APPLICATION OF A LOCALITY PRESERVING DISCRIMINANT ANALYSIS APPROACH TO ASR

Vikrant Singh Tomar, Richard C. Rose
McGill University

telecommunications & signal processing laboratory

Presented at,
ISSPA 2012
Montreal, QC, Canada

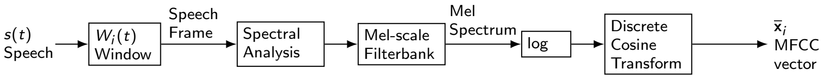
July 3, 2012

Outline

- **Background:** Feature analysis for HMM (hidden Markov model) based ASR
- **Problem:** Capturing spectral dynamics requires high dimensional feature vectors (dim > 100, typically)
- **Solution:** Dimensionality reducing linear transformations
- **Approach:** Locality preserving discriminant analysis (LPDA)
  - maximize discrimination between model classes
  - preserve local structure of the within-class data
- **Experimental Study:** Compare ASR performance for a speech in noise task using LPDA with performance obtained using more well known approaches

ASR Feature Analysis

- Mel-frequency Cepstrum Coefficients (MFCC)



- Captures the static spectral information over a ∼20 msec analysis frame.
- What about surrounding speech context (evolution of speech spectrum)?

Capturing Spectrum Evolution

- Concatenate multiple speech frames (typically ∼100 msec of speech):

$$\mathbf{x}_i = \begin{bmatrix} \overline{\mathbf{x}}_{i-k} \\ \vdots \\ \overline{\mathbf{x}}_i \\ \vdots \\ \overline{\mathbf{x}}_{i+k} \end{bmatrix}$$

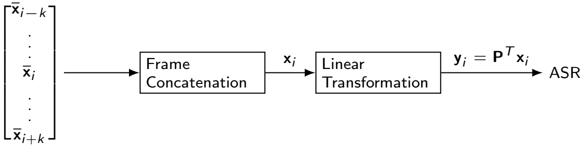
  [Eisele and Haeb-Umbach, 1996]

- Issues:
  - High dimensionality of the resultant feature vectors ($dim = \mathbf{117}$ for $k = \mathbf{4}$)
  - High inter-frame correlation among feature vectors
- **Solution:** Dimensionality reducing linear transformations

Feature-space Transformations

- Project high dimensional feature vectors to a lower dimensional space

$$\mathbf{y}_i = \mathbf{P}^T \mathbf{x}_i$$



- Optimization criteria for estimating **P**:
  - Improved class separability – use a discriminant criterion – Linear Discriminant Analysis (LDA) [Duda et al., 2000]
  - Preserve underlying geometrical relationships among the feature vectors – use a manifold learning approach – Locality Preserving Projections (LPP)[He and Niyogi, 2002, Tang and Rose, 2008]

Manifold Learning

- Find a low-dimensional basis for describing high dimensional data
- Assumption: High dimensional data can be considered as a set of geometrically related points rest- ing on or close to the surface of a lower dimensional manifold.
- **Why:** Local relationships among feature vectors can be constrained by the manifold

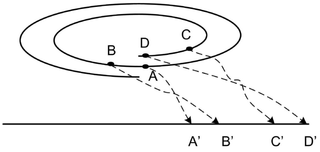


Illustration of dimensionality reduction for two-dimensional data embedded in a nonlinear manifold space with relative position information reserved. [Tang and Rose, 2008]

An Alternative Optimization Criterion

- Motivation:
  - Discriminant approaches like LDA do not account for the geometric structure of the data
  - Locality preserving approaches like LPP do not enhance class discrimination
- Locality preserving discriminant approach (LPDA):
  – Combines manifold learning with inter-class discrimination
  – Multiple class specific sub-manifolds
  - Maximize class separability : Discriminate between sub-manifolds
  - Preserve local within class relationships : Preserve local sub-manifold structures

Locality Preserving Discriminant Analysis (LPDA)

- Embed feature vectors $\mathbf{X}$ into graph(s) $\mathcal{G}$ defined over *some* geometric measure $\mathbf{W} = [w_{ij}]_{N \times N}$ [Yan et al., 2007]
  - The idea is to manipulate the geometry of the graph nodes while preserving important relationships between them
- For a graph $\mathcal{G} = \{\mathbf{X}, \mathbf{W}\}$, graph scatter measure in the transformed space is defined as:

$$F(\mathbf{P}) = \sum_{i \neq j} ||\mathbf{y}_i - \mathbf{y}_j||^2 w_{ij} = \sum_{i \neq j} ||\mathbf{P}^\mathbf{T}\mathbf{x}_i - \mathbf{P}^\mathbf{T}\mathbf{x}_j||^2 w_{ij}$$

The goal of LPDA is to minimize the within class scatter, and maximize the between class scatter while preserving local relationships

LPDA – Graph Embedding

- Embed the feature vectors belonging to the same class into intrinsic graph $\mathcal{G}_{int} = \{\mathbf{X}, \mathbf{W}_{int}\}$
  - $\mathbf{X}$ = Nodes of the graphs = features vectors
  - $\mathbf{W_{int}}$ = Intrinsic affinity matrix; $\mathbf{W_{int}} = [w_{ij}^{int}]_{N \times N}$

$$w_{ij}^{int} = \begin{cases} exp(-||\mathbf{x}_i - \mathbf{x}_j||^2)/\rho & ; \quad \mathbf{x}_i \text{ \& } \mathbf{x}_j \textit{ are close and in same class} \\ 0 & ; \quad \textit{otherwise} \end{cases}$$

- Embed the feature vectors belonging to different classes into penalty graph $\mathcal{G}_{pen} = \{\mathbf{X}, \mathbf{W}_{pen}\}$
  - $\mathbf{W_{pen}}$ = Penalty affinity matrix; $\mathbf{W_{pen}} = [w_{ij}^{pen}]_{N \times N}$

$$w_{ij}^{pen} = \begin{cases} exp(-||\mathbf{x}_i - \mathbf{x}_j||^2)/\rho & ; \quad \mathbf{x}_i \text{ \& } \mathbf{x}_j \textit{ are close but NOT in same class} \\ 0 & ; \quad \textit{otherwise} \end{cases}$$

LPDA – Optimization Criterion

- Minimize the scatter of the intrinsic graph $F_{int}(\mathbf{P})$ (preserve within-class manifold based relationships)
- Maximize the scatter of the penalty graph $F_{pen}(\mathbf{P})$ (maximize inter-class discrimination)

$$\mathbf{P}_{lpda} = \arg\max_{\mathbf{P}} \frac{F_{pen}(\mathbf{P})}{F_{int}(\mathbf{P})}$$

- $\mathbf{P}_{lpda}$ can be obtained by solving the generalized eigenvalue problem:

$$(\mathbf{X}(\mathbf{D}_{pen} - \mathbf{W}_{pen})\mathbf{X}^T)\mathbf{p}^j_{lpda} = \lambda_j(\mathbf{X}(\mathbf{D}_{int} - \mathbf{W}_{int})\mathbf{X}^T)\mathbf{p}^j_{lpda}$$

$\mathbf{D} = [d_{ij}]$ is a diagonal matrix whose elements correspond to the column sum of the affinity matrix $\mathbf{W}$, *e.g.*, $d^{int}_{ii} = \sum_j w^{int}_{ij}$ etc.

Experimental Study

- Evaluate feature-space dimensionality reducing transformations in terms of ASR word error rate (WER) on a speech in noise task domain
- Compare:
  - Linear discriminant analysis (LDA)
  - Locality preserving projections (LPP)
  - Locality preserving discriminant analysis (LPDA)
- After projection, feature-decorrelation (diagonal covariances) is no longer guaranteed
  - Most ASR systems assume diagonal covariances
  - Combine with semi-tied covariance (STC) transformations [Gales, 1999]

Task Domain

- Aurora2 speech corpus:
  - 8440 noise corrupted utterances from 55 male and 55 female speakers for training
  - 4004 utterances; four different noise types for testing
- Baseline:
  - 12-dimensional MFCC + Energy + Δ + $\Delta\Delta$ features used for baseline
  - Whole word continuous density HMM model
  - 11 words + sil + sp, 16 states per word $\Rightarrow$ 180 states, 3 Gaussians per state
- Feature-space transformations:
  - 9 frames stacked for feature concatenation
  - Continuous density HMM states used as classes
  - Semi-tied covariance adaptation is performed

ASR (% WER) for Aurora2 Corpus

Noise Type	Technique	SNR (dB)			
		20	15	10	5
Car	Baseline	2.77	3.36	5.45	**12.31**
	LDA	3.82	4.26	6.74	17.15
	LDA + STC	2.83	3.45	5.69	15.92
	LPP+STC	2.71	3.61	6.08	14.97
	LPDA+STC	**2.30**	**2.77**	**5.19**	12.73
Airport	Baseline	3.42	4.88	8.49	16.58
	LDA	5.67	7.07	10.26	19.83
	LDA+STC	3.18	4.11	7.72	15.65
	LPP+STC	4.35	6.95	10.38	21.15
	LPDA+STC	**3.10**	**4.09**	**7.49**	**15.09**

- Use of semi-tied covariance (STC) is critical for all approaches.

ASR (% WER) for Aurora2 Corpus

Noise Type	Technique	SNR (dB)			
		20	15	10	5
Car	Baseline	2.77	3.36	5.45	**12.31**
	LDA	3.82	4.26	6.74	17.15
	LDA + STC	2.83	3.45	5.69	15.92
	LPP+STC	2.71	3.61	6.08	14.97
	LPDA+STC	**2.30**	**2.77**	**5.19**	12.73
Airport	Baseline	3.42	4.88	8.49	16.58
	LDA	5.67	7.07	10.26	19.83
	LDA+STC	3.18	4.11	7.72	15.65
	LPP+STC	4.35	6.95	10.38	21.15
	LPDA+STC	**3.10**	**4.09**	**7.49**	**15.09**

- All approaches are effective (better than baseline) at high and medium SNR's
- All approaches are not effective at low SNR

ASR (% WER) for Aurora2 Corpus

Noise Type	Technique	SNR (dB)			
		20	15	10	5
Car	Baseline	2.77	3.36	5.45	**12.31**
	LDA	3.82	4.26	6.74	17.15
	LDA + STC	2.83	3.45	5.69	15.92
	LPP+STC	2.71	3.61	6.08	14.97
	LPDA+STC	**2.30**	**2.77**	**5.19**	12.73
Airport	Baseline	3.42	4.88	8.49	16.58
	LDA	5.67	7.07	10.26	19.83
	LDA+STC	3.18	4.11	7.72	15.65
	LPP+STC	4.35	6.95	10.38	21.15
	LPDA+STC	**3.10**	**4.09**	**7.49**	**15.09**

- LPDA+STC provides highest WER reduction in most noise conditions

Conclusions

- LPDA: a feature-space dimensionality reduction approach that combines discriminant and manifold learning criteria

Conclusions

- LPDA: a feature-space dimensionality reduction approach that combines discriminant and manifold learning criteria
  - Graph embedding:
    - A generalized framework
    - No assumption about the distribution of data
  - Manifold learning: Preserve within-class nonlinear structure of the data
  - Between class discrimination
  - Soft-weights: the closer the two vectors the higher the penalty upon misclassification

Conclusions

- LPDA: a feature-space dimensionality reduction approach that combines discriminant and manifold learning criteria
  - Graph embedding:
    - A generalized framework
    - No assumption about the distribution of data
  - Manifold learning: Preserve within-class nonlinear structure of the data
  - Between class discrimination
  - Soft-weights: the closer the two vectors the higher the penalty upon misclassification
- Provides from $6 - 27\%$ reduction in WER relative to LDA

Conclusions

- LPDA: a feature-space dimensionality reduction approach that combines discriminant and manifold learning criteria
  - Graph embedding:
    - A generalized framework
    - No assumption about the distribution of data
  - Manifold learning: Preserve within-class nonlinear structure of the data
  - Between class discrimination
  - Soft-weights: the closer the two vectors the higher the penalty upon misclassification
- Provides from $6-27\%$ reduction in WER relative to LDA
- Populating the affinity matrices $\mathbf{W}_{int}$ and $\mathbf{W}_{pen}$ is a very computationally intensive task
  - Future work will include reducing the relatively high computation cost of estimating the LPDA transformation matrix

References

[Duda et al., 2000]
Duda, R. O., Hart, P. E., and Stork, D. G. (2000)
Pattern Classification
Wiley Interscience, 2nd edition

[Eisele and Haeb-Umbach, 1996]
Eisele, T. and Haeb-Umbach, R. (1996)
A comparative study of linear feature transformation techniques for automatic speech recognition
Spoken Language, 1996, pages 1–4

[Gales, 1999]
Gales, M. J. F. (1999)
Semi-tied covariance matrices for hidden markov models
IEEE Transactions on Speech and Audio Processing, 7(3):272 – 281

[He and Niyogi, 2002]
He, X. and Niyogi, P. (2002)
Locality Preserving Projections
In *Neural Information Processing Systems (NIPS)*

[Tang and Rose, 2008]
Tang, Y. and Rose, R. (2008)
A study of using locality preserving projections for feature extraction in speech recognition
In *ICASSP: IEEE International Conference on Acoustics, Speech, and Signal Processing*

[Yan et al., 2007]
Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q., and Lin, S. (2007)
Graph embedding and extensions: a general framework for dimensionality reduction
IEEE transactions on pattern analysis and machine intelligence, 29(1):40–51

Why STC?

- Only a limited number of parameters can be *robustly* estimated for each CDHMM state
  - Modeling full covariances (when correlation exists) results in a dramatic increase in such parameters
  - Hence, independence between feature vector components is assumed in ASR
  - But not explicitly modeling the full-covariance results in ASR performance degradation
- Dimensionality reduction generally results in a highly correlated feature space, *i.e.*, full covariance matrices
  - Discarding this information results in performance degradation
- Semi-tied covariances [Gales, 1999]:
  - Approximates full covariance modeling by allowing few full covariance matrices to be shared across many distributions
  - Effectively each distribution maintains its own diagonal covariance