On the use of KPCA for feature extraction in speech recognition

*Eurospeech 2003*

A. Lima, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura

{amaro, zen, nankaku, chiyomi, tokuda, kitamura}@ics.nitech.ac.jp.

Nagoya Institute of Technology
Department of Computer Science
Nagoya, Japan
Introduction

Traditional × Proposed Approach

\[ w_i = [w_{i,1}, w_{i,2}, \ldots, w_{i,n}]^T \]: mel-cepstral coefficients of the i-th frame

\[ W_{i}^{kpc} = [W_{i,1}^{kpc}, W_{i,2}^{kpc}, \ldots, W_{i,k}^{kpc}]^T \]: KPCA representation of \( w_i \)

Higher dimensional representation

↓

More discriminative features
Overview

- Kernel-based techniques
  - Support Vector Machines (SVMs)
  - Kernel Discriminant Analysis (KDA)
  - Kernel Principal Component Analysis (KPCA)
- Have not been carefully investigated in speech recognition

Objective

- Represent the speech features in a higher dimensional space
- A 520-word-vocabulary recognition experiment
Principal Component Analysis

General aspects

- Well-known and established technique for dimensionality reduction
- It represents a linear transformation
- The data is expressed in a new coordinate basis
- It corresponds to the maximum variance “direction”
Principal Component Analysis

Equations

- Centered data
  \[ x_k \in \mathbb{R}^n, \ k = 1, \ldots, M \text{ and } \sum_{k=1}^{M} x_k = 0 \]

- Covariance matrix
  \[ C = \frac{1}{M} \sum_{j=1}^{M} x_j x_j^T \]

- Eigenvalue equation
  \[ \lambda v = Cv \text{ where } \lambda \geq 0, \ v \in \mathbb{R}^n \backslash \{0\} \]

- All the solutions for \( v \) must be a linear combination of \( x_1, \ldots, x_M \)

- \( \lambda (x_k \cdot v) = (x_k \cdot Cv), \ k = 1, \ldots, M \)
Kernel Functions

General aspects

- Solve the problem of nonlinear mapping to a higher dimensional space

\[ \Phi : \mathbb{R}^n \mapsto \mathcal{F} \]

\[ \mathbf{x} \mapsto \mathbf{X} \]

- Without using an explicit mapping

- Using kernel functions as the dot product of the mapped variables

\[ K(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}) \]
Kernel Functions

Examples

- **Polynomial:** $K(x, y) = (x \cdot y + 1)^p$
- **Gaussian RBF (GRBF):** $K(x, y) = e^{-\frac{|x-y|^2}{2\sigma^2}}$
- **Sigmoidal:** $K(x, y) = \tanh(\chi x \cdot y - \delta)$
General aspects
- It applies the kernel function to PCA
- The representation of PCA in a higher dimensional space
Extention to Kernel PCA

Diagram

Training Samples

Test Samples

Input space
Extention to Kernel PCA

Diagram

Training Samples

Test Samples

Input space

Feature space

$\Phi$

Principal components

$\Phi(w_1)$ $\Phi(w_2)$ $\Phi(w_3)$
Diagram

Extention to Kernel PCA
Extention to Kernel PCA

Equations

- Centered mapped data
  \[ \Phi(x_1), \ldots, \Phi(x_M) \text{ and } \sum_{k=1}^{M} \Phi(x_k) = 0 \]

- Covariance matrix
  \[ \Sigma = \frac{1}{M} \sum_{j=1}^{M} \Phi(x_j)\Phi(x_j)^T \]

- Eigenvalue equation
  \[ \lambda \mathbf{V} = \Sigma \mathbf{V}, \text{ where } \lambda \geq 0 \text{ and } \mathbf{V} \in \mathcal{F}\backslash\{\mathbf{0}\} \]

- All the solutions for \( \mathbf{V} \) must be a linear combination of
  \[ \Phi(x_1), \ldots, \Phi(x_M) \]

- \[ \lambda (\Phi(x_k) \cdot \mathbf{V}) = (\Phi(x_k) \cdot \Sigma \mathbf{V}), \ k = 1, \ldots, M \]
Extension to Kernel PCA

Equations

\[ \lambda(\Phi(x_k) \cdot V) = (\Phi(x_k) \cdot \Sigma V) \]
Extention to Kernel PCA

Equations

\[ \lambda(\Phi(x_k) \cdot V) = (\Phi(x_k) \cdot \Sigma V) \]

\[ \downarrow \begin{cases} 
V = \sum_{i=1}^{M} \alpha_i \Phi(x_i), \text{ where } \alpha_i, \ i = 1, \ldots, M 
\end{cases} \]
Extention to Kernel PCA

Equations

\[
\lambda (\Phi(x_k) \cdot V) = (\Phi(x_k) \cdot \Sigma V)
\]

\[
\downarrow \left\{ \begin{array}{l}
V = \sum_{i=1}^{M} \alpha_i \Phi(x_i), \text{ where } \alpha_i, i = 1, \ldots, M \\
M \lambda \alpha = K \alpha
\end{array} \right\}
\]

where \( K \) is an \( M \times M \) matrix formed by \( K_{ij} = (\Phi(x_i) \cdot \Phi(x_j)) \)

and \( \alpha = [\alpha_1, \ldots, \alpha_M]^T \)
Extention to Kernel PCA

Equations

\[ \lambda (\Phi(x_k) \cdot V) = (\Phi(x_k) \cdot \Sigma V) \]

\[ \Downarrow \left\{ \begin{array}{l}
V = \sum_{i=1}^{M} \alpha_i \Phi(x_i), \text{ where } \alpha_i, i = 1, \ldots, M \\
M \lambda \alpha = K \alpha \\
\text{where } K \text{ is an } M \times M \text{ matrix, formed by } K_{ij} = (\Phi(x_i) \cdot \Phi(x_j)) \\
\text{and } \alpha = [\alpha_1, \ldots, \alpha_M]^T \\
\text{The solutions for non-zero eigenvalues are } \\
\alpha^k, k = m, \ldots, M, \text{ where } \lambda_m \text{ is the } 1^{st} \lambda \neq 0
\end{array} \right. \]
Extention to Kernel PCA

Equations

\[ \lambda(\Phi(x_k) \cdot V) = (\Phi(x_k) \cdot \Sigma V) \]

\[ \downarrow \begin{cases} V = \sum_{i=1}^{M} \alpha_i \Phi(x_i), \text{ where } \alpha_i, i = 1, \ldots, M \\ M \lambda \alpha = K \alpha \end{cases} \]

where \( K \) is an \( M \times M \) matrix formed by \( K_{ij} = (\Phi(x_i) \cdot \Phi(x_j)) \) and \( \alpha = [\alpha_1, \ldots, \alpha_M]^T \)

The solutions for non-zero eigenvalues are

\[ \alpha^k, k = m, \ldots, M, \text{ where } \lambda_m \text{ is the } 1^{st} \lambda \neq 0 \]

\[ V^k \cdot \Phi(w) = \sum_{i=1}^{M} \alpha_i^k (\Phi(x_i) \cdot \Phi(w)) \]
Extention to Kernel PCA

Training

- $x_i, \ i = 1, \ldots, M$ : training vectors (mel-cepstral coefficients)

$$M \lambda \alpha = \begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) & \cdots & K(x_1, x_M) \\
K(x_2, x_1) & K(x_2, x_2) & \cdots & K(x_2, x_M) \\
\vdots & \vdots & \ddots & \vdots \\
K(x_M, x_1) & K(x_M, x_2) & \cdots & K(x_M, x_M) \end{bmatrix} \begin{bmatrix} \alpha_1 \\
\alpha_2 \\
\vdots \\
\alpha_M \end{bmatrix}$$

Test

- $w$ : test vector (mel-cepstral coefficients)

$$W_{k^{\text{pca}}} = \begin{bmatrix} W_{1}^{\text{pca}} & \cdots & W_{M-m+1}^{\text{pca}} \end{bmatrix}^T$$

$$W_{k-m+1}^{\text{pca}} = V^k \cdot \Phi(w) = \sum_{i=1}^{M} \alpha_i^k (\Phi(x_i) \cdot \Phi(w)), \text{ where } k = m, \ldots, M$$

Eurospeech 2003 – p.14/19
### General features

<table>
<thead>
<tr>
<th>Task</th>
<th>An isolated word recognition task</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Database</strong></td>
<td></td>
</tr>
<tr>
<td>Utterances</td>
<td>520 words (C set - ATR Japanese speech database)</td>
</tr>
<tr>
<td>Speakers</td>
<td>80 speakers (40 males/40 females)</td>
</tr>
<tr>
<td>Training data</td>
<td>10,600 utterances (20 speakers)</td>
</tr>
<tr>
<td>Test data</td>
<td>31,200 utterances (60 speakers)</td>
</tr>
</tbody>
</table>
Experimental Work

Baseline conditions

<table>
<thead>
<tr>
<th>Sampling rate</th>
<th>10 kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Window</td>
<td>25.6 ms Blackman window with 10 ms shifts</td>
</tr>
<tr>
<td>Features</td>
<td>13 Mel-cepstral coefficients + 13 $\Delta$ + 13 $\Delta\Delta$</td>
</tr>
<tr>
<td>HMM</td>
<td>12 states left-right with single mixture</td>
</tr>
<tr>
<td>Feasibility</td>
<td>$N$ frames randomly picked up ($N = 1024$)</td>
</tr>
</tbody>
</table>

Results

- The baseline error rate: 8.36% ($\frac{2608}{31200}$)
### Experimental Work

#### Results

**Error rate (%)**

\[ K(x, y) = (x \cdot y + 1)^p \]

<table>
<thead>
<tr>
<th>dim\kernel</th>
<th>( p = 1^* )</th>
<th>( p = 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>8.82</td>
<td>7.65</td>
</tr>
<tr>
<td>13</td>
<td>7.45</td>
<td>6.71</td>
</tr>
<tr>
<td>16</td>
<td>8.19</td>
<td>6.84</td>
</tr>
<tr>
<td>32</td>
<td>10.37</td>
<td>6.36</td>
</tr>
<tr>
<td>64</td>
<td>*</td>
<td>8.96</td>
</tr>
<tr>
<td>128</td>
<td>*</td>
<td>16.31</td>
</tr>
<tr>
<td>256</td>
<td>*</td>
<td>36.97</td>
</tr>
</tbody>
</table>

* Does not apply.
* It represents the use of PCA.

---

Eurospeech 2003 – p.17/19
Experimental Work

Results

![Bar chart showing error rates for Baseline, PCA, and KPCA. The error rates are 8.36%, 7.45%, and 6.53%, respectively.](chart.png)
Conclusions and Future Work

A kernel based technique was applied to a 520-word-vocabulary speech recognition task

- The use of PCA and KPCA generated some improvement in accuracy for this speech recognition task

- The results presented in this work have confirmed the effectiveness of KPCA

- Further experiments are being conducted to observe the impact of KPCA in speech recognition

- Further research on kernel-based techniques should be performed to observe their effectiveness in speech recognition with differing levels of complexity