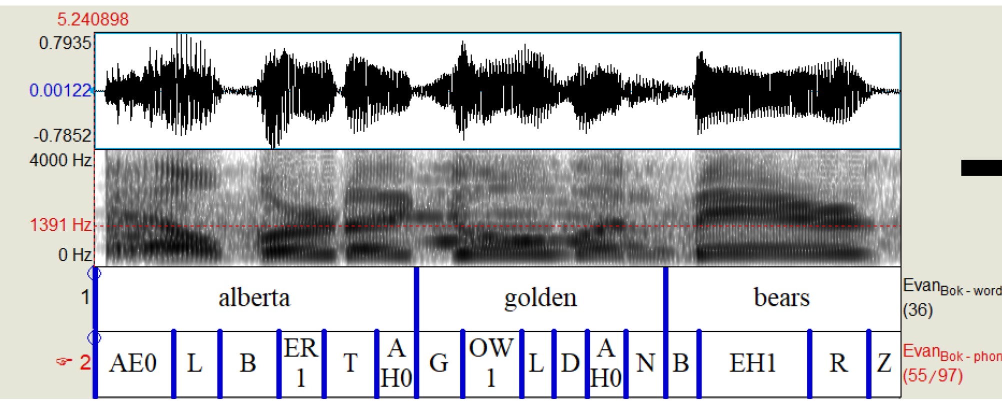
adjacent segments

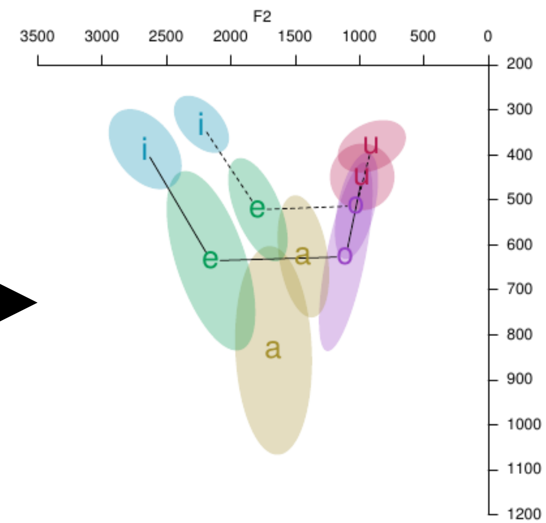
Vowel duration

Source: http://drammock.github.io/phonR/

Raw data

Analysis





(R, Goldvarb...)

How?

Data file (CSV)

Gender

formants

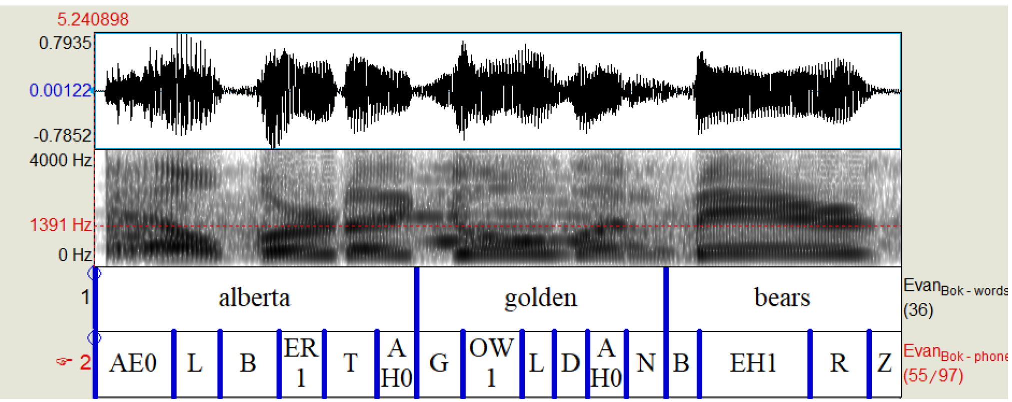
Vowel duration

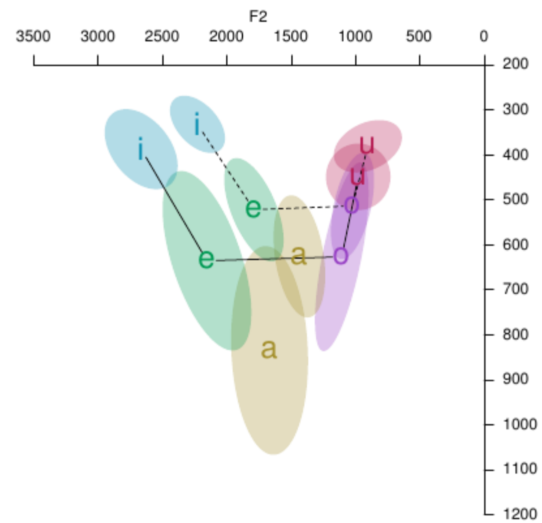
Age

adjacent segments

Raw data

Analysis





- scripts
- manual entry
- software

Data file (CSV)

Gender

formants

Vowel duration

Age

adjacent segments

Why automate this process?

- Practical reasons
  - Technical skill
  - Time/duplication of effort
  - Availability
- Methodological reasons
  - <u>Standardized</u>, customizable linguistic measures

- Harder with 1+ corpora…

Raw data to CSV file: steps

1. Process raw data

2. Make measures

3. Find relevant tokens

4. End up with usable spreadsheet

Software goals

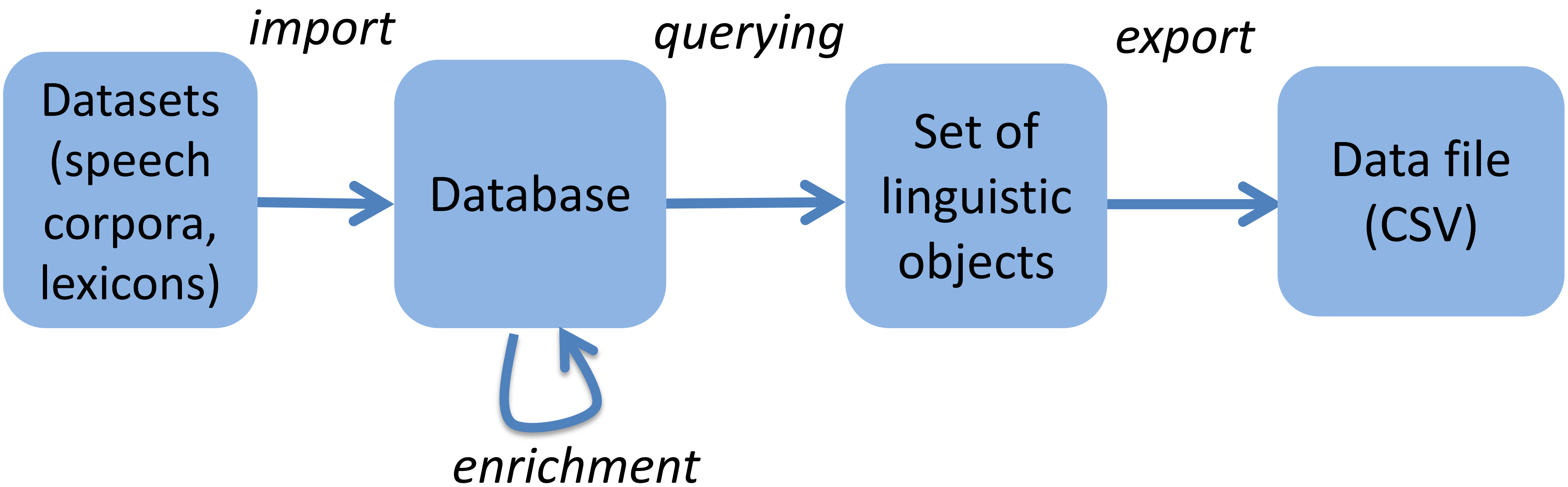
- Scalable & fast
- Require minimal technical skill from user
- Abstraction away from dataset format
- Querying dataset without access to raw data

- ⇒ Easier large-scale studies using speech corpora

ISCAN: A SYSTEM FOR INTEGRATED PHONETIC ANALYSES ACROSS SPEECH CORPORA

Michael McAuliffe[a], Arlie Coles[a], Michael Goodale[a], Sarah Mihuc[a], Michael Wagner[a], Jane Stuart-Smith[b], Morgan Sonderegger[a]

Proc. ICPhS 2019



- Implementation
  - Python API
  - Graphical User Interface

- (show GUI here)

- Note:
  - Server-client architecture enables analysis without access to raw data
  - Permissions system controls who can see/hear tokens
  - Can be installed on web server (default) or personal computer

Documentation exists

- GUI: https://iscan.readthedocs.io/
  - Can sign up as tutorial user
- Python API: https://polyglotdb.readthedocs.io/

SPADE: datasets

- 43+ datasets, 4 countries, 115 years
- heterogeneous corpus formats
- public and private

SPADE: datasets

To date:

- Acquired: ~22
- Measurements generated: 10-13
- ~ 10 dialect regions
  - ~500 hours (?)

SPADE: ethics and credit

- For private datasets (data guardians): ethics complex: GDPR + US laws
- Data transfer agreement
  - data use in keeping with original permissions, as far as is possible

- 'SPADE consortium': author on everything
- Datasets of measures → data guardians at end of project

Case studies

Done	Sibilants	Vowels: formants
In progress	Vowels: voicing effect Stops	Vowels 3
Planned	Vowels (dynamic)	r, l

Case studies

Done	Sibilants Stuart-Smith et al. *Proc. ICPhS 2019*	Vowels: formants Mielke et al. *Proc. ICPhS 2019*
In progress	Vowels: voicing effect Tanner et al. *Toronto WP Ling 2019*	Stops Vowels 3
Planned	Vowels (dynamic)	r, l

Caveats

- For all case studies:
  - We can largely analyze ‘internal factors’
  - Little social information (only gender, age)
    - Very limited range for ethnicity, race, SES…
  - Almost all ‘transcriptions’ are phonological/forced alignment
    - Little manual checking

/s/-retraction in English

- /s/ → [ʃ]-like sound in /str/
  - *string, street*

Stuart-Smith et al.
Proc. ICPhS 2019

Updated analysis

- Sound change, varies:
  - by dialect
    - London, Philadelphia, Raleigh vs. Australian English, RP
  - By speaker

(e.g. Baker et al, 2011; Stevens and Harrington, 2016)

Questions

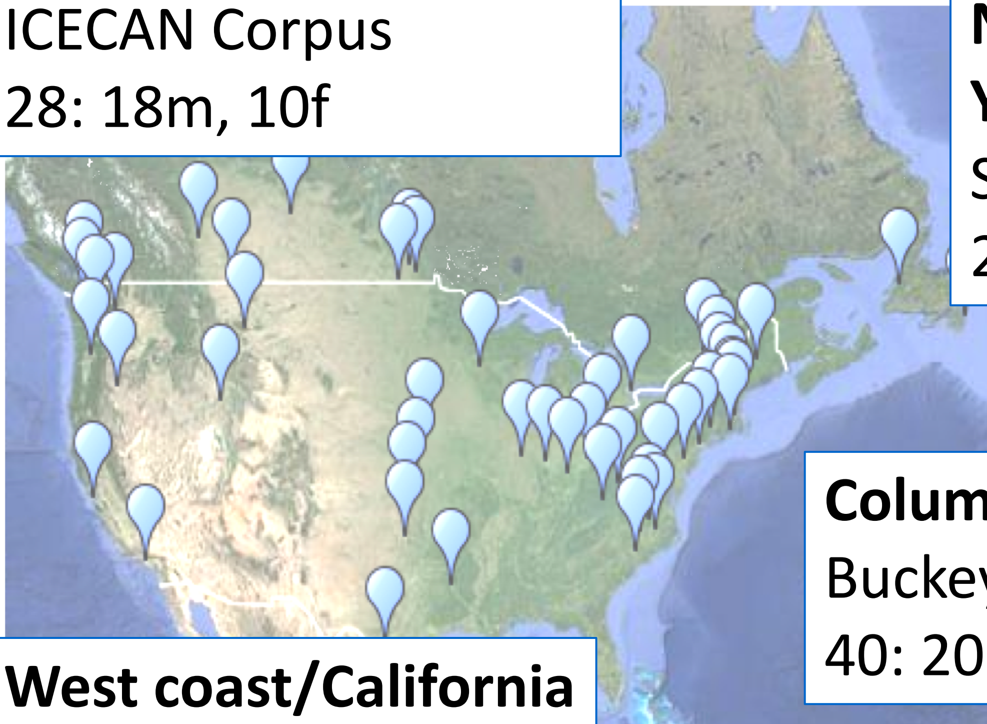
- Basic dialect differences in /s/, /ʃ/
- What is the evidence for /s/-retraction across English dialects?
  - dichotomous pattern or a continuum?
- Speakers `` ``
- Received wisdom:
  - Sibilants don't differ much by dialect
  - there are 'retracting'/non-retracting dialects and speakers

SPADE

Sample for this study: New World

Canada
ICECAN Corpus
28: 18m, 10f

Northern Cities, e.g. New York, Philadelphia
Santa Barbara Corpus
20: 9m, 11f



Columbus, Ohio
Buckeye Corpus
40: 20m, 20f

West coast/California
Santa Barbara Corpus
46: 20m, 26f

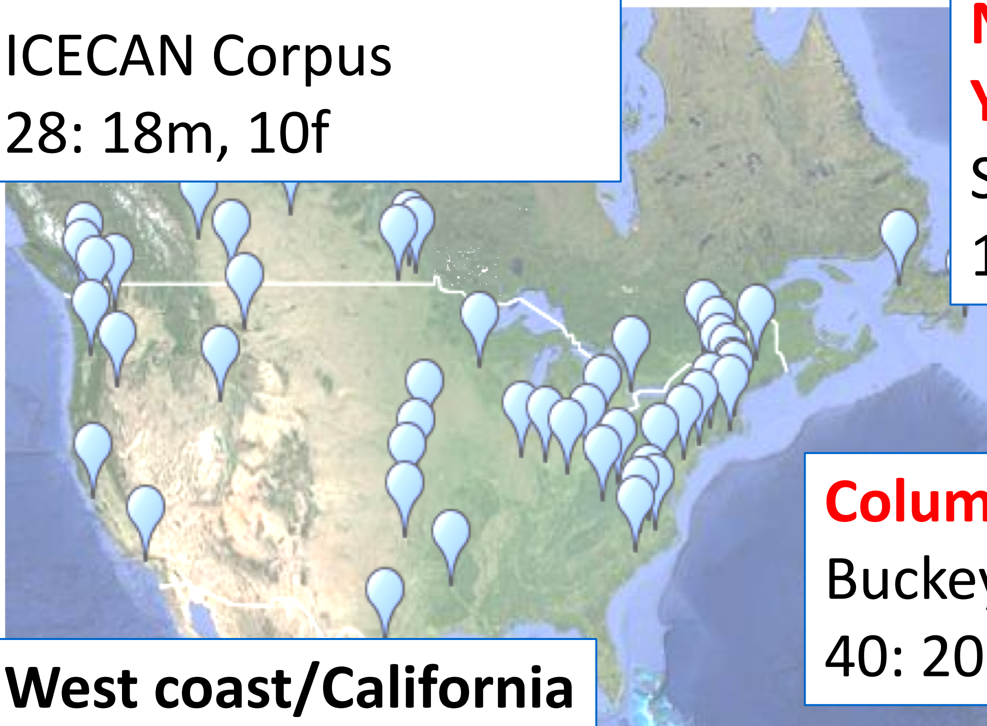
www.google.com/maps/

Raleigh, North Carolina
Raleigh Corpus
101: 50m, 51f

235 speakers

SPADE

Sample for this study: New World



www.google.com/maps/

Canada
ICECAN Corpus
28: 18m, 10f

Northern Cities, e.g. New York, Philadelphia
Santa Barbara Corpus
19: 8m, 11f

Columbus, Ohio
Buckeye Corpus
40: 20m, 20f

West coast/California
Santa Barbara Corpus
43: 20m, 23f

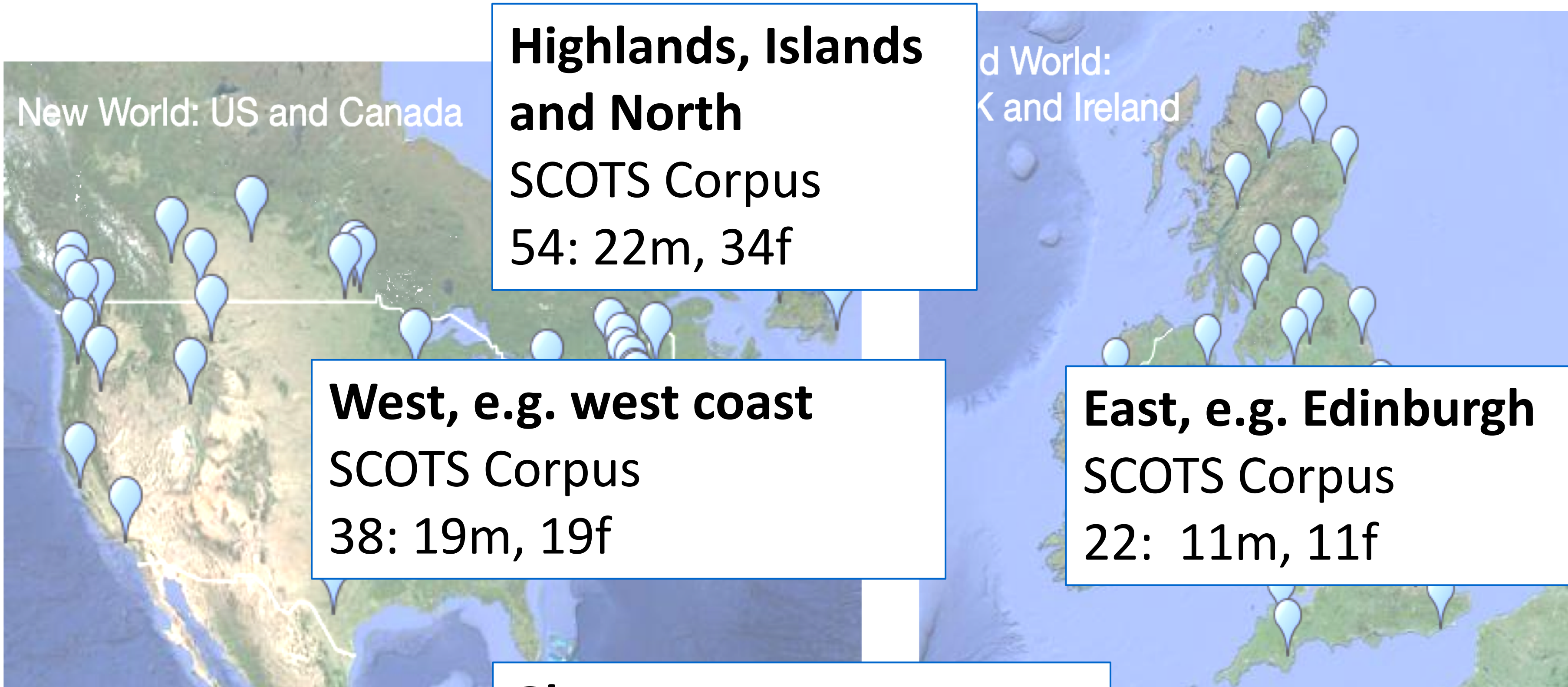
Raleigh, North Carolina
Raleigh Corpus
101: 50m, 51f

reported to show /s/-retraction

235 speakers

SPADE

Sample for this study: Old World



Highlands, Islands and North
SCOTS Corpus
54: 22m, 34f

West, e.g. west coast
SCOTS Corpus
38: 19m, 19f

East, e.g. Edinburgh
SCOTS Corpus
22: 11m, 11f

Glasgow
Sounds of the City
70: 35m, 36f

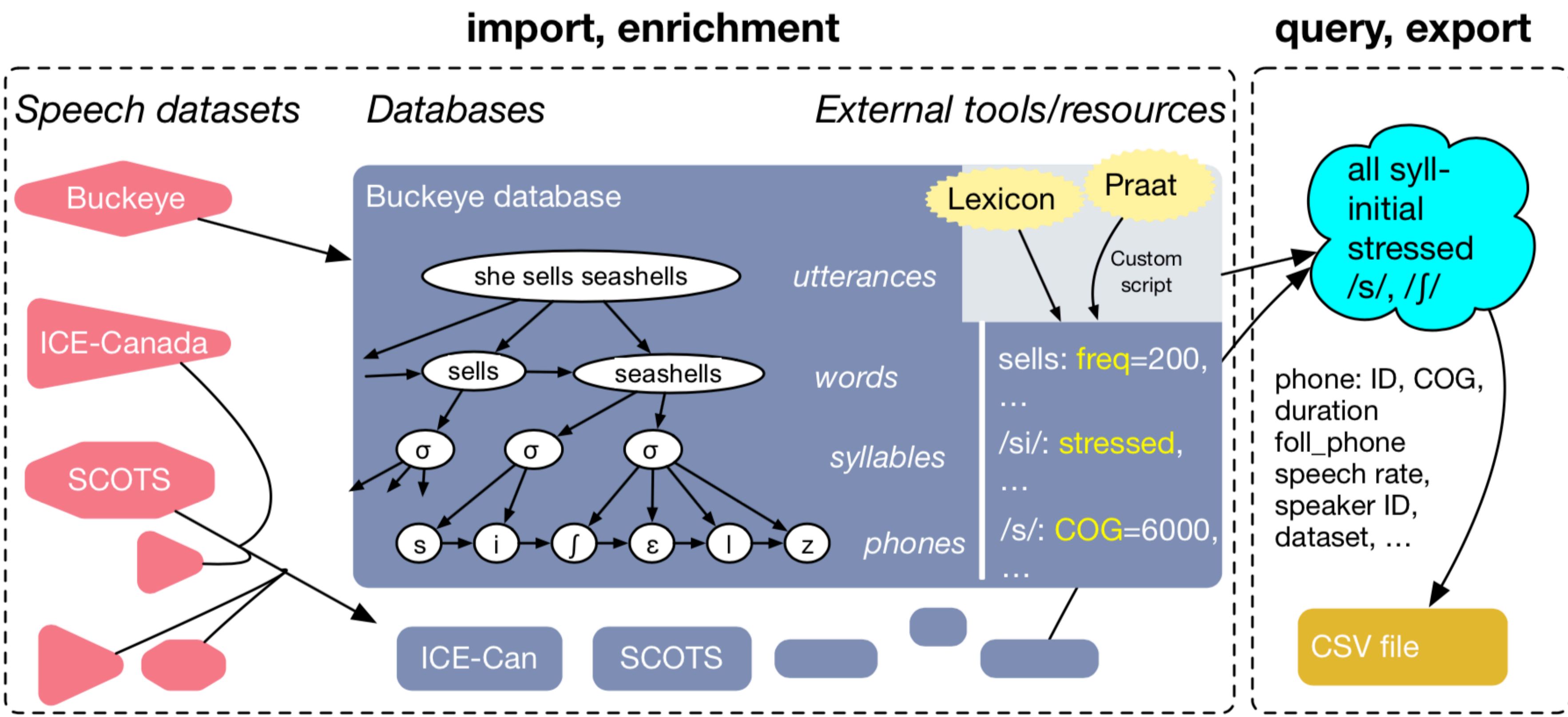
www.google.com/maps/

185 speakers

Data

- All instances of stressed, word-initial /s/
- Acoustic measure: spectral Centre of Gravity (CoG)
  - 1-16 kHz, middle 50%
  - Data cleaning
- N = 98k
- Prediction: **/s/ > /str/ > /ʃ/**

ISCAN usage



import, enrichment
query, export
Speech datasets
Databases
External tools/resources
Buckeye
ICE-Canada
SCOTS
Buckeye database
she sells seashells
sells
seashells
σ
σ
σ
s
i
ʃ
ɛ
l
z
utterances
words
syllables
phones
Lexicon
Praat
Custom script
sells: freq=200,
…
/si/: stressed,
…
/s/: COG=6000,
…
ICE-Can
SCOTS
all syll-initial stressed /s/, /ʃ/
phone: ID, COG, duration
foll_phone
speech rate,
speaker ID,
dataset, …
CSV file

Model

- Bayesian linear mixed-effects model
  - `Stan/brms` (Carpenter et al., 2017; Bürkner 2018)

- Random effects:
  - Dialect (10) ← Dialect mean
  - Speaker, <u>within dialect</u> (418) ← Speaker <u>offset</u> from dialect mean
  - Word (4k)

- Predictors:
  - /sV/ /sp, st, sk, spr skr/ /str/ /ʃV/

s	sC	str	ʃ

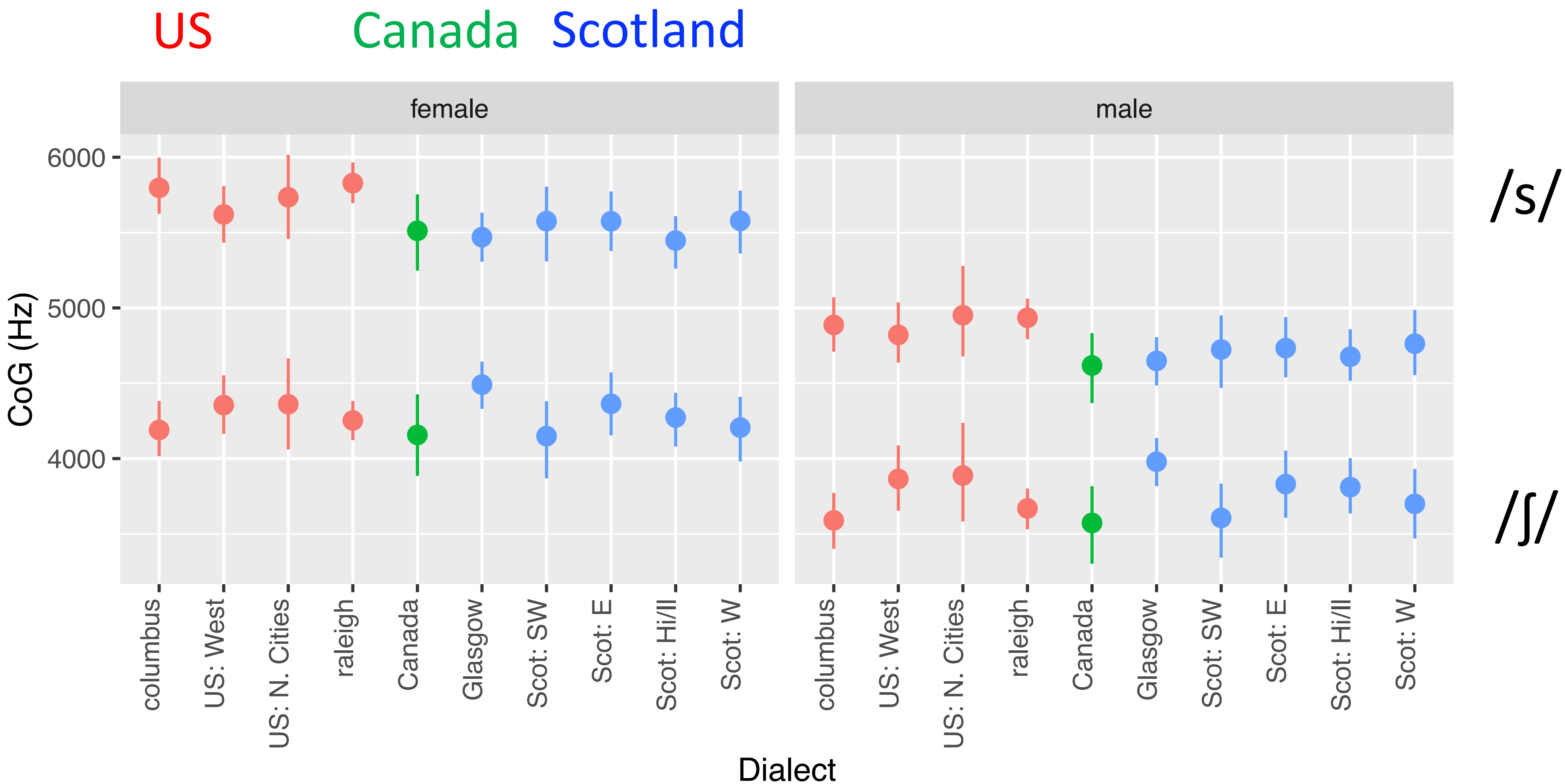
onset

  - Gender (M/F)
  - gender:onset
  - Phone duration

varies by dialect

varies by speaker (w/in dialect)

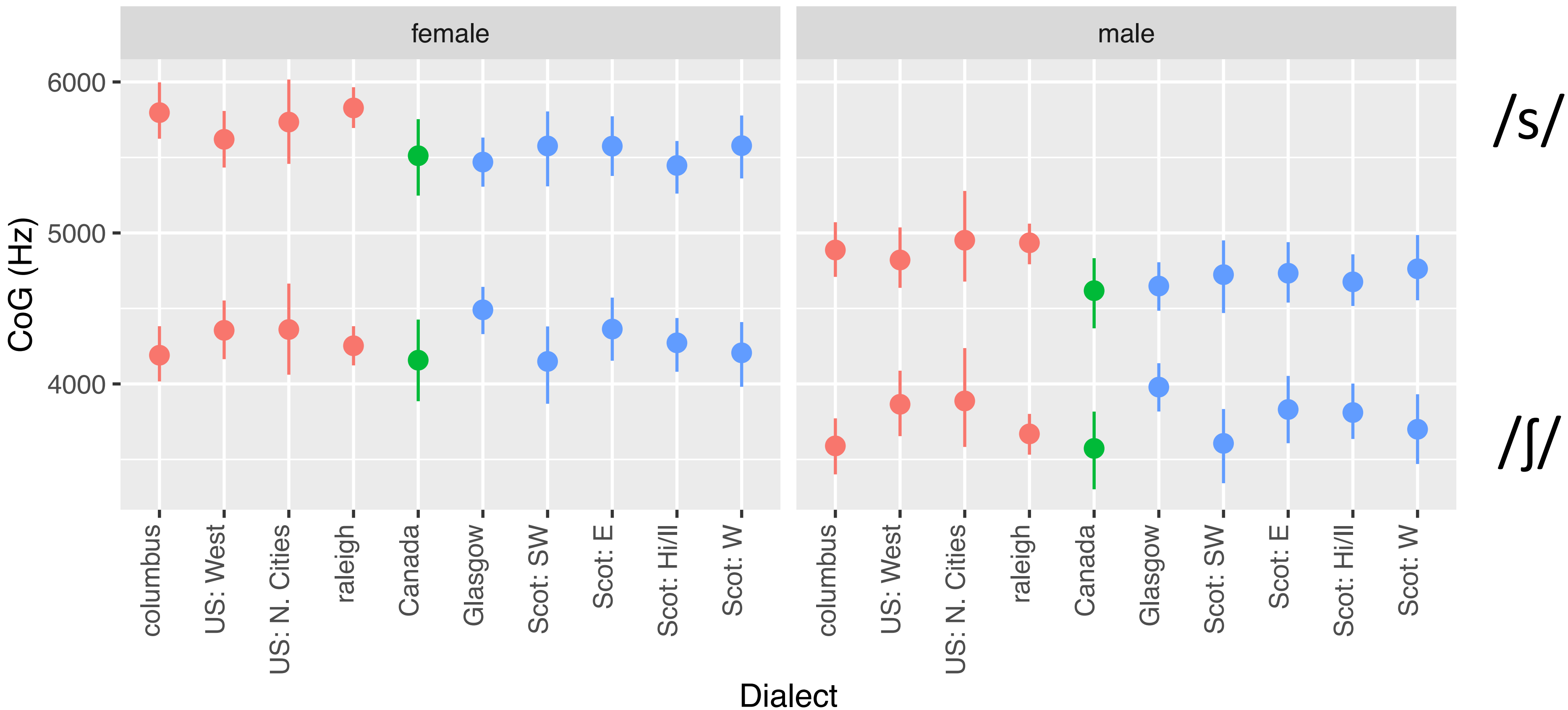
Results: s+V, ʃ+V



95% HPD intervals

Results: s+V, ʃ+V

US Canada Scotland



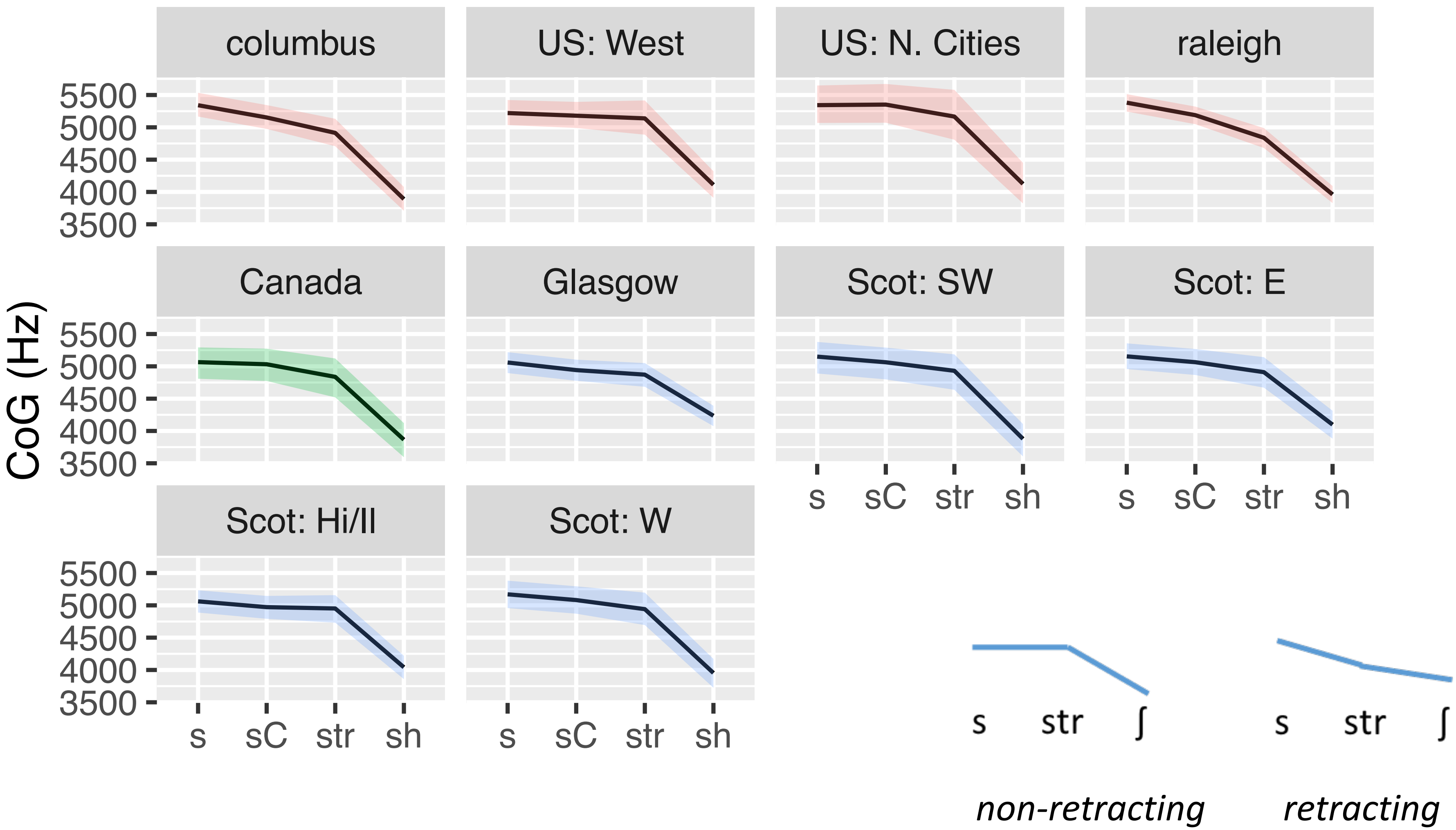
Dialects vary somewhat

Gender effect: overall, s/ʃ difference

Results: s+V, ʃ+V

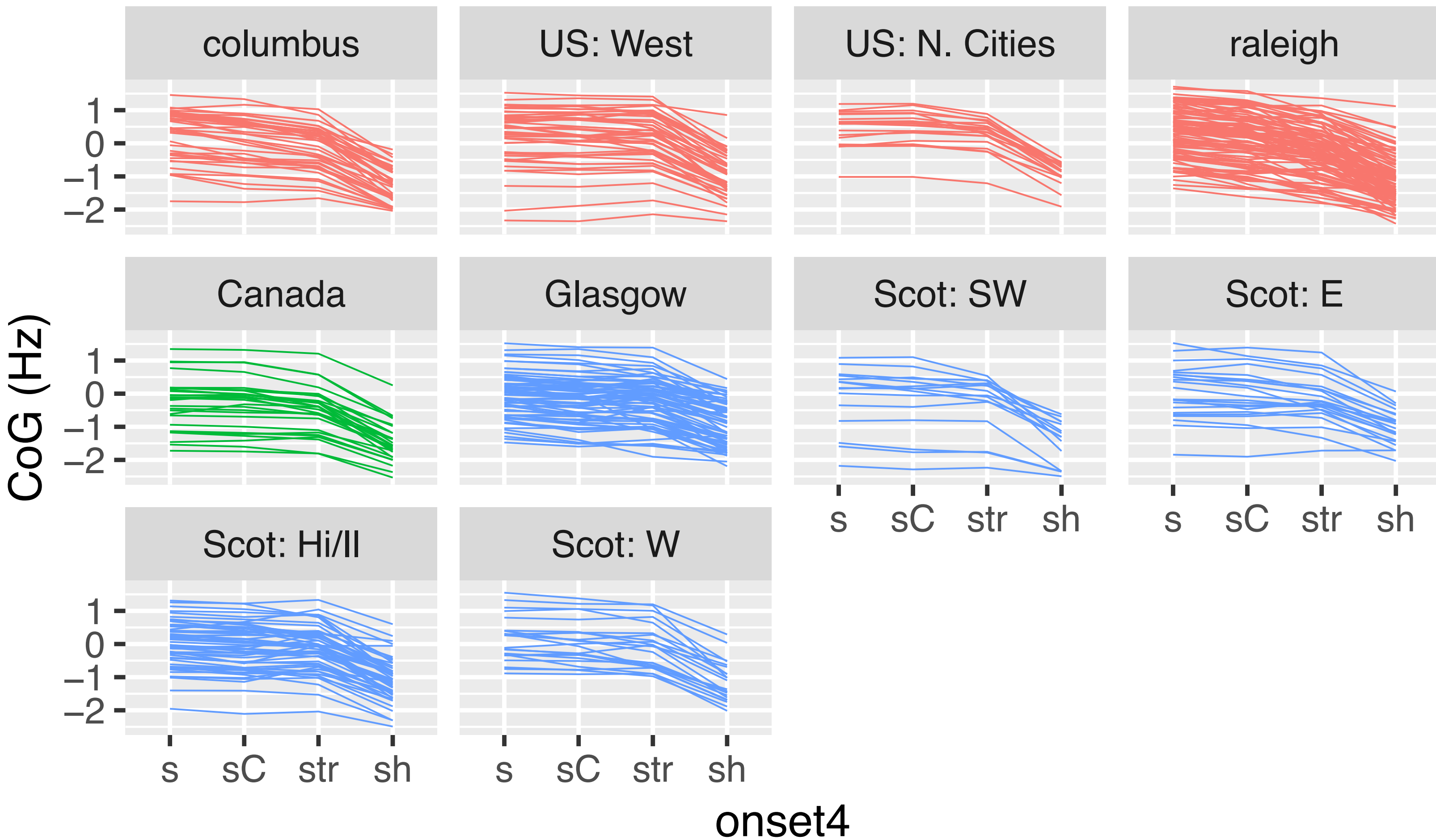
- Gender effect dialect differences:
  - Unclear / small

Results: onset



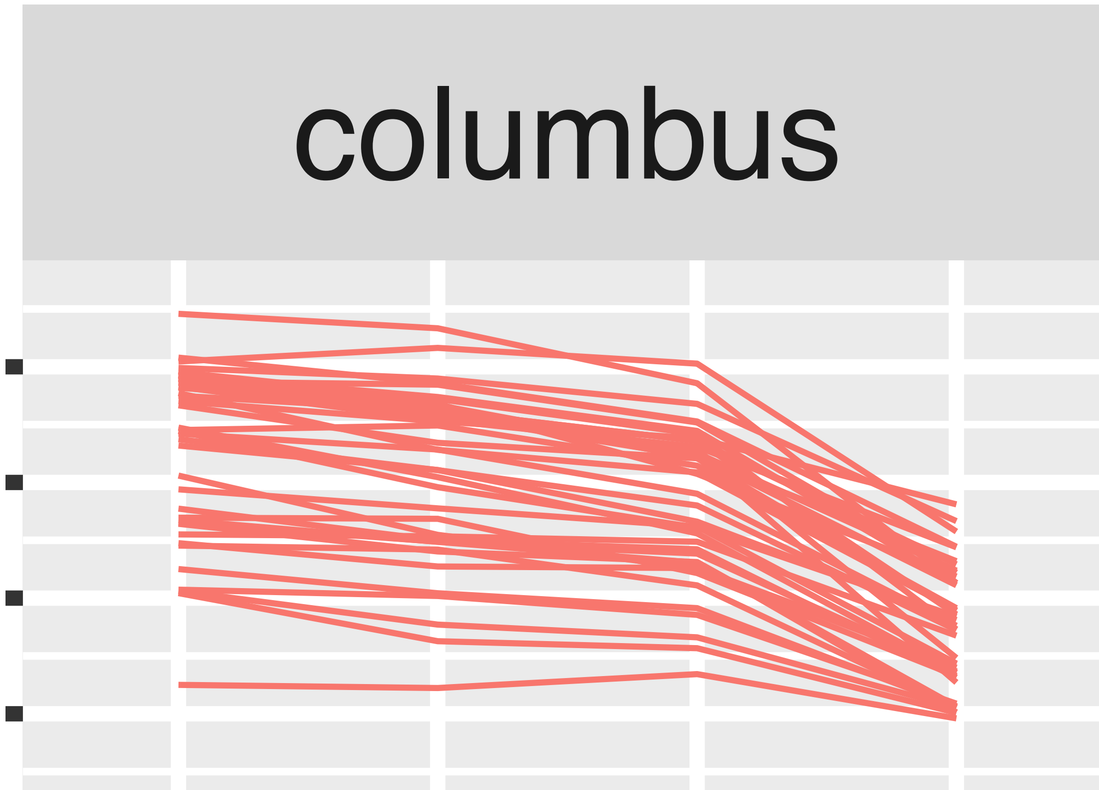
Dialects differ

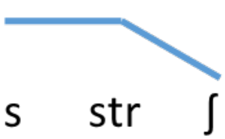
Results: onset



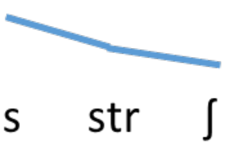
Speakers differ more

Results: onset





non-retracting



retracting

Retraction ratio

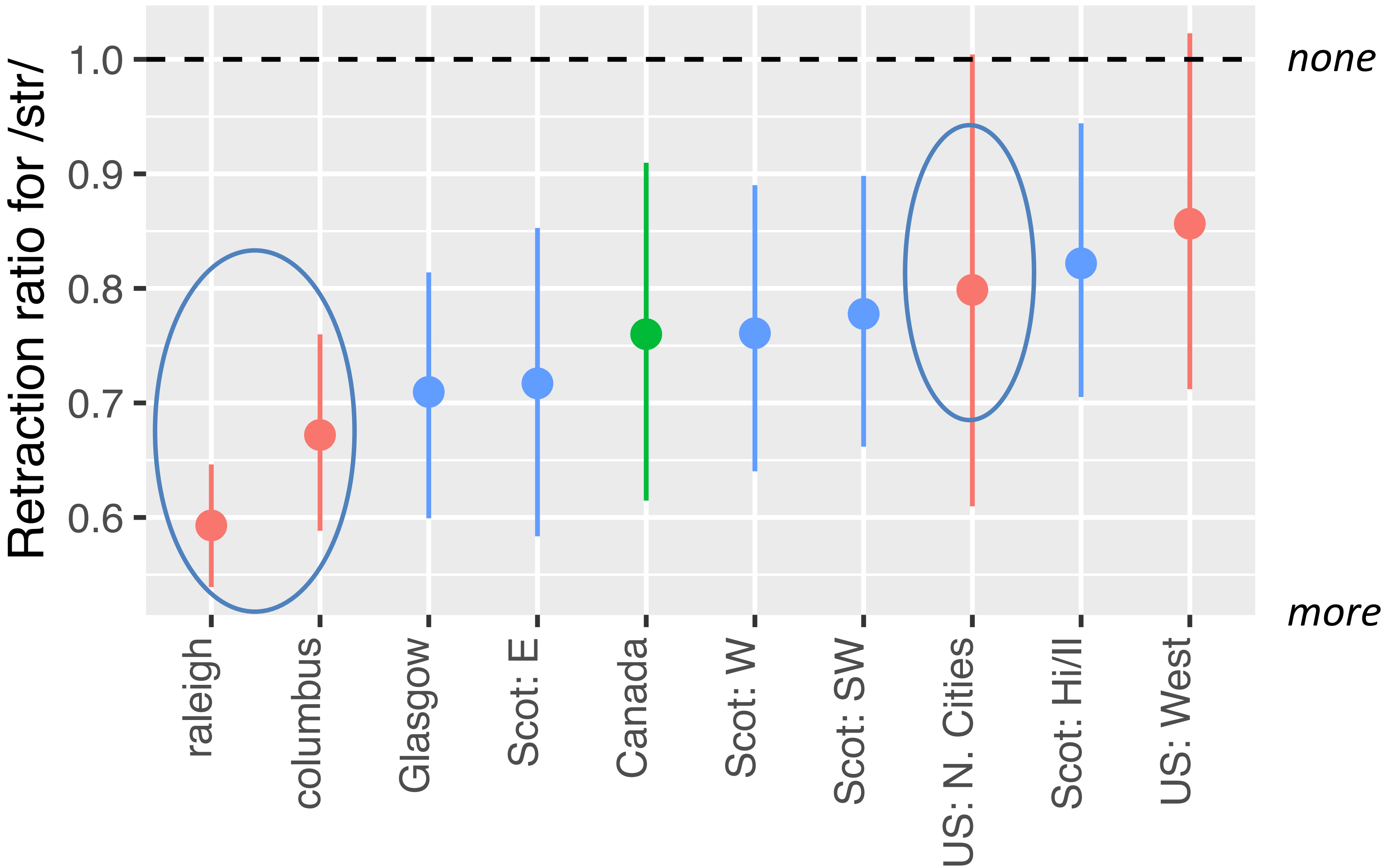
- Measure of degree of /s/ retraction:

$$\frac{(\text{/str/ mean} - \text{/sh/ mean})}{(\text{/s/ mean} - \text{/sh/ mean})}$$

0: "st(r)ew" = "shrew"

1: "st(r)ew" = "stew"

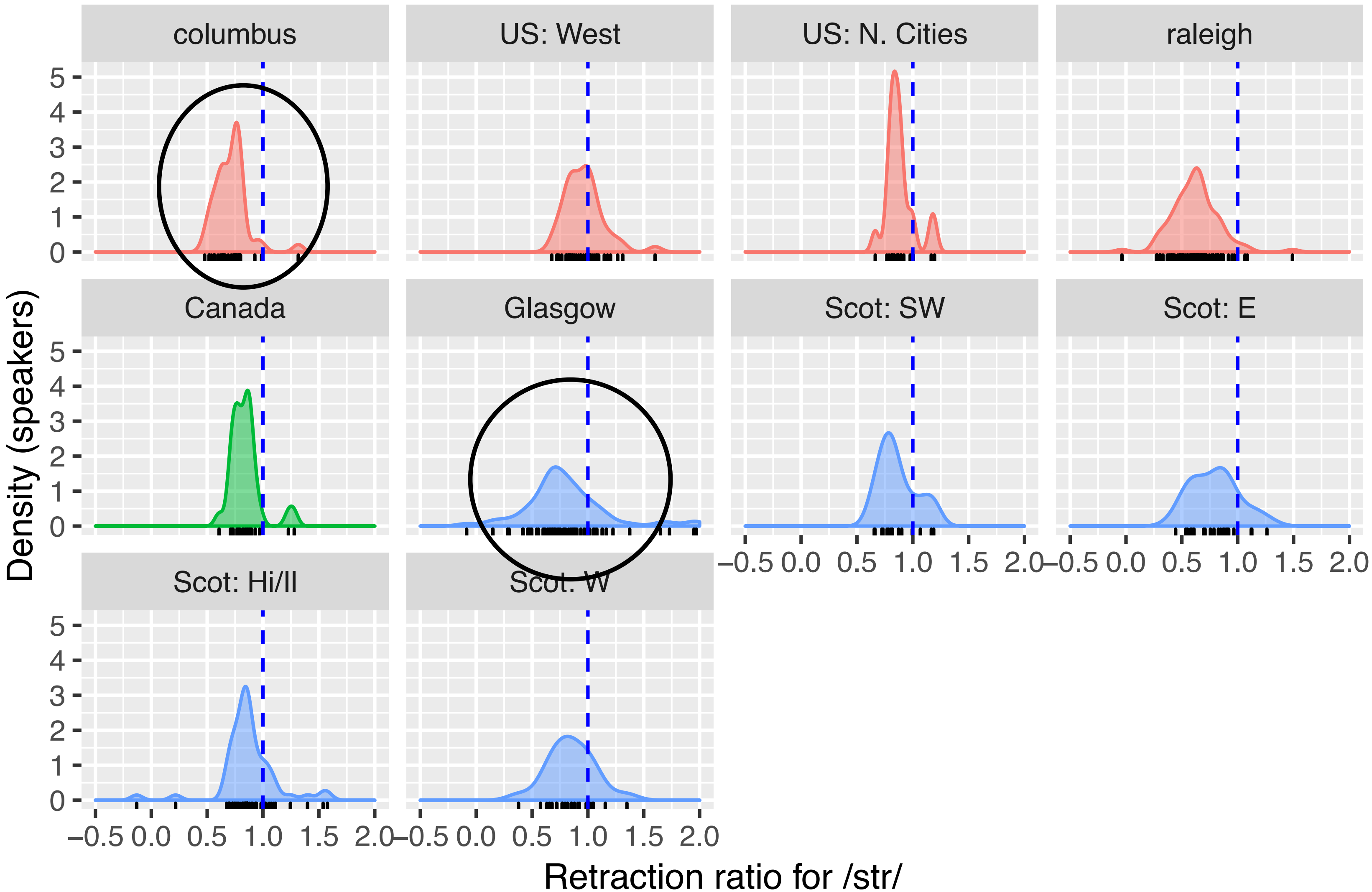
Results: retraction



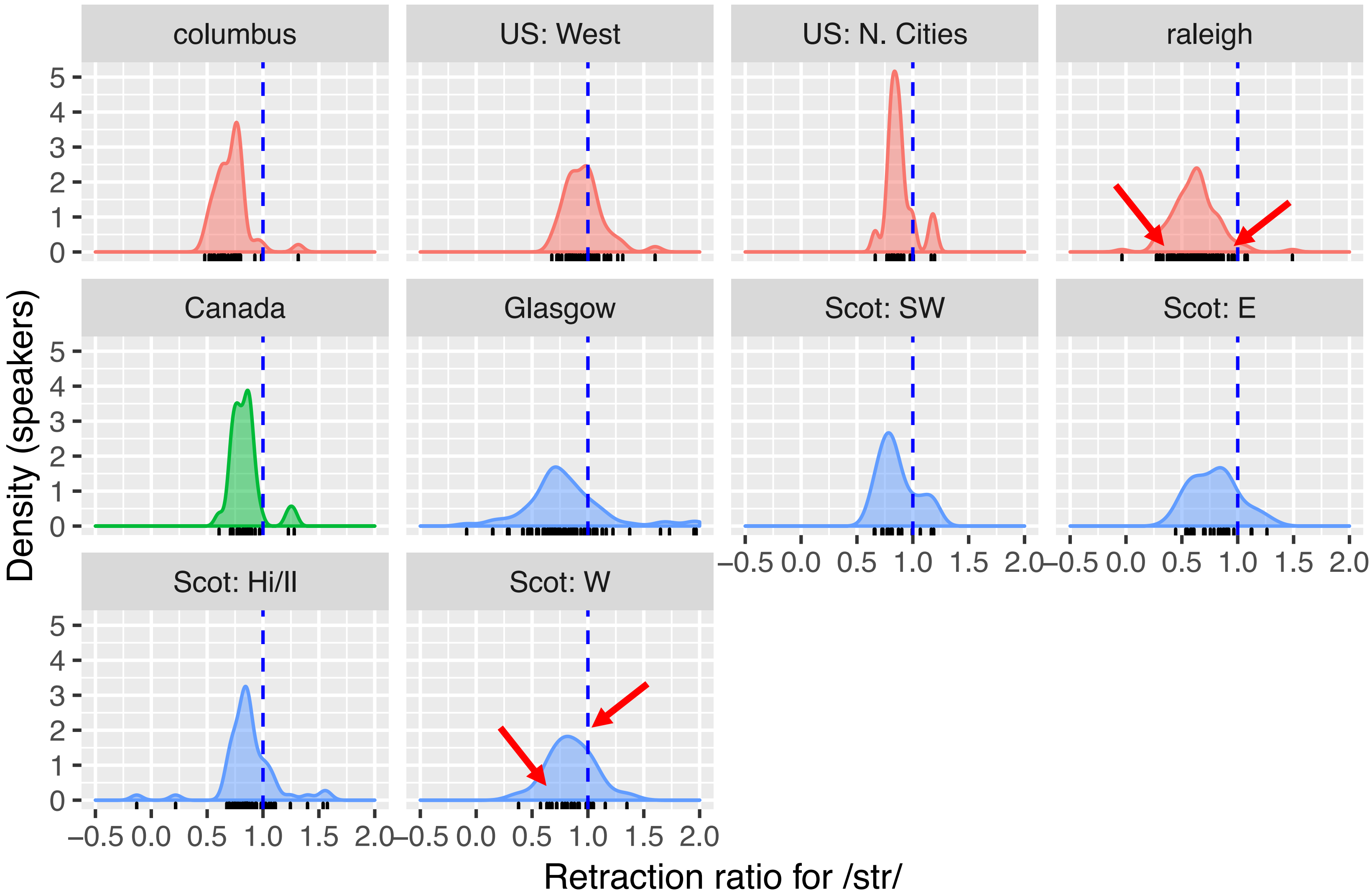
Dialects differ

no dichotomy

Distribution of speaker retraction ratios:



Degree of speaker variation varies by dialect



‘Retracting’/non-retracting?

Results: retraction

- Female speakers : more retracted
  - Across dialects

Discussion

- Dialects differ in sibilant production generally, and in degree of retraction
- Speakers `` ``
- More within-dialect than across-dialect variation
- Retracting vs. non-retracting?
  - Little evidence for dichotomy

Discussion

- /str/-retraction maybe not a primary axis along which individuals/dialects vary

- Scaling up analysis allows identification of new patterns

Case studies

	Sibilants	Vowels: formants
Done	Stuart-Smith et al. *Proc. ICPhS 2019*	Mielke et al. *Proc. ICPhS 2019*
In progress	Vowels: voicing effect Stops	Vowels 3
Planned	Vowels (dynamic)	r, l

Case study: Vowel duration

- Voicing effect (*p**a**d* > *p*<u>*a*</u>*t*)
  - Primary cue to "voicing" word-finally
    ⇒ large effect ?
    (Chen, 1970)

lab speech, two dialects

  - Robust effect?
    - Across dialects, speakers
    - Spontaneous speech

James Tanner
PhD thesis

Toronto Working Papers in Ling. 2019

Voicing effect: data

- Utterance-final CVC words
  - n = 59k

- New World:
  - Midwest (Buckeye)
  - Washington DC AAVE (CORAAL)
  - Raleigh
  - New Eng, Lower South, Northern Cities, NYC, 'West' (Santa Barbara Corpus)

Possible enhanced VE
(Holt et al 2016, Farrington, 2018)

Voicing effect: data

- Old world:
  - RP (Fabricus, 2000)
  - Scottish: (SCOTS corpus, Sounds of the City)
    - Central
    - Edinburgh
    - Northern
    - Islands
    - Glasgow

Scottish Vowel Length Rule: overrides VE?

Voicing effect: analysis

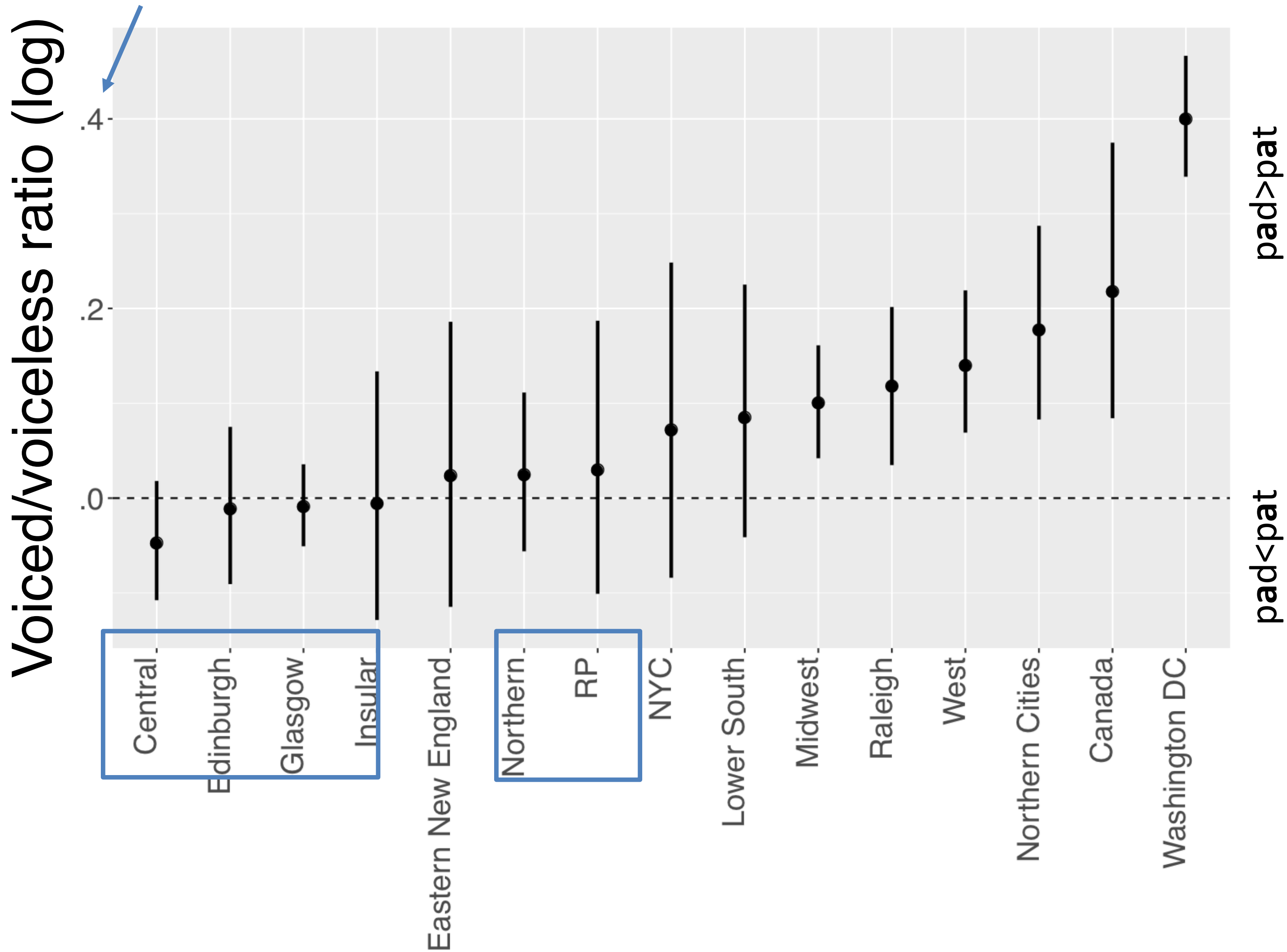
- Similar Bayesian linear mixed-effects model

- Effect of interest: following C voicing
  - Varies by dialect
  - By speaker within dialect

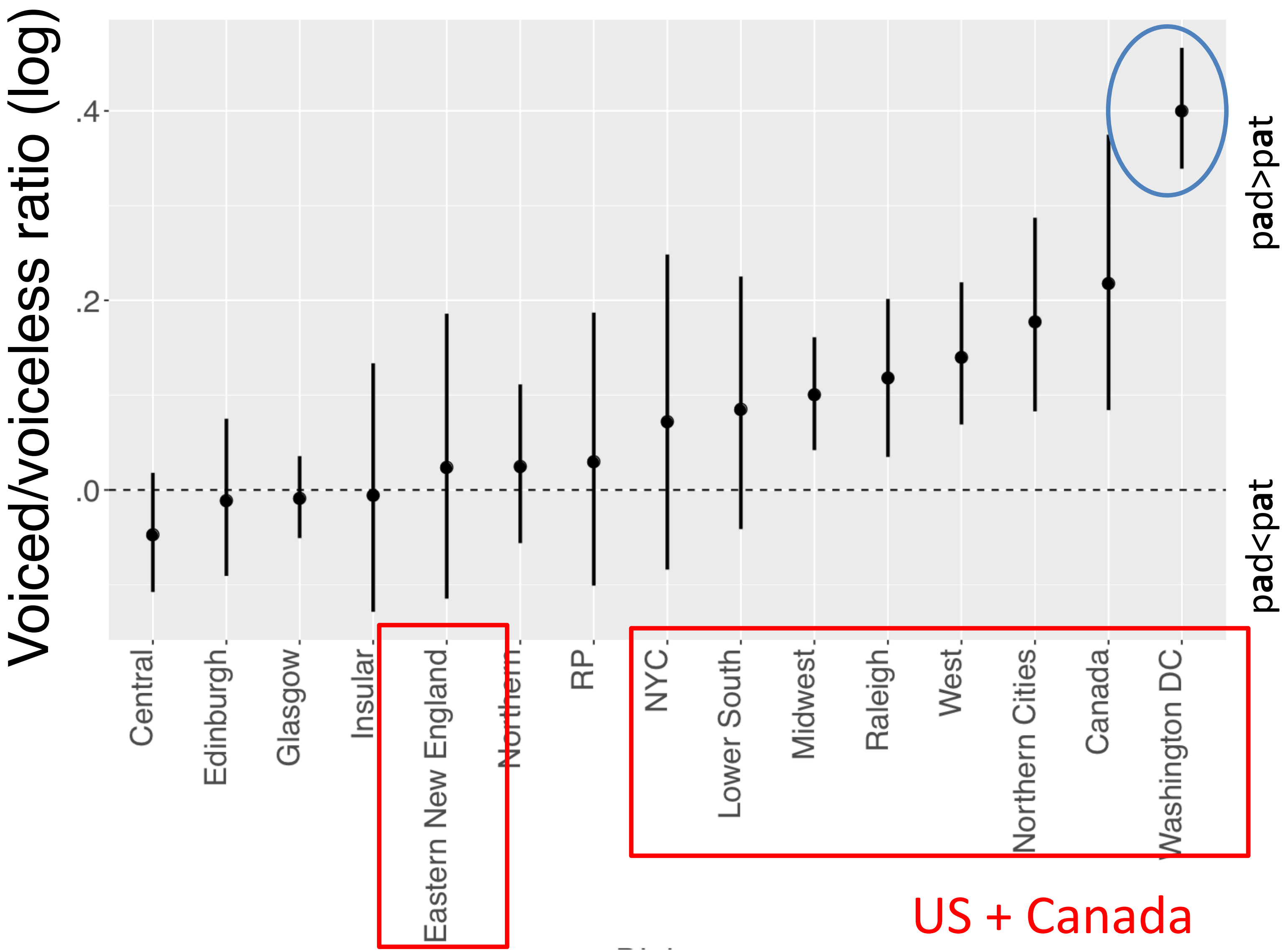
- + controls (speech rate, word frequency, vowel height..)

Results

~ minimum value from lab studies

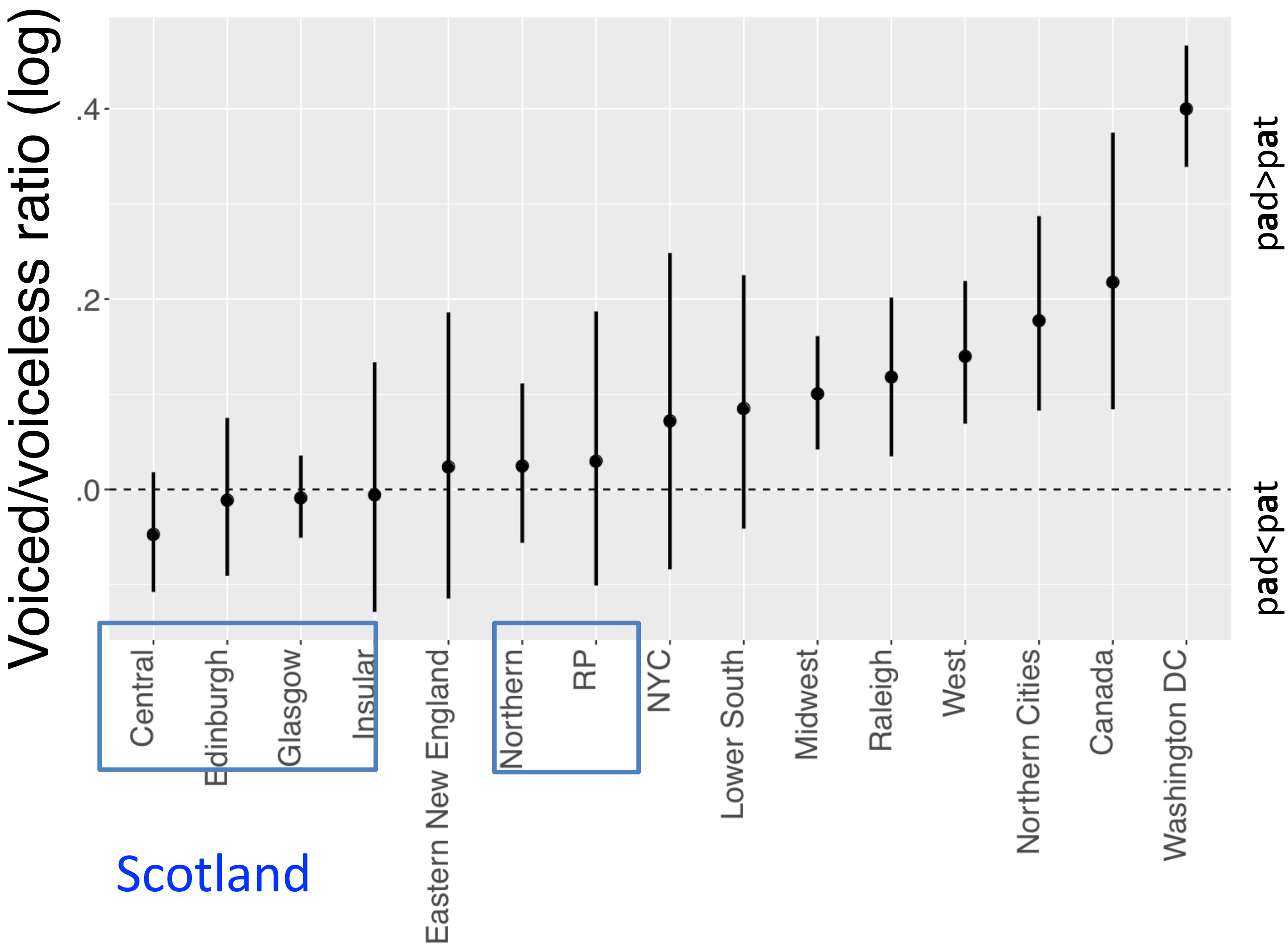


Results



Voiced/voiceless ratio (log)
.4
.2
.0
pad>pat
pad<pat
Central
Edinburgh
Glasgow
Insular
Eastern New England
Northern
RP
NYC
Lower South
Midwest
Raleigh
West
Northern Cities
Canada
Washington DC
US + Canada

Results



Voiced/voiceless ratio (log)
.4
.2
.0
pad>pat
pad<pat
Central
Edinburgh
Glasgow
Insular
Eastern New England
Northern
RP
NYC
Lower South
Midwest
Raleigh
West
Northern Cities
Canada
Washington DC
Scotland

Results

- Speakers differ in Voicing Effect, within dialect
  - Gender effect: small/absent

- Dialects differ more

Results

- VE very sensitive to context and style:

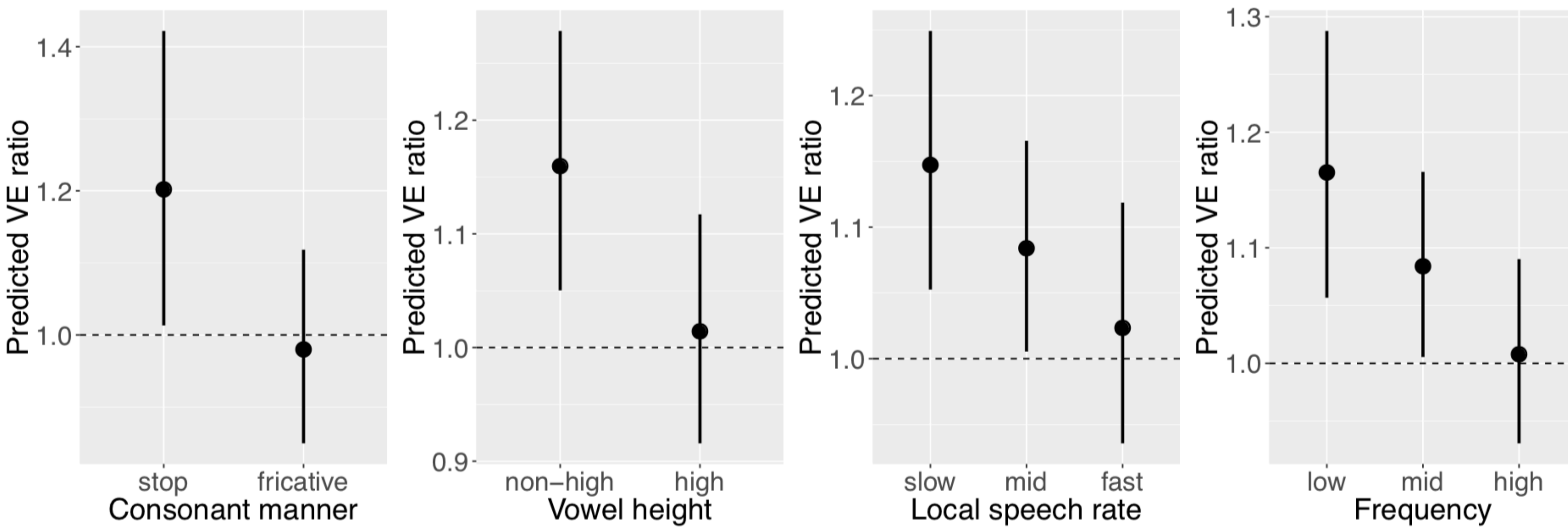


Figure 1: Predicted voicing effect (median and 95% CrI) as a function of consonant manner, vowel height, local speech rate (at -1, 0, 1), and word frequency (at -1, 0, 1). Predictions based on regression lines computed from the model's posterior, marginalising over all other covariates.

Discussion

- Voicing effect:
  - Robustness
  - Effect size: spontaneous vs. read speech
- High dialect variability
  - Partially due to dialect-specific [voice] processes?
- Scaling up analysis allows new perspective

Discussion

- Speaker vs. dialect variability:
  - VE: speakers < dialects
  - /s/ retraction: speakers > dialects
- New result from 'integrated' corpus analysis
  - but ??
- Ideas?

Case study: vowel formants

- Hypothesis from sociolinguistics: (Labov, 1994)
  - Intraspeaker variation in vowel production ~ same axis as diachronic change in community

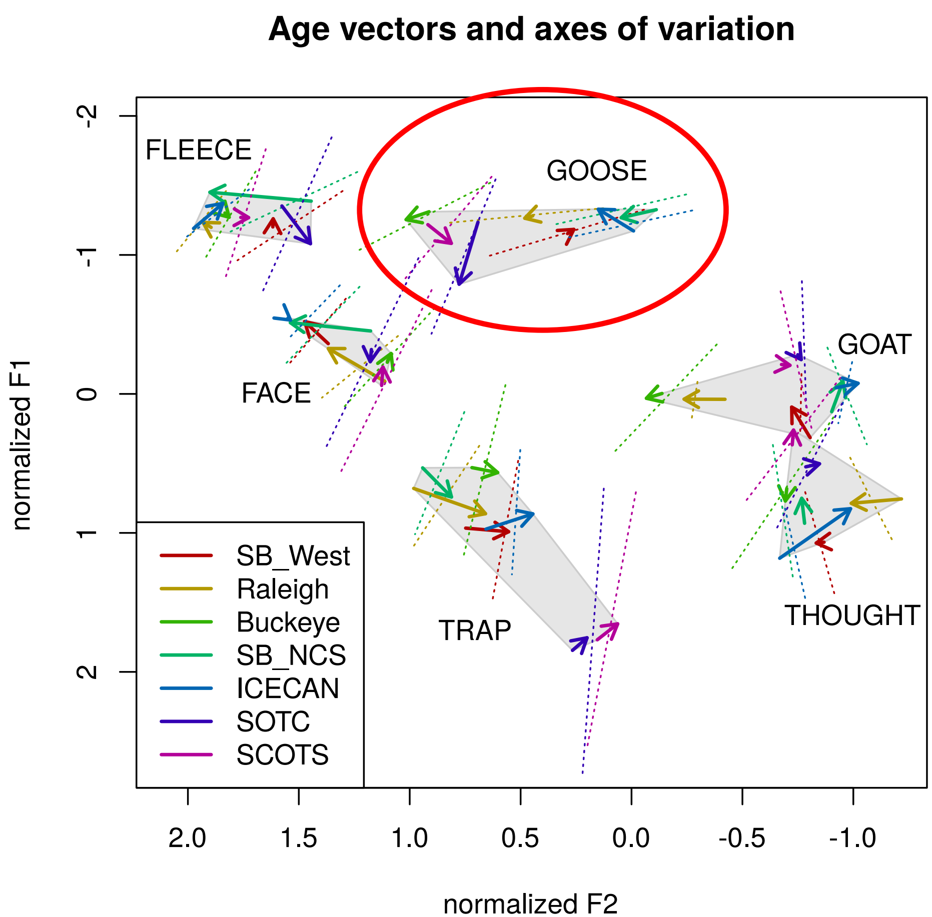


Figure 3: Age vectors (reflecting change in apparent time; arrows) and axes of intra-speaker variation (dotted lines) for six vowels across seven groups of speakers from six corpora.

Mielke et al. *Proc. ICPhS 2019*

Lead: Jeff Mielke, Erik Thomas

Case studies

Done	Sibilants Stuart-Smith et al. *Proc. ICPhS 2019*	Vowels Mielke et al. *Proc. ICPhS 2019*
In progress	Vowels 2 Stops: VOT	Vowels: low vowel spaces (F1/F2)
Planned	Vowels: dynamic F1, F2	r, l: ?

Thanks!

- SPADE Team, especially
  - Jane Stuart-Smith, Michael McAuliffe, James Tanner, Vanna Willerton, Jeff Mielke, Rachel MacDonald
- MCQLL lab programming RAs
  - Michael Goodale, Arlie Coles, Elias Stengel-Eskin
- Funding
  - Digging Into Data, SSHRC, NSERC

Questions