Sheep, Goats, Lambs and Wolves
An Analysis of Individual Differences in Speaker Recognition Performance

11/23/98

These are the slides of the NIST coordinated presentation for the ICSLP '98 Conference in Sidney, Australia, November 1998. They are not part of the official Proceedings of this Conference.

Click here to start

Table of Contents

Sheep, Goats, Lambs and Wolves, An Analysis of Individual Differences in Speaker Recognition Performance

The Hypothetical Menagerie

The Question

The Task and the Data

Speaker Recognition Performance

The Goat Test Statistic, nonparametric Kruskal-Wallis rank sum test

The Hypothetical Menagerie

Lamb/Wolf Test Statistics

Statistical Significance

Distribution of Errors versus Animal Rankings

Can a Lamb be also a Wolf?

Can a Sheep be also a Wolf?

Author: George R. Doddington

Home Page: http://www.nist.gov/speech
Can a Sheep be also a Lamb?

Conclusion: It’s a ZOO out there!
Sheep, Goats, Lambs and Wolves

An Analysis of Individual Differences in Speaker Recognition Performance

George Doddington\textsuperscript{1,2,3,5}, Walter Liggett\textsuperscript{1}, Alvin Martin\textsuperscript{1}, Mark Przybocki\textsuperscript{1}, Douglas Reynolds\textsuperscript{3,4}

\textsuperscript{1} National Institute of Standards and Technology
\textsuperscript{2} The Johns Hopkins University
\textsuperscript{3} U.S. Department of Defense
\textsuperscript{4} MIT Lincoln Laboratory
\textsuperscript{5} SRI International
The Hypothetical Menagerie

- Typical speakers: The well-behaved majority.
  - Sheep: Speakers who exhibit **good true speaker acceptance**.

- Problem speakers: The troublesome minorities.
  - Goats: Speakers who are **exceptionally unsuccessful at being accepted**.
  - Lambs: Speakers who are **exceptionally vulnerable to impersonation by others**.
  - Wolves: Speakers who are **exceptionally successful at impersonating others**.
Conclusion: It’s a zoo out there!

Goats, Lambs and Wolves exist!

Recommendation:

• Direct research toward the minimization of these undesirable animals.

• Include measurement of speaker differences as part of performance evaluation.

An html copy of this presentation is available at URL http://www.nist.gov/speech/icslp_98
Sheep, Goats, Lambs and Wolves An Analysis of Individual Differences in Speaker Recognition Performance

George Doddington1,2,3,5, Walter Liggett1, Alvin Martin1, Mark Przybocki1, Douglas Reynolds3,4

1 National Institute of Standards and Technology 2 The Johns Hopkins University 3 U.S. Department of Defense 4 MIT Lincoln Laboratory 5 SRI International
The Question

- Do these animals really exist? That is, do the speakers in the population of speakers truly exhibit individual differences?

- OR -

- Are the observed performance differences merely a result of statistical variance (due to handset, message content, acoustic environment, etc.)?
The Hypothetical Menagerie

● **Typical speakers: The well-behaved majority.**
  - Sheep: Speakers who exhibit good true speaker acceptance.

● **Problem speakers: The troublesome minorities.**
  - Goats: Speakers who are exceptionally unsuccessful at being accepted.
  - Lambs: Speakers who are exceptionally vulnerable to impersonation by others.
  - Wolves: Speakers who are exceptionally successful at impersonating others.
The Task and the Data

Text-independent speaker verification: accept speaker if score \( \geq \) threshold (if \( x \geq \theta \)).

- 30-second test segments (3400 segments, total)
- Conversational telephone speech
- 200 men and 220 women (no cross-sex trials)
- Different handsets used in training and test
- Data restricted to electret handsets

The Question

- Do these animals really exist? That is, do the speakers in the population of speakers truly exhibit individual differences?

- OR -

- Are the observed performance differences merely a result of statistical variance (due to handset, message content, acoustic environment, etc.)?
The Task and the Data

Text-independent speaker verification: accept speaker if score $\geq$ threshold (if $x \geq q$).

- 30-second test segments (3400 segments, total)
- Conversational telephone speech
- 200 men and 220 women (no cross-sex trials)
- Different handsets used in training and test
- Data restricted to electret handsets

The Goat Test Statistic

nonparametric Kruskal-Wallis rank sum test

- Each model speaker, $M_i$, contributes $n_i$ test segments ($n_i \geq 5$).
- Each of these segments is scored against its respective model.

<table>
<thead>
<tr>
<th>Model Speaker</th>
<th>Segment Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_1$</td>
<td>$x_{11}$ $x_{12}$ $\cdots$</td>
</tr>
<tr>
<td>$M_2$</td>
<td>$x_{21}$ $x_{22}$ $\cdots$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$M_i$</td>
<td>$x_{li}$ $x_{li}$ rank sum $x_{lim}$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$M_m$</td>
<td>$x_{m1}$ $x_{m2}$ $\cdots$</td>
</tr>
</tbody>
</table>

Test Procedure

- Replace each score by its rank among all scores: $x_{ij} \Rightarrow r_{ij}$
- Sum the segment ranks for each model speaker: $R_i = \sum_{j=1}^{n_i} r_{ij}$
- Compute the Kruskal-Wallis test statistic:
  \[
  H = \left\{ \frac{12}{N(N+1)} \cdot \frac{\sum_{i=1}^{m} R_i^2}{n_i} \right\} - 3(N+1)
  \]
  where $N = \sum_{i=1}^{m} n_i$
The Hypothetical Menagerie

- **Typical speakers:** The well-behaved majority.
  - *Sheep:* A speaker who exhibits **good true speaker acceptance.**

- **Problem speakers:** The troublesome minorities.
  - *Goat:* A speaker who is **exceptionally unsuccessful at being accepted.**
  - *Lamb:* A **model** speaker who is **exceptionally vulnerable to impersonation by others.**
  - *Wolf:* A **segment** speaker who is **exceptionally successful at impersonating others.**
The Goat Test Statistic nonparametric Kruskal-Wallis rank sum test

- Each model speaker, $M_i$, contributes $n_i$ test segments ($n_i \geq 5$).
- Each of these segments is scored against its respective model.

Test Procedure

- Replace each score by its rank among all scores: $x_{ij} \rightarrow r_{ij}$.
- Sum the segment ranks for each model speaker:
- Compute the Kruskal-Wallis test statistic:
**Lamb/Wolf Test Statistics**

- Each segment speaker, $S_i$, contributes $n_i$ test segments
- Each of these segments is scored against all models

<table>
<thead>
<tr>
<th>Model Speaker</th>
<th>$M_1$</th>
<th>$M_2$</th>
<th>$M_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>$x_{11}$</td>
<td>$x_{21}$</td>
<td>$x_{m1}$</td>
</tr>
<tr>
<td>$S_2$</td>
<td>$x_{12}$</td>
<td>$x_{22}$</td>
<td>$x_{m2}$</td>
</tr>
<tr>
<td>$S_s$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Test Procedure**

- Compute average impostor scores, $x_{ij}$, for all $S_i$ and $M_j$
- Replace each score by its rank among all scores.
- Sum the ranks for each putative animal:
  - **Lamb**: Row sum
  - **Wolf**: Column sum
The Hypothetical Menagerie

- **Typical speakers: The well-behaved majority.**
  - Sheep: A speaker who exhibits good true speaker acceptance.

- **Problem speakers: The troublesome minorities.**
  - Goat: A speaker who is exceptionally unsuccessful at being accepted.
  - Lamb: A model speaker who is exceptionally vulnerable to impersonation by others.
  - Wolf: A segment speaker who is exceptionally successful at impersonating others.
Statistical Significance

Goatish, lambish and wolfish behaviors were all demonstrated with confidence > 99%

<table>
<thead>
<tr>
<th>The Kruskal-Wallis rank sum test statistic ($\chi^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D.F.</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td><strong>Goat Test</strong></td>
</tr>
<tr>
<td><strong>Lamb Test</strong></td>
</tr>
<tr>
<td><strong>Wolf Test</strong></td>
</tr>
</tbody>
</table>
Lamb/Wolf Test Statistics

- Each segment speaker, $S_i$, contributes $n_i$ test segments
- Each of these segments is scored against all models

Test Procedure

- Compute average impostor scores, $x_{ij}$, for all $S_i$ and $M_j$.
- Replace each score by its rank among all scores.
- Sum the ranks for each putative animal:
  - Lamb: Row sum
  - Wolf: Column sum
Distribution of Errors

versus Animal Rankings

Cumulative Errors
Misses for Model Speakers
False Alarms for Model Speakers
False Alarms for Segment Speakers

Cumulative Trials
ordered by Goat/Lamb/Wolf rank
Goatish, lambish and wolfish behaviors were all demonstrated with confidence > 99%
Can a Lamb be also a Wolf?

**Wolf Average**
Impostor scores for a segment speaker averaged over the best scoring 10% of model speakers

**Lamb Average**
Impostor scores for a model speaker averaged over the best scoring 10% of segment speakers

Slide 11 of 14

Distribution of Errors versus Animal Rankings

Cumulative Errors
Misses for Model Speakers False Alarms for Model Speakers False Alarms for Segment Speakers
Cumulative Trials
ordered by Goat/Lamb/Wolf rank
Can a Sheep be also a Wolf?

Wolf Average
Imposter scores for a segment speaker averaged over the best scoring 10% of model speakers

Model Average
True Speaker scores for a model speaker averaged over all model speaker segments
Can a Lamb be also a Wolf?

Wolf Scores versus Lamb Scores
Wolf Average
Impostor scores
for a segment speaker
averaged over
the best scoring 10%
of model speakers
Lamb Average
Impostor scores for a model speaker averaged over
the best scoring 10% of segment speakers
R2 = 0.63
R2 = 0.34
Can a Sheep be also a Lamb?

Lamb Average
Impostor scores for a model speaker averaged over the best scoring 10% of segment speakers

Model Average
True Speaker scores for a model speaker averaged over all model speaker segments

Slide 13 of 14
Can a Sheep be also a Wolf?

Wolf Average
Impostor scores
for a segment speaker averaged over the best scoring 10% of model speakers

Model Average
True Speaker scores for a model speaker
averaged over all model speaker segments

Wolf Scores versus Model Scores
R2 = 0.31
R2 = 0.03
Can a Sheep be also a Lamb?

Lamb Average
Impostor scores
for a model speaker averaged over the best scoring 10% of segment speakers
Model Average
True Speaker scores for a model speaker
averaged over all model speaker segments
Lamb Scores versus Model Scores
R² = 0.01
R² = 0.09
Conclusion: It’s a out there!

Goats, Lambs and Wolves exist!

Recommendation:

- Direct research toward the minimization of these undesirable animals.
- Include measurement of speaker differences as part of performance evaluation.

An html copy of this presentation is available at URL http://www.nist.gov/speech/icslp_98