The HKCUPU System for The NIST 2010 Speaker Recognition Evaluation

Weiwu JIANG1, Man-Wai MAK2, Wei RAO2 and Helen MENG1

1Dept. of System Engineering and Engineering Management, The Chinese University of Hong Kong
2Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University

Introduction

This paper presents the HKCUPU speaker recognition system submitted to NIST 2010 speaker recognition evaluation (SRE). The system comprises five subsystems, each with different acoustic features, session-variability reduction methods, speaker modeling and scoring methods and classifiers. This paper reports the results of individual and fusion systems for the core test and highlights the improvements made by our newly proposed JFA-Fishervoicematrix (FSH) subsystem. Results show that FSH outperforms JFA when its projection matrix is channel-dependent (telephone or microphone) and that FSH is complementary to other state-of-the-art techniques. It was also found that VAD is an important pre-processing step for interview speech.

Subsystems

<table>
<thead>
<tr>
<th>Subsystem</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>JFA</td>
<td>Joint Factor Analysis</td>
</tr>
<tr>
<td>JSV</td>
<td>Linear SVM on JFA-GMM supervectors</td>
</tr>
<tr>
<td>JSF</td>
<td>Cosine-kernel SVM on JFA speaker factors</td>
</tr>
<tr>
<td>GSV</td>
<td>Linear SVM on GMM-supervectors with NAP</td>
</tr>
<tr>
<td>FSH</td>
<td>Cosine distance on Fishervoicematrix-projected speaker factors</td>
</tr>
</tbody>
</table>

Voice Activity Detection

<table>
<thead>
<tr>
<th>Type of Speech</th>
<th>VAD Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telephone</td>
<td>Energy-based VAD for GSV subsystem</td>
</tr>
<tr>
<td>ETSI Adaptive Multi-Rate (AMR) VAD</td>
<td>for other four subsystems</td>
</tr>
<tr>
<td>Microphone and Interview</td>
<td>Spectral subtraction followed by energy-based VAD with crosstalk removal</td>
</tr>
</tbody>
</table>

Acoustic Features

<table>
<thead>
<tr>
<th>Subsystem</th>
<th>Features and Dimension</th>
<th>Frame Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>JFA</td>
<td>MFCC, Delta, Delta-Delta (51Dim)</td>
<td>25ms</td>
</tr>
<tr>
<td>JSV</td>
<td>MFCC, Delta, Delta-Delta (51Dim)</td>
<td>25ms</td>
</tr>
<tr>
<td>JSF</td>
<td>PLP, Delta, Delta-Delta (52 Dim)</td>
<td>20ms</td>
</tr>
<tr>
<td>GSV</td>
<td>PLP, Delta, Delta-Delta (24 Dim)</td>
<td>25ms</td>
</tr>
<tr>
<td>FSH</td>
<td>PLP, Delta, Delta-Delta (52 Dim)</td>
<td>20ms</td>
</tr>
</tbody>
</table>

- GSV: Cepstral mean normalization followed by feature warping.
- Others: Mean-variance-normalization followed by feature warping.

Subsystem Description

JFA Subsystem:
- 2048-Gaussian GMM-supervector of speaker $s$:
  \[ m(s) = m + \frac{1}{2}V(s) + Ux(s,h) \]
- \[ m(c) = m + \frac{1}{2}V(c) \]
- Linear SVM Scoring:
  \[ S_{\text{JFA}}(s,c) = a_s(m(s),m(c)) + \sum_i a_i(m_i(b_i),m(c)) + d_i \]
  where $h_i$ is the $i$th background speaker and $SV$ contains the support vector indexes of background factors.

JSV Subsystem:
- Use GMM-supervectors of target speaker $s$ and claimant $c$ obtained from JFA as feature vectors:
  \[ m(s) = m + \frac{1}{2}V(s) \]
  \[ m(c) = m + \frac{1}{2}V(c) \]
- Cosine-kernel SVM Scoring:
  \[ S_{\text{JSV}}(s,c) = \alpha_sK(y(s),y(c)) + \sum_i \alpha_iK(y(h_i),y(c)) + d_i \]

JSF Subsystem:
- Use speaker-factors of target speaker $s$ and claimant $c$ obtained from JFA as feature vectors:
  \[ m(s) = m + \frac{1}{2}V(s) + Ux(s,h) \]
  \[ m(c) = m + \frac{1}{2}V(c) + Ux(c,h) \]
- Cosine-kernel SVM Scoring:
  \[ S_{\text{JSF}}(s,c) = \alpha_sK(y(s),y(c)) + \sum_i \alpha_iK(y(h_i),y(c)) + d_i \]

GSV Subsystem:
- Use MAP adapted 512-Gaussian GMM-supervectors.
- Transformed by NAP matrices (Corank = 16 for tel; Corank = 128 for mic/interview).
- Linear SVM Scoring with T-norm.

FSH Subsystem:
- Apply JFA speaker factors $y(s)$ as feature vectors to estimate a nonparametric Fisher discriminant projection matrix $W$:
  \[ W = W_1 W_2 W_3 \]
  where $W_1$ is PCA projection matrix, $W_2$ is whitened-class projection matrix, and $W_3$ is nonparametric between-class projection matrix. $W_3$ focuses on the boundaries between speakers without using parametric models to approximate the distribution of $y(s)$.
  \[ W_3 = \arg \max_i \| W_1^T W_2 \Sigma W_2^T W_1 \| \]
  \[ W_2 = \arg \max_s \| W_1^T S W_1 \| \]
  \[ W_2^T S W_2 = I \]
  \[ W_3 = \Phi \Lambda^{-\frac{1}{2}} \]
- Cosine distance scoring:
  \[ S_{\text{FSH}}(s,c) = \frac{\langle W_3 y(s), W_3 y(c) \rangle}{\| W_3 y(s) \| \| W_3 y(c) \|} \]

Training Data

UBM:
- Tel – NIST04, 05, 06 tel speech
- Mic – NIST 05, 06 mic speech

JFA:
- Matrix $F$ – NIST04, 05, 06, Switchboard Phase2, Phase 3, Cellular Parts 2 (300 speaker factors)
- Tel Matrix $U$ – NIST04, 05, 06 tel speech (100 Ch factors)
- Mic Matrix $U$ – NIST05, 06 mic speech (75 Ch factors)
- Interview Matrix $U$ – NIST08 interview speech (75 Ch factors)
- Totally rank of $U$ = 100 tel + 75 mic + 75 interview

Fishervoice:
- Projection matrix $W$ – NIST04, 05, 06 tel speech (400 gender-dependent speakers with 8 different utterances)

NAP:
- Tel – NIST04, 05, 06 tel speech (Corank = 16)
- Mic/Interview – NIST05, 06 mic speech, NIST08 interview speech (Corank = 128)

SVSM Imposter-Class:
- JSV and JSF – NIST05, 06, 08, Switchboard Cellular Parts 2
- GSV – NIST 06 tel speech and NIST 06 mic speech

Score normalization:
- Tnorm for JSV, JSF and GSV – NIST04-06 tel/mic speech
- Tnorm for JFA and FSH – NIST04-06 tel/mic speech

Conclusions

The HKCUPU system submitted to NIST 2010 SRE is composed of 5 subsystems. The fusion system reduces the EER by 42% and minimumDCF by 56% when compared with the best individual subsystems. It was also found that the newly proposed FSH subsystem is complementary to JFA and performs significantly better than JFA when its projection matrices were trained by the type of speech that matches the evaluation conditions.