Learning speaker normalization using semisupervised manifold alignment

Andrew R. Plummer[1], *Mary E. Beckman*[1], *Mikhail Belkin*[2], *Eric Fosler-Lussier*[1,2],*Benjamin Munson*[3]

[1]Department of Linguistics, The Ohio State University, Columbus, OH, USA
[2]Dept. of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA
[3]Department of Speech-Language-Hearing Sciences, University of Minnesota, USA
{plummer,mbeckman}@ling.osu.edu, {mbelkin,fosler}@cse.osu.edu, munso005@umn.edu

Abstract

As a child acquires language, he or she: perceives acoustic information in his or her surrounding environment; identifies portions of the ambient acoustic information as language-related; and associates that language-related information with his or her perception of his or her own language-related acoustic productions. The present work models the third task. We use a semisupervised alignment algorithm based on manifold learning. We discuss the concepts behind this approach, and the application of the algorithm to this task. We present experimental evidence indicating the usefulness of manifold alignment in learning speaker normalization.

Index Terms: speaker normalization, manifold alignment, language acquisition

1. Introduction

As infants acquire language, they: perceive acoustic information in their surrounding environment; identify portions of the ambient acoustic information as language-related; and associate that language-related information with their perception of their own language-related acoustic productions. The present work focuses on the third action, called the *association task*, composed of the following subtasks:

- GROUPING. Infants need to group together adult signals they perceive as similar. Infants also need to group together their own signals that they perceive as similar.
- MAPPING. Infants need to map groups of adult signals to groups of infant signals in a reasonable way. For example, an adult $/a/$ must be mapped to an infant $/a/$.

In this paper we describe an acquisition model that incorporates both GROUPING and MAPPING. That is, we provide a full model of the association task, focusing primarily on MAPPING.

Prior acquisition models address GROUPING, whereas MAPPING is treated in preprocessing, or simply omitted. For example, [1] uses a neural network-based approach wherein infant and adult acoustic productions are modeled using a neural map. The map is trained on synthetic data intended to represent the babbling of an infant, and acoustic data derived from vowels segmented from actual adult speech. Peaks in the babbling output more or less coincide with peaks in the adult input. Since the synthesized data ranges over the adult vowel space, the natural differences between infant acoustics and adult acoustics are not accounted for. That is, MAPPING is omitted from the model. The acquisition model in [2], which builds on [1], models disjoint infant and adult vowel spaces using a preprocessing normalization step that maps the two spaces into a common space. That is, MAPPING is treated prior to learning.

The difficulty of the MAPPING task appears in other domains, such as speech recognition and sociophonetics [3, 4], that handle data drawn from multiple speakers. Typically, a normalization algorithm is applied to the acoustic data before analysis is carried out. The fact of cross-language differences in speaker effects indicates that normalization is a learned process [5]. In this light, MAPPING is performing speaker normalization. The challenge is to model the learning of this process.

We approach this challenge using recent developments in manifold learning in speech analysis [6, 7, 8]. We conceptualize the adult and infant vowel spaces as geometrically similar low-dimensional manifolds (Section 2), and use a manifold-based laplacian classifier [9] to perform the GROUPING task (Section 3). Given the geometric similarity of the adult and infant spaces, we model MAPPING as a continuous transformation using a semisupervised manifold alignment algorithm [10] (Section 3). We describe a model that incorporates both components of the association task in vowel acquisition (Section 4). Finally, we measure the model's potential usefulness and discuss improvements and feasible expansions on the present work (Section 5).

We conclude this section by formalizing the (semisupervised) association task. We make the following simplifying assumptions: (i) The ambient acoustic information is only that produced by an adult woman (adult) and a six-month-old infant. (ii) The acoustic information is vocalic (thus GROUPING yields groups of vowels). (iii) The number of adult and infant vowel groups is the same. (iv) Geometric relationships between adult vowel groups are preserved by MAPPING. This ensures that adult vowel groups are reasonably mapped to infant vowel groups. (v) There is a set of exemplar adult acoustic vectors, each corresponding to an infant acoustic vector.

We assume two distinct articulatory systems A and I that output signals. The system A represents an adult female articulatory system, and I an infant articulatory system. Our data set is a collection of vectors of acoustic information derived from vocalic signals s output by A and I as follows:

- Given a vocalic signal s, we recover formant values F1, F2, F3, and F4 (in Barks) from s at a single point in time.
- We denote the formant values as $f_1^s, f_2^s, f_3^s, f_4^s$ (usually omitting the superscript), and use them to form a vector $\langle f_1, f_2, f_3, f_4 \rangle$.

The vectors derived from vocalic signals are called *acoustic vectors*. The *adult vowel space* V_A is the collection of acoustic vectors derived from A. Similarly, the *infant vowel space* V_I is the collection of acoustic vectors derived from I. Both spaces are represented in Figure 1a).

We focus on acoustic vectors with four formant values $\langle f_1, f_2, f_3, f_4 \rangle$, and simplified acoustic vectors with values for

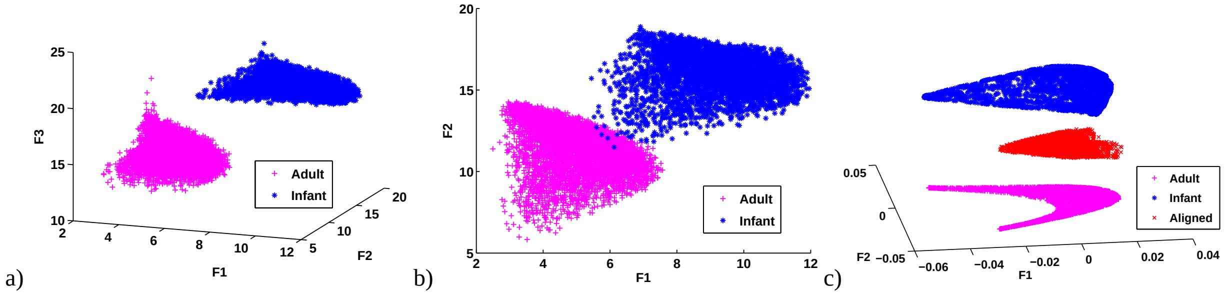

Figure 1: The adult and infant acoustic vectors in magenta and blue, respectively. a) depicts the simplified adult and infant vowel spaces. b) depicts the simplified spaces collapsed into the F1 and F2 space, and the significant lack of overlap. c) depicts the two-dimensional adult (blue) and infant (magenta) vowel manifolds in a normalized F1 and F2 space. We use their similar geometry to align them (red).

only the first three formants $\langle f_1, f_2, f_3 \rangle$, Given this data representation, we formulate the subtasks of GROUPING and MAPPING as follows:

- Group the acoustic vectors in V_A into v-many categories $\mathsf{a}_1, \ldots, \mathsf{a}_v$, called *adult vowel categories*. Group the acoustic vectors in V_I into v-many categories $\mathsf{i}_1, \ldots, \mathsf{i}_v$, called *infant vowel categories*.
- Map the categories $\mathsf{a}_1, \ldots, \mathsf{a}_v$ to the categories $\mathsf{i}_1, \ldots, \mathsf{i}_v$, preserving local geometric relationships. That is, achieve a reasonable correspondence between $\mathsf{a}_1, \ldots, \mathsf{a}_v$ and $\mathsf{i}_1, \ldots, \mathsf{i}_v$ (using exemplar pairs).

2. Manifolds

We briefly present the conceptual basis for the manifold learning approach we intend to use. For a detailed introduction to manifolds see [11]. Intuitively, a *manifold* is a subset of a higher-dimensional space that is similar to a lower-dimensional Euclidean space. For example, the surface of a sphere in three-dimensional space, though not perfectly flat, is similar to the Euclidean plane (we say that the surface is a *two-dimensional manifold*). This similarity yields useful lower-dimensional representations of the higher-dimensional data. For example, the approximately spherical Earth is represented using two-dimensional (planar) maps. The simplified vowel spaces V_I and V_A (Figure 1a)) are situated in three-dimensional space, yet they are both similar to the Euclidean plane (i.e., they are two-dimensional manifolds). We want to take advantage of their lower-dimensional representations, depicted in Figure 1c), for GROUPING and MAPPING.

Consider the manifold V_A. Being similar to the plane means that there is a family of functions on V_A, each smoothly mapping a portion of V_A to the Euclidean plane. In practice, we do not have these smooth functions immediately in hand.

1. We need a way to approximate these smooth, dimensionality-reducing functions, as well as other smooth functions defined on our manifolds,
2. We need a way to measure the smoothness of these approximations.

For details on the mathematical development of these desiderata, see [12]. We briefly recount the needed components. Denote by $\mathcal{L}^2(V_A)$ the set of smooth functions on V_A. The *laplace-beltrami operator* is defined as the divergence of the gradient ($\Delta f = div(grad(f))$) of functions $f \in \mathcal{L}^2(V_A)$. This operator yields both a simple approximation of smooth functions on V_A, and a simple method for measuring their smoothness.

For computational purposes, V_A is discretely modeled by a weighted graph G. Points in V_A are represented by the vertices of G. The weight between vertices i and j represents the closeness of those points on the manifold. The laplace-beltrami operator is discretely modeled as $L = D - W$ where:

- W is the weight matrix of G,
- D is the diagonal matrix with $D_{ii} = \sum_j W_{ij}$.

The matrix L is the called *graph Laplacian* of G.

The graph Laplacian yields a smoothness measure as follows. Assume the vertices of G are numbered $1, \ldots, n$, and let $f = (f_1, \ldots, f_n)$ be a function on G. Suppose w_{ij} is large (i.e., i and j are close). For f to be smooth, we want $(f_i - f_j)^2$ to be very small. That is, we want $w_{ij}(f_i - f_j)^2 \approx 0$ for each i, j. Define the smoothness measure of f on G as:

$$S_G(f) = \sum_{i \sim j} w_{ij}(f_i - f_j)^2 = fLf^T \qquad (1)$$

A function f is *smooth* if $S_G(f) \approx 0$. The alignment algorithm we use to model MAPPING crucially depends on the smoothness measure, since it measures the preservation of local geometry on V_A.

3. Algorithms

We present the manifold algorithms used in our model. We begin with a simple laplacian classifier [9] used to model GROUPING. We present the algorithm for a general vowel space $V \subseteq \mathbb{R}^n$ with vowel categories $\mathsf{z}_1, \ldots, \mathsf{z}_v$. Let $x_1, \ldots x_k \in V$ and let the natural numbers $1, \ldots, v$ represent the categories $\mathsf{z}_1, \ldots, \mathsf{z}_v$. We assume that the categories of the first $s < k$ points are known, and denote the category of x_i as b_i ($i \leq s$). The algorithm is as follows:

1. Construct a weighted graph G which has k vertices representing the points $x_1, \ldots, x_k$. Compute the graph Laplacian L of G.
2. Let L_1 be a $s \times s$ matrix, L_2 a $s \times (k - s)$ matrix, and L_3 a $(k - s) \times (k - s)$ matrix such that

$$L = \begin{pmatrix} L_1 & L_2 \\ L_2^T & L_3 \end{pmatrix} \qquad (2)$$

3. Compute $\bar{b} = L_3^{-1} L_2^T (b_1, \ldots, b_s)^T$. Then $\bar{b} = (\bar{b}_1, \ldots, \bar{b}_{k-s})$, and the category of x_{s+j} is $\bar{b}_j$.

The theoretical aspects of this algorithm, and comparisons to other semisupervised approaches to data classification are presented in [9]. Manifold-based classification of speech sounds is also analyzed in [8, 7].

For the MAPPING task, we want to learn a transformation T from V_A to V_I, that preserves the similarities of their geometries. Rather than learning T directly, we learn a *factorization* of T. We factor T into composite functions f and g, with an intermediate space V_Z, where $f : V_A \rightarrow V_Z$ and $g : V_I \rightarrow V_Z$. Conceptually, we are collapsing the manifolds V_I and V_A onto another manifold V_Z (see Figure 1c)). We then learn T by learning f and g.

We learn f and g using the semisupervised manifold alignment algorithm in [10]. Suppose we have a set of exemplar vectors $V_X \subseteq V_A$ that correspond to a set of vectors $V_Y \subseteq V_I$, both of cardinality ℓ. We assume this correspondence is a bijection $x_i \mapsto y_i$ between V_X and V_Y that is representative of the geometry-preserving transformation T. We want to factor T in functions f and g. Since x_i corresponds to y_i, we want f_i close to g_i, that is, we want $(f_i - g_i)^2$ to be small. Moreover, we want local geometries to be preserved; that is, we want f to be smooth on V_A, and g to be smooth on V_I. Let L_A, and L_I be the laplacian on the graphical representation of V_A and V_I. We use the smoothness measure in (1) to minimize:

$$C(f, g) =_{def} \mu \sum_{i=1}^{\ell} (f_i - g_i)^2 + f L_A f^T + g L_I g^T \qquad (3)$$

where μ is a weight indicating the importance of the alignment. To properly minimize (3), we need it to be invariant under simultaneous scaling of f and g. Thus, we minimize:

$$\bar{C}(f, g) =_{def} \frac{C(f, g)}{f^T f + g^T g} \qquad (4)$$

Let $\{x_1, \ldots, x_k\} = V_A \subseteq \mathbb{R}^n$ and $\{y_1, \ldots, y_q\} = V_I \subseteq \mathbb{R}^n$ such that x_i corresponds to y_i for $1 \leq i \leq \ell$ and $\ell < min(k, q)$. The algorithm is as follows:

1. Define $U^{\alpha\beta}$ as an $\alpha \times \beta$ matrix such that

$$U_{ij}^{\alpha\beta} = \begin{cases} \mu & i = j \leq \ell \\ 0 & otherwise. \end{cases} \qquad (5)$$

2. Compute the laplacians L_A and L_I, and let

$$L_Z = \begin{pmatrix} L_A + U^{kk} & -U^{kq} \\ -U^{qk} & L_I + U^{qq} \end{pmatrix} \qquad (6)$$

3. Compute the eigenvectors h of L_Z, where $h = [f^T g^T]^T$, which is equivalent to minimizing (4).

4. Let $h_1, \ldots, h_r$ be the eigenvectors corresponding to the r smallest non-zero eigenvalues of L_Z. We have the following factorization:

$$x_i \in V_A \mapsto \langle f_1(i), \ldots, f_r(i) \rangle$$
$$y_i \in V_I \mapsto \langle g_1(i), \ldots, g_r(i) \rangle$$

The number of eigenvectors used (r) depends on the precision desired in approximating f and g (and thus T).

4. Data and model

Our model requires representation of acoustic productions of both an infant and an adult woman. We use data corresponding to the following two subcomponents of the Variable Linear Articulatory Model (VLAM, [13]): (1) the formant frequencies yielded by simulation of a six-month-old male vocal tract, and (2) the formant frequencies yielded by the simulation of a 16-year-old male vocal tract, which corresponds to the range of vowels that an adult woman might produce. Each data set contains approximately 4000 acoustic data points which are vectors whose components are the first four formant frequencies (F1, F2, F3, F4). We convert all of the formant values into the Bark scale to approximate their psychoacoustic value.

The exemplar data we use for semisupervised learning were perceptually categorized [14]. The stimuli were 408 synthesized vowels. These represented 38 tokens generated by the VLAM model for each of seven different vocal tract growth stages. These included the six-month-old and 16-year-old vocal tracts which yielded the data sets described above. The 38 tokens were meant to represent 38 commonly occurring acoustic prototypes, as determined by previous cross-linguistic surveys. Each of these was fit to an appropriate location in the maximal vowel space for each vocal-tract stage.

21 Greek listeners, 20 Korean listeners, and 21 English-speaking listeners were tested by a same-language-speaking research assistant, with all instructions translated to the native language. Stimuli were blocked by vocal-tract age. Presentation was randomized within blocks, and block order was randomized. On each trial, the listener heard a vowel, then clicked on a vowel category label on a computer screen to indicate the chosen vowel type. The number of labels reflected the number of monophthongal vowels in the language (i.e., 5 for Greek, 7 for Korean, and 12 for English). The labels were then replaced by a line visual analog scale, on which the listener clicked to record the “goodness” or confidence in the identification. A data point received a consensus label (reflecting the “community norm”) if inter-listener agreement on that label was above chance on a binomial test and there were no ties for the top-ranked choice.

We take the adult VLAM data to be V_A, and the infant VLAM data to be V_I. Let $V_X = \{x_1^A, \ldots, x_{38}^A\}$ and $V_Y = \{y_1^I, \ldots, y_{38}^I\}$ be the set of perceptually categorized adult and infant exemplar vectors. Let a_i^j be the category of x_j^A, and let i_i^j be the category of y_j^I.

1. For each x_j^A, compute its n nearest neighbors in V_A and assign them category a_i^j. Call the derived set of categorized adult acoustic vectors A_{cat}. Carry this process out for each y_i^I, and call the derived set of categorized infant acoustic vectors I_{cat}.

2. Run the laplacian classifier (GROUPING) on V_A using A_{cat} as training data. Similarly, run GROUPING on V_I using I_{cat} as training data.

3. For each adult vowel category a_i, let $A_{\text{cat}}^{\mathsf{a}_i}$ be the set of adult acoustic vectors in A_{cat} with category a_i. Compute the mean of each $A_{\text{cat}}^{\mathsf{a}_i}$, and compute the m nearest neighbors of this mean. Call this set of nearest neighbors $A_{\text{align}}^{\mathsf{a}_i}$. In similar fashion, compute each $I_{\text{align}}^{\mathsf{i}_i}$. Let

$$A_{\text{align}} = \begin{pmatrix} A_{\text{align}}^{\mathsf{a}_1} \\ \vdots \\ A_{\text{align}}^{\mathsf{a}_v} \end{pmatrix} \quad I_{\text{align}} = \begin{pmatrix} I_{\text{align}}^{\mathsf{i}_1} \\ \vdots \\ I_{\text{align}}^{\mathsf{i}_v} \end{pmatrix} \qquad (7)$$

4. Run the semisupervised alignment algorithm (MAPPING) on V_A and V_I, using A_{align} and I_{align} for alignment training.

5. Results and discussion

The perceptual data yields three sets of categorized exemplar data, allowing us three sets of experiments. The model algorithm was run using the Greek, Korean, and English perceptual data, in turn, as V_X and V_Y. The Greek perceptual categorization yielded five adult and infant vowel categories. The Korean perceptual categorization yielded seven adult vowel categories, and five infant vowel categories. In some cases, a category had only one representative point. We excluded such points, resulting in a simplified categorization, with four adult and infant vowel categories. The English perceptual categorization yielded ten adult vowel categories, and seven infant vowel categories. Again, in some cases a category had only one representative point. We excluded such points, resulting in a simplified categorization, with four adult and infant vowel categories (dividing the vowel spaces in a different way from the four categories of the simplified Korean grouping).

We performed the experiments in two phases, one using acoustic vectors with three formant components $\langle f_1, f_2, f_3 \rangle$, the other using acoustic vectors with four formant components $\langle f_1, f_2, f_3, f_4 \rangle$. In both cases, training parameters were as follows: we computed the five nearest neighbors of each vector in Step 1., and 100 nearest neighbors in Step 3. In testing, we use 'leave-one-out' cross-validation on V_X (the perceptually categorized adult data) for each language. Both GROUPING and MAPPING were tested on the held-out vector. For GROUPING, we simply compared the known category of the held-out vector against the category assigned by the laplacian classifier. Correctness is defined as having the same category. Correctness results for GROUPING on vectors with three (four) formant values were – Greek: 75% (75%); Korean: 88% (85%); English: 93% (90%). These accuracy rates are likely to be a coarser representation of the infant's sensory map than we are trying to model with the two manifolds. We might achieve better GROUPING if we used individual results from the perception experiment instead of consensus results and referred to the goodness ratings. We might also need a more densely sampled set of perceptual responses. The accuracies nevertheless seem high enough to attempt an initial test of MAPPING.

Table 1: *Mapping Results*. Values indicate percent of correctly mapped test cases.

Eigenvectors	Greek		Korean		English	
2	54	51	67	58	60	83
3	64	51	64	67	80	63
4	62	67	58	67	80	67
Num. of Formants	3	4	3	4	3	4

For MAPPING, we use the manifold V_Z (the space in which V_A and V_I are aligned), and the functions $f : V_A \rightarrow V_Z$ and $g : V_I \rightarrow V_Z$ learned by the alignment algorithm. We use f to map the held-out vector to a point z on V_Z. We then computed the 15 nearest neighbors of z in V_Z, mapped these back to V_I using the inverse of g, recovered their categories assigned by the laplacian classifier, selected the mode category, and compared it to that of the held-out point. Correctness is defined as for GROUPING. Correctness results for MAPPING are shown in Table 1. The number r of smallest non-constant eigenvectors used in the alignment algorithm determines the precision of the approximations f and g.

Using $r > 4$ did not improve performance. The coarseness of GROUPING likely precludes improved performance of finer-grained approximations of MAPPING. Although these accuracy rates are considerably lower than what we might expect from adult perception of vowels in isolation (e.g. [15]), they are a promising start. Using more eigenvectors would almost surely result in greater accuracy but requires a more densely sampled perceptual space. We are currently obtaining more perceptually categorized data.

6. Acknowledgments

We are thankful for the contributions of Ilana Heintz and Patrick Reidy. This work was funded by NSF grant BCS-0739206.

7. References

[1] G. Westermann and E. R. Miranda, "A new model of sensorimotor coupling in the development of speech," *Brain and Language*, vol. 89, pp. 393–400, 2004.

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

[3] 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.

[4] P. Adank, R. Smits, and R. van Hout, "A comparison of vowel normalization procedures for language variation research," *Journal of the Acoustical Society of America*, vol. 116, pp. 3099–3107, 2004.

[5] K. Johnson, "Resonance in an exemplar-based lexicon: The emergence of social identity and phonology," *Journal of Phonetics*, vol. 34, pp. 485–499, 2006.

[6] 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.

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

[8] A. Errity and J. McKenna, "An investigation of manifold learning for speech analysis," in *Proceedings of The International Conference on Spoken Language Processing*, 2006, pp. 2506–2509.

[9] J. Shawe-Taylor and Y. Singer, Eds., *Regularization and Semi-supervised Learning on Large Graphs*, ser. Lecture Notes in Computer Science, vol. 3120. Springer, 2004.

[10] J. Ham, D. D. Lee, and L. K. Saul, "Semisupervised alignment of manifolds," in *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence*, Z. Ghahramani and R. Cowell, Eds., vol. 10, 2005, pp. 120–127.

[11] M. Spivak, *Calculus on Manifolds*. W. A. Benjamin, Inc., 1965.

[12] M. Belkin and P. Niyogi, "Semi-supervised learning on riemannian manifolds." *Machine Learning*, vol. 56, no. 1-3, pp. 209–239, 2004.

[13] 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.

[14] 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 categorization study (A)," *Journal of the Acoustical Society of America*, vol. 127, p. 2018, 2010.

[15] T. M. Neary and P. F. Assmann, "Modeling the role of inherent spectral change in vowel identification," *Journal of the Acoustical Society of America*, vol. 80, pp. 1297–1308, 1986.