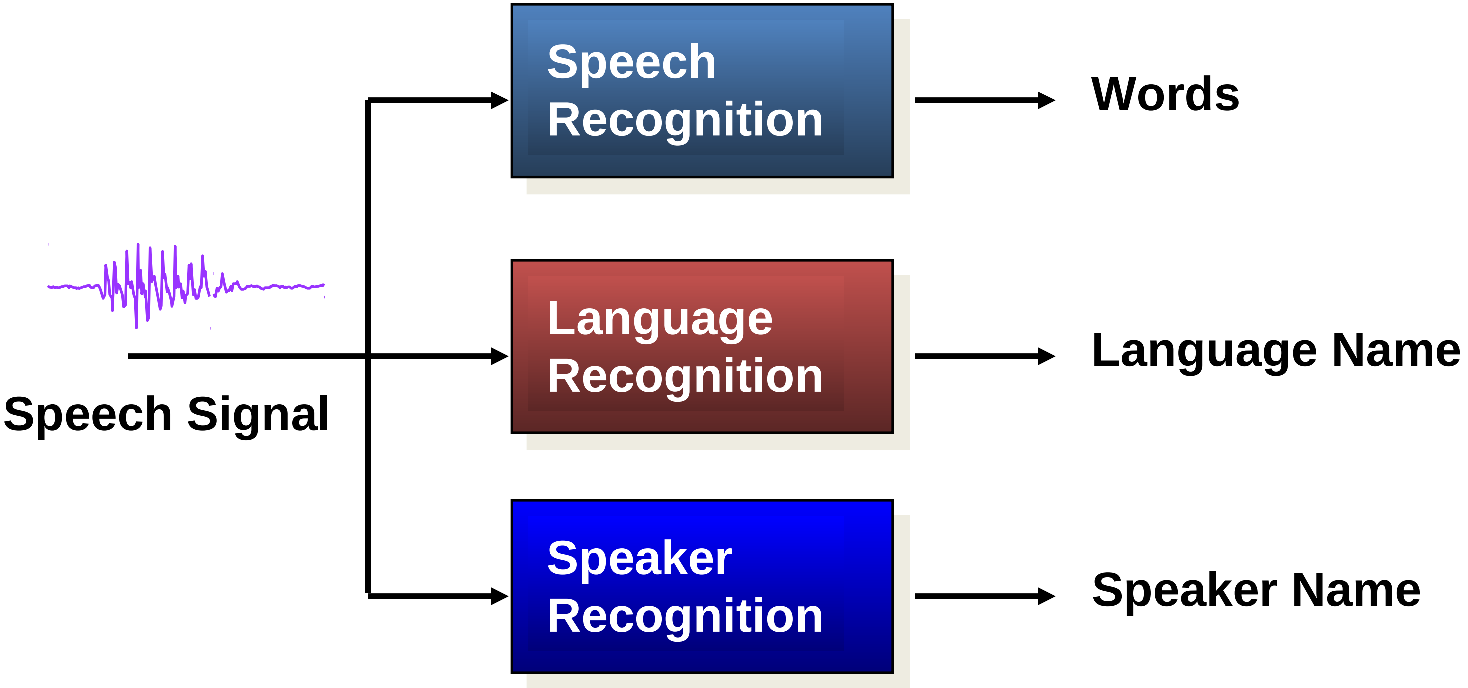
Speaker Recognition

Dr. Anil Kumar Vuppala

IIIT-Hyderabad

Introduction



Speaker Recognition: Man v/s Machine

- Recognizing people solely by their speech
- Who has spoken rather than what is spoken?
- Recognizing people by listening – Speaker recognition by man
- Recognizing people by signal processing and pattern recognition – Speaker recognition by machine
- Speaker recognition by machine is also termed as Automatic speaker recognition

Speaker Recognition by Machine

- Some signal processing features are extracted from the speech signal using the existing signal processing tools
- Reference speaker models are built using the pattern recognition tools for extracted features
- Speaker models are used for recognizing speakers
- It is a fact that machine depends more on physiological aspect of speaker information and less on behavioral aspect of speaker information

Automatic Speaker Recognition

- Biometric Speaker Recognition
  - Speech as a biometric feature for person authentication
  - Security purpose
    - Shall I allow this person to do his/her business?
  - Commercial and Defense Applications
  - Banking transactions, entry to protected areas

- Forensic Speaker Recognition
  - Speech as a forensic evidence for person identification
  - Criminal investigation
    - Is the given speech data really spoken by this person?
    - Who among the suspects has spoken this message?

Classification of Speaker Recognition

- Speaker Recognition: Recognizing speakers by extracting and modeling signal processing features from the speech signal
- Classification
  - Speaker verification v/s Speaker identification
  - Text-dependent v/s Text-indenpendent
  - Closed set v/s Open set

Speaker Verification v/s Identification

- Verifying the identity claim of a speaker
  - I am so and so, please allow me to use

- Identifying the speaker of the speech signal
  - I will not tell who I am, bet you identify me

Speech Modalities

- **Text-dependent recognition**
  - Recognition system knows text spoken by person
  - Examples: fixed phrase, prompted phrase
  - Used for applications with strong control over user input
  - Knowledge of spoken text can improve system performance

- **Text-independent recognition**
  - Recognition system does not know text spoken by person
  - Examples: User selected phrase, conversational speech
  - Used for applications with less control over user input
  - More flexible system but also more difficult problem
  - Speech recognition can provide knowledge of spoken text

Closed-set v/s Open-set

- Closed-Set: Speech during testing is always from one of the enrolled speakers
  - Identify who is the speaker among the enrolled
- Open-Set: Speech during testing may be from the speaker who is not enrolled
  - Identify whether he/she belongs to the enrolled set or not
  - If so, identify who is the speaker among the enrolled

Speaker Recognition by Pattern Recognition Approach

- Pattern recognition task
  - Feature extraction
  - Training/Pattern classification
  - Testing/Pattern comparison
- Feature extraction
  - Digital signal processing tools
- Training
  - Pattern recognition tools
- Testing
  - Spectral dissimilarity measures

Feature Extraction

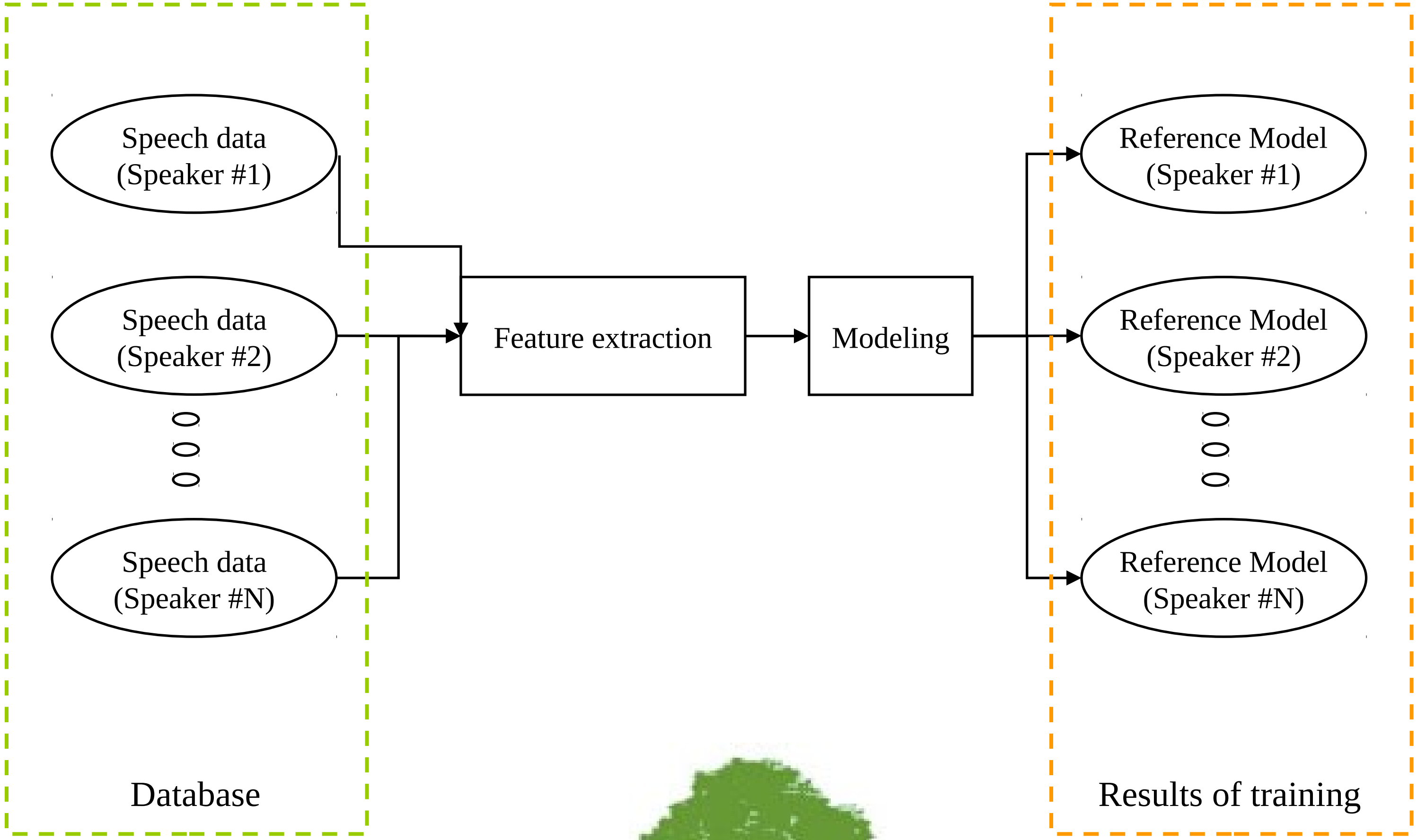
- Reduced data rate and enhance relevant information
- The accuracy of classification is strongly determined by its selection

- **Speech analysis**
  - Segmental (1-5 ms)
  - Sub-segmental (10-40 ms)
  - Supra-segmental (100 ms)

- **Speech signal processing**
  - FFT-implemented filterbanks
  - LP analysis
  - Cepstral Analysis
  - Sinusoidal Analysis

Speaker Modelling

- Speaker Recognition system (training process)

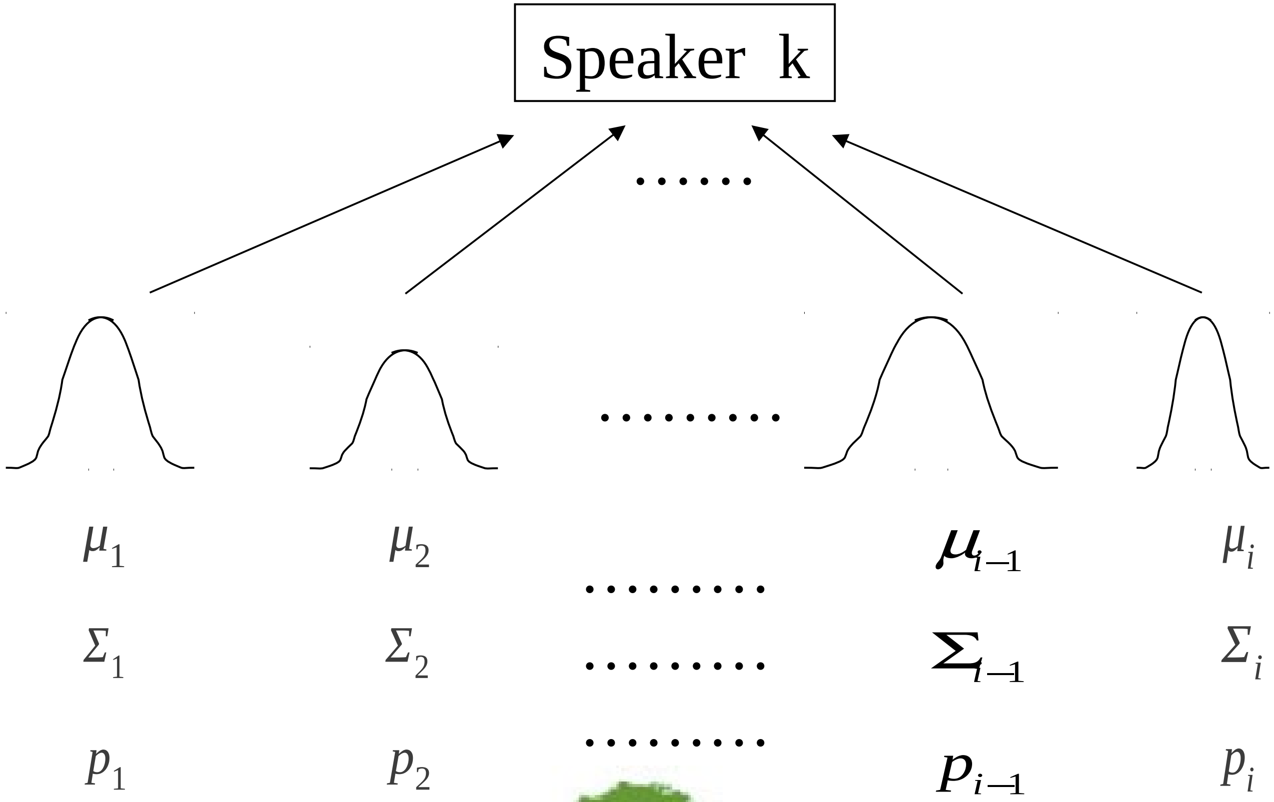


Speaker Models

- Dynamic Time Warping (DTW)
- Vector Quantization (VQ)
- Gaussian Mixture Model (GMM)
- Hidden Markov Models (HMM)
- Neural Network Models (NN)
- Support Vector machine (SVM)

Gaussian Mixture Model (GMM)

- Each speaker is modeled by a sum of different Gaussians



Gaussian mixture models (cont.)

For a *D*-dimensional feature vector $\vec{x}$, the mixture density used for the likelihood function is defined as follows:

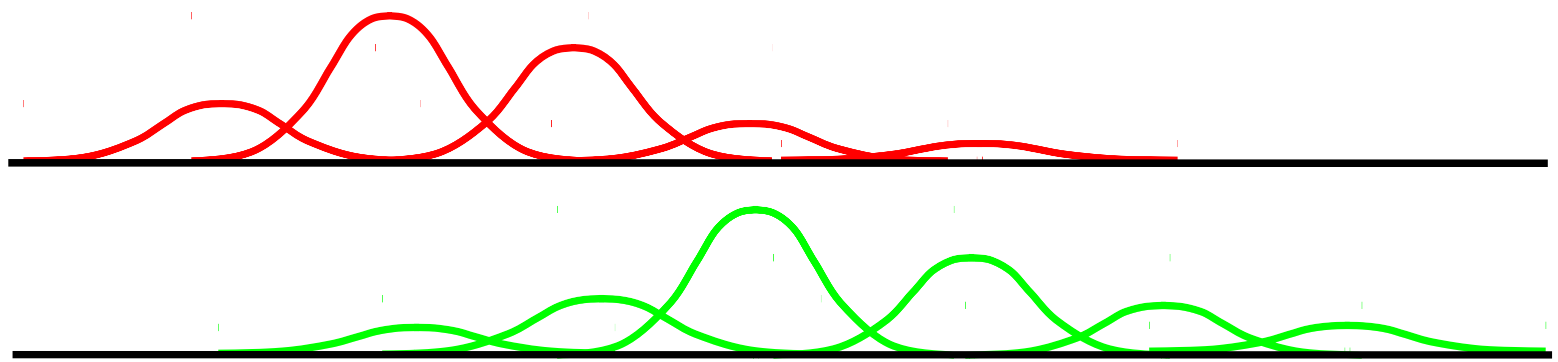
$$p(\vec{x}|\lambda)=\sum_{i=1}^{M} w_i p_i(\vec{x}) \qquad \sum w_i = 1$$

Gaussian densities $p_i(\vec{x})$, each parameterized by a *D* × 1mean vector $\vec{u}_i$ and a *D* × *D* covariance matrix Σ_i:

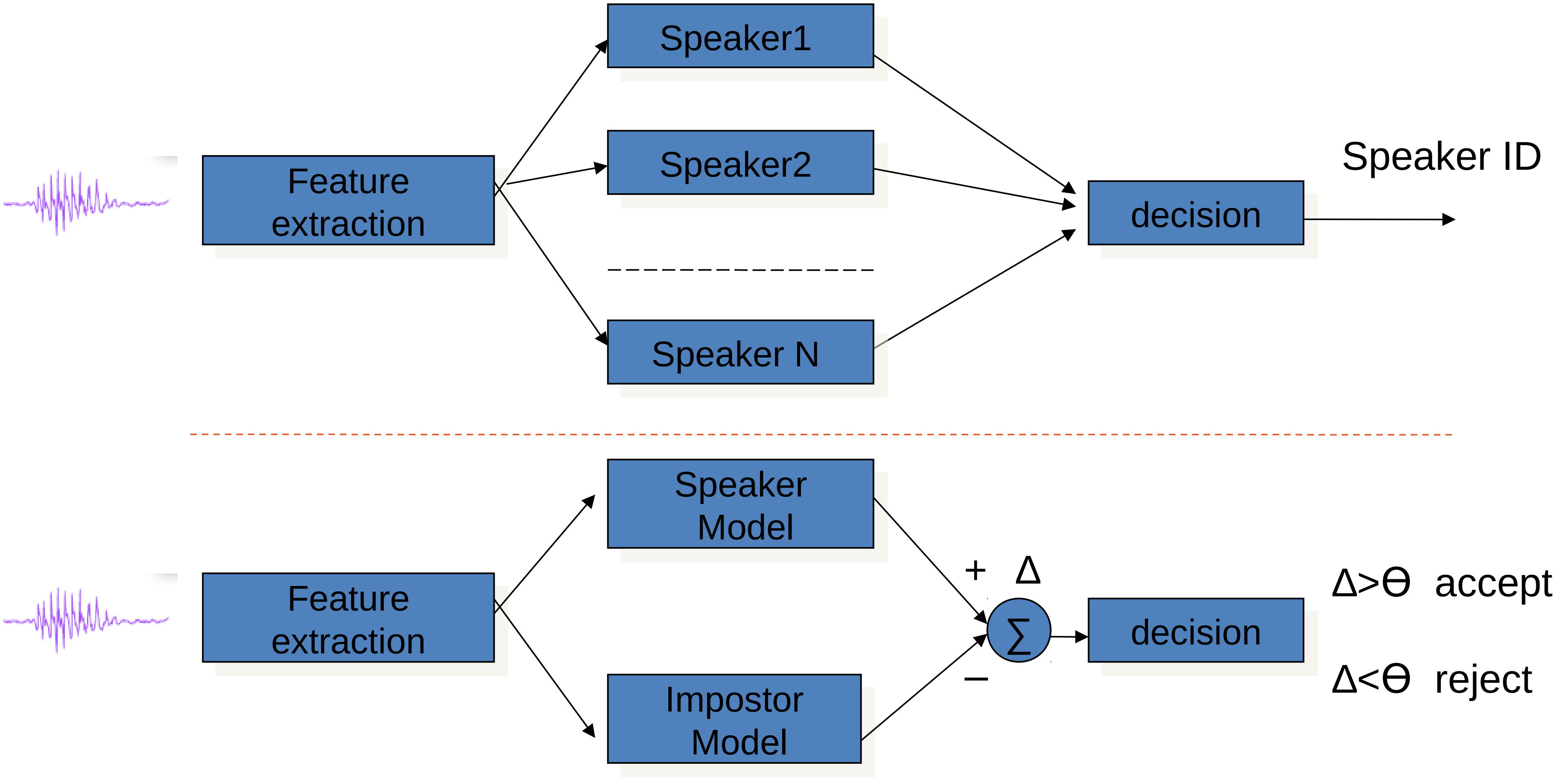
$$p_i(\vec{x})=\frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} e^{-(1/2)(\vec{x}-\vec{\mu}_i)' \sum (\vec{x}-\vec{\mu}_i)}$$

Collectively, the parameters of the density model are denoted as

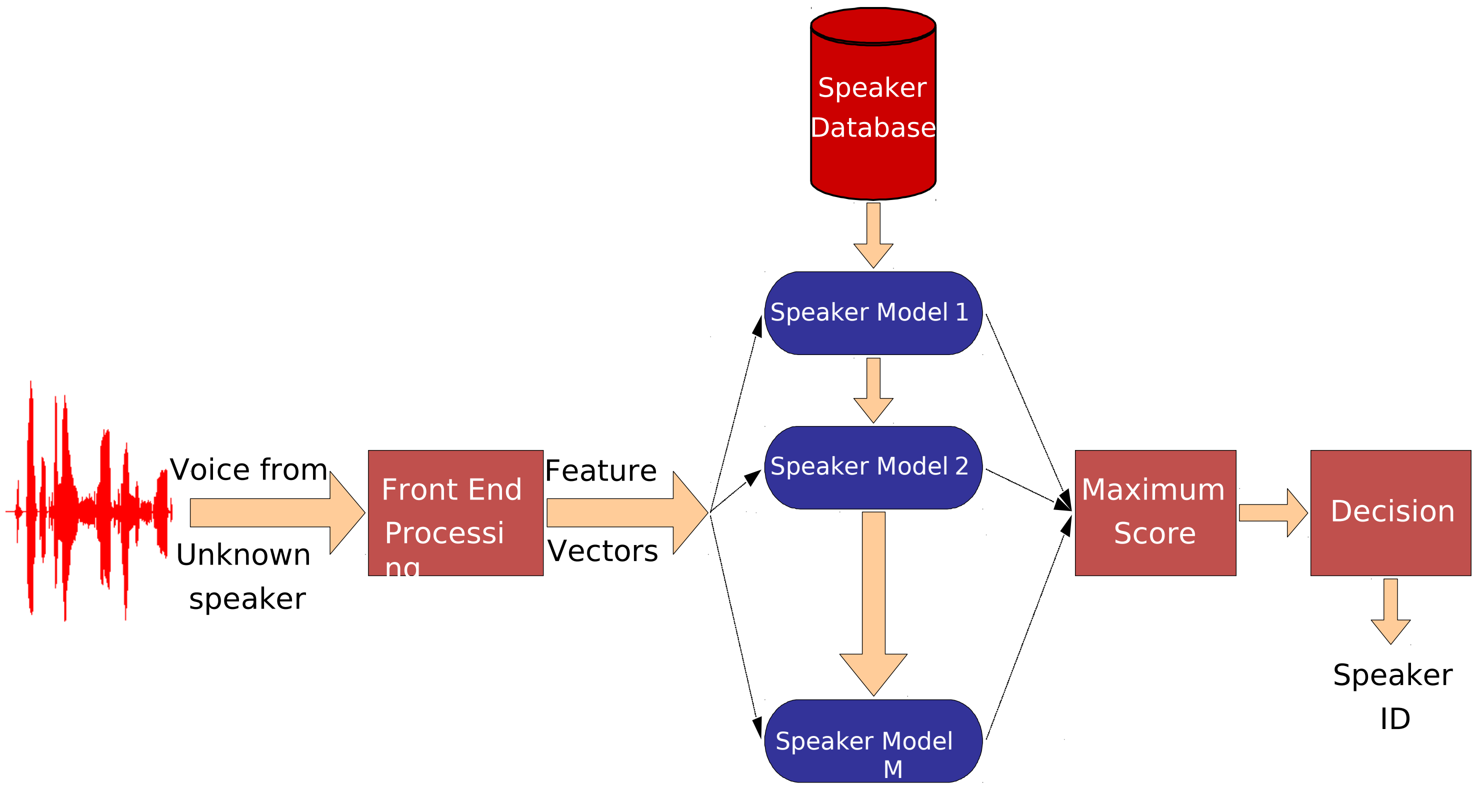
$$\lambda=(w_i, \vec{u}_i, \Sigma_i)$$



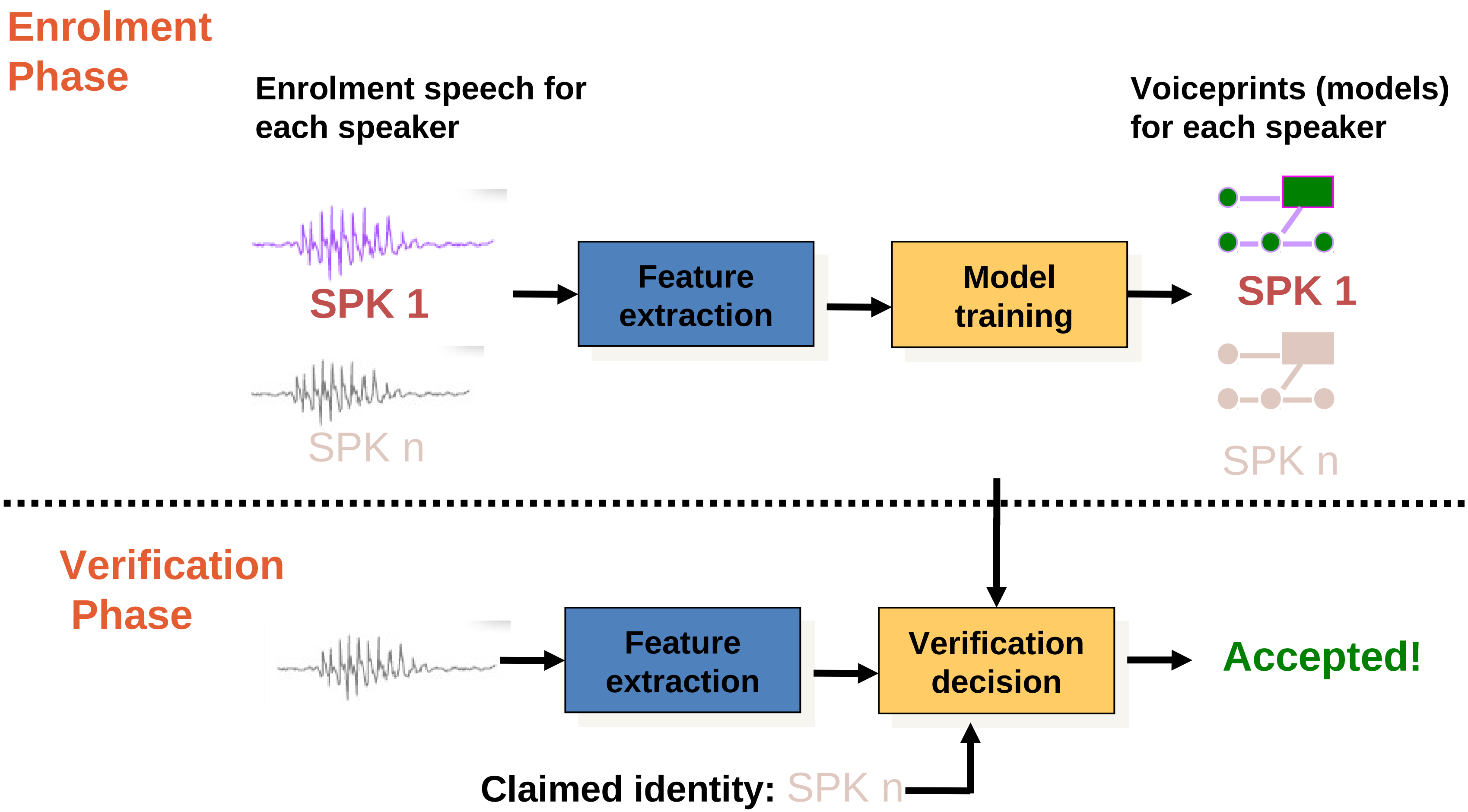
Identification vs verification



Phases of Speaker Identification System



Phases of Speaker Verification System



Speaker Recognition

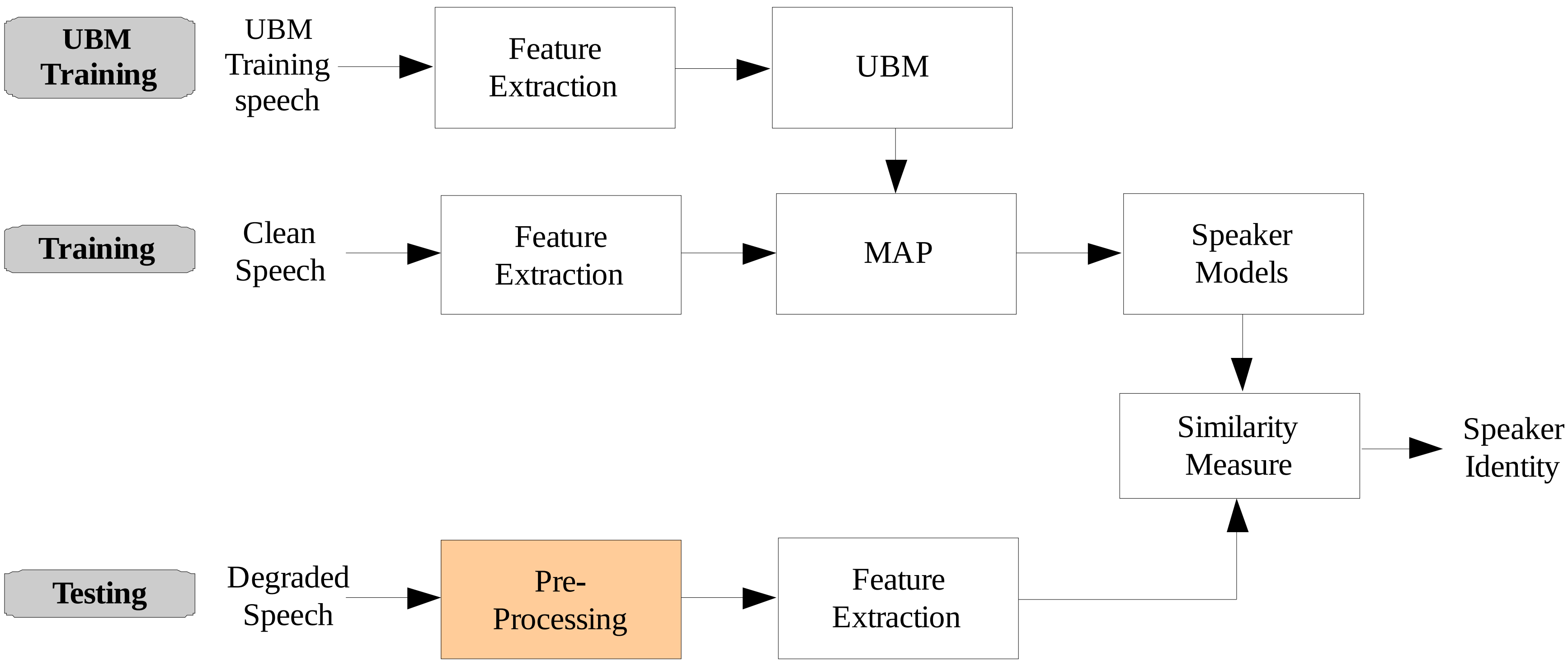


Fig.: Block diagram of speaker recognition using GMM-UBM

Experimental Description

- **Database**
  - TIMIT (630 speakers, 438 males and 192 females)
  - 10 sentences for each speaker, 3 s each.
  - Subset of 100 speakers
  - Training: First 8 sentences, Testing: Last 2 sentences
- **Speaker-Specific Feature**
  - MFCC
- **Modelling**
  - GMM-UBM
- **UBM Training**
  - 1 Hour of speech
  - Performance is 99% for TIMIT database by using 512 mixtures.

Speaker Recognition Results

Table: **Speaker recognition performance (percentage of identification) under noisy environment.** In table abbreviations DEG, TP, SP1, SP2, TSP1 and TSP2 refer to degraded speech, temporal processing, multi band spectral subtraction, MMSE-STSA estimator, combined temporal and multi-band spectral subtraction, and combined temporal and MMSE-STSA estimator, respectively.

SNR	**0 dB**	**3 dB**	**6 dB**	**9 dB**	**12 dB**	**15 dB**	**20 dB**	**30 dB**
DEG	1.50	2.00	2.00	3.50	10.50	23.50	51.50	89.00
TP	3.00	5.00	16.00	21.00	33.50	42.50	78.50	87.50
SP1	6.00	19.00	32.00	49.50	49.50	70.50	87.00	92.00
SP2	5.50	13.00	21.00	36.00	57.00	77.00	86.50	91.50

Thank you