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

Andrew R. Plummer

Department of Linguistics, The Ohio State University, Columbus, Ohio, USA

plummer@ling.osu.edu

Abstract

We argue that infants perform an abstraction over their auditory representations of the vowels of individual speakers, by mapping them to a mediating space of speaker-independent representations, guided by vocal imitative interaction with their caretakers, as a first step in the phonological acquisition process. Furthermore, we proffer a methodology for modeling this abstraction which involves the alignment of the cognitive structures, or manifolds, that the infant builds from the auditory representations of the vowels of individual speakers. As a demonstration of the methodology, we show that higher-dimensional "excitation pattern" representations facilitate modeling of the influence of the imitative process on perception and abstraction more so than lower-dimensional formant representations.

Index Terms: vowel normalization, manifold alignment, vocal imitation, phonological abstraction.

1. Introduction

Recent cognitive models of vowel normalization [1, 2, 3] 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. The models make ontological commitments to an external or universalist, direct transformation interpretation of the mapping, or that of an internal, potentially idiosyncratic alignment-based interpretation. That is, the "direct transformation" approach [1, 3] assumes that the infant learns a pre-specified transformation, whereas the "alignment" approach [2] uses a set of infant-caretaker auditory representation pairs from which a full transformation is inferred.

In this paper, we extend a particular model [2] 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. More generally, we proffer a methodology for investigating normalization that provides for the incorporation of relevant biological and social phenomena.

2. Conceptual basis

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" [4, p. 230]. 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.

Recent work [5] 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). This in turn suggests that phonological abstraction is an integral component of the spoken language acquisition process. Indeed, infants appear to be reconciling the absolute differences between their perceptual representations of adult vowels and their own by six months of age [6, 7, 8, 9]. Vowel normalization is often taken to be *the* phonological abstraction by which this is achieved [2, 10, 11, 12, 13, 14]. We likewise assume that vowel normalization plays a significant role in the abstraction process, though we take it to be more complex than is typically assumed.

More specifically, in contrast to previous models [11, 12, 13, 14], our model is based on the following two assumptions. First, vowel normalization is malleable with respect to short-term contextual information [4, 15], as well as long-term distributional and ontogenetic information [16]. Second, we assume that 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. The second assumption is based on the following reasoning from evidence about more general aspects of learning to perceive others in relation to the self.

Experimental results in speech perception [17] and pathology [18] suggest the importance of individual

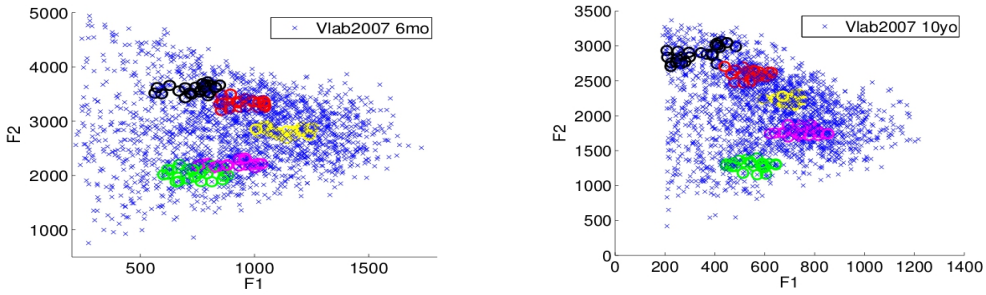

Figure 1: *Blue x's represent the infant (left) and caretaker (right) vowel tokens generated by the VLAM (Vlab2007 version) in F1-F2 space. The infant and caretaker o's of the same color, corresponding to one another via an imitation process, are used to guide the manifold alignment algorithm.*

speaker identification, hence the representation of "other" conspecifics, to the formation of more abstract phonetic representations. More general modeling of the cognitive development of intentional agents (e.g., the 'like me' framework [19]) adduces the need for a representation of "self," as distinct from others. The essential distinction between self and others entails carrying out an abstraction process that "allows the infant to see the behavior of others as commensurate with their own" [19, p. 26]. Applying the general model to the domain of spoken language acquisition [20] suggests the importance of the distinction between self and others in vowel normalization. This social component, along with vowel data demonstrating the cross-language differences in the abstraction process [21], strongly suggests that vowel normalization is indeed learned.

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 [20, 22, 23, 24, 25]. Experimental results suggest that these vocal imitation interactions involve: (i) structured turn-taking between the infant and caretaker [22], and (ii) caretaker responses differentiated according to the nature of infant vocalizations [24, 25]. These richly-structured individuated instances of vocal imitation provide "evidence for [the child] to deduce a correspondence between his output and the speech sound equivalent within [the mother's] L1 that she produces" [20, (p. 87)].

3. Modeling framework

We make the simplifying assumption that the vowels experienced by the infant are restricted to those produced by the infant and one caretaker. These vowels are modeled as follows. The Variable Linear Articulatory Model (VLAM, [26]) generates articulatory configurations and their corresponding speech signals, simulating speech productions of humans ranging in age from early infancy to adulthood. 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). 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 [27].

In demonstrating our method, we use two kinds of representations for the infant and caretaker vowels: a standard formant frequency representation, and an auditory representation. Each vowel signal output by the VLAM is synthesized using the first four formant frequencies determined by the signal's articulatory configuration. These *formant representations* for the infant and caretaker vowels are depicted as blue x's in Figure 1. The *auditory representations* we use are "excitation patterns" derived using the transformations described in [28] applied to the vowel signals generated by the VLAM. We use the term *vowel representation* when distinguishing between representation types is not needed.

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

Vowel signals from the caretaker and infant vocalizations are taken to lie on "vowel manifolds" [29, 30]. These geometric structures are hypothesized to motivate the infant's formation of cognitive structures, called "perceptual manifolds" [31], used for representation and normalization. These perceptual manifolds are modeled as weighted graphs, and their geometric properties are represented by the weights. A *weighted graph* is a triple $G = (N, E, W)$ where N is a set of *nodes*, E is a set of *edges* connecting nodes, and W is a *weight function* which assigns a nonnegative value to each edge in E.

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. That is,

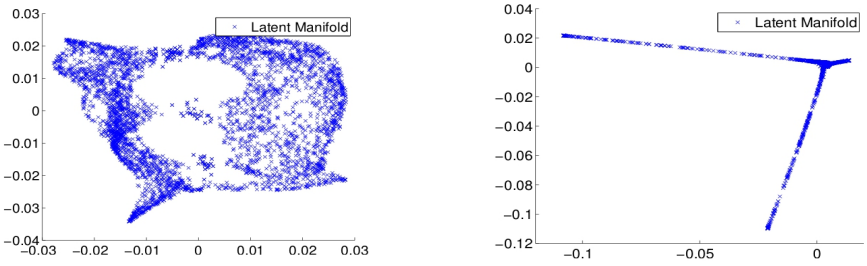

Figure 2: *The speaker-independent latent manifold yielded by the alignment algorithm carried out over formant (left) and excitation pattern (right) representations of tokens.*

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

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

We take both to be nearest-neighbor functions [32] based on Euclidean distance, greatly simplifying M_C and M_I.

The model of vowel normalization is a manifold alignment computation [33], implemented as a correspondence-based algorithm that maps points on two (or more) manifolds to a common "latent space" [34]. The alignment requires methods for (i) combining the geometric information of the manifolds to facilitate their alignment, and (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 and guide the alignment.

Toward (i), let G be a weighted graph with weight matrix W. The *graph Laplacian* of G is the matrix $L = D - W$ where D is a diagonal matrix such that $D_{ii} = \sum_j W_{ij}$. The graph Laplacian [35] is a discrete approximation of the Laplace-Beltrami operator on a Riemannian manifold [36]. 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 [32]. Let L_C and L_I be the graph Laplacians for M_C and M_I, respectively. The algorithm [34] 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.

The infant needs to populate this collection of infant-caretaker representation pairs. To exposit the methodology, we manually selected prototypes (the open circles in Figure 1) representing productions of (i) an agent with full command of a canonical 5-vowel system (right), and (ii) the infant's vocal imitations that receive contingent response (left). Thus the infant tokens of a given color are assumed to correspond to the caretaker tokens of the same color via vocal imitative interaction between the agents as described in Section 2. Although we manually selected the imitation data, we can incorporate this selection process within the cognitive model as the algorithmic determination of a characteristic function over $V_C \times V_I$, denoted

$$\chi_{voc} : V_C \times V_I \to \{0, 1\}.$$

Essentially, χ_{voc} models the identification of vocal imitative interactive experience that affects normalization.

Finally, the infant computes the alignment of M_C and M_I. The alignment algorithm [34] constructs 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.

4. Discussion

Using this modeling framework, we carried out two simulations, one using formant representations, and the other using auditory representations. In both simulations, the infant-caretaker pairs guide the learning of a mapping from the infant and caretaker manifolds to a "latent," speaker-independent space (see Figure 2). When higher-dimensional auditory representations are used, the latent space reflects the ">" shape of the given five-vowel system (Figure 2, right), capturing the influence of the imitative process on perception and abstraction. No such shape is observed in the latent space derived from the lower-dimensional formant representations (Figure 2, left).

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 informa-

tion, long-term distributional and ontogenetic information, and social development.

5. Acknowledgements

We thank Mary Beckman, Eric Fosler-Lussier, Misha Belkin, and Patrick Reidy for their contributions to this paper. Funding was provided by NSF grant BCS-0739206 and the Cognitive Science Center at The Ohio State University.

6. References

[1] H. Ishihara, Y. Yoshikawa, K. Miura, and M. Asada, "How caregiver's anticipation shapes infant's vowel through mutual imitation," *Autonomous Mental Development, IEEE Transactions on*, vol. 1, no. 4, pp. 217 –225, Dec. 2009.

[2] A. R. Plummer, M. E. Beckman, M. Belkin, E. Fosler-Lussier, and B. Munson, "Learning speaker normalization using semisupervised manifold alignment," in *Proceedings of INTERSPEECH 2010*, Tokyo, September 2010.

[3] G. Ananthakrishnan and G. Salvi, "Using imitation to learn infant-adult acoustic mappings," in *Proceedings of INTERSPEECH 2011*, 2011, pp. 765–768.

[4] K. Johnson, "Contrast and normalization in vowel perception," *Journal of Phonetics*, vol. 18, pp. 229–254, 1990.

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

[6] P. K. Kuhl, "Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories," *Journal of the Acoustical Society of America*, vol. 66, no. 6, pp. 1668–1679, 1979.

[7] ——, "Perception of auditory equivalence classes for speech in early infancy," *Infant Behavior and Development*, vol. 6, pp. 263–285, 1983.

[8] P. K. Kuhl and A. N. Meltzoff, "Infant vocalizations in response to speech: Vocal imitation and developmental change," *Journal of the Acoustical Society of America*, vol. 100, no. 4, pp. 2425–2438, 1996.

[9] L. Ménard, J.-L. Schwartz, and L.-J. Boë, "Auditory normalization of French vowels synthesized by an articulatory model simulating growth from birth to adulthood," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1892–1905, 2002.

[10] D. Hindle, "Approaches to vowel normalization in the study of natural speech," in *Linguistic Variation: Models and Methods*, D. Sankoff, Ed. New York: Academic, 1978, pp. 161–171.

[11] H. M. Sussman, "A neuronal model of vowel normalization and representation," *Brain and Language*, vol. 28, pp. 12–23, 1986.

[12] D. Smith, R. Patterson, R. Turner, H. Kawahara, and T. Irino, "The processing and perception of size information in speech sounds," *Journal of the Acoustical Society of America*, vol. 117, pp. 305–318, 2005.

[13] H. Ames and S. Grossberg, "Speaker normalization using cortical strip maps: A neural model for steady-state vowel categorization," *Journal of the Acoustical Society of America*, vol. 124, no. 6, pp. 3918–3936, 2008.

[14] I. Heintz, M. Beckman, E. Fosler-Lussier, and L. Ménard, "Evaluating parameters for mapping adult vowels to imitative babbling," in *INTERSPEECH 09*, Brighton, UK, 2009, pp. 688–691.

[15] P. Ladefoged and D. E. Broadbent, "Information conveyed by vowels," *JASA*, vol. 29, no. 1, pp. 98–104, 1957.

[16] M. E. Kohn and C. Farrington, "Evaluating acoustic speaker normalization algorithms: Evidence from longitudinal child data," *Journal of the Acoustical Society of America*, vol. 131, pp. 2237–2248, 2012.

[17] S. J. Winters, S. V. Levi, and D. B. Pisoni, "Identification and discrimination of bilingual talkers across languages," *Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4524–4538, 2008.

[18] T. K. Perrachione, S. N. D. Tufo, and J. D. E. Gabrieli, "Human voice recognition depends on language ability," *Science*, vol. 333, p. 595, 2011.

[19] A. Meltzoff, "The 'like me' framework for recognizing and becoming an intentional agent," *Acta Psychologica*, vol. 124, pp. 26–43, 2007.

[20] I. S. Howard and P. Messum, "Modeling the development of pronunciation in infant speech acquisition," *Motor Control*, vol. 15, pp. 85–117, 2011.

[21] K. Johnson, "Speaker normalization in speech perception," in *The Handbook of Speech Perception*, R. E. Remez and D. B. Pisoni, Eds. Blackwell, 2005, pp. 363–389.

[22] N. Masataka, *The Onset of Language*. Cambridge, UK: Cambridge University Press, 2003.

[23] W. T. Fitch, *The Evolution of Language*. Cambridge University Press, 2010.

[24] J. Gros-Louis, M. J. West, M. H. Goldstein, and A. P. King, "Mothers provide differential feedback to infants' prelinguistic sounds," *International Journal of Behavioral Development*, vol. 30, no. 5, pp. 112–119, 2006.

[25] M. H. Goldstein and J. A. Schwade, "Social feedback to infants' babbling facilitates rapid phonological learning," *Psychological Science*, vol. 19, no. 5, pp. 515–523, 2008.

[26] S. Maeda, "On articulatory and acoustic variabilities," *Journal of Phonetics*, vol. 19, pp. 321–331, 1991.

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

[28] B. C. J. Moore, B. G. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *Journal of the Audio Engineering Society*, vol. 45, no. 4, pp. 224–240, 1997.

[29] A. Jansen and P. Niyogi, "Intrinsic fourier analysis on the manifold of speech sounds," in *in IEEE Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2006, pp. 241–244.

[30] ——, "Semi-supervised learning of speech sounds," in *Proceedings of INTERSPEECH 2007*, 2007.

[31] P. Niyogi, "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)*, 2004.

[32] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, pp. 1373–1396, 2003.

[33] C. Wang, "A geometric framework for transfer learning using manifold alignment," Ph.D. dissertation, University of Mass. Amherst, 2010.

[34] J. Ham, D. D. Lee, and L. K. Saul, "Semisupervised alignment of manifolds," in *Proc. of the Ann. Conf. on Uncertainty in AI*, Z. Ghahramani and R. Cowell, Eds., vol. 10, 2005, pp. 120–127.

[35] F. R. K. Chung, *Spectral Graph Theory*, ser. Regional Conference Series in Mathematics. American Mathematical Society, 1997, number 92.

[36] S. Rosenberg, *The Laplacian on a Riemannian manifold: an introduction to analysis on manifolds*. Cambridge University Press, 1997.