A tutorial on speaker verification

Student: Jun Wang
Supervisor: Thomas Fang Zheng
Tsinghua University, 2014, 4
Outline

• Introduction

• GMM-UBM framework of speaker verification

• The ivector methodology of speaker verification

• Intersession compensation and scoring method for ivector

• Toolkits and database

• Some of my previous work

• References
1 Introduction

• Speaker recognition is a technique to recognize the identity of a speaker from a speech utterance.

  - Spk recognition
    - Spk identification
      - Spk verification
    - Text dependent
      - Text independent
      - Close set
      - Open set

• My research area focus on the open-set, text-independent speaker verification.
A multitude of researches have been conducted to address the following three fields:

- Speech parameterization
- Pattern matching
- Scoring method

fig1 main research fields in speaker recognition
Speech parameterization consists in transforming the speech signal to a set of feature vectors. Most of the speech parameterizations used in speaker verification systems rely on a cepstral representation of speech. [F. Bimbot, 2004]

fig2 modular representation of mfcc feature extractor
Main approaches in pattern matching for speaker recognition

- Template matching
  - Nearest neighbor [A. Higgins, 1993]
  - Vector quantization [F. Soong, 1985]
  - Gaussian Mixture Model [A. Reynolds, 2003]
  - Joint factor analysis [P. Kenny, 2006]
  - ivector [N. Dehak, 2011]
  - time delay neural work [Y. Bennani, 1991]

- Probabilistic model
  - Artificial Neural Network
Performance measure

- For speaker identification:
  \[
  \text{Recognition Rate} = \frac{\text{number of correct recognition}}{\text{total number of trials}}
  \]

- For speaker verification:
  \[
  \text{False Reject Rate} = \frac{\text{number of rejective true speaker}}{\text{total number of true speaker}}
  \]
  \[
  \text{False Acceptance Rate} = \frac{\text{number of accepted imposter}}{\text{total number of imposter}}
  \]
  \[
  \text{EER} = \text{False Reject Rate} = \text{False Acceptance Rate}
  \]

Detection error tradeoff (DET) curve is often used to describe the performance.
Cost function (\(C_{DET}\)) is also defined as a weighted sum of FAR and FRR. [NIST, 2008]
Speaker verification [S. Furui, 1981; D. A. Reynolds, 2003]: to verify a speech utterance belongs to a specified enrollment, accept or reject.
• GMM-UBM framework [D. A. Reynolds, 2000]

- Gaussian Mixture Model is used to modeling the probability density function of a multi-dimensional feature vector.

- Given a speech feature vector $X=\{x_i\}$ of dimension $F$, the probability density of $x_i$ given a C GMM speaker model $\lambda$ is given by:

$$p(x_i|\lambda) = \sum_{c=1}^{C} w_c g(x_i, \mu_c, \Sigma_c)$$

$$\sum_{c=1}^{C} w_c = 1$$
The UBM is trained using EM algorithm and a speaker GMM is established by adjusting the UBM parameters by MAP.

fig4 modeling methods for GMM-UBM
• From distribution:
  
  ➢ A speaker utterance is represented by GMM which is adapted from the UBM via MAP.
  
  \[ M = m + Dz \]

  ➢ UBM \( m \) represents all acoustic and phonetic variations in speech data where \( m \) is a supervector with dimension \( CF \).

  ➢ \( D \) is diagonal matrix in full space \( (CF \times CF) \) and \( z \) is normally distributed random vector with dimension \( CF \).

  ➢ \( M \sim N(m, DD^T) \).
Over recent years, ivector has demonstrated state-of-the-art performance for speaker verification.
• Jonit factor analysis [P. Kenny, 2007]

- JFA is a model of speaker and session variability in GMMs.

\[ M = m + Vy + Ux + Dz \]

- where \( m \) is a speaker- and session-independent supervector with CF dimension. (UBM)

- \( M \) is a speaker- and channel- dependent supervector.

\[ m = [:]_{CF \times 1} \quad M = [:]_{CF \times 1} \]
\[ M = m + V y + U x + D z \]

- \( V \) and \( D \) define a speaker subspace, and \( U \) defines a session subspace.

\[
V = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_c \end{bmatrix}_{CF \times R} \quad U = \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_c \end{bmatrix}_{CF \times L} \quad D = \begin{bmatrix} \Sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_c \end{bmatrix}_{CF \times CF}
\]

- The vector \( y, z \) and \( x \) are assumed to be a random variable with a normally distribution \( N(0, I) \).

- \( z \) is a normally distributed CF dimension random vector.
• i-vector [N. Dehak, 2011]

- make no distinction between speaker effects and session effects in GMM supervector space.
- define a total variability space, contains speaker and session variabilities simultaneously.

\[ M = m + Tw \]

- \( M \sim N(m, TT^T) \)
- \( w \sim N(0, I) \)
\[ M = m + Tw \]

\[ T = \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_c \end{bmatrix}_{CF \times R} \]

\[ m = [i]_{CF \times 1}, \quad M = [i]_{CF \times 1}, \quad w = [i]_{R \times 1} \]

T is a low rank \( CF \times R \) subspace that contains the eigenvectors with the largest eigenvalues of total variability covariance matrix.

\[ w \sim N(0, I) \]
Training and testing procedure for i-vector

1. **Train UBM**
2. **Train T matrix**
3. **Extractor i-vector**
4. **Training speech i-vector**
5. **Enrollment i-vector**
6. **Test i-vector**

**Steps:**
- **Training speech**
- **Enrollment speech**
- **Test speech**

**Flow:**
- Training speech → Feature extractor → Extractor i-vector → Training speech i-vector
- Enrollment speech → Feature extractor → Extractor i-vector → Enrollment i-vector
- Test speech → Feature extractor → Extractor i-vector → Test i-vector

**Diagram:**
- Fig6 training and testing procedure for i-vector
• Object function

- $M = m + Tw$
- $M \sim N(m, TT^T)$
- Suppose $x_i \sim N(M, \Sigma)$, $x_i = m + Tw + \varepsilon$
- For Gaussian Mixture Model, $x_{i,c} = m_c + T_c w + \varepsilon_c$
- $\mathcal{L} \sim p(x_i | \lambda)$
- Define object function: $\mathcal{L} = \prod_c p(x_{i,c} | \lambda)$
• i-vector extraction [N. Dehak, 2011]

- The Baum Welch statistics needed to estimate a given speech utterance:

  \[ N_c = \sum_t P(c|x_t) \]

  \[ F'_c = \sum_t P(c|x_t)x_t \]

  \[ F_c = \sum_t P(c|x_t)(x_t - m_c) \]
• **i-vector extraction [N. Dehak, 2011]**

  - The ivector of a speech segment $X$ is computed as the mean of the posterior probability $P(w|X)$.

  $$P(w|X) \sim N(\bar{w}, \Xi)$$

  $$\bar{w} = \Xi T^T \Sigma^{-1} F$$

  $$\Xi = (I + \sum_c T_c^T \Sigma_c^{-1} N_c T_c)^{-1}$$
• **T matrix training [N. Dehak, 2011]**

  ➢ T matrix can be trained by an EM procedure.
  ➢ E steps computes the posterior probability $P(w|X)$.
  ➢ M step optimizes $T$ by updating following formula:

  $$T_c = \left( \sum_u F_c(u) \bar{w}^T \right) \left( \sum_u N_c(u) (\bar{w} \bar{w}^T + \Xi) \right)$$
T matrix training [N. Dehak, 2011]

\[
T_c = (\sum_u F_c(u) \bar{w}^T)(\sum_u N_c(u)(\bar{w}\bar{w}^T + \Xi))
\]

\[
T_c = \begin{bmatrix}
\vdots \\
\vdots \\
\vdots \\
\end{bmatrix}_{F \times R} \\
T = \begin{bmatrix}
T_1 \\
T_2 \\
\vdots \\
T_C \\
\end{bmatrix}_{CF \times R}
\]
4 Intersession compensation and scoring method for i-vector

- WCCN
- LDA
- PLDA
- NAP
- EFR
- sphNorm

Cosine distance scoring
PLDA scoring

fig7 intersession compensation and scoring method for i-vector
• Cosine distance [N. Dehak, 2009]

Using cosine kernel between the target speaker ivector and test speaker ivector.

\[ \text{score}(\omega_1, \omega_2) = \frac{\omega_1^T \omega_2}{\sqrt{\omega_1^T \omega_1 \omega_2^T \omega_2}} \]
• **WCCN [A. Hatch, 2006]**

- to minimize the classification error.

- \( k(\omega_1, \omega_2) = \omega_1^t R \omega_2 \)

- \( R = W^{-1} \quad W^{-1} = BB^T \)

- \( W = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (\omega_i^s - \overline{\omega_s})(\omega_i^s - \overline{\omega_s})^t \)

- \( \omega' = B^t \omega \)
LDA [K. Fukunaga, 1990; N. Dehak, 2009]

- to seek new orthogonal axes to better discriminate different classes.

- a linear transformation that maximizes the between-class variation while minimizing the within-class variances.

- fisher criterion is used for this purpose.
• LDA [K. Fukunaga, 1990; N. Dehak, 2009]

- $S_b$ is between-class covariance matrix, and $S_w$ is the within-class covariance matrix. The solution $v$ is generalized eigenvectors.

\[
J(v) = \frac{v^t S_b v}{v^t S_w v} \quad \text{Reyleigh coefficient}
\]

- $S_b = \sum_{s=1}^{S} (w_s - \bar{w})(w_s - \bar{w})^t$

- $S_w = \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (\omega_i^s - \bar{\omega}_s)(\omega_i^s - \bar{\omega}_s)^t$

- $S_b v = \lambda S_w v$

- $\omega' = A^t \omega$
• PLDA [S. J. D. Prince, 2007]

Technically, assuming a factor analysis (FA) model of the i-vectors of the form:

\[ w = \mu + Fh + Gy + \varepsilon \]

in practice G always equals to zero.

First computes the maximum likelihood estimate (MLE) of the factor loading matrix F (the Eigenvoice subspace).

Here, \( w \) is the i-vector, \( \mu \) is the mean of training i-vectors, and \( h \sim N(0, I) \) is a vector of latent factors. The full covariance residual noise term \( \varepsilon \) explains the variability not captured through the latent variables.
PLDA [S. J. D. Prince, 2007]

- Given a pair of ivectors \( D = \{w_1, w_2\} \), \( H_1 \) means two vectors from the same speaker and \( H_0 \) means two vectors from different speakers. [P. Kenny, 2010]

- The verification score is computed for all possible model-test i-vector trials. The scores are computed as the log-likelihood ratio between the same (\( H_1 \)) versus different (\( H_0 \)) speaker models hypotheses:

\[
llr = \ln \frac{p(w_1, w|H_1)}{p(w_1|H_0) \cdot p(w_2|H_0)}
\]
5 Toolkits and database

- Kaldi toolkits [D. Povey, 2011]
- Database:
  - trials: NIST SRE08 female core test, contains 1997 females, 59343 trials.
  - lda/plda training data: fisher English database, contains 7196 females, 13827 sessions.
  - UBM training data: fisher English database, 6000 sessions female speech data.
setup:

mfcc features, extracting with 20ms hamming window, every 10ms, 19 mel-frequency cepstral coefficient together with log energy were used. Delta and delta-delta coefficient were then calculated to produce 60-dimensional feature vector.

2048 Gaussian Mixtures, gender-dependent.

400-dimensional ivector.

150-dimensional lda/plda.
SRE 8 results with kaldi: core test, female

<table>
<thead>
<tr>
<th>EER(%)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>cosine</td>
<td>28.77</td>
<td>4.78</td>
<td>28.60</td>
<td>21.32</td>
<td>20.43</td>
<td>11.36</td>
<td>7.35</td>
<td>7.63</td>
</tr>
<tr>
<td>LDA</td>
<td>24.10</td>
<td>1.79</td>
<td>24.18</td>
<td>14.56</td>
<td>14.42</td>
<td>10.25</td>
<td>6.46</td>
<td>6.58</td>
</tr>
<tr>
<td>PLDA</td>
<td>20.09</td>
<td>2.09</td>
<td>20.43</td>
<td>17.87</td>
<td>13.34</td>
<td>8.37</td>
<td>4.44</td>
<td>4.74</td>
</tr>
</tbody>
</table>

condition:
1. All trials involving only interview speech in training and test
2. All trials involving interview speech from the same microphone type in training and test
3. All trials involving interview speech from different microphones types in training and test
4. All trials involving interview training speech and telephone test speech
5. All trials involving telephone training speech and noninterview microphone test speech
6. All trials involving only telephone speech in training and test
7. All trials involving only English language telephone speech in training and test
8. All trials involving only English language telephone speech spoken
6 Some of my previous work

- Sequential Model adaptation for Speaker Verification
- Block-wise training for ivectors
- Phone-based alignment for channel robust speaker verification ......
- Mlp classification for ivector ......
- ......
References

THANK YOU!