Computing low-dimensional representations of speech from socio-auditory structures for phonetic analyses

Andrew R. Plummer[a,*], Patrick F. Reidy[b]

[a]*Ohio State University, Columbus, OH, USA*
[b]*Callier Center for Communication Disorders, University of Texas at Dallas, Dallas, Texas, USA*

Abstract

Low-dimensional representations of speech data, such as formant values extracted by linear predictive coding analysis or spectral moments computed from whole spectra viewed as probability distributions, have been instrumental in both phonetic and phonological analyses over the last few decades. In this paper, we present a framework for computing low-dimensional representations of speech data based on two assumptions: that speech data represented in high-dimensional data spaces lie on shapes called *manifolds* that can be used to map speech data to low-dimensional coordinate spaces, and that manifolds underlying speech data are generated from a combination of language-specific lexical, phonological, and phonetic information as well as culture-specific socio-indexical information that is expressed by talkers of a given speech community. We demonstrate the basic mechanics of the framework by carrying out an analysis of children's productions of sibilant fricatives relative to those of adults in their speech community using the phoneigen package – a publicly available implementation of the framework. We focus the demonstration on enumerating the steps for constructing manifolds from data and then using them to map the data to a low-dimensional space, explicating how manifold structure affects the learned low-dimensional representations, and comparing the use of these representations against standard acoustic features in a phonetic analysis. We conclude with a

*Corresponding author
Email address: arplummer.research@gmail.com (Andrew R. Plummer)

discussion of the framework's underlying assumptions, its broader modeling potential, and its position relative to recent advances in the field of representation learning.

Keywords: manifold alignment, Laplacian Eigenmaps, socio-indexical, phonetic categories, low-dimensional representations of speech

---

1. Introduction

The formulation of the source-filter model within the acoustic theory of speech production (Chiba & Kajiyama, 1941; Fant, 1960) provided both the conceptual and technical basis for characterizing speech sound segments in terms of a small number of acoustically grounded features (as in Jakobson et al., 1951). These theoretical advances helped frame the development of digital signal processing techniques for recovering source and filter information from audio recordings of speech signals. Linear predictive coding (LPC) techniques in particular (see Burg, 1967; Makhoul, 1973, inter alia) yielded implementations of efficient algorithms for extracting formant and fundamental frequency values from segments of recorded speech. These low-dimensional representations of speech data are widely used in phonetic analyses due to the incorporation of LPC implementations in software packages for computational analysis (e.g., Praat; Boersma, 2001) that have been adopted by a large portion of the phonetics research community. While criticisms of LPC have been noted, particularly in the case of formant extraction from speech data with high fundamental frequencies (see e.g., Story & Bunton, 2016, for discussion as well as an alternative method for estimating formant values in the speech of young children), LPC techniques nevertheless underlie much of the computational analysis carried out by phoneticians.

For sound types that do not have a glottal source, such as voiceless obstruents, different types of methods have been developed that avoid decomposing the spectrum into source and filter components. For example, the method of spectral moments (e.g., Forrest et al., 1988) treats a power spectrum as a discrete

probability mass function and computes from it a small number of parameters that index its gross distributional properties. The quantities most commonly used as spectral moments are the first moment (mean or centroid), which indicates the center of gravity of the distribution along the frequency scale; the second (central) moment (variance) or its square root (standard deviation), both of which indicate the spread of the distribution; the third standardized moment (skewness), a unitless quantity that indicates the (a)symmetry of the distribution; and the fourth standardized moment (kurtosis), a unitless quantity that indicates the heaviness of the tails of the distribution. Another example is discrete cosine transform (DCT) coefficients, which projects the spectrum onto a basis of cosine functions (e.g., Nossair & Zahorian, 1991). This latter method yields a lower-dimensional representation of the spectrum (e.g., 10 or fewer coefficients) that may be used to reconstruct a smoothed version of the spectrum.

In this paper, we extend the long line of approaches to computing low-dimensional representations of speech data (e.g., those described above, inter alia) by putting forward a computational framework based on the following two assumptions. The first, termed the *manifold assumption*, holds that speech data represented in high-dimensional data spaces lie on shapes called *manifolds* that are embedded within these data spaces and that can be mapped to low-dimensional coordinate spaces that facilitate phonetic analyses. The second, termed the *socio-indexical assumption*, concerns the types of information that may be used to characterize the structure of manifolds underlying data. Stated explicitly, the socio-indexical assumption entails that the manifolds underlying speech data are generated from a combination of language-specific lexical, phonological, and phonetic information and culture-specific socio-indexical information that is expressed by talkers of a given speech community during speech production. The framework is currently implemented via the phoneigen [1] pack-

[1]The source code for the phoneigen package is publicly available online at https://github.com/patrickreidy/phoneigen, along with a description of the package and in-

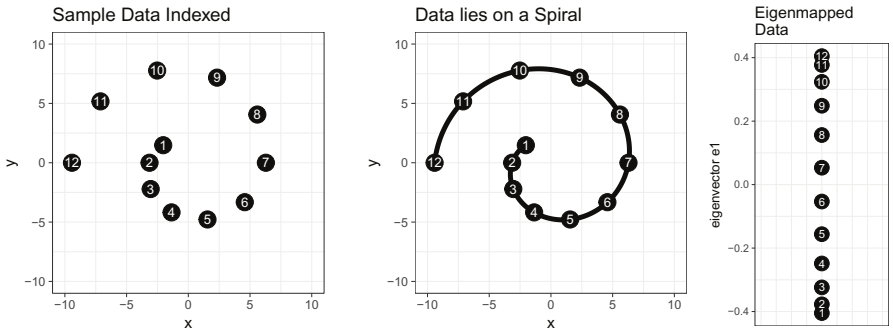


Figure 1: (left) Two-dimensional data lies on the one-dimensional Archimedes Spiral manifold (middle). Data eigenmapped to lower-dimensional representations (right).

age, which is demonstrated in detail in Section 2. Specifically, we use phoneigen in an analysis of the acquisition of sibilant fricatives produced by young children with respect to the productions of adults in their speech community. Before proceeding to the example analysis, we briefly discuss the motivation for each of the two assumptions underlying our framework.

Motivation for the manifold assumption stems (in part) from the emergence of computational techniques for computing representations of data that lie on manifolds within the field of representation learning (see Bengio et al., 2013, for a review). In particular, several graph-based methods (see Tenenbaum et al., 2000; Roweis & Saul, 2000; Seung & Lee, 2000; Belkin & Niyogi, 2003, inter alia) were put forward for computing low-dimensional representations of data sets situated in higher-dimensional spaces that reflect their intrinsic dimensionality. To illustrate the main idea, the data shown in Figure 1 (left) is situated in a two-dimensional data space, while the intrinsic dimensionality of the data is one-dimensional, as it lies on an Archimedes Spiral embedded within the data space (middle). These graph-based techniques provide the computational means for mapping the ostensibly two-dimensional data to a one-dimensional coordinate space that reflects the underlying manifold structure. In brief, data

stallation instructions.

points correspond to nodes in a graph, while edges that connect nodes are constructed based on a specified distance computation. The resulting graph is used to map the data to a lower-dimensional space. Figure 1 (right) shows the output of the *Laplacian Eigenmaps* technique (Belkin & Niyogi, 2003) applied to the data[2], which is derived from eigenvectors of a matrix representation of the graph called the *Laplacian matrix*, *graph Laplacian*, or simply the *Laplacian* (see Chung, 1997; Cvetković et al., 2010).

The Laplacian Eigenmaps technique is implemented as a part of the phoneigen package used in Section 2 to demonstrate the applicability of our framework in phonetic analysis. The computations used in the demonstration are reviewed in Appendix A (hence, readers unfamiliar with graph-based methods may want to read material therein before proceeding with Section 2). In general, graph-based manifold methods have been usefully applied across a range of speech science modeling problems including phonetic category learning (e.g., Jansen & Niyogi, 2007; Plummer et al., 2010; Plummer, 2014) and automatic speech recognition (e.g., Errity & McKenna, 2006; Jafari & Almasganj, 2010; Zhao & Zhang, 2012; Norouzian et al., 2013; Tomar & Rose, 2014; Huang et al., 2016b,a). Section 3 discusses broader motivation for the manifold assumption as well as how our framework is situated with respect to recent advances in representation learning within the machine learning community.

Broader motivation for the socio-indexical assumption is reserved for Section 3, which reviews several experiments and analyses that inform our burgeoning understanding of the relationships between language-specific lexical and phonological representations and phonetic categories. However, the main idea behind the assumption can be illustrated by again using the data and the coordinate space in Figure 1 as a toy model of a sensory space. The Archimedes Spiral manifold structure underlying the data points is characterized by a fixed formal mathematical rule—i.e., the points in the xy-plane that lie on the Archimedes spiral are those points that satisfy the equation $r = a\theta$ (expressed in polar

[2]The exact procedure for generating the output is covered in Appendix A.

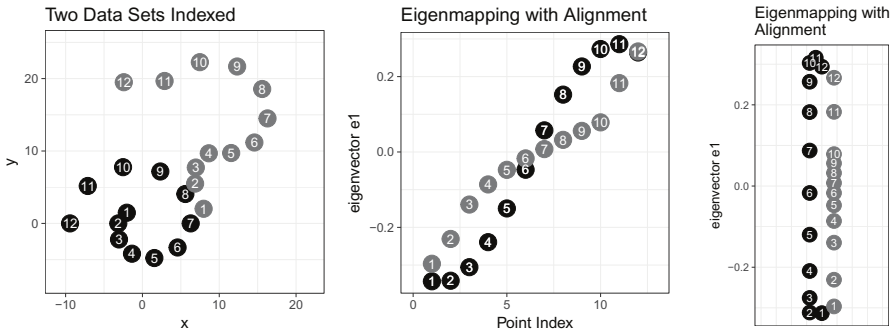


Figure 2: (left) The two data sets P (black) and Q (gray). Eigenmapping in terms of point indices in the original data sets (middle) and a one-dimensional analysis space (right) where representations are slightly translated for visibility.

coordinates). Yet, we depart from this type of characterization of underlying manifold structure. Rather, within our framework, the data points need not lie on a spiral or any other shape described by a fixed formula. The points may instead lie on one or any number of manifolds constructed from language-specific phonological or lexical information, socially- or culturally-derived information, or broader cognitive information.

Indeed, a key aspect of the approach is that language-specific and socio-cultural-specific information can be used to "align" data sets that occupy different areas of a sensory space. For example, suppose P and Q in Figure 2 (left) are sets of auditory data produced by two different talkers. Moreover, suppose the indices of the points encode lexical items. Graphs (i.e., manifolds) can be constructed over P and Q and then combined by adding edges that connect a subset of the nodes based on like lexical items. This computation, called *manifold alignment* (see, e.g., Hamm et al., 2005, as well as Chapter 5 in Ma & Fu, 2012), is applied in the phoneigen example analysis in Section 2 and reviewed in Appendix A. Crucially, the auditory information, the socio-indexical information, and the lexical information are all reflected in the resulting *socio-auditory manifold*. Applying the Laplacian Eigenmaps technique to the combined graph

yields the aligned representations[3] in Figure 2 (middle and right). Under this approach, the superposing language-specific and socio-cultural-specific information and its interaction with the sensory space information are thus made an integral part of the computations that map speech data to low-dimensional spaces for phonetic analyses. In this connection, Section 2 provides a concrete analytic example in support of the socio-indexical assumption. In the analysis, during graph construction the nodes of the graphs still correspond to speech data points, but crucially edges that connect nodes are constructed based on socio-indexical and lexical information in addition to a specified distance computation. Details of the procedure are provided in the demonstration that follows.

2. Socio-auditory manifolds and eigenmapping for phonetic analyses

This section demonstrates how to use functions from the phoneigen package in order to compute low-dimensional phonetic representations for the voiceless sibilant fricatives /s/ and /ʃ/ produced by native English-speaking adults and native English-acquiring two- and three-year-old children. These consonants are articulated by raising the tongue toward the roof of the mouth, so as to form a narrow constriction in the oral cavity. Turbulence noise sources are generated when the air flowing through this linguopalatal constriction becomes turbulent and when the turbulent jet impinges on the teeth downstream from the constriction. These noise sources excite the cavity anterior to the constriction. Consequently, the difference in place of articulation between sibilant fricatives is represented in the distribution of energy in their respective spectra: e.g., in an adult's fluent productions, the more anterior place of articulation for /s/, compared with /ʃ/, entails a smaller front-cavity volume and thus resonances at higher frequencies (see Figure 3).

While an energy spectrum is a convenient representation that is easily computed from the frication noise of a sibilant, one problem posed by such a rep-

[3] Again, the exact procedure for generating the output is covered in Appendix A.

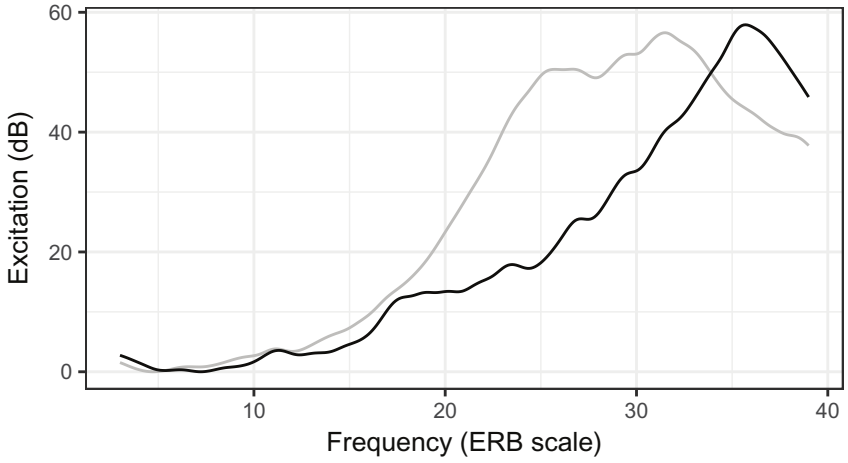


Figure 3: Excitation patterns computed from participant A50N's productions of /s/ (black) and /ʃ/ (gray).

resentation is its high dimensionality (e.g., a spectrum estimated from a 20-ms window of frication noise sampled at 44.1 kHz will be a 441-dimensional vector). In the examples that follow, high-dimensional representations of /s/ and /ʃ/ are mapped to a low-dimensional representation that denotes the phonetic differences in place of articulation between these fricatives. Moreover, this mapping is learned by constructing a *socio-auditory manifold*, a graph structure that represents acoustic properties of the high-dimensional spectral representations of the speech data within a superposing interpersonal network structure that links the speakers who produced the tokens of /s/ and /ʃ/. The goals of this section are: first, to demonstrate how to compose a sequence of functions from the phoneigen package in order to take a data set comprising high-dimensional observations as input, construct a socio-auditory manifold, and derive a low-dimensional representation as output; and, second, to demonstrate that the derived low-dimensional representation depends crucially on the structure of the socio-auditory manifold. Readers unfamiliar with graph-based methods may wish to read Appendix A before proceeding.

2.1. Overview of the data set

The productions of /s/ and /ʃ/ used in this section were provided by the Learning to Talk Project (NIDCD grant R01-02932; Principal Investigators: Jan Edwards, Mary E. Beckman, and Benjamin Munson), a longitudinal study of phonological and lexical development in preschool children. As part of this project, adult speakers were also tested in order to assess the adult production norms in the ambient community in which the child participants were being raised. Additionally, the data and some of the results reported here have been previously reported in Reidy et al. (2017).

The data set is accessed by calling[4]:

```
SibilantFricatives()
```

This returns a tibble object that comprises 2475 observations on 14 variables. These productions of target /s/ and /ʃ/ were elicited from 69 children (33 girls, 36 boys), between the ages of 28 and 39 months, and 16 adults (10 women, 6 men) with a picture-prompted word repetition task. The test words for this task were selected in order to elicit multiple attempts of each target fricative in word-initial position, across a variety of following vowel contexts. On each trial, a picture of the referent of the test word was displayed on a computer screen, and then an audio recording of the test word, spoken by a female adult native speaker, was played over loudspeakers. The participants were asked to repeat, into a microphone, the word that they heard (see Edwards & Beckman, 2008a,b for discussion of the word-repetition task) . Each experiment session was recorded digitally at 44.1 kHz sampling rate and 16 bit resolution. Each attempted production of a target was transcribed by one of a team of

[4]The content of this section of the paper is also available as a vignette within the phoneigen package. After installing the phoneigen package, the vignette can be retrieved by calling in an R environment vignette("socio-auditory-manifolds", "phoneigen"). The vignette itself includes additional expository code that has been omitted here due to considerations of space. All source code for the vignette (i.e., for all analyses and figures) is available in the file https://github.com/patrickreidy/phoneigen/vignettes/socio-auditory-manifolds.Rmd.

phonetically-trained research assistants. For those productions whose manner was judged to be a sibilant fricative, the place of articulation was denoted along a four-point scale: [s], [s]:[ʃ] (intermediate, but closer to [s]), [ʃ]:[s] (intermediate, but closer to [ʃ]), and [ʃ]. If a production was judged to be a sibilant fricative, then the research assistant also manually annotated the onset of frication and the onset of voicing.

Each production by a child that was transcribed as a sibilant fricative (including substitution errors, such as [s] for target /ʃ/) was used as a stimulus in a visual-analog-scaling (VAS) perceptual rating task: the initial consonant-vowel sequence was extracted, beginning 5 ms prior to the onset of frication and ending 150 ms after the onset of voicing in the vowel. Batches of these stimuli were played to native adult English-speaking listeners who were asked to rate the frication noise in a production by clicking some point on a double-headed arrow that was presented visually on a computer monitor. This arrow was anchored by the text "the 's' sound" at one end and by "the 'sh' sound" at the other. The click location in pixels was logged automatically, and the pixel locations were normalized to fall within the $[0, 1]$ interval, with lesser ratings indicating more [s]-like sounds and greater ratings indicating more [ʃ]-like sounds (see Edwards et al., 2015 for further discussion of the VAS method).

Comprehensive documentation for the data set is available by calling:

```
help(SibilantFricatives, phoneigen)
```

We call attention to a handful of variables that will be used below when constructing the socio-auditory manifolds. Each production is uniquely identified by its Session and Trial within the session. The speaker of each production is identified by Participant; adult participants may be associated with two Sessions. The target word used to elicit the production is denoted by Orthography, and some target words were elicited multiple times in a given Session. The mean VAS ratings for the children's productions are found in the Rating variable; these ratings fall within $[0, 1]$, with lesser ratings indicating more [s]-like sounds and greater ratings indicating more [ʃ]-like sounds.

Finally, the ExcitationPattern variable in the data set is a list whose elements are 361-component numeric vectors. Each such vector denotes the values of an *excitation pattern*, whose values were derived by passing an input sound through a filter bank that comprised 361 fourth-order gammatone filters. The center frequencies of these filters were evenly spaced from 3 to 39 (i.e., 0.1 inter-channel separation) along the equivalent rectangular bandwidth (ERB) scale, a logarithmic transformation of the physical hertz scale that models the tonotopic mapping of the basilar membrane (see Moore, 2012 for a review). Each channel in this filter bank can be thought of as modeling the frequency tuning properties of a narrow cross-section of the basilar membrane, and the filter bank, as a whole, may be thought of as a sequence of such cross-sections spaced evenly along the length of the basilar membrane. An excitation pattern results from applying this filter bank to an input sound, summing the energy within the signal output by each channel, and associating those energy values to the center frequencies of the channels in the filter bank (see Reidy, 2015 for a detailed description of the excitation pattern construction process). In what follows, we take as fixed the values along the ERB scale on which an excitation pattern is defined, and thus represent each excitation pattern as the 361-dimensional vector of its values.

The SibilantFricatives() data set exemplifies how data should be structured when using the phoneigen package: the observations of high-dimensional data (here, excitation patterns) occur as a list-column of numeric vectors (here, the ExcitationPattern variable); metadata about the high-dimensional data (here, the talker, the target word, etc.) occur as columns of basic data types (here, Participant, Orthography, etc.).

2.2. A socio-auditory manifold for the adults' productions

As a 361-dimensional vector, each excitation pattern may be construed as a point in $\mathbb{R}^{361}$. In traditional phonetic analyses, the dimensionality of each excitation pattern would be reduced by mapping it through a small number of functions that were chosen *a priori*—e.g., functions that compute values of

statistical-distributional parameters (such as spectral moments) or that project onto a different basis (such as the discrete cosine transform). By contrast, the current method (based on manifold alignment and Laplacian Eigenmaps) learns such a mapping into a low-dimensional space (say $\mathbb{R}^1$ or $\mathbb{R}^2$) in a data-dependent manner by constructing a socio-auditory manifold, which involves the following five steps:

1. uniquely identify each excitation pattern as a node in the socio-auditory manifold;
2. define a symmetric, nonnegative function that takes two excitation patterns as inputs, and outputs the distance between them;
3. add edges between excitation patterns produced by a given speaker, in order to construct speaker-internal submanifolds, and then weight these edges according to the function defined in step (2);
4. add edges between excitation patterns produced by different speakers, in order to align or register the multiple speaker-internal submanifolds into a (connected) socio-auditory manifold, and then weight these edges;
5. compute the Laplacian eigenvectors of the constructed socio-auditory manifold.

2.2.1. Step 1: Identify the nodes in the manifold

For the SibilantFricatives() data set, the first step is accomplished automatically by virtue of the Session and Trial variables, which together identify each excitation pattern. In general, having one or more variables in a data set that jointly identify the high-dimensional data observations is necessary for the bookkeeping involved when constructing a socio-auditory manifold. In particular, to add edges between nodes it is necessary to take the Cartesian square over the set of nodes in the socio-auditory manifold, all while keeping track of multiple data and metadata variables associated with each node. The function CartesianSquare facilitates this bookkeeping. When calling CartesianSquare it is vital to include all variables that are necessary for identifying nodes and for adding and weighting edges between nodes in the socio-auditory manifold. For the adults'

data, these variables are Participant, Session, Trial, Orthography, and ExcitationPattern:

```
adults_node_pairs <-
  CartesianSquare(
    x = dplyr::filter(SibilantFricatives(), Adult),
    Participant, Session, Trial, Orthography, ExcitationPattern
  )
```

2.2.2. Step 2: Define a distance function over the nodes in the manifold

The second step requires us to choose a distance function that will be used to compare two excitation patterns. A great number of functions are viable candidates: the Euclidean and Manhattan distances may be familiar to the reader, and Deza & Deza (2009) presents an encyclopedic catalog of distance functions. Previous applications of manifold learning to speech data have had success with other functions (e.g., Jansen & Niyogi, 2006; Plummer et al., 2010); however, in this example, we choose to use an information-theoretic distance, the *Jeffrey divergence*, which is defined on two vectors x_i and x_j whose values have respectively been normalized so as to sum to 1:

$$\mathsf{Jeffrey}(x_i, x_j) = \sum_n (x_i[n] - x_j[n]) \cdot \log\left(\frac{x_i[n]}{x_j[n]}\right)$$

While different distance functions have proven useful on comparable high-dimensional representations of speech data (see differences between Jansen & Niyogi, 2006; Plummer et al., 2010; Reidy et al., 2017), we emphasize that the choice of distance function is not trivial. Practitioners are encouraged to choose a distance function that captures meaningful differences between observations, given how their data are represented. For example, the Jeffrey divergence would not be an appropriate notion of distance in applications where the speech data are represented as time series and where the researchers sought to map the realizations of those time series into a low-dimensional space. It is an ongoing research question to understand how the different distance functions behave on speech data, in the context of manifold learning; hence, our use of the Jeffrey

divergence here should not be taken as an assurance that this notion of distance will yield acceptable results in all instances, even when the speech data are represented as *static* excitation patterns or spectra as in the current example.

The Jeffrey divergence can be straightforwardly implemented as a generic function and associated methods, in order to facilitate vectorization over lists of excitation patterns:

```
Jeffrey <- function(i, j) {
  UseMethod("Jeffrey", i)
}

Jeffrey.numeric <- function(i, j) {
  i <- i/sum(i)
  j <- j/sum(j)
  sum((i-j) * log(i/j))
}

Jeffrey.list <- function(i, j) {
  purrr::map2_dbl(i, j, Jeffrey.numeric)
}
```

2.2.3. Step 3: Construct individual submanifolds

The third step involves the construction of speaker-internal submanifolds by adding and weighting edges between productions by the same talker. To simplify the demonstration, we will add edges in a programmatic fashion: all distinct productions by a given talker will be connected, yielding speaker-internal submanifolds that are complete graphs (this procedure is in contrast to the example in Appendix A, where edges are added between nodes based on the results of a nearest-neighbor search). The weight assigned to each edge will be based on the Jeffrey divergence between the two excitation patterns (read: nodes) connected by the edge. Given adjacent nodes i and j, with corresponding excitation patterns x_i and x_j, respectively, the weight of the edge connecting i and j is $w(i,j) = e^{-\mathsf{Jeffrey}(x_i,x_j)}$. By exponentiating the inverse of the Jeffrey divergence, a relatively large weight will be assigned to an edge between two nodes whose

corresponding excitation patterns have relatively small divergence (see left panel of Figure 4 below). Hence, the weighting function w may be thought of as a similarity function.

These weighted edges are added between pairs of nodes with a call to WeightEdgesIf. The edges argument takes an expression that evaluates to a *logical* vector within the data set passed to x; hence, this expression will often reference *metadata* variables within the data set x. In this example, the subexpression !(Session_i == Session_j & Trial_i == Trial_j) ensures that there are no loops (i.e., that no edges connect a node to itself); the other subexpression Participant_i == Participant_j adds an edge between all distinct productions by the same adult. The weights argument takes an expression that evaluates to a *numeric* vector within the data set x; hence, this expression should reference the two *data* variables. In the data set returned by WeightEdgesIf, the edge weights are stored in a variable named W_ij; only for pairs of nodes (i.e., on rows) where the edges expression evaluates to TRUE is the W_ij variable nonzero.

```
adults_submanifolds <-
  WeightEdgesIf(
    x = adults_node_pairs,
    edges = Participant_i == Participant_j &
      !(Session_i == Session_j & Trial_i == Trial_j),
    weights = exp(-Jeffrey(ExcitationPattern_i, ExcitationPattern_j))
  )
```

2.2.4. Step 4: Align the individual submanifolds

In the fourth step, edges are added between productions by different talkers in order to align the speaker-specific components into a connected socio-auditory manifold. As in the previous step, these between-speaker edges are added via a call to WeightEdgesIf. In this example, the submanifolds are aligned to each other according to the lexical information associated with the nodes: two excitation patterns are connected by an edge if the respective speakers are different, but the respective target words are the same. An appropriate expression for this alignment scheme is Participant_i == Participant_j & Orthography_i == Or-

thography_j. Note that this expression indeed adds no loops into the graph.

```
adults_manifold <-
  WeightEdgesIf(
    x = adults_submanifolds,
    edges = Participant_i != Participant_j &
      Orthography_i == Orthography_j,
    weights = exp(-Jeffrey(ExcitationPattern_i, ExcitationPattern_j))
  )
```

Note that in both calls to WeightEdgesIf the weights were assigned using the same formula. Consequently, all edges in the socio-auditory manifold could have been added with a single call to WeightEdgesIf. We nonetheless chose to add these edges in two steps in order to emphasize a conceptual difference in the work being done by the two types of edge: the speaker-internal edges *construct spaces* of speech productions; the between-speaker edges *align spaces*. This distinction is important to bear in mind since not all applications of manifold learning will warrant a common weighting scheme for the two types of edge. For example, in some applications, it is desirable to scale the weights on the space-constructing edges by a parameter $\mu \in [0, 1]$ and to scale the weights on the space-aligning edges by $1 - \mu$, in order to differentially adjust the relative importance of the internal structure of each space versus the correspondences between spaces. In other applications, the spaces to be aligned may not be commensurable (i.e., it may not be possible to define a coherent distance function on points from different spaces), in which case the weights applied to the space-aligning edges must be a constant or derived from some source other than the observations from the two spaces (see, Plummer, 2014; Wang & Mahadevan, 2009).

2.2.5. Step 5: Eigenmap the aligned manifold into a low-dimensional space

The final step asks us to map the aligned manifold into a low-dimensional space. In order to carry out this mapping by computing the Laplacian eigenvectors of the constructed manifold, it will first be necessary to reshape the edge-weight values in adults_manifold$W_ij into the adjacency matrix of the manifold. That is, the weights variable W_ij has the one-dimensional structure of

a vectorized matrix that needs to be recast into a two-dimensional array. This reshaping is accomplished by calling AdjacencyMatrix, which should be called with a comma-separated list of unquoted variable names whose values will be used to construct row- and column-names for the resulting matrix.

```
adults_adjacency <-
  AdjacencyMatrix(
    x = adults_manifold,
    Participant, Session, Trial, Orthography
  )
```

The subsequent computations for computing the degree matrix, the Laplacian matrix, and the Laplacian eigenvectors of the adults_adjacency matrix are encapsulated in the function LaplacianEigenmaps. The returned data table comprises thee variables: Eigenvector, a vector of eigenvector names e0, e1, e2, ...; Eigenvalue, a vector of the eigenvalues associated with the eigenvectors; and Projection, a list of data tables that associate each original data observation (identified by the row-names of the adjacency matrix) to its projected value on a given eigenvector.

```
adults_eigenmaps <-
  LaplacianEigenmaps(adults_adjacency)
```

The optimal n-dimensional embedding of the data is given by the Laplacian eigenvectors associated with the n least nonzero eigenvalues. By construction, the adults_manifold is a connected graph (i.e., there is a sequence of edges from any node to any other node); whence, it follows that only the first eigenvalue (i.e., the one associated with e0) is zero.[5] Consequently, the first eigenvector e0 is ignored, and only eigenvectors e1 and above should be considered when determining the low-dimensional space into which the data are eigenmapped.[6]

[5] In actuality, the first eigenvalue of the adults_adjacency matrix is not identically zero, as the reader may verify by calling adults_eigenmaps$Eigenvalue[1]. This discrepancy is due to quantization error; however, we note that the magnitude of the first eigenvalue (2.35×10^{-15}) is well less than a reasonable threshold of sqrt(.Machine$double.eps) ($= 1.49 \times 10^{-8}$).

[6] Note that LaplacianEigenmaps returns all Laplacian eigenvectors, not just those judged to

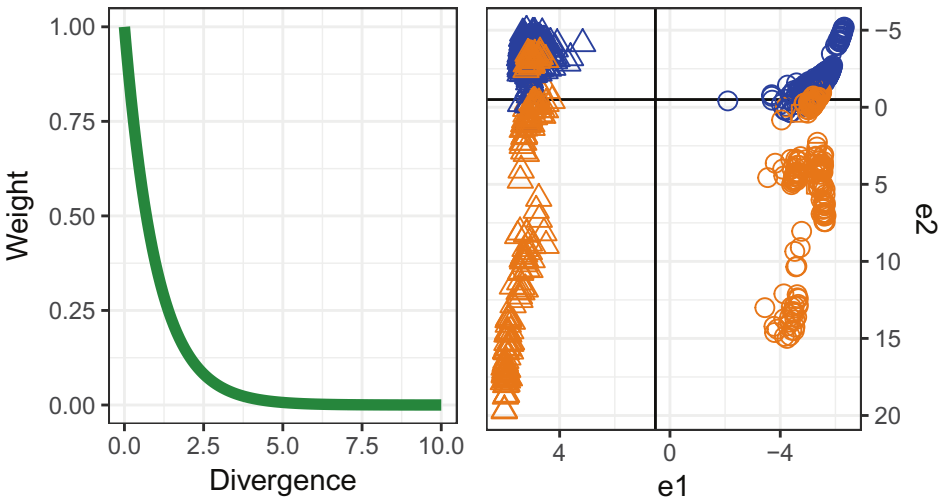


Figure 4: *Left panel*: The exponentially-decaying relationship between Jeffrey divergence and the edge weight derived from it. *Right panel*: The low-dimensional representation for the adults' productions derived by eigenmapping the adults_manifold. Target consonant is indicated by point shape: circles for /s/ and triangles for /ʃ/. Talker sex is indicated by color: blue for female and orange for male. Decision boundaries for logistic regression classifiers that predict target consonant and talker sex are indicated by black vertical and horizontal lines, respectively.

Once the eigenmapping projections have been computed, a low-dimensional representation of the data can be constructed with the function ReduceDimensions, which should be called with a comma-separated list of the unquoted Eigenvector names that determine the low-dimensional space:

```
adults_e1_e2 <-
  ReduceDimensions(
    x = adults_eigenmaps,
    e1, e2
  )
```

The right panel of Figure 4 below shows the distribution of the adults' sibilant fricative productions in the two-dimensional space defined by the eigenvectors adults_eigenmaps$e1 and adults_eigenmaps$e2. Visual inspection of this figure suggests that e1 encodes the linguistic place-of-articulation difference between /s/ and /ʃ/ and that e2 encodes indexical differences between men and women. Construction of the socio-auditory manifold did not explicitly leverage information about the target consonant or the talker sex associated with the sibilant fricative productions; rather, it leveraged information about the target *word* and the talker *identity*, which suggests that by eigenmapping the manifold into a lower-dimensional space, more abstract categories were learned (i.e., target consonant abstracts over the multiple target words in which the fricatives were produced; talker sex abstracts over the individual male and female talkers). Furthermore, because the learned eigenvectors do not encode the more granular categories of target word and target identity, the acoustic similarity information that was encoded as edge weights seems to be crucial to the learned abstractions.

---

be nonzero or greater than some threshold. Furthermore, depending on how a manifold is constructed, it is possible for more than one eigenvalue to be zero. Specifically, the number of zero-valued eigenvalues will be equal to the number of connected components in the manifold. Hence, in practice, it is important to inspect the eigenvalues output by LaplacianEigenmaps, and not to assume that all eigenvalues other than e0 should be used to determine the low-dimensional embedding of the data.

2.2.6. Comparison with spectral moments

Although spectral moments have been criticized for being ambiguously interpretable in terms of the underlying articulation (see Koenig et al., 2013) and for not capturing subtle spectral characteristics that may be perceptually salient (see Jannedy & Weirich, 2017), we nonetheless take them as an acceptable point of comparison with the low-dimensional representation derived from Laplacian eigenmapping, because of their sheer pervasiveness in the literature. For example, spectral moments have been used to represent sibilant fricatives in studies that have investigated *inter alia* place-of-articulation differences (e.g., Forrest et al., 1988; Jongman et al., 2000), sex-related differences (e.g., Fox & Nissen, 2005; Romeo et al., 2013), and developmental differences (e.g., Li, 2012; Nissen & Fox, 2005; Nittrouer et al., 1989).

The literature suggests that centroid and skewness are the two most reliable moments for differentiating /s/ and /ʃ/ (e.g., Forrest et al., 1988; Jongman et al., 2000). Centroid indexes the location of energy concentration along the frequency scale, which is expected to vary between /s/ and /ʃ/ due to differences in front cavity volume. Skewness tends to be (negatively) correlated with centroid, which is likely due to the finite support of spectral representations, imposed by the acoustic waveform being sampled at a finite rate. The left panel of Figure 5 shows the distribution of centroid and skewness values computed from the adults' sibilant fricative productions. In this plot, it is evident that the two underlying dimensions of variation learned from Laplacian eigenmapping (i.e., target consonant and talker sex) are simultaneously discernible in either of the spectral moments shown. For example, the centroid values could be (very) roughly divided into four intervals: men's /ʃ/ < women's /ʃ/ < men's /s/ < women's /s/. This situation is an example of a commonly encountered problem with purely physical representations of speech data: the underlying dimensions of variation that are of principal interest do not differentially map onto orthogonal physical attributes of speech.

A traditional solution to this problem is to batch normalize physical feature

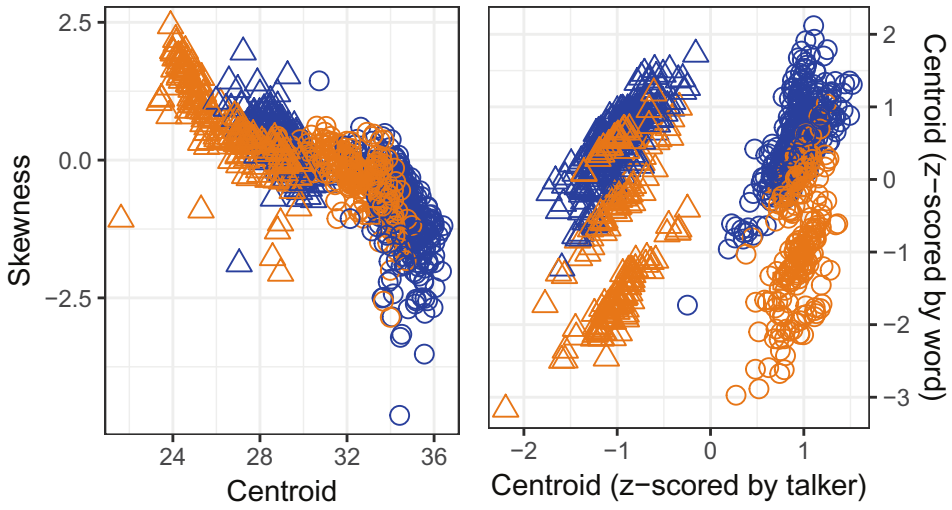


Figure 5: *Left panel*: The distribution of raw centroid and skewness values computed from the adults' productions. Target consonant is indicated by point shape: circles for /s/ and triangles for /ʃ/. Talker sex is indicated by color: blue for female and orange for male. *Right panel*: The distribution of centroid values computed from the adults' productions, after these values have been normalized within talker or within word. Point shape and color indicate target consonant and talker sex, respectively.

values in order to marginalize out some underlying dimension of variation. For example, data may be normalized within the levels of an indexical variable in order to marginalize out the underlying indexical variation, and likewise for a linguistic variable. While this traditional method of batch normalization has been, for the purposes of model fitting, superseded by more contemporary and appropriate approaches, such as fitting a mixed-effects model with the indexical and/or the linguistic variable as random-effects grouping factors, the ability to visualize the batch normalized data is an instructive intuition-building exercise for helping to understand the role played by the edge-structure in the socio-auditory manifold constructed above. The right panel of Figure 5 demonstrates two examples of batch normalization: when z-scored by talker (x-axis), the residual variation indicates the linguistic place-of-articulation dimension; when z-scored by word (y-axis), the residual variation indicates the indexical dimension of talker sex. The similarity between the right panels of Figures 4 and 5 suggest that the socio-auditory manifold pools data points within talker and target word, the two variables used to add edges to the graph structure.

To quantitatively compare the learned eigenvectors e1 and e2 with centroid frequency as a predictor of either target consonant or talker sex, we fit a series of logistic regression classifiers. For the eigenvectors, classical (fixed-effects-only) regression models were fit. For centroid frequency, mixed-effects logistic regression models were fit with either Participant (when predicting target consonant) or Orthography (when predicting talker sex) as the random-effects grouping factor. Target consonant category was predicted almost perfectly by either method: 100% accuracy with the Laplacian eigenvector e1, and 99.792% accuracy with a mixed-effects classifier with fixed and random effects of centroid, grouped within Participant. Likewise, talker sex was predicted with high accuracy by both methods, but again Laplacian eigenmapping yielded greater accuracy: 90.938% accuracy with e2, and 86.354% accuracy with a mixed-effects classifier comprising fixed and random effects of centroid, grouped within Orthography.

2.2.7. Interim summary and discussion

This example demonstrated how to implement the computations needed to construct a socio-auditory manifold from high-dimensional speech data and then eigenmap that manifold into a low-dimensional space—a procedure borrowed from the field of manifold learning. The specific edge-structure used in the manifold for the adults' productions was found to function similarly to post-hoc batch normalization within linguistic (target word) and indexical (talker identity) variables. Compared with normalized centroid values, the Laplacian eigenvectors better predicted target consonant and talker sex, although these improvements were very modest.

An instructive interpretation of manifold learning, in this example, is that eigenmapping projects a gestalt representation of speech (e.g., an excitation pattern) into a space whose dimensions are interpretable as the underlying, *independent* linguistic and indexical dimensions of acoustic-phonetic variation for sibilant fricatives. In this way, the low-dimensional values output by eigenmapping should not necessarily be construed as "acoustic features" in the traditional meaning of the term: The image of the speech observations under eigenmapping represent the distribution of these observations across the underlying dimensions of meaningful variation (e.g., linguistic and indexical dimensions). By contrast, such dimensions of variation are often (over)loaded onto individual acoustic features—construed as variables computed solely from individual speech observations—as was seen in the raw centroid values.

2.3. A socio-auditory manifold for representing children's productions

The preceding example focused on enumerating the computational steps that together construct a socio-auditory manifold and eigenmap it to a low-dimensional space. The current example, by contrast, focuses squarely on the third and fourth steps in this process—the addition of edges to the manifold—in order to explicate how the structure of the socio-auditory manifold affects the low-dimensional representation learned from eigenmapping.

For this demonstration, the goal will be to learn for the children's productions a one-dimensional representation that predicts the associated VAS ratings. Because these VAS ratings indicate adult perceptual judgments along a *one-dimensional* continuum and because the linguistic place-contrast between /s/ and /ʃ/ is indicated by the *first* Laplacian eigenvector of the manifold of the adults' productions, the manifold for deriving a feature for the children's productions will have the following graph structure: within a given talker, all distinct productions will be connected to each other; between any two adults, productions of the same target word will be connected; between a child and an adult, productions of the same target word will be connected; between any two children, no productions will be connected. This particular construction is motivated by the facts that it represents the structure of each intratalker production space and that it puts each child's production space in correspondence with the community-norm set by the adults' production spaces.[7]

The listing below shows how the community_norm_manifold can be implemented:[8]

```
community_norm_manifold <-
  SibilantFricatives() %>%
  CartesianSquare(
    Participant, Adult, Session, Trial, Orthography,
    ExcitationPattern
```

[7]This construction assumes that the adult community-norm production categories are similar in structure to the community-norm perception categories, and that the perceptual processes at play in the VAS task involve something like comparing the presented stimuli to these community-norm categories. By omitting child-to-child connections in the community_norm_manifold, we do not intend to assume that the phonological and speech categories of children develop without interaction with their peers. Rather, this omission only assumes that the information in the child-to-child alignments may not be helpful in learning a representation that predicts *adults'* perceptions of children's productions.

[8]In this and subsequent listings, the calls to functions from the phoneigen package have been chained together using the pipe operator %>% from the magrittr package. This operator takes the value of the expression on its left-hand side and passes it as the first unnamed argument of the expression on its right-hand side.

```
  ) %>%
  WeightEdgesIf(
    edges = Participant_i == Participant_j &
      !(Session_i == Session_j & Trial_i == Trial_j),
    weights = exp(-Jeffrey(ExcitationPattern_i, ExcitationPattern_j))
  ) %>%
  WeightEdgesIf(
    edges = Participant_i != Participant_j &
      Orthography_i == Orthography_j &
      (Adult_i | Adult_j),
    weights = exp(-Jeffrey(ExcitationPattern_i, ExcitationPattern_j))
  )
```

A one-dimensional representation of the children's (and the adult's) productions can then be learned by eigenmapping the community_norm_manifold:

```
community_norm_e1 <-
  community_norm_manifold %>%
  AdjacencyMatrix(
    Participant, Session, Trial, Orthography
  ) %>%
  LaplacianEigenmaps() %>%
  ReduceDimensions(e1)
```

Figure 6 makes the case that the Laplacian eigenvector e1 denotes a linguistic place-continuum learned from the community_norm_manifold. The left panel shows the distribution of e1 values stratified by the transcription and target categories. The top and bottom strata correspond to the adults' productions of /s/ and /ʃ/, respectively, and the intermediate strata correspond to the children's productions. The mean of the values in each stratum is indicated by an oversized point. The most extreme mean values are found in the adults' strata, suggesting a continuum anchored by the adults' productions as endpoints. The children's productions tend to fall between these endpoints, suggesting that the values of e1 represent the incipient consonant contrast developing in these children. Furthermore, within each target consonant category (i.e., blue points for target /s/ and orange points for target /ʃ/), the mean values of each stratum form an orderly scale according to transcription category: [ʃ] → [ʃ]:[s] → [s]:[ʃ]

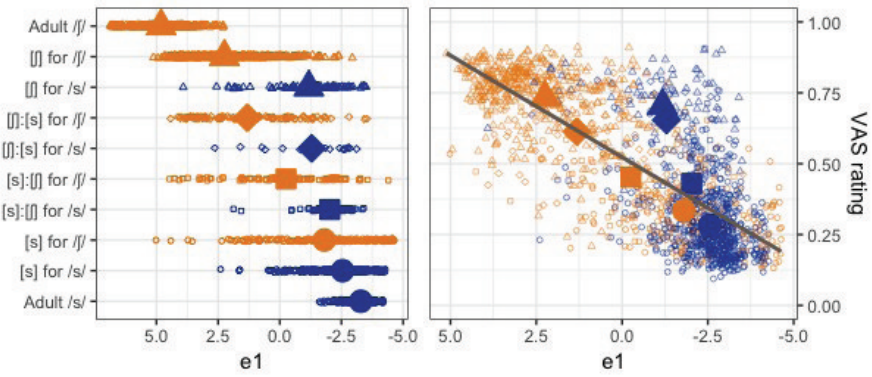


Figure 6: *Left panel*: The distribution of Laplacian eigenvector e1 of the community_norm_manifold across transcription categories. Target consonant is indicated by point color: blue for /s/ and orange for /ʃ/. Transcription category is indicated by point shape: circles for [s], squares for [s]:[ʃ], diamonds for [ʃ]:[s], and triangles for [ʃ]. The large filled points denote the mean of each subset of points. *Right panel*: The relationship between VAS rating of each item with the values of the Laplacian eigenvector e1 of the community_norm_manifold.

→ [s].[9] The right panel shows a measure of adults' perceptual judgments that is more continuous than transcription categories, plotting the VAS ratings against the values of Laplacian eigenvector e1.

2.3.1. Comparison with a different manifold structure

In order to demonstrate how the structure of the socio-auditory manifold affects the eigenmapped representation of the speech observations, we present a manifold structure that is conceptually similar to the community_norm_manifold but that yields an eigenmapping that is starkly different from the one learned above. Specifically, we will construct a manifold that is similar to the commu-

[9] One curiosity about the distribution of e1 values is that values for the children's productions of target /s/ span a much narrower range than the values for their productions of target /ʃ/. It is not clear whether this difference in range reflects a greater sensitivity by transcribers to misarticulations of target /s/ as opposed to target /ʃ/, or whether it is an artifact of the graph structure (e.g., that the excitation patterns for children's productions of target /ʃ/ words were adjacent to a greater number of adults' productions in the community_norm_manifold, as compared with those for children's productions of target /s/ words).

nity_norm_manifold except that it omits all edges that connect different productions by a given child talker. Consequently, in this manifold, the neighborhood of each node associated with a production of some word by some child comprises all and only nodes associated with a production of that same word by some adult. The following code constructs this word_neighborhoods_manifold and then eigenmaps it into a one-dimensional space:

```
word_neighborhoods_manifold <-
  SibilantFricatives() %>%
  CartesianSquare(
    Participant, Adult, Session, Trial, Orthography,
    ExcitationPattern
  ) %>%
  WeightEdgesIf(
    edges = Participant_i == Participant_j & Adult_i & Adult_j &
      !(Session_i == Session_j & Trial_i == Trial_j),
    weights = exp(-Jeffrey(ExcitationPattern_i, ExcitationPattern_j))
  ) %>%
  WeightEdgesIf(
    edges = Participant_i != Participant_j &
      Orthography_i == Orthography_j &
      (Adult_i | Adult_j),
    weights = exp(-Jeffrey(ExcitationPattern_i, ExcitationPattern_j))
  )

word_neighborhoods_e1 <-
  word_neighborhoods_manifold %>%
  AdjacencyMatrix(
    Participant, Session, Trial, Orthography
  ) %>%
  LaplacianEigenmaps() %>%
  ReduceDimensions(e1)
```

Figure 7 demonstrates the importance of the within-child edges for learning a low-dimensional embedding that represents a linguistic place-of-articulation continuum. Without these edges, the observations are pooled into regions defined by the acoustic properties of the adults' productions of a given word. The

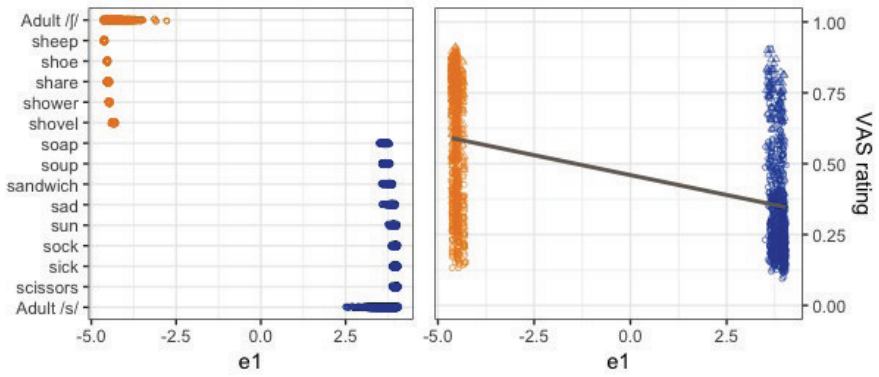


Figure 7: *Left panel*: The distribution of Laplacian eigenvector e1 of the word_neighborhoods_manifold across target words. Target consonant is indicated by point color: blue for /s/ and orange for /ʃ/. *Right panel*: The relationship between VAS rating of each item with the values of the Laplacian eigenvector e1 of the word_neighborhoods_manifold.

reason for the disparity between the eigenmapped embeddings of the community_norm_manifold and the word_neighborhoods_manifold is that the values in the manifolds' respective adjacency matrices act as penalties on the distance between points in the eigenmapped embedding. Hence, if two adjacent nodes in a manifold have a high edge weight (due to low divergence between their excitation patterns), then an embedding will incur a large penalty if it maps these two nodes far from each other in the low-dimensional space. Conversely, if two nodes are not adjacent, then their corresponding entries in the adjacency matrix will be zero, and an embedding will incur no penalty for mapping these nodes to points that are far from each other.

Consequently, because a given production by a child is adjacent only to productions by adults in the word_neighborhoods_manifold, the eigenmapped embedding will map the children's productions near to the adults' productions of the same words in order to avoid any large penalties, without any regard as to where a given child's productions are mapped in relation to each other. By contrast, the neighborhood of a given production by a child in the community_norm_manifold comprises productions from adults and productions from that child; hence, the eigenmapped embedding of a given observation results

from a competition between being near to productions by the same talker and being near to adults' productions of the same word—with this competition being settled by the acoustic similarity between all of these adjacent productions as encoded in the edge weights of the manifold.

2.3.2. Comparison with centroid frequency

To compare the one-dimensional eigenmapped representation of the children's productions from the community_norm_manifold with an acoustic feature that is commonly used to characterize children's productions of sibilant fricatives, we computed the centroid frequency of the excitation patterns. Figure 8 plots the distribution of (unnormalized) centroid values across the transcription categories (left panel), as well as the relationship between the VAS ratings and these centroid values (right panel). As with the Laplacian eigenvector learned from the community_norm_manifold, the centroid values stratified by the transcription and target category form an orderly scale from [ʃ] to [s]. As a quantitative comparison of the community_norm_e1 Laplacian eigenvector with centroid, the Pearson product-moment correlation coefficient was computed between each variable and the VAS ratings: community_norm_e1 was found to have a stronger relationship with VAS ratings ($|r| = 0.757$) than centroid does ($|r| = 0.685$).

2.3.3. Concluding remarks

The preceding examples have demonstrated how to construct and eigenmap socio-auditory manifolds, using functions from the phoneigen package, in order to derive low-dimensional phonetic representations of both adults' and children's speech. In both of these examples, the eigenmapping yielded representations that were sensible and, by some measure, improved upon a conventional representation for sibilant fricatives—spectral moments. While such results do indicate the utility of the proposed method, we close by emphasizing that we recognize that the socio-auditory manifold and eigenmapping approach may not be suitable for all instances of speech research. Given the flexibility available to the researcher in constructing a socio-auditory manifold (e.g., the choice of

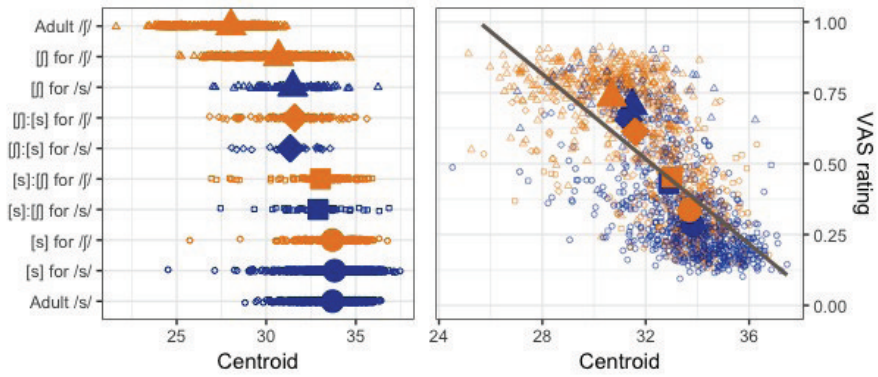


Figure 8: *Left panel*: The distribution of centroid values computed from the children's productions, across transcription categories. Target consonant is indicated by point color: blue for /s/ and orange for /ʃ/. Transcription category is indicated by point shape: circles for [s], squares for [s]:[ʃ], diamonds for [ʃ]:[s], and triangles for [ʃ]. The large filled points denote the mean of each subset of points. *Right panel*: The relationship between VAS rating of each item with the centroid frequency values.

metadata to be used to add edges to the manifold, the choice of differentially weighting edges), there is still much applied research to be done in discovering applications of this method to new problems in speech science and phonetics. However, we ultimately view this flexibility as a strength of the method since it gives the researcher the opportunity to bring her content-knowledge of the many facets (i.e., articulatory, acoustic, sociolinguistic) of speech to bear when constructing a socio-auditory manifold that is best suited to the problem at hand.

3. Broader impact of the analytic framework

We close this paper with a brief discussion of the potential impact of our framework beyond the analytic applications covered in Section 2. The discussion stems from two key issues suggested by the analyses. The first issue is that representations of speech data computed from methods such as LPC, spectral moments, and the DCT (see Section 1) are physicalist in nature, i.e., they do not directly correspond to cognitive representations of language-specific speech

sound categories, nor do they model the socio-cultural information that is part and parcel of speech. Moreover, socio-phonetic models of speech perception that rely on these analyses of speech data are perforce built (at least, in part) on top of physicalist representations that reflect neither the cognitive mechanisms that speakers use in listening to speech nor the social systems within which hearers interpret speech. The second issue stems from demonstrations that methods such as LPC have difficulties representing child speech (see Story & Bunton, 2016, inter alia). The core of the issue is illustrated by the fact that LPC is based on an adult model of speech production. Yet, speech and language are developmental phenomena that take shape as individuals progress from infants into adult members of an ambient speech community. Thus models of speech that underlie the computational techniques that yield representations used in phonetic analyses should reflect the social, physiological, and other developmental processes individual undergo. Below, we address both of these issues in relation to the socio-indexical assumption and the manifold assumption. We conclude the discussion by situating our framework within the larger setting of representation learning.

We begin our discussion of the first issue with respect to the socio-indexical assumption. Indeed, a growing number of studies suggest that speech and language models may benefit from data representations that can more directly reflect the interaction between phonetic categories and social categories (e.g., Strand & Johnson, 1996; Johnson et al., 1999; Hay et al., 2006; Warren et al., 2007; Warren, 2017). For example, Strand & Johnson's (1996) study of the integration of audio-visual information in the perception of tokens along an /s/–/ʃ/ showed that listeners' categorical judgments of acoustically identical fricatives changed based on whether they are paired with a male or female face articulating the tokens. More recently, Warren's (2017) New Zealand English study investigated the interaction between socio-indexical and phonetic information in listeners' parsing of the speech signal. The results showed that listeners' perception of a word containing a vowel realized with the near-square merger (associated with younger speakers) early in an utterance affected their sensitiv-

ity to the perception of a rising intonation at the end of the utterance as uptalk (also associated with younger speakers), as well as the process by which listeners' differentiate between uptalk and question intonation. This work not only supports the position that socio-indexical information factors into computing phonetic categorical information from the speech signal, but that socio-indexical information also impacts interpretation of the signal across simultaneous levels of speech (see Ladd, 2012).

The need for data representations that reflect the interaction between language-specific phonetic categories and culture-specific social categories comes into sharper relief when viewed from a developmental vantage point. In particular, there is a growing collection of evidence that the interaction between the two types of category is present throughout the language acquisition process. At the latter end of the process, Drager's (2011) study of the productions of "like" phrases by New Zealand high school girls between 16–18 years of age shows that the phonetic realization of /k/ differed across the girls' social groups in relation to the grammatical function of the word in their utterances. These results suggest that older children and adolescents learn to express gender identities and other production-based indications of group membership, which may, in turn, engender socially-motivated variation in the speech signal. Further support comes from Li et al.'s (2016) study of Canadian English-speaking children between 4–16 years of age wherein boys' gender identity was shown to affect the acoustic quality of their productions of /s/. Moreover, the earliest stages of phonological and lexical acquisition are known to proceed in lockstep with the dynamics of age-differentiated social engagement between children and their caretakers. A striking component of this complex relationship is observed in infants transitioning from the dyadic mutual attention between themselves and their caretakers during the initial stages of phonetic category learning to the triadic joint attention between infants, caretakers, and other worldly objects that is a prerequisite for early word learning (see Vihman, 2014 for a review). This set of results as a whole suggests that language acquisition involves the formation and emulation of models of socially salient speakers that children engage

with during the acquisition process, as well as formation of mutable models of the different types of social engagement.

Results of this nature strongly suggest that representations of speech data that more directly encode the social components of speech and language described above may provide the basis for more fruitful analyses than those based solely on representations computed from a strictly physicalist point of view. Moreover, they support the design of representation computations that take into account the social developmental nature of speech and language phenomena. Of course, social development takes place alongside physiological development, and it is in this connection that we revisit the manifold hypothesis, within the context of its role in models of speech articulation over the last few decades.

Given the constraints of the physical systems involved, articulatory modeling is a natural setting for the use of manifolds, especially in deriving low-dimensional representations of speech that reflect primary degrees of freedom of the articulatory system (see, e.g., the "uncontrolled manifold" method Scholz & Schöner, 1999; Saltzman et al., 2006; Saltzman & Holt, 2014; Szabados & Perrier, 2016). A key example is Maeda's (1991) use of a statistical "linear manifold" method (factor analysis, in this case) to map high-dimensional midsagittal vocal tract representations to a lower-dimensional parameter space whose degrees of freedom correspond to the "articulatory blocks" characterized by Lindblom & Sundberg (1971). These lower-dimensional representations facilitate user control of the articulatory synthesizer built over the midsagittal vocal tract data while also providing linguistically meaningful interpretations of the vocal tract shapes generated by the synthesizer.

Maeda's articulatory model has since been updated to model vocal tract sizes and shapes (and the corresponding acoustic output) ranging from those of young infants to those of mature adults using a fixed mapping method that scales an adult vocal tract model (Boë & Maeda, 1998). While potentially useful for developmental modeling and analysis (see, e.g., Oohashi et al., 2017; Rvachew et al., 2006), it is crucial to recognize that aspects of adult-based articulatory

models, e.g., control strategies and articulator dynamics, fail to extend to the speech of young children and the vocalizations of infants under scaling (see de Boer & Fitch, 2010, and Plummer, TBD for broader discussion)[10]. While the analysis in Section 2 constructed manifolds from auditory data points it is of course possible to use articulatory data instead to construct "socio-articulatory" manifolds. The use of articulatory data in our framework offers a potentially useful step away from fixed scaling, allowing the mappings that relate adult articulatory models to child articulatory models to be learned via manifold alignment.

More generally, the data points that manifolds are constructed over need not come from a single sensory space. To illustrate, it is possible to construct one graph over auditory representations derived from productions of a single speaker as well as a second graph over corresponding articulatory representations of those productions and then align the two graphs. The representations derived from the aligned graph's Laplacian will reflect the manifold structures of the data points from each respective sensory space rather than that of any one of them.

From a developmental perspective, manifold alignment both within and across different sensory spaces is a promising approach to modeling the physiological and socio-cultural developmental complexities involved in phonological acquisition. The alignment computation provides the basis for modeling how infants (i) form relations across sensory spaces as their early vocal experiences begin to shape their emerging vowel spaces, (ii) form abstract relations that reconcile the radical differences between the sizes and shapes of their own vocal tract and those of their caretakers, and (iii) organize their culture-specific socio-vocal interactions with caretakers that further shape infant vowel space development. A number of computational models of acquisition have been put

[10]Note that this statement applies to other models of articulation, e.g., Rubin et al.'s (1981) geometric model of articulation, incorporated within the TADA model, which can be configured to model different vocal tract sizes and shapes; see Iskarous et al., 2003).

forward that attempt to shed light on these issues (see Guenther, 1995; Oudeyer, 2005; Howard & Messum, 2011; Rasilo et al., 2013; Warlaumont et al., 2013; Messum & Howard, 2015, inter alia). However, they model relations between auditory and articulatory representations in terms of direct mappings between them, making it difficult to incorporate the influence of the dynamics of both the infant's physiological growth and their social interactions with caretakers on the emergence of phonetic categories over the course of development, especially across cultures. In contrast, our framework provides room for modeling the infant's cognitive representation of not just modality-specific articulation and audition representations, but also constructed "intermodal" representations, via manifold alignment. These separate intermodal representations leave the modality-specific representations intact to support the learning of other mappings.

In light of the aforementioned efforts to carefully model the myriad layered aspects of language learning, we conclude this discussion by situating our framework with respect to recent advances in deep learning. The set of graph-based manifold methods (Laplacian Eigenmaps included) that can be leveraged within our framework for computing low-dimensional representations of speech data falls within the domain of representation learning (see Bengio et al., 2013, for a review). Representation learning is a subfield of machine learning concerned with building representations of data that facilitate other tasks such as classification. Over the last decade, deep learning has emerged as a subfield of representation learning wherein architectures built using a variety of neural net models and probabilistic graphical models learn hierarchies of features that (attempt to) enhance or even negate the use of domain-specific features engineered by researchers (see Hinton et al., 2012, for a review and perspective on deep learning in automatic speech recognition). Although manifold methods can be incorporated within these deep learning architectures, the methods themselves are contraposed as "shallow learning" methods that are focused more on data exploration and interpretation.

A full discussion of the contrast between deep and shallow learning methods

is beyond the scope of this paper, though we close with one key consequence. While the recent technological advances yielded by deep learning architectures are quite impressive, insight into learning problems may be obscured by the generic hierarchical computations these architectures carry out. To illustrate, the layer-wise "abstraction" carried out by deep neural net architectures purports to generically eliminate variation in raw sensory input as stable categorical representations are computed. Yet, socio-phonetic analyses are typically based on specific kinds of variation being preserved in representations that are output from normalization computations. At present, it is unclear how to relate the kinds of representations computed by deep learning architectures with those that facilitate socio-phonetic analyses. In contradistinction, smaller-scale analyses and modeling efforts aimed at carefully parsing and analyzing variation in speech data remain fundamental to the conceptual advances and technical formulations that are driving the science forward. The manifold-based framework we have proposed herein is broad enough in principle to model a multitude of speech data phenomena while maintaining the interpretability of model components.

Acknowledgements

We thank Jan Edwards, Mary Beckman, and Benjamin Munson for providing the data from the Learning 2 Talk Project, which were used for the demonstrations in Section 2. We also thank the too-many-to-name research assistants who collected and annotated these data, as well as the families who donated their time to participate in the study.

Appendix A. Manifold computations

Appendix A.1. Laplacian Eigenmaps

In this section we briefly review three concepts and computations used in the sequence of computations specified in Section 2: weight matrices, the Laplacian, and the Laplacian Eigenmaps method. To facilitate the review, we focus only the

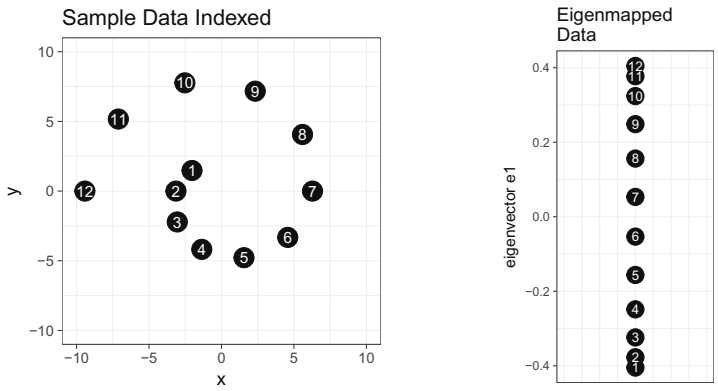
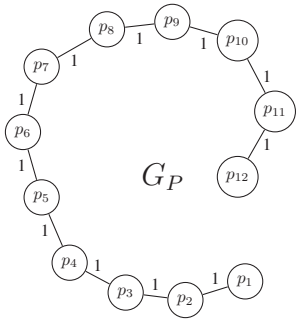


Figure A.9: Data eigenmapped to lower-dimensional representations.

data set P, shown in Figure A.9 (left). Moreover, we assume the graph structure G_P over the data shown in Figure A.10 (left) has already been constructed, and as a result of the construction each data point P_i corresponds to a node p_i, and each node p_j is connected by an edge to node p_{j-1} and p_{j+1} for $2 \leq j \leq 11$. The graph construction procedure is covered in the following section.

The *weight matrix* of a graph simultaneously encodes the set of nodes a graph contains, which nodes are connected to each other by an edge, as well as the weights that are assigned to those edges. Figure A.10 (right) shows a weight matrix W_P that encodes the nodes of G_P as its row and column indices, and the edges of G_P and their weights as entries. For example, the entry at row p_{11} and column p_4, denoted $W_P(p_{11}, p_4)$, is a 0, which indicates that no edge exists between these two nodes, while the entry $W_P(p_4, p_3)$ is a 1. This nonzero entry indicates both that an edge exists between p_4 and p_3 and that its weight is 1.

Given the weight matrix of a graph, we can construct its *Laplacian* representation. To illustrate the process, we use the weight matrix W_P representing our graph G_P. First, we construct a matrix D_P with the same number of rows and columns as W_P. The diagonal entry of D_P for row i is the sum of all the entries in row p_i in W_P, with all non-diagonal entries set to 0. We then subtract W_P from D_P. That is, the graph Laplacian of G_P, denoted by L_P, is computed as $L_P = D_P - W_P$. The graph Laplacian L_P for the graph G_P is shown in



W_P

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}	p_{12}
p_1	0	1	0	0	0	0	0	0	0	0	0	0
p_2	1	0	1	0	0	0	0	0	0	0	0	0
p_3	0	1	0	1	0	0	0	0	0	0	0	0
p_4	0	0	1	0	1	0	0	0	0	0	0	0
p_5	0	0	0	1	0	1	0	0	0	0	0	0
p_6	0	0	0	0	1	0	1	0	0	0	0	0
p_7	0	0	0	0	0	1	0	1	0	0	0	0
p_8	0	0	0	0	0	0	1	0	1	0	0	0
p_9	0	0	0	0	0	0	0	1	0	1	0	0
p_{10}	0	0	0	0	0	0	0	0	1	0	1	0
p_{11}	0	0	0	0	0	0	0	0	0	1	0	1
p_{12}	0	0	0	0	0	0	0	0	0	0	1	0

Figure A.10: The graph G_P (left) and its weight matrix W_P (right).

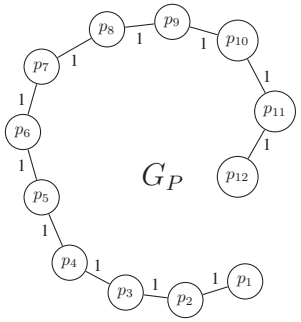
Figure A.11 (indices are nodes as in weight matrices).

In general, let L_G be the Laplacian of a graph G with n-many nodes uniquely corresponding to data points. If e is a nonzero vector in $\mathbb{R}^n$ such that $L_G e = \lambda e$ for some scalar value λ, then λ is an *eigenvalue* of L_G with corresponding *eigenvector* e.[11] The collection of eigenvalues of L_G is called the *spectrum* of L_G, which is denoted $\text{spec}(L_G)$. It is important to note that $\text{spec}(L_G)$ for a graph with n nodes will always contain n eigenvalues; however, some eigenvalues may be repeated in $\text{spec}(L_G)$ while still corresponding to distinct eigenvectors. Moreover, for every graph G, 0 is in $\text{spec}(L_G)$ and each eigenvalue λ in $\text{spec}(L_G)$ is a real number with $\lambda \geq 0$. This fact makes it convenient to order the eigenvalues in $\text{spec}(L_G)$ as follows: $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1}$.

Regarding the eigenvectors corresponding to the eigenvalues in $\text{spec}(L_G)$, those corresponding to the same eigenvalue can be chosen to be orthogonal, and those corresponding to different eigenvalues are orthogonal. Therefore, we are guaranteed to have a set of n orthogonal[12] eigenvectors corresponding to the eigenvalues of L_G. Moreover, we can scale each eigenvector without impacting orthogonality, thus we assume that the eigenvectors have unit length.

[11] Eigenvectors of real symmetric matrices which may always be chosen to contain real values.

[12] Two vectors u and v are *orthogonal* if $\sum_i u(i) \cdot v(i) = 0$.



L_P

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}	p_{12}
p_1	1	-1	0	0	0	0	0	0	0	0	0	0
p_2	-1	2	-1	0	0	0	0	0	0	0	0	0
p_3	0	-1	2	-1	0	0	0	0	0	0	0	0
p_4	0	0	-1	2	-1	0	0	0	0	0	0	0
p_5	0	0	0	-1	2	-1	0	0	0	0	0	0
p_6	0	0	0	0	-1	2	-1	0	0	0	0	0
p_7	0	0	0	0	0	-1	2	-1	0	0	0	0
p_8	0	0	0	0	0	0	-1	2	-1	0	0	0
p_9	0	0	0	0	0	0	0	-1	2	-1	0	0
p_{10}	0	0	0	0	0	0	0	0	-1	2	-1	0
p_{11}	0	0	0	0	0	0	0	0	0	-1	2	-1
p_{12}	0	0	0	0	0	0	0	0	0	0	-1	1

Figure A.11: The graph G_P (left) and its Laplacian matrix L_P (right).

For each graph G we thus have $\text{spec}(L_G) = \{\lambda_0, \lambda_1, \ldots, \lambda_{n-1}\}$, and we have corresponding eigenvectors $e_0, e_1, \ldots, e_{n-1}$ that are pairwise orthogonal and of unit length. Crucially, each of the eigenvectors e_i can be viewed as a function from the nodes of G, and hence from the data used to construct G, to the set of real numbers. These functions, called *eigenmappings*, provide the means to map data to a lower-dimensional analysis space that reflects the structure encoded by G.

Returning to our example data set P, using the eigenvectors of the Laplacian L_P, we can map our data points to a one-dimensional space that reflects the underlying manifold structure. We ignore the eigenvector corresponding to the eigenvalue 0 (since it is a scaled version of the vector $\mathbf{1}$ and hence uninformative), and use the eigenvector e_1 corresponding to the smallest nonzero eigenvalue λ_1. In this case, each data point P_i maps to the ith component of e_1, denoted $e_1(i)$, and the mapping is shown in Figure A.9 (right). Note that the representations yielded by the eigenmapping reflect the intrinsic one-dimensional nature of the data set P. Note also that the computation described above yields the representations in Figure 1 (right).

Appendix A.2. Manifold alignment using Laplacian Eigenmaps

This section is focused on explicating the sequence of computational steps used in the analysis in Section 2, hence we keep to a small example where data

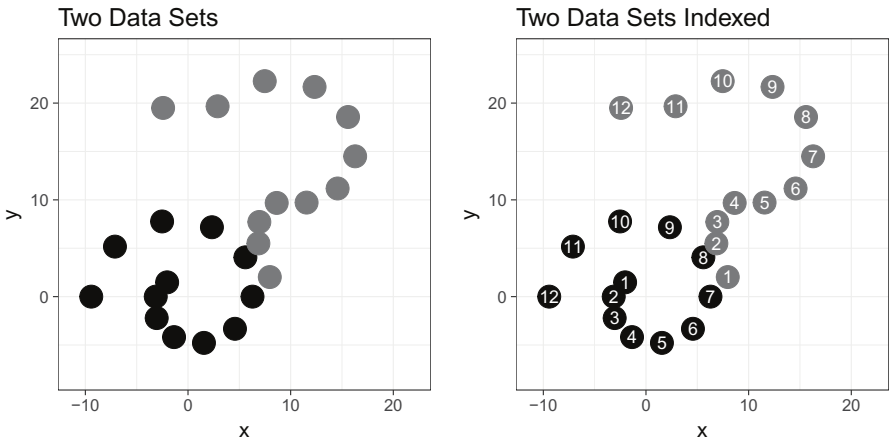


Figure A.12: The two data sets P (black) and Q (gray).

and corresponding graphs are visualizable. We take the two data sets P and Q in Figure 2 (shown again in Figure A.12) from Section 1 as our point of departure. We assume that the points in P lie on a spiral, and suppose we assume that the points in Q lie on a similar manifold whose middle section is denser and whose ends curl outward instead of inward. Moreover, we suppose that the measurements that have yielded the points in Q render larger x and y values than the measurements that yielded the points in P. Suppose we have additional information that point P_1 and point Q_1 are similar to each other in that they are close to the ends of their respective manifolds. The same holds for points P_{12} and Q_{12}. Moreover, the point P_6 and point Q_6 are similar to each other in that they are "in the middle" of their respective manifolds. Given this similarity between data points across the two manifolds, our goal is to construct a mapping where the learned representation of P_i is close to that of Q_i in a learned analysis space.

In order to construct such a mapping, we use manifold alignment and the Laplacian Eigenmaps technique as follows:

1. uniquely identify each data point as a node in the graph;
2. specify a symmetric, nonnegative function that takes two data points as

inputs, and outputs the distance between them;

3. add edges between data points in a given data set, and then weight these edges according to the function defined in step (2);
4. add edges between data points from different data sets, in order to align the separate graphs into a single (connected) graph, and then weight these edges;
5. compute the Laplacian eigenvectors of the single constructed graph.

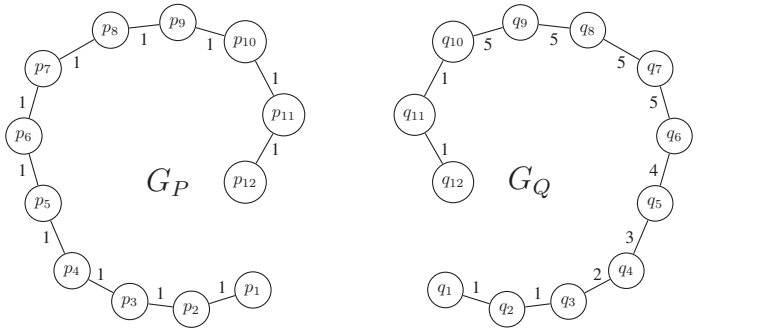
This procedure mirrors that specified in Section 2.

Step (1): We simply map each data point P_i to a node p_i and each data point Q_i to a node q_i.

Step (2): We use the Euclidean metric as our distance computation between data points.

Step (3): Beginning with the data set P, we add edges using the k-nearest neighbors method: for each data point P_i, we compute its k-nearest neighbors using the Euclidean metric. Each of these neighbors has a corresponding node in the graph, and we add an edge connecting these nodes to the node p_i. To keep the example simple, we set $k = 1$. Moreover, we use a constant weight function that maps each edge to the value 1. The resulting graph, which we denote by G_P, is shown in Figure A.13 (top, left). To complete step (3), we proceed in kind for the data in Q, however, to reflect the nature of manifold Q is assumed to lie on, the weight function assigns higher weights to edges in the middle of the graph, G_Q, which is shown in Figure A.13 (top, right). Weight matrices W_P and W_Q are shown in Figure A.13 (bottom). Note that the graphs G_P and G_Q are similar (but not identical) models of the manifolds underlying the data sets P and Q, respectively.

Step (4): We can capture the fact that points P_1 and Q_1 are similar to each other, although not close to each other, by adding an edge that connects p_1 with q_1. Similarly, we can add an edge that connects p_{12} and q_{12}, and say, p_6



W_P

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}	p_{12}
p_1	0	1	0	0	0	0	0	0	0	0	0	0
p_2	1	0	1	0	0	0	0	0	0	0	0	0
p_3	0	1	0	1	0	0	0	0	0	0	0	0
p_4	0	0	1	0	1	0	0	0	0	0	0	0
p_5	0	0	0	1	0	1	0	0	0	0	0	0
p_6	0	0	0	0	1	0	1	0	0	0	0	0
p_7	0	0	0	0	0	1	0	1	0	0	0	0
p_8	0	0	0	0	0	0	1	0	1	0	0	0
p_9	0	0	0	0	0	0	0	1	0	1	0	0
p_{10}	0	0	0	0	0	0	0	0	1	0	1	0
p_{11}	0	0	0	0	0	0	0	0	0	1	0	1
p_{12}	0	0	0	0	0	0	0	0	0	0	1	0

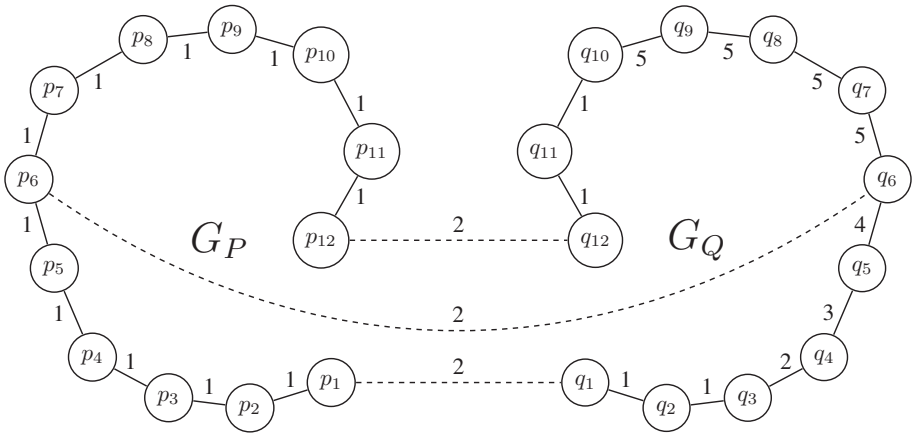
W_Q

	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9	q_{10}	q_{11}	q_{12}
q_1	0	1	0	0	0	0	0	0	0	0	0	0
q_2	1	0	1	0	0	0	0	0	0	0	0	0
q_3	0	1	0	2	0	0	0	0	0	0	0	0
q_4	0	0	2	0	3	0	0	0	0	0	0	0
q_5	0	0	0	3	0	4	0	0	0	0	0	0
q_6	0	0	0	0	4	0	5	0	0	0	0	0
q_7	0	0	0	0	0	5	0	5	0	0	0	0
q_8	0	0	0	0	0	0	5	0	5	0	0	0
q_9	0	0	0	0	0	0	0	5	0	5	0	0
q_{10}	0	0	0	0	0	0	0	0	5	0	1	0
q_{11}	0	0	0	0	0	0	0	0	0	1	0	1
q_{12}	0	0	0	0	0	0	0	0	0	0	1	0

Figure A.13: (top) Manifolds G_P (left) and G_Q (right) constructed from the data sets P and Q, respectively. (bottom) The corresponding weight matrices W_P (left) and W_Q (right).

and q_6, capturing the similarity between these pairs of points across manifolds. Crucially, we are free to choose weights that reflect the degree of similarity between the points. In this case, we assign each new edge a weight of 2, as shown in Figure A.14 (top). The weight matrix for the newly formed graph is shown in Figure A.14 (bottom). Note that the upper left quadrant of this weight matrix is simply the weight matrix W_P for the graph G_P, and the lower right quadrant is simply the weight matrix W_Q for the graph G_Q. The newly added edges between the nodes of G_P and G_Q are represented in the lower left and upper right quadrants.

Step (5): Now that we have created a single graph from two, we can use its weight matrix to compute its Laplacian, and then compute the eigenvalues and corresponding eigenvectors of that Laplacian. The eigenvector e_1 is now an eigenmapping from the nodes of G_P and G_Q to a one-dimensional analysis space, and hence a mapping of the data points in P and Q to the analysis



	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}	p_{12}	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9	q_{10}	q_{11}	q_{12}
p_1	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
p_2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p_3	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p_4	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p_5	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p_6	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
p_7	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p_8	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p_9	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p_{10}	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
p_{11}	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
p_{12}	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2
q_1	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
q_2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
q_3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0
q_4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	3	0	0	0	0	0	0	0
q_5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	4	0	0	0	0	0	0
q_6	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	4	0	5	0	0	0	0	0
q_7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	5	0	0	0	0
q_8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	5	0	0	0
q_9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	5	0	0
q_{10}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	1	0
q_{11}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
q_{12}	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0

Figure A.14: Alignment of two manifolds G_P and G_Q. (top) Graphs G_P and G_Q are aligned to form a new manifold. (bottom) The alignment computation represented as a combination of the weight matrices W_P and W_Q.

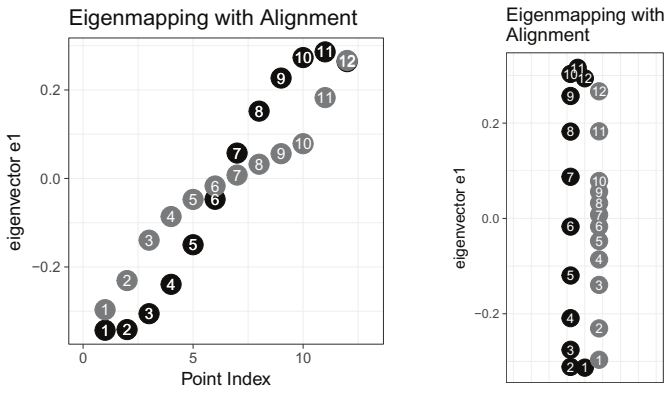


Figure A.15: Eigenmapping from the nodes of G_P and G_Q in terms of point indices in the original data sets (left) and a one-dimensional analysis space (right).

space. In this case, the first 12 components of e_1 act as the eigenmapping for the points in P, and the last 12 components the eigenmapping for the points in Q. The eigenmapping is shown in Figure A.15 (left and right), with points jittered horizontally in the latter case for clarity. Note that the computation described above yields the representations in Figure 2 (middle and right).

References

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, *15*, 1373–1396. doi:10.1162/089976603321780317.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 1798–1828. doi:doi.ieeecomputersociety.org/10.1109/TPAMI.2013.50.

Boë, L.-J., & Maeda, S. (1998). Modélisation de la croissance du conduit vocal. Espace vocalique des nouveaux–nés et des adultes. Conséquences pour l'ontegenèse et la phylogenèse. *Journées d'Études Linguistiques: "La Voyelle dans Tous ces États"*, (pp. 98–105).

de Boer, B., & Fitch, W. T. (2010). Computer models of vocal tract evolution: An overview and critique. *Adaptive Behavior*, *18*, 36–47. doi:10.1177/1059712309350972.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*, 341–345.

Burg, J. P. (1967). Maximum entropy spectral analysis. In *Proc. 37th Meeting, Society of Exploration Geophysicists*. Oklahoma City, OK. Reprinted in D. G. Childers, Ed., Modern Spectrum Analysis (IEEE Press, New York, 1978), pp. 34–41.

Chiba, T., & Kajiyama, M. (1941). *The vowel, its nature and structure*. Tokyo: Kaiseikan.

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

Cvetković, D., Rowlinson, P., & Simić, S. (2010). *An Introduction to the Theory of Graph Spectra*. London Mathematical Society Student Texts 75. Cambridge University Press.

Deza, M. M., & Deza, E. (2009). Encyclopedia of distances. In *Encyclopedia of Distances* (pp. 1–583). Springer.

Drager, K. K. (2011). Sociophonetic variation and the lemma. *Journal of Phonetics*, *39*, 694–707.

Edwards, J., & Beckman, M. E. (2008a). Methodological questions in studying consonant acquisition. *Clinical Linguistics and Phonetics*, *22*, 937–956.

Edwards, J., & Beckman, M. E. (2008b). Some cross-linguistic evidence for modulation of implicational universals by language-specific frequency effects in phonological development. *Language Learning and Development*, *4*, 122–156.

Edwards, J. R., Beckman, M. E., & Munson, B. (2015). Cross-language differences in acquisition. In M. A. Redford (Ed.), *Handbook of Speech Production* chapter 23. (pp. 530–554). Chichester, UK: Wiley-Blackwell.

Errity, A., & McKenna, J. (2006). An investigation of manifold learning for speech analysis. In *Proceedings of The International Conference on Spoken Language Processing* (pp. 2506–2509).

Fant, G. (1960). *Acoustic theory of speech production, with calculations based on X-ray studies of Russian articulations*. The Hague: Mouton.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America*, *84*, 115–123.

Fox, R. A., & Nissen, S. L. (2005). Sex-related acoustic changes in voiceless English fricatives. *Journal of Speech, Language, and Hearing Research*, *48*, 753–765. doi:10.1044/1092-4388(2005/052).

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, *102*, 594–621.

Hamm, 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* (pp. 120–127). volume 10.

Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, *34*, 458–484. doi:https://doi.org/10.1016/j.wocn.2005.10.001.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*, 82–97. doi:10.1109/MSP.2012.2205597.

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

Huang, H., ten Bosch, L., Cranen, B., & Boves, L. (2016a). Phone classification via manifold learning based dimensionality reduction algorithms. *Speech Communication*, *76*, 28 – 41. URL: http://www.sciencedirect.com/science/article/pii/S016763931500120X. doi:https://doi.org/10.1016/j.specom.2015.10.005.

Huang, H., Liu, Y., ten Bosch, L., Cranen, B., & Boves, L. (2016b). Locally learning heterogeneous manifolds for phonetic classification. *Computer Speech and Language*, *38*, 28 –45. URL: http://www.sciencedirect.com/science/article/pii/S0885230815001114. doi:https://doi.org/10.1016/j.csl.2015.12.002.

Iskarous, K., Goldstein, L. M., Whalen, D. H., Tiede, M. K., & Rubin, P. E. (2003). CASY: the haskins configurable articulatory synthesizer. In *ICPhS-15* (pp. 185–188).

Jafari, A., & Almasganj, F. (2010). A family of discriminative manifold learning algorithms and their application to speech recognition. *IEEE Transactions on Audio Speech and Language Processing*, *52*, 725–735.

Jakobson, R., Gunnar, F., & Halle, M. (1951). *Preliminaries to speech analysis: The distinctive features and their correlates*. Cambridge, MA: MIT Press.

Jannedy, S., & Weirich, M. (2017). Spectral moments vs discrete cosine transformation coefficients: Evaluation of acoustic measures distinguishing two merging German fricatives. *Journal of the Acoustical Society of America*, *142*, 395–405. doi:10.1121/1.4991347.

Jansen, A., & Niyogi, P. (2006). Intrinsic fourier analysis on the manifold of speech sounds. In *in IEEE Proceedings of International Conference on Acoustics, Speech, and Signal Processing* (pp. 241–244).

Jansen, A., & Niyogi, P. (2007). Semi–supervised learning of speech sounds. In *Proceedings of INTERSPEECH 2007*.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*, 359–384.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*, 1252–1263.

Koenig, L. L., Shadle, C. H., Preston, J. L., & Mooshammer, C. R. (2013). Toward improved spectral measures of /s/: Results from adolescents. *Journal of Speech, Language, and Hearing Research*, *56*, 1175–1189.

Ladd, D. R. (2012). What is duality of patterning, anyway? *Language and Cognition*, *4*, 261–273. doi:10.1515/langcog-2012-0015.

Li, F. (2012). Language-specific developmental differences in speech production: A cross-language acoustic study. *Child Development*, *83*, 1303–1315.

Li, F., Rendall, D., Vasey, P. L., Kinsman, M., Ward-Sutherland, A., & Diano, G. (2016). The development of sex/gender-specific /s/ and its relationship to gender identity in children and adolescents. *Journal of Phonetics*, *57*, 59–70. doi:10.1016/j.wocn.2016.05.004.

Lindblom, B. E. F., & Sundberg, J. E. F. (1971). Acoustical consequences of lip, tongue, jaw, and larynx movement. *Journal of the Acoustical Society of America*, *50*, 1166–179.

Ma, Y., & Fu, Y. (2012). *Manifold Learning Theory and Applications*. CRC Press.

Maeda, S. (1991). On articulatory and acoustic variabilities. *Journal of Phonetics*, *19*, 321–331.

Makhoul, J. (1973). Spectral analysis of speech by linear prediction. *IEEE Transactions on Audio and Electroacoustics*, *21*, 140–148. doi:10.1109/TAU.1973.1162470.

Messum, P., & Howard, I. (2015). Creating the cognitive form of phonological units: The speech sound correspondence problem in infancy could be solved by mirrored vocal interactions rather than by imitation. *Journal of Phonetics*, *53*, 125–140. doi:10.1016/j.wocn.2015.08.005.

Moore, B. C. J. (2012). *An Introduction to the Psychology of Hearing*. (Sixth ed.). Emerald.

Nissen, S. L., & Fox, R. A. (2005). Acoustic and spectral characteristics of young children's fricative productions: A developmental perspective. *Journal of the Acoustical Society of America*, *118*, 2570–2578.

Nittrouer, S., Studdert-Kennedy, M., & McGowan, R. S. (1989). The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech and Hearing Research*, *32*, 120–132. doi:10.1044/jshr.3201.120.

Norouzian, A., Rose, R., & Jansen, A. (2013). Semi-supervised manifold learning approaches for spoken term verification. In *Proceedings of INTERSPEECH 2013*.

Nossair, Z. B., & Zahorian, S. A. (1991). Dynamic spectral shape features as acoustic correlates for initial stop consonants. *Journal of the Acoustical Society of America*, *89*, 2978–2991.

Oohashi, H., Watanabe, H., & Taga, G. (2017). Acquisition of vowel articulation in childhood investigated by acoustic-to-articulatory inversion. *Infant Behavior and Development*, *46*, 178–193. doi:10.1016/J.INFBEH.2017.01.007.

Oudeyer, P.-Y. (2005). How phonological structures can be culturally selected for learnability. *Adaptive Behavior*, *13*, 269–280. doi:10.1177/105971230501300407.

Plummer, A. R. (2014). *The acquisition of vowel normalization: Theory and computational framework*. Ph.D. thesis The Ohio State University.

Plummer, A. R. (TBD). The challenges of developing articulatory synthesis models of early vocal production in humans. In *175th Meeting of the Acoustical Society of America, POMA*. Minneapolis, MN.

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.

Rasilo, H., Räsänen, O., & Laine, U. K. (2013). Feedback and imitation by a caregiver guides a virtual infant to learn native phonemes and the skill of speech inversion. *Speech Communication*, *55*, 909–931. doi:10.1016/j.specom.2013.05.002.

Reidy, P. F. (2015). *The spectral dynamics of voiceless sibilant fricatives in English and Japanese*. Ph.D. thesis The Ohio State University.

Reidy, P. F., Kristensen, K., Winn, M. B., Litovsky, R. Y., & Edwards, J. R. (2017). The acoustics of word-initial fricatives and their effect on word-level intelligibility in children with bilateral cochlear implants. *Ear and Hearing*, *38*, 42–56. doi:10.1097/AUD.0000000000000349.

Romeo, R., Hazan, V., & Pettinato, M. (2013). Developmental and gender-related trends of intra-talker variability in consonant production. *Journal of the Acoustical Society of America*, *134*, 3781–3792. doi:10.1121/1.4824160.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*, 2323–2326. doi:10.1126/science.290.5500.2323.

Rubin, P., Baer, T., & Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America*, *70*, 321–328. doi:10.1121/1.386780.

Rvachew, S., Mattock, K., Polka, L., & Ménard, L. (2006). Developmental and cross-linguistic variation in the infant vowel space: The case of Canadian English and Canadian French. *Journal of the Acoustical Society of America*, *120*, 2250–2259.

Saltzman, E., & Holt, K. (2014). Movement forms: A graph-dynamic perspective. *Ecological Psychology*, *26*, 60–68.

Saltzman, E., Kubo, M., & Tsao, C.-C. (2006). Controlled variables, the uncontrolled manifold, and the task–dynamic model of speech production. In P. Divenyi, S. Greenberg, & G. Meyer (Eds.), *Dynamics of Speech Production and Perception* NATO Science Series (pp. 21–31). Amsterdam: IOS Press.

Scholz, J. P., & Schöner, G. (1999). The uncontrolled manifold concept: Identifying control variables for a functional task. *Experimental Brain Research*, *126*, 289–306.

Seung, H. S., & Lee, D. D. (2000). The manifold ways of perception. *Science*, *290*, 2268–2269.

Story, B. H., & Bunton, K. (2016). Formant measurement in children's speech based on spectral filtering. *Speech Communication*, *76*, 93–111. doi:10.1016/j.specom.2015.11.001.

Strand, E. A., & Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In D. Gibbon (Ed.), *Natural language processing and speech technology: Results of the 3rd KONVENS Conference, Bielfelt* (pp. 14–26). Berlin: Mouton de Gruyter.

Szabados, A., & Perrier, P. (2016). Uncontrolled manifolds in vowel production: Assessment with a biomechanical model of the tongue. In *Interspeech 2016* (pp. 3579–3583). URL: http://dx.doi.org/10.21437/Interspeech.2016-1579. doi:10.21437/Interspeech.2016-1579.

Tenenbaum, J. B., de Sliva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*, 2319–2323. doi:10.1126/science.290.5500.2319.

Tomar, V., & Rose, R. (2014). Using laplacian eigenmaps latent variable model and manifold learning to improve speech recognition accuracy. *Speech Communication*, *22*. doi:doi-org.proxy.lib.ohio-state.edu/10.1016/j.specom.2010.04.005.

Vihman, M. M. (2014). *Phonological development: The first two years*. (2nd ed.). Malden, MA: Wiley-Blackwell.

Wang, C., & Mahadevan, S. (2009). Manifold alignment without correspondence. In *IJCAI'09: Proceedings of the 21st international jont conference on Artifical intelligence* (pp. 1273–1278). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Warlaumont, A. S., Westermann, G., Buder, E. H., & Oller, D. K. (2013). Prespeech motor learning in a neural network using reinforcement. *Neural Networks*, *38*, 64–75.

Warren, P. (2017). The interpretation of prosodic variability in the context of accompanying sociophonetic cues. *Laboratory Phonology*, *8*, 11. doi:10.5334/labphon.92.

Warren, P., Hay, J. B., & Thomas, B. (2007). The loci of sound change effects in recognition and perception. In J. S. Cole, & J. I. Hualde (Eds.), *Laboratory Phonology 9* (pp. 87–111). Mouton de Gruyter.

Zhao, X., & Zhang, S. (2012). Phoneme recognition using an adaptive supervised manifold learning algorithm. *Neural Computing and Applications*, *21*, 1501 – 1515. URL: http://proxy.lib.ohio-state.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=79956059&site=ehost-live.