ON THE USE OF AUDITORY AND AUTOMATIC SYSTEMS TO HANDLE MISMATCHED CONDITIONS IN FORENSIC SPEAKER RECOGNITION

Anil Alexander †, Damien Dessimoz‡, Filippo Botti‡, and Andrzej Drygajlo †

† Swiss Federal Institute of Technology, Lausanne  
   Signal Processing Institute  
‡ University of Lausanne  
   School of Criminal Sciences
• Bayesian interpretation in Forensic Automatic Speaker Recognition
• Strength of evidence in Aural and Automatic Speaker Recognition
• Evaluating the strength of evidence
• Comparing performances of Aural and Automatic Speaker Recognition Systems
• Necessity of adapting the automatic systems to different conditions
• Discussion of the perceptual cues used by laypersons to recognize speakers
• Conclusion
Bayesian Interpretation in Forensic Automatic Speaker Recognition

- **Evidence (E)**
  
The score obtained comparing statistical model of the suspect’s voice and a questioned recording (trace)

\[ H_0 \] – The two recordings have the same source

\[ H_1 \] – The two recordings have a different source

**Likelihood Ratio (LR)**

The relative probability of observing a particular score "E", with respect to two competing hypotheses

\[
LR = \frac{p(E | H_0)}{p(E | H_1)}
\]
Experimental Framework

• **Test Database** *(subset of <<polyphone IPSC-02>>)*
  - **Speakers** *(Swiss-French)*
    - 5 traces for each speaker and condition (PSTN, GSM and Noisy PSTN)
    - 1 suspect reference recording for each speaker and condition (PSTN and GSM)

• **Recording Lengths**
  - 15 second / trace: simulation of real cases (undisguised hoaxes, menacing calls etc)
  - 90 seconds / reference recording

• **Testing Scenarios Evaluated**
  - Reference PSTN vs Traces PSTN
  - Reference PSTN vs Traces GSM
  - Reference PSTN vs Traces Noisy PSTN
  - Reference GSM vs Traces GSM
Aural Speaker Recognition

Experimental Framework

• Listeners
  » 90 listeners whose mother-tongue is French
  » Laypersons with no phonetic training
  » Same computer and headphones

• Training
  » No limitation on the number of listening trials

• Testing
  » Verbal scores scale from 1 through 7
  » Perceptual cues
Perceptual Verbal Scale

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I am sure that the two speakers are not the same</td>
</tr>
<tr>
<td>2</td>
<td>I am almost sure that the two speakers are not the same</td>
</tr>
<tr>
<td>3</td>
<td>It is possible that the two speakers are not the same</td>
</tr>
<tr>
<td>4</td>
<td>I cannot decide</td>
</tr>
<tr>
<td>5</td>
<td>It is possible that the two speakers are the same</td>
</tr>
<tr>
<td>6</td>
<td>I am almost sure that the two speakers are the same</td>
</tr>
<tr>
<td>7</td>
<td>I am sure that the two speakers are the same</td>
</tr>
</tbody>
</table>

Perceptual Cues

Subjects asked to note factors they considered in recognizing speakers at the end of each session.
Strength of Evidence for Aural Recognition

Likelihood Ratio ($LR$) = Ratio of the heights on the histograms for the two hypotheses at the point "E"

- Discrete scores
- Histograms used to estimate the probabilities of scores for each hypothesis
Automatic Speaker Recognition System

• Recognition System
  – Feature extraction: RASTA -PLP
  – Statistical modeling: Gaussian Mixture Modeling (GMM)
  – Likelihood ratio (Kernel Density Estimation)

• Methodology
  – Single questioned recording and single suspect recording
  – Suspect Reference (R) and Traces (T) databases – Subsets of « Polyphone IPSC-02 »)
Strength of Evidence for the Automatic System

- Scores are continuous
- Kernel-density-based estimate of the probability density of scores for each hypothesis

\[ LR = \frac{P(E|H_0)}{P(E|H_1)} \]

LR = Ratio of heights on the curves for the two hypotheses at the point "E"
Tippett Plots

- Representation of the proportion of likelihood ratios greater than a given LR, for cases corresponding to hypotheses $H_0$ and $H_1$, i.e. $P(LR(H_1) > LR)$

- Separation between curves representing $H_0$ and $H_1$ indicates how well the system differentiates between cases in which each of the two hypotheses is known to be true
Evaluating Strength of Evidence in Matched Conditions

Ref. PSTN vs Traces PSTN

# Similar separations between curves for aural and automatic systems
Evaluating Strength of Evidence in Mismatched Conditions

Ref. PSTN vs Traces Noisy PSTN

Better curve separation in aural recognition

Better evaluation of LR for aural recognition in mismatched conditions
Adaptation for noisy conditions results in the improvement of performance of automatic recognition.
Performance Measurement
(in terms of Recognition Errors)

• Are the trends shown by the Tippett plots for the evaluation of the strength of evidence also shown by measuring the recognition errors in speaker verification?

  – Detection Error Tradeoff (DET) Curves
    • Relative plot of False Match Rate and False Non-Match Rate varying a decision point

  – Equal Error Rate
    • when False Match Rate = False Non-Match Rate on DET curve (useful to compare system performances)
Matched Conditions: Comparing Aural and Automatic

Ref PSTN – Trace PSTN  
Ref GSM - Trace GSM

In matched conditions automatic performs better than aural recognition.
Mismatched Conditions: Comparing Aural and Automatic

Mismatched Conditions → Automatic recognition shows similar or slightly degraded performance as compared to Aural recognition → Adaptation necessary

Mismatched Conditions:

- PSTN - GSM
  - Aural (PSTN-GSM)
  - Automatic (PSTN-GSM)

- PSTN – Noisy PSTN
  - Aural (PSTN-Noise)
  - Automatic (PSTN-Noise)
After adaptation for noisy conditions, automatic system shows similar or better performance compared to aural recognition.

# Adapted using Spectral Subtraction Based on Minimum statistics (R Martin 1994)
EER : Aural and Automatic Recognition

<table>
<thead>
<tr>
<th>EER</th>
<th>PSTN-PSTN</th>
<th>PSTN-GSM</th>
<th>PSTN-Noise</th>
<th>GSM-GSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aural</td>
<td>16%</td>
<td>30%</td>
<td>26%</td>
<td>21%</td>
</tr>
<tr>
<td>Automatic</td>
<td>12%</td>
<td>36%</td>
<td>36%</td>
<td>4%</td>
</tr>
</tbody>
</table>

- Both aural and automatic recognition perform better in matched than in mismatched conditions
- Automatic recognition performs better than aural in matched conditions
- Aural recognition performs better than automatic in mismatched conditions

Adaptability to different conditions - Necessity for automatic systems
Perceptual Cues Used by Laypersons

• Human beings use a lot of perceptual cues to recognize speakers
  e.g. pronunciation, timbre, intonation, rate of speech
  breathing, loudness, imagined physiognomy, etc.

• These cues are used in aural recognition, to recognize speakers in different conditions

• Evaluating relative perceived importance of these perceptual cues
Relative Importance of Perceptual Cues

<table>
<thead>
<tr>
<th></th>
<th>PSTN-PSTN</th>
<th>PSTN-GSM</th>
<th>PSTN-Noise</th>
<th>GSM-GSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accent, pronunciation, articulation 30%</td>
<td>34%</td>
<td>31%</td>
<td>30%</td>
<td>26%</td>
</tr>
<tr>
<td>Timbre 24%</td>
<td>25%</td>
<td>25%</td>
<td>22%</td>
<td>22%</td>
</tr>
<tr>
<td>Intonation 20%</td>
<td>16%</td>
<td>24%</td>
<td>18%</td>
<td>21%</td>
</tr>
<tr>
<td>Rate of Speech 10%</td>
<td>9%</td>
<td>7%</td>
<td>12%</td>
<td>12%</td>
</tr>
<tr>
<td>Speech Defects 7%</td>
<td>6%</td>
<td>7%</td>
<td>8%</td>
<td>5%</td>
</tr>
</tbody>
</table>

# The perceived importance of these perceptual cues is relatively constant in different conditions for human listeners
Conclusions

- In matched recording conditions of training and testing, automatic recognition systems performed better than aural systems.

- In mismatched conditions, baseline automatic systems showed comparable or slightly degraded performance compared to aural systems.

- Aural recognition relies on high-level perceptual cues to recognize speakers.

- Baseline automatic speaker recognition systems should be adapted to each of the mismatched conditions to improve performance.
Questions?

Thank you for your attention.