The INRIA-IRIT system for NIST'SRE-2010

Reda Jourani [1,3], Khalid Daoudi [2], Jérôme Farinas [1] and Régine André-Obrecht [1]

[1] SAMoVA Group, IRIT - UMR 5505 du CNRS
[2] INRIA Bordeaux-Sud Ouest
[3] Laboratoire LRIT. Faculty of Sciences, Mohammed 5 Agdal University

System description

- First participation at NIST-SRE campaigns.
- The speaker detection system jointly developed by INRIA (Bordeaux, France) and IRIT (Toulouse, France).
- Mainly based on the open source software ALIZE/Spkdet.

Feature extraction

- The feature extraction is carried out by Spro.
- Bandwidth is limited to the 300-3400Hz range.
- 20 ms Hamming window length, 10 ms window shift.
- 24 filter bank coefficients ⟶ LFCCs:

19 statics + 19 Δ + 11 ΔΔ + ΔE.

- Cepstral Mean Subtraction and variance normalization.

- Zero mean and unit variance normalization of the energy coefficients.
- VAD based on tri-Gaussian model of log-energy.
- keeping only the most energized (informative) frames.

- Interview segments processing:

-The estimated intervals where the target speaker is speaking are determined based on the VAD.
-The VAD on the B channel determines the time intervals where the interviewer is speaking.
-The estimated intervals are removed from the A channel speech segments.

- Post-processing:

-To deal with the energized non-speech segments, the different distances of speakers from the microphones and possible time shifts.
-Tests done on the tarball of the NIST-SRE' 2010 development data and NIST-SRE' 2008 data.

↓

Cleaning speech segments of interviews A channels from the segments shorter than 20ms.
Purging telephone and microphone data from respectively, speech segments shorter than 40ms and 20ms.

- Normalization to fit a zero mean and unit variance distribution.

Universal Background Models

- 2 gender-dependent UBMs with 512 Gaussian components and diagonal covariance matrices.
- Trained by the LIA laboratory (Laboratoire d'Informatique d'Avignon, France) using telephone data from the Fisher English Training Speech Part 1, and microphone data from the NIST-SRE' 2005 data.

Session variability modeling

- Performing the Latent Factor Analysis modeling and retaining only the speaker dependent components.
- Estimation of 2 gender-dependent 40 rank session variability matrices on NIST-SRE'2004 and 2005 data.
- 194 male speakers and 134 female speakers. 27 different sessions per speaker in average.

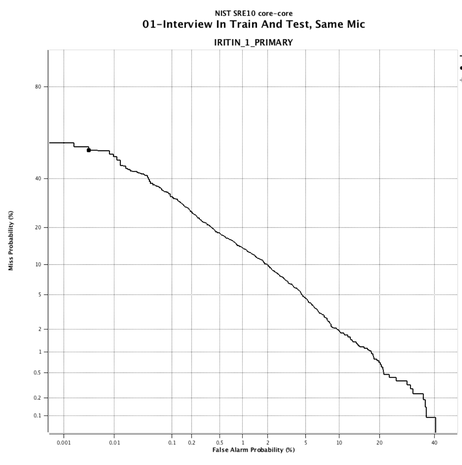
Score normalization

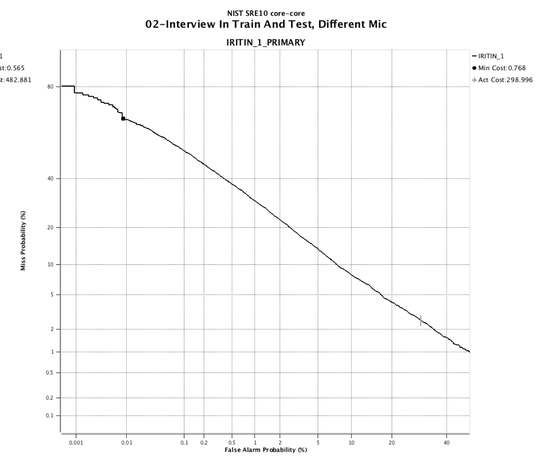
- Gender-dependent T-norm.
- 200 male speakers and 200 female speakers from NIST-SRE' 2006.
- 50% on telephone data, and 50% on microphone data.

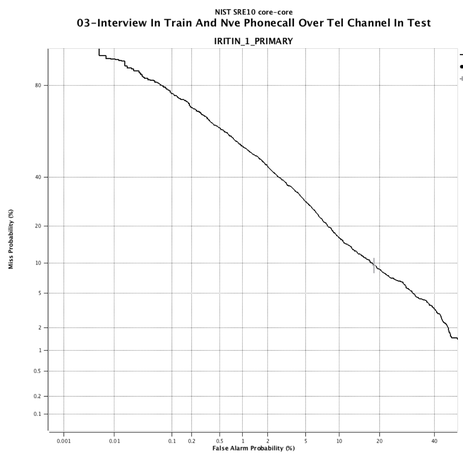
Decision

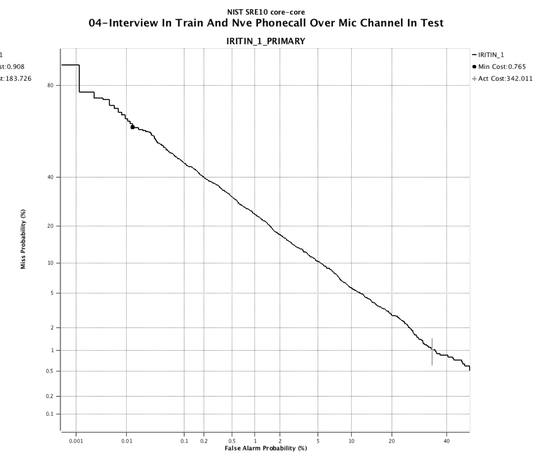
- Unique gender-independent threshold set on the EER point estimated on the male part of the NIST-SRE' 2008 primary task.

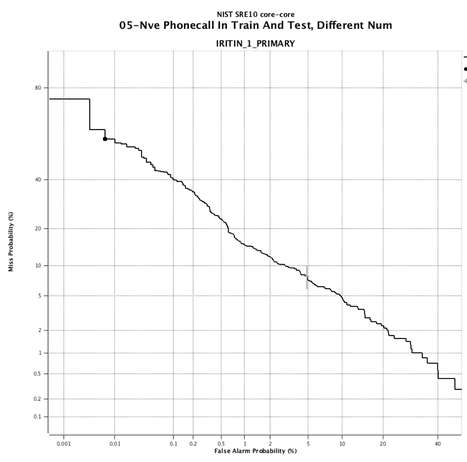
Results

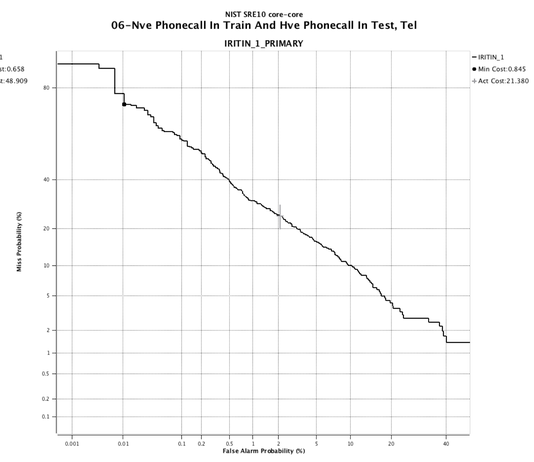


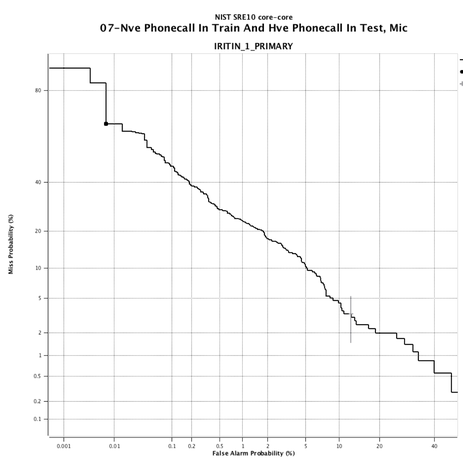


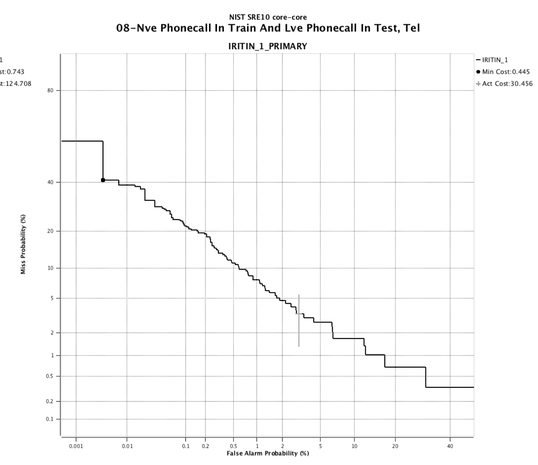


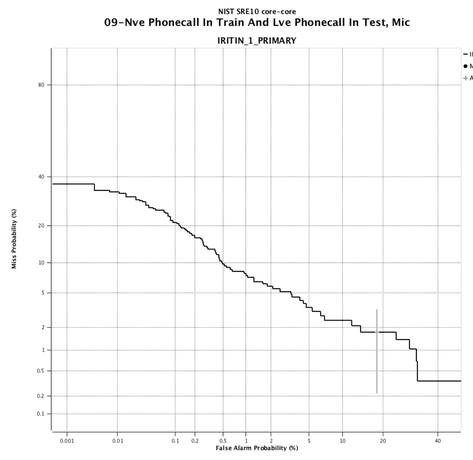












Speed and resources

- 8 x Intel XEON 64bits 3.16GHz, with 6MB of L2 cache per processor and 24GB of RAM.
- Client models training : 0.086xRT, 7.5 GB.
- Test segments processing : 0.208xRT, 3.28 GB.

Thanks to the LIA laboratory, particularly J-F. Bonastre, A. Larcher and D. Matrouf, for providing the UBMs and for their continuous and valuable help during the whole campaign period.