Eigenvoice Speaker Adaptation via Composite Kernel PCA

James Kwok  Brian Mak  Simon Ho

Department of Computer Science
Hong Kong University of Science and Technology
Hong Kong
Speaker Adaptation

• A well-trained speaker-dependent (SD) model generally achieves a significantly lower word error rate than a speaker-independent (SI) model on recognizing speech from the specific speaker

• Hard to acquire a large amount of data from a user to train the SD model
  – adapt the SI model with a relatively small amount of SD speech
    • maximum a posteriori (MAP) adaptation
    • maximum likelihood linear regression (MLLR) adaptation
  – when the amount of available adaptation speech is really small (e.g., only a few seconds): eigenvoice-based adaptation
Eigenvoice vs Kernel Eigenvoice

- **Eigenvoice (EV)**
  - use principal component analysis (PCA) to find the eigenvoices
  - represent the new speaker as a linear combination of the leading eigenvoices
  - estimate the (small) set of weights by using maximum likelihood
  - linear PCA $\rightarrow$ captures only linear relationships

- **Kernel** eigenvoice (KEV)
  - kernel PCA
  - issues:
    - do all computations rely only on kernel evaluations?
    - how to compute the observation likelihood?
Eigenvoice Speaker Adaptation via Composite Kernel PCA

PCA

supervector

adaptation data

PC−1
PC−2
PC−M

ML

w

w1

wM

...
Eigenvoice: Training

- A set of speaker-dependent (SD) acoustic hidden Markov models (HMMs) are trained from each speaker
  - in general, the HMM states are GMMs

- A speaker’s voice is represented by a speaker supervector that is composed by concatenating the mean vectors of all his HMM Gaussian distributions
  - $R$ states in each HMM
  - $x_i = [x'_i 1, \ldots, x'_i R]'$

- PCA is then performed on a set of training speaker supervectors and the resulting eigenvectors are called eigenvoices
The new speaker’s supervector $s$ is assumed to be a linear combination of the $M$ leading eigenvoices $\{v_1, \ldots, v_M\}$

$$s = s^{(ev)} = \sum_{m=1}^{M} w_m v_m$$

Given the adaptation data $O = \{o_1, o_2, \ldots, o_T\}$, estimate the eigenvoice weights $(w = [w_1, \ldots, w_m]')$ by maximum likelihood

$$\max_w Q(w) \equiv -\frac{1}{2} \sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_t(r) \|o_t - s_r(w)\|^2_{C_r}$$

- $\gamma_t(r)$: posterior probability of observation sequence being at state $r$ at time $t$
- $C_r$: covariance matrix of the Gaussian at state $r$
- $s_r$: $r$th constituent of $s$
Kernel Principal Component Analysis

- Kernel PCA: linear PCA in the feature space

- Given \( \{x_1, \ldots, x_N\} \), construct \( K = [k(x_i, x_j)] = [\varphi(x_i)'\varphi(x_j)] \)

- \( K = U\Lambda U' \) (assume that \( \{\varphi(x_1), \ldots, \varphi(x_N)\} \) has been centered)
  - \( U = [\alpha_1, \ldots, \alpha_N] \) with \( \alpha_i = [\alpha_{i1}, \ldots, \alpha_{iN}]' \)
  - \( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_N) \)

- \( k \)th orthonormal eigenvector: \( v_k = \sum_{i=1}^{N} \frac{\alpha_{ki}}{\sqrt{\lambda_k}} \varphi(x_i) \)
Problem

- Estimation of the eigenvoice weights requires the evaluation of the distances between adaptation data $o_t$ and Gaussian means of the new speaker in the observation space.

- EV: breaks up the speaker-adapted (SA) model found by EV adaptation into its constituent HMM Gaussians:
  
  \[ s^{(ev)} \rightarrow s_1^{(ev)}, \ldots, s_R^{(ev)} \rightarrow \text{Gaussian means} \]

- KEV: the SA model found by KEV adaptation resides in the feature space, not in the input speaker supervector space:
  
  - cannot access each constituent Gaussian directly.
Composite Kernel

\[ k(x_i, x_j) = f(k_1(x_{i1}, x_{j1}), \ldots, k_R(x_{iR}, x_{jR})) \]
Examples

- **Direct sum kernel:**
  \[
  k(x_i, x_j) = \sum_{r=1}^{R} k_r(x_{ir}, x_{jr})
  \]
  - corresponding feature: \( \varphi(x_i) = [\varphi_1(x_{i1})', \ldots, \varphi_R(x_{iR})']' \)

- **Tensor product kernel:**
  \[
  k(x_i, x_j) = \prod_{r=1}^{R} k_r(x_{ir}, x_{jr})
  \]

- If \( k_r(\cdot, \cdot)'s \) are valid Mercer kernels, so is \( k(\cdot, \cdot) \)
New Speaker in the Feature Space

\[
\varphi(s) = \sum_{m=1}^{M} w_m v_m = \sum_{m=1}^{M} \sum_{i=1}^{N} \frac{w_m \alpha_{mi}}{\sqrt{\lambda_m}} \varphi(x_i)
\]

- \(r\)th constituent: \(\varphi_r(s_r) = \sum_{m=1}^{M} \sum_{i=1}^{N} \frac{w_m \alpha_{mi}}{\sqrt{\lambda_m}} \varphi_r(x_{ir})\)

- Similarity between \(\varphi_r(s_r)\) and \(\varphi_r(o_t)\):

\[
k_r(s_r, o_t) = \varphi_r(s_r)' \varphi_r(o_t) = A(r, t) + \sum_{m=1}^{M} \frac{w_m}{\sqrt{\lambda_m}} B(m, r, t)
\]

\[- A(r, t) = \frac{1}{N} \sum_{j=1}^{N} k_r(x_{jr}, o_t)\]

\[- B(m, r, t) = \left( \sum_{i=1}^{N} \alpha_{mi} k_r(x_{ir}, o_t) \right) - A(r, t) \left( \sum_{i=1}^{N} \alpha_{mi} \right)\]
Maximum Likelihood Adaptation

- $k_r(\cdot, \cdot)$: e.g., isotropic kernels $k_r(s_r, o_t) = \kappa(\|o_t - s_r\|^2_C)$
  - e.g., Gaussian kernels: $k_r(s_r, o_t) = \exp(-\beta\|o_t - s_r\|^2_C)$
  - if $\kappa$ is invertible, $\|o_t - s_r\|^2_C \rightarrow$ function of $k_r(s_r, o_t) \rightarrow$ function of $w$

- Substitute back to $Q(w)$ and differentiate to obtain $\partial Q / \partial w_j$

- No closed form solution for the optimal $w$
  - use generalized EM algorithm (GEM)

- $w(0)$: eigenvoice weights of the supervector composed from the speaker-independent model $x^{(si)}$
  - $w_m(0) = v'_m \varphi(x^{(si)})$ (can be obtained from kernel evaluations)
Incorporate the SI Model

- Interpolate $\varphi(s)$ with the $\varphi$-mapped SI supervector $\varphi(x^{(si)})$ to obtain the final SA model (in the feature space):

  $$\varphi^{(rkev)}(s) = w_0 \varphi(x^{(si)}) + (1 - w_0) \varphi(s), \quad 0 \leq w_0 \leq 1$$

  - $w_0$ estimated in the same manner as the other $w_m$’s
  - robust kernel eigenvoice

- $\varphi^{(rkev)}(s)$ contains components in $\varphi(x^{(si)})$ from eigenvectors beyond the $M$ selected kernel eigenvoices for adaptation

  - preserve the speaker-independent projections on the remaining less important but robust eigenvoices in the final speaker-adapted model
Experimental Setup: Data Set and HMM Models

- TIDIGITS corpus
  - 163 speakers (of both genders) in each (training and test) set, each pronouncing 77 utterances of 1-7 digits (out of: “0”, “1”, ..., “9”, and “oh”)

- 12 mel-frequency cepstral coefficients and the normalized frame energy from each speech frame of 25 ms at every 10 ms

- Digit model
  - strictly left-to-right HMM with 16 states
  - one Gaussian with diagonal covariance per state

- A 3-state “sil” model to capture silence speech and a 1-state “sp” model to capture short pauses between digits
Adaptation

• SD digit model
  – one for each training speaker
  – variances and transition matrices are borrowed from SI models (only the Gaussian means are estimated)

• The “sil” and “sp” models are simply copied to the SD model

• 5, 10, 20 digits for adaptation (∼ 2.1s, 4.1s, and 9.6s of speech)

• Results are averages of 5-fold cross-validation over all test speakers

• (Testing) word accuracy of SI model: 96.25%
**Experiment 1: Number of Kernel Eigenvoices**

- KEV outperforms the SI model even with only two eigenvoices
- Robust KEV significantly improves KEV
Experiment 2: KEV vs. EV

![Graph showing word recognition accuracy vs. amount of adaptation data for robust KEV, KEV, robust EV, and EV.]
• (Robust) KEV always performs better than (robust) EV

• When only 2.1s or 4.1s of adaptation data are available
  \[ \text{EV} \approx \text{MAP} \approx \text{MLLR} < \text{SI} \approx \text{robust EV} < \text{KEV} < \text{robust KEV} \]

• With 9.6s of adaptation data
  – MLLR works marginally better than robust KEV (by an absolute 0.06%)

• Word error rate reduction over SI

<table>
<thead>
<tr>
<th></th>
<th>KEV</th>
<th>robust KEV</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1s</td>
<td>16.0%</td>
<td>27.5%</td>
</tr>
<tr>
<td>4.1s</td>
<td>21.3%</td>
<td>31.7%</td>
</tr>
<tr>
<td>9.6s</td>
<td>21.3%</td>
<td>33.3%</td>
</tr>
</tbody>
</table>
Eigenvoice Speaker Adaptation via Composite Kernel PCA

NIPS 2003
Conclusion and Future Work

- (Nonlinear) kernel PCA + composite kernel
  - better eigenvoices $\rightarrow$ improved speaker adaptation

- Interpolate the SI model with the speaker model found by KEV

- In the TIDIGITS task
  - standard EV does not help
  - KEV outperforms SI by 16–21% (word error rate reduction)
  - robust KEV: 28–33% word error rate reduction over SI

- Disadvantage: KEV is slower than EV
  - online computation of many kernel functions required during subsequent speech recognition
  - currently investigating speed-up techniques