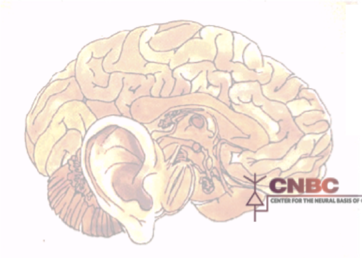
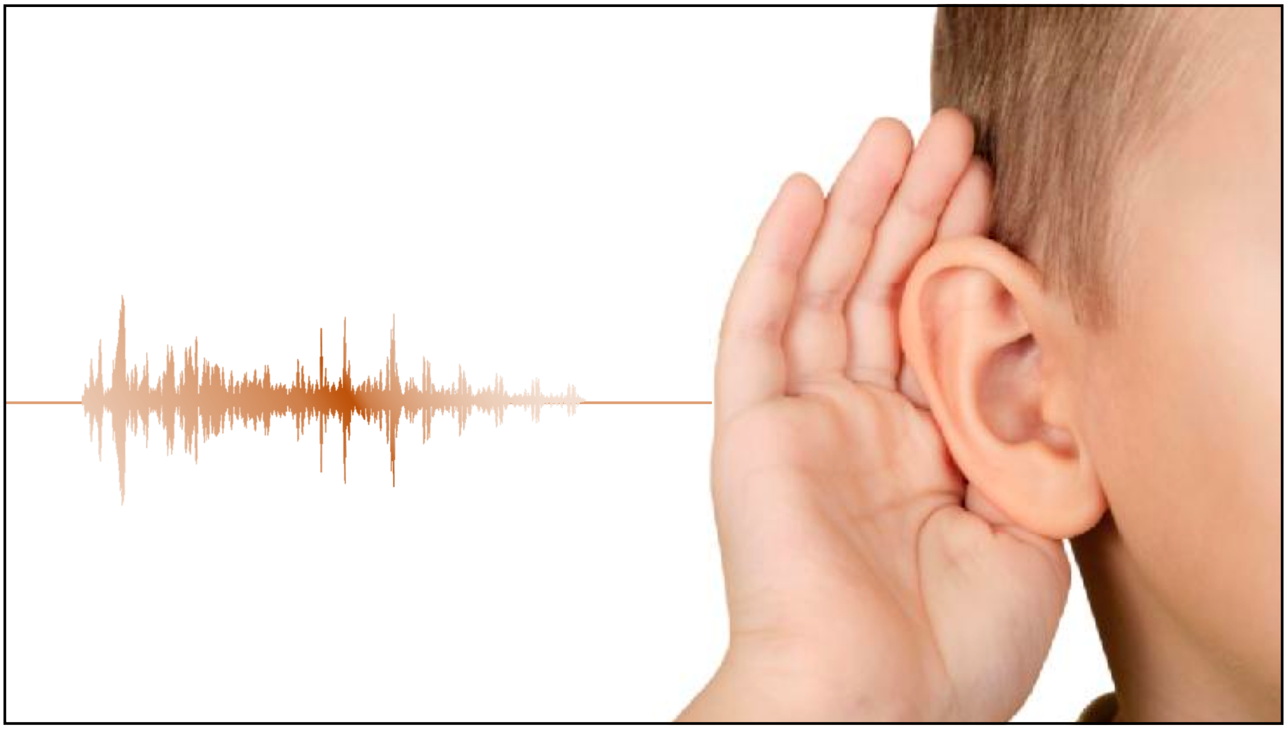
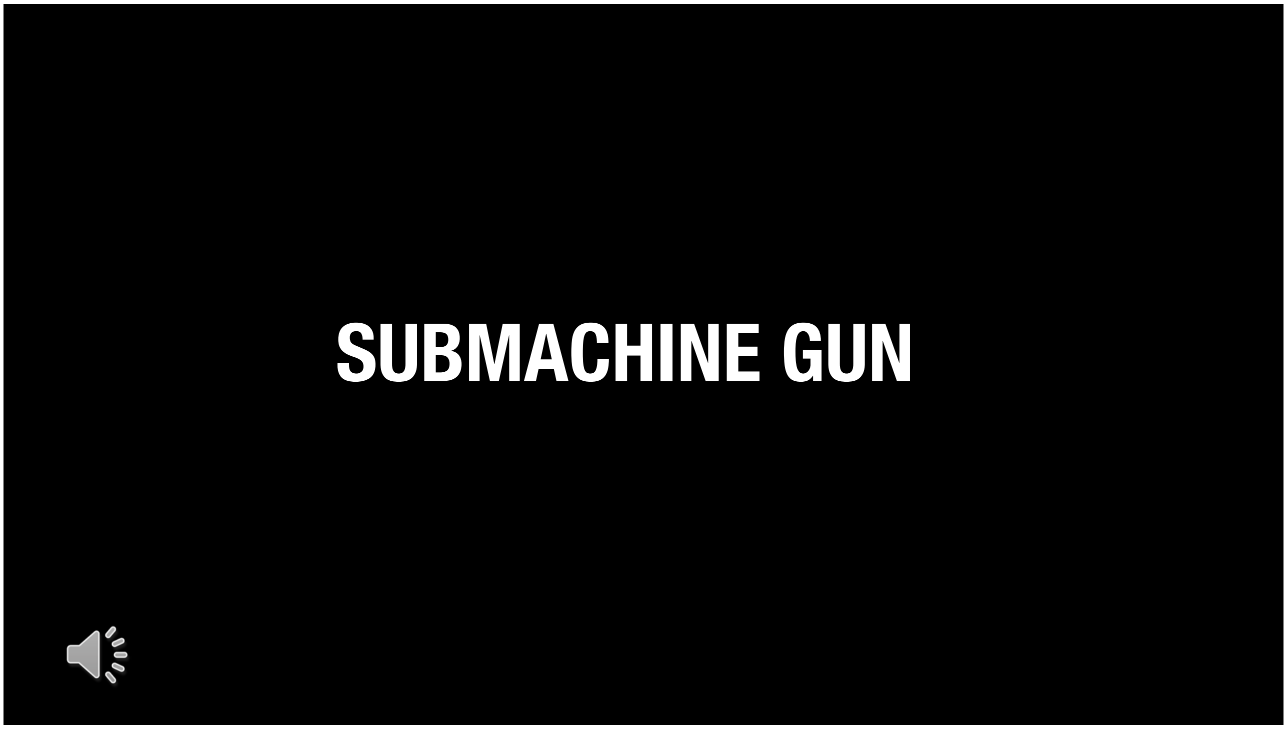
Understanding how humans interpret the complexity of spoken language

Part I: Cracking the Speech Code with Learning



Lori L. Holt
Professor, Department of Psychology
Carnegie Mellon University





SUBMACHINE GUN

Learning

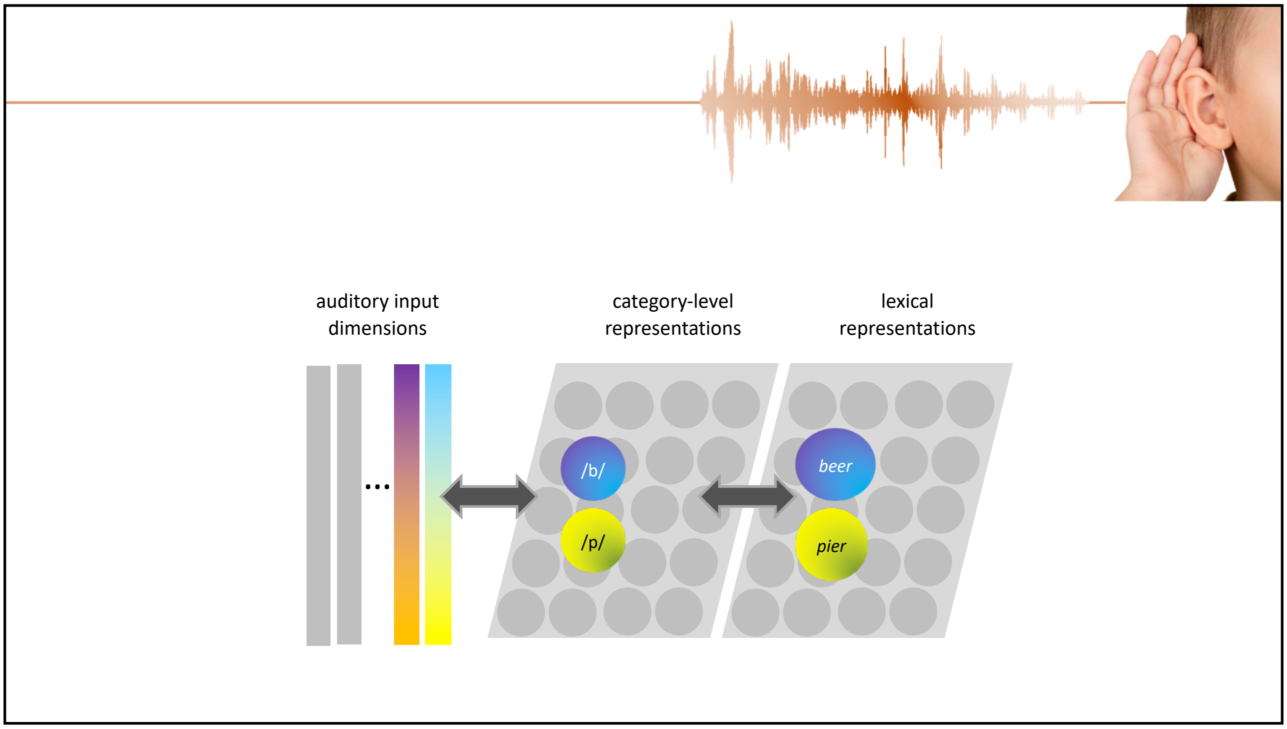
across speech signals

Part I

Learning Across Longer-term to **Develop New** Representations

Part II

Learning Across Shorter-term to **Adapt Existing** Representations



auditory input dimensions

category-level representations

lexical representations

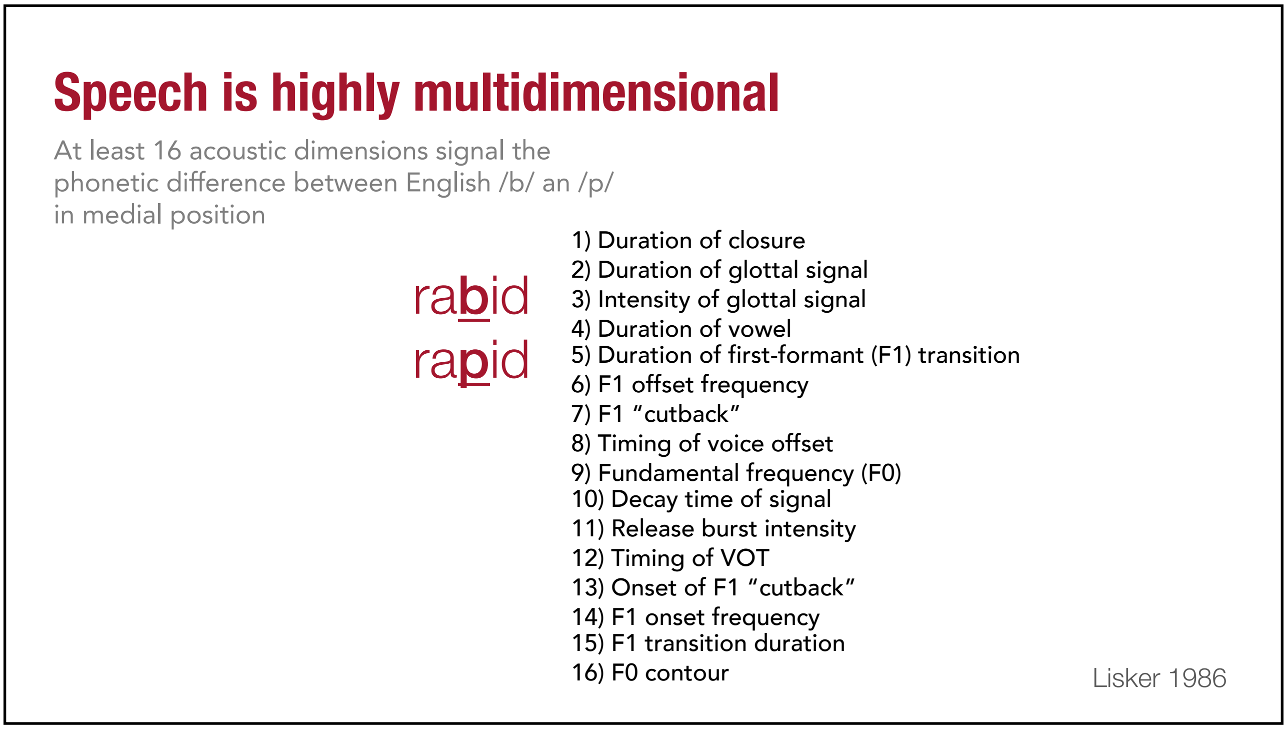
…

/b/

/p/

beer

pier



Speech is highly multidimensional

At least 16 acoustic dimensions signal the phonetic difference between English /b/ an /p/ in medial position

rabid

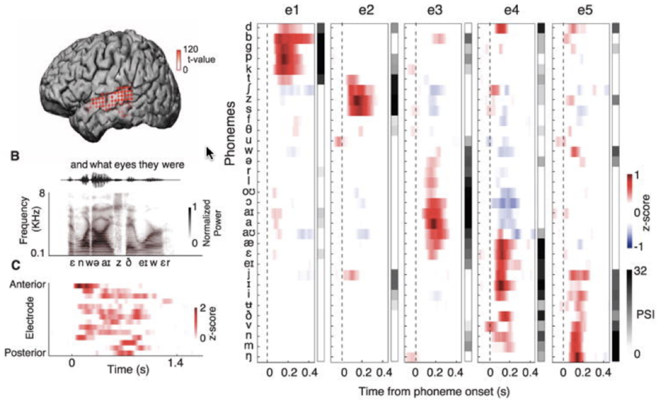
rapid

1) Duration of closure
2) Duration of glottal signal
3) Intensity of glottal signal
4) Duration of vowel
5) Duration of first-formant (F1) transition
6) F1 offset frequency
7) F1 "cutback"
8) Timing of voice offset
9) Fundamental frequency (F0)
10) Decay time of signal
11) Release burst intensity
12) Timing of VOT
13) Onset of F1 "cutback"
14) F1 onset frequency
15) F1 transition duration
16) F0 contour

Lisker 1986

Human Superior Temporal Gyrus Selectivity to Speech

Mesgarani, Cheung & Chang, 2014



Well-known acoustic features of phonemes are explicitly encoded in population responses

Speech is highly multidimensional

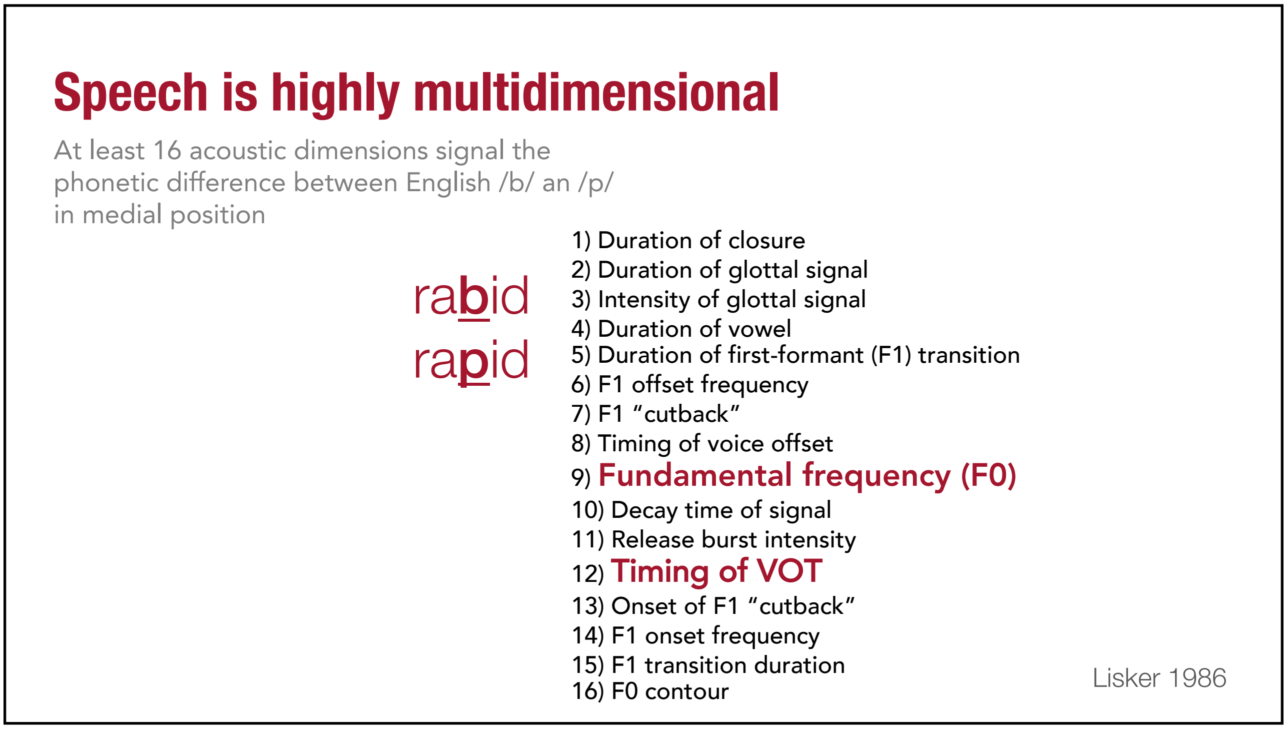
At least 16 acoustic dimensions signal the phonetic difference between English /b/ an /p/ in medial position

ra**b**id

ra**p**id

1) Duration of closure
2) Duration of glottal signal
3) Intensity of glottal signal
4) Duration of vowel
5) Duration of first-formant (F1) transition
6) F1 offset frequency
7) F1 "cutback"
8) Timing of voice offset
9) Fundamental frequency (F0)
10) Decay time of signal
11) Release burst intensity
12) Timing of VOT
13) Onset of F1 "cutback"
14) F1 onset frequency
15) F1 transition duration
16) F0 contour

Lisker 1986



Speech is highly multidimensional

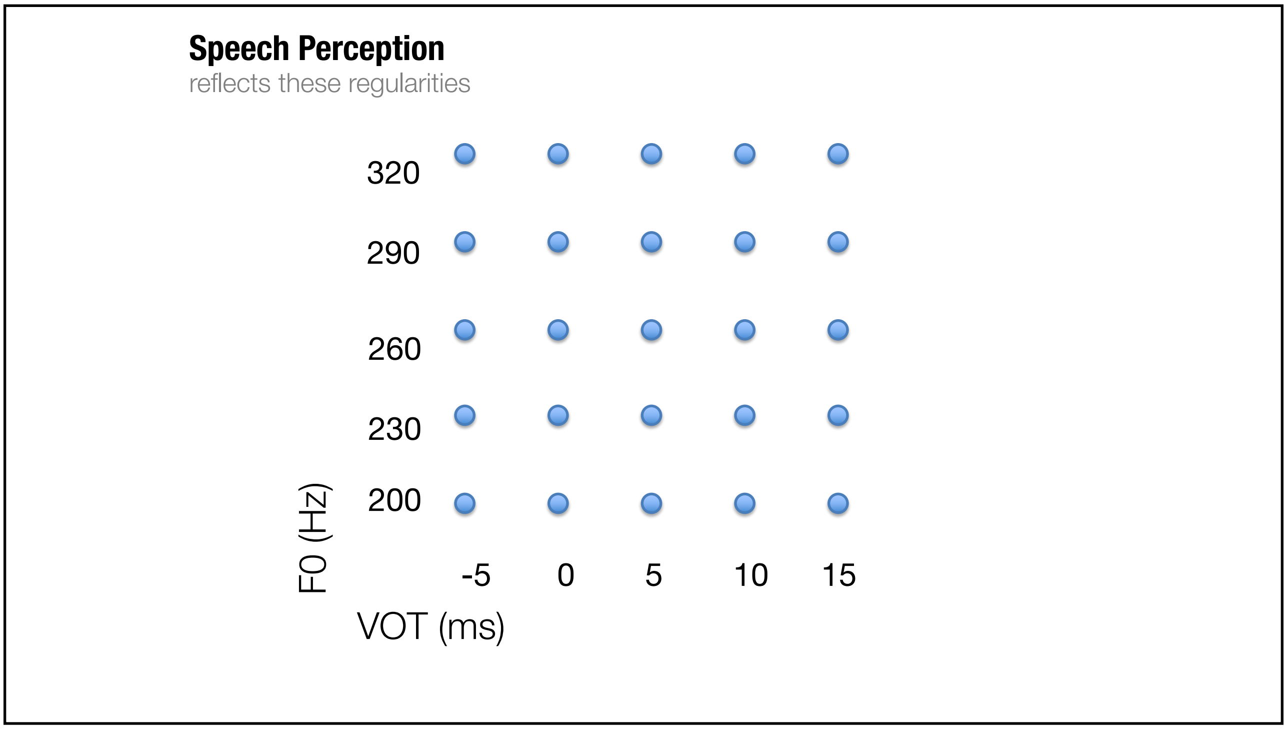
At least 16 acoustic dimensions signal the phonetic difference between English /b/ an /p/ in medial position

ra**b**id

ra**p**id

1) Duration of closure
2) Duration of glottal signal
3) Intensity of glottal signal
4) Duration of vowel
5) Duration of first-formant (F1) transition
6) F1 offset frequency
7) F1 “cutback”
8) Timing of voice offset
9) **Fundamental frequency (F0)**
10) Decay time of signal
11) Release burst intensity
12) **Timing of VOT**
13) Onset of F1 “cutback”
14) F1 onset frequency
15) F1 transition duration
16) F0 contour

Lisker 1986

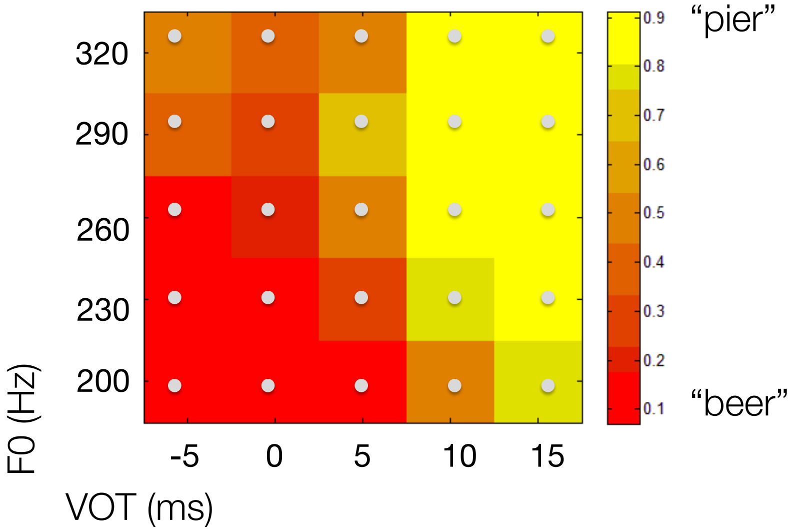


Speech Perception

reflects these regularities

Speech Perception

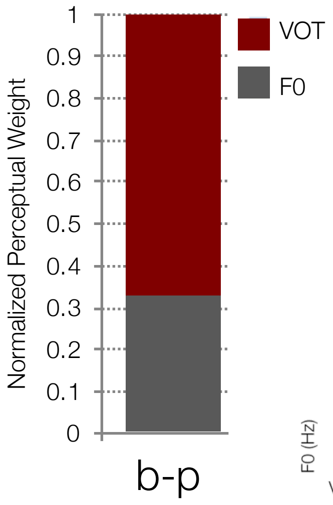
reflects these regularities



Both F0 and VOT influence perception

Perceptual Cue Weighting

dimensions are not equivalent
some better signal phonetic category membership

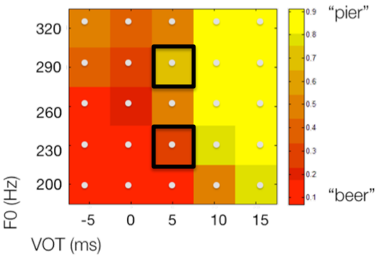


VOT is more strongly predictive of beer-pier responses than F0

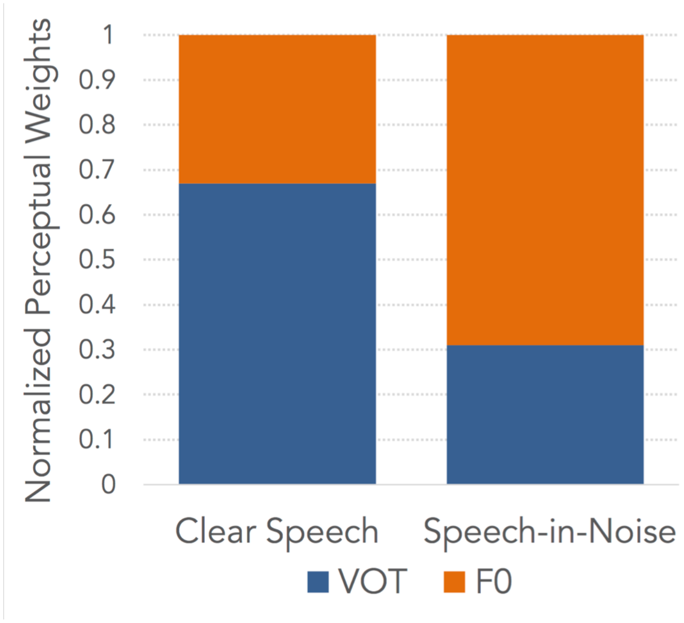
greater perceptual weight

F0 does contribute

especially when VOT is ambiguous



e.g., Idemaru & Holt 2011
Schultz et al., 2012
Francis et al., 2008
see Holt & Lotto 2006



In Clear Speech →
Rely more on VOT

In Noise →
Rely more on F0

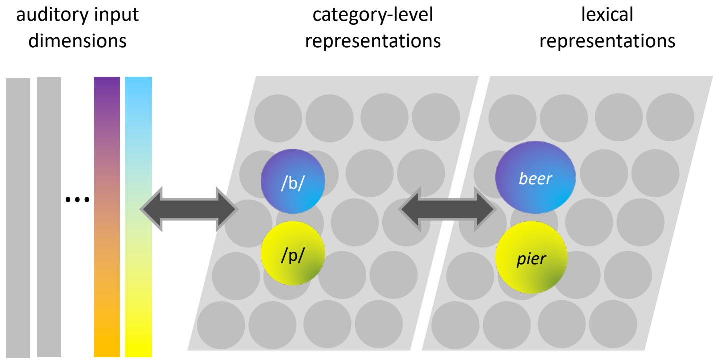
Holt, Tierney, Guerra, Laffere, Dick, 2018

Part I

Learning Across Longer-term to **Develop New** Representations

Part II

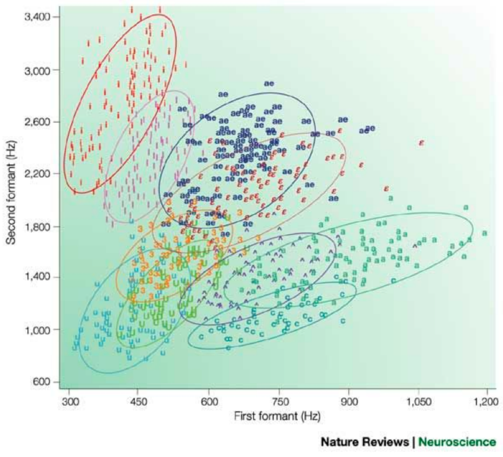
Learning Across Shorter-term to **Adapt Existing** Representations



auditory input dimensions

category-level representations

lexical representations



Nature Reviews | Neuroscience

Acoustics of speech input are notoriously variable

But there is distributional regularity, too

from Kuhl, 2004

Category Learning…

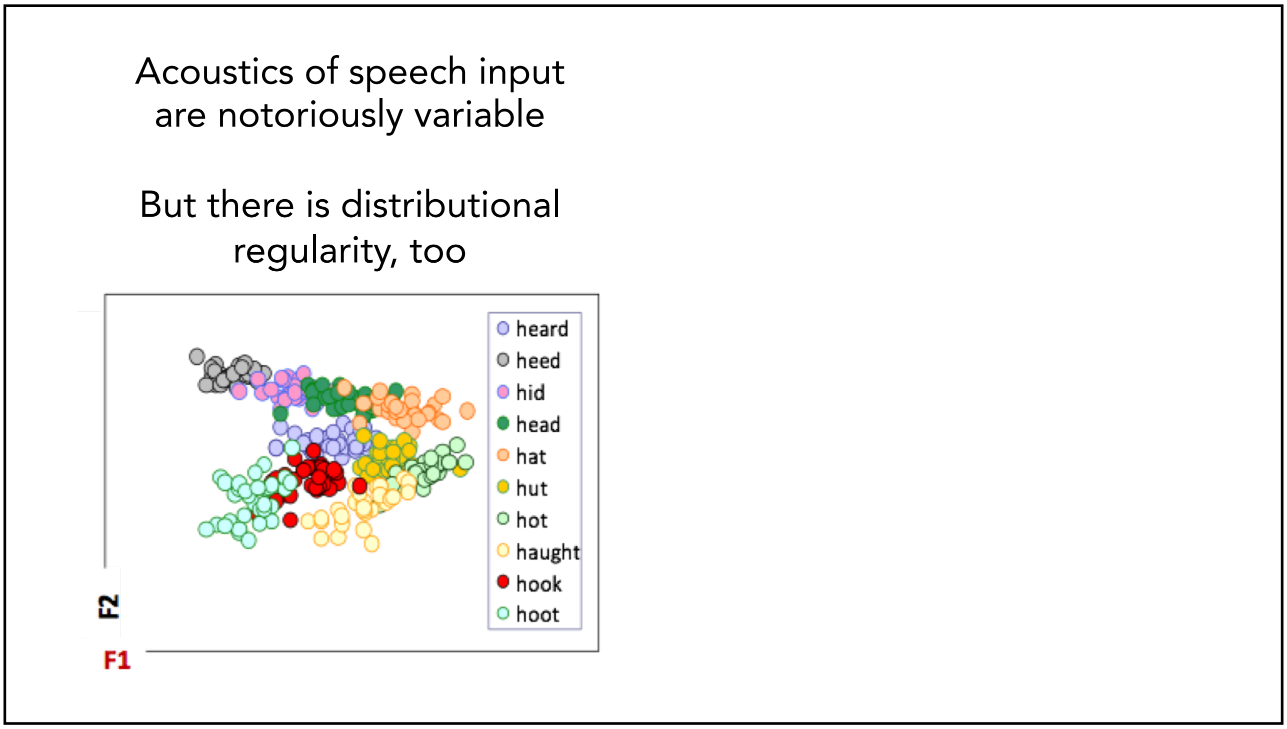
involves learning to treat physically-distinct experiences as functionally equivalent

supports **generalization** of knowledge to new, unfamiliar, experiences that share statistical structure with the category

Categorization

distinct experiences
as functionally-equivalent

Categorization allows us to generalize to understand new experiences



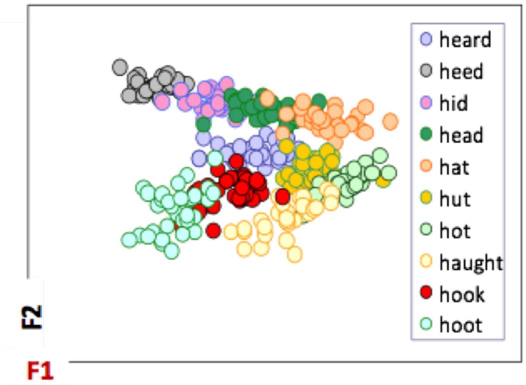
Acoustics of speech input are notoriously variable

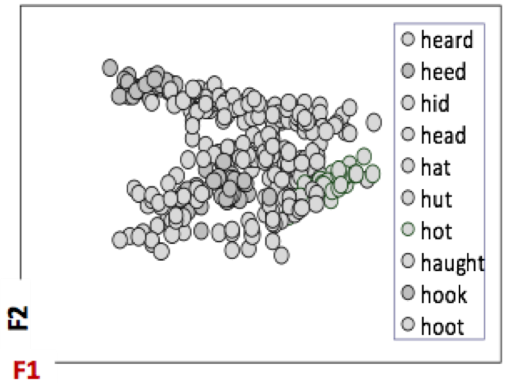
But there is distributional regularity, too

Acoustics of speech input are notoriously variable

But there is distributional regularity, too

But the learner does not have access to labeled instances

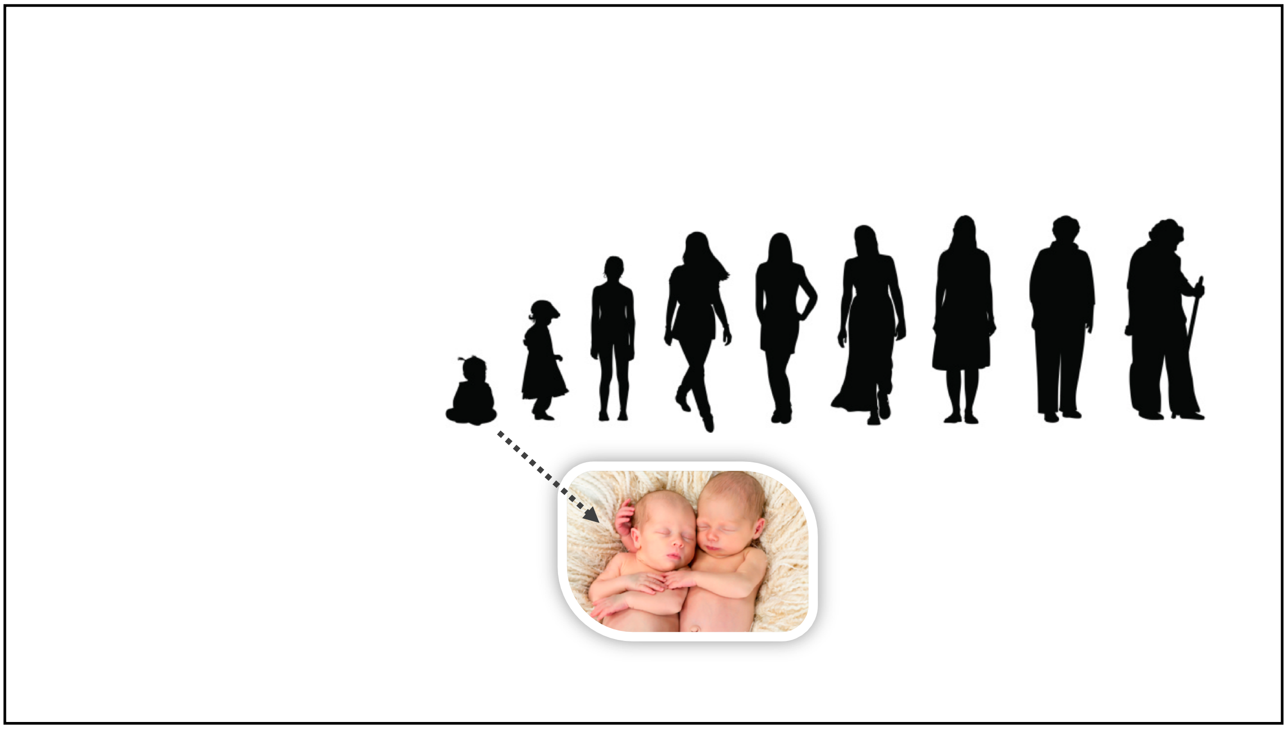


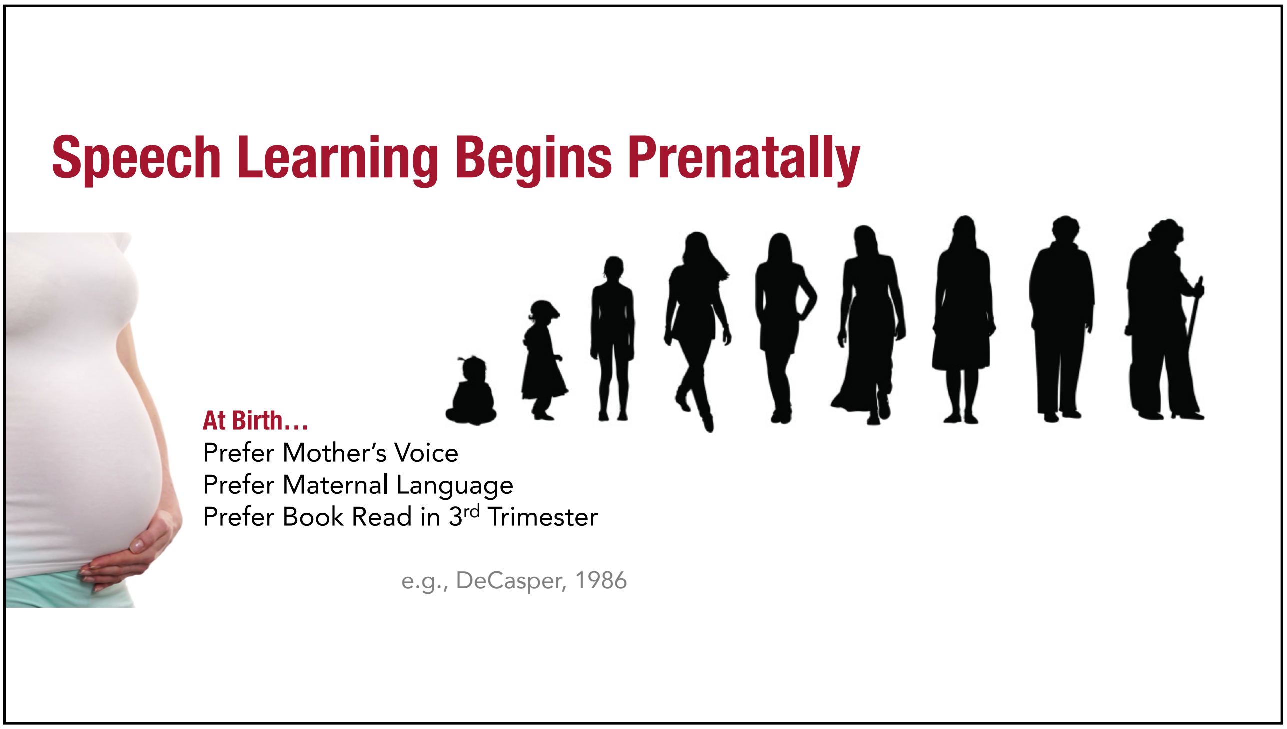


Part I

Learning Across Longer-term to **Develop New** Representations

How do we learn to map complex, multidimensional distributions of sounds to form categories?





Speech Learning Begins Prenatally

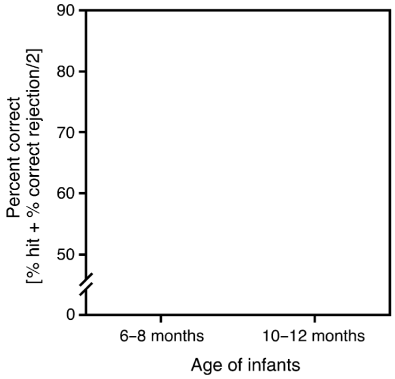
At Birth…
Prefer Mother's Voice
Prefer Maternal Language
Prefer Book Read in 3rd Trimester

e.g., DeCasper, 1986

Speech Learning Begins Prenatally

Learning Continues in Infancy

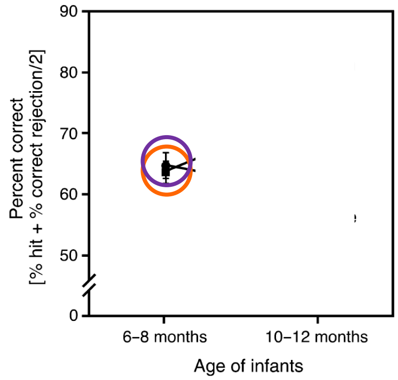
Infants' Discrimination of English /ra/ vs. /la/



Kuhl et al., 2006

Learning Continues in Infancy

Infants' Discrimination
of English /ra/ vs. /la/

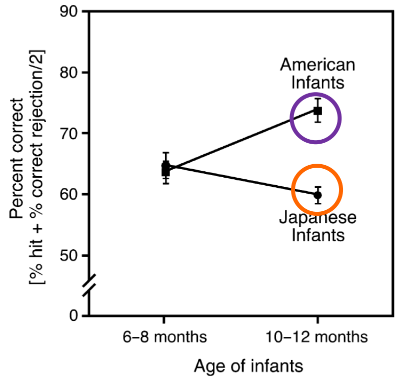


Kuhl et al., 2006

Early Infancy:
perception based
on acoustic differences

Learning Continues in Infancy

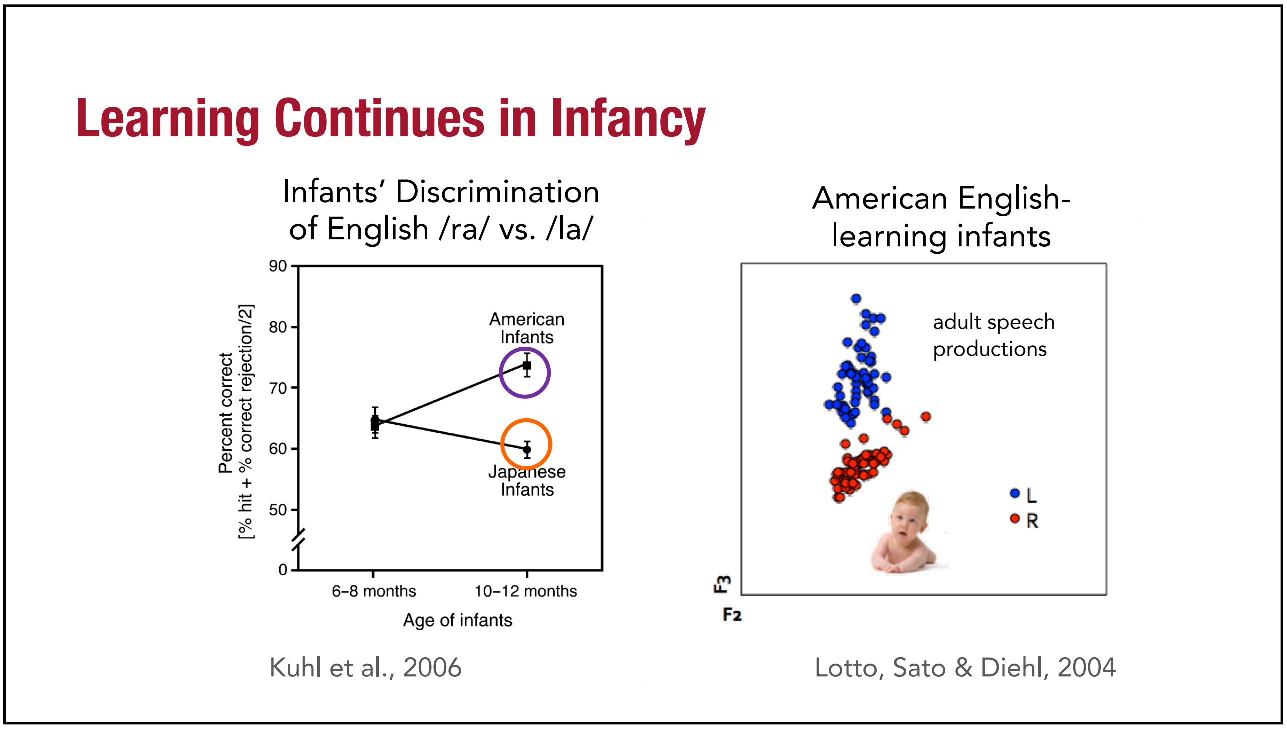
Infants' Discrimination
of English /ra/ vs. /la/



Kuhl et al., 2006

Early Infancy:
perception based
on acoustic differences

Later in Year 1:
native-language experience
affects perception



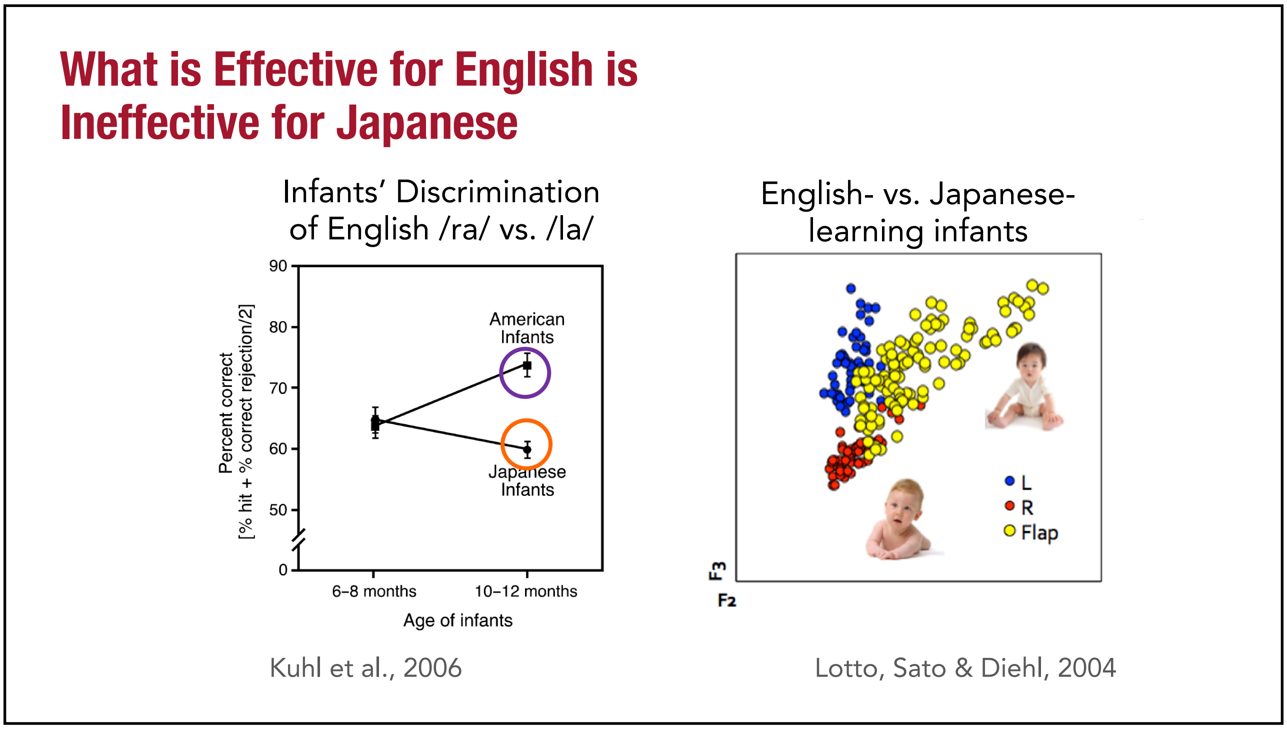
Learning Continues in Infancy

Infants' Discrimination of English /ra/ vs. /la/

Kuhl et al., 2006

American English-learning infants

Lotto, Sato & Diehl, 2004



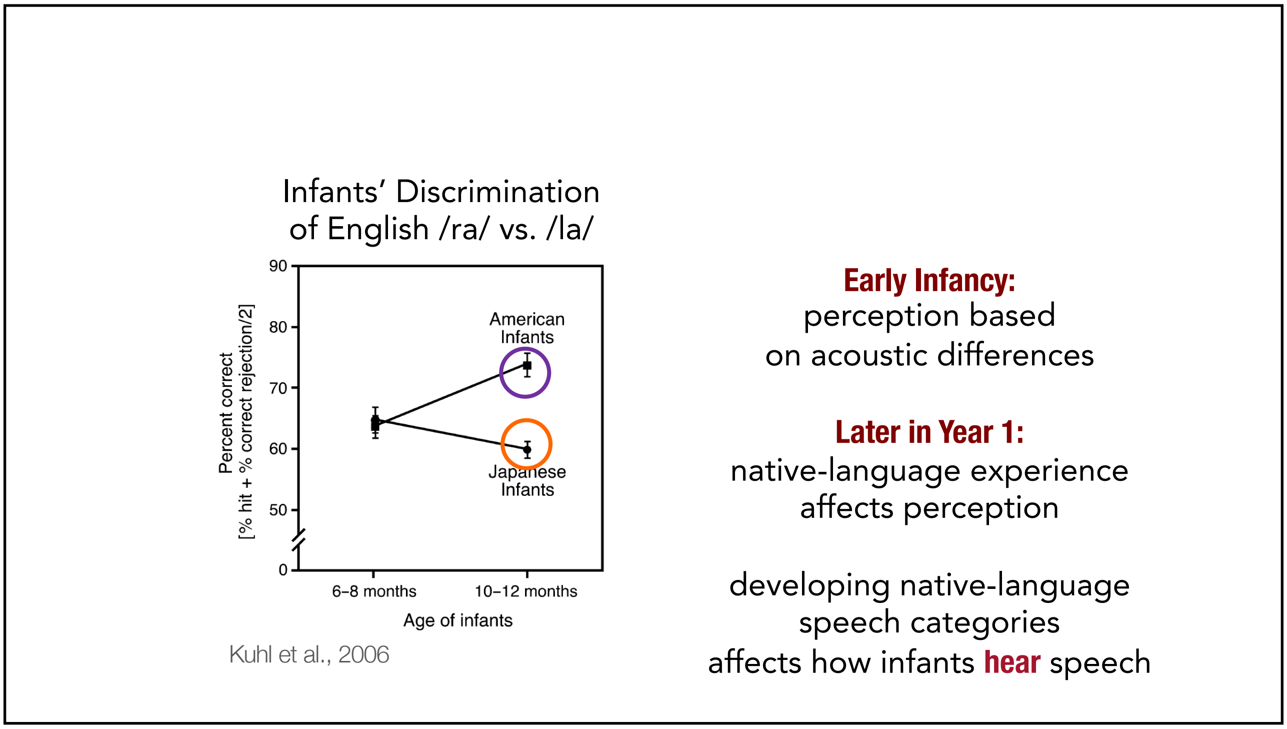
What is Effective for English is Ineffective for Japanese

Infants' Discrimination of English /ra/ vs. /la/

Kuhl et al., 2006

English- vs. Japanese-learning infants

Lotto, Sato & Diehl, 2004



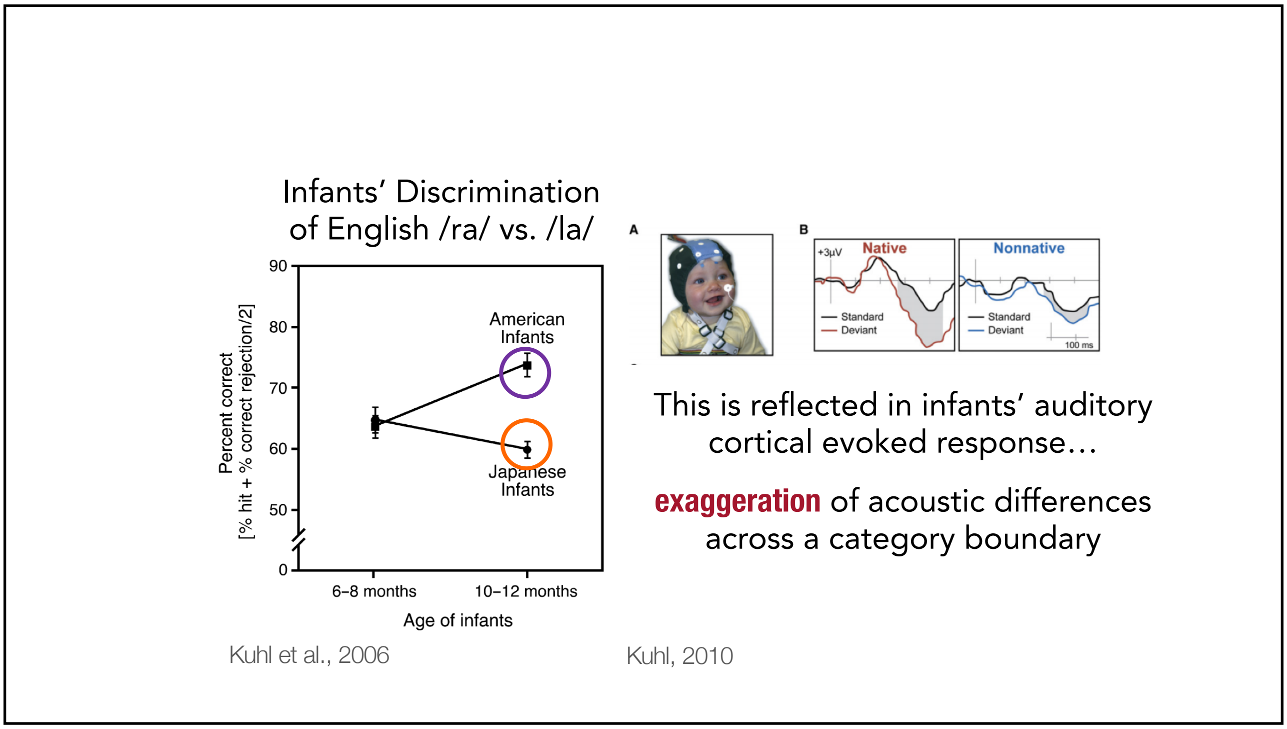
Infants' Discrimination of English /ra/ vs. /la/

Kuhl et al., 2006

Early Infancy:
perception based
on acoustic differences

Later in Year 1:
native-language experience
affects perception

developing native-language
speech categories
affects how infants **hear** speech



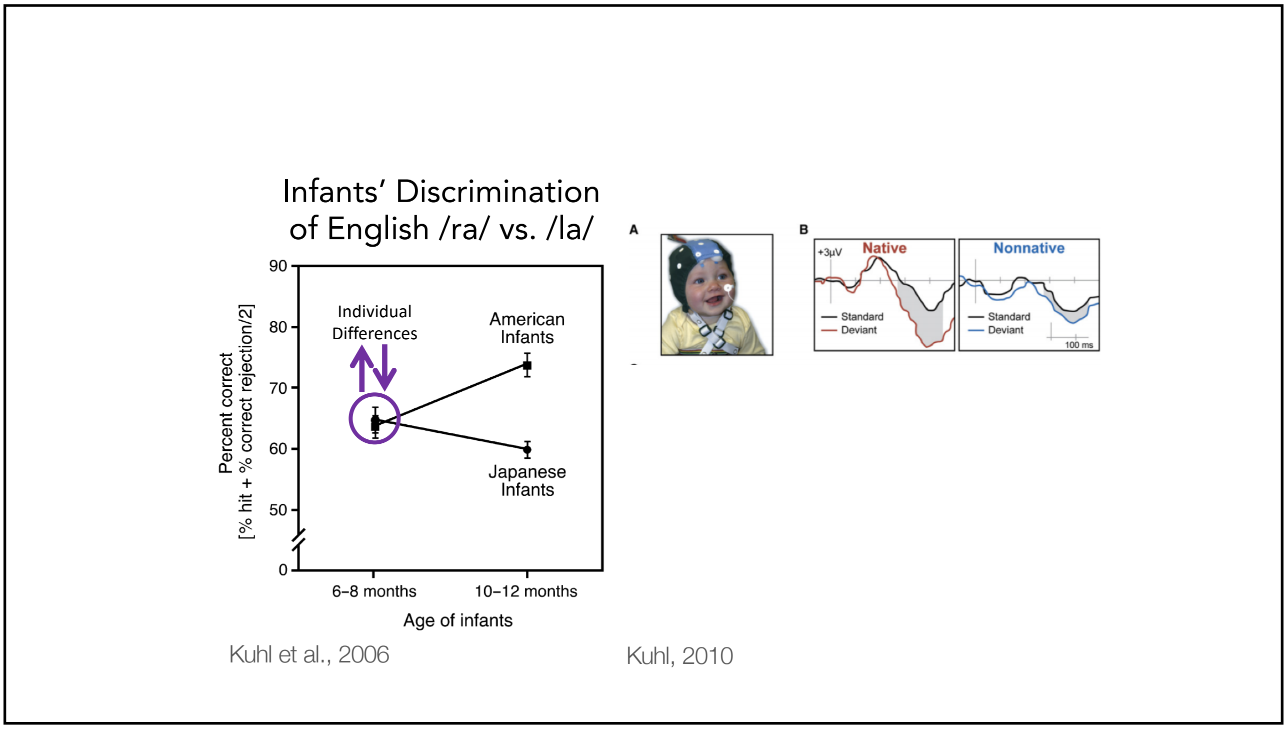
Infants' Discrimination of English /ra/ vs. /la/

Kuhl et al., 2006

This is reflected in infants' auditory
cortical evoked response…

exaggeration of acoustic differences
across a category boundary

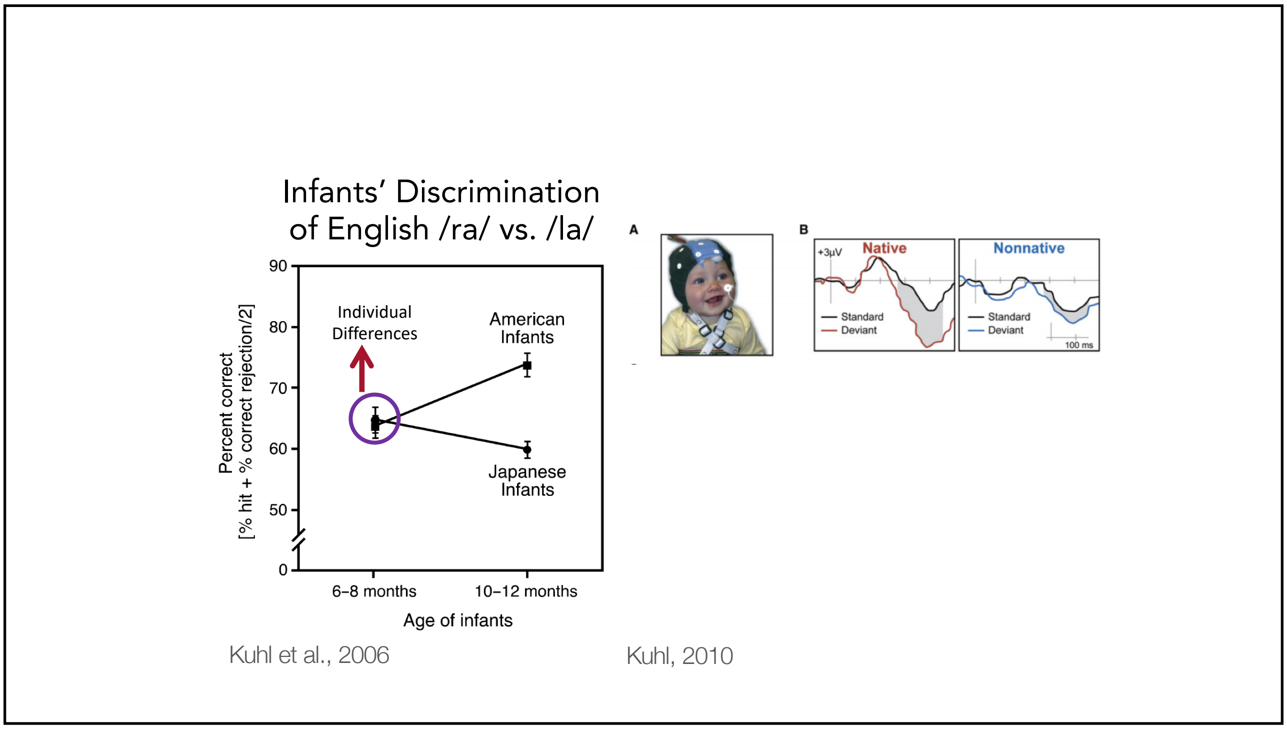
Kuhl, 2010



Infants’ Discrimination of English /ra/ vs. /la/

Kuhl et al., 2006

Kuhl, 2010

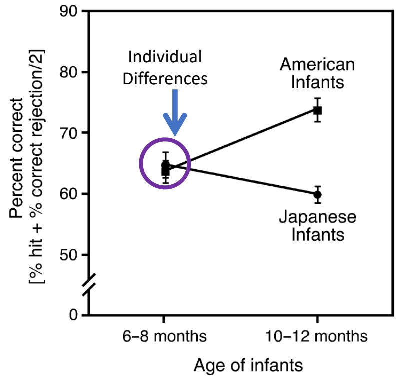


Infants’ Discrimination of English /ra/ vs. /la/

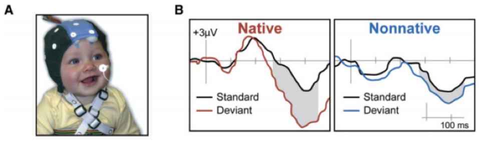
Kuhl et al., 2006

Kuhl, 2010

Infants' Discrimination of English /ra/ vs. /la/



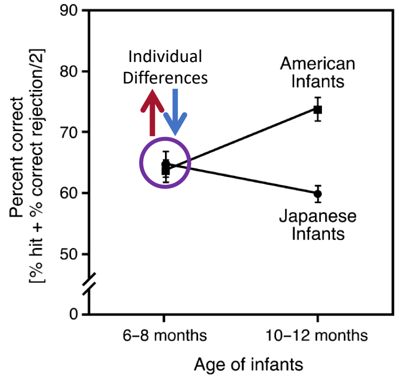
Kuhl et al., 2006



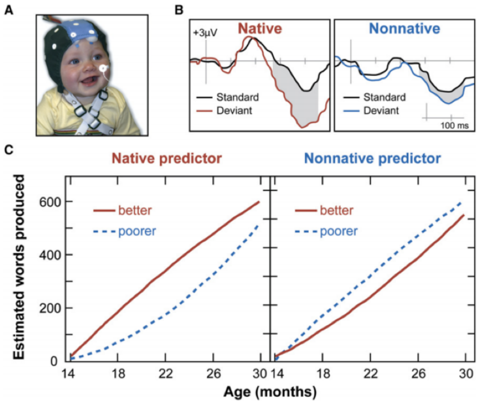
Kuhl, 2010

Better native-language categories at 6-8 months predicts vocabulary at 30 months

Infants' Discrimination of English /ra/ vs. /la/



Kuhl et al., 2006

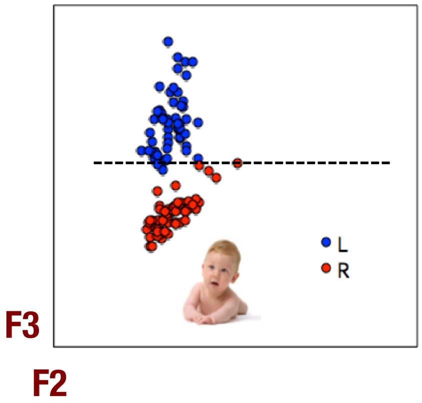


Kuhl, 2010

Classic textbook understanding…
speech category learning is largely complete in infancy

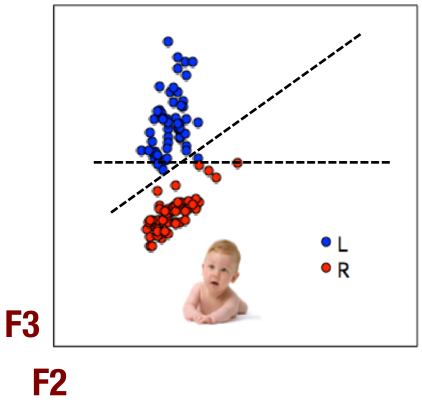
Classic textbook understanding…
speech category learning is largely complete in infancy

But…



F3 is single best predictor of English /r/-/l/ category membership

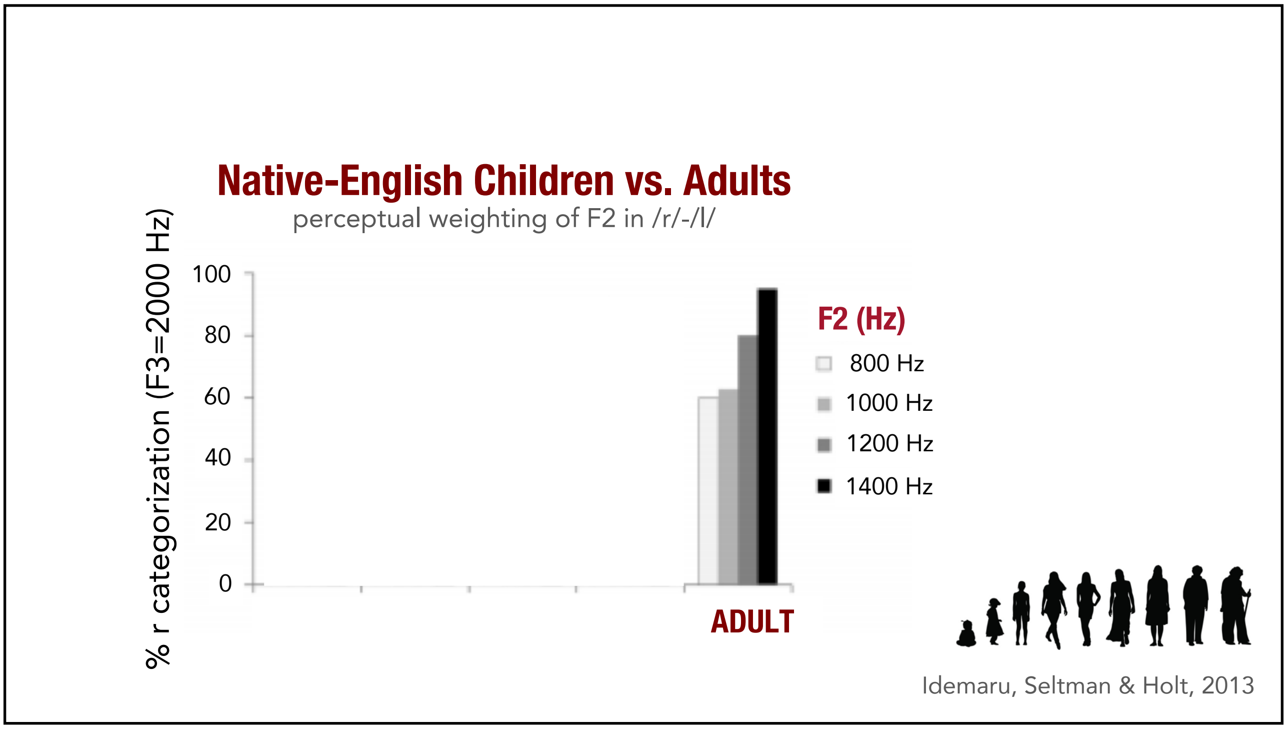
Idemaru, Seltman & Holt, 2013



F3 is single best predictor of English /r/-/l/ category membership

…but F2 is an important secondary cue

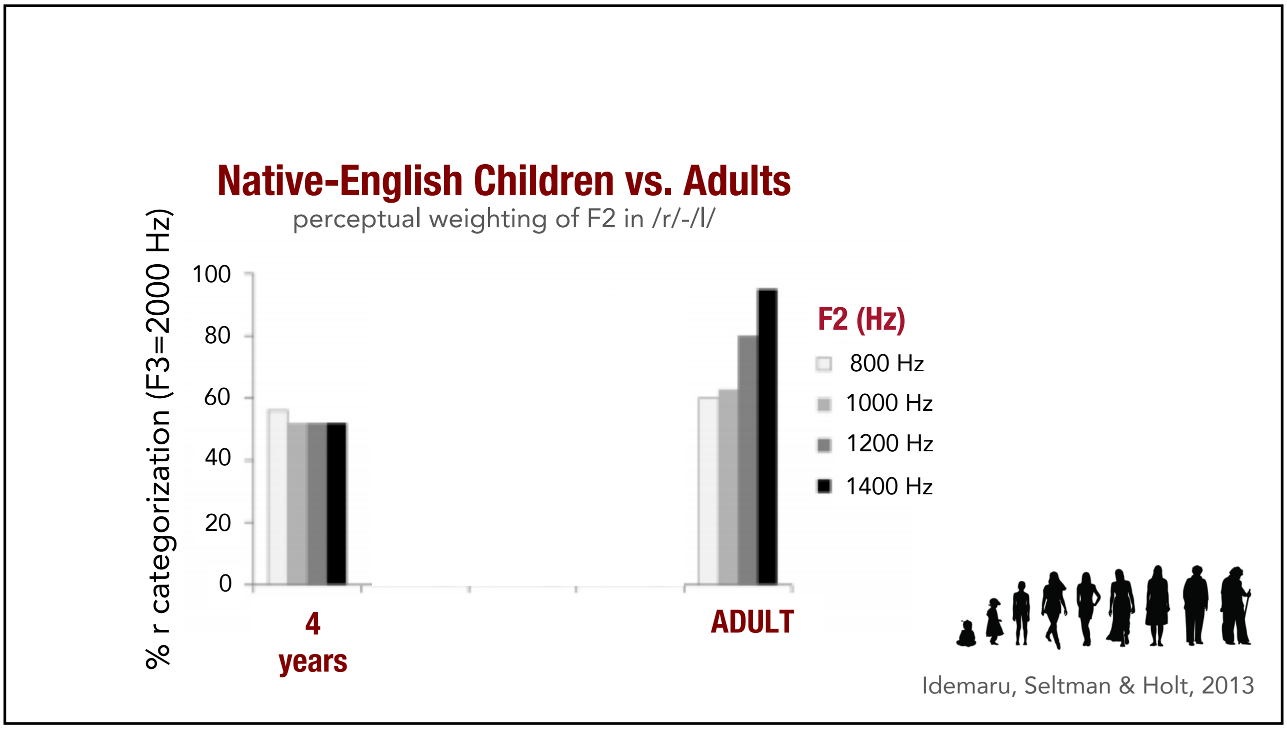
Idemaru, Seltman & Holt, 2013



Native-English Children vs. Adults

perceptual weighting of F2 in /r/-/l/

Idemaru, Seltman & Holt, 2013



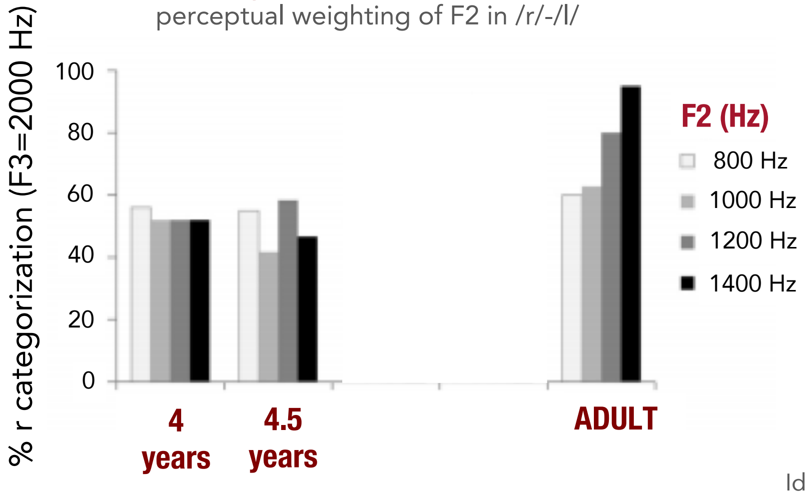
Native-English Children vs. Adults

perceptual weighting of F2 in /r/-/l/

Idemaru, Seltman & Holt, 2013

Native-English Children vs. Adults

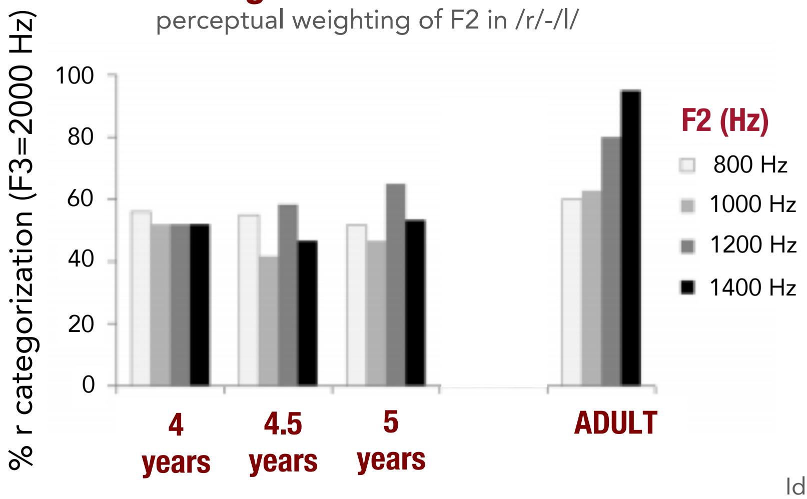
perceptual weighting of F2 in /r/-/l/



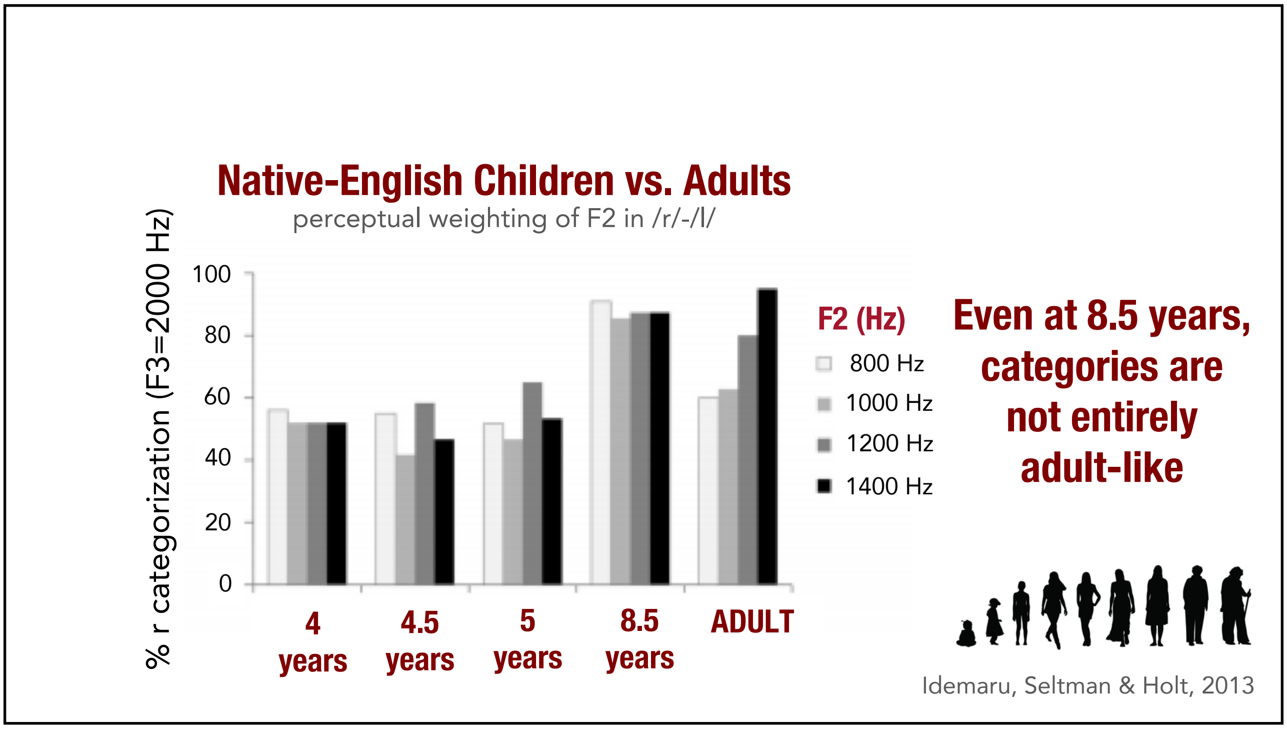
Idemaru, Seltman & Holt, 2013

Native-English Children vs. Adults

perceptual weighting of F2 in /r/-/l/



Idemaru, Seltman & Holt, 2013

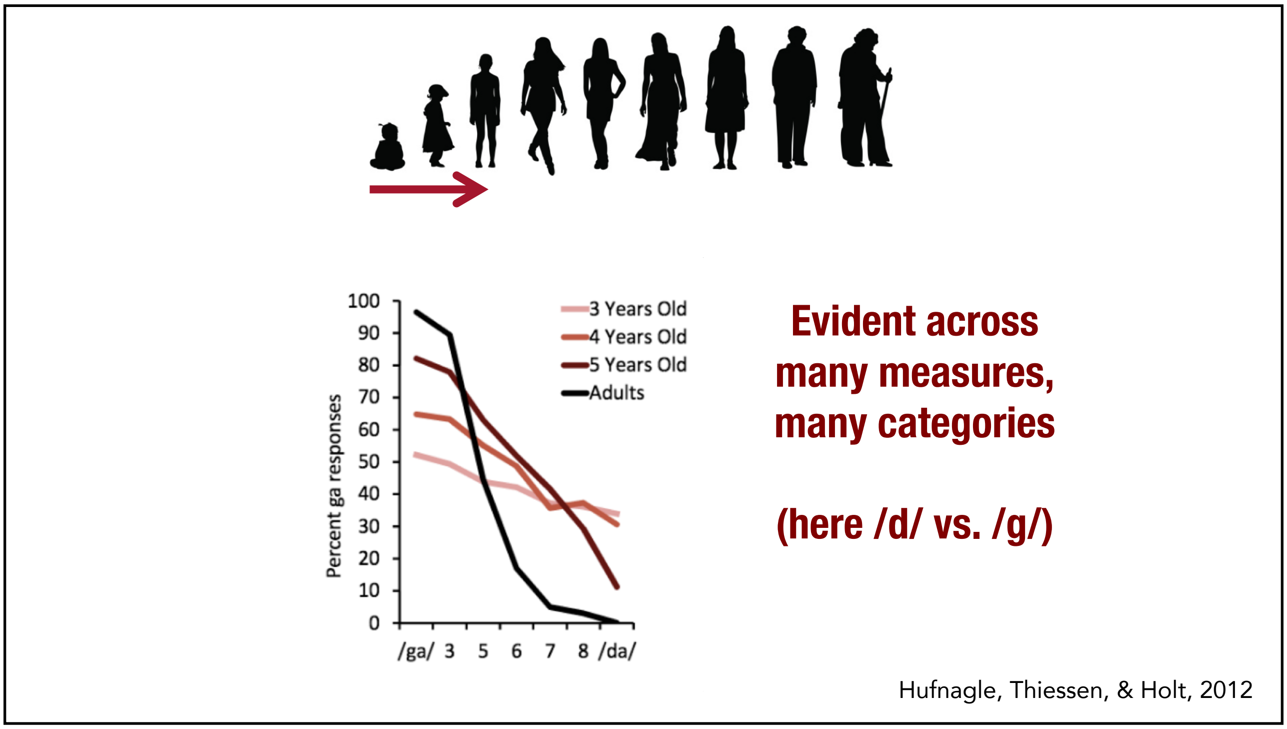


Native-English Children vs. Adults

perceptual weighting of F2 in /r/-/l/

Even at 8.5 years, categories are not entirely adult-like

Idemaru, Seltman & Holt, 2013



Evident across many measures, many categories

(here /d/ vs. /g/)

Hufnagle, Thiessen, & Holt, 2012

There is a long developmental tail to speech category development

There is a long developmental tail to speech category development

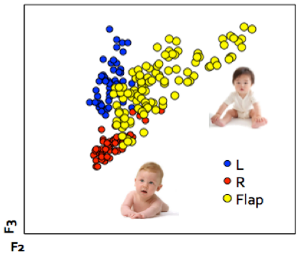
Learning affects listening

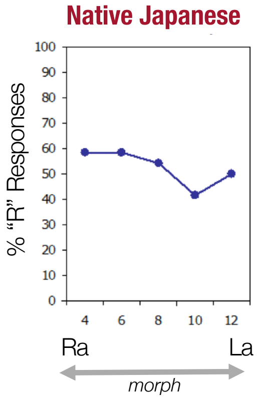
Category learning "warps" perceptual space

Exaggerates differences between categories

There is a long developmental tail to speech category development

This can have profound effects into adulthood

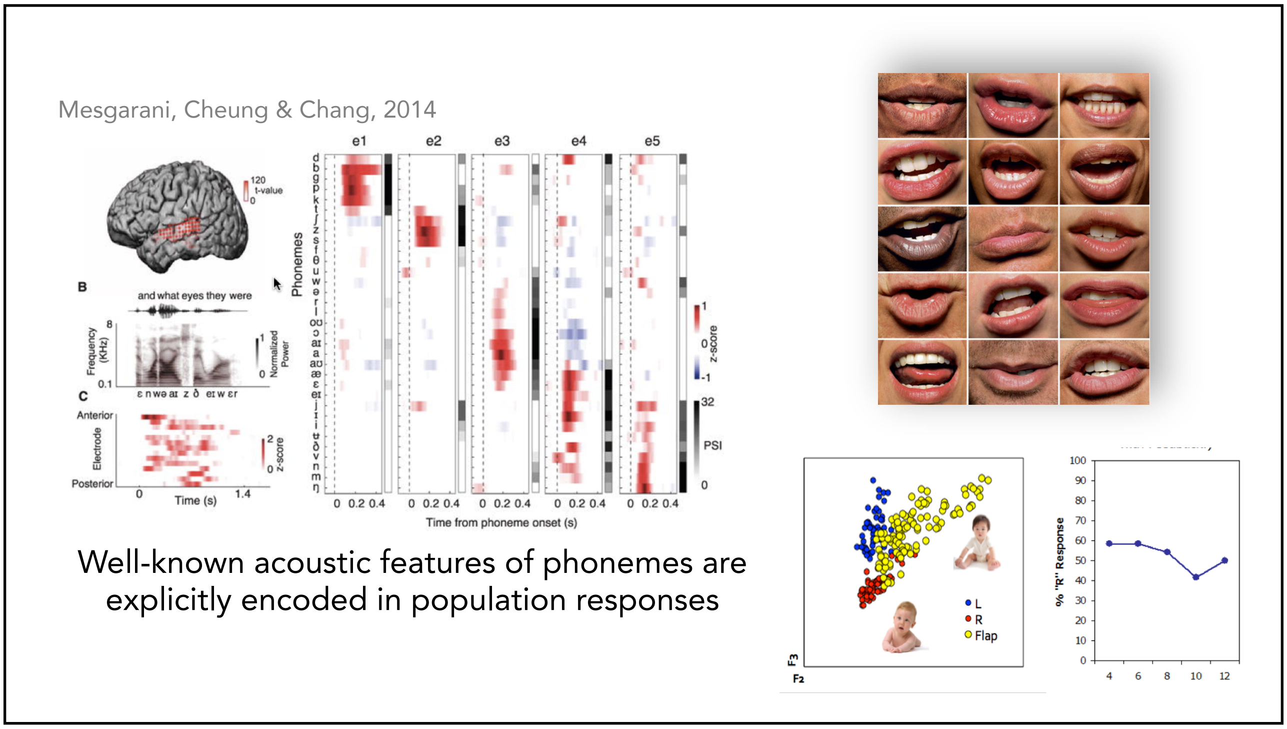




Living / working in US and
4000 trials of explicit training
on endpoint stimuli,
with feedback!

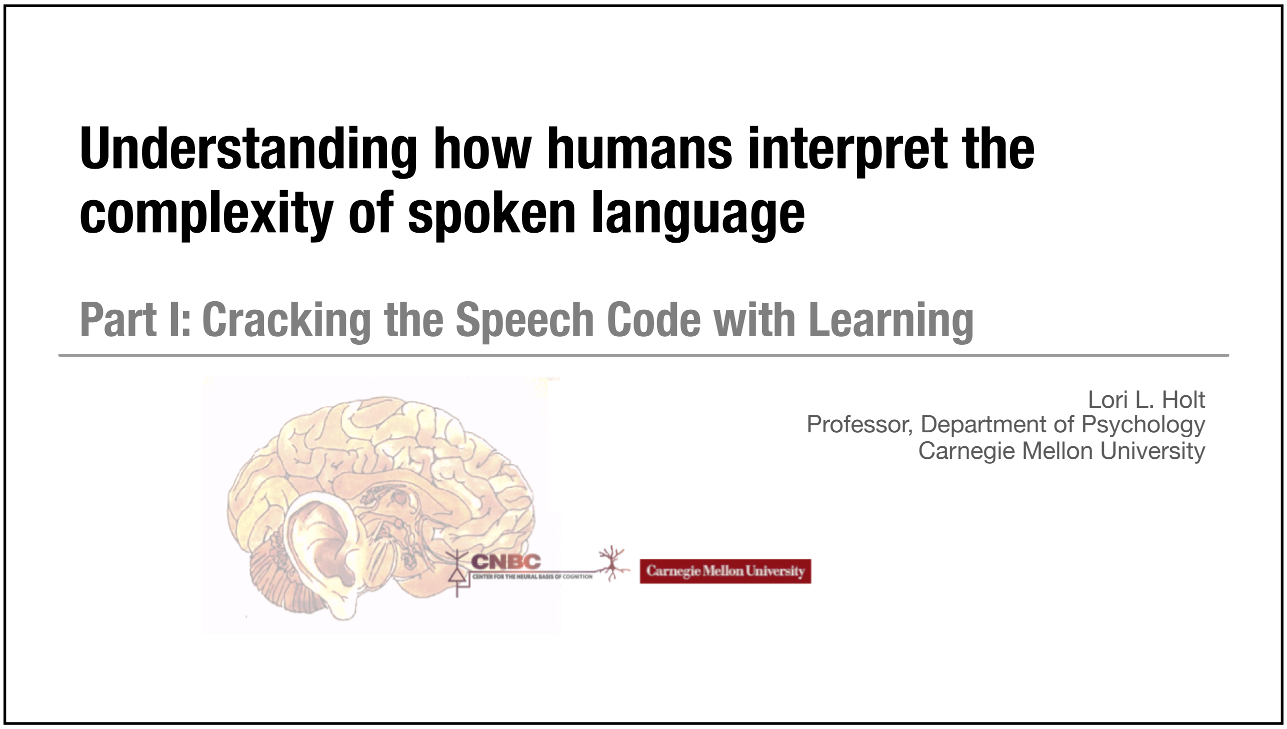
Textbook example of
'lack of plasticity'
among adult learners

Ingvalson, Holt & McClelland, 2012



Mesgarani, Cheung & Chang, 2014

Well-known acoustic features of phonemes are explicitly encoded in population responses



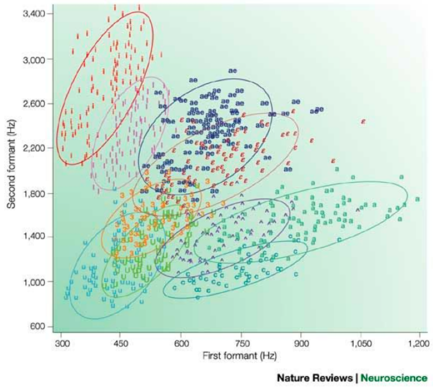
Understanding how humans interpret the complexity of spoken language

Part I: Cracking the Speech Code with Learning

Lori L. Holt
Professor, Department of Psychology
Carnegie Mellon University

CNBC
CENTER FOR THE NEURAL BASIS OF COGNITION
Carnegie Mellon University

We have snapshots at different ages
But… no real understanding of the category learning mechanism(s)

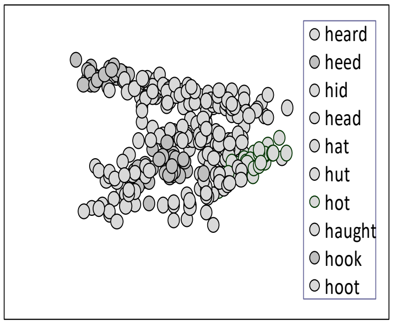


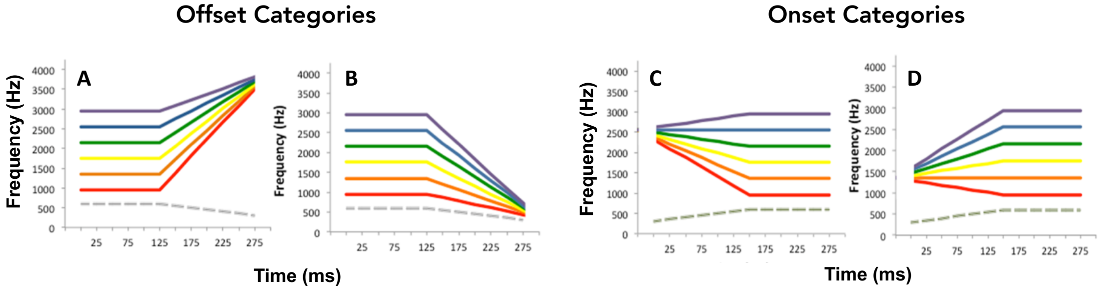
Considering infants' limited behavioral repertoire, unsupervised, passive learning across regularities in the input has been a favored model

of course, difficult to test…
Maye, Werker, Gerken, 2002

Mechanisms of Change

- How do listeners learn across unlabeled categories?
- What is the form of this learning? Is this sensitivity unique to speech?
- Is there intermediate ground between purely passive, unsupervised learning and instruction?

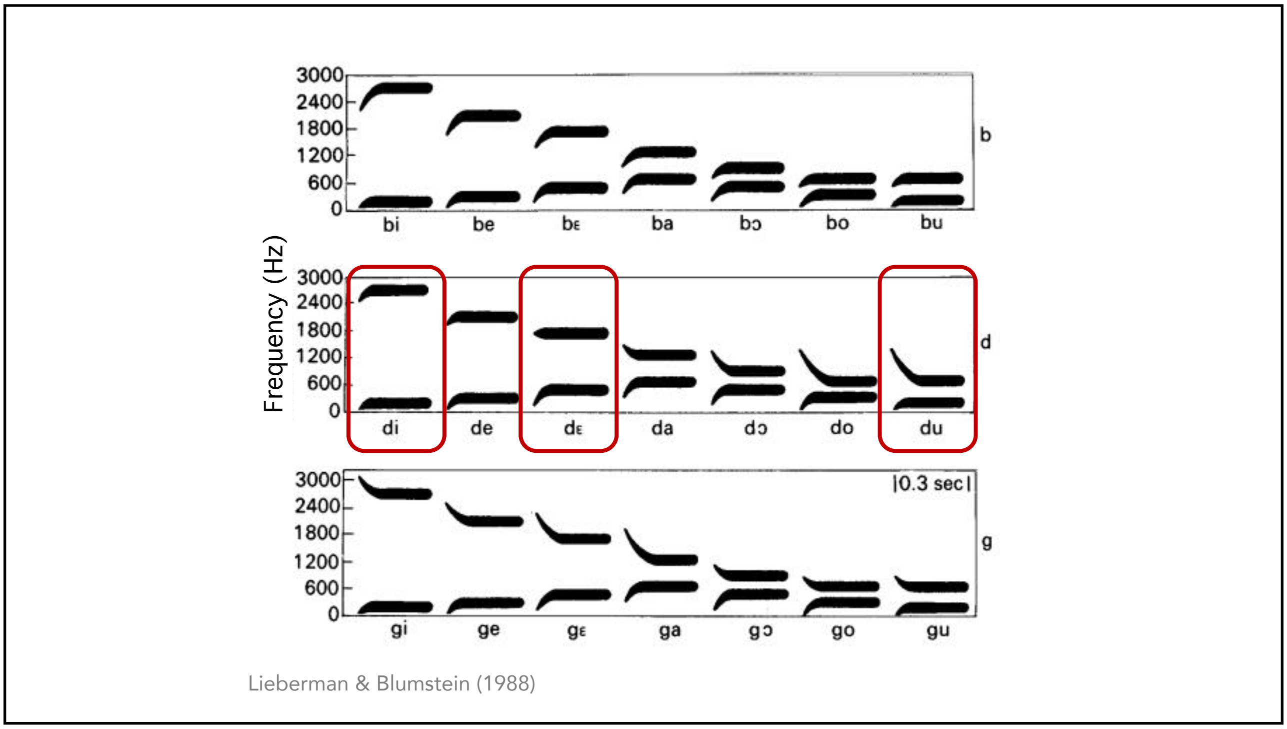




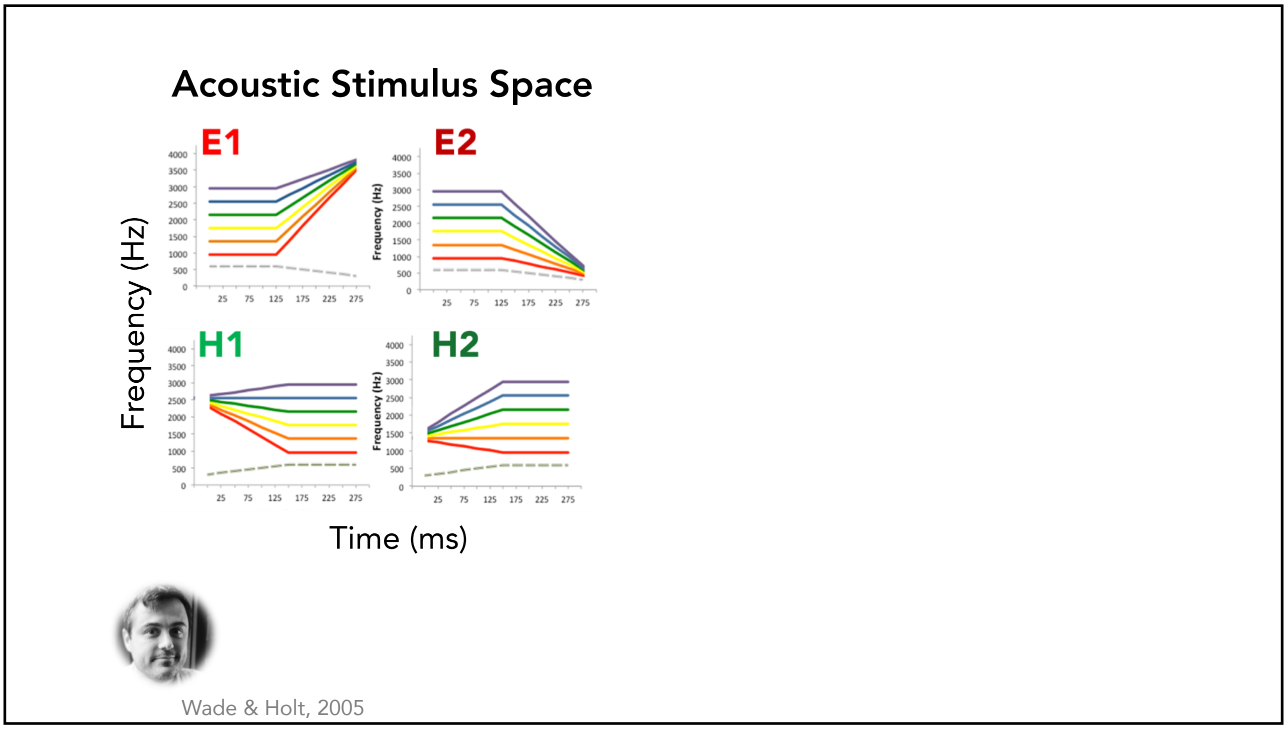
Few exemplars
Simple, unidimensional regularity

Few exemplars
No unidimensional regularity

Wade & Holt, 2005



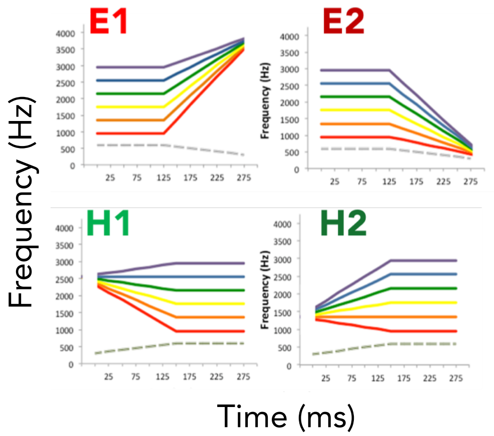
Lieberman & Blumstein (1988)



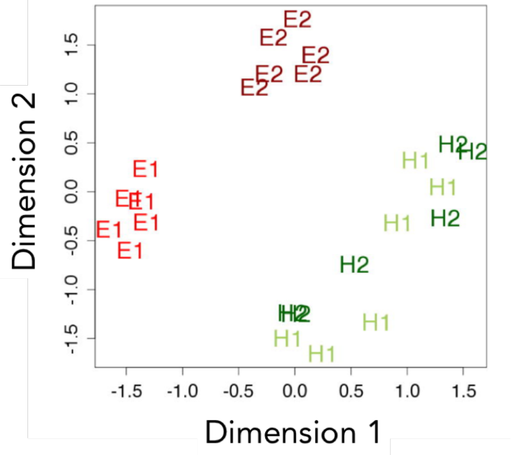
Acoustic Stimulus Space

Wade & Holt, 2005

Acoustic Stimulus Space



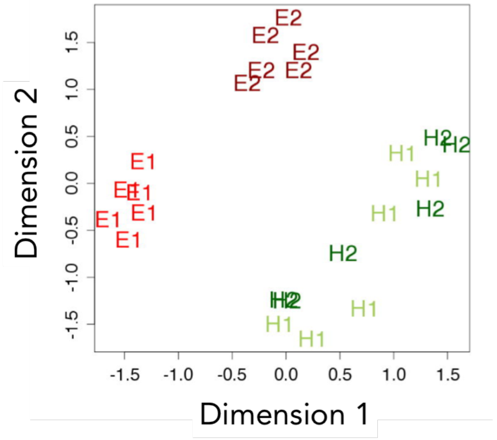
Perceptual Space (MDS)



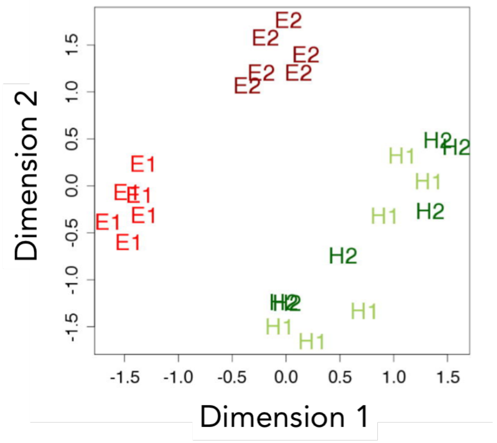
Wade & Holt, 2005

Emberson, Liu & Zevin, 2013

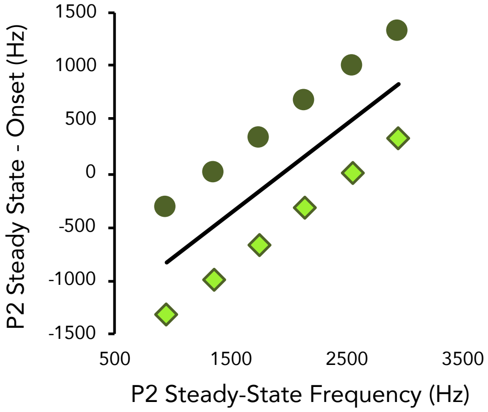
Overlapping in Perceptual Space



Overlapping in Perceptual Space



But, Statistically Structured



"We did not find evidence that exposure facilitated perceptual distinction between H1 and H2" [9 min]

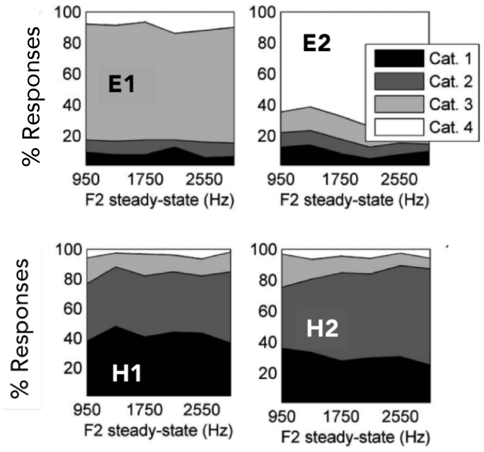
Emberson, Liu & Zevin, 2013

"We did not find evidence that exposure facilitated perceptual distinction between H1 and H2" [9 min]

Emberson, Liu & Zevin, 2013

Exposure via an unsupervised sorting task did not differentiate H1 and H2

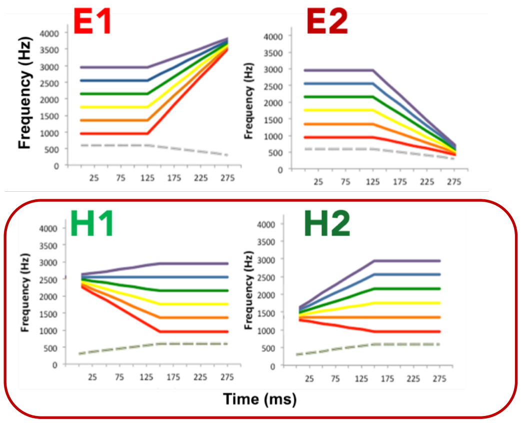
[30 min]

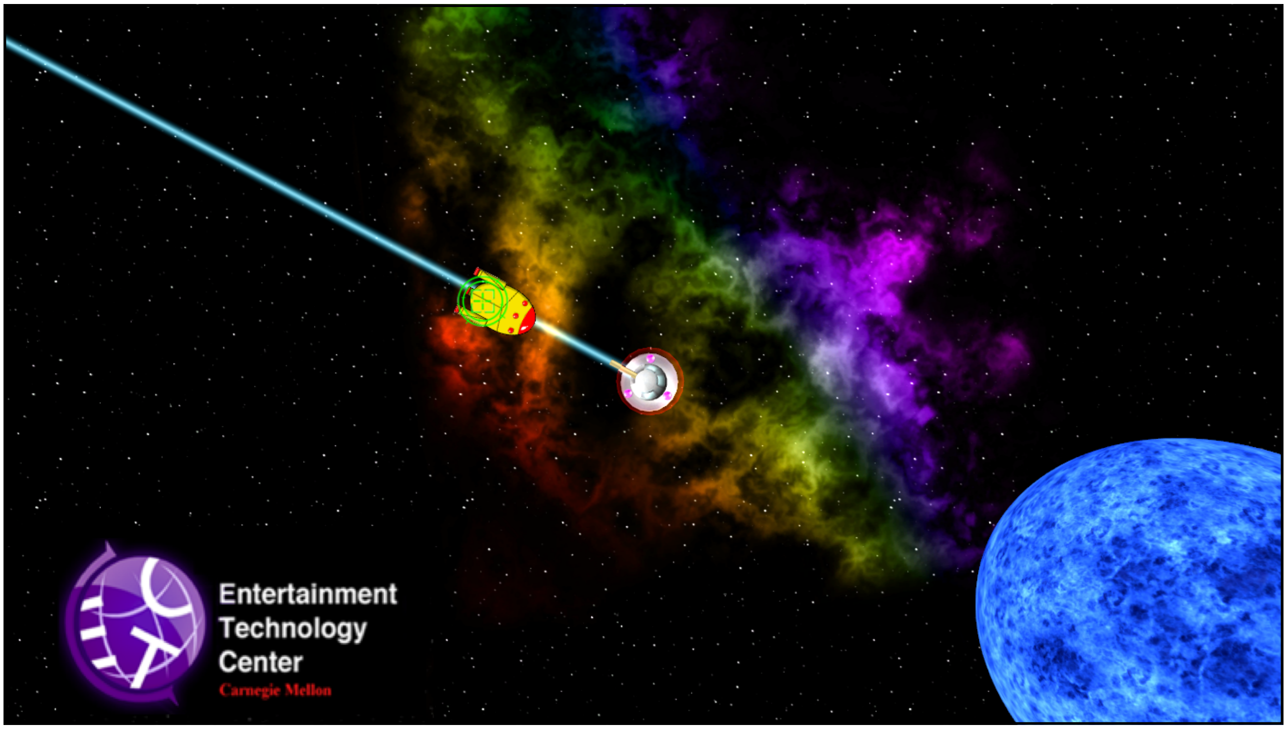


Wade & Holt, 2005

The Puzzle

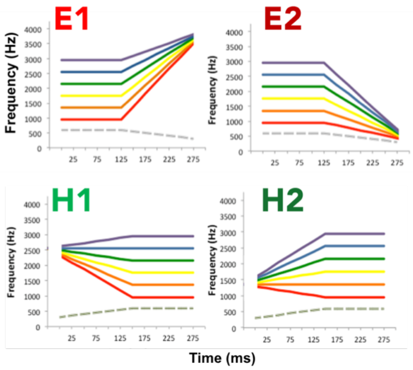
- Statistical structured, but not learned across passive exposure
- Simplifying a challenge present in speech categories
- Few exemplars
- Modest acoustic complexity




Entertainment
Technology
Center
Carnegie Mellon

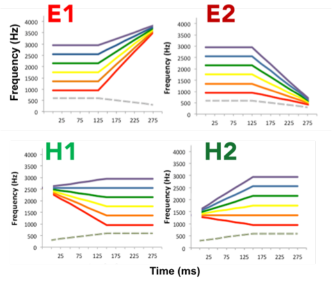
Incidental Learning

Alignment of behaviorally-relevant environmental events with statistically-structured input promote learning above-and-beyond passive exposure



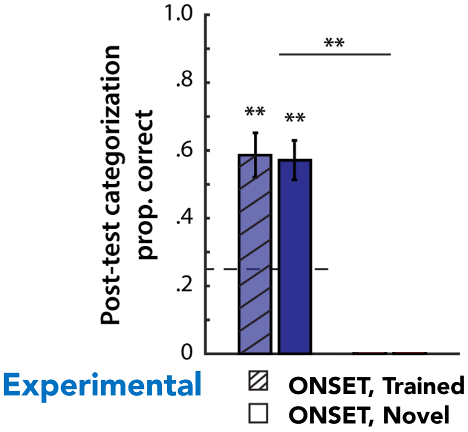
36 minutes of game play

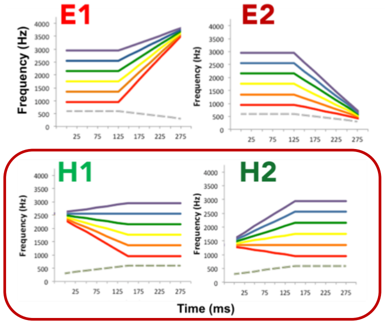
After, an explicit labeling task (novel generalization sounds)



Lim, Fiez, & Holt, PNAS, 2019
also: Wade & Holt, 2005; Leech et al. 2009; Gabay et al. 2015

Behavioral Post-test Results



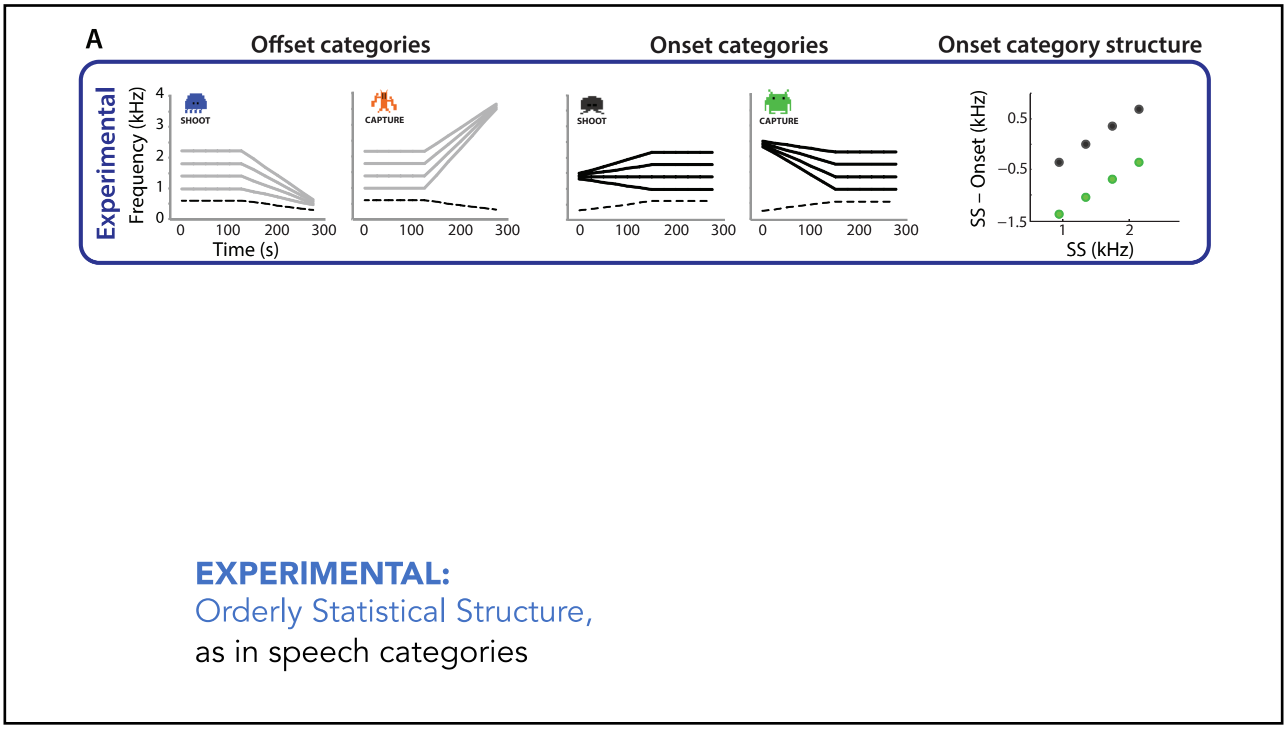


Lim, Fiez, & Holt, PNAS, 2019
also: Wade & Holt, 2005; Leech et al. 2009; Gabay et al. 2015

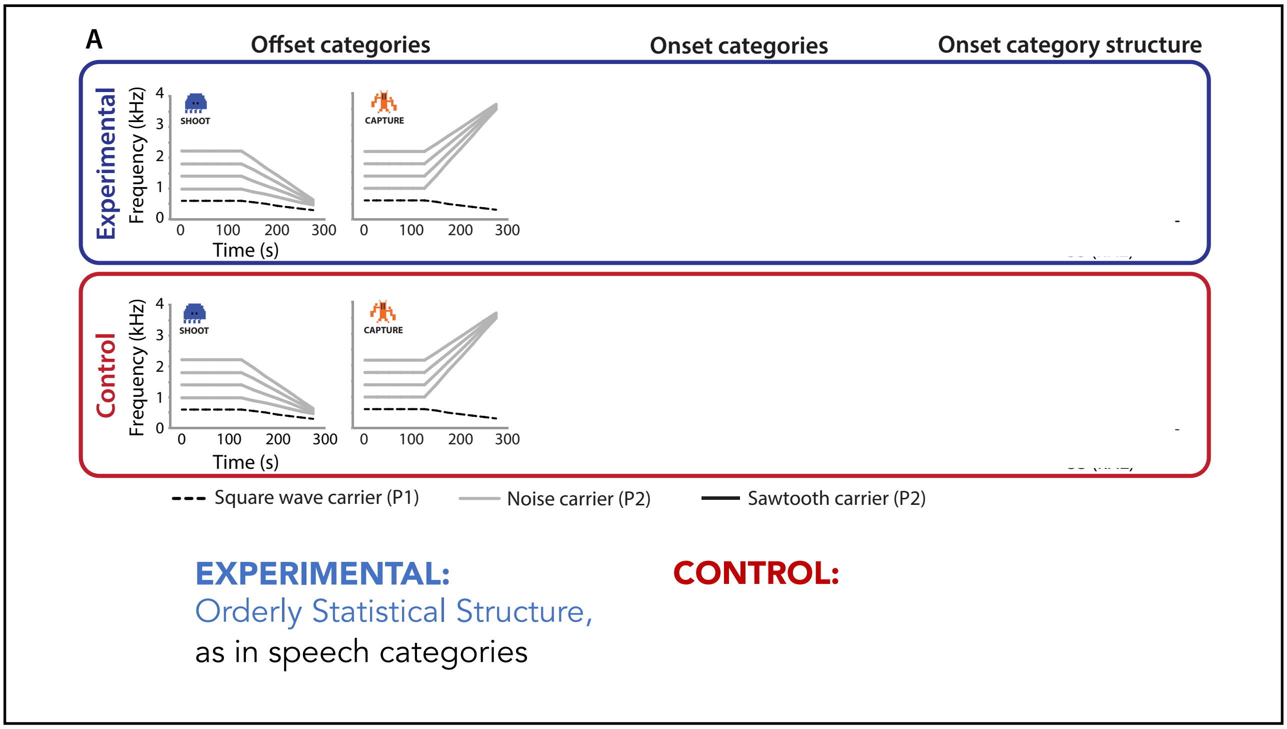
1 **Listeners can learn auditory categories incidentally**

1 **Listeners can learn auditory categories incidentally**

2 **Is this learning ‘statistical’?**

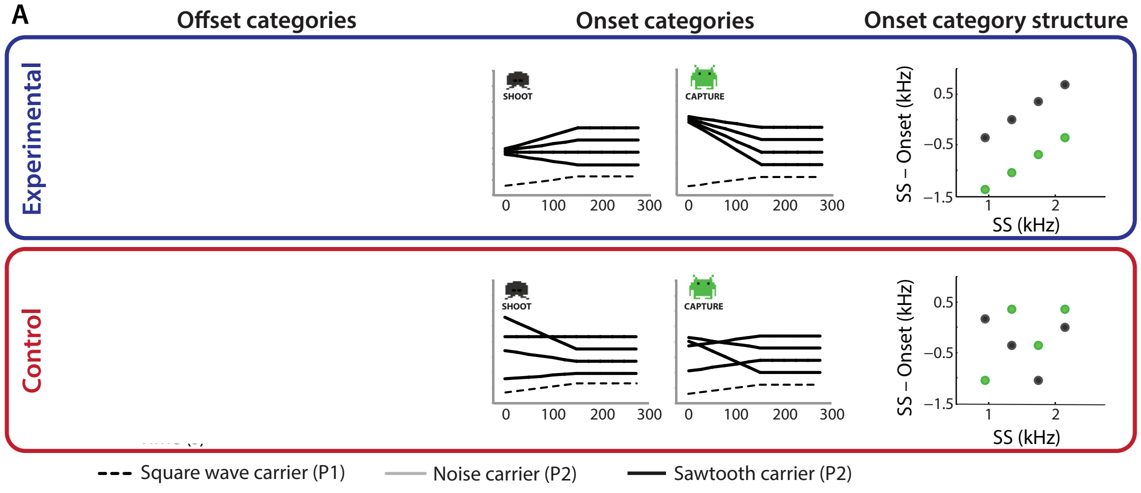


EXPERIMENTAL:
Orderly Statistical Structure,
as in speech categories



EXPERIMENTAL:
Orderly Statistical Structure,
as in speech categories

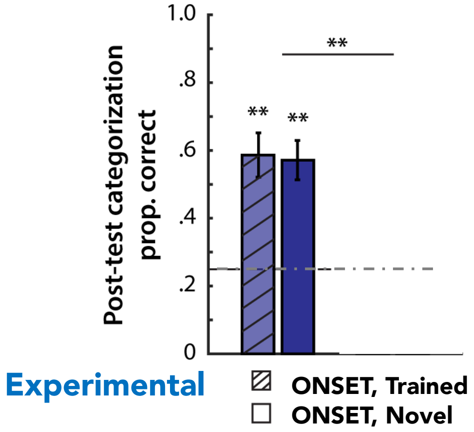
CONTROL:



EXPERIMENTAL:
Orderly Statistical Structure,
as in speech categories

CONTROL:
No Orderly Statistical Structure
Sampling same acoustic space
Equated exposure with Experimental

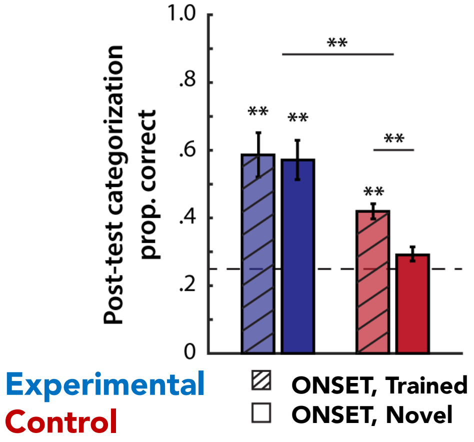
Behavioral Post-test Results



Statistically-structured input is learned when aligned with behaviorally-relevant actions and events.

Lim, Fiez, & Holt, PNAS, 2019
also: Wade & Holt, 2005; Leech et al. 2009; Gabay et al. 2015

Behavioral Post-test Results



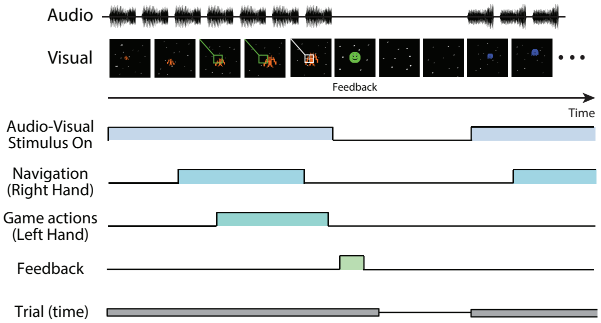
When input is less statistically-structured, there is poor incidental learning.

Lim, Fiez, & Holt, PNAS, 2019
also: Wade & Holt, 2005; Leech et al. 2009; Gabay et al. 2015

1 **Listeners can learn auditory categories incidentally**

2 **Incidental learning is sensitive to the statistical regularity in the input**

1. **Listeners can learn auditory categories incidentally**
2. **Incidental learning is sensitive to the statistical regularity in the input**
3. **What supports incidental learning?**



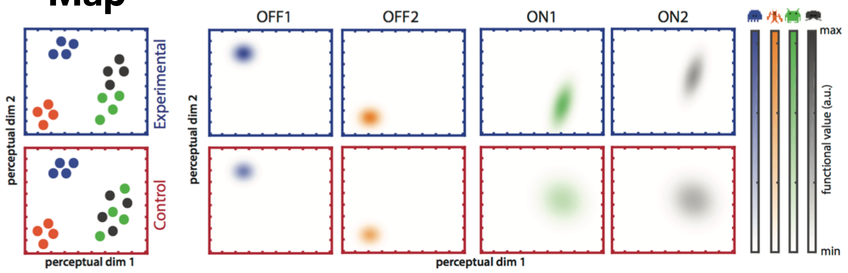
36 minutes of game play

After, an explicit labeling task (novel generalization sounds)

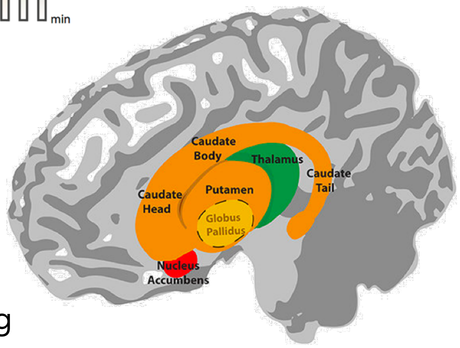
Lim, Fiez & Holt
PNAS 2019

Stimulus Map

Value-Marking Behavioral Outcomes



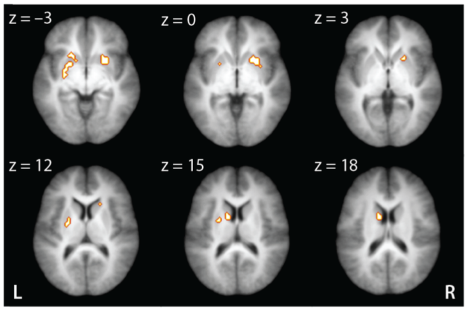
Reward value is 'incidental' from time-alignment with other behaviors



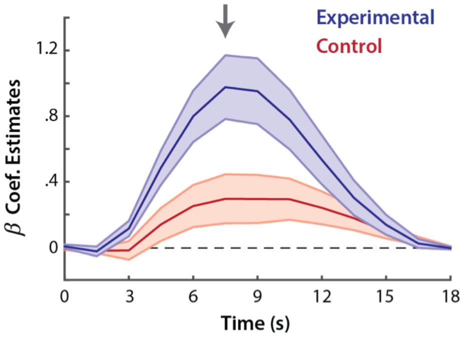
Posterior Striatum – caudate body/tail and putamen
Widely implicated in non-declarative, implicit learning

Striatal Engagement in Incidental Learning

Average Activation by Condition
left caudate body and
bilateral putamen



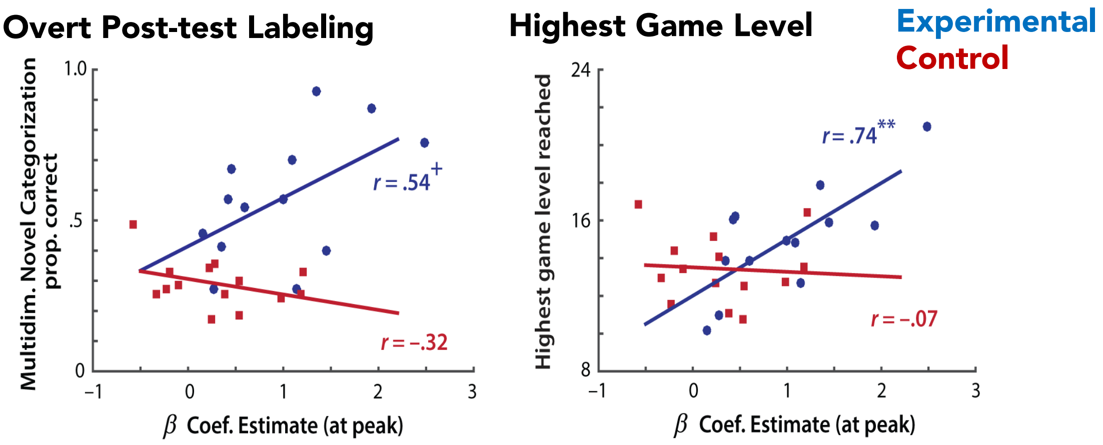
Average Activation by Condition
across striatal voxels



Lim, Fiez, & Holt, PNAS 2019

Behavioral measures of incidental category learning are correlated with striatal activation

for the Experimental Condition

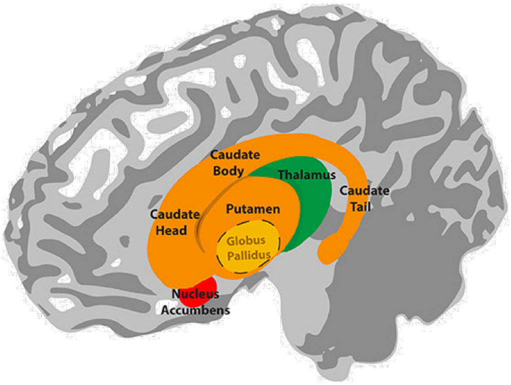


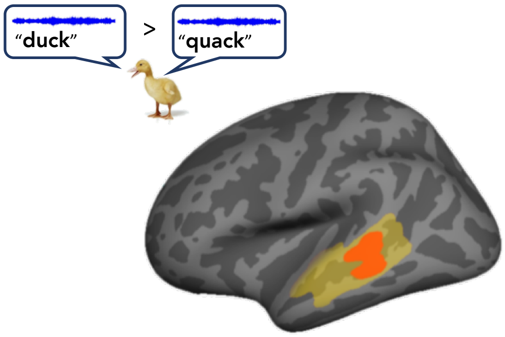
Overt Post-test Labeling

Highest Game Level

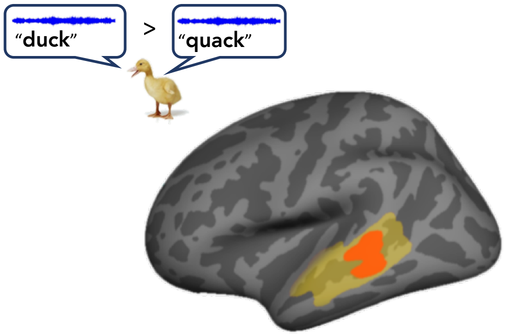
Lim, Fiez, & Holt, PNAS 2019

What is the Role of Striatum?

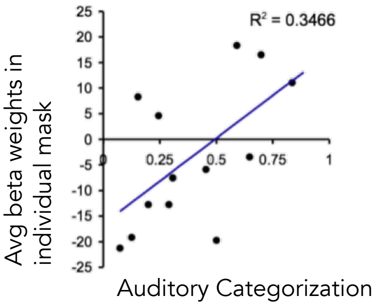




Leech, Holt, Devlin & Dick, 2009

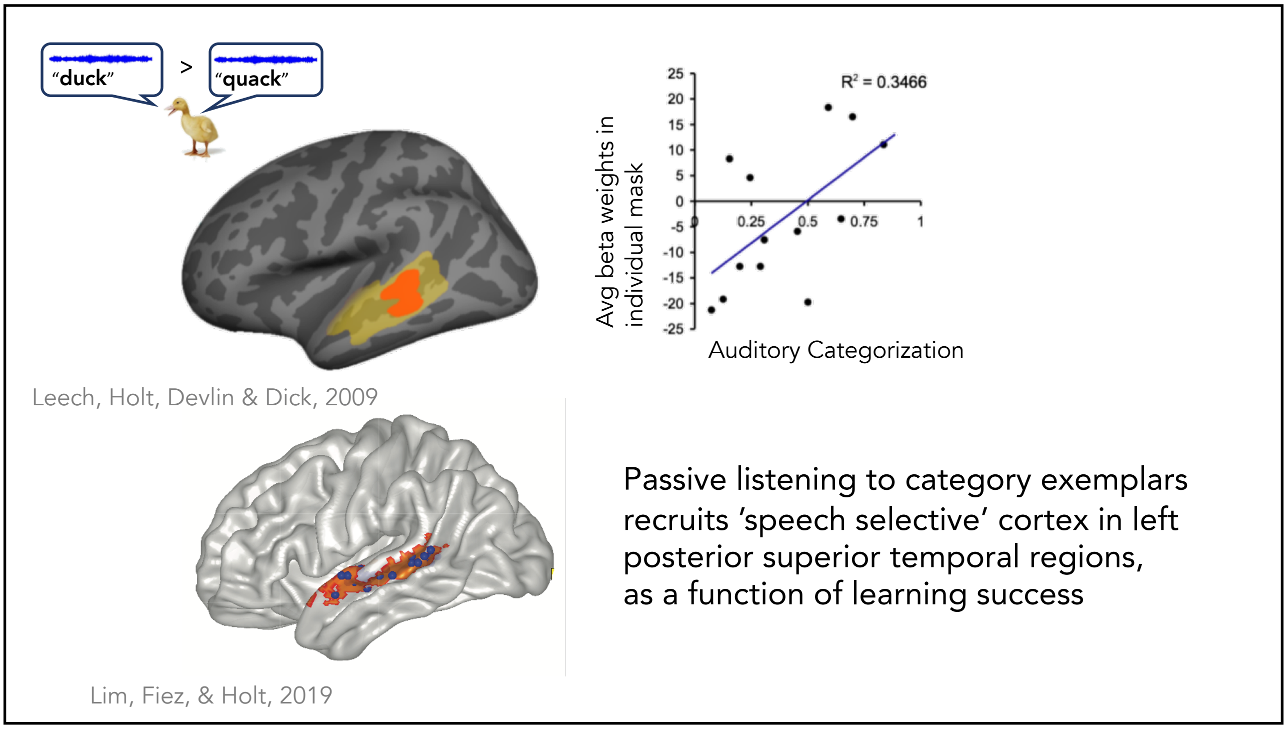


Leech, Holt, Devlin & Dick, 2009



Passive listening to category exemplars recruits 'speech selective' cortex in left posterior superior temporal regions at post-test,

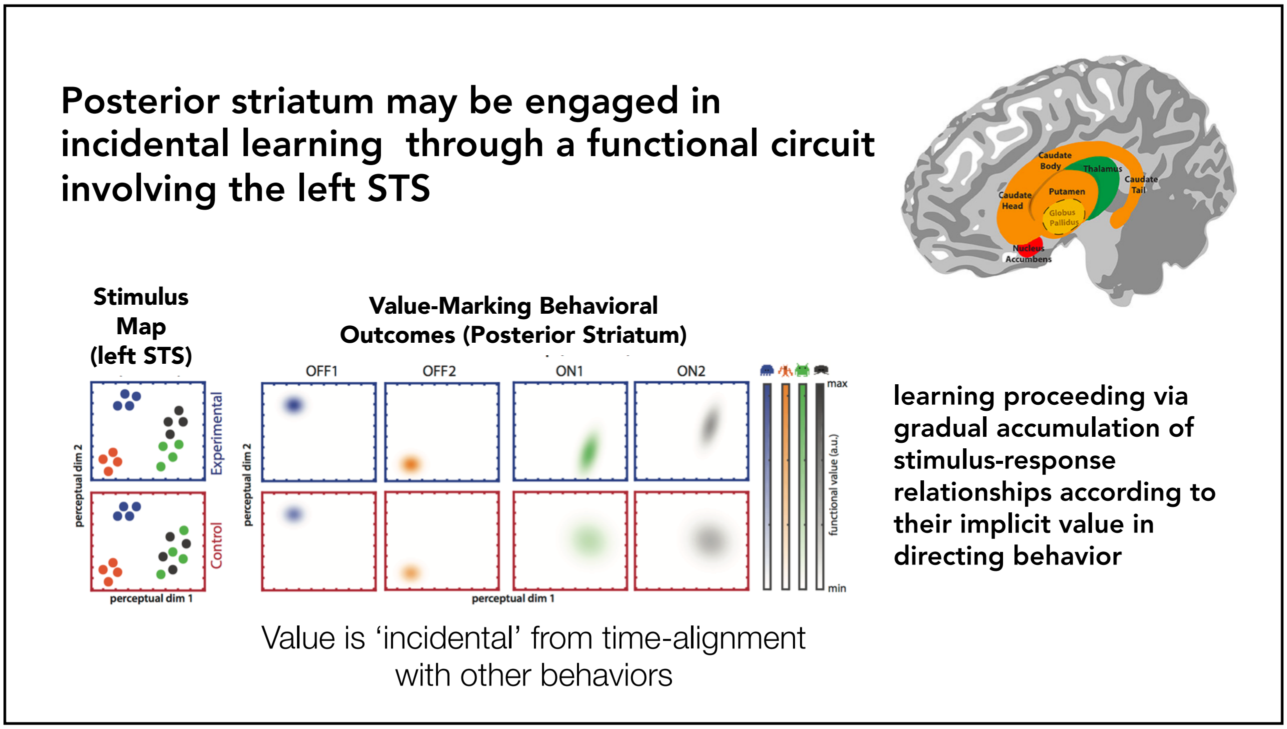
as a function of category learning success



Leech, Holt, Devlin & Dick, 2009

Lim, Fiez, & Holt, 2019

Passive listening to category exemplars recruits 'speech selective' cortex in left posterior superior temporal regions, as a function of learning success

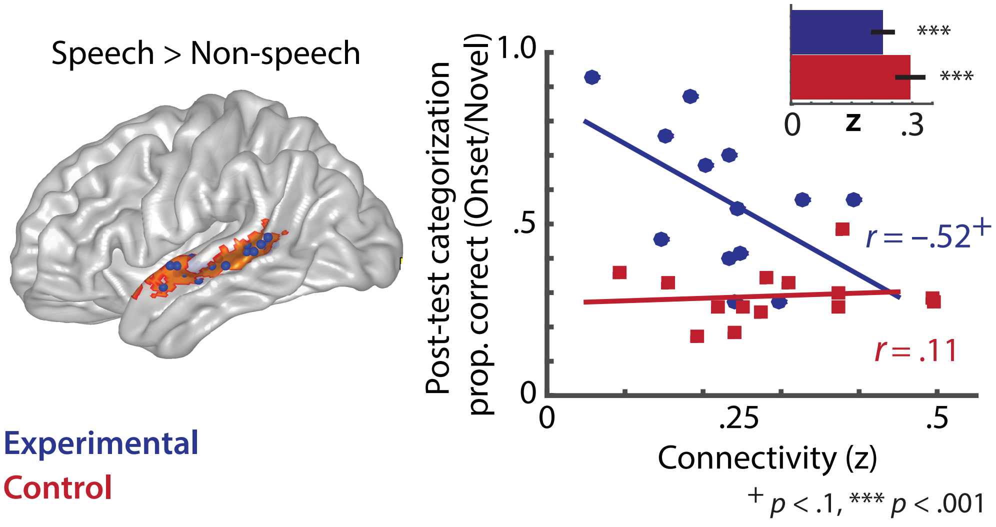


Posterior striatum may be engaged in incidental learning through a functional circuit involving the left STS

Value is 'incidental' from time-alignment with other behaviors

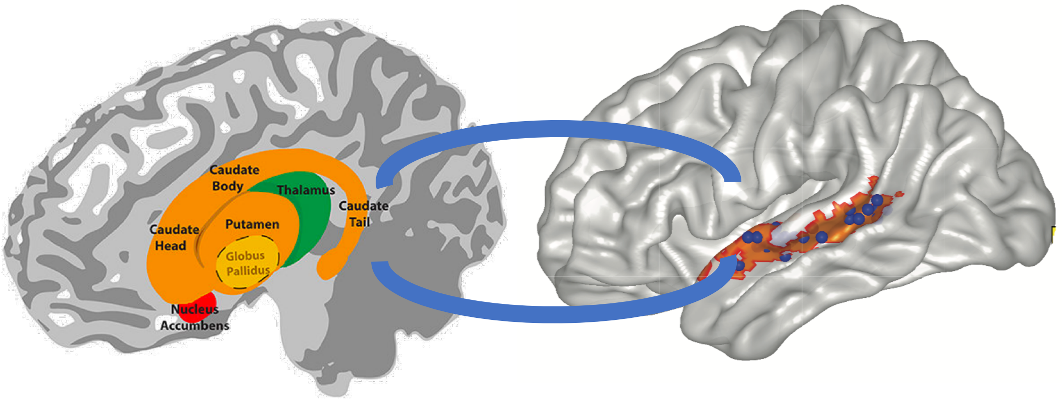
learning proceeding via gradual accumulation of stimulus-response relationships according to their implicit value in directing behavior

Striatum–left STS connectivity

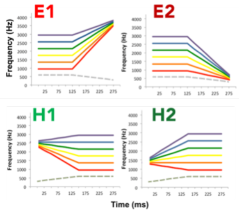


$^{+}p < .1$, $^{***}p < .001$

Cortico-striatal Interaction

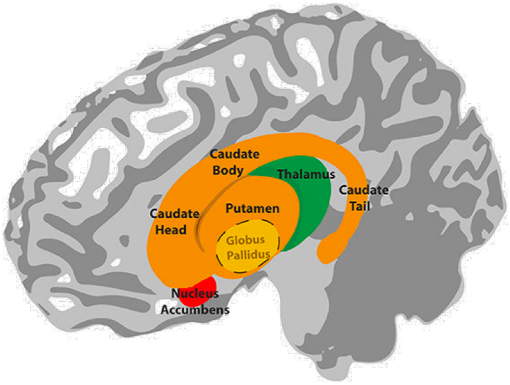


Lim, Fiez, & Holt, PNAS 2019



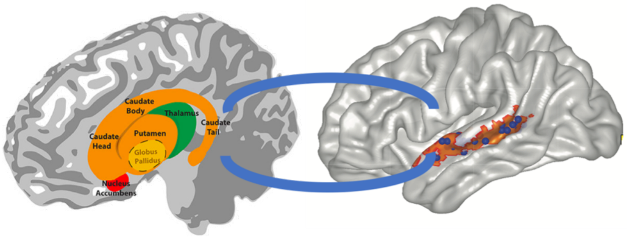
Active engagement in an environment aligned with the statistical structure provides an 'assist' to learning distributional regularities

Recruitment of the striatum may be essential in learning across distributions of input that are difficult to acquire through unsupervised learning

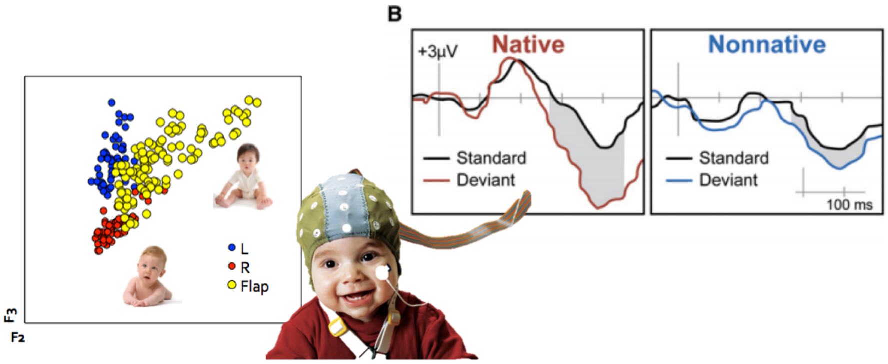


Incidental Learning

Learning across statistical regularities can be incidental, and not overtly driven by an intention to learn, while still taking place in the context of an active task that generates valuable predictions and rewarding outcomes.



But is there incidental learning of speech categories?

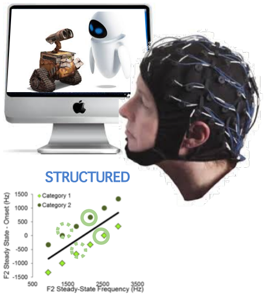


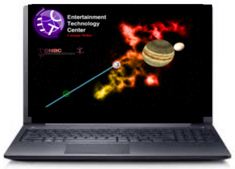
5 Days Video Game Play

30-min/day

Pre-Training

Passive Listening to Sounds

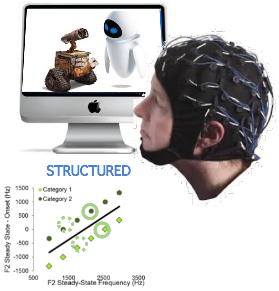




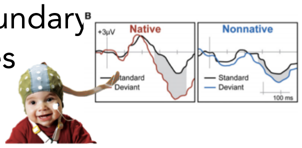
Liu & Holt, 2011

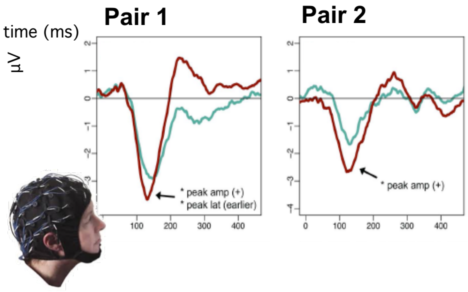
Post-Training

Passive Response to Sounds



Mismatch Negativity (MMN) for stimuli that cross a newly-learned category boundary just as in infant speech studies





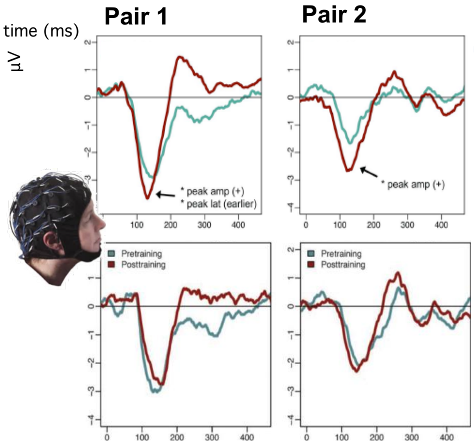
▬ Pre-training
▬ Post-training

category learning warps perceptual space

Good Learners

generalization accuracy >50%

Liu & Holt (2011)



▬ Pre-training
▬ Post-training

category learning warps perceptual space

Good Learners

generalization accuracy >50%

not mere exposure

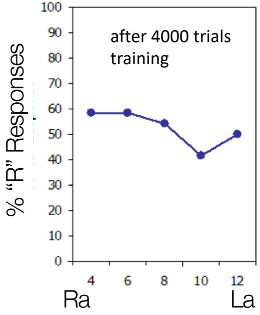
Poor Learners

generalization accuracy <50%

Liu & Holt (2011)

But is there incidental learning of speech categories?

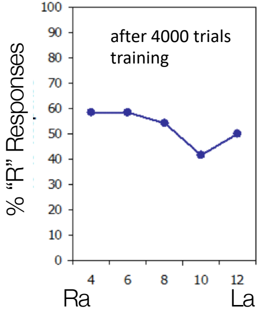
Native Japanese



Ingvalson, McClelland, & Holt, 2012

But is there incidental learning of speech categories?

Native Japanese

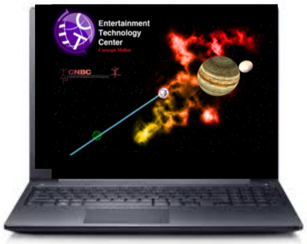


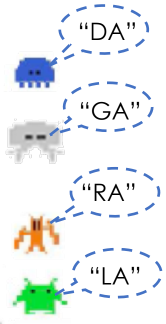
PARTICIPANTS
Native Japanese
Late learners of English
<2 years in US

Pretest/Posttest
Battery of English /r/-/l/ perception tests

TRAINING
2.5 hours of video game
across 5 days

Lim & Holt, 2011





CONTROL CATEGORIES
exist in Japanese (easy)

TEST CATEGORIES
not in Japanese (difficult)

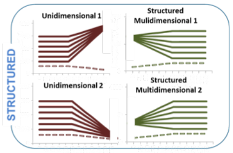
Lim & Holt, 2011

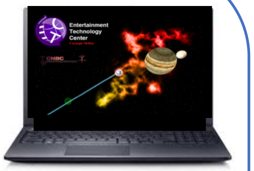
TRAINING
2.5 hours of video game
across 5 days

Native Japanese Adults
Late learners of English
<2 years in US

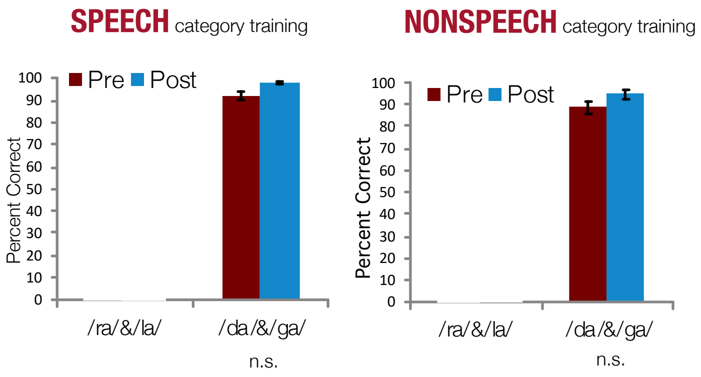
Pretest/Posttest
Battery of English
/r/-/l/ perception tests

CONTROL: Nonspeech



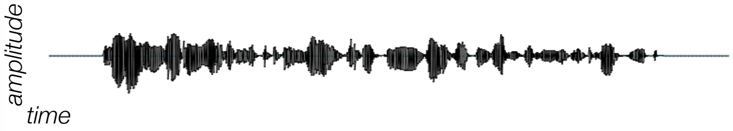


Lim & Holt, 2011



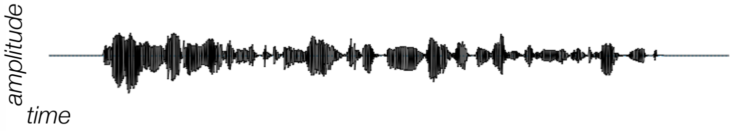
Lim & Holt, 2011

Yet…



therearenosilencesbetweenwordsastherearewhitespacesinwrittentext

there are no silences between words as there are white spaces in written text

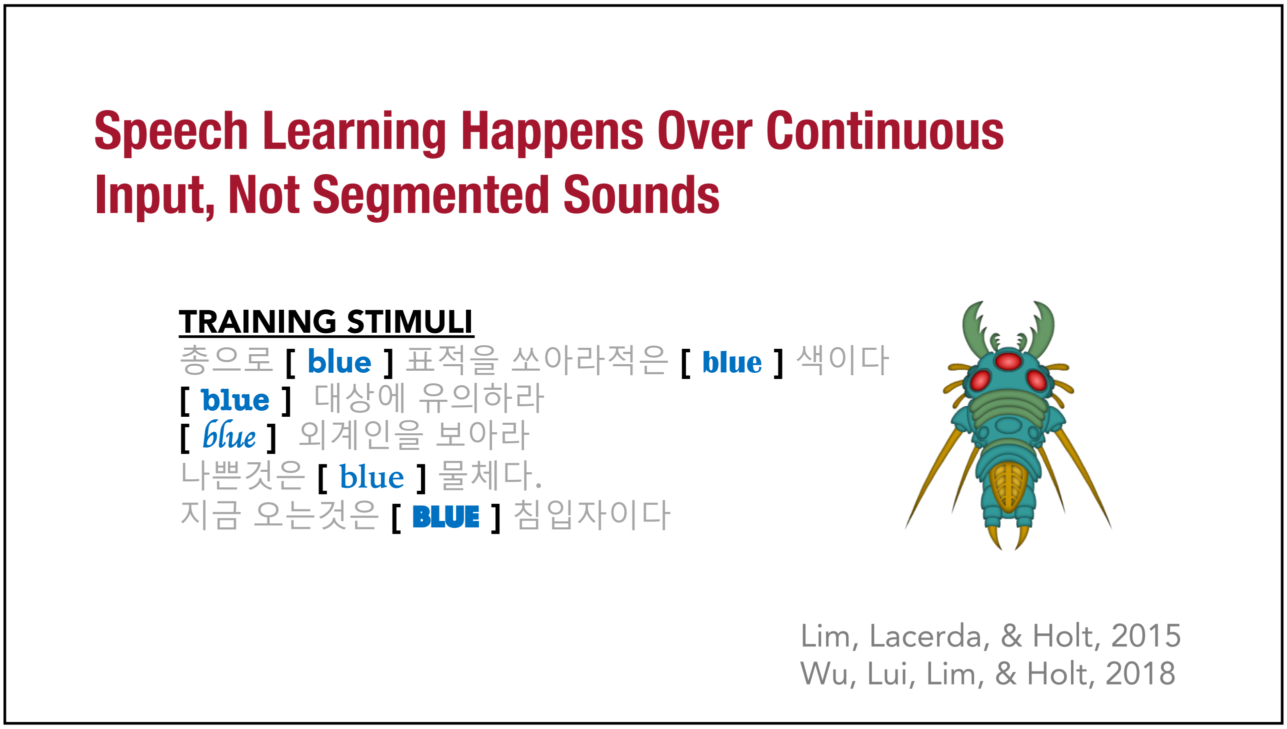


therearenosilencesbetweenwordsastherearewhitespacesinwrittentext

there are no silences between words as there are white spaces in written text

Even if you discover a 'unit' in continuous sound, it varies across instances

Categorization happens in the context of segmentation; each requires learning

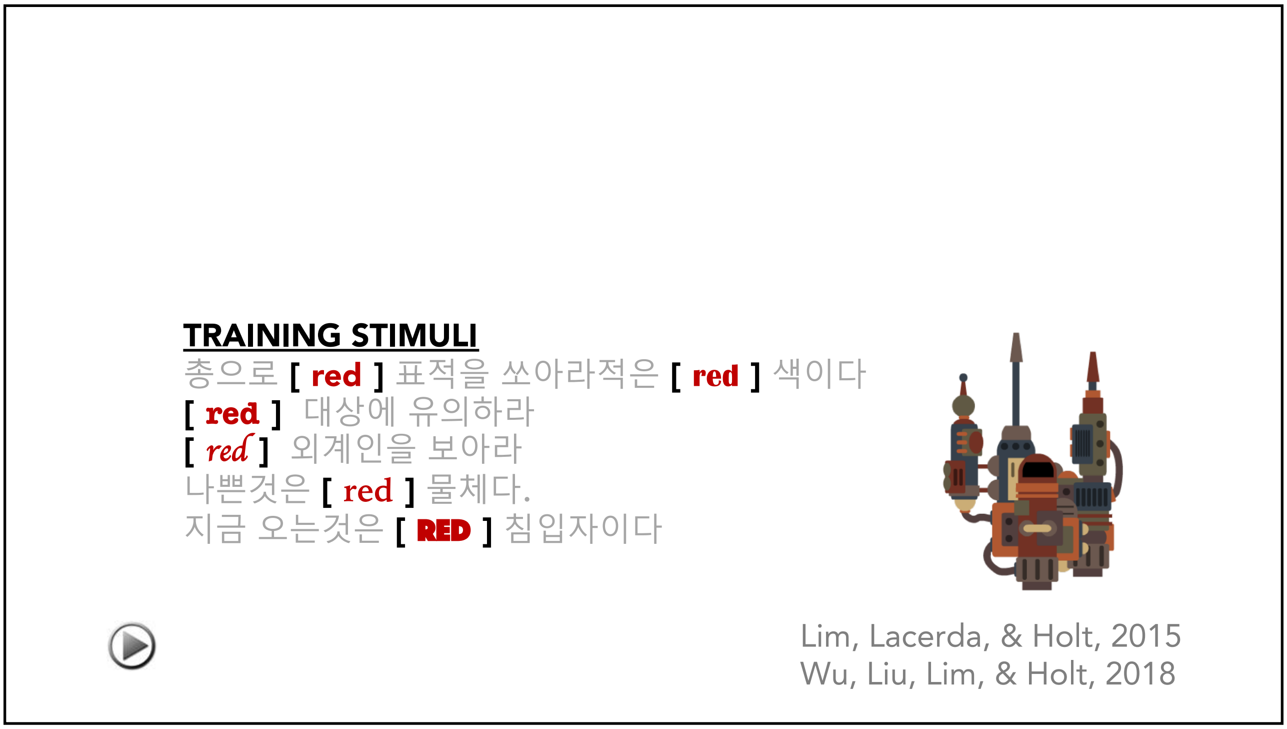


Speech Learning Happens Over Continuous Input, Not Segmented Sounds

TRAINING STIMULI

총으로 [**blue**] 표적을 쏘아라적은 [**blue**] 색이다
[**blue**] 대상에 유의하라
[*blue*] 외계인을 보아라
나쁜것은 [blue] 물체다.
지금 오는것은 [**BLUE**] 침입자이다

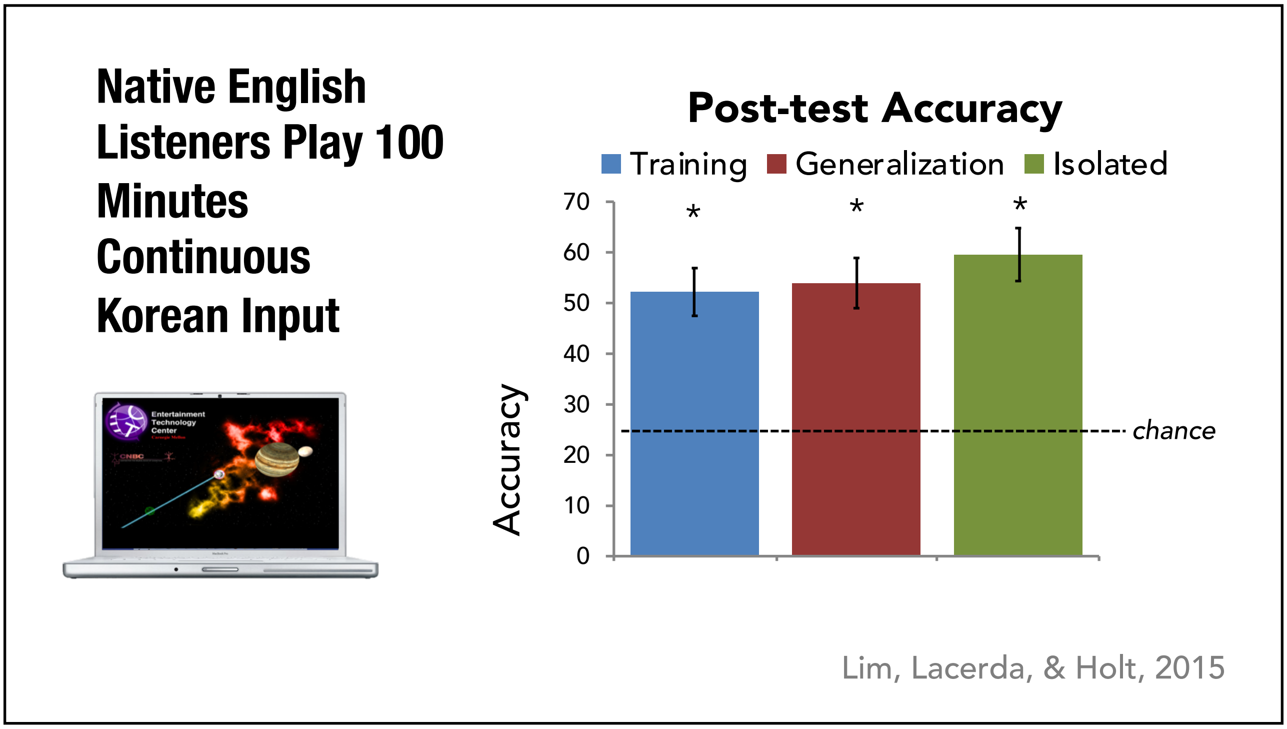
Lim, Lacerda, & Holt, 2015
Wu, Lui, Lim, & Holt, 2018



TRAINING STIMULI

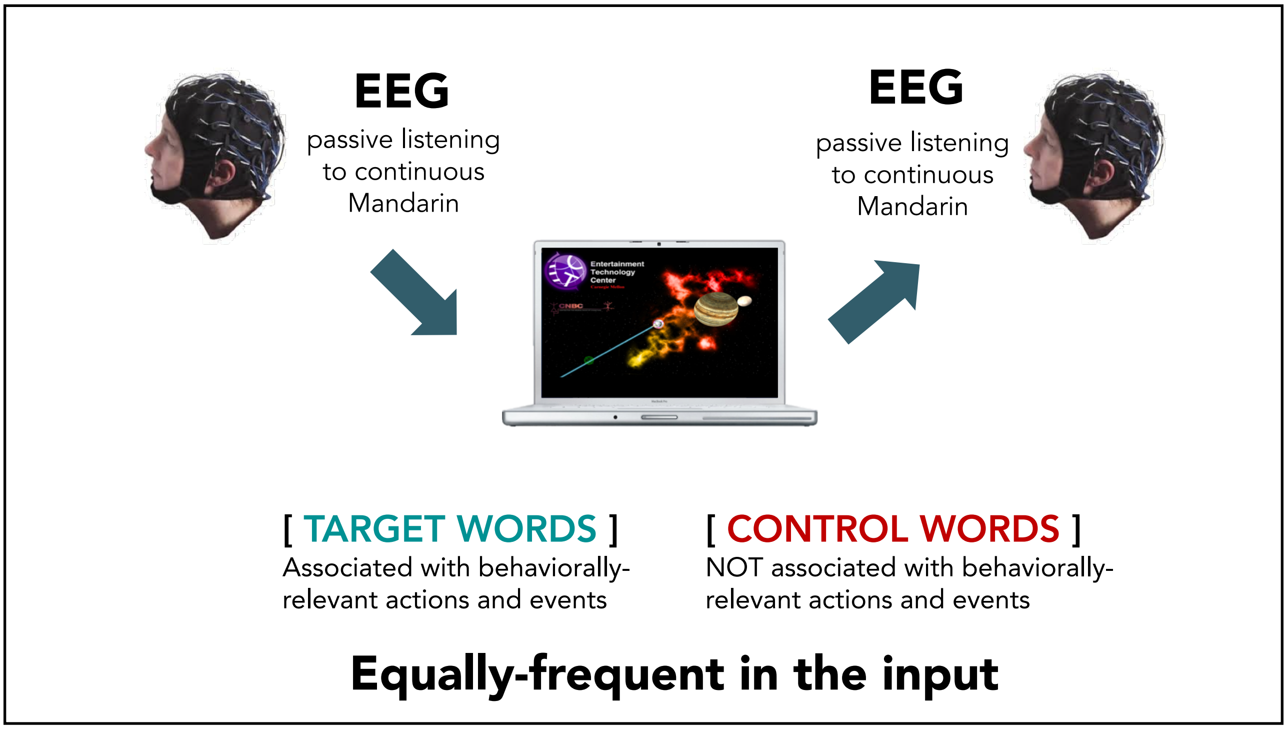
총으로 [**red**] 표적을 쏘아라적은 [**red**] 색이다
[**red**] 대상에 유의하라
[*red*] 외계인을 보아라
나쁜것은 [red] 물체다.
지금 오는것은 [**RED**] 침입자이다

Lim, Lacerda, & Holt, 2015
Wu, Liu, Lim, & Holt, 2018



Native English Listeners Play 100 Minutes Continuous Korean Input

Lim, Lacerda, & Holt, 2015

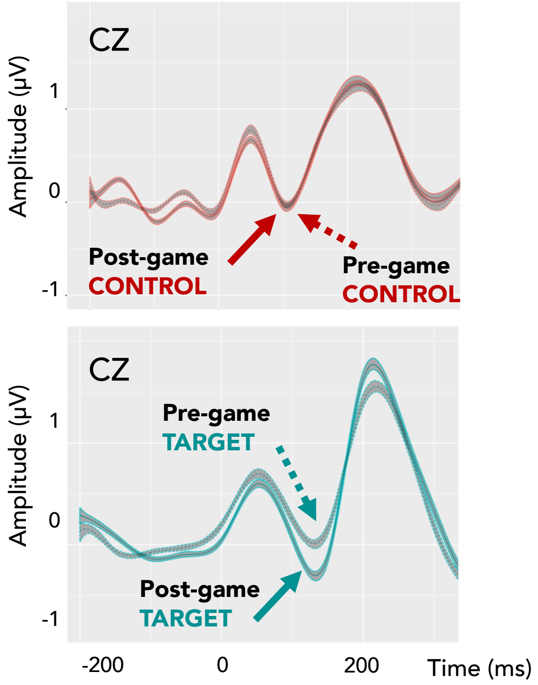


EEG
passive listening to continuous Mandarin
EEG
passive listening to continuous Mandarin

[TARGET WORDS]
Associated with behaviorally-relevant actions and events

[CONTROL WORDS]
NOT associated with behaviorally-relevant actions and events

Equally-frequent in the input

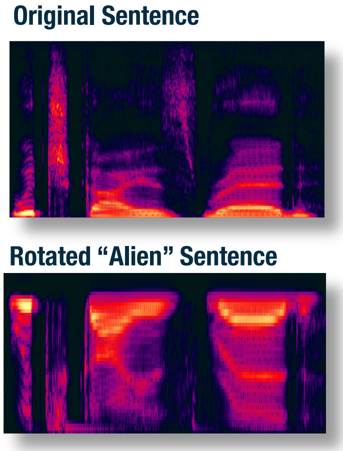


Alignment with behaviorally-relevant actions and events (**TARGET**) leads to different outcomes than passive exposure (**CONTROL**)

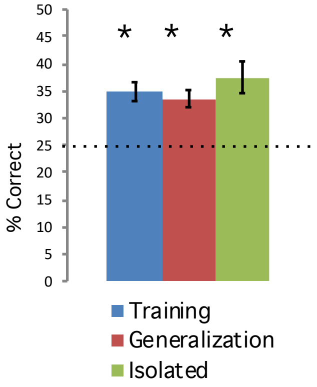
Exaggerated N100 associated with segmentation in continuous speech

(Sanders, Newport & Neville, 2002)

Wu, Liu, Lim & Holt, 2018



Post-Training Categorization Accuracy



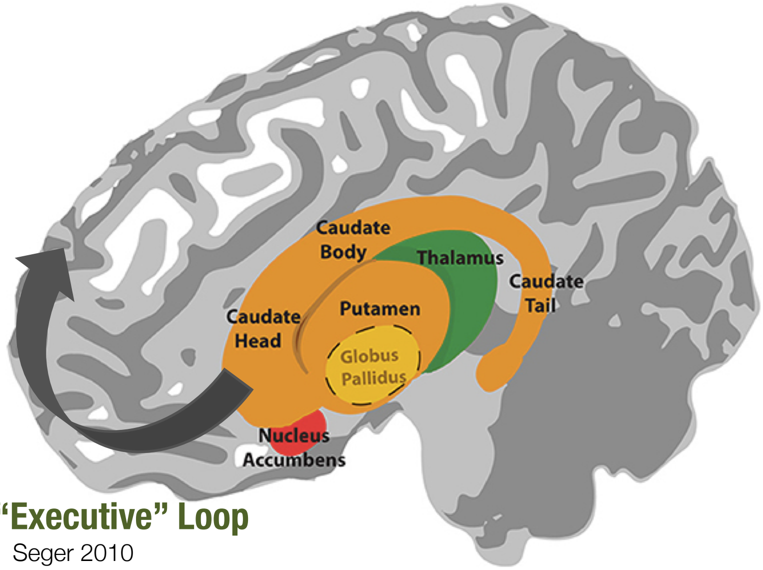
FOLLOW-UP STUDY: 6-9 days of training
One listener achieved 83%
Average accuracy = 51%

Lim, Lacerda, & Holt, 2015

How do listeners learn across unlabeled categories?

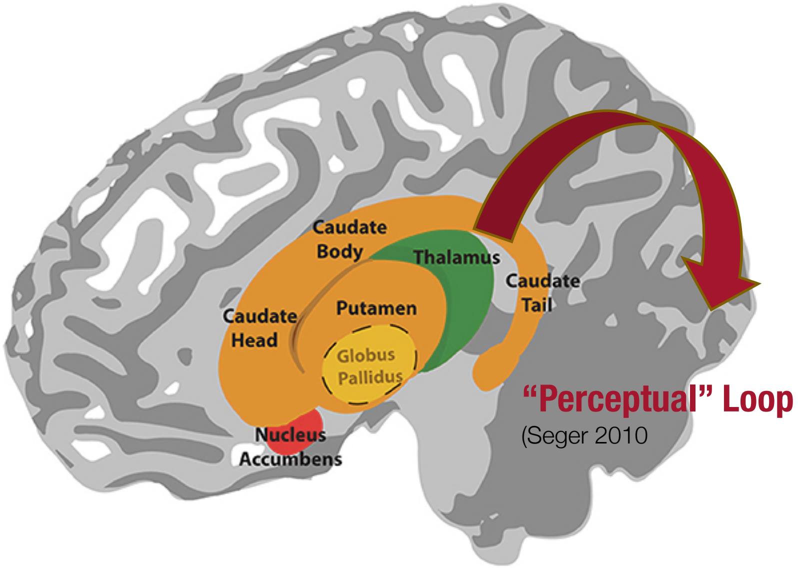
What is the form of this learning? Is this sensitivity unique to speech?

Is there intermediate ground between purely passive, unsupervised learning and instruction?



“Executive” Loop
Seger 2010

Stimulus – Response - Feedback
Tricomi et al. 2006



"Perceptual" Loop

(Seger 2010

Incidental Learning in Videogame

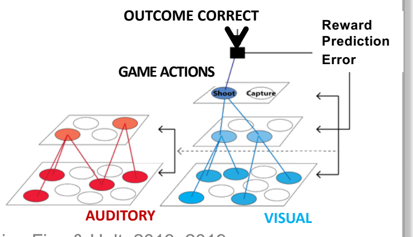
Lim, Fiez & Holt, 2013

Procedural Knowledge

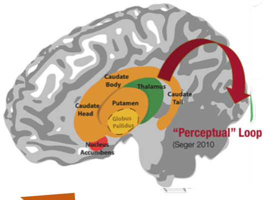
characterized by multiple, difficult-to-verbalize dimensions

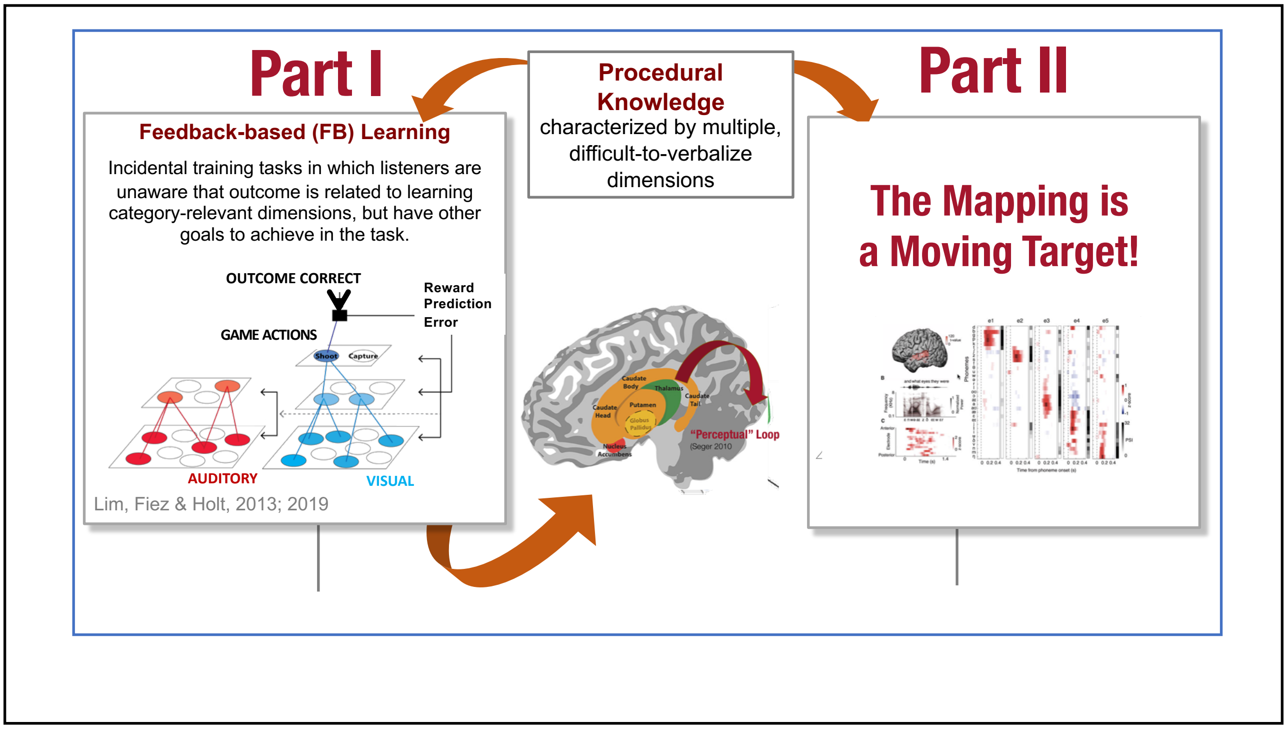
Feedback-based (FB) Learning

Incidental training tasks in which listeners are unaware that outcome is related to learning category-relevant dimensions, but have other goals to achieve in the task.



Lim, Fiez & Holt, 2013; 2019





Part I

Feedback-based (FB) Learning

Incidental training tasks in which listeners are unaware that outcome is related to learning category-relevant dimensions, but have other goals to achieve in the task.

Lim, Fiez & Holt, 2013; 2019

Procedural Knowledge characterized by multiple, difficult-to-verbalize dimensions

Part II

The Mapping is a Moving Target!

National Science Foundation WHERE DISCOVERIES BEGIN

NIDCD National Institute on Deafness and Other Communication Disorders
Improving the lives of people who have communication disorders

Riksbankens Jubileumsfond

Dr. Travis Wade
CTO, Posit Science

Wade & Holt, 2005

Dr. Sung-joo Lim
Boston University

Lim & Holt 2011
Lim, Lacerda, Holt, 2015
Lim, Fiez, Holt. 2013; 2019

Dr. Ran Liu
MaRI, Chief Scientific Officer

Liu & Holt, 2011

Dr. Kaori Idemaru
University of Oregon

Idemaru, Selman & Holt (2013)

Dr. Fred Dick
Birkbeck College,
University of London

Leech et al. 2009

Dr. Julie Fiez
University of Pittsburgh

Lim, Fiez, Holt. 2013; 2019

papers available at www.psy.cmu.edu/~holtlab/