

Individual differences and
attentional effects on cue weighting
for prosody perception

Jennifer Cole

Northwestern University

My collaborators-

Joseph Roy

José I. Hualde

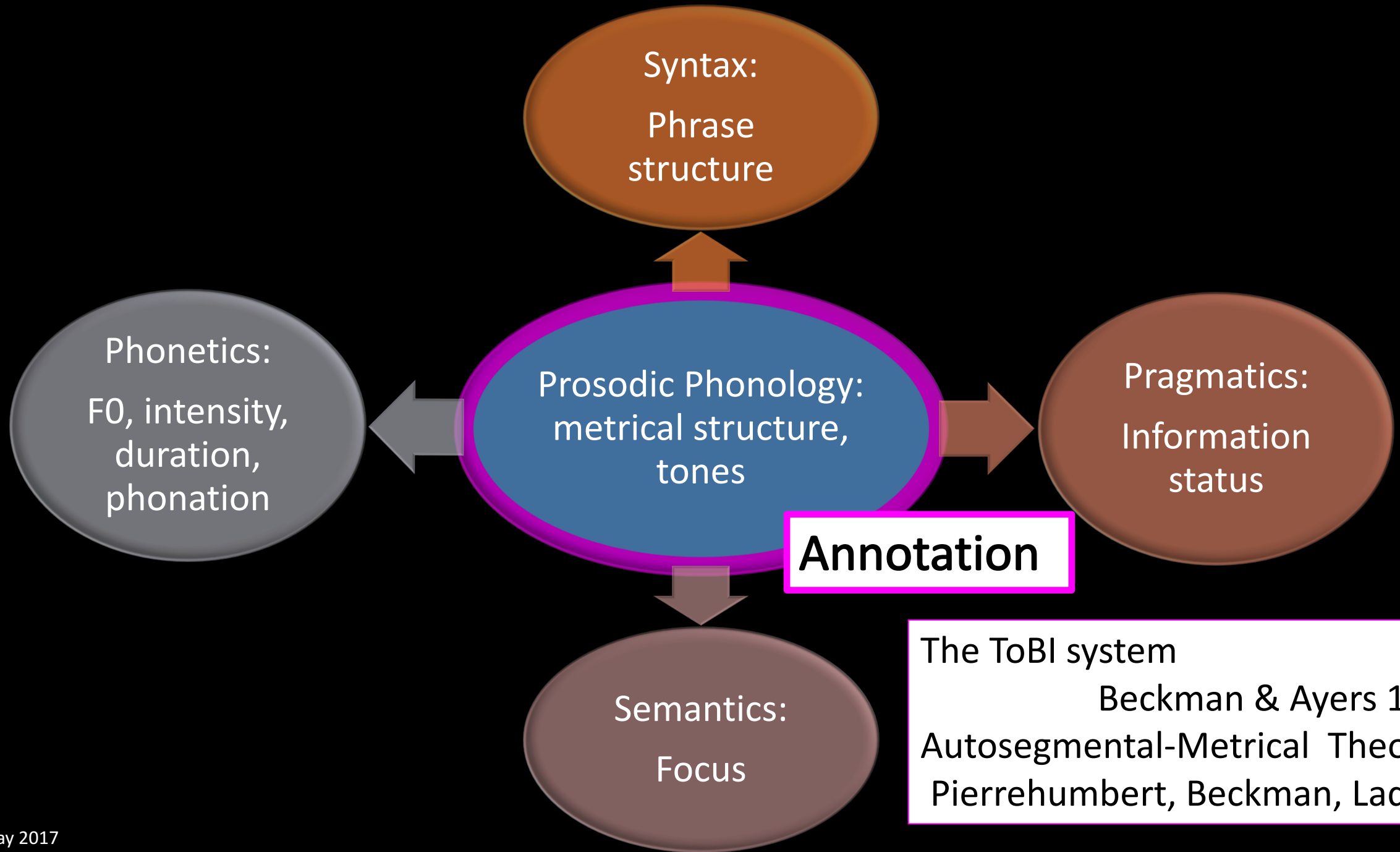
Ph.D. students-

Christopher Eager

Timothy Mahrt

Suyeon Im

What counts as data for a theory
of prosody?



Annotation

The ToBI system
Beckman & Ayers 1997
Autosegmental-Metrical Theory
Pierrehumbert, Beckman, Ladd,...

Annotation

Training examples:



banana2

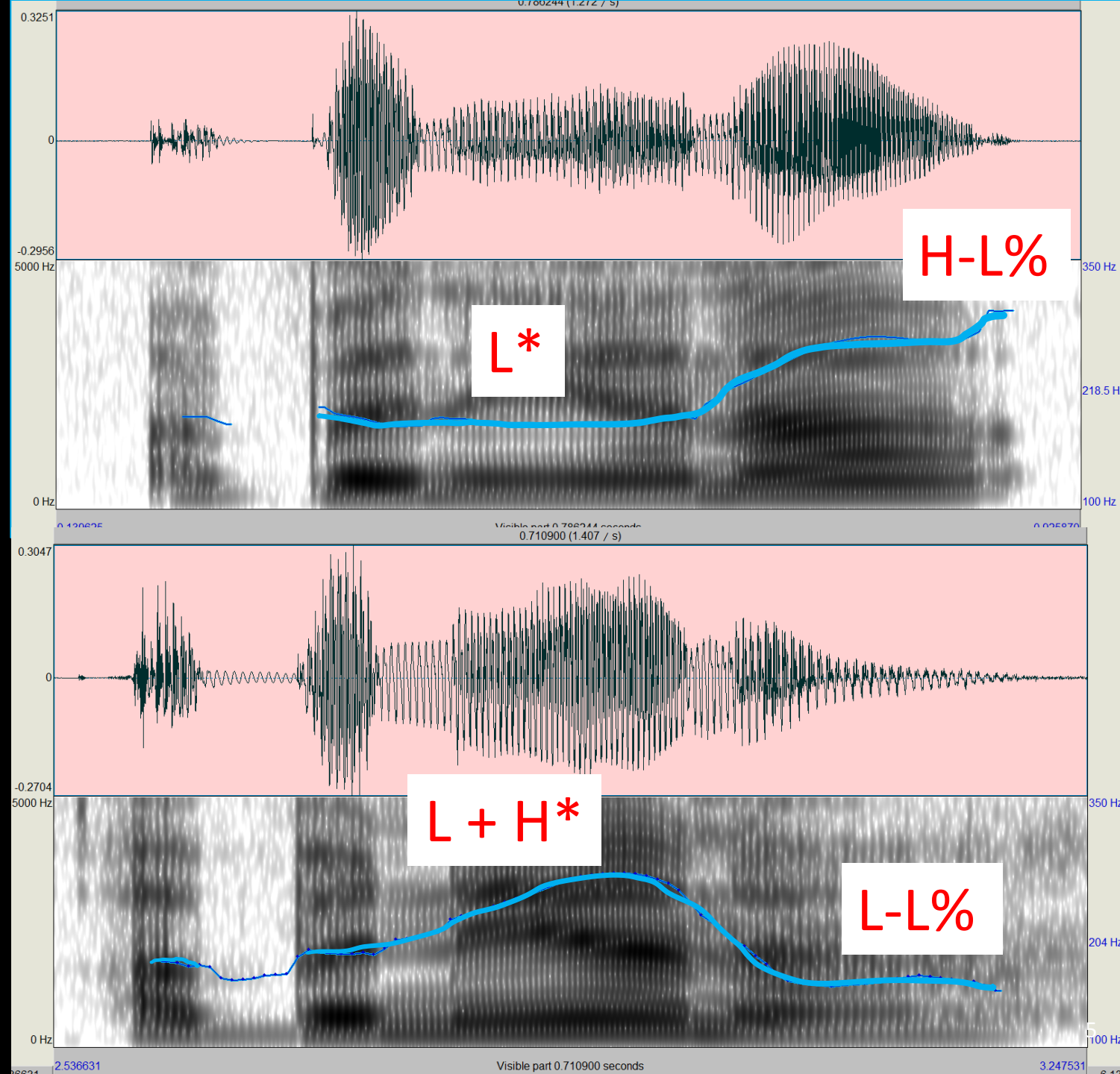
Veilleux, Shattuck-Hufnagel, Brugos. 2006.
Transcribing Prosodic Structure of Spoken Utterances
with ToBI.

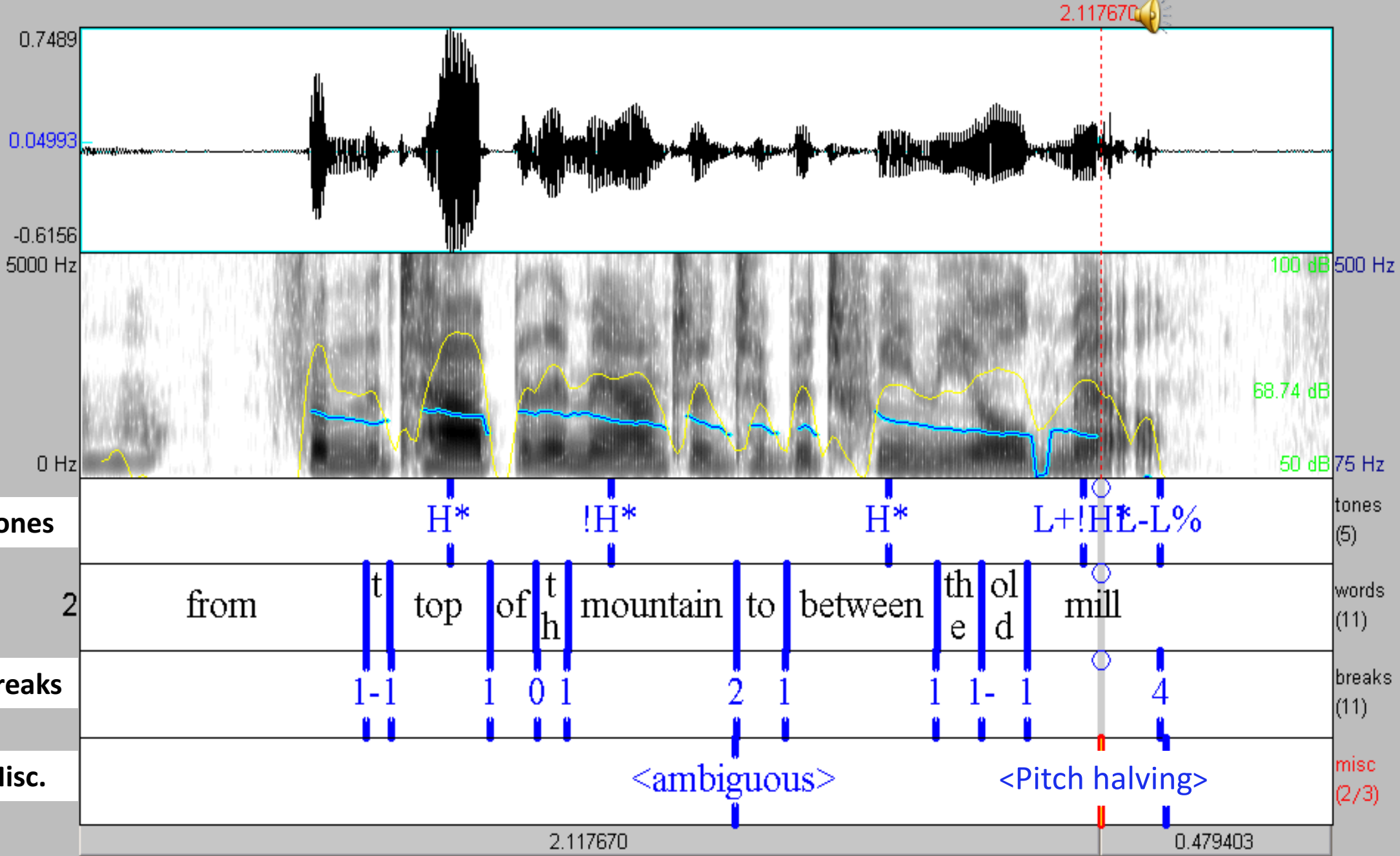
<https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-911-transcribing-prosodic-structure-of-spoken-utterances-with-tobi-january-iap-2006/>



ex5b1bananas

5 May 2017



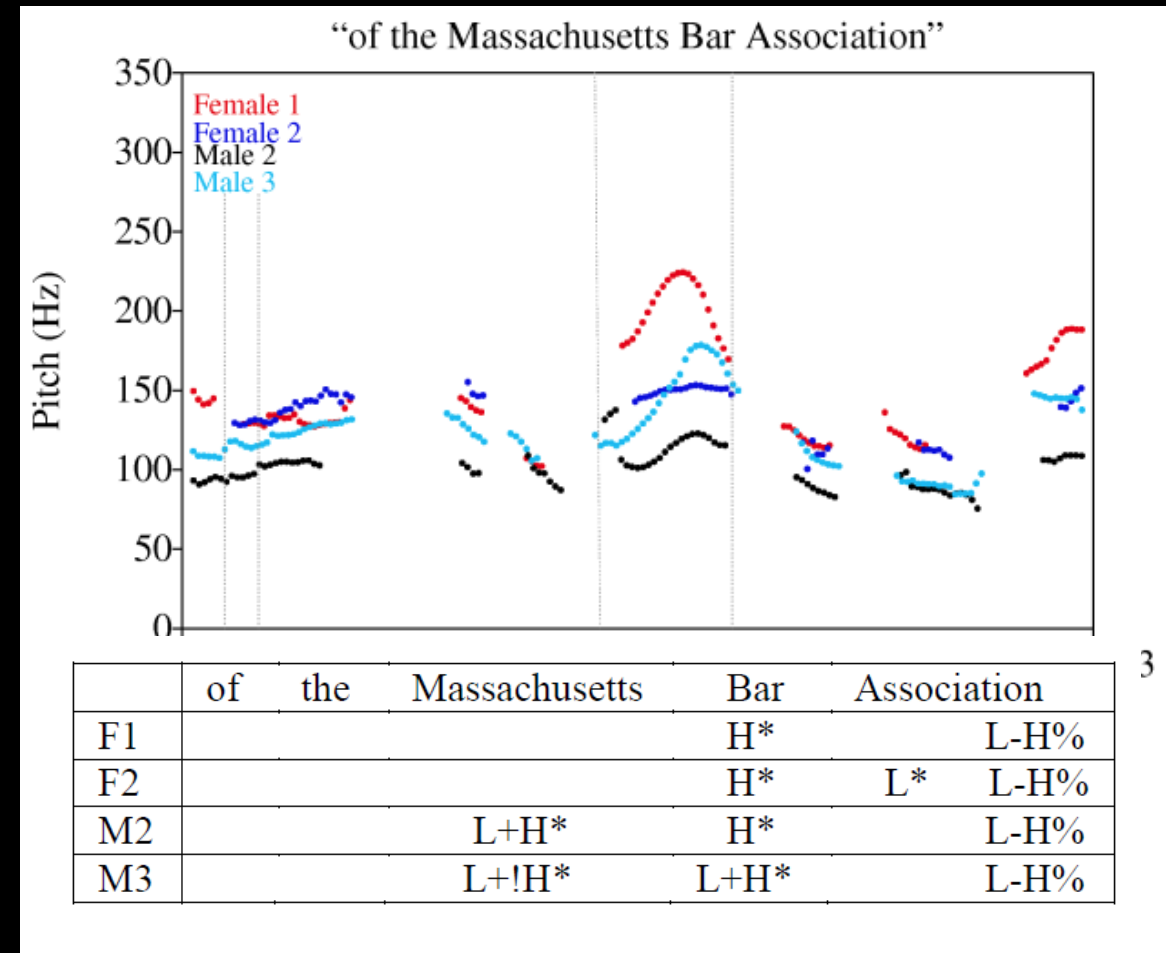


Variability in production

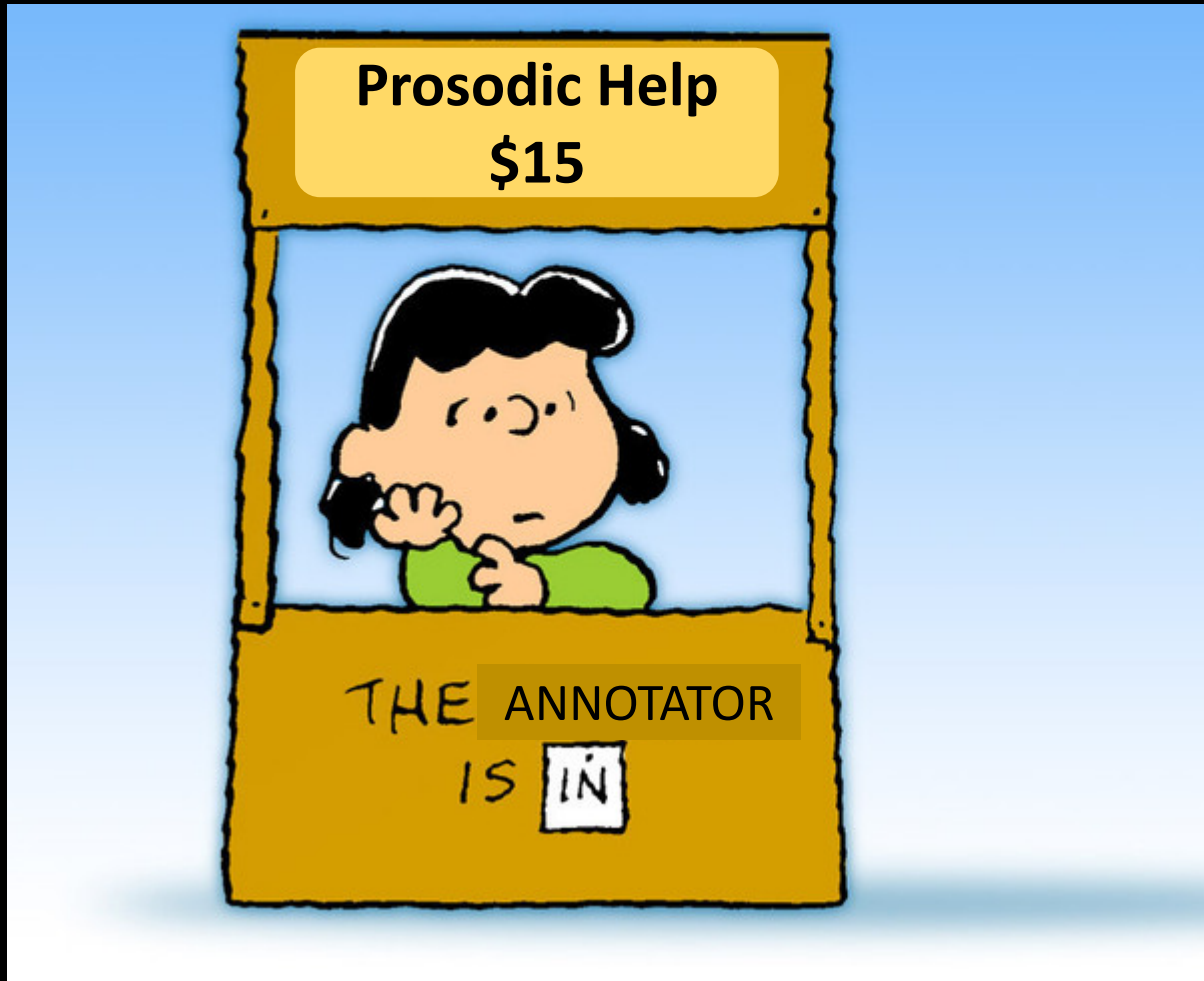
Speaker differences in

- phonological specification (ToBI)
- detailed acoustic cues

See also Grice et al., 2017; Cangemi et al. 2015;
Cole & Shattuck-Hufnagel 2011; Yoon 2010;
Peppé et al. 2000; Grabe 2004; Cole et al. 2007



The Annotator's Dilemma



- Ambiguity in acoustic cues
- Conflict between signal and top-down expectations
- Result: Annotator uncertainty

A solution? Multiple annotators, consensus, arbitration, majority rule



~~Impact~~ Opportunity for research



Search for signal in the noise...
Is there information in annotator disagreement?
Are inter-annotator differences systematic?

A search for individual differences in the perception of prosodic features

Perceptual sensitivity to acoustic cues to prosody



Sensitivity to contextual factors that predict prosodic features

Relative weighting among acoustic cues to prosody

Attentional focus on cues to linguistic vs. indexical information

This talk

- An exploratory study of individual listener differences in the perception of prosodic features in conversational speech, viewed through lens of prosodic annotation, by untrained listeners, performing real-time annotation.
- How does an individual listener's prosodic annotation relate to the presence of acoustic cues and other properties of the linguistic context?

What we (hope to) learn from noisy annotations

- Are untrained annotators **systematic** in their rating of the prosodic features of a word?
- Are there individual **differences** in cue selection? In relative cue weighting?

Spoiler alert: differences arise in cue selection and/or cue weighting, but they do not qualitatively restructure the mapping between sound and meaning.

Methods

- 32 participants (native, monolingual, American English). No training.
- Speech materials: Buckeye Corpus (Pitt et al. 2007), short excerpts (13-24 s) from 16 speakers. 932 words total
- Rapid Prosody Transcription: real-time, auditory annotation of prominence and boundary, using custom web-based tool (LMEDS: Mahrt 2016)
- PLUS: one ToBI annotation of same materials, performed by trained annotators (J. Cole & J.I. Hualde)

Attentional focus

The RPT task was performed twice, by the same annotators.

Instructions called the annotator's **attention** to sound- or meaning-related criteria in rating prominence and boundary:

<SOUND> “Listen for words that stand out due to pitch, loudness, tempo, or discontinuities or breaks in the speech stream.”

These are the primary data presented here...

Attentional focus

The annotation task was performed twice, by the same annotators. Instructions called the annotator's **attention** to sound- or meaning-related criteria in rating prominence and boundary:

<SOUND> “Listen for words that stand out due to pitch, loudness, tempo, or discontinuities or breaks in the speech stream.”

<MEANING> “Listen for words that convey the main points of information, or places where the speech stream could be segmented with minimal disruption of meaning.”

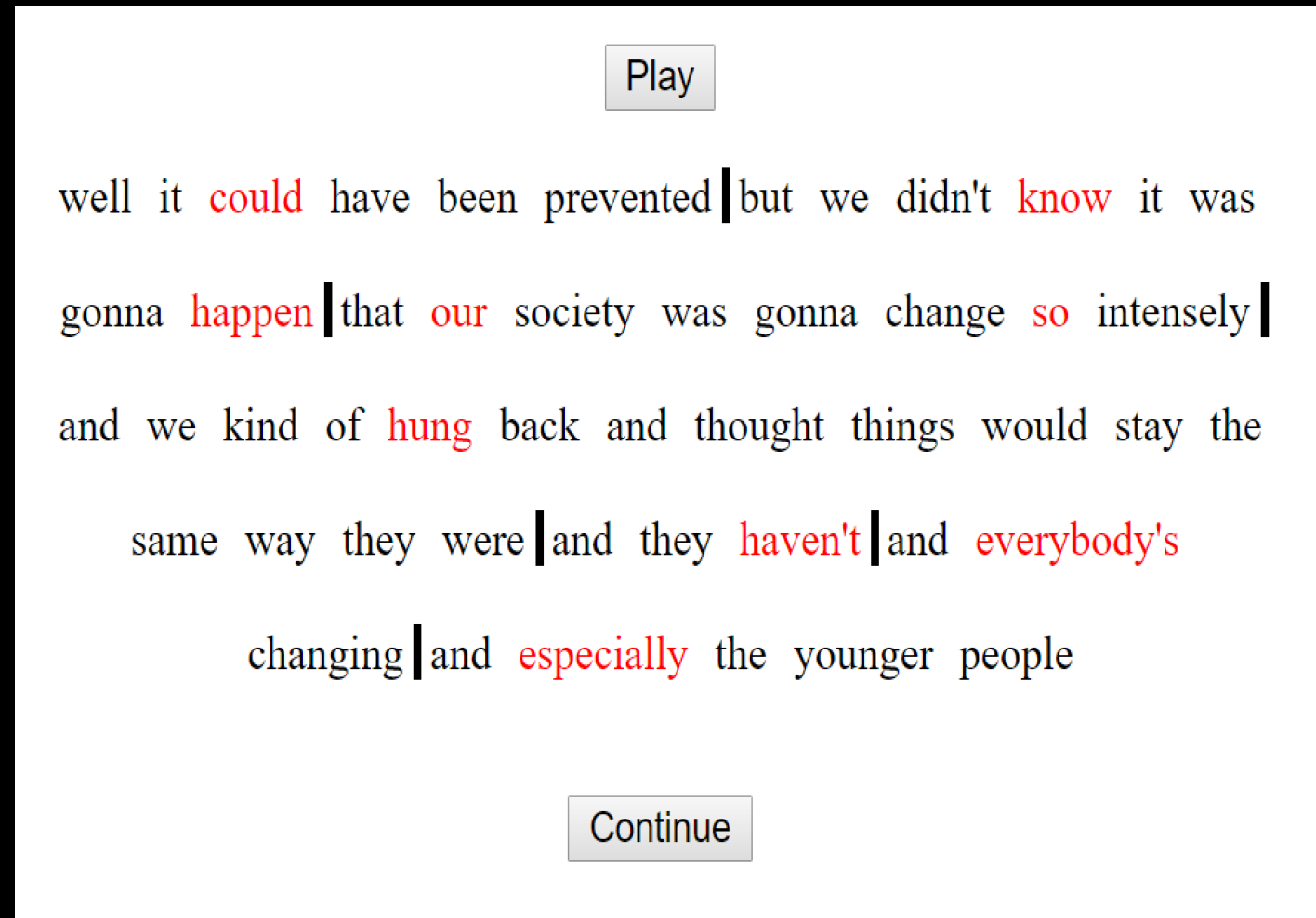
--we'll briefly turn to these data at the end

Rapid Prosody Transcription

Screenshot of LMEDS interface for prosodic annotation.

Annotator selects a word perceived as prominent (red), and/or a following boundary (vertical bar).

Note: annotator does not view pitch track or other acoustic display



Play

well it **could** have been prevented|but we didn't **know** it was gonna **happen**|that **our** society was gonna change **so** intensely|and we kind of **hung** back and thought things would stay the same way they were|and they **haven't**|and **everybody's** changing|and **especially** the younger people

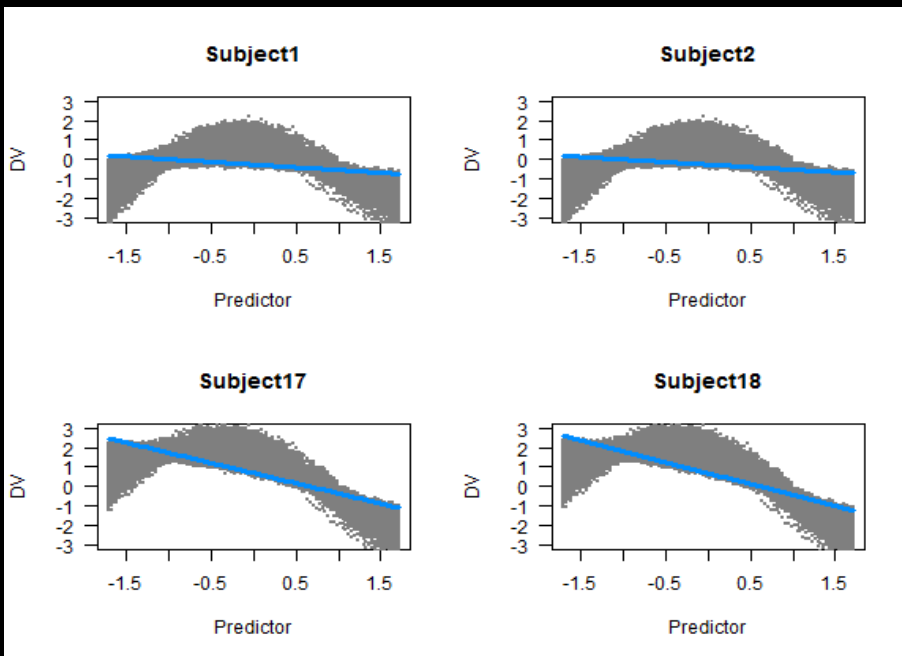
Continue

Statistical methods

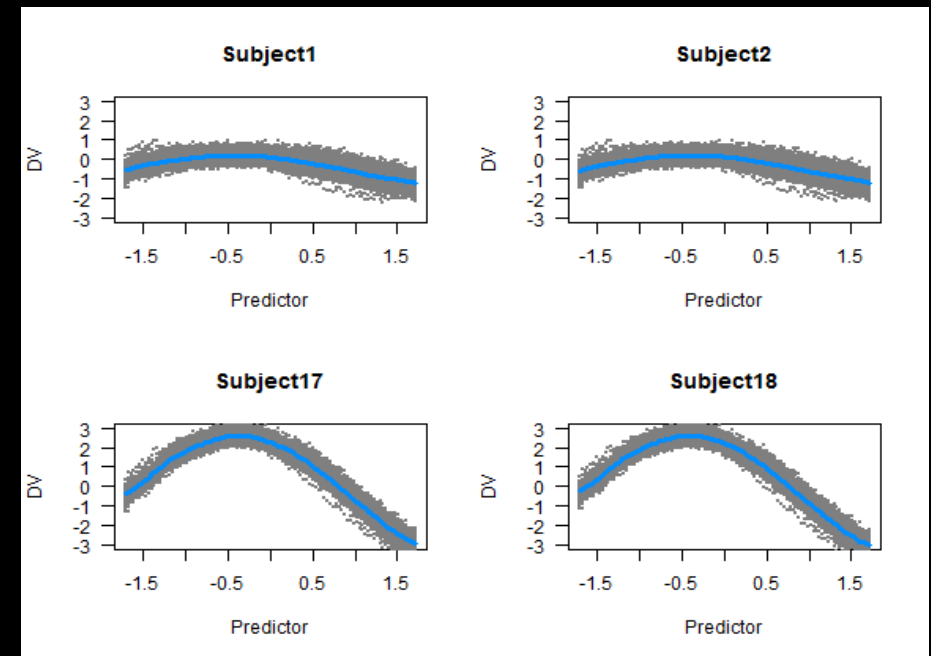
- We analyze the prominence and boundary label assigned to each content word in corpus:
 - Individual annotator rating: 0 or 1
 - Average rating from pooled annotators: 0-1
- Fleiss' Kappa, for inter-annotator agreement (reliability)
 - measures observed agreement in relation to expected agreement, controlling for the overall frequency of each label
- Generalized Additive Mixed Models (GAMMs), logistic regression for individual differences in cue selection and weighting
 - Modeling in R with the *bam* function: GAMs for very large data sets

- GAMMs can model non-linear effects of a predictor on prosody ratings, and an individual participant's non-linear deviations from the overall fixed effect.
- Compare linear model and GAMM on simulated data:

Linear Mixed effect Model (simulation)



GAMM (simulation)



Predictor variables

Acoustic measures:

- **Local RMS intensity**
- **Local max F0** (log)



These measures taken from the primary stressed vowel of each word, locally normalized using z-transform in window of 5 stressed syllables

Predictor variables

Acoustic measures:

- **Local RMS intensity**
- **Local max F0** (log)
- **Local tempo** -- measure based on Pfitzinger's (1998) algorithm. Tracks local changes in phone rate that result in lengthening or shortening of word duration; *statistically independent of word frequency*

Predictor variables

Acoustic measures:

- **Local RMS intensity**
- **Local max F0 (log)**
- **Local tempo**
- **Post-pause duration**

Predictor variables

Acoustic measures:

- **Local RMS intensity**
- **Local max F0** (log)
- **Local tempo**
- **Post-pause duration**

Non-acoustic:

- **Word frequency** (*in Switchboard corpus*)
- **Part of Speech**
- **Boundary marked** by same annotator (*for Prominence*)

These expectation-driven factors shown to influence prosodic ratings in our prior work

(Cole et al., 2010a, b)

Predictor variables

Acoustic measures:

- **Local RMS intensity**
- **Local max F0** (log)
- **Local tempo**
- **Post-pause duration**

Plus:

- Random intercept for lexical item
- Random smooths for each predictor, by annotator

Non-acoustic:

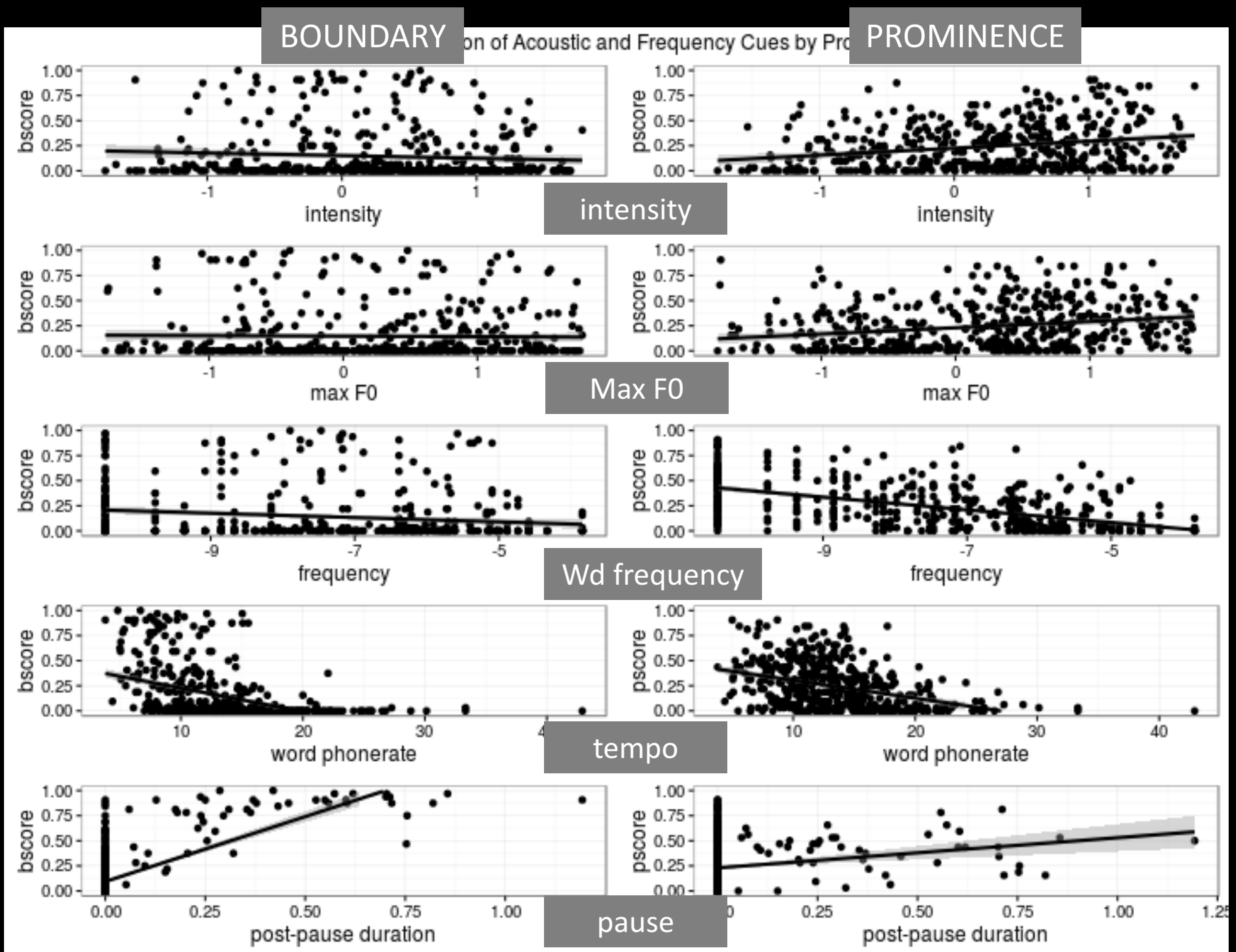
- **Word frequency** (*in Switchboard corpus*)
- **Part of Speech**
- **Boundary marked** by same annotator (*for Prominence*)

Results (1): Inter-annotator agreement

- 477 content words x 32 annotators = 15,264 observed ratings for boundary, same number of prominence ratings
- **Fleiss' Kappa: .52 for boundary** (high-moderate)
.28 for prominence (low-moderate)

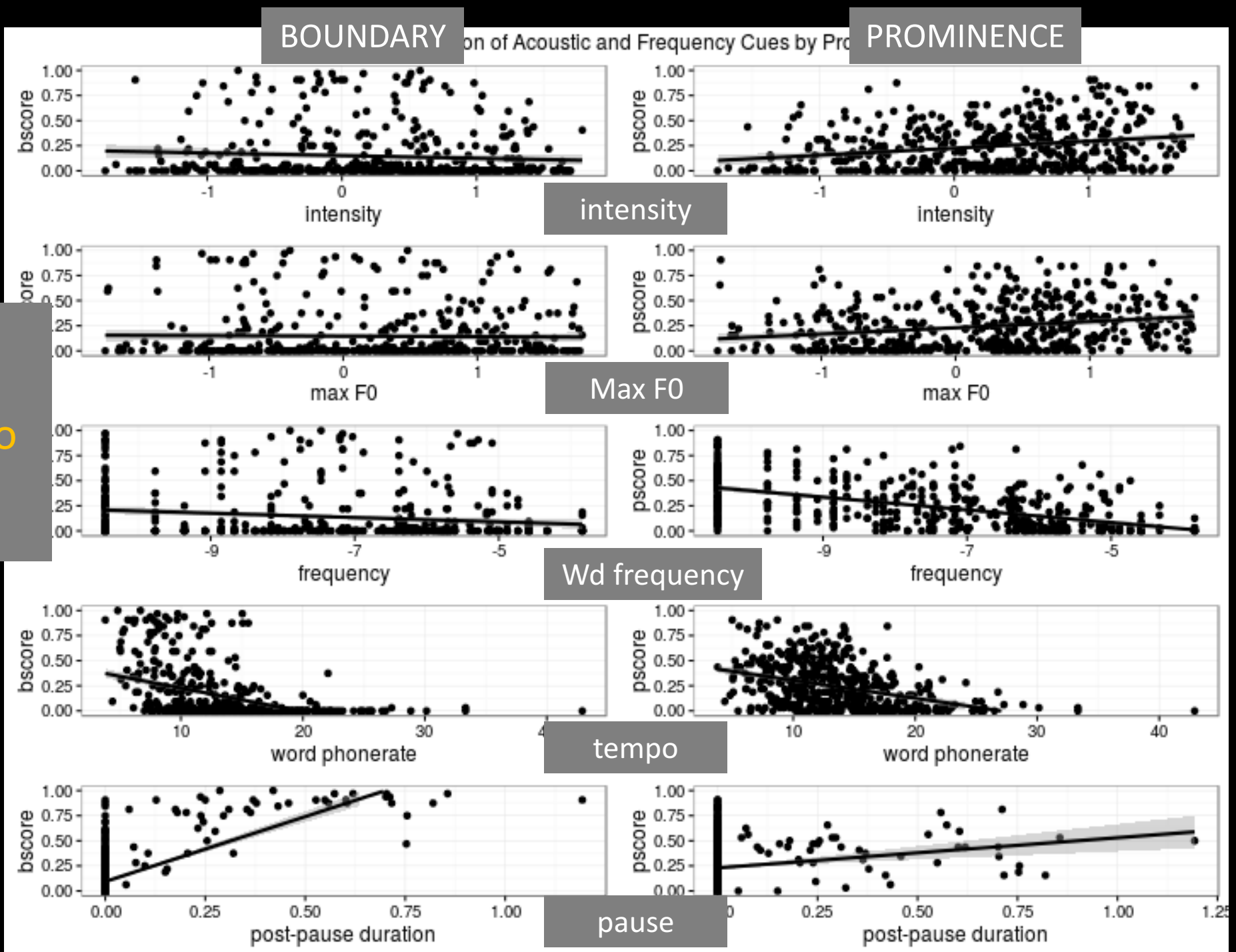
Results (2)

Distributions of average boundary (left) and prominence (right) ratings across range of each predictor variable.



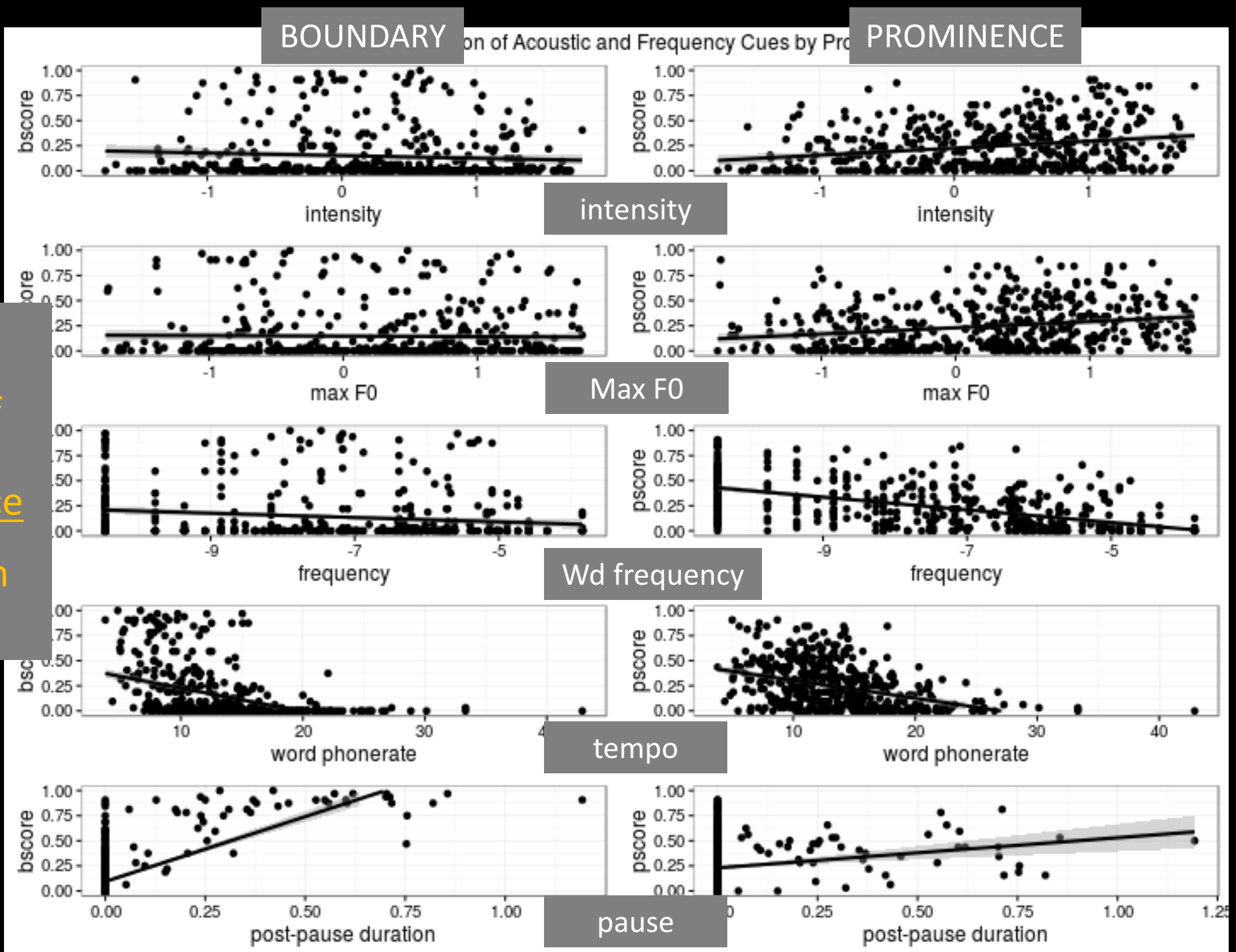
Results (2)

Not bimodal...
Suggests that no individual cue suffices to signal prosodic distinction



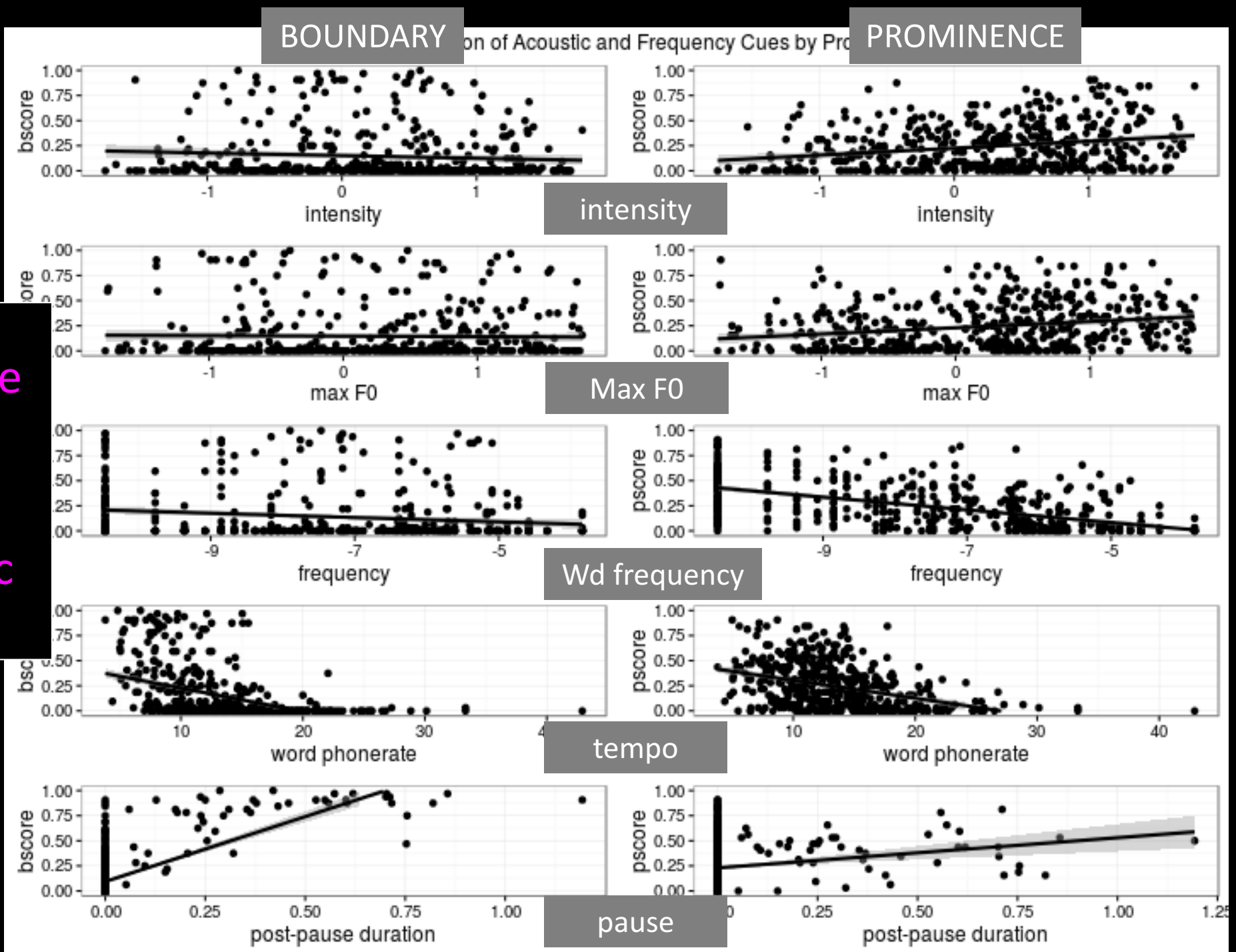
Results (2)

Average prosody rating skewed towards 0
(data denser at bottom of plots)
Agreement on the absence of prosodic feature is higher than agreement on its presence

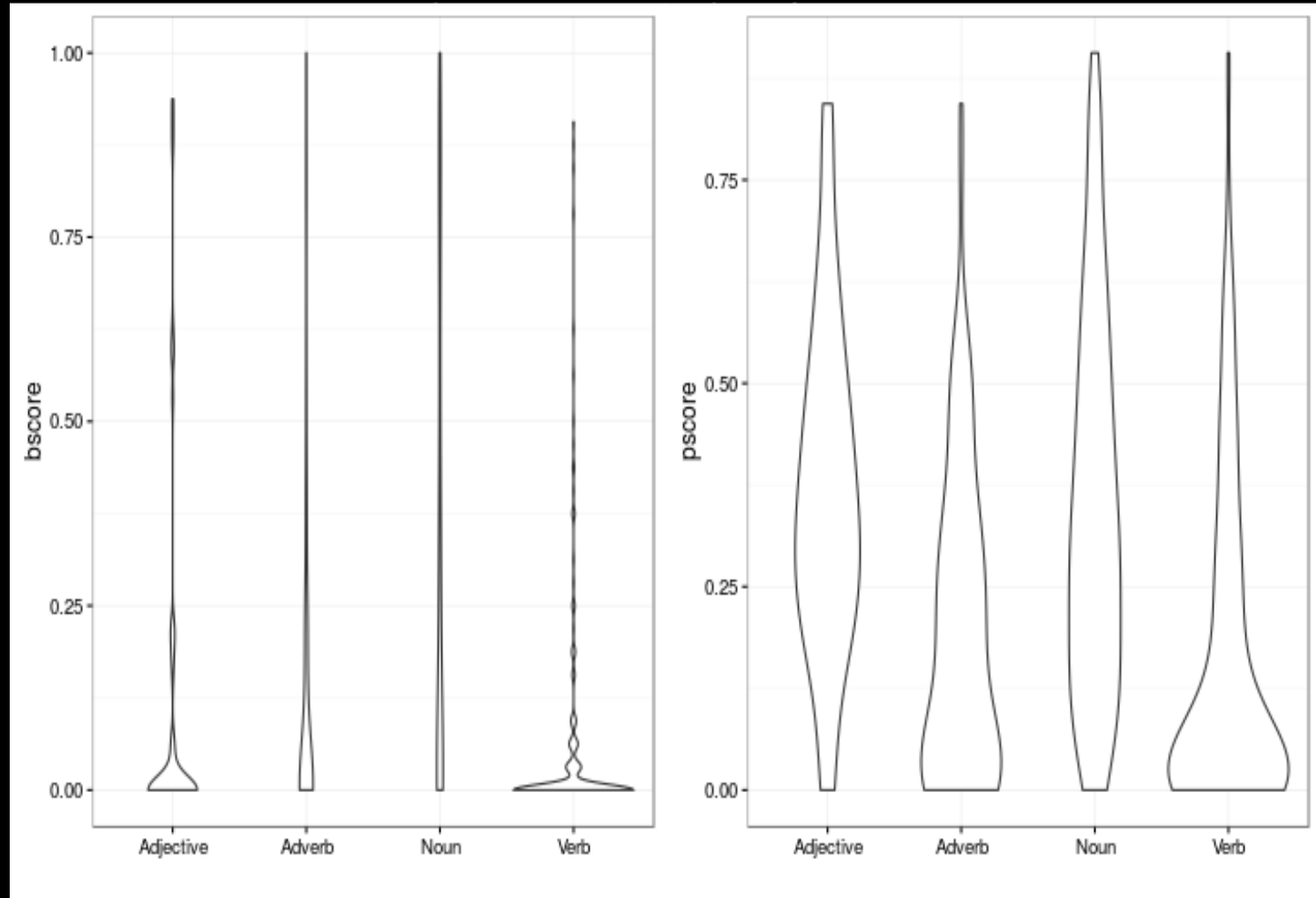


Results (2)

Is any of the noise in these distributions due to individual differences in how acoustic continua are mapped onto prosodic distinctions?



Distributions of average boundary (left) and prominence (right) ratings across categorical predictors.



Results (3): GAMM

Modeling individual annotator ratings

Predictor variables:

- Local RMS intensity
- Local max F0 (log)
- Local tempo
- Post-pause duration
- Word frequency
- Part of Speech
- (Boundary marked)

Random factors: word, annotator

Results (3): GAMM

Modeling individual annotator ratings

- **Deviance explained: Boundary 63%; Prominence 37%**

Results (3): GAMM

Modeling individual annotator ratings

- Deviance explained: Boundary 63%; Prominence 37%
- All predictors show significant effects:
 - tempo, max F0, intensity, pause, word frequency, POS
- Significant random effect of word
- Significant random smooths for annotator, with all predictors (except POS)

Non-parametric tests for smooth terms

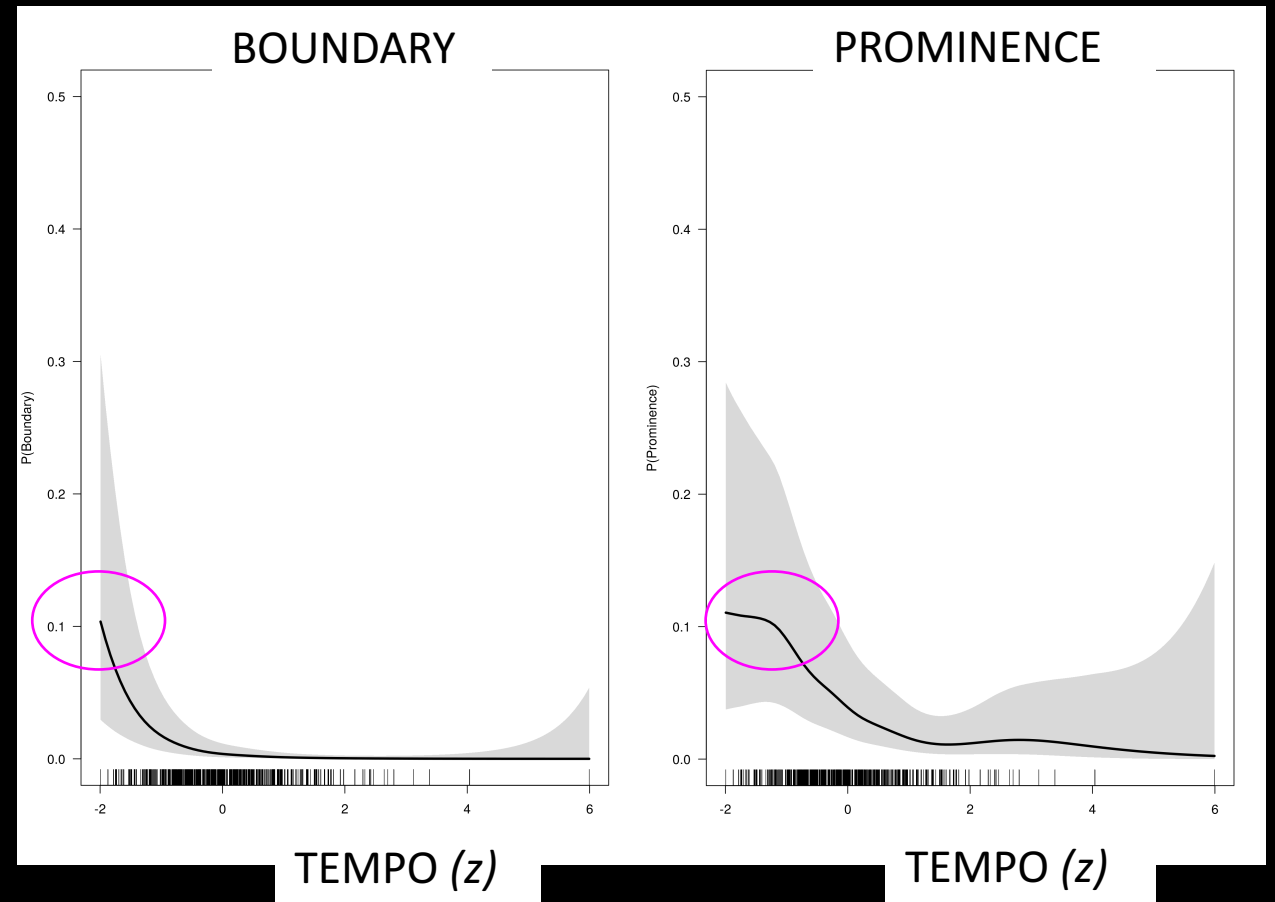
	Boundary (Deviance Explained = 63%)				Prominence (Dev. Explained = 37 %)			
Fixed Effects: (smooth terms)	Effective DF	Residual Effect DF	Chi.sq	p-value	Edf	Ref DF	Chi.sq	p-value
Tempo	2.96	3.60	175.57	<.001	5.13	6.09	78.52	<.001
Max F0	8.12	8.66	38.19	<.001	7.35	8.12	128.90	<.001
Intensity	7.65	8.32	61.13	<.001	6.86	7.71	22.73	.003
Word Frequency	1.00	1.00	5.06	0.02	1.00	1.00	40.88	<.001
Pause	4.26	4.87	249.54	<.001	5.74	6.45	16.39	.02
Random Effects: (smooth terms)								
Word	178.33	270	1253.24	<.001	225.75	270	1458.47	<.001
Subject x Tempo	34.42	287	55.34	<.001	63.439	287	168.76	<.001
Subject x F0	2.26	287	2.45	<.001	28.80	287	42.09	<.001
Subject x Intensity	2.25	287	2.45	<.001	40.74	287	71.35	<.001
Subject x Wd Frequency	16.54	287	21.63	0.004	75.90	287	234.19	<.001
Subject x Pause	57.39	287	184.71	<.001	2.54	287	2.81	<.001
Subject x Part of Speech	0.01	124	0.01	0.39	28.37	124	41.95	.001
Subject x Boundary mark					22.08	62.00	64.66	<.001

Parametric tests for categorical predictors

	Boundary				Prominence			
Fixed effect	Estimate	Std. Error	z	P-value	Estimate	Std. Error	z	P-value
(Intercept)	-4.5389	0.3756	-12.085	<.0001	-0.56	0.21	-2.62	.01
POS 1 [Adjective vs Noun]	1.2896	0.3539	3.644	0.0003	-1.28	0.21	-6.18	<.0001
POS 2 [Adverb vs Noun]	1.3413	0.3526	3.804	0.0001	-1.01	0.18	-5.63	<.0001
POS 3 [Verb vs Noun]	0.2479	0.3839	0.646	0.5184	-1.23	0.19	-6.18	<.0001
Boundary Marked					0.47	0.12	3.92	.0001

Visualizations show estimated effects (and CIs) of predictor variables on increasing/decreasing the likelihood of prominence or boundary marking.

Estimated likelihood of prosodic marking



Predictor variables can be ranked in terms of their peak effect in increasing/decreasing the likelihood of prominence or boundary marking.

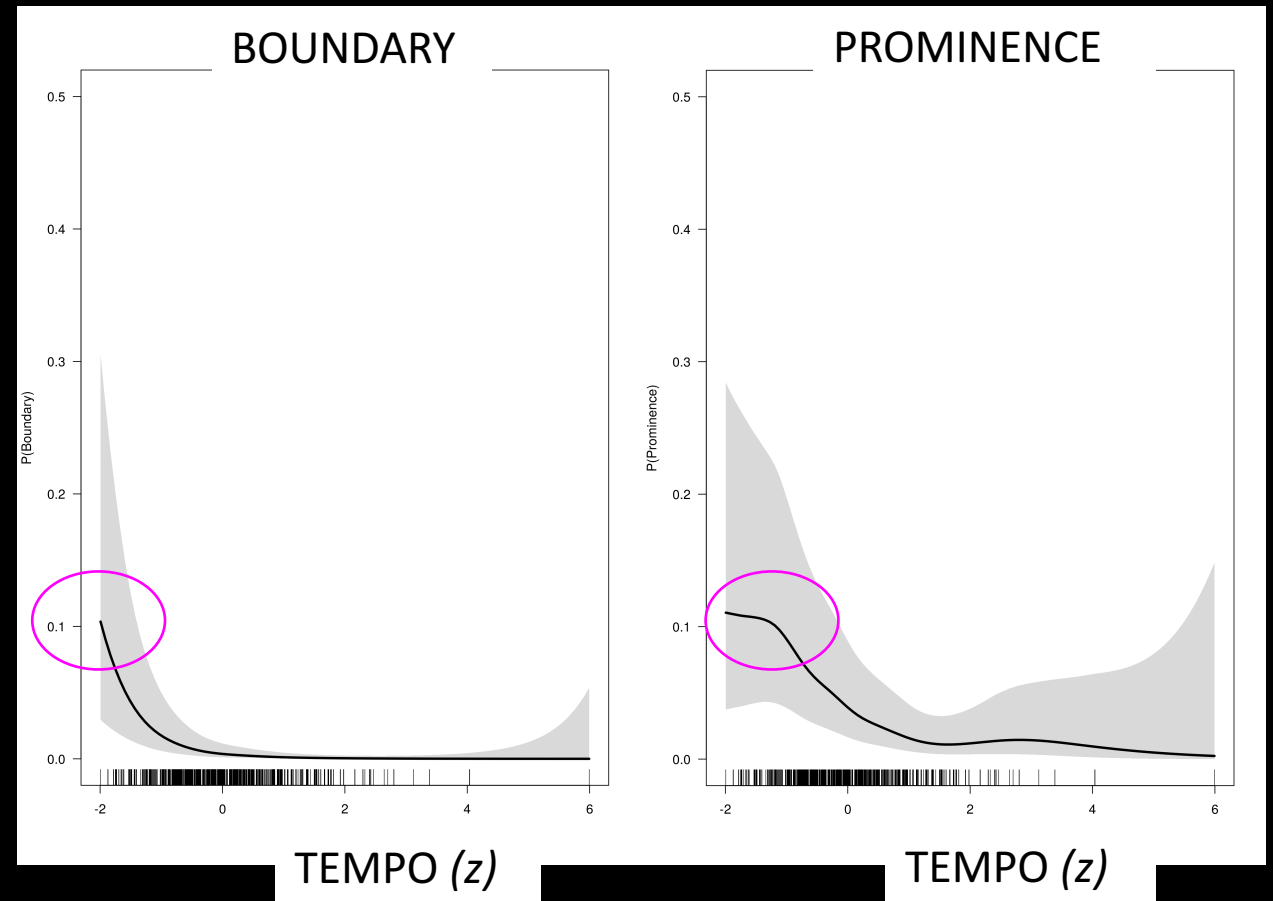
Boundary predictors:

Pause >> Tempo

Prominence predictors:

Word Freq. >> POS > Tempo

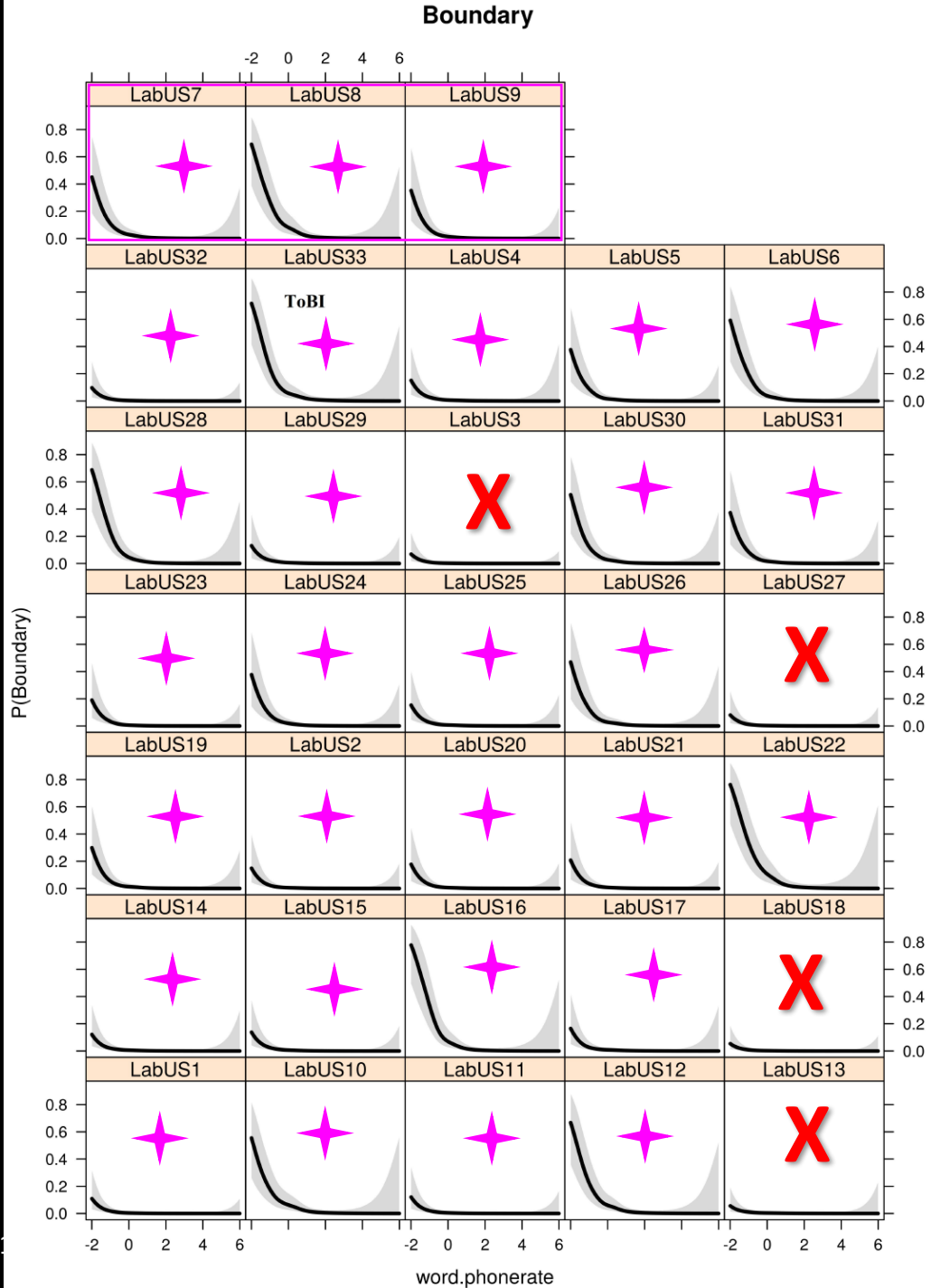
Estimated likelihood of prosodic marking



Results (4): Individual annotator differences

- Analysis based on GAMM visualization (as just seen)
- Next slides:

Effects of Tempo on Boundary marking

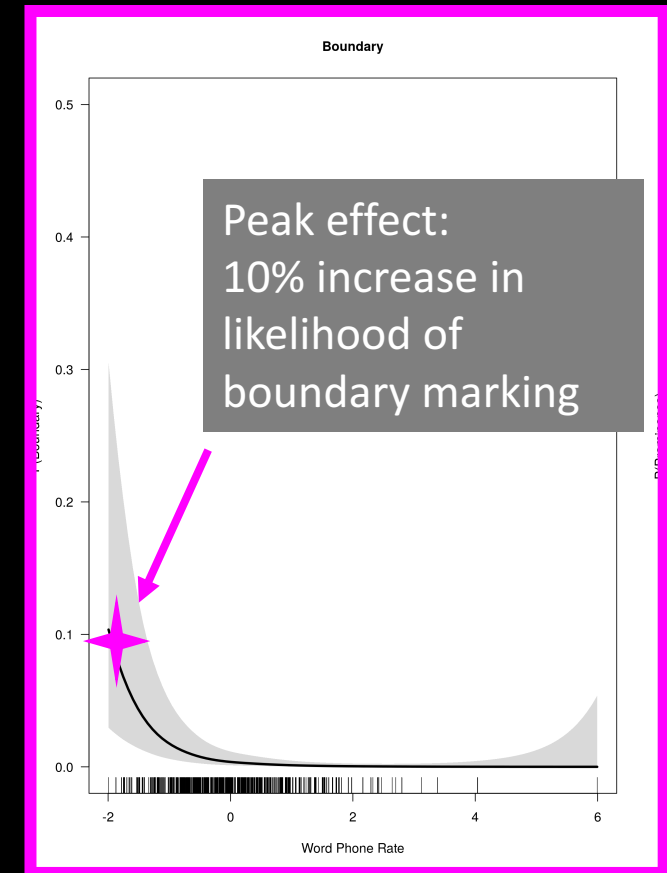


Words with the lowest **tempo** are more likely to be marked for a following prosodic **boundary**

Graphs of model estimates

← *By participant*

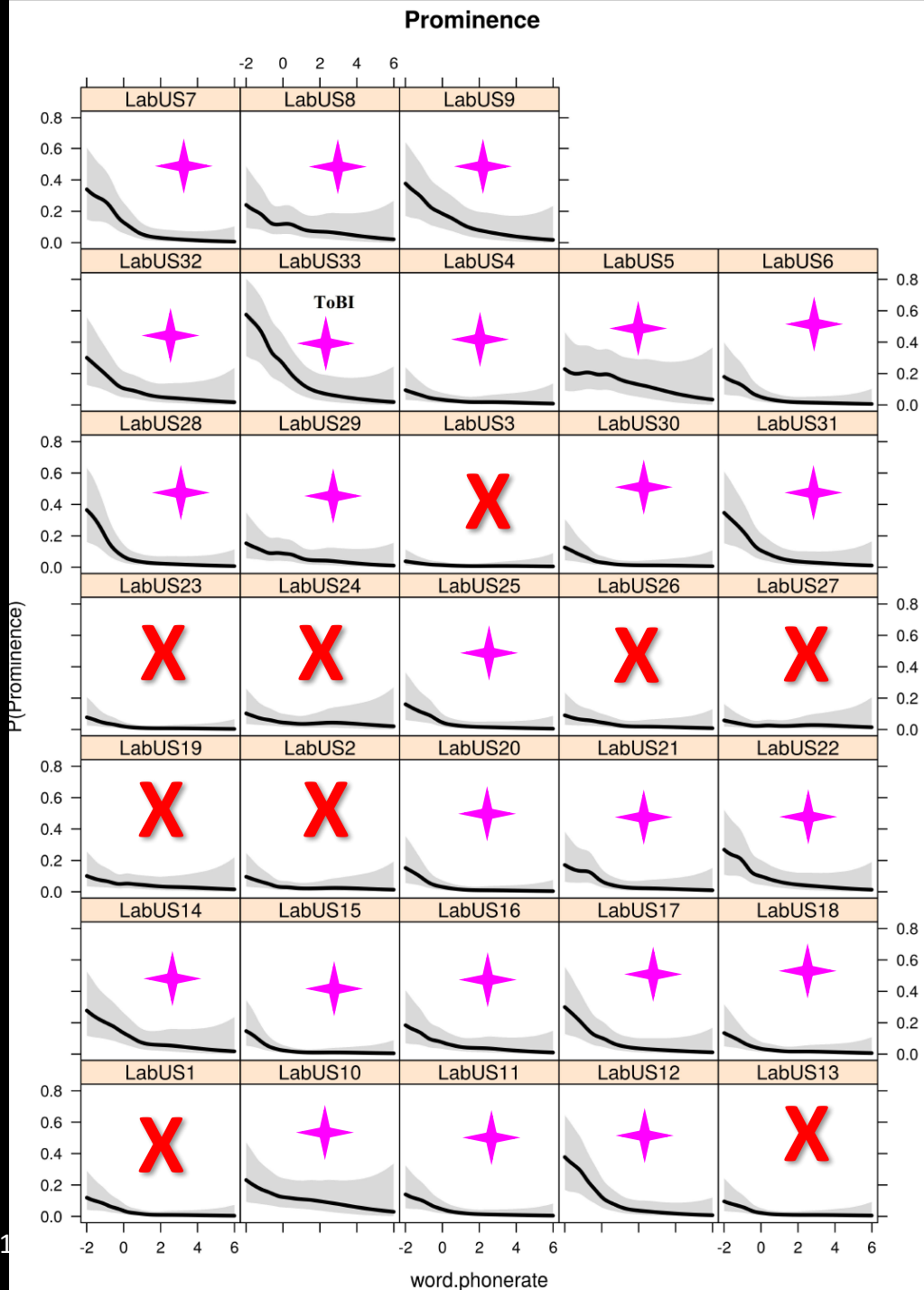
full model →



Results (4): Individual annotator differences

- Next slides:

Effects of Tempo on Prominence marking

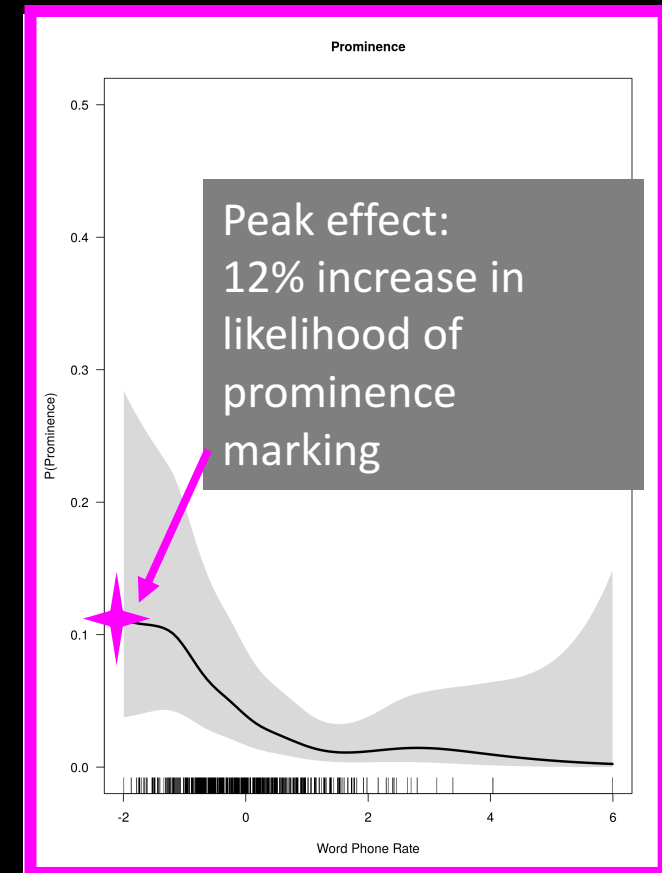


Words with the lowest **tempo** are more likely to be marked for a following prosodic **prominence**

Graphs of model estimates

← *By participant*

full model →



Individual differences in cue weighting

- Visualizations of GAMM estimates confirm that effect patterns across annotators vary in magnitude, but the overall effect pattern is the same.

Individual differences in cue selection

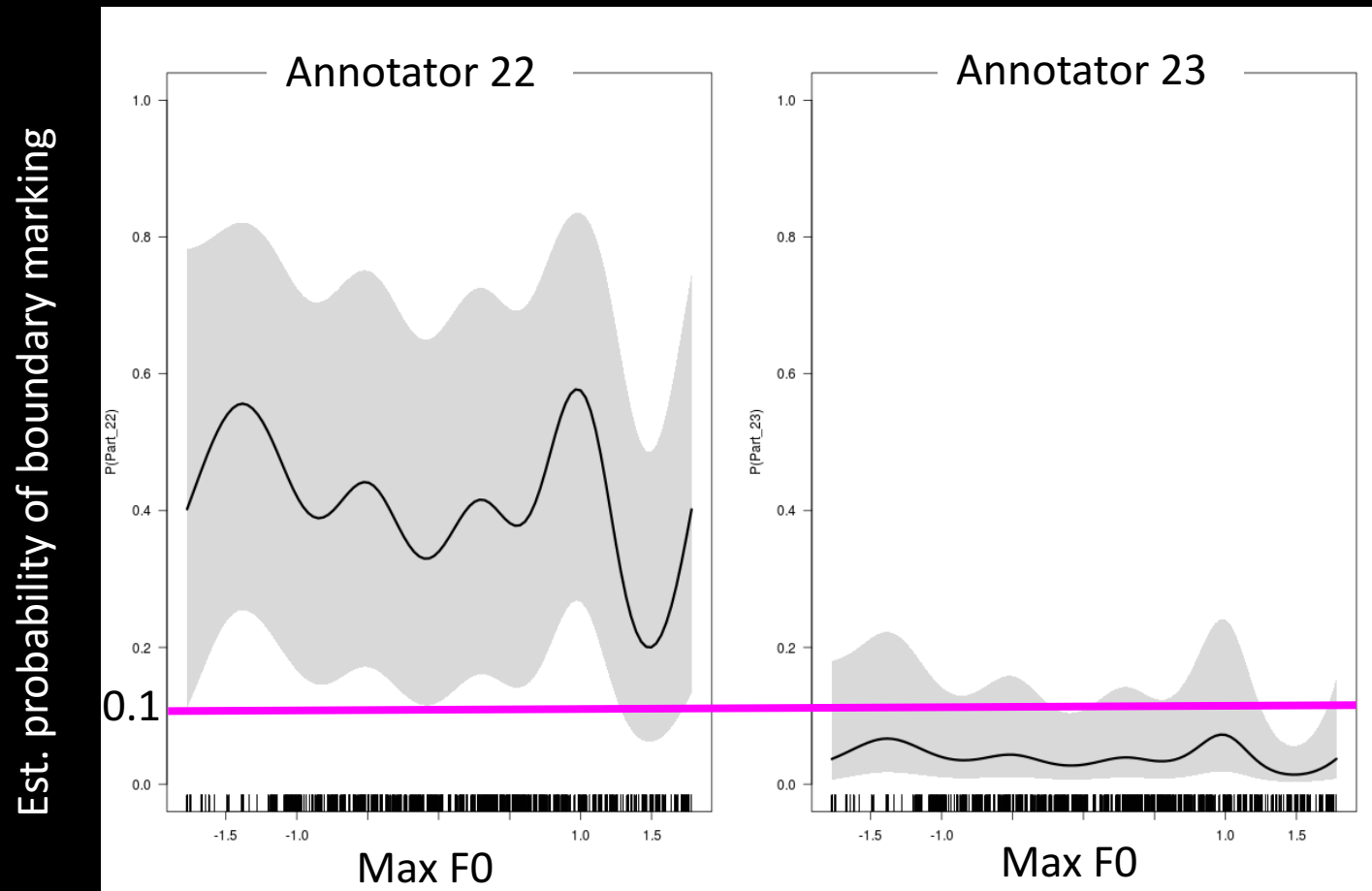
Informally assessed by coding each predictor (cue) for its (non-)selection by each annotator:

- A cue is “selected” by an individual annotator if it boosts prominence/boundary marking by 10% or more, across any range of its values.
- A predictor that fails to reach the threshold effect size is coded as “not selected”

Illustration of 10% boost threshold for estimated effect of Max F0 on the probability of boundary marking, for two annotators

“Selected”

“Not selected”



BOUNDARY

Selected
(light) and
unselected
(gray) cues for
32
annotators

Annotator	Intensity	Max F0	Wd. Freq.	Tempo	Pause	POS
ToBI 33	+	+	+	+	+	+
8	+	+	+	+	+	+
10	+	+	+	+	+	+
12	+	+	+	+	+	+
16	+	+	+	+	+	+
22	+	+	+	+	+	+
28	+	+	+	+	+	+
6	+	-	+	+	+	+
26	+	-	-	+	+	+
24	+	-	-	+	+	-
30	+	-	-	+	+	-
31	+	-	-	+	+	-
1	-	-	-	+	+	-
2	-	-	-	+	+	-
4	-	-	-	+	+	-
5	-	-	-	+	+	-
7	-	-	-	+	+	-
9	-	-	-	+	+	-
11	-	-	-	+	+	-
14	-	-	-	+	+	-
15	-	-	-	+	+	-
17	-	-	-	+	+	-
19	-	-	-	+	+	-
20	-	-	-	+	+	-
21	-	-	-	+	+	-
23	-	-	-	+	+	-
25	-	-	-	+	+	-
29	-	-	-	+	+	-
32	-	-	-	+	+	-
3	-	-	-	-	+	-
13	-	-	-	-	+	-
18	-	-	-	-	+	-
27	-	-	-	-	+	-

BOUNDARY

6 RPT
annotators
and the ToBI
annotator
select all cues

Annotator	Intensity	Max F0	Wd. Freq.	Tempo	Pause	POS
ToBI 33	+	+	+	+	+	+
8	+	+	+	+	+	+
10	+	+	+	+	+	+
12	+	+	+	+	+	+
16	+	+	+	+	+	+
22	+	+	+	+	+	+
28	+	+	+	+	+	+
6	+	-	+	+	+	+
26	+	-	-	+	+	+
24	+	-	-	+	+	-
30	+	-	-	+	+	-
31	+	-	-	+	+	-
1	-	-	-	+	+	-
2	-	-	-	+	+	-
4	-	-	-	+	+	-
5	-	-	-	+	+	-
7	-	-	-	+	+	-
9	-	-	-	+	+	-
11	-	-	-	+	+	-
14	-	-	-	+	+	-
15	-	-	-	+	+	-
17	-	-	-	+	+	-
19	-	-	-	+	+	-
20	-	-	-	+	+	-
21	-	-	-	+	+	-
23	-	-	-	+	+	-
25	-	-	-	+	+	-
29	-	-	-	+	+	-
32	-	-	-	+	+	-
3	-	-	-	-	+	-
13	-	-	-	-	+	-
18	-	-	-	-	+	-
27	-	-	-	-	+	-

BOUNDARY

17
annotators
select only
Tempo and
Pause as cues

Annotator	Intensity	Max F0	Wd. Freq.	Tempo	Pause	POS
ToBI 33	+	+	+	+	+	+
8	+	+	+	+	+	+
10	+	+	+	+	+	+
12	+	+	+	+	+	+
16	+	+	+	+	+	+
22	+	+	+	+	+	+
28	+	+	+	+	+	+
6	+	-	+	+	+	+
26	+	-	-	+	+	+
24	+	-	-	+	+	-
30	+	-	-	+	+	-
21	+	-	-	+	+	-
1	-	-	-	+	+	-
2	-	-	-	+	+	-
4	-	-	-	+	+	-
5	-	-	-	+	+	-
7	-	-	-	+	+	-
9	-	-	-	+	+	-
11	-	-	-	+	+	-
14	-	-	-	+	+	-
15	-	-	-	+	+	-
17	-	-	-	+	+	-
19	-	-	-	+	+	-
20	-	-	-	+	+	-
21	-	-	-	+	+	-
23	-	-	-	+	+	-
25	-	-	-	+	+	-
29	-	-	-	+	+	-
32	-	-	-	+	+	-
3	-	-	-	-	+	-
13	-	-	-	-	+	-
18	-	-	-	-	+	-
27	-	-	-	-	+	-

BOUNDARY

Everyone
selects
Pause

Annotator	Intensity	Max F0	Wd. Freq.	Tempo	Pause	POS
ToBI 33	+	+	+	+	+	+
8	+	+	+	+	+	+
10	+	+	+	+	+	+
12	+	+	+	+	+	+
16	+	+	+	+	+	+
22	+	+	+	+	+	+
28	+	+	+	+	+	+
6	+	-	+	+	+	+
26	+	-	-	+	+	+
24	+	-	-	+	+	-
30	+	-	-	+	+	-
31	+	-	-	+	+	-
1	-	-	-	+	+	-
2	-	-	-	+	+	-
4	-	-	-	+	+	-
5	-	-	-	+	+	-
7	-	-	-	+	+	-
9	-	-	-	+	+	-
11	-	-	-	+	+	-
14	-	-	-	+	+	-
15	-	-	-	+	+	-
17	-	-	-	+	+	-
19	-	-	-	+	+	-
20	-	-	-	+	+	-
21	-	-	-	+	+	-
23	-	-	-	+	+	-
25	-	-	-	+	+	-
29	-	-	-	+	+	-
32	-	-	-	+	+	-
3	-	-	-	-	+	-
13	-	-	-	-	+	-
18	-	-	-	-	+	-
27	-	-	-	-	+	-

PROMINENCE

Selected
(light) and
unselected
(gray) cues for
32
annotators

*Nb, rows not
same as
Boundary table*

Annotator	Intensity	Max F0	Wd. Freq.	Tempo	Pause	POS	Boundary
ToBI 33	+	+	+	+	+	+	+
5	+	+	+	+	+	+	+
7	+	+	+	+	+	+	+
8	+	+	+	+	+	+	+
9	+	+	+	+	+	+	+
10	+	+	+	+	+	+	+
12	+	+	+	+	+	+	+
14	+	+	+	+	+	+	+
16	+	+	+	+	+	+	+
17	+	+	+	+	+	+	+
22	+	+	+	+	+	+	+
28	+	+	+	+	+	+	+
29	+	+	+	+	+	+	+
31	+	+	+	+	+	+	+
32	+	+	+	+	+	+	+
6	-	+	+	+	+	+	+
11	-	+	+	+	+	+	+
21	-	+	+	+	+	-	-
4	-	-	+	+	-	+	+
18	-	-	+	+	-	+	+
20	-	-	+	+	-	+	+
25	-	-	+	+	-	+	-
26	-	-	+	-	+	+	-
24	-	-	+	-	-	+	+
1	-	-	+	-	-	+	-
15	-	-	-	+	-	-	+
30	-	-	-	+	-	-	-
23	-	-	-	-	-	+	-
27	-	-	-	-	-	+	-
13	-	-	-	-	-	-	+
19	-	-	-	-	-	-	+
2	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-

PROMINENCE

14 RPT
 annotators
 and the ToBI
 annotator
 select all cues

Annotator	Intensity	Max F0	Wd. Freq.	Tempo	Pause	POS	Boundary
ToBI 33	+	+	+	+	+	+	+
5	+	+	+	+	+	+	+
7	+	+	+	+	+	+	+
8	+	+	+	+	+	+	+
9	+	+	+	+	+	+	+
10	+	+	+	+	+	+	+
12	+	+	+	+	+	+	+
14	+	+	+	+	+	+	+
16	+	+	+	+	+	+	+
17	+	+	+	+	+	+	+
22	+	+	+	+	+	+	+
28	+	+	+	+	+	+	+
29	+	+	+	+	+	+	+
31	+	+	+	+	+	+	+
32	+	+	+	+	+	+	+
11	-	+	+	+	+	+	+
21	-	+	+	+	+	-	-
4	-	-	+	+	-	+	+
18	-	-	+	+	-	+	+
20	-	-	+	+	-	+	+
25	-	-	+	+	-	+	-
26	-	-	+	-	+	+	-
24	-	-	+	-	-	+	+
1	-	-	+	-	-	+	-
15	-	-	-	+	-	-	+
30	-	-	-	+	-	-	-
23	-	-	-	-	-	+	-
27	-	-	-	-	-	+	-
13	-	-	-	-	-	-	+
19	-	-	-	-	-	-	+
2	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-

PROMINENCE

No single cue selected by all annotators.

Less clustering of selected cues across annotators compared to Boundary

Annotator	Intensity	Max F0	Wd. Freq.	Tempo	Pause	POS	Boundary
ToBI 33	+	+	+	+	+	+	+
5	+	+	+	+	+	+	+
7	+	+	+	+	+	+	+
8	+	+	+	+	+	+	+
9	+	+	+	+	+	+	+
10	+	+	+	+	+	+	+
12	+	+	+	+	+	+	+
14	+	+	+	+	+	+	+
16	+	+	+	+	+	+	+
17	+	+	+	+	+	+	+
22	+	+	+	+	+	+	+
28	+	+	+	+	+	+	+
29	+	+	+	+	+	+	+
31	+	+	+	+	+	+	+
32	+	+	+	+	+	+	+
6	-	+	+	+	+	+	+
11	-	+	+	+	+	+	+
21	-	+	+	+	+	-	-
4	-	-	+	+	-	+	+
18	-	-	+	+	-	+	+
20	-	-	+	+	-	+	+
25	-	-	+	+	-	+	-
26	-	-	+	-	+	+	-
24	-	-	+	-	-	+	+
1	-	-	+	-	-	+	-
15	-	-	-	+	-	-	+
30	-	-	-	+	-	-	-
23	-	-	-	-	-	+	-
27	-	-	-	-	-	+	-
13	-	-	-	-	-	-	+
19	-	-	-	-	-	-	+
2	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-

Cue hierarchies

Boundary cue selection implicational hierarchy:

$F0 \Rightarrow \text{Word Freq.} \Rightarrow \text{POS} \Rightarrow \text{Intensity} \Rightarrow \text{Tempo} \Rightarrow \text{Pause}$

Prominence cue selection implicational hierarchy:

$\text{Intensity} \Rightarrow \text{Max } F0 \Rightarrow \text{Pause} \Rightarrow \text{Word Freq.}$

Rational cue selection (frequency \Leftrightarrow strength)

Boundary cues ranked by selection frequency:

Pause > Tempo >>> Intensity > POS > Word Freq. > F0

Boundary cues ranked by effect size:

Pause >> Tempo > *others*

Rational cue selection (frequency \Leftrightarrow strength)

Boundary cues ranked by selection frequency:

Pause > Tempo >>> Intensity > POS > Word Freq. > F0

Boundary cues ranked by effect size:

Pause >> Tempo > others

Prominence cues ranked by selection frequency:

POS > Word Freq. > Boundary, Tempo > Pause > F0 > Intensity

Prominence cues ranked by effect size:

Word Freq. >> POS > Tempo

Rational cue selection (frequency \Leftrightarrow strength)

Boundary cues ranked by selection:

Pause > Tempo >>> *others*

Boundary cues ranked by effect:

Pause >> Tempo > *others*

BOUNDARY:
TIMING CUES

Prominence cues ranked by selection:

POS > Word Freq. > Bo

Prominence cues ranked by effect:

Word Freq. >> POS > Tempo

PROMINENCE:
NON-ACOUSTIC CUES

Attentional factors?

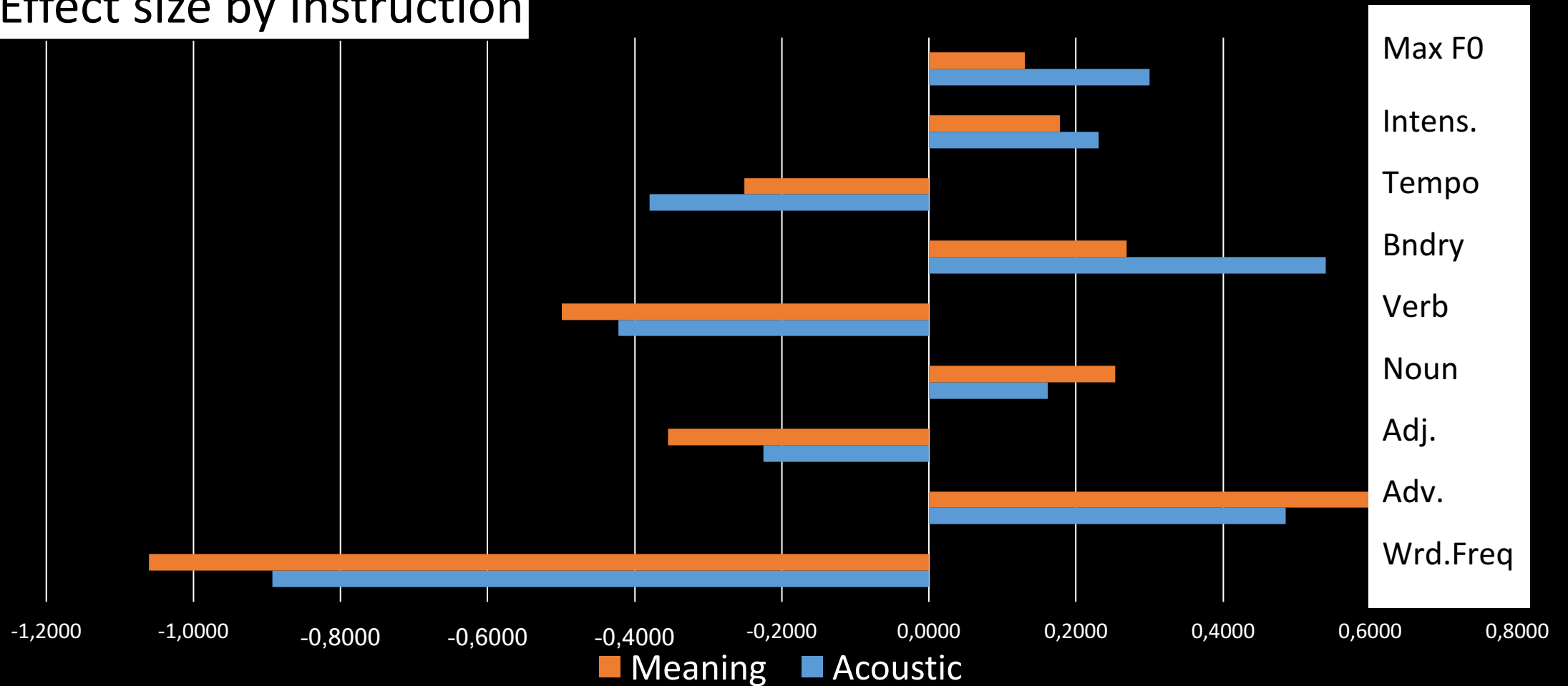
- Might individual differences in cue selection be driven by attentional focus? Do annotators attend more/less to some cues?
- Possibly... evidence for attentional effects comes from the second annotation task, where annotators were explicitly instructed to attend to factors related to meaning.

Attentional factors?

- Regression model shows significant effect of Instruction on prominence/boundary rating, and significant interactions of Instruction with most predictors.
- → Attentional focus determines cue weighting in full model (all annotators)

Attention to sound vs. meaning

Effect size by Instruction



Conclusions

- **Untrained annotators are systematic** in their rating of the prosodic features of a word.
 - Boundaries rating is more reliable than Prominence rating
 - Prosodic distinctions do not rest on individual cues; all cues tested contribute to prosodic rating.

Conclusions

- **Untrained annotators are systematic** in their rating of the prosodic features of a word.
- **Individual annotators differ** in cue selection, and in cue weighting, but the general pattern of effects for a selected cue are the same for all annotators.

Conclusions

- **Untrained annotators are systematic** in their rating of the prosodic features of a word.
- **Individual annotators differ** in cue selection, and in cue weighting
- **Cue selection is rational.** Stronger cues are selected more frequently than weaker cues

Conclusions

- **Untrained annotators are systematic** in their rating of the prosodic features of a word.
- **Individual annotators differ** in cue selection, and in cue weighting
- **Cue selection is rational.**
- **Attentional focus** influences cue selection and weighting (at group level).

Conclusions

- **Untrained annotators are systematic** in their rating of the prosodic features of a word.
- **Individual annotators differ** in cue selection, and in cue weighting
- **Cue selection is rational.**
- **Attentional focus** influences cue selection and weighting (at group level).

Future work:

What factors influence attentional focus in prosody perception?

Processing of linguistic message vs. social information?

Cognitive load?

Other?

THANKS!