ON THE USE OF AUDITORY AND AUTOMATIC SYSTEMS TO HANDLE MISMATCHED CONDITIONS IN FORENSIC SPEAKER RECOGNITION

Anil Alexander †, Damien Dessimoz‡, Filippo Botti‡, and Andrzej Drygajlo †

† **Swiss Federal Institute of Technology, Lausanne**
Signal Processing Institute
‡ **University of Lausanne**
School of Criminal Sciences

Outline

- Bayesian interpretation in Forensic Automatic Speaker Recognition
- Strength of evidence in Aural and Automatic Speaker Recognition
- Evaluating the strength of evidence
- Comparing performances of Aural and Automatic Speaker Recognition Systems
- Necessity of adapting the automatic systems to different conditions
- Discussion of the perceptual cues used by laypersons to recognize speakers
- Conclusion

Bayesian Interpretation in Forensic Automatic Speaker Recognition

- **Evidence (E)**

  The score obtained comparing statistical model of the suspect's voice and a questioned recording (trace)

H_0 – The two recordings have the same source

H_1 – The two recordings have a different source

$$LR = \frac{p(E \mid H_0)}{p(E \mid H_1)}$$

Likelihood Ratio (LR)

The relative probability of observing a particular score "E", with respect to two competing hypotheses

Experimental Framework

- **Test Database** ***(subset of <<polyphone IPSC-02>>)***
  - **Speakers (Swiss-French)**
    - 5 traces for each speaker and condition (PSTN, GSM and Noisy PSTN)
    - 1 suspect reference recording for each speaker and condition (PSTN and GSM)
  - **Recording Lengths**
    - 15 second / trace : simulation of real cases (undisguised hoaxes, menacing calls etc)
    - 90 seconds / reference recording
- **Testing Scenarios Evaluated**
    - Reference PSTN vs Traces PSTN
    - Reference PSTN vs Traces GSM
    - Reference PSTN vs Traces Noisy PSTN
    - Reference GSM vs Traces GSM

Aural Speaker Recognition

Experimental Framework

- **Listeners**
  - » 90 listeners whose mother-tongue is French
  - » Laypersons with no phonetic training
  - » Same computer and headphones
- **Training**
  - » No limitation on the number of listening trials
- **Testing**
  - » Verbal scores scale from 1 through 7
  - » Perceptual cues

Perceptual Verbal Scale and Perceptual Cues

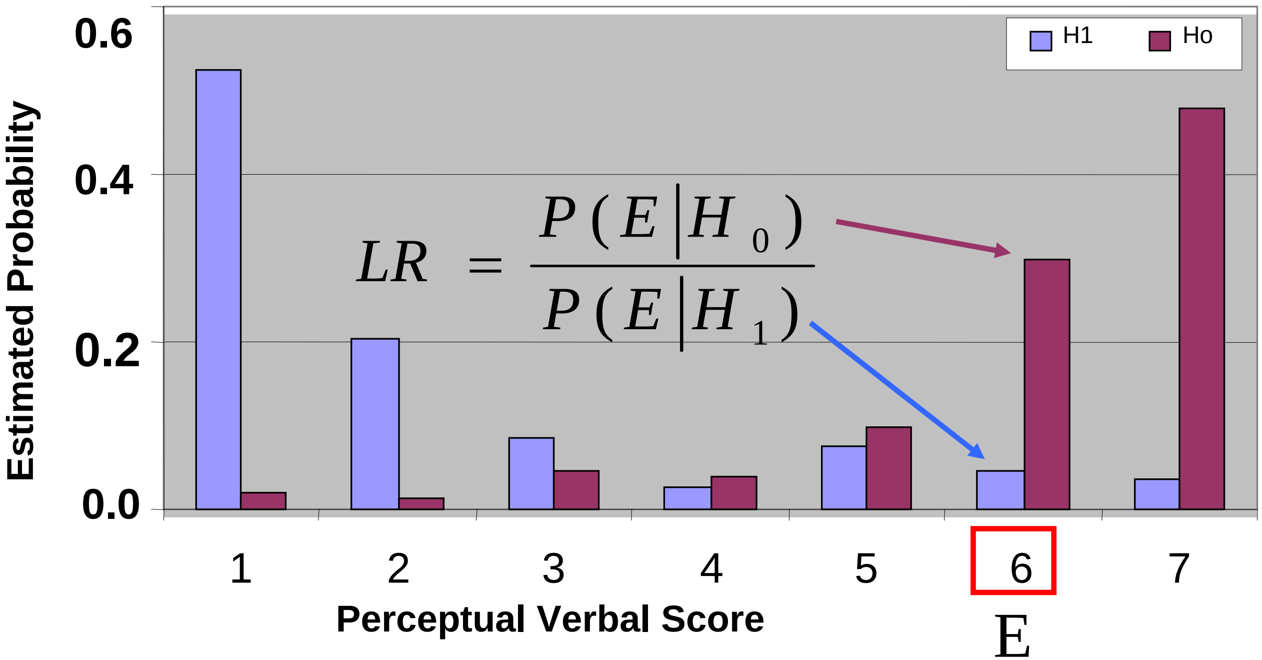
Perceptual Verbal Scale

Score 1 I am sure that the two speakers are not the same
Score 2 I am almost sure that the two speakers are not the same
Score 3 It is possible that the two speakers are not the same
Score 4 I cannot decide
Score 5 It is possible that the two speakers are the same
Score 6 I am almost sure that the two speakers are the same
Score 7 I am sure that the two speakers are the same

Perceptual Cues

Subjects asked to note factors they considered in recognizing speakers at the end of each session

Strength of Evidence for Aural Recognition



- *Discrete scores*
- *Histograms used to estimate the probabilities of scores for each hypothesis*

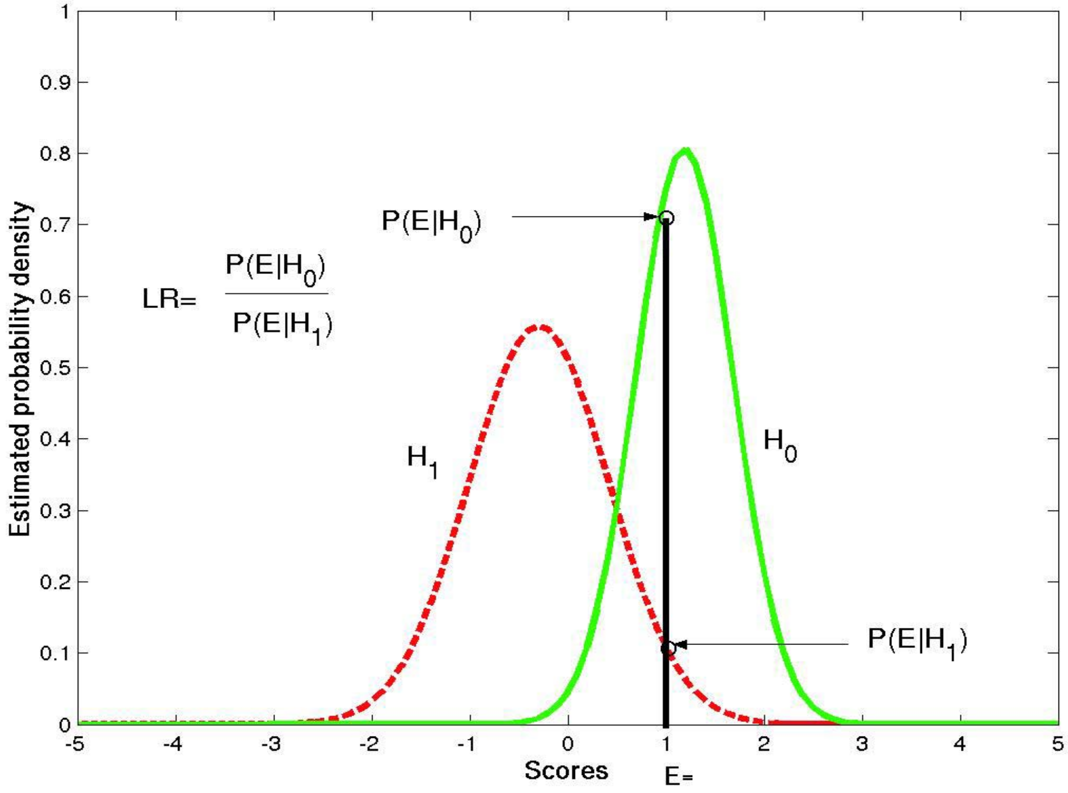
Likelihood Ratio (LR) = Ratio of the heights on the histograms for the two hypotheses at the point "E"

Automatic Speaker Recognition System

- **Recognition System**
  - **Feature extraction:** RASTA -PLP
  - **Statistical modeling:** Gaussian Mixture Modeling (GMM)
  - **Likelihood ratio** (Kernel Density Estimation)

- **Methodology**
  - Single questioned recording and single suspect recording
    (F. Botti., A. Alexander, and A. Drygajlo, "An Interpretation Framework for the Evaluation of Evidence in Forensic Automatic Speaker Recognition with Limited Suspect Data", in Proceedings of 2004: A Speaker Odyssey, Toledo, Spain, 2004, pp. 63–68.)
  - Suspect Reference (R) and Traces (T) databases – Subsets of « Polyphone IPSC-02 »)

Strength of Evidence for the Automatic System



- *Scores are continuous*
- *Kernel-density-based estimate of the probability density of scores for each hypothesis*

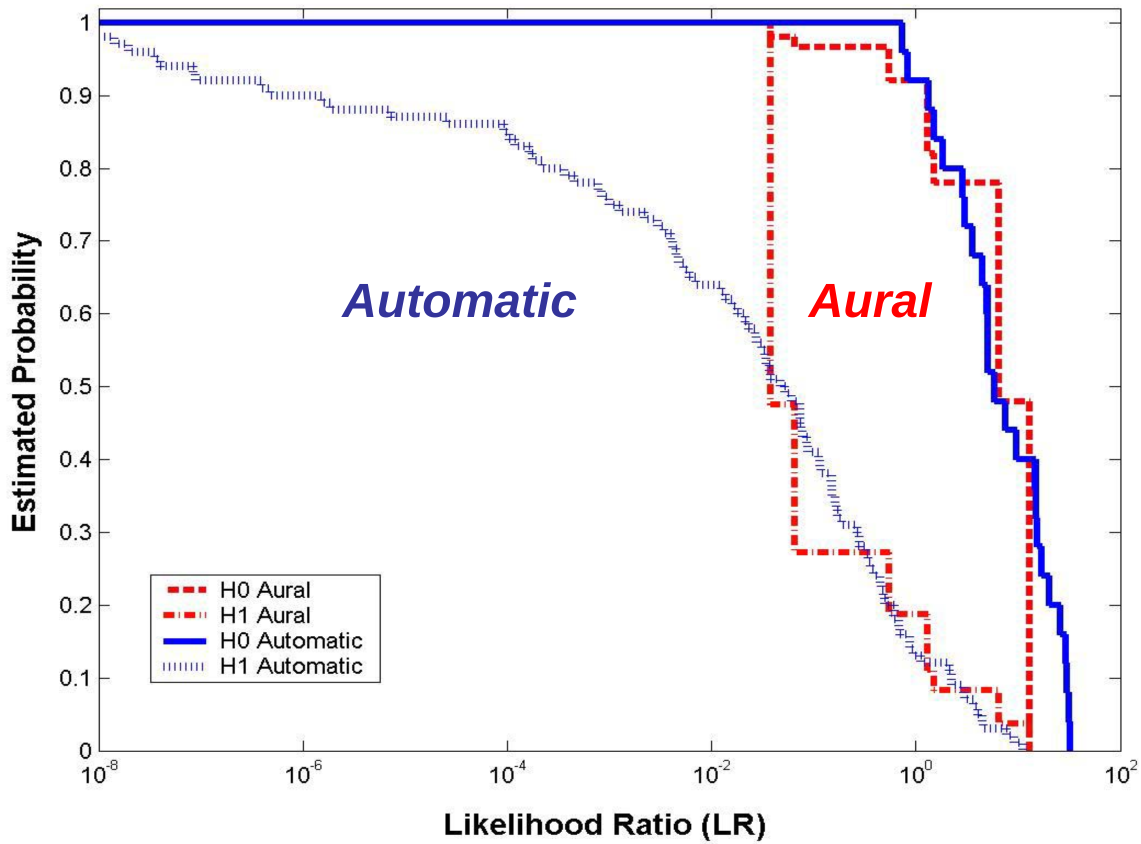
LR = Ratio of heights on the curves for the two hypotheses at the point "E"

Evaluating Strength of Evidence

- **Tippett Plots**
  - Representation of the proportion of likelihood ratios greater than a given LR, for cases corresponding to hypotheses H_0 and H_1, i.e. $P(LR(H_i)>LR)$
  - Separation between curves representing H_0 and H_1 indicates how well the system differentiates between cases in which each of the two hypotheses is known to be true

Evaluating Strength of Evidence in Matched Conditions

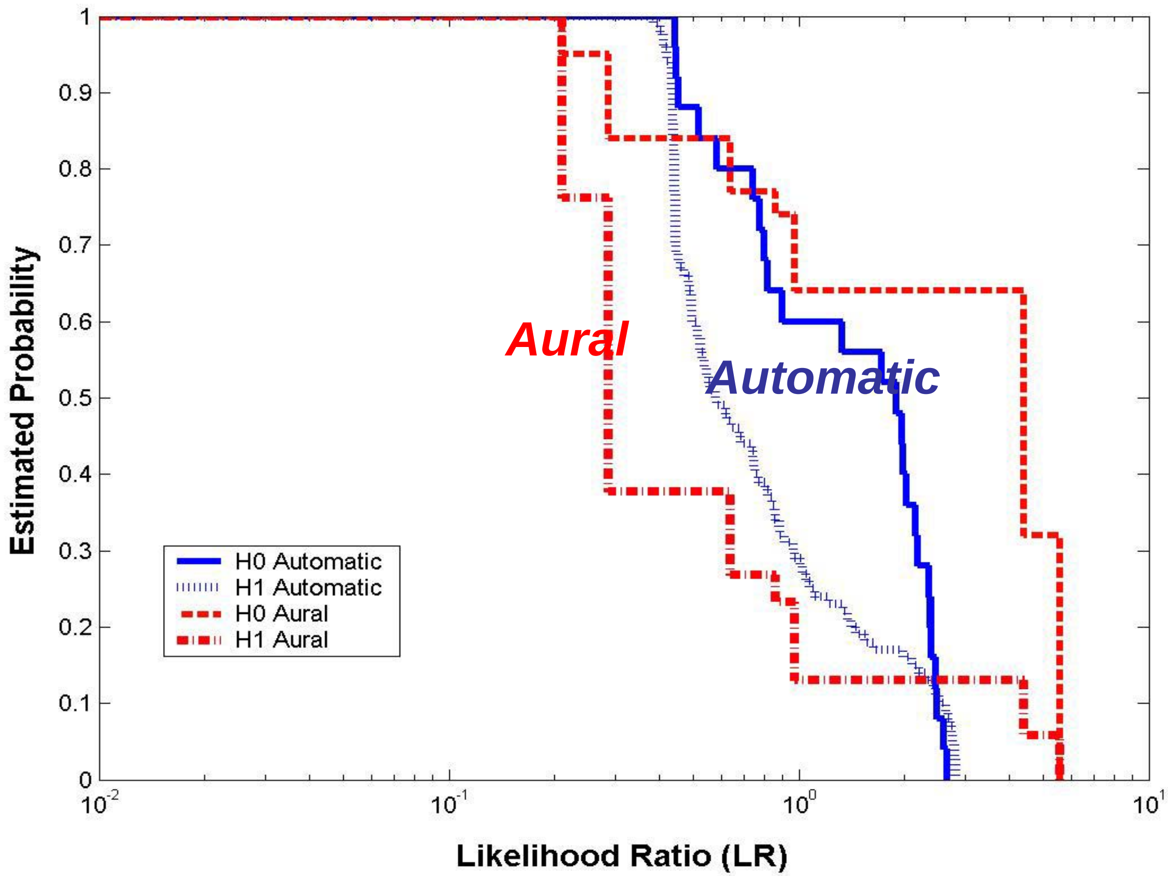
Ref. PSTN vs Traces PSTN



 # Similar separations between curves for aural and automatic systems

Evaluating Strength of Evidence in Mismatched Conditions

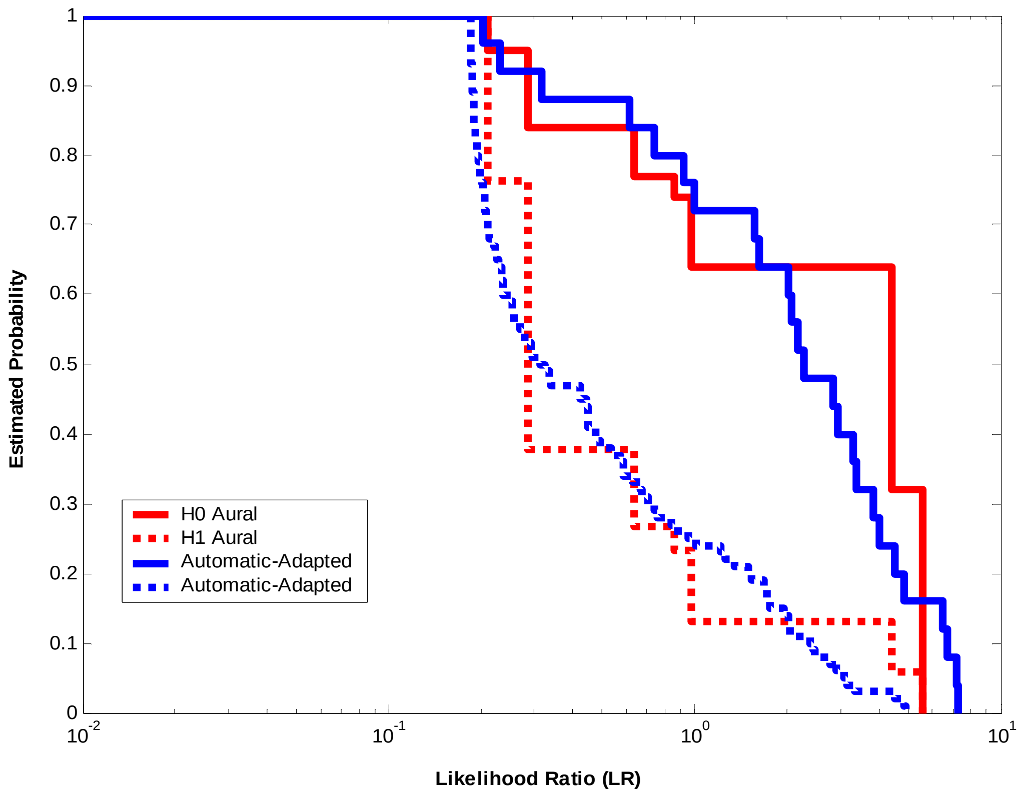
Ref. PSTN vs Traces Noisy PSTN



Better curve separation in aural recognition → *Better evaluation of LR for aural recognition in mismatched conditions*

Evaluating Strength of Evidence in Adapted Conditions

Ref. PSTN vs Traces Adapted Noisy PSTN



Adaptation for noisy conditions results in the improvement of performance of automatic recognition

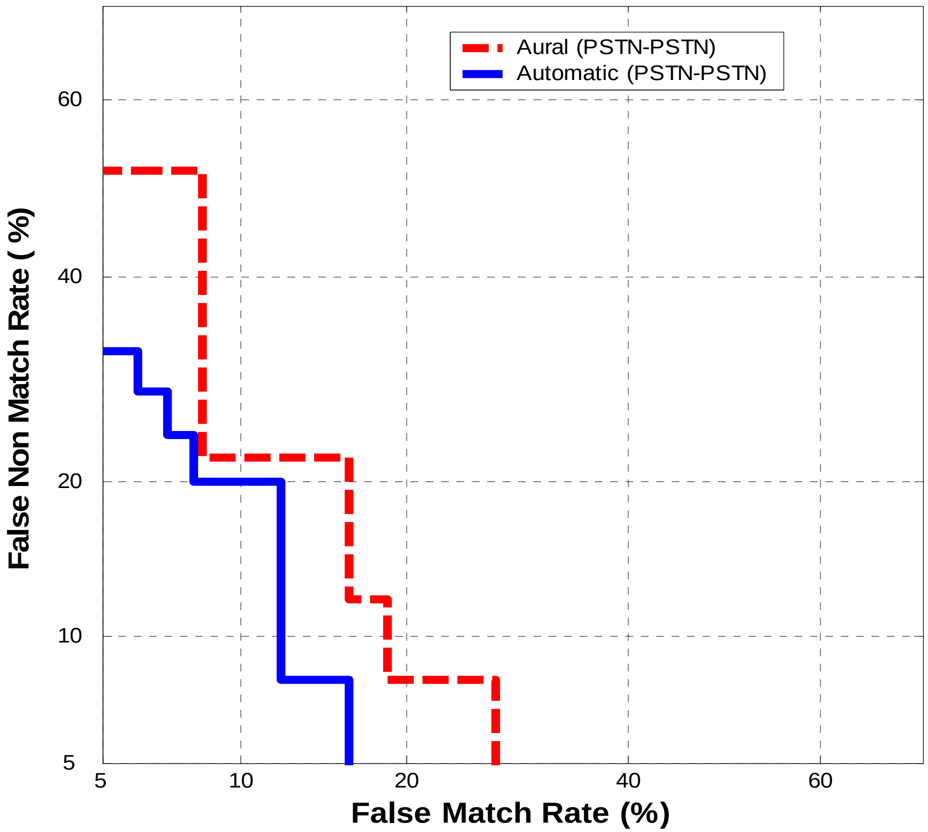
Performance Measurement
(in terms of Recognition Errors)

- Are the trends shown by the Tippett plots for the evaluation of the strength of evidence also shown by measuring the recognition errors in speaker verification ?

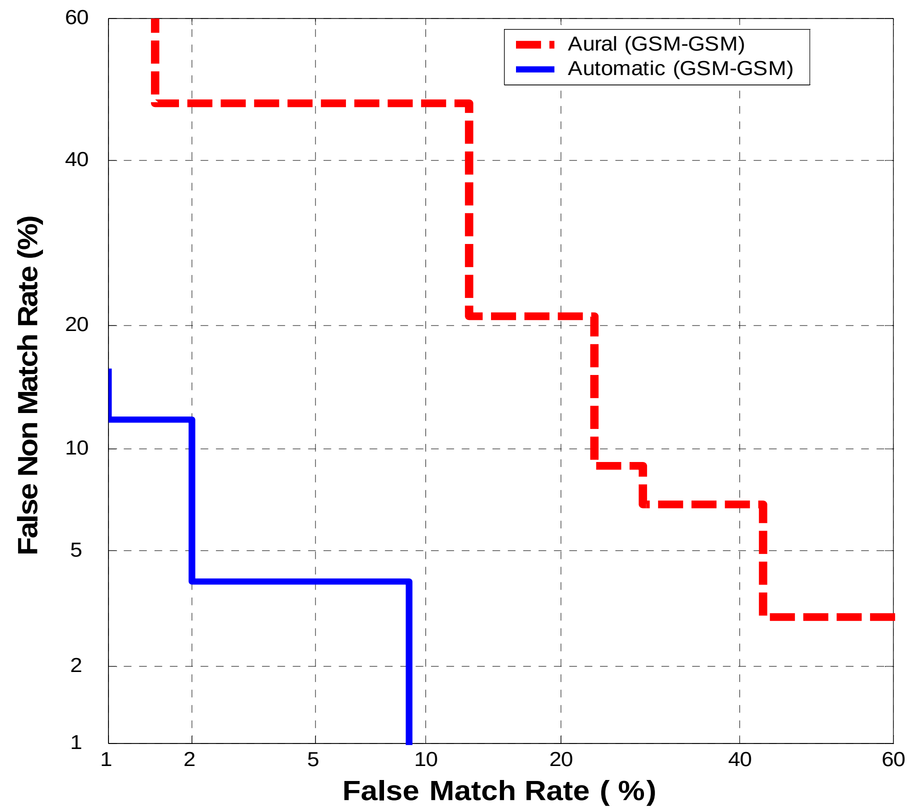
  - Detection Error Tradeoff (DET) Curves
    - Relative plot of False Match Rate and False Non-Match Rate varying a decision point
  - Equal Error Rate
    - when False Match Rate = False Non-Match Rate on DET curve *(useful to compare system performances)*

Matched Conditions : Comparing Aural and Automatic

Ref PSTN – Trace PSTN



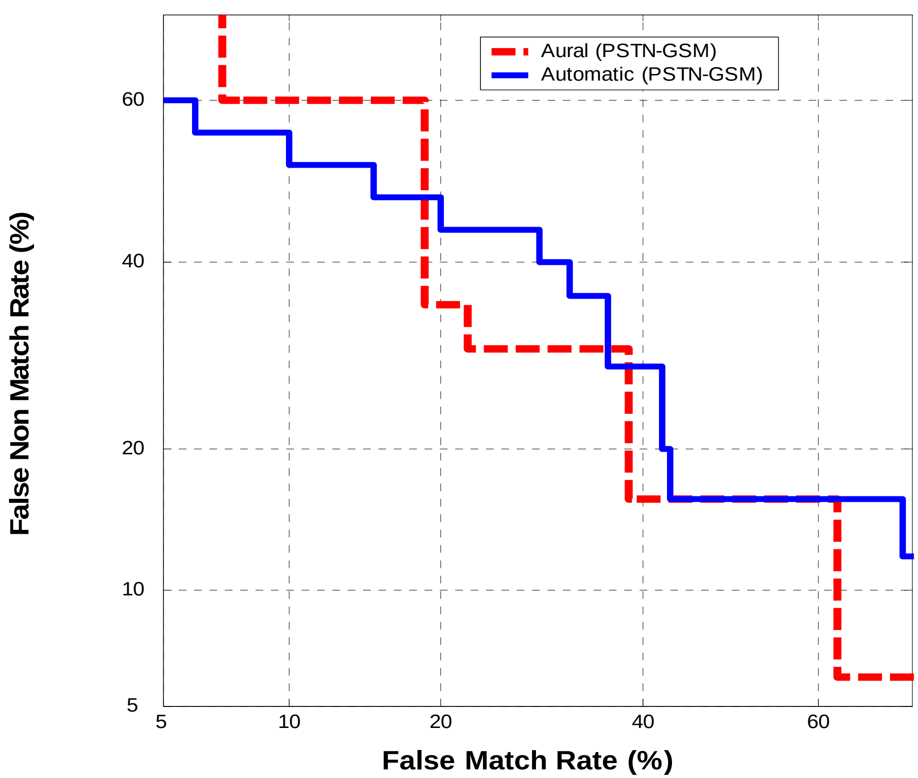
Ref GSM- Trace GSM



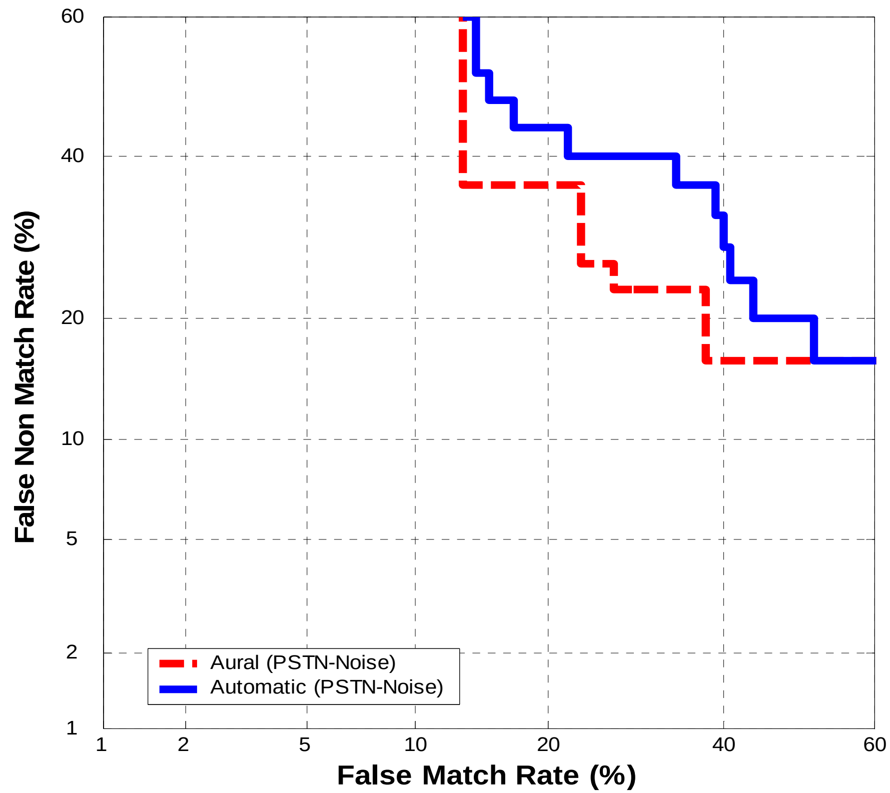
In matched conditions automatic performs better than aural recognition

Mismatched Conditions : Comparing Aural and Automatic

PSTN - GSM



PSTN – Noisy PSTN

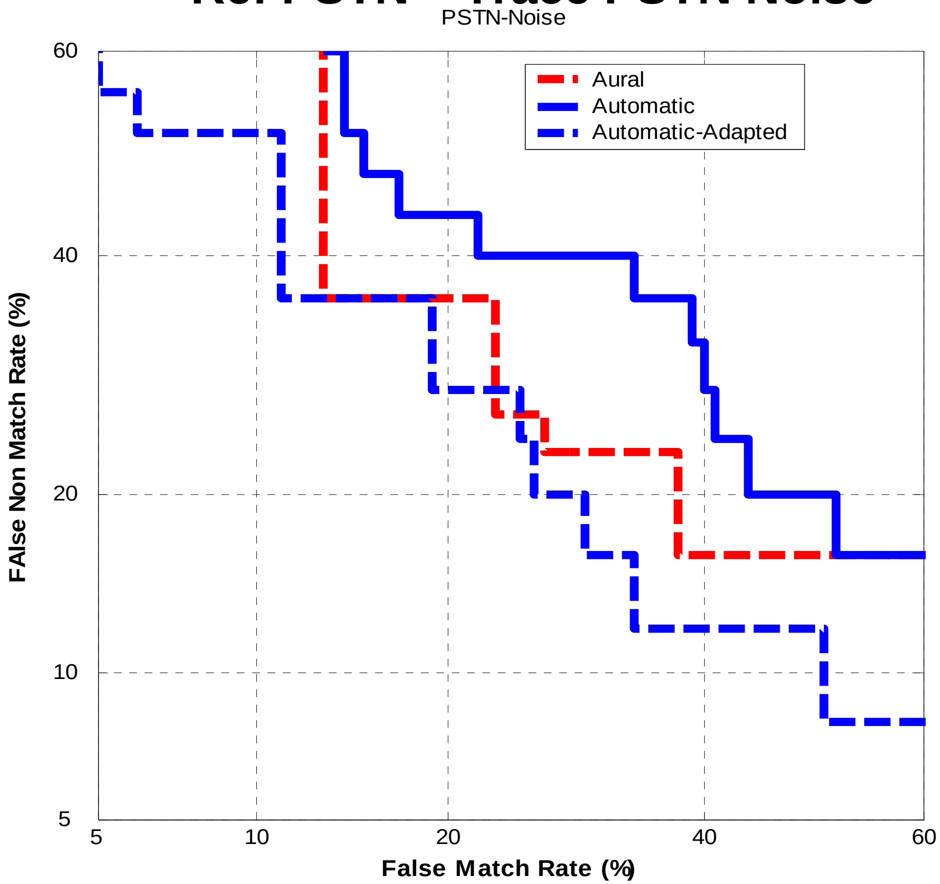


Mismatched Conditions → Automatic recognition shows similar or slightly degraded performance as compared to Aural recognition

⇨ Adaptation necessary

Adapting for Mismatched Conditions

Ref PSTN – Trace PSTN Noise



 # Adapted using ***Spectral Subtraction*** *Based on Minimum statistics (R Martin 1994)*

After adaptation for noisy conditions, automatic system shows similar or better performance compared to aural recognition

Performance Measurement

EER : Aural and Automatic Recognition

EE R	PSTN-PSTN	PSTN-GSM	PSTN-Noise	GSM-GSM
Aural	16%	30%	26%	21%
Automatic	12%	36%	36%	4%

– Both aural and automatic recognition perform better in matched than in mismatched conditions

– Automatic recognition performs better than aural in matched conditions

– Aural recognition performs better than automatic in mismatched conditions

Adaptability to different conditions - Necessity for automatic systems

Perceptual Cues Used by Laypersons

- Human beings use a lot of perceptual cues to recognize speakers

  e.g. pronunciation, timbre, intonation, rate of speech breathing, loudness, imagined physiognomy, etc.

- These cues are used in aural recognition, to recognize speakers in different conditions

- Evaluating relative perceived importance of these perceptual cues

Relative Importance of Perceptual Cues

	PSTN-PSTN	PSTN-GSM	PSTN-Noise	GSM-GSM
Accent, prononciation, articulation 30%	34%	31%	30%	26%
Timbre 24%	25%	25%	22%	22%
Intonation 20%	16%	24%	18%	21%
Rate of Speech 10%	9%	7%	12%	12%
Speech Defects 7%	6%	7%	8%	5%

 # The perceived importance of these perceptual cues is relatively constant in different conditions for human listeners

Conclusions

- In matched recording conditions of training and testing, automatic recognition systems performed better than aural systems.
- In mismatched conditions, baseline automatic systems showed comparable or slightly degraded performance compared to aural systems.
- Aural recognition relies on high-level perceptual cues to recognize speakers.
- Baseline automatic speaker recognition systems should be adapted to each of the mismatched conditions to improve performance.

Questions?

Thank you for your attention.