INTRINSIC FOURIER ANALYSIS ON THE MANIFOLD OF SPEECH SOUNDS

Aren Jansen

The University of Chicago
Dept. of Computer Science
Chicago, IL 60637

Partha Niyogi

The University of Chicago
Depts. of Computer Science and Statistics
Chicago, IL 60637

ABSTRACT

Recently, there has been much interest in geometrically motivated dimensionality reduction algorithms. These algorithms exploit low-dimensional manifold structure in certain natural datasets to reduce dimensionality while preserving categorical content. This paper has two goals: (i) to motivate the existence of a low-dimensional curved manifold structure to voiced speech sounds, and (ii) to present a new intrinsic (manifold-based) spectrogram technique founded on the existence this manifold structure. We find that the intrinsic representation allows phonetic distinction in fewer dimensions than required by a traditional spectrogram.

1. INTRODUCTION

Let $\mathcal{M}$ denote the set of possible vocal tract transfer functions that correspond to the articulatory configurations used in speech production. Since each $g(\omega) \in \mathcal{M}$ is a function in $\mathcal{L}^2$, it follows that $\mathcal{M} \subset \mathcal{L}^2$ is a low-dimensional submanifold if the space of articulatory configurations is also low-dimensional (*i.e.*, the articulators have few degrees of freedom). A speech signal $x(t)$ has a corresponding trajectory in the space of articulatory configurations, and therefore also has a corresponding trajectory $p(t)$ on the manifold $\mathcal{M}$. A traditional spectrogram is obtained by projection maps $f_\omega : \mathcal{M} \rightarrow \mathbb{R}$ applied to the trajectory, namely $f_\omega[p(t)]$.

Since $\mathcal{M}$ is a submanifold of $\mathcal{L}^2$, it inherits a Riemannian structure from $\mathcal{L}^2$ and is, as we will see, non-Euclidean. Therefore, a natural intrinsic basis exists on this manifold (from harmonic analysis), giving rise to a corresponding set of projection maps $f_j : \mathcal{M} \rightarrow \mathbb{R}$. These $\{f_j\}$ are the eigenfunctions of the Laplacian operator on the manifold. Thus, the speech trajectory can be represented under this projection, $f_j[p(t)]$, giving rise to an alternative *intrinsic* spectrogram that is faithful to the geometry of speech sounds.

Our development of intrinsic Fourier analysis for speech sounds has two parts. First, we must motivate the manifold assumption for speech if the distinct intrinsic coordinate system is to exist at all. Second, in the absence of an analytical coordinate chart for the manifold, we must provide an algorithmic means for approximating the set of projection maps for intrinsic coordinate systems.

The articulatory parameterizations of phoneme production developed by Fant and others [1, 2] strongly indicate the existence of a low-dimensional manifold structure to certain classes of speech. To formally motivate this, we present a brief derivation of sounds generated by series of concatenated tubes. This system serves as a simple model for the vocal tract proven useful in the tradition of acoustic modelling of voiced phoneme production. We formally demonstrate that the set of transfer functions for such finitely parameterized acoustic tube filters forms a low-dimensional curved manifold that spans a higher dimensional space.

We continue by presenting a new method for approximating the intrinsic coordinate system derived from the graph Laplacian spectral clustering methods of [3]. Our method provides the means to cast a traditional extrinsic spectrogram onto the intrinsic basis of the manifold. We conclude with a preliminary example of using this method and its phonetic interpretation.

2. 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 [2]: (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 axis of the tube.

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

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

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,

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

is the approximate form of the radiation impedance given by Stevens [2] (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 is the speed of sound, $A = \pi a^2$ is the area of the piston, and $k = \omega/c$ is the wave number. The term K_s is a real-valued frequency-dependent factor (see [2] or [4] for details).

Now consider generalized the 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 Equation 1 with the boundary conditions of Equation 2 takes the form $\hat{p}(\omega) = \hat{s}(\omega)g(\omega, \{L_i\}, \{A_i\})$, where

$$g(\omega, \{L_i\}, \{A_i\}) = \frac{\mathbf{Z}_r(\omega)}{M_{22} - \mathbf{Z}_r(\omega)M_{12}}. \qquad (3)$$

is the transfer function for the entire N-tube filter. Here, $M = \prod_{i=1}^{N} C_i$ with

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

3. THE SPEECH MANIFOLD

Consider the single tube acoustic tube filter with length L and cross-sectional area A. The transfer function g, which determines the acoustic signal, is given by Equation 3. 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^{h} be the range of human vocal tract lengths. Then, the set $\mathcal{M}_1$ has the following properties (for details, see [4]):

1. The map $\phi : (L, A) \in \mathbb{R}^2 \rightarrow \mathcal{M}_1$ defined by g is a diffeomorphism for $L \in I_L^{\text{h}}$.
2. There exists $l_1, l_2 \in I_L^{\text{h}}$ at which the tangent vectors $\partial g_1/\partial L$ are linearly independent.

It follows from Item 1 that ϕ^{-1} is a coordinate chart on the set $\mathcal{M}_1$. Therefore, $\mathcal{M}_1$ is formally a smooth two-dimensional manifold embedded in the ambient space $\mathcal{L}^2$. Furthermore, Item 2 implies the existence of extrinsic manifold curvature. Finally, the manifold $\mathcal{M}_1$ spans a subspace of $\mathcal{L}^2$ that is much larger than the dimension of the manifold.

We saw in Equation 3 that the N-tube transfer function solutions involve one matrix multiplication per tube segment. It follows that solution complexity monotonically increases with N. Therefore, the curvature and spanning properties described above for the single tube case will apply for the N-tube generalization as well. In general, the dimension of the N-tube solution manifold is 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 in the model. It follows that an approximate low-dimensional manifold structure exists for the class of voiced speech sounds.

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 additional 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, individual obstruent phonemes still cluster naturally, and the algorithm that we develop in the following section still applies under a clustering interpretation. Developing this theory lies outside the scope of this paper.

4. AN INTRINSIC SPECTROGRAM REPRESENTATION

A traditional spectrogram is the short time Fourier spectrum of an audio signal. The 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 H-dimensional discrete Fourier transform. The spectrogram is then given by $\hat{s}(t_i, \omega_j) = \hat{s}_i(\omega_j) \in \mathbb{R}^H$.

Consider instead the formulation $\hat{s}(t_i, \omega_j) = f_j(\hat{s}_i(\omega))$, where each $f_j : \mathbb{R}^H \rightarrow \mathbb{R}$ is the Cartesian projection map defined by $f_j(v) = v_j$ for $v \in \mathbb{R}^H$. 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 projection maps,

$\{f_j\}$, that reflects the intrinsic geometry of the speech manifold.

4.1. The Laplacian and graph Laplacian operators

The Laplacian operator on a Riemannian manifold $\mathcal{M}$ is the second order differential operator typically denoted $\Delta_{\mathcal{M}}$. It is a positive semidefinite operator whose eigenfunctions form an orthogonal basis for $\mathcal{L}^2(\mathcal{M})$. If $\{\lambda_i\}$ and $\{e_i\}$ are the sorted eigenvalues and corresponding eigenfunctions of the Laplacian, 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 a natural basis, the Laplacian operator also provides a measure of smoothness for functions. Given a measure μ on $\mathcal{L}^2(\mathcal{M})$, the functional

$$S[f] = \int_{\mathcal{M}} \|\nabla_{\mathcal{M}} f\|^2 d\mu = \langle \Delta_{\mathcal{M}} f, f \rangle_{\mathcal{L}^2(\mathcal{M})} \qquad (4)$$

increases as smoothness of f decreases [5]. 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. Furthermore, each $e_i : \mathcal{M} \to \mathbb{R}$ varies smoothly with *geodesic* distance on $\mathcal{M}$ and is therefore faithful to the geometry of the manifold.

In practice, we are not given the precise form of the manifold $\mathcal{M}$, so the Laplacian operator cannot be used directly. Instead, we must implement the graph theory analogue as follows: Consider a manifold $\mathcal{M}$ embedded in $\mathbb{R}^H$ and N data points $x_1, \ldots x_N \in \mathcal{M}$. We can construct an adjacency graph with one vertex V_i per data point x_i. We connect vertices V_i and V_j with an edge of weight one if x_i is one of the n nearest neighbors of x_j or x_j is one of the n nearest neighbors of x_i. This graph can be represented by the adjacency matrix W, which is symmetric and binary-valued. From this, we can determine the so-called graph Laplacian, $L = W - D$, where D is the diagonal matrix with elements $D_{ii} = \sum_j W_{ji}$.

The graph Laplacian is a positive semidefinite $N \times N$ matrix that satisfies all the properties given above for the continuous Laplacian operator [5]. However, there are two main differences in the graph analogue. First, we are now limited to functions that are defined on the graph, not the entire manifold. Second, the $\mathcal{L}^2$ inner product is now replaced by an $\mathbb{R}^H$ inner product. The recent Laplacian eigenmaps dimensionality reduction algorithm [5] proceeds by solving the eigenvalue problem $L\mathbf{e}_i = \lambda_i \mathbf{e}_i$, and projecting the points $\{x_i\}$ onto $m \leq H$ lowest eigenvectors according to $P_m(x_i) = (\mathbf{e}_2(i), \ldots, \mathbf{e}_m(i))$. However, this method does not allow out-of-sample extension as the projection is only defined on the graph.

4.2. Unsupervised manifold learning

To achieve our goal of learning the intrinsic projection maps, we can extend the Laplacian eigenmap approach out-of-sample by using a modified form of an unsupervised manifold regularization algorithm, presented in [3]. In the unsupervised learning setting, the algorithm input is a set of unlabeled training data, $x_1, \ldots, x_N \in \mathbb{R}^H$, that forms a mesh of data points that lie on the manifold. The optimization problem takes the form

$$f^* = \arg \min_{f \in \mathcal{H}_K} \|f\|_K^2 + \xi \mathbf{f}^T L \mathbf{f}, \qquad (5)$$

where $\mathcal{H}_K$ is the reproducing kernel Hilbert space (RKHS) for the kernel K, L is the graph Laplacian as defined in Section 4.1, and $\mathbf{f} = \langle f(x_i), \ldots, f(x_N) \rangle^T$ is the vector of values of f 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 smoothness functional of Equation 4. The single parameter ξ, then, determines the intrinsic smoothness of the functions determined. By the RKHS representer theorem [3], the j-th component of our new projection map is then given by

$$f_j^*(v) = \sum_{i=1}^{N} \alpha_i^j K(x_i, v), \qquad (6)$$

where $\{x_i\}$ are the input unlabeled data, and $\alpha^j \in \mathbb{R}^N$ is the j-th eigenvector (sorted by eigenvalue) to the generalized eigenvalue problem $(\xi I + LK)\alpha = \lambda K \alpha$. In this eigenvalue problem expression, K is the $N \times N$ Gram matrix defined on the input unlabeled data by $K_{ij} = K(x_i, x_j)$. Unlike the unsupervised learning algorithm of [3], we are now interested in all of the $\{f_j\}$, not just one for binary classification. Note that this set of projection maps is defined for all points in $\mathcal{M}$, not just those used to define L.

4.3. Intrinsic spectrogram algorithm

Returning to the goal of an intrinsic spectrogram representation, we need to supply as unlabeled data, $\{x_i\}$, a large set of phoneme Fourier spectra across all phonetic categories. This provides the mesh on the speech manifold needed to create the adjacency graph that defines the graph Laplacian. Solving the optimization problem of Equation 5 determines the projection onto a basis that will differentiate a signal's intrinsic trajectory on the manifold rather than its extrinsic one in the ambient space. Given the low-dimensional curved manifold structure to speech motivated in previous sections, we expect phonetic content to be differentiated with fewer components in this basis than with a traditional spectrogram. We also expect a clustering of phonetic content to be reflected in the intrinsic basis as well.

We created a dataset consisting of 10 examples of each of the 58 phonemes, randomly chosen from the TIMIT database.

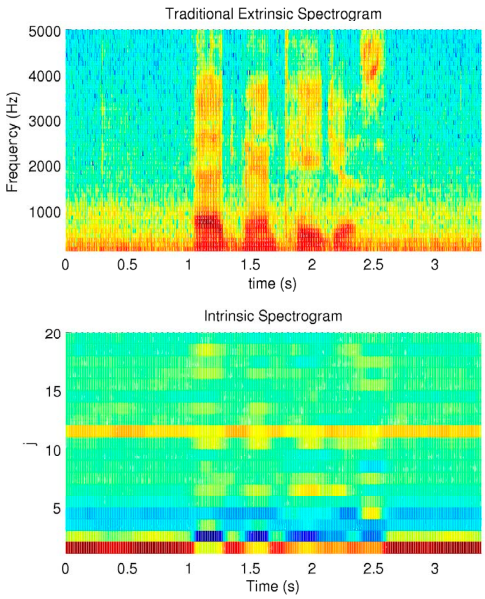

Fig. 1. Traditional (top) and intrinsic (bottom) spectrograms for the word "advantageous".

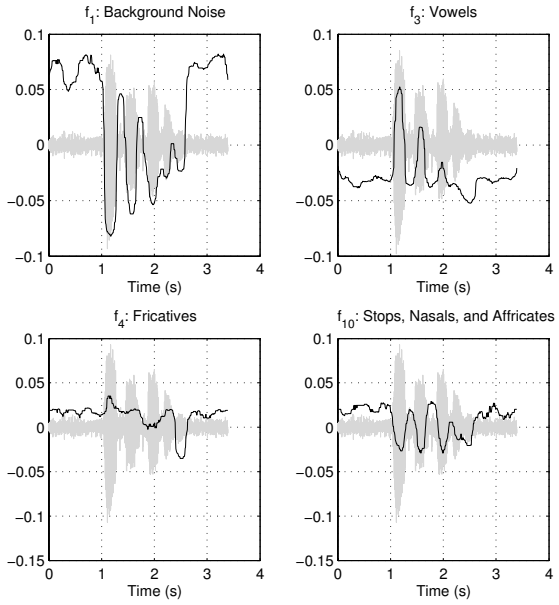

Fig. 2. Four components of the intrinsic spectrogram (shown in black) overlaid on the wave form.

This is the requisite unlabeled training data required by the unsupervised algorithm to determine the intrinsic basis functions as described above. Since the $\{\alpha^j\}$ can be precomputed offline using this standard set, converting a traditional spectrogram into this intrinsic representation requires only a computation of Equation 6 for each time window.

Fig. 1 shows the extrinsic ($H = 50$) and intrinsic spectrograms for a noisy recording of the word "advantageous". Here we construct the adjacency graph with $n = 6$ nearest neighbors, take as the smoothness parameter $\xi = 1$, and implement the linear kernel $K(x, y) = x^T y$ (see Equation 5). Notice that most of the activity in the intrinsic representation is contained in the first 10 components. This indicates that the intrinsic representation is compressing relevant information more efficiently than in the extrinsic case.

Fig. 2 shows various components of the intrinsic spectrogram overlaid on the recording's waveform. These examples demonstrate the intrinsic spectrogram ability to efficiently reflect phonetic distinction. The first component distinguishes between signal and background noise. The third component picks out the two occurrences of the vowel phoneme /æ/. The fourth component picks out the fricative /s/. Finally, the 10th component indicates the stops /t/ and /d/, nasal /n/, and affricate /j/. The general trend is increasing distinction as the component index increases. The lower eigenvectors, then, provide broad class distinction, while the higher eigenvectors serve to differentiate smaller nuances. It is not the case that each component makes a single classification. However, the number of components involved in a particular classification is much smaller than if we attempt to determine phonetic content using a standard spectrogram.

5. CONCLUSION

We have argued that speech sounds form a low-dimensional curved manifold. From this, we have motivated the utility of an intrinsic spectrogram representation. We have presented an algorithm to approximate this intrinsic form and provided an example demonstrating this method's compact representation of the phonetic dimensions.

6. REFERENCES

[1] G. Fant, *Acoustic Theory of Speech Production*, Mouton and Co., Paris, 1970.

[2] K. N. Stevens, *Acoustic Phoenetics*, MIT Press, Cambridge, MA, 1998.

[3] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold Regularization: A Geometric Framework for Learning from Examples," Tech. Rep. TR-2004-06, U. of Chicago, Aug. 2004.

[4] A. Jansen and P. Niyogi, "A Geometric Perspective on Speech Sounds," Tech. Rep. TR-2005-08, U. of Chicago, June 2005.

[5] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," Tech. Rep. TR-2002-01, U. of Chicago, Jan. 2002.