High-Performance Session Variability Compensation in Forensic Automatic Speaker Recognition

Daniel Ramos, Javier Gonzalez-Dominguez,
Eugenio Arevalo and Joaquin Gonzalez-Rodriguez
ATVS – Biometric Recognition Group
Universidad Autonoma de Madrid
daniel.ramos@uam.es
http://atvs.ii.uam.es

Outline

Forensic Automatic Speaker Recognition:

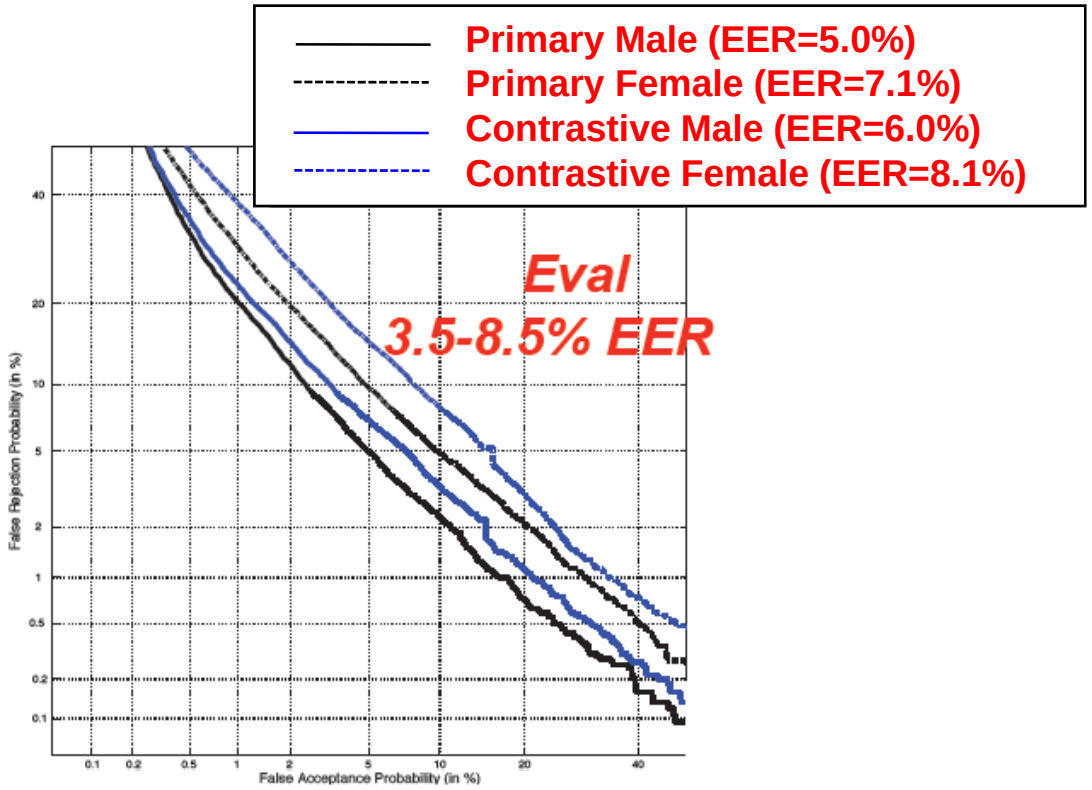
- Where are we?
  - State of the art dominated by high-performance session variability compensation
- Some challenges affecting session var. comp.
  - Database mismatch
  - Sparse background data
  - Duration variability
- Research trends
  - Facing the challenges

Where Are We?

- Automatic Speaker Recognition (ASpkrR) technology
  - Driven by NIST Speaker Recognition Evaluations (SRE)
- State Of The Art dominated by
  - Spectral systems
  - High-performance session variability compensation
    - Factor Analysis, flavors and evolutions
    - Data-driven
- Currently a mature technology
  - Usable in many applications

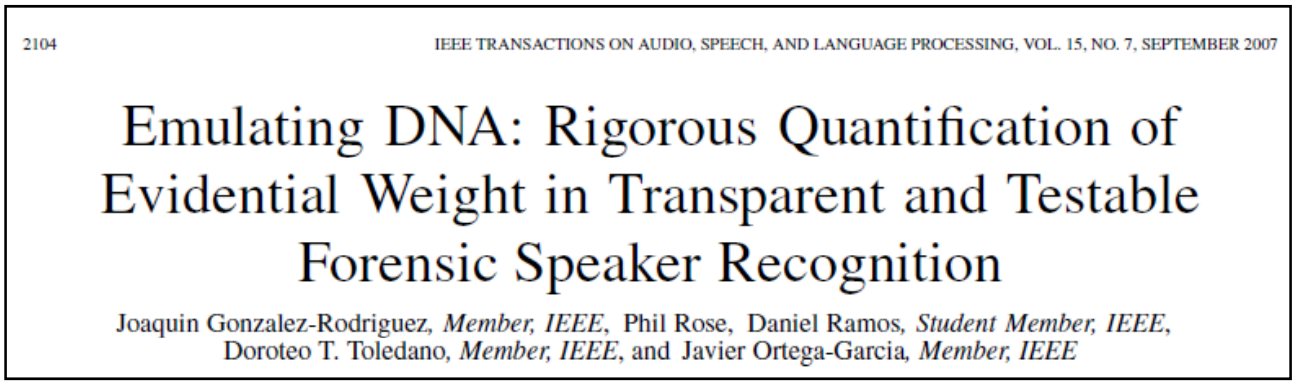
Where Are We?

- Discrimination performance (DET plots)
  - ATVS single spectral system in NIST SRE 2010
  - i-Vectors, session variability compensation



Where Are We?

- To consider in Forensic ASpkrR
  - Convergence to scientific standards
    - "Emulating DNA", Likelihood Ratio (LR) paradigm



2104 IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 15, NO. 7, SEPTEMBER 2007

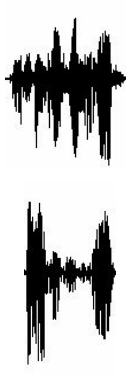
Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition

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

  - Unfavorable environment
    - Mostly uncontrolled conditions
    - Sparse amount of speech (comparison and background)

Where Are We?

- LR paradigm in Forensic ASpkrR



Speaker Recognition System

Score to LR Transformation (calibration)

$$LR$$

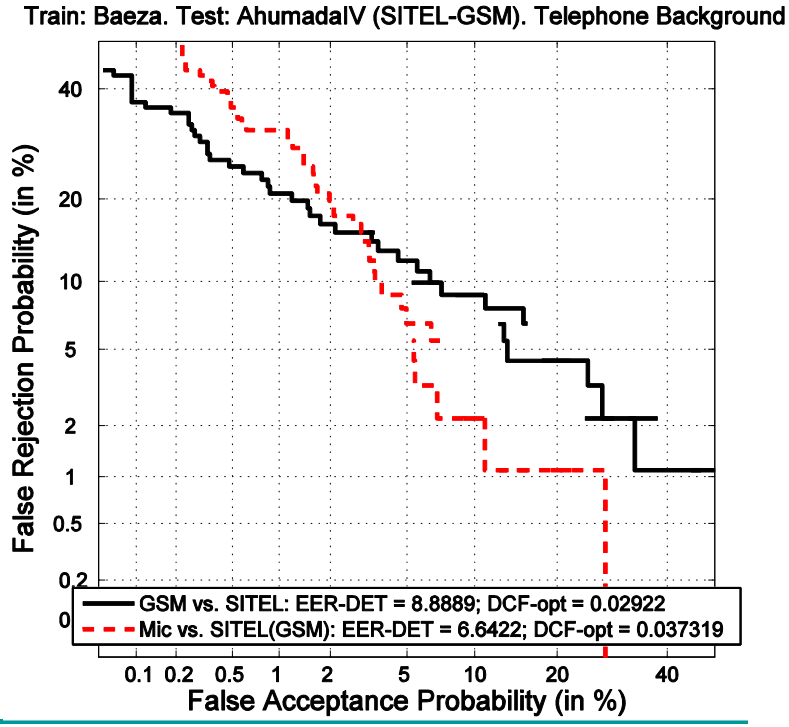
$$LR = \frac{p\left(E \middle| \theta_p, I\right)}{p\left(E \middle| \theta_d, I\right)}$$

Score taken as Evidence (E)

- Two stages
  - Discrimination stage (standard, score-based architecture)
  - Calibration stage (LR computation)

Where Are We?

- Discrimination performance
  - Example with AhumadaIV-Baeza database
    - Thanks to Guardia Civil Española
  - NIST-SRE-like task: comparison between
    - 120s of GSM or microphone (controlled) speech
      - Acquired following Guardia Civil protocols
    - 120s GSM-SITEL speech
      - Acquired using the SITEL Spanish National wire-tapping system



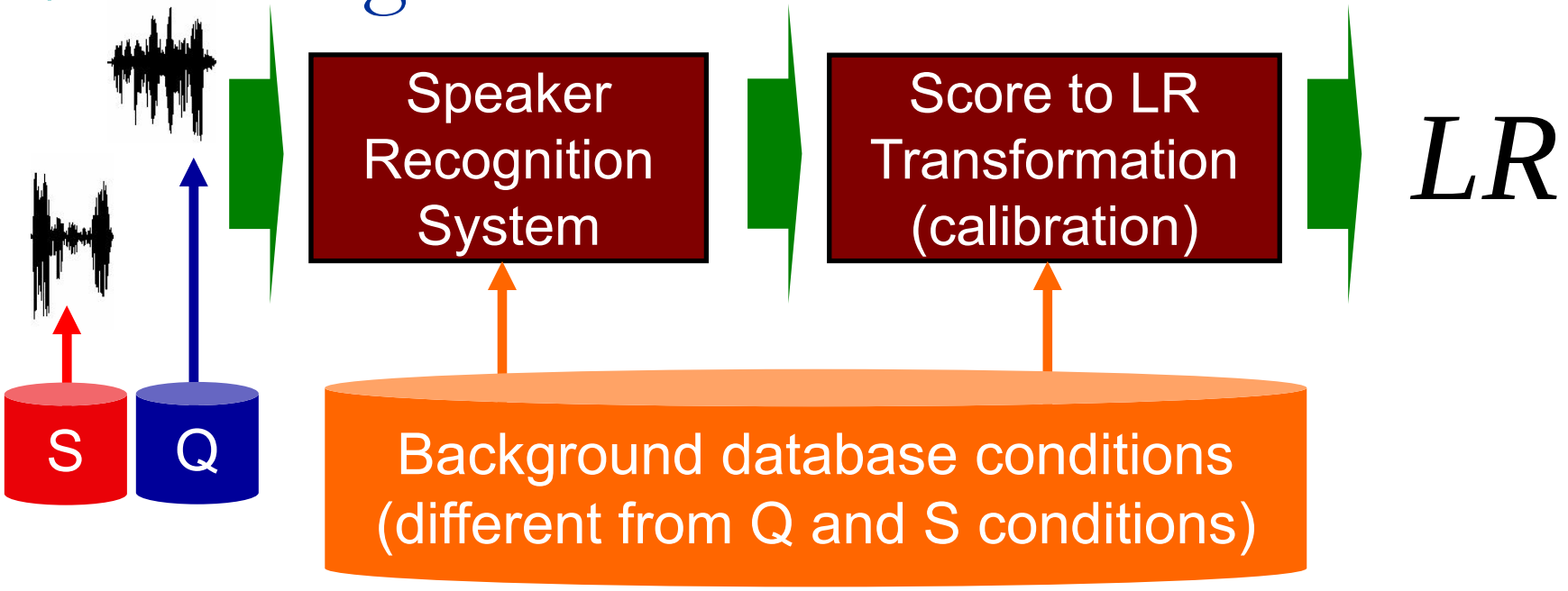
NIST SRE vs. Forensic ASpkrR

- Main commonalities
  - Highly variable environment (telephone, different microphones, interview, etc.)
  - LR paradigm
    - NIST SRE allow LR calibration (assessed by C_{llr})…
    - …although we believe this should be further encouraged
- But in Forensic ASpkrR (and not in NIST SRE)
  - Typical lack of representative background data
    - NIST SRE: lots of speech from past SRE
  - Utterance duration is uncontrolled
    - NIST SRE: conditions of fixed, controlled duration

Challenges of Session Variability Comp.

- Some typical forensic scenarios where session variability compensation degrades
  - Strong database mismatch
  - Sparse background data
  - Extreme duration variability
- Scenarios not present in NIST SRE
  - Minor attention to these problems

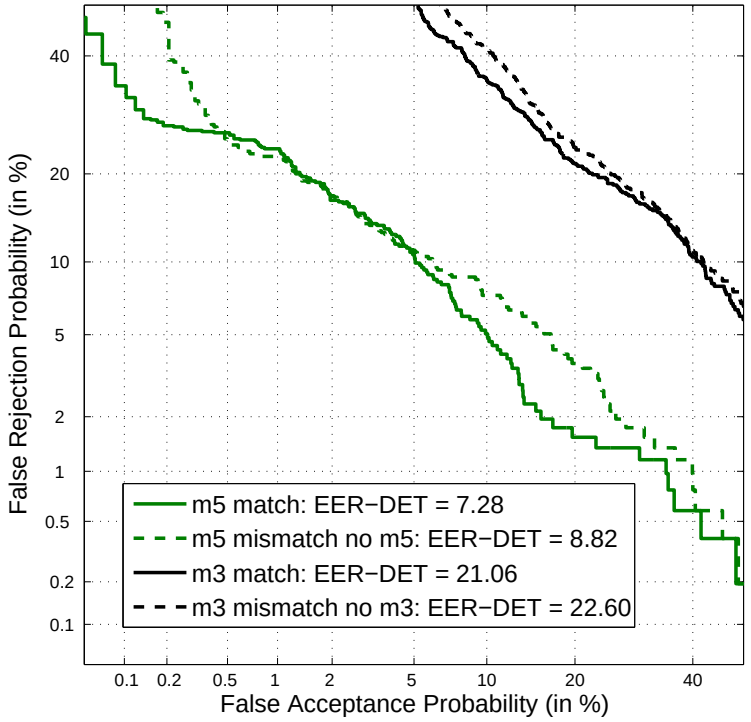
Challenges: Database Mismatch



- Database mismatch: background and comparison (Questioned Q, Suspect S) databases are different
  - Additional problem to mismatch among Q and S
  - Degrades performance of session variability compensation
    - Subspaces are not representative of comparison speech

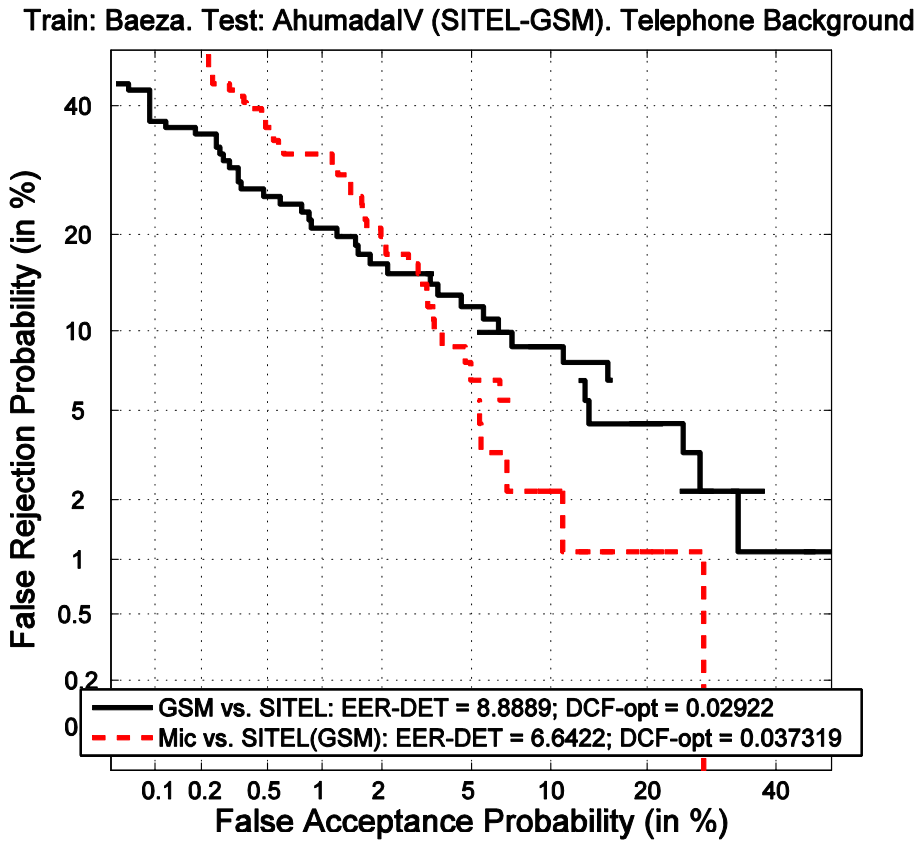
Challenges: Database Mismatch

- Example in NIST SRE 2008
  - Comparison of two speech utterances
    - Speech from a single channel (microphone m3 or m5)
    - Speech from any channel in SRE08
  - Speech from m3/m5 included or not in background
    - UBM, normalization and session variability compensation



Challenges: Database Mismatch

- Example: AhumadaIV-Baeza
  - Background: NIST SRE telephone-only speech
  - Bad performance for low FA rates when microphonic speech is used for training
    - Even when microphone speech is controlled and of higher quality
    - Following the standard acquisition procedures of Guardia Civil Española



Database Mismatch: Research

- Need of collection of more representative databases
- Case study: continuous efforts of Guardia Civil Española
  - Ahumada-Gaudi (2000, spontaneous speech, landline telephone and microphone)
  - AhumadaIII (2008, real forensic cases, multidialect, GSM over magnetic tape)
  - AhumadaIV (2009, speech from SITEL)
  - …


ELSEVIER

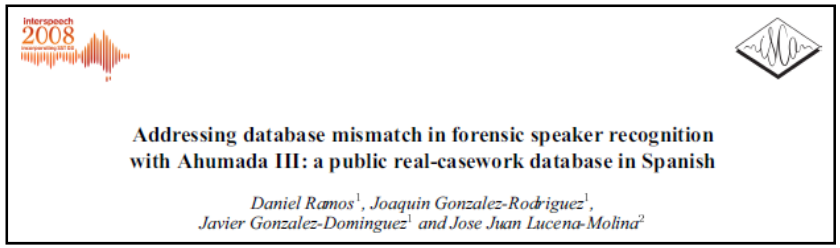
Speech Communication 31 (2000) 255–264

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

AHUMADA: A large speech corpus in Spanish for speaker characterization and identification ☆

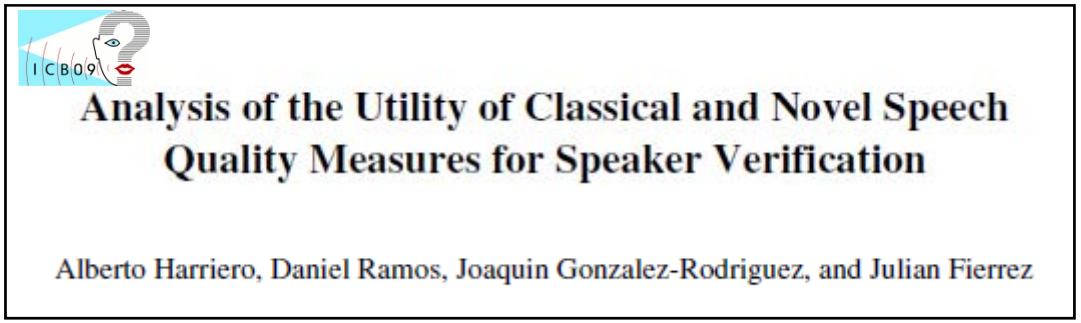
Javier Ortega-Garcia [a,*], Joaquin Gonzalez-Rodriguez [a], Victoria Marrero-Aguiar [b]

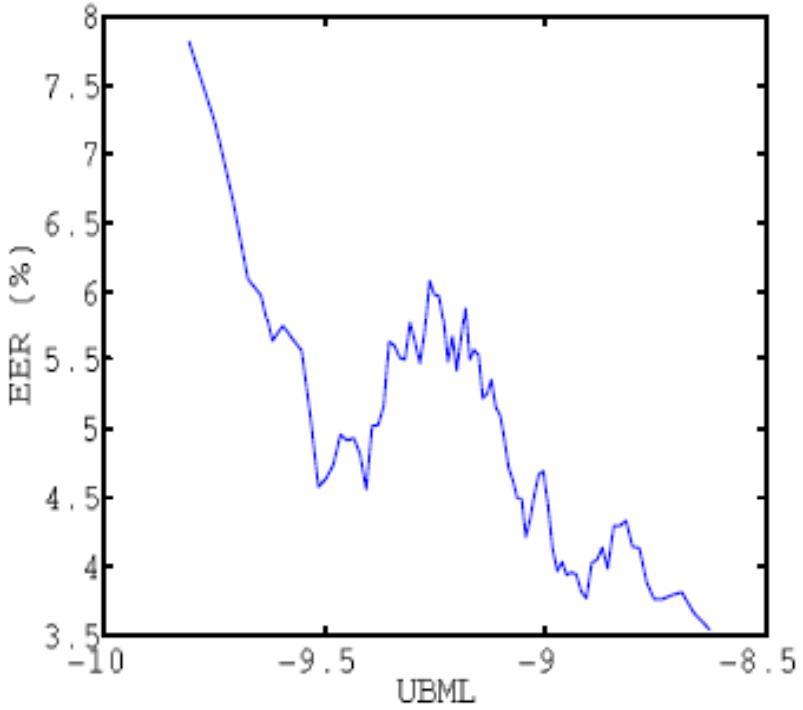

Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-casework database in Spanish

Daniel Ramos[1], Joaquin Gonzalez-Rodriguez[1], Javier Gonzalez-Dominguez[1] and Jose Juan Lucena-Molina[2]

Database Mismatch: Research

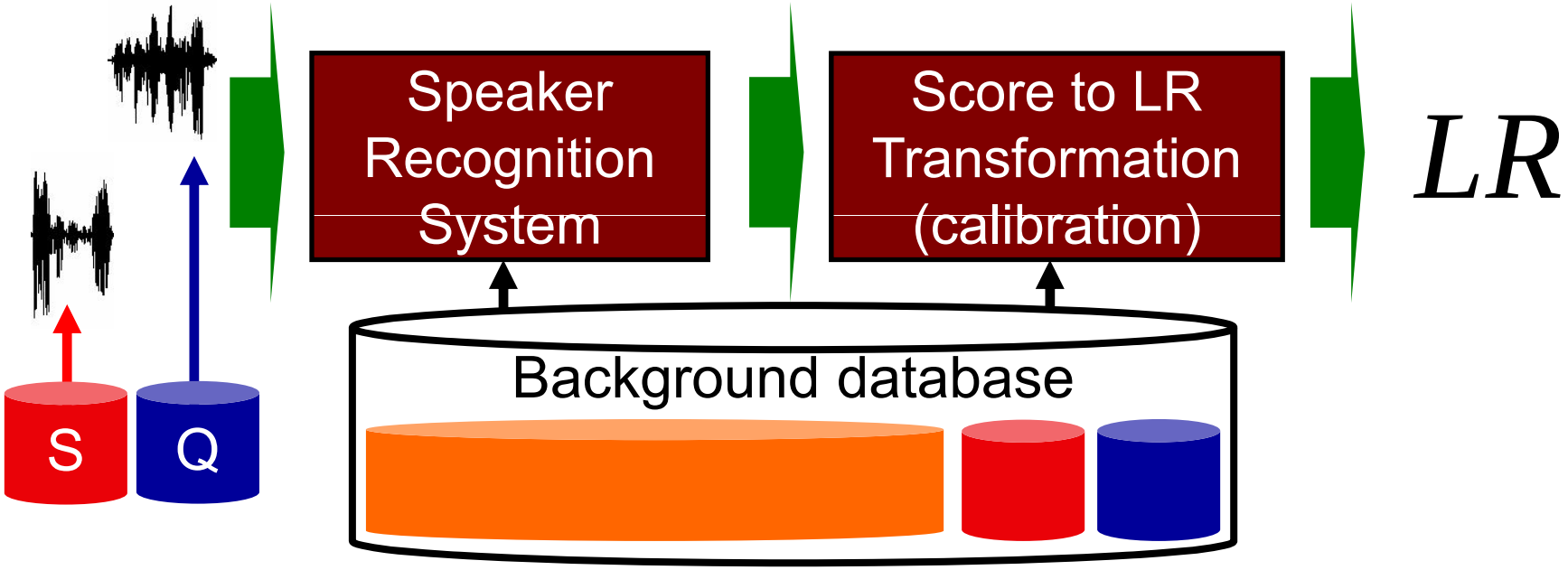
- Predictors of database mismatch
  - *E. g.*: log-likelihood with respect to UBM (UBML)
  - Low UBML indicates database mismatch
    - Performance degrades





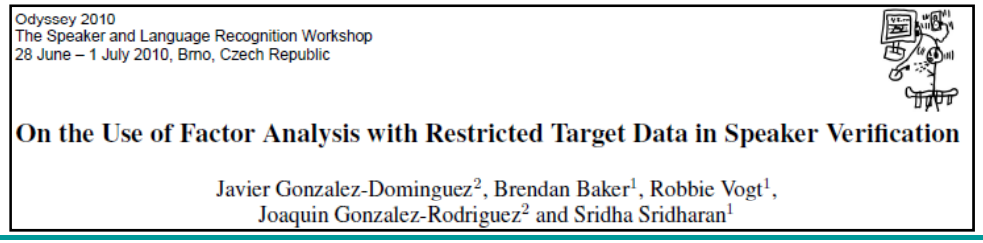
Challenges: Sparse Background Data

- Typical in forensics: some representative background data is available
  - But typically a sparse corpus
- Optimal use of this background data for session variability compensation



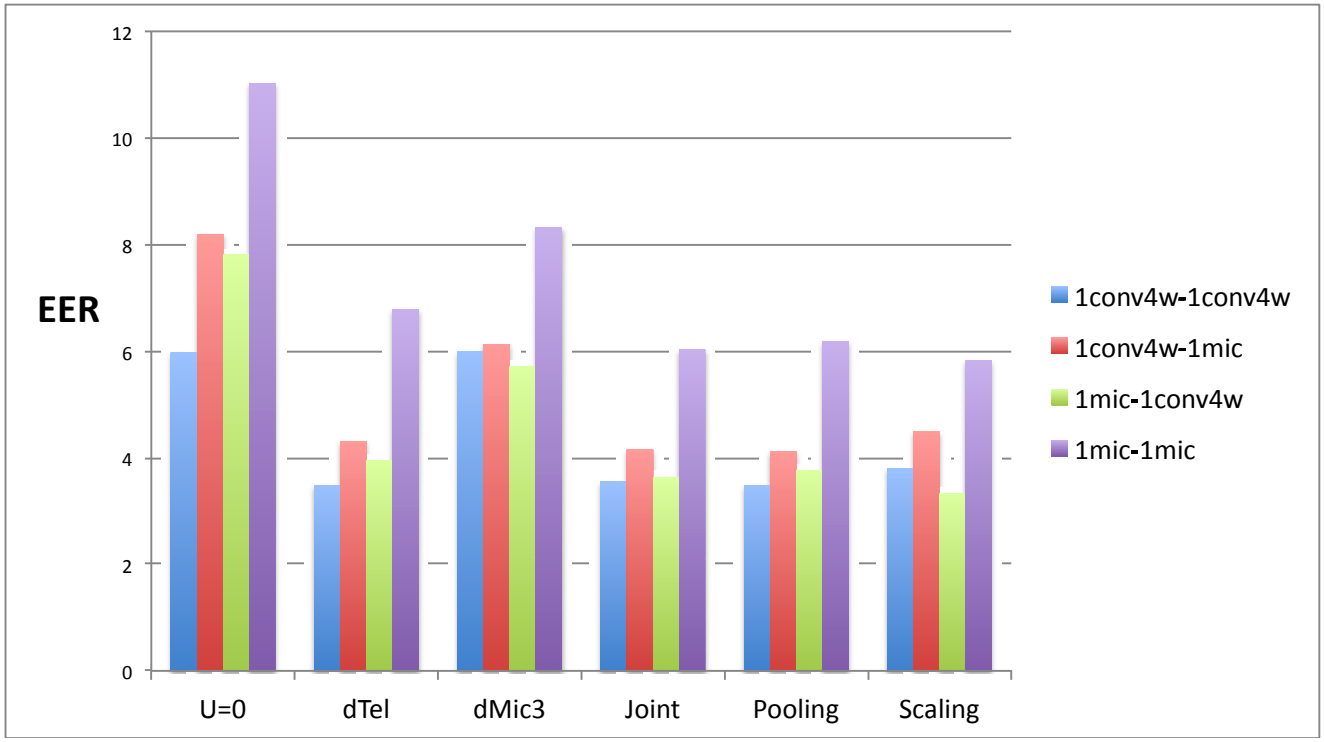
Sparse Background Data: Research

- Example: simulation using NIST SRE 2008
  - Wealth background corpus of telephone data
  - Sparse background corpus of microphone data
  - Microphone and telephone data to be compared
- Session variability compensation strategies
  - Joining compensation matrices
  - Pooling Gaussian statistics
  - Scaling Gaussian statistics



Sparse Background Data: Research

- Combination strategies of available data
  - Wealth corpus, telephone data (dTel)
  - Small corpus, sparse microphone data (dMic3)



Challenges: Duration Variability

- Impact in session variability compensation and score normalization
  - Subspaces/cohorts trained with long utterances
  - Comparison with short utterances
- Other effects
  - Misalignment in the scores due to duration variability
    - Degrades global discrimination performance
    - Seriously affects calibration

Challenges: Duration Variability

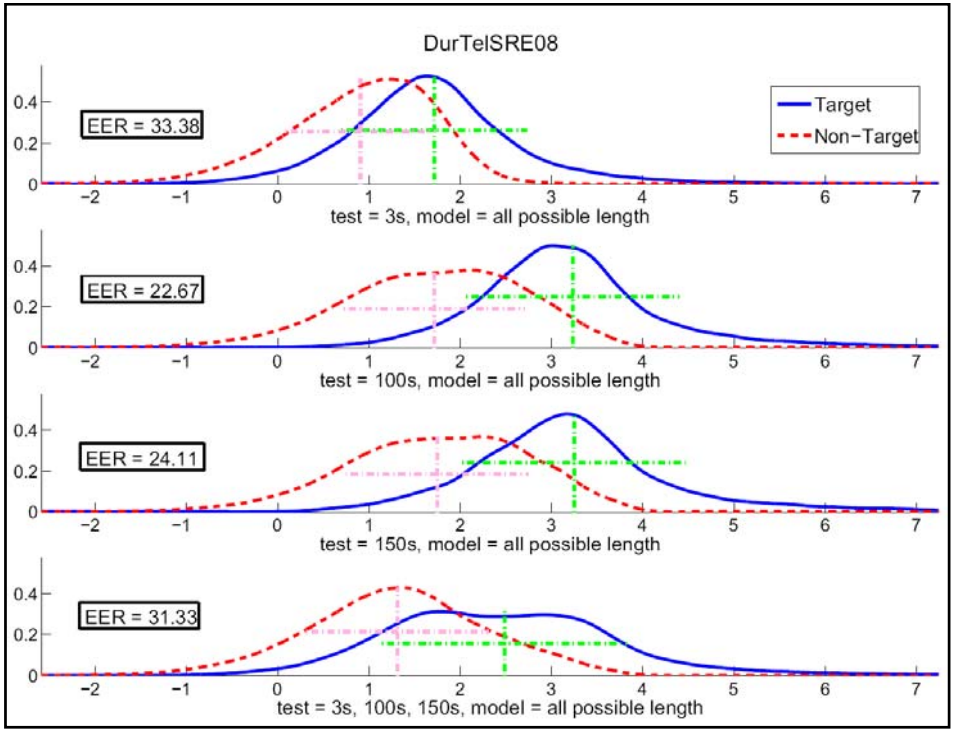
- Impact on score normalization
  - Cohorts trained with fixed-length utterances
- Example in Ahumada III (10s. test utterances)

	NIST SRE cohorts (roughly 150s)	Cohorts adjusted in length
EER with ZT-Norm	12,48%	10,46%

- Impact on session variability compensation
  - More difficult to avoid
    - Supervectors from short utterances are highly variable
  - More research needed

Challenges: Duration Variability

- Duration variability: misalignment effects
  - Different ranges for different test segment durations
  - Even after score normalization (T-Norm)



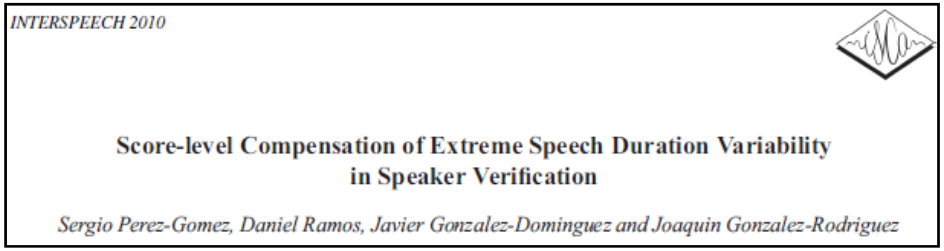
INTERSPEECH 2010

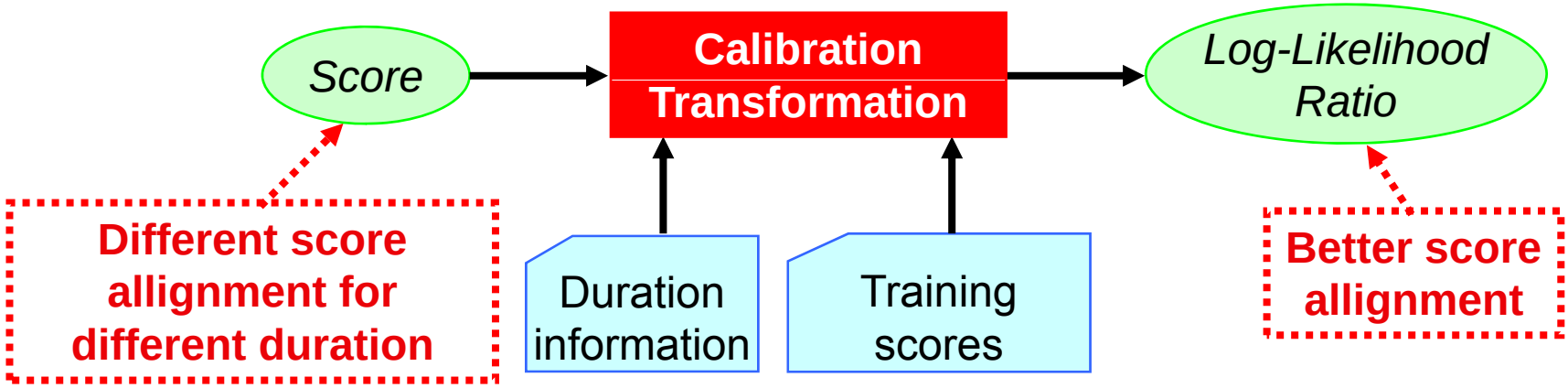
Score-level Compensation of Extreme Speech Duration Variability in Speaker Verification

Sergio Perez-Gomez, Daniel Ramos, Javier Gonzalez-Dominguez and Joaquin Gonzalez-Rodriguez

Duration Variability: Research

- Calibration incorporating duration variability
  - Corrects misalignments due to fixed-cohort normalizations
  - Improves overall discrimination performance





Duration Variability: Research

- Calibration incorporating duration variability
  - Corrects misalignments due to fixed-cohort normalizations
  - Improves overall discrimination performance

	EER (%)	EER improvement (%)						
Test segment length (sec.)		1D-GM	2D-GM	1D-LLR	2D-LLR	BLR 1	BLR 2	BLR 3
3, 10	31.34	2.79	7.17	2.68	7.41	2.63	6.86	4.04
3, 100	31.81	9.61	16.99	9.31	17.23	9.34	15.73	4.59
3, 150	33.08	5.21	10.26	5.15	10.54	4.75	8.73	1.62
10, 100	26.39	3.47	15.67	3.33	15.79	3.57	15.49	2.29
10, 150	27.41	1.97	11.98	1.95	12.28	1.89	11.67	0.22
3, 10, 150	31.11	4.55	10.42	4.43	10.73	2.14	6.32	−0.16
3, 100, 150	31.33	10.28	19.01	10.00	19.29	8.38	13.89	6.37
3, 30, 60, 100, 150	27.96	7.59	19.54	7.44	19.80	3.38	11.67	2.02
All Durations	26.55	4.30	16.37	4.20	16.62	1.84	12.21	0.16

Exception…

Conclusions

- High-performance session variability compensation
  - Works for NIST SRE scenarios
  - Works for forensic scenarios comparable to NIST
- Forensic scenarios where session var. comp. degrades
  - Database mismatch
  - Sparse background data
  - Duration of utterances
- Research directions
  - Predicting and compensating database mismatch
  - Robustness to the lack of backgrund data
  - Robustness to variability in the duration of the utterances

High-Performance Session Variability Compensation in Forensic Automatic Speaker Recognition

Daniel Ramos, Javier Gonzalez-Dominguez,
Eugenio Arevalo and Joaquin Gonzalez-Rodriguez
ATVS – Biometric Recognition Group
Universidad Autonoma de Madrid
daniel.ramos@uam.es
http://atvs.ii.uam.es

ATVS

UAM UNIVERSIDAD AUTONOMA DE MADRID