*AI*IA 2009: International Conference on Artificial Intelligence*
EVALITA Workshop

Reggio Emilia, 12 Dicembre 2009

The Unifi-EV2009-1 Protocol for Evalita 2009

Prof. Ing. Monica Carfagni, Ing. Matteo Nunziati
{monica.carfagni,matteo.nunziati}@unifi.it

Agenda

- Our goals
- Protocol Description
- Decision Threshold
- Background and Development Populations
- Results
- Conclusions

Our goals

- Methodology: use standard tools and populations

- The LT aim is to slowly introduce ASR in Italian forensics
  => we need careful and documented steps in order to make ASR widely accepted in courts !!!

- 2009 goal: adopt classical, well assessed (even if obsolete) technologies
  => huge amount of comparable tests in literature

Protocol Description 1/3

- Software based on Alize/Mistral project + SPro
- Front-end: 13 MFCC + Δ + ΔΔ + loge
- Silence/bg noise removal: 2-component GMM against loge.
  =>Cluster with lowest energy is discarded

Protocol Description 2/3

- Engine: GMM-UBM model against MFCC + Δ + ΔΔ
- Post-1: Scores T-normed by using a "best 10" approach
- Post-2: LR is retrieved from target and non-target score distributions

=> score sets approximated with Gaussian distrib.

Protocol Description 3/3

- CTS: we apply NIST rules
- OTS: we can't apply NIST rules. A diarization tool is required but missing !!
- IDEA (ugly patch again): perform a "DMTI-standard" session with SNR measurements and so on... then apply ASR

Decision Threshold 1/2

- Decision Threshold is a non-problem: it is province of the court
- But... it is mandatory for the track !!! so we have to define it!
- We use a really neutral threshold: P_{post}>50% means "same speaker"

Decision Threshold 2/2

- P_{prior}?! It is an interesting problem…
- We have fixed an arbitrary flat distribution: given *n* speakers in the test set, each one has $P_{prior}=1/n$
- According to Bayes, this leads to: "LR > n-1" means "same speaker"

Background and Development Populations

Both are:

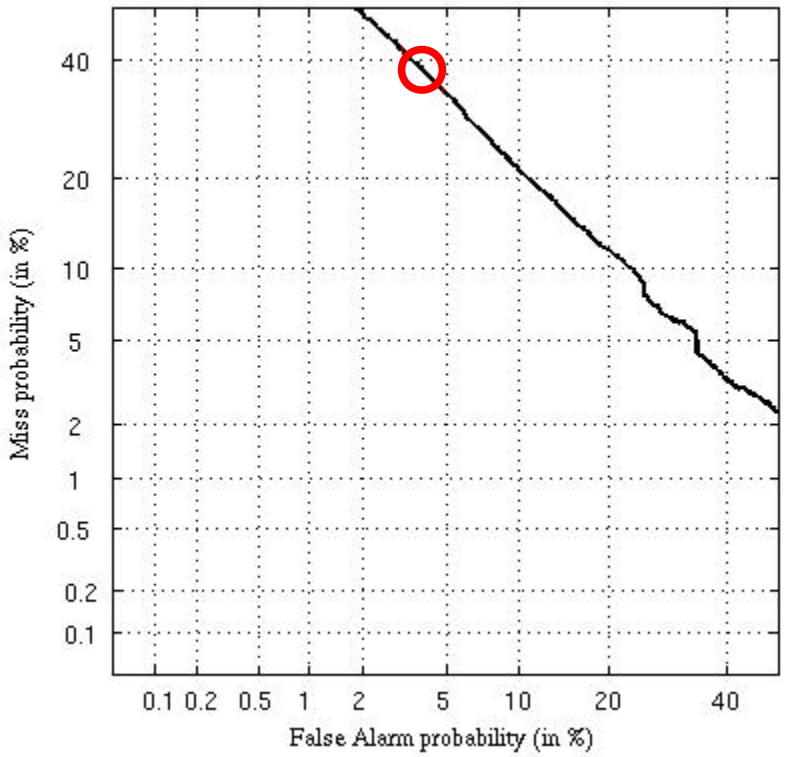
- distributed by CSLU
- Gender independent (50% male - 50% female)
- Spontaneous speech for 20 s (more or less)
- PSTN - 8kHz – mu-law

- Background: 94 speakers x 21 languages (inc. En, It)

- Development: 8 recordings x 131 speakers (all En)

Results 1/4

- Devel. performance is aligned to literature: EER is 15%



- Devel. performance at our CTS threshold is:
  FA 4.1% - FR 37.2%

Results 2/4

We have made a cross comparison between questioned recordings:

- 4 recs x 2 speakers – 8kHz – linear enc – Silent Room
- 56 non trivial comparisons

Performance at our CTS threshold is:
FA 12.5% - FR 0% (FA 4.1% - FR 37.2%)

- 56 not a really wide set...
- Mismatching recording conditions

Results 3/4

comparison between questioned recs and CTS recs:

- 16 recs – 8kHz – A-law – real forensic (GSM + wiretape)
- 128 non trivial comparisons

Performance at our CTS threshold is:

FA 0% - FR 100% (FA 4.1% - FR 37.2% / FA 12.5% - FR 0%)

- 128 not a really wide set...
- MRC are not the cause... not the major one IMHO

Results 4/4

Looking at the OST:

- Same recording conditions of CST (according to guidelines)
- Average SNR <=3dB: we refused any comparison

this is what we do in real cases

Is noise the answer to our disaster in CST?! Maybe...

- 128 not a really wide set... again ;-)

...but SNR-driven compensation of LR seems a must!

Conclusions

- We have completely failed the EVALITA test
- Noise is the most probable cause of our disaster – being MRC a second cause
- SNR-driven compensation of LR at frame level is our next priority
  =>It is not the-next-big-thing (ATVS published it in 2006)
- We do not want to improve the performance: we want to automatically weaken support to any hypothesis according to SNR.

Thank you very much !!!