An Overview of the NIST Series of Speaker Recognition Evaluations and Technologies (1996-2014)

Joaquin Gonzalez-Rodriguez
ATVS-Universidad Autonoma de Madrid
joaquin.gonzalez@uam.es     http://atvs.ii.uam.es
Valletta, yesterday
Speaker Recognition ... in dogs!
An Overview of the NIST Series of Speaker Recognition Evaluations and Technologies (1996-2014)

Joaquin Gonzalez-Rodriguez
ATVS-Universidad Autonoma de Madrid
joaquin.gonzalez@uam.es http://atvs.ii.uam.es
Outline

- Speaker Recognition (SpkRec) and Biometrics
- Assessment and NIST SpkRec evaluation (SRE) protocols
- Early NIST SREs (1996-2001)
  - GMM-UBM
- Going Higher (2002-2005)
  - Supra-segmental systems
  - High-dimensionality spectral systems
- Big Data (2006-2012)
  - Factor Analysis and i-vectors
- NIST i-vector Challenge (2014)
- Conclusions
Speaker Recognition and Biometrics
Speaker information in the speech signal

“... no es la primera vez que me hacen este chiste ...”
Multilevel Speaker Information

- Dependent on speech and linguistic acts
  - Analog to writer identification from written text
- Skpr information embedded at multiple levels
  - *Short-term* (20-40ms): acoustic/spectral, particular realization of “sounds”
  - *Coarticulation* (50-300ms): spectro-temporal contours
  - *Prosody*: intonation (pitch) & energy contours, durations, speech rate
  - *Voice quality*: voice settings (modal, harsh, breathy, creaky …) and glottal source characterization
  - *Idiolect*: use of phrases, words, …
Sources of Variability of Spkr Info

- **Speaker intrinsic factors**
  - Time lapse between recordings
  - Linguistic content and type of conversation
  - Speaking style and emotional conditions
  - Health (cold, dysphonia ...) and aging

- **Speaker extrinsic factors**
  - Microphone/handset and channel/coding effects
  - Noise and reverberation

- **Intersession variability**
  - Accounts for main channel and speaker variability between recording sessions
Assessment and NIST Speaker Recognition Evaluation (SRE) Protocols
NIST Speaker Recognition Evaluations

- Organized by the Speech Group at the (US) National Institute of Standards and Technology

- Main task: speaker DETECTION
  - Whether or not a given target speaker is speaking in a given speech segment
  - Other tasks: segmentation, adaptation, HASR …

- NIST provides:
  - Evaluation plans
  - Development and test sets on specific speech corpora
  - Unified measurements of error (cost)
  - Forum for discussion among participants
Telephone Corpora: Switchboard & Mixer I

- Recorded and distributed by LDC (Linguistic Data Consortium, University of Pennsylvania)
- Conversational telephone speech, assigned topics
- Five minutes calls, ~25 calls/speaker
- Handset variability (carbon vs electret)
- Different phases of Switchboard & Mixer I:
  - Landline (local and long distance calls)
  - cellular
  - wireless
Corpora for NIST SRE's (e.g., 2004)

<table>
<thead>
<tr>
<th>Year</th>
<th>Corpus</th>
<th>Detection Tasks</th>
<th>Unique Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>SWBD I</td>
<td>1sp lim</td>
<td>USA coverage</td>
</tr>
<tr>
<td>1997</td>
<td>SWBD II phase 1</td>
<td>1sp lim</td>
<td>Mid-Atlantic</td>
</tr>
<tr>
<td>1998</td>
<td>SWBD II phase 2</td>
<td>1sp lim</td>
<td>Mid-West</td>
</tr>
<tr>
<td>1999</td>
<td>SWBD II phase 3</td>
<td>1sp lim</td>
<td>South</td>
</tr>
<tr>
<td>2000</td>
<td>recycled p1 &amp; p2</td>
<td>1sp lim</td>
<td>---</td>
</tr>
<tr>
<td>2000</td>
<td>AHUMADA</td>
<td>1sp lim</td>
<td>Spanish</td>
</tr>
<tr>
<td>2001</td>
<td>repeat of 2000</td>
<td>1sp lim</td>
<td>---</td>
</tr>
<tr>
<td>2001</td>
<td>SWBD I</td>
<td>1sp ext</td>
<td>---</td>
</tr>
<tr>
<td>2002</td>
<td>SWBD cellular p1</td>
<td>1sp lim</td>
<td>Cellular GSM</td>
</tr>
<tr>
<td>2002</td>
<td>SWBD p2 &amp; p3</td>
<td>1sp ext</td>
<td>---</td>
</tr>
<tr>
<td>2002</td>
<td>FBI Voice DB</td>
<td>1sp mm</td>
<td>Multi Modal</td>
</tr>
<tr>
<td>2003</td>
<td>SWBD cellular p2</td>
<td>1sp lim</td>
<td>Cellular CDMA</td>
</tr>
<tr>
<td>2003</td>
<td>repeat of 2002</td>
<td>1sp ext</td>
<td>---</td>
</tr>
<tr>
<td>2004</td>
<td>MIXER</td>
<td>1sp var</td>
<td>Multi-language/</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>transmission types</td>
</tr>
</tbody>
</table>

Table 1: Corpora used for various NIST Speaker Recognition evaluations. Abbreviations: “lim” for limited-data, “ext” for extended-data, “var” for limited and extended combined, “mm” for multi-modal, and “p” for phase.

<table>
<thead>
<tr>
<th>Other Language</th>
<th>Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>52</td>
</tr>
<tr>
<td>Mandarin</td>
<td>46</td>
</tr>
<tr>
<td>Russian</td>
<td>48</td>
</tr>
<tr>
<td>Spanish</td>
<td>79</td>
</tr>
<tr>
<td>English only</td>
<td>85</td>
</tr>
<tr>
<td>Total</td>
<td>310</td>
</tr>
</tbody>
</table>

Table 6: Target speakers included in the 2004 evaluation data by language spoken in addition to English.

<table>
<thead>
<tr>
<th>Language</th>
<th>Training Sides</th>
<th>Test Sides</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>2515</td>
<td>907</td>
</tr>
<tr>
<td>Arabic</td>
<td>300</td>
<td>96</td>
</tr>
<tr>
<td>Mandarin</td>
<td>238</td>
<td>64</td>
</tr>
<tr>
<td>Russian</td>
<td>274</td>
<td>61</td>
</tr>
<tr>
<td>Spanish</td>
<td>99</td>
<td>48</td>
</tr>
</tbody>
</table>

Table 7: Numbers of conversation sides, by language being spoken, included as training or test segment data in the 2004 evaluation.
Evaluation measures

- **Test trials:** given a speech file, determine if a given speaker is actually speaking
  - **Target trials:**
    - The specified speaker is speaking in the test segment
  - **Impostor (non-target) trials:**
    - The specified speaker is NOT speaking in the test segment

- **Participants submit two outputs per trial:**
  - **Actual decision**
    - True / False
  - **Likelihood Score**
    - The higher the score, the higher the support to the same-speaker hypothesis
DET plots

- $P_{\text{miss}}$ and $P_{\text{FA}}$ are plotted on the $x$ and $y$ axes on a normal deviate scale.
- Gaussian score distributions plotted as straight lines.
- Allows easier system comparison (e.g. low error rates).
Cost Function (1996-2008)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost of a miss</td>
<td></td>
</tr>
<tr>
<td>Cost of a false alarm</td>
<td></td>
</tr>
<tr>
<td>Probability of a target</td>
<td></td>
</tr>
<tr>
<td>Probability of a non-target</td>
<td></td>
</tr>
</tbody>
</table>
Evaluation Measures

- $C_{\text{DET}}$ is the primary measure to rank systems
  - FA errors are 10 times more relevant than Misses
    - Even though $C_{\text{miss}} = 10 \times C_{\text{FA}}$ (as $P_{\text{non-target}} = 0.99$)
  - Measures the goodness of both discrimination and threshold setting

- As participants provide:
  - Actual decision $\Rightarrow$ allow NIST to compute $C_{\text{DET}}$
  - Confidence scores $\Rightarrow$ allow NIST to compute $\min C_{\text{DET}}$
    - $\min C_{\text{DET}}$ provides best possible cost for the given system discrimination with optimal threshold

- Calibration loss: $C_{\text{DET}} - \min C_{\text{DET}}$
  - Measure the goodness of threshold selection
  - Better discrimination does not mean lower cost
NIST’04 1s1s Calibration Results

Common Evaluation Condition
Primary Systems
1side training -1side test
ATVS
Short-term parameterization (MFCC)
### Gaussian Mixture Models (GMM)

**$x_t$: LPCC/MFCC coefficients vector**

**Mean vector:**

$$\mu_p = \{\mu_{ip}\}$$

**Covariance matrix:**

$$\Sigma_p = \{\Sigma_{ip}\}$$

**Weights vector:**

$$\omega_p = \{\omega_{ip}\}, \sum_i \omega_{ip} = 1$$

**Speaker model $p$:**

$$\lambda_p = \{\mu_{ip}, \Sigma_{ip}, \omega_{ip}\}$$

\[
p(x | \lambda_p) = \sum_{i=1}^{M} \omega_{ip} g_{ip} (x)
\]

\[
g_{ip} (x) = N\left(\mu_{ip}, \Sigma_{ip}\right)
\]

![GMM Diagram](image)
MAP adaptation from UBM

- Only means are usually adapted with target data
  - Weights and covariances shared across speakers
The GMM-UBM Framework: Score Normalization

- Score misalignment
The GMM-UBM Framework: Score Normalization

- **T-Norm**: input speech versus a cohort of impostor models

- **Z-Norm**: target speaker model tested with a cohort of impostor trials

- **ZT-Norm**: T-Norm + Z-Norm

- **ZT-norm has been the bottleneck in NIST evals**
Going Higher (2002-2005): High-level Supra-segmental & High-dimensionality Systems
Higher Level Speaker Identification


Exploiting High-Level Information for High-Performance Speaker Recognition

Douglas Reynolds¹, Walter Andrews², Joseph Campbell¹, Jiří Navrátil³, Barbara Peskin⁴, Andre Adami⁵, Qin Jin⁶, David Klusáček⁷, Joy Abramson⁸, Radu Mihaescu⁹, John Godfrey¹, Douglas Jones¹, Bing Xiang¹⁰

¹ MIT Lincoln Laboratory, ² US Department of Defense, ³ IBM, ⁴ International Computer Science Institute, ⁵ Oregon Graduate Institute, ⁶ Carnegie Mellon University, ⁷ Charles University, ⁸ York University, ⁹ Princeton University, and ¹⁰ Cornell University
SuperSID Results

- With Switchboard II data and common protocols:

<table>
<thead>
<tr>
<th>System</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Acoustic baseline (GMM-UBM cepstral features)</td>
<td>0.7</td>
</tr>
<tr>
<td>2. Pitch and energy distributions</td>
<td>16.3</td>
</tr>
<tr>
<td>3. Pitch and energy slopes + durations + phoneme context</td>
<td>5.2</td>
</tr>
<tr>
<td>4. Prosodic statistics</td>
<td>8.1</td>
</tr>
<tr>
<td>5. Phone n-grams (5 PPRLM phone sets)</td>
<td>4.8</td>
</tr>
<tr>
<td>6. Phone binary trees (5 PPRLM phone sets)</td>
<td>3.3</td>
</tr>
<tr>
<td>7. Phone cross-stream + temporal (5 PPRLM phone sets)</td>
<td>3.6</td>
</tr>
<tr>
<td>8. Pronunciation modeling (SRI prons + 5 PPRLM phone sets)</td>
<td>2.3</td>
</tr>
<tr>
<td>9. Word n-grams/Idiolect (Dragon transcripts)</td>
<td>11.0</td>
</tr>
</tbody>
</table>
SuperSID Fusion

Evaluation conditions (e.g., 2004)

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Speakers</th>
<th>Models</th>
<th>Target Trials</th>
<th>Impostor Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 sides</td>
<td>121</td>
<td>123</td>
<td>470</td>
<td>4594</td>
</tr>
<tr>
<td>8 sides</td>
<td>307</td>
<td>398</td>
<td>1498</td>
<td>15482</td>
</tr>
<tr>
<td>3 sides</td>
<td>310</td>
<td>458</td>
<td>1778</td>
<td>17703</td>
</tr>
<tr>
<td>3 convs</td>
<td>309</td>
<td>538</td>
<td>2068</td>
<td>20880</td>
</tr>
<tr>
<td>1 side</td>
<td>310</td>
<td>417</td>
<td>2392</td>
<td>23832</td>
</tr>
</tbody>
</table>

Table 8: For each model type, numbers of target speakers, individual models, target trials, and non-target (impostor) trials in each test. The figures for the 10 and 30 second model types are identical to those for the 1 side type. (The figures on trials given apply for the three single channel test segment types, and are slightly different for the summed channel single conversation test segments.)

<table>
<thead>
<tr>
<th>Type of Transmission</th>
<th>Training Sides</th>
<th>Test Sides</th>
</tr>
</thead>
<tbody>
<tr>
<td>Landline</td>
<td>257</td>
<td>580</td>
</tr>
<tr>
<td>Cellular</td>
<td>178</td>
<td>361</td>
</tr>
<tr>
<td>Cordless</td>
<td>176</td>
<td>219</td>
</tr>
<tr>
<td>Other/unknown</td>
<td>5</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 9: Phone transmission types of the training and test conversation sides for the core test condition included in the NIST 2004 evaluation data.

<table>
<thead>
<tr>
<th>Type of Handset</th>
<th>Training Sides</th>
<th>Test Sides</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speakerphone</td>
<td>37</td>
<td>67</td>
</tr>
<tr>
<td>Headset</td>
<td>107</td>
<td>116</td>
</tr>
<tr>
<td>Ear-bud</td>
<td>42</td>
<td>63</td>
</tr>
<tr>
<td>Regular (hand-held)</td>
<td>452</td>
<td>914</td>
</tr>
<tr>
<td>Other/unknown</td>
<td>5</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 10: Phone handset types of the training and test conversation sides for the core test condition included in the NIST 2004 evaluation data.
SVMs and SuperVectors

- SVMs: efficient training and scoring in high-dimensional space
- Targets & non-targets cepstral coefficients are not separable by an hyperplane
- Generalized Linear Discriminant Sequence Kernel (GLDS):
  - e.g. orig dim ~ 30 $\Rightarrow$ expanded dim ~ 3300
- GMM-Supervectors (GSV):
  - Dimension: feature dimension (e.g. 39 MFCC) x number of
NAP (Nuissance Attribute Projection)

**Hypothesis**
Channel distortion is focused in a reduced set of dimensions (eigenvalues)

**Objective**
Find and cancel such directions (through eigenvectors)

*Eigenvectors: most affected distortions*

**TRAIN:** finding the eigenvectors.

*Channel compensation:* cancelling the projection of input vectors ($x$) over distorted directions (eigenv.)

\[ \tilde{x}' = \tilde{x} - \langle \tilde{x} | \text{eigenv} \rangle \]
The inference of identity in forensic speaker recognition

Christophe Champod *, Didier Meuwly

Institut de Police Scientifique et de Criminologie, University of Lausanne, CH-1015 Lausanne, Switzerland
Received 28 October 1998; received in revised form 13 September 1999

Abstract

The aim of this paper is to investigate the ways of interpreting evidence within the field of speaker recognition. Several methods – speaker verification, speaker identification and type I and type II errors statement – will be presented and evaluated in the light of judicial needs. It will be shown that these methods for interpreting evidence unfortunately force the scientist to adopt a role and to formulate answers that are outside his scientific province. A Bayesian interpretation framework (based on the likelihood ratio) will be proposed. It represents an adequate solution for the interpretation of the aforementioned evidence in the judicial process. It fills in the majority of the gaps of the other inference frameworks and allows likening the speaker recognition to the same logic than the other forensic identification evidences. © 2000 Published by Elsevier Science B.V. All rights reserved.
Testability

- How have we assessed LR performance?
  - Tippett plots, which are case-independent

![Graphic Representation](image)

- $H_p$ true $\sim$ targets
- $H_d$ true $\sim$ impostors
ATVS Generative LRs @ NIST’04

Five submitted systems

NIST 04 Submitted LR system

---

Tarea 1side-1side (All Trials)

- MFCC-1 + MFCC-2 EER: 13.64
- MFCC-1 EER: 15.94
- MFCC-2 EER: 14.39
- LR MFCC-1 EER: 16.53
- LR TDLRA MFCC-1 EER: 17.20

Proportion of cases (%)

LR greater than

False Alarm probability (in %)
Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition

Joaquin Gonzalez-Rodriguez a,*, Andrzej Drygajlo b, Daniel Ramos-Castro a, Marta Garcia-Gomar c, Javier Ortega-Garcia a

a ATVS (Speech and Signal Processing Group), Computer Science, Escuela Politecnica Superior, Universidad Autonoma de Madrid, Ctra. Colmenar km. 15 Campus de C., E-28049 Madrid, Spain
b Signal Processing Institute, Swiss Federal Institute of Technology Lausanne (EPFL), CH-1015, Lausanne, Switzerland
c Agfis, Centro de Empresas La Arboleda, Ctra. A-3 Km. 7, E-28031 Madrid, Spain

Received 4 November 2004; received in revised form 26 July 2005; accepted 20 August 2005
Available online 26 September 2005

Abstract

In this contribution, the Bayesian framework for interpretation of evidence when applied to forensic speaker recognition is introduced. Different aspects of the use of voice as evidence in the court are addressed, as well as the use by the forensic expert of the likelihood ratio as the right way to express the strength of the evidence. Details on computation procedures of likelihood ratios (LR) are given, along with the assessment tools and methods to validate the performance of these Bayesian forensic systems. However, due to the practical scarcity of suspect data and the mismatched conditions between traces and reference populations common in daily casework, significant errors appear in LR estimation if specific robust techniques are not applied. Original contributions for the robust estimation of likelihood ratios are fully described, including TDLRA (target dependent likelihood ratio alignment), oriented to guarantee the presumption of innocence of suspected but non-perpetrators speakers. These algorithms are assessed with extensive Switchboard experiments but moreover through blind LR-based submissions to both NFI-TNO 2003 Forensic SRE and NIST 2004 SRE, where the strength of the evidence was successfully provided for every questioned speech-suspect recording pair in the respective evaluations.

© 2005 Elsevier Ltd. All rights reserved.
Application-independent evaluation of speaker detection

Niko Brümmer a,b,*, Johan du Preez b

a Spescom DataVoice, Stellenbosch 7600, South Africa
b University of Stellenbosch DSP Group, South Africa

Received 1 November 2004; received in revised form 1 August 2005; accepted 3 August 2005
Available online 6 September 2005

Abstract

We propose and motivate an alternative to the traditional error-based or cost-based evaluation metrics for the goodness of speaker detection performance. The metric that we propose is an information-theoretic one, which measures the effective amount of information that the speaker detector delivers to the user. We show that this metric is appropriate for the evaluation of what we call application-independent detectors, which output soft decisions in the form of log-likelihood-ratios, rather than hard decisions. The proposed metric is constructed via analysis and generalization of cost-based evaluation metrics. This construction forms an interpretation of this metric as an expected cost, or as a total error-rate, over a range of different application-types. We further show how the metric can be decomposed into a discrimination and a calibration component. We conclude with an experimental demonstration of the proposed technique to evaluate three speaker detection systems submitted to the NIST 2004 Speaker Recognition Evaluation.

© 2005 Elsevier Ltd. All rights reserved.
Any speaker recognition approach can be decomposed in two sequential stages:

- Questioned speech
- Control speech

**Distance / Similarity / Likelihood Measures**

**Presentation Stage**

- Identification/exclusion
- Verbal scale of probabilities

**score**

- Likelihood Ratio

**C_{DET}** is an *application-dependent* simultaneous measure of discrimination and calibration

**C_{llr} =** discrimination loss + calibration loss
Effects of miscalibration: an example
ATVS Calibration in SRE 2005

Excellent calibration performance for all conditions evaluated
Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition

Joaquin Gonzalez-Rodriguez, Member, IEEE, Phil Rose, Daniel Ramos, Student Member, IEEE, Doroteo T. Toledano, Member, IEEE, and Javier Ortega-Garcia, Member, IEEE

Abstract—Forensic DNA profiling is acknowledged as the model for a scientifically defensible approach in forensic identification science, as it meets the most stringent court admissibility requirements demanding transparency in scientific evaluation of evidence and testability of systems and protocols. In this paper, we propose a unified approach to forensic speaker recognition (FSR) oriented to fulfill these admissibility requirements within a framework which is transparent, testable, and understandable, both for scientists and fact-finders. We show how the evaluation of DNA evidence, which is based on a probabilistic similarity-typicality metric in the form of likelihood ratios (LR), can also be generalized to continuous LR estimation, thus providing a common framework for phonetic-linguistic methods and automatic systems. We highlight the importance of calibration, and we exemplify with LRs from diphthongal F-pattern, and LRs in NIST-SRE06 tasks. The application of the proposed approach in daily casework remains a sensitive issue, and special caution is enjoined. Our objective is to show how traditional and automatic FSR methodologies can be transparent and testable, but simultaneously remain conscious of the present limitations. We deal with issues in the inclusion of probability, expressed in the form of individualisation (hard match), as categorical opinion of identity of sources, exclusion (nonmatch) statements, or making use of verbal scales of probability of hypothesis, given evidence. The process leading from evidence to conclusion is often opaque, either because it lacks scientific rigor and is inherently unfalsifiable, or because the approach is inadequately tested, and thus cannot quote random match probabilities or estimate the chance of error. Not surprisingly, this has often resulted in legal discussion about the acceptance of expert testimony. Contrasting with this, DNA profiling [1], [5], [65] has solid and well-known scientific foundations, and is probabilistic [18], [60]. Avoiding individualization or exclusion statements for the determination of the source of the evidence, DNA evidence is often presented using frequencies, match probabilities, and inclusion or exclusion probabilities [24], but many influential forensic scientists [1], [16], [27], [26] advocate assessing the weight of the evidence with likelihood ratios.
Alternate microphones in SRE06

- Ear-bud/lapel mike
- Mini-boom mike
- Courtroom mike
- Conference room mike
- Distant mike
- Near-field mike
- PC stand mike
- Micro-cassette mike

Information on the microphone type used in each non-telephone test segment data will not be available to recognition systems.
Factor Analysis

**GMM-UBM (MAP):**

\[ \mu_{sh} = \mu + Dz_{sh} \]

**D:** Full-rank diagonal (scaling factor)

**z:** speaker component

**Eigenvoices:**

\[ \mu_s = \mu + V y_s \]

**V:** speaker variability subspace (low-rank)

**y:** corresponding weights for a given speaker, speaker factors

**Eigenchannels**

\[ \mu_{sh} = \mu_s + U x_h \]

**U:** session variability subspace (low-rank)

**x:** corresponding weights for a given utterance/speaker, channel factors
Learning Speaker & Channel Variability

\[ A = \begin{pmatrix} \mu_{11} & \mu_{12} \\ \mu_{s1} & \mu_{s2} \\ \mu_{n1} & \mu_{n2} \\ \mu_{sn} \end{pmatrix} \]

\[ \Theta_b = \begin{pmatrix} \mu_{s1} - \mu_0 \\ \mu_{s2} - \mu_0 \\ \mu_{sn} - \mu_0 \end{pmatrix} \]

\[ \Theta_w = \begin{pmatrix} \mu_{11} - \mu_{s1} \\ \mu_{21} - \mu_{s2} \\ \mu_{n1} - \mu_{sn} \end{pmatrix} \]

Between scatter matrix \( S_b = \Theta_b \Theta_b^T \)

Within scatter matrix \( S_w = \Theta_w \Theta_w^T \)

\( \sim 2000 \text{ speakers} \)
\( \sim 5 \text{ utts/speaker} \)
\( \sim 10k \text{ columns} \)
Factor Analysis flavours: ATVS with SRE06 data
The core (required) condition includes:

- Conversational telephone speech
  - Telephone channel
  - Microphone channel
- Conversational speech data in an interview scenario
  - Microphone channel

Systems know:

- Telephone / microphone channel
- Interview scenario / telephone conversation
6.2 Numbers of Test Segments

Table 4 provides estimated upper bounds on the numbers of segment to be included in the evaluation for each test condition.

Table 4: Upper bounds on numbers of segments by test condition

<table>
<thead>
<tr>
<th>Test Conditions</th>
<th>Max Segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>10sec</td>
<td>5000</td>
</tr>
<tr>
<td>short3</td>
<td>8000</td>
</tr>
<tr>
<td>long</td>
<td>2000</td>
</tr>
<tr>
<td>summed</td>
<td>5000</td>
</tr>
</tbody>
</table>

6.3 Numbers of Trials

The trials for each of the speaker detection tests offered will be specified in separate index files. These will be text files in which each record specifies the model and a test segment for a particular trial. The number of trials for each test condition is expected not to exceed 100,000.
NIST SRE 08: new test data

- Two types of test speech:
  - Phonecall conversational speech (Mixer 3)
    - Phonecall-phn: telephone recording
    - Phonecall-mic: multiple simultaneous microphone recording
  - Interview speech (Mixer 5)
    - Interview-mic: multiple simultaneous microphone recording

<table>
<thead>
<tr>
<th>Training channel</th>
<th>Test channel</th>
<th># trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telephone</td>
<td>Telephone</td>
<td>24128</td>
</tr>
<tr>
<td></td>
<td>Microphone</td>
<td>8746</td>
</tr>
<tr>
<td>Microphone</td>
<td>Telephone</td>
<td>6693</td>
</tr>
<tr>
<td></td>
<td>Microphone</td>
<td>19776</td>
</tr>
<tr>
<td><strong>Total male</strong></td>
<td><strong>59343</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Training channel</th>
<th>Test channel</th>
<th># trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telephone</td>
<td>Telephone</td>
<td>12922</td>
</tr>
<tr>
<td></td>
<td>Microphone</td>
<td>7025</td>
</tr>
<tr>
<td>Microphone</td>
<td>Telephone</td>
<td>5048</td>
</tr>
<tr>
<td></td>
<td>Microphone</td>
<td>14270</td>
</tr>
<tr>
<td><strong>Total female</strong></td>
<td><strong>39265</strong></td>
<td></td>
</tr>
</tbody>
</table>

1788 Mixer 3 (conversational) spk models
1475 Mixer 5 (interview) spk models
devATVS08: simulating SRE08 x-chan

Evolution of ATVS GMM system in SRE08 development (male subset)

- Telephone dev., raw scores: EER-DET = 19.72; DCF-opt = 0.0776
- + Microphone dev., EER-DET = 15.01; DCF-opt = 0.0572
- + Channel compensation in enrolled models: EER-DET = 11.32; DCF-opt = 0.0432
- + Wiener filtering: EER-DET = 6.28; DCF-opt = 0.0309
- + Channel compensation in test segments: EER-DET = 5.85; DCF-opt = 0.0286
- + condition-independent TNorm: EER-DET = 5.40; DCF-opt = 0.0250
- + condition-dependent TNorm: EER-DET = 4.80; DCF-opt = 0.0246

¡GMM SRE06! (no Tnorm)
SRE 2010

- New operating point at very low FA rates

Table 2: Speaker Detection Cost Model Parameters for the core and 8conv/core test conditions

<table>
<thead>
<tr>
<th>$C_{\text{Miss}}$</th>
<th>$C_{\text{FalseAlarm}}$</th>
<th>$P_{\text{Target}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.001$^4$</td>
</tr>
</tbody>
</table>

- Relative Cost of FA to Miss errors:
  - From 10 (old point) to 1000 (new point)
- Huge amounts of trials for reliable error estimation

- High and low vocal effort

- Variable length segments, english only
6.1 Numbers of Models

Table 4 provides estimated upper bounds on the numbers of models (target speakers) to be included in the evaluation for each training condition.

Table 4: Upper bounds on numbers of models by training condition

<table>
<thead>
<tr>
<th>Training Condition</th>
<th>Max Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>10sec</td>
<td>3,000</td>
</tr>
<tr>
<td>core</td>
<td>6,000</td>
</tr>
<tr>
<td>8conv</td>
<td>1,000</td>
</tr>
<tr>
<td>8summed</td>
<td>1,000</td>
</tr>
</tbody>
</table>

6.2 Numbers of Test Segments

Table 5 provides estimated upper bounds on the numbers of segments to be included in the evaluation for each test condition.

Table 5: Upper bounds on numbers of segments by test condition

<table>
<thead>
<tr>
<th>Test Conditions</th>
<th>Max Segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>10sec</td>
<td>6,000</td>
</tr>
<tr>
<td>core</td>
<td>25,000</td>
</tr>
<tr>
<td>summed</td>
<td>6,000</td>
</tr>
</tbody>
</table>

6.3 Numbers of Trials

The trials for each of the speaker detection tests offered will be specified in separate index files. These will be text files in which each record specifies the model and a test segment for a particular trial. The number of trials for each test condition is expected not to exceed 750,000.
Total variability and i-vectors

- Channel factors to be discarded
  - But still contain residual information from the speaker

- Instead of two different subspaces
  - Speaker (V)
  - Channel (U)

- A single subspace T
  - Contains both speaker and channel variability

- i-vector as feature extractor (front end)

- Factor analysis in the T subspace
  - PLDA (Probabilistic LDA)
SRE 2012 Overview

- “a significant departure from previous NIST SREs”
- Most target speakers from previous corpora
  - Train from all speech available
  - Large number of segments from multiple recordings (mic int, mic phn, tel phn)
- Training data provided at registration (not eval)
- Knowledge of all targets IS allowed in computing each trial’s detection score
  - Tests on both known and unknown impostors
SRE 2012: more news

- logLR as system output score REQUIRED !!!
- Additive and environmental noise
  - Improved VAD criticals to performance !!!
- Durations: 300, 100 and 30 sec
- No ASR provided
- New $C_{\text{primary}}$
  - Intended for “wide” (not single point) calibration
  - Averages two operating points
    - $P_{\text{target}_A1}=0.01$ and $P_{\text{target}_A2}=0.001$ ($C_{\text{miss}}=C_{\text{FA}}=1$)
SRE 2012 Common (required) conditions

Five common conditions were specified for SRE12

They were all trials involving multiple segment training and

1) interview speech in test without added noise in test
2) telephone channel speech in test without added noise in test
3) interview speech in test with added noise in test
4) telephone channel speech in test with added noise in test
5) telephone channel speech intentionally collected in a noisy environment* in test and without added noise

*this was self-reported by the speaker
### SRE 2012: #trials

<table>
<thead>
<tr>
<th>Common Condition</th>
<th>Core (target / known non-target / unknown non-target)</th>
<th>Extended Trials (target / known non-target / unknown non-target)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2,897 / 46,601 / 61,871</td>
<td>3,860 / 10,985,377 / 11,349,426</td>
</tr>
<tr>
<td>2</td>
<td>7,354 / 445,041 / 105,196</td>
<td>7,354 / 10,312,118 / 2,088,834</td>
</tr>
<tr>
<td>3</td>
<td>3,851 / 49,032 / 20,048</td>
<td>5,127 / 12,444,672 / 4,804,500</td>
</tr>
<tr>
<td>4</td>
<td>7,176 / 411,843 / 4,872</td>
<td>7,176 / 9,471,219 / 124,830</td>
</tr>
<tr>
<td>5</td>
<td>3,883 / 209,532 / 2,406</td>
<td>3,883 / 5,119,130 / 77,745</td>
</tr>
</tbody>
</table>

**Total (core) = 1,381,603 trials !!!**
**Total (extended) ~67,000,000 trials !!!**
Why “Big Data” Evals?

- Transformation of the “simple” speaker detection task
  - Comparison of two utterances

- Systems trained on hundreds of hours of speech
  - Learning proper subspaces & fusion transformations
  - From thousands of speakers
  - Tens of thousands of utterances
  - Varied conditions of channel, speaking style, duration, noise …
Demystifying SREs (2014): NIST i-vector Challenge
NIST i-vector Challenge 2014

- Foster participation from out-of-NIST_SRE
- Instead of speech, i-vectors are provided
- Unknown development and evaluation i-vectors (Spk Ids unknown)
  - Unknown number of i-vectors per speaker
  - Duration provided (side-info)
  - Reference system (cosine-scoring) provided
- Challenge: improve the reference system
- On-line submission and real-time scoring
  - Over 40% evaluation data, open for ~6 months
  - Horse-race effect
    - Knowledge of other participants minimum cost
Discussion & Conclusions
Conclusions

- The NIST series of SRE evals have fostered enormous progress in Speaker Recognition
- Participation is highly recommended to be part and keep track of progress
  - Gap from concepts to large scale implementations
- Drawbacks:
  - Standardization
  - Innovation is indirectly penalized
  - “Unrealistic” fusions of large number of subsystems
  - Data engineering consumes 90% effort during eval period
Challenges

- Adaptation to new application domains with little or no development data
  - Clustering of unlabeled data
- Calibration in highly mismatched conditions
- Goodness of individual comparisons
- Integration of knowledge from linguists & phoneticians
  - Human Aided Speaker Recognition
Enjoy COST … and Malta (Marsaxlokk)
An Overview of the NIST Series of Speaker Recognition Evaluations and Technologies (1996-2014)

Joaquin Gonzalez-Rodriguez
ATVS-Universidad Autonoma de Madrid
joaquin.gonzalez@uam.es http://atvs.ii.uam.es
The Framework: Testability

- Useful representation: APE curve and Cllr

- For priors given (case dependent) we can obtain Probabilities of Error associated with the elicited LR