DT2118
Speech and Speaker Recognition
Speaker Recognition

Giampiero Salvi

KTH/CSC/TMH giampi@kth.se

VT 2015

Outline

Outline

Person Identification

Methods rely on:

- something you posses:
  key, magnetic card, ...
- something you know:
  PIN-code, password, ...
- something you are:
  physical attributes, behaviour (biometrics)

Biometric identification features

physical attributes	**activity/behaviour**
height and weight	
finger print	handwriting
hand shape	typing patterns
retina	gestures
face	facial expressions
speech	
vocal tract size	speech rate
nasal cavities	intonation
glottal folds	vocabulary, grammar

Recognition, Verification, Identification

Recognition: general term

Speaker verification:

- an identity is claimed and is verified by voice
- binary decision (accept/reject)
- performance independent of number of users

Speaker identification:

- choose one of N speakers
- close set: voice belongs to one of the N speakers
- open set: any person can access the system
- problem difficulty increases with N

Speaker Recognition: Advantages

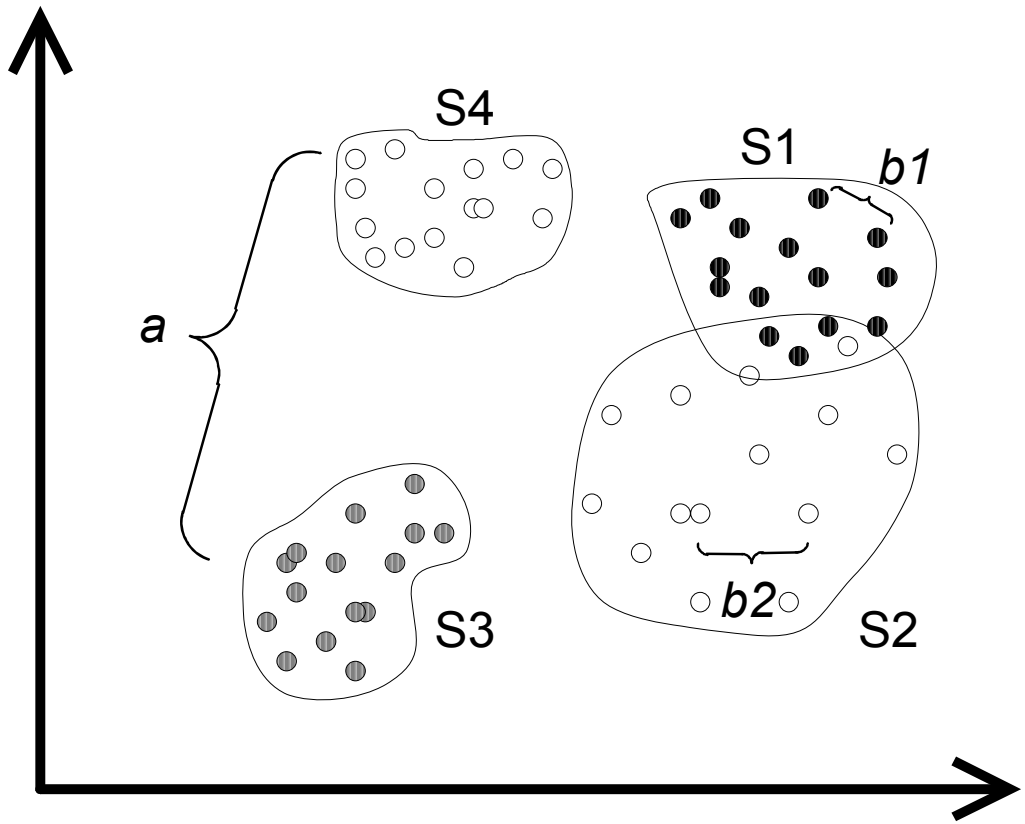
- speech is natural
- simple to record (cheap equipment)
- speech may already be used in the application

Speaker Recognition: Limitations

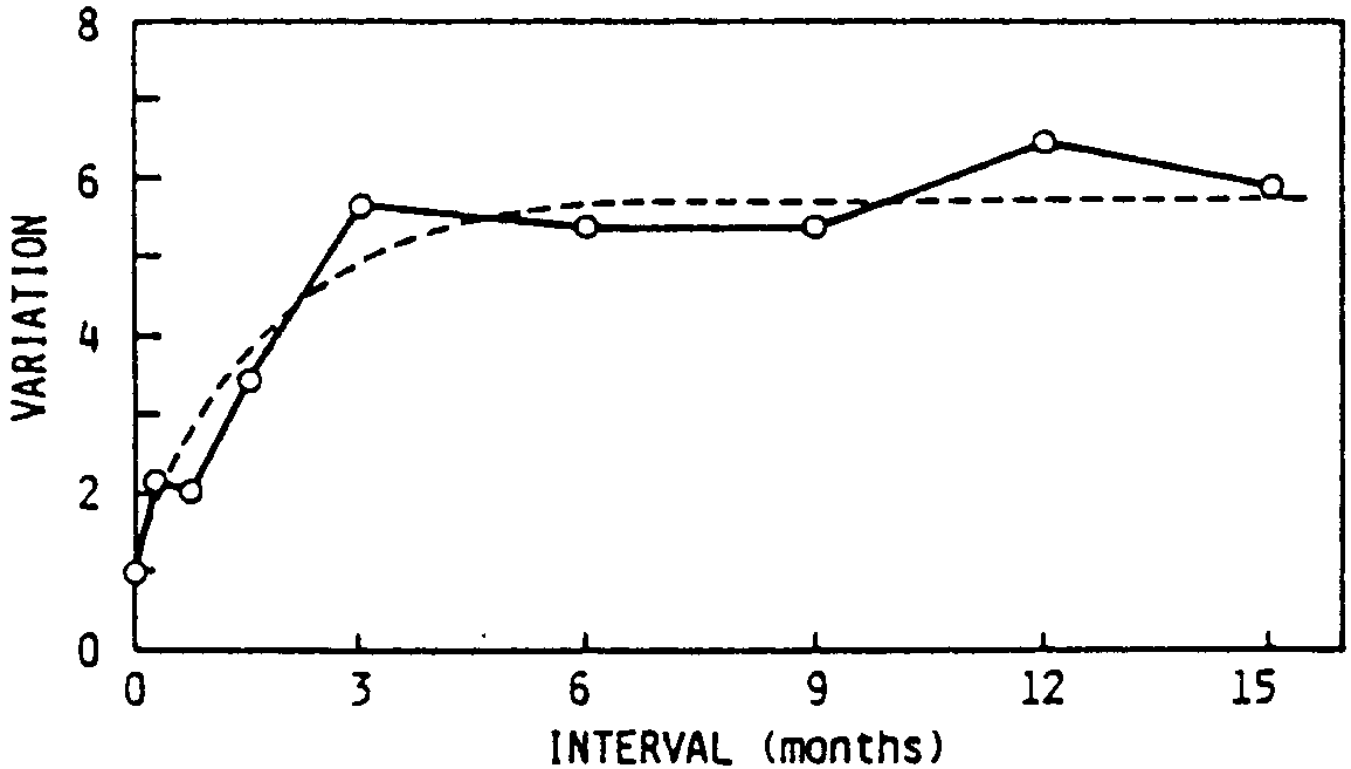
- not 100% security (but that's true for other techniques)
- large variability in speech
- behaviour, different microphones, physical and mental condition

Outline

The Speaker Space



Voice Variability in Time



Within-speaker variability (identical utterance)
average over 9 male speakers [1]

---

[1] S. Furui. "Research of individuality features in speech waves and automatic speaker recognition techniques". In: *Speech Communication* 5.2 (1986)

Influence of the Channel

- different microphones (e.g.: telephones)
- transmission: line, equipment, coding, noise
- little control over the speaker and environment if remotely connected

Challenge: separate speaker characteristics from environment (both are long-time properties of the signal)

Representations

Speech Recognition:

- represent speech content
- disregard speaker identity

Speaker Recognition:

- represent speaker identity
- disregard speech content

Representations

Speech Recognition:

- represent speech content
- disregard speaker identity

Speaker Recognition:

- represent speaker identity
- disregard speech content

Surprisingly:

- MFCCs used for both
- suggests that feature extraction could be improved

Text Dependence

Either fix the content or recognise it. Examples:

- Fixed password (text dependent)
- User-specific password
- System prompts the text (prevents impostors from recording and playing back the password)
- any word is allowed (text independent)

Modelling Techniques

HMMs

- Text dependent systems
- state sequence represents allowed utterance

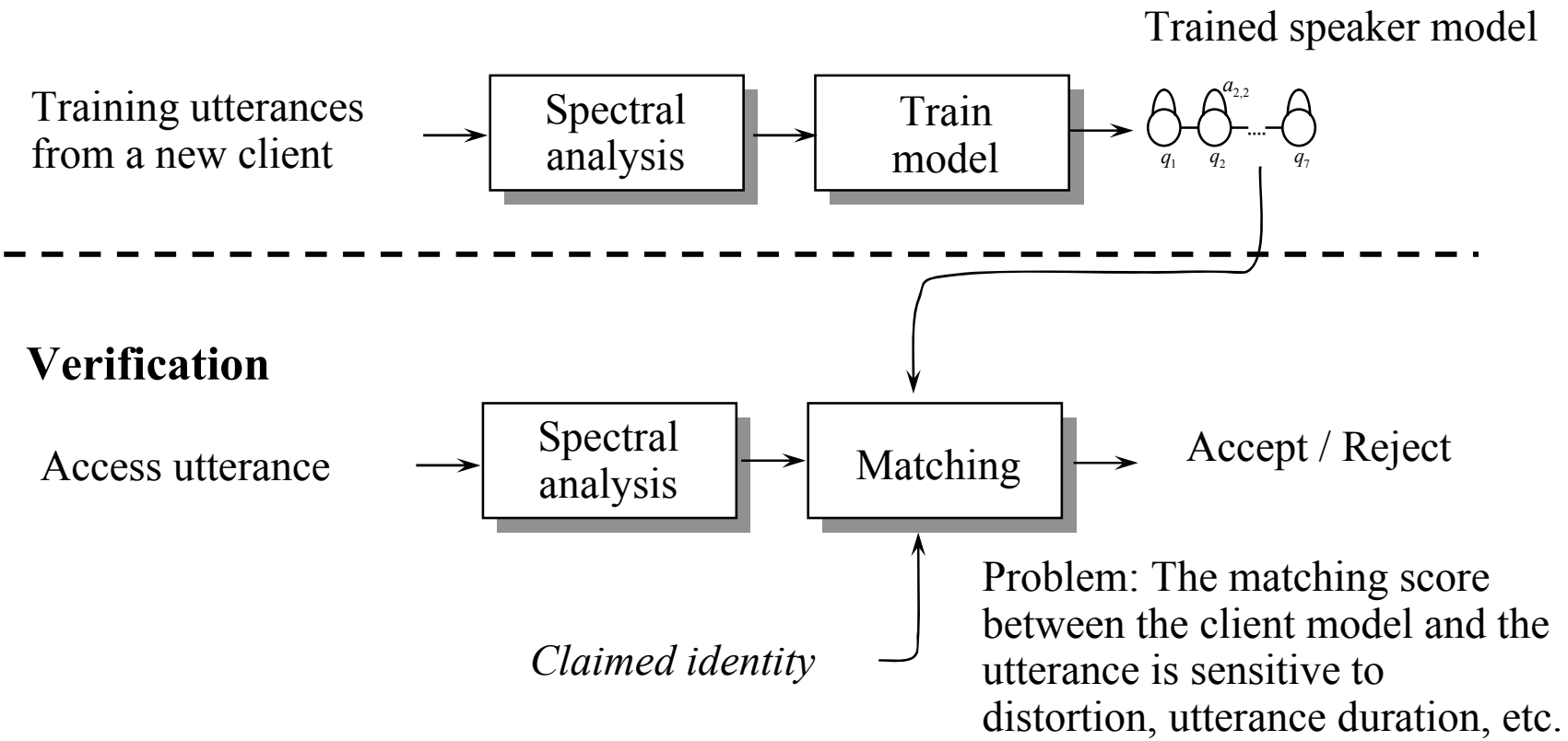
GMMs (Gaussian Mixture Models)

- Text independent systems
- large number of Gaussian components
- sequential information not used

SVM (Support Vector Machines)

Combined models

Speaker Verification



Probabilistic Approach

Bayes decision theory (C: client, $\bar{C}$: not client)

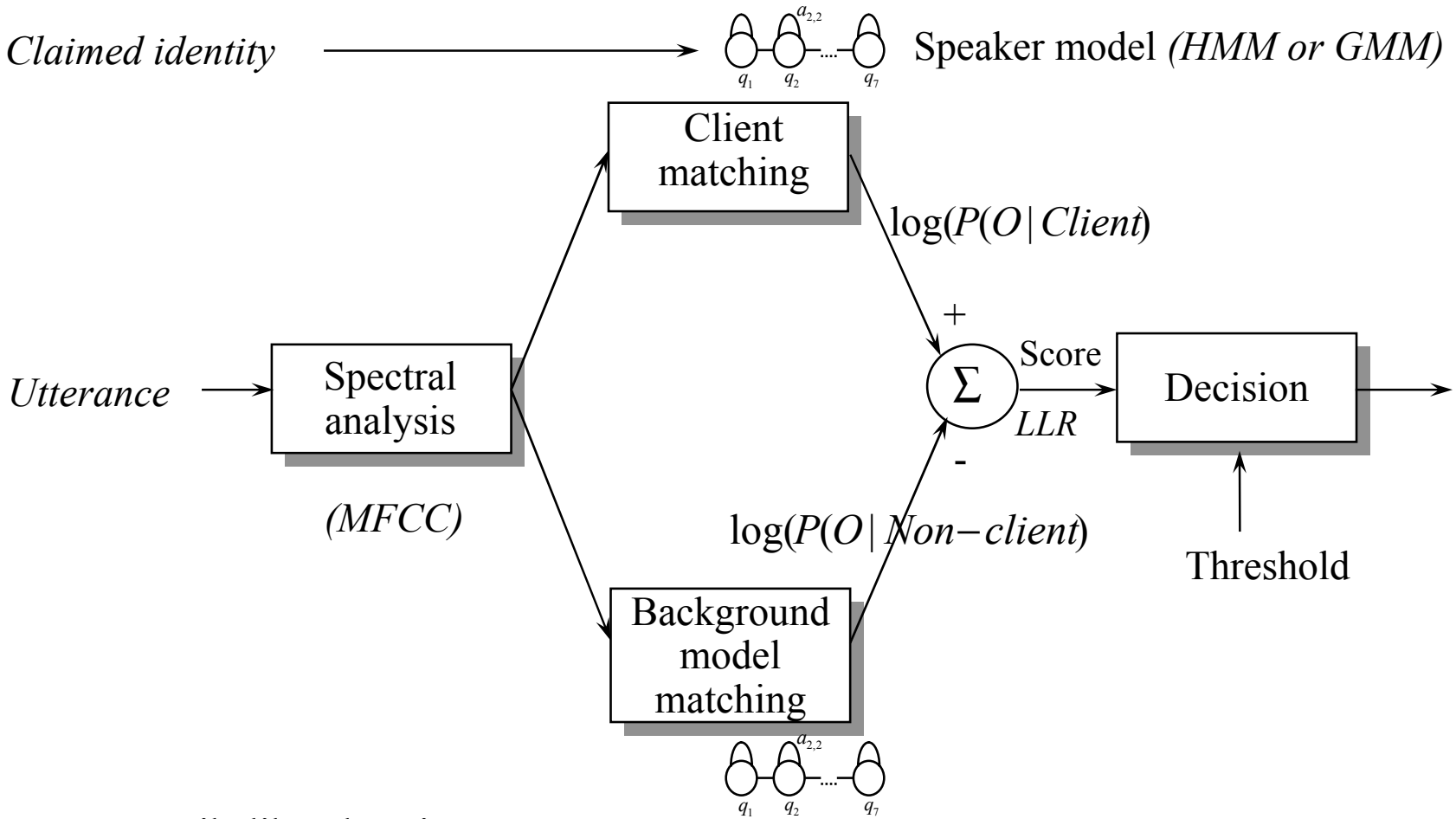
$$\frac{\text{client sounds like this}}{\text{anybody sounds like this}} = \frac{P(C|O)}{P(\bar{C}|O)} =$$

$$= \frac{P(O|\theta_C)P(C)}{P(O|\theta_{\bar{C}})P(\bar{C})} > R$$

Optimal Threshold:

$$R = \frac{\text{Cost of False Accept}}{\text{Cost of False Reject}}$$

Standard System



LLR: Log Likelihood Ratio

Background model(-s) *(HMM or GMM)* from many speakers

Client model estimation in text-independent system

Not realistic to train the GMM for each client

- risk of unreliable estimation

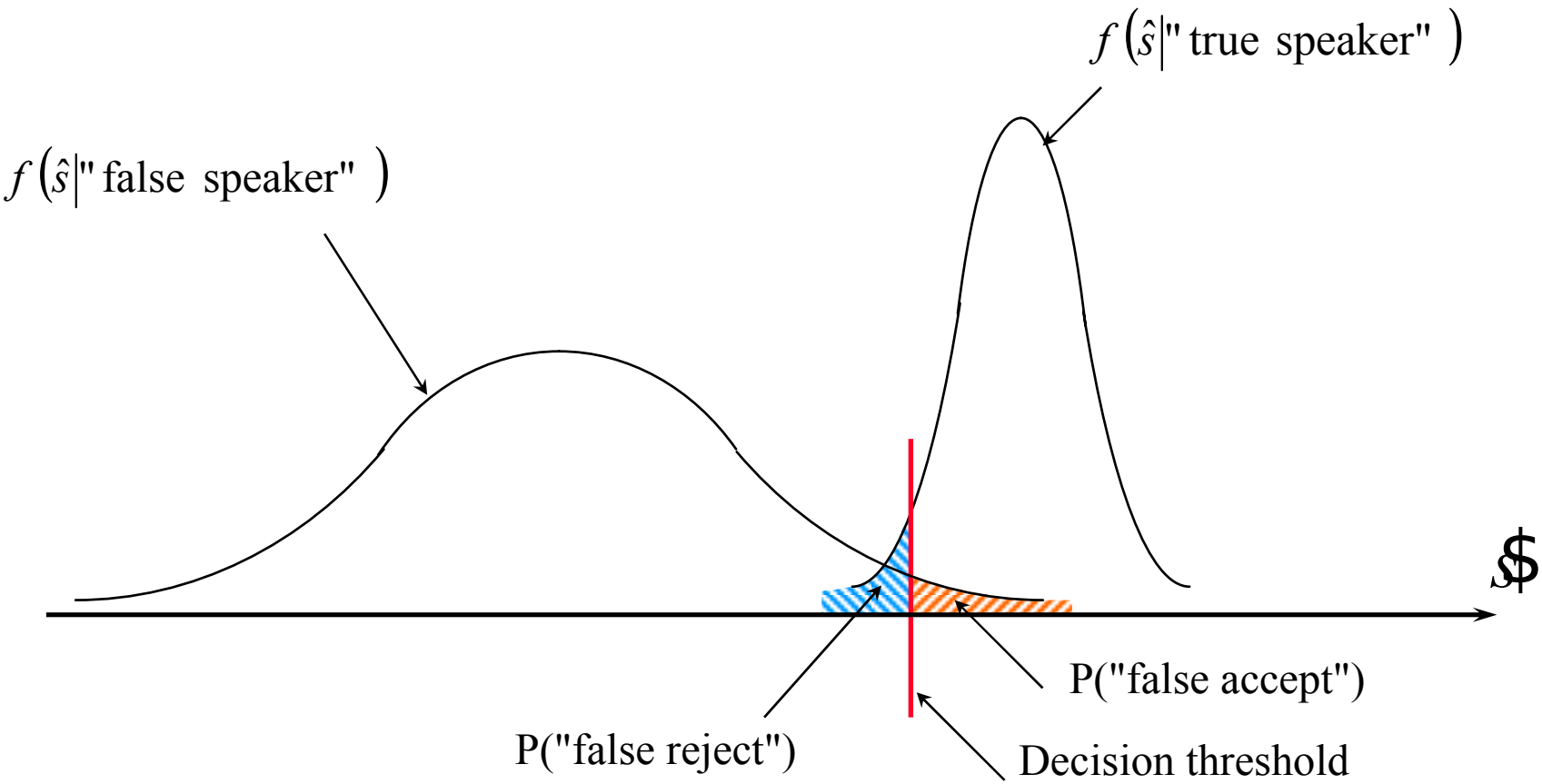
Instead adaptation of background model (multi-speaker)

- non-observed components in adaptation are unchanged
- they do not contribute in the matching probability ratio
- only well trained components contribute

Evaluation

Claimed Identity	Decision: Accept	Decision: Reject
True	OK	False Reject (FR)
False	False Accept (FA)	OK

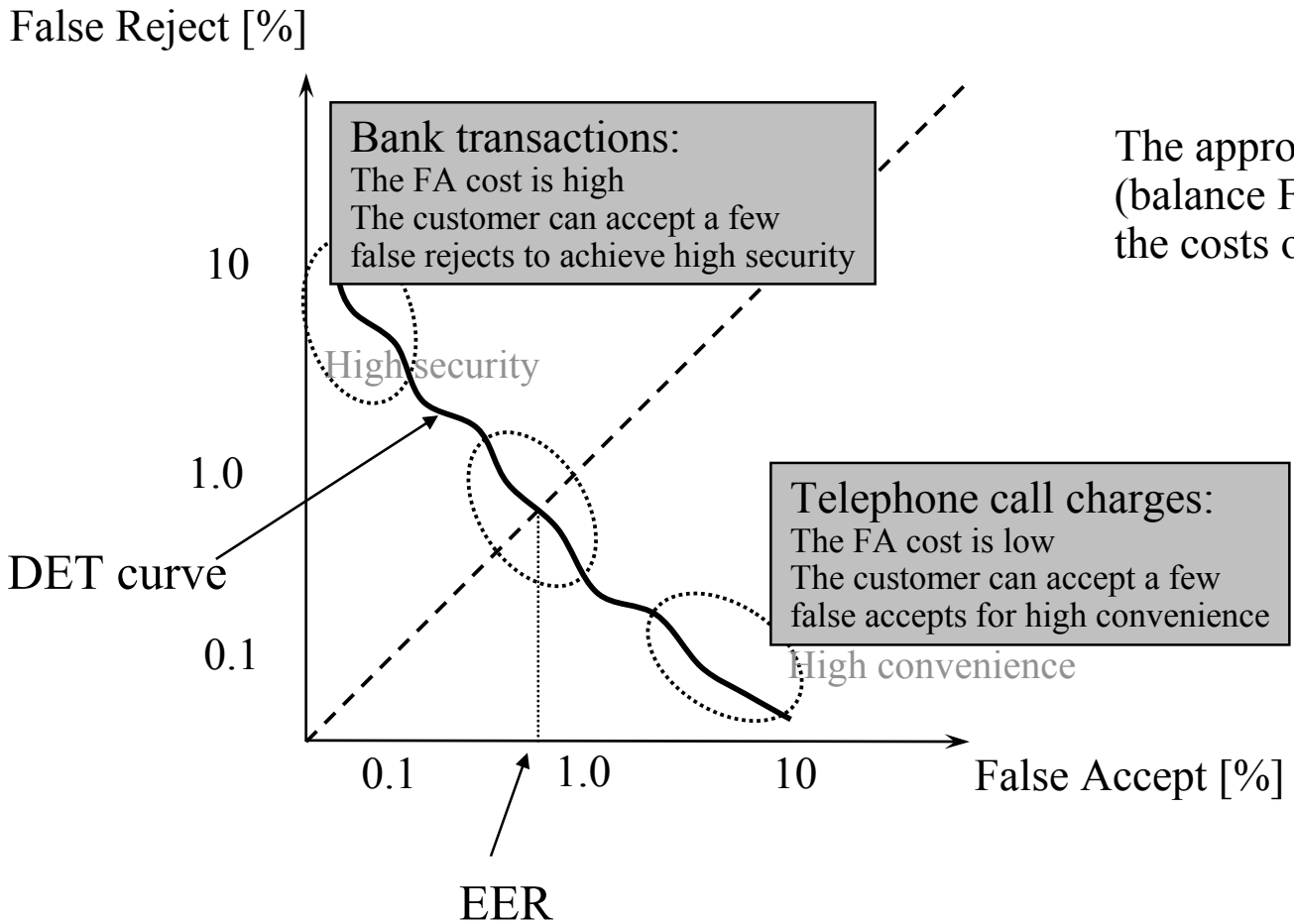
Score Distribution and Error Balance



Performance Measures

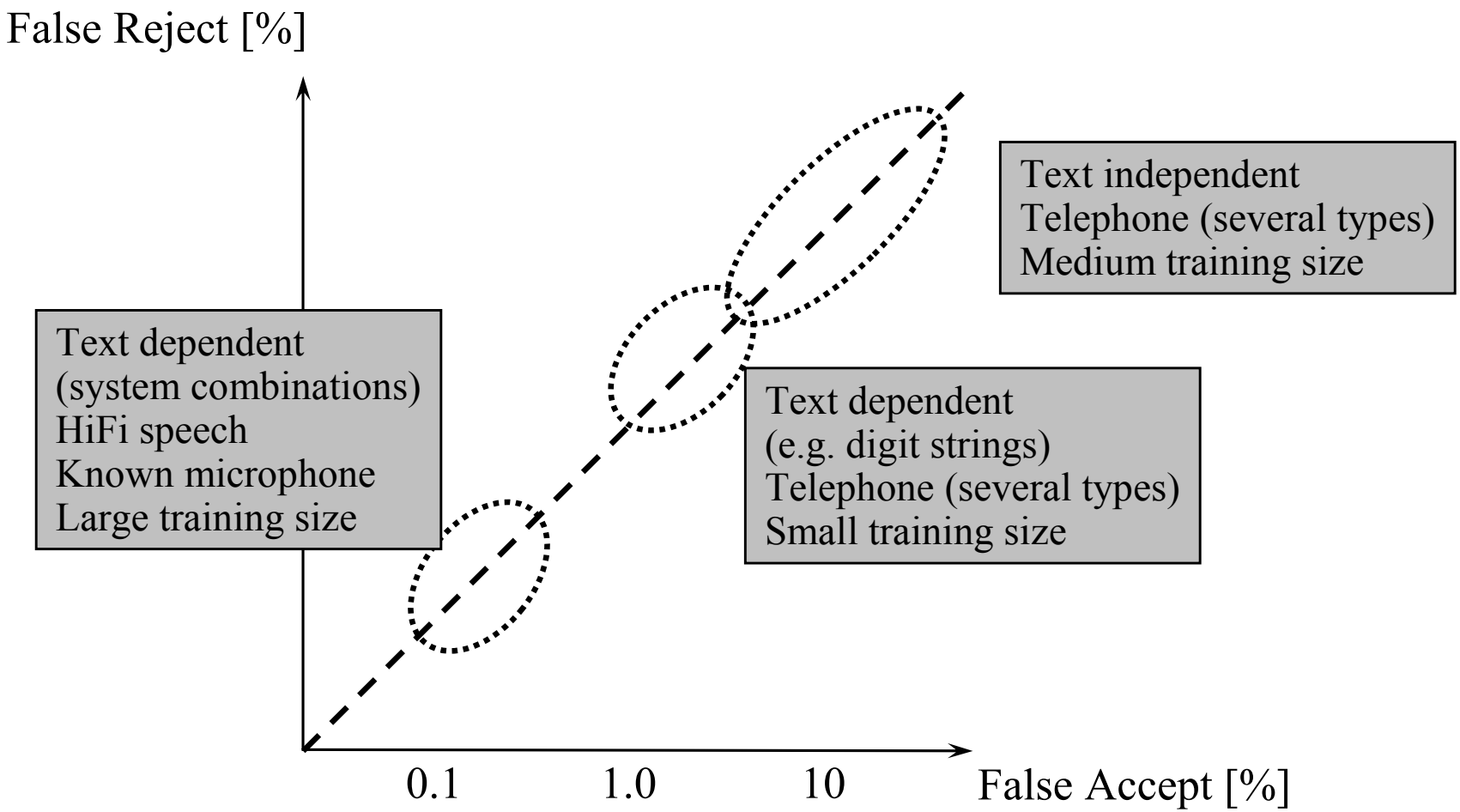
- False Rejection Rate (FR)
- False Acceptance Rate (FA)
- Half Total Error Rate (HTER = (FR+FA)/2)
- Equal Error Rate (EER)
- Detection Error Trade-off (DET) Curve

Application-Dependent Operating Point

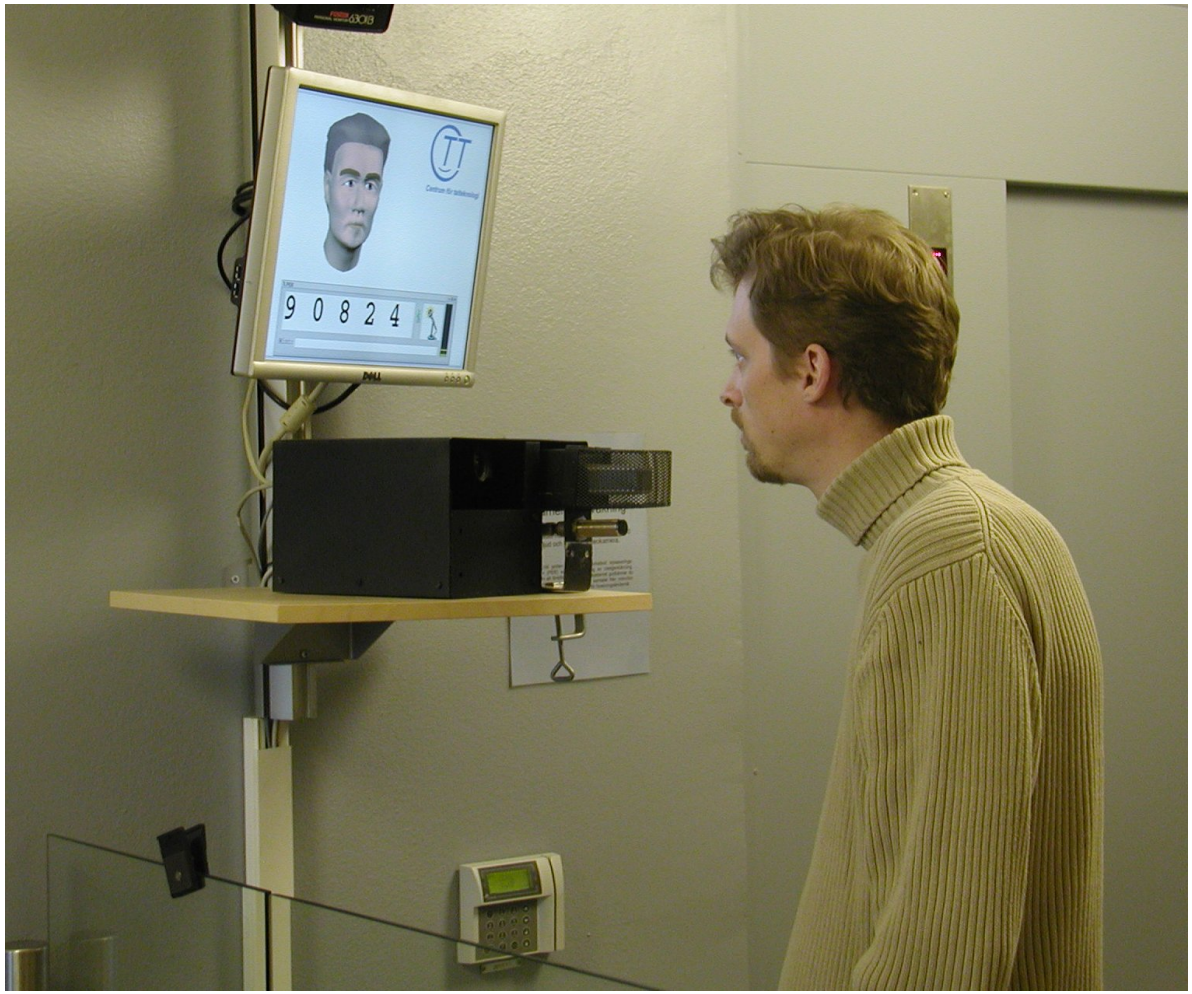


The appropriate operating point (balance FA/FR) depends on the costs of each error type

Performance in Different Applications

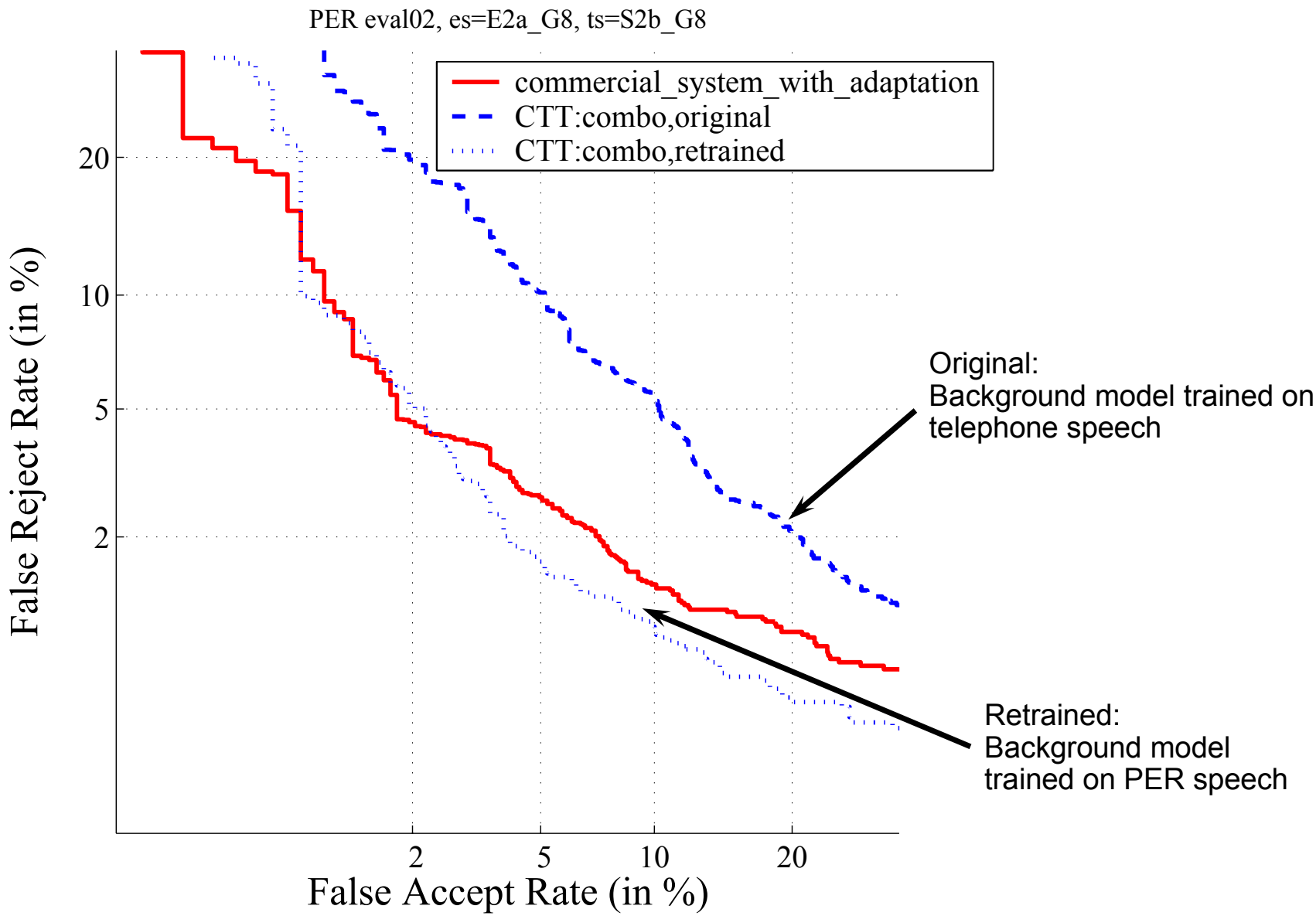


In-House Example



Created by Håkan Melin

PER vs Commercial System



The Animal Park

Categorisation of speakers by the system performance

- Sheep: "harmless" users with low error rate
- Goats: "non-reliable", high variability, high error rate
- Lambs: vulnerable, easy to impersonate
- Wolves: potentially successful impostors

Impostors

- Performance usually measured on random speakers as impostors
- how different are real impostors?
- might have knowledge of client's voice
- technical impostors

Technical impostors

Varying technical sophistication

- Playback of recorded speech
- Concatenative synthesis
- Voice transformation
- Trainable speaker dependent speech synthesis

Preventive techniques

- Detect artificial features (typical features of speech synthesis)
- Detect if repetitions of the same text are identical

Competition development race between imposture and prevention techniques

Outline

Multi-Speaker Recordings

n-speaker detection: is a speaker present in a conversation

speaker tracking: same as above plus time positioning

speaker segmentation: determine the number of speakers and when they speak

Outline

Forensic Speaker Recognition

Determine if a suspect of a crime has spoken the recorded utterance

Difficulties

- Unknown and uncontrollable recording conditions
- High degree of variability
- Incooperative speakers: The speaker does not want to be identified as the target speaker, the opposite to speaker verification
- May try to disguise his/her voice

Risk of Incorrect Use

Example:

- False Acceptance Rate $= 1\%$
- possible prosecutor conclusion: 99% probability the suspect is guilty
- possible defense conclusion: if in the city there are 100.000 inhabitants, 1000 would match. 0.1% probability the suspect is guilty

Neither is right. Use Bayesian decision theory (similar to differential diagnosis)