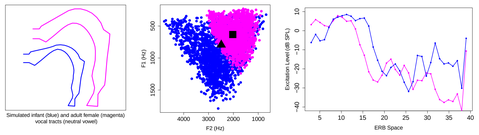
A Virtual Environment for Modeling the Acquisition of Vowel Normalization

Andrew R. Plummer
The Ohio State University

What is this project all about?

Lack of Invariance Problem for Vowel Category Acquisition

- In order to speak, we must have a goal in mind of how we would like to sound. We also have to know how to configure our vocal tracts to meet that goal.
- Prior to learning their first words, children must learn both of these types of knowledge - how they should sound, which they learn from listening to and watching their parents and caretakers, and how to configure their vocal tracts to sound that way.
- Results of decades of research on vowels support the conclusion that perception and production of language-specific vowel categories cannot be based on invariant targets that are represented directly in either the auditory domain or the articulatory (sensorimotor) domain.



- For example, an infant's and an adult female's productions of the neutral vowel [ə] differ when they are schematically represented (**a**) in the articulatory (sensorimotor) domain using VLAM simulations of vocal tract growth [left], (**f**) in the acoustic domain as points in the F1/F2 formant space [middle], or (**e**) in the auditory domain using ERB-transformed excitation patterns [right].
- This raises a number of questions about how an infant can acquire the cognitive representations relevant for learning the vowels of the ambient language.

The Acquisition of Vowel Normalization

- Vowel normalization is a computation that is meant to account for the differences in the absolute direct (physical or psychophysical) representations of qualitatively equivalent vowel productions that arise due to differences in speaker properties such as body size types, age, gender, and other socially interpreted categories that are based on natural variation in vocal tract size and shape.
- The main focus of this project is the establishment of computational methods for investigating the acquisition of vowel normalization during early infancy that complement, enhance, and ultimately drive theoretical and experimental advances.

Modeling the acquisition of vowel normalization

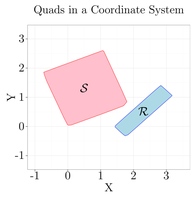
Previous Approaches

- Some models of acquisition assume a fixed auditory transform to normalize for talker vocal tract size (e. g., Callan et al., 2000), ignoring evidence that normalization must be culture-specific (e. g., Johnson, 2005).
- Others assume learning is based on statistical regularities solely within the auditory domain (e. g., Assmann and Nearey, 2008), ignoring evidence that articulatory experience also shapes vowel category learning (e.g., Kamen and Watson, 1991).
- More recent models assume that learning is based primarily on statistical regularities within the auditory domain (e.g., Ishihara et al., 2009, Ananthakrishnan and Salvi, 2011) or articulatory domain (e.g., Rasilo et al., 2013), as revealed by interaction with a caretaker, ignoring developmental complexities of internal representation of the interaction, including:
  - the gradual development of representation of the self and others (e.g., Mead, 1909, Hsu et al., 2013);
  - the creation of intermodal representations (Meltzoff and Kuhl, 1994) and multisensory perceptual narrowing (Lewkowicz and Ghazanfar, 2009).

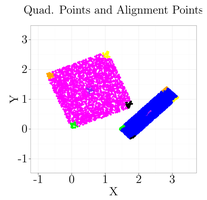
Our Virtual Environment

- We present a virtual environment for vocal learning which provides the means to model the acquisition of vowel normalization within a developmental framework encompassing a suite of vocal learning phenomena, including language-specific caretaker vocal exchanges, perceptual warping, and multisensory matching/narrowing.
- The virtual environment consists of the following components:
  - Caretaker agents from five different language communities – American English, Cantonese, Greek, Japanese, and Korean – derived from vowel category perception experiments (Munson et al., 2010, Plummer et al., 2013), which are used to model their social/vocal signaling in response to infant vowel productions.
  - Infant agents that "vocally interact" with their caretakers via turn-taking exchanges. Infants internalize the social and vocal signals accumulated during these exchanges and carry out computations over them (Plummer, 2012, 2013).

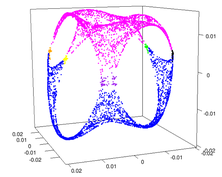
Aligning quadrilaterals: a motivating example illustrating the basic concepts



- Suppose we want to align the quadrilaterals $\mathcal{S}$ and $\mathcal{R}$ situated within a coordinate system (left).
- The computation involves constructing pairs of points that guide the alignment (orange, yellow, green, black, purple, at right).
- Pairs consist of one point from $\mathcal{S}$ and a corresponding point from $\mathcal{R}$.

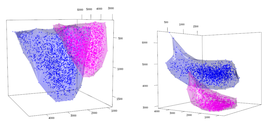


- The quadrilaterals $\mathcal{S}$ and $\mathcal{R}$ are represented using weighted graph structures called "manifolds," inspired by the topological objects.
- The manifold representations of $\mathcal{S}$ and $\mathcal{R}$ are "aligned" using the constructed pairing (left), yielding a set of aligned representations in a new reference frame (right).



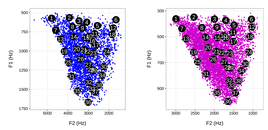
Virtual environment modeling components

Maximal Vowel Spaces and Prototypes



- 38 vowel stimuli were generated by the *Variable Linear Articulatory Model* (VLAM, Boe and Maeda, 1998), for each of seven ages, including 6 months and 10 years (left).
- Each set of stimuli is situated within a **maximal vowel space** (MVS, Boe et al., 1989, Schwartz et al., 2007) producible by the model at the corresponding age.

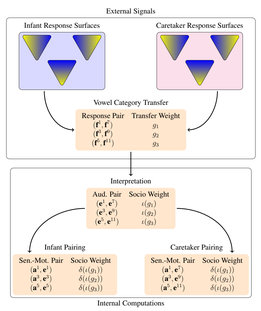
Perceptual Categorization Experiments



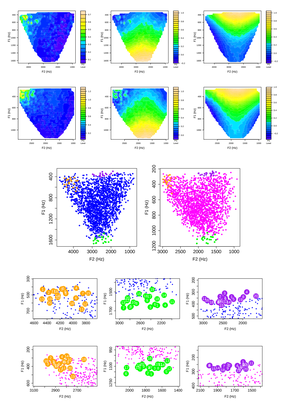
- The stimuli were categorized by members of 5 different language communities: Cantonese (n=15), English (n=21), Greek (n=21), Japanese (n=21), and Korean (n=20).
- Each listener assigned each stimulus a vowel category from the listener's native language, along with a "goodness rating" (Miller, 1994, 1997) indicating how good the listener felt that stimulus was as an example of the assigned category.

Caretaker Models: Representation Pairing and Transfer

- Formant pattern 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.

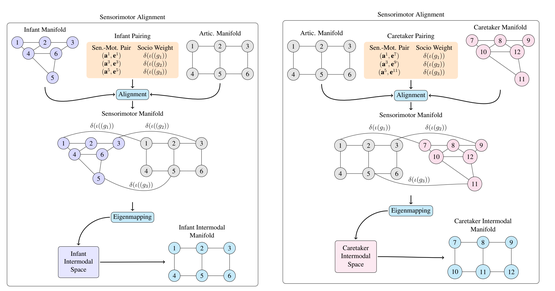


Vowel Category Response Surfaces and Response Pairings

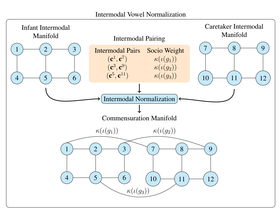


- Goodness values are modeled using a statistical methodology based on analysis of a set of cross-language vowel categorization experiments (Munson et al., 2010).
- The statistical methodology, based on a smoothing spline approach (Wahba, 1990, Gu, 2002) to additive modeling (see Hastie and Tibshirani, 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 patterns 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 goodness ratings.

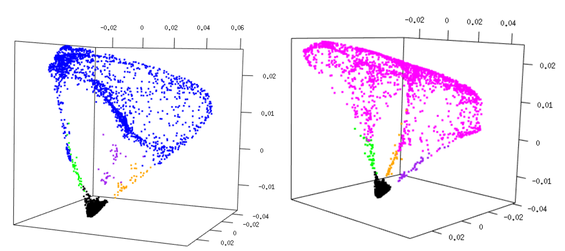
Infant Models: Sensorimotor and Intermodal Alignment

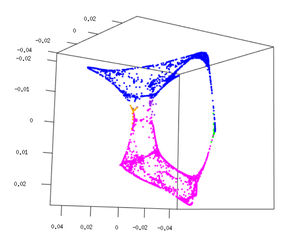


- Manifolds formed over representations within the articulatory and auditory domains are aligned using the weighted sensorimotor pairings. These **sensorimotor manifolds** yield **intermodal representations** (**c**) of the articulatory representations and excitation patterns.
- The intermodal representations 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 and Commensuration Representations

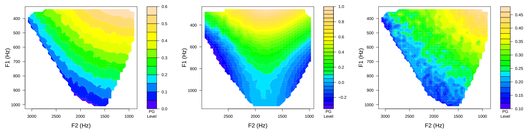




- The response pairs are internalized by the infant as socio-auditory pairs, each having a socio-auditory weight $\iota(g)$.
- The socio-auditory pairs yield two sets of sensorimotor pairs, where each pair is assigned a socio-sensorimotor weight $\delta(\iota(g))$.
- Manifolds formed over representations 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).

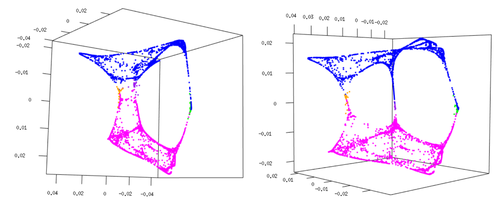
What's been done with this framework so far?

Using VCRSs for Within- and Cross-language Comparison



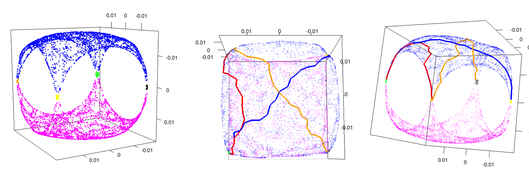
- The VCRSs above depict the /u/ responses from a Japanese listener (left), another Japanese listener (center), and a Cantonese listener (right).
- Quantitative analysis in Plummer, et al. (2013) shows that "distances" between VCRSs can capture (i) cross-language differences between the five-vowel systems of Greek and Japanese, and (ii) within-language sociolinguistic differences concerning Japanese /u/.

Using Manifolds for Within- and Cross-language Comparison



- The commensuration representations above are computed by an infant following vocal learning from a model Japanese listener (left), and a model Greek listener (right).
- Preliminary analysis in Plummer (2014) shows that the infant computations reflect the cross-language differences observed in VCRSs, capturing the language-specificity of vowel category acquisition, while suggesting that acquisition is also dyad-specific.

Where are we going from here?



Vowel Dynamics: Manifolds over representations can be used to model paths through reference frames. We are currently investigating the language-specific nature of path formation as a result of language-specific vowel normalization.

Social Categories: In addition to vowel category responses, listeners also provided age and gender ratings for a subset of the vowel prototypes. We are currently investigating relationships between these ratings and the vowel category responses, and their influence on the acquisition of normalization.

Acknowledgements

The perceptual categorization data are from a study by Benjamin Munson, using stimuli provided by Lucie Ménard and subjects recruited by Catherine McBride-Chang, Chanelle Mays, Asimina Syrika, Kiyoko Yoneyama, and Hyunju Chung. Work supported by NSF grants BCS 0729277 (to Benjamin Munson) and BCS 0729306 (to Mary Beckman). We thank Pat Reidy for the auditory modeling code.

Contact information

Email: plummer@ling.ohio-state.edu
Web page: http://www.ling.ohio-state.edu/~plummer/
Project Page: http://learningtotalk.org