Using Second Order Statistics for Text Independent Speaker Verification

Ran D. Zilca

Work performed while with Amdocs R&D

June 19th 2001
Problem Description

- Open set speaker identification
  - Text independent
  - Telephony conversational speech
  - Large population (>1000 speakers)
- Large call volume
- Rapid scoring required (normalized)
Proposed Solution

2 Phase Scoring

M Speakers → Rapid Scoring CM/ULS → N Best → GMM Scoring → Best Score

June 19th 2001
2-Phase Accuracy

Speaker Detection – N Best 2-Phase

- Accurate System (GMM) (N = all speakers)
- Rapid Scoring (CM/ULS) (N = 1)
CM / ULS

Rapid Scoring (Speaker Verification)

June 19th 2001
CM / ULS Rapid Scoring

- Covariance Modeling (CM)
  - Single Gaussian models
  - Shape only = Covariance only
  - Zero mean (CMS) ➔ Covariance only

- CM / FLS = Frame Level Scoring
  - Single Gaussian GMM (likelihood ratio)
  - Scoring against a new model = repeating frame by frame scoring
\[ S = - \frac{D(C_{utt}, C_s)}{D(C_{utt}, C_w)} \]

- \( C_{utt} \) = Utterance Covariance
- \( C_s \) = Speaker Covariance
- \( C_w \) = World Covariance (UBM)
CM / ULS

- Distortion measure between Covariance matrices
- Previously used for closed set ID (Gish 90, Bimbot and Mathan 93, Campbell 97, Faundez-Zanuy 2000)
- No reported switchboard/NIST experiments
- Extension to normalize scores required (verification)
Distortion Measures

- Divergence Shape (Campbell)

\[ DS_{1,2} = DS(C_1, C_2) = \frac{1}{2} tr[(C_1 - C_2)(C_2^{-1} - C_1^{-1})] \]

- Sphericity Measure (Bimbot et. al.)

\[ SM_{1,2} = SM(C_1, C_2) = \frac{1}{2} tr(C_1 C_2^{-1}) tr(C_2 C_1^{-1}) \]
Orientation vs. Exact Position

June 19th 2001

Observation
Computational Complexity

Some Examples

June 19th 2001
Voiceprint Size

- **18th order feature vector:**
  - CM: 171
  - GMM-512: 18,944
  - GMM-2048: 75,776

- **24th order feature vector:**
  - CM: 300
  - GMM-512: 25,088
  - GMM 2048: 100,352

June 19th 2001
<table>
<thead>
<tr>
<th></th>
<th>Enrollment</th>
<th>Verification</th>
</tr>
</thead>
<tbody>
<tr>
<td>SG (GMM-1)</td>
<td>0.07</td>
<td>1.67</td>
</tr>
<tr>
<td>DSR</td>
<td>0.07</td>
<td>0.01</td>
</tr>
<tr>
<td>GMM - 512</td>
<td>1320</td>
<td>13.44</td>
</tr>
</tbody>
</table>

Comparison of CPU time [seconds], for enrollment and verification
Settings

- 18\textsuperscript{th} order MFCC (from 19 filter banks)
- No band-limiting
- No Delta’s
- No Pre-emphasis
- “Aggressive” voice detection based on spectral peaks (LSF differences - IHM)
- Gender and handset dependent UBM’s trained on NIST-99 test data
Tested Systems

- GMM
  - Bayesian adaptation (Reynolds 97)
  - 512 components
  - 5 component approximation
- SG (Single Gaussian)
- DSR (Divergence Shape Ratio)
- SMR (Sphericity Measure Ratio)
Electret only (models + trials)
Handset Type

- Se – Te
- Se – Tc
- Sc – Te
- Sc – Tc

June 19th 2001
Cohort vs. UBM

- Motivation: inexpensive scoring
- Method:
  \[ S = - \frac{D(C_{utt}, C_s)}{\frac{1}{B} \sum_{b=1}^{B} D(C_{utt}, C_b)} \]
- Cohort Speakers from NIST-99:
  - 857 male-electret
  - 591 male-carbon
  - 1449 female-electret
  - 523 female-carbon
Cohort vs. UBM

Speaker Detection Performance

- SMR - UBM
- SMR - Cohort

Miss probability (in %)
- False Alarm probability

June 19th 2001
Recent Experiments

Different Front End
ZNORM, TNORM

June 19th 2001
Different Front End

Speaker Detection Performance

- MFCC 19 – 18, spectral envelope VOX
  - DCF = 0.0807
  - EER = 21%

- MFCC 23 – 12 + Δ’s, 2-GMM logE VOX
  - DCF = 0.0832
  - EER = 21%

June 19th 2001
June 19th 2001

ZTNORM

Speaker Detection Performance: $E_t / E_t$

Miss probability (in %)

- MFCC 19 – 18, spectral envelope VOX
  - DCF = 0.0807
  - EER = 21%

- MFCC 23 – 12 + $\Delta$’s, 2-GMM logE VOX
  - DCF = 0.0789
  - EER = 20%

False Alarm probability
Conclusions

- CM/ULS rapid scoring method
- ~20% EER (15% EER for GMM)
- Applications:
  - Open set ID (standalone, two-phase)
  - Strict computational environments
  - Quick experimental evaluation of new features
- Robustness to CB handsets
- SMR Better than 1-component GMM!
Future Work

- Expand to multi-modal / GMM (SMR better than 1-component GMM!)
  - Faundez-Zanuy (Icassp-2001): VQ-CM (with Sphericity Measure)
    Closed set ID, no priors/weights
- Improve SMR accuracy – different features / RASTA / Warping