# Speaker vérification by inexperienced and experienced listeners vs speaker verification system

Juliette Kahn[1,2] , Nicolas Audibert[3], Solange Rossato[2] and Jean-François Bonastre[1]

[1]University of Avignon – CERI/LIA – France
[2]University of Grenoble – LIG – France
[3]CNRS, Paris 3 Sorbonne Nouvelle – LPP – France

# Goals of HASR

- For the first time in 2010
- **« How can human experts effectively utilize automatic speaker recognition technology? »**
- Participation open to all who might be interested, ranging from “experts” to “naïve” listeners

# Task

- Verification Task with 2.5 minutes samples extracted from the SRE10 core test
- A selection of **difficult trials** done by NIST based on scores given by a particular system
- 2 sets:
  - HASR 1 : 15 trials
  - HASR 2 : 150 trials (include HASR1)

# LIA-LIG participation in HASR

# Listening and scoring protocol

- 3 native French listeners (2 female, 1 male)
- Allowed to examine spectrograms and band-pass filtered signals
- True/false decision and confidence rating
- Submitted decision = majority voting

# Mapping

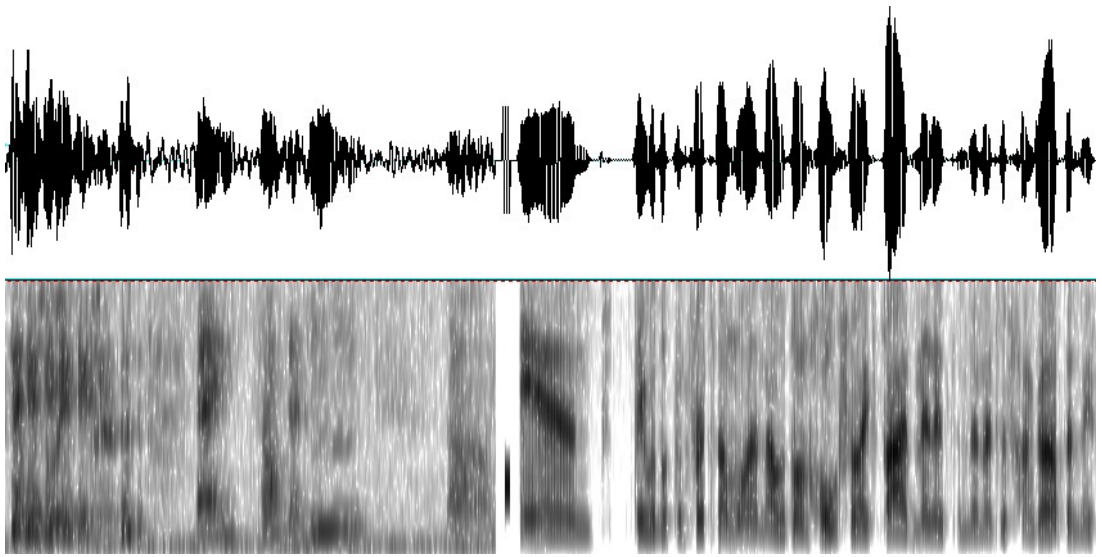
- Submitted score = mapping of human decisions with SVM-GMM automatic system score distributions
- Purpose : comparing the automatic system results and the humans submission

# Speech material

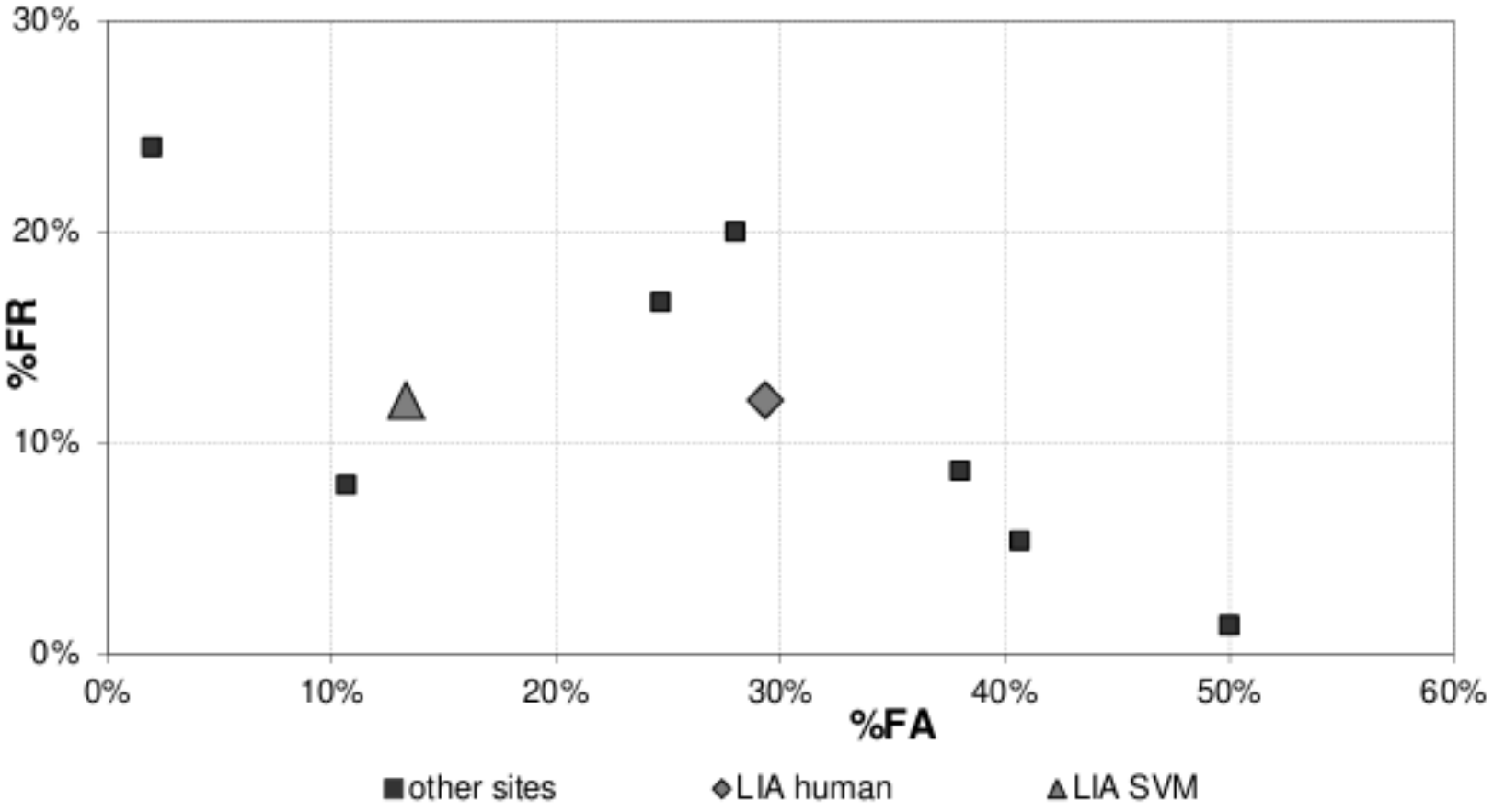
- NIST provided pairs of 2.5 minutes speech samples (like for automatic systems)
- Too long for an auditory comparison of non-familiar voices
    - Usually around 6 and 10 second extract in perception test

# Automatic stimuli generation

- Selection of 6 seconds-long extracts from the model and test segments based on energy detection (MISTRAL/ALIZE tools)
- Concatenation of beep-separated energy-normalized extracts alternated between model and test => 60 seconds-long stimuli



# Results for HASR 2 (150 trials)



# Human performance analysis

- Inter-listener agreement

| decision | non-target (99) | target (51) |
|---|---|---|
| false | 37 | 8 |
| true | 16 | 15 |

- Mean confidence ratings

| decision | non-target | target |
|---|---|---|
| false | 2.5 | 2.3 |
| true | 2.2 | 1.8 |

# Discussion

- Listeners feeling: evaluation of human ability to normalize for channel differences rather than voice similarity
- No actual acoustic analysis performed in this evaluation: might help human decision making
- Limitations of the protocol
  - Enough trials?
  - Trials release procedure does not allow randomization

# Extended study

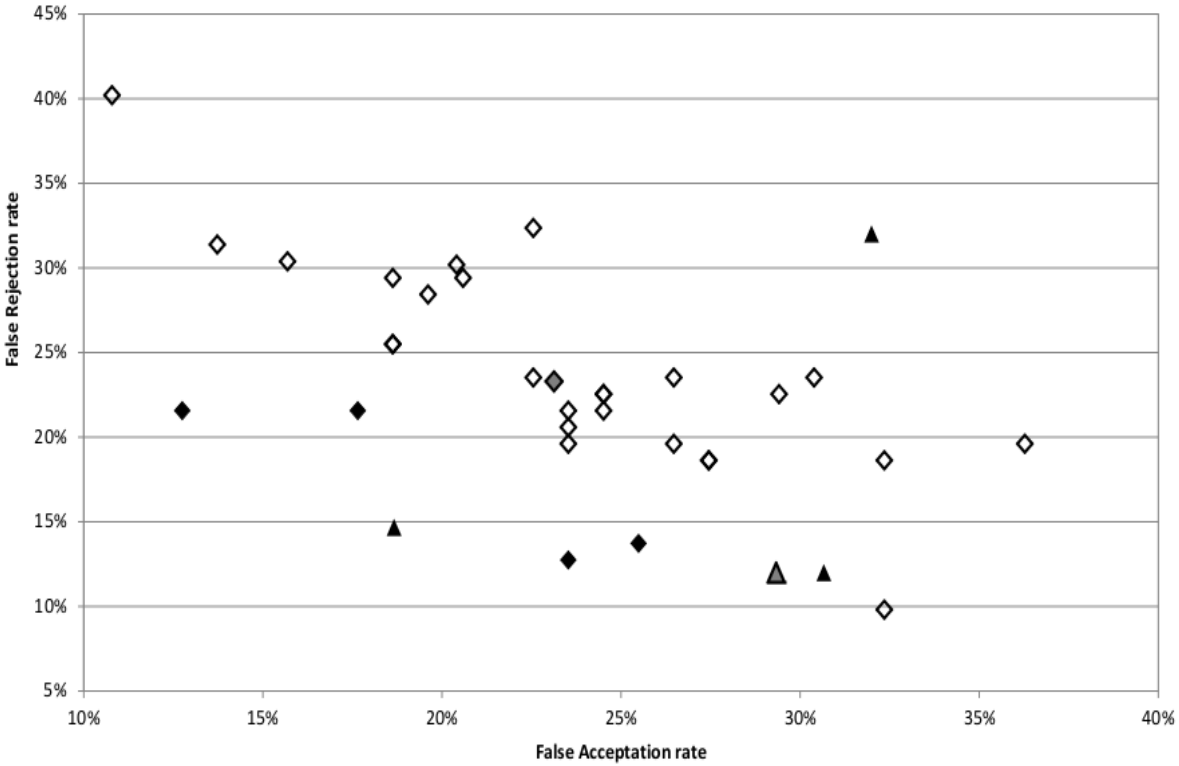
# Questions

- Influence of the number of listeners ?
- Differences between experienced and non-experienced listeners ?
- Complementarity between the humans and the system's decisions ?

# What changed ?

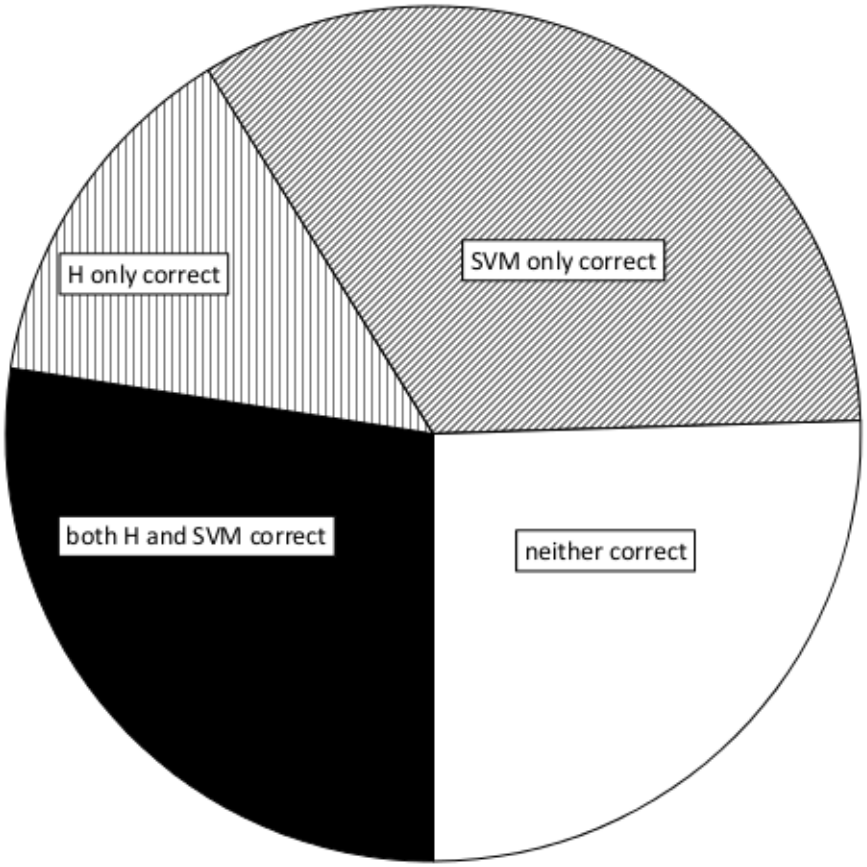
- More listeners (all native French)
  - 30 non-experienced listeners
  - 10 experienced listeners
- Randomized presentation of the trials
- Balanced number of non-target and target trials
  - Natural prior is 0.5
- Only one listening per trial allowed

# Results : Non-Experienced listeners vs NIST submission

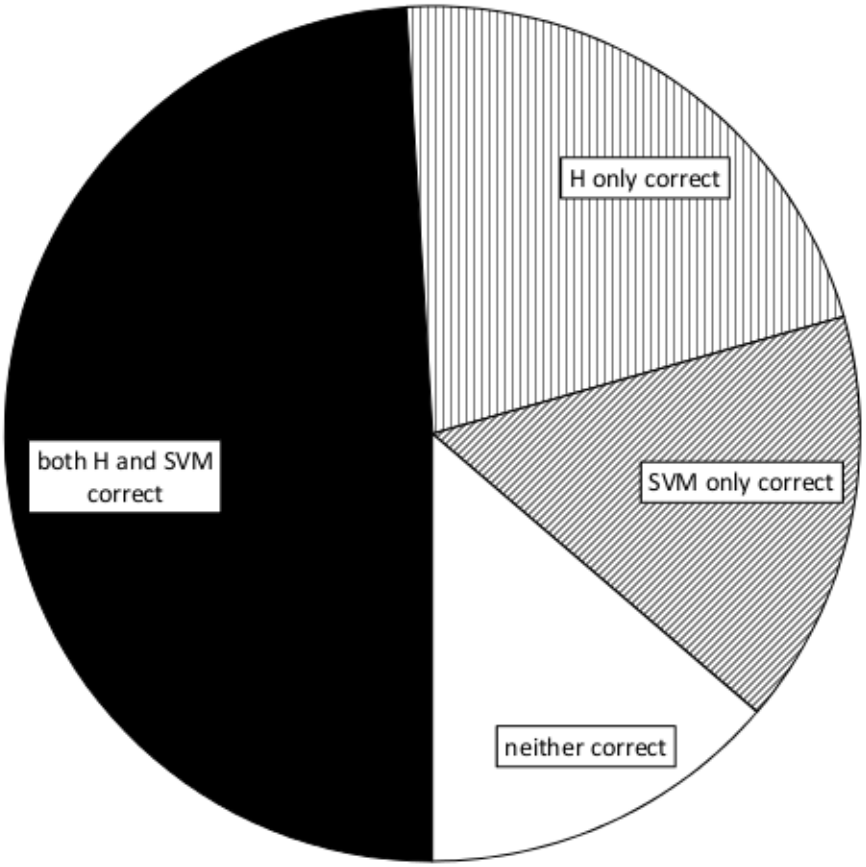



- Only 4 non-experienced listeners performed above chance level
- Very large gap of performance accross trials (3% to over 90% correct answers)
- No difference between male and female trials
- Different behaviours : « yes-listeners » vs « no-listeners »
- Correlation between performance and the English level

# Human and SVS complementarity




Non-target trials



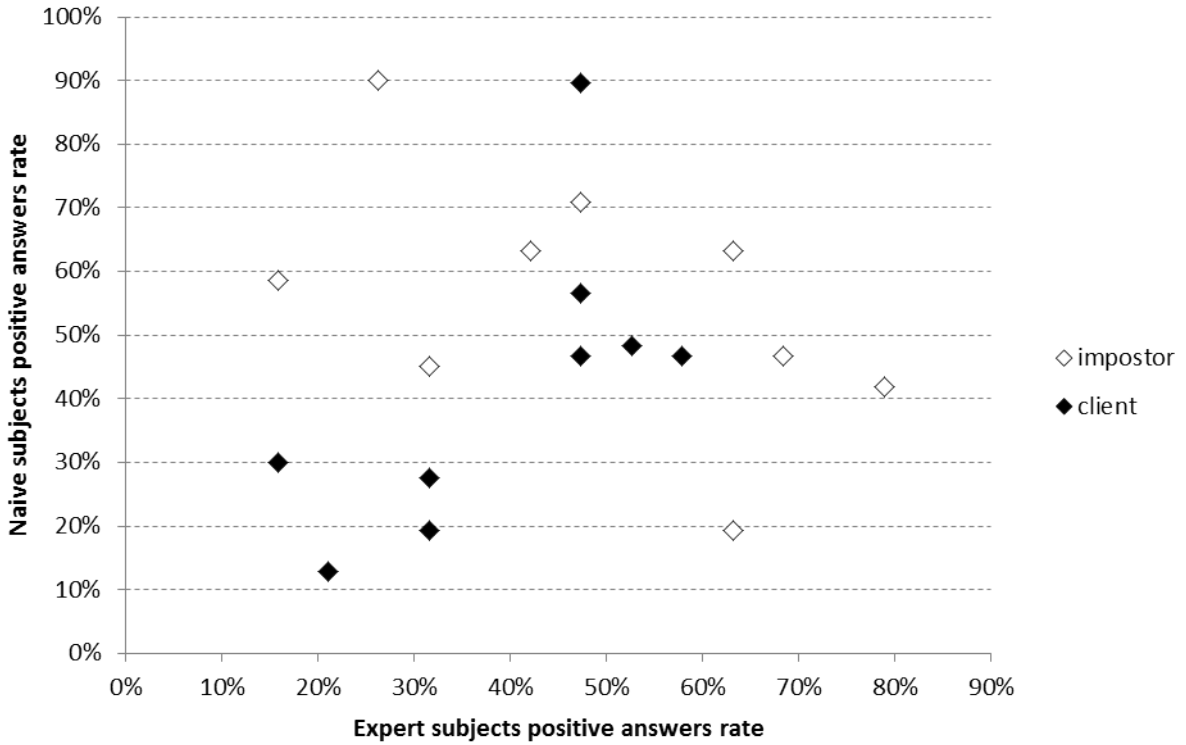

Target trials

# Non-experienced vs experienced listeners

Compared on a smaller set of trials

Equivalent performance between the two groups

- 39% vs 33% of correct answers



# Suggestions and future work

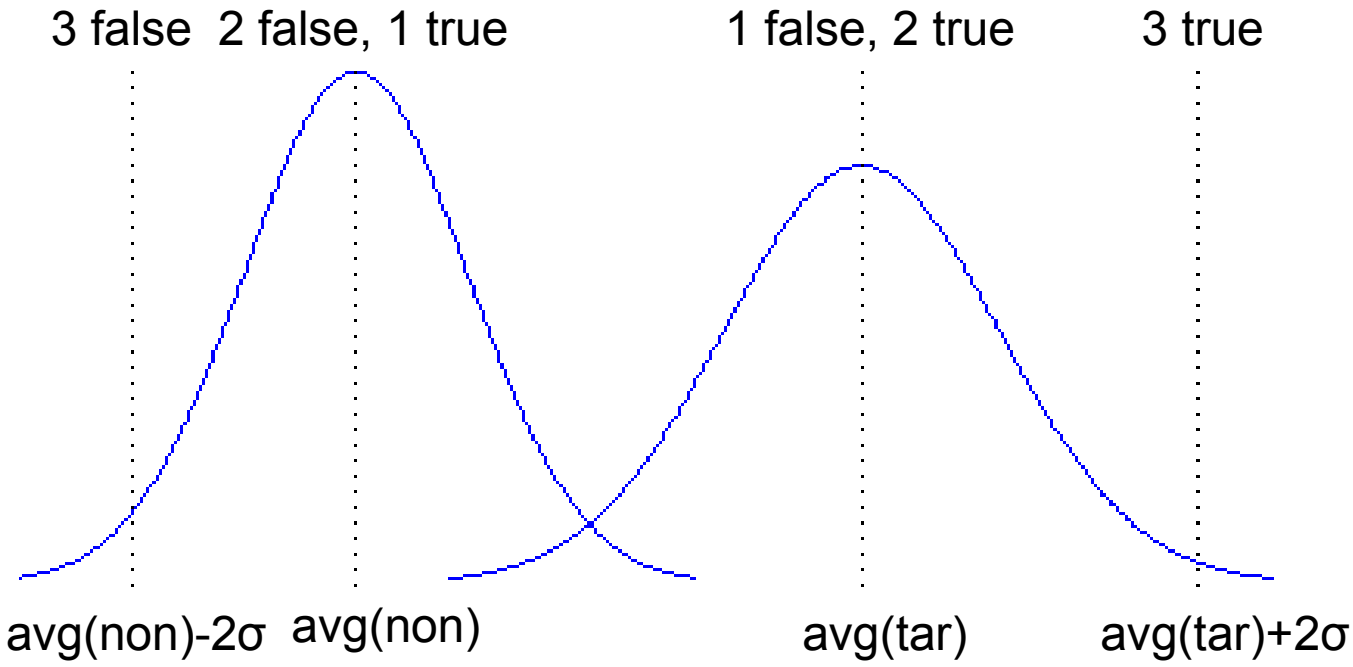
- How the human can help the system ?
  - Examine the trials with the scores near the threshold of the system
- How such performance variation according to the trial can be explained?
- Replicable with native listeners?

Thank you

Questions ?

# Decision to score mapping

- GMM-SVM system with FA (cf. Larcher *et al*.)
- System developed on NIST SRE 2008 data
- Mapping according to human decisions



# Mapping between listeners ratings and SVM scores