The Evaluation of Speaker Recognition Technology

• a challenge
• an opportunity


**Presentation Outline**

- **The Game**
  - Applications
  - Task definition

- **The Challenge**
  - Problem dimensions
  - Evaluation factors

- **The Opportunity**
  - Technology development
  - Task definition
Types of Speaker Recognition Applications

• Those that benefit *the speaker*
  – Granting a personal privilege.
    This typically occurs in physical entry control or information access control applications.

• Those that benefit *someone else*
  – Gathering information.
    This typically occurs in forensic applications and other intelligence gathering kinds of applications.
Types of Speaker Recognition Tasks

- **Speaker Verification**
  - 2-class decision problem
  - Given a reputed identity
  - Did the reputed person say it?

- **Speaker Identification**
  - N-class decision problem
  - Who said it?
The Technical Task for Access Control Applications

- **Speaker Verification**
  - Training
    - Build a model of each user’s speech data
  - Usage
    1. User requests access and offers identity
    2. User speaks
    3. System accepts (or rejects) proffered identity
  - Performance
    - Measure error probabilities, $P_{\text{miss}}$ and $P_{\text{false\_alarm}}$, as a function of acceptance threshold
The Technical Task for Forensic Applications

• Speaker **Verification**
  - Training
    • Build a model of the target’s speech data
  - Usage
    • System computes confidence of the target hypothesis
  - Performance
    • Measure error probabilities, $P_{\text{miss}}$ and $P_{\text{false_alarm}}$, as a function of confidence
Types of Speech

• Text-dependent
  – Access control applications
  – The speaker is cooperative
  – usually little speech data (time is precious)

• Text-\textit{in}dependent
  – Forensic applications
  – The speaker is \textit{not} cooperative
  – often lots of speech data
The Technical Challenge: Robust Recognition

😊 Similarity 😊 versus 😞 Variability 😞

– Among Speakers
  • sex
  • dialect
  • size
  • age

– The Speaker
  • health
  • emotions
  • metabolism
  • bio-drift, aging

– The Channel
  • microphone, noise and distortion
Evaluation Objectives

• To support R&D
  – What are the important issues?
  – Which of my modeling/algorithmic “improvements” actually improve performance?

• To assess application readiness
  – Will speaker recognition technology support this application?

• To measure operational performance
  – Why isn’t the system working well enough?
**Evaluation Design**

- **Define** the speaker recognition task.
- **Create** a test corpus to *accurately represent* the actual speaker recognition problem.
  - represent all factors and conditions of interest
- **Collect** a *sufficient sample* of data to provide statistically significant results for all factors and conditions of interest.
- **Measure** performance and analyze for all factors and conditions of interest.
How much test data is required?

- The “Rule of 30”:

  To be 90 percent confident that the true error rate is within +/- 30 percent of the observed error rate, there must be at least 30 errors.

- This assumes statistically independent trials. But how is this done?
  - Speaker selection
  - Microphone selection

- And which factors are to be evaluated?
Key Evaluation Factors

• Speakers
  – to study population performance characteristics

• Sessions

• Microphones

• Amount of training data
  – # of seconds, # of sessions

• Amount of test data
  – # of seconds
It’s a Zoo out there...
**The Speaker 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.**
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
Conditional Evaluation of Performance

• Measure true speaker performance as a function of:
  ✓ amount of test/training data
  ✓ sex
  • health, voice pitch
  • noise, channel conditions

• Condition impostor trials on:
  • the sex, pitch, age, dialect, size . . .
    . . . of the true speaker
    . . . of the impostor
Speaker Recognition

Evaluation Measures

- Speaker Verification is a Detection Problem
- Evaluation is in Terms of Detection Errors

- Detection Error Trade-off – the DET plot
- Equal Error Rate – EER
- Geometric Mean Error – GME
- Detection Cost – $C_{DET}$
Detection Error Trade-off: “The DET Plot”
**Equal Error Rate**

\[ E_{\text{miss}} = E_{\text{false\_alarm}} \]
Geometric Mean Error

\[ E_{GM} = \sqrt{P_{\text{miss}} \cdot P_{\text{false\_alarm}}} \]
Detection Cost

• Model the Expected Cost (Value) of a Detection:

\[
C_{\text{DET}} = C_{\text{miss}} \cdot P_{\text{miss}} \cdot P_{\text{target\_spkr}} \\
+ C_{\text{false\_alarm}} \cdot P_{\text{false\_alarm}} \cdot P_{\text{impostor}}
\]

where

- \(C_{\text{miss}}\) = the cost of a miss
- \(C_{\text{false\_alarm}}\) = the cost of a false alarm
- \(P_{\text{miss}}\) = the conditional probability of a miss
- \(P_{\text{false\_alarm}}\) = the conditional probability of a false alarm
- \(P_{\text{target\_spkr}}\) = the \textit{a priori} probability of the target speaker
- \(P_{\text{impostor}}\) = 1 - \(P_{\text{target\_spkr}}\)
Constant Cost Lines on the DET plot
Pooling Results across Speakers

- Speaker-Specific decision thresholds -- post facto choice of decision thresholds for each speaker -- **NO!**
  - Estimating correct thresholds is part of the *task*
  - Optimistic bias from limited data
- Speaker Normalization -- *compute normalization from training data*
  - One global decision threshold
  - Post facto choice of a single global threshold is less of a factor, but doing so still gives results an optimistic bias and ignores the essential and nontrivial task of choosing the threshold.
The NIST Open Evaluations

• Text-independent speaker detection
• $\leq 2$ minutes of training
• $\leq 1$ minute test segment duration
• Hundreds of speakers
• Conversational telephone speech
• **Detection Cost** used as evaluation measure
General NIST findings

• Performance improves with
  – more training data
  – longer test segments

• Performance degrades with
  – channel variations (microphone and line)
  – channel degradation (noise and distortion)
  – voice pitch deviations of the true speaker

• Performance is independent of sex
What does the Future Offer?†

- Lots of Potential Application Opportunities, courtesy of *the information age*!
- Advanced Speaker Recognition Technology, by more comprehensive speaker modeling.
  - Capitalize on explosion in computing *power*.
  - Use *more* speaker training data.
  - Exploit *temporal* speaker characteristics.
  - Become *familiar* with the target speaker.

† Avignon 1998
Speaker Information of Word Bigrams

The graph shows the relationship between the number of occurrences of bigrams and the speaker information of these bigrams (in bits). Common phrases such as "how shall", "shall I", "you bet", "it were", "in terms", "so forth", "uh-huh uh-huh", "um-hum um-hum", "<start> sure", "sort of", "<start> yeah", "<start> you know", and "<end>" are highlighted on the graph.
The NIST Extended Data Task

- Text-independent speaker detection
- > 10 minutes of training
- > 2 minutes test segment duration
- Hundreds of speakers
- Conversational telephone speech
- Detection Cost used as evaluation measure
Speaker Detection Performance versus # of training conversations

Speaker Detection based on Word bigram Statistics -- bigram-count $\geq$ 200

- Ntrain=1
- Ntrain=2
- Ntrain=4
- Ntrain=8
- Ntrain=16

Miss probability (in %)

False Alarm probability (in %)