2001 A Speaker Odyssey:
The Speaker Recognition Workshop

"On the Application of the Bayesian Framework to Real Forensic Conditions with GMM-based Systems"

J. Gonzalez-Rodriguez(1), J. Ortega-Garcia(1), and J.-J. Lucena-Molina(2)

(1) ATVS-Biometric Research Lab., www.atvs.diac.upm.es
Universidad Politécnica de Madrid, SPAIN

(2) Dirección General de la Guardia Civil, Madrid, SPAIN

Outline

- Spk. Recognition in Forensic Cases
- The Likelihood Ratio (LR) Approach
- LR Computation
- LR-based System Description
- Experiments and Results
- Conclusions

The Forensic approach

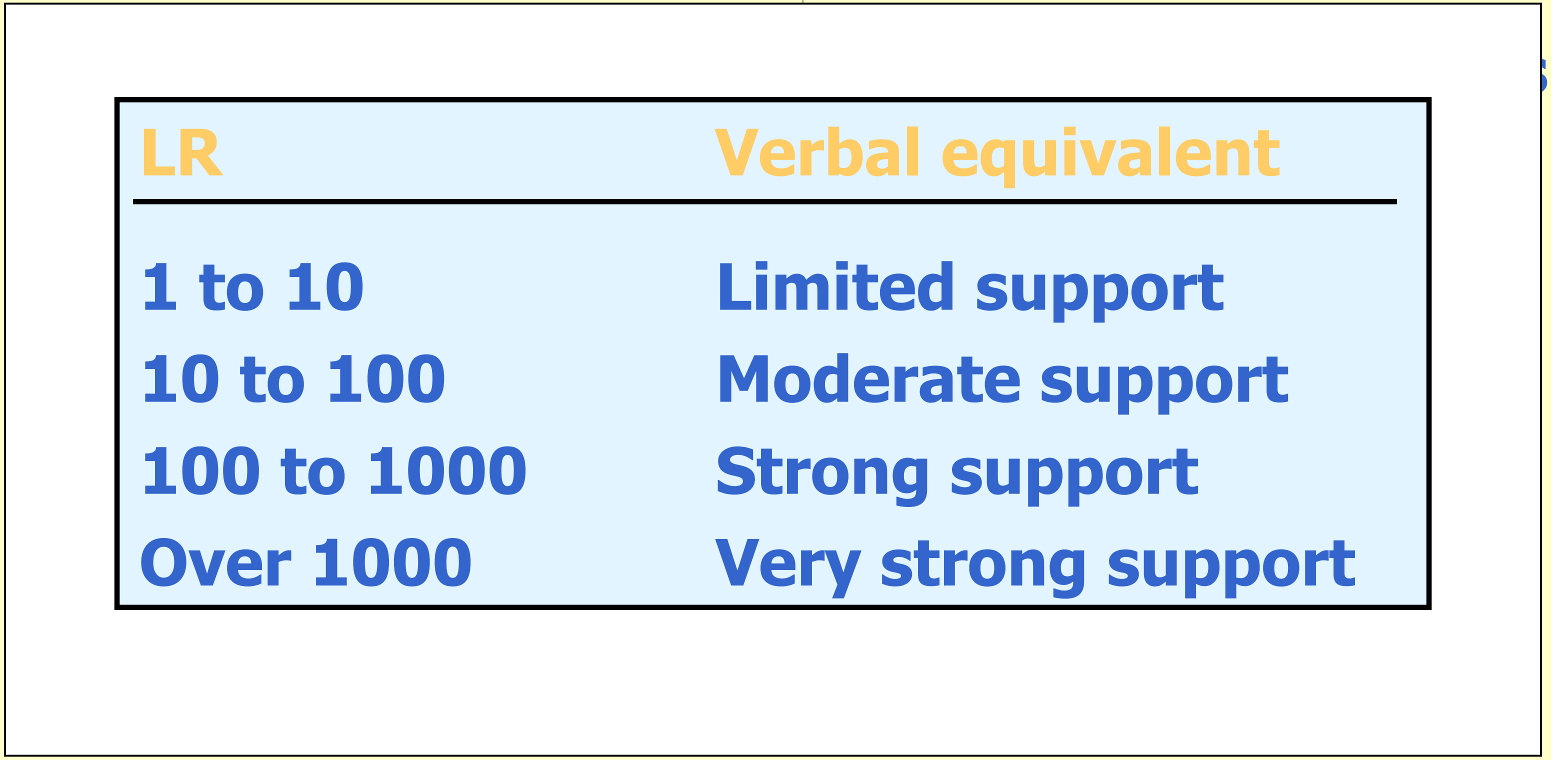
- **Bayesian approach firmly established as theoretical framework in forensic disciplines [Evett, 98].**

- **Roles of judge/jury (judgement/verdict) and scientist (speech processing/interpretation of results) clearly separated.**

- **In court room: odds in favor of prosecution proposition (*"the questioned voice has been uttered by the suspect",* C), given the circumstances of the case (I) and observations made by forensic expert (E).**

- These odds can be expressed as:

$$O(C|E,I) = \frac{\Pr(E|C,I)}{\Pr(E|\bar{C},I)} \cdot O(C|I)$$



LR	Verbal equivalent
1 to 10	Limited support
10 to 100	Moderate support
100 to 1000	Strong support
Over 1000	Very strong support

- Evett suggests (in DNA) a scale of LRs with linguistic qualifier of strenght of verbal support for evidence:

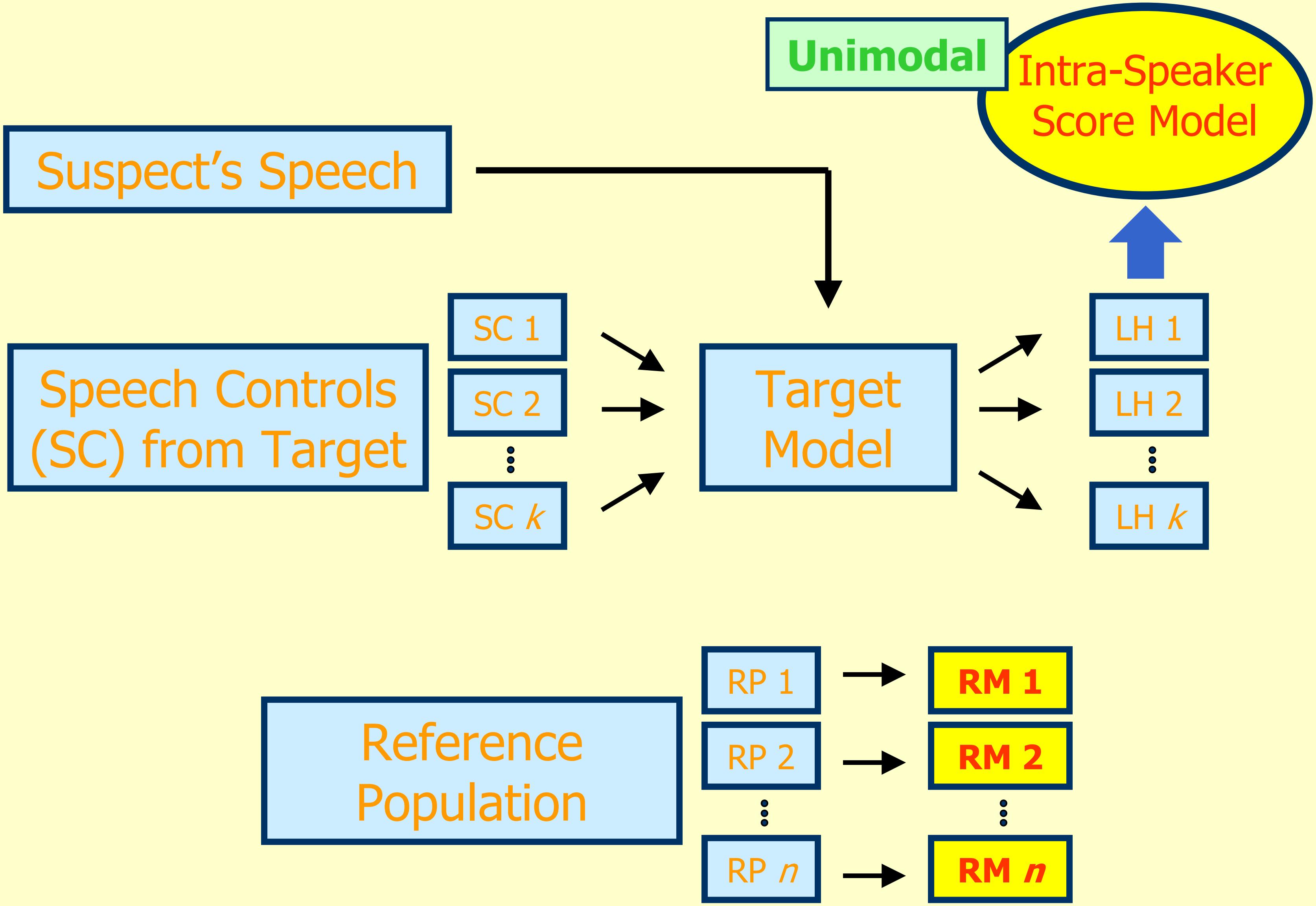
LR approach in Forensic Spk Recognition

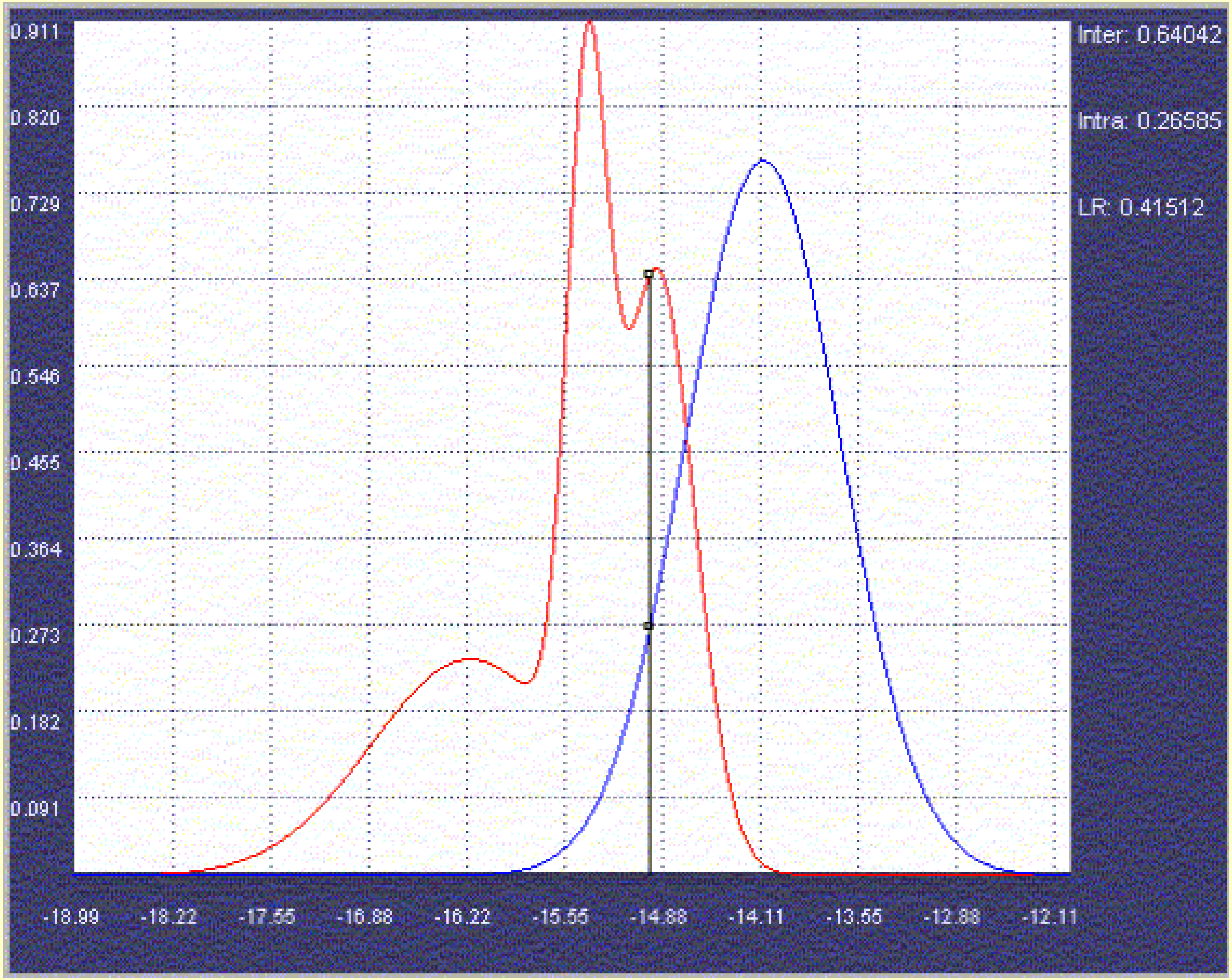
- Output of conventional SR systems (Spk. Verif., Errors type I & II, Spk ID) scarsely provide conclusions to the Court.

- Strong recommendation of using LRs as scientific information in Forensic Speaker Recognition Cases [Champod & Mewly, 98]

- Role of the scientist NOT infering Spk identity BUT showing and interpreting LRs of the opposite hypotheses, C and $\overline{C}$

LR Computation: Training

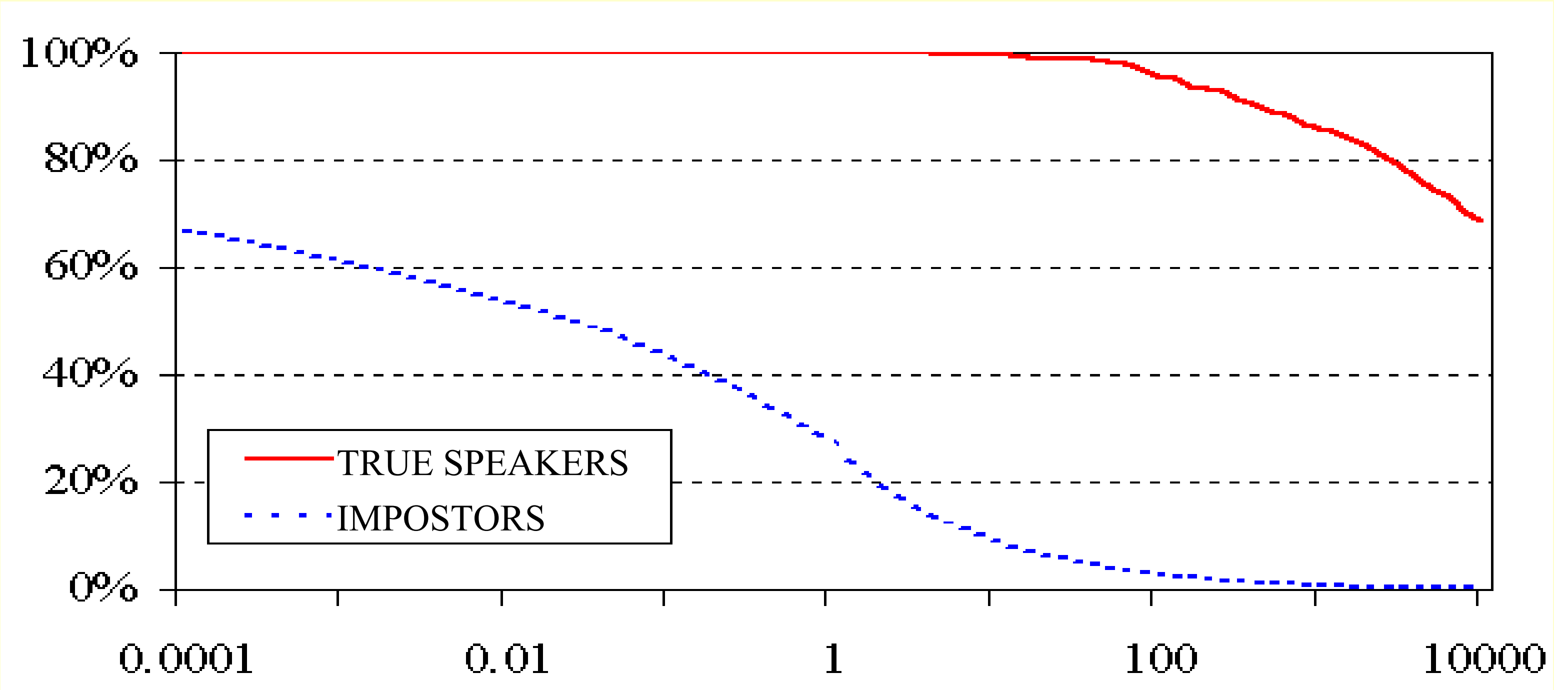




Inter: 0.64042

Intra: 0.26585

LR: 0.41512

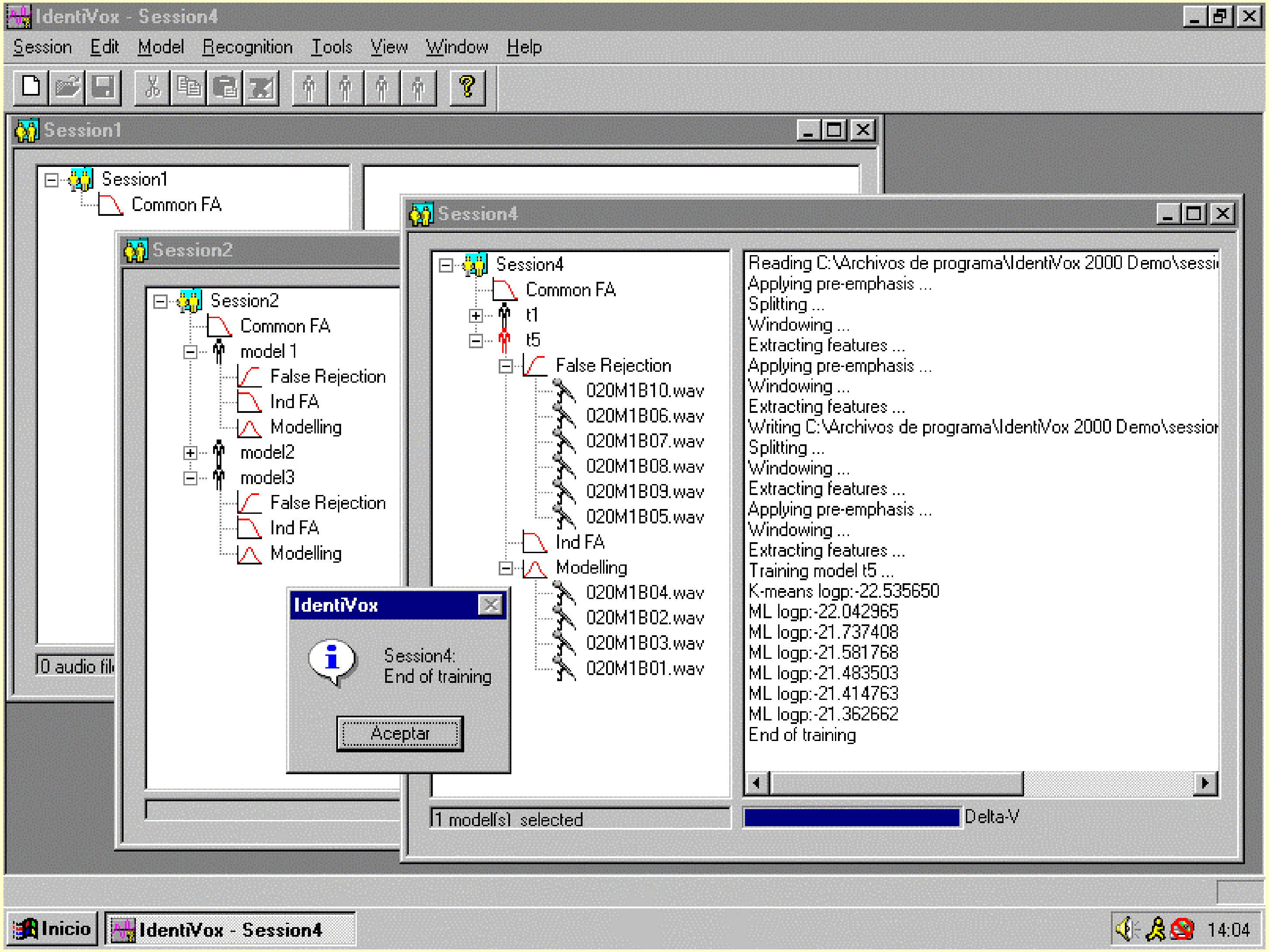


- LR calibration expressed in terms of proportion of cases with *"LR values greater than …"* [Tippet, 68; Evett, 96], that is, for any *x*-axis value each curve shows proportion of cases with LR greater than *x*.

- Tippet plots expressing opposite hypotheses: **C**, the system providing high LRs, and $\overline{C}$, the system providing low LRs.

- The greater the separation between curves, the higher the discriminating power of the technique.

LR-based System Design: *IdentiVox LR*



```
Reading C:\Archivos de programa\IdentiVox 2000 Demo\sessi
Applying pre-emphasis ...
Splitting ...
Windowing ...
Extracting features ...
Applying pre-emphasis ...
Windowing ...
Extracting features ...
Writing C:\Archivos de programa\IdentiVox 2000 Demo\sessior
Splitting ...
Windowing ...
Extracting features ...
Applying pre-emphasis ...
Windowing ...
Extracting features ...
Training model t5 ...
K-means logp:-22.535650
ML logp:-22.042965
ML logp:-21.737408
ML logp:-21.581768
ML logp:-21.483503
ML logp:-21.414763
ML logp:-21.362662
End of training
```

Session4:
End of training

What does *"Real Forensic Conditions"* mean in system evaluation?

- **Real Forensic Procedure: Not just standard SV system and binary decision (accepted/rejected, match/no match), but rather, estimation of LR through an appropiate method.**

- **Real Forensic Tasks: Target Speech for training and testing, Speech Controls, Reference population; and single/multi-session availability.**

- **Real Forensic Speech: Type of Speech and Conditions usually found in real cases.**

Real Forensic Speech

- Telephone speech
  - Long Duration recordings
    - (Wire)Tapping (organized crime: illegal traffic of drugs and other substances, illegal immigrations, job or sentimental conflicts...)
      - Conditions: Very Long Term (minutes of speech), multi-session, multi-channel, 10s of speakers involved, Standard telephone-channel conditions
    - Threats (person-to-person revenge, extortion...)
      - Conditions: long-term (> 1 min.), 1-5 speakers involved, emotional variability
  - Short duration recordings
    - Terrorist Threats
      - Conditions: short duration (< 20 s), 1-2 speakers involved, single session, use of automatic answering machines. 100s of potencial suspects.

- Microphone speech (Hidden body-mic. recordings):
  - Police investigation guidance (i.e., monitoring in cafeteria, prison, ...). Mainly Speech Enhancement
  - Private interviews (i.e., threats, extortion)

Experimental conditions and Database

- Real land-line telephone spontaneous multi-session data from AHUMADA/GAUDI database [Ortega, 00].
- Hamming windows of 32 ms., 50% overlapping, MFCCs+Δ+ΔΔ, CMN channel compensation.
- Speaker & Ref. population models: 1 minute of read speech, GMMs obtained through 32-gaussian ML training.
- Speech controls (SC, target speaker) and Test Files (TF): extracted from phonetically balanced utterances and 10-digit strings from different sessions.
- Reference population: 249 (122M+127F) separate spks.
- Suspects: 116 (52M+64F) speakers acting as “true” for own target model, and “false” for other targets (1,000 “true” LRs + 20,000 “false” LRs per task).

Tasks and Real Forensic Correlation

•Task 1 (T1): Single session speech in both training and testing.

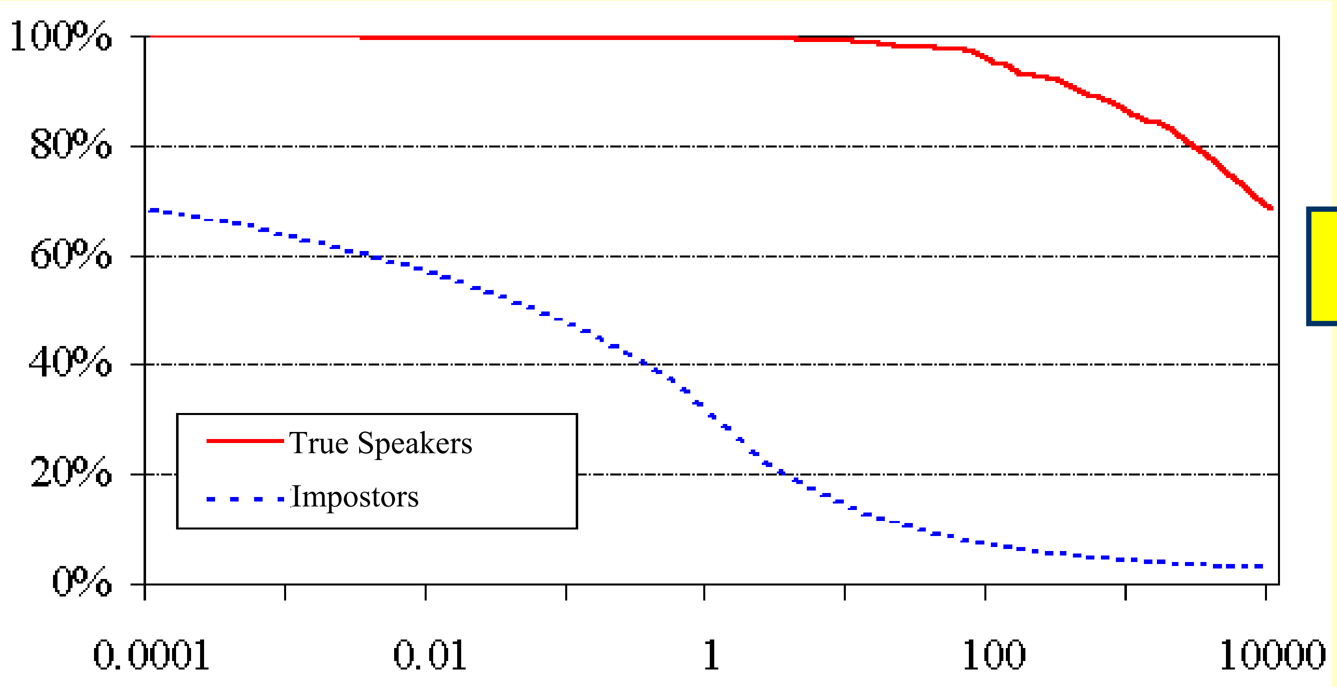
Forensic correlation: Suspect acknowledges his own voice except for some "sensitive" utterances, all in the same conversation.

•Task 2 (T2): Multisession training /single session testing.

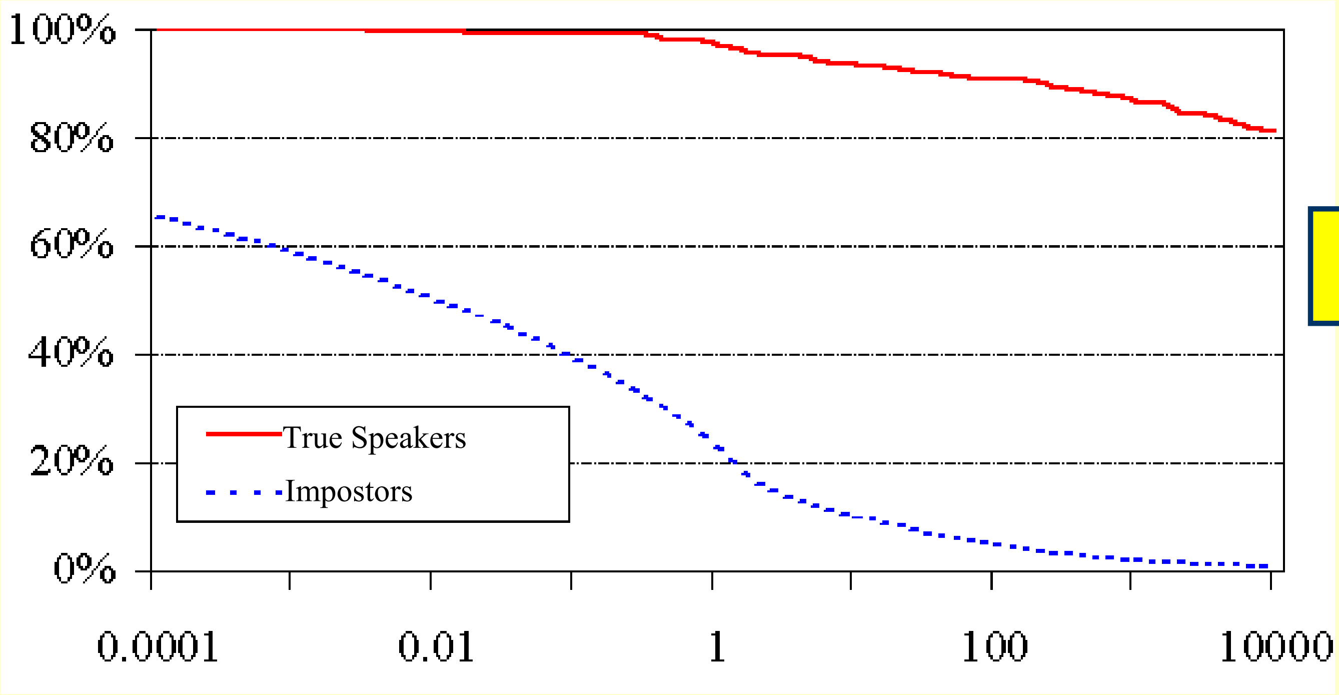
Forensic correlation: Suspect acknowledges several "irrelevant" conversations, but not other(s) "sensitive" one(s).

•Task 3 (T3): Single session training / multisession testing.

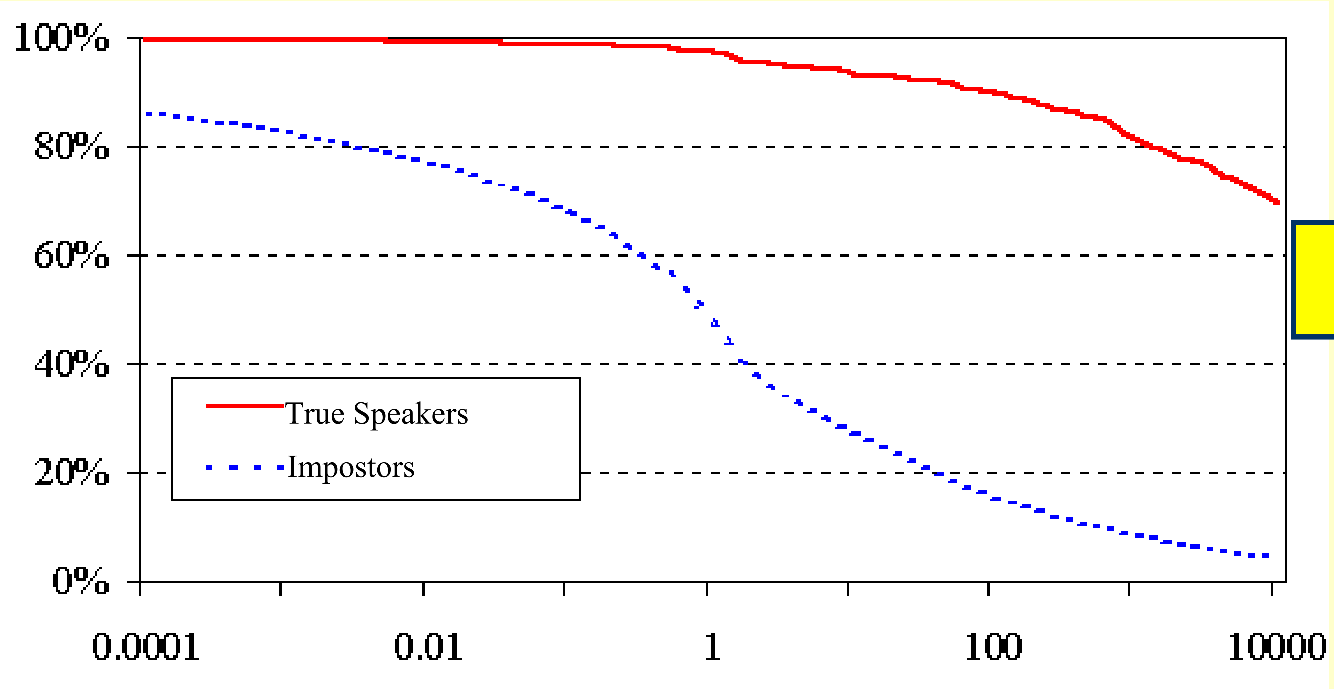
Forensic correlation: Suspect is recorded in Court and is not recognizing any other speech evidence(s) as belonging to him.


100%
80%
60%
40%
20%
0%
0.0001
0.01
1
100
10000
True Speakers
Impostors

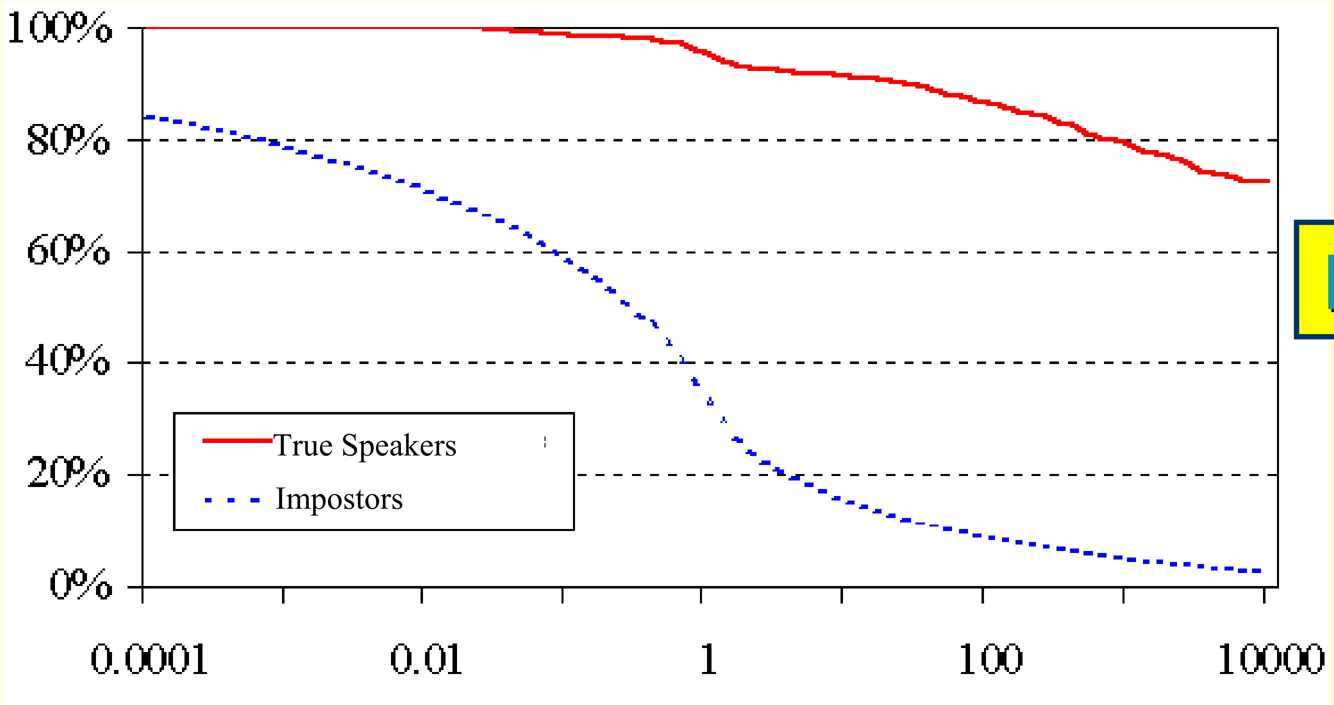
Male


100%
80%
60%
40%
20%
0%
0.0001
0.01
1
100
10000
True Speakers
Impostors

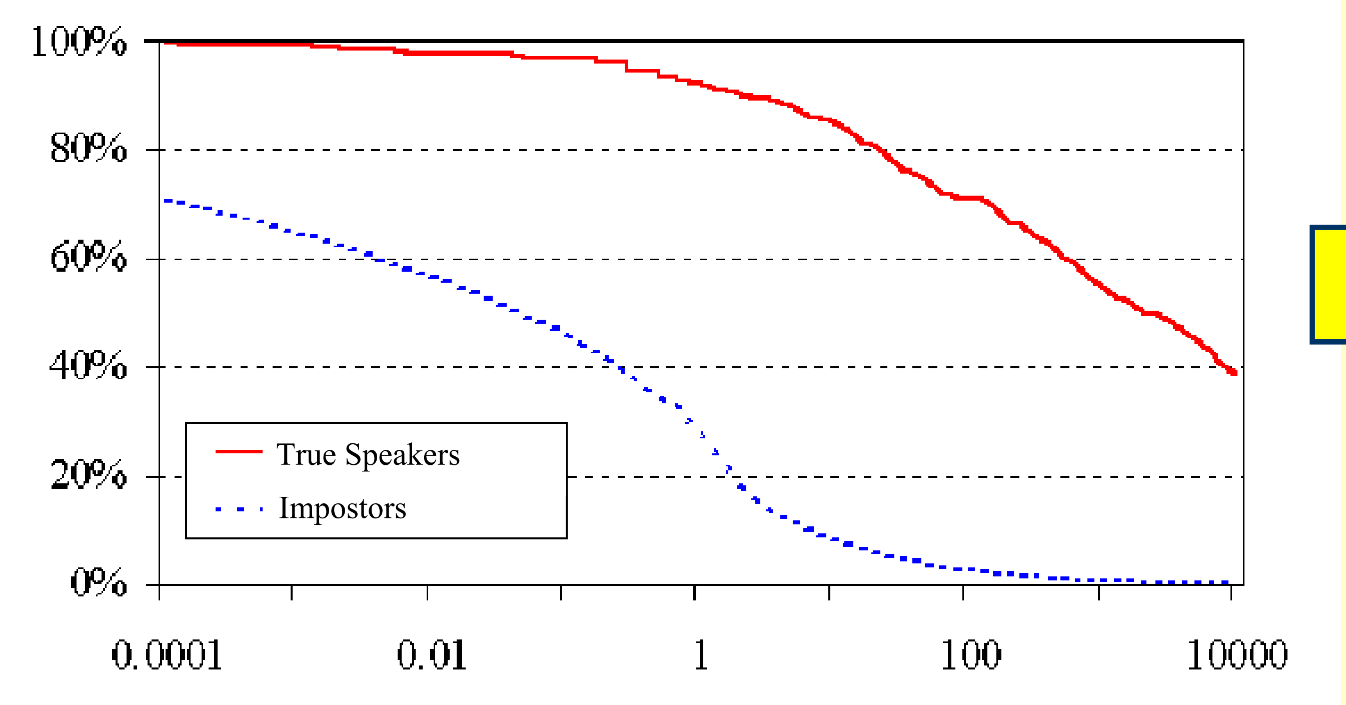
Female


100%
80%
60%
40%
20%
0%
0.0001
0.01
1
100
10000
True Speakers
Impostors

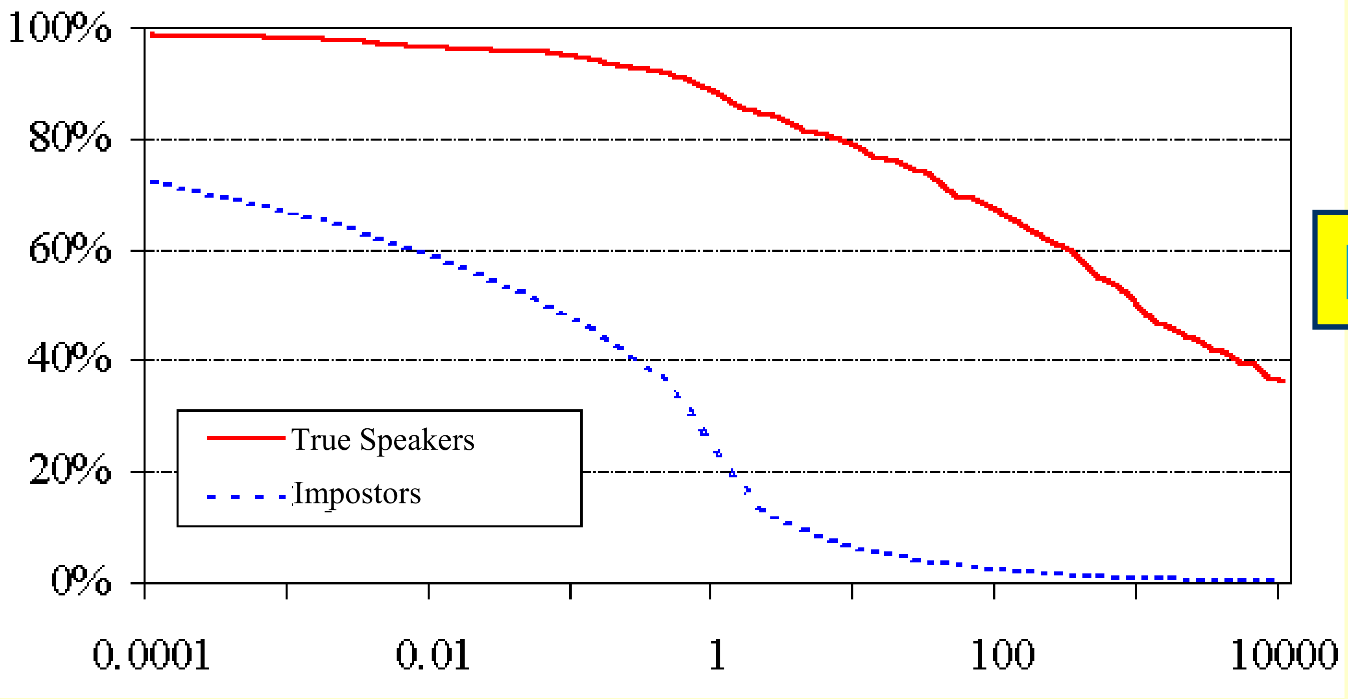
Male


100%
80%
60%
40%
20%
0%
0.0001
0.01
1
100
10000
True Speakers
Impostors

Female


100%
80%
60%
40%
20%
0%
0.0001
0.01
1
100
10000
True Speakers
Impostors

Male


100%
80%
60%
40%
20%
0%
0.0001
0.01
1
100
10000
True Speakers
Impostors

Female

Conclusions

- **High discriminating abilities when the system is tested/calibrated with real multisession telephone speech.**

- **In every single LR test a separate big reference population is employed, reinforcing statistical significance of results.**

- **Multiple post-processing of real testing evidences is possible as multiple short-length tests are available from questioned recording.**

- **Results based on LR approach demonstrate usefulness and reliability of this automatic procedure for forensic science in spk. recognition cases.**

Future Trends

•Search of optimal population sets for forensic cases, considering size, channel variability, and speech contents.

•Although standard long duration speech is found in many forensic cases, degraded and/or short duration training speech should be included in future tests ⇒ Real forensic database? Legal and definition limits.

•Combined LR scores with recent new approaches: Idiolects [Doddington 01], “Magic”GMMs [Reynolds 01], Phonetic SR [Campbell 01], Intonation [Weber, 01], etc.