Speaker Recognition

Mats Blomberg

Speech, Music and Hearing
KTH/CSC

Literature

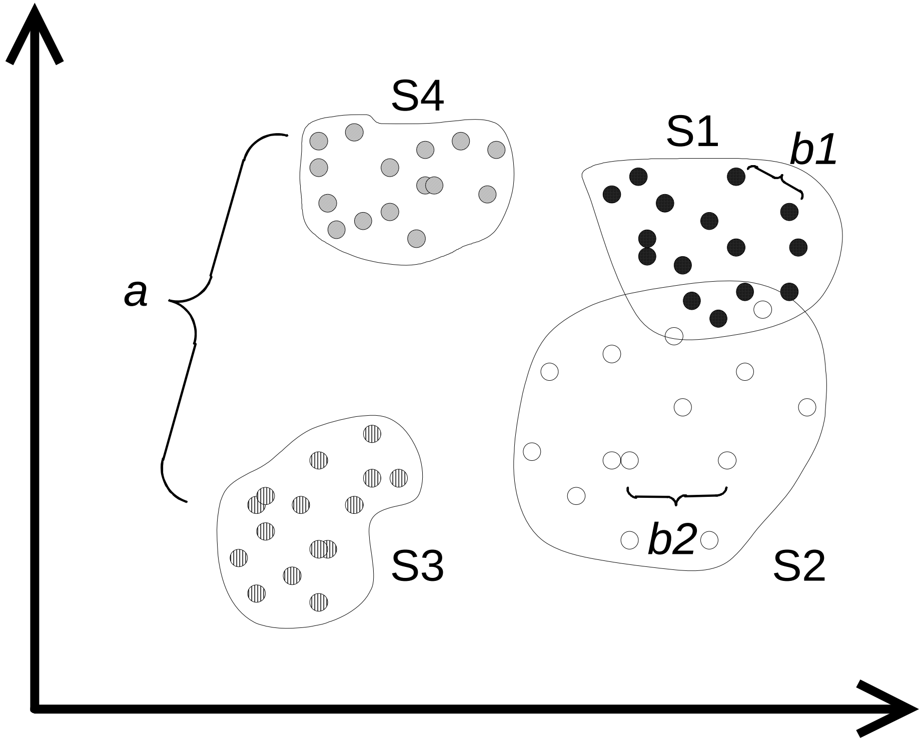
- Bimbot, F. et al. (2004). A Tutorial on Text-Independent Speaker Verification. EURASIP J. on Appl. Sig. Proc. 2004:4
- Melin, H. (2006). Automatic speaker verification on site and by telephone: methods, applications and assessment. Doct. Thesis, Speech, Music and Hearing, KTH.
- Champod, C. and Meuwly, D. (1998). The inference of identity in forensic speaker recognition. RLA2C Workshop, Avignon.
- Zetterholm, E., Blomberg, M., Elenius, D. (2004) A comparison between human perception and a speaker verification system score of a voice imitation, Proc Australian Int. Conf. on Speech Science & Technology.

Lecture overview

- Repetition of basic theory and techniques
- Score normalisation techniques
- PER entrance system
- Imitation
- Forensics

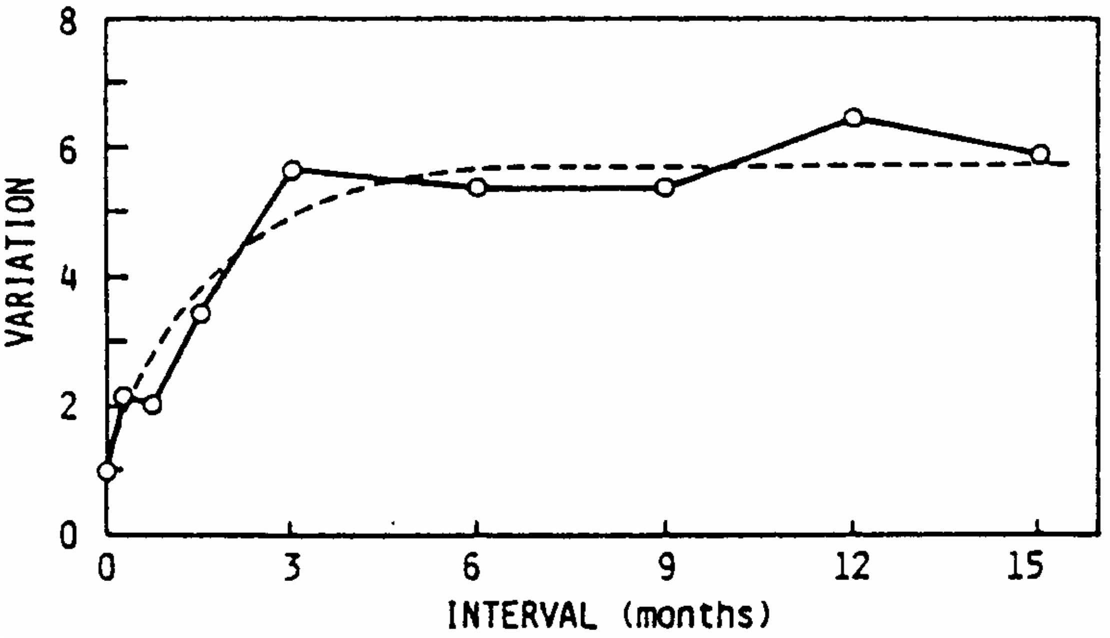
The speaker space

(Intra- vs. interspeaker distance)



Voice characteristics vary with time

Variability within one speaker



Acoustic variation among identical utterances as a function of the duration of the recordings. Average for nine male speakers. (Furui, 1986).

Influence of the telephone network

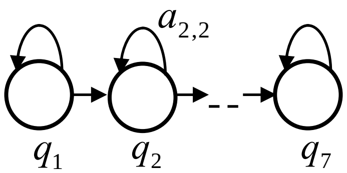
- Different telephones (microphones)
- Transmission
  - Variability among lines and equipment
  - Digital coding
  - Noise
- Little control or knowledge over the speaker and the speaker's environment
- *Challenge: To separate speaker specific from environmental specific parameters!*

Same or different analysis as in speech recognition?

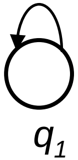
- SPEECH recognition should be SPEAKER independent
  - Should extract phonetic information but not speaker information
- SPEAKER recognition should be SPEECH independent
  - Should extract speaker information but not speech information
- This suggests that the optimal acoustic features are different between speech and speaker recognition
- However, experiments have shown that the best SPEECH representation is at the same time one of the best SPEAKER representations
- Why? Maybe the optimal representation contains both SPEECH and SPEAKER information

Modelling techniques

- HMM
  - Text-dependent systems
  - The state sequence represents allowed utterances



- GMM (Gaussian Mixture Models)
  - Text-independent systems
  - Single-state HMM with large number of Gaussian mixture components (~ 1000) representing any utterance by the speaker
  - Sequential information is not used



- Combined GMM + HMM systems

- Discrimination-based learning
  - Support Vector Machines (SVM)
  - Artificial Neural Networks (ANN)

Support Vector Machines (SVM)

- Increasingly popular
- Separates complex regions between two classes through an optimal nonlinear decision boundary
- A kernel function transforms the complex speaker space to a space more suited for linear discrimination
- Limitation: Can’t handle the temporal structure of speech
- Combination with GMM possible
  - Use GMM likelihood values for each frame and mixture component as input vector to SVM

Speaker Verification

- Most applied application in speaker recognition
- Binary decision Accept/Reject claimed identity

- Speaker Identification
  - Identify the speaker as one out of N candidates

Probabilistic approach

- Bayes’ decision theory
  - The ratio between the probability scores of a client and an anti-client model is compared with a decision threshold

If $$\frac{\text{P(The client sounds like this)}}{\text{P(Anybody could sound like this)}} > \text{R}$$

$$\boxed{\frac{P(O\,|\,\theta_C)}{P(O\,|\,\theta_{\overline{C}})} \ge R}$$

Then accept, else reject

O: utterance
θ_C: model of client C

optimal threshold $$R = \frac{P(\overline{C})}{P(C)}\frac{Cost_{FA}}{Cost_{FR}}$$

The optimal threshold is dependent on the a priori probability of impostors and the cost ratio between False Accept and False Reject

If the costs are equal and the impostors and true clients are qually probable,
The optimal threshold is 1, i.e. accept the speaker if the probability of
the client model is larger than that of “anybody else”. Makes sense.

Probabilistic approach (2)

Two variants of anti-clients (background models)

-Method 1: “**Universal (global) model**”

-One model is trained on a large number of speakers;

-client independent

$$P\left(O \mid \theta_{\overline{C}}\right) \approx P\left(O \middle| \theta_M\right)$$

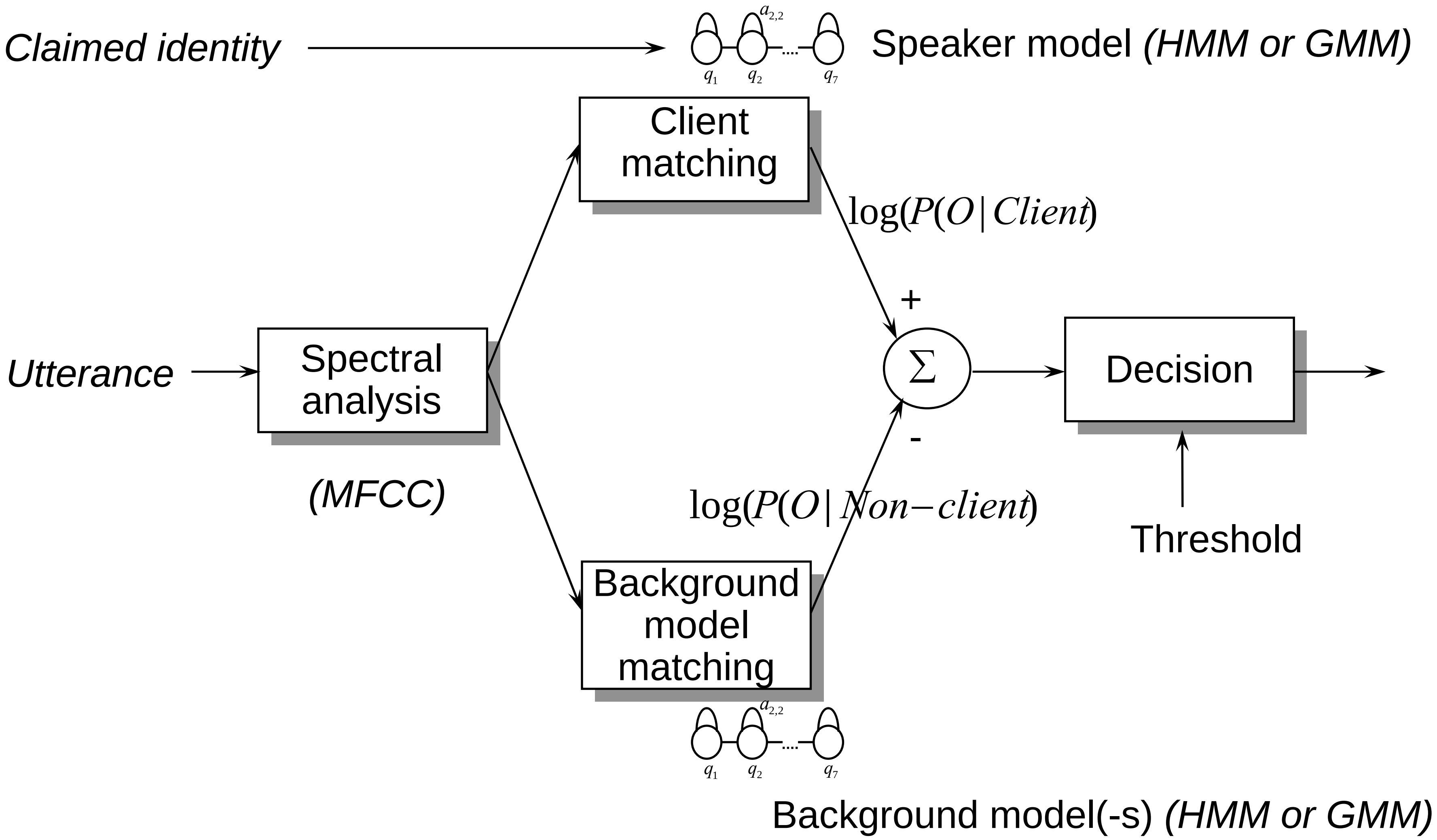
-Method 2: “**Cohort-model**”

-Several submodels are trained on small speaker groups “close” to the client;

-client specific

$$P\left(O \mid \theta_{\overline{C}}\right) \approx \frac{1}{\left|W_M\right|} \sum_{i \in W_M} P\left(O \middle| \theta_i\right) \approx \frac{1}{N} \sum_{\substack{i \in W_M \\ N \text{ "closest"}}} P\left(O \middle| \theta_i\right)$$

Standard system



Claimed identity

Speaker model *(HMM or GMM)*

Client matching

$\log(P(O \mid Client))$

+

Utterance

Spectral analysis

(MFCC)

Σ

Decision

-

$\log(P(O \mid Non-client))$

Threshold

Background model matching

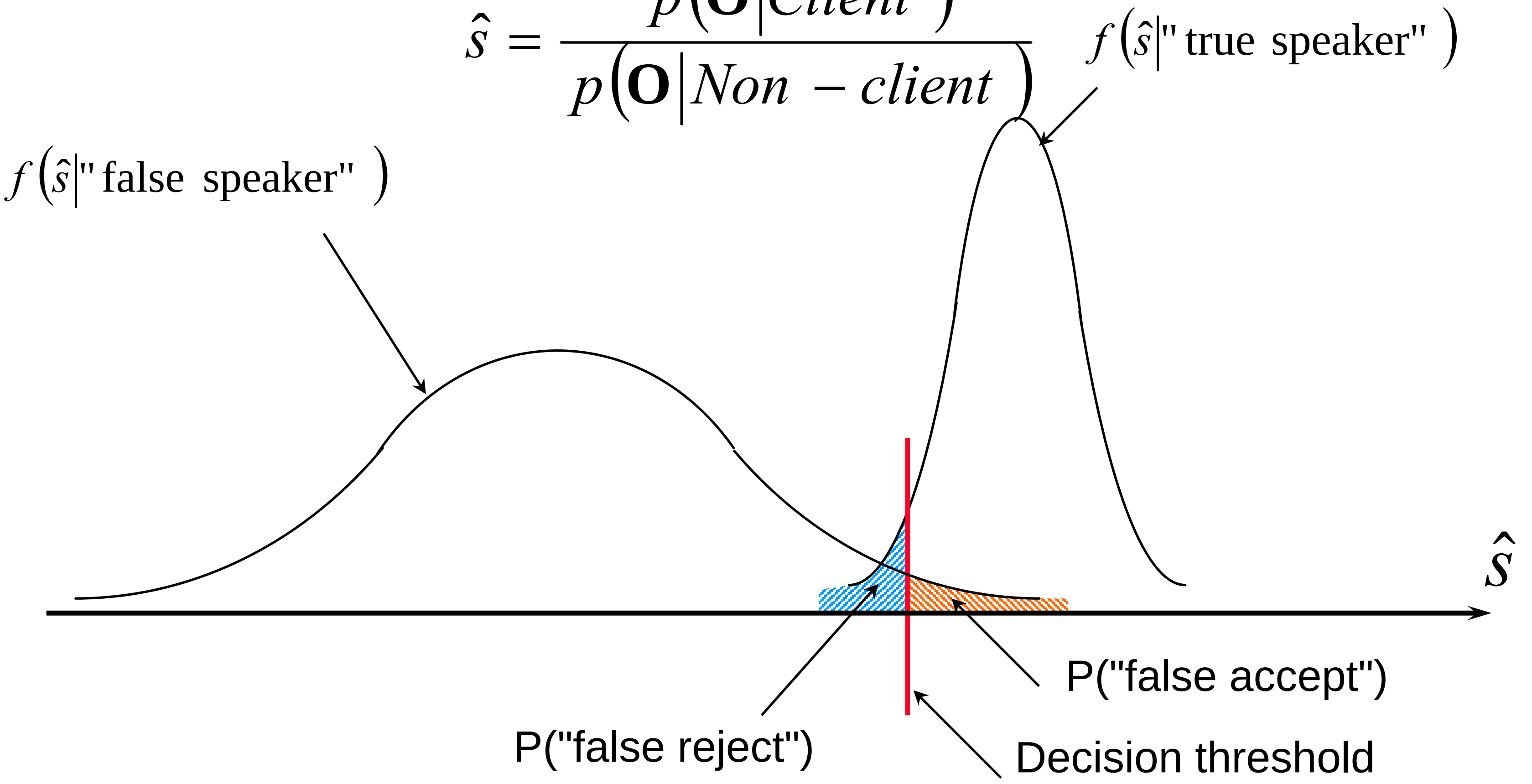
Background model(-s) *(HMM or GMM)*

Two types of errors

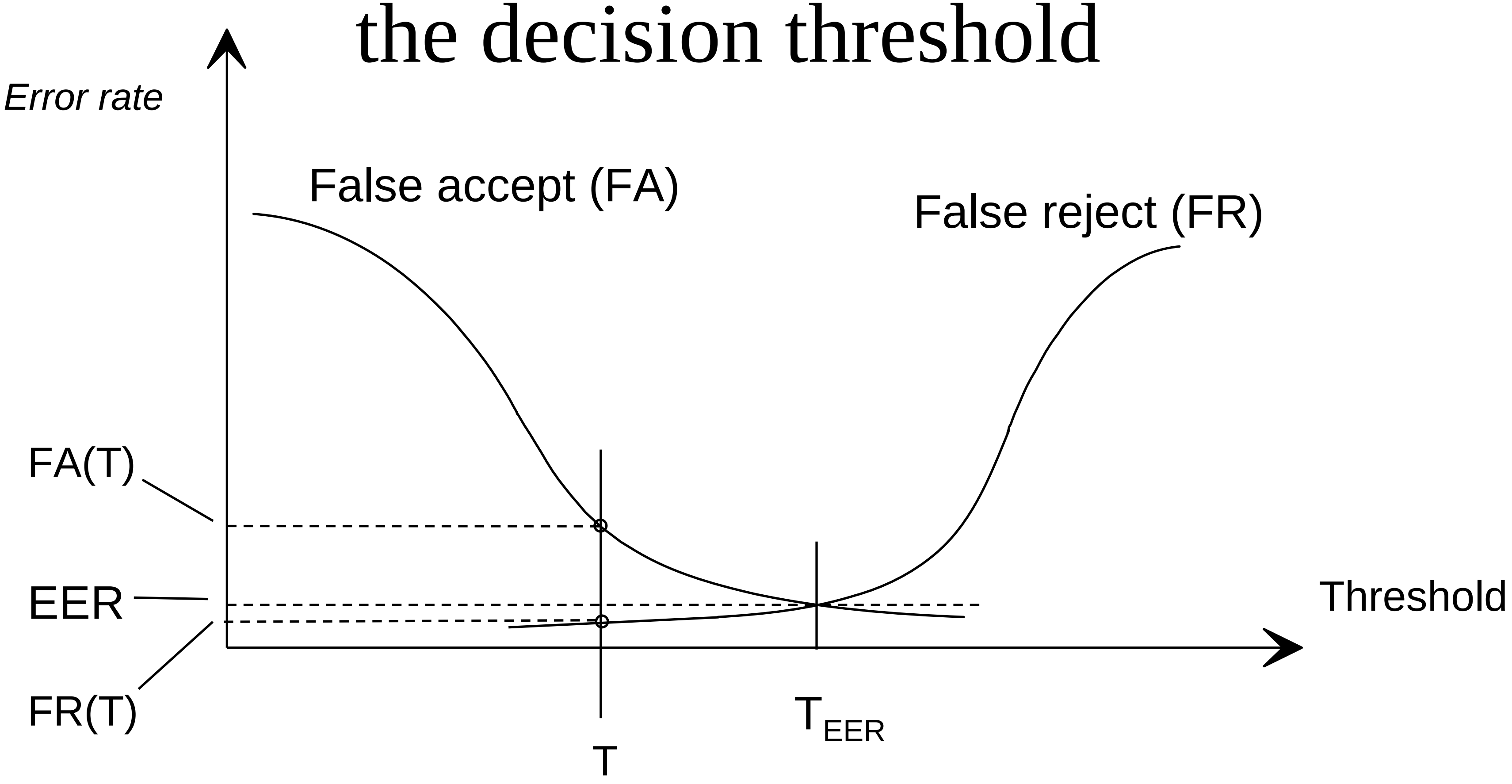
		Claimed identity:	
		True	*False*
Decision:	*Accept*	OK	False Accept (FA)
	Reject	False Reject (FR)	OK

Score distribution
for true and false speaker identities

$$\hat{s} = \frac{p\left(\mathbf{O} \middle| Client\right)}{p\left(\mathbf{O} \middle| Non-client\right)}$$


$f\left(\hat{s}\middle|\text{"true speaker"}\right)$
$f\left(\hat{s}\middle|\text{"false speaker"}\right)$
$\hat{s}$
P("false accept")
P("false reject")
Decision threshold

The error balance depends on the decision threshold

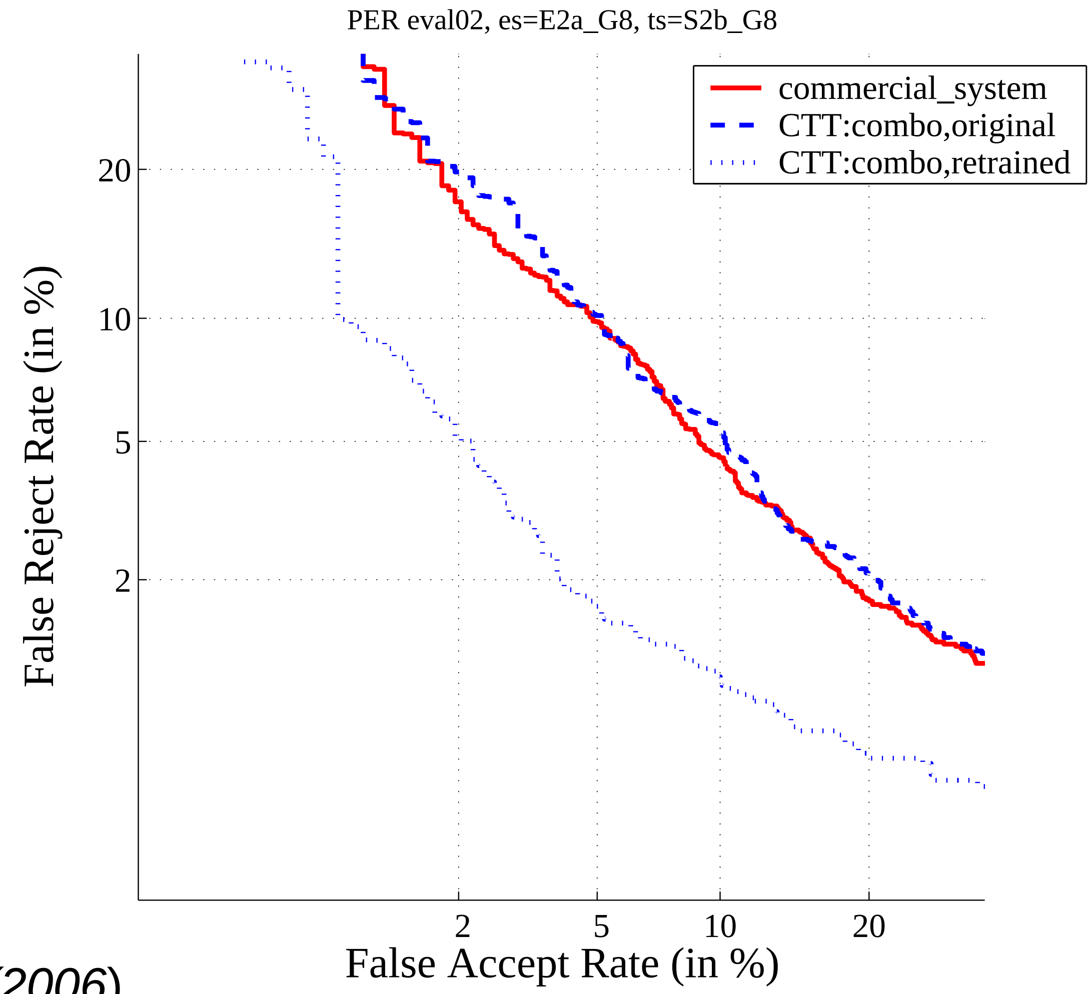


EER: Equal Error Rate, $EER = FA(T_{EER}) = FR(T_{EER})$
at an *a posteriori* determined threshold

Performance measures

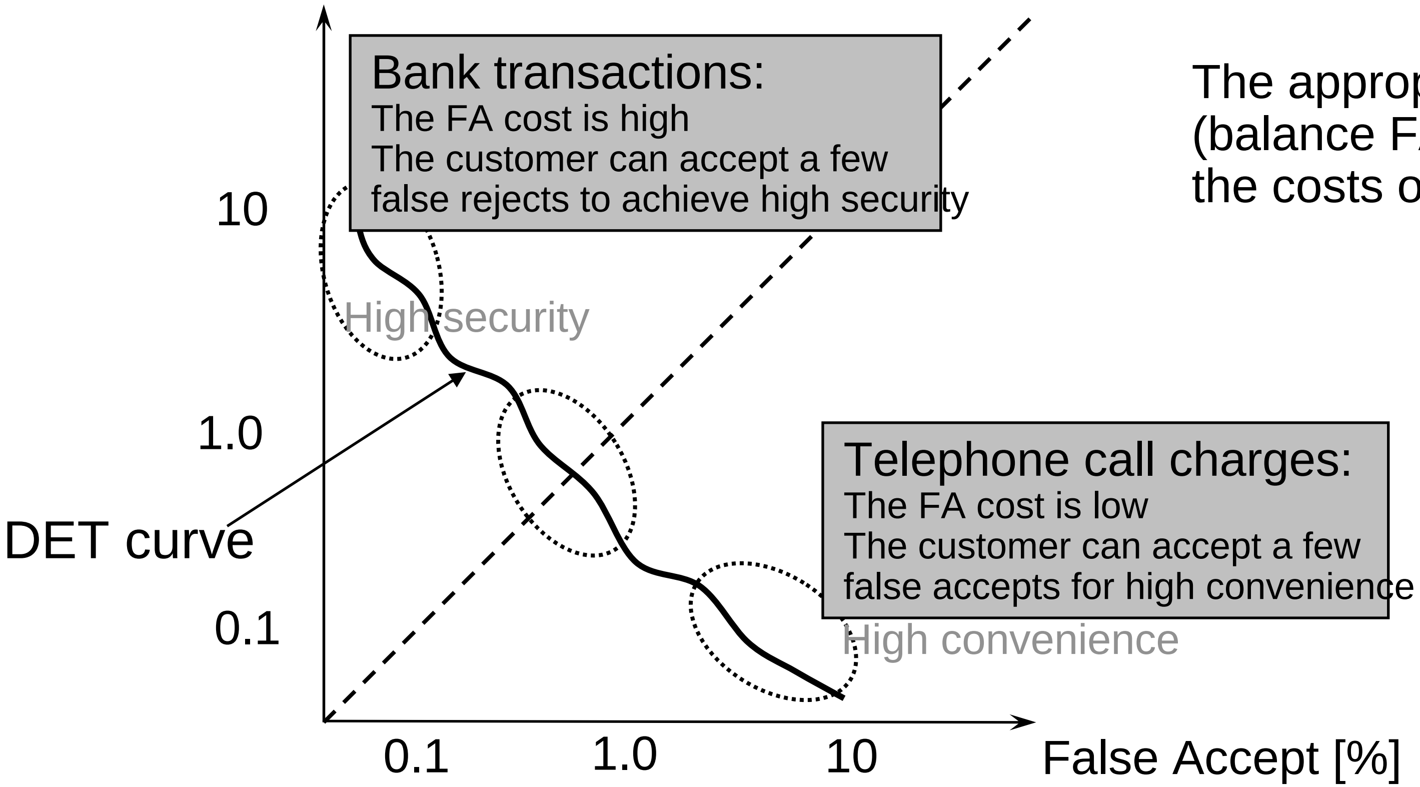
- False Rejection rate (FR)
  - FR = (Nbr *false reject utterances*) / (Nbr true ID attempts)
- False Acceptance rate (FA)
  - FA = (Nbr *false accept utterances*) / (Nbr impostor attempts)
- Half Total Error Rate (HTER)
  - HTER = (FR + FA) /2
- Equal Error Rate (EER)
  - EER = FR = FA at an a posteriori determined threshold
  - Well defined measure, but cannot be selected in practice
- Detection Error Trade-off (DET)
  - Exhibits FR and FA at different thresholds
  - Similar to “Receiver Operating Characteristics” (ROC)

Detection Error Trade-off (DET)



From Melin (2006)

Application-dependent operating point



Bank transactions:
The FA cost is high
The customer can accept a few false rejects to achieve high security

Telephone call charges:
The FA cost is low
The customer can accept a few false accepts for high convenience

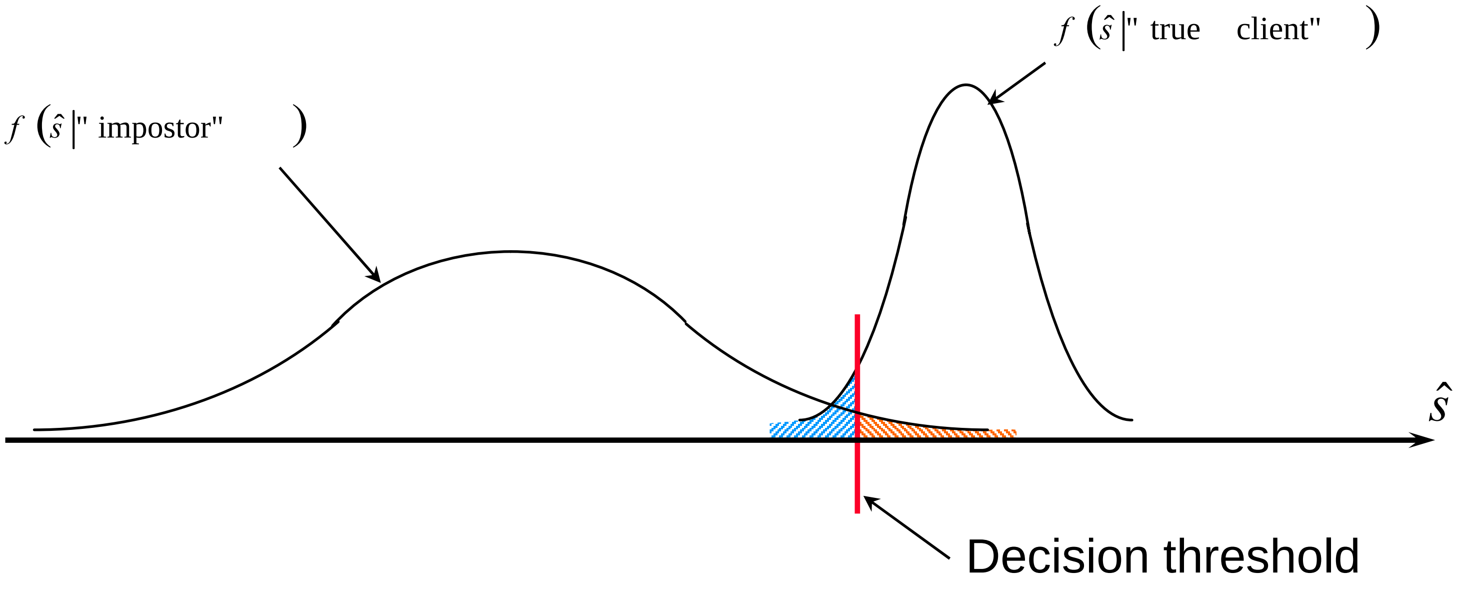
The appropriate operating point (balance FA/FR) depends on the costs of each error type

Score Normalization

- The score variability between trials makes it difficult to set the decision threshold.
  - The enrollment material may differ between speakers
    - Phonetic content, duration, noise, etc.
  - Mismatch between enrollment and test data
    - Client model mismatch
      - **Intra-speaker variability**
        - **Health, emotion, etc.**
        - **High variability during training lowers the score during test**
        - **Low variability during training increases the risk of mismatch during test**
      - **Environment condition changes**
        - **Transmission channel, acoustic environment, speech material**
    - Background model match
      - **The speaker’s position in the background model distribution affects the score**
        - **Speaker-independent threshold is not optimal but unavoidable**
        - **Normalization of the score value is required**
        - **Corresponds to threshold adjustment**

Score Normalization (2)

- Client score and Impostor score
  - The impostor score variability is the largest and its distribution is available
  - The client score distribution is rarely available
- Normalize the impostor score distribution



$f\left(\hat{s}\,|\,\text{" true client"}\right)$
$f\left(\hat{s}\,|\,\text{" impostor"}\right)$
$\hat{s}$
Decision threshold

Normalization techniques

- Acoustic features
  - CMS, feature variance normalization, feature warping, etc.
- Background model
  - Bayesian hypothesis test, likelihood ratio
  - Mismatch in both client and background models is cancelled
    - Example: environmental noise during test
- Normalization of the obtained score
  - Znorm – zero normalization
  - Hnorm – handset normalization
  - Tnorm – test-normalization
  - HTnorm – handset variation of Tnorm
  - Cnorm - Clustering
  - Dnorm – distance normalization
  - WMAP – MAP approach on likelihood ratio

Znorm

- Transform the score distributions for non-clients to zero average and unit variance, N(0,1)

$$\tilde{L}_\lambda(X) = \frac{L_\lambda(X) - \mu_{\lambda,\,\text{Non-clients}}}{\sigma_{\lambda,\,\text{Non-clients}}}$$

- The transformed score represents the number of standard deviations above the impostor average score. Same with the decision threshold.
- Assuming Gaussian distribution of the impostor scores, the False Accept Rate is directly defined by the decision threshold
- Estimation of speaker dependent average and variance can be done offline.

Hnorm

- Handset normalization to deal with handset mismatch between training and testing
- Important for telephone applications, different microphone types cause many errors
- Variant of Znorm
- Detect the type of handset used and apply its normalization
- μ_{CARB_MIC} and σ_{CARB_MIC} are estimates of the score distribution of non-client speech using the detected microphone type using the client model

$$\tilde{L}_{\lambda}(X) = \frac{L_{\lambda}(X) - \mu_{\lambda,CARB\_MIC}}{\sigma_{\lambda,CARB\_MIC}}$$

Tnorm

- Test-normalization matches the input utterance against a large number of non-client models. From these, the impostor mean and variances are estimated.
- The normalization equation is the same as in Znorm
- Avoids the possible acoustic mismatch between test and normalization utterances, since the test utterance itself is used for normalization. This is a problem in Znorm.
- The test utterance is matched against several models, which delays the decision.

HTnorm

- Combination of Hnorm and Tnorm
- Handset-dependent normalization parameters are estimated by testing the input utterance against handset-dependent impostor models
- During testing, the type of handset relating to the claimed speaker determines the normalization parameter

Cnorm

- Used when there are several unidentified handsets (as in mobile telephones)
- Blind clustering of the normalization data.
- Hnorm algorithm using each cluster as a different handset

Dnorm

- Takes client model into account
- If the client model is very dissimilar to the world model, one would expect large score difference between client and impostor utterances
- Normalize the score with the distance between the client and the world models.
- Generate client and pseudo-impostor data using the world model
- The distance is derived by comparing scores from client model scores on world model generated data with world model scores on client model generated data

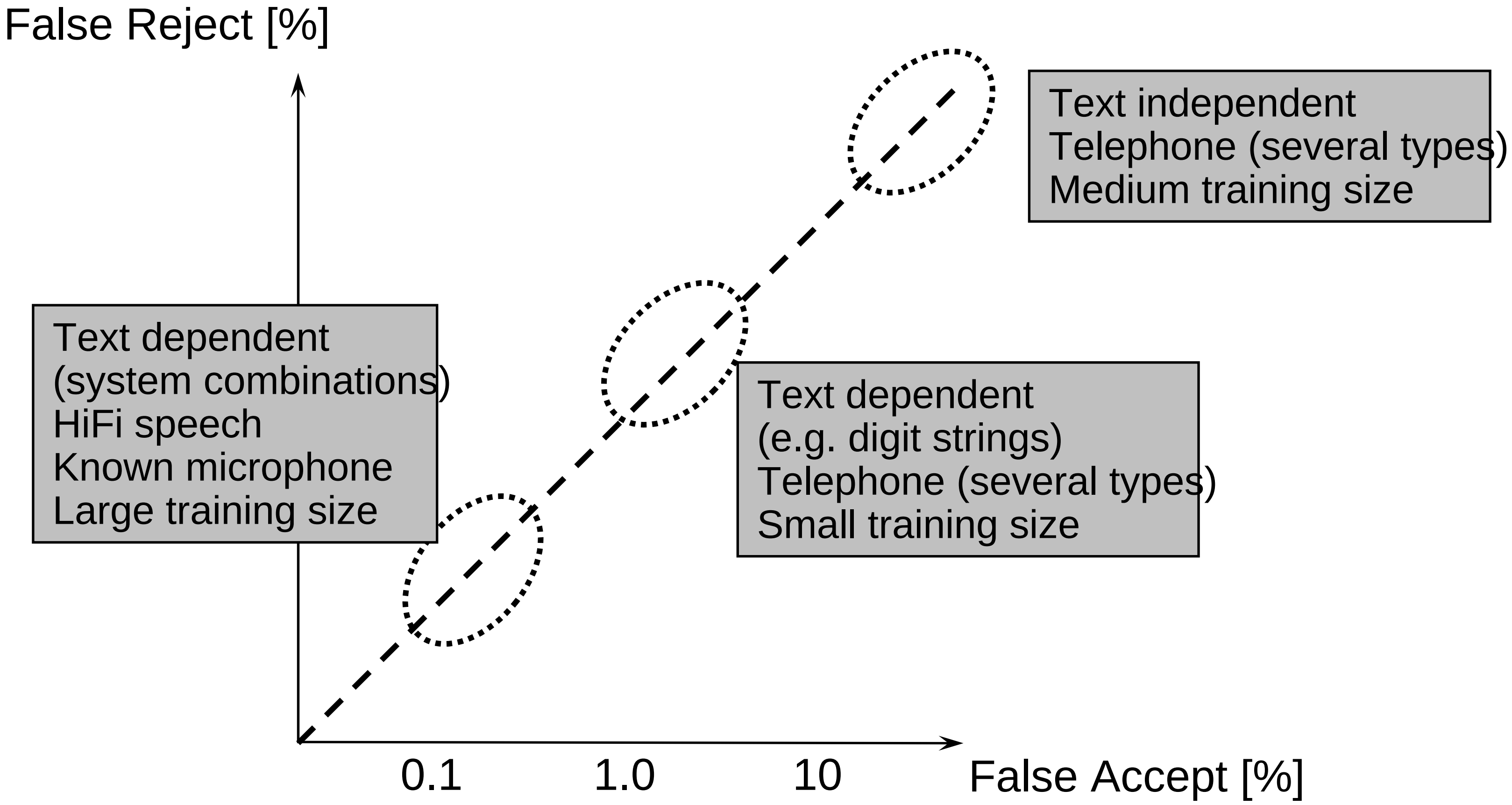
Normalization:

$$\tilde{L}_\lambda(X) = \frac{L_\lambda(X)}{ModelDist(Client, World)} = \frac{L_\lambda(X)}{KL2(\lambda, \bar{\lambda})}$$

Normalization summary

- HTnorm seems better than the other
  - Unfortunately, most computationally expensive
- Combination of normalization techniques improves the performance
  - Especially when combining ”learning condition” normalization with ”test-based normalization”

Performance in different applications



Text independent
Telephone (several types)
Medium training size

Text dependent
(system combinations)
HiFi speech
Known microphone
Large training size

Text dependent
(e.g. digit strings)
Telephone (several types)
Small training size

Key problems

- **Optimisation**
  - Minimal error rate with minimal amount of enrolment speech
  - Determine the decision threshold
  - Improve/Combine features and methods
  - Improved resistance against (human or technical) impersonation
- **User acceptance:**
  - How to design an application
- **Evaluation:**
  - how to measure the technical performance?
  - how to estimate the rate of impostor attempts

Security aspects

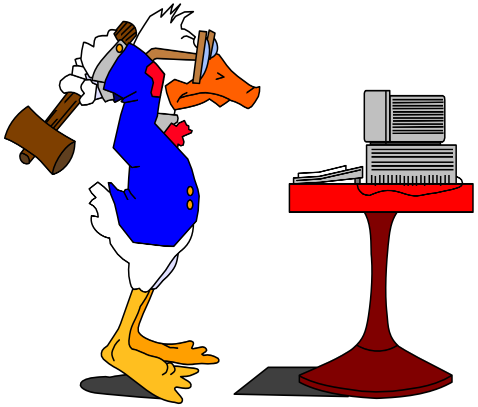
- Performance is measured using casual impostors
- What is the immunity against real impostor attempts?
  - Imitations? Recordings? “Personal” speech synthesis?
- The security of conventional systems can be raised by combination with voice
  - E.g. protection if credit card + PIN code is stolen
- Preventive effect by
  - Recordings can be saved for later manual control

Simple combination of methods

- How to combine speaker verification and PIN-code?
- PIN-code:
  - Assume FR = 1% (Estimated frequency of forgetting the code)
  - FA = 0.01% if the code is unknown (1 out of 10 000 combinations)
  - If the code is known: FA = 100%
- Perform speaker verification if the PIN-code is correct
- Should not increase FR => FR(voice) should be <= 1%
  - If used once a week, False Reject occurs once every two years
- The corresponding FA can be picked from a DET-diagram
  - Using PER results: FA ~= 20% Prevents 80% of impostor attempts

User aspects

- As little training as possible, preferrably nothing
  - The speaker’s variability cannot be measured
- Speaker verification should simplify for the user, preferrably transparent
- Door guard or warning bell?
- What balance FA / FR?
  - Depends on the security demands and the costs
  - True clients should not be disturbed



“The animal park"

“Categorisation" of a speaker 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

Impostor - who?

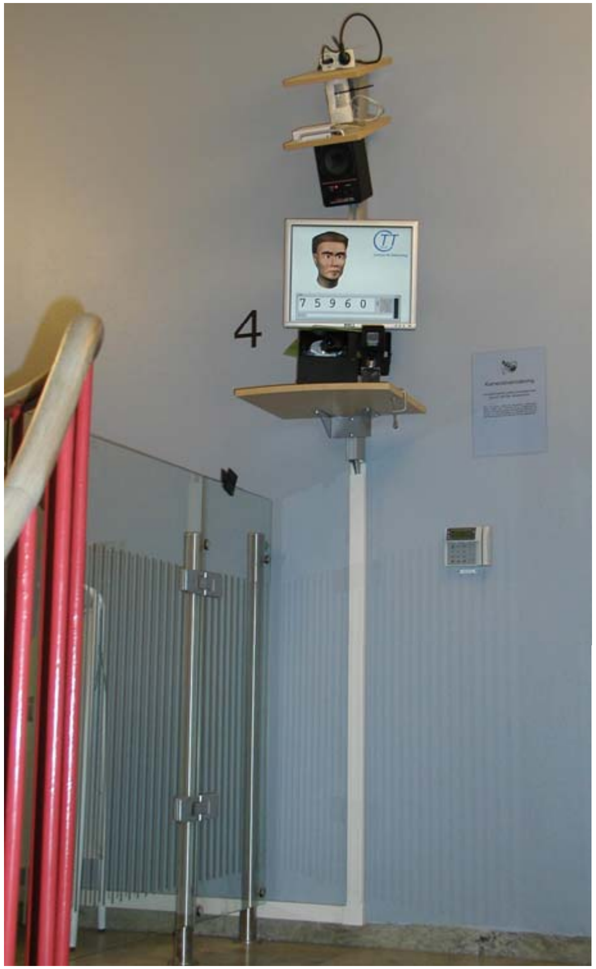
- Risk factors:
  - May know a password
  - Has recordings
  - Can buy information
  - Impersonation
  - Family member, twin
  - How much damage can be done?
- Professional and naïve impersonation
- Technical impostors

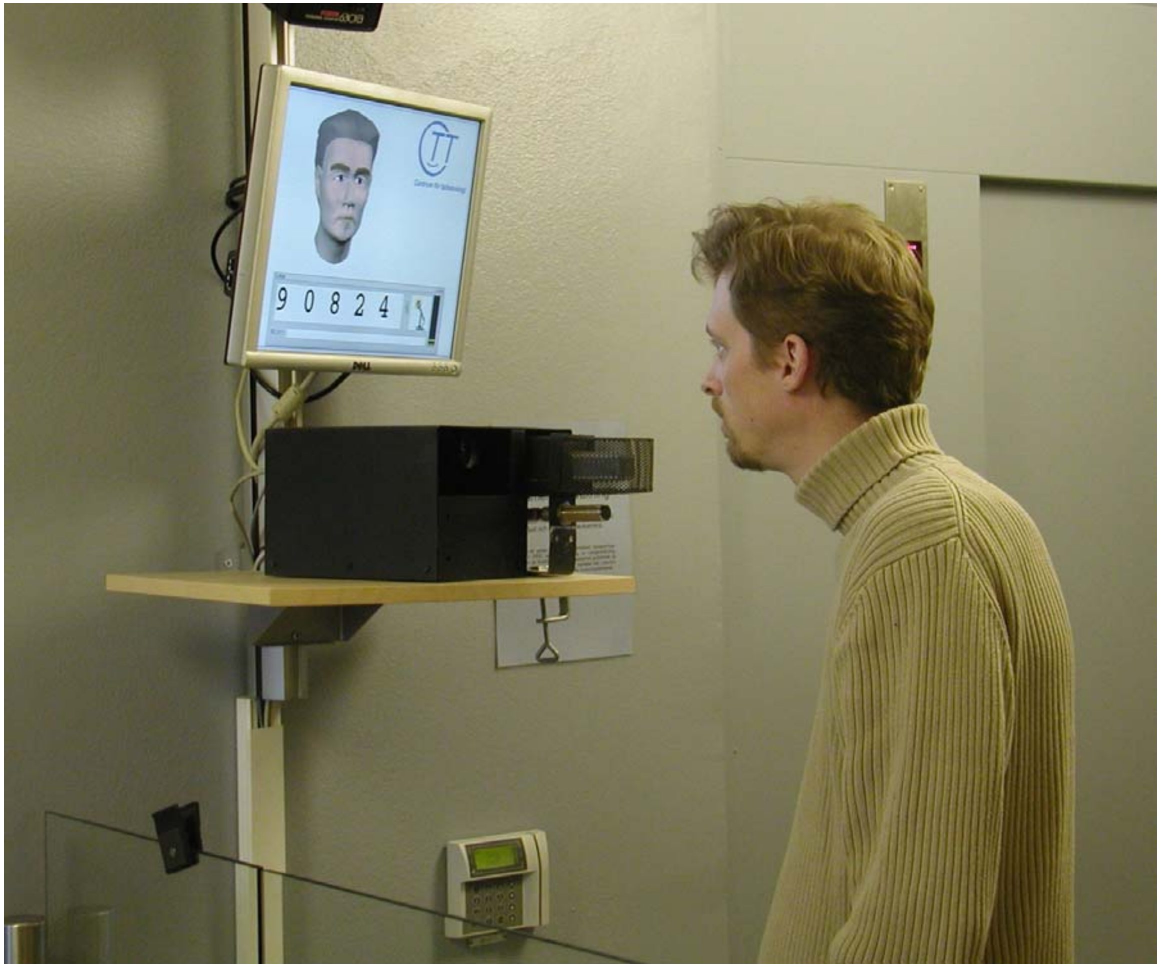
The CTT project PER (Prototype Entrance Receptionist)

- Visually detects the presence of a person at the TMH entrance
- Identifies personnel using speaker verification and unlocks the gate
  - Say your name and a prompted digit sequence
  - Animated talking face
- Combined HMM and GMM system
  - Comparable performance with commercial system
- In practical use since 1998
- Developer Håkan Melin defended his PhD in 2006

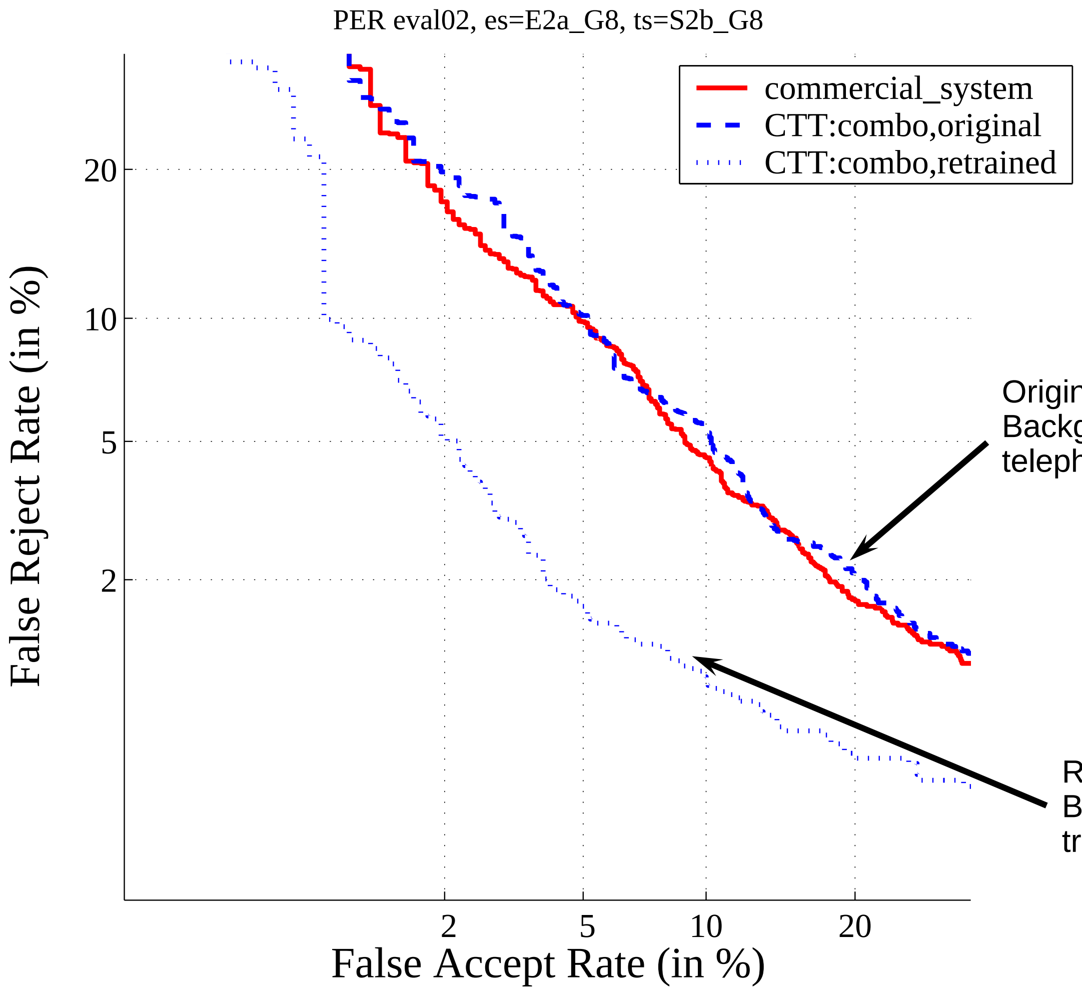
PER at the TMH entrance

Subject: the developer Håkan Melin





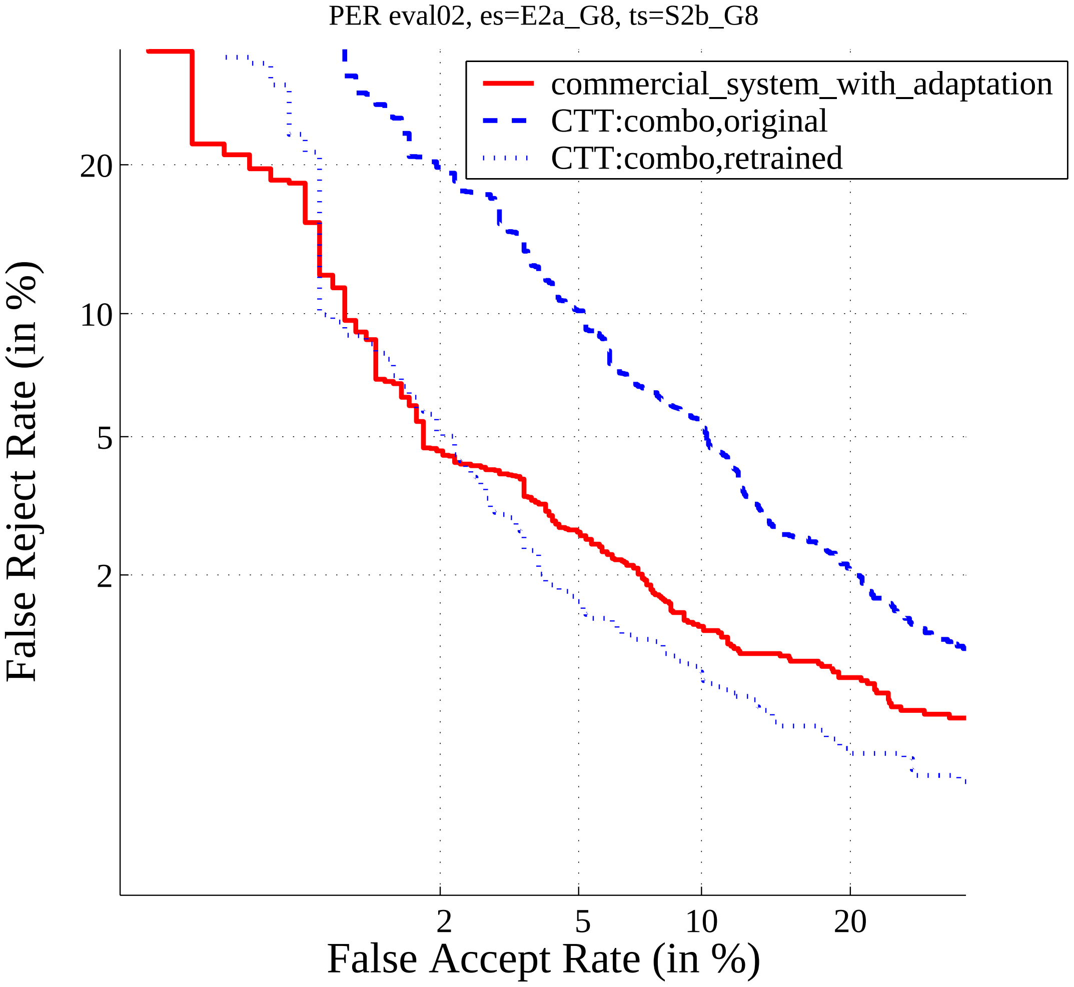
PER and unadapted commercial system



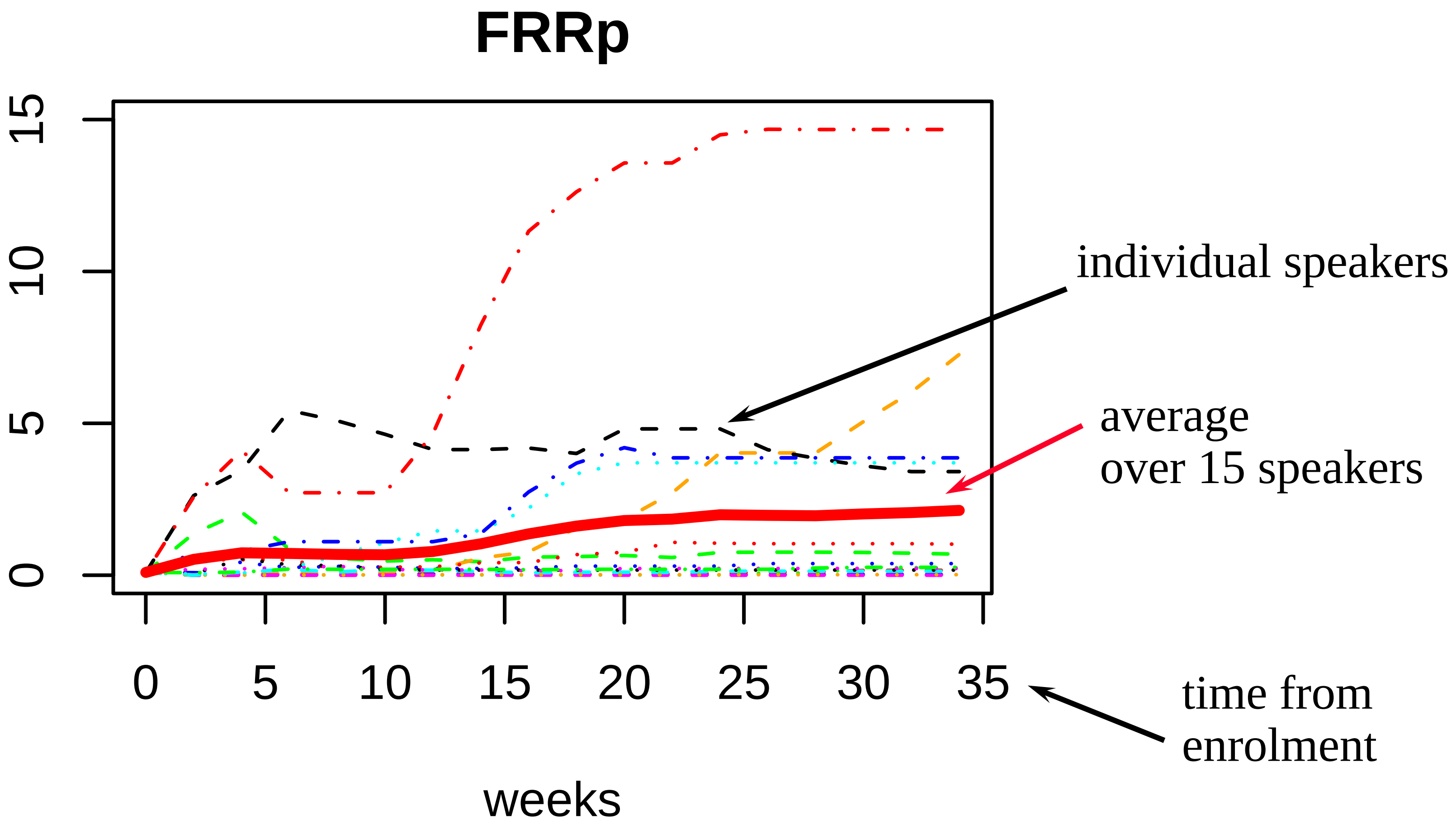
Original:
Background model trained on telephone speech

Retrained:
Background model trained on PER speech

PER and adapted commercial system



PER: estimated individual speaker FRR



individual speakers

average
over 15 speakers

time from
enrolment

Impersonation

- Study by Daniel Elenius (Master Thesis project, 2001)
- Naive subjects and one professional
- Imitation of a similar speaker (training by listening) increased the FAR from 18% to 48%. Combination with score feedback from the SV system during training increased FAR further to 57%
- Imitation by listening of a speaker with average similarity increased FAR from 0% to 17%. No improvement by score feedback.
- Conclusion
  - Unacceptable risk of False Accept using imitation of a similar target voice.
  - Recommended precaution
    - Combine with other methods. Use voice for extra security
  - No case has happened (to my knowledge)

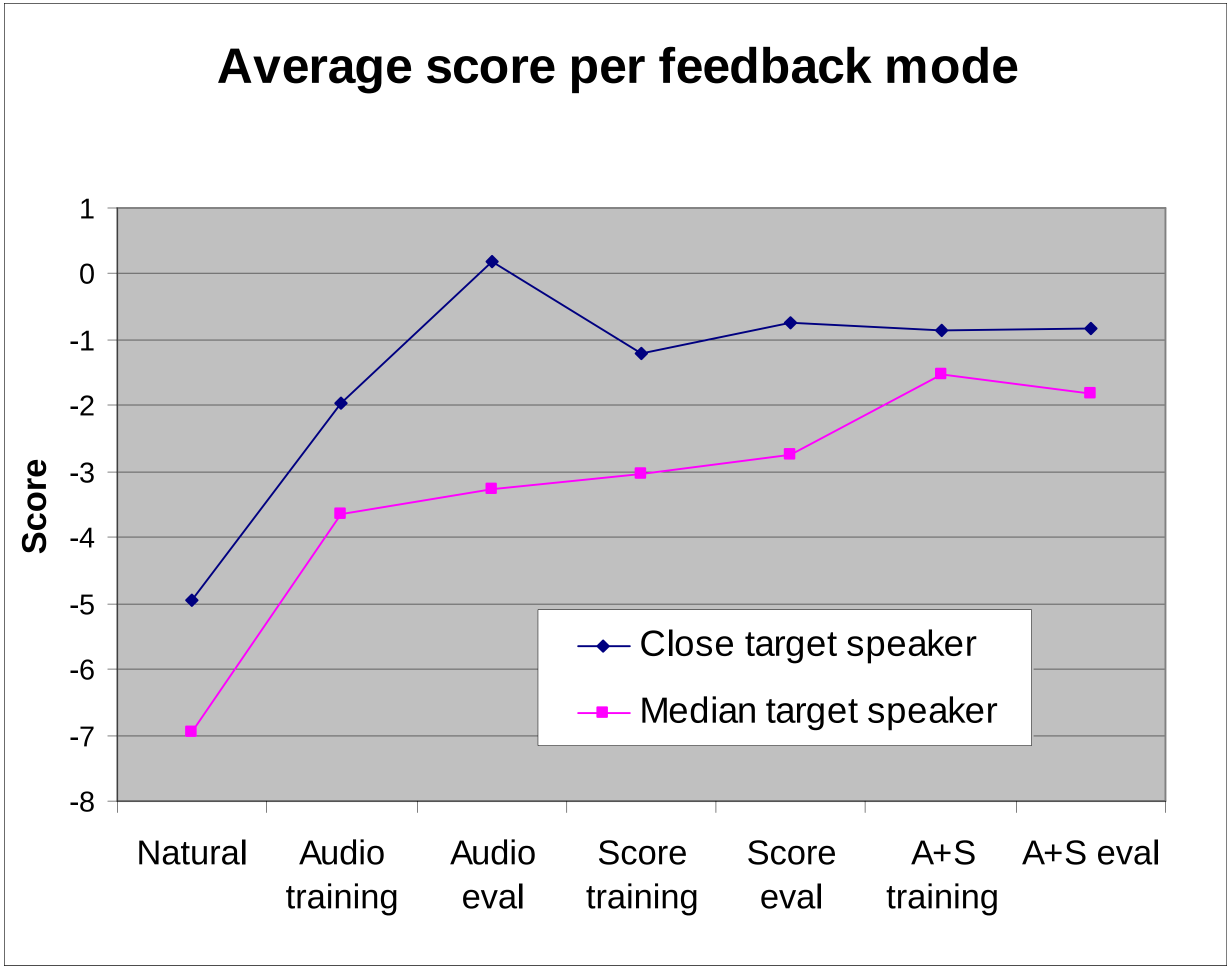
Speaker verification scores and acoustic analysis of a professional impersonator

Elisabeth Zetterholm
Mats Blomberg
Daniel Elenius

Australian Int. Conf. on Speech Science & Technology, 2004

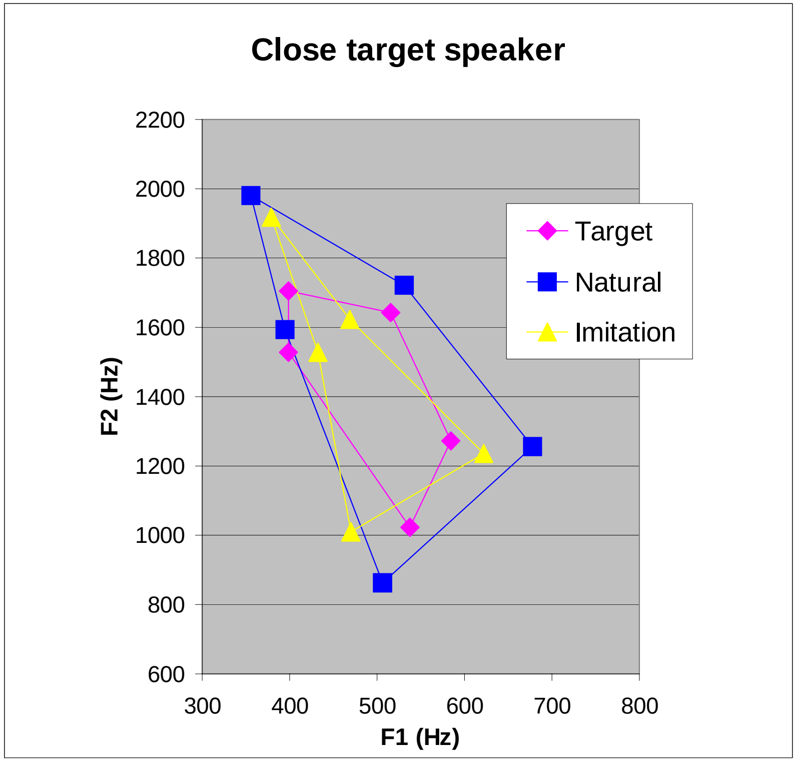
Effect of Imitation Training

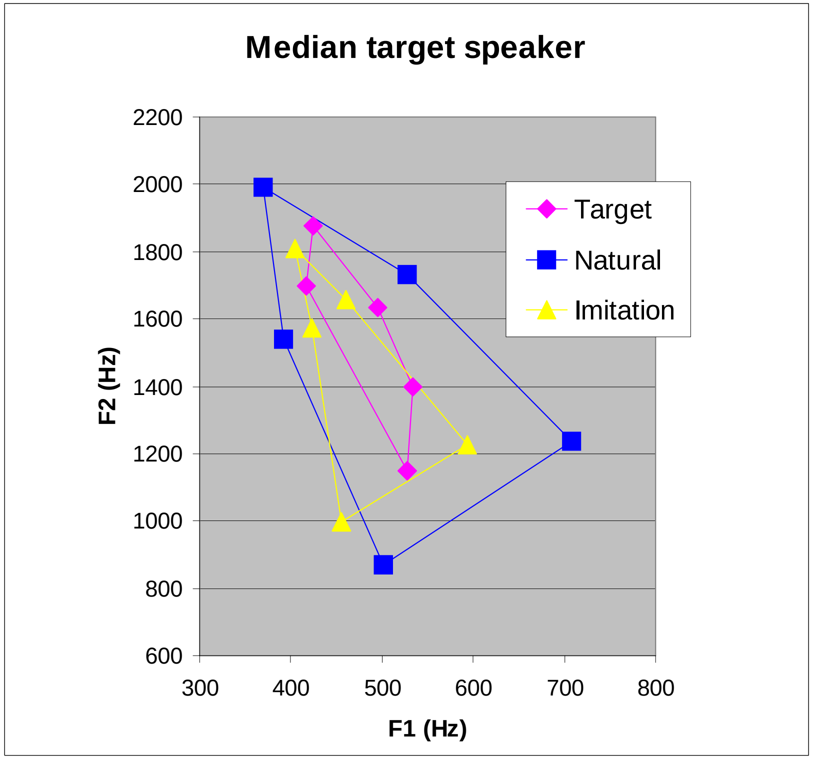
Professional impersonator



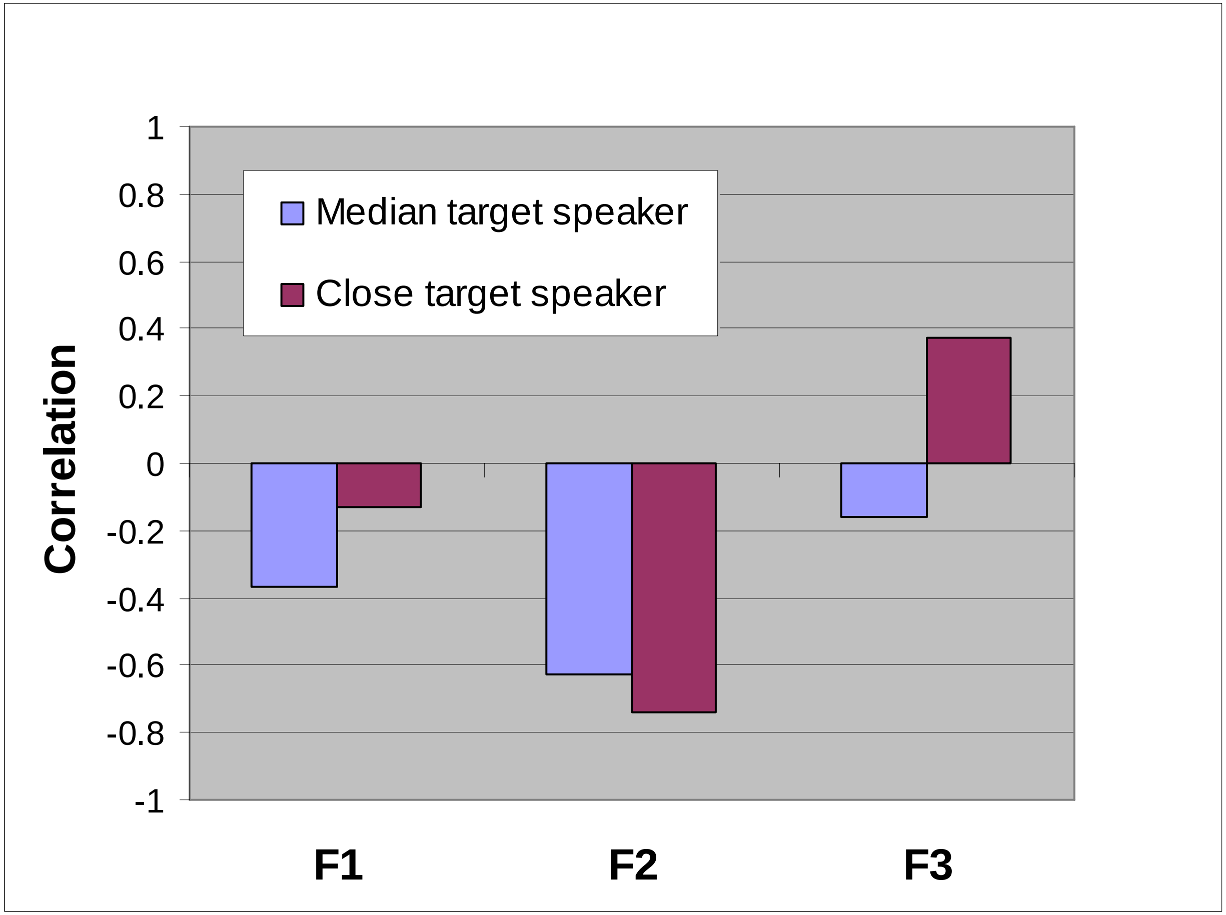
Vowel formants

Professional impersonator

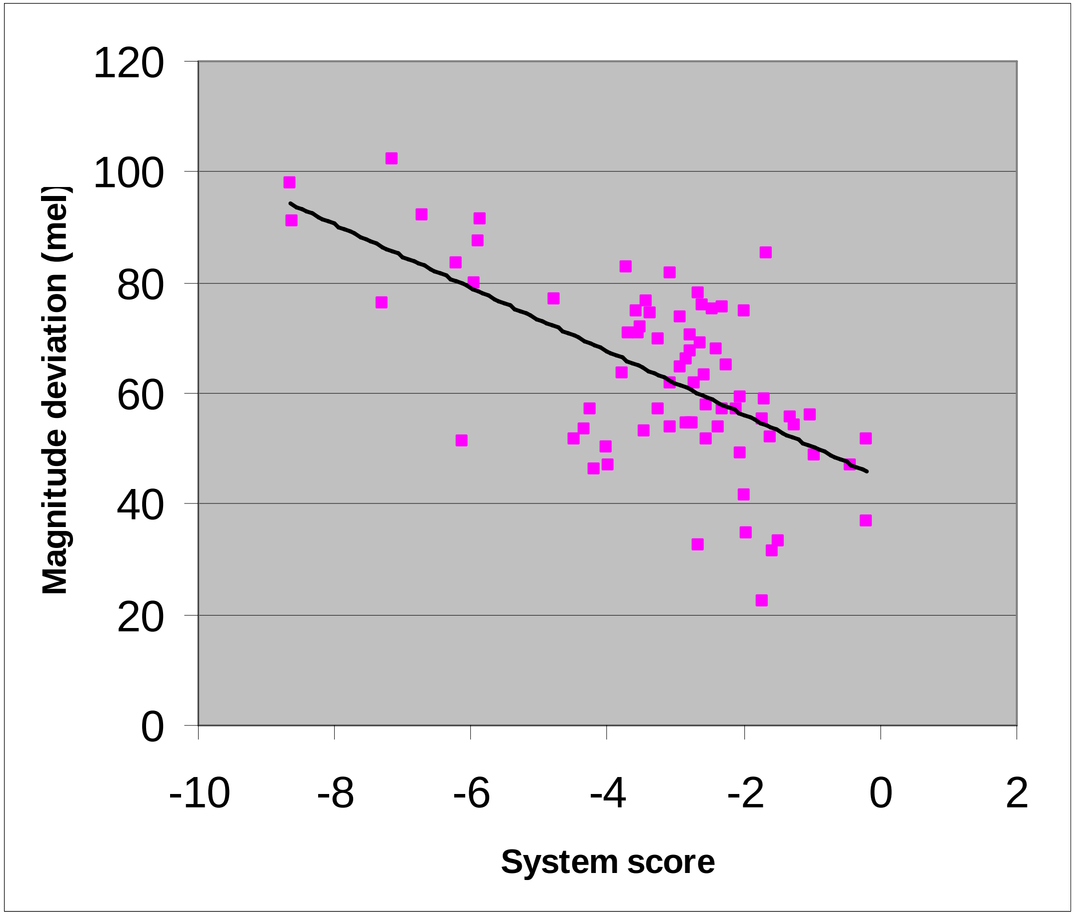




Formant deviation correlation

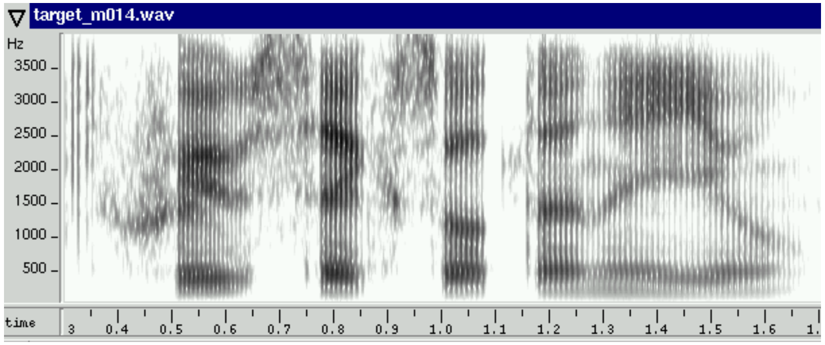


F2 deviation vs score

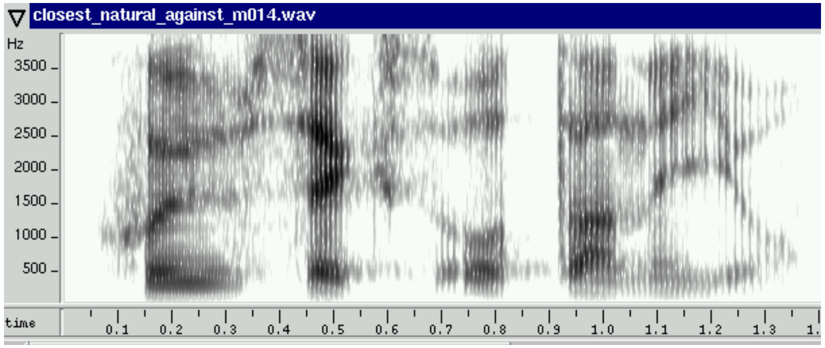


Median target speaker

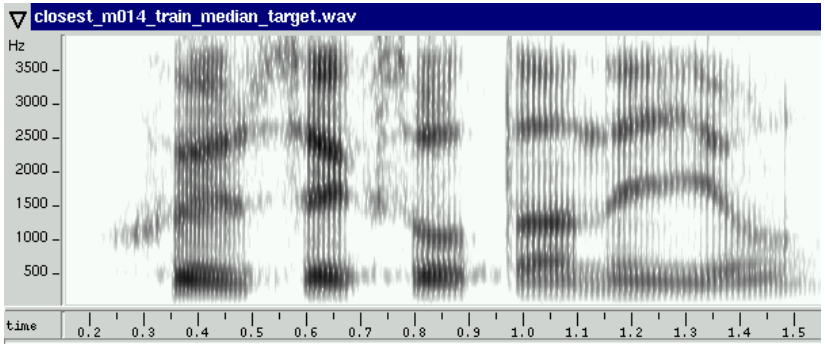
Target speaker


target_m014.wav

Closest natural utterance


closest_natural_against_m014.wav

Closest mimic utterance


closest_m014_train_median_target.wav

Technical impostors

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

- Real problem for speaker verification

- Preventive techniques
  - Detect artificial character
  - Deterministic (repetitions of the same text are identical)

Extensions of Speaker Verification

- SV assumes that training and test recordings are monospeaker recordings
- Some applications need to detect the presence of a given speaker in a multispeaker recording
    - n-speaker detection: Is a target speaker present in a conversation within a group of speakers
    - Speaker tracking: n-speaker detection plus time positioning
    - Speaker segmentation: Determine the number of speakers and when they speak

Forensic Speaker Recognition

- Difficulties not present in speaker verification
  - Unknown and uncontrollable (realistic) recording conditions
  - Impose high degree of variability
    - Peculiar inter- and intraspeaker variability
      - **Type of speech, gender, time separation, age, dialect, sociolect, jargon, emotional state, use of narcotics, etc.**
    - Forced intraspeaker variability
      - **Lombard effect, stress, cocktail-party effect**
    - Channel-dependent variability
      - **Type of handset, landline/mobile, channel, bandwidth, electrical and acoustical noise, reverberation, distortion, etc.**
  - Incooperative speakers
    - Rather the opposite, trying to disguise his/her voice

Incorrect usage of probabilistics in forensic speaker recognition

- Consider example:
- The expert:
  - The probability that another person than the suspect has the features of the recorded utterance is 1%
- Prosecutor:
  - Then there is 99% probability that the suspect is guilty
- Defense:
  - In this city with 100 000 citizens, there are 0.01*100 000 = 1000 persons with these features. Accordingly, the probability that the defendant is guilty is 1/1000.
- Who is correct?
  - Neither
- Use Bayesian framework
  - Since the a priori probability is often unknown, it is only possible to say how the likelihood ratio between two hypotheses is changed upon the analysis of the utterance, not the absolute value of one hypothesis

Automatic forensic speaker recognition

- Semi-automatic systems
  - Several have been developed
  - Require use by expert phoneticians
  - Lacks generalization
- Automatic systems
  - ”Appears to have reached a sufficient level” (Nakasone & Beck, 2001)
  - Produces binary decisions
  - Its performance can be evaluated
- Forensic methodology aspects
  - Non-zero error rate – how should it be used?
  - Should a decision be made? Then what about the jury?
  - How to take prior probabilities (circumstances) into account?
  - How to quantify the cost of errors (innocents convicted and guilty freed)?

Forensic speaker recognition: Conclusion

- Systems for commercial use need to be modified
  - Assign a confidence measure of binary decisions
  - Bayesian approach to include prior probability (circumstances related with evidence)

- ”Automatic speaker recognition systems constitute a milestone in forensic speaker recognition”
(Bimbot et al 2004)

- Remaining unsolved issues
  - Real forensic speech databases
  - Evaluation methodology
  - Role of the expert