SECURITY 2015

23. ročník konference o bezpečnosti v ICT

Ověřování osob pomocí hlasu

Honza Černocký

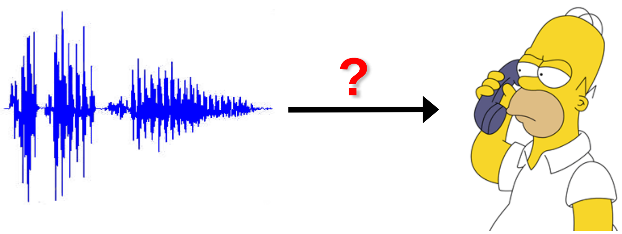
Vysoké učení technické v Brně,
BUT Speech@FIT

Agenda

- Basis of speaker verification
- Evaluation of performance
- Speaker verification system architecture
- Enemy No. 1: Intersession variability
- State of the art
- Data needed to adapt the system
- User scenarios
- Enemy No. 2: Attacks
- Vendors
- Conclusion

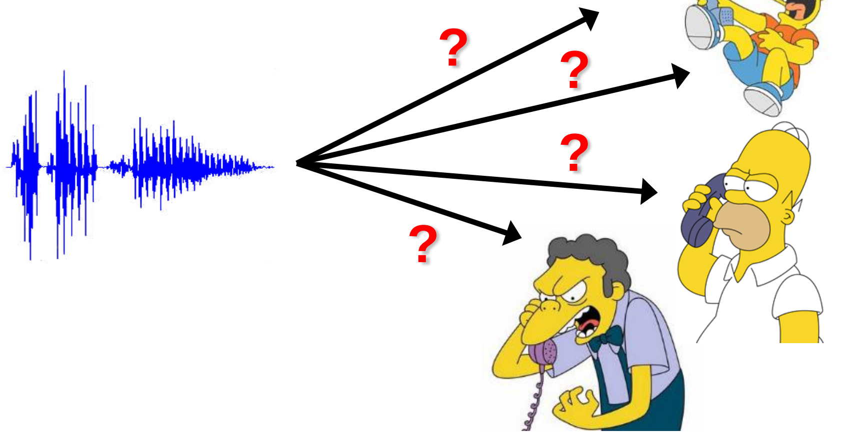
Verification

Is this Homer’s voice?



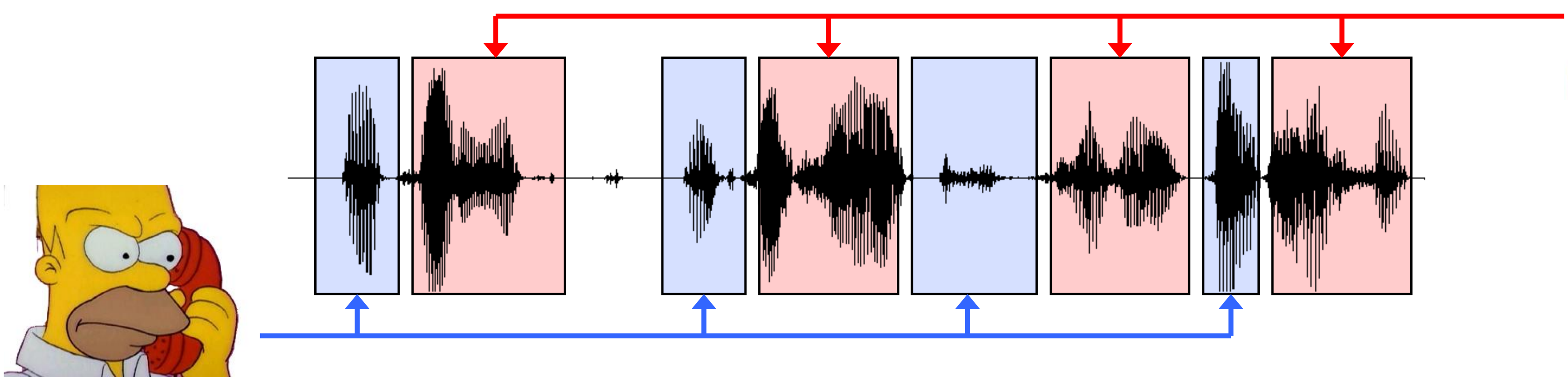
Identification

Whose voice is this?



Diarization (Segmentation and Clustering)

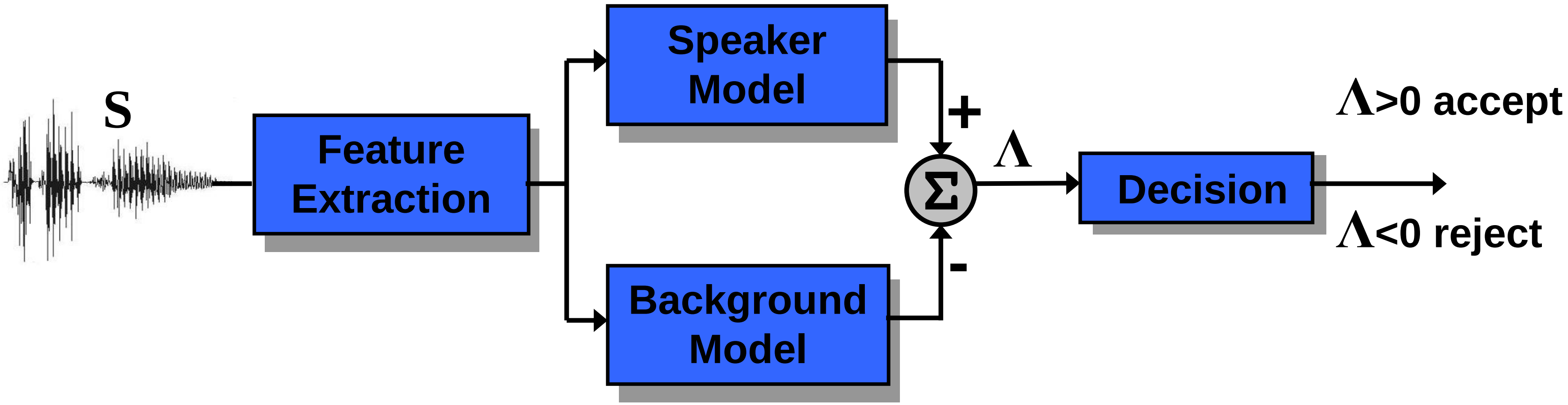
Where are speaker changes? Which segments are from the same speaker?



Speaker detection decision approaches have roots in signal detection theory

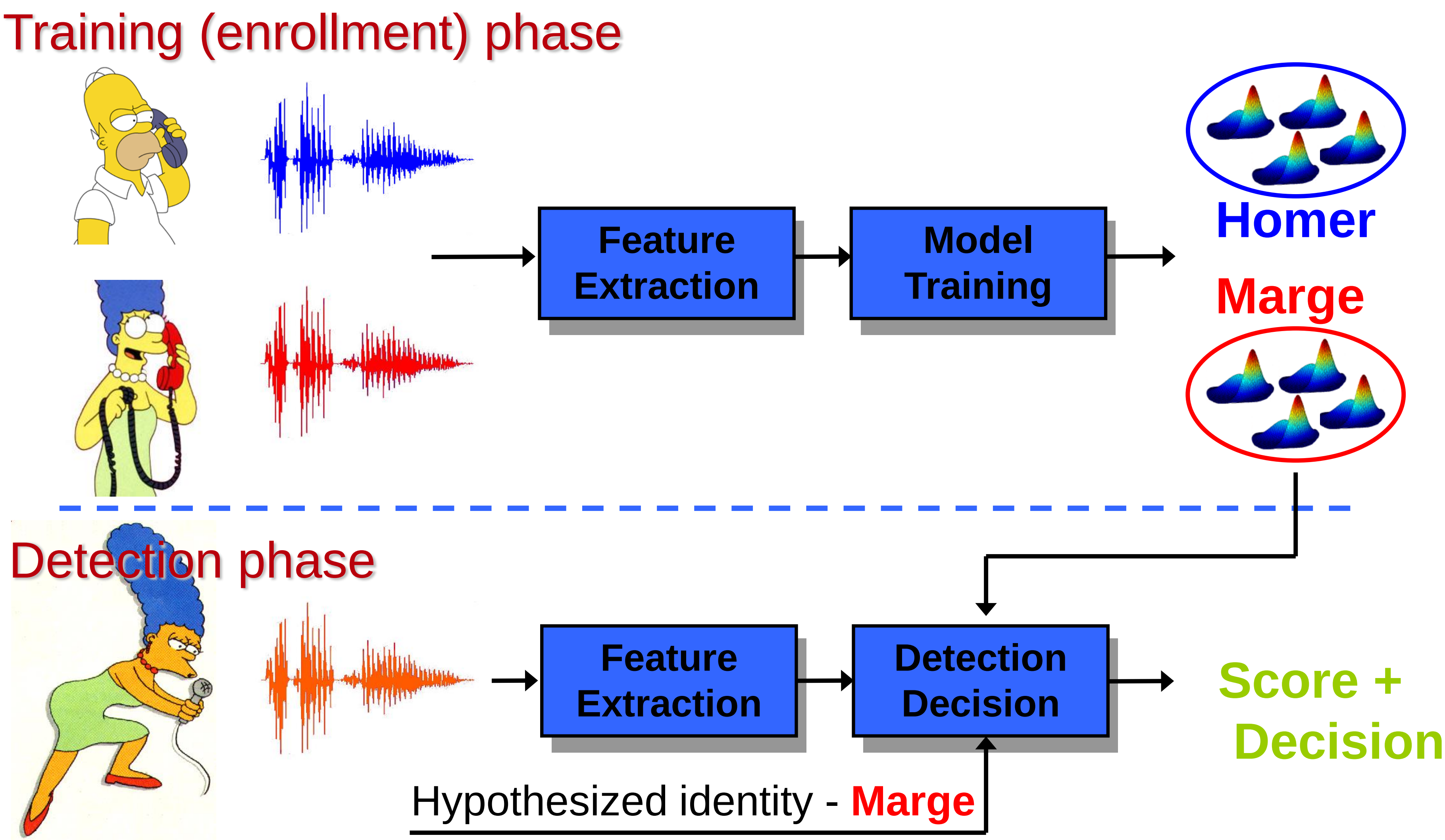
- **2 class Hypothesis test**
  - H0: the speaker is **not** the target speaker
  - H1: the speaker is the target speaker
- **Statistic computed on test utterance S as likelihood ratio**

$$\Lambda = \log \frac{\text{Likelihood } \mathbf{S} \text{ came from speaker model}}{\text{Likelihood } \mathbf{S} \text{ did } \underline{\mathbf{not}} \text{ come from speaker model}}$$



Phases of Speaker Detection System

Two distinct phases to any speaker detection system

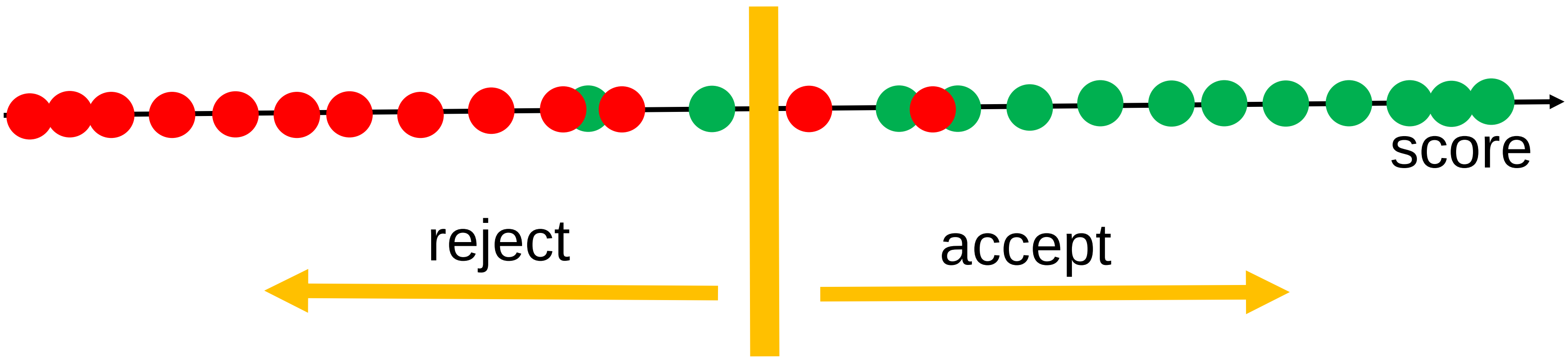


Agenda

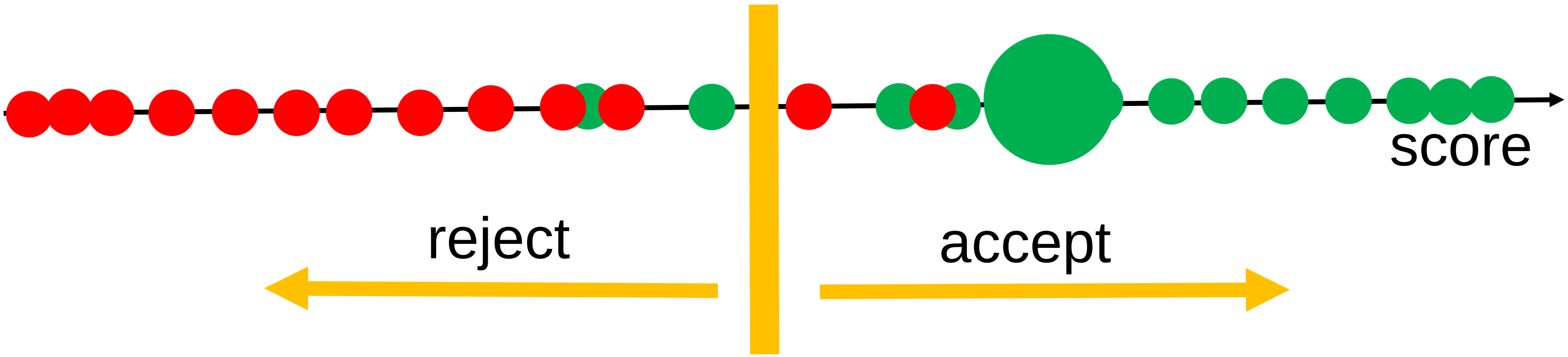
- Basis of speaker verification
- Evaluation of performance
- Speaker verification system architecture
- Enemy No. 1: Intersession variability
- State of the art
- Data needed to adapt the system
- User scenarios
- Enemy No. 2: Attacks
- Vendors
- Conclusion

Does the system work well ?

- We need some (lots of) data – pairs model speaker – test speaker.
  - **Target-trials** (test speaker = model speaker)
  - **Non-target-trials** (test speaker ≠ model speaker)
- We run them thru the system and record the scores
- We need to set the detection **threshold**

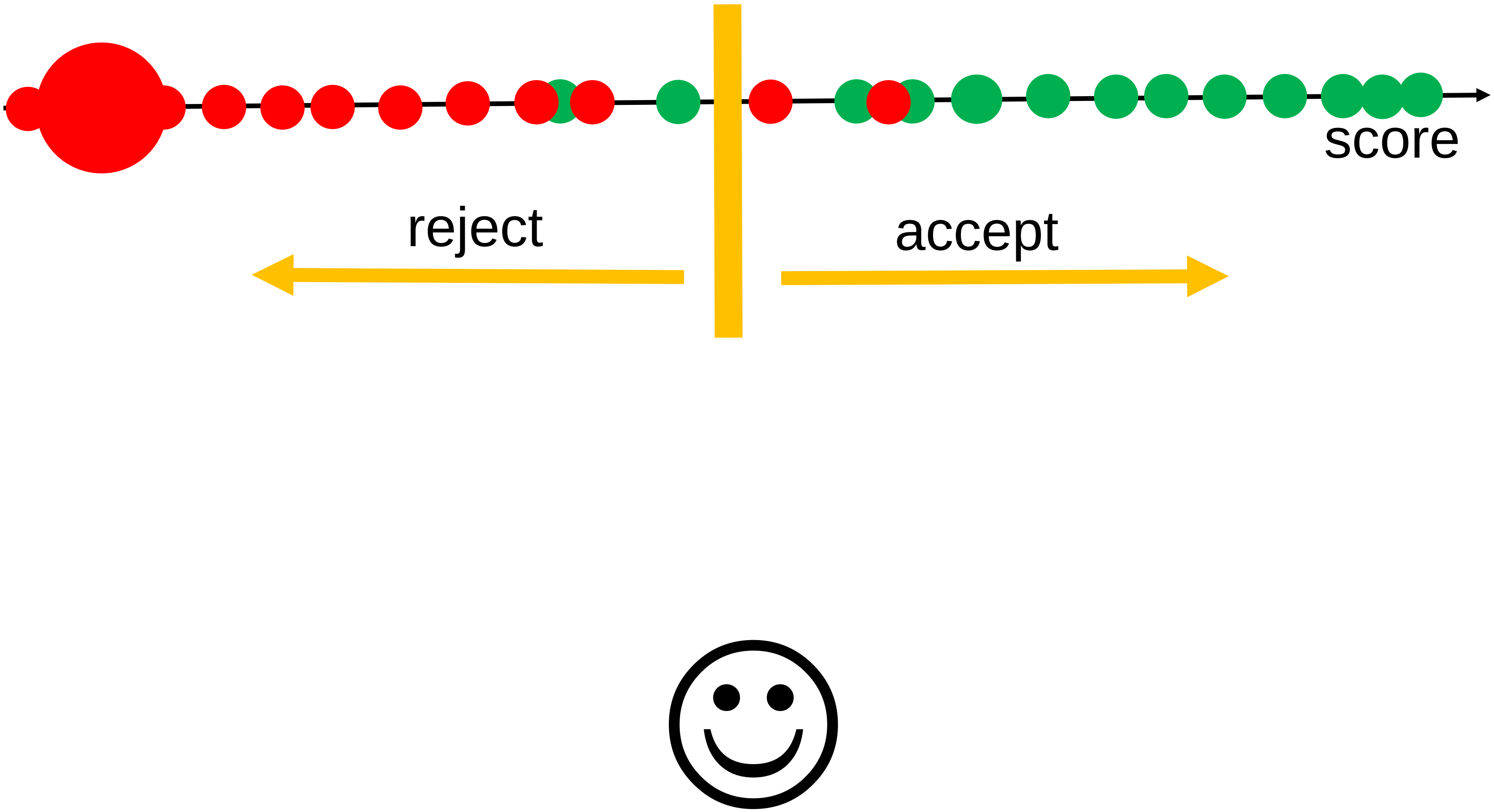


True accept

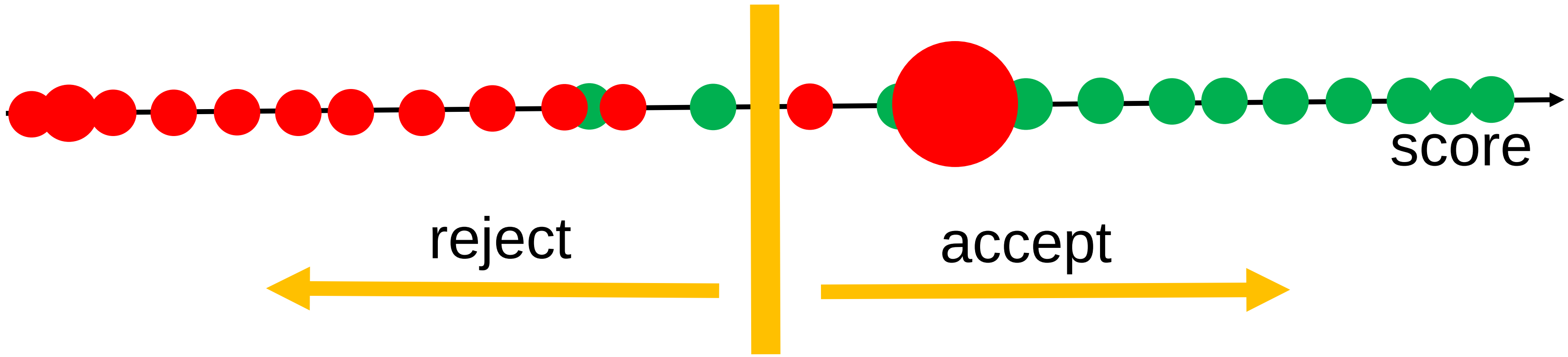


☺

True reject

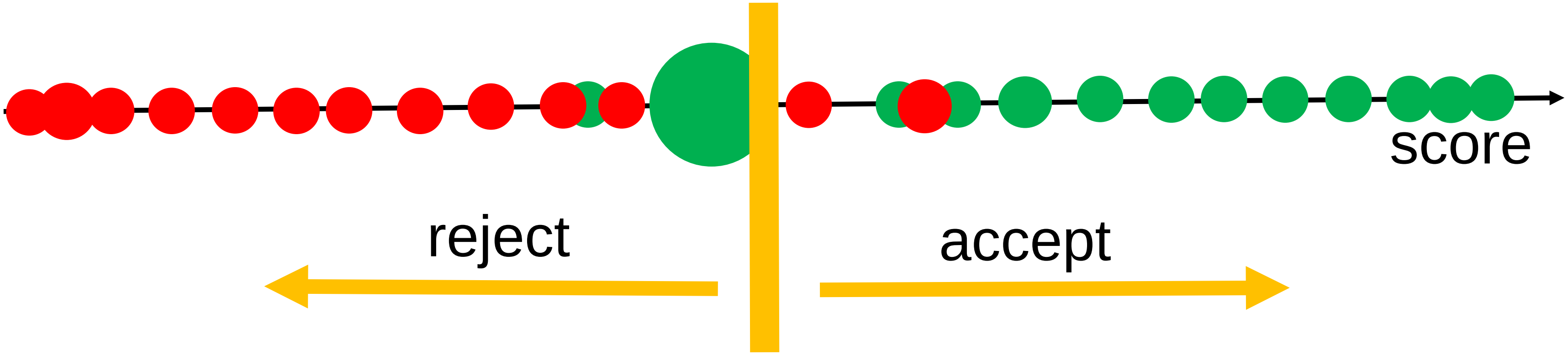


False accept



☹

False reject

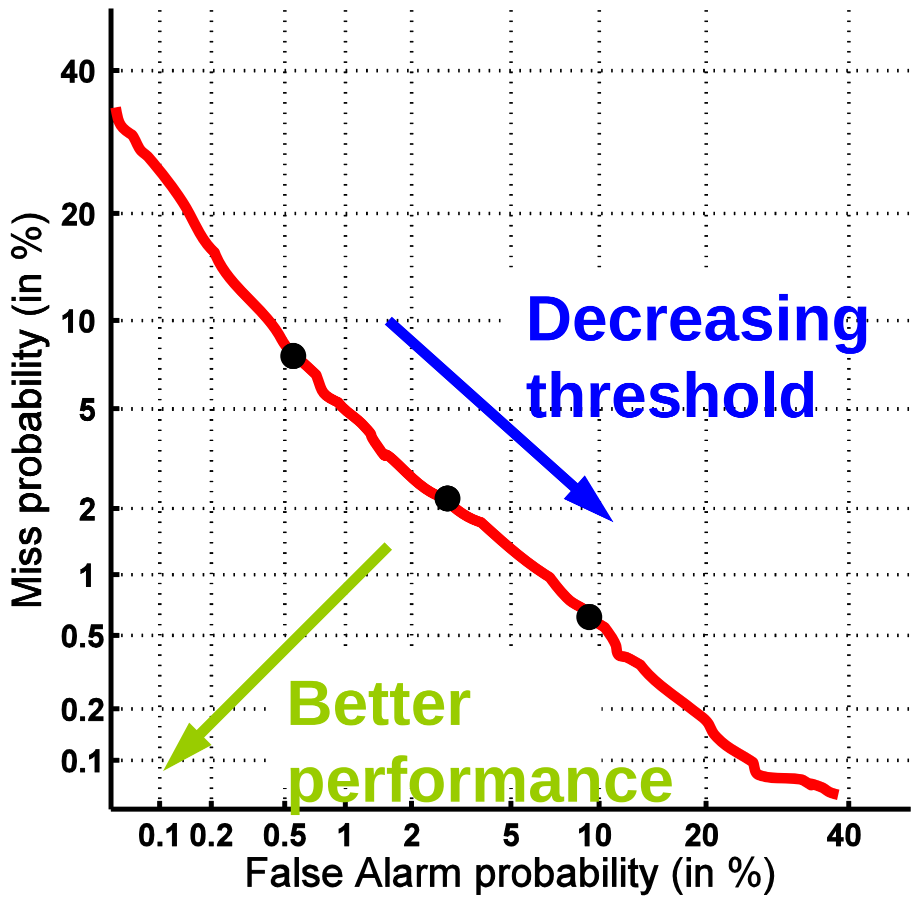


DET – Detection Error Tradeoff I.

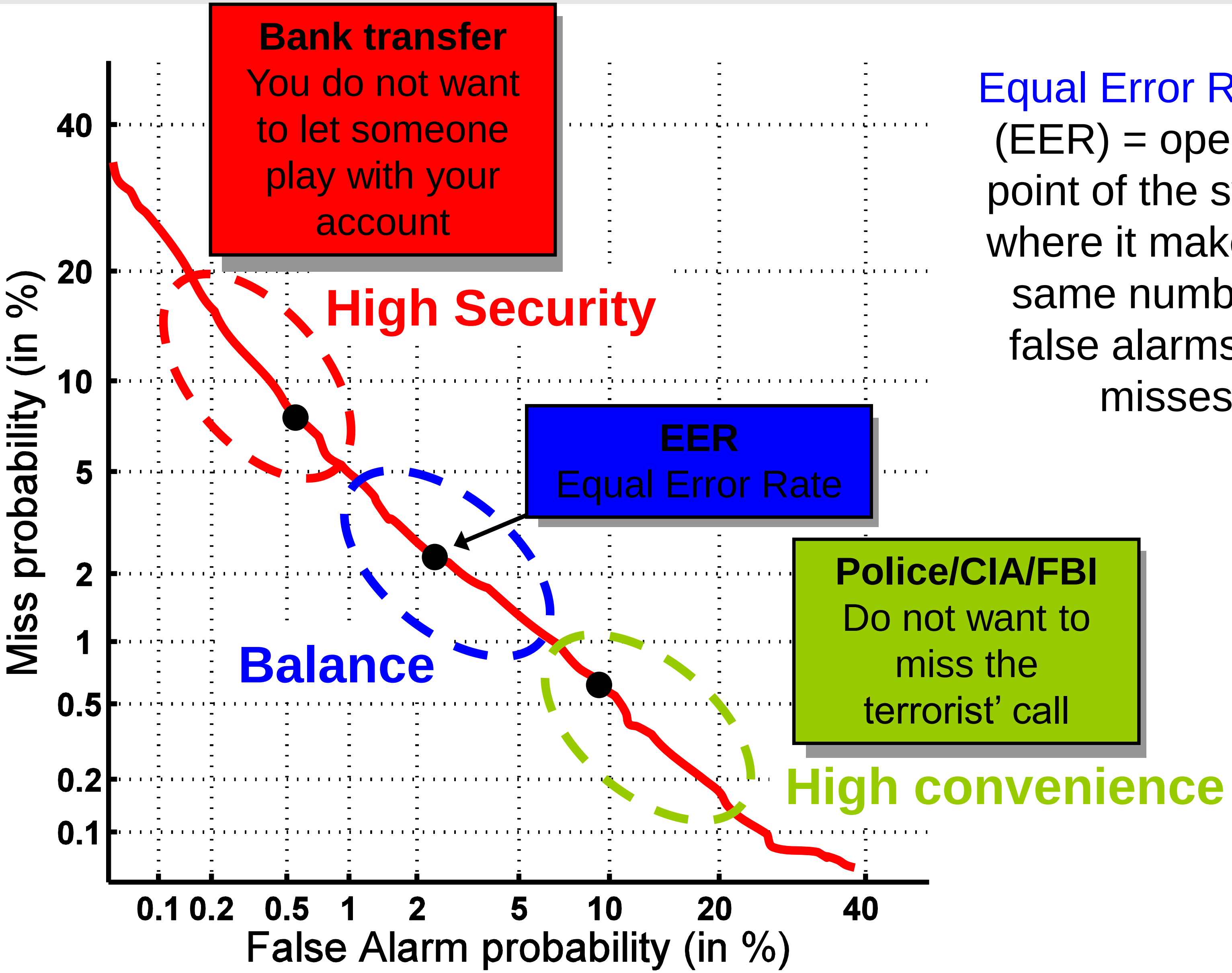
The performance of a detection system is measure of the trade-off between these two errors – is controlled by adjustment of the decision threshold

$$\Pr(\text{miss}|\text{thr}) = \frac{\#\ \text{target trials scores} < \text{thr}}{N_{\text{target}}}$$

$$\Pr(\text{false alarm}|\text{thr}) = \frac{\#\ \text{non-target trials scores} > \text{thr}}{N_{\text{non-target}}}$$



DET – Detection Error Tradeoff II.

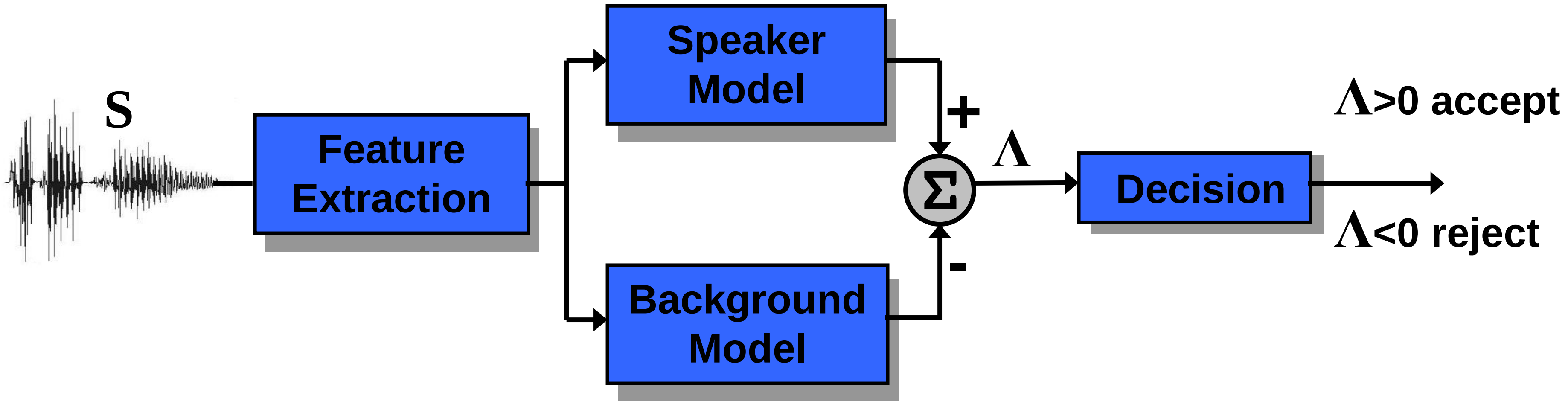


Equal Error Rate (EER) = operating point of the system where it makes the same number of false alarms and misses

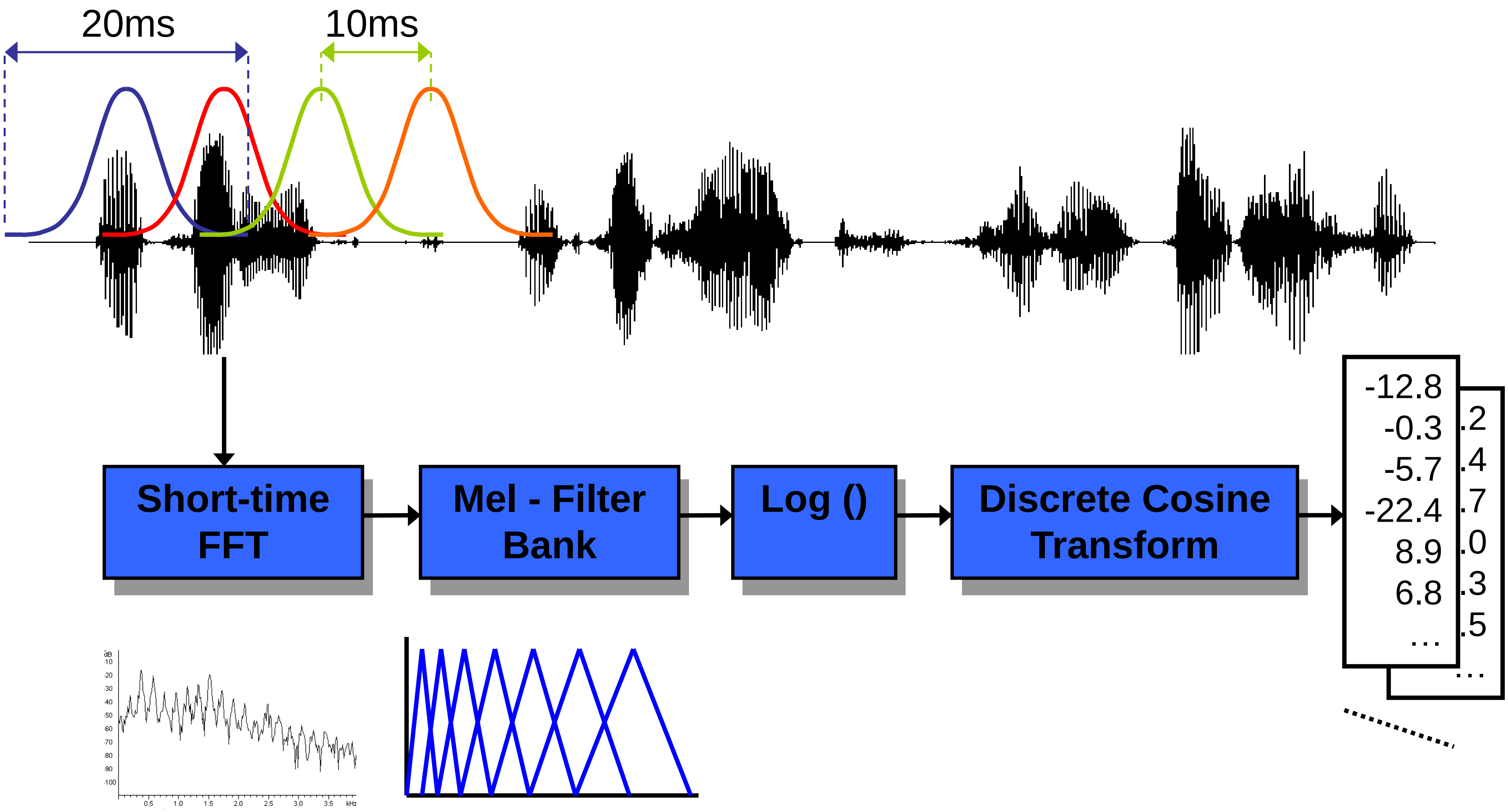
Agenda

- Basis of speaker verification
- Evaluation of performance
- Speaker verification system architecture
- Enemy No. 1: Intersession variability
- State of the art
- Data needed to adapt the system
- User scenarios
- Enemy No. 2: Attacks
- Vendors
- Conclusion

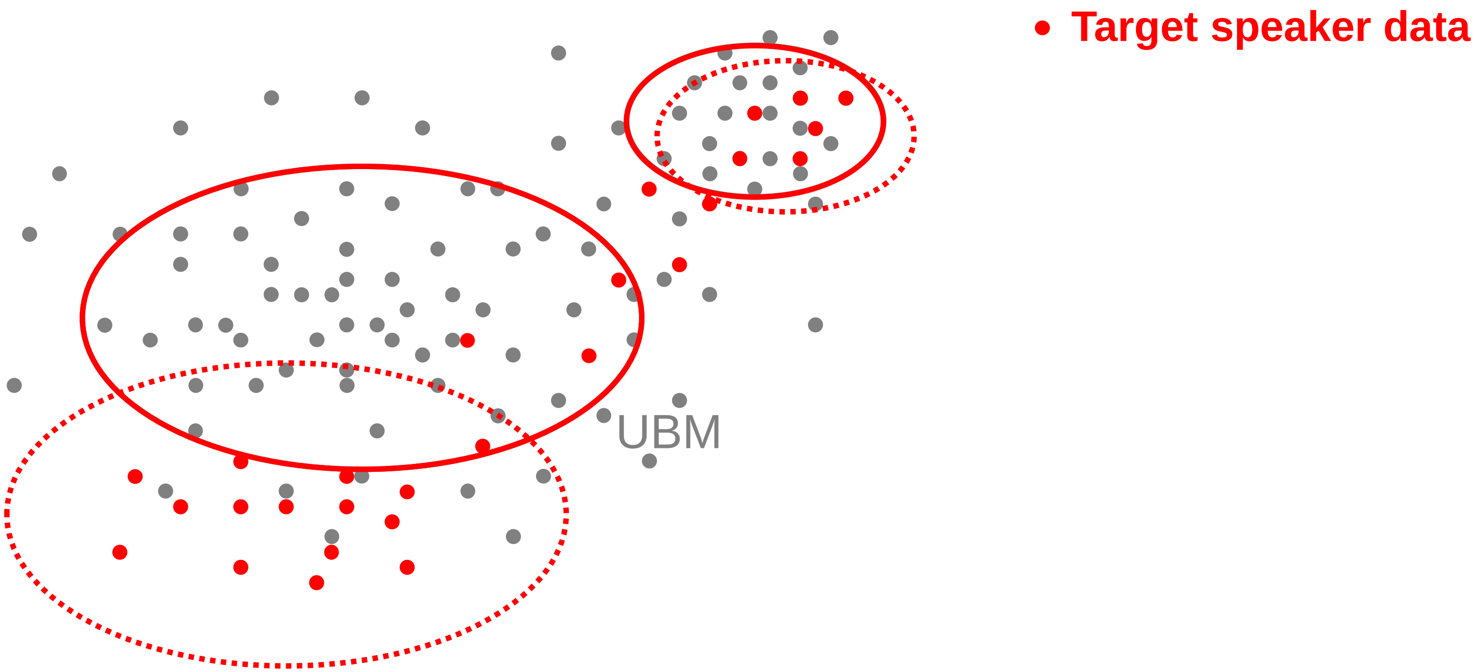
$$\Lambda = \log \frac{\text{Likelihood } \mathbf{S} \text{ came from speaker model}}{\text{Likelihood } \mathbf{S} \text{ did } \underline{\mathbf{not}} \text{ come from speaker model}}$$



Spectral features - MFCC



MAP adaptation – How to create speaker model



- **UBM model** - 2 Gaussians
- **Speaker model adapted from UBM**

Agenda

- Basis of speaker verification
- Evaluation of performance
- Speaker verification system architecture
- Enemy No. 1: Intersession variability
- State of the art
- Data needed to adapt the system
- User scenarios
- Enemy No. 2: Attacks
- Vendors
- Conclusion

Channel/session effects

The largest challenge to practical use of speaker detection systems is channel/session variability

- **Variability** – refers to changes in channel between enrolment and successive detection attempts
- **Channel/session** effects encompasses several factors
  - The microphones

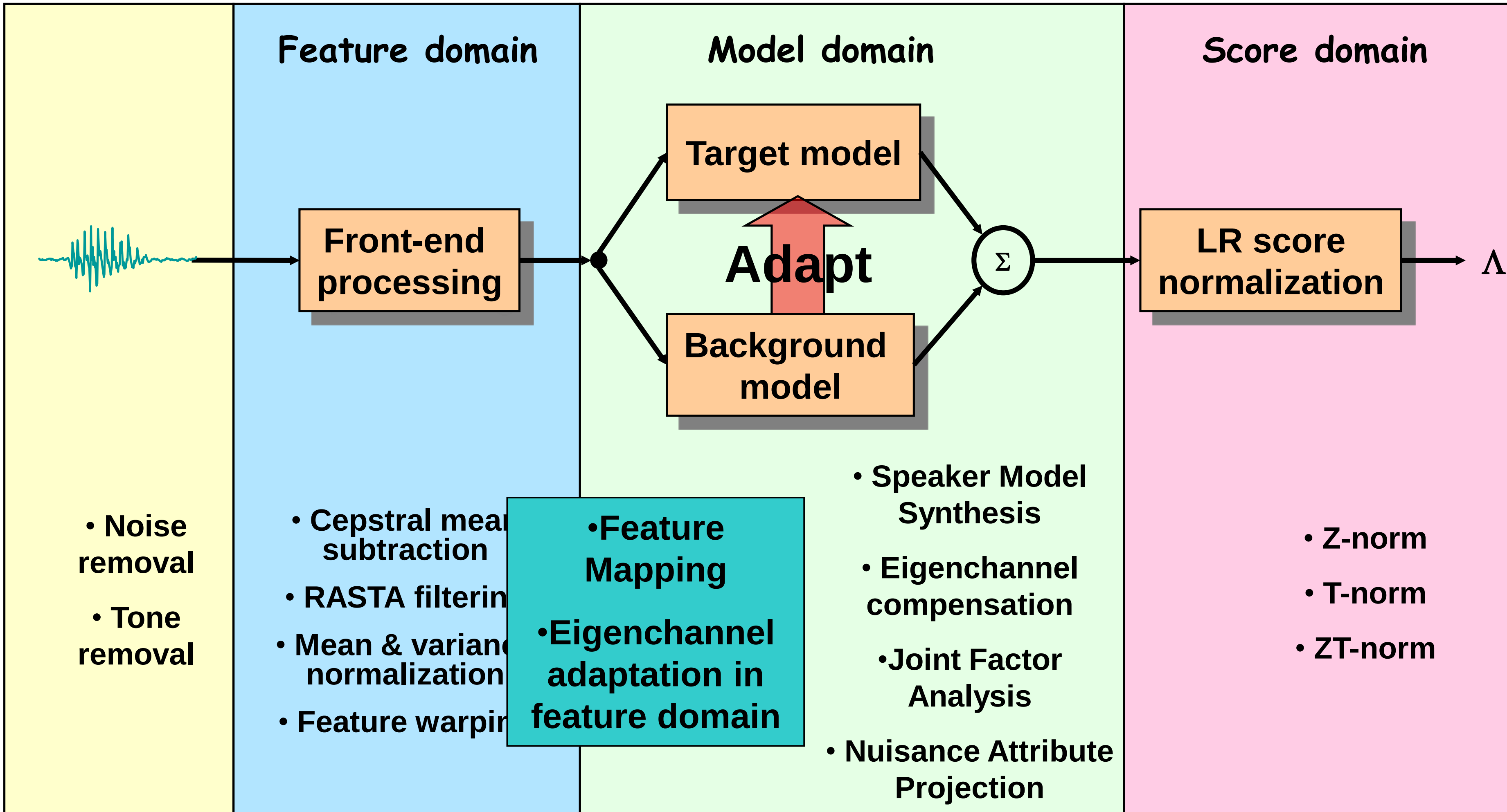
    Carbon-button, electret, hands-free, array, …
  - The acoustic environment

    Office, car, airport, street, restaurant, …
  - The transmission channel

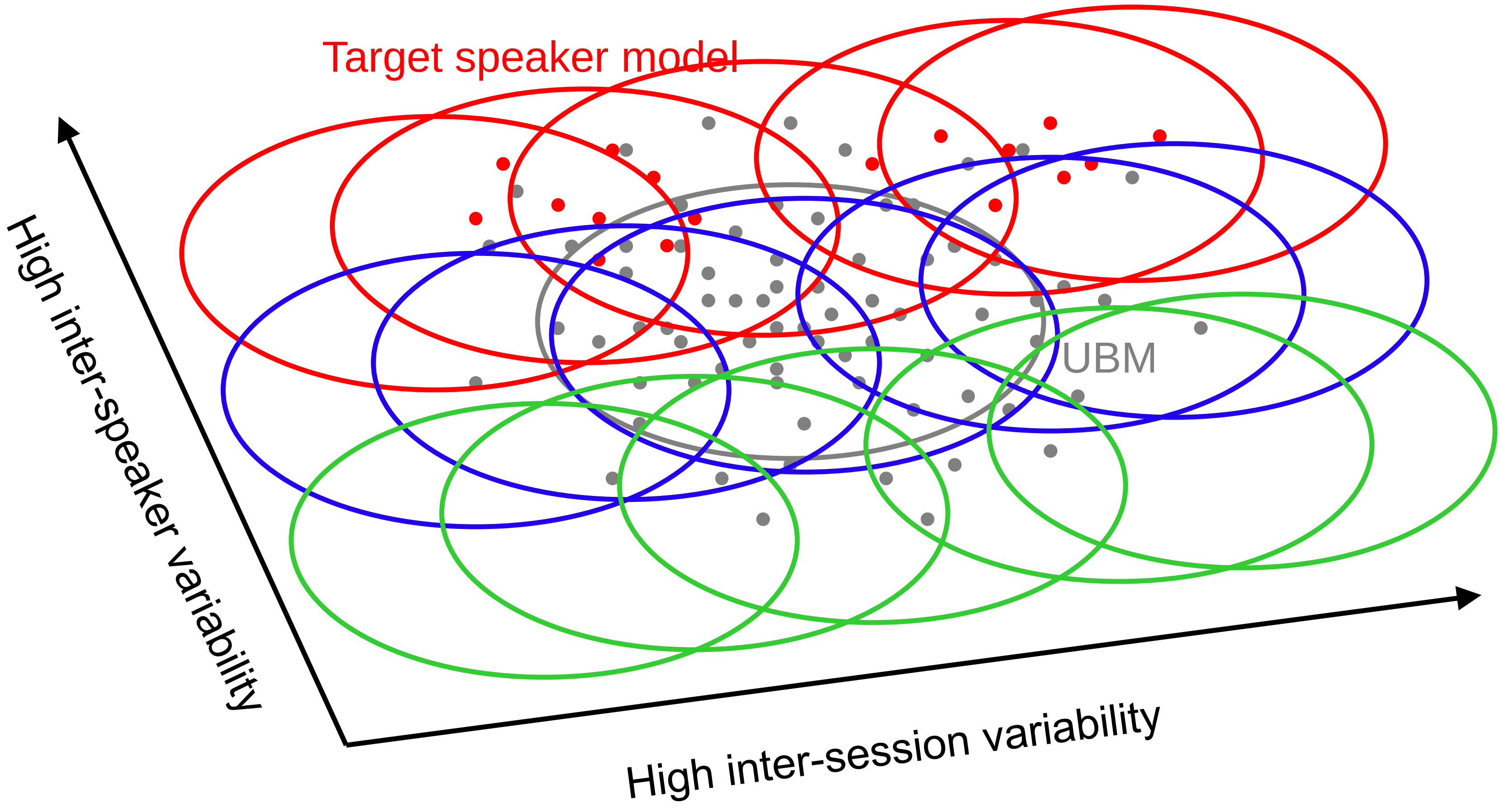
    Landline, cellular, VoIP,…
  - The speaker him/herself

    emotion state, language, content, politeness, stress, alcohol

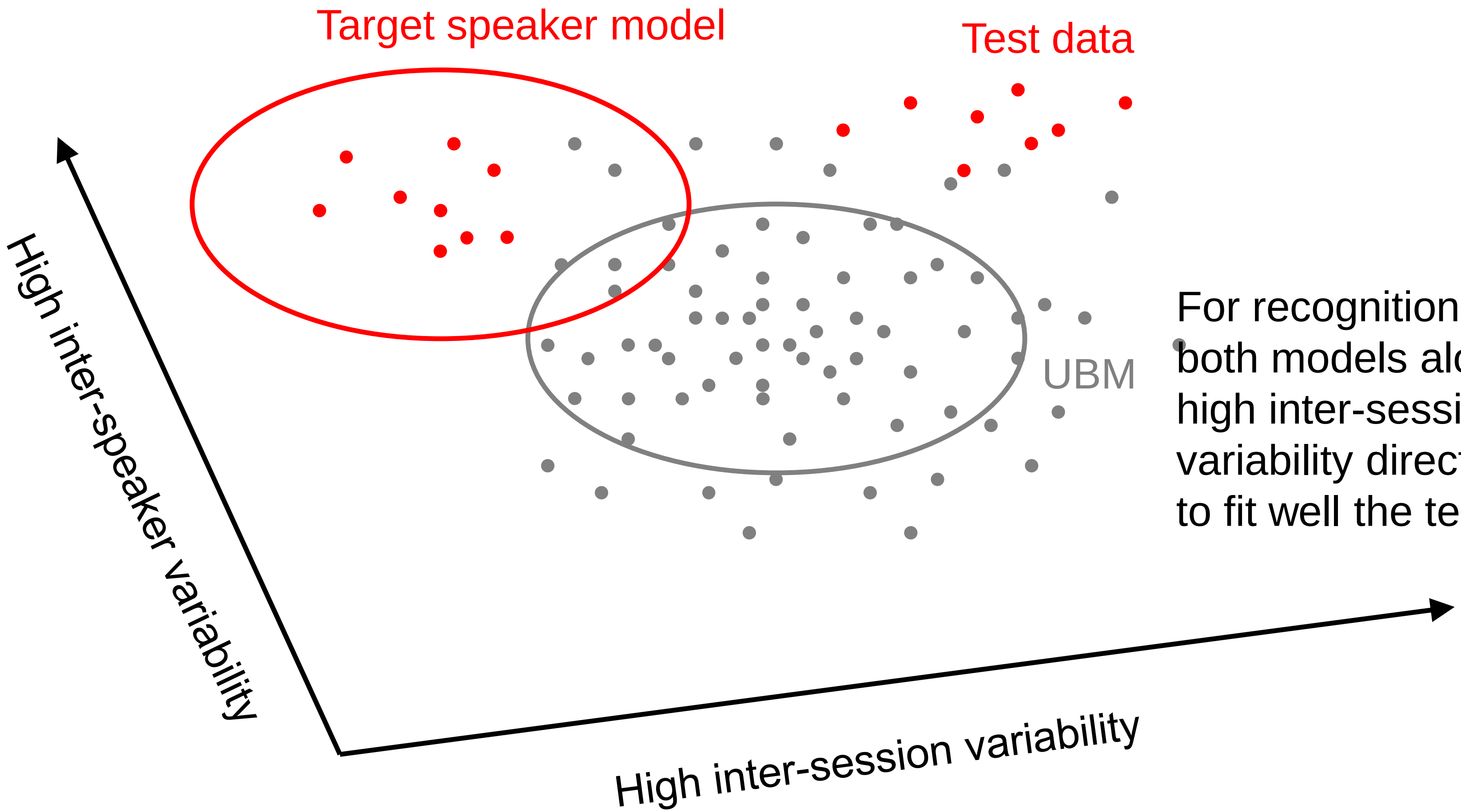
Years of SRE R&D fighting the variability …



Inter-session variability



Inter-session variability compensation

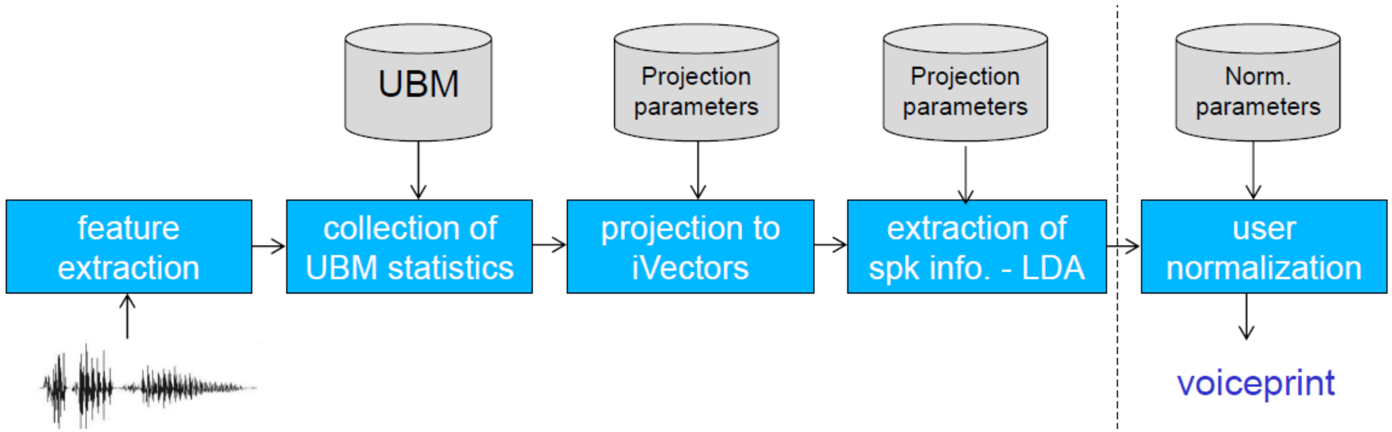


For recognition, move both models along the high inter-session variability direction(s) to fit well the test data

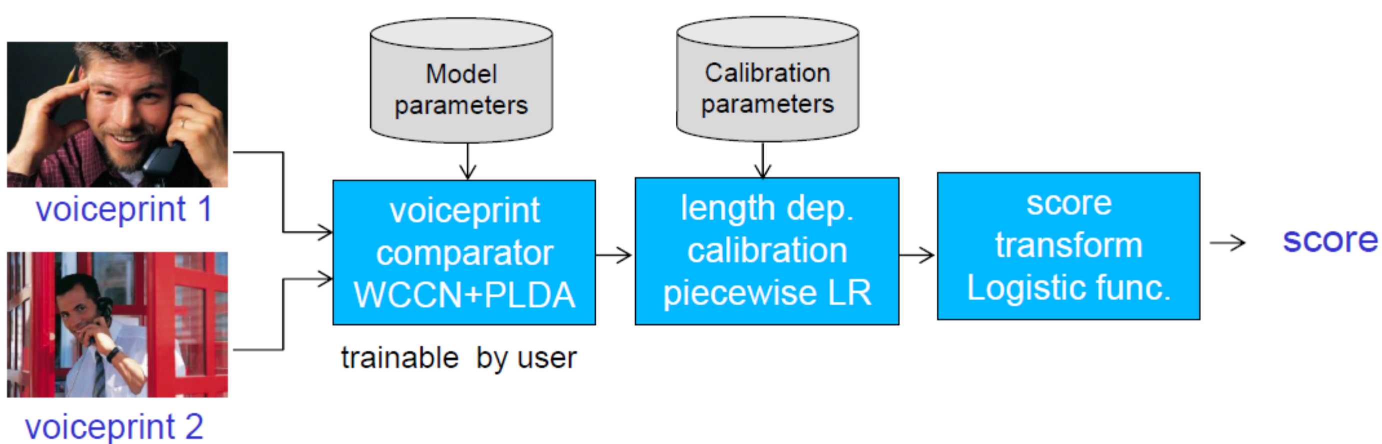
Agenda

- Basis of speaker verification
- Evaluation of performance
- Speaker verification system architecture
- Enemy No. 1: Intersession variability
- State of the art
- Data needed to adapt the system
- User scenarios
- Enemy No. 2: Attacks
- Vendors
- Conclusion

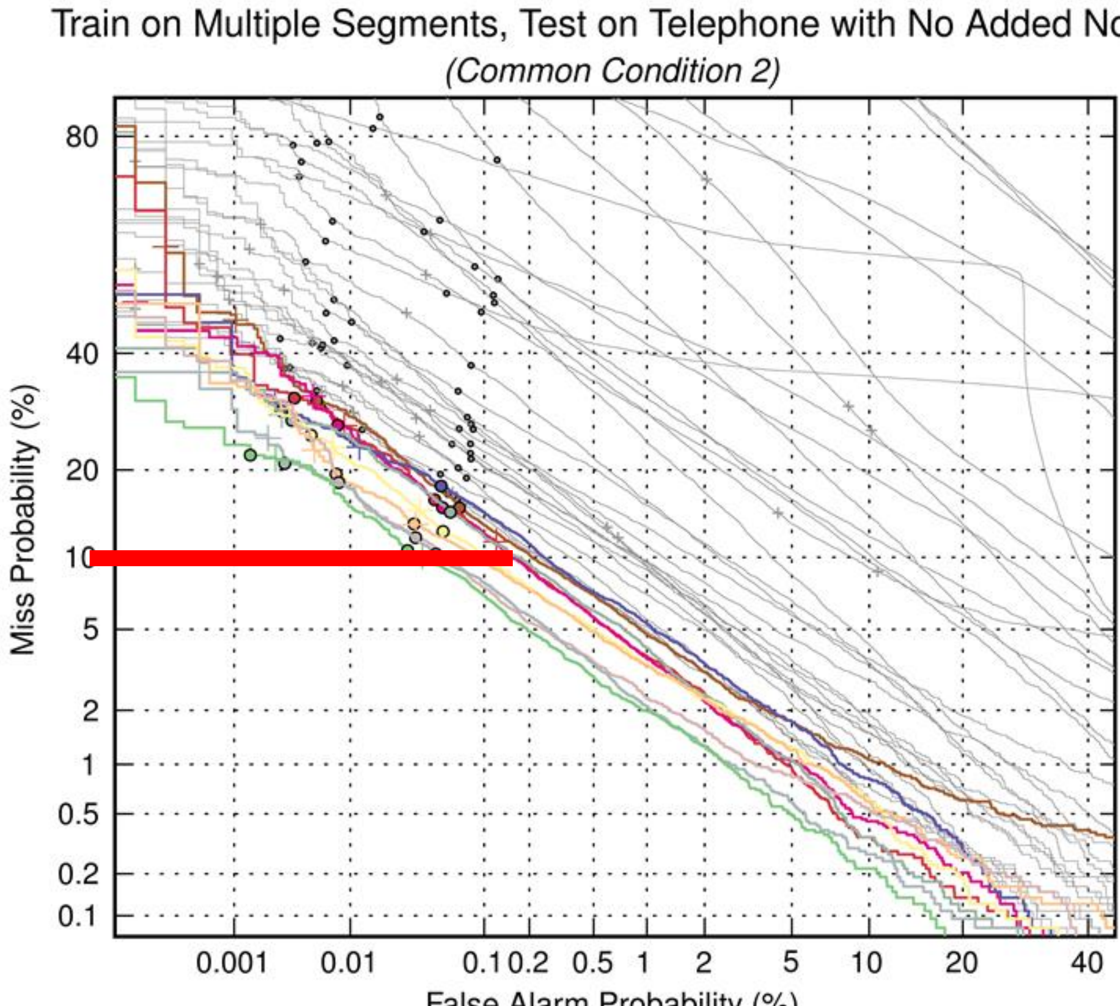
- Low-dimensional representation of whole recordings
  - i-Vectors (for R&D), Voiceprints (for business)



- Allows for very fast scoring.



- Works very nicely for long telephone recordings (EER ~2%) – multiple successes in NIST evaluations.
- Examples …



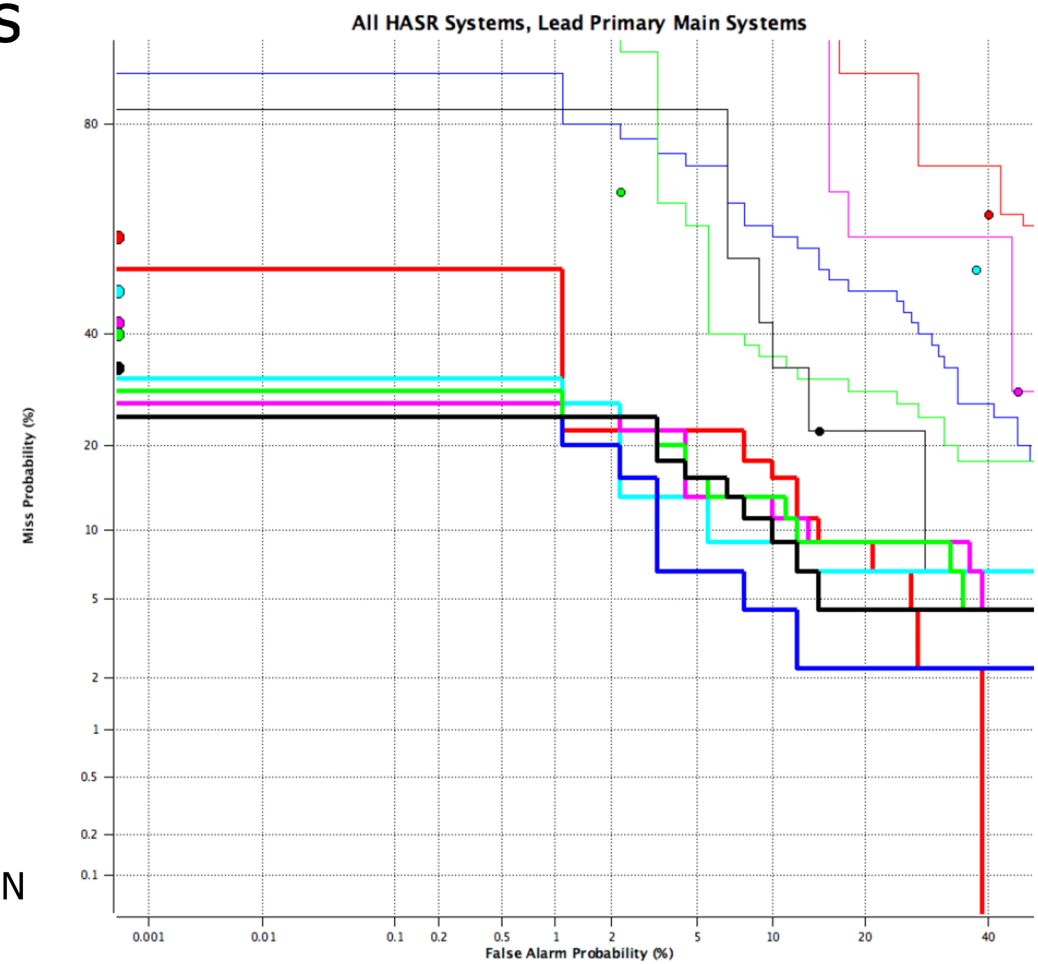
What to expect II.

- Noise, varying communication channels, short recordings (10s) still a problem – DARPA RATS program
- Examples …

fa@miss10%	miss@fa1.5%	EER
5.17	28.26	7.26
6.56	30.86	8.18
6.92	32.26	8.31
8.81	33.73	9.22
8.42	33.88	9.18
8.74	35.43	9.37
7.77	33.70	8.79
7.91	33.34	8.89

Comparison with human performance

- For **known voices**, humans are unbeatable.
- For unknown ones, machines are superior (especially for unfamiliar languages and environments…)



CRAIG S. GREENBERG, ALVIN F. MARTIN, MARK A. PRZYBOCKI: Human Assisted Speaker Recognition, NIST, INFORMATION TECHNOLOGY LABORATORY, INFORMATION ACCESS DIVISION, 2012.

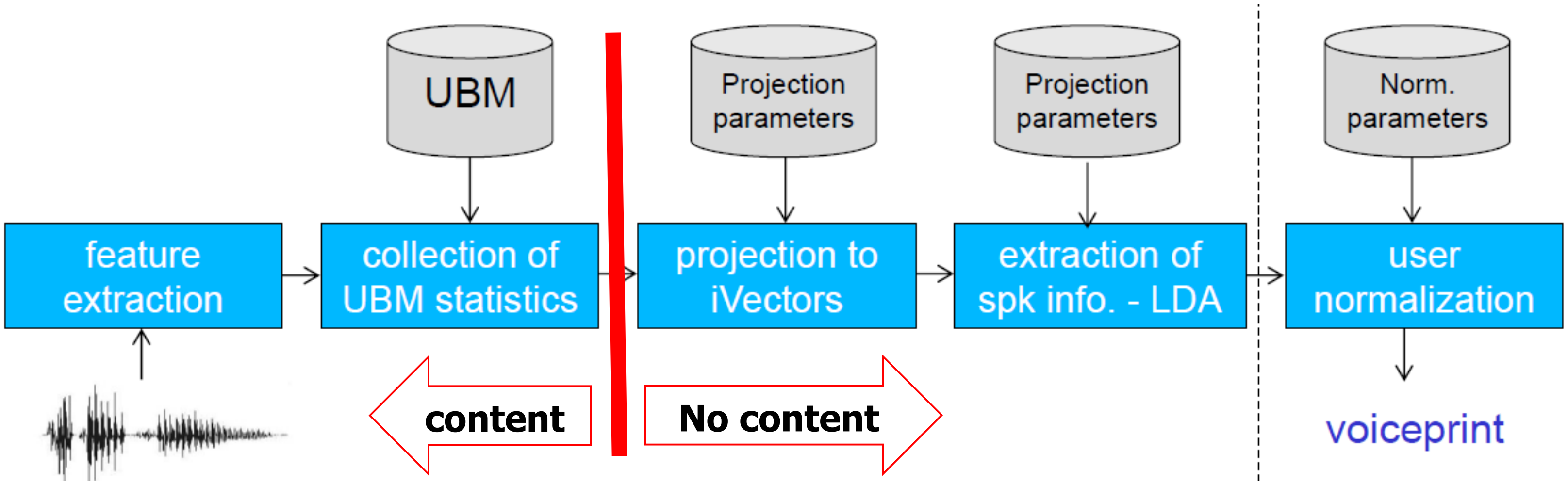
Agenda

- Basis of speaker verification
- Evaluation of performance
- Speaker verification system architecture
- Enemy No. 1: Intersession variability
- State of the art
- Data needed to adapt the system
- User scenarios
- Enemy No. 2: Attacks
- Vendors
- Conclusion

SRE – user data

- The performance of the SRE system crucially depends on how the training data is close to the deployment.
- UBM – needs lots (100s of hours) of unannotated data, not very sensitive.
- VoicePrint extractor – dtto.
- Scoring done by PLDA
  - Voice-prints with speaker labels (`A,B,C, …`) needed
  - Even 50 speakers help to increase the accuracy by 30%.
  - It might be problematic to collect even these 50 speakers (if possible on different communication channels…)
  - Work running on unsupervised adaptation on unannotated data.

- Allowing for transfer of speaker identities
  - without giving out the original WAV
  - Without possibility to reconstruct what was said.



Opening a range of opportunities for

- Cooperation between customers
- Cooperation with R&D teams.
- Standardization started !

Agenda

- Basis of speaker verification
- Evaluation of performance
- Speaker verification system architecture
- Enemy No. 1: Intersession variability
- State of the art
- Data needed to adapt the system
- User scenarios
- Enemy No. 2: Attacks
- Vendors
- Conclusion

Speaker Recognition Applications

Security and defense
Forensic
Looking for suspect in quantity of audio
Waiting online for suspect

Access Control
Physical facilities
Computer networks & websites

Transaction Authentication
Telephone banking
Remote purchases
Fraud detection

Speech Data Management
Voice mail browsing
Search in audio archives

Personalization
Voice-web/device customization
Intelligent answering machine

Application dictates different speech modalities

Text-dependent

- Recognition system knows text spoken by person
- Example: fixed or prompted phrases
- Used for applications with strong control over user
- Knowledge of spoken text can improve system performance

Text-independent

- Recognition system does not know text spoken by person
- Example: User selected phrases, conversational speech
- Used for applications with less or no control over user
- More flexible system but more difficult problem
- Speech recognition can provide knowledge of spoken text

Agenda

- Basis of speaker verification
- Evaluation of performance
- Speaker verification system architecture
- Enemy No. 1: Intersession variability
- State of the art
- Data needed to adapt the system
- User scenarios
- Enemy No. 2: Attacks
- Vendors
- Conclusion

Impresonation

- Impostor modifies his/her voice to sound as the genuine speaker.
- Good for humans, systems almost insensitive

Replay attack

- presenting recorded speech data from the genuine speaker
- Easy due to broad availability of high quality recording and playback devices (smartphones)
- Very difficult defense
- Text-dependent systems.

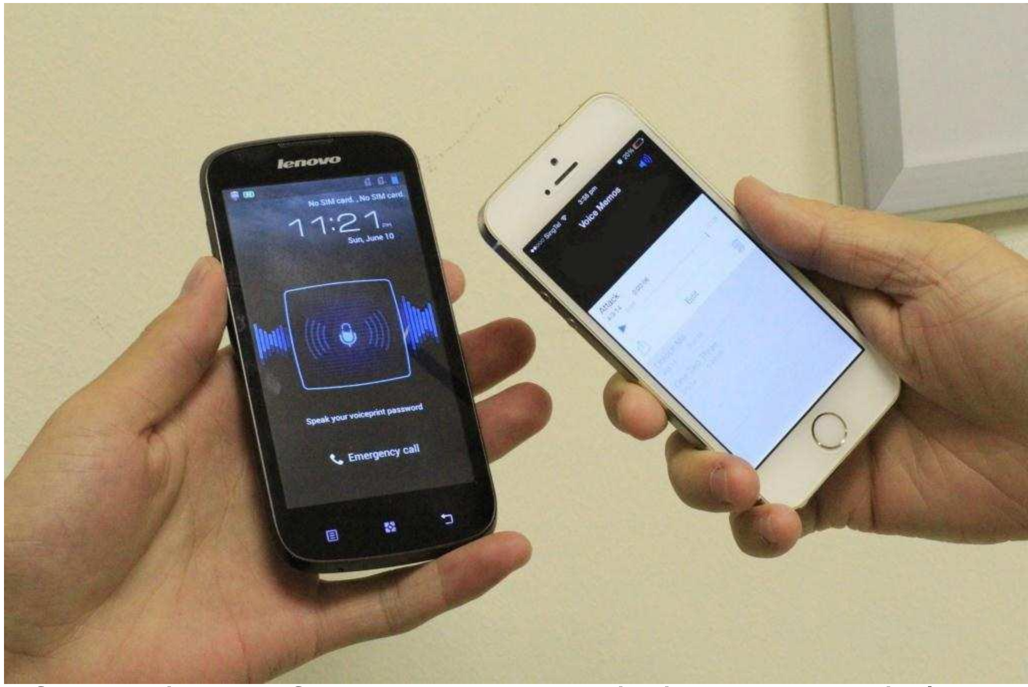


Illustration from Spoofing and countermeasures for speaker verification: a survey, Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, Haizhou Li, Speech Communication, Feb 2015

Speech synthesis

- speech synthesis systems can nowadays be modified to the voice of a particular speaker and used to attack even text-dependent systems
- Research works aiming at the detection of artificial speech, to the best of our knowledge, nothing done in production systems.
- Still requires speech processing skills but there’s a “democratization” of know-how and tools…

Voice modification

- modifying source impostor's voice to genuine speaker's voice allowing for "speaking as your mother in law"
- Same as for speech synthesis:
  - Research works aiming at the detection of artificial speech, to the best of our knowledge, nothing done in production systems.
  - Still requires speech processing skills but there's a "democratization" of know-how and tools…

Summary of attacks

Spoofing technique	Accessibility (practicality)	Effectiveness (risk)		Countermeasure availability
		Text-independent	Text-dependent	
Impersonation	Low	Low	Low	Non-existent
Replay	High	High	Low to high	Low
Speech synthesis	Medium to high	High	High	Medium
Voice conversion	Medium to high	High	High	Medium

Recommended reading:

Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, Haizhou Li: Spoofing and countermeasures for speaker verification: a survey, *Speech Communication*, Feb 2015.

Agenda

- Basis of speaker verification
- Evaluation of performance
- Speaker verification system architecture
- Enemy No. 1: Intersession variability
- State of the art
- Data needed to adapt the system
- User scenarios
- Enemy No. 2: Attacks
- Vendors
- Conclusion

Commercial vendors

- Nuance/Loquendo http://www.nuance.com/for-business/customer-service-solutions/voice-biometrics/index.htm
- Agnitio http://www.agnitio-corp.com/products/commercial/voice-authentication
- Speech Technology Center http://speechpro.com/product/voice-authentication/voicekey
- VoiceTrust http://www.voicetrust.com/
- Phonexia http://phonexia.com/technologies/sid

Notes on commercial systems

- All mentioned companies have state-of-the-art technology
- all of them are the best on the market!
- When acquiring Voice Biometry, you should ask:
  1. Where does the **core engine** come from ? From you or over 3 re-sellers ?
  2. Can we obtain a **functioning trial/demo** version to be evaluated by ourselves on our data ?
  3. How easily can we **adapt the system** on our data ?
  4. Plus the usual questions on integration, support, price …
- Good vendors will tell you what their engines are based on, everything is published !

Agenda

- Basis of speaker verification
- Evaluation of performance
- Speaker verification system architecture
- Enemy No. 1: Intersession variability
- State of the art
- Data needed to adapt the system
- User scenarios
- Enemy No. 2: Attacks
- Vendors
- Conclusion

Voice biometry summary

- Non-invasive, naturally available, no additional devices or hassle for users.
- Can operate in the background (during the call to the agent)
- Can be adapted to target conditions.
- Voice is spreading in the industry (SIRI, etc)

- Can't be recommended as the only modality for authentication
- No established evaluation methodology
- Attacks yet to come, tools are around.

Děkujeme za pozornost.

Honza Černocký
FIT VUT v Brně
cernocky@fit.vutbr.cz
http://speech.fit.vutbr.cz/