A human benchmark for automatic speaker recognition

Milou van Dijk, Rosemary Orr, David van der Vloed and David A. van Leeuwen

University College Utrecht, The Netherlands
Netherlands Forensic Institute, The Hague
Radboud University Nijmegen, The Netherlands

15 October 2013
Two main areas of research where identity of speaker plays a role

- Automatic Speaker Recognition
- Forensic Speaker Comparison
Two main areas of research where identity of speaker plays a role

- Automatic Speaker Recognition
- Forensic Speaker Comparison

In recent years, we have reached the insight that the

- task
- presentation of the answer

are, in fact, the same for both areas.
Task and presentation of the answer

Task

Compare two segments of speech w.r.t. the identity of the speaker

- forensic: trace and suspect
- automatic: train and test

Presentation of the answer

Probabilistically, as a likelihood ratio

\[ r = \frac{P(\text{speech segments} \mid H_1)}{P(\text{speech segments} \mid H_2)} \]

Hypotheses \( H_1, H_2 \):

- forensic: Prosecutor's and Defense hypothesis \( H_p, H_d \)
- automatic: target and non-target hypothesis
Task and presentation of the answer

Task

Compare two segments of speech w.r.t. the identity of the speaker

- forensic: trace and suspect
- automatic: train and test

Presentation of the answer

probabilistically, as a *likelihood ratio*

\[ r = \frac{P(\text{speech segments} \mid H_1)}{P(\text{speech segments} \mid H_2)} \]

Hypotheses \(H_1, H_2\):  

- forensic: *Prosecutor’s* and *Defense* hypothesis \(H_p, H_d\)
- automatic: *target* and *non-target* hypothesis
How the answer is used

Forensic speaker comparison

To separate contribution of the evidence $E$ from speech from all the other evidence $I$ in the posterior odds

$$
\frac{P(H_p | E, I)}{P(H_d | E, I)} = \frac{P(E | H_p, I)}{P(E | H_d, I)} \times \frac{P(H_p, I)}{P(H_d, I)}
$$

judge/jury wants to know given by expert other evidence
How the answer is used

**Forensic speaker comparison**

To separate contribution of the evidence $E$ from speech from all the other evidence $I$ in the posterior odds

$$\frac{P(H_p \mid E, I)}{P(H_d \mid E, I)} = r \times \frac{P(H_p, I)}{P(H_d, I)}$$

judge/jury wants to know given by expert other evidence

**In automatic speaker recognition**

To minimize the Bayes’ risk: expected cost of decisions using cost function $DCF(C_{FA}, C_{miss}, P_{tar})$

decide same speaker iff

$$r > \frac{1 - P_{tar}}{P_{tar}} \frac{C_{FA}}{C_{miss}}$$
This is the *same* likelihood ratio!
Forensic Speaker Comparison

- Mostly qualitative statements
- In search of typicalities (higher weight of evidence)
- By human listening (auditory) or visual inspection spectogram (acoustic)
- Aware of electro-acoustical differences trace and reference samples
- If LR calculation, based on feature population frequencies
The differences

Forensic Speaker Comparison

- Mostly qualitative statements
- In search of typicalities (higher weight of evidence)
- By human listening (auditory) or visual inspection spectogram (acoustic)
- Aware of electro-acoustical differences trace and reference samples
- If LR calculation, based on feature population frequencies

Automatic Speaker Recognition

- Concentrating on large homogeneous databases
- Most more interested in *discrimination* than *calibration* aspects
- Blind application, little sample quality control
- If LR calculation, based on development test statistics
Since *task* and *presentation* are the same, can we compare the two, given the same data?

### Human benchmark

<table>
<thead>
<tr>
<th>Year</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>1937</td>
<td>McGehee. Line-up size 5, live material</td>
</tr>
<tr>
<td>2011</td>
<td>Ramos. NIST SRE 2010 Human Assisted Speaker Recognition (HASR) hard material</td>
</tr>
<tr>
<td>2011</td>
<td>Kahn. NIST SRE 2010 HASR</td>
</tr>
</tbody>
</table>

Only Ramos (2011) considered calibration aspects
Goals

• Same material as in NIST SRE
• Same overall difficulty of trials
• Study the score-to-likelihood-ratio characteristics
• Method

Realistic experimental boundary conditions

• Naïve listeners (no forensic experts in this study)
• Mostly non-native listeners
• Pooled scores: joint performance characteristics
• Just quick holistic auditory comparison, no detailed forensic auditory-acoustic analysis
Speech material characteristics

General
- NIST SRE 2010 material
- Male speakers
- Telephone
- English
- Conversations
- $\sim$ 5 min
Trial selection

Controlling ‘difficulty’

- Compute all train-test scores using RUN automatic system
- Self-calibrate the scores to give well-calibrated log-likelihood-ratios
- Choose target and non-target trials around three regions:
  1. Hard trials, around log $r = 0$
  2. Representative trials, in mode of distribution
  3. Easy trials
Distribution

- Target prior is 50%
  - \( \Rightarrow \) likelihood ratio becomes posterior odds
    - Subjects are told this
    - Subjects are not expected to count their answers
- 1280 unique trials, 50% target, 50% non-target
- 40 subjects each doing 32 trials
  - 4 hard, 24 representative, 4 easy trials
  - same ratios as overall PDF
- every subject same difficulty of trials
Task

- Comparison of 2 speech segments
- Control over current playing samples
- Play-back continues where left off
- No further play-back control
- No spectrographic information
- Subject-paced
- 10-point answer scale
Responses: a verbal scale

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>certain the same</td>
<td>the same</td>
</tr>
<tr>
<td>4</td>
<td>very confident the same</td>
<td>the same</td>
</tr>
<tr>
<td>3</td>
<td>confident the same</td>
<td>the same</td>
</tr>
<tr>
<td>2</td>
<td>uncertain the same</td>
<td>the same</td>
</tr>
<tr>
<td>1</td>
<td>very uncertain the same</td>
<td>the same</td>
</tr>
<tr>
<td>1</td>
<td>very uncertain different</td>
<td>different</td>
</tr>
<tr>
<td>2</td>
<td>uncertain different</td>
<td>different</td>
</tr>
<tr>
<td>3</td>
<td>confident different</td>
<td>different</td>
</tr>
<tr>
<td>4</td>
<td>very confident different</td>
<td>different</td>
</tr>
<tr>
<td>5</td>
<td>certain different</td>
<td>different</td>
</tr>
</tbody>
</table>
Subject recruitment

- 40 subjects from University College Utrecht
  - naïve w.r.t. speaker recognition
  - most non-native English speakers
  - living/studying in English-speaking community
- Recruitment using contemporary social media
- No payments, sessions lasted $\sim 30$–45 minutes
- No hearing problems reported, abilities not tested
Experimental protocol

- Experiment conducted using ordinary laptop
- Audio through high quality headset
- Quiet office environment
- Protocol

  - Instructions on screen
  - One screen emphasizing *50% trials are “same speaker”*
  - 6 trials habituation (easy trials, no feedback)
  - 32 trials for experiment, no feedback
  - Short debriefing with performance feedback in terms of Equal Error Rate
Experiential results

Graph

- Responses transformed to numeric scale $-4.5, \ldots, 4.5$
- Pooled subjects
- Pooled difficulty

Observe

- Response ratios increase with score
- Not monotonously rising
- Not many “very uncertain” responses
Pooling over subject and difficulty

- Set thresholds at $-5, \ldots, 5$
- Compute false alarm and miss probabilities
- Plot the Convex Hull of $(P_{FA}, P_{miss})$
- Lower-Left: lower errors is better
Graph

- Same information as ROC
- Axes transformed using probit()
- Data split along difficulty

Observe

- Easier: lower errors
- All $\approx$ representative
Influence of decision speed

Graph
- EER per subject, vs average decision time $t$ per trial

Observe
- Small dependence

$$E_e = a - bt$$
$$b = 0.35\%/s$$

Not shown in graph

<table>
<thead>
<tr>
<th>Time</th>
<th>difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.8 s</td>
<td>hard</td>
</tr>
<tr>
<td>17.8 s</td>
<td>representative</td>
</tr>
<tr>
<td>15.6 s</td>
<td>easy</td>
</tr>
</tbody>
</table>
The results allow us to compute the LR for every response value $s$

1. **Maximum Likelihood estimates**

$$r = \frac{P(s | \text{same speaker})}{P(s | \text{different speaker})} = \frac{\# \text{ responses same speaker score } s}{\# \text{ same speaker trials}} / \frac{\# \text{ responses different speaker score } s}{\# \text{ different speaker trials}}$$
The results allow us to compute the LR for every response value $s$

1. Maximum Likelihood estimates

\[ r = \frac{P(s \mid \text{same speaker})}{P(s \mid \text{different speaker})} \]

\[ = \frac{\# \text{ responses same speaker score } s}{\# \text{ same speaker trials}} \div \frac{\# \text{ responses different speaker score } s}{\# \text{ different speaker trials}} \]

2. Constraining monotonous score-to-likelihood

\[ r = -\text{slope of ROC-CH line segment corresponding to } s \]
Score-to-likelihood-ratio mapping

Graph
- Black: ROC-CH constrained to monotonous
- Red: Maximum Likelihood probabilities
- Log of LR!

Observe
- Methods give similar LR
- Fairly *linear* score-to-log-likelihood-ratio
- *Low magnitude* log LRs
Why are the log-likelihood-ratios so small?

Log LRs are in range $-1.5$ to $1.5$

$\Rightarrow$ LR in range $0.23 < r < 5.4$.

Several reasons for this:

- Overall performance is relatively poor ($E_\text{=} = 13\%$ for easy trials)
- Subjects use individually different calibration
  - Subject 1’s “confident” may mean something else than Subject 2’s “confident”
  - Pooling these judgements leads to poorer discrimination performance
  $\Rightarrow$ Lower magnitude log LRs after calibration
- Psychological effect to want to use full scale of confidence (“certain”) despite little reason for such outspoken confidence
- Limited number of trials
Per-subject calibration: self-calibration

Method 1: self-calibration
- calibrate per-subject ROC on all subject’s data
- pooling calibrated LRs
- *cheating*, really

Method 2: cross validation
- Use first 16 trials per subject to train calibration of responses
- apply these to second 16 trials
- and *vise versa*
Cross-validation experiment:

- Still limited LR range \(0.21 < r < 7.3\)
- Consistent with findings Daniel Ramos (2011)

Further:

- EER not smaller than with uncalibrated responses

\[\Rightarrow\] 16 trials too few for individual calibration
Do subjects want to use full response scale?

If so, we’d expect them to gradually use wider range of responses

- Split trials for every subject
- 1st and 2nd half
- expect *increase*
  - range / width distribution
  - variance responses
  - mean absolute value
- none found
- also not in quarters or other time partitions
Limitations in the number of trials

There were ‘only’ 1280 trials in the test

- ROC-CH calibration (a.k.a. PAV) limits the magnitude to the log LR
- “Laplace’s’ rule of succession” effectively adds 1 response “very very certain” in ML probability
  - to both ends of the response scale
  - for both target and non-target trials
- this means, the minimum / maximum LRs are $\frac{1}{641}$ and 641

- but we observe much smaller magnitude LLRs
Conclusions

- Naïve listener’s speaker discrimination performance is not great
  - \( E_\text{=} = 26.5 \% \)
  - pooled responses, average 18 s exposure time
- Response-to-log-likelihood-ratio transformation is remarkably linear
- verbal extreme of “certainty” only corresponds to \( LR \sim 5 \)