Albayzin 2010 Evaluation Campaign: Speaker Diarization

Martin Zelenák, Henrik Schulz and Javier Hernando

FALA 2010, November 10th, Vigo, Spain

Centre de Tecnologies i Aplicacions del Llenguatge i la Parla
Universitat Politècnica de Catalunya

Outline

Speaker diarization task

Database

Participants

Results

Conclusions

Speaker diarization task: Definition

- Task:

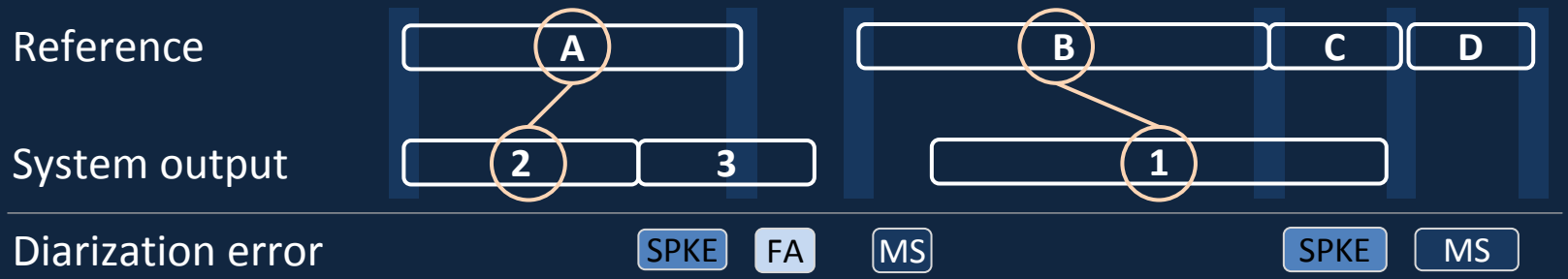
  “Who spoke when?”

- Domain:

  Broadcast news

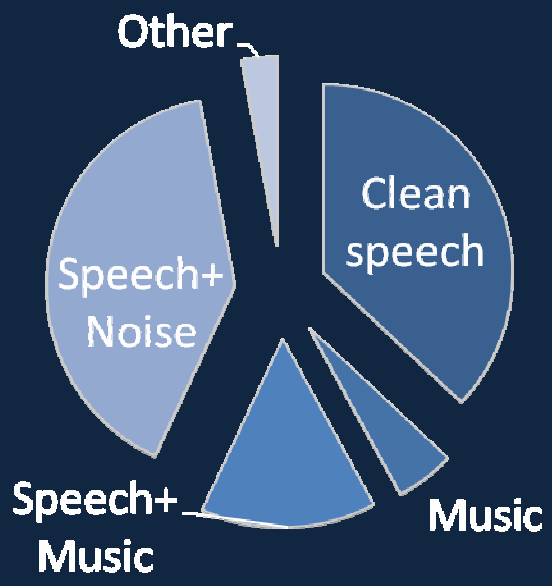
Speaker diarization task: Evaluation method

- 1-to-1 mapping: Ref. clust. IDs ↔ System clust. IDs
- Forgiveness collar ±0.25s around ref. seg. boundaries
- Scoring metric: Diarization Error Rate (DER)
  - (Speaker Diarization Error Time) / (Total Actual Time)



Database

- Evaluation data: Broadcast news from Catalan 3/24 TV channel
  - 8 sessions – 30 hours
  - 66–120 speakers/session
- Reference transcriptions:
  - Originally w/o silences
  - Simple low-energy silence detector + smoothing (min. duration 0.5s)



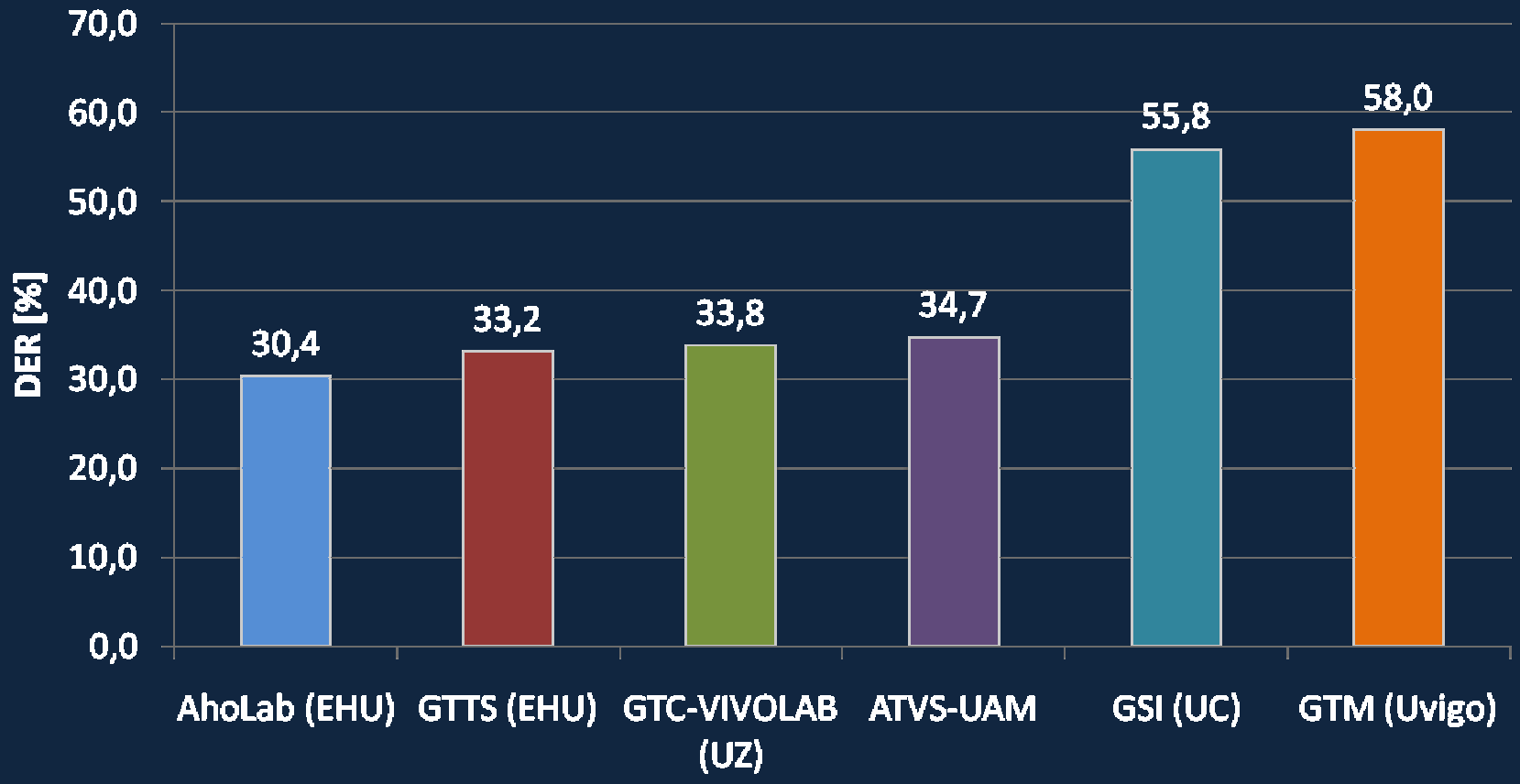
Participating teams

Team ID	Research Institute
AhoLab	University of the Basque Country (EHU)
GSI	University of Coimbra (UC)
GTM	University of Vigo (UVigo)
GTC-VIVOLAB	University of Zaragoza (UZ)
GTTS	University of the Basque Country (EHU)
ATVS-UAM	Autonomous University of Madrid (UAM)

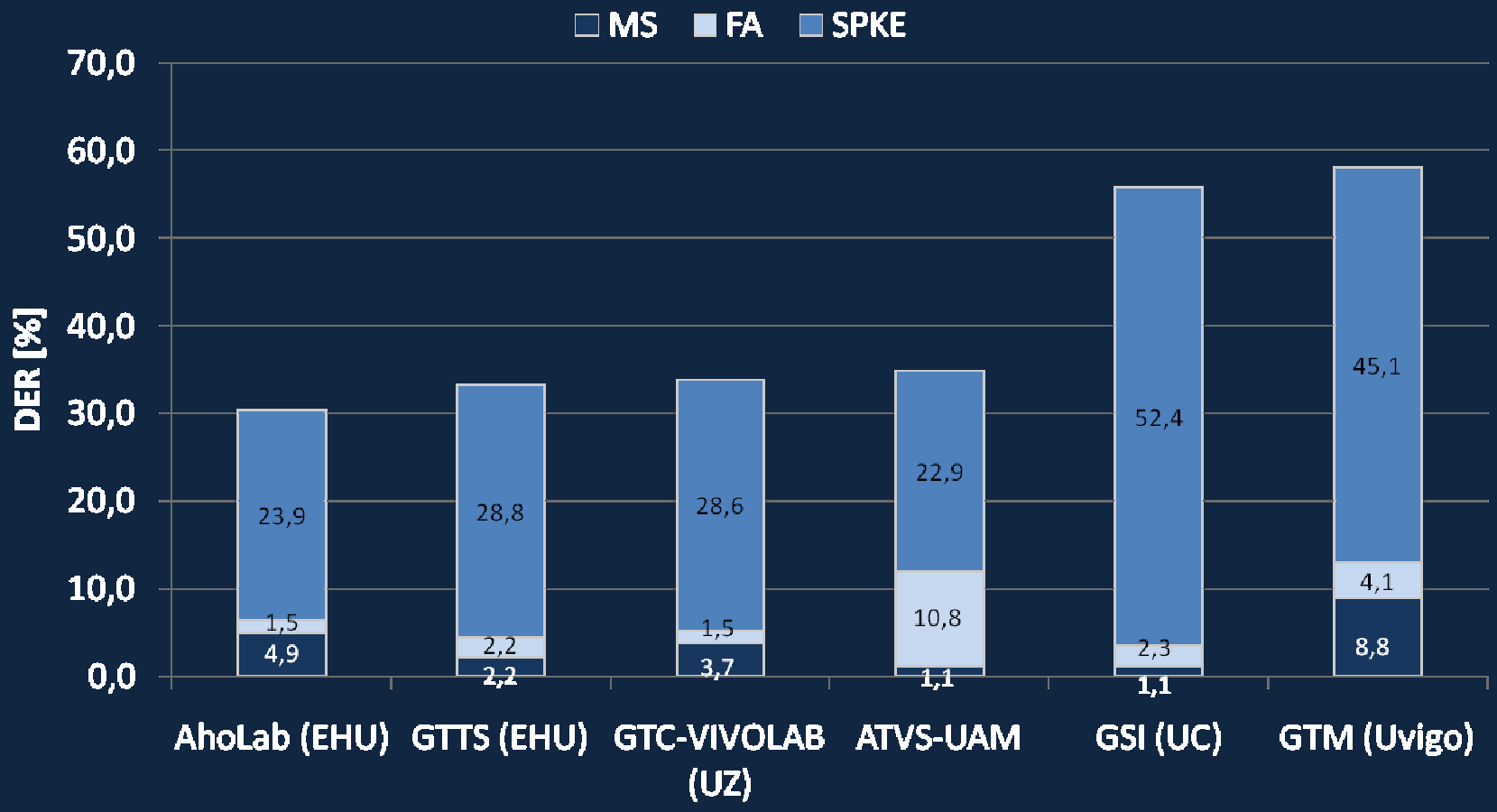
Participating systems

Team	System characteristics
AhoLab (EHU)	MFCCs, growing window speaker-change det. (voiced only, BIC), online clustering (BIC) => single pass
GSI (UC)	MFCCs + other, UBM-GMM decod. with high mix. transit. penalties: 1 mix. seq. <-> 1 cluster, audio fingerprint and BIC verif.
GTM (UVigo)	MFCCs + logE, DCTS, change-point refinement (BIC, Poisson), male/female, neighbor CLR-merging, AHC (score/all clust. models, cosine dist.)
GTC-VIVOLAB (UZ)	MFCCs, eigenvoice factor analysis (FA) speaker segmentation (10 clust./5 min.), BIC-based AHC
GTTS (EHU)	MFCCs (cl. + Δ + $\Delta\Delta$), XBIC-based speaker segm., AHC (MAP-adapt. GMM sufficient stats., dot-scoring similarity)
ATVS-UAM	MFCCs + Δ (CMN, RASTA, FWarp), 2-step clustering (FA iVectors/1.5 min., cosine dist.), Viterbi decoding (prior norm. speaker centroids)

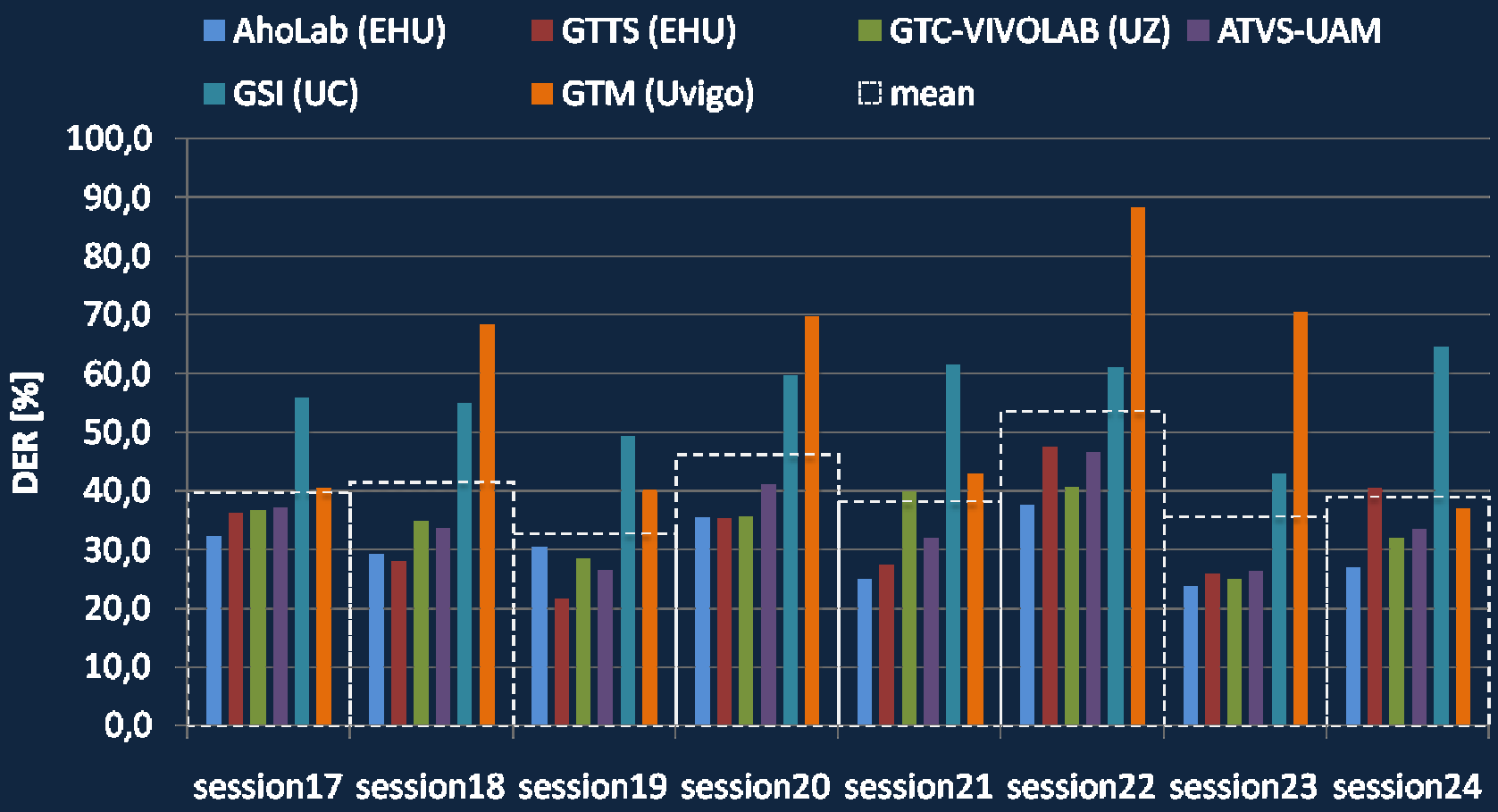
Results: Overall performance (1)



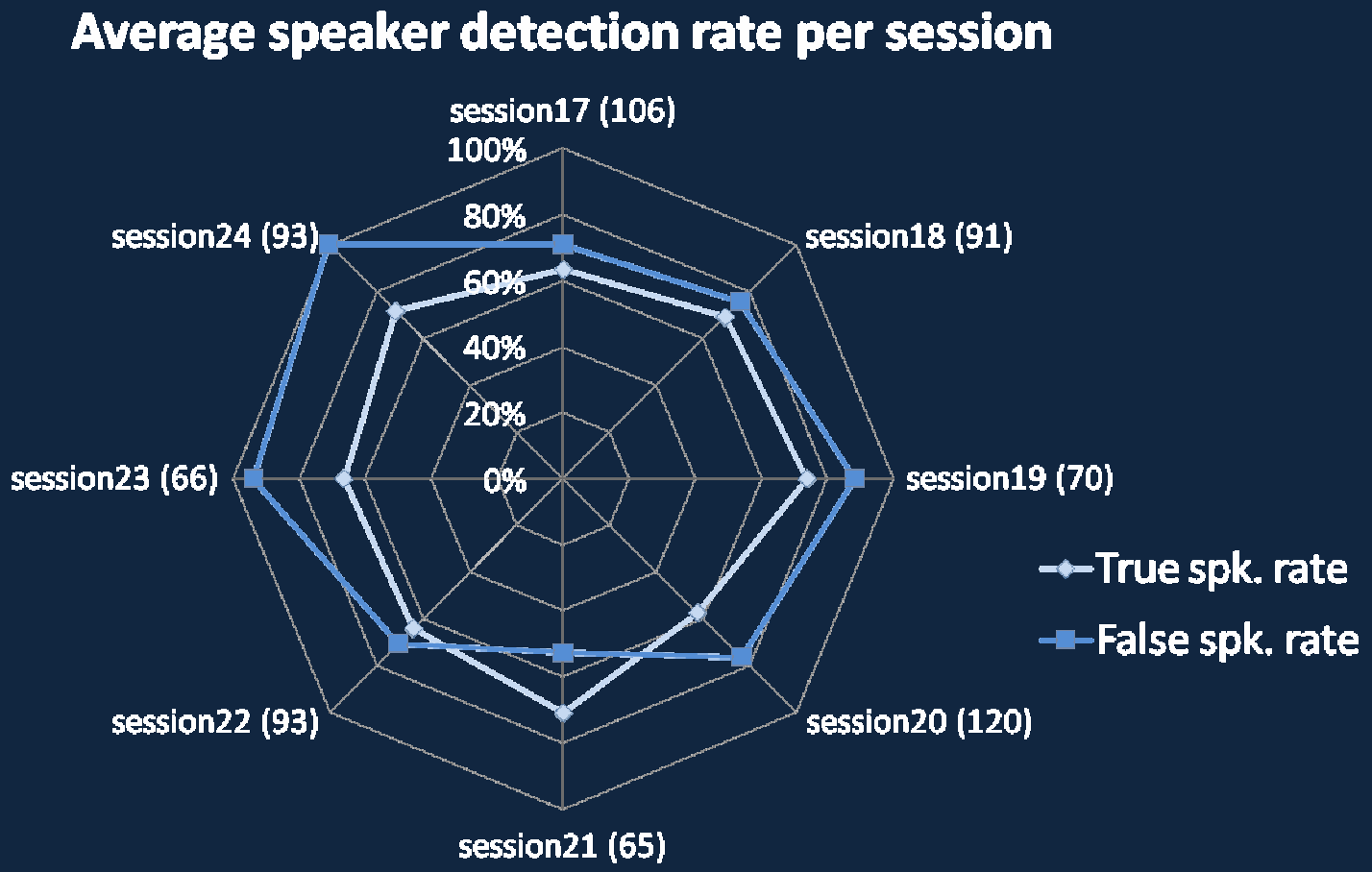
Results: Overall performance (2)



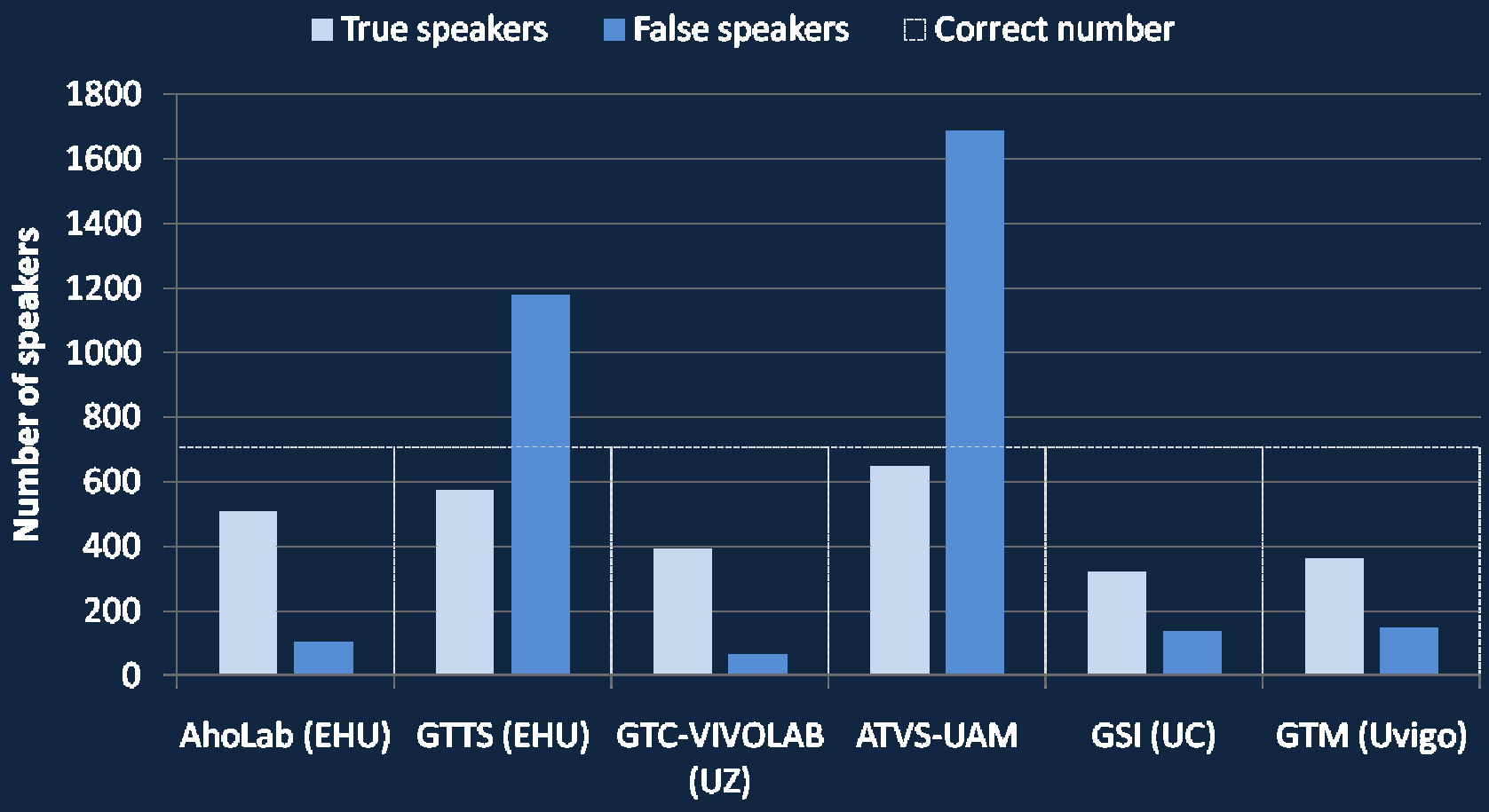
Results: Performance per session



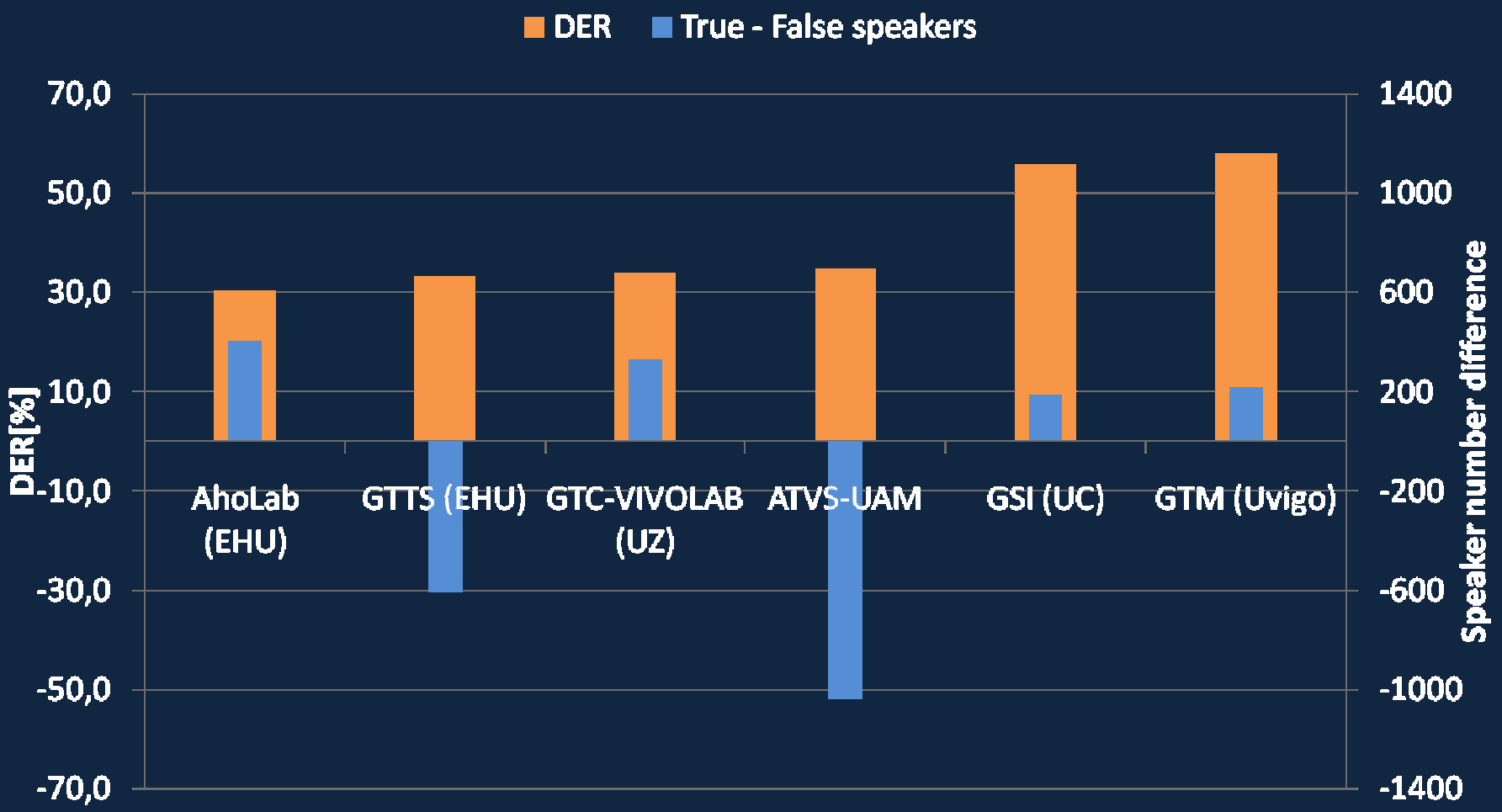
Results: Speaker number detection per session



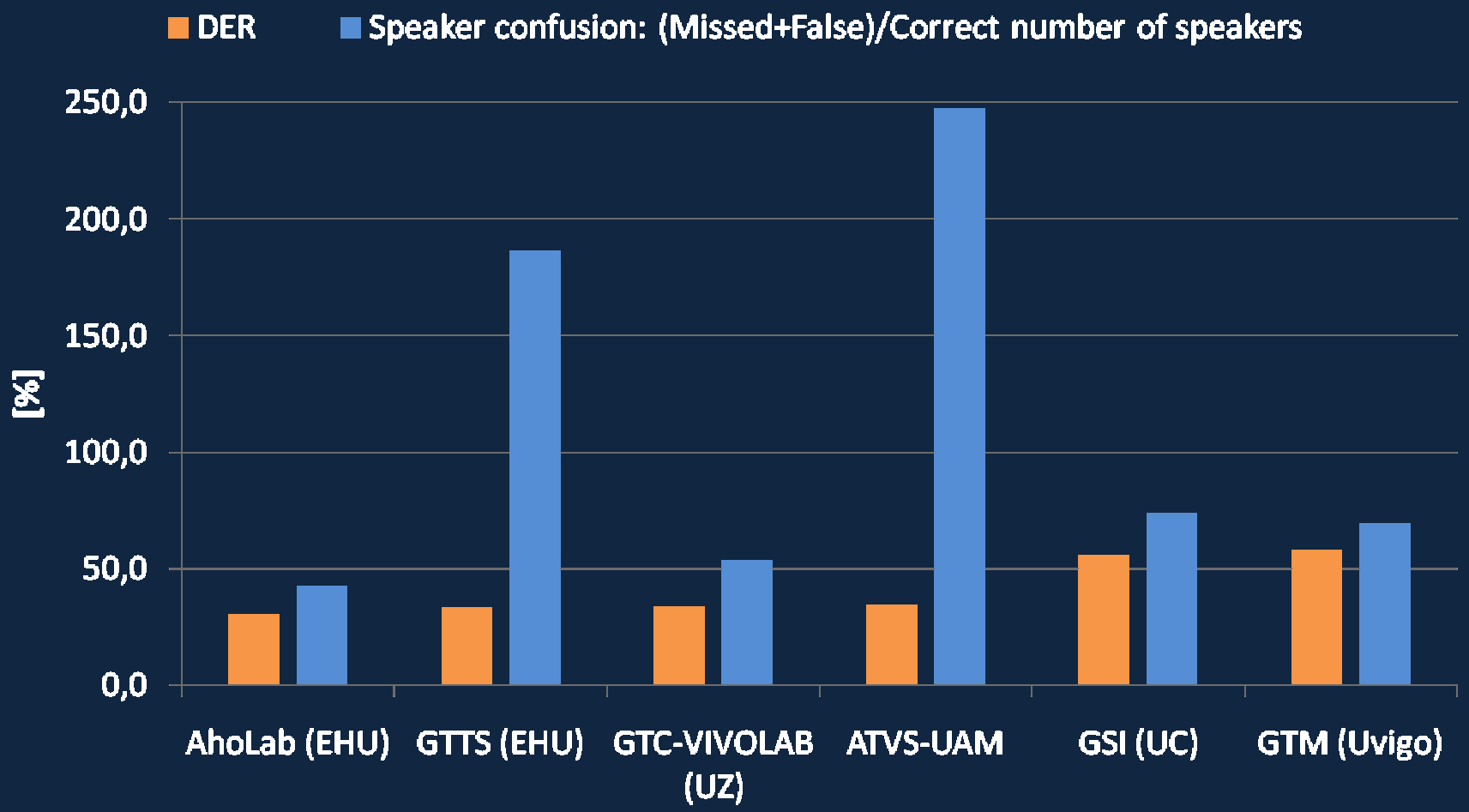
Results: Speaker number detection



Results: Summary



Results: Summary



Conclusions

- Challenging conditions (in general, worse results than NIST RT or ESTER evaluations)
- Best result (evaluation winner): AhoLab team
- Comparable performance: GTTS, GTC-VIVOLAB and ATVS-UAM
- Very different performance concerning the number of detected speakers

Thank you for your attention