Lecture

Speech and Audio Signal Processing

Lecture 10: Speaker recognition

TECHNISCHE UNIVERSITÄT DARMSTADT

Content

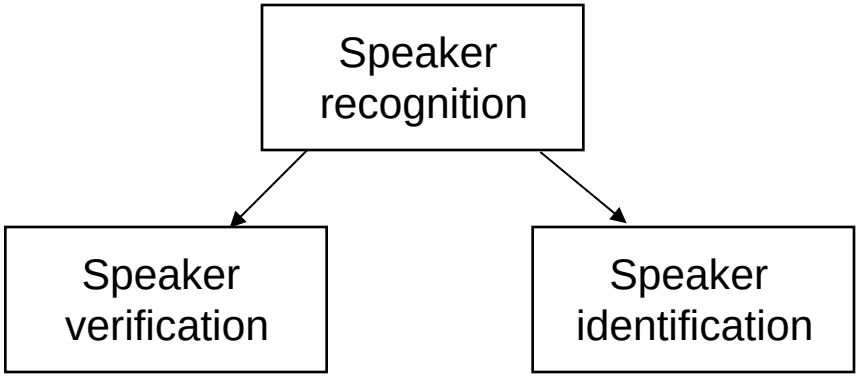
TECHNISCHE UNIVERSITÄT DARMSTADT

- Applications of speaker recognition
- Types of speaker recognition
  - Speaker verification
  - Speaker identification
- Preprocessing and feature extraction
- Speaker recognition based on GMMs
- Speaker verification => error analysis
- Speaker identification => a practical example
- Model adaptation
- Introduction to speech recognition: Generals about Hidden Markov Models (HMMs).

Applications

- Access control (Access to buildings, start cars, use electronic devices, etc. => Alternative to other bionic access control (fingerprint, face, retina, etc.).
- Telephone based applications, e.g., recognition of preferred speaker with individualized services.
- Improving speech enhancement methods, e.g., speaker specific bandwidth extension or focusing a beamformer to a preferred speaker.
- Recognition of speakers in a conference call and appropriate control of speaker focusing beamformers.
- Individualized training for speech recognition systems used by several speakers.
- Hearing aids: Detect own voice: Control gain settings and beamformer adaptation.

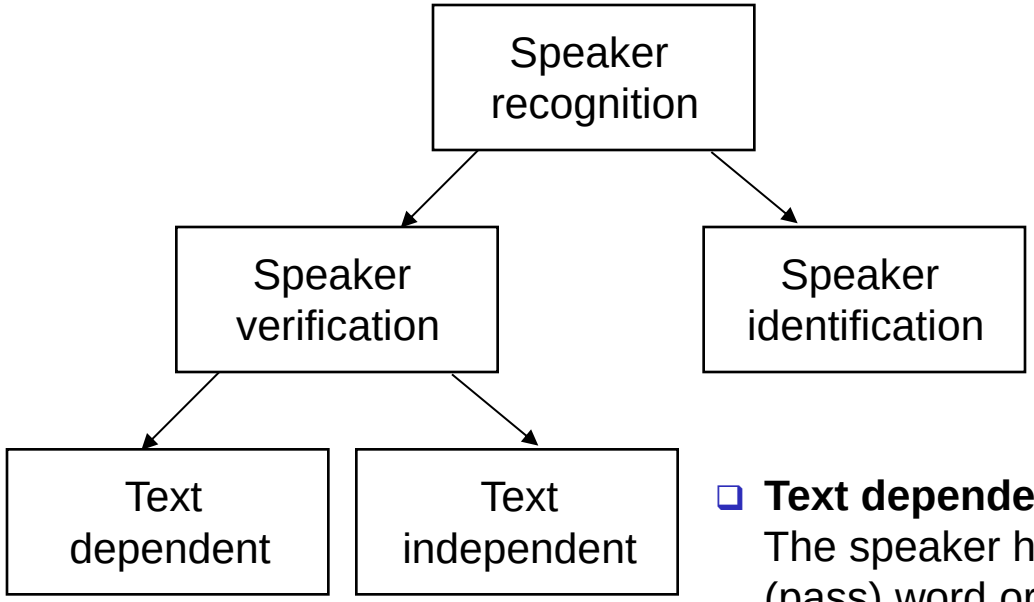
Types of speaker recognition



- **Speaker verification:**
  Binary decision: Is this the voice of a claimed speaker?

- **Speaker identification:**
  1 of N decision: Which of N speakers is active?

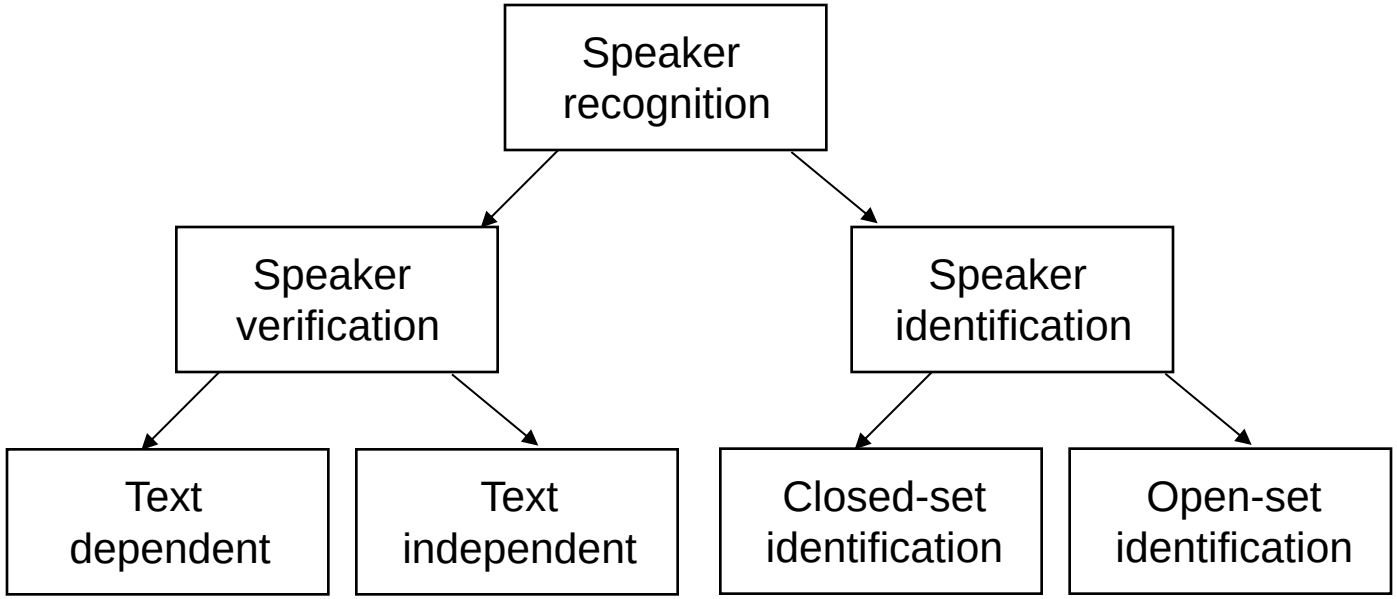
Types of speaker recognition



- **Text dependent:**
  The speaker has to pronounce a known (pass) word or text

- **Text independent :**
  The speaker pronounces an unknown text. Typical application in a hidden context.

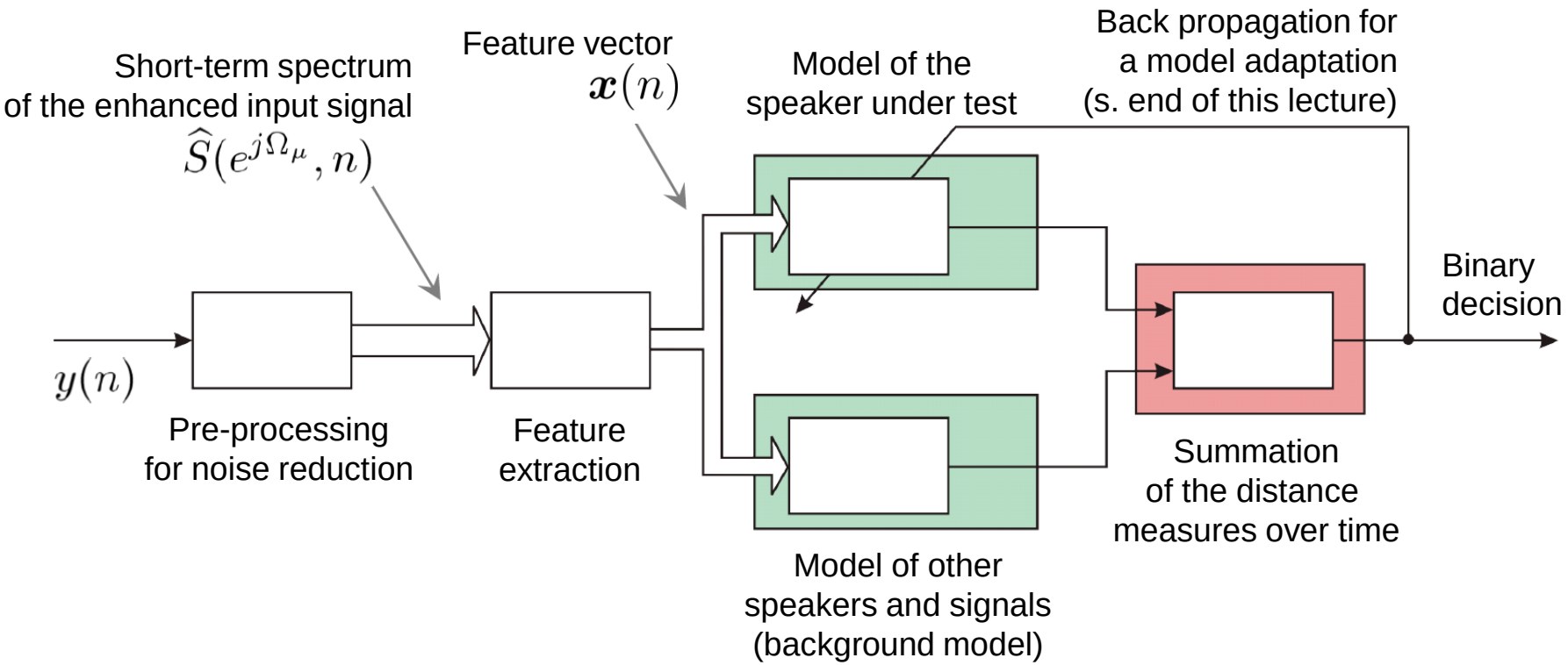
Types of speaker recognition



- **Closed-set:**
  A set of known speakers
- **Open-set:**.
  Unknown speakers, also the number of speakers can be unknown. An initial set of speakers can be known at the start which may increase during operation.

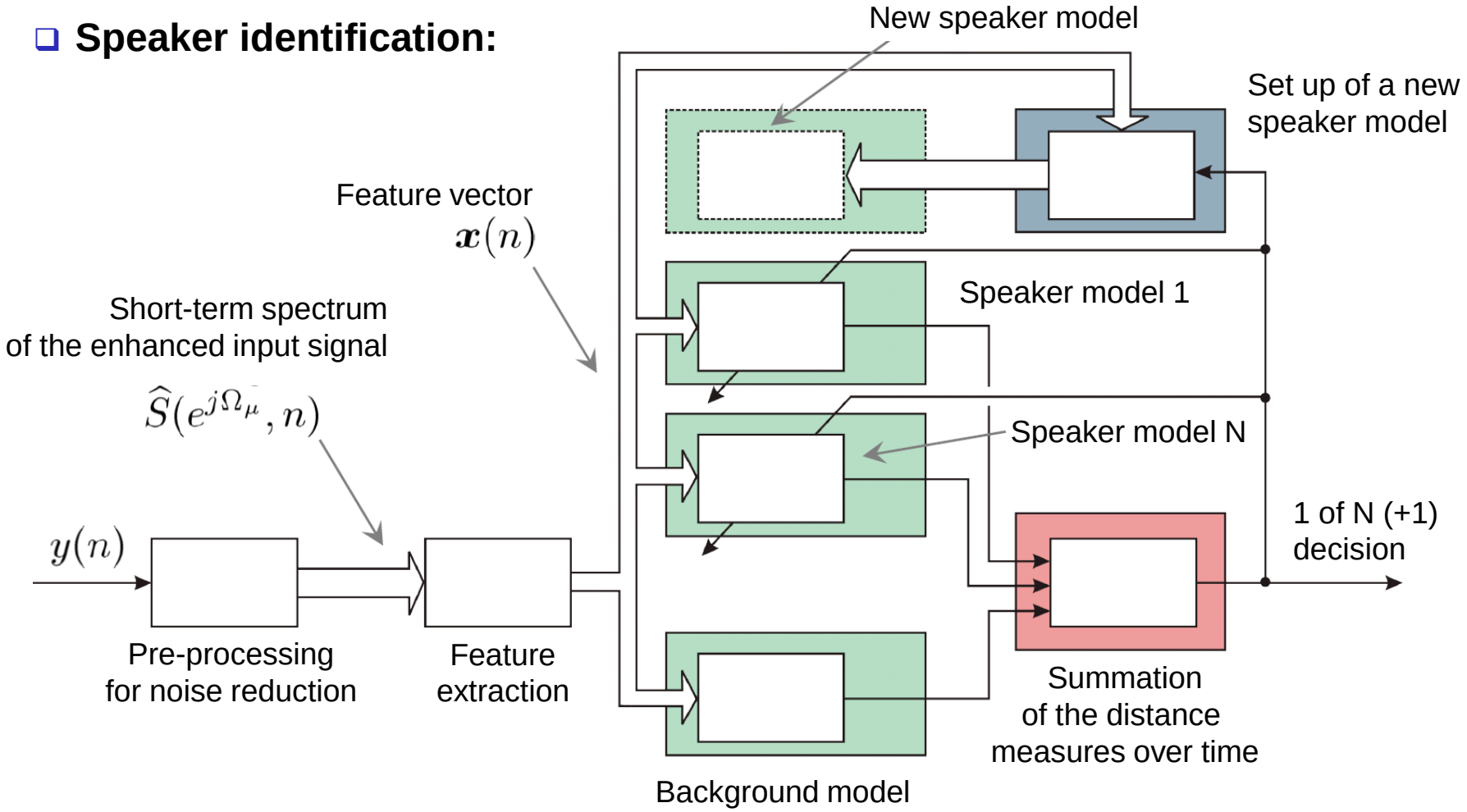
Speaker **verification** – basic structure

- **Speaker verification:**



Speaker **identification** – basic structure

- **Speaker identification:**



Pre-processing

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Pre-processing I:**
  Reduction of interfering noise components: Two alternatives:
  - 1) Use the known Wiener filter approach:

$$\widehat{H}_{\text{opt}}(e^{j\Omega}, n) \quad = \quad \max\left\{1 - K_{\text{over}} \frac{\widehat{S}_{bb}(\Omega, n)}{\widehat{S}_{yy}(\Omega, n)},\ H_{\text{min}}\right\}$$

  - 2) Apply a “cepstral mean subtraction (CMS)”:

    - CMS concept: Perform a long-term smoothing of the cepstral coefficients or MFCCs, respectively. → Reduce effect of reverberant room!

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Pre-processing II:**
  Segmentation of speech components: Two alternatives:
  - 1) Apply a segmentation or voice activity detection based on the un-constrained Wiener filter:

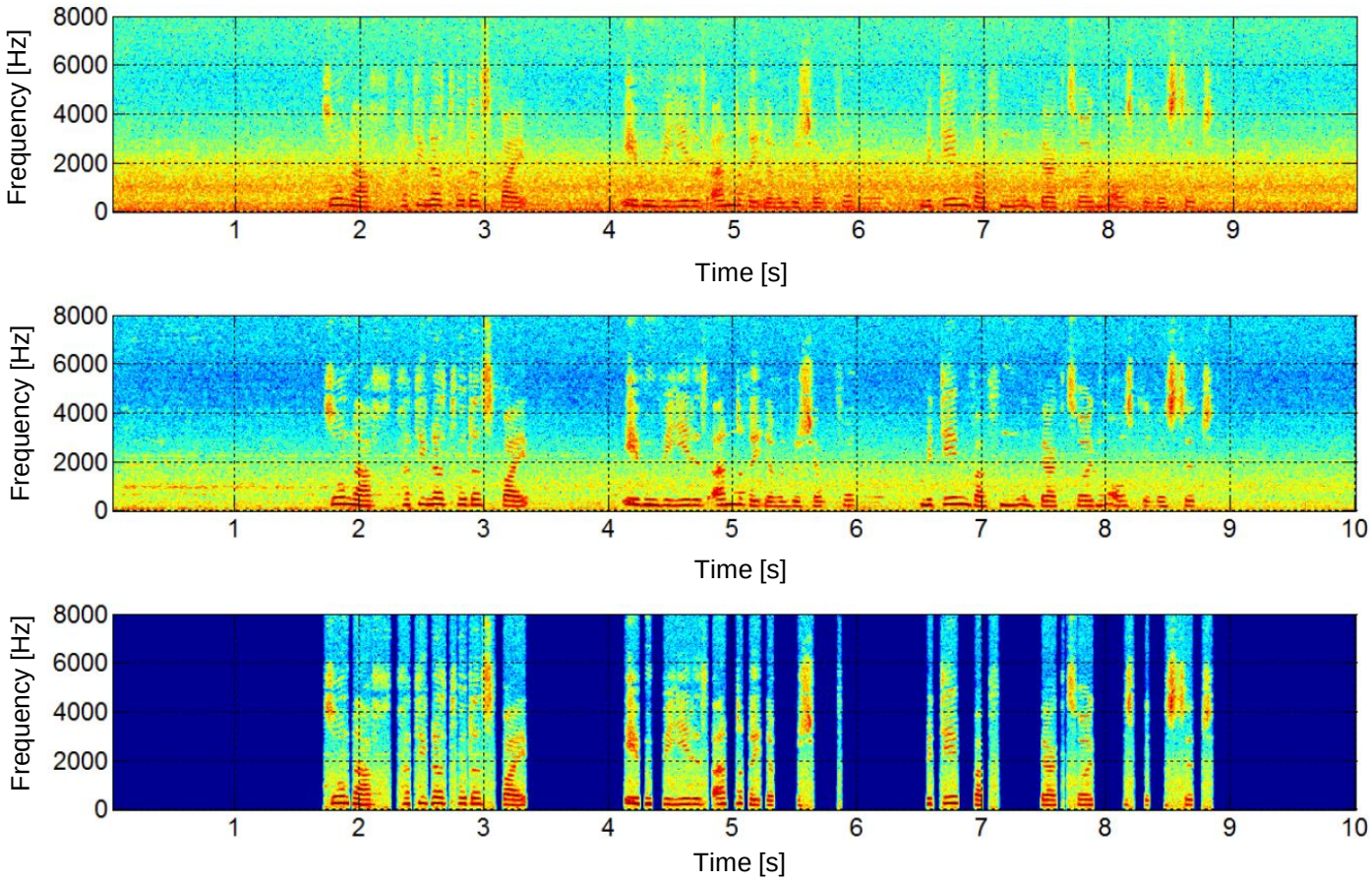
$$\widehat{H}_{\text{uc}}(e^{j\Omega}, n) \quad = \quad \max\left\{1 - K_{\text{over}} \frac{\widehat{S}_{bb}(\Omega, n)}{\widehat{S}_{yy}(\Omega, n)},\, 0\right\}$$

In case in 10% to 30% of the spectral coefficients open the filter the corresponding time frame is considered as active:

$$\text{seg}(n) \;=\; \begin{cases} 1 & : \quad \text{if} \quad \frac{1}{N} \sum_{\mu=0}^{N-1} \widehat{H}_{\text{uc}}(e^{j\Omega_\mu}, n) \;>\; 0.1 \cdots 0.3 \\ 0 & : \quad \text{else} \end{cases}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Pre-processing II: Segmentation of speech components – Example:**



- **Input signal:**
- **Signal after noise reduction**
- **Signal after segmentation**

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Pre-processing II:**
  Segmentation of speech components

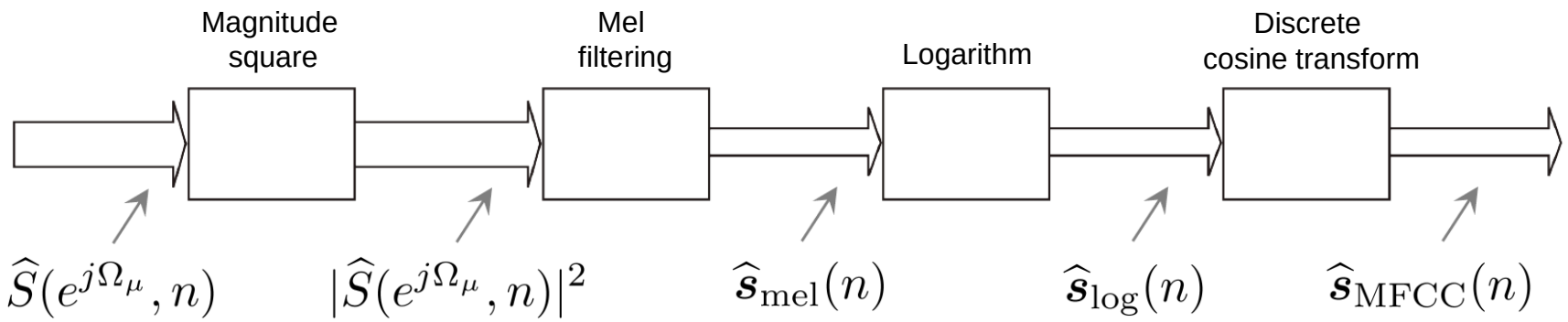
  - 2) Use a bi-Gaussian model classifier:

    Train a GMM with two Gaussians based on the cepstral feature vectors or the log energy.

    Signal frames assigned to the Gaussian with lower values are disregarded for the model training and during the evaluation steps.

Feature extraction

- Mel-filtered cepstral coefficients:



  - Typically, a power normalization is performed by setting the cepstral coefficient with index “0” to a normalized value.

  - Also, usually the Delta and Delta-Delta feature vectors are added.

MFCC: application for speaker and speech recognition

TECHNISCHE UNIVERSITÄT DARMSTADT

- The MFCC values have shown to be useful for speaker and speech recognition.
- It is remarkable that this is the case since these applications have different targets:
  - Speech recognition: Detect speech independent of the speaker. The stronger a speaker dependency can be removed the better for the recognition.
  - Speaker recognition: Here especially speaker dependent features are of interest.
- Early speaker recognition systems therefore performed a speech recognition based on phoneme detection and then evaluated speaker dependent features for these phonemes, separately.
- However, in comparison to phoneme independent speaker recognition no real benefit, but a much higher computational complexity was observed.
- Concluding, MFCCs contain speaker (and speech) dependent information. A speaker adaptation of speech recognition systems typically increases the recognition rates.

Background models

TECHNISCHE UNIVERSITÄT DARMSTADT

- The background model which is necessary for speaker verification can be designed according to two ways:

- 1) Train several models of alternative speakers and verify the target speaker against all those models.

- 2) Use a pool of speakers and train one single large GMM based on those signals. Here, also non-speech signals can be contained in order to design a robust method applicable for non-ideal voice activity detection.

  The second method has shown advantages and is therefore the currently preferred procedure.

Speaker verification

- 1) GMM training:

  Here the GMMs of the speaker(s) to verify are trained as well as background models.

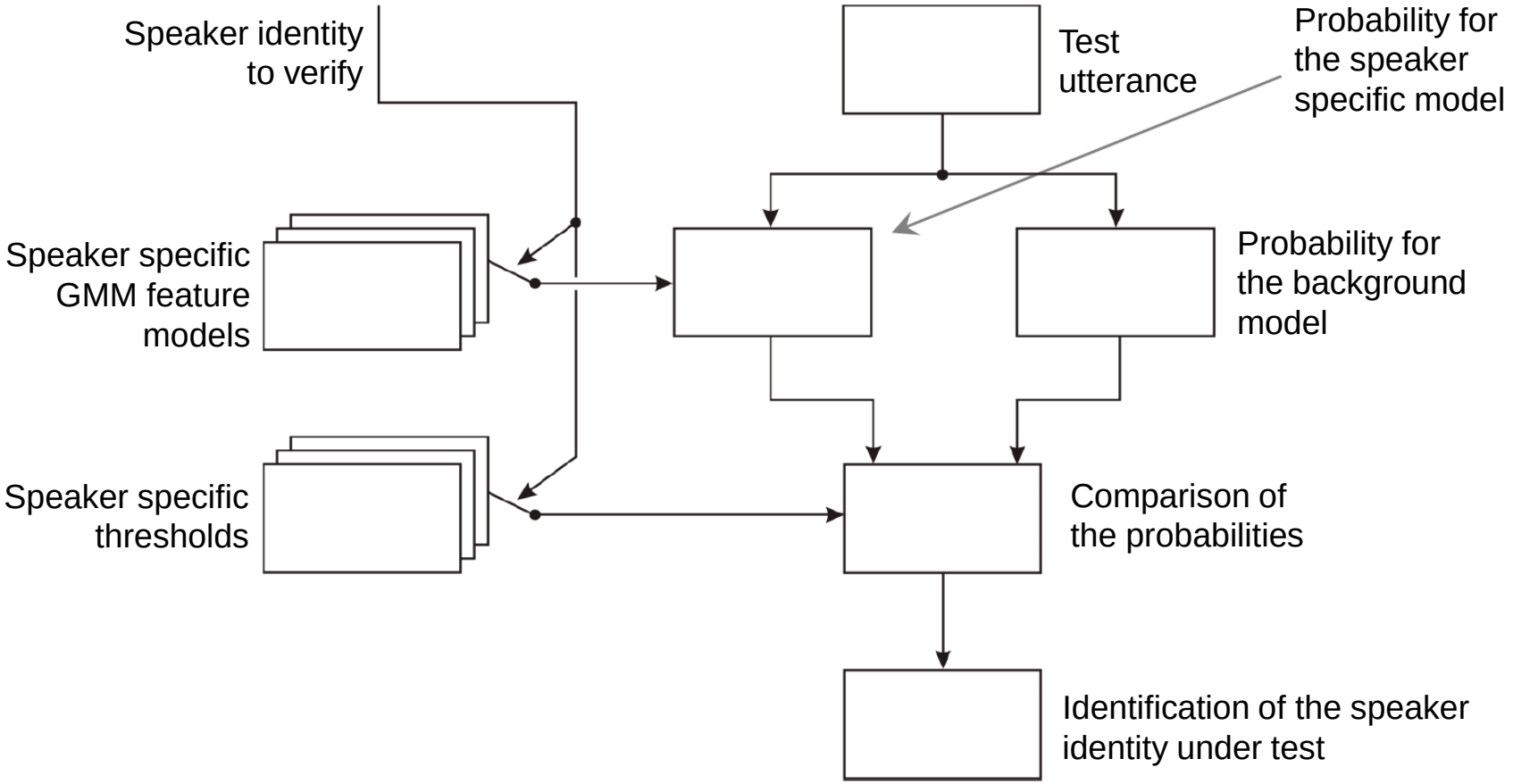
  Training data for the speakers of 30 sec to 3 min. is typically used and sufficient => typical model order: 64 – 256 Gaussians.

  For the background model large speech and audio data bases can be used => typical model order: 512 – 2048 Gaussians.

  The GMMs are trained based on the iterative method shown in the last lecture. => typical: 5 – 10 iteration steps.

Speaker verification

- 2) Verification:



Speaker verification

TECHNISCHE UNIVERSITÄT DARMSTADT

- Bayes detection approach:
  - Two hypothesis:

    H_0 : Activity of the verified speaker

    H_1 : No activity of the verified speaker

  - Detect for the speaker to verify based on the higher a posteriori probability:

    $$p(H_0|\boldsymbol{X}) > p(H_1|\boldsymbol{X})$$

    with the Bayes relation:

    $$p(H_i|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|H_i)\,p(H_i)}{p(\boldsymbol{X})}$$

  - the decision criterion can be formulated as

    $$p(\boldsymbol{X}|H_0)\,p(H_0) > p(\boldsymbol{X}|H_1)\,p(H_1)$$

    with the feature vector sequence:

    $$\boldsymbol{X} = [\boldsymbol{x}(0), \boldsymbol{x}(1), \cdots, \boldsymbol{x}(N-1)]$$

Speaker verification

TECHNISCHE UNIVERSITÄT DARMSTADT

- Feature vector sequence:

$$\boldsymbol{X} = [\boldsymbol{x}(0), \boldsymbol{x}(1), \cdots, \boldsymbol{x}(N-1)]$$

- Product probability:

$$p(\boldsymbol{X}|H_i) = \prod_{n=0}^{N-1} p(\boldsymbol{x}(n)|H_i)$$

- Feature vector sequence,
  with (s) indicating the speaker and (b) the background model:

$$\begin{aligned}
\log p(\boldsymbol{X}|H_0) &= \log p(\boldsymbol{X}|\boldsymbol{g}^{(s)}, \boldsymbol{\mu}^{(s)}, \boldsymbol{\Sigma}^{(s)}) \\
&= \sum_{n=0}^{N-1} \log \left\{ \sum_{k=0}^{K-1} g_k^{(s)} \mathcal{N}(\boldsymbol{x}(n)|\boldsymbol{\mu}_k^{(s)}, \boldsymbol{\Sigma}_k^{(s)}) \right\} \\
\log p(\boldsymbol{X}|H_1) &= \log p(\boldsymbol{X}|\boldsymbol{g}^{(b)}, \boldsymbol{\mu}^{(b)}, \boldsymbol{\Sigma}^{(b)}) \\
&= \sum_{n=0}^{N-1} \log \left\{ \sum_{k=0}^{K-1} g_k^{(b)} \mathcal{N}(\boldsymbol{x}(n)|\boldsymbol{\mu}_k^{(b)}, \boldsymbol{\Sigma}_k^{(b)}) \right\}
\end{aligned}$$

Speaker verification

- Detection criterion

$$p(\boldsymbol{X}|H_0)\,p(H_0) \;>\; p(\boldsymbol{X}|H_1)\,p(H_1)$$

- can be written as:

$$\sum_{n=0}^{N-1} \log \left\{ \sum_{k=0}^{K-1} g_k^{(s)}\, \mathcal{N}(\boldsymbol{x}(n)|\boldsymbol{\mu}_k^{(s)}, \boldsymbol{\Sigma}_k^{(s)}) \right\}$$
$$> \sum_{n=0}^{N-1} \log \left\{ \sum_{k=0}^{K-1} g_k^{(b)}\, \mathcal{N}(\boldsymbol{x}(n)|\boldsymbol{\mu}_k^{(b)}, \boldsymbol{\Sigma}_k^{(b)}) \right\} + \log p(H_1) - \log p_(H_0)$$

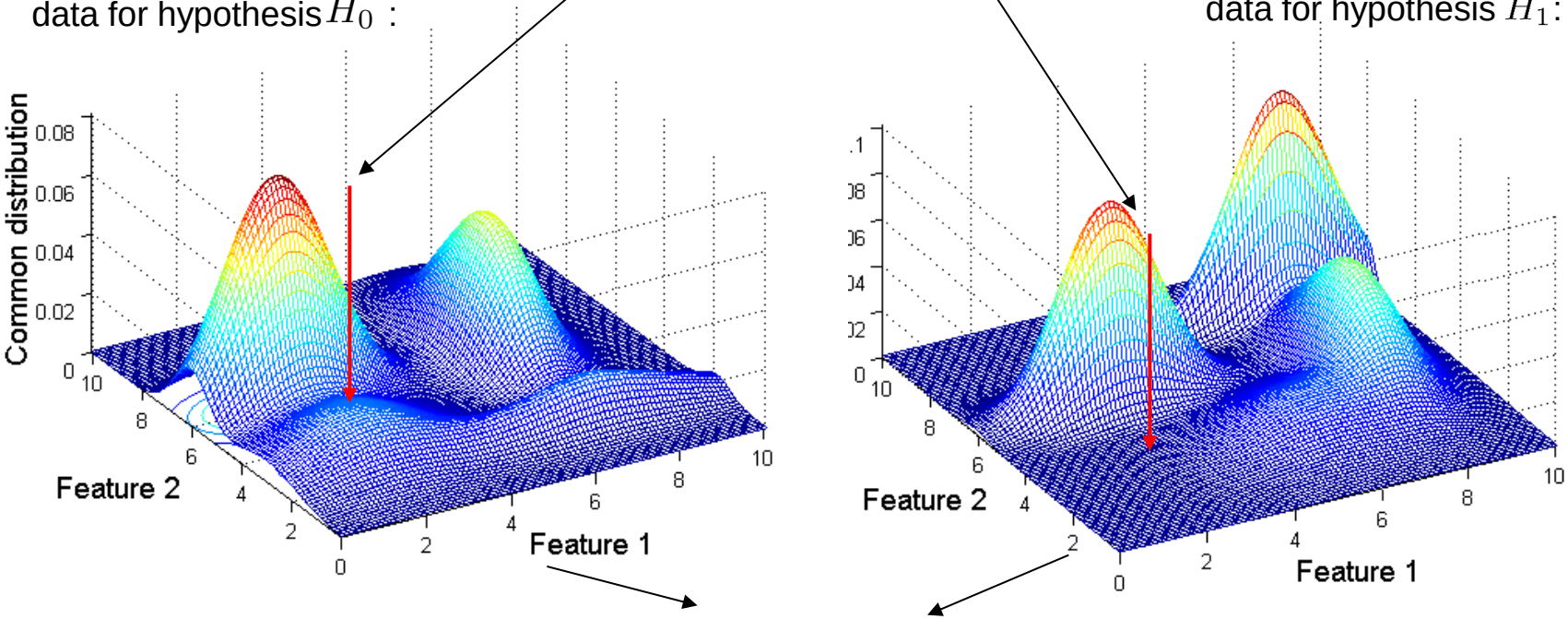
Verification procedure

- Detection for one of two hypotheses,
  based on GMM models of a multi-dimensional probability density function.

Observed data => current feature vector:

Probability model trained with data for hypothesis H_0 :

Probability model trained with data for hypothesis H_1:



Decision for the model with the higher probability for the current feature vector (red).

Detection sampling rates

TECHNISCHE UNIVERSITÄT DARMSTADT

- A detection is typically evaluated for one complete feature vector

$$\boldsymbol{X} = [\boldsymbol{x}(0), \boldsymbol{x}(1), \cdots, \boldsymbol{x}(N-1)]$$

  i.e., after N samples of the feature vector.

- Typically, the longer the feature vector is, the more reliable is the detection.
- However, the detection is then rather slow and cannot track well speaker changes.
- Dependent on the application, typically decisions are made every 1 – 10 sec.

Speaker verification

- **Error evaluation for the hypothesis H_0:**

  - Detection: $f_\mathrm{d} = p(\mathrm{det} = H_0 | H_0)$
  - False rejection: $f_\mathrm{fr} = p(\mathrm{det} = H_1 | H_0)$
  - False alarm: $f_\mathrm{fa} = p(\mathrm{det} = H_0 | H_1)$

  $$f_\mathrm{d} + f_\mathrm{fr} = 1$$

  Equal error rate (EER): $f_\mathrm{fr} = f_\mathrm{fa}$

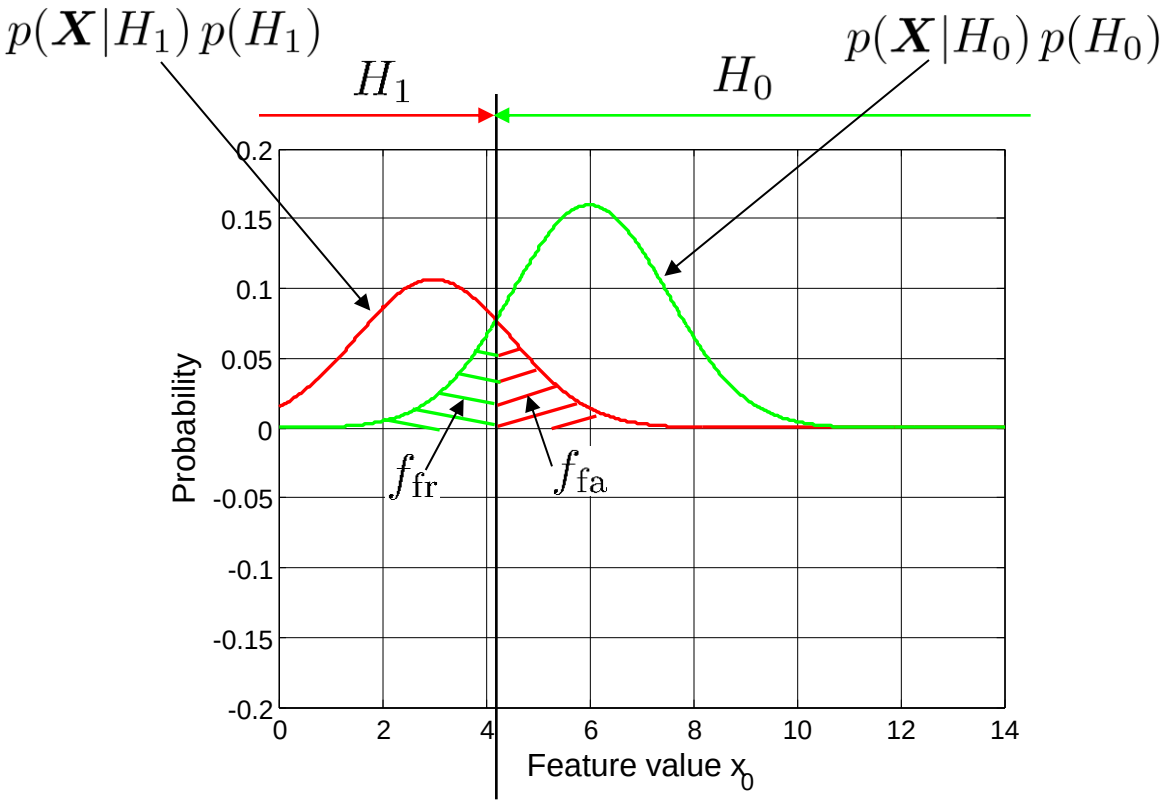
- A threshold is defined in order to adjust the missed detection and false alarm rates.

- Typically, the threshold is chosen such that:

$$p(\boldsymbol{X}|H_0)\, p(H_0) \;=\; p(\boldsymbol{X}|H_1)\, p(H_1)$$

Speaker verification

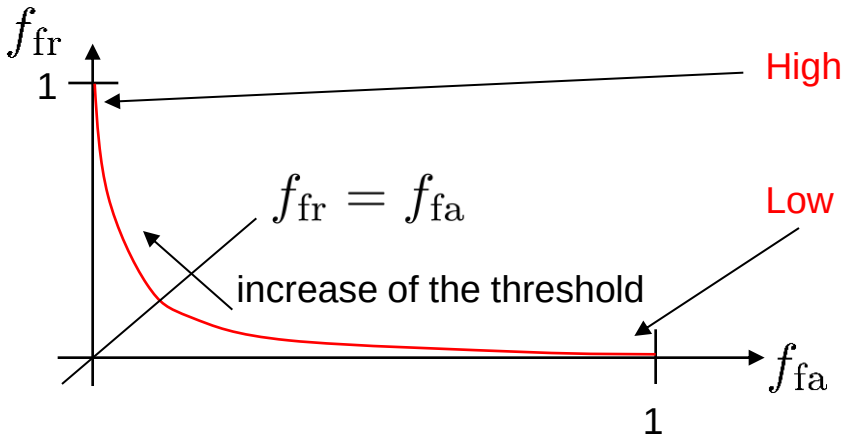
- **The threshold allows to adjust the false alarm and false rejection rates.**



Speaker verification

- **Threshold adjustment**:
  The threshold should be chosen to optimize the design criteria depending on the application. => Is a false alarm or a false detection more critical?

- The **ROC (receiver operator characteristic) or DET (detection error trade-offs)** curve defines operating points dependent on the chosen threshold.



High threshold => low false alarm rate but also low detection rate (=> high false rejection)

Low threshold => high detection rate (=> low false rejection) but also high false alarm rate

- An operating point where the detection systems are typically evaluated is the **"equal error rate" EER**, where $f_{\mathrm{fr}} = f_{\mathrm{fa}}$
- Typically, the thresholds are speaker specific. Then the threshold has to be adjusted during the training.

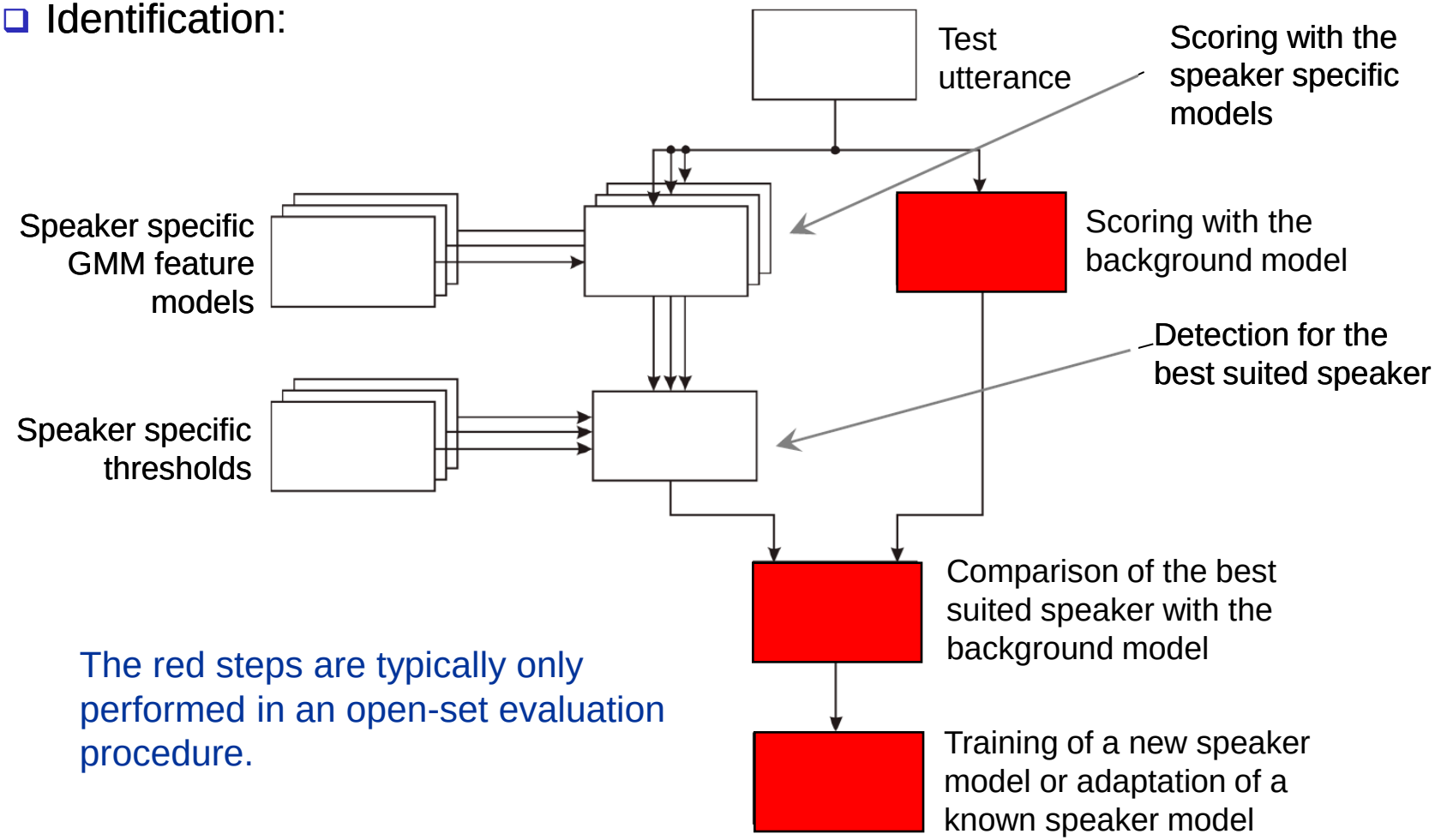
Variability of scores

TECHNISCHE UNIVERSITÄT DARMSTADT

- The quality of the detection depends on many factors:
  - **The quality of the training data:**
    Acoustic quality, phonetic content, natural and typical speech of the speaker, etc.
  - **Intra-speaker variability:**
    Difference of the speech of the test and training samples, e.g. by emotions, Lombard effect (other pronunciation in loud environments)
  - **Transmission channel:**
    Echo, reverberation, bandwidth limitation (e.g. by telephones).
  - **Environmental noise:**
    Car noise, cocktail party noise, etc.

Speaker identification

- Identification:



The red steps are typically only performed in an open-set evaluation procedure.

Results for a speaker identification procedure [3]

TECHNISCHE UNIVERSITÄT DARMSTADT

- Setup:
  - Database of 51 male speakers.
  - 10 conversations recorded in 10 different sessions for approximately 45 seconds.
  - Use of data with a telephone bandwidth, sampling rate 8 kHz.
  - MFCC features were used with 25-dimensional cepstral vectors.
  - Experiments were conducted on a 16-speaker subset.
  - GMM models were used with diagonal covariance matrices.

Results for a speaker identification procedure [3]

TECHNISCHE UNIVERSITÄT DARMSTADT

- Results: Recognition rate (correct identified segments / # segments) dependent on the **training data length, the test data length**, and the **number of Gaussians in the GMM** models.

Training data length	GMM model order	Test data length 1 sec	5 sec	10 sec
30 sec	8	54.6 %	79.8 %	86.6 %
	16	63.7 %	87.3 %	90.5 %
	32	64.6 %	85.3 %	88.4 %
60 sec	8	66.1 %	91.5 %	97.3 %
	16	74.9 %	95.7 %	98.8 %
	32	78.6 %	95.6 %	98.3 %
90 sec	8	71.5 %	95.5 %	98.8 %
	16	79.0 %	98.0 %	99.7 %
	32	84.7 %	98.8 %	99.6 %

Model adaptation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Procedure:
  - When a specific speaker has been detected, his model can be updated during the detection process. This may lead to a better detection in the following since the model can then adapt to
    - the current pronunciation of the speaker
    - the current transmission channel of the speaker (between his mouth and the microphone) or the telephone connection.
  - Generally, all parameters could be adapted. However, experiments showed that the adaptation of the means allows a good trade-off between complexity and gain in performance.

Model adaptation

- Procedure:
  - Supposing the new feature vector $\boldsymbol{X} = [\boldsymbol{x}(0), \boldsymbol{x}(1), \cdots, \boldsymbol{x}(N-1)]$ for which a specific speaker has been detected.
  - First a soft allocation of these feature vectors is calculated for each of the K Gaussian centers:

$$\gamma_k(n) = \frac{g_k \, \mathcal{N}(\boldsymbol{x}(n)|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=0}^{K-1} g_j \, \mathcal{N}(\boldsymbol{x}(n)|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

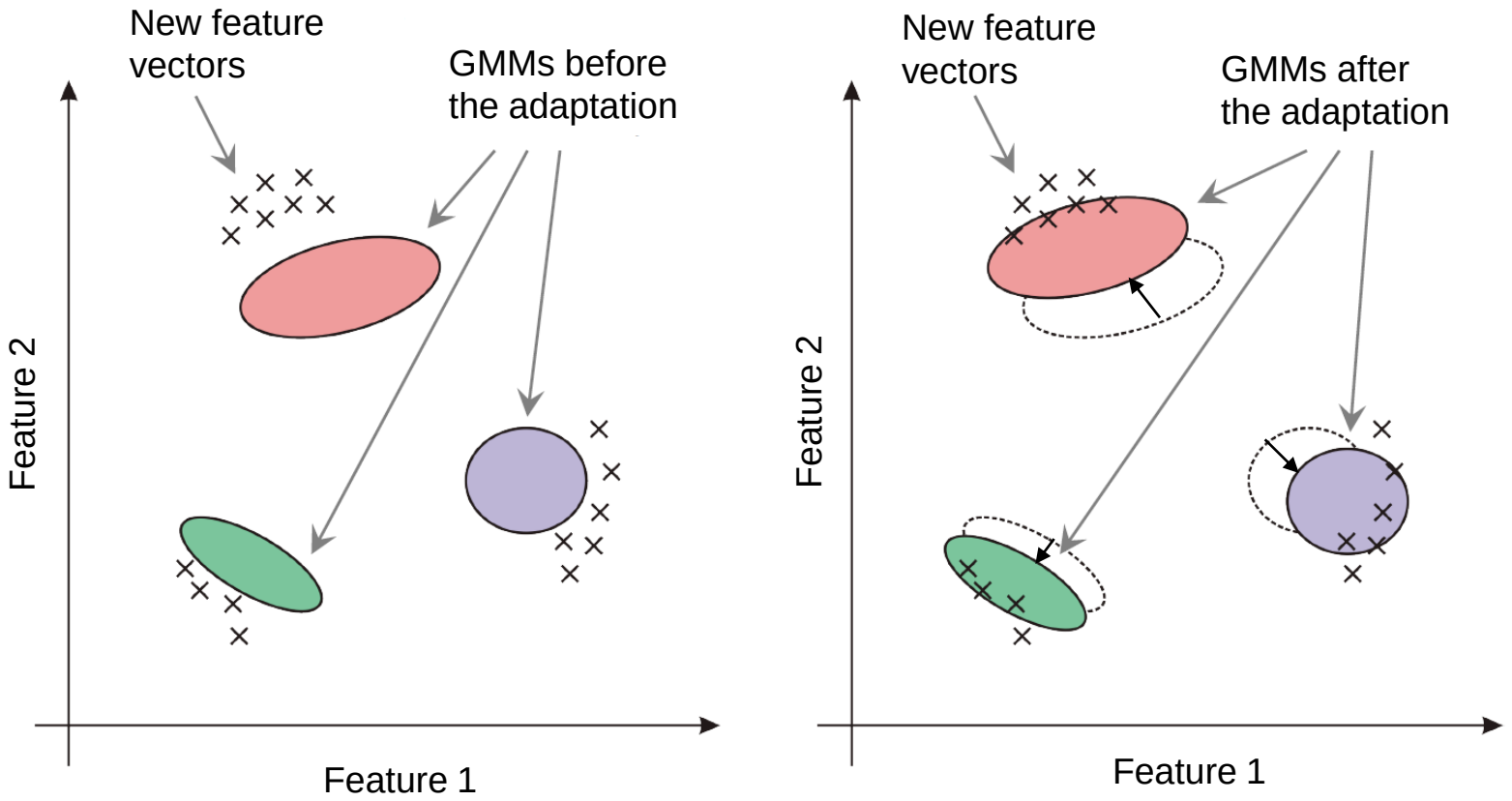
  - Then the mean values are updated

$$\boldsymbol{\mu}_k^{(\text{new})} = \frac{\sum_{n=0}^{N-1} \gamma_k(n) \, \boldsymbol{x}(n) + N_k \, \boldsymbol{\mu}_k^{(\text{old})}}{\sum_{n=0}^{N-1} \gamma_k(n) + N_k}$$

    where N_k describes the number of elements during the training step.

Model adaptation

- Example:



Summary

- Applications of speaker detection
- Types of speaker detection
  - Speaker verification
  - Speaker identification
- Preprocessing and feature extraction
- Speaker detection based on GMMs
- Speaker verification => error analysis
- Speaker identification => a practical example
- Model adaptation

=> next week: Hidden Markov Models (HMMs) and speech recognition

References

Speaker detection :

[1] Frédéric Bimbot, et al.: *A Tutorial on Text-Independent Speaker Verification,* Eurasip Journal on Applied Signal Processing 2004:4, 430-451

[2] Douglas A. Reynolds: *An overview of Automatic Speaker Recognition,* Technology, In Proc. ICASSP, IEEE, vol. IV, pp. 4072-4075, 2002

[3] D.A. Reynolds, R.C. Rose: *Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models*, IEEE Trans. on Speech and Audio Processing, vol.3, no.1, pp. 72-83, 1995