Intrinsic Fourier Analysis on the Manifold of Speech Sounds

Aren Jansen    Partha Niyogi

Department of Computer Science

THE UNIVERSITY OF CHICAGO

ICASSP 2006
Objectives

1. Motivate the existence of a low-dimensional manifold structure for speech

2. Present an algorithm to compute the intrinsic spectrogram
The Physics of Speech Production

Aren Jansen, Partha Niyogi

Intrinsic Fourier Analysis on the Manifold of Speech Sounds
The Physics of Speech Production

Acoustic Tube Model

The map \( \phi_N : \{L_i\} \times \{A_i\} \rightarrow \mathcal{M}_N \) is a diffeomorphism.

Inverse map \( \phi^{-1} \) is a coordinate chart on \( \mathcal{M}_N \).

\( \mathcal{M}_N \) is a smooth, \( 2N \)-dimensional submanifold of \( \mathcal{L}^2 \).

\( \mathcal{M}_N \) is extrinsically curved and spans the ambient space.

\[
M = \prod_{i=1}^{N} \left[ \cos \frac{\omega L_i}{c} \frac{A_i}{\rho_0 c} \sin \frac{\omega L_i}{c} \right] \]

\[
g(\omega) = \frac{Z_r(\omega)}{M_{22} - Z_r(\omega) M_{12}}
\]
The Speech Manifold

- $\mathcal{A} = \text{set of vocal tract articulatory configurations}$
- $\mathcal{M} = \text{set of vocal tract transfer functions}$
- Physics $\Rightarrow \phi : \mathcal{A} \rightarrow \mathcal{M}$ is a diffeomorphism
- Low $\dim(\mathcal{A}) \Rightarrow \mathcal{M}$ is a low-dimensional manifold
Intrinsic Spectrogram Representation

For a signal $x(t)$, let

$$\vec{x}_i = i^{th} \text{ signal window}$$

$$\vec{y}_i = \| \text{DFT}(\vec{x}_i) \| \in \mathcal{M} \subset \mathbb{R}^H$$

Traditional spectrogram, $S(t_i, \omega_j) = \vec{y}_i[j]$

Rewrite: $S(t_i, \omega_j) = f_j(\vec{y}_i)$ where $f_j : \mathbb{R}^H \rightarrow \mathbb{R}$, $f_j(\vec{v}) = \vec{v}[j]$

Our Goal
Implement intrinsic projection maps

Aren Jansen, Partha Niyogi
Intrinsic Fourier Analysis on the Manifold of Speech Sounds
The Laplacian Operator, $\Delta_M$

- Second-order differential operator on manifold $\mathcal{M}$
- Normalized eigenfunctions $\{e_i\}$ form orthogonal basis for $L^2(\mathcal{M})$ $(i.e. \ f = \sum_i a_i e_i)$
- Define smoothness functional:

$$S[f] = \int_{\mathcal{M}} \|\nabla_M f\|^2 d\mu = \langle \Delta_M f, f \rangle_{L^2(\mathcal{M})}$$

$$S[e_i] = \lambda_i$$

- Low $\lambda_i \Rightarrow e_i$ varies smoother with geodesic distance along manifold
The Graph Laplacian Operator, $L_G$

- Given $x_1, x_2, \ldots, x_N \in \mathcal{M}$ construct $k$-nearest neighbor adjacency graph $G$, with adjacency matrix $W$

- $L_G = W - D$. where $D_{ii} = \sum_j W_{ij}$

- Analogous to $\Delta_M$, but restricted to functions on graph

- $S_G[f] = f^T L_G f$, where $f = \langle f(x_1), \ldots, f(x_N) \rangle^T$

Aren Jansen, Partha Niyogi 
Intrinsic Fourier Analysis on the Manifold of Speech Sounds
Computing Intrinsic Maps

- Solve optimization problem:

\[ f^* = \arg \min_{f \in \mathcal{H}_K} \| f \|_K^2 + \xi f^T L f \]

- Admits solutions of form:

\[ f_j^*(v) = \sum_{i=1}^{N} \alpha_j^i K(x_i, v) \]

where

- \( \alpha_j^i \in \mathbb{R}^N \) is the \( j \)-th eigenvector to \((\xi I + LK)\alpha = \lambda K\alpha\)
- \( K \) is the \( N \times N \) Gram matrix with \( K_{ij} = K(x_i, x_j) \)
Algorithm Summary

1. Supply a large set of Fourier amplitude spectra, \( \{ x_i \} \), across all phonetic classes
2. Calculate the graph Laplacian over \( \{ x_i \} \)
3. Solve the optimization problem to recover intrinsic basis projection maps
4. Project traditional spectrogram on the intrinsic basis
An Example: “advantageous”

\[
\{x_1, \ldots, x_N\}: \text{10 examples of 58 TIMIT phonemes}
\]

\[
k = 6, \xi = 1, \quad K(x, y) = x^T y
\]

Activity limited to first ten components

Efficient information encoding
Intrinsic Component Behavior

- Initial components provide broad class distinctions
- Higher $\lambda$ components differentiate smaller nuances
Semi-supervised Classification

- Given \( l \) labelled and \( u \) unlabelled examples
- Construct intrinsic basis with \( u \) unlabelled examples
- \( E(l) = \text{extrinsic test error} \)
  \( I(l) = \text{intrinsic test error} \)
  \( O(u + l) = \text{optimal test error} \)
- Performance Gap Improvement:
  \[ G(l) = \frac{E(l) - I(l)}{E(l) - O} \]
Conclusions

- Speech sounds have an underlying low-dimensional manifold structure.
- Manifold structure can be exploited for novel speech representations.
- Intrinsic Fourier analysis provides a compact representation for speech signals.
- Approach shows promise in speech recognition applications.