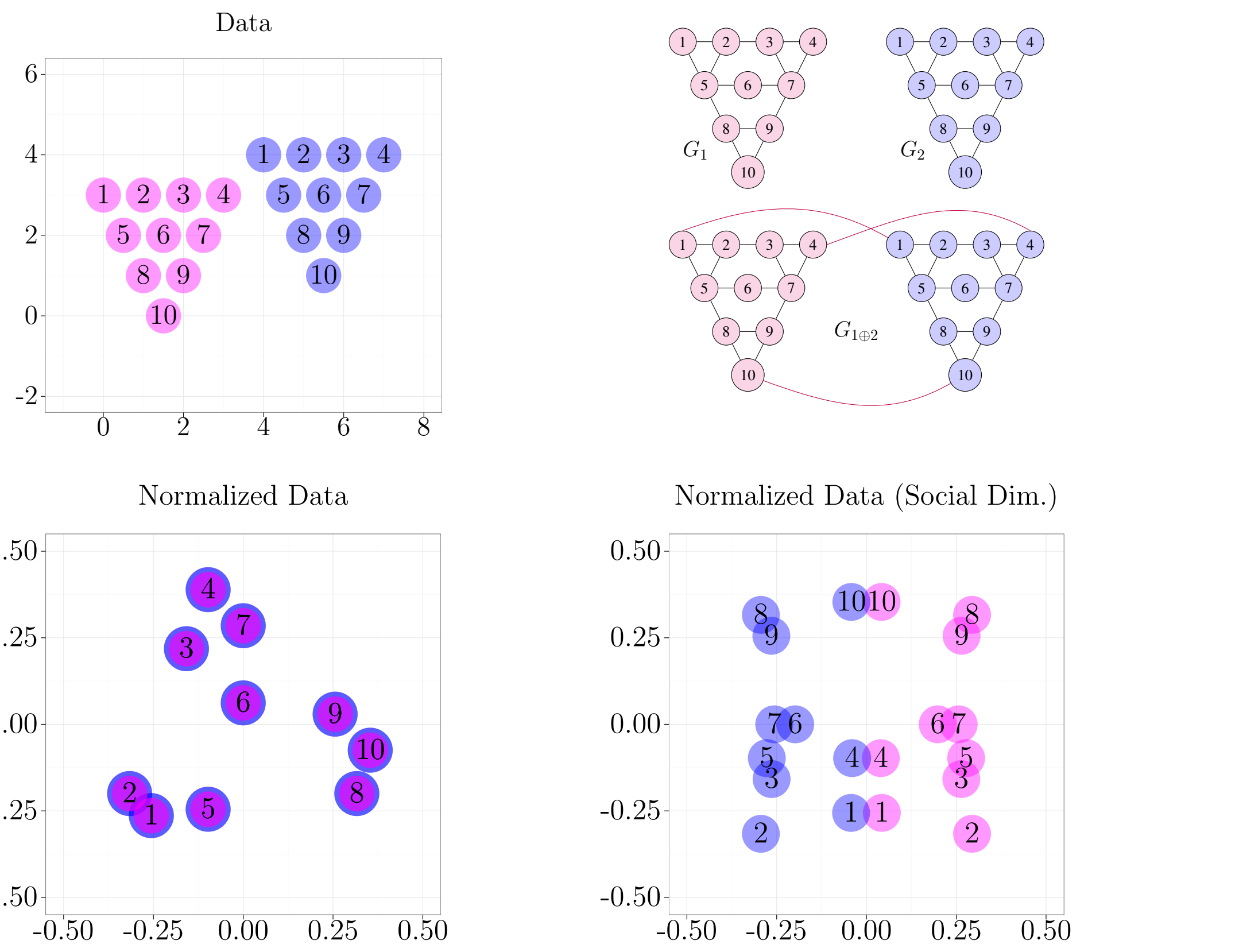
Modeling the acquisition of vowel normalization as cognitive manifold alignment

Andrew R. Plummer
The Ohio State University, Columbus, OH, USA

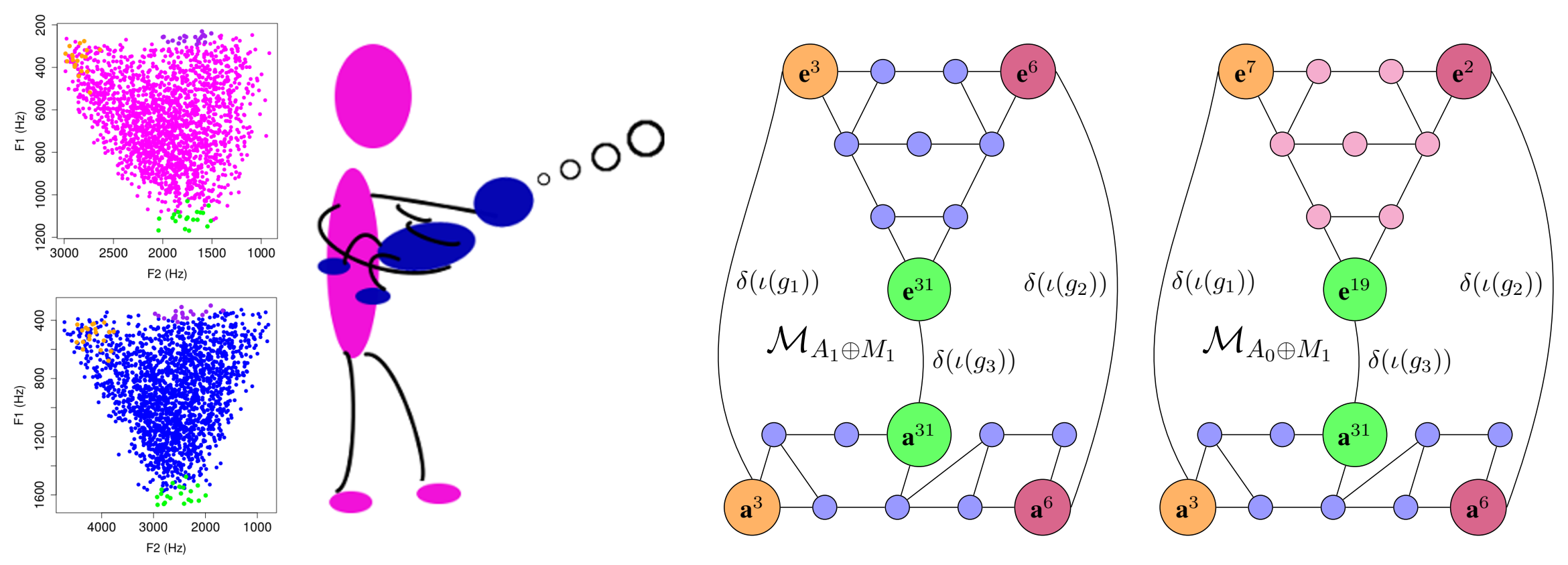
Speech Recognition Virtual Kitchen

Introduction



- Within the context of phonological acquisition, vowel normalization is typically construed as a fixed, reductive computational prelude to vowel category acquisition.
- Yet, research over the last few decades suggests that, to the contrary, vowel normalization is:
  - acquired over the course of ontogeny,
  - sensitive to cross-language differences, and
  - generative in nature, rather than reductive.
- We take vowel normalization to be an acquired computation involving an infant's generation of models of the self and others, and relations between models, based on language-specific vocal interaction with caretakers during early infancy.
- In this presentation we aim to take a first step in identifying and characterizing structural and representational aspects that reflect the influence of language-specific vocal interaction on the acquisition of vowel normalization.

The Acquisition of Vowel Normalization

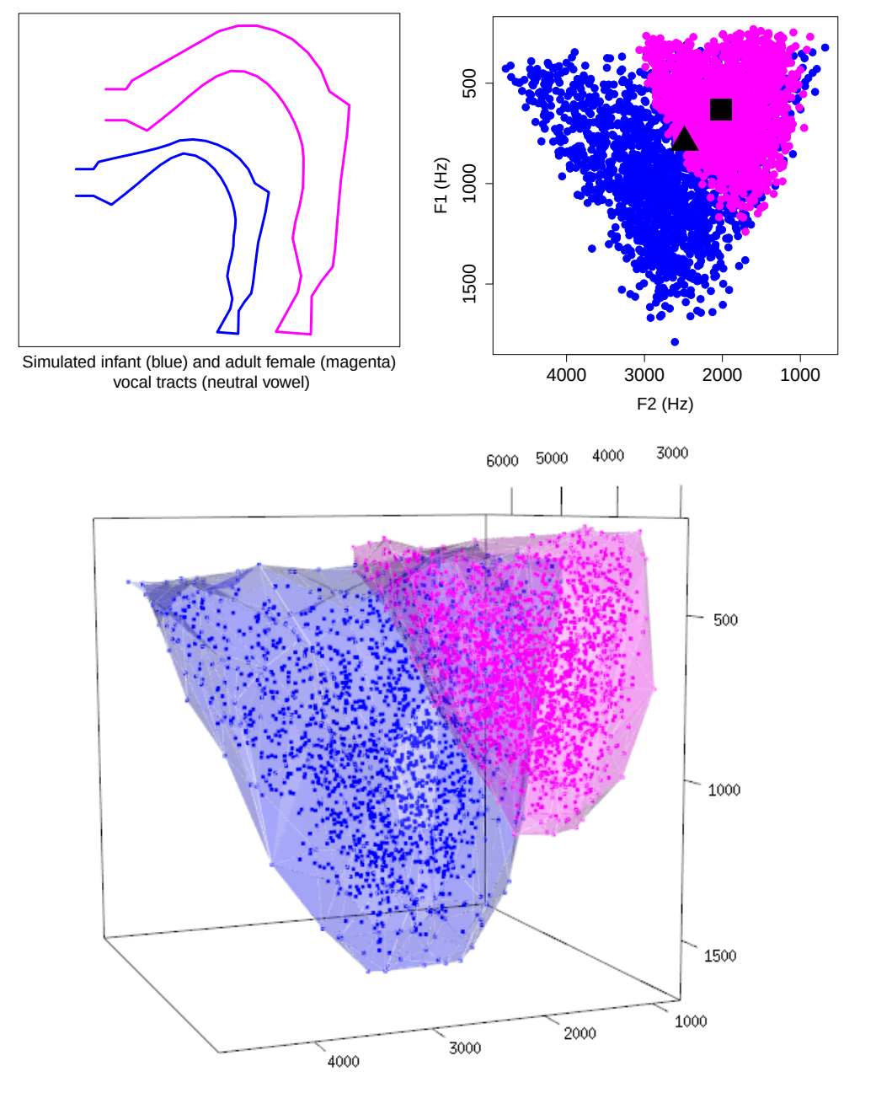


- Manifolds are topological structures which potentially aid in the organization of perception (Seung and Lee, 2000) and speech production (Saltzman, Kubo, and Tsao, 2006).
- We take infants' models of the self and others to be cognitive manifolds formed over auditory and articulatory representations, as well as derived intermodal and other higher-order representations.
- We take infants' formation of relations between manifolds to be a cognitive alignment computation which combines two (or more) manifolds into a new aligned manifold based on internalized vocal interaction with a caretaker.
- Aligned manifolds provide the basis for computing higher-order representations, which may reflect multisensory narrowing (Lewkowicz and Ghazanfar, 2009) and serve as input to other computations, e. g., vowel category acquisition.

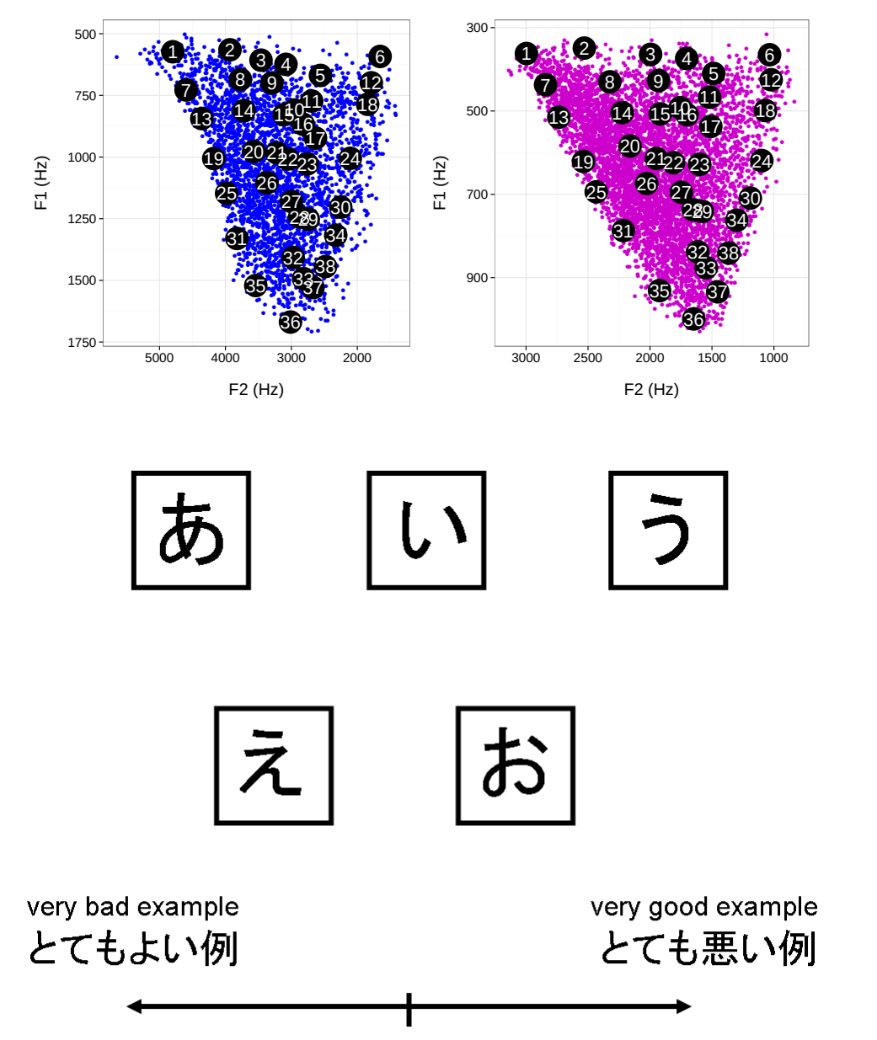
Theoretical and Computational Modules of the Modeling Framework

Maximal Vowel Spaces

- The VLAM (Boë and Maeda, 1998) is a computational model of the articulatory system and its speech production capacities.
- The VLAM is age-varying and capable of representing vocal tract lengths ranging from those of infants to young adults
- Given an age in years, the set of all articulatory configurations of the VLAM at that age that do not result in occlusion of the oral cavity yield a maximal vowel space (Schwartz et al., 2007) for that age.
- We take each MVS to be identified by a set of articulatory vectors, each of which has a corresponding formant pattern which, in turn, is characterized as a formant vector whose components are the first three formant frequencies of that formant pattern.


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

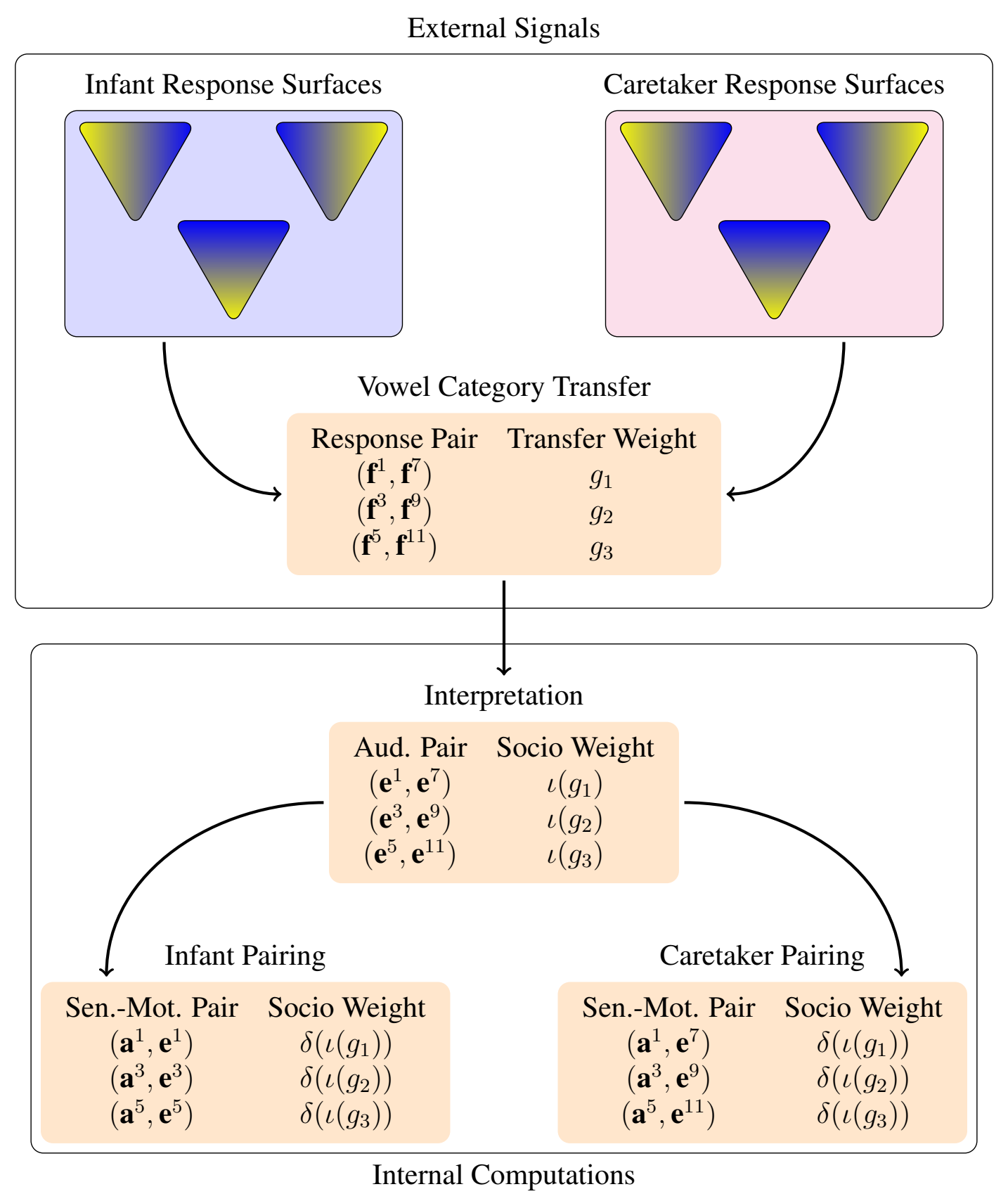
Perceptual Categorization Experiments



- 38 vowel stimuli were generated by the VLAM for each of seven ages, including 6 months (left, blue) and 10 years (left, magenta), situated within the appropriate maximal vowel space producible by the model at the corresponding age.
- Each of the stimuli for each of the vocal tract ages was categorized by members of five different language communities: Cantonese (n=15), American English (n=21), Greek (n=21), Japanese (n=21), and Korean (n=20).
- Each listener assigned each stimulus a vowel category from the listener's native language, along with a "goodness rating" (Miller, 1994, 1997) indicating how good the listener felt that stimulus was as an example of the assigned category.

Vocal Exchange and Cognitive Pairing

- Formant vector representations of each vowel signal (**f**) are assigned goodness values reflecting a caretaker's intuitions about the categorical status of the signal within the caretaker's vowel system.
- Representations of infant vowels with high goodness values are paired with caretaker vowels with high goodness values. Each of these response pairs is assigned a transfer weight *g*.
- Formant pattern pairs are internalized as pairs of excitation patterns (**e**), each of which is assigned a socio-auditory weight $\iota(g)$.
- These socio-auditory pairs in turn yield two sets of sensorimotor pairs. Each sensorimotor pair is composed of an excitation pattern and an articulatory representation (**a**), and is assigned a socio-sensorimotor weight $\delta(\iota(g))$.
- Each set of sensorimotor pairs represents the infant's creation of a preliminary representation of a social agent in the infant's vocal learning environment.


External Signals

Infant Response Surfaces

Caretaker Response Surfaces

Vowel Category Transfer

Response Pair	Transfer Weight
$(\mathbf{f}^1, \mathbf{f}^7)$	g_1
$(\mathbf{f}^3, \mathbf{f}^9)$	g_2
$(\mathbf{f}^5, \mathbf{f}^{11})$	g_3

Interpretation

Aud. Pair	Socio Weight
$(\mathbf{e}^1, \mathbf{e}^7)$	$\iota(g_1)$
$(\mathbf{e}^3, \mathbf{e}^9)$	$\iota(g_2)$
$(\mathbf{e}^5, \mathbf{e}^{11})$	$\iota(g_3)$

Infant Pairing

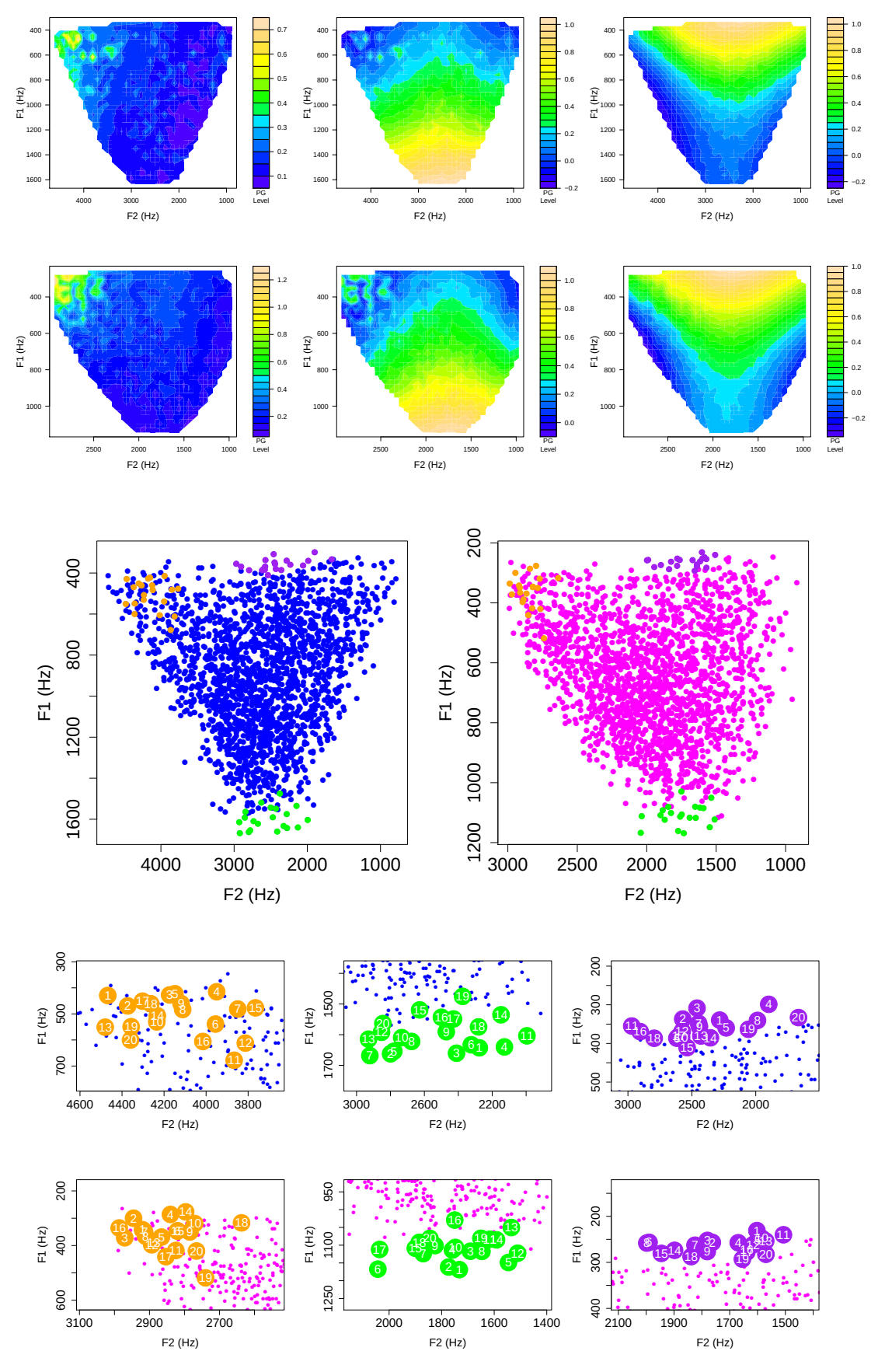
Sen.-Mot. Pair	Socio Weight
$(\mathbf{a}^1, \mathbf{e}^1)$	$\delta(\iota(g_1))$
$(\mathbf{a}^3, \mathbf{e}^3)$	$\delta(\iota(g_2))$
$(\mathbf{a}^5, \mathbf{e}^5)$	$\delta(\iota(g_3))$

Caretaker Pairing

Sen.-Mot. Pair	Socio Weight
$(\mathbf{a}^1, \mathbf{e}^7)$	$\delta(\iota(g_1))$
$(\mathbf{a}^3, \mathbf{e}^9)$	$\delta(\iota(g_2))$
$(\mathbf{a}^5, \mathbf{e}^{11})$	$\delta(\iota(g_3))$

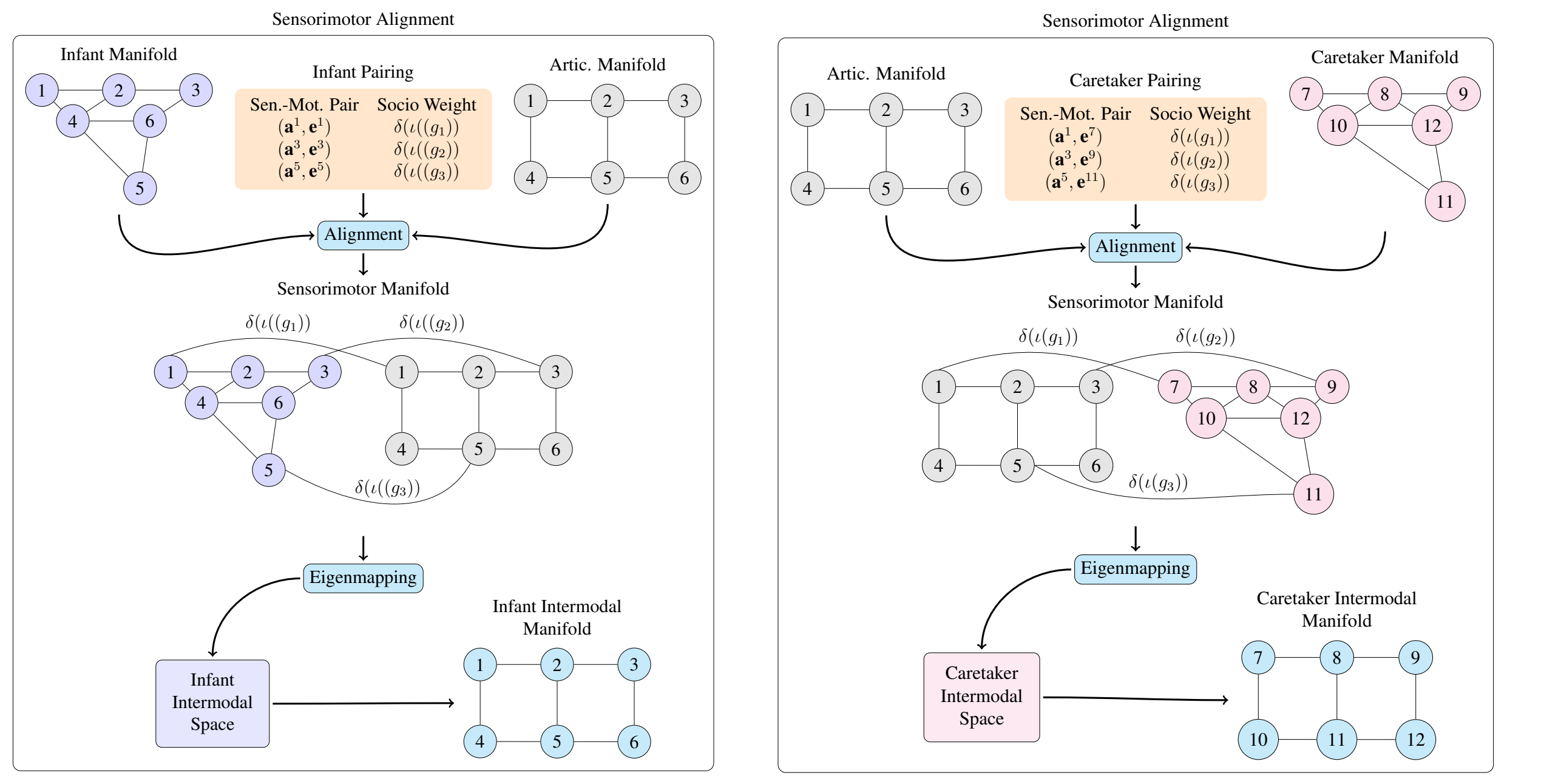
Internal Computations

Vowel Category Response Surface Functions



- Goodness values are modeled using a statistical methodology based on analysis of a set of cross-language vowel categorization experiments (Munson et al., 2010).
- The statistical methodology, based on a smoothing spline approach (Wahba, 1990, Gu, 2002) to additive modeling (see Hastie and Tibshirani, 1990), provides a set of vowel category response surfaces over the MVS for each age, based on a listener's identification responses and associated goodness ratings for the 38 stimuli.
- The surfaces to the left for vowels [i,a,u] for ages 6 months (first row) and 10 years (second row) are derived from goodness ratings provided by a Japanese subject.
- Vowel category response surfaces for a given subject over the 6 month old and 10 year old MVSs provide a model of vocal exchanges between a caretaker and infant.
- For each category in the caretaker's language, pairs are formed over formant vectors with high goodness ratings from the 6 month old MVS and the 10 year old MVS, and assigned a goodness value *g* based on the ratings.
- The response pairs are internalized by the infant as socio-auditory pairs, each with a socio-auditory weight $\iota(g)$.
- The socio-auditory pairs yield two sets of sensorimotor pairs, where each pair is assigned a socio-sensorimotor weight $\delta(\iota(g))$.

Sensorimotor and Intermodal Manifold Alignment


Sensorimotor Alignment

Infant Manifold

Infant Pairing

Sen.-Mot. Pair	Socio Weight
$(\mathbf{a}^1, \mathbf{e}^1)$	$\delta(\iota((g_1))$
$(\mathbf{a}^3, \mathbf{e}^3)$	$\delta(\iota((g_2))$
$(\mathbf{a}^5, \mathbf{e}^5)$	$\delta(\iota((g_3))$

Artic. Manifold

Alignment

Sensorimotor Manifold

$\delta(\iota((g_1))$ $\delta(\iota((g_2))$ $\delta(\iota((g_3))$

Eigenmapping

Infant Intermodal Space

Infant Intermodal Manifold

Sensorimotor Alignment

Artic. Manifold

Caretaker Pairing

Sen.-Mot. Pair	Socio Weight
$(\mathbf{a}^1, \mathbf{e}^7)$	$\delta(\iota(g_1))$
$(\mathbf{a}^3, \mathbf{e}^9)$	$\delta(\iota(g_2))$
$(\mathbf{a}^5, \mathbf{e}^{11})$	$\delta(\iota(g_3))$

Caretaker Manifold

Alignment

Sensorimotor Manifold

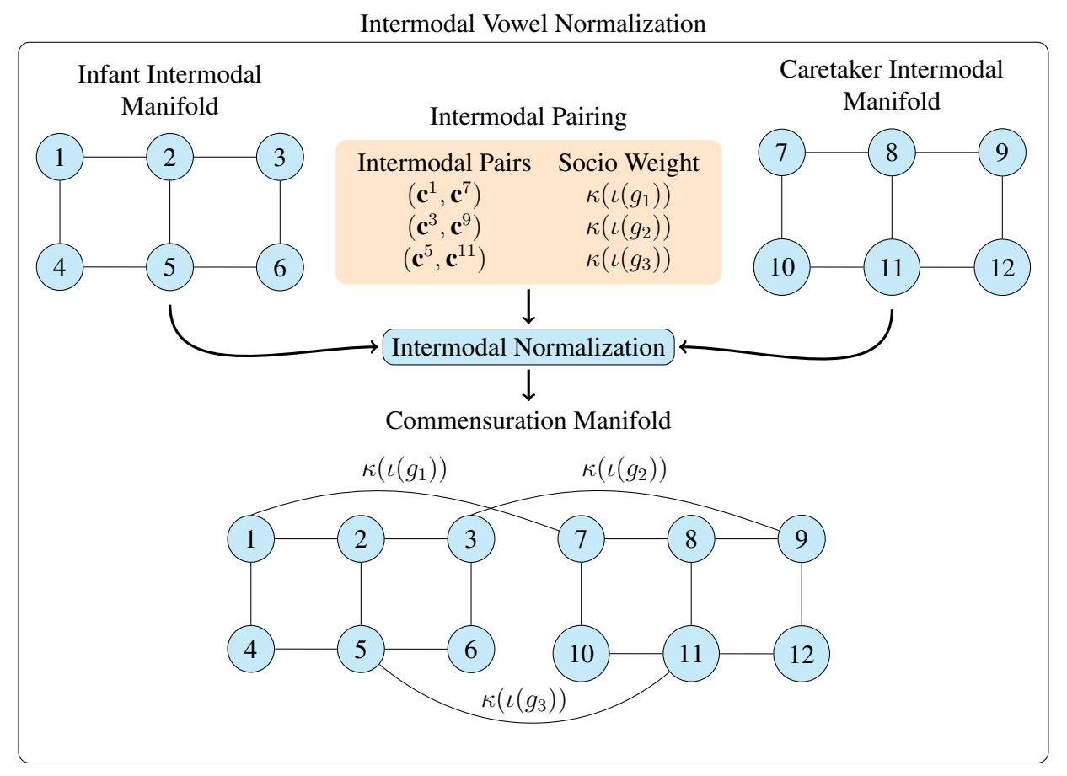
$\delta(\iota(g_1))$ $\delta(\iota(g_2))$ $\delta(\iota(g_3))$

Eigenmapping

Caretaker Intermodal Space

Caretaker Intermodal Manifold

- Manifolds formed over representations within the articulatory and auditory domains are aligned using the weighted sensorimotor pairings. These sensorimotor manifolds yield intermodal representations (**c**) of the articulatory representations and excitation patterns.
- The intermodal representations provide intermodal pairs corresponding to the internalized socio-auditory pairs, where each pair is assigned a socio-intermodal weight $\kappa(\iota(g))$.
- Manifolds formed over intermodal representations within the intermodal domain are aligned using the weighted intermodal pairings, providing a commensuration structure used for vowel categorization, inter alia.


Intermodal Vowel Normalization

Infant Intermodal Manifold

Intermodal Pairing

Intermodal Pairs	Socio Weight
$(\mathbf{c}^1, \mathbf{c}^7)$	$\kappa(\iota(g_1))$
$(\mathbf{c}^3, \mathbf{c}^9)$	$\kappa(\iota(g_2))$
$(\mathbf{c}^5, \mathbf{c}^{11})$	$\kappa(\iota(g_3))$

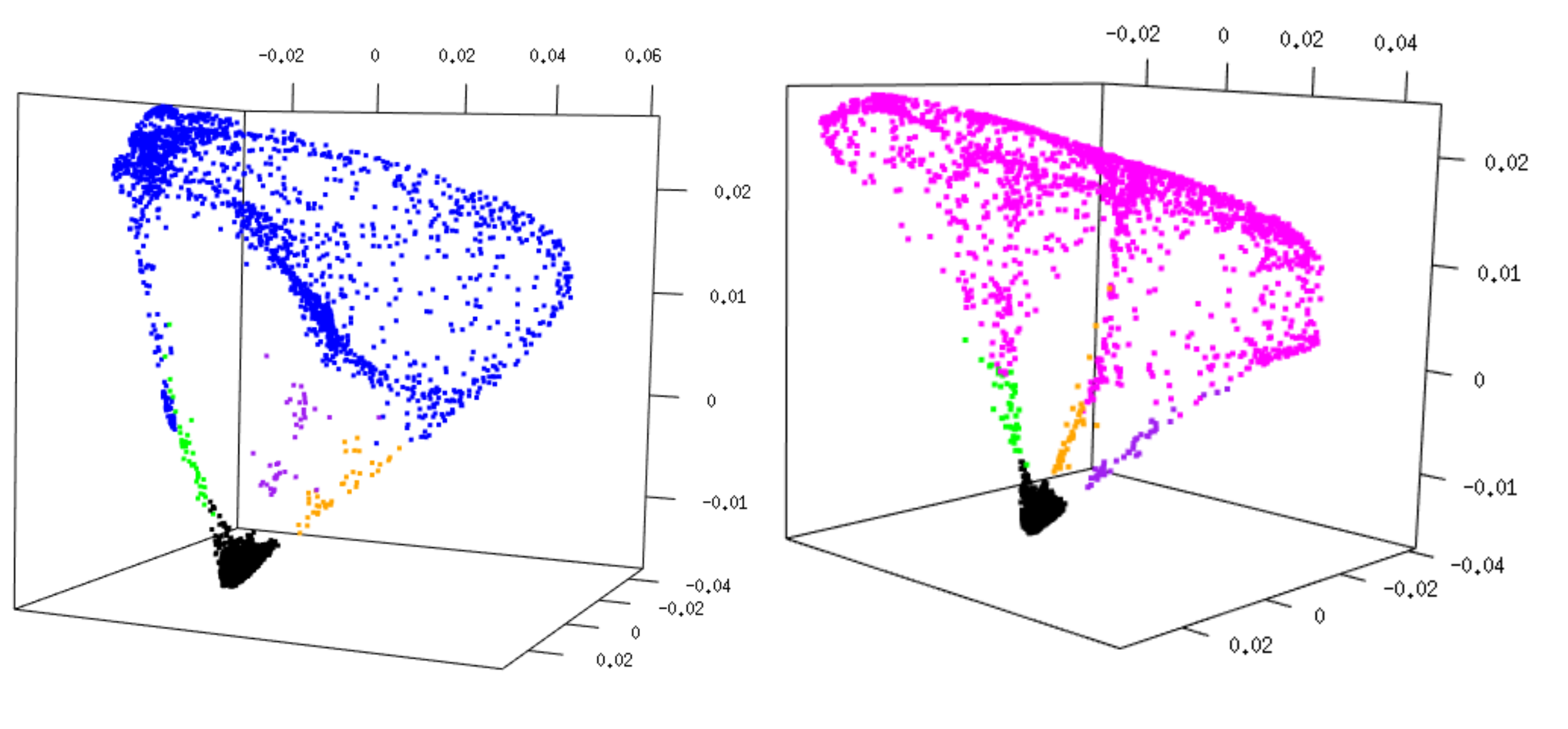
Caretaker Intermodal Manifold

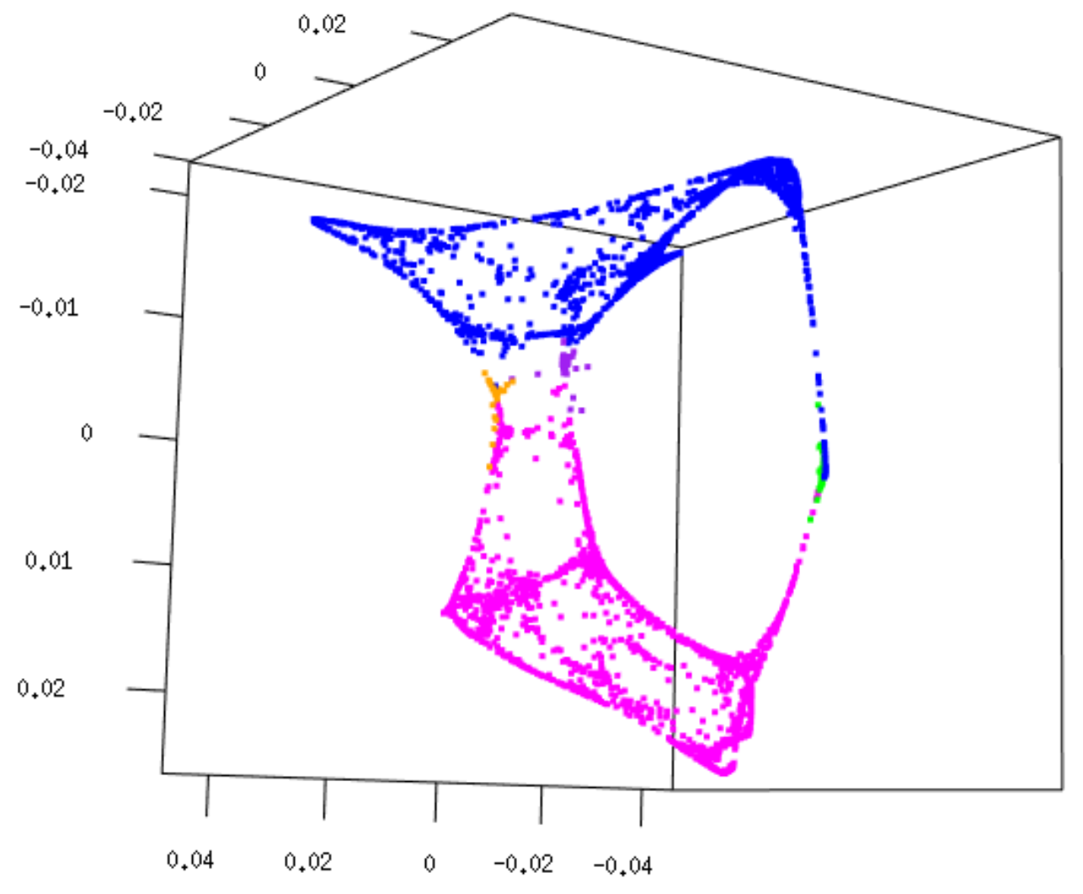
Intermodal Normalization

Commensuration Manifold

$\kappa(\iota(g_1))$ $\kappa(\iota(g_2))$ $\kappa(\iota(g_3))$

Generating Representations through Manifold Alignment

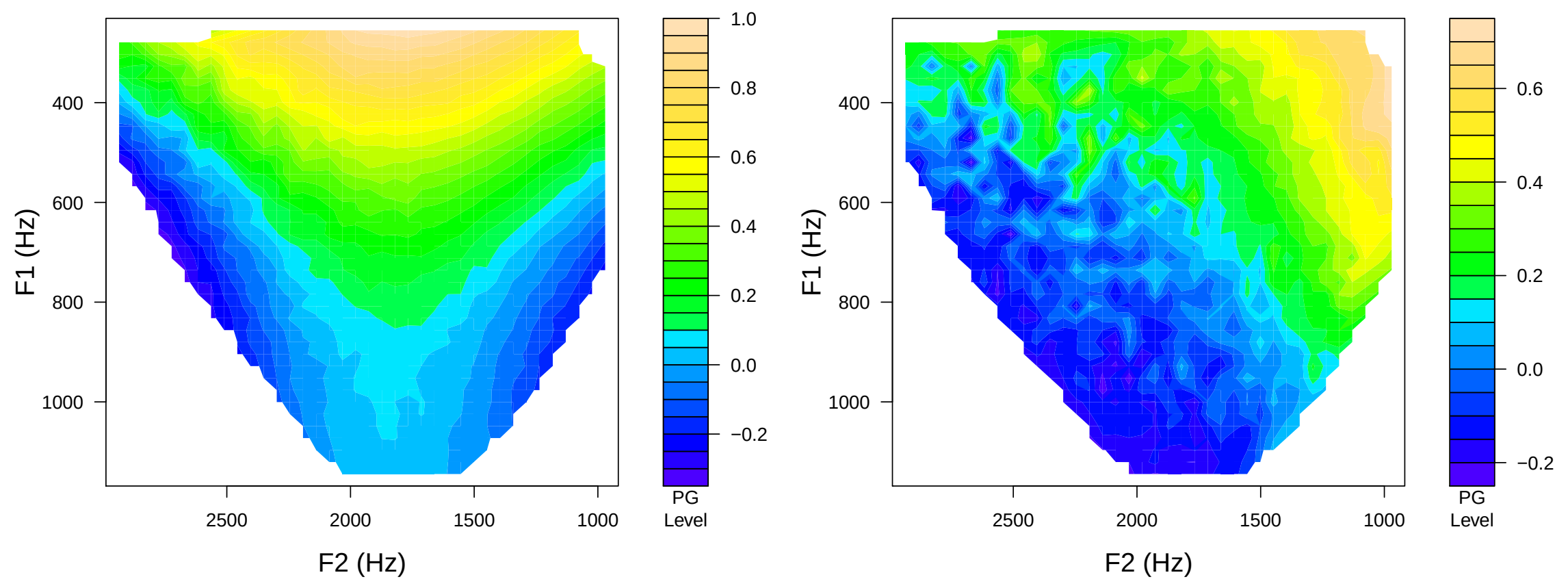




- Manifolds formed over representations within the articulatory and auditory domains are aligned using the weighted sensorimotor pairings, yielding two sets of intermodal representations (above).
- Intermodal pairs corresponding to the internalized socio-auditory pairs are each assigned a socio-intermodal weight $\kappa(\iota(g))$.
- Manifolds formed over intermodal representations are aligned using weighted intermodal pairings, yielding commensuration representations (left).

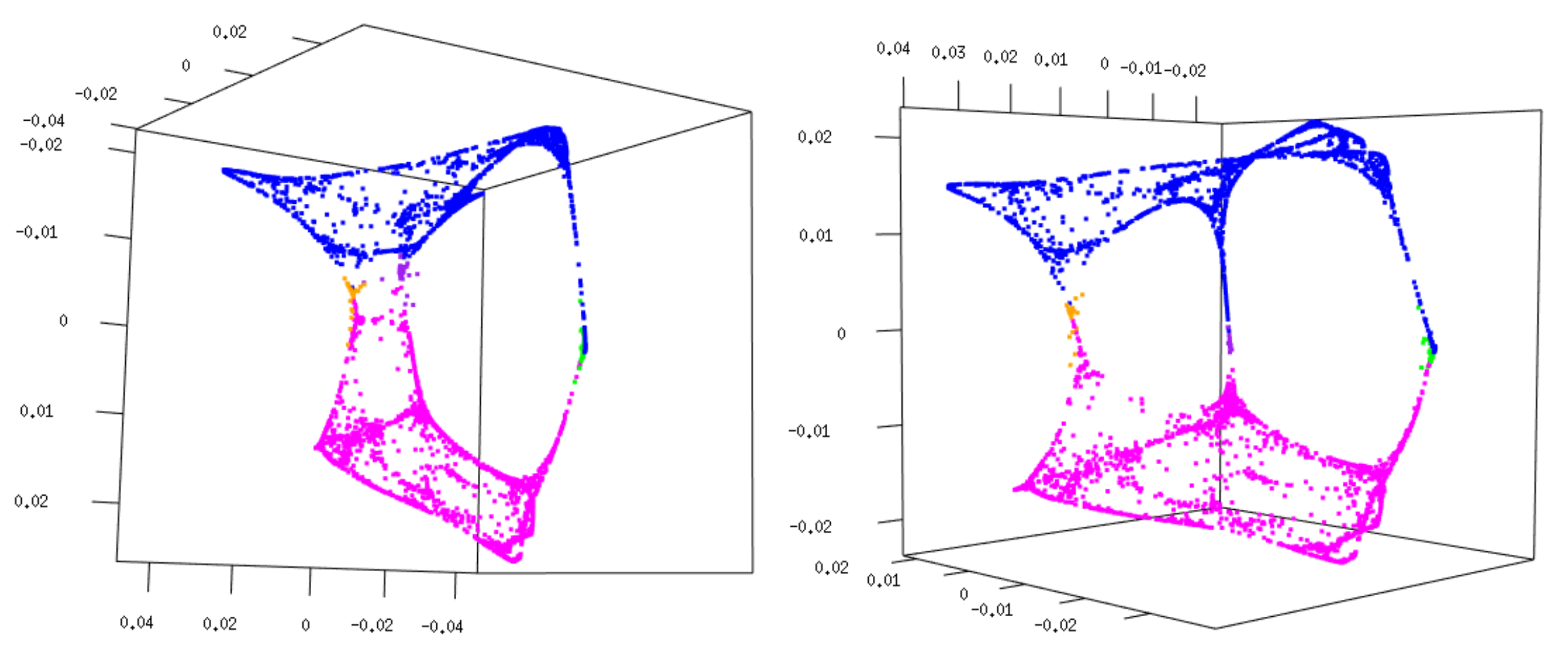
Avenues of Inquiry

Response Comparison Within and Across Languages



- VCRSs 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 VCRSs can capture cross-language differences between the five-vowel systems of Japanese (left) and Greek (right), and within-language sociolinguistic differences concerning Japanese /u/.

Structural Analysis Within and Across Languages



- The computational cognitive framework provides inroads toward quantitative analysis, e.g., parameter space testing, comparison across algorithmic variants, etc.
- For example, the representations and structures above derived from Japanese (left) and Greek (right) VCRSs reflect an infant's internalization of the language-specific, and potentially dyad-specific socio-vocal experience during early infancy.

Discussion

- We view vowel normalization as an acquired computation involving an infant's formation of manifold models of the self and others, and relations between them, based on language-specific vocal interaction with caretakers during early infancy.
- 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 renewed relevance to ideas suggested by early social psychologists concerning the phatic aspects of 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.

Acknowledgments

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, and Catherine McBride-Chang for providing the resources to recruit the Cantonese-speaking subjects. We also wish to thank Ben Munson, Lucie Ménard, and Mary Beckman for the underlying project infrastructure. Work supported by NSF grants BCS-0729306 and BCS-0729277.

Corresponding author: Andrew R. Plummer - Department of Computer Science and Engineering - The Ohio State University - Columbus, Ohio, USA

email: plummer@ling.ohio-state.edu

www: http://www.learningtotalk.org