Aligning manifolds to model the earliest phonological abstraction in infant-caretaker vocal imitation

Andrew R. Plummer

Department of Linguistics
The Ohio State University

September 13, 2012

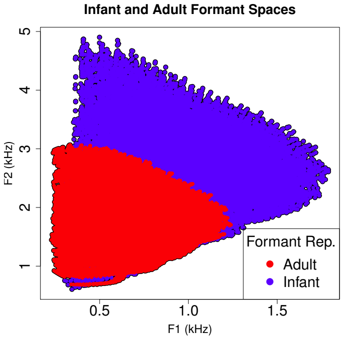
Approaches to Vowel Normalization

Recent Work

Recent cognitive models of vowel normalization (Ishihara et al., 2009; Plummer et al., 2010; Ananthakrishnan & Salvi, 2011) take the mapping between auditory representations of the infant's own vocalizations and those of a caretaker to be a transformation that the infant builds during vocal imitative interactions with the caretaker.

I) That is, the "direct transformation" approach (Ishihara et al., 2009; Ananthakrishnan & Salvi, 2011) assumes that the infant learns a pre-specified transformation.

II) In contrast, the "alignment" approach (Plummer et al., 2010) uses a set of infant-caretaker auditory representation pairs from which a full transformation is inferred.



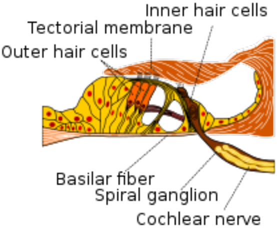
The Point of this Presentation

Contributions

We extend a particular model (Plummer et al., 2010) within the alignment approach in two key ways:

(i) by assigning an interpretation to the given set of infant-caretaker pairs that casts the need to abstract as an asset, rather than a liability, and

(ii) by using a more suitable model for the infant's representations of infant and caretaker vowels.



Formulating Vowel Normalization

Object of Study

We take vowel normalization to be a cognitive process “in which interspeaker vowel variability is reduced in order that perceptual vowel identification may then be performed by reference to relative vowel quality rather than absolute [psychophysical] parameters of vowels” (Johnson, 1990, p. 230).

Phonological Abstraction

Vowel normalization in this sense can be viewed as a particular instance of the more general notion of phonological abstraction with respect to vowels (hereafter, simply phonological abstraction), defined as the computation of an abstract representation of a vowel, from one (or more) of its perceptual representations, to facilitate some further computation.

Phonological Abstraction and Acquisition

Phonological Abstraction facilitates Lexical Processing

Recent work (Cutler et al., 2010) suggests that phonological abstraction facilitates lexical processing. Specifically, “prelexical phonemic categories are an essential part of word recognition” since they “allow the listener to map distinct acoustic events onto the same underlying lexical representations” (pp. 93-4).

Phonological Abstraction in Acquisition

i) This in turn suggests that phonological abstraction is an integral component of the spoken language acquisition process.

ii) Indeed, infants appear to be reconciling the absolute differences between their perceptual representations of adult vowels and their own by six months of age (Kuhl, 1979, 1983; Kuhl & Meltzoff, 1996; Ménard et al., 2002).

Acquisition of Vowel Normalization

Key Assumptions

In contrast to previous models (Sussman, 1986; Smith et al., 2005; Ames & Grossberg, 2008; Heintz et al., 2009), our model is based on the following two assumptions.

i) Vowel normalization is malleable with respect to short-term contextual information (Johnson, 1990; Ladefoged & Broadbent, 1957), as well as long-term distributional and ontogenetic information (Kohn & Farrington, 2012).

ii) The representational structures or transforms that underpin normalization are not pre-specified but are themselves learned, as part of acquiring the phonology of a spoken language.

Learning by Vocal Imitative Interaction

The assumption that normalization is learned entails that there is a process by which it is learned. We assume that this process is social interaction between the infant and a caretaker characterized by specific types of vocal imitation (Howard & Messum, 2011; Masataka, 2003; Fitch, 2010; Gros-Louis et al., 2006; Goldstein & Schwade, 2008).

Aspects of Imitative Interaction

Aspects of Vocal Imitation

Experimental results suggest that these vocal imitation interactions involve:

I) structured turn-taking between the infant and caretaker (Masataka, 2003), and

II) caretaker responses differentiated according to the nature of infant vocalizations (Gros-Louis et al., 2006; Goldstein & Schwade, 2008).

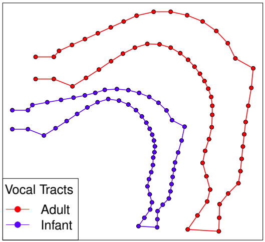
These structured individuated instances of vocal imitation are “evidence for [the child] to deduce a correspondence between his output and the speech sound equivalent within [the mother’s] L1 that she produces” (Howard & Messum, 2011, p. 87).

Vowel Modeling: Formant Representations

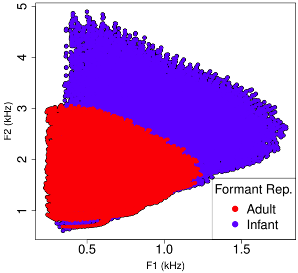
VLAM Formant Representations

I) The Variable Linear Articulatory Model (VLAM, Boë & Maeda, 1997) generates articulatory configurations and their corresponding speech signals, simulating speech productions of humans ranging in age from early infancy to adulthood.

II) Each vowel signal output by the VLAM is synthesized using the first four formant frequencies determined by the signal's articulatory configuration.



Infant and Adult Midsagittal Vocal Tracts (Neutral)



Infant and Adult Formant Spaces

Vowel Modeling: Infant and Caretaker Data

Approximating Infant and Caretaker Productions

I) We make the simplifying assumption that the vowels experienced by the infant are restricted to those produced by the infant and one caretaker.

II) We pseudorandomly generated 2,000 vowel signals using the VLAM set at 6 months of age (10 years of age, respectively) to represent the infant's vowels (caretaker's vowels, respectively).

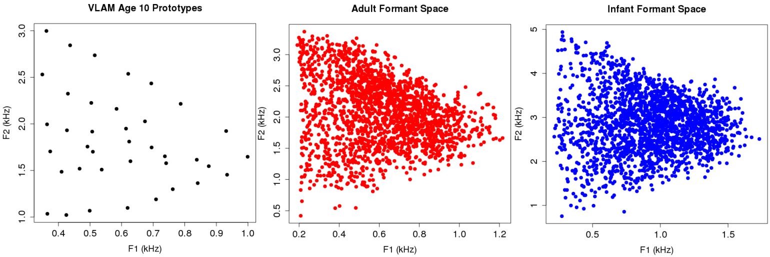


Figure: We used the 10 year-old setting to model the caretaker as it was perceived to be most similar to a young female adult in a cross-language perception study (Munson et al., 2010).

Vowel Modeling: Caretaker Vowel System

Approximating Caretaker Vowel System: Greek

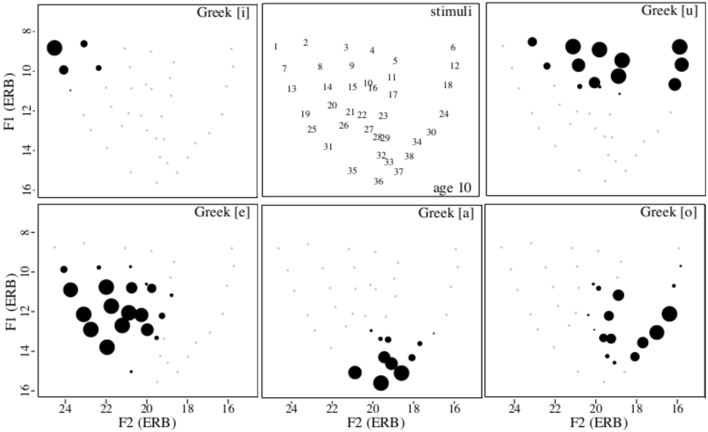
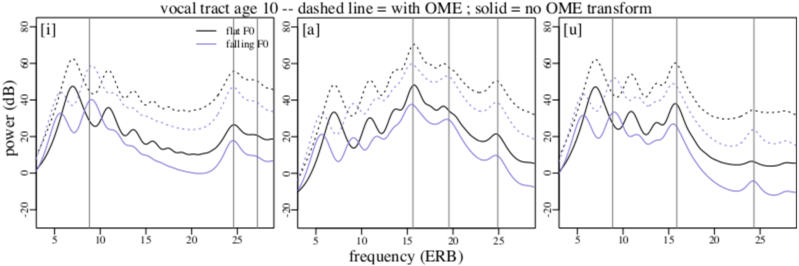


FIGURE: We approximate the vowel system of the caretaker using the categorizations of the VLAM age 10 stimuli by native speakers of Greek.

AUDITORY REPRESENTATIONS

The *auditory representations* we use are "excitation patterns" derived using the transformations described in Moore et al. (1997) applied to the vowel signals generated by the VLAM.



We will primarily use auditory representations for the remainder of this presentation, though we use the neutral notation V to denote collections of representations.

Cognitive Modeling: Perceptual Manifolds

Let V denote the set of generated representations, and let V_C and V_I denote the partition cells of V consisting of the caretaker and infant vowel representations, respectively.

Perceptual Manifolds

I) Perceptual manifolds are modeled as weighted graphs $G = (N, E, W)$, and their geometric properties are represented by the weights W on the edges E connecting the nodes in N.

II) Exposed to the vowels in V, the infant creates perceptual manifolds M_C and M_I which are complete graphs whose nodes are (in one-to-one correspondence with) the representations in V_C and V_I, respectively.

III) That is, $M_C = \langle V_C, E_C, W_C \rangle$ and $M_I = \langle V_I, E_I, W_I \rangle$. The geometric structure of the perceptual manifolds is determined by their weight functions

$$W_C : E_C \to \mathbb{R}_+ \qquad W_I : E_I \to \mathbb{R}_+.$$

both taken to be nearest-neighbor functions (Belkin & Niyogi, 2003) based on Euclidean distance, greatly simplifying M_C and M_I.

Cognitive Modeling: Manifold Alignment

Vowel Normalization as Manifold Alignment

The model of vowel normalization is a manifold alignment computation (Wang, 2010), implemented as a correspondence-based algorithm that maps points on two (or more) manifolds to a common "latent space" (Ham et al., 2005). The alignment requires methods for

i) combining the geometric information of the manifolds to facilitate their alignment,

ii) populating equal-length arrays of points from each manifold, such that, given an index, the points in each array at that index correspond to each other, to further facilitate alignment.

Graph Laplacian

i) The *graph Laplacian* of G is the matrix $L = D - W$ where D is a diagonal matrix such that $D_{ii} = \sum_j W_{ij}$ (Chung, 1997).

ii) The graph Laplacian L of a graph G is a principled choice for approximating geometry-preserving functions on G in terms of the eigenvectors of L (Belkin & Niyogi, 2003).

Cognitive Modeling: Manifold Alignment

Combining Geometric Information

Let L_C and L_I be the graph Laplacians for M_C and M_I, respectively. The alignment algorithm (Ham et al., 2005) combines L_C and L_I to facilitate the alignment of M_C and M_I with respect to a set of corresponding points drawn from V_C and V_I.

Populating arrays of corresponding points

i) Population of the arrays of corresponding points amounts to the specification of a characteristic function over $V_C \times V_I$, denoted $\chi_{voc} : V_C \times V_I \rightarrow \{0, 1\}$.

ii) Essentially, χ_{voc} models the identification of vocal imitative interactive experience that affects normalization, which may be derived in a number of different ways.

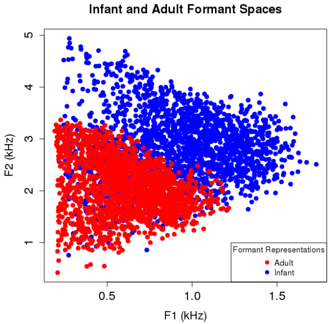
iii) To exposit the methodology, we use "good" productions of i) a caretaker with full command of the Greek 5-vowel system, and ii) an infant's vocalizations that receive contingent response.

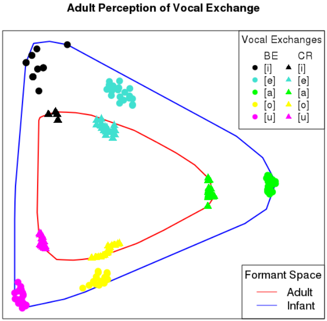
Cognitive Modeling: Caretaker-centric Mapping

Response Surface Projection

I) We use a simple response surface projection method over the Greek perceptual categorization responses to approximate the caretaker's contingent responses to infant vocalizations.

II) The infant's vocalizations that are perceived as "good" examples of a vowel category receive "good" caretaker responses from that category. Each good infant vocalization and its caretaker response constitute members of χ_{voc}.



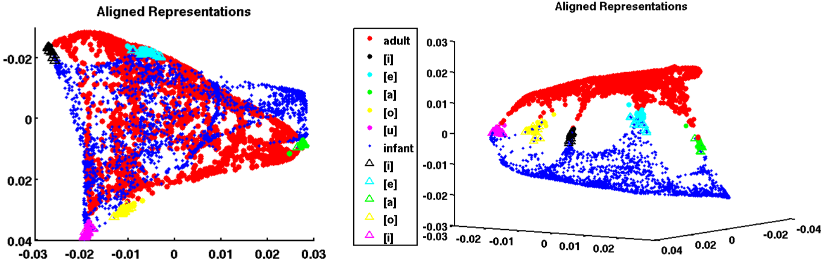


Cognitive Modeling: Alignment Algorithm

The infant computes the alignment of M_C and M_I by constructing a “combined Laplacian” from L_C and L_I, using χ_{voc}, and infers a *normalization transformation*

$$N(L_C, L_I, \chi_{voc}) : V \to V_Z.$$

from the vowel representations in V to a latent space V_Z whose points are “abstract representations” of the representations in V.



NARROWING THE INVESTIGATION

The framework we proffered thus allows for investigation of the effects of different vowel representations on normalization. More generally, it allows for straightforward investigation of the effects of the following:

I) different weight functions, and thus different geometrical structures over the vowel representations,

II) different methods of combining the graph Laplacians, and of combining them with infant-caretaker pairs, and

III) different arrays of infant-caretaker pairs, which vary in number of pairs, and in the characteristics of the pairs included.

The framework allows for all of these components to be modified in accordance with short-term contextual information, long-term distributional and ontogenetic information, and social development.

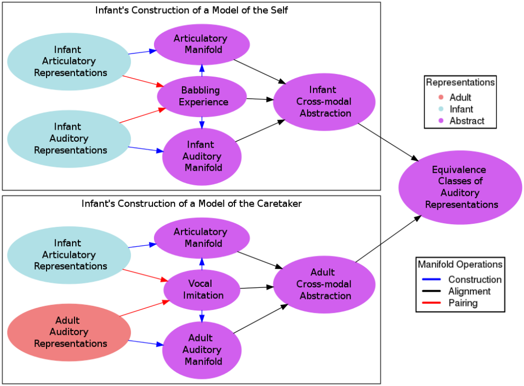
Broadening the Investigation

The alignment method may be viewed as an "intramodal" mapping over a single reference frame of representations, though it can easily be extended to handle "cross-modal" mappings between reference frames over differing representations.

i) In addition to M_C and M_I, the infant creates a manifold M_A over infant articulatory representations.

ii) The infant aligns M_I and M_A to create representations of self, while aligning M_C and M_A to create representations of the caretaker.

iii) Manifolds are then created over these representations, respectively, and then aligned, to yield equivalence classes of vowel representations.



Infant's Construction of a Model of the Self

Infant Articulatory Representations

Articulatory Manifold

Babbling Experience

Infant Auditory Representations

Infant Auditory Manifold

Infant Cross-modal Abstraction

Infant's Construction of a Model of the Caretaker

Infant Articulatory Representations

Articulatory Manifold

Vocal Imitation

Adult Auditory Representations

Adult Auditory Manifold

Adult Cross-modal Abstraction

Equivalence Classes of Auditory Representations

Representations

- ● Adult
- ● Infant
- ● Abstract

Manifold Operations

- — Construction
- — Alignment
- — Pairing

Ames, H., & Grossberg, S. (2008). Speaker normalization using cortical strip maps: A neural model for steady-state vowel categorization. *Journal of the Acoustical Society of America*, *124*(6), 3918–3936.

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

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

Boë, L.-J., & Maeda, S. (1997). 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.

Chung, F. R. K. (1997). *Spectral Graph Theory*. Regional Conference Series in Mathematics. American Mathematical Society. Number 92.

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.) *Laboratory Phonology 10*, (pp. 91–111). Berlin: de Gruyter.

Fitch, W. T. (2010). *The Evolution of Language*. Cambridge University Press.

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

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.

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).

Heintz, I., Beckman, M., Fosler-Lussier, E., & Ménard, L. (2009). Evaluating parameters for mapping adult vowels to imitative babbling. In *INTERSPEECH 09*, (pp. 688–691). Brighton, UK.

Howard, I. S., & Messum, P. (2011). Modeling the development of pronunciation in infant speech acquisition. *Motor Control*, *15*, 85–117.

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.

Johnson, K. (1990). Contrast and normalization in vowel perception. *Journal of Phonetics*, *18*, 229–254.

Kohn, M. E., & Farrington, C. (2012). Evaluating acoustic speaker normalization algorithms: Evidence from longitudinal child data. *Journal of the Acoustical Society of America*, *131*, 2237–2248.

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America*, *66*(6), 1668–1679.

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development*, *6*, 263–285.

Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *Journal of the Acoustical Society of America*, *100*(4), 2425–2438.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *JASA*, *29*(1), 98–104.

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

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.

Moore, B. C. J., Glasberg, B. G., & Baer, T. (1997). A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society*, *45*(4), 224–240.

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.

Plummer, A. R., Beckman, M. E., Belkin, M., Fosler-Lussier, E., & Munson, B. (2010). Learning speaker normalization using semisupervised manifold alignment. In *Proceedings of INTERSPEECH 2010*. Tokyo.

Smith, D., Patterson, R., Turner, R., Kawahara, H., & Irino, T. (2005). The processing and perception of size information in speech sounds. *Journal of the Acoustical Society of America*, *117*, 305–318.

Sussman, H. M. (1986). A neuronal model of vowel normalization and representation. *Brain and Language*, *28*, 12–23.

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