Prerequisites for computational models of the early emergence of phonological categories

Mary E. Beckman & Andrew R. Plummer
(Ohio State University)
http://www.learningtotalk.org

Saffran, Newport, & Aslin (1996)

Training: 8-month-old babies exposed to just 2 minutes of spliced together syllables, such as [bidakupado …]

Test: looking time to 3-syllable “utterances” that either (a) were encountered (such as [bidaku]) or (b) were not (such as [bikudo])

Result: novelty effect for (b)

Authors’ interpretation: the babies could segment out and learn such long sequences, without support of word-level prosodic cues

Little statisticians?

How human infants might learn the components of language, and what recent research results mean in light of Noam Chomsky’s theories about language acquisition, are debated. And readers continue to express their concern about “misplaced” crabs, along with a plant that is not a grass.

MICHAEL AND JEANETTE TWA

Bates & Elman (1996) on Saffran et al.

Interpretation: "This result means that infants can use simple statistics to discover word boundaries in connected speech, right at the age when systematic evidence of word recognition starts to appear in real life."

Significance: result "flies in the face of received wisdom"

(1) It is "surprising [that] a purely inductive, statistically driven process, based on only 2 min of incidental input, with no reward or punishment other than the pleasure of listening to a disembodied human voice."

(2) It "contradicts the widespread belief that humans cannot and do not use generalized statistical procedures to acquire language."

Bates & Elman (1996) conclude ...

"Although we now know that linguistic regularities are learnable by neural networks with an imperfect but very large database, it has been argued that human infants do not learn in this way, and even if they did, their memory and attention span are insufficient to support the kind of statistical learning required to get language off the ground. This conclusion was premature. The new work has shown that infants are capable of extracting statistical regularities from only 2 min of spoken input with little effort."

Pinker (1997) rebuts Bates & Elman

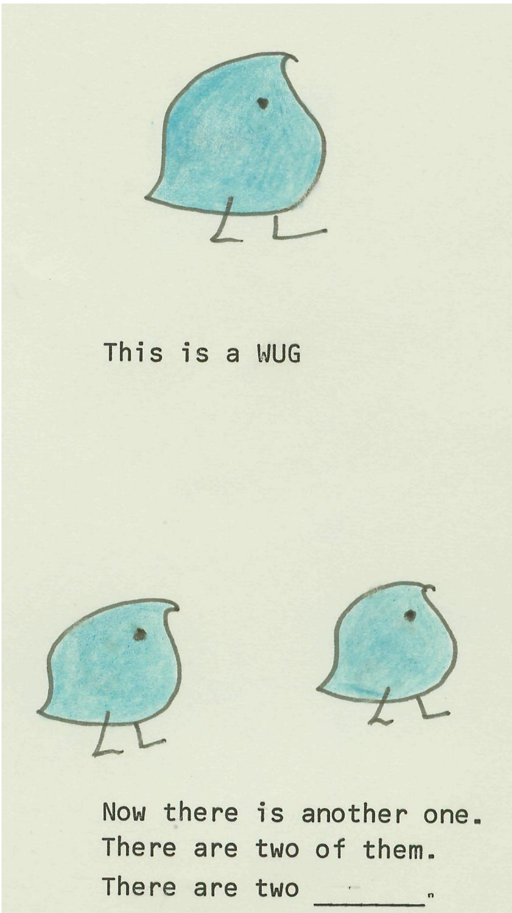
‘Bates and Elman suggest that if children can learn words by recording frequent sound sequences, they might learn grammar the same way. But words and grammar are different. The sequence of sounds making up a word is not capturable by rules (“monkey” cannot be understood as a combination of “mon” and “key”), but must be memorized. And because there are a finite number of words, they all can be recorded.

The sequence of words making up a sentence, however, is capturable by rules. (For example, “the eggplant ate Chicago,” though an improbable word sequence, can be understood from the meanings of “eggplant,” “ate,” and “Chicago” and the way they are combined). Word sequences need not and cannot be memorized, because they form an open-ended set.’

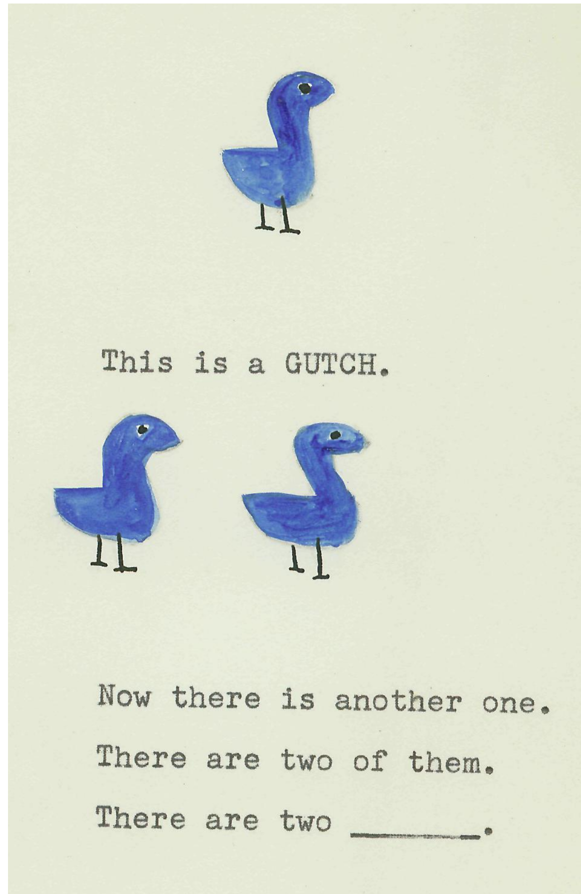
Plan for this talk

- Outline a more realistic model of the “grammar of words”
  - Review evidence that phonology has a “grammar” that makes the word *monkey* more than just the sequence of sounds [m]+[ʌ]+[ŋ]+[k]+[i]
  - Review evidence that phonological categories such as the “sounds” [m], [ʌ], [ŋ], [k], and [i] in the word *monkey* themselves are very complex grammatical abstractions
  - Review evidence regarding earliest “learning” of them
- Review (the unwarranted assumptions underlying) two classes of models of how these categories can be learned
- Outline steps we are proposing to take to overcome these assumptions and show some preliminary modeling results

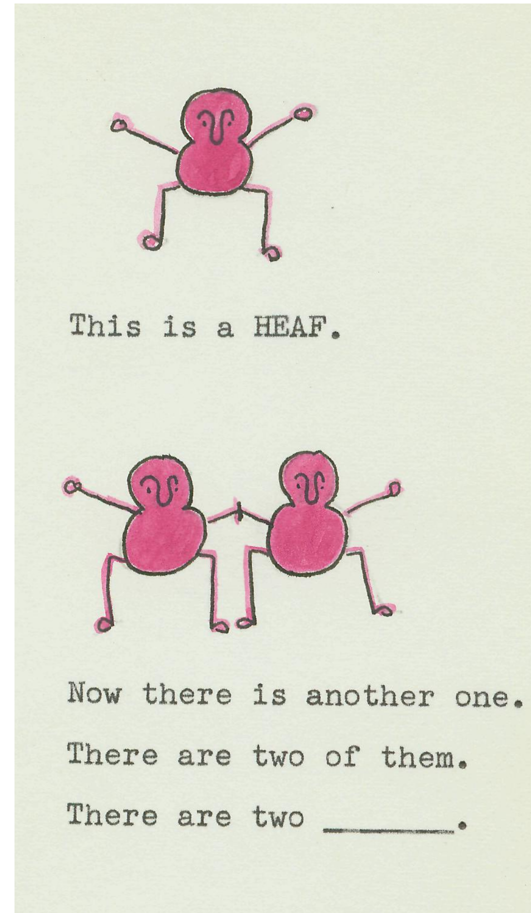
The grammar of sound sequences



[wʌgz]



[gʌʧɪz]
OCP epenthesis



[hifs]
voicing harmony

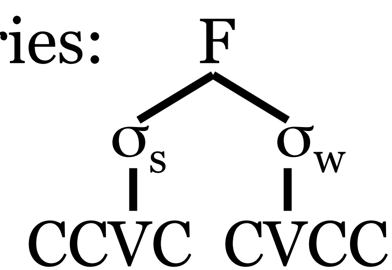
Productive "rules" for cut/join points

English blends cut and join at stressed C-V boundary:

smoke /smok/ + *fog* /fɔg/ = *smog* /smɔg/

breakfast /bɹɛkfəst/ + *lunch* /lʌntʃ/ = *brunch* /bɹʌntʃ/

The "rule" references structural categories:



Productively applies to nonwords (Treiman, 1995):

/hɹk/ + /jɪg/ → /hɪg/ favored 43:1 over /hɹg/

Preference for cut point modulated by a sensitivity to phonotactic probability of VC sequence (Treiman, Kessler, Knewasser, Tincoff, & Bowman, 2000):

/hɹp/ + /jɪʤ/ → /hɪʤ/ favored 19 to 1 over /hɹʤ/

Phonotactics and duality of patterning

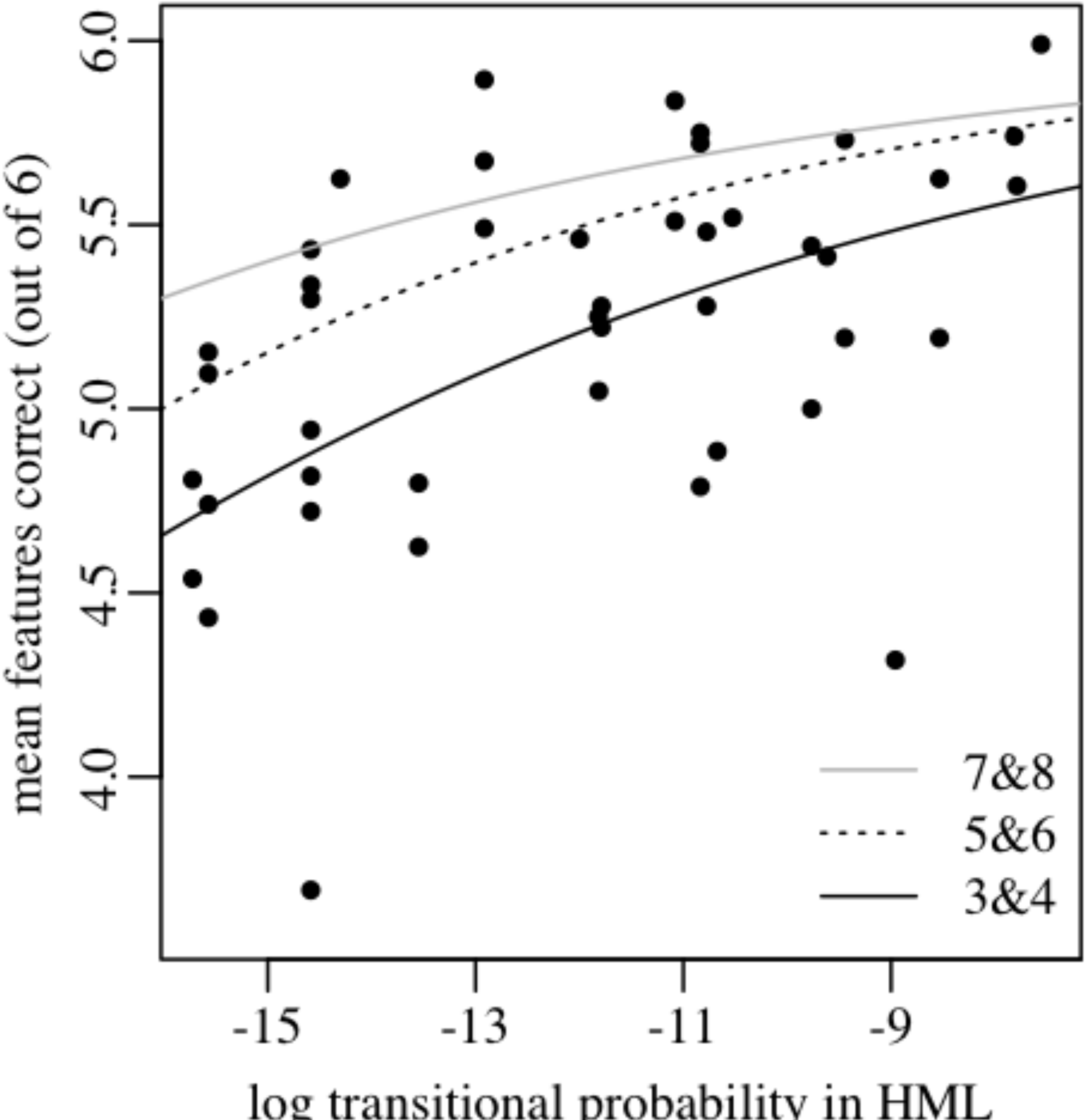
- The (recordable) real word *brick* and nonword **blick* are both better than **bnick* (Chomsky & Halle, 1965)

Judgments are continuous and related to type frequency

- Adult speakers judge nonsense words containing phoneme sequences that occur in many words of English to be "more like a word of English" (Coleman & Pierrehumbert, 1997; Vitevich, Luce, Charles-Luce, & Kemmerer, 1997; many others)

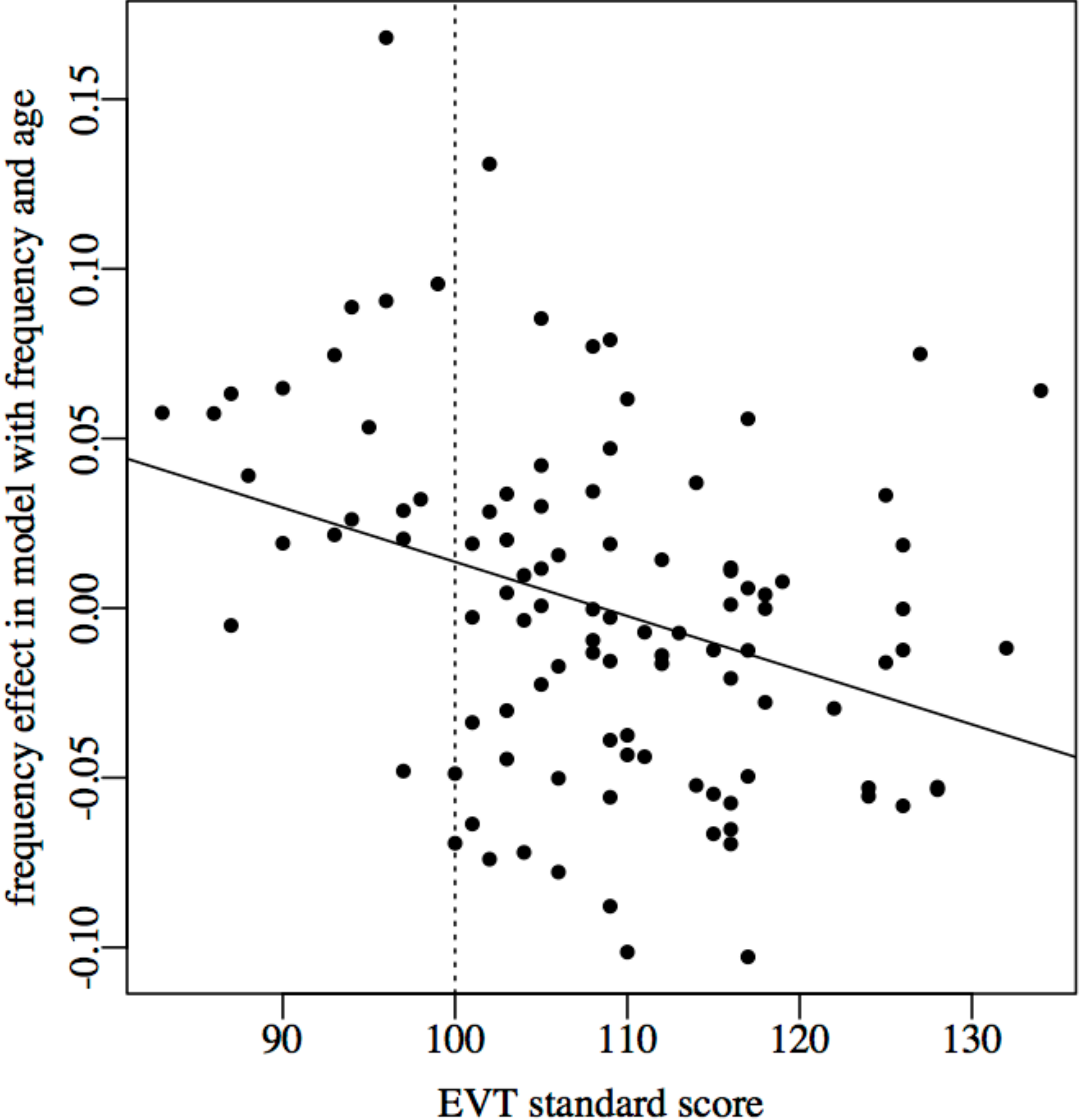
Knowing more words allows for a richer grammar:

- Cut-off probability for forms judged to be absolutely bad varies from speaker to speaker, and is correlated with the size of the speaker's lexicon (Frisch, Large, Zawaydeh, & Pisoni (2001)



Edwards, Beckman, and Munson (2004)

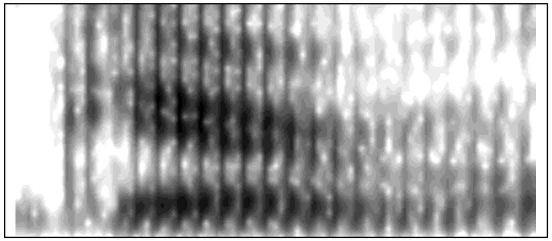
Accuracy of production of consonant and vowel sound sequences in a nonword repetition task related to both (1) child's age & (2) phonotactic probability of the sound sequence



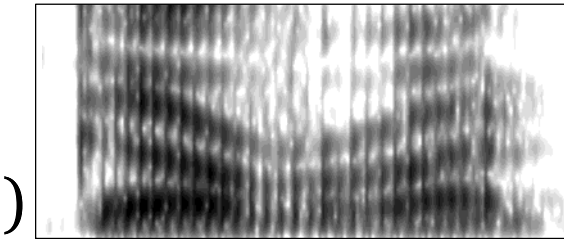
No effect of age, once expressive vocabulary size included in model.
And size of phonotactic probability effect (here individual-level slope in mixed effects model) is related to vocabulary size.

Lower-level abstractions over tokens

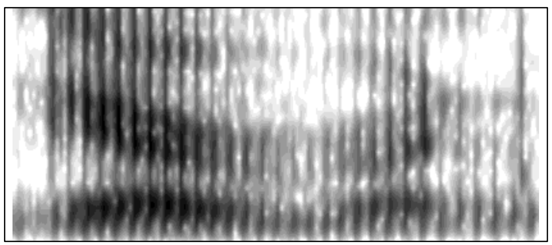
s3c
(202 ms)



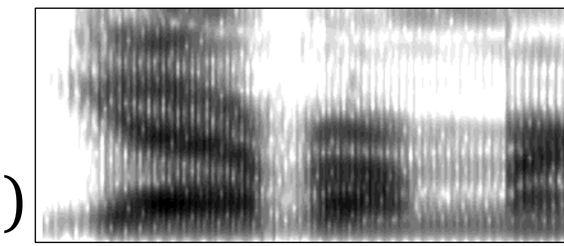
s2b
(347 ms)



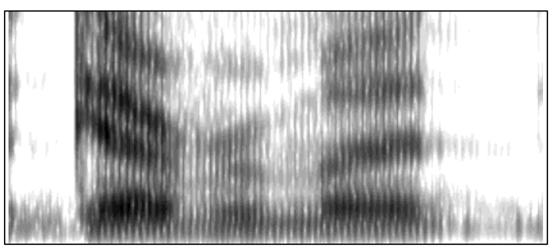
s3b
(281 ms)



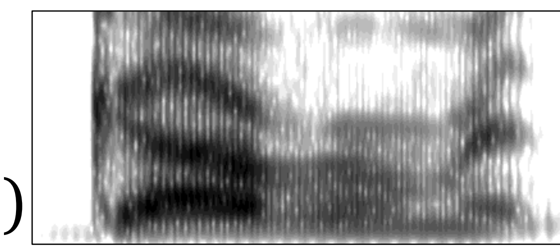
s1b
(374 ms)



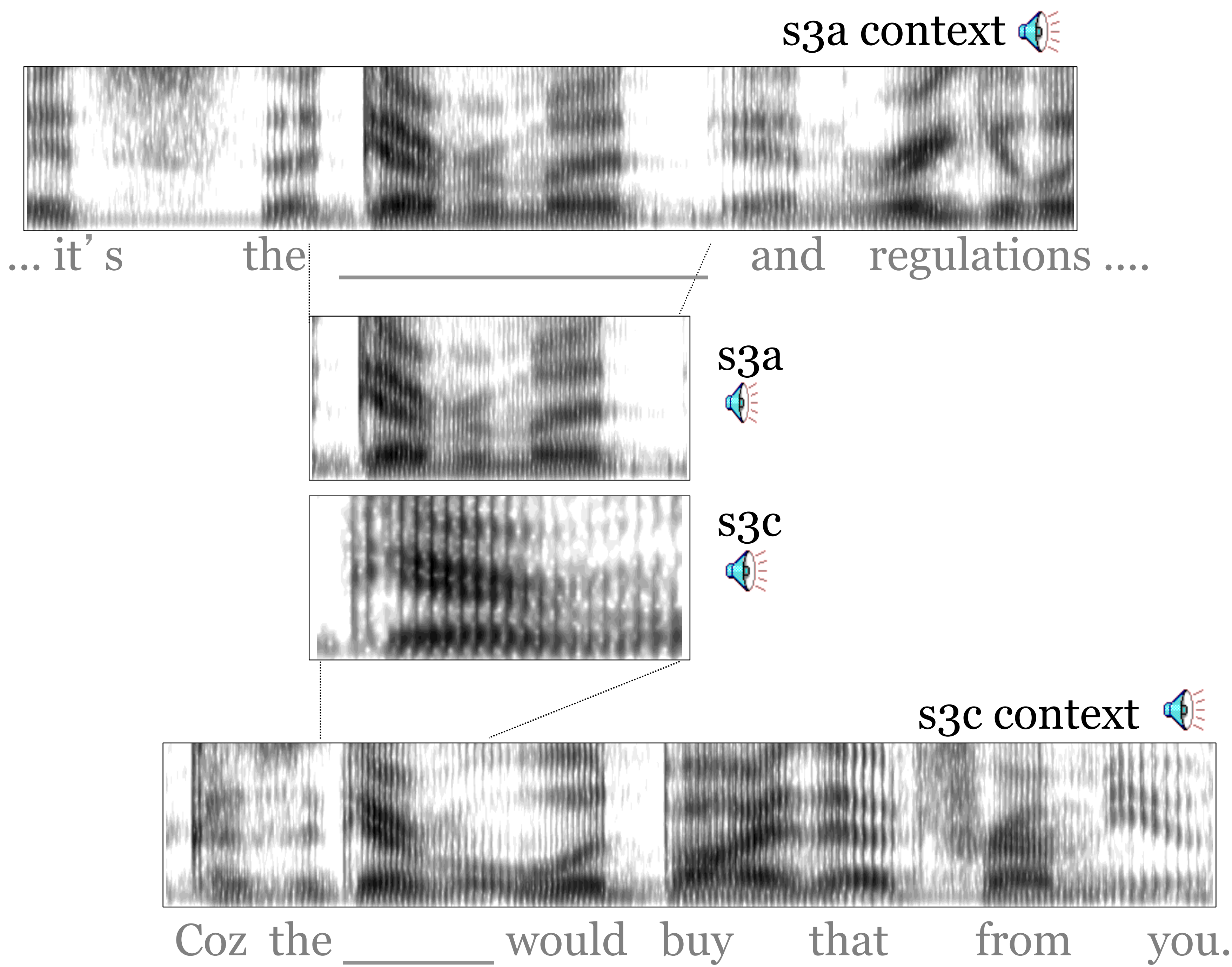
s3a
(563 ms)



s4a
(393 ms)

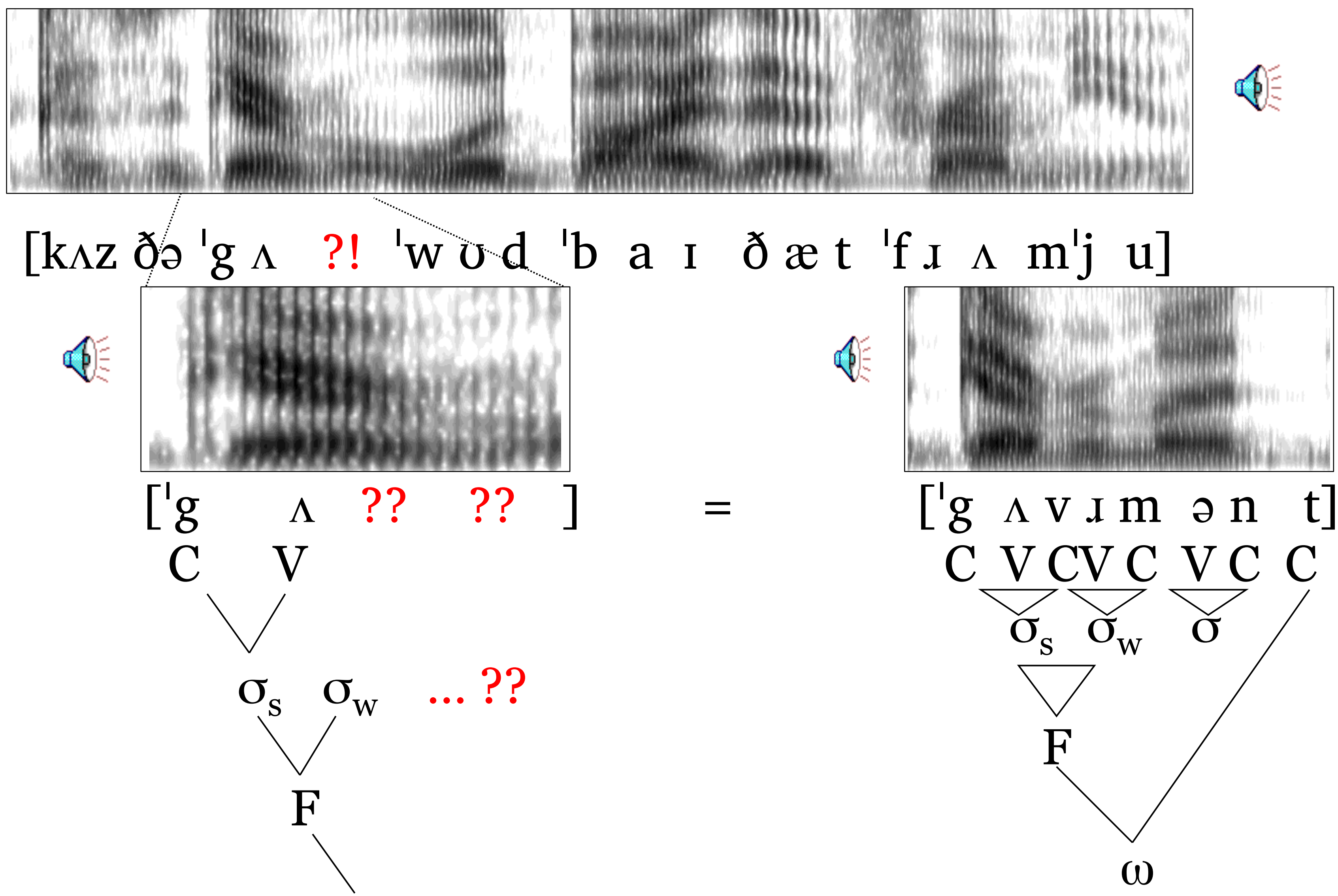


Kenyon & Knott (1951): GA [ˈgʌvɹmənt] ~ [ˈgʌvɹnmənt]; SE (includes Kentucky) [ˈgʌvəmənt] ~ [ˈgʌvənmənt] ~ [ˈgʌvm̩ənt]



s3a context
... it' s the ______ and regulations ....
s3a
s3c
s3c context
Coz the _____ would buy that from you.

Top-down parse from islands of reliability



Head turn preference paradigm

Child sits on parent's lap and orients toward "interesting" sounds

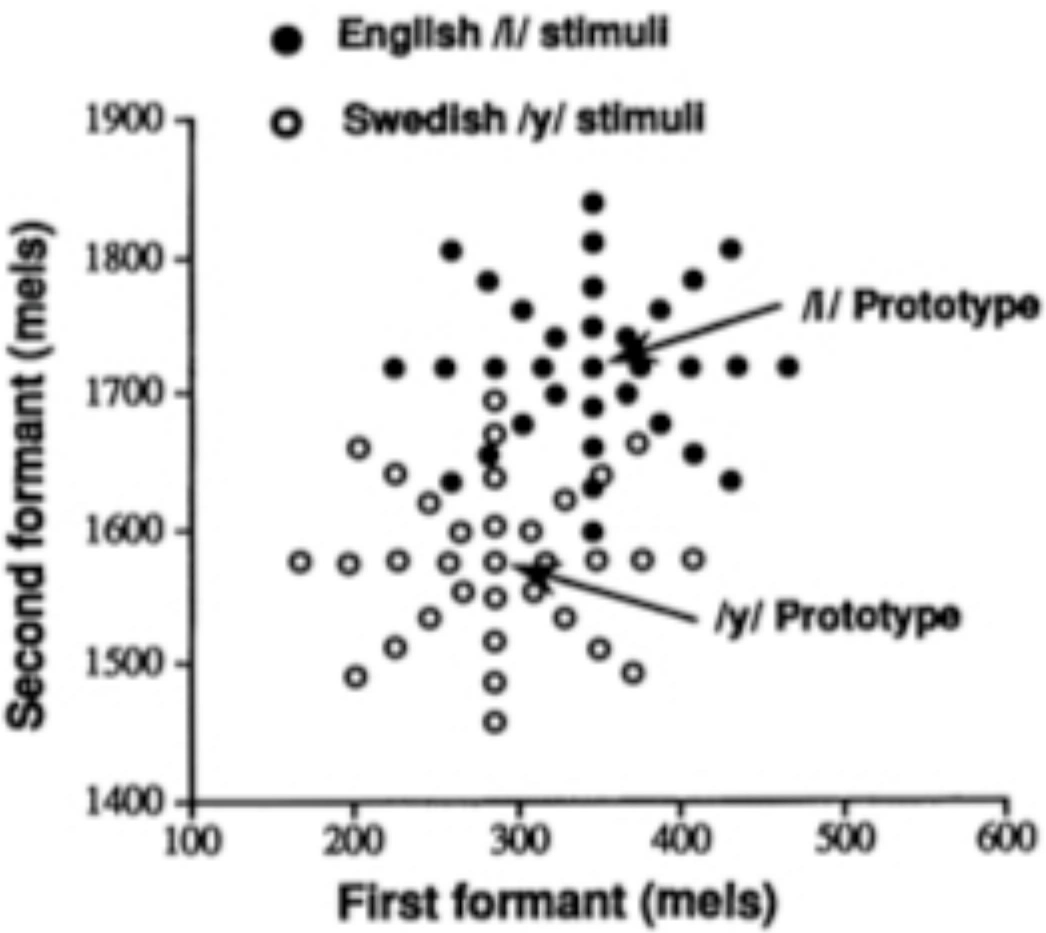
<u>Measure</u>: looking time

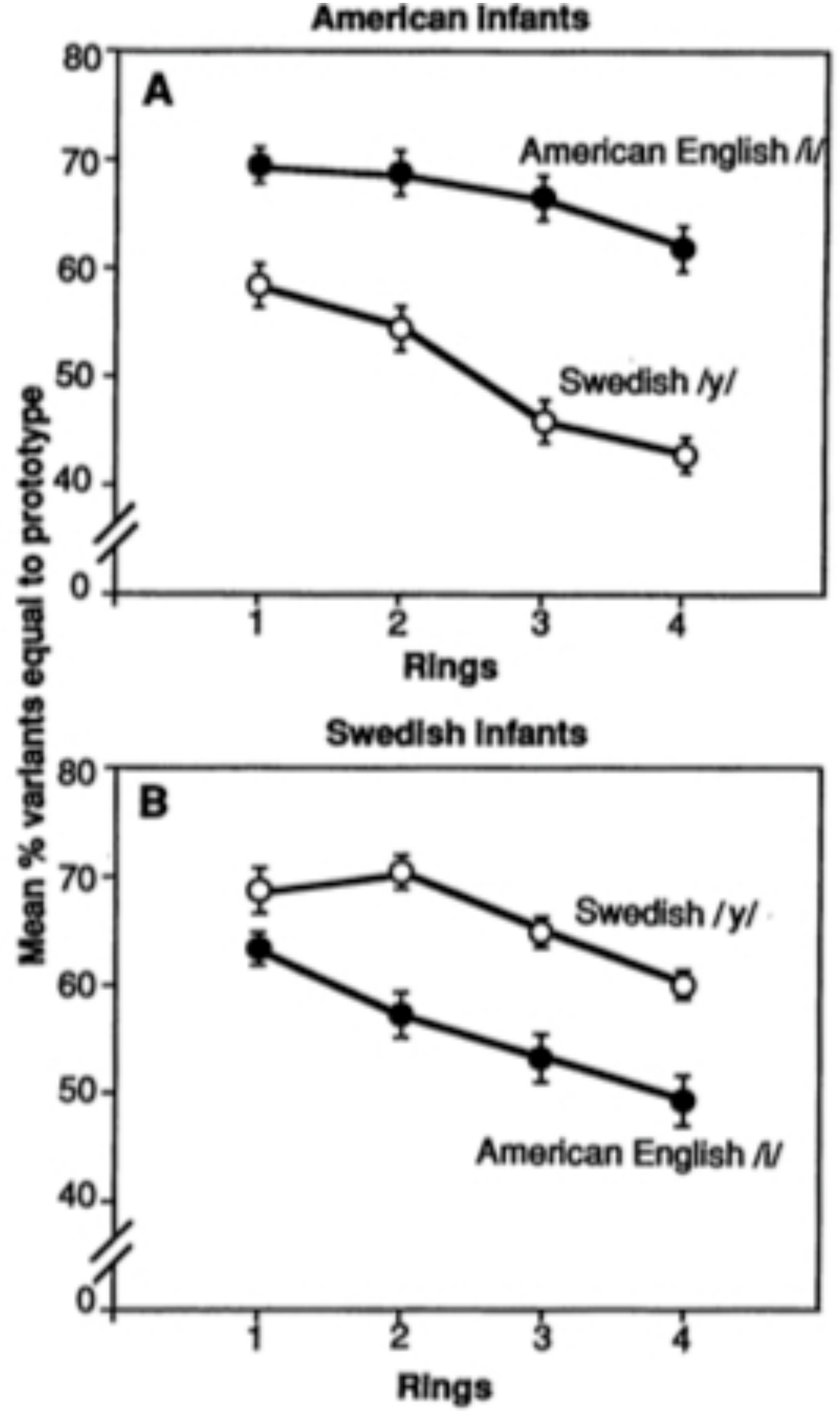
A common set up for this paradigm: Sounds played on left or right

Early evidence for abstraction over tokens

- At 6 months, both English- and Japanese-learning infants differentiate both [ni:kusu] and [ni:k] from [ni:ks], but after 12 months Japanese-learning infants no longer differentiate [ni:kusu] from [ni:ks] (Kajikawa, Fais, Mugitani, Werker, & Amano, 2006)
- At 6, 12, and 18 months, Japanese-learning infants fail to differentiate [ki:t] from [ki:ts], although starting at 12 months they differentiate [ki:tsu] from [ki:ts], (Mugitani, Fais, Kajikawa, Werker, & Amano, 2007)
- At 6, 12, and 18 months, English-learning infants differentiate [ni:k] and [ni:ks] and [ki:t] from [ki:ts] (Fais, Kajikawa, Amano, & Werker, 2009)

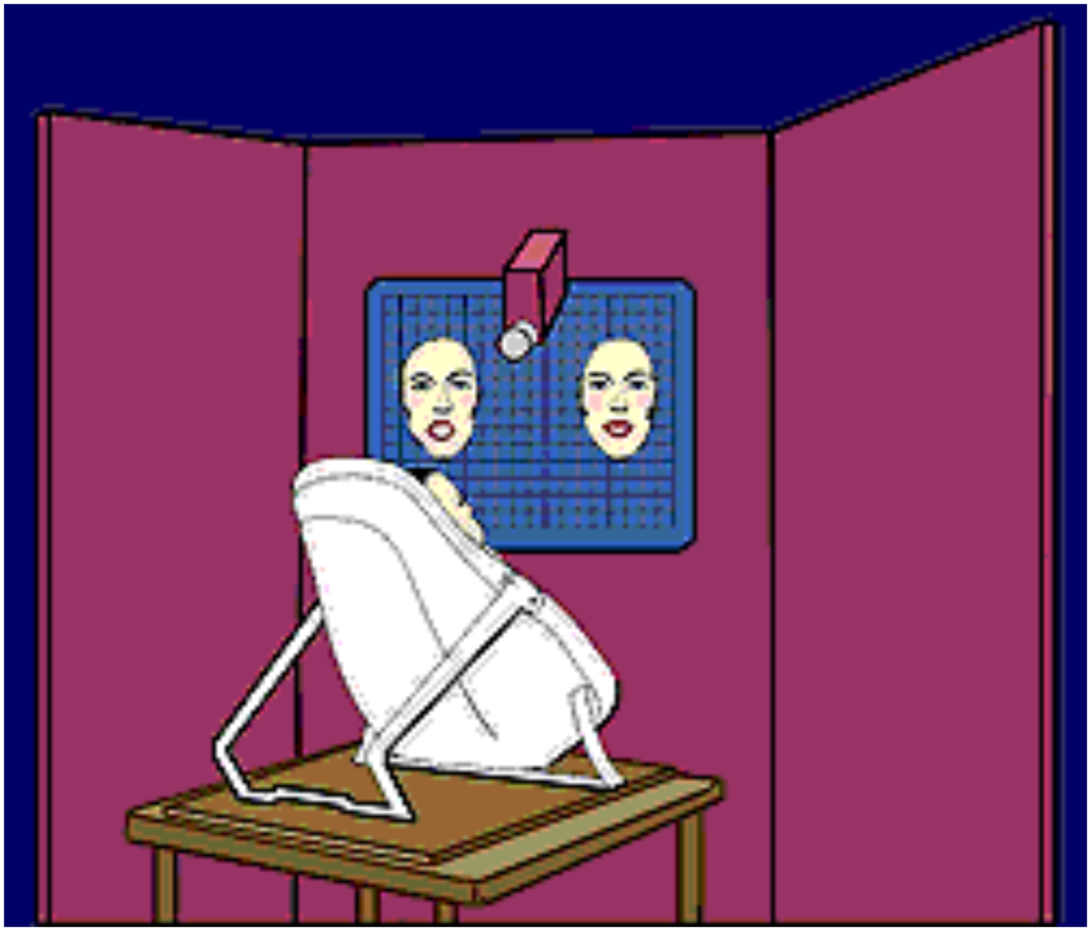
Earlier evidence of abstraction for vowels





Kuhl, Williams, Lacerda, Stevens, & Lindblom (1992) Figs 1-2

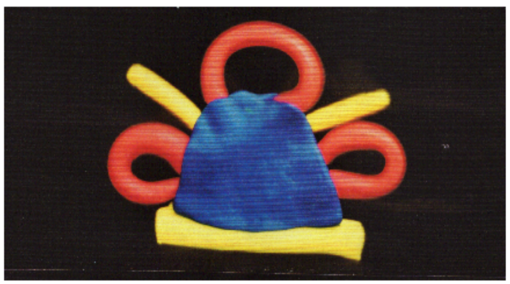
Even earlier evidence for vowel categories



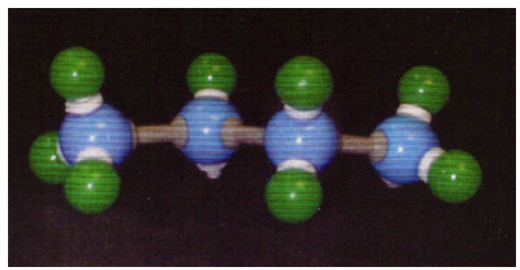
- Infant looks longer at the face that matches /a/ or /i/ being played over the loud speaker (Kuhl & Meltzoff, 1982)
- Listeners judge the infant's coos as more like the vowel that the infant watches (Kuhl & Meltzoff, 1996)

Top-down processing as vocabulary grows

- At 14 months, infants no longer able to differentiate fine ambient-language contrasts such as /bɪ/ versus /dɪ/ when associated with pictured objects (Werker & Stager, 2000)



(a) Pokey



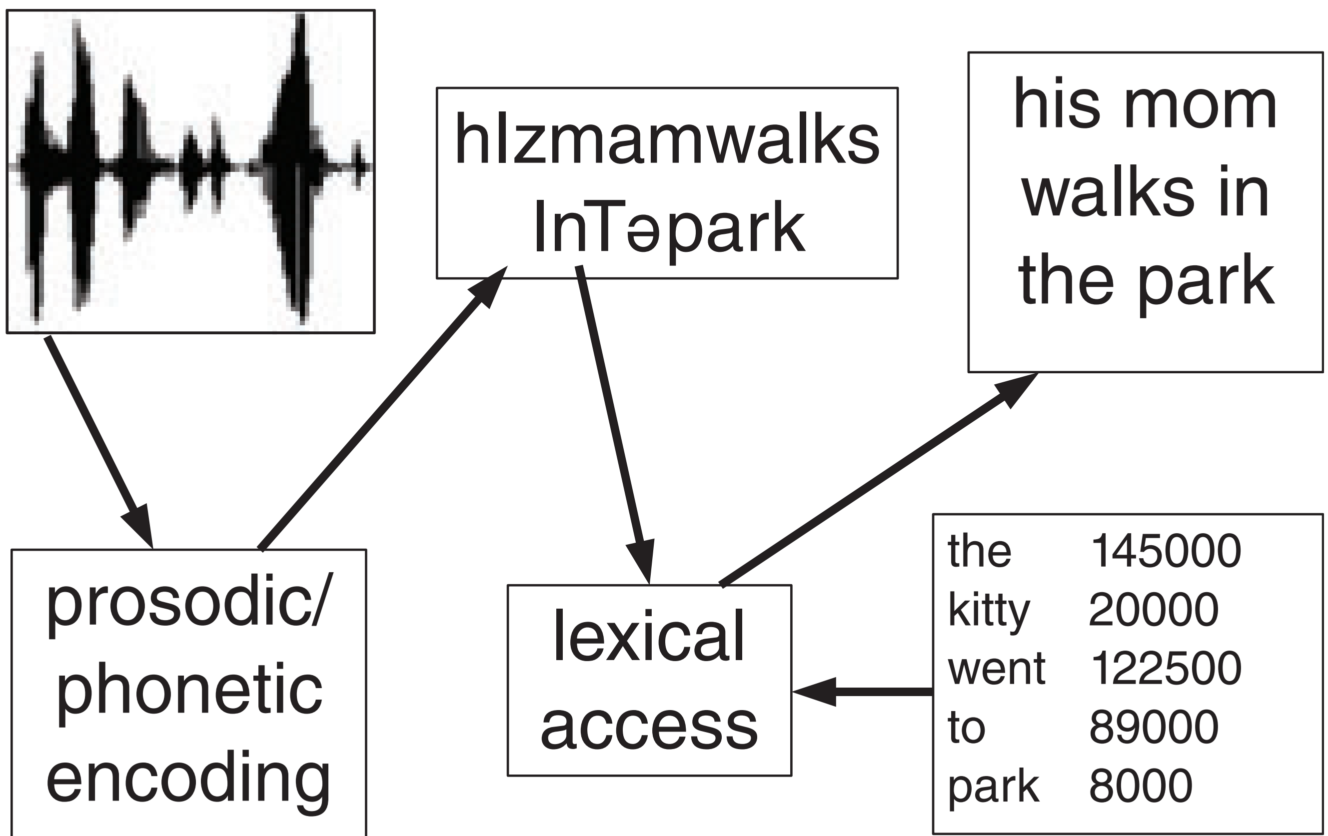
(b) Molecule

(c) Water Wheel

- Recovery of ability to discriminate by 22 months, correlated with the size of the child's vocabulary (Werker, Corcoran, Fennell, & Stager, 2000)

Model of phonotactics, word segmentation

- Daland & Pierrehumbert (2011) train Bayesian learner to segment phonetic transcriptions of running speech by optimizing the resulting lexicon.



Other similar models include …

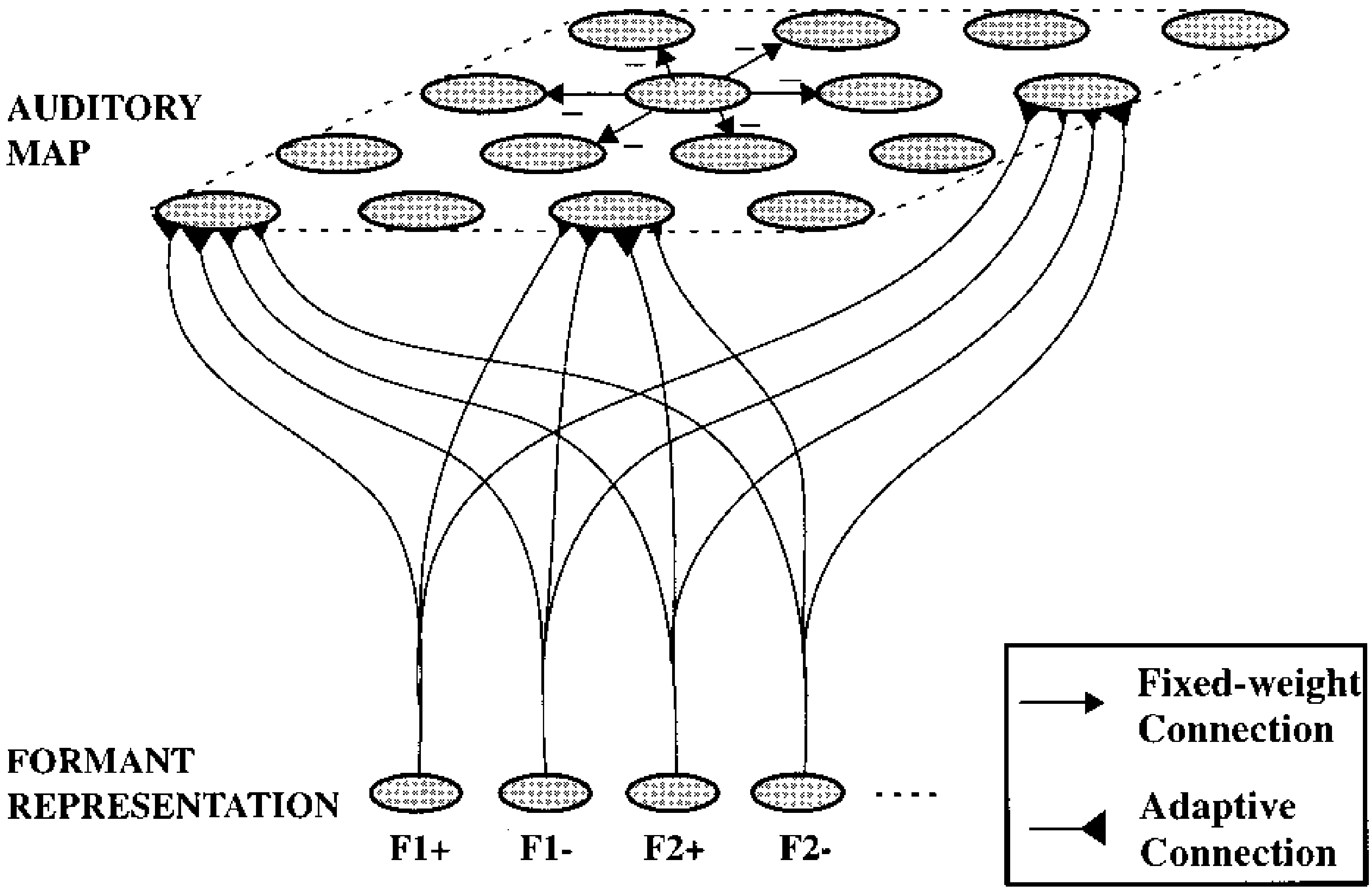
Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61: 93-125.

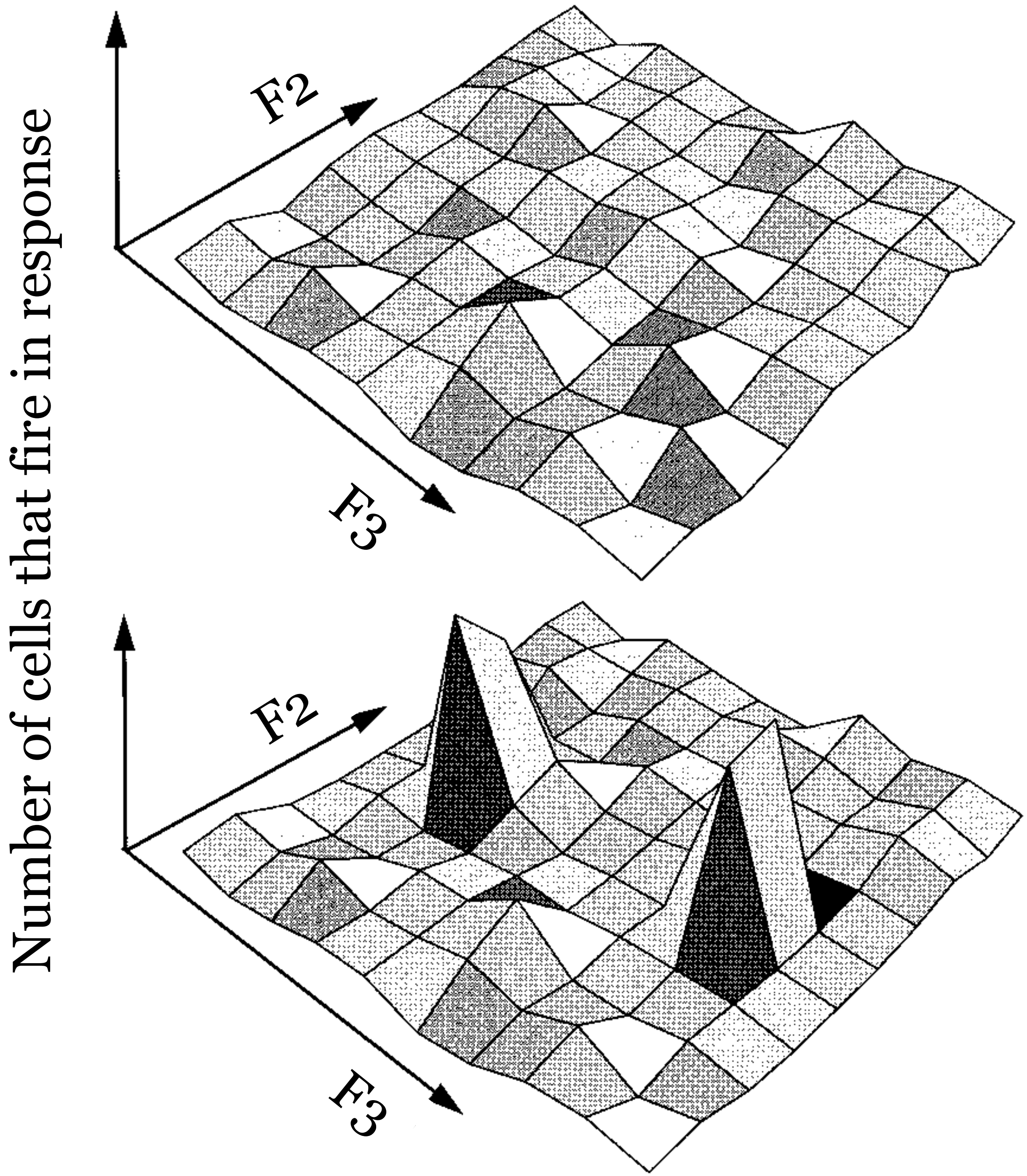
Cairns, P., Shillcock, R. C., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology*, 33: 111-153.

Elsner, M., Goldwater, S., & Eisenstein, J. (2012). Bootstrapping a Unified Model of Lexical and Phonetic Acquisition. *Proceedings of the Association for Computational Linguistics*.

☞ All assume infant has consonant and vowel segments.

Guenther & Gjaja (1996) Kohonen map



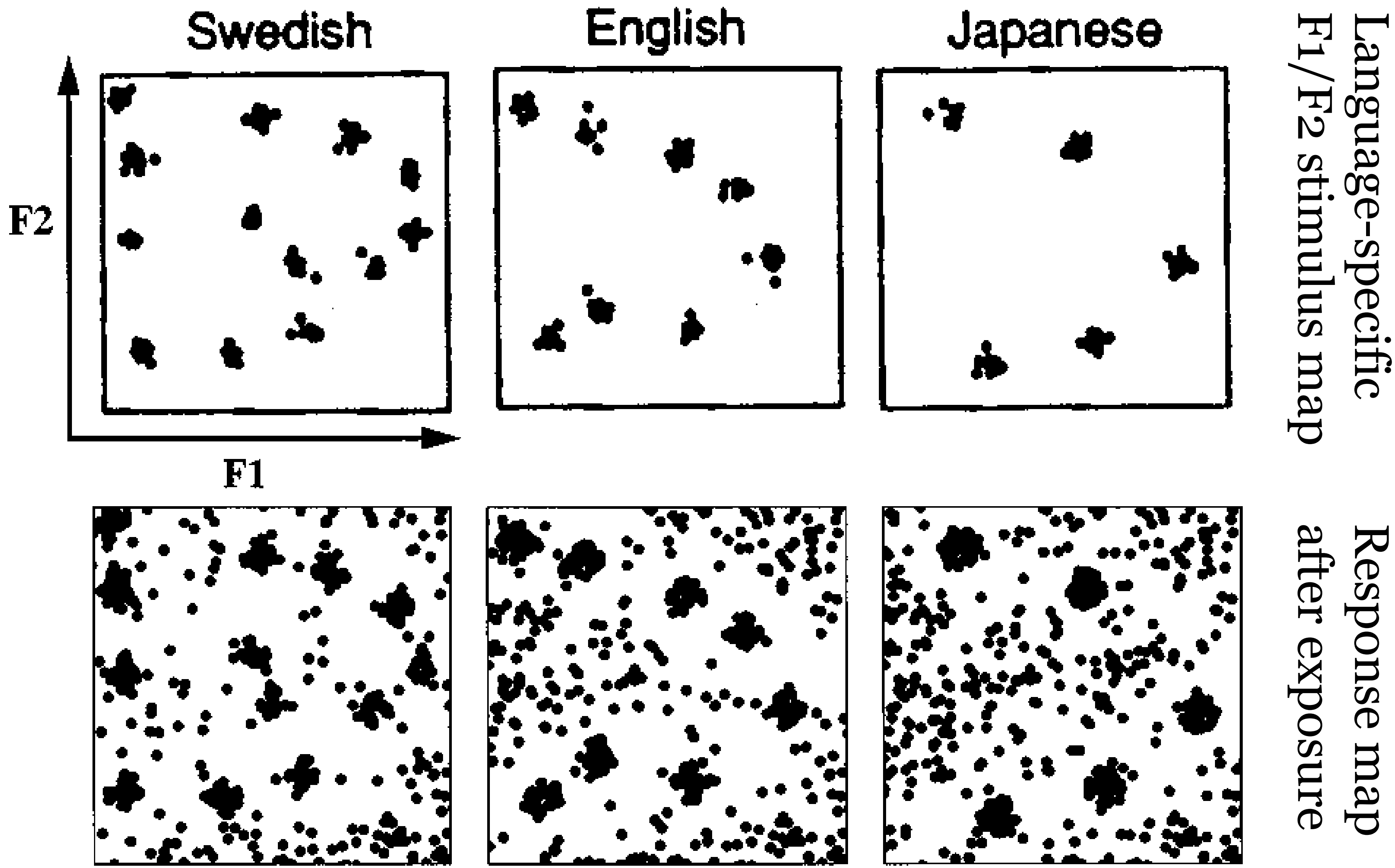


Guenther & Gjaja (1996) Fig. 3.

Top panel: Distribution of the preferred stimuli of auditory map cells over F2/F3 space before training

Bottom panel: Distribution after training with American English /r/ & /l/ inputs

Perceptual magnet effect (G&J’96 Fig. 4)



Other similar models include …

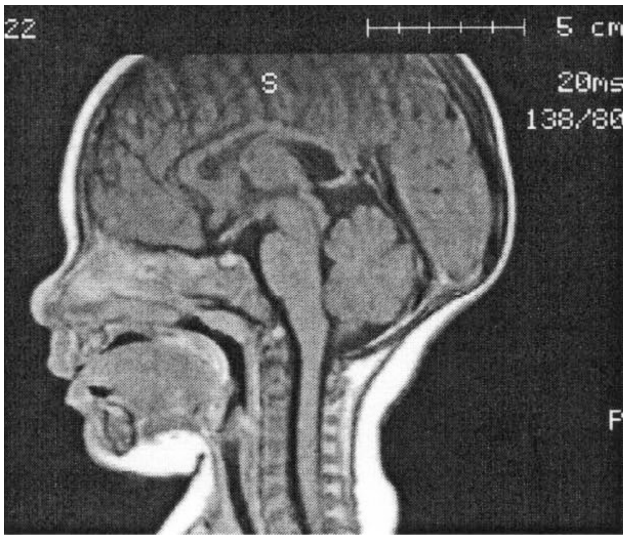
Vallabha, V. G., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, 104: 13273-13278.

Westermann, G., & Miranda, E. R. (2004). A new model of sensorimotor coupling in the development of speech. *Brain and Language*, 89: 394-400.

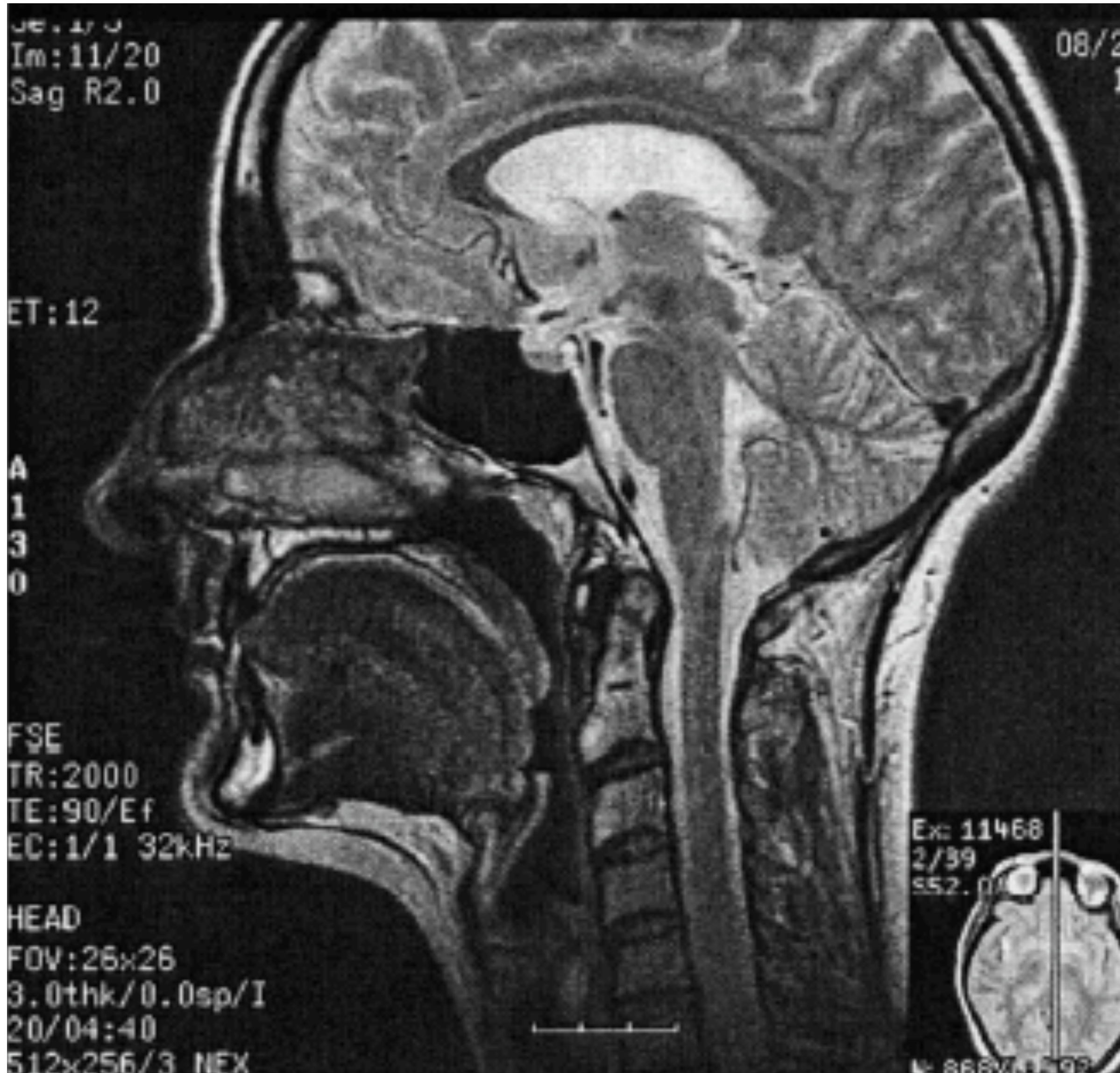
☞ All assume that abstraction over any given sensory space occurs directly in the neural representation of that sensory space, and that an auditory space for adult vowel productions maps in a straightforward way onto the auditory sensory space for the infant's productions.

Vorperian, Kent, Lindstrom, Kalina, Gentry, and Yandell (2005)

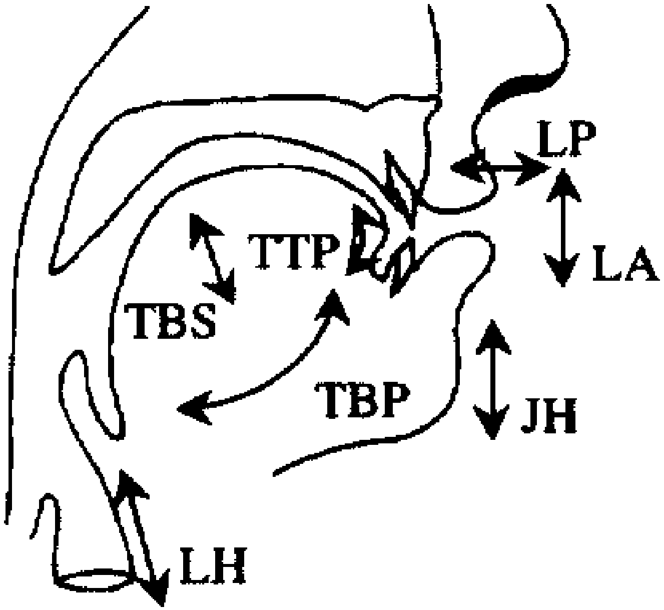
MRI of 7-month-old (left) & adult females (right)



Infant’s vocal tract is ½ the adult’s length, w/ pharynx much shorter

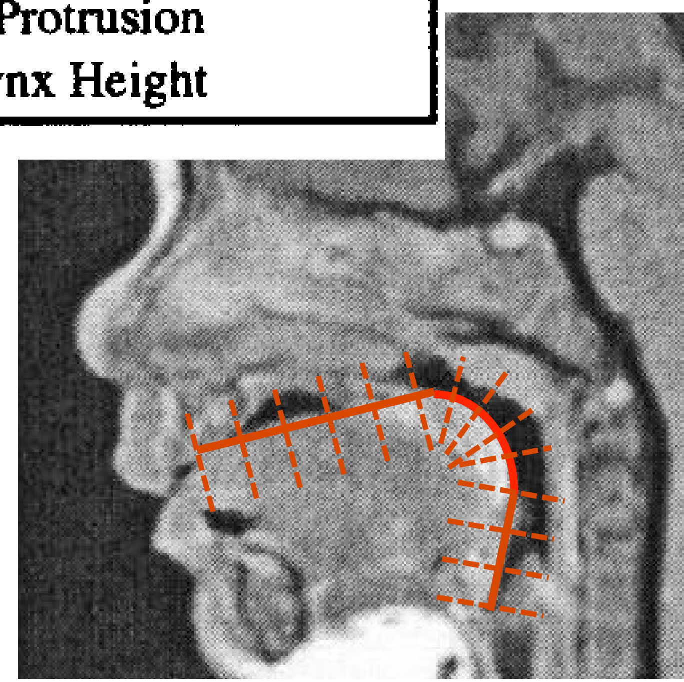


Callan, Kent, Guenther, Vorperian (2000)

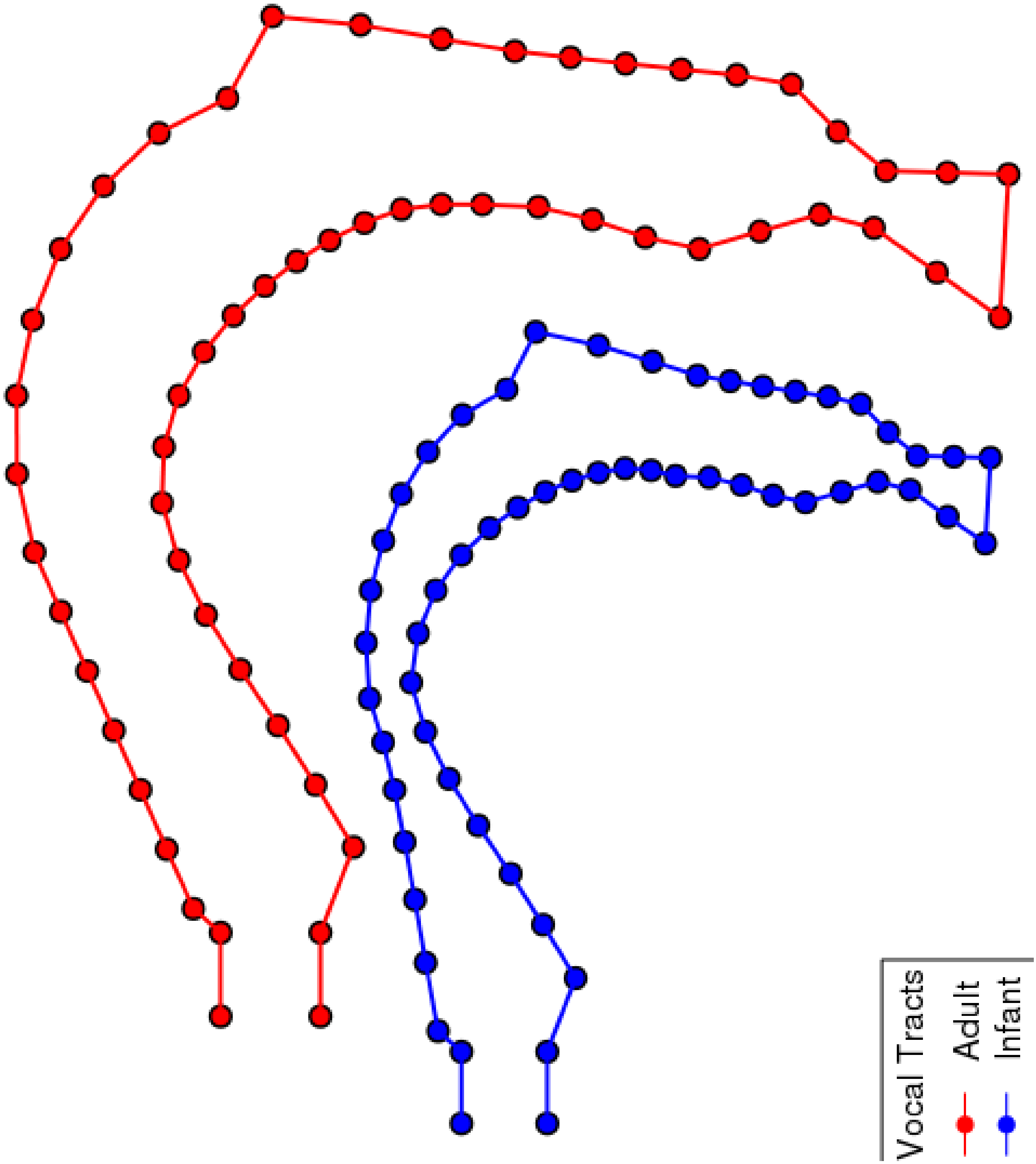


Maeda Articulation Parameters	
JH	Jaw Height
TBP	Tongue Body Position
TBS	Tongue Body Shape
TTP	Tongue Tip Position
LA	Lip Aperture
LP	Lip Protrusion
LH	Larynx Height

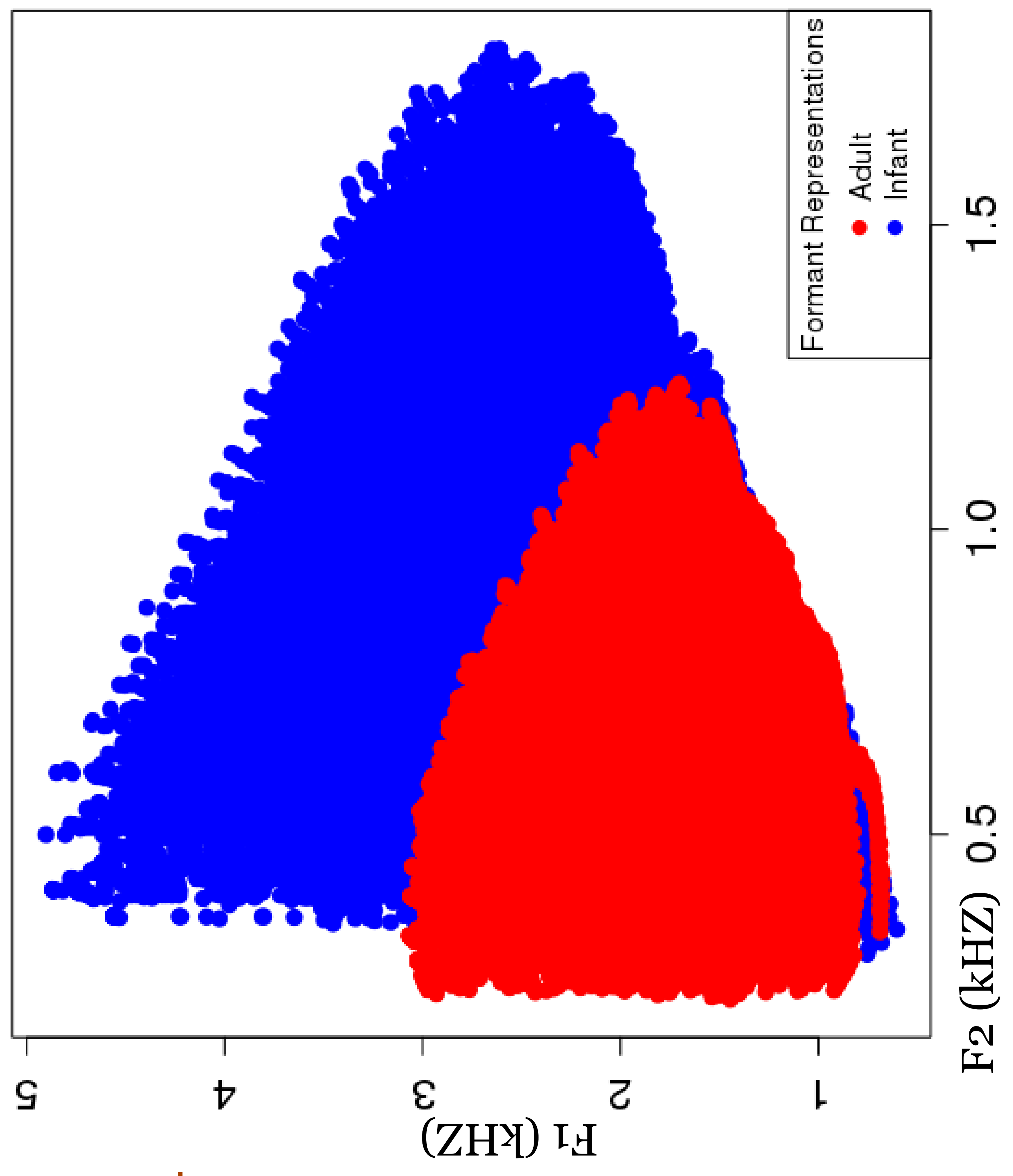
Build articulatory synthesis models from MRIs for 3-, 7-, 15- 24-, 36- and 45-month old infants, based on VTcalcs – i.e. Maeda (1990) PCA-based adult vocal tract model.



Plummer (2012) simulates infant's and mother's vocal tracts using Boë & Maeda's (1997) Variable Linear Articulatory Model



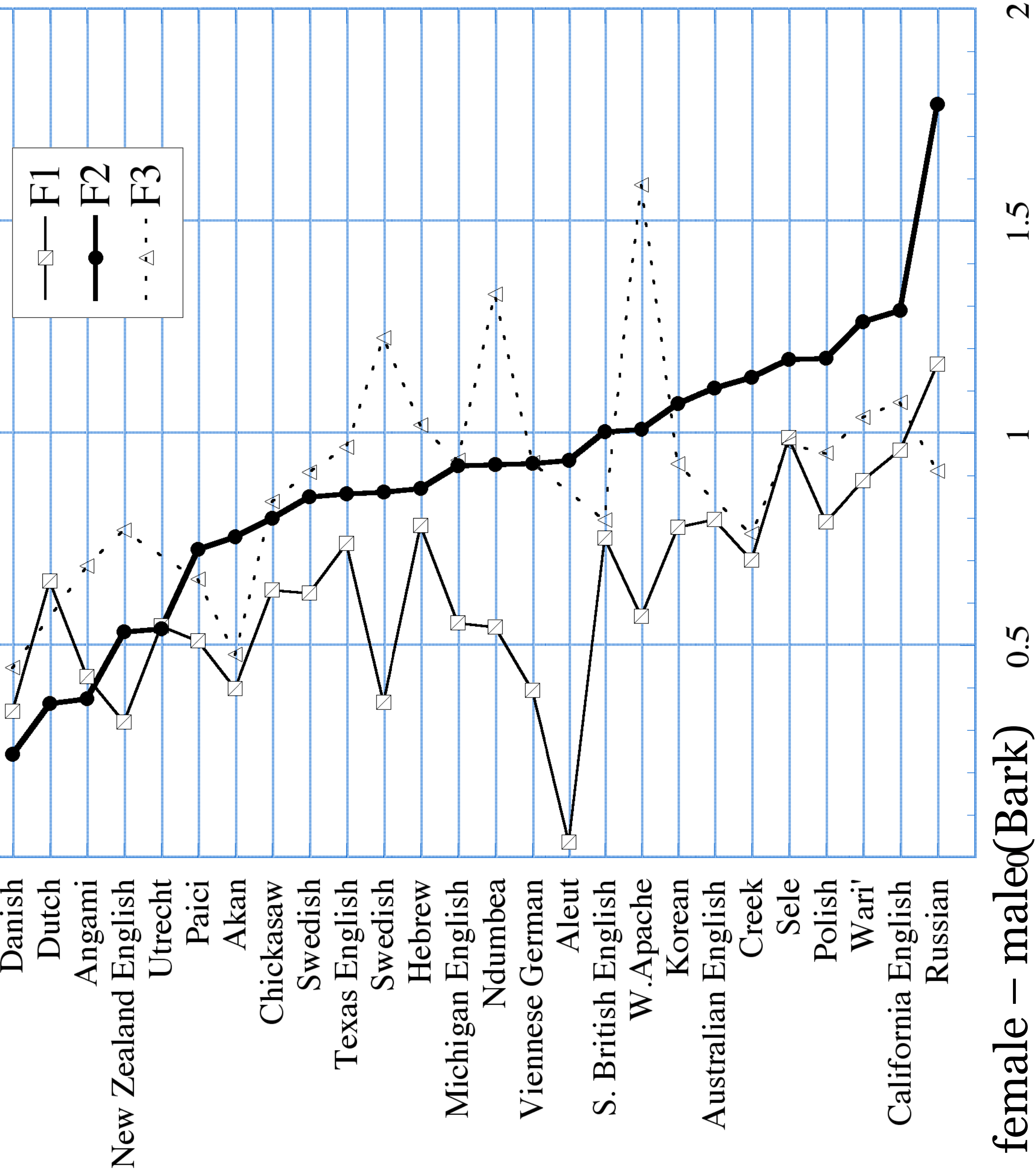
Infant's and mother's maximal vowel spaces generated by VLAM (Plummer, 2012)

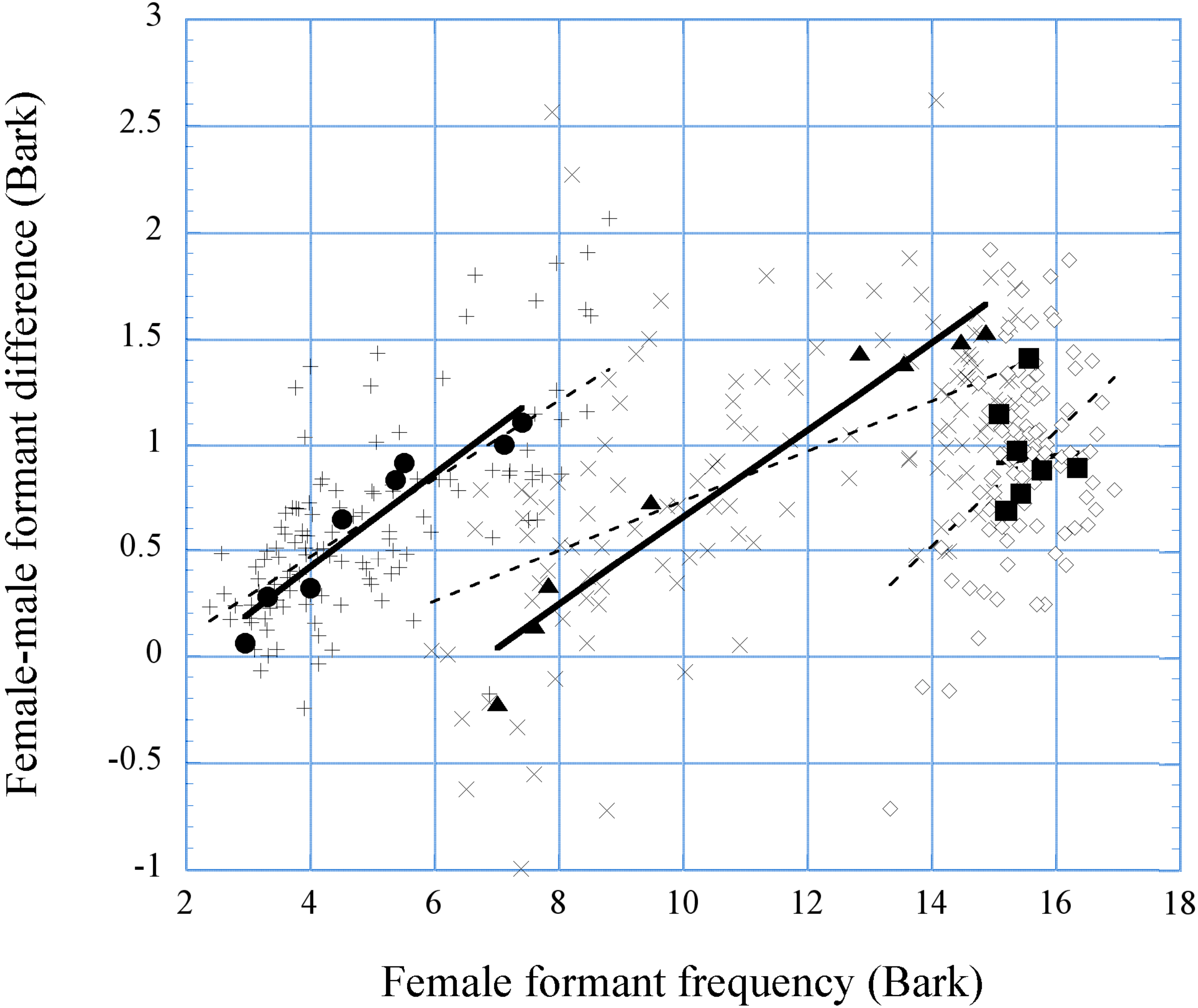


Callan et al. (2007) neural net modeling

- Used the modified VTcalcs models to explore what formant patterns can be produced at which ages.
- Showed that different articulatory configurations needed at different age to make F-pattern that is appropriate for each of the American English vowels.
- Trained neural net to build age-specific mappings from articulatory patterns to formant patterns.
- Simulated learning and subsequent adaptation of vowel categories by supervised learning using labeled vowel regions projected from adult auditory vowel space map directly onto the infant or child auditory sensory space.

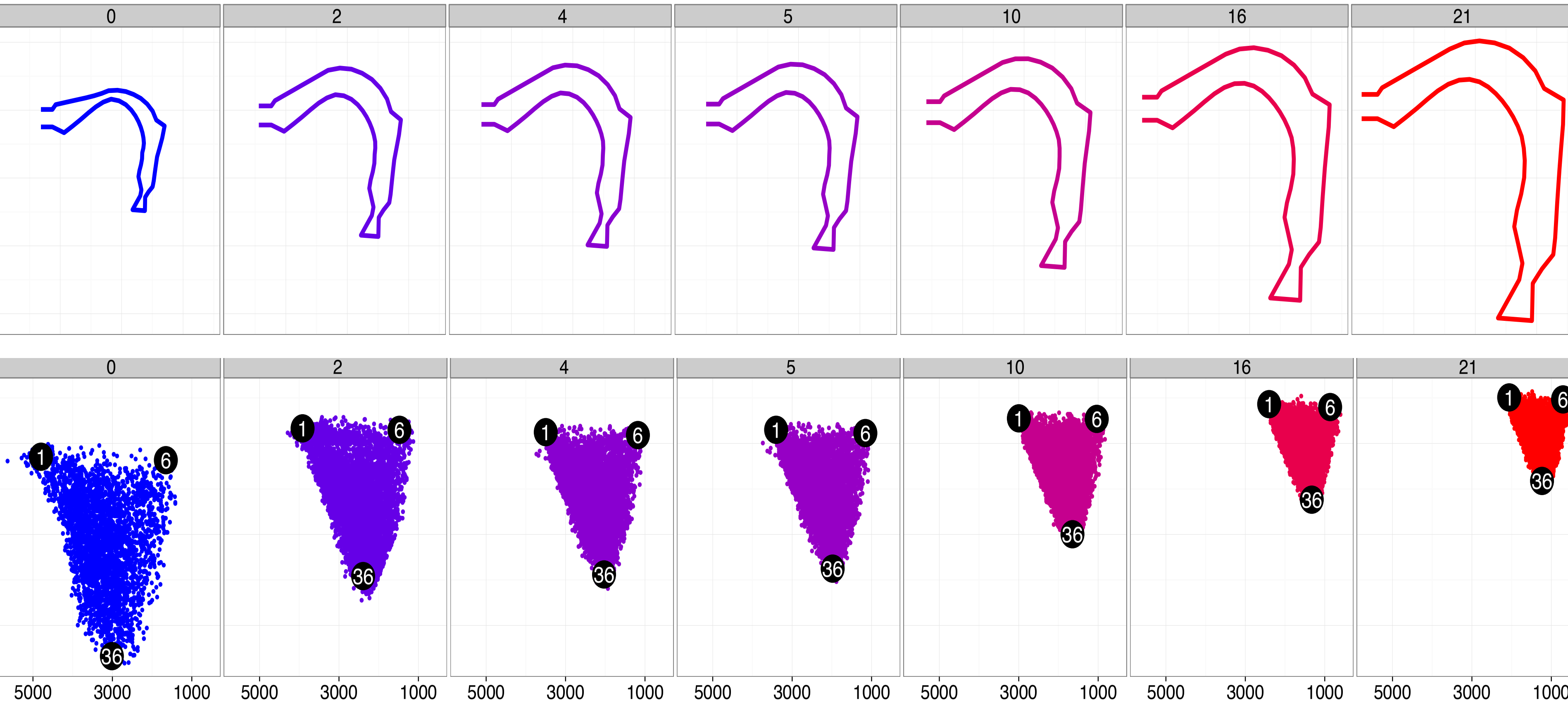
Evidence that speaker normalization is learned includes cultuer-specific differences in formant values between men and women (Johnson, 2005)





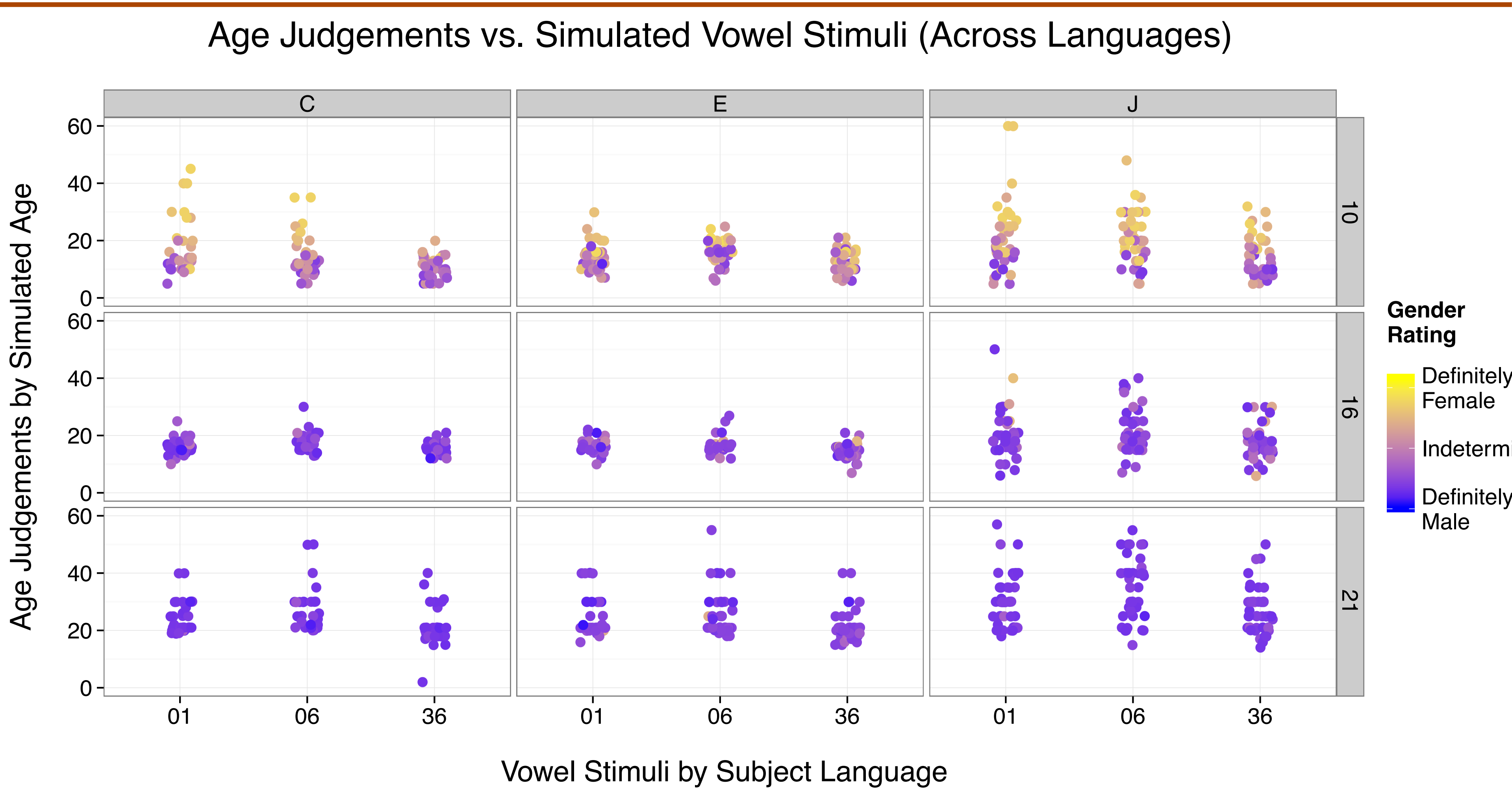
Mean formant difference (female – male) is not predictable from mean female values.

Plummer, Munson, Ménard, & Beckman (2013)



MVS for 7 vocal tract ages; corner vowels rated for age and gender by speakers of Cantonese, English, & Japanese.

Culture-specific age/gender ratings



Plummer, Munson, Ménard, & Beckman (2013) results.

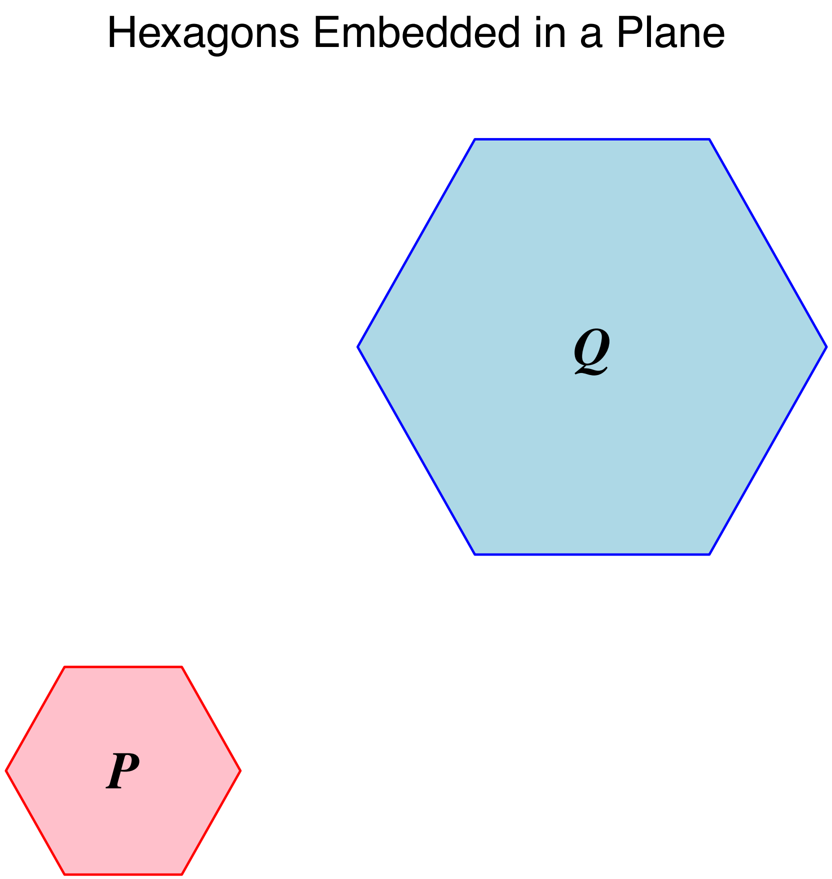
Idea: build mediating cognitive manifolds

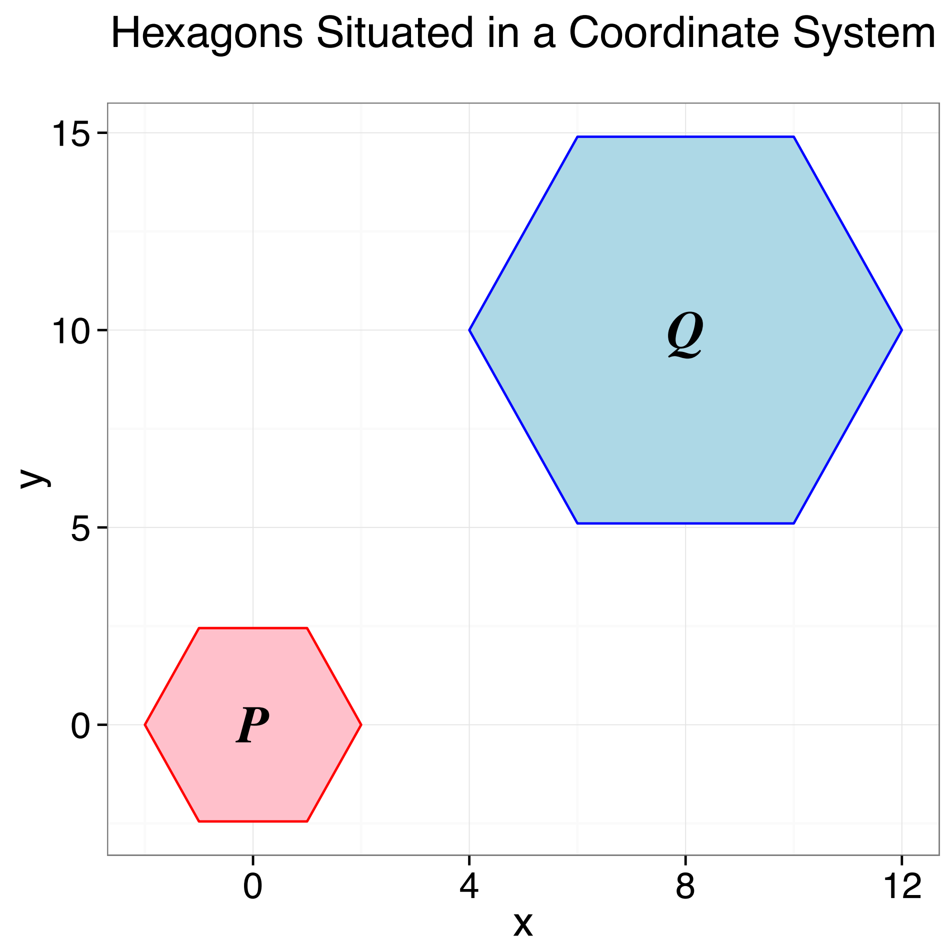
A cognitive manifold describes what our brains might know about something that is very complex and multi-dimensional by building a much lower-dimensional "map" of it.

Ex., a map of some region of the world is a 2-dimensional manifold designed to capture what we need to know to navigate the 3-dimensional surface of our planet.

(If time, work through extract from tutorial "Geometric methods and manifold learning" Belkin & Niyogi, 2003)

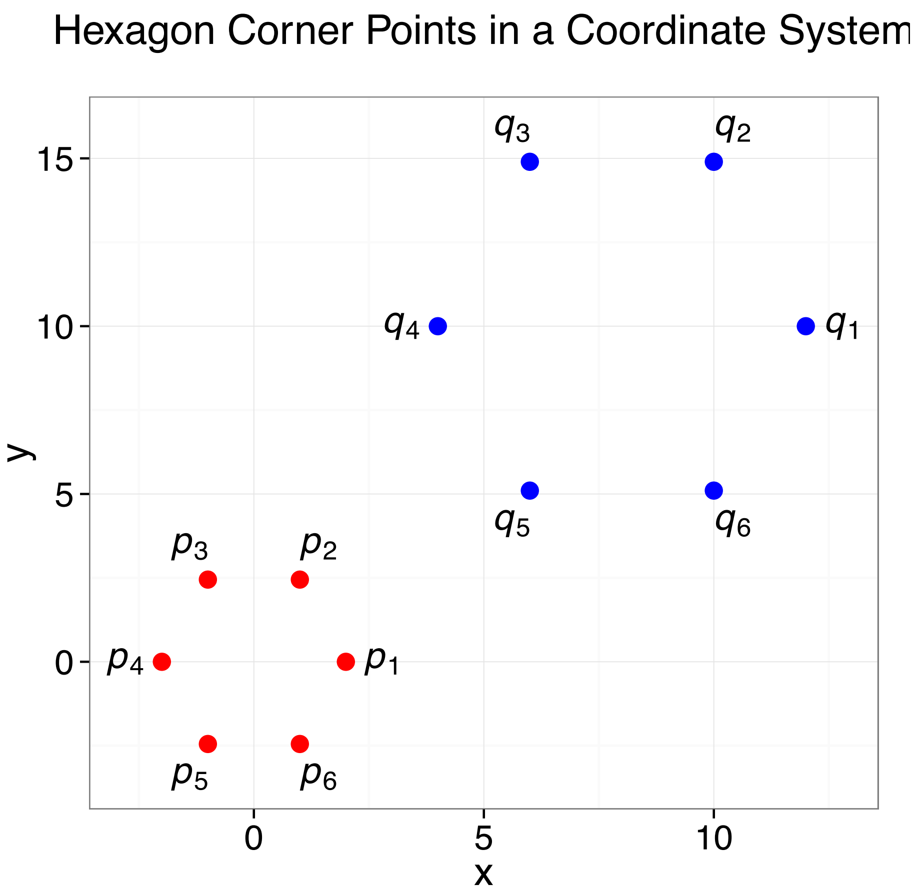
Manifold alignment, example 1

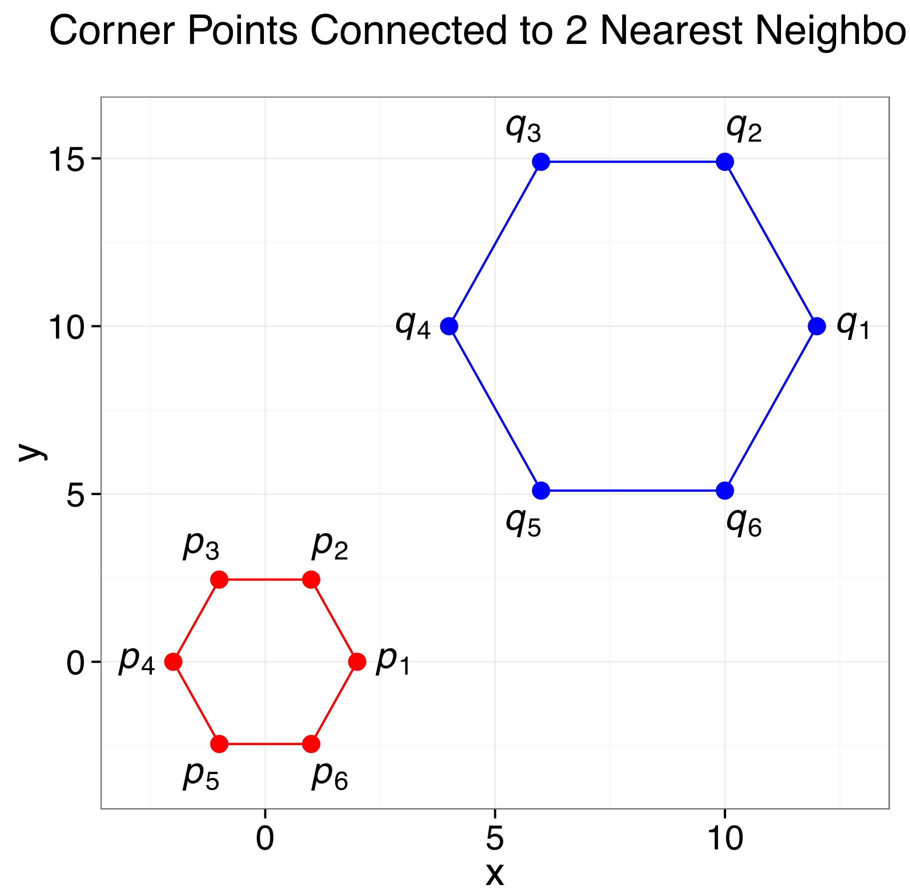




How to learn the mapping between analogous points on the large (Q) and small (P) hexagonal surfaces?

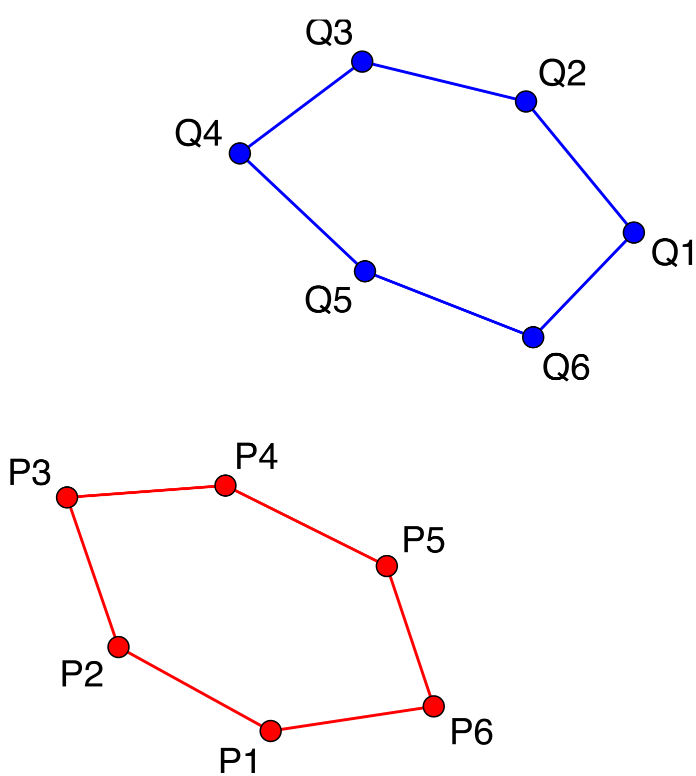
Example 1, step 1

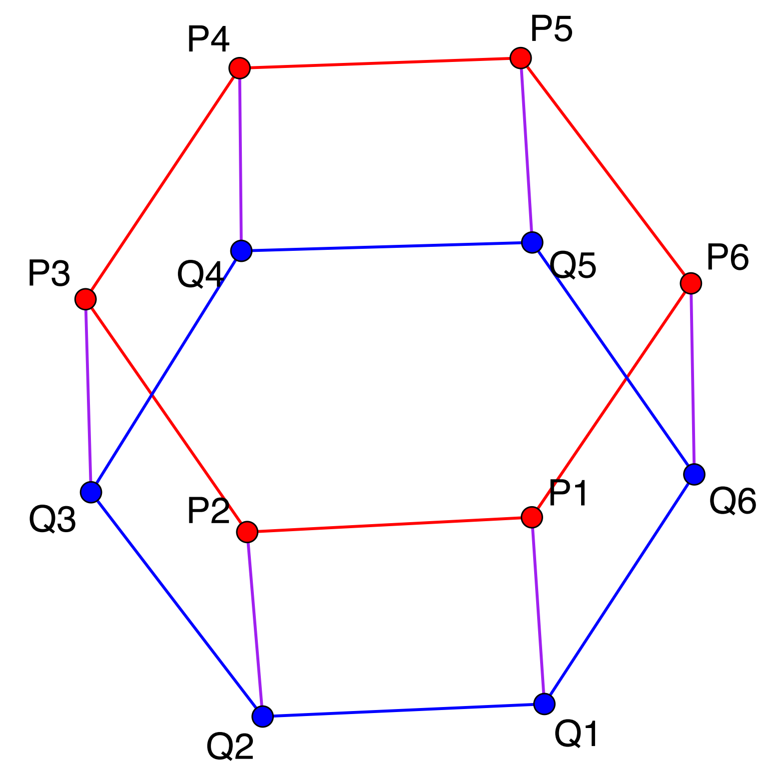




Build manifolds to represent each hexagon in terms of its corner points and each point's two nearest neighbors.

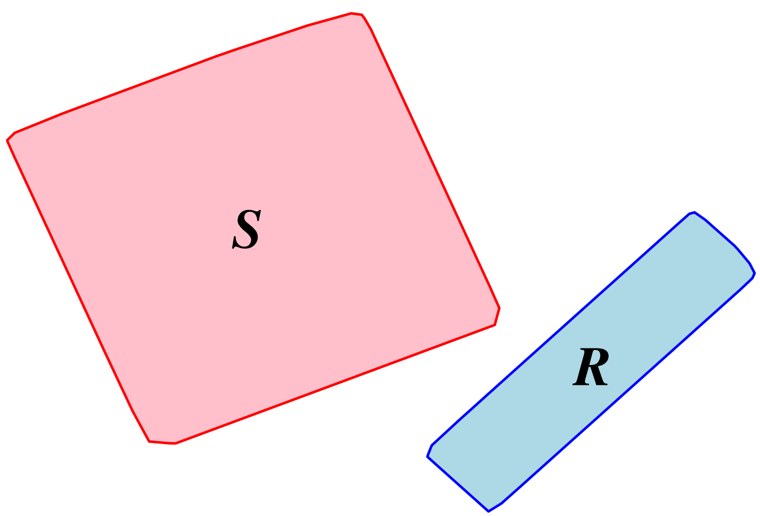
Example 1, step 2

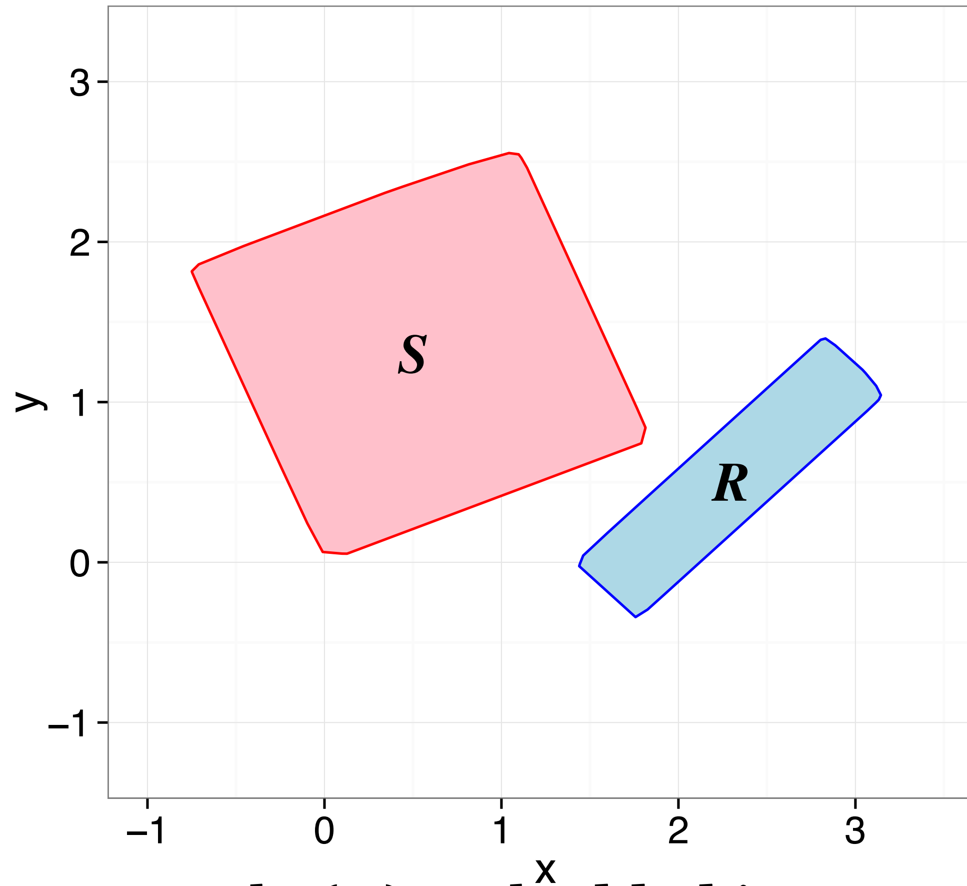




Each shape can be summarized in terms of an adjacency matrix (left) and the two matrices combined in an alignment reference frame (right).

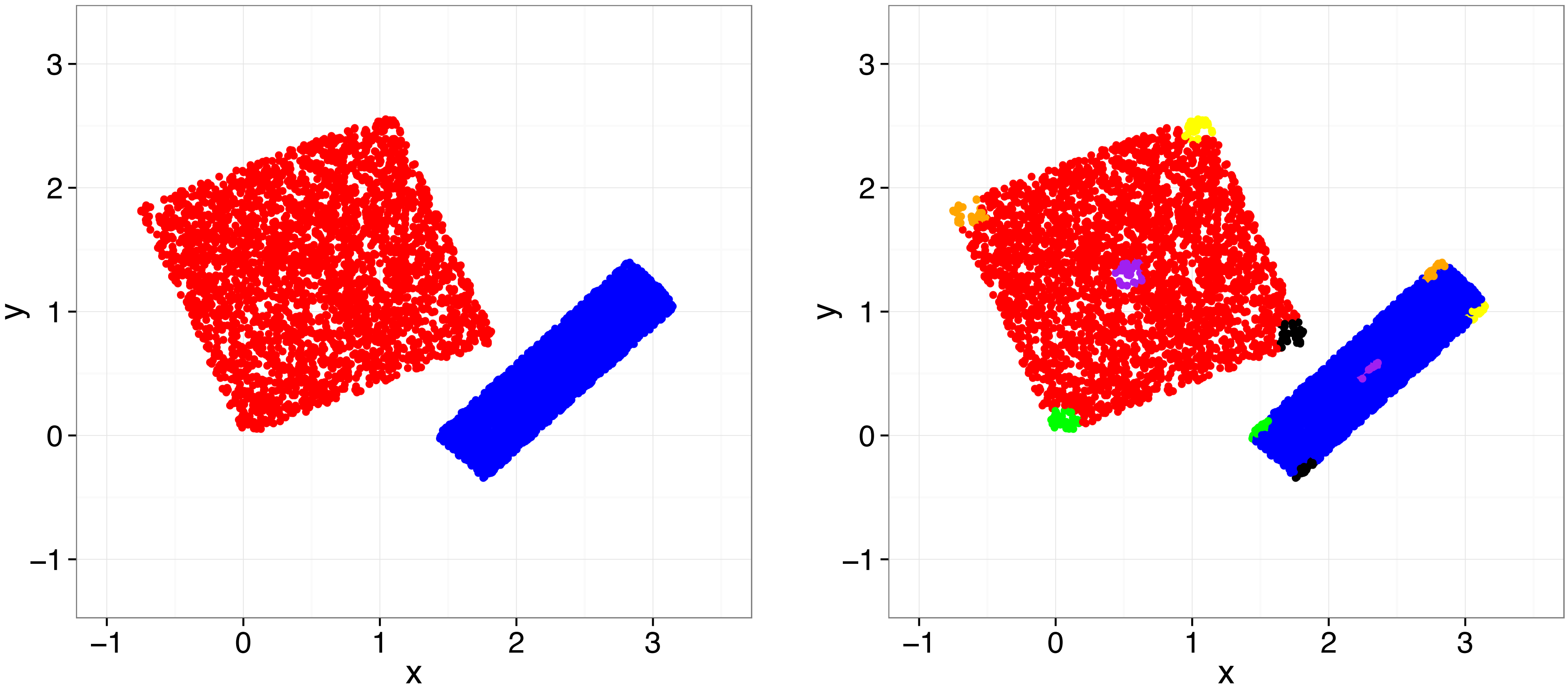
Manifold alignment, example 2





How to align a square (S) and a rectangle (R) embedded in a plane to map between analogous positions on their surfaces?

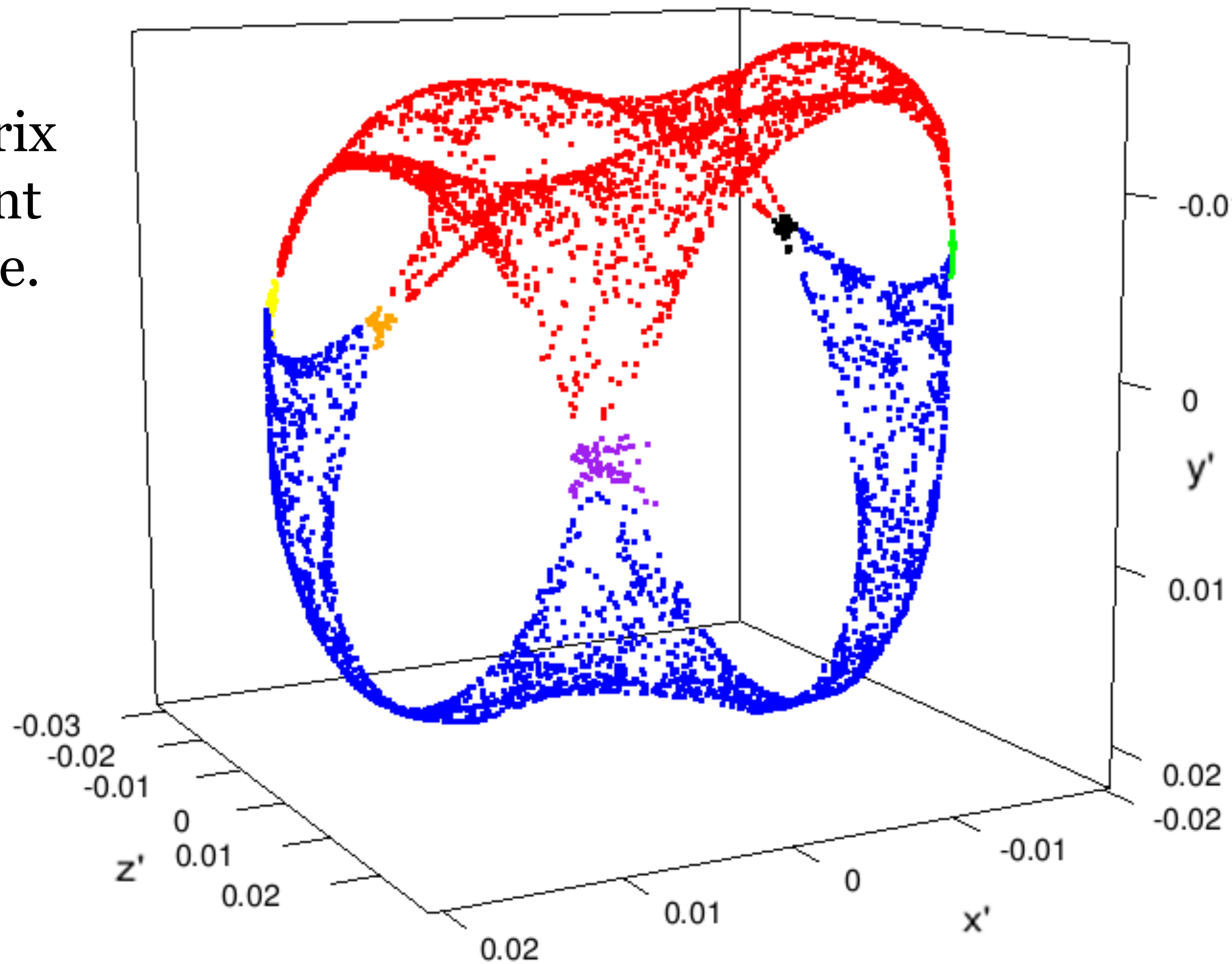
Example 2, steps 1 and 2



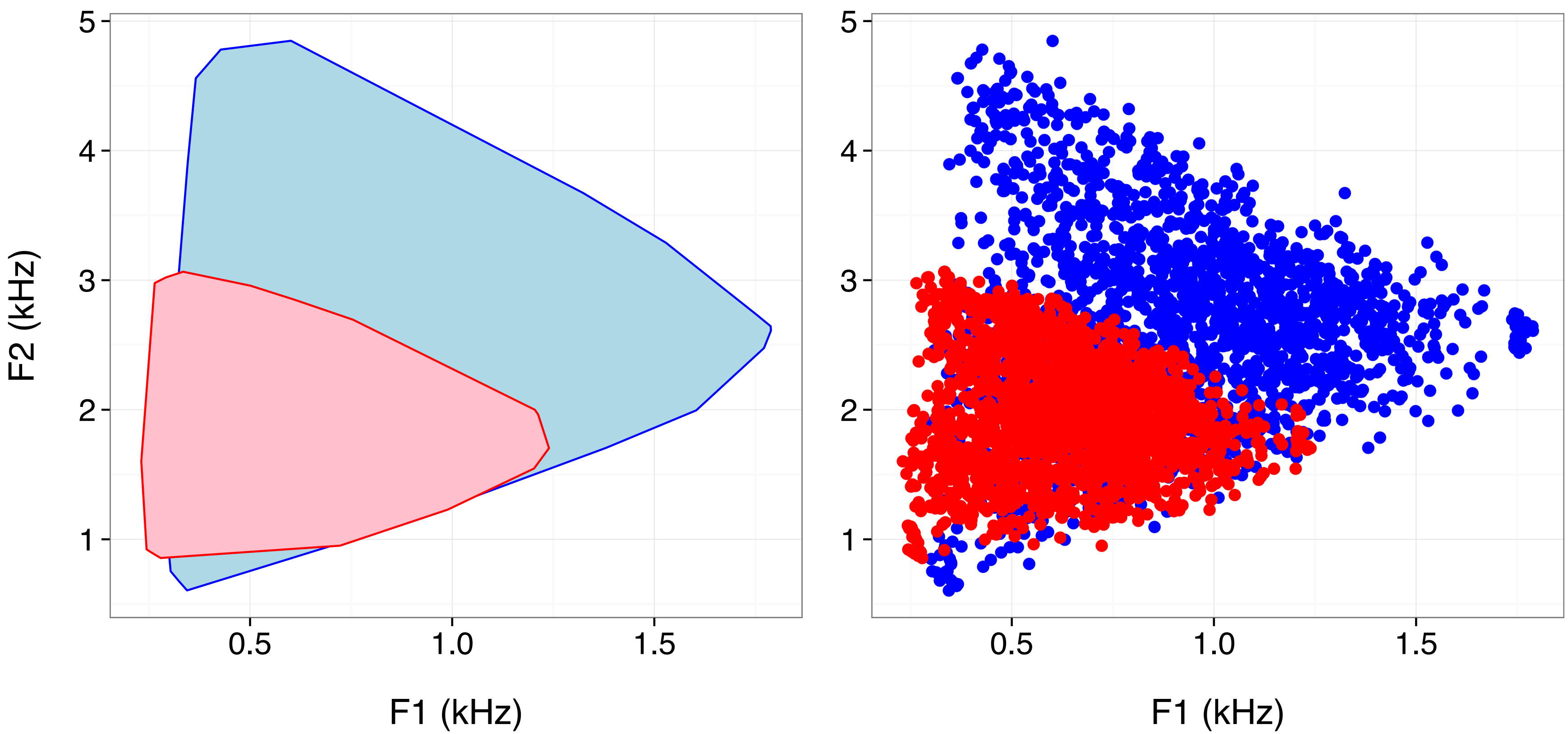
Make a dense sampling of points on each of the two surfaces and weight the adjacency matrices to emphasize (alternatively, to de-emphasize) the alignment of some pairs.

Example 2, the result

The combined adjacency matrix in the alignment reference frame.

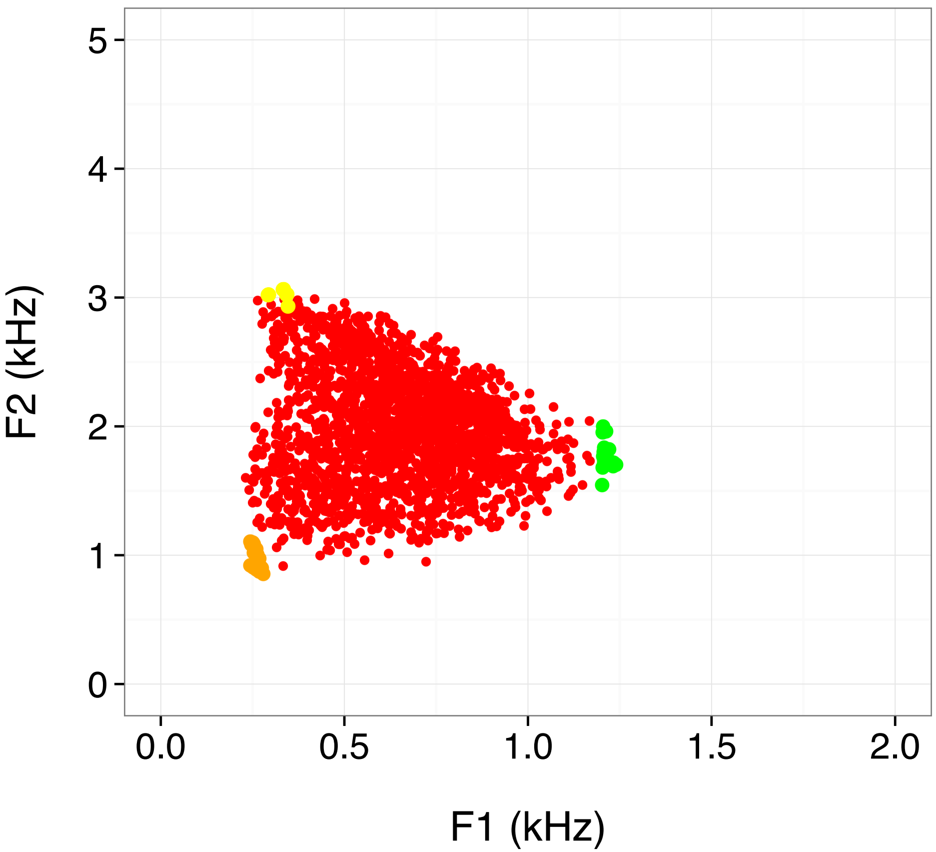


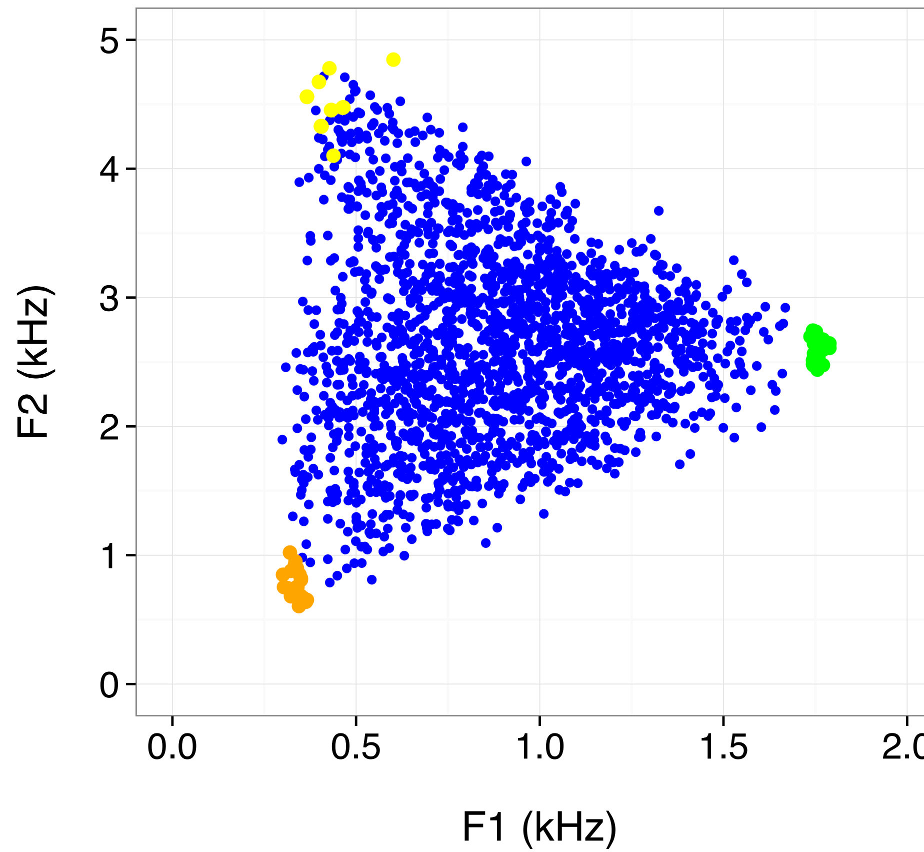
Manifold alignment, example 3



How to align an infant's vowel space (blue triangle) and a caretaker's (pink triangle)? Step 1, make a dense sampling of the two spaces in the F1 / F2 reference frame.

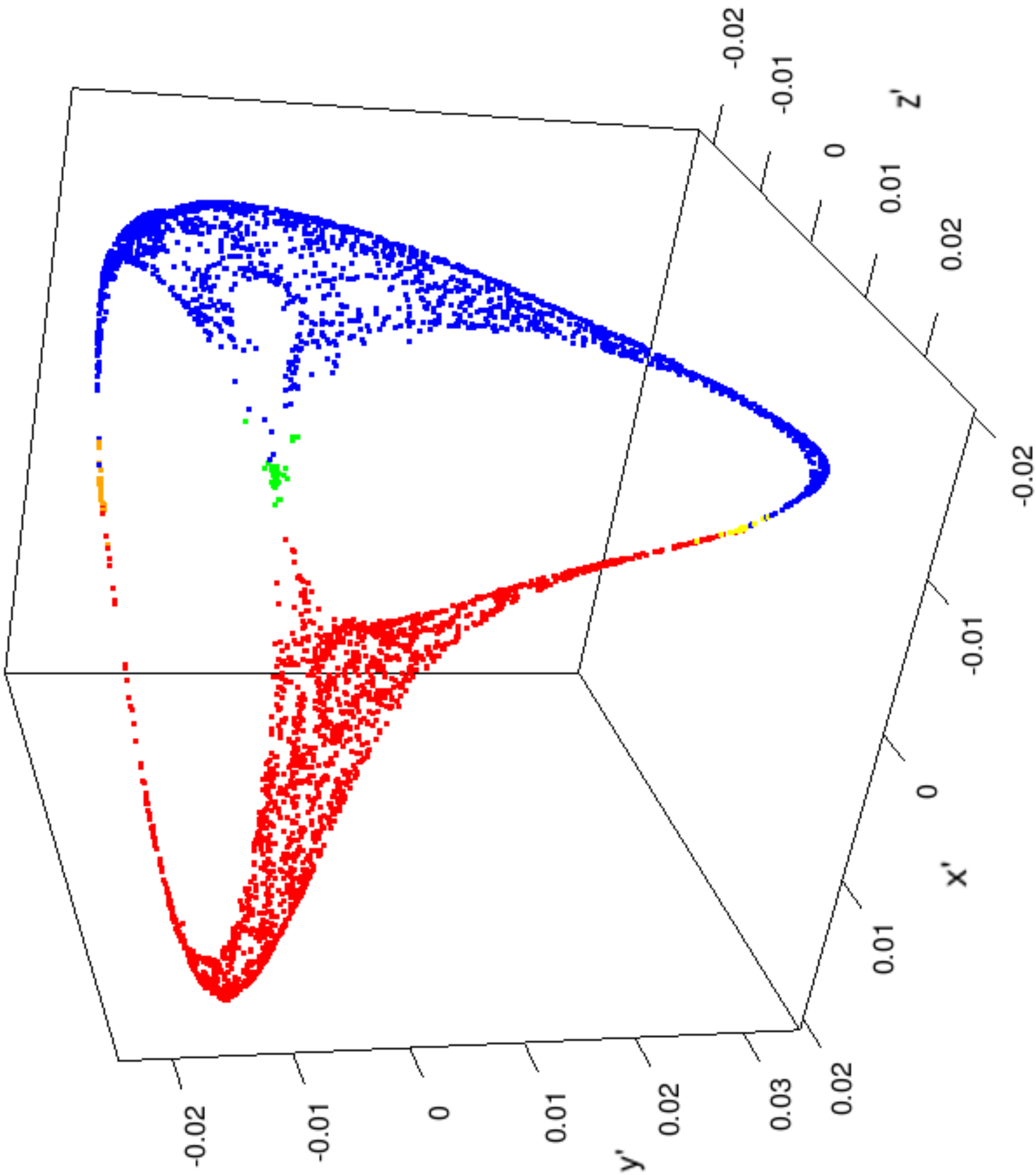
Example 3, step 2



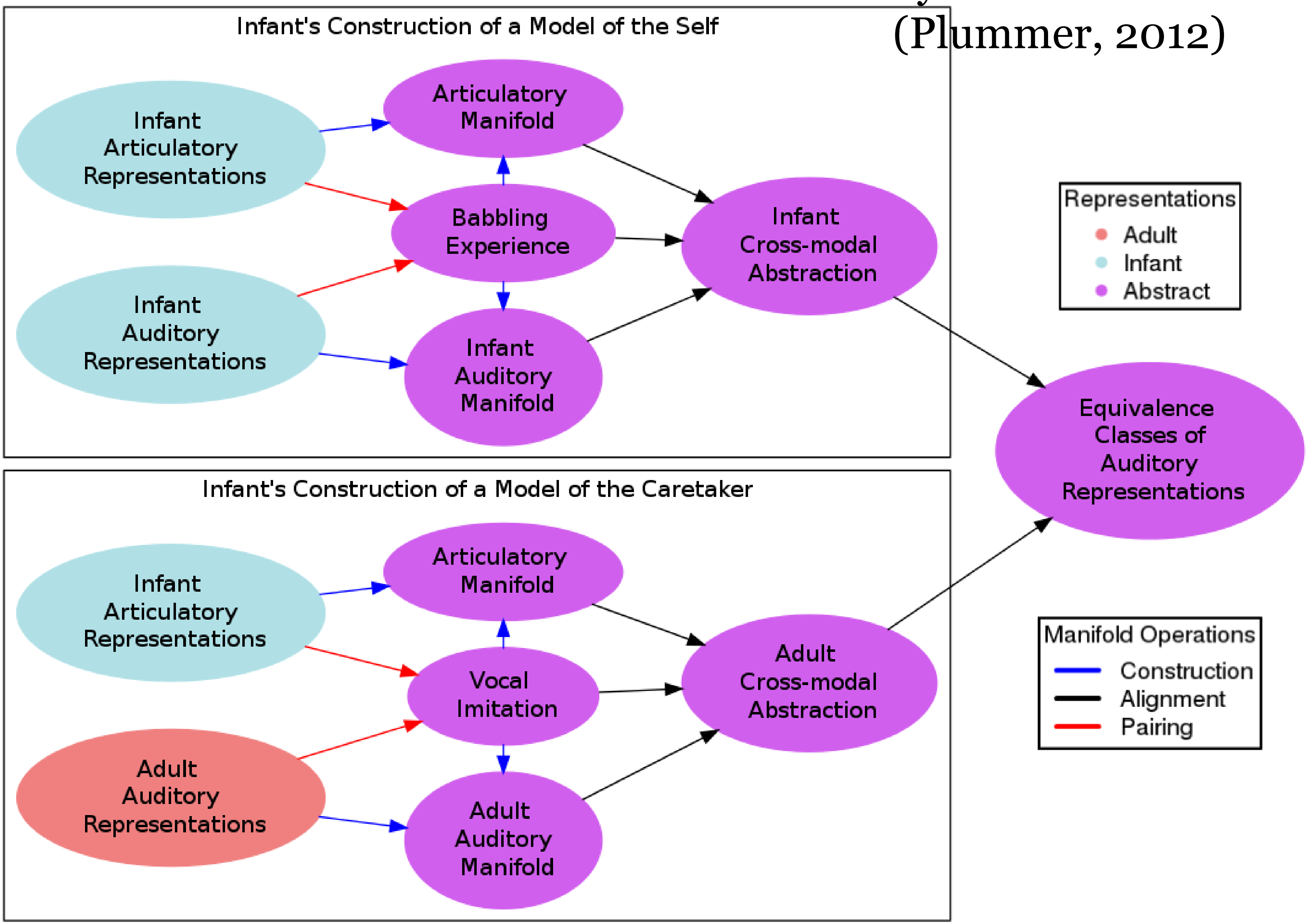


Weighting the adjacency matrix to emphasize points that the caretaker recognizes and gives feedback (e.g., by cooing back).

The resulting vowel space in a combined adjacency matrix reference frame that is abstracted away from the auditory sensory F1/F2 reference frame.



System architecture (Plummer, 2012)



Work in progress

- Building language-specific caretaker response reference frames, using vowel category judgments and goodness ratings (Plummer, Ménard, Munson, & Beckman, submitted).
- Simulate babbling by building the cross-modal manifold alignment between the infant's articulatory reference frame and the infant's auditory reference frame (Plummer, in progress).
- Test the caretaker feedback model (Plummer, in progress).

謝謝 (part 1)

Acknowledgement of funding sources

- NSF Collaborative Research grants:
  - BCS 0729306 to Ohio State University Principal Investigators Mary Beckman & Eric Fosler-Lussier
  - BCS 0729140 to U. of Wisconsin, Madison, PI Jan Edwards
  - BCS 0729277 to U. Minnesota PI Benjamin Munson
- OSU Cognitive Science Seed Grant to Mary Beckman, Mikhail Belkin, and Eric Fosler-Lussier
- Grants from SSHRC, NSERC, and Fonds Québécois de Recherche sur la Société et la Culture to Lucie Ménard
- NIDCD grant RO1 02932 to Jan Edwards

謝謝!

- to 國立中正大學 for inviting me to give this talk
- to 國家科學委員會 for their generous travel support
- to my collaborators, especially those listed on the acknowledgements slide, and
- to you for your kind attention

Ευχαριστώ πολύ

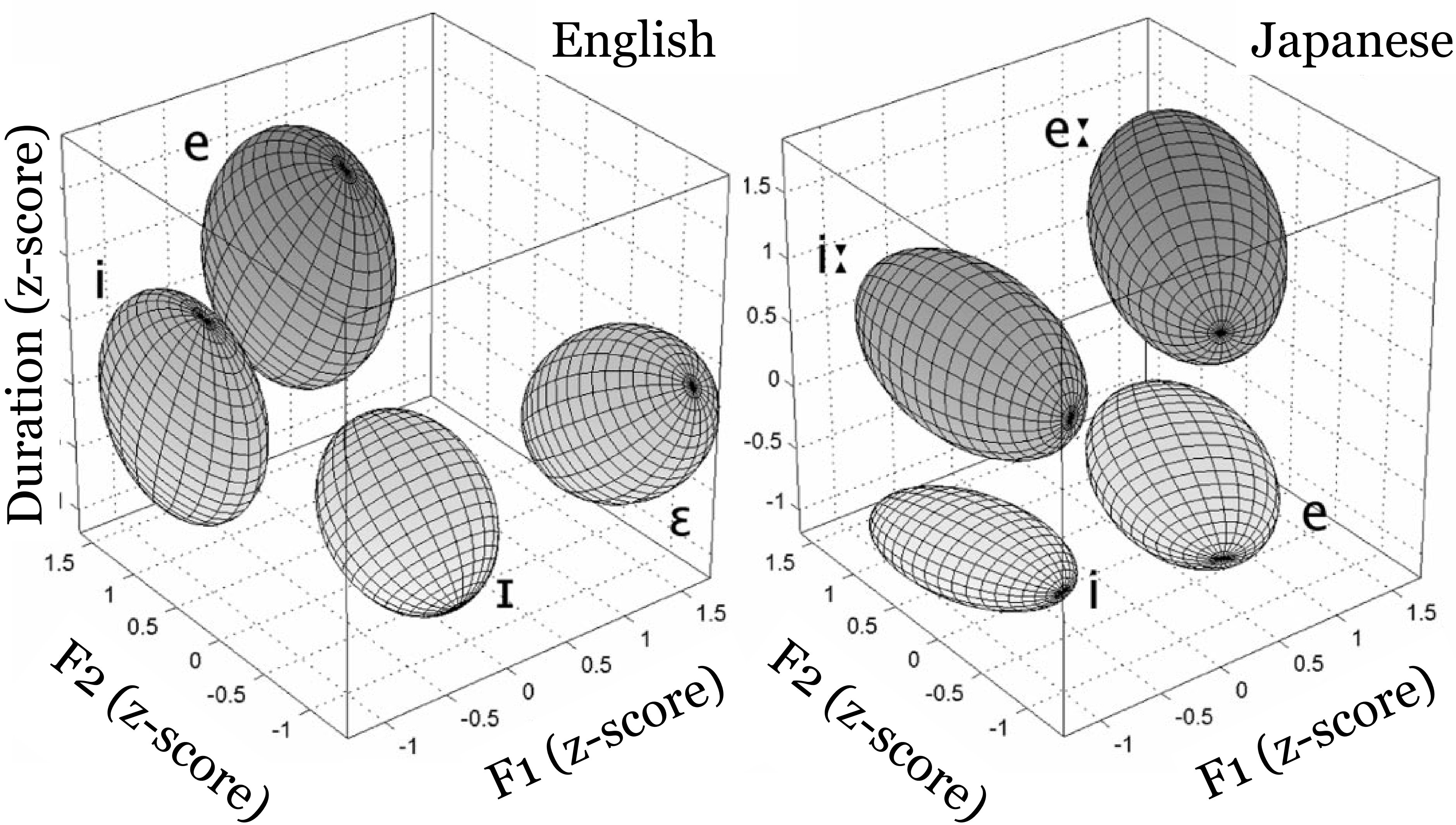
唔該

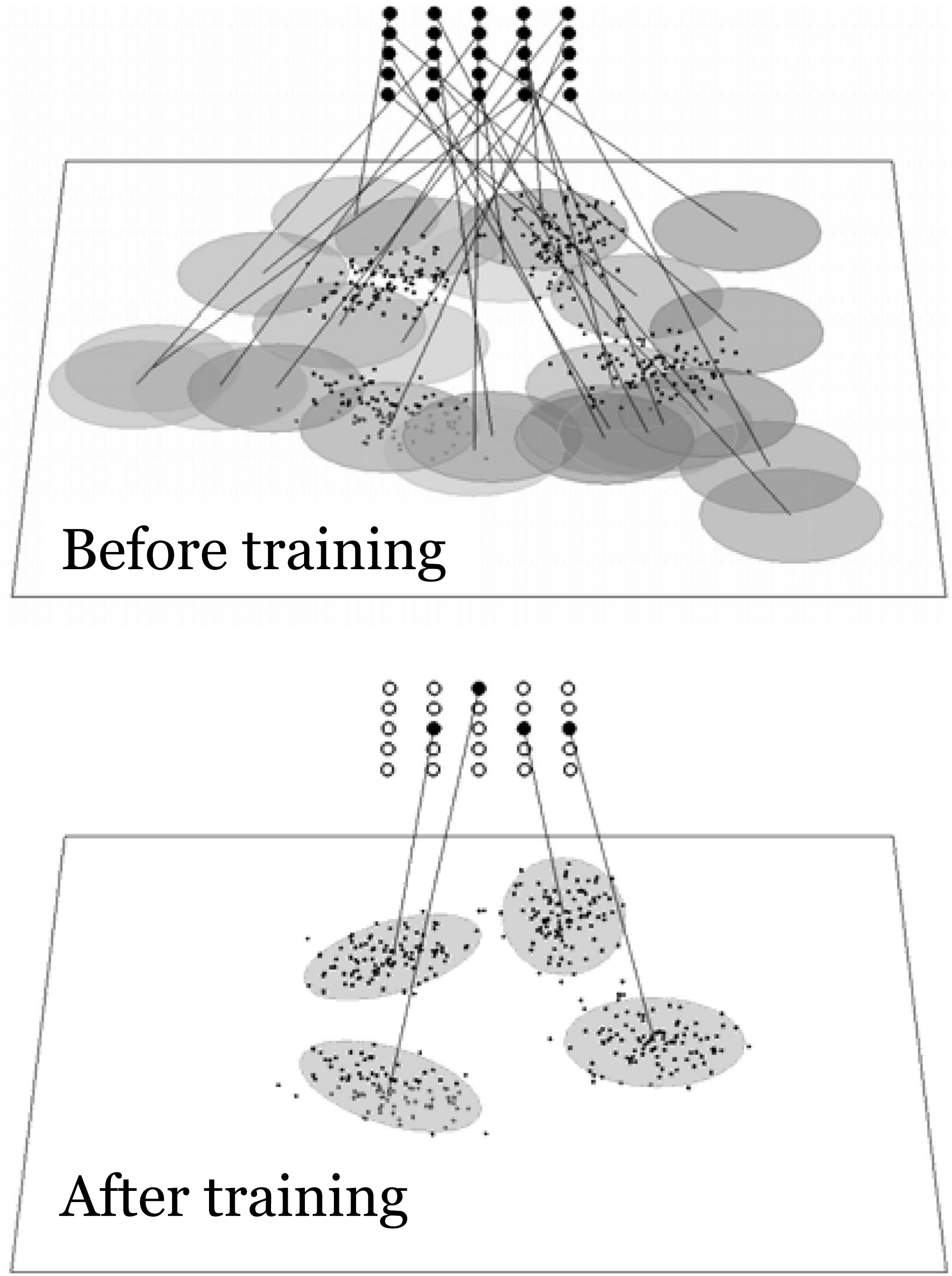
감사합니다

Thank you

ありがとう

Vallabha, McClelland, Pons, Werker, & Amano (2007): infant-directed speech





Vallabha et al. (2007) Fig. 2:

Example of unsupervised learning of vowel categories in F1 / F2 space, using the Online Expectation Maximization algorithm for warping the response map through exposure to Gaussian models of infant-directed speech input

Plummer (2012) results: Caretaker responds in imitative interaction, which rewards infant productions that are closer to adult category.