Graph alignment and cross-modal learning during early infancy

Andrew R. Plummer

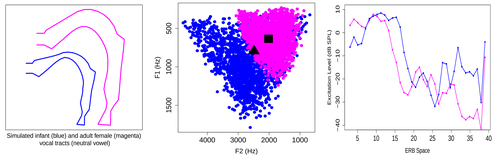
The Ohio State University, Columbus, OH, USA

www: http://www.learningtotalk.org

email: plummer@ling.ohio-state.edu

Introduction

- 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.

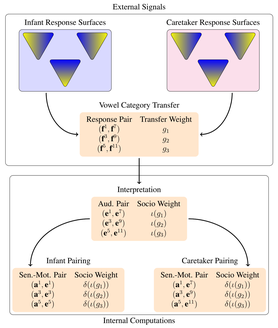
Previous Modeling

- 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 that learning can be 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).

Cross-modal Learning as Graph Alignment

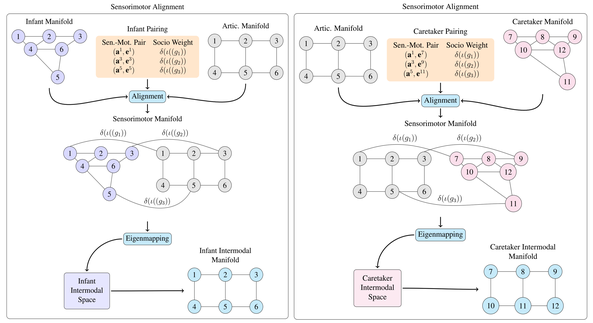
- We outline an alternative approach that models cross-modal learning using graph structures, called "manifolds," which organize sensory information in the auditory and in the articulatory domain, so that information can be linked across the two domains via graph alignment.
- Graph alignment is guided by perceptual targets that are internalized in early infancy through social/vocal interaction with caretakers, so that vowel categories (**c**) can be identified with the abstractions that mediate between the two domains during alignment, rather than with domain-internal representations (**a**), (**f**), or (**e**).

Internalization and Pairing Computations

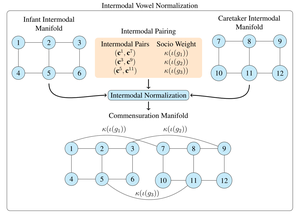


- Formant pattern representations of each vowel signal (**f**) are assigned goodness values reflecting a caretaker's intuitions about the categorical status of the signal within the caretaker's vowel system.
- Representations of infant vowels with high goodness values are paired with caretaker vowels 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 exictation 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.

Sensorimotor Alignment Computations

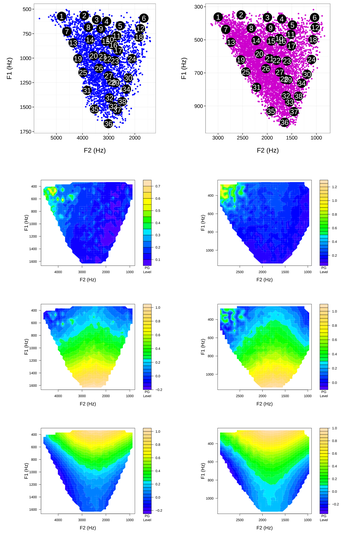


Intermodal Vowel Normalization



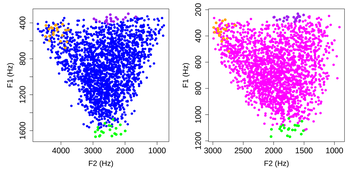
- Manifolds formed over representations within the articulatory and auditory domains are aligned using the weighted sensorimotor pairings. These sensorimotor manifolds yield intermodal representations (**c**) of the articulatory representations and excitation patterns.
- The intermodal representations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

Acoustic and Social Signals

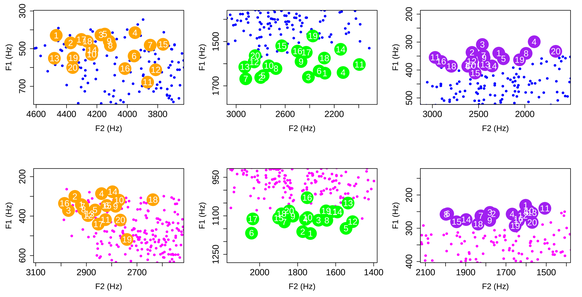


- Goodness values are modeled using a statistical methodology based on analysis of a set of cross-language vowel categorization experiments (Munson et al., 2010).
- 38 vowel stimuli were generated by the *Variable Linear Articulatory Model* (VLAM, Boe and Maeda, 1998), for each of seven ages, including 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.
- 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.
- 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 (left) and 10 years (right) are derived from goodness ratings provided by a Japanese subject.

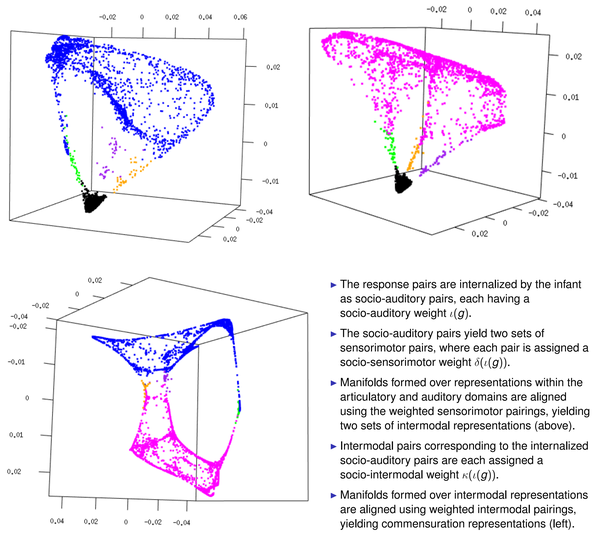
Response Pairings



- 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.



Multisensory Representational Output



- 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).

Summary

- On our approach vowel normalization is a generative procedure, rather than a reductive invariance computation or statistical summary.
- Acoustic and social signals derived from interaction with a caretaker provide the raw material for the normalization computation.
- Auditory representations are computed over the acoustic and social signals, providing the targets for sensorimotor alignment and the computation of rich representations of the self and the caretaker.
- Higher-order intermodal representations of the self and caretaker are computed from the sensorimotor alignments, which reflect multisensory perceptual narrowing.
- The intermodal representations are then aligned to yield a commensuration structure that provides the basis for vowel categorization, and other cognitive computations.

Acknowledgments

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).

Corresponding author: Andrew R. Plummer - Department of Linguistics - The Ohio State University - Columbus, Ohio, USA

email: plummer@ling.ohio-state.edu

www: http://www.learningtotalk.org