The acoustics of [voice] in infant-directed speech and implications for phonological learning

Chandan Narayan
University of Toronto

Kyle Gorman
University of Pennsylvania

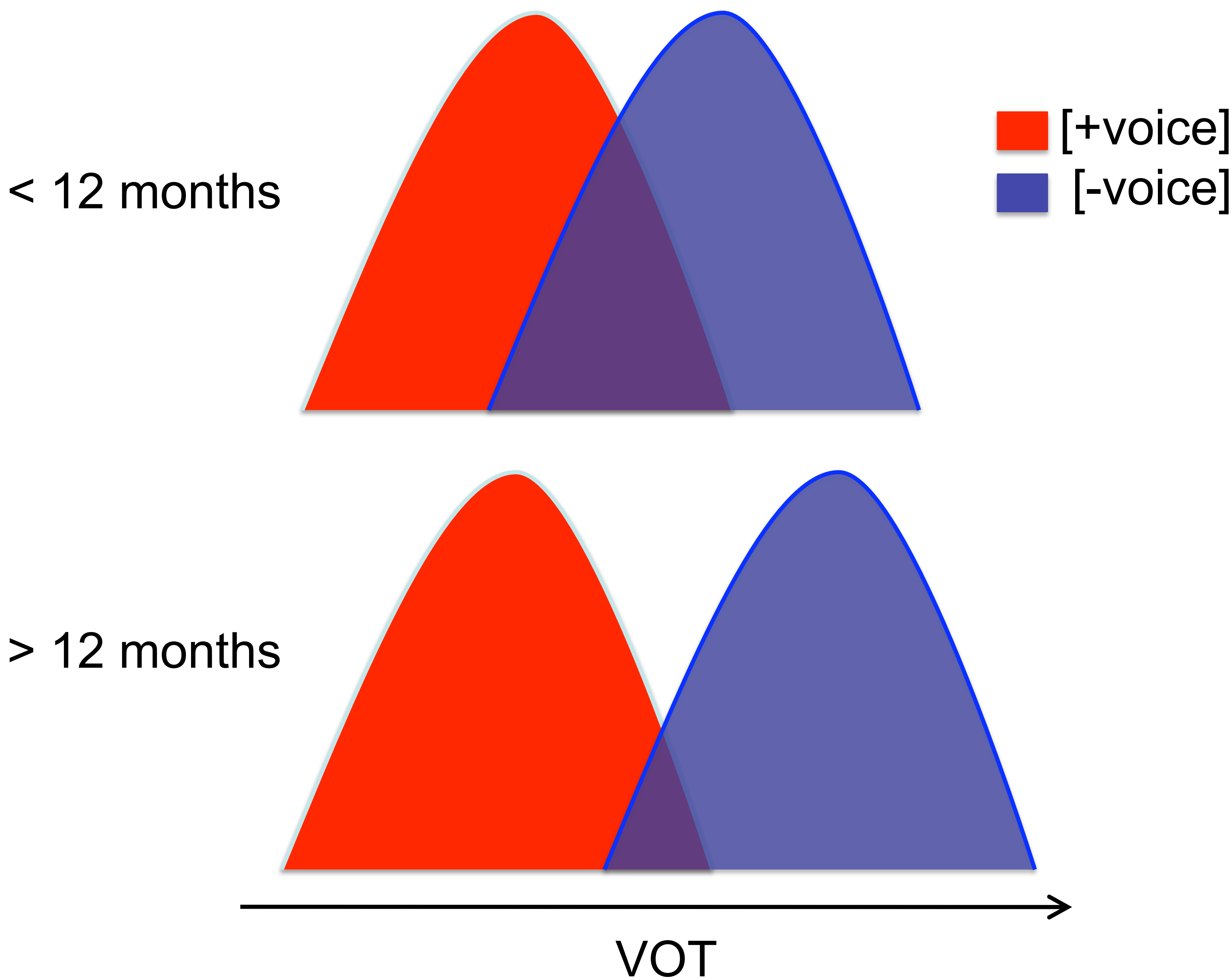
Daniel Swingley
University of Pennsylvania

BUCLD 33

Conflicting facts

- Infant-directed speech (IDS) provides the learner with *enhanced* cues to phonological contrast
  - Expanded *F*1 x *F*2 space (Kuhl et al., 1997)
  - Larger f_0 range in tones (Liu et al. 2007)
  - Modally distributed cues to vowel length and quality (Werker et al., 2007)

- In early infancy, caregivers' production of VOT shows more overlap between [+voice] and [-voice] than in late infancy (Sundberg & Lacerda, 1999; Baran et al., 1977)



< 12 months
> 12 months
[+voice]
[-voice]
VOT

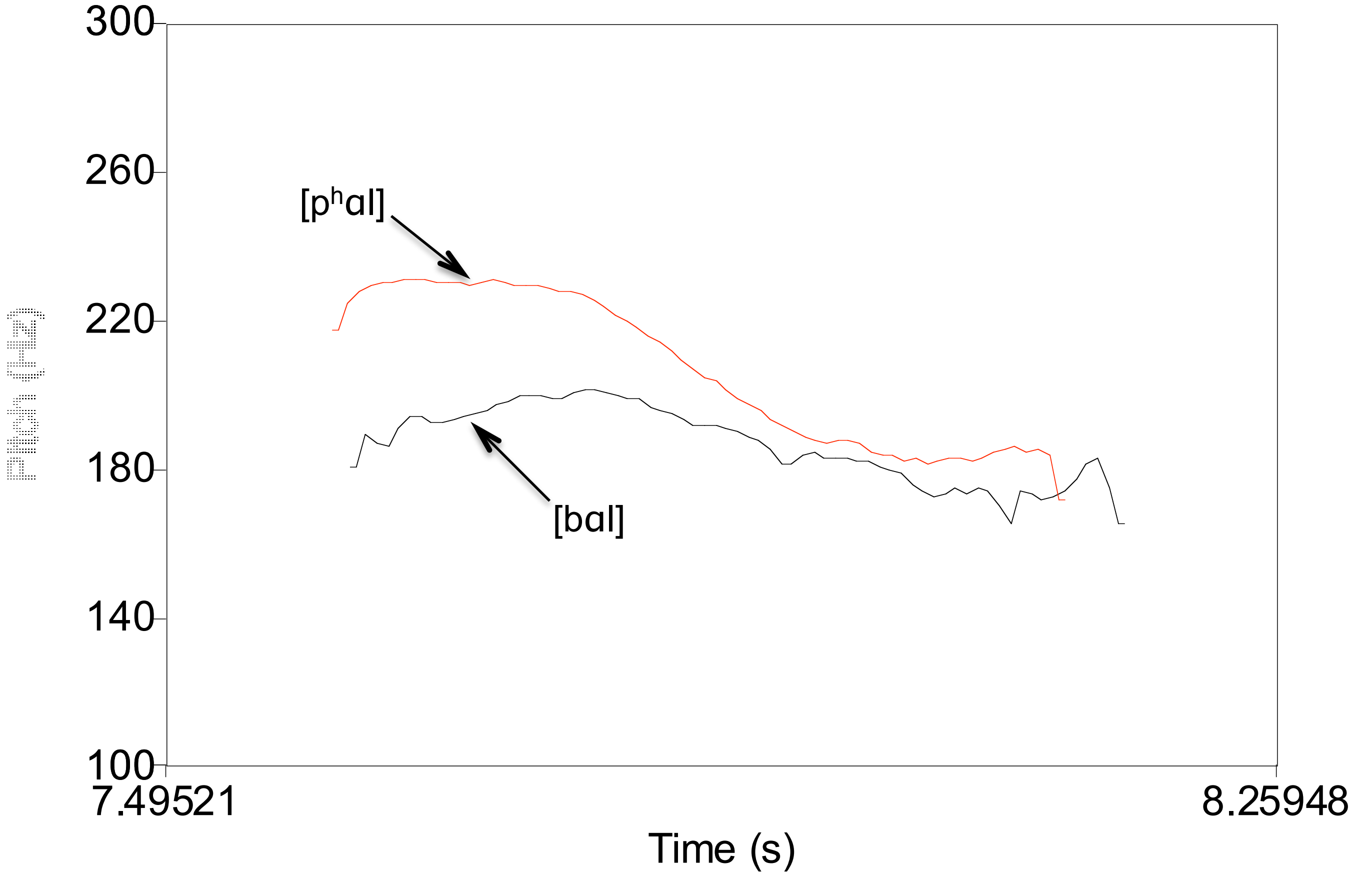
How do infants get [voice]? A challenge to enhancement

- Plenty of evidence that by 10-12 months, infants “know” the phonological categories (including [voice]) of their ambient language (Werker & Tees, 1984; numerous others)
- So, if IDS is providing infants with *sloppy* acoustic cues to phonological contrast, how might infants get to be so good at what they do?

Secondary cue to [voice]: f_0 perturbation

- ***Acoustics***
  - Following [+voice] stops, the fundamental frequency (f_0) of the vowel in a CV syllable is lower than when following [-voice] stops

- ***Perception***
  - When VOT is ambiguous, listeners report [+voice] when V has low f_0 and [-voice] when V has high f_0 (Lisker & Abramson, 1970; Abramson & Lisker, 1980; many others)



300
260
220
180
140
100
Pitch (Hz)
[pʰal]
[bal]
7.49521
8.25948
Time (s)

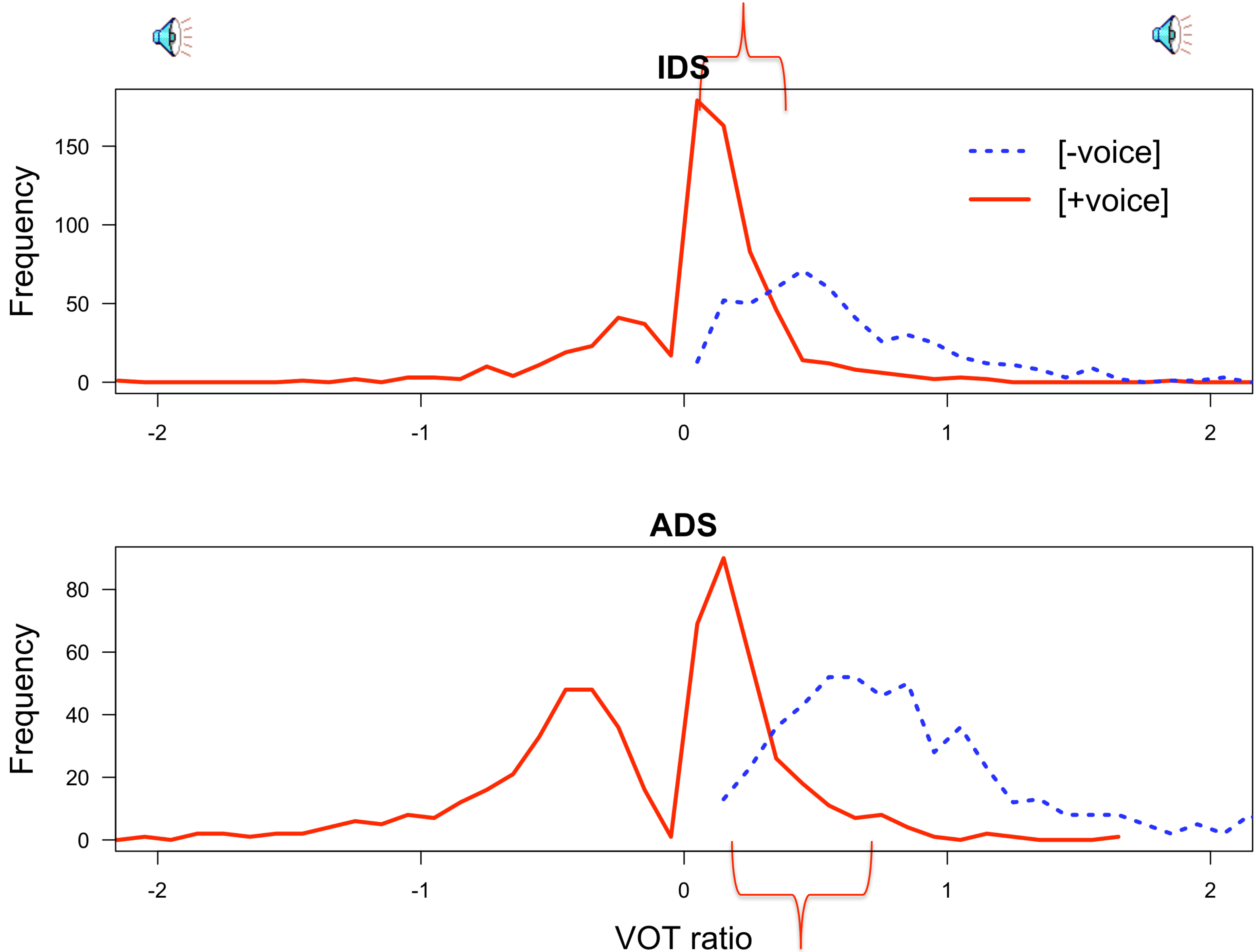
f_0 control

- Kingston & Diehl (1994; Francis et al., 2007) suggest that listeners control f_0, giving listeners extra low-frequency information in the vicinity of the stop closure for [+voice] perception

- f_0 is an “enhance-able” acoustic feature that mothers might exploit when conveying [voice] information to their children

Corpora

- ***Brent Corpus* (IDS)** (Brent and Siskind, 2001)
  - 4 mothers using infant-directed English speech
  - Natural mother-infant interactions at 9 months of age
  - 500 utterances/mother → over 1200 word-initial CVs
- ***Buckeye Corpus* (ADS)** (Pitt et al., 2007)
  - 4 women (3 with young infants, 1 with an older toddler) speaking a Midlands dialect with an adult
  - 20 minutes of speech/speaker → over 1000 CVs



IDS
Frequency
150
100
50
0
-2
-1
0
1
2
[-voice]
[+voice]
ADS
Frequency
80
60
40
20
0
-2
-1
0
1
2
VOT ratio

Results: VOT

- Register (IDS vs. ADS) x Voicing ([+voice] vs. [-voice]) x Place (bilabial, alveolar, velar) ANOVA

- *Voicing x Register interaction*
  – Suggests that the difference in VOT between [+voice] and [-voice] is **greater in ADS** than in IDS

Results: VOT

- VOT in each register was used in a linear discriminant analysis to classify [voice]

IDS	Predicted (%) [+voice]	[-voice]
[+voice]	72.7 ($n = 365$)	27.3 ($n = 137$)
[-voice]	11.2 ($n = 78$)	88.8 ($n = 620$)

ADS	[+voice]	[-voice]
[+voice]	88.3 ($n = 432$)	11.7 ($n = 57$)
[-voice]	10.9 ($n = 62$)	89.1 ($n = 507$)

Results: f_0 perturbation

- Peak f_0 (*z*-Mel) following release of stop
- Voicing x Register ANOVA
- Significant effect of [voice] on f_0 but no interactions
  - Mean f_0 is higher following [-voice] stops
    [+voice] = 0.16 vs. [-voice] = 0.38

Modeling [voice]

- Predict presence of [voice] given VOT and f_0
- Hierarchical logistic regression
  - Allow speakers to vary in implementations of VOT and f_0 cues to [voice]
  - Random (per-subject) slopes for VOT and f_0
  - We'll attempt to interpret subject effects to show the relative degree to which subjects implement the two cues

Hierarchical model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.4272	0.1166	20.82	<2e-16
VOT	-6.3426	0.4264	-14.87	<2e-16
f_0	-0.0872	0.1352	-0.64	0.519
VOT:f_0	-1.1257	0.5419	-2.08	0.038

- No main effect of f_0 but it inversely covaries with VOT, as expected

- The effect size for f_0 is equivalent to about a semitone in mean pitch region

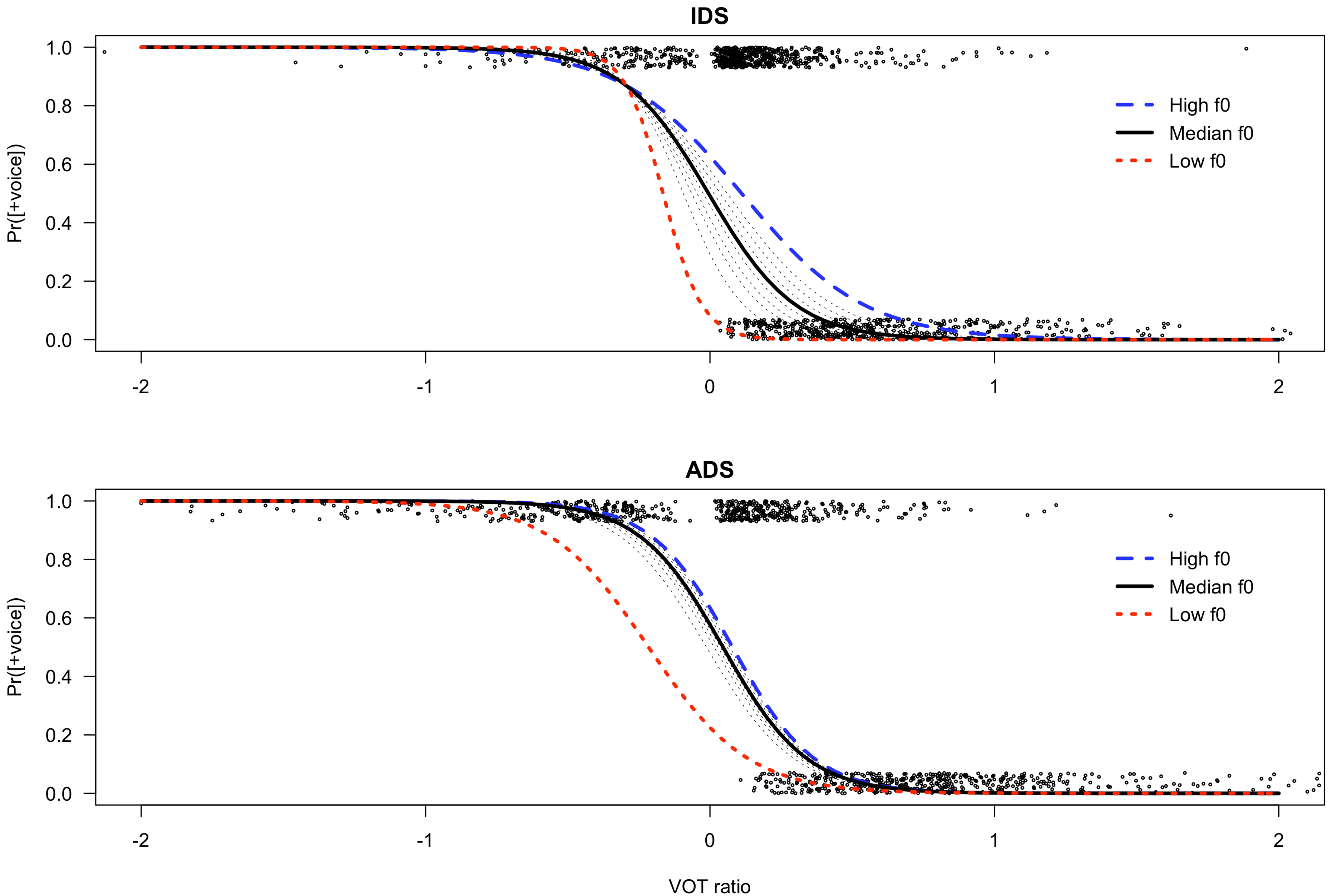
Random effects

	VOT	f_0	VOT:f_0	
1	-0.3072	0.193	-1.10	
2	0.3237	0.260	-1.47	
3	-0.4676	0.073	-0.38	
4	-0.1164	-0.218	1.20	IDS
5	-0.0031	0.085	-0.49	ADS
6	-1.3383	-0.232	1.31	
7	0.9212	-0.119	0.68	
8	1.3105	-0.076	0.43	

Register-specific (pooled) models of [voice]

IDS	β	Std. Error	95% CI	z	Sig.
(Intercept)	0.01	0.08	(-0.16, 0.17)	0.11	0.91
VOT ratio	-6.39	0.40	(-7.21, -5.64)	-15.91	<0.0001
Peak f_0	-0.59	0.10	(-0.80, -0.40)	-5.82	<0.0001
VOT ratio x Peak f_0	-2.03	0.44	(-2.91, -1.17)	-4.59	<0.0001

ADS					
(Intercept)	0.35	0.11	(0.15, 0.56)	3.32	<0.001
VOT ratio	-6.82	0.47	(-7.79, -5.95)	-14.58	<0.0001
Peak f_0	-0.36	0.13	(-0.55, -0.09)	-2.62	<0.01
VOT ratio x Peak f_0	0.30	0.60	(-0.92, 1.41)	0.41	0.61


IDS
ADS
High f0
Median f0
Low f0
Pr([+voice])
VOT ratio
1.0
0.8
0.6
0.4
0.2
0.0
-2
-1
0
1
2

Predicting [voice]: Discussion

- Given VOT and f_0:
  - VOT contributes to a [voice] prediction in *both* IDS and ADS
  - f_0 contributes more to a [voice] prediction in IDS than in ADS!

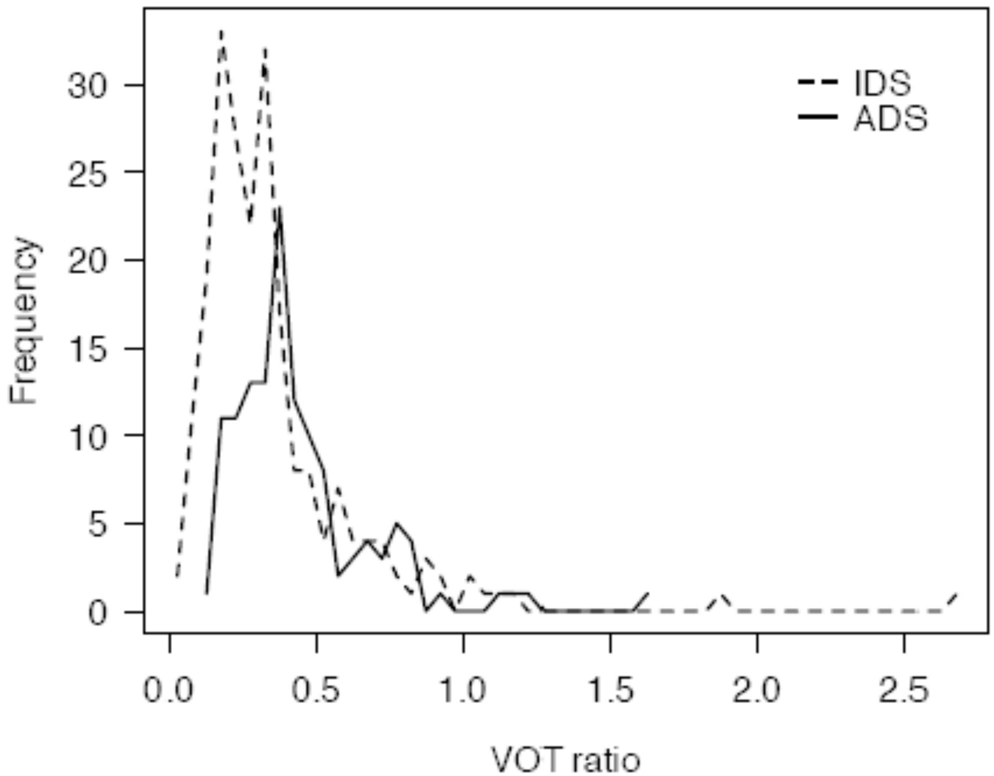
- There is enough consistent VOT information in ADS which essentially overrides the f_0 regularity

- <u>When VOT is highly variable (with more overlap between categories) as in the case of IDS, f_0 information is useful in predicting [voice]</u>

Misclassification analysis

- What is the practical import of the logistic models?
- The LR models were used as a [voice] classifier with *only* VOT as a predictor

Misclassifications using only VOT



- Twice as many misclassifications in IDS than in ADS
- Majority of misclassifications occur in the 0-0.5 range, which corresponds to the region of overlap in VOT

When misclassified tokens were ***re-classified*** **using *f*0**,

69% of IDS tokens were correctly classified ($p < 0.001$)

52% of ADS tokens were correctly classified ($p = 0.79$)

Take home message

- f_0 in the IDS sample preserves [voice] information when VOT information alone is ambiguous
- The emergence of f_0 as a stable contributor to [voice] prediction suggest a *covert contrast* that the learner might recover

Take this home too!

- So far, distributional models of phonetic category learning, that rely on *enhancement* as a hallmark of learning, are dimensionally *flat* when it comes to multiple cues to phonologically relevant features such as [voice]

A comprehensive theory of phonetic category learning (in infancy) must consider:

- The changing nature of the IDS across development: [voice] as presented to 9 month olds is different from [voice] to 14 month olds
- Multiple cues to relevant features: such as covarying VOT and f_0 for [voice], $F2$ onset and burst frequency for place, etc.

Thank you

- **Chandan Narayan**
  chandan.narayan@utoronto.ca
- **Kyle Gorman**
  kgorman@ling.upenn.edu
- **Daniel Swingley**
  swingely@psych.upenn.edu

Funded by IGERT(NSF) to CN & KG and the NIH to DS.
Thanks to Penn Baby Lab.