Intrinsic Spectral Analysis

Aren Jansen, *Member, IEEE*, and Partha Niyogi

***Abstract*—It has long been posited that the space of speech sounds is inherently low dimensional, the result of a relatively small number of degrees of freedom involved in the human vocal apparatus. We attempt to formalize this notion by analyzing a simple physical model of the vocal tract and demonstrating that it produces transfer functions whose spectra are restricted to low dimensional manifolds embedded in an infinite dimensional space of square integrable functions. While source convolution and channel distortion precludes analytic recovery of the articulatory configuration from the observed signal, we present a data-driven unsupervised learning algorithm called Intrinsic Spectral Analysis designed to recover from a stream of unannotated and unsegmented audio a set of nonlinear basis functions for the speech manifold. Projecting a traditional spectrogram onto this nonlinear basis defines a novel acoustic representation that is demonstrated to have phonological significance, improved phonetic separability, inherent speaker independence, and complementarity with standard acoustic front-ends.**

***Index Terms*—Manifold learning, speech processing, speech recognition, unsupervised learning.**

I. Introduction

A WIDE range of acoustic signals are generated via a complex physical process with relatively few degrees of freedom. The melody of a musical instrument is controlled via a small number of mouth or finger positions; the hum of a CPU fan is controlled via the electrical current passed though the electric motor; and the message in one's speech is controlled via articulators such as the glottis, tongue, and lips. In each case, the production mechanisms define maps from a discrete or continuous low dimensional parameter or configuration space to a high dimensional observation space, commonly taken to be frequency spectra. Recovering the low dimensional parametrization is often of substantial interest since it can provide insight into the underlying state or intent of the system. However, in most cases of interest, observations are corrupted by noise and other channel distortions that may preclude the application of analytic solutions towards this end. As such, we are interested in data-driven approaches to estimating from potentially large amounts of unannotated and unsegmented audio a set of nonlinear projection maps that recover the underlying configuration space.

Manifold learning is a popular statistical framework for geometric data analysis that attempts to address this problem and has been well studied in the machine learning community over the past decade. This class of techniques are united by a common assumption that high-dimensional natural data sets with relatively few underlying degrees of freedom may reside on (or near, if noise is considered) low dimensional manifolds embedded in the ambient observation space. A number of algorithmic frameworks are based on this perspective, including locally linear embedding (LLE) [1]–[3], Laplacian eigenmaps [4], manifold regularization [5], diffusion maps [6], ISOMAP [7], and local tangent space alignment [8]. The success of this class of algorithms rests on the validity of the manifold structural assumption for each application area.

In this paper we focus on speech, where it has long been understood [9] that there exists a low-dimensional parametrization that both linguists and engineers have sought to accurately characterize for scientific and technological purposes. The articulatory parametrization of speech production developed by Fant [10], Stevens [11], and others strongly point to the existence of the low-dimensional manifold structure required by the above-mentioned manifold learning algorithms. In attempt to formalize these long-maintained intuitions, we consider the source-filter model of speech production [10], which involves filtering a glottal or turbulent source by the vocal tract resonator. The precise form of the vocal tract transfer function and, consequently, the phonetic content of the acoustic signal produced, is controlled by the spatial configuration of the vocal tract. The map from the observation space to articulatory configuration space, if smooth and bijective, is the coordinate chart necessary to formally substantiate the existence of an underlying manifold structure.

In real speech applications, where we are without the precise analytic form of this coordinate chart, we can still resort to data-driven means of approximating it. Toward this end, we detail Intrinsic Spectral Analysis (ISA), a casting of a manifold learning algorithm as a signal processing technique. First introduced in [12] and [13], ISA is designed to approximate a set of nonlinear projection maps onto an intrinsic coordinate system using the machinery of graph Laplacian spectral clustering [5]. For a d-dimensional manifold, an intrinsic coordinate system provides a more fundamental embedding in $\mathbb{R}^d$, where Euclidean distance becomes a proxy for geodesic distance on the manifold (the length of the shortest path restricted to the manifold). For the above-described vocal tract model, the intuition is that this intrinsic space would explicitly encode a useful parametrization of the articulatory configuration space.

Manuscript received December 20, 2011; revised June 25, 2012, October 11, 2012, and November 30, 2012; accepted December 10, 2012. Date of publication January 11, 2013; date of current version March 12, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Konstantinos Slavakis. Part of the work described in this paper was conducted while A. Jansen was a postdoctoral scholar at the University of Chicago.

A., Jansen was with the University of Chicago, Chicago, IL, USA. He is now with the Human Language Technology Center of Excellence and Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: aren@jhu.edu).

P. Niyogi, deceased, was with the Department of Computer Science, University of Chicago, Chicago, IL, 60637 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2013.2238931

Unlike the related Laplacian eigenmaps dimensionality reduction algorithm [4], ISA admits out-of-sample extension and thus may be applied online for large scale speech technologies. Like cepstral analysis and linear prediction, ISA is applied on top of traditional short time Fourier analysis, using the data itself to recover an transformation of the spectrogram into alternative basis. While we evaluate the usefulness of the embedding produced by ISA in a range of downstream applications, we are also interested in interpreting its behavior in linguistic terms, given the acoustic phonetic motivations from which we begin. The phonological theory of distinctive features [14]–[16] is a natural point of contact given its explicit articulatory motivations and their explanatory power for the confusion patterns observed in both human [17] and machine [18] recognition of speech sounds. We will demonstrate that many ISA dimensions are correlates to individual distinctive features, a set of binary production characteristics that are independent of the phoneme they are expressed in or the vocal characteristics of the speaker.

The following sections present an interdisciplinary investigation into the rich structure underlying natural speech signals in an effort to engage interrelated but rarely connected concepts from linguistics (acoustic phonetics and phonology), machine learning (unsupervised and manifold learning), signal processing (short time Fourier analysis), and speech recognition (zero resource speech recognition, speaker independence, and phone recognition). In Section II, we provide detailed motivation for the existence a low dimensional manifold structure for speech sounds using a simple parametric model of the vocal tract. In Section III we briefly review manifold learning techniques as a prerequisite for the ensuing presentation of the ISA algorithm in Section IV. Finally, we evaluate ISA on both synthetic and natural speech audio, and demonstrate the proposed technique recovers a set of nonlinear projection maps that have both linguistic and technological significance. Note that while our treatment focuses squarely on the application to speech audio, it is important to note that the proposed analysis technique is applicable as defined to any signal (audio or otherwise) where short time Fourier analysis is employed.

II. The Geometric Structure of Speech Sounds

The laws governing acoustic physics determine a map $\phi : \mathcal{A} \rightarrow \mathcal{M}$ from the space of possible articulatory configurations, $\mathcal{A}$, to the space of vocal tract transfer functions, $\mathcal{M} \subset \mathcal{L}^2$, where, $\mathcal{L}^2$ is the space of square integrable functions. Each speech utterance, evolving as a function of time t, is produced according to a path a_t in $\mathcal{A}$ that is mapped under ϕ to a corresponding path $g_t = \phi(a_t)$ in the space of vocal tract transfer functions $\mathcal{M}$ (see Fig. 1). The observed speech waveform is produced by (i) modulating the filter transfer function at time t by the corresponding source spectrum and (ii) applying the inverse Fourier transform to recover the signal.

The analysis presented in this section centers around a brief derivation of sounds generated by series of concatenated tubes with a configuration space $\mathcal{A}$ defined by the lengths and cross sectional areas of the tube segments. This system serves as a simple model for the vocal tract proven useful in the tradition of acoustic phonetics for capturing the percept of certain vowel sounds. We argue that if the resulting map ϕ is a diffeomorphism, which we motivate without formal proof, it follows that the set of transfer functions for these finitely-parametrized acoustic tube filters forms a low-dimensional manifold that is diffeomorphic to $\mathcal{A}$. Using the derived coordinate chart, we generate and visualize selected synthetic data manifolds, which provide a compelling alternative to the canonical examples (e.g., the swiss roll) for toy experimentation in Section III.

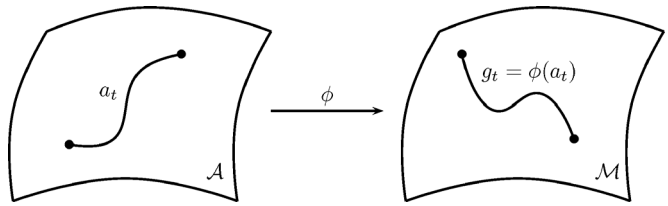

Fig. 1. A path a_t in articulatory configuration space $\mathcal{A}$ is mapped to a corresponding path g_t in the space vocal tract transfer functions $\mathcal{M}$. If ϕ is a diffeomorphism, $\mathcal{M}$ is a manifold with dimension $\dim(\mathcal{A})$.

A. The Physics of Acoustic Tubes

The acoustic analysis of the vocal tract resonator can be reduced to the tractable problem of concatenated uniform acoustic tubes with the introduction of several approximations, valid for human speech up to 5 kHz [11]: (i) rigid vocal tract walls, (ii) small transverse vocal tract dimensions, (iii) small pressure, density, and velocity perturbations, and (iv) transient perturbations. Under these approximations the one-dimensional equations of compressible fluid flow continuity and conservation of momentum reduce to the acoustic equations for the uniform tube filter,

$$\frac{\partial p}{\partial t} = \frac{\gamma p_0}{A}\frac{\partial U}{\partial x} \quad \text{and} \quad \frac{\partial p}{\partial x} = \frac{\rho_0}{A}\frac{\partial U}{\partial t}, \tag{1}$$

where p_0 and ρ_0 are the equilibrium air pressure and density, p and U are the wave pressure and volume velocity deviations from equilibrium, $\gamma = 5/3$, and A is the cross-sectional area of the tube. The independent variable x is the position along the tube axis.

The first boundary condition specifies the volume velocity input into the system at $x = 0$. The second condition exploits the fact that a sound wave faces the acoustic impedance of the surrounding environment at the open end of the tube ($x = L$). These constraints are formulated by the relations

$$\hat{U}(0,\omega) = \hat{s}(\omega) \quad \text{and} \quad \hat{U}(L,\omega) = \frac{\hat{p}(L,\omega)}{\mathbf{Z}_r(\omega)}, \tag{2}$$

where $\hat{U}$ and $\hat{p}$ are the Fourier transforms of U and p, $\hat{s}$ is the Fourier transform of the volume velocity source, and ω is the angular frequency. Here,

$$\mathbf{Z}_r(\omega) = \frac{\rho_0 c k^2 K_s(\omega)}{4\pi} + i\frac{4ck\rho a}{5A}, \tag{3}$$

is the approximate form of the radiation impedance given by Stevens [11] (valid up to 6 kHz), based on a model of a circular piston of air (the mouth) on the surface of a sphere with radius 9 cm (the head). Here $c = \sqrt{\gamma p_0/\rho_0}$ is the speed of sound in air, $A = \pi a^2$ is the cross sectional area of the piston, and

$k = \omega/c$ is the wave number. The term K_s is a real-valued frequency-dependent factor (see [11] for details).

Now consider the generalized filter composed of a series of N tubes with lengths $\{L_i\}$ and cross-sectional areas $\{A_i\}$. Relying on continuity of pressure and volume velocity at inter-tube boundaries, the solution for N concatenated tubes is equivalent to determining N single tube solutions. Given this filter geometry, the output pressure spectrum that satisfies (1) with the boundary conditions of (2) takes the form $\hat{p}(\omega) = \hat{s}(\omega)g(\omega, \{L_i\}, \{A_i\})$, where

$$g(\omega, \{L_i\}, \{A_i\}) = \frac{\mathbf{Z}_r(\omega)}{M_{22} - \mathbf{Z}_r(\omega)M_{12}}. \tag{4}$$

is the transfer function for the entire N-tube filter. Here, M is the product of 2×2 transformation matrices of the form $M = \prod_{i=1}^{N} C_i$, where the transformation matrix C_i for the ith tube is given by

$$C_i = \begin{bmatrix} \cos kL_i & i\frac{A_i}{\rho_0 c}\sin kL_i \\ i\frac{\rho_0 c}{A_i}\sin kL_i & \cos kL_i \end{bmatrix}.$$

We have made available [19] a software package that generates speech sounds for an arbitrary N-tube filter and source function and is capable of real-time synthesis as model parameters are continuously varied.

B. The Speech Manifold

Consider the single tube acoustic tube filter with length L and cross-sectional area A, a simple model for the production of the neutral vowel /ə/. The transfer function g is given by (4) with $N = 1, L_1 = L$, and $A_1 = A$. Let $\mathcal{M}_1$ be the subset of $\mathcal{L}^2$ defined by

$$\mathcal{M}_1(I_L, I_A) = \{g(\omega, L, A) | L \in I_L, A \in I_A\},$$

where I_x is an open interval of parameter x. Let $I_L^{\mathrm{h}} \approx (10, 30)$ cm be the range of human vocal tract lengths and $I_A^{\mathrm{h}} \approx (0, 20)$ cm^2 be the range of human vocal tract cross sectional areas during production of /ə/. If the map $\phi : (L, A) \in \mathbb{R}^2 \rightarrow \mathcal{M}_1$ defined by g is a diffeomorphism (bijective and infinitely differentiable, along with the inverse) for $L \in I_L^{\mathrm{h}}$ and $A \in I_A^{\mathrm{h}}$, it follows that ϕ^{-1} is a coordinate chart on the set $\mathcal{M}_1$ and $\mathcal{M}_1$ is formally a two-dimensional smooth manifold embedded in the observed ambient space $\mathcal{L}^2$. Note that most of the diffeomorphism prerequisites are clear upon inspection of the denominator in (4), provided $\mathrm{Re}(\mathbf{Z}_r) \neq 0$ and $A > 0$. Further details can be found in [12].

To visualize these acoustic tube filter transfer function manifolds, which are embedded in the function space $\mathcal{L}^2$, we can begin by approximating each vocal tract transfer function g as the equivalent infinite sequence in $g^* \in \ell^2$. In particular, given a frequency sampling interval $\Delta\omega$, we define $g_n^* = g(n\Delta\omega)$. Moreover, if we assume bounded support such that $g(\omega) = 0$ for $\omega > \omega_c, g^*$ becomes finite and we approximate each function $g \in \mathcal{L}^2$ with an element in $\mathbb{R}^d$, where $d = \omega_c/\Delta\omega$. Note that under this approximation, the $\mathcal{L}^2$ inner product between two functions is replaced with a scalar product between the two corresponding vectors in $\mathbb{R}^d$.

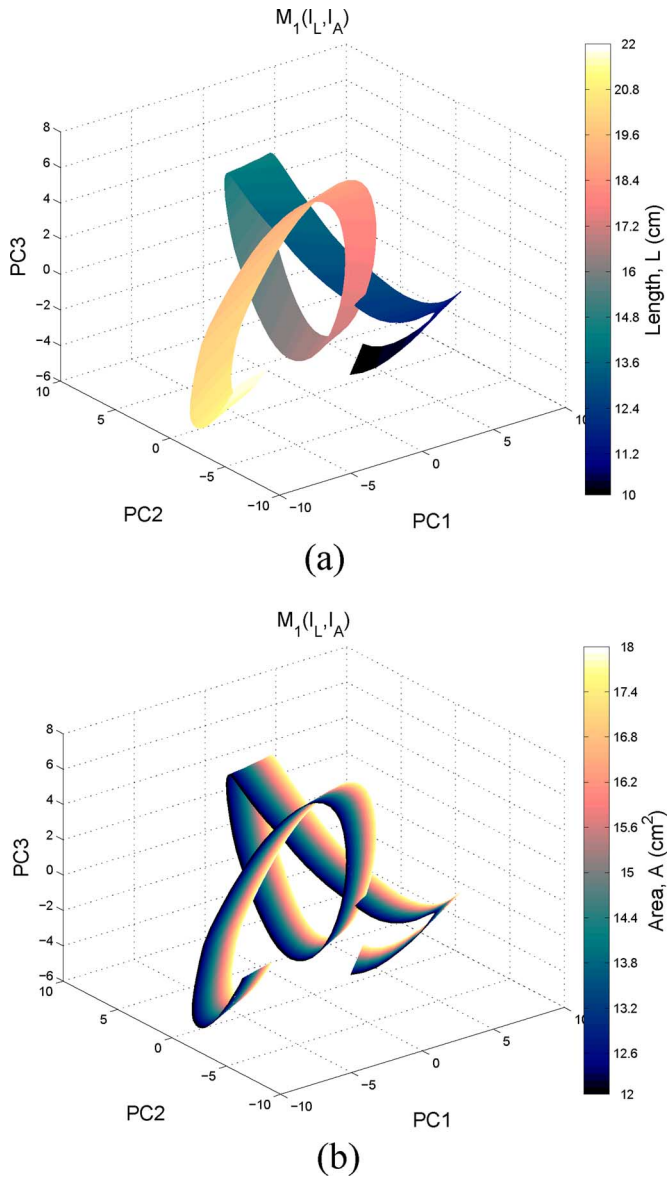

Fig. 2. Three-dimensional principal component projection of the manifold $\mathcal{M}_1(I_L, I_A)$. Color is used to distinguish (a) length L and (b) cross-sectional area A parameter values.

Fig. 2 uses this visualization strategy to display a 3-dimensional principal component projection of the single tube manifold $\mathcal{M}_1(I_L, I_A)$ for the length and cross-sectional area ranges $I_L = (10, 22)$ cm and $I_A = (12, 18)$ cm^2, respectively. Here, we use the frequency cutoff $\omega_c = 2\pi \times 5$ kHz and sampling interval $\Delta\omega = 2\pi \times 50$ Hz, resulting in a manifold embedded in 100-dimensional Euclidean space (one dimension per frequency channel for $\omega_c/\Delta\omega = 100$ total bins) and projected onto its three principal axes. This 2-manifold may be interpreted as the space of transfer functions for the neutral vowel /ə/ corresponding to the full range of human vocal tract sizes. The manifold takes on a ribbon-like structure with clear extrinsic curvature (i.e., curvature evident when embedded in the ambient observation space). Figs. 2(a) and (b) use color to differentiate tube length and cross-sectional area, respectively.

We saw in (4) that the N-tube transfer function solutions involve one matrix multiplication per tube segment. It follows that

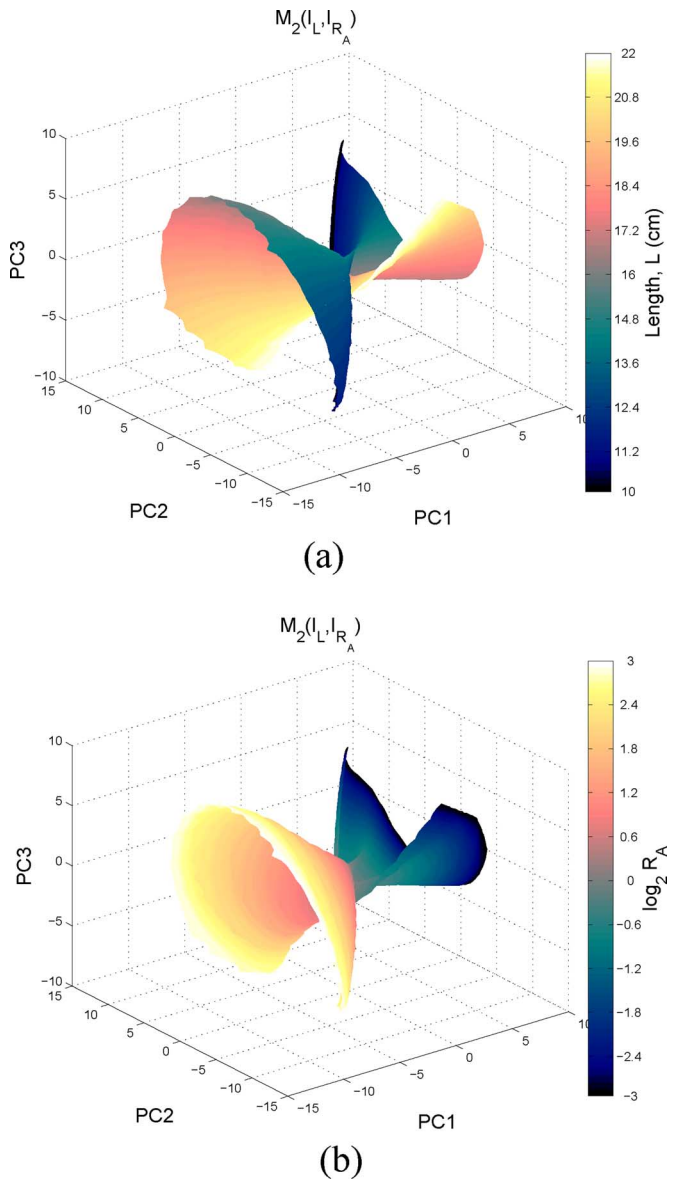

Fig. 3. Three-dimensional principal component projection of the manifold $\mathcal{M}_2(I_L, I_{R_A})$. Color is used to distinguish (a) length L and (b) cross-sectional area ratio R_A parameter values.

solution complexity monotonically increases with N, making analysis even more cumbersome than is evident for the single tube case. However, if the N-tube transfer function is also a diffeomorphism, then it too would assume a formal manifold structure with dimension equal to the number of configuration parameters independently varied (at most $2N$). For N sufficiently large, the acoustic tube filter can simulate the vocal tract to an accuracy that is limited only by the approximations used by the physical model. It would thus follow that a low-dimensional manifold structure exists for the space of speech sounds, assuming of course that the source spectrum is also bijective and infinitely differentiable.

The N-tube model is capable of capturing detailed approximations to true vocal tract profiles, which can now be safely measured in humans during production using magnetic resonance imaging [20]. However, there is a rich literature in acoustic phonetics, dating back to Fant in 1960 [10] and further developed by Stevens [11], demonstrating that two concatenated tubes are sufficient to capture the dominant formant structure variation present within the class of vowel sounds. With this motivation, we next consider a synthetic speech manifold designed to represent variation across both speaker (vocal tract length) and phonetic class (cross sectional area ratio between the two tubes).

Fig. 3 displays the 3-dimensional principal component projection of the 2-tube manifold

$$\mathcal{M}_2(I_L, I_{R_A}) = \{g(\omega, L_1, L_2, A_1, A_2) | L \in I_L, R_A \in I_{R_A}\}$$

as the total tract length $L = L_1 + L_2$ and area ratio $R_A = A_1/A_2$ are varied in the human ranges $I_L = [10, 22]$ cm and $I_{R_A} = [1/8, 8]$, respectively. We fix $L_1/L_2 = 1$ and $\max(A_1, A_2) = 15$ cm^2, leaving two degrees of freedom. We again use the frequency cutoff $\omega_c = 2\pi \times 5$ kHz and sampling interval $\Delta\omega = 2\pi \times 50$ Hz, resulting in a 100-dimensional Euclidean space approximation. Note that the apparent degeneracy is simply an artifact of the principal component projection (the map is one-to-one). This is another 2-manifold that may be roughly interpreted as the space of transfer functions across the full range of vowel height ($R_A = 1/8$ for the low vowel /a/ and $R_A = 8$ for the high vowel /i/) and speaker vocal tract length. Figs. 3(a) and (b) use color to differentiate tract length and $\log_2 R_A$ coordinates, respectively. Extrinsic curvature of this two-dimensional manifold is again evident.

We have presented our discussion of the manifold properties in terms of the filter transfer function. However, any work with actual speech data will be in the form of a product of the transfer function with an independent source spectrum. In the case of sonorants, the vocal tract filter is driven by a combination of periodic glottal vibration and stochastic, but statistically regular, processes. For these sounds we can analytically model the source spectrum, allowing the set of output pressure spectra to inherit the manifold properties of the transfer function. For turbulence-driven obstruents, the manifold interpretation cannot be formally developed. However, noise source modulation by the filter transfer function will cause individual obstruent phonemes to cluster naturally, and the algorithms that we develop in the following sections can still apply under a clustering interpretation.

III. Graph Laplacian Manifold Learning

Now that we have explored the underlying manifold structure in the speech domain, it remains to develop techniques that might exploit it for practical applications. The geometric signal processing approach we develop below relies on the graph Laplacian operator. In this section, we provide a brief discussion of the operator theory and the existing manifold learning frameworks derived from it.

A. The Laplace-Beltrami and Graph Laplacian Operators

The Laplace-Beltrami operator on a Riemannian[1] manifold $\mathcal{M}$ is the second order differential operator typically denoted

[1]Riemannian manifolds have inner products defined in each local tangent space such that a notion of paths, curvature, and geodesic distance may be defined.

$\Delta_{\mathcal{M}}$. It is a positive semidefinite operator whose eigenfunctions form an orthogonal basis for $\mathcal{L}^2(\mathcal{M})$ [4]. If $\{\lambda_i\}$ and $\{e_i\}$ are the sorted eigenvalues and corresponding eigenfunctions of the Laplace-Beltrami operator, respectively, then any function $f : \mathcal{M} \to \mathbb{R}$ may be written $f = \sum_i a_i e_i$ for some $\{a_i\}$. In addition to providing a potentially useful basis for square integrable functions, the Laplace-Beltrami operator may be used to quantify the smoothness of functions defined on the manifold. Given a measure μ on $\mathcal{L}^2(\mathcal{M})$, the functional

$$S[f] = \int_{\mathcal{M}} \|\nabla_{\mathcal{M}} f\|^2 d\mu = \langle \Delta_{\mathcal{M}} f, f \rangle_{\mathcal{L}^2(\mathcal{M})} \tag{5}$$

increases as smoothness of f decreases [4]. Here, $\nabla_{\mathcal{M}}$ is the gradient operator on $\mathcal{M}$ and $\langle \cdot, \cdot \rangle_{\mathcal{L}^2(\mathcal{M})}$ is the $\mathcal{L}^2$ inner product on $\mathcal{M}$. It follows that the smoothness of an eigenfunction is determined by the magnitude of the corresponding eigenvalue, since $S[e_i] = \lambda_i$. Therefore, if we limit an eigenbasis expansion of a function f to finite terms, we can impose any desired level of smoothness in the approximation. Thus, initial eigenfunctions $e_i : \mathcal{M} \to \mathbb{R}$ vary most smoothly with *geodesic* distance on $\mathcal{M}$ and best preserve locality as defined on the manifold regardless of the particular form of the original embedding [4]. In this way, they define an optimal embedding and a natural choice of coordinate system for the manifold. We refer to these maps e_i as intrinsic basis functions below.

In practice, we are not given an analytical form of the manifold $\mathcal{M}$, so the Laplace-Beltrami operator cannot be used directly. However, if we are presented with a collection of points sampled from $\mathcal{M}$, we can implement the graph theoretic analogue as follows: Consider a manifold $\mathcal{M}$ embedded in $\mathbb{R}^d$ and a collection of n data points $X = \{x_1, \ldots x_n\} \subset \mathcal{M}$. We construct a weighted, undirected adjacency graph $G = (V, E)$ with one vertex $V_i \in V$ per data point $x_i \in X$. We connect vertices V_i and V_j with an edge of similarity weight W_{ij} if x_i is one of the κ nearest neighbors of x_j or x_j is one of the κ nearest neighbors of x_i. From this, we can determine the so-called graph Laplacian, $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is the diagonal vertex degree matrix with elements $D_{ii} = \sum_j W_{ji}$. One can also consider a normalized variant, $\mathbf{L}_{\text{norm}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, where $\mathbf{I}$ is the identity matrix. This normalization reduces the effect of large variation in vertex degree arising from sampling sparsity.

The graph Laplacian is a positive semidefinite $n \times n$ matrix that satisfies, in discrete analogue, all the properties given above for the continuous Laplace-Beltrami operator [4]. It follows that if we regard the graph as a mesh on the manifold, the basis determined by the graph Laplacian serves as an approximation to an intrinsic basis for the manifold that the sample was drawn from [4]. It is relevant to note two differences between the continuous and discrete cases. First, the graph analogue is limited to functions that are defined on the graph, not the entire manifold. We denote such functions by the column vector $\mathbf{f} = \langle f(x_1) f(x_2) \ldots f(x_n) \rangle^T$ containing the function image of X. Second, the $\mathcal{L}^2$ inner product is now replaced by an $\mathbb{R}^n$ inner product. It follows that the graph analogue of the smoothness functional of (5) becomes

$$S_G[\mathbf{f}] = \mathbf{f}^T \mathbf{L} \mathbf{f} = \frac{1}{2} \sum_{i,j} W_{ij} (f(x_i) - f(x_j))^2. \tag{6}$$

From the rightmost form of this smoothness functional, the connection of this framework to spectral clustering is highlighted. Consider a logical partition of the graph into two parts $V = V_1 \bigcup V_2, V_1 \bigcap V_2 = \emptyset$. If the graph is faithful to this partition, with vanishing edge weights between V_1 and V_2, then the function that minimizes S_G will reflect the minimum cut (after the constant function, which is ignored).

We see the graph Laplacian approach has dual purpose. In the case the data X provides a sufficiently dense mesh on a manifold, this approach defines a set of basis functions that define the (approximate) intrinsic coordinate system for the manifold. Even when the manifold structure is not present, there still may exist a logical partition of the data into clusters and the basis can reflect this structure as well. In this case, the eigenfunctions of the graph Laplacian with small corresponding eigenvalues each provide a projection map from data points to a real value that permits isolation of one of the data clusters. Note that there will be one zero eigenvalue for each cluster present in the data that produces a distinct connected component in the nearest neighbor graph. Near-zero eigenvalues will result from clusters that are weakly connected.

B. Previous Graph Laplacian Frameworks

The graph Laplacian was first introduced to the field of machine learning as the theoretical centerpiece of the Laplacian eigenmaps dimensionality reduction algorithm [21]. The reduction in dimension is accomplished by projecting points onto a small set of graph Laplacian eigenfunctions. As above, let $X = x_1, \ldots, x_n$ be a set of points sampled from the manifold $\mathcal{M} \subset \mathbb{R}^d$ and let $\mathbf{L}$ be the corresponding $n \times n$ graph Laplacian matrix constructed according to Section III.A. Then, if $\{\lambda_i\}$ and $\{\mathbf{e}_i\}$ are the sorted non-zero eigenvalues ($\lambda_{i+1} > \lambda_i$ and $\lambda_i > 0$ for all i) and corresponding eigenvectors of $\mathbf{L}$, respectively, the d'-dimensional Laplacian eigenmaps projection $\mathcal{P}^{\text{lap}}_{d'}$ for each data point $x_i \in X$ is defined by

$$\mathcal{P}^{\text{lap}}_{d'}(x_i) = \langle e_{1i}, e_{2i}, \ldots, e_{d'i} \rangle, \tag{7}$$

where e_{ji} is the ith component of $\mathbf{e}_j$.

Fig. 4 displays a two-dimensional Laplacian eigenmaps projection ($d' = 2$) of the 2-tube manifold $\mathcal{M}_2(I_L, I_{R_A})$ from Section II.B (pictured in Fig. 3). Here, we constructed a $\kappa = 200$ nearest neighbor adjacency graph over 10,000 samples from the manifold using heat kernel graph weights ($W_{ij} = e^{-|x_i - x_j|^2/2\sigma^2}$, with $\sigma = 1$). Notice that the Laplacian eigenmaps algorithm functions to recover the two-dimensional intrinsic parameter space, unrolling the extrinsic curvature out to a simple plane with orthogonal principle axes corresponding to variation in L and R_A.

Now, the most significant shortcoming of Laplacian eigenmaps is that it does not allow for out-of-sample extension, as the eigenfunctions recovered are only defined on the nodes of the graph. Therefore, any application to novel data would require recomputation of the adjacency graph and Laplacian matrix. However, the Laplacian eigenmaps algorithm was extended to handle out-of-sample application with the development of locality preserving projections (LPP) [22]. Here, the eigenfunctions are restricted to linear functions defined on the whole space. This results in a set of linear functions that

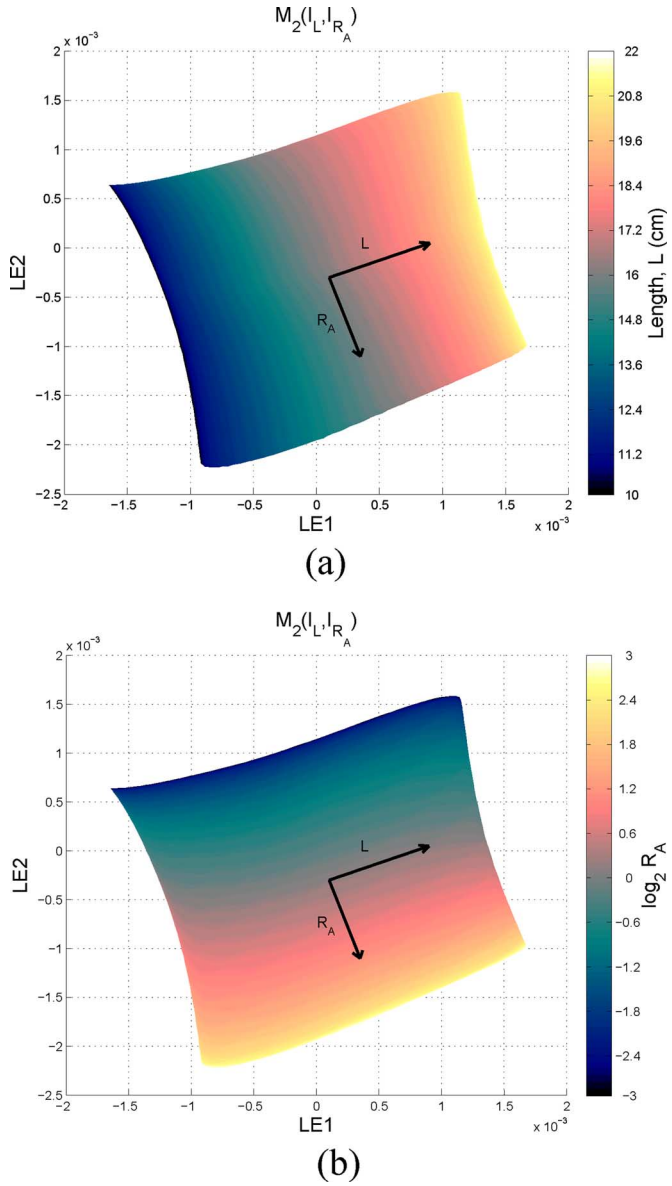

Fig. 4. Two-dimensional Laplacian eigenmaps projection of the manifold $\mathcal{M}_2(I_L, I_{R_A})$ defined in Section II.B (cf. Fig. 3). Color is used to distinguish (a) length L and (b) cross-sectional area ratio R_A parameters values.

are most similar to the actual eigenbasis, providing the best approximation on which to project novel data. This method was extended to more exotic function classes with kernelized locality preserving projections [23], which loosened the search to arbitrary reproducing kernel Hilbert spaces.

These dimensionality reduction techniques were followed by the development of a coherent framework for graph Laplacian-based semi-supervised learning known as manifold regularization [5]. Here, the graph Laplacian smoothness functional ((6)) is incorporated into a traditional optimization problem as a regularization term. In Section IV, we will use a special case of this manifold regularization framework to define our algorithm for ISA.

C. Other Manifold Learning Frameworks

As indicated in the introduction, there has been no shortage of general purpose manifold learning algorithms developed by the machine learning community in the past decade [1]–[8], though next-to-none have been seriously evaluated in the signal processing domain in the manner described here (one exception is ISOMAP, which was recently applied to speech in [24]). These algorithms could easily be used in place of graph Laplacian-based techniques, provided a method for out-of-sample extension is available. A recent paper has presented a generic approach for out-of-sample extension [25] that requires only a limited re-optimization for novel data points and may pave the way for future evaluation. However, the purpose of the present paper is to formally consider the question of manifold structure underlying speech sounds, present a viable algorithm for exploiting it, and demonstrate this point of view has practical merit. Given the experimental success demonstrated in the following sections, we invite others interested in manifold learning techniques to translate other techniques into this domain, though this sort of comprehensive evaluation falls outside the scope of this paper.

IV. Intrinsic Spectral Analysis

A traditional spectrogram is the discrete short time Fourier amplitude spectrum of an audio time series. The amplitude spectrum at time t_i is determined using a short window of the signal centered about t_i. Let $s_i(t)$ be the i-th signal window and let $\hat{s}_i(\omega)$ be the corresponding d-dimensional discrete amplitude spectrum. The spectrogram is then given by $\hat{s}(t_i, \omega_j) = \hat{s}_i(\omega_j) \in \mathbb{R}^d$. Consider instead the formulation $\hat{s}(t_i, \omega_j) = f_j(\hat{s}_i(\omega))$, where each $f_j : \mathbb{R}^d \to \mathbb{R}$ is the Cartesian projection map defined by $f_j(v) = v_j$ for $v \in \mathbb{R}^d$. This formulation isolates the choice of the Cartesian basis for the standard spectrogram and emphasizes the role of alternative bases. In this light, our goal in this section is to determine a set of intrinsic projection maps, $\{f_j\}$, that reflect the geometry of the speech manifold.

A. Unsupervised Manifold Learning

To achieve our goal of learning the intrinsic projection maps, we can extend the above-mentioned Laplacian eigenmaps approach out-of-sample by introducing a modified variant of the unsupervised manifold regularization algorithm, presented in [5]. In the unsupervised learning setting, the algorithm input is a set of unlabeled training data, $X = x_1, \ldots, x_n \in \mathbb{R}^d$, that forms a mesh of data points that lie on the manifold. The optimization problem takes the form

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}_K} \|f\|_K^2 + \xi \mathbf{f}^T \mathbf{L} \mathbf{f}, \tag{8}$$

where $\mathcal{H}_K$ is the reproducing kernel Hilbert space (RKHS) for some positive semi-definite kernel function K, $\mathbf{L}$ is the graph Laplacian as defined in Section III.A, and $\mathbf{f} = \langle f(x_1) f(x_2) \ldots f(x_n) \rangle^T$ is the vector of values of f computed on the graph. The first term is the extrinsic norm, limiting the complexity of the solution in the ambient space. The second term is graph analogue of the intrinsic smoothness functional of (5). The single parameter ξ, then, determines the balance between extrinsic and intrinsic smoothness of the functions determined. By the RKHS representer theorem [5], the solutions of (8) can be written as

$$f^*(v) = \sum_{i=1}^{n} a_i K(x_i, v), \tag{9}$$

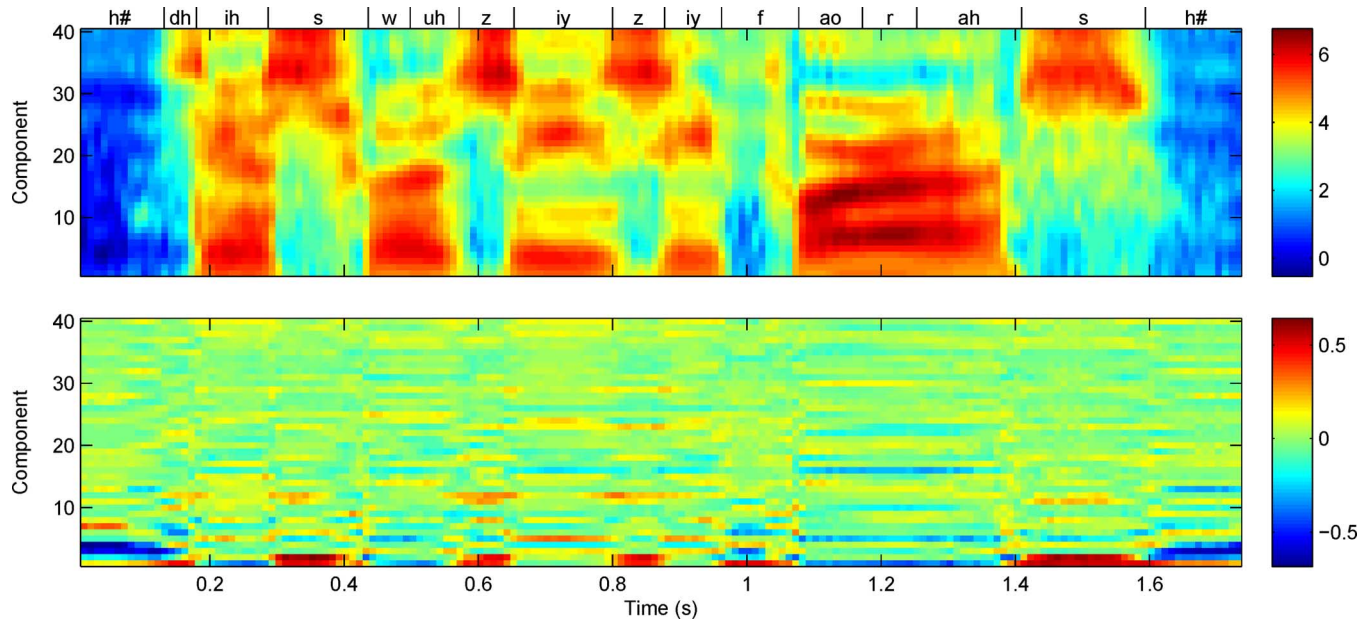

Fig. 5. Extrinsic (top) and intrinsic (bottom) spectral representations for the utterance "This was easy for us." Note that a nonlinear mel-scale frequency warping was used.

where $\{x_i\}$ are the input unlabeled data and $\mathbf{a} \in \mathbb{R}^n$ is the new parametrization of the function we need to estimate. To proceed, we plug the functional form of (9) into the optimization problem of (8). Taking the gradient with respect to the parameter vector $\mathbf{a}$ and setting it to zero sets up the following generalized eigenvalue problem:

$$(\mathbf{I} + \xi \mathbf{L}\mathbf{K})\mathbf{a} = \lambda \mathbf{K}\mathbf{a}. \tag{10}$$

Here, $\mathbf{K}$ is the $n \times n$ Gram matrix defined on the input unlabeled data by $K_{ij} = K(x_i, x_j)$. This eigenvalue decomposition will produce a full spectrum of eigenvectors, each defining its own intrinsic projection map f_j^* defined by the jth eigenvector $\mathbf{a}^{(j)}$. Unlike the unsupervised learning algorithm of [5], we are now interested in several of the $\{f_j\}$, not just one for binary classification or clustering. Recall that the intrinsic basis functions produced by the Laplacian eigenmaps algorithm were defined only on the points used to construct the graph Laplacian. Our new set of projection maps is now defined out-of-sample, i.e., f_j^* may be computed for arbitrary points on the manifold and may also be used more generally for any point in $\mathbb{R}^d$.

B. Intrinsic Spectrogram Algorithm

Given the nomenclature define above, the algorithm for computing the intrinsic spectrogram is comprised of three steps:

1) Given a set of unlabeled data $X = \{x_i\}_{i=1}^n$ sampled from the manifold, construct a κ nearest neighbor graph and compute the graph Laplacian $\mathbf{L}$ (either normalized or unnormalized).
2) Given a kernel K, solve the generalized eigenvalue problem of (10) for the weights $\{\mathbf{a}^{(j)}\}_{j=1}^{d'}$.
3) Project amplitude spectrum at each time point of the extrinsic spectrogram onto the first d' intrinsic basis functions (sorted by increasing eigenvalue) according to (9).

Note that steps 1 and 2 are computed offline using the standard training set X. Thus, converting the extrinsic spectrogram of a novel utterance into this intrinsic representation requires only the computation of Equation (9) across the utterance.

Fig. 5 shows an example extrinsic $(d = 40)$ and intrinsic $(d' = 40)$ spectrograms for the TIMIT utterance "This was easy for us" (TIMIT sentence sx3). Here, we constructed the dataset X with 200 examples of each of the 48 phonetic categories specified in [26].[2] Each example was extrinsically represented by a 40-dimensional, homomorphically smoothed, auditory (log) spectrum (40 mel scale bands, from 0–8 kHz) computed from a 25 ms signal window centered in each phonetic segment. The adjacency graph was constructed using $\kappa = 3$ nearest Euclidean neighbors and binary-valued edge weights. For the optimization problem of (8), we take as the intrinsic smoothness parameter $\xi = 1$. Finally, to accommodate nonlinear intrinsic projections maps, we employ the radial basis function (RBF) kernel, $K(x, y) = e^{-|x-y|^2/2\sigma^2}$, where σ is taken to be 1/3 of the mean Euclidean distance between the graph vertices. Note that optimal settings of κ, ξ, and σ depend on the intended application and manifold sampling density; we investigate the role this parameter in the experiments described below. Given the low-dimensional curved manifold structure motivated in previous sections, one might expect phonetic content to be more transparently differentiated in the intrinsic basis than in a traditional spectrogram. It is clear from Fig. 5 that the intrinsic representation redistributes much of the spectral variation to the lower eigenvalued components. It is also clear that these initial components do not each covary with the presence of a single speech sound. In the next section, we examine whether this alternative organization may have a natural linguistic interpretation.

V. Intrinsic Spectral Analysis Interpretation

The intrinsic representation is a projection of spectral information onto a set of basis functions ordered by their smooth-

[2]Note that while we use a class balanced sample here, balancing was not required to obtain good performance in the experiments in Section VII in which we randomly selected examples from the entire corpus (ignoring class).

ness on the adjacency graph and thus, to an approximation, their smoothness on the manifold. The question that remains is whether this new intrinsic representation has a linguistically meaningful interpretation. Before we can quantitatively address this question, we must first consider which phonological units are likely to be expressed by the intrinsic coordinates given the qualitative characteristics. The properties observed in Fig. 5 above rule out a direct phonetic interpretation, but instead are suggestive of a hierarchical or coarse-to-fine ordering. In this light, a natural linguistic framework with which to engage is the theory of distinctive features.

A. The Theory of Distinctive Features

In contrast to a traditional phonemic nomenclature, the theory of distinctive features posits that the phoneme is not the atomic building block of a language. Instead, it states that each speech sound may be more naturally represented as an array of sub-phonetic, binary-valued distinctive features. Each feature represents the minimal distinction between groups of phonemes that share some set of linguistic, articulatory, perceptual, and/or acoustic attributes. There are two main types of distinctive features: articulator-free and articulator-bound.

Articulator-free features describe high-level properties that can be used to specify the broad classes of speech sounds. In the context of physical modeling, these features will determine the details of the model architecture, including whether a turbulent or periodic source is used and whether nasal resonator coupling is introduced. For example, the articulator-free distinctive feature [son] distinguishes sonorant sounds ($[+\mathrm{son}]$), produced with a largely open vocal tract (such as vowels), from obstruent sounds ($[-\mathrm{son}]$) such as fricatives, produced with a significant enough tract constriction to generate turbulent flow. *Articulator-bound* distinctive features describe the behavior of various speech articulators during the production of speech sounds. It is this type of feature that provide a highly discretized specification of the articulatory configurations for a given physical model of the vocal tract. For example, the feature [high] distinguishes between those vowels produced with the tongue close ($[+\mathrm{high}]$) to the roof of the mouth (e.g., the /i/ in *beat*) from those that are not ($[-\mathrm{high}]$). In terms of the concatenated tube model presented in Section II.A, $[+\mathrm{high}]$ vowels will require smaller cross sectional areas for the later tube segments corresponding to the mouth region.

The articulator-free and articulator-bound categories suggest that distinctive features have an natural hierarchical structure [27]. Nodes closer to the root of the tree correspond to articulator-free distinctions that are more perceptually salient and whose acoustic correlates are less context dependent. As we progress down the hierarchy, articulator-bound features make finer distinctions between individual or small groups of phonemes.

B. Intrinsic Components as Distinctive Feature Correlates

Since distinctive features provide a binary-valued parametrization of the architecture and configuration of the physical models of speech production that are responsible for the manifold structure, it is reasonable to expect a numerical correlation between the binary feature values and intrinsic spectral components. Moreover, the intrinsic representation of Fig. 5 indicates that the hierarchical organization of the distinctive features may be manifested in the coarse-to-fine ordering of the intrinsic basis functions when sorted by eigenvalue. In order to evaluate the existence of a distinctive feature interpretation, we need to determine whether individual intrinsic components can be used to robustly infer the value of individual binary features in a relatively simple (e.g., linear) manner. We began by constructing a test dataset consisting of one example for each occurrence of each phone (across speakers and contexts) in the testing portion of the TIMIT database, for a total of 92,500 examples. As was the case for the training set described in Section IV.B, each test example was represented by a 40-dimensional discrete log auditory spectrum (40 mel scale bands, from 0–8 kHz) computed from a 25 ms signal window centered on each phonetic segment. Here, we used the same nonlinear intrinsic projection maps described in Section IV.B.

Fig. 6 shows the probability densities, as computed using uniform kernel density estimation, of several intrinsic coordinate values for various natural classes of speech sounds.[3] In each subplot, the natural class pair differs by a single distinctive feature. For example, the bottom-right subplot displays separate f_{10} densities for $[+\mathrm{anterior}]$ (/s z/) and $[-\mathrm{anterior}]$ (/ʃ ʒ/) strident fricatives. In this way, this subplot provides a visualization of the tenth intrinsic coordinate's correlation with the distinctive feature [anterior]. Also displayed in each plot is the classification equal error rate[4] (EER) of the associated classes using the single component shown. For example, using just the first intrinsic dimension f_1^*, sonorant ([+son]) phones can be distinguished from obstruent ($[-\mathrm{son}]$) phones with an accuracy of $86.2\% = 100 - 13.8\%$.

Another important property to notice is that for several cases in Fig. 6, the optimal decision threshold on the intrinsic component is at or near zero. In these cases, the intrinsic coordinate value is not only correlated with the expression of the given distinctive feature, but also provides a direct estimate of the feature value itself. This fact is particularly striking when you recall these components were computed in a completely unsupervised fashion. Now, it should be noted that there can exist individual extrinsic components (i.e., frequency bands) that can discriminate some of these natural classes reasonably well. However, these bands are not ordered according to the any phonological hierarchy and they can at best be correlated with feature values, as the optimal decision threshold will not be located at the origin.

Fig. 7 shows the probability densities of several intrinsic coordinate values, computed from vowel data alone, for various natural classes of vowel sounds. As above, each natural class pair differs by a single distinctive feature. We find that the first three intrinsic components are strongly correlated with height, backness, and tenseness, the three principal distinctive features that distinguish English vowels. Vowel height refers to the vertical position of the tongue within the mouth, and can be seen as correlate to the cross sectional area R_A ratio of Section II.B. Similarly, vowel backness refers to the axial (horizontal) position relative to the rear of the mouth. Thus, the first two intrinsic

[3]The examples shown were chosen after limited manual inspection; certainly other phonological correlations exist and could be easily uncovered with an automated analysis.

[4]The threshold was chosen to equalize the error rates for each class.

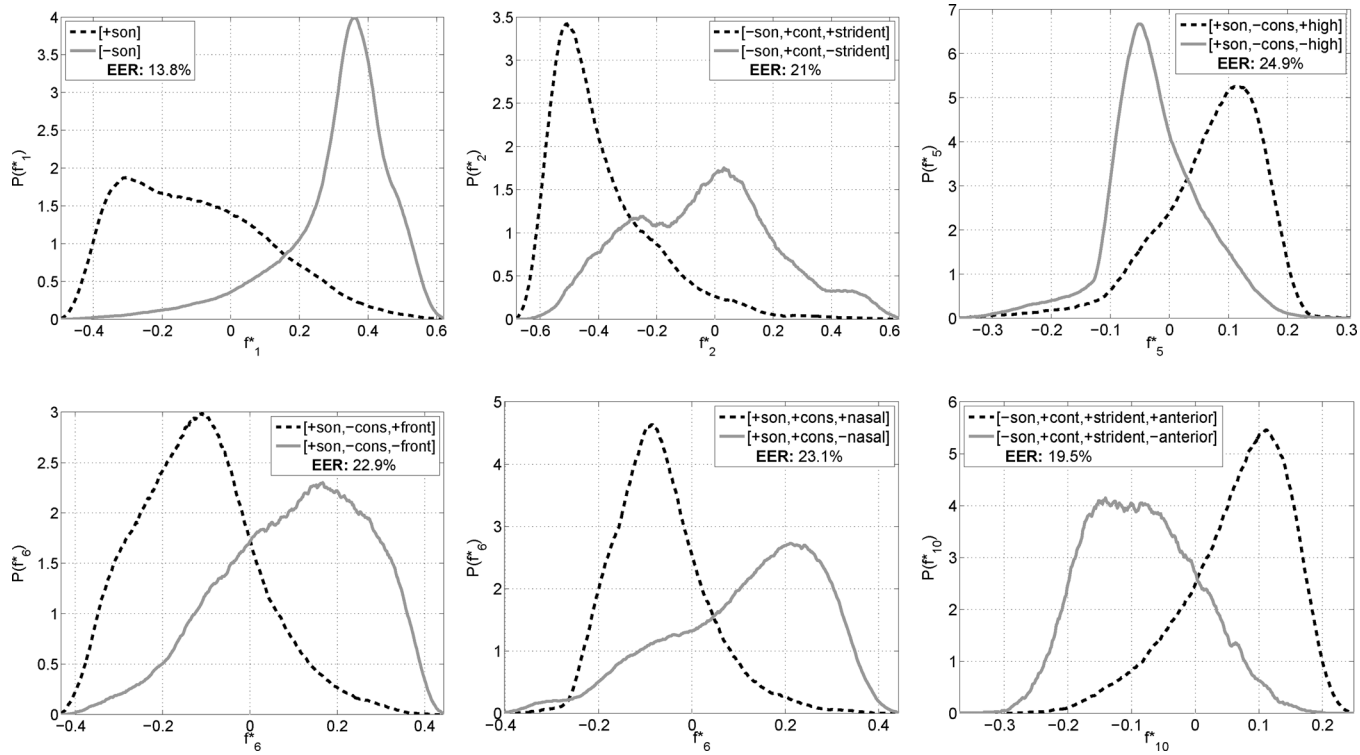

Fig. 6. Estimated probability densities $P(f_j^*)$ of various intrinsic coordinates for various natural classes of phonemes. Each plot shows the densities for two classes that differ by a single distinctive feature. The binary classification equal error rate (EER) using the single coordinate is shown in the legend.

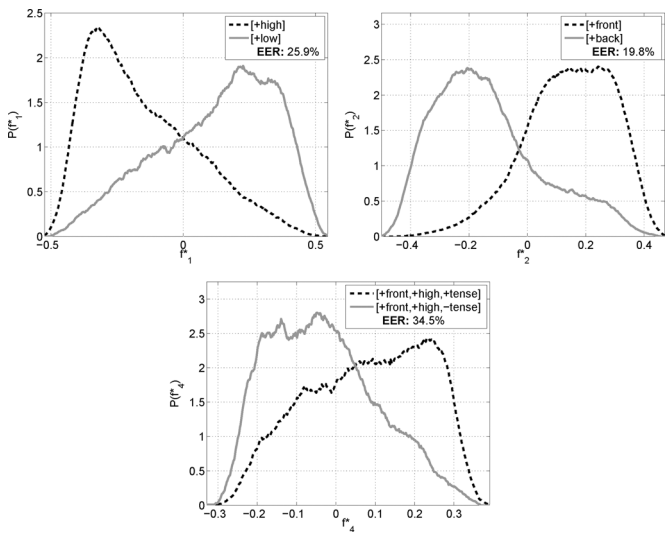

Fig. 7. Estimated probability densities $P(f_j^*)$ of various intrinsic coordinates for various natural classes of vowel phonemes. Here, the adjacency graph is constructed using vowel examples only, providing a more local and specialized intrinsic coordinate system.

coordinates correspond to the physical location of the tongue during vowel production, indicating ISA has performed an effective inversion of the production mechanism. Note that in the global intrinsic representation, these features were instead correlated with the higher coordinates, while in the vowel centric coordinates they are the dominant dimensions of variability.

VI. Phonetic Separability

In Fig. 4, we saw that the eigenfunctions of the graph Laplacian operator can approximate a coordinate chart on the manifold, recovering a representation corresponding to the underlying parameter space. Moreover, in Section V, we found single intrinsic components that, when appropriately thresholded, can discriminate between natural classes of speech sounds motivated by distinctive feature theory. These results suggest that the intrinsic spectrogram may provide an acoustic representation that simplifies the geometric organization of individual speech sounds as realized across speakers. In order to quantitatively evaluate this intuition, we consider the linear separability of phonetic categories of the extrinsic and intrinsic spectrogram representations.

The test dataset defined in Section V.B, consisting of $\sim$ 1000–2000 examples of each of the 48 phonetic classes defined by [26], was used to set up $\binom{48}{2} = 1128$ binary phone classification problems. For each phone pair, we use (supervised) linear discriminant analysis to compute the optimal separating hyperplane between the examples of each class, using both extrinsic and intrinsic representations. To compute the intrinsic representation, we used the same basis functions specified in Sections IV.B and V.B. Fig. 8 displays the equal training set error rate of the optimal separating hyperplanes as a measure of inherent linear separability for the intrinsic versus the extrinsic representation for the 1128 binary classification problems. Those points falling below the equal performance line (shown in black) correspond to phone pairs where the intrinsic representation improved linear separability over the traditional log mel spectrogram.

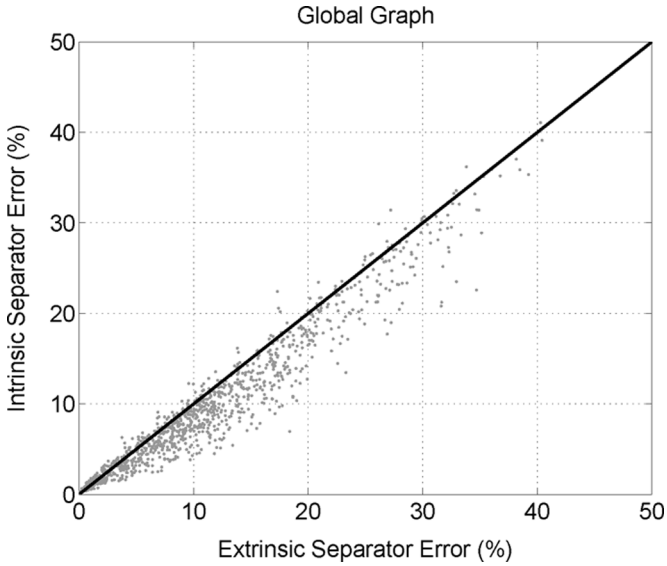

Fig. 8. Optimal separating hyperplane error rate of the extrinsic versus the *global* intrinsic representation for the 1128 binary phone classification problems. The equal performance line is shown in black. Points above/below the line indicate reduced/improved linear separability with ISA for the corresponding phone pair. The large majority of points are below the line.

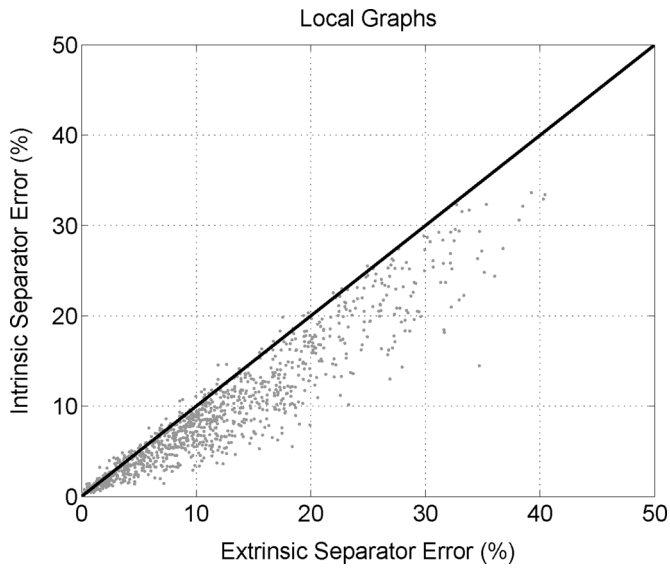

Fig. 9. Optimal separating hyperplane error rate of the extrinsic versus the *local* intrinsic representation for the 1128 binary phone classification problems. The equal performance line is shown in black. Points above/below the line indicate reduced/improved linear separability with ISA for the corresponding phone pair. Points fall further below the line on average than in Fig. 8.

Averaged over the 1128 binary classification problems, the intrinsic spectral representation leads to an 11.5% relative reduction in the EER of the optimal linear separator. For the phone pairs helped most by ISA, we observe as much as 50% relative reduction. Improved linear separability suggests that acoustic models for speech recognition applications may succeed with reduced model complexity and less training data. In the case of Gaussian mixture monophone or triphone model, improved linear separability may reduce the number of mixture components required. For example, the experiments performed in [28] applied locality preserving projections, a *linear* intrinsic projection technique, to a continuous speech recognition task and observed a reduction in word error rate. The improved class separability facilitated by our *nonlinear* projection maps promise even more significant gains, an item we will explore in Section VII.

In Section V.B, we considered a local intrinsic representation for the vowel region of the speech manifold and found it provided a set of components corresponding to distinctive features of vowel height and backness. We generalized this notion of a local intrinsic representation to the case of binary classification. In particular, using just the data for each phone pair, we constructed separate adjacency graphs and computed a set of local intrinsic basis functions for each using the techniques introduced in Section IV. Then, for each of the 1128 binary classification problems, we project the extrinsic spectra onto the specialized intrinsic representation for that phone pair. Fig. 9 shows the extrinsic vs. intrinsic performance of the 1128 classification problems. These specialized intrinsic representations produce an average relative reduction of 19% in linear separator error rate (and as much as 80%), significantly better than that observed when using the global intrinsic representation. We will return to this in Section VIII.

VII. Speech Recognition Experiments

As demonstrated in Section VI, the increased linear separability of phoneme classes when using ISA indicates that improved phone recognition may be possible with reduced model complexity. In this section, we summarize the results of a recent study [29] evaluating nonlinear ISA on two mainstream speech recognition problems at different extremes of model complexity: (i) the zero resource speech recognition task of unsupervised term discovery, where there is no transcribed training materials, and (ii) the high resource speech recognition task of phone recognition using neural networks with an ample hidden layer and plenty of training data.

A. Unsupervised Term Discovery

As massive internet collections of untranscribed speech audio emerge, there is a clear opportunity for unsupervised speech technologies. Unsupervised spoken term discovery is the task of searching untranscribed speech collections for repeated words and phrases without using any language specific knowledge. The prevailing approaches [30]–[32] rely on exhaustive dynamic time warping-based searches across all pairs of speech intervals drawn from the collection. While a computationally challenging endeavor, recent efficient approximation strategies [33] have made processing hundreds of hours of speech tractable. At the core of the task is the inherent speaker independence of the representation. Indeed, the ability to associate acoustic realizations of a given phoneme or word spoken by different speaker is exactly what acoustic model supervision is meant to accomplish. In [34], it was found that signal processing-derived front-ends like perceptual linear prediction (PLP) and mel frequency cepstral coefficients (MFCC) were of comparable utility, both falling far short of supervised representations like matched-language posteriorgrams. We have already demonstrated the ability of ISA to

TABLE I
SPOKEN TERM DISCOVERY EVALUATION RESULTS (IN % AVERAGE PRECISION). DISTANCE METRICS LISTED ARE THOSE USED IN NEAREST NEIGHBOR GRAPH CONSTRUCTION

Features	Dimension	AP
Mel Spectrogram	40	10.7
Log Mel Spectrogram	40	7.6
Linear ISA (cosine)	39	25.6
PLP	39	34.8
MFCC	39	33.8
Nonlinear ISA (Euclidean)	39	38.6
Nonlinear ISA (cosine)	39	**48.5**
English Posteriorgram	40	75.4

unravel complicated speaker and message dependencies in the observed transfer functions (e.g., Figs. 3 and 4) and to improve phonetic separability. The next question is whether we would have similar luck on a fully unsupervised application.

We sought to answer this question with the evaluation defined in [34], which we adapted to the TIMIT corpus. The evaluation protocol is as follows. First, using the time aligned word transcriptions, we extracted all word examples that were at least 0.35 s in duration and at least 6 characters as text, yielding 11,000 word examples across 3745 types. Second, we computed all pairwise dynamic time warping (DTW) similarities (one minus the DTW distance normalized by the sum of the two example durations in frames), involving some 60 million DTW distance calculations (completes in approximately 10 minutes on a cluster of 100 CPUs). Finally, treating DTW similarity as a same/different classifier score, we computed the average precision (AP) for the task of separating same word type pairs from different word type pairs. Due to its proven success for this task [34], we considered cosine distance as the DTW frame level distance metric for all features evaluated. Note that this choice removes any effect of principal component analysis (PCA) as cosine distance is invariant under rotations.

Our definition of ISA involves three parameters that must be chosen by the user: κ, the number of nearest neighbors for graph construction; ξ, the weight on the intrinsic norm in the optimization problem; and σ, the width of the RBF kernel (for nonlinear ISA only). Ideally, the dependence of downstream performance on the values of these parameters would be weak, precluding the need for labeled development data. In practice, we found $\kappa = 10, \xi = 0.03$, and $\sigma = 0.4m$, where m is the mean distance between samples in X, to be a good universal choice across the experiments. Performance was moderately dependent on κ and σ; varying κ from 4 to 12 and σ from $0.1m$ to $1.0m$ resulted in at most 4% loss in average precision.

Table I lists this average precision (AP) for several front-ends. All features were computed in 25 ms windows sampled every 10 ms. The mel-scale spectrogram used 40 mel channels and the PLP used a 12th-order linear predictive coding-smoothed, bark-scale spectrograms. Thirteen cepstral coefficients (including constant zeroth order component) were used for both PLP and MFCC. For the spectrograms, PLP, and MFCC, each feature dimension was normalized to zero mean and unit variance. Intrinsic spectrograms were derived from the log mel spectrogram, where n = 10,000 randomly selected samples (from the whole corpus for a class-unbalanced sample) were used to construct the graph Laplacian. We kept only the first 13 intrinsic components (skipping the first trivial dimension) to maintain the same feature dimension as used for PLP and MFCC. Velocity (Δ) and acceleration ($\Delta\Delta$) features were included for all of these features. Finally, to set a supervised performance ceiling, we also considered English posteriorgram features generated using a multi-layer perceptron with a single hidden layer of size 2000 nodes trained on all TIMIT sx/si sentences (PLP with 9-frame context used as input).

The primary result is striking: nonlinear ISA substantially improved representational speaker independence relative to both PLP and MFCC features, the speech recognition mainstays. The improvement in cross-speaker word matching average precision is nearly 15% absolute (more than 40% relative), representing a third of the gap between the original PLP/MFCC and the *supervised* performance ceiling. This was accomplished without any supervision whatsoever and relied only on the manifold assumption considered in this study. While linear ISA improves upon the log mel spectrogram, it falls short of standard features like PLP and MFCC. Nonlinear ISA is substantially better than the linear version, further substantiating the claim that curved manifolds require nonlinear kernels to recover an optimal embedding.

B. Supervised Phone Recognition

Next, we considered the more familiar task of supervised phone recognition on the TIMIT corpus, employing a state-of-the-art hidden Markov model/multi-layer perceptron (HMM/MLP) back-end to evaluate ISA against traditional front-ends. In particular, we used the hierarchical MLP system defined in [35], previously demonstrated capable of producing highly accurate recognition in an almost completely bottom-up fashion (no top down grammatical constraints beyond a simple bigram phonetic language model). The hierarchical architecture included two levels of neural networks. The first estimated posterior probabilities of three tri-state classes for each of the 48 speech sounds defined in [26]. Here, 3-layer MLPs using a 9 frame context was used. The second neural network took as input the tri-state posteriors over a large 23-frame context and estimated a new set of tri-state posteriors. Finally, these tri-state posteriors were used to generate scaled likelihoods that provide state emission likelihoods for an HMM decoder with a bigram phonetic language model. The 48 phone classes were reduced to the standard 39 unit set [26] for scoring. We also considered the combination of feature types, which we performed on the final output posteriorgram of each constituent stream using the Dempster-Shafer (DS) theory of evidence [36]. DS combination employs an inverse entropy weighting, with the rationale that high entropy frames indicate an acoustic model that is uncertain of which phonetic class is occurring. Thus, by downweighting the high entropy frames in the combination, the combined output will rely more on models that are more confident about the phonetic identity. While this combination strategy has been proven successful for this setting, several others are explored in [36].

Table II lists the phone recognition accuracies using PLP, nonlinear ISA (RBF kernel), and the Dempster-Schafer combination of the two. Note that for ISA, we determined suitable parameter values using cross-validation ($\kappa = 6, \xi = 0.03, \sigma = 1.0m$) on the training corpus.

TABLE II
SUPERVISED PHONE RECOGNITION RESULTS (IN % PHONE RECOGNITION ACCURACY; DS = Dempster − Shafer)

Features	Accuracy
PLP	77.0
Nonlinear ISA	76.0
DS: PLP + Nonlinear ISA	**78.5**

Nonlinear ISA falls only one point behind traditional PLP, indicating that the nonlinear embedding learned maintains sufficient information to bear the full representational burden for downstream phone classification. Moreover, the improvement observed with DS combination indicates that the manifold features encode information that is complementary to the standard PLP front-end, reaching state-of-the-art phone recognition performance for this given back-end architecture.

Thus, we conclude that without supervision of any kind, intrinsic spectral analysis is able to recover a nonlinear transformation that, when applied to the mel spectrogram, produces features as useful for subsequent supervised recognition as linear prediction and cepstral analysis. This is a remarkable result, since once we construct the nearest neighbor graph with binary weights, all absolute notions of locality have been obscured to the optimization. It follows that the topological structure alone of our unlabeled data sample is sufficient to recover a representation that encodes the necessary information for successful recognition downstream.

VIII. DISCUSSION

The experimental results above indicate that ISA provides a suitable alternative to standard front-ends like MFCC and PLP for both unsupervised and supervised applications. This claim is substantiated by greatly improved representational speaker independence, as demonstrated in the experiments described in Section VII.A, while maintaining the information necessary to enable state-of-the-art supervised phonetic recognition, as demonstrated in Section VII.B. However, ISA is not without its limitations. The main computational burden of ISA is in solving the generalized eigenvalue problem of (10). While eigensolver complexity depends on the sparsity and topology of the nearest neighbor graph used to construct the Laplacian, runtime for exact methods is at least quadratic in the number of graph vertices. This means that there are practical limitations on the manifold sampling density and, as a consequence, limitations on the degree of curvature that may be unraveled by the projection maps. Randomized approximation methods may provide a straightforward way around this [37]–[39], but application of speech specific strategies may also be explored.

One such possibility for speech is to construct multiple local coordinate systems for various distinct regions of the speech manifold in the manner explored in Sections V.B and VI. For example, suppose we are provided a collection of points on the manifold that are known to be vowels. Then, we can use this data to construct a vowel-specific intrinsic coordinate system that can discern finer structure in this local region of the manifold. However, constructing the specialized adjacency graphs requires supervision, but broad class distinctions are less language dependent and can be more easily ported from high to low resource languages.

Since the projection maps need only be estimated once offline, the only runtime computational burden is the calculation of (9) for each frame in the test utterance. However, for a large number n of training samples, the use of an RBF kernel requires a nontrivial amount of additional processing. Here, approximate nearest neighbor techniques are the clear solution (see [33] for an example), capable of reducing the number of kernel function evaluations from n to $O(\log n)$.

In the realm of manifold learning, signal noise has the potential of introducing substantial complications. First, in the presence of noise, the manifold structure can be obscured, counteracting the benefits that this class of nonlinear transformations can provide. Moreover, if the training data sample has heterogeneous noise conditions, disconnected noise-specific subgraphs can emerge, resulting in separate ISA subspaces for each condition. Finally if the projection maps are estimated on clean data and evaluated in noisy conditions, there is a potential for nonlinear amplification of the distortion. Despite all of these potential stumbling blocks, there is a silver lining. As mentioned above, when using binary edge weights in the nearest neighbor graph, the ISA optimization produces a new embedding of the data that depends only on the topology of the graph and not on the absolute distances in the original space. This means that the only property that need be preserved across noise conditions to achieve noise invariance would be this topological structure. Coupled to a suitable graph alignment procedure [40], this fact may provide a promising means to adapt to changing noise conditions.

Finally, while the discrete cosine transform that converts log power spectra to cepstral coefficients is parameter-free, ISA introduces several parameters that need to be optimally chosen. In Section VII we observed a modest performance dependence on the various parameters, but we do remain above PLP and MFCC performance over a wide range of values. Still, automatic parameter selection remains an important research goal for this and other manifold learning methods.

IX. CONCLUSION

We have argued that speech sounds form a low-dimensional, extrinsically curved manifold and have visualized synthetic speech data to support this claim. With this motivation, we presented the Intrinsic Spectral Analysis algorithm to recover a set of intrinsic projections maps that define a novel spectral embedding for downstream application. We found that individual intrinsic components are correlated with the expression of various distinctive features. When extending the intrinsic representation out-of-sample using a nonlinear kernel-based approximation, we found an improved linear separability of the phonetic classes. Moreover, we demonstrated that ISA admits dramatically improved representational speaker independence compared with standard front-ends and supports state-of-the-art supervised phone recognition. This justifies their standing as a superior all-purpose speech representation. While we have focused our attention in this paper on speech recognition applications, ISA may be applied in any domain where traditional Fourier analysis is useful and there is reason to believe the space of possible signals is naturally constrained.

ACKNOWLEDGMENT

The authors would like to thank Samuel Thomas of the Center for Language and Speech Processing at Johns Hopkins University for his assistance in running experiments in Section VII.B. We also thank three anonymous reviewers for the very detailed and helpful comments, helping us to improve the readability and technical soundness of the manuscript.

Author P. Niyogi, deceased, contributed to all aspects of the original manuscript, which consisted of Sections I–VI. The experiments described in Section VII and revisions during peer review were performed after his passing in 2010.

REFERENCES

[1] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.
[2] D. L. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," in *Proc. NAS*, 2003, vol. 100, pp. 5591–5596.
[3] S. T. Roweis and L. K. Saul, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," *J. Mach. Learn. Res.*, vol. 4, pp. 119–155, 2003.
[4] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data represenation," *Neural Comput.*, vol. 16, pp. 1373–1396, 2003.
[5] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, 2006.
[6] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, "Diffusion maps, spectral clustering and reaction coordinates of dynamical systems," *J. Appl. Comput. Harmon. Anal.*, vol. 21, pp. 113–127, 2006.
[7] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.
[8] Z. Zhang and Z. Hongyuan, "Principal manifolds and nonlinear dimension reduction via local tangent space alignment," *SIAM J. Scientif. Comput.*, vol. 26, no. 1, pp. 313–338, 2005.
[9] R. Togneri, M. D. Alder, and Y. Attikiouzel, "Dimension and structure of the speech space," in *Proc. IEE Conf. Commun., Speech, Vis.*, 1992.
[10] G. Fant, *Acoustic Theory of Speech Production*. Paris, France: Mouton, 1970.
[11] K. N. Stevens, *Acoustic Phoenetics*. Cambridge, MA, USA: MIT Press, 1998.
[12] A. Jansen and P. Niyogi, "A geometric perspective on speech sounds," U. of Chicago, Chicago, IL, USA, Tech. Rep. TR-2005-08, Jun. 2005.
[13] A. Jansen and P. Niyogi, "Intrinsic Fourier analysis on the manifold of speech sounds," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2006, pp. 241–244.
[14] R. Jakobson, G. Fant, and M. Halle, *Preliminaries to Speech Analysis: The Distinctive Features*. Cambridge, MA, USA: MIT Press, 1952.
[15] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York, NY, USA: Harper & Row, 1968.
[16] M. Halle, , D. Osherson and H. Lasnik, Eds., "Phonology," in *Language Volume 1*. Cambridge, MA, USA: MIT Press, 1990, pp. 43–68.
[17] G. A. Miller and P. Nicely, "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Amer.*, vol. 27, no. 2, pp. 338–352, 1955.
[18] A. Lovitt, J. Pinto, and H. Hermansky, "On confusions in a phoneme recognizer," Idiap Res. Inst., Martigny, Switzerland, IDIAP Res. Rep. IDIAP-RR-07-10, 2007.
[19] A. Jansen, "Spgen v1.0," 2005 [Online]. Available: http://www.clsp.jhu.edu/ajansen/research.html
[20] E. Bresch, Y.-C. Kim, K. S. Nayak, D. Bryd, and S. S. Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging," *IEEE Signal Process. Mag.*, vol. 25, no. 3, pp. 123–132, 2008.
[21] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, vol. 14, pp. 585–591.
[22] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, vol. 16, pp. 153–160.
[23] A. Jansen and P. Niyogi, "Semi-supervised learning of speech sounds," presented at the Interspeech, Antwerp, Belgium, Aug. 28, 2007.
[24] L. ten Bosch, A. Hamalainen, and M. Ernestus, "Assessing acoustic reduction: Exploiting local structure in speech," in *Proc. Interspeech*, 2011, pp. 2665–2668.
[25] H. Strange and R. Zwiggelaar, "A generalized solution to the out-of-sample extension problem in manifold learning," in *Proc. AAAI*, 2011, pp. 471–476.
[26] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 11, pp. 1641–1648, 1989.
[27] J. J. McCarthy, "Feature geometry and dependency: A review," *Phonetica*, vol. 43, pp. 84–108, 1988.
[28] Y. Tang and R. C. Rose, "A study of using locality preserving projections for feature extraction in speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pp. 1569–1572.
[29] A. Jansen, S. Thomas, and H. Hermansky, "Intrinsic spectral analysis for zero and high resource speech recognition," presented at the Interspeech, Portland, OR, USA, 2012.
[30] A. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 186–197, 2008.
[31] A. Muscariello, G. Gravier, and F. Bimbot, "Audio keyword extraction by unsupervised word discovery," presented at the Interspeech, Brighton, U.K., Sep. 2009.
[32] A. Jansen, K. Church, and H. Hermansky, "Towards spoken term discovery at scale with zero resources," presented at the Interspeech, Makuhari, Japan, Sep. 2010.
[33] A. Jansen and B. V. Durme, "Efficient spoken term discovery using randomized algorithms," in *Proc. ASRU*, 2011, pp. 401–406.
[34] M. Carlin, S. Thomas, A. Jansen, and H. Hermansky, "Rapid evaluation of speech representations for spoken term discovery," presented at the Interspeech, Florence, Italy, Sep. 2011.
[35] J. Pinto, B. Yegnanarayana, H. Hermansky, and M. Magimai-Doss, "Exploiting contextual information for improved phoneme recognition," presented at the Interspeech, Antwerp, Belgium, Aug. 2007.
[36] F. Valente and H. Hermansky, "Combination of acoustic classifiers based on Dempster-Shafer theory of evidence," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2007, pp. 1129–1132.
[37] V. Rokhlin, A. Szlam, and M. Tygert, "A randomized algorithm for principal component analysis," *SIAM J. Matrix Anal. Appl.*, vol. 31, no. 3, pp. 1100–1124, 2009.
[38] F. Tompkins and P. J. Wolfe, "Approximate intrinsic Fourier analysis of speech," presented at the Interspeech, Brighton, U.K., Sep. 2009.
[39] J. Sun, Y. Xie, H. Zhang, and C. Faloutsos, "Less is more: Compact matrix decomposition for large space graphs," presented at the SDM, Minneapolis, MN, USA, Apr. 2007.
[40] C. Priebe, D. Marchette, Z. Ma, and S. Adali, "Manifold matching: Joint optimization of fidelity and commensurability," *Brazilian J. Probab. Statist.*, 2012.

Aren Jansen (M'13) received the B.A. degree in physics from Cornell University, Ithaca, NY, USA, in 2001 and the M.S. degree in physics as well as the M.S. and Ph.D. degrees in computer science from the University of Chicago, Chicago, IL, USA, in 2003, 2005, and 2008, respectively.

He is currently a Research Scientist at the Human Language Technology Center of Excellence and an Assistant Research Professor in the Department of Electrical and Computer Engineering, both at The Johns Hopkins University, Baltimore, MD, USA. His research explores various aspects of the speech recognition problem, with a focus on whole word acoustic modeling, sparse representations and models, and unsupervised/semi-supervised learning of words and speech sounds.

Partha Niyogi received the B.Tech. degree from the Indian Institute of Technology (IIT), Delhi, India, and the S.M. degree and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA.

He was the Louis Block Professor in Computer Science and Statistics at The University of Chicago. Before joining the University of Chicago, he worked at Bell Laboratories as a Member of the Technical Staff for several years. His research interests were in pattern recognition and machine learning problems that arise in the computational study of speech and language.

Dr. Niyogi died in 2010. During his short career, he made major contributions to a broad range of theoretical and applied problems in statistical learning, language acquisition and evolution, speech recognition, and computer vision.