# Modeling gestural coordination in infant-caregiver dyads in the earliest stages of phonological acquisition

Andrew R. Plummer

Dept. of Comp. Sci. & Eng., The Ohio State University, Columbus, OH, USA

plummer@ling.ohio-state.edu

July 28, 2014

# Gestures and Gestural Coordination

Figure: “Trudeau shrug” from http://trudeauphotogalleryportraits.blogspot.com/#ottawa130 (left), and the “Renaissance Elbow” depicted in *Family Group* by Anonymous (right).

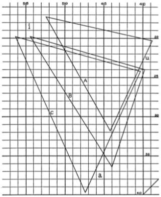
## Gestural Coordination in Communication

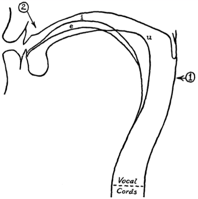
- Gestural communication between individuals can take many forms – e.g. manual gestures and vocal gestures in humans (and in surprisingly many other primates including wild chimpanzees and bonobos).
- Still, gesturing is “a subject on which everyone is something of an expert and most of us are also pretty ignorant” (http://www.robertfulford.com/Gesture.html), and its component concepts are very difficult to define in a nontrivial way.

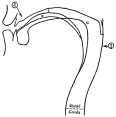
## Gestural Coordination in Phonetics and Phonology

- In formulating descriptive and theoretical accounts of speech and language, phonologists have long made use of gestures, construed in terms of vocal tract morphology, organization, and control during speech production, taken, often by authority, to be invariant across the species.
- The advent of acoustic phonetics in the 19th century brought widespread attention to the acoustic properties of speech and their relation to aspects of speech perception during the first half of the 20th century (see Chiba & Kajiyama, 1941).
- This development directed attention to two key issues which greatly impacted the study of gestural coordination:
  1. the substantial variation in speech signals both within and across speakers and the consequent challenges for speech perception, and
  2. the importance of perception-production links in the development and use of phonological knowledge.

# Gestures and Gestural Coordination







## Gesture Theory (Joos, 1948)

- Gestures within Joos’ approach are “articulatory movements and adjustments (including those of the breathing muscles and the glottis)” (p. 62).
- Regarding the perception-production link within an individual, “there can be no doubt that [speech sounds and articulation are] correlated in the speech center of the brain, for how else does anyone learn to speak except by building up such correlating neural patterns?” (p. 61)
- Gestural coordination between individuals with different vocal tracts is facilitated by “[t]he brain, [which] contains an equivalent of Fig. 29,...and if it has two of [these], it can shift and distort one of them to fit the other” (p. 63).
- Moreover, “[l]earning to do this trick is part of learning to talk. It seems to be one of the last things learned, and its proper management is the outstanding characteristic of adult speech behavior” (p. 63).

# Gestures and Gestural Coordination

## Influential relatives that have appeared over the decades

- Motor Theory (Liberman et al., 1967; Liberman & Mattingly, 1985)
  - Gestures are “class[es] of movements by one or more articulators that results in a particular, linguistically significant deformation, over time, of the vocal-tract configuration” (Liberman & Mattingly, 1985, p. 21), invariant under some criteria.
  - Gestural coordination between individuals is mediated by an innate perception-production link, and phonological acquisition is primarily selectionist.
- Articulatory Phonology (Browman & Goldstein, 1989, 1990a,b)
  - Gestures are “abstract characterisation[s] of coordinated task-directed movements of articulators within the vocal tract” (Browman & Goldstein, 1989, p. 206), aimed at producing constrictions, with an invariant core of some kind.
  - Gestural coordination between infants and caregivers is taken to influence two developmental processes with respect to the infant: “(1) differentiation and tuning of individual gestures and (2) coordination of the individual gestures” (p. 204).
- Distributional/Statistical Learning
  - A number of distributional/statistical learning approaches have appeared in recent years (e.g., de Boer, 2000; Ishihara et al., 2009; Ananthakrishnan & Salvi, 2011; Miura et al., 2012; Rasilo et al., 2013).
  - Gestures are vectors/functions in a parameter space whose dimensions correspond to morphological aspects and motor control of the vocal tract.
  - Gestural coordination between infant and caregiver is approximated by the infant via statistical summaries of their gestures based on caregiver feedback.

## UNPACKING GESTURAL COORDINATION

- Within each approach, core aspects of gestures are assumed to be invariant across individuals, which simplifies both modeling and description of an infant's knowledge of equivalence classes over gestures across individuals, and description of such equivalence classes generally.
- While these kinds of simplifications have resulted in a substantial amount of progress in the field, and outside of it, they leave little room for modeling important classes of phenomena that are once again emerging from new experimental and technical advances, e.g.,:

1. the highly complex socio-cognitive and morphological development that begins in utero and continues well into puberty, with all its intricacies along the way, and
2. the socio-cognitive aspects of animal communication that provide the means for comparative understanding of the potential phylogenetic origins of the ontogenetic aspects we observe in infants as they develop and acquire language.

# Gestures and Gestural Coordination

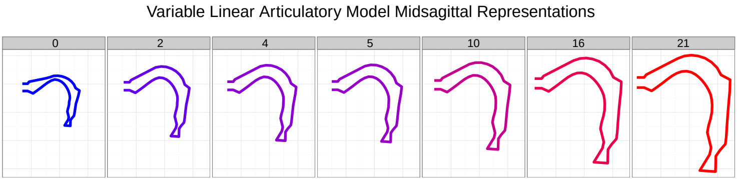
## Phatic foundations of phonological acquisition

- In the interest of forward progress, we have to begin re-examining the conceptual bases of our approaches in light of the tremendous progress that has been made in understanding the phatic functions underlying phonological acquisition.
- In this connection, we take as our point of departure the proposition that phonological acquisition is built on a foundation formed by the phatic functions of the infant's early communicative efforts with caregivers.

## Modeling the acquisition of vowel systems

- The complex nature of these phatic functions has only recently come to light, and, as such, their impact on phonological acquisition is scarcely understood (Gros-Louis et al., 2006; Goldstein & Schwade, 2008; Goldstein et al., 2009; Hsu et al., 2013).
- In order to facilitate understanding, we put forward a modeling framework focused on how infants acquire vowel systems via socio-vocal interaction with their caregivers and the role of phatic functions in doing so.

# Modeling Framework: Conceptual Basis

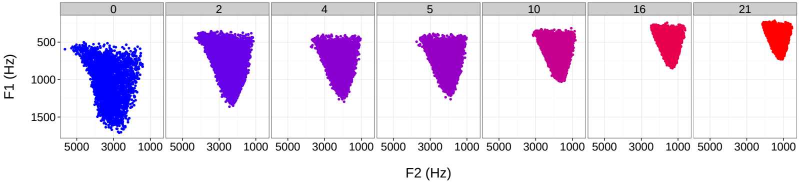

Variable Linear Articulatory Model Midsagittal Representations
0
2
4
5
10
16
21

## Using an Articulatory Model

- The VLAM (Boë & Maeda, 1998) is an age-varying computational model of the articulatory system that is capable of representing vocal tract lengths ranging from those of infants to young adults, calibrated in accordance with age-related “organic variation” (Goldstein, 1980; Beck, 1996).
- Midsagittal representations are wrought by configuring “articulatory blocks” (Lindblom & Sundberg, 1971; Maeda, 1990, 1991) whose components correspond to jaw height, tongue body position, tongue dorsum position, tongue apex position, lip protrusion, lip height, and larynx height.
- Parameters corresponding to each component take on numeric values within intervals centered at 0, and each parameter setting is called an articulatory vector.

# Modeling Framework: Conceptual Basis

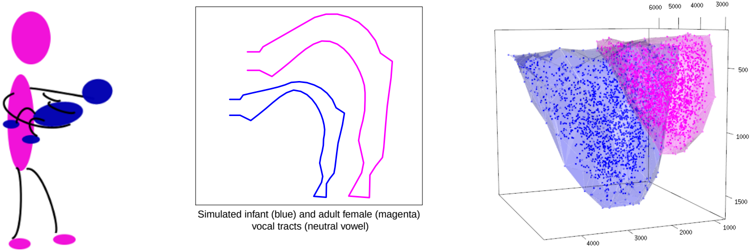
Variable Linear Articulatory Model Maximal Vowel Spaces



## Maximal Vowel Spaces

- Given an age in years, the set of all articulatory vectors for the VLAM at that age that do not result in occlusion of the oral cavity yield a corresponding maximal vowel space (MVS, Boë et al., 1989; Schwartz et al., 2007) for that age.
- For each vocal tract age $a$, we take the MVS to be characterized by its set of articulatory vectors, or maximal articulatory space (MARS).
- Each maximal articulatory space has a corresponding maximal formant space (MFS) composed of formant vector whose components are the first three formant frequencies corresponding to an articulatory vector in the MARS.
- Given a vocal tract age $a$, let MARS($a$) be a dense sampling of the MARS for age $a$, and let MFS($a$) be the corresponding dense sampling of the MFS.

# Modeling Framework: Conceptual Basis



Simulated infant (blue) and adult female (magenta) vocal tracts (neutral vowel)

FIGURE: (left) Dyad composed of an infant (blue) and caregiver (magenta), with simulated vocal tracts (neutral vowel, center), and maximal formant spaces MFS(0.5) and MFS(10) (right).

## Infant-Caregiver Dyad Modeling

- We fix two vocal tract ages $a_0$ and $a_1$ where $a_0$ is the age of a model caregiver, and $a_1$ that of a model infant.
- We take $a_1$ to be six months (denoted 0.5), and $a_0$ to be 10 years, based on age and gender judgments provided during perceptual experiments using VLAM stimuli which indicated that the 10 y.o. vocal tract sounded most like an adult female.

# Modeling Framework: Conceptual Basis

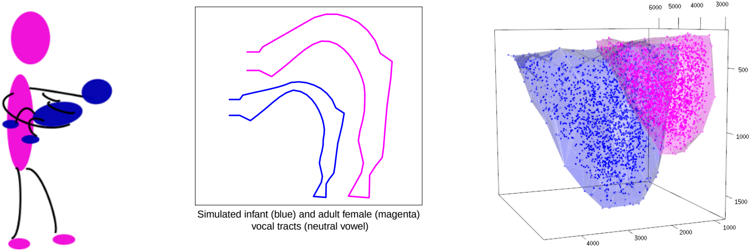
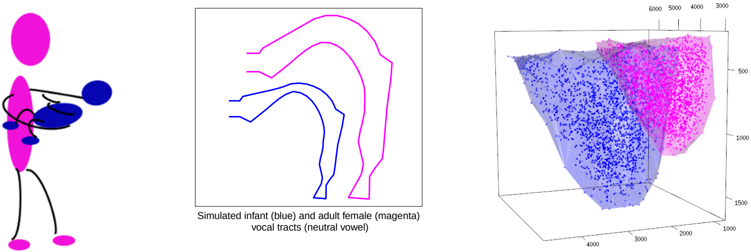

Simulated infant (blue) and adult female (magenta) vocal tracts (neutral vowel)

FIGURE: (left) Dyad composed of an infant (blue) and caregiver (magenta), with simulated vocal tracts (neutral vowel, center), and maximal formant spaces MFS(0.5) and MFS(10) (right).

## Constructing Models of the Self

- Infants require sensory access to their own productions to ensure normal development during phonological acquisition, and in the case of spoken language, this access is through the auditory and articulatory systems.
- While this early experience is typically construed as an exploratory practice phase in which the infant is learning motor control and the perceptual consequences of vocal productions, it seems that infants are also developing complex models of the self as they begin to form affiliative relations with caregivers.

# Modeling Framework: Conceptual Basis



Simulated infant (blue) and adult female (magenta) vocal tracts (neutral vowel)

FIGURE: (left) Dyad composed of an infant (blue) and caregiver (magenta), with simulated vocal tracts (neutral vowel, center), and maximal formant spaces MFS(0.5) and MFS(10) (right).

## Constructing Models of Caregivers

- Phonological systems appear to be more complex than is necessary for “efficient communication” (see Fitch, 2004), suggesting that speech signals contain room for passing on phatic information not directly tied to referential content.
- Caregivers may be exploiting this room for passing on socio-indexical information to infants during socio-vocal exchanges (see Masataka, 2003, for a review).
- Moreover, infants may be internalizing this information with a high level of fidelity in order to construct rich models of their caregivers that facilitate social learning.

# Modeling Framework: Conceptual Basis

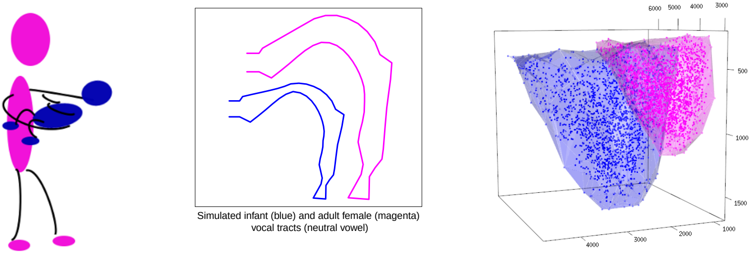
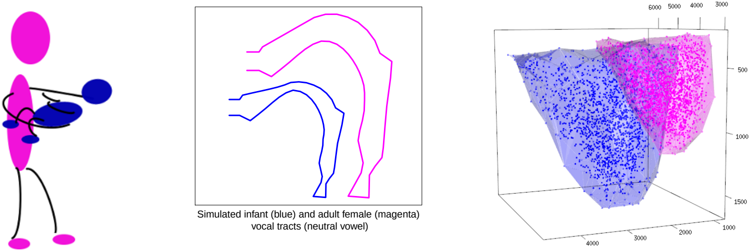

Simulated infant (blue) and adult female (magenta) vocal tracts (neutral vowel)

Figure: (left) Dyad composed of an infant (blue) and caregiver (magenta), with simulated vocal tracts (neutral vowel, center), and maximal formant spaces MFS(0.5) and MFS(10) (right).

## Model Construction: Multisensory Computations

- Perception-production links serve as more than conduits for the passage of information from one sensory domain to another, and their development likely aids in the construction of models of the self and others.
- This development involves numerous multisensory computations and abstractions that take places during early infancy (Meltzoff & Kuhl, 1994; Lewkowicz & Ghazanfar, 2009).

# Modeling Framework: Conceptual Basis



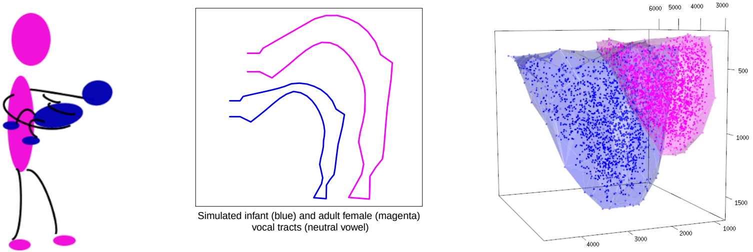
Simulated infant (blue) and adult female (magenta) vocal tracts (neutral vowel)

FIGURE: (left) Dyad composed of an infant (blue) and caregiver (magenta), with simulated vocal tracts (neutral vowel, center), and maximal formant spaces MFS(0.5) and MFS(10) (right).

## Model Construction: Socio-vocal Exchanges

- One of the principle phatic functions of an infant's early communicative efforts with caregivers involves the formation of commensuration relations between models of the self and models of caregivers constructed by the infant (e.g., Meltzoff, 2007).
- Both construction and relation of models is likely facilitated by specialized interaction between infants and caregivers during which complex information concerning caregiver socio-vocal properties may be internalized by the infant with high fidelity.

# Modeling Framework: Conceptual Basis



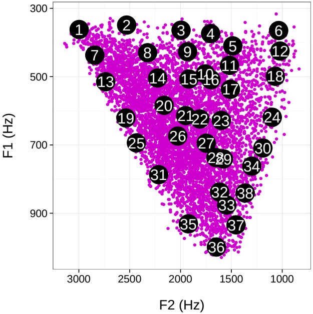
Simulated infant (blue) and adult female (magenta) vocal tracts (neutral vowel)

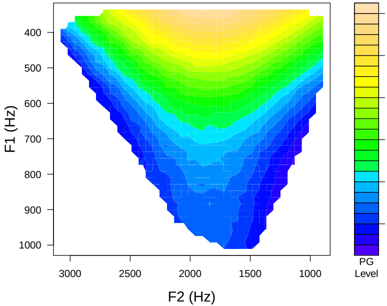
FIGURE: (left) Dyad composed of an infant (blue) and caregiver (magenta), with simulated vocal tracts (neutral vowel, center), and maximal formant spaces MFS(0.5) and MFS(10) (right).

## Model Construction: Socio-vocal Exchanges

- We limit our consideration of this specialized interaction to turn-taking vocal exchanges between infants and their caregivers, which are known to play a role in forming affiliative relations which influence phonological acquisition (see Masataka, 2003).
- There is growing evidence for the existence of phylogenetic precursors in other primates (e.g., Crockford & Boesch, 2005; Genty et al., 2014), suggesting that turn-taking vocal exchanges act as a general compression mechanism for passing socio-indexical information between signaler and receiver.

# Modeling Framework



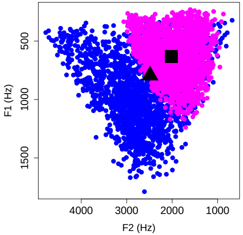


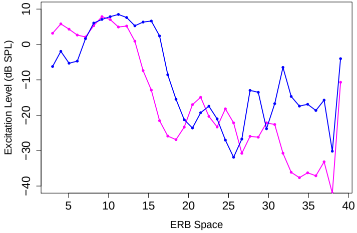
## External Data

We use results from perceptual categorization experiments (Munson et al., 2010) in building vowel category response fields over maximal formant spaces that capture

- a caregiver's knowledge of their language's vowel system and
- how they might pass that knowledge on to infants during vocal exchanges.

# Modeling Framework

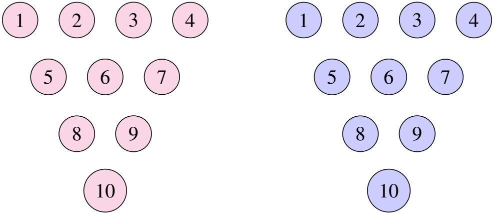




## Internalization

- We use a psychophysical transformation (Moore & Glasberg, 1996; Moore, 1997) that maps vowels to excitation vectors that capture the auditory system's influence on sound internalization.
- Given an MFS $\textsc{mfs}(a)$, the transformation yields a maximal auditory space (MAUD), denoted $\textsc{maud}(a)$.
- We also use a phatic transformation that captures an infant's internalization of caregiver social representations.

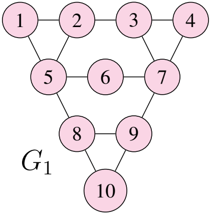
# Modeling Framework

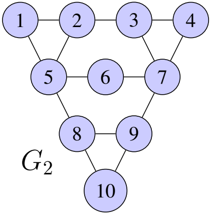


## Cognitive Computations

- We use cognitive manifolds (Plummer, 2014, derived from Seung & Lee, 2000; Gallese, 2001; Niyogi, 2004, inter alia) formed over representations within two psychophysical reference frames (auditory and articulatory) to model an infant's initial creation of models of the self and their caregivers.
- We use manifold alignment (see Ham et al., 2005; Wang, 2010) to model an infant's creative generation of representations that facilitate commensuration of models of the self and others.

# Modeling Framework

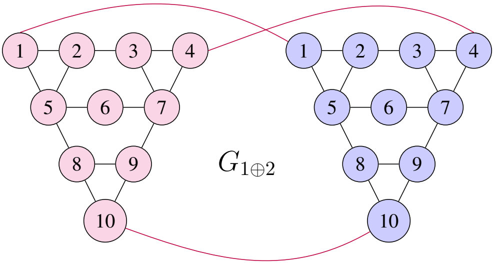




## Cognitive Computations

- We use cognitive manifolds (Plummer, 2014, derived from Seung & Lee, 2000; Gallese, 2001; Niyogi, 2004, inter alia) formed over representations within two psychophysical reference frames (auditory and articulatory) to model an infant's initial creation of models of the self and their caregivers.
- We use manifold alignment (see Ham et al., 2005; Wang, 2010) to model an infant's creative generation of representations that facilitate commensuration of models of the self and others.

# Modeling Framework



## Cognitive Computations

- We use cognitive manifolds (Plummer, 2014, derived from Seung & Lee, 2000; Gallese, 2001; Niyogi, 2004, inter alia) formed over representations within two psychophysical reference frames (auditory and articulatory) to model an infant's initial creation of models of the self and their caregivers.
- We use manifold alignment (see Ham et al., 2005; Wang, 2010) to model an infant's creative generation of representations that facilitate commensuration of models of the self and others.

# Modeling External Data

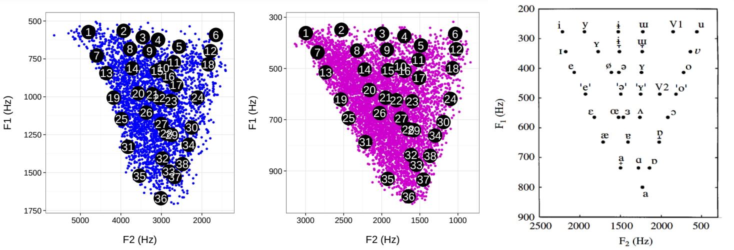
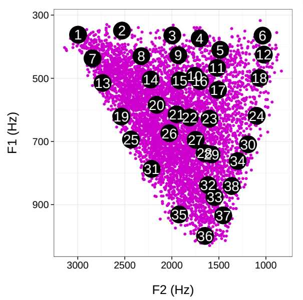


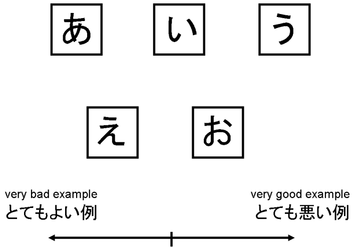
Figure: Prototype sets P(0.5) (left), P(10) (center), and those from Schwartz et al. (1997a).

## Stimuli

- The stimuli used in the perceptual experiments (Munson et al., 2010) were vowel prototypes (Vallée et al., 1995), or simply prototypes, selected from the MFSs for selected vocal tract ages.
- The selection process (Ménard & Boë, 2000; Ménard et al., 2002) is meant to yield a set of cross-linguistically relevant prototypes in accordance with the dispersion-focalization theory (Schwartz et al., 1997a,b).
- For each selected age $a$, a set of 38 prototypes, $\mathrm{P}(a)$, were selected from $\mathrm{MFS}(a)$.
- Prototypes in $\mathrm{P}(a)$ are indexed $\mathbf{p}^i$, for $1 \leq i \leq 38$, though the superscript is often dropped, or used in place of $\mathbf{p}$.

# Modeling External Data





あ い う

え お

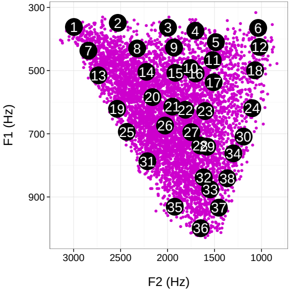
very bad example
とてもよい例

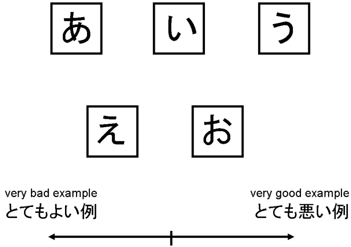
very good example
とても悪い例

## Vowel Category Responses

- Based on the vowel categories within their native languages, Cantonese-, American English-, Greek-, Japanese-, and Korean-speaking listeners categorized seven sets of synthetic vowels generated from models of the vocal tract ranging in age from six months to 21 years of age.
- “Cantonese- and American English-speaking listeners categorized prototypes by clicking on keywords representing the monophthongal vowels in each language, while Greek-, Japanese-, and Korean- speaking listeners categorized by clicking on a symbol string that unambiguously represented a (short monophthongal) vowel in isolation” (Edwards and Beckman, in press).

# Modeling External Data





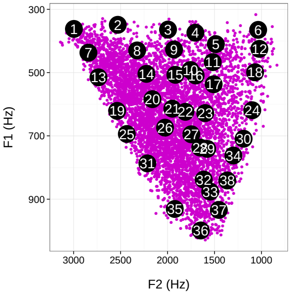
very bad example
とてもよい例

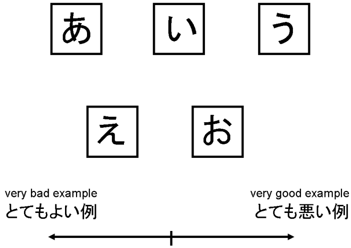
very good example
とても悪い例

## Goodness Ratings

- After selecting a vowel category for a given stimulus, listeners provided a goodness rating along a visual analog scale (VAS, Massaro & Cohen, 1983; Miller, 1994, 1997) indicating how good the listener felt that stimulus was as an example of the selected category.
- Goodness ratings are numeric values normalized to fall between 0 (poorest) and 1 (best) to facilitate interpretation.
- Goodness ratings provide fine-grained information about listeners' categorization of prototypes that allow for the construction of “response surface” models that are otherwise unavailable to the modeler.

# Modeling External Data





あ い う

え お

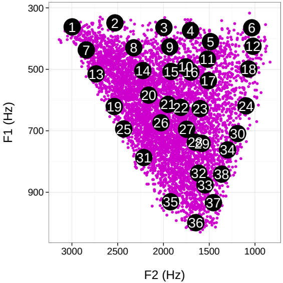
very bad example
とてもよい例

very good example
とても悪い例

## Using categorizations and goodness ratings

- For each language, and for each subject, and for each vocal tract age, each prototype has a category associated with it, and a goodness rating.
- We want to use these responses to build models of each subject's knowledge of their language's vowel system.
- To illustrate the approach, we'll focus on the category responses provided by a Japanese subject, named $J_{20}$, and the 10 year old vocal tract age.

# Modeling External Data



| Proto. | Cat. | Good. |
|---|---|---|
| 1 | i | 0.88 |
| 2 | u | 0.78 |
| 3 | u | 0.82 |
| 4 | u | 0.85 |
| 5 | u | 0.92 |
| 6 | u | 0.84 |
| 7 | e | 0.16 |
| ⋮ | ⋮ | ⋮ |

## Organizing and Generalizing Subject Responses

- Vowel category judgements and goodness ratings for subject $J_{20}$ with respect to the age 10 prototypes are shown above.
- Each response gives us a hint about a subject's knowledge of their language's vowel system, but we need to fill in some gaps.
- For example, according to $J_{20}$ prototype 1 is a really good example of /i/, but what about /u/, /o/, or the other vowels?
- Similarly, the response tells us that prototype 2 is a good example of /u/, but what about /i/, /o/, or the other vowels?
- And what about all of the other prototypes?

# Modeling External Data

## Expanding the responses across the vowel space

- We can use these responses to model what a subject might know about each prototype with respect to each vowel in the subject's vowel system.
- For each prototype, we create a vector with as many components as there are vowels in the subject's vowel system:

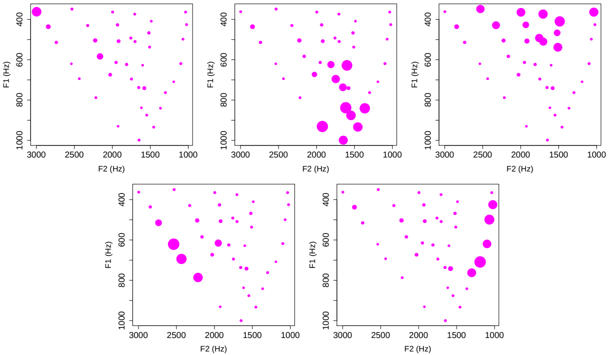
  | i | u | e | o | a |
  |---|---|---|---|---|
  | _ | _ | _ | _ | _ |

  whose components are a function of the subject's response to that prototype.
- To exemplify, for prototype 1 with response (i, 0.88), the vector is

  | i | u | e | o | a |
  |---|---|---|---|---|
  | 0.88 | 0.06 | 0.06 | 0.06 | 0.06 |

- This function is specified in generality in Plummer et al. (2013).
- Thus, for each prototype, we have a model of a vowel category goodness for each vowel.

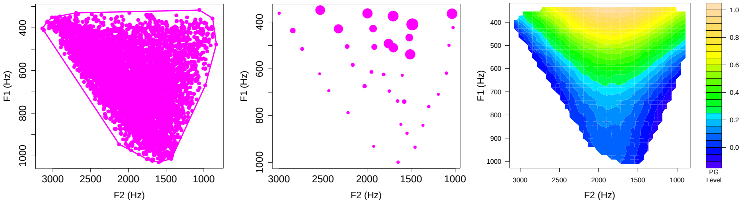
# Modeling External Data



## Individual Category Response Functions

- For each subject, and for each age, using these vectors, we can define functions from prototypes to goodness ratings for each vowel category.
- We can depict these individual category response functions in terms of weighted scatter plots, like the ones shown above for age 10, for subject $J_{20}$.

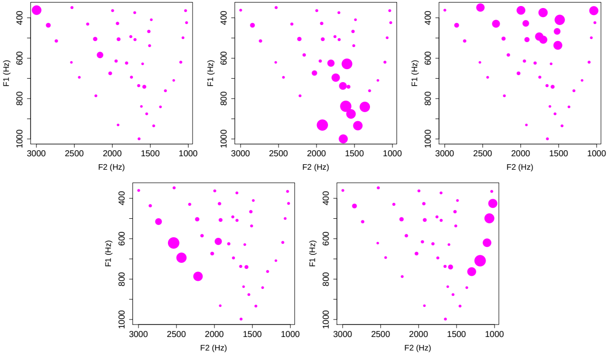
# Modeling External Data



## Additive Modeling

- Now that we have models of vowel category responses for each prototype for each vocal tract age, we can extend these models to the entire maximal formant space for each age.
- The basic idea is to use additive models (Friedman & Stuetzle, 1981; Buja et al., 1989; Hastie & Tibshirani, 1990; Wood, 2003, 2006) based on smoothing splines (Wahba, 1990; Gu, 2002) to construct a field of responses over a maximal formant space using an individual category response function.
- For a given vowel category, the response field value for a formant vector in a maximal formant space is meant to approximate a subject’s goodness rating of that vector as an example of that vowel category.

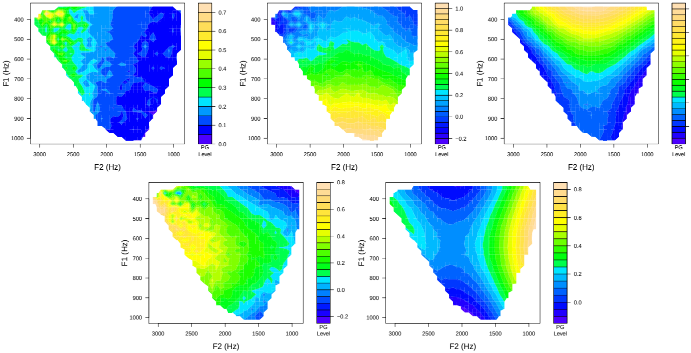
# Modeling External Data



## Vowel Category Response Fields

- For each subject, and for each age, using individual category response functions, we can define functions from maximal vowel spaces to goodness ratings for each vowel category.
- We can depict these vowel category response fields (VCRF) in terms of contour plots, like the ones shown above for age 10, for subject $J_{20}$.

# Modeling External Data



## Vowel Category Response Fields

- For each subject, and for each age, using individual category response functions, we can define functions from maximal vowel spaces to goodness ratings for each vowel category.
- We can depict these vowel category response fields (VCRF) in terms of contour plots, like the ones shown above for age 10, for subject $J_{20}$.

# Modeling External Data

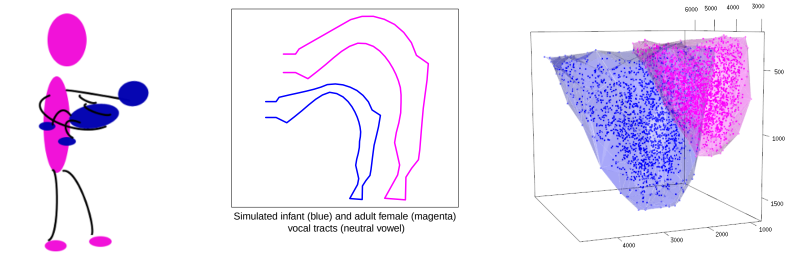
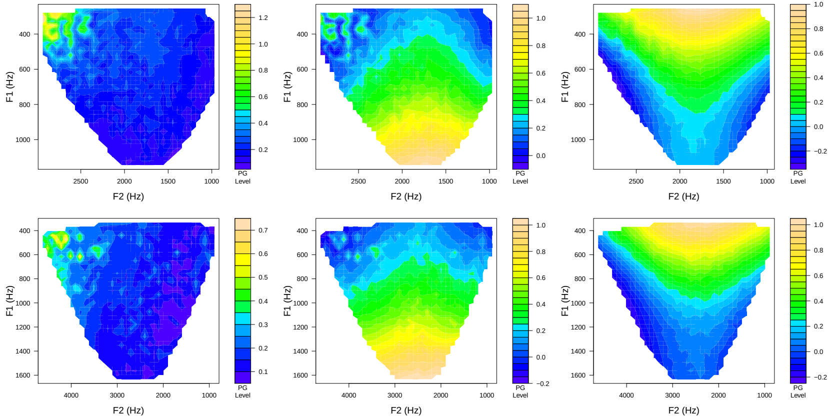

Simulated infant (blue) and adult female (magenta) vocal tracts (neutral vowel)

FIGURE: (left) Dyad composed of an infant (blue) and caregiver (magenta), with simulated vocal tracts (neutral vowel, center), and maximal formant spaces MFS(0.5) and MFS(10) (right).

## Infant-Caregiver Dyad Modeling

- Recall that we fixed two VLAM vocal tract ages $a_0 = 10$ and $a_1 = 0.5$ in order to model a caregiver and infant dyad.
- We now complicate these models, first by assigning $J_{20}$'s VCRFs to our model caregiver, yielding an agent that has knowledge of the Japanese vowel system.
- This knowledge is modeled as the ability to pick out formant vectors that have high goodness ratings for each vowel category across vocal tract ages, including our model infant's.

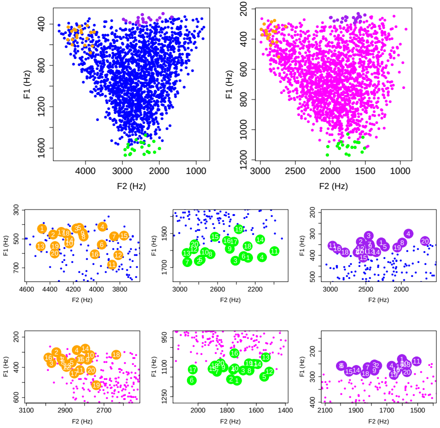
# Modeling External Data



## Corner vowel VCRFs for caregiver $J_{20}$

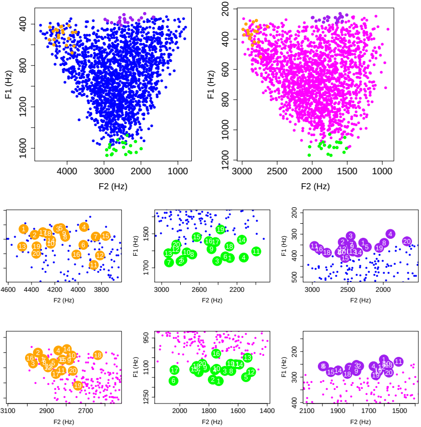
- To facilitate presentation, we'll focus on the corner vowels [i], [a], and [u].
- $J_{20}$'s VCRFs for corner vowels [i] (left), [a] (center), and [u] (right), over MFS(10) (top) and MFS(0.5) (bottom) are depicted above.

## MODELING VOCAL INTERACTION



- The corner vowel VCRFs for $J_{20}$ yield weighted pairings of formant vectors which model vocal interaction.
- The formation of the pairings occurs during idealized interaction between infant and caregiver.
- Representations of infant vowels with high goodness values are paired with caregiver vowels with high goodness values.
- Each of these response pairs is assigned a transfer weight $g$.

## MODELING VOCAL INTERACTION



**Response Pairings**

| INF | CAR | *weight* |
|---|---|---|
| $\mathbf{f}^{3}$ | $\mathbf{f}^{7}$ | $g_1$ |
| $\mathbf{f}^{6}$ | $\mathbf{f}^{2}$ | $g_2$ |
| $\mathbf{f}^{31}$ | $\mathbf{f}^{19}$ | $g_3$ |
| ⋮ | ⋮ | ⋮ |

These weighted pairs are grouped together into response pairing structures which model the collection of vocal exchanges involving turn-taking.

# Modeling Internalization

## Infant-Caregiver Dyad Modeling

We now complicate our model infant with the following computational attributes:

### Internalization

- Infants internalize the turn-taking socio-vocal exchanges with caregivers, and use these internalized exchanges to construct inchoate models of the self and of their caregivers.
- Infants also internalize more generic auditory and articulatory experience outside of the turn-taking setting, which yields broader sets of representations of their own productions and those of their caregiver.

# Modeling Cognitive Computations

## Infant-Caregiver Dyad Modeling

We now complicate our model infant with the following computational attributes:

### Cognitive Computations

- Manifolds are formed over representations within the articulatory and auditory reference frames and aligned neighbors computation aligned using the internalized socio-vocal experience.
- The resulting aligned manifolds constitute emerging models of the self and caregiver which provide a basis for further computations.
- These manifolds map articulatory and auditory representations to an intermodal reference frame where manifolds are again formed and aligned using internalized socio-vocal experience, with the aligned manifolds mapping intermodal representations to a commensuration frame where models of the self and caregiver are comparable.

# Modeling Internalization

Response Pairings

| INF | CAR | *weight* |
|---|---|---|
| $\mathbf{f}^3$ | $\mathbf{f}^7$ | $g_1$ |
| $\mathbf{f}^6$ | $\mathbf{f}^2$ | $g_2$ |
| $\mathbf{f}^{31}$ | $\mathbf{f}^{19}$ | $g_3$ |
| $\vdots$ | $\vdots$ | $\vdots$ |

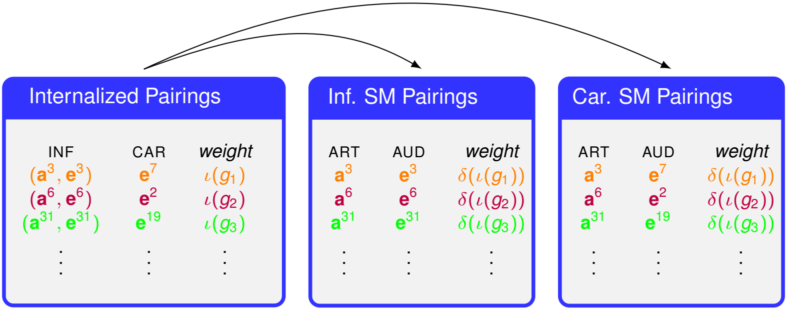
Internalized Pairings

| INF | CAR | *weight* |
|---|---|---|
| $(\mathbf{a}^3, \mathbf{e}^3)$ | $\mathbf{e}^7$ | $\iota(g_1)$ |
| $(\mathbf{a}^6, \mathbf{e}^6)$ | $\mathbf{e}^2$ | $\iota(g_2)$ |
| $(\mathbf{a}^{31}, \mathbf{e}^{31})$ | $\mathbf{e}^{19}$ | $\iota(g_3)$ |
| $\vdots$ | $\vdots$ | $\vdots$ |

## Internalized Pairing Structures

- Each response pair with weight $g$ in a response pairing structure corresponds to an internalized pair with weight $\iota(g)$.
- For example, $(\mathbf{f}^3, \mathbf{f}^7)$ with weight $g_1$ corresponds to $((\mathbf{a}^3, \mathbf{e}^3), \mathbf{e}^7)$ with weight $\iota(g_1)$ where
  - $\mathbf{f}^3 \mapsto (\mathbf{a}^3, \mathbf{e}^3)$ (an articulatory vector that produced $\mathbf{f}^3$ paired with an excitation vector derived from $\mathbf{f}^3$),
  - $\mathbf{f}^7 \mapsto \mathbf{e}^7$ (an excitation vector derived from $\mathbf{f}^7$),
  - $g_1 \mapsto \iota(g_1)$

# Modeling Internalization



### Internalized Pairings

| INF | CAR | *weight* |
|---|---|---|
| $(\mathbf{a}^3, \mathbf{e}^3)$ | $\mathbf{e}^7$ | $\iota(g_1)$ |
| $(\mathbf{a}^6, \mathbf{e}^6)$ | $\mathbf{e}^2$ | $\iota(g_2)$ |
| $(\mathbf{a}^{31}, \mathbf{e}^{31})$ | $\mathbf{e}^{19}$ | $\iota(g_3)$ |
| ⋮ | ⋮ | ⋮ |

### Inf. SM Pairings

| ART | AUD | *weight* |
|---|---|---|
| $\mathbf{a}^3$ | $\mathbf{e}^3$ | $\delta(\iota(g_1))$ |
| $\mathbf{a}^6$ | $\mathbf{e}^6$ | $\delta(\iota(g_2))$ |
| $\mathbf{a}^{31}$ | $\mathbf{e}^{31}$ | $\delta(\iota(g_3))$ |
| ⋮ | ⋮ | ⋮ |

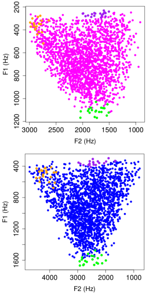
### Car. SM Pairings

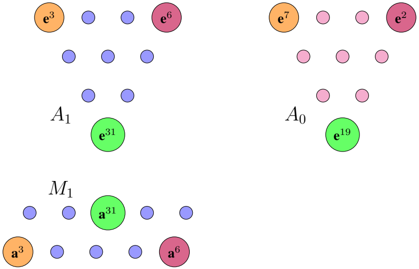
| ART | AUD | *weight* |
|---|---|---|
| $\mathbf{a}^3$ | $\mathbf{e}^7$ | $\delta(\iota(g_1))$ |
| $\mathbf{a}^6$ | $\mathbf{e}^2$ | $\delta(\iota(g_2))$ |
| $\mathbf{a}^{31}$ | $\mathbf{e}^{19}$ | $\delta(\iota(g_3))$ |
| ⋮ | ⋮ | ⋮ |

## Creating Sensorimotor Pairing Structures

- Each internalized pair with weight $\iota(g)$ in an internalized pairing structure yields two sensorimotor pairs with weight $\delta(\iota(g))$.
- For example, $((\mathbf{a}^3, \mathbf{e}^3), \mathbf{e}^7)$ with weight $\iota(g_1)$ yields the following pairs:
  - $(\mathbf{a}^3, \mathbf{e}^3)$ (infant sensorimotor pair),
  - $(\mathbf{a}^3, \mathbf{e}^7)$ (caregiver sensorimotor pair),
  - $\iota(g_1) \mapsto \delta(\iota(g_1))$ (sensorimotor weight).

# Modeling Internalization

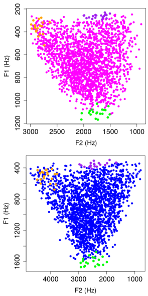


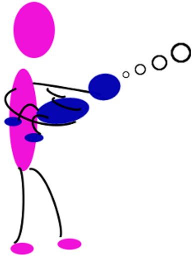


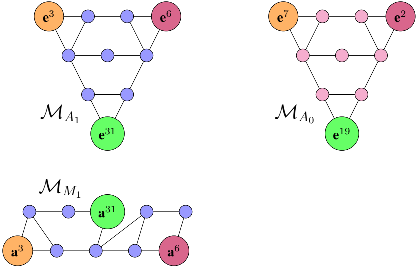
## Broad sets of internal representations

- Broad auditory and articulatory experience of the self is modeled in terms of subsets of MARS(0.5) and MAUD(0.5), while broad auditory experience of the caregiver is modeled in terms of a subset of MAUD(10).
- These subsets may also include representations derived from internal models that provide the means to approximate representations in reference frames already populated with representations derived from experience.
- Representations within these subsets are heuristically depicted above together with those derived from turn-taking exchanges in sets denoted $A_0$, $A_1$, and $M_1$.

# Modeling Cognitive Computations



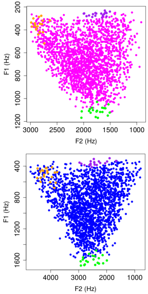


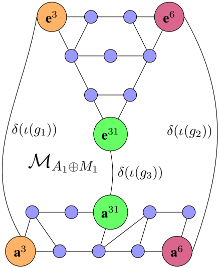


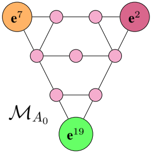
## Formation of auditory and articulatory manifolds

- Given a set $X$, a manifold $M_X = (V_X, E_X)$ is formed over $X$ by
  - bijectively mapping $X$ to a set of vertices $V_X$,
  - forming an edge set $E_X$ over $V_X$ using a $k$-nearest neighbors computation on $X$.
- The auditory manifolds $\mathcal{M}_{A_0}$ and $\mathcal{M}_{A_1}$, and the articulatory manifold $\mathcal{M}_{M_1}$ are respectively formed over the sets $A_0$, $A_1$, and $M_1$ in this fashion.
- At this stage the relational structure of the manifolds encodes the local relationships between representations derived from the infant, within each reference frame, and those derived from the caregiver.

# Modeling Cognitive Computations



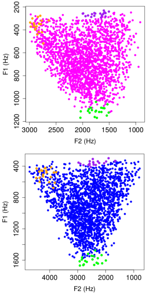


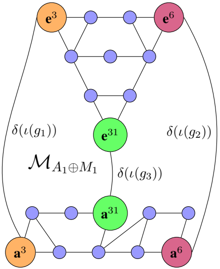


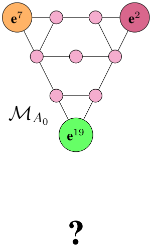
## Sensorimotor models of the self and others

- The auditory manifold $\mathcal{M}_{A_1}$ and the articulatory manifold $\mathcal{M}_{M_1}$ are aligned using the infant sensorimotor pairing to form the sensorimotor manifold $\mathcal{M}_{A_1 \oplus M_1}$, which constitutes a model of the self.
- The auditory manifold $\mathcal{M}_{A_0}$ and the articulatory manifold $\mathcal{M}_{M_1}$ are aligned using the caregiver sensorimotor pairing to form the sensorimotor manifold $\mathcal{M}_{A_0 \oplus M_1}$, which constitutes a model of the caregiver.

# Modeling Cognitive Computations



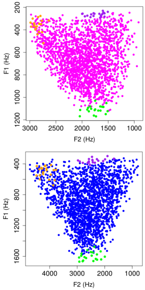


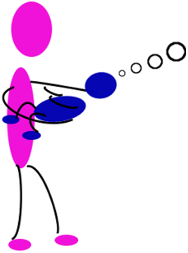


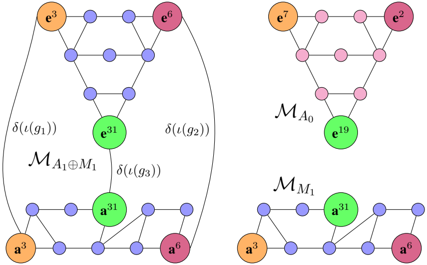
## Sensorimotor models of the self and others

- The auditory manifold $\mathcal{M}_{A_1}$ and the articulatory manifold $\mathcal{M}_{M_1}$ are aligned using the infant sensorimotor pairing to form the sensorimotor manifold $\mathcal{M}_{A_1 \oplus M_1}$, which constitutes a model of the self.
- The auditory manifold $\mathcal{M}_{A_0}$ and the articulatory manifold $\mathcal{M}_{M_1}$ are aligned using the caregiver sensorimotor pairing to form the sensorimotor manifold $\mathcal{M}_{A_0 \oplus M_1}$, which constitutes a model of the caregiver.

# Modeling Cognitive Computations



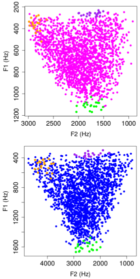


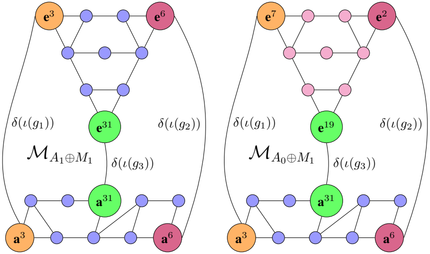


## Sensorimotor models of the self and others

- The auditory manifold $\mathcal{M}_{A_1}$ and the articulatory manifold $\mathcal{M}_{M_1}$ are aligned using the infant sensorimotor pairing to form the sensorimotor manifold $\mathcal{M}_{A_1 \oplus M_1}$, which constitutes a model of the self.
- The auditory manifold $\mathcal{M}_{A_0}$ and the articulatory manifold $\mathcal{M}_{M_1}$ are aligned using the caregiver sensorimotor pairing to form the sensorimotor manifold $\mathcal{M}_{A_0 \oplus M_1}$, which constitutes a model of the caregiver.

# Modeling Cognitive Computations

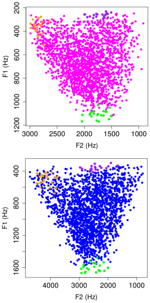


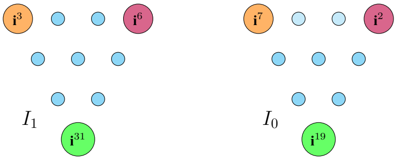


## Sensorimotor models of the self and others

- The auditory manifold $\mathcal{M}_{A_1}$ and the articulatory manifold $\mathcal{M}_{M_1}$ are aligned using the infant sensorimotor pairing to form the sensorimotor manifold $\mathcal{M}_{A_1 \oplus M_1}$, which constitutes a model of the self.
- The auditory manifold $\mathcal{M}_{A_0}$ and the articulatory manifold $\mathcal{M}_{M_1}$ are aligned using the caregiver sensorimotor pairing to form the sensorimotor manifold $\mathcal{M}_{A_0 \oplus M_1}$, which constitutes a model of the caregiver.

# Modeling Cognitive Computations

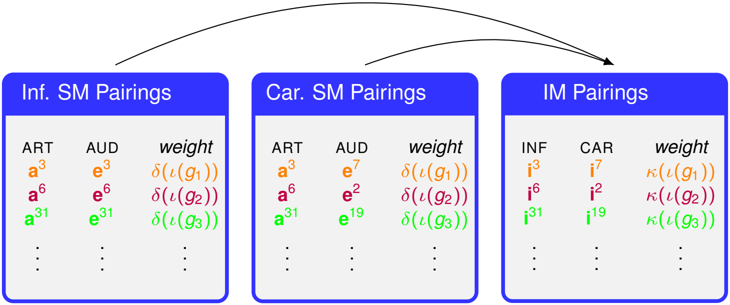




## Computing Intermodal Representations

- The infant's models of the self, $\mathcal{M}_{A_1 \oplus M_1}$, and caregiver $\mathcal{M}_{A_0 \oplus M_1}$, serve as input to a Laplacian eigenmapping (Belkin & Niyogi, 2003) that maps the auditory and articulatory representations that these models are built from to newly generated intermodal representations within an intermodal reference frame.
- That is, the sets $A_0$ and $A_1$ are eigenmapped to sets of corresponding intermodal representations $I_0$ and $I_1$, depicted heuristically above, based on the models $\mathcal{M}_{A_0 \oplus M_1}$ and $\mathcal{M}_{A_1 \oplus M_1}$.

# Modeling Cognitive Computations



Inf. SM Pairings

| ART | AUD | *weight* |
|---|---|---|
| $\mathbf{a}^3$ | $\mathbf{e}^3$ | $\delta(\iota(g_1))$ |
| $\mathbf{a}^6$ | $\mathbf{e}^6$ | $\delta(\iota(g_2))$ |
| $\mathbf{a}^{31}$ | $\mathbf{e}^{31}$ | $\delta(\iota(g_3))$ |
| $\vdots$ | $\vdots$ | $\vdots$ |

Car. SM Pairings

| ART | AUD | *weight* |
|---|---|---|
| $\mathbf{a}^3$ | $\mathbf{e}^7$ | $\delta(\iota(g_1))$ |
| $\mathbf{a}^6$ | $\mathbf{e}^2$ | $\delta(\iota(g_2))$ |
| $\mathbf{a}^{31}$ | $\mathbf{e}^{19}$ | $\delta(\iota(g_3))$ |
| $\vdots$ | $\vdots$ | $\vdots$ |

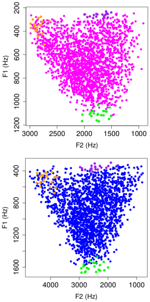
IM Pairings

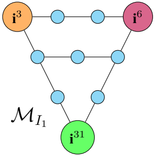
| INF | CAR | *weight* |
|---|---|---|
| $\mathbf{i}^3$ | $\mathbf{i}^7$ | $\kappa(\iota(g_1))$ |
| $\mathbf{i}^6$ | $\mathbf{i}^2$ | $\kappa(\iota(g_2))$ |
| $\mathbf{i}^{31}$ | $\mathbf{i}^{19}$ | $\kappa(\iota(g_3))$ |
| $\vdots$ | $\vdots$ | $\vdots$ |

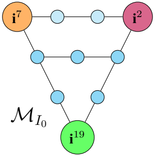
## Creating Intermodal Pairing Structures

- Based on the eigenmapping, each pair of sensorimotor pairs with weight $\delta(\iota(g))$ within a pair of sensorimotor pairing structures yields an intermodal pair with weight $\kappa(\iota(g))$.
- For example, the pairs $(\mathbf{a}^3, \mathbf{e}^3)$ and $(\mathbf{a}^3, \mathbf{e}^7)$ with weight $\delta(\iota(g_1))$ yield the following:
  - $((\mathbf{a}^3, \mathbf{e}^3), (\mathbf{a}^3, \mathbf{e}^7)) \mapsto (\mathbf{i}^3, \mathbf{i}^7)$ (an intermodal pair),
  - $\delta(\iota(g_1)) \mapsto \kappa(\iota(g_1))$ (intermodal weight).

# MODELING COGNITIVE COMPUTATIONS



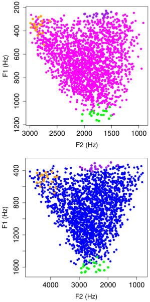


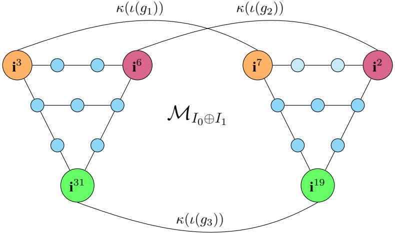


## INTERMODAL MODELS AND COMMENSURATION MANIFOLDS

- Intermodal manifolds $\mathcal{M}_{I_0}$ and $\mathcal{M}_{I_1}$ are constructed over intermodal representations $I_0$ and $I_1$ using the same graph formation computation that yields auditory and articulatory manifolds.
- The intermodal manifolds $\mathcal{M}_{I_0}$ and $\mathcal{M}_{I_1}$ are aligned using the intermodal pairing to form the commensuration manifold $\mathcal{M}_{I_0 \oplus I_1}$, which reflects an emerging socio-cognitive convergence between the model of the self and the caregiver.

# MODELING COGNITIVE COMPUTATIONS

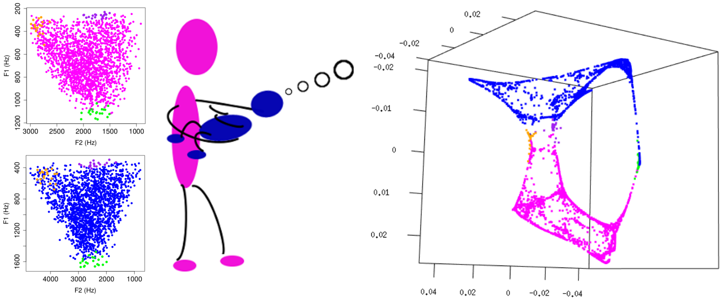




## INTERMODAL MODELS AND COMMENSURATION MANIFOLDS

- Intermodal manifolds $\mathcal{M}_{I_0}$ and $\mathcal{M}_{I_1}$ are constructed over intermodal representations $I_0$ and $I_1$ using the same graph formation computation that yields auditory and articulatory manifolds.
- The intermodal manifolds $\mathcal{M}_{I_0}$ and $\mathcal{M}_{I_1}$ are aligned using the intermodal pairing to form the commensuration manifold $\mathcal{M}_{I_0 \oplus I_1}$, which reflects an emerging socio-cognitive convergence between the model of the self and the caregiver.

# Modeling Cognitive Computations



## Computing Commensuration Representations

- The commensuration manifold serves as input to an eigemapping that maps the intermodal representations to newly generated commensuration representations within a commensuration reference frame.
- These commensuration representations can be used in other computations, e.g., those that compute vowel categorization, or in further cognitive computations.

# Summary

## Merits of the Approach

- Vowel Category Response Fields
  - VCRFs allow us to make within and cross-language comparisons concerning each subject's knowledge of the vowel system of their native language.
  - Quantitative analysis in Plummer et al. (2013) shows that "distances" between VCRFs can capture cross-language differences between the five-vowel systems of Greek and Japanese, and within-language sociolinguistic differences concerning Japanese /u/.
- Manifolds and Manifold Alignment
  - While the computational cognitive framework provides inroads toward quantitative analysis, e.g., parameter space testing, comparison of different algorithmic instantiations of the computations, etc., its principle role at present is at the conceptual-interpretive level.
  - Manifolds and manifold alignment focus attention on the rich generative computations that takes place during early infancy in the creation of representations of both the self and of others, bringing new relevance to ideas suggested by early social psychologists concerning the phatic aspects of language acquisition.
  - Moreover, the conceptualization provides the basis for reasoning about aspects of phonological acquisition that have yet to truly be brought to light, e.g., the acquisition of cognitive structures for representing vowel dynamics.

# SUMMARY

## TAKE HOME POINTS

- There's plenty of room at the beginning:

  `www.youtube.com/watch?v=4eRCygdW--c`

- Progress/success is more variegated than we think:

  `http://www.youtube.com/watch?v=27X8NFHxuFk`

# Thank you

## Acknowledgements

We wish to thank

- Asimina Syrika and Kiyoko Yoneyama for recruiting and testing the Greek- and Japanese-speaking participants,
- Hyunju Chung and Chanelle Mays for recruiting and testing the Korean- and the English-speaking participants,
- Catherine McBride-Chang for providing the resources to recruit the Cantonese-speaking subjects,
- Mary Beckman and Ben Munson for numerous important contributions to various aspects of this presentation.
- Lucie Ménard, Pat Riedy, and Jeff Holliday for various supporting technical components of the computational modeling framework, and
- Eric Fosler-Lussier and Misha Belkin for providing initial direction concerning the computational modeling framework.
- Work supported by NSF grants BCS-0729306 and BCS-0729277.

Ananthakrishnan, G., & Salvi, G. (2011). Using imitation to learn infant–adult acoustic mappings. In *Proceedings of INTERSPEECH 2011*, (pp. 765–768).

Beck, J. M. (1996). Organic variation of the vocal apparatus. In W. J. Hardcastle, & J. Laver (Eds.) *Handbook of Phonetic Sciences*, (pp. 256–297). Cambridge, England: Blackwell.

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, *15*(6), 1373–1396.

Boë, L.-J., & Maeda, S. (1998). Modélization de la croissance du conduit vocal. Éspace vocalique des nouveaux–nés et des adultes. Conséquences pour l'ontegenèse et la phylogenèse. In *Journée d'Études Linguistiques: "La Voyelle dans Tous ces États"*, (pp. 98–105). Nantes, France.

Boë, L.-J., Perrier, P., Guérin, B., & Schwartz, J.-L. (1989). Maximal vowel space. In *EUROSPEECH 09*, (pp. 281–284). Paris, France.

Browman, C., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*, 201–251.

Browman, C., & Goldstein, L. (1990a). Gestural specification using dynamically–defined articulatory structures. *Journal of Phonetics*, *18*(3), 299–320.

Browman, C., & Goldstein, L. (1990b). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston, & M. E. Beckman (Eds.) *Papers in Laboratory Phonology: Vol. 1. Between the Grammar and Physics of Speech*, (pp. 341–376). Cambridge, England: Cambridge University Press.

Buja, A., Hastie, T., & Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, *17*(2), 453–510.

Chiba, T., & Kajiyama, M. (1941). *The Vowel, its Nature and Structure*. Tokyo: Tokyo–Kaiseikan Publishing Company, Ltd.

Crockford, C., & Boesch, C. (2005). Call combinations in wild chimpanzees. *Behaviour*, *142*(4), 397–421.

de Boer, B. (2000). Self–organization in vowel systems. *Journal of Phonetics*, *28*(4), 441–465.

Fitch, W. T. (2004). Evolving honest communication systems: Kin selection and "mother tongues". In D. K. Oller, & U. Griebel (Eds.) *Evolution of Communication Systems: A Comparative Approach*, (pp. 275–296). Cambridge, Massachusetts: MIT Press.

Friedman, J. H., & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American statistical Association*, *76*(376), 817–823.

Gallese, V. (2001). The 'shared manifold' hypothesis. from mirror neurons to empathy. *Journal of Consciousness Studies*, *8*(5–7), 33–50.

Genty, E., Clay, Z., Hobaiter, C., & Zuberbühler, K. (2014). Multi-modal use of a socially directed call in bonobos. *PLoS-ONE*, *9*(1), e84738.

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*(5), 515–523.

Goldstein, M. H., Schwade, J. A., & Bornstein, M. H. (2009). The value of vocalizing: Five–month–old infants associate their own noncry vocalizations with responses from caregivers. *Child development*, *80*(3), 636–644.

Goldstein, U. G. (1980). *An articulatory model of the vocal tract of the growing child*. Ph.D. thesis, Massachusetts Institute of Technology.

Gros-Louis, J., West, M. J., Goldstein, M. H., & King, A. P. (2006). Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development*, *30*(5), 112–119.

Gu, C. (2002). *Smoothing spline ANOVA models*. Springer.

Ham, J., Lee, D. D., & Saul, L. K. (2005). Semisupervised alignment of manifolds. In Z. Ghahramani, & R. Cowell (Eds.) *Proc. of the Ann. Conf. on Uncertainty in AI*, vol. 10, (pp. 120–127).

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized Additive Models*. New York: Chapman and Hall.

Hsu, H. C., Iyer, S. N., & Fogel, A. (2013). Effects of social games on infant vocalizations. *Journal of Child Language*, *41*(1), 1–23.

Ishihara, H., Yoshikawa, Y., Miura, K., & Asada, M. (2009). How caregiver's anticipation shapes infant's vowel through mutual imitation. *Autonomous Mental Development, IEEE Transactions on*, *1*(4), 217 –225.

Joos, M. (1948). Acoustic phonetics. *Language*, *24*(2), 5–136.

Lewkowicz, D. J., & Ghazanfar, A. A. (2009). The emergence of multisensory systems through perceptual narrowing. *Trends in cognitive sciences*, *13*(11), 470–478.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, *74*(6), 431.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36.

Lindblom, J., & Sundberg, J. E. F. (1971). Acoustical consequences of lip, tongue, jaw, and larynx movement. *Journal of the Acoustical Society of America*, *50*, 1166–179.

Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal–tract shapes using an articulatory model. In W. Hardcastle, & A. Marchal (Eds.) *Speech Production and Speech Modeling*, (pp. 131–149). The Netherlands: Kluwer Academic Publishers.

Maeda, S. (1991). On articulatory and acoustic variabilities. *Journal of Phonetics*, *19*, 321–331.

Masataka, N. (2003). *The Onset of Language*. Cambridge, UK: Cambridge University Press.

Massaro, D. W., & Cohen, M. M. (1983). Categorical or continuous speech perception: A new test. *Speech Communication*, *2*, 15–35.

Meltzoff, A. (2007). The ‘like me’ framework for recognizing and becoming an intentional agent. *Acta Psychologica*, *124*, 26–43.

Meltzoff, A. N., & Kuhl, P. K. (1994). Faces and speech: Iintermodal processing of biologically relevant signals in infants and adults. In D. J. Lewkowicz, & R. Lickliter (Eds.) *The Development of Intersensory Perception: Comparative Perspectives*, (pp. 335–369). Lawrence Erlbaum Associates, Inc.

Ménard, L., & Boë, L.-J. (2000). Exploring vowel production strategies from infant to adult by means of articulatory inversion of formant data. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, (pp. 465–468). Beijing, China.

Ménard, L., Schwartz, J.-L., & Boë, L.-J. (2002). Auditory normalization of French vowels synthesized by an articulatory model simulating growth from birth to adulthood. *Journal of the Acoustical Society of America*, *111*(4), 1892–1905.

Miller, J. L. (1994). On the internal structure of phonetic categories: A progress report. *Cognition*, *50*(1–3), 271–284.

Miller, J. L. (1997). Internal structure of phonetic categories. *Language and cognitive processes*, *12*(5/6), 865–869.

Miura, K., Yoshikawa, Y., & Asada, M. (2012). Vowel acquisition based on an auto–mirroring bias with a less imitative caregiver. *Advanced Robotics*, *26*(1–2), 23–44.

Moore, B. C. J. (1997). *An Introduction to the Psychology of Hearing*. Academic Press.

Moore, B. C. J., & Glasberg, B. G. (1996). A revision of Zwicker's loudness model. *Acta Acoustica*, *82*, 335–345.

Munson, B., Ménard, L., Beckman, M. E., Edwards, J., & Chung, H. (2010). Sensorimotor maps and vowel development in English, Greek, and Korean: A cross–linguistic perceptual categorizaton study (A). *Journal of the Acoustical Society of America*, *127*, 2018.

Niyogi, P. (2004). Towards a computational model of human speech perception. In *Proceedings of the Conference on Sound to Sense, MIT (In Honor of Ken Stevens' 80th birthday)*.

Plummer, A. R. (2014). *The Acquisition of Vowel Normalization: Theory and Computational Framework*. Ph.D. thesis, The Ohio State University.

Plummer, A. R., Ménard, L., Munson, B., & Beckman, M. E. (2013). Comparing vowel category response surfaces over age–varying maximal vowel spaces within and across language communities. In *Proceedings of INTERSPEECH 2013*.

Rasilo, H., Räsänen, O., & Laine, U. K. (2013). Feedback and imitation by a caregiver guides a virtual infant to learn native phonemes and the skill of speech inversion. *Speech Communication*, *55*(9), 909–931.

Schwartz, J.-L., Boë, L.-J., & Abry, C. (2007). Linking dispersion–focalization theory and the maximum utilization of the available distinctive features principle in a perception–for–action–control theory. In M.-J. Sole, P. S. Beddor, & M. Ohala (Eds.) *Experimental Approaches to Phonology*, (pp. 104–124). Oxford University Press.

Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997a). The dispersion–focalization theory of vowel systems. *Journal of Phonetics*, *25*, 255–286.

Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997b). Major trends in vowel system inventories. *Journal of Phonetics*, *25*, 233–253.

Seung, H. S., & Lee, D. D. (2000). The manifold ways of perception. *Science*, *290*(5500), 2268–2269.

Vallée, N., Boë, L.-J., & Payan, Y. (1995). Vowel prototypes for UPSID's 33 phonemes. In *Proceedings of ICPhS 2*, (pp. 424–427). Stockholm.

Wahba, G. (1990). *Spline models for observational data*. Society for Industrial and Applied Mathematics.

Wang, C. (2010). *A Geometric Framework For Transfer Learning Using Manifold Alignment*. Ph.D. thesis, University of Mass. Amherst.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC.

Wood, S. N. (2003). Thin–plate regression splines. *Journal of the Royal Statistical Society (B)*, *65*(1), 95–114.