Semi-Supervised Learning of Speech Sounds

Aren Jansen **Partha Niyogi**

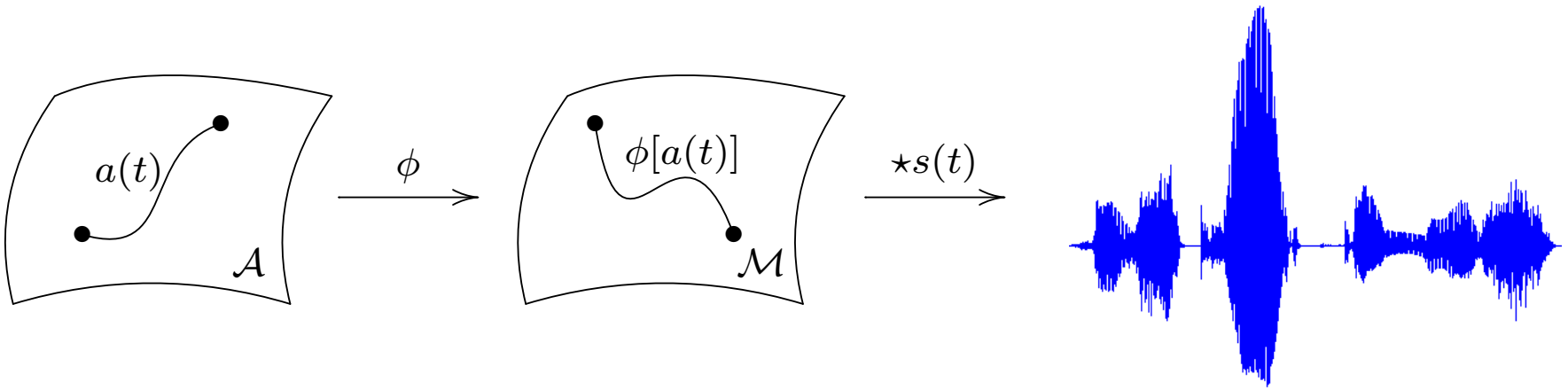
Department of Computer Science

Interspeech 2007

Objectives

1. Present a manifold learning algorithm based on locality preserving projections for semi-supervised phone classification (LPPSSL)
2. Perform toy classification experiments designed to isolate the role of geometric structure in the speech domain
3. Demonstrate that exploiting both manifold and cluster structure is necessary for semi-supervised success

The Speech Manifold



- $\mathcal{A} =$ space of vocal tract articulatory configurations
- $\mathcal{M} =$ space 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

The Laplacian Operator, $\Delta_{\mathcal{M}}$

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

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

$$S[e_i] = \lambda_i$$

- Low $\lambda_i \Rightarrow e_i$ varies more smoothly 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 $\Delta_{\mathcal{M}}$, but restricted to functions on graph
- $S_G[\mathbf{f}] = \frac{\mathbf{f}^T L_G \mathbf{f}}{\mathbf{f}^T D \mathbf{f}}$, where $\mathbf{f} = \langle f(x_1), ..., f(x_N) \rangle^T$
- Function that minimizes S_G is minimum cut for G

Computing an Ordered Intrinsic Basis with LPP

- Solve optimization problem:

$$\mathbf{f}^* = \arg \min_{\mathbf{f}^T \mathbf{D} \mathbf{f} = 1} \mathbf{f}^T L_G \mathbf{f} \iff L_G \mathbf{f}_k = \lambda_k D \mathbf{f}_k$$

- Extend k^{th} eigenfunction, $\mathbf{f}_k$, out of sample ($f_k \in \mathcal{H}_K$):

$$f_k(v) = \sum_{i=1}^{N} \alpha_i^{(k)} K(x_i, v)$$

where $\alpha^{(k)} = \mathcal{K}^+ \mathbf{f}_k$ and $\mathcal{K}^+ =$ pseudoinverse of the $N \times N$ Gram matrix with $\mathcal{K}_{ij} = K(x_i, x_j)$

- Sort eigenfunctions according to

$$m(\mathbf{f}) = \frac{\operatorname{sgn}(\mathbf{f})^T L_G \operatorname{sgn}(\mathbf{f})}{\operatorname{sgn}(\mathbf{f})^T D \operatorname{sgn}(\mathbf{f})}$$

LPPSSL: Incorporating Labelled Data

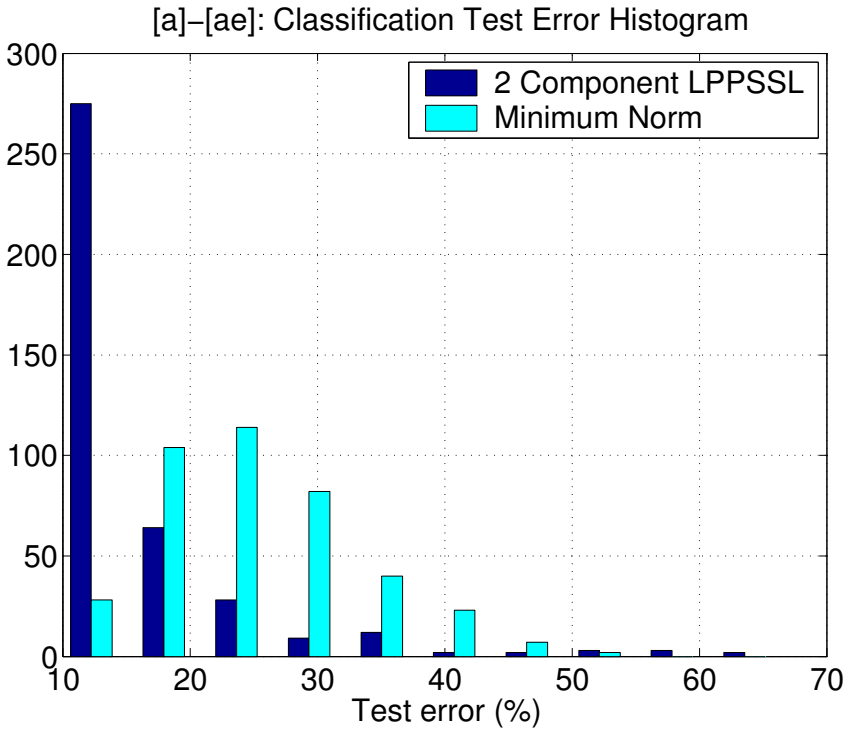
- Given d non-trivial basis functions, $\{f_1, \ldots, f_d\}$, cast labelled examples $\{x_i\}_{i=1}^{l}$ into intrinsic representation:

$$x' = \langle f_1(x), \ldots, f_d(x) \rangle$$

- Determine map from intrinsic representation to labels using any machine learning method
- For a linear map (min-norm solution):
  1. Let F be $(l \times d)$ matrix with $F_{ij} = f_i(x_j)$
  2. Let $\mathbf{y}_l \in \{-1, 1\}^l$ be the vector of training labels
  3. Solve $\mathbf{y}_l = F\beta$ (for $l \neq d$, use Moore-Penrose inverse)

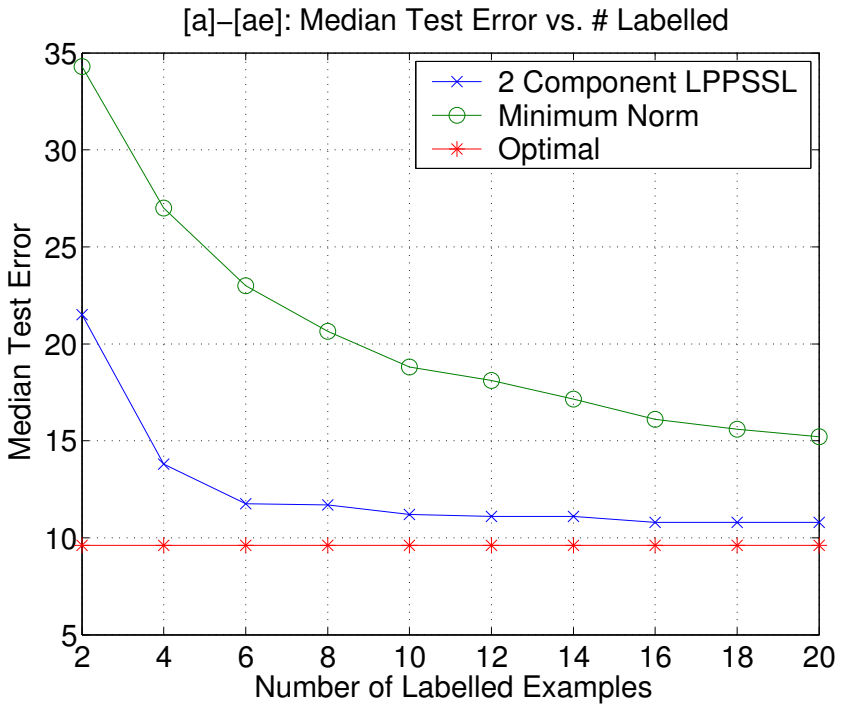
A Detailed Example: /a/-/æ/ Classification

- 50-dim DFT representation for each example
- 500 training examples from each class
- Test procedure (repeat 400 times):
  1. Randomly label $l/2$ of each class
  2. Compute linear classifiers with l labelled and $u = 1000 - l$ unlabelled examples
  3. Test on additional 1000 examples



Optimal RLS Error = 9.6%
Median LPPSSL Error = 11.8%
Median Min-norm Error = 24.3%

A Detailed Example: /a/-/æ/ Classification



- Define gap improvement:

$$G(l) = \frac{\text{Min-norm error} - \text{LPPSSL error}}{\text{Min-norm error} - \text{Optimal error}} \approx \text{const}$$

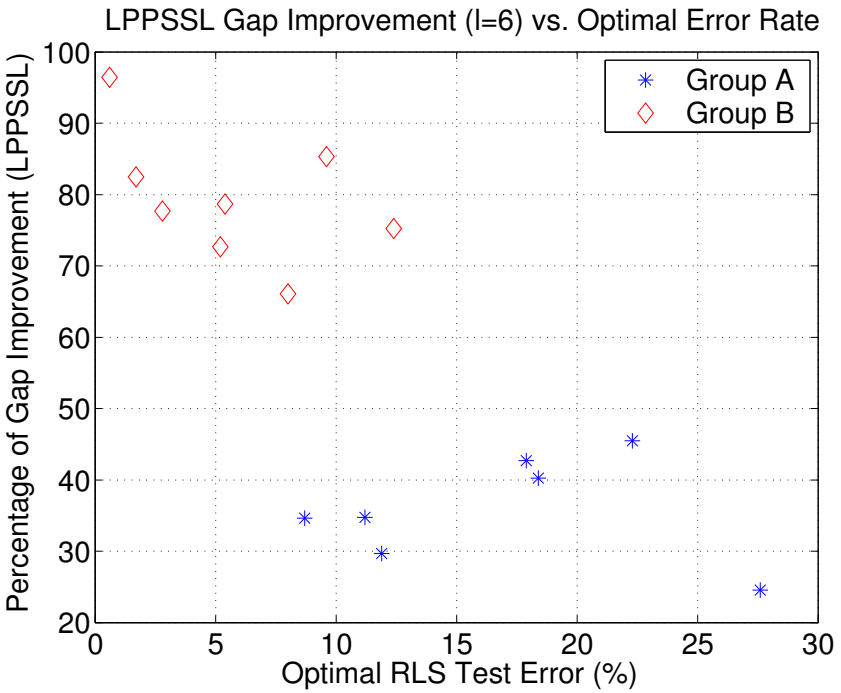
Performance Across the Vowel Manifold

Pair	E_{opt}	E_{mn}	E_{ssl}	$G(6)$
ə-o	27.6	44.3	40.2	24.6
ɪ-i	22.3	37.7	30.7	45.5
o-ɑ	18.4	36.9	29.5	40.3
ə-ɪ	17.9	30.2	25.0	42.7
ə-ɑ	12.4	29.8	16.7	75.2
ɪ-æ	11.9	30.1	24.7	29.7
ə-æ	11.2	25.5	20.5	34.7
ɑ-æ	9.6	24.3	11.8	85.3
o-æ	8.7	24.3	18.9	34.6
o-ɪ	8.0	21.7	12.7	66.1
ə-i	5.4	14.3	7.3	78.7
i-æ	5.2	14.9	7.9	72.7
ɑ-ɪ	2.8	11.1	4.7	77.7
o-i	1.7	7.4	2.7	82.5
ɑ-i	0.6	3.4	0.7	96.4

- $E_{\text{opt}} =$ optimal RLS error
- $E_{\text{mn}} =$ median min-norm error ($l = 6$)
- $E_{\text{ssl}} =$ median 2-comp. LPPSSL error ($l = 6$)
- $G(6) = \frac{E_{\text{mn}} - E_{\text{ssl}}}{E_{\text{mn}} - E_{\text{opt}}}$

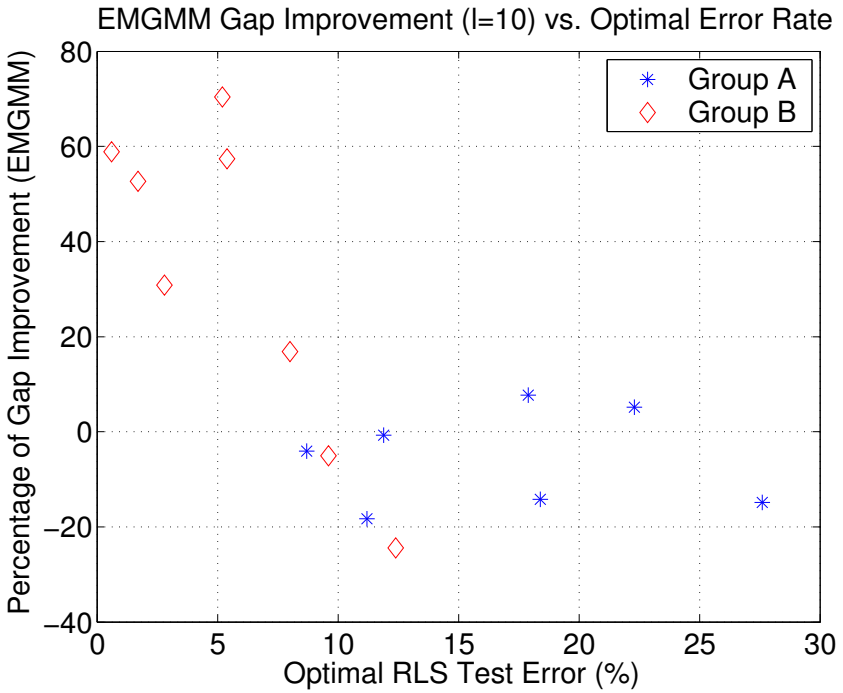
Performance Across the Vowel Manifold

- **Group A** pairs are poorly separable/minimally clustered (e.g. close vs. near close)
- **Group B** pairs are highly clustered with distinct articulator configurations (e.g. close vs. open, front vs. back)
- Manifold structure admits significant gap improvements for **Group A** pairs



Isolating the Role of Cluster Structure

- Semi-supervised EMGMM algorithm:
  1. Train 2-mixture GMM with $l = 10$ labelled examples
  2. Classify unlabelled examples and iterate
- Optimal GMM and RLS error rates clearly correlated
- EMGMM fails on **Group A** problems



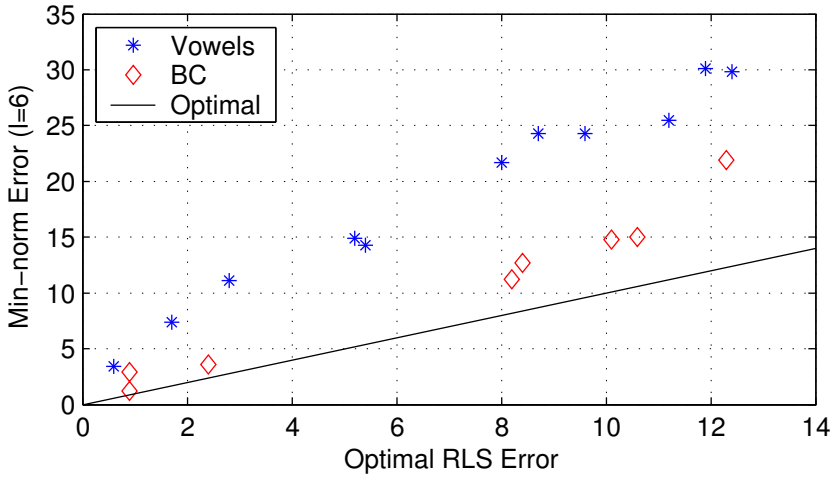
Broad Class Performance

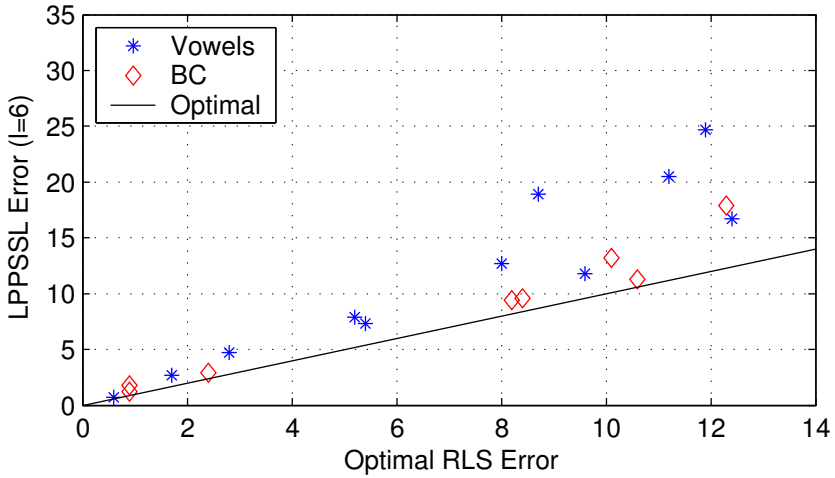
Pair	E_{opt}	E_{mn}	E_{ssl}	$G(6)$
Ap-V	34.9	42.5	39.6	37.7
St-F	24.7	30.7	28.2	41.7
St-Ap	21.1	28.4	24.8	49.3
N-Ap	20.6	37.6	28.7	52.6
St-V	19.5	34.0	28.7	36.9
Af-F	17.1	27.2	21.5	56.9
N-V	14.8	34.0	18.2	82.3
St-N	12.3	21.9	17.9	41.7
F-Ap	10.6	15.0	11.3	85.2
St-Af	10.1	14.8	13.2	33.3
F-V	8.4	12.7	9.6	72.0
F-N	8.2	11.2	9.4	60.0
Af-Ap	2.4	3.6	2.9	58.3
Af-V	0.9	2.9	1.8	55.0
Af-N	0.9	1.2	1.2	0.0

- Classes: Vowels, Approximants, Nasals, Fricatives, Affricates, Stops
- 500 train/test examples for each class
- Individual phones represented according to their occurrence rate in TIMIT

Broad Class vs. Vowel Performance

- Broad class clusters more separated than vowels
- Min-norm broad class outperforms min-norm vowels
- LPPSSL performance roughly the same
- Accomodating both manifold and cluster structure provides invariance to cluster separation





Conclusions

- Speech sounds have an approximate low-dimensional manifold structure
- Presented LPPSSL algorithm to leverage manifold structure for semi-supervised learning
- Cluster structure alone is insufficient for the speech domain
- Manifold structure can be beneficial even with minimal supervision