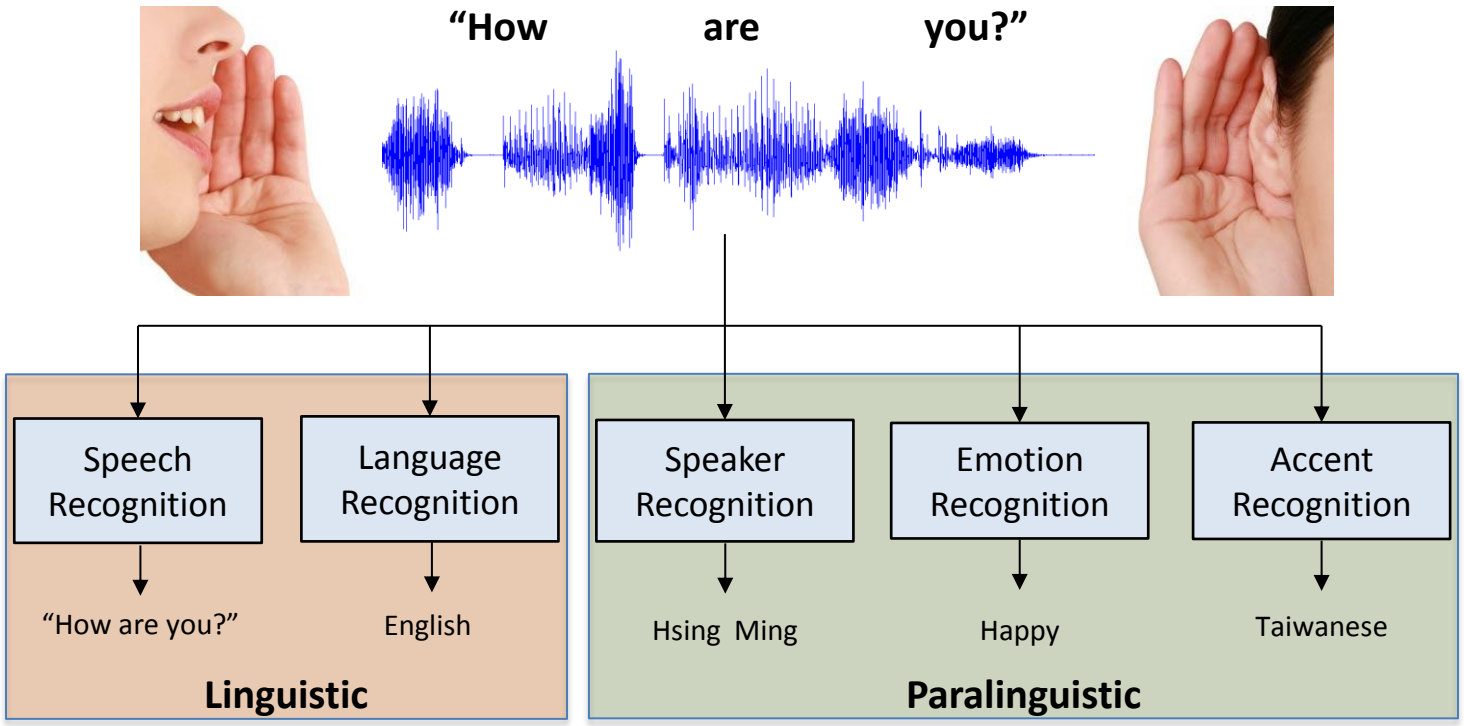
Speaker Verification – The present and future of voiceprint based security

Prof. Eliathamby Ambikairajah
Head of School of Electrical Engineering & Telecommunications,
University of New South Wales, Australia

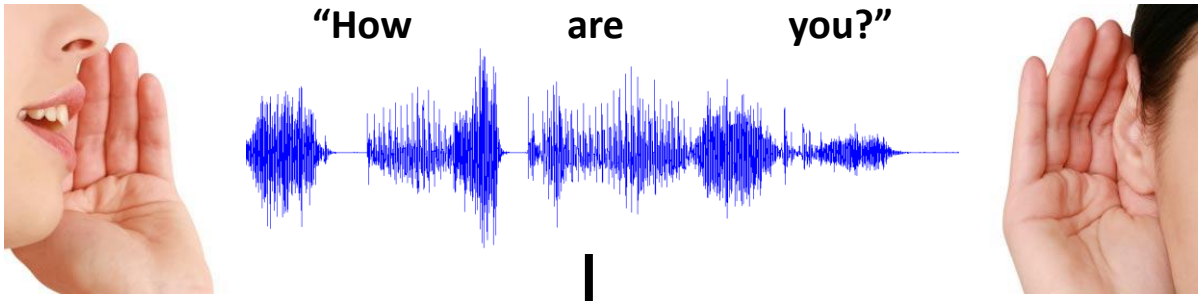
Outline

- Introduction
- Speaker Verification Applications
- Speaker Verification System
- Performance measure
- NIST Speaker Recognition Evaluation (SRE)
- Discussion

Introduction



- Speech conveys several types of information
  - Linguistic: message and language information
  - Paralinguistic : emotional and physiological characteristics



Speech Recognition

Language Recognition

Speaker Recognition

Emotion Recognition

Accent Recognition

Speaker Identification

determines who is speaking given a set of enrolled speakers

Speaker Verification

determines if the unknown voice is from the claimed speaker

Speaker Diarization

partition an input audio stream into homogeneous segments according to the speaker identity

Speaker Identification

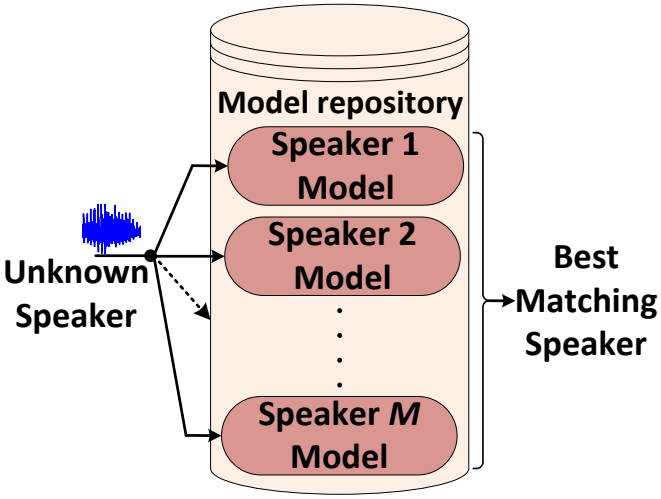
determines who is speaking given a set of enrolled speakers

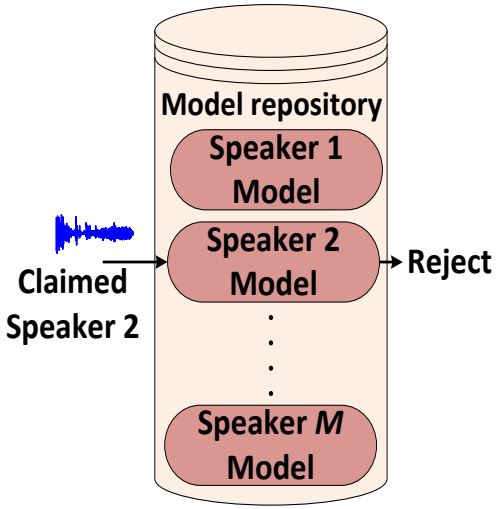
Speaker Verification

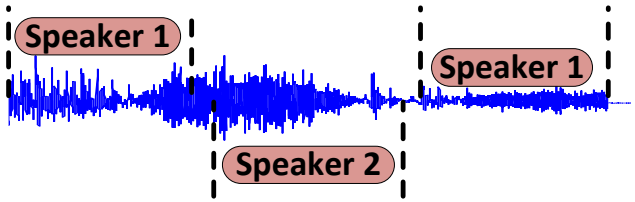
determines if the unknown voice is from the claimed speaker

Speaker Diarization

partition an input audio stream into homogeneous segments according to the speaker identity







Speaker Verification Applications - Biometrics

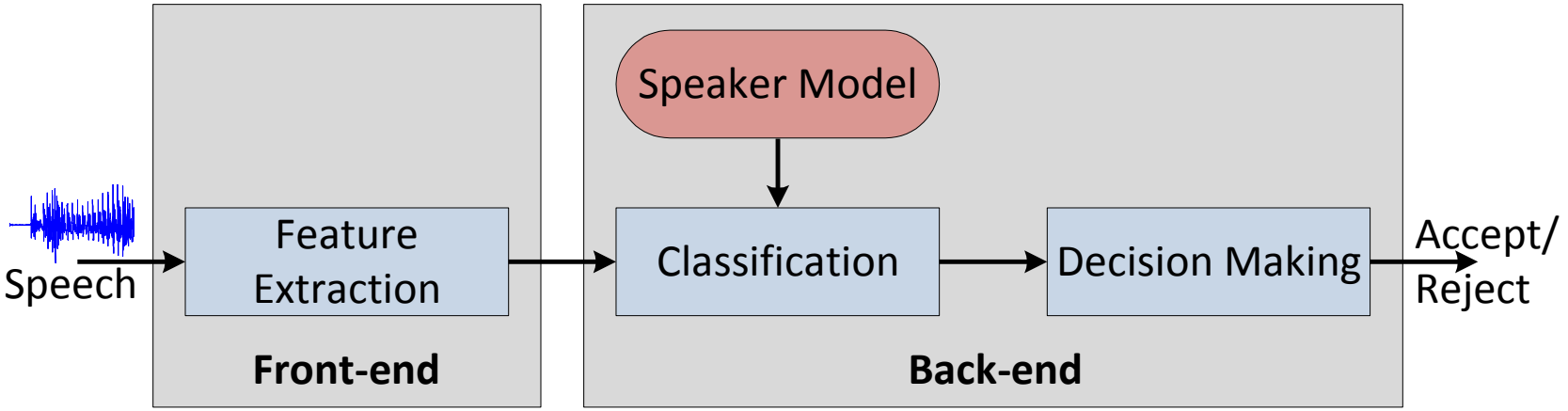
Access control

Physical facilities

Transaction authentication

Telephone credit card purchases

Speaker Verification System – Basic Overview



- In automatic speaker verification,
  - The front-end converts speech signal into a more convenient representation (typically a set of feature vectors)
  - The back-end compares this representation to a model of a speaker to determine how well they match

Speaker Verification System

I am John

Feature Extraction

$C_0 | C_1 | C_2 | \cdots | C_n$

Universal Background Models (UBM)

Generic Male

Generic Female

Determine level of Match

Determine level of Match

Likelihood of Generic Male

Likelihood of John

Speaker Models

Speaker 1 Model

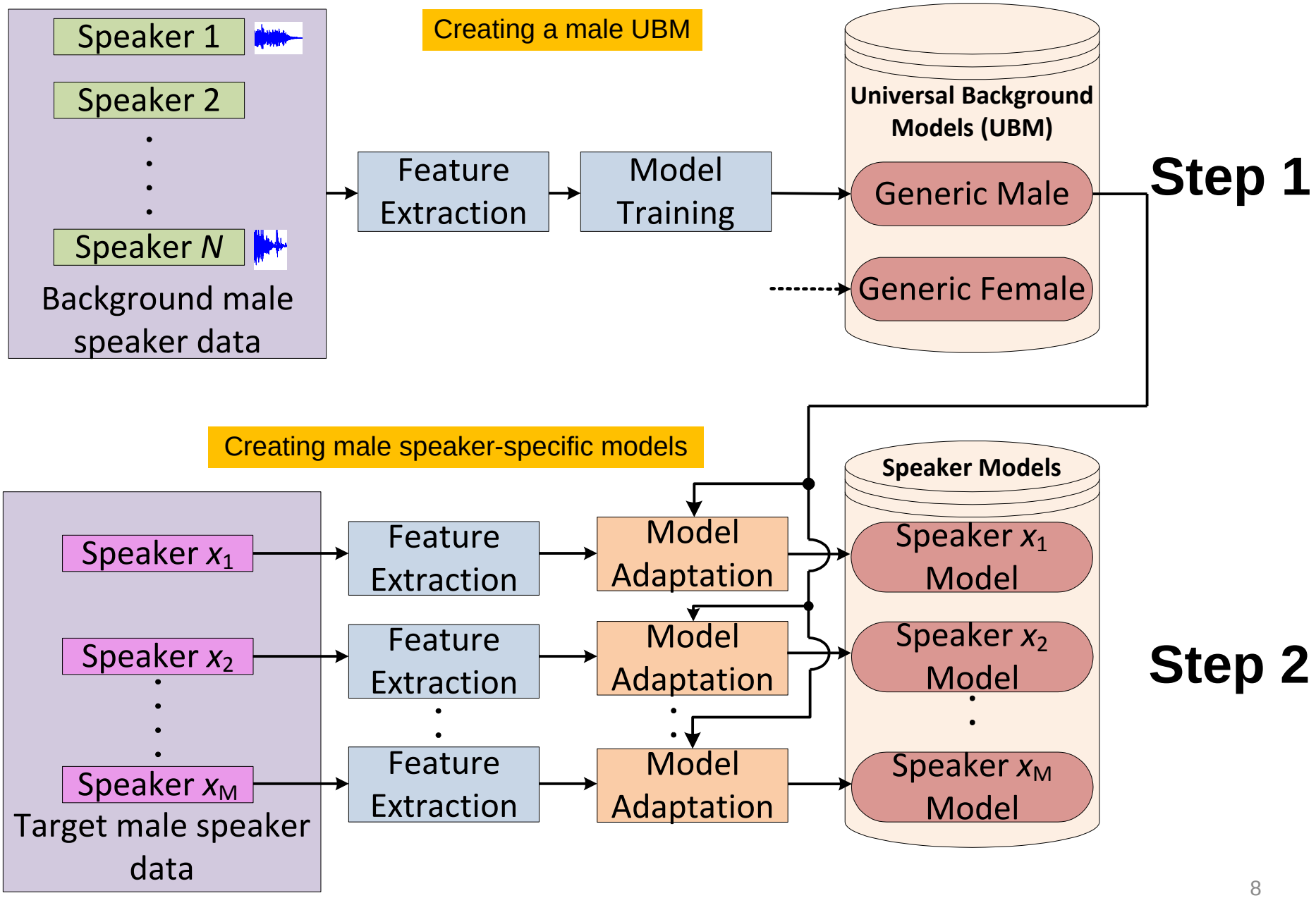
John's Model

Likelihood Ratio Decision Making

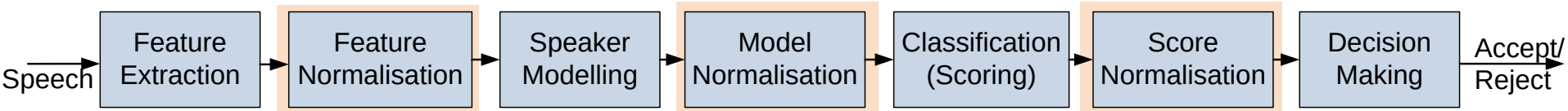
NOT JOHN

UBM: represent general, speaker independent model to be compared against a person-specific model when making an accept or reject decision.

Speaker Verification System – Speaker Enrolment



Detailed Speaker Verification System

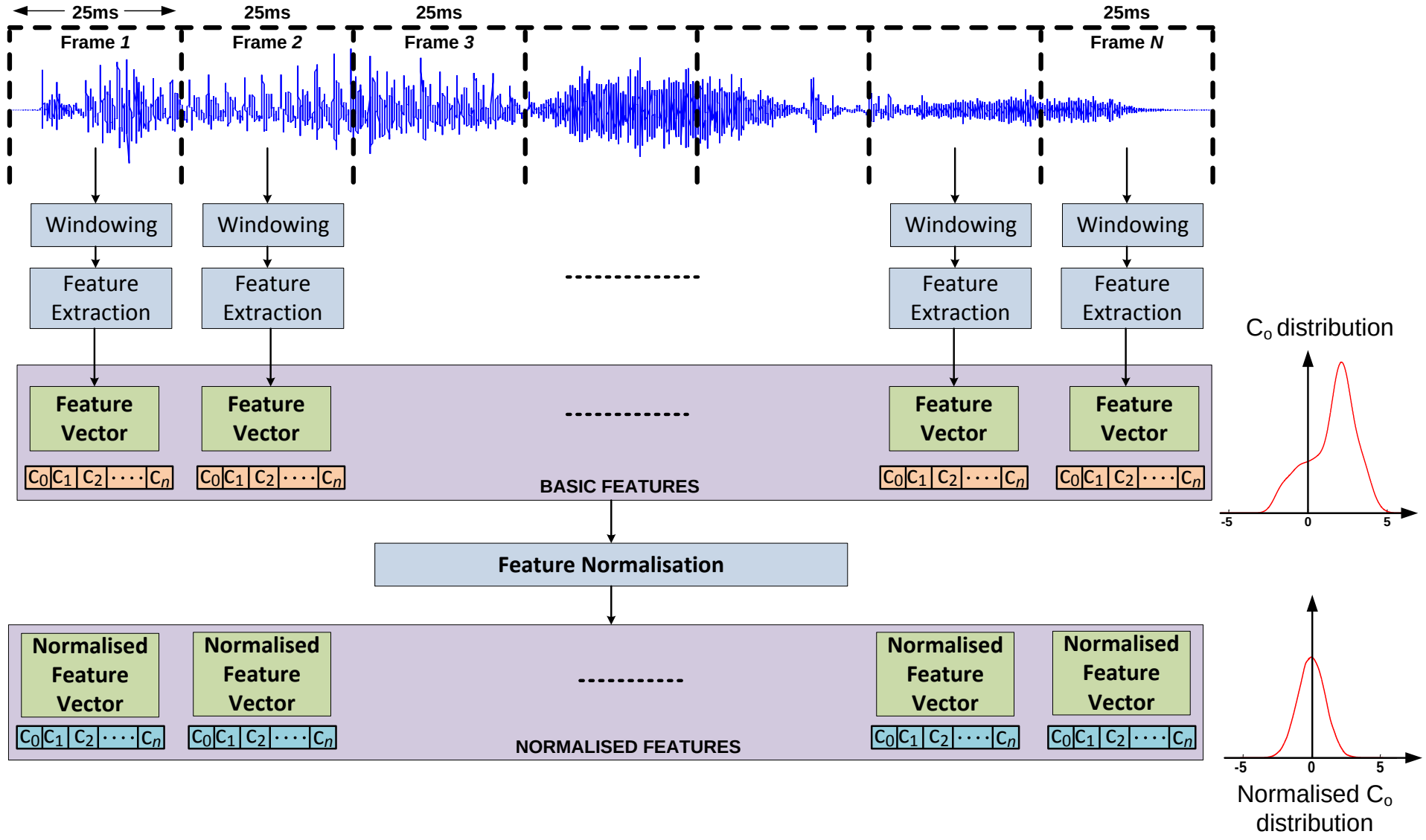


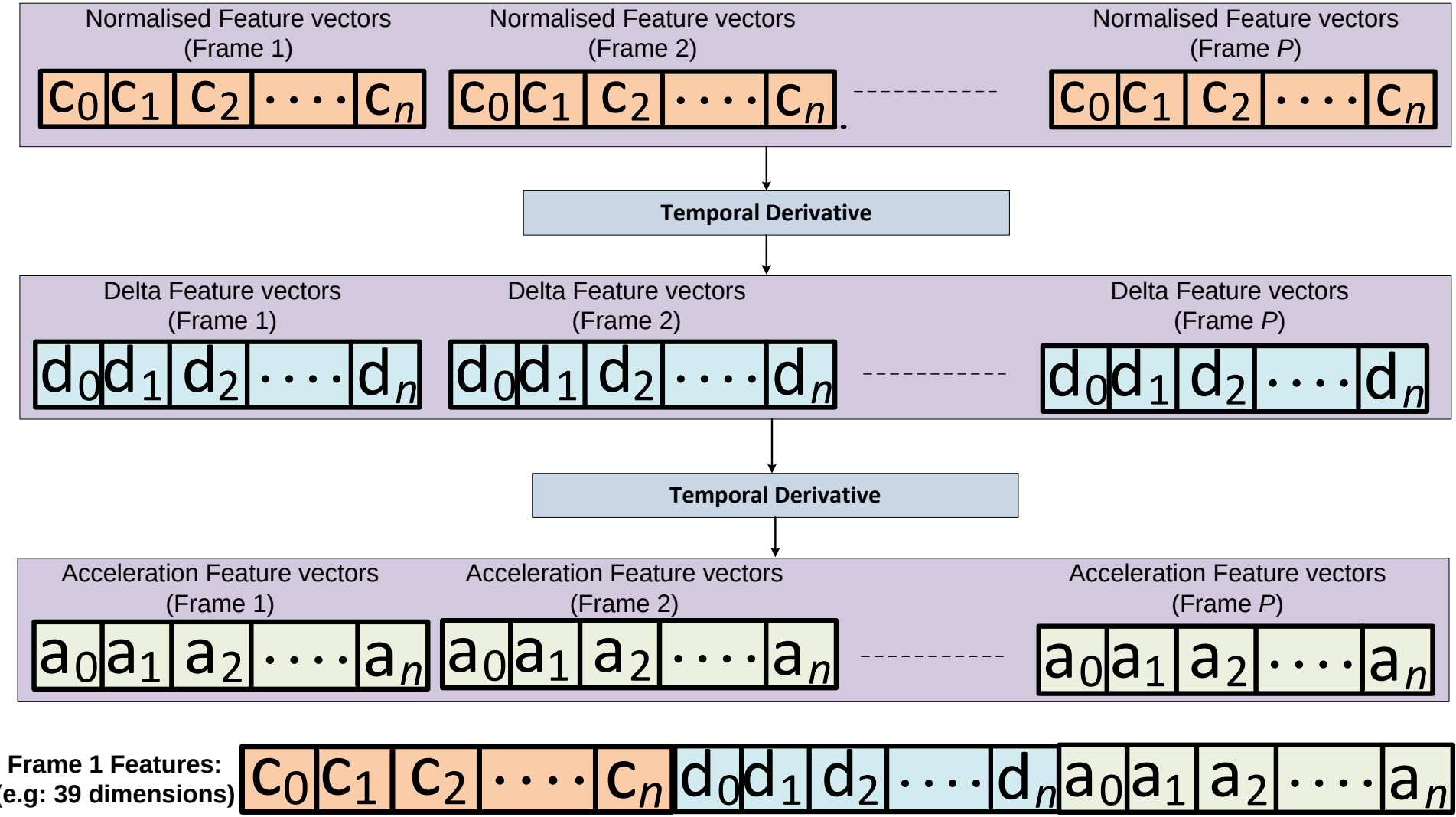
- Cepstral Mean Subtraction (CMS)
- RelAtive SpecTrAl (RASTA)
- Feature Warping
- Feature Mapping

- Nuisance Attribute Projection (NAP)
- Joint Factor Analysis (JFA)
- i-vectors
- Within Class Covariance Normalisation (WCCN)
- Linear Discriminant Analysis (LDA)
- Probabilistic Linear Discriminant Analysis (PLDA)

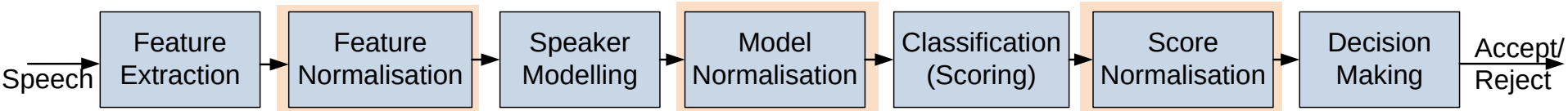
- Zero-normalisation (Z-norm)
- Test-normalisation (T-Norm)

Front-end: Feature Extraction




Normalised Feature vectors (Frame 1)
Normalised Feature vectors (Frame 2)
Normalised Feature vectors (Frame P)
c0 c1 c2 ···· cn
Temporal Derivative
Delta Feature vectors (Frame 1)
Delta Feature vectors (Frame 2)
Delta Feature vectors (Frame P)
d0 d1 d2 ···· dn
Temporal Derivative
Acceleration Feature vectors (Frame 1)
Acceleration Feature vectors (Frame 2)
Acceleration Feature vectors (Frame P)
a0 a1 a2 ···· an
Frame 1 Features: (e.g: 39 dimensions)
c0 c1 c2 ···· cn d0 d1 d2 ···· dn a0 a1 a2 ···· an

Detailed Speaker Verification System

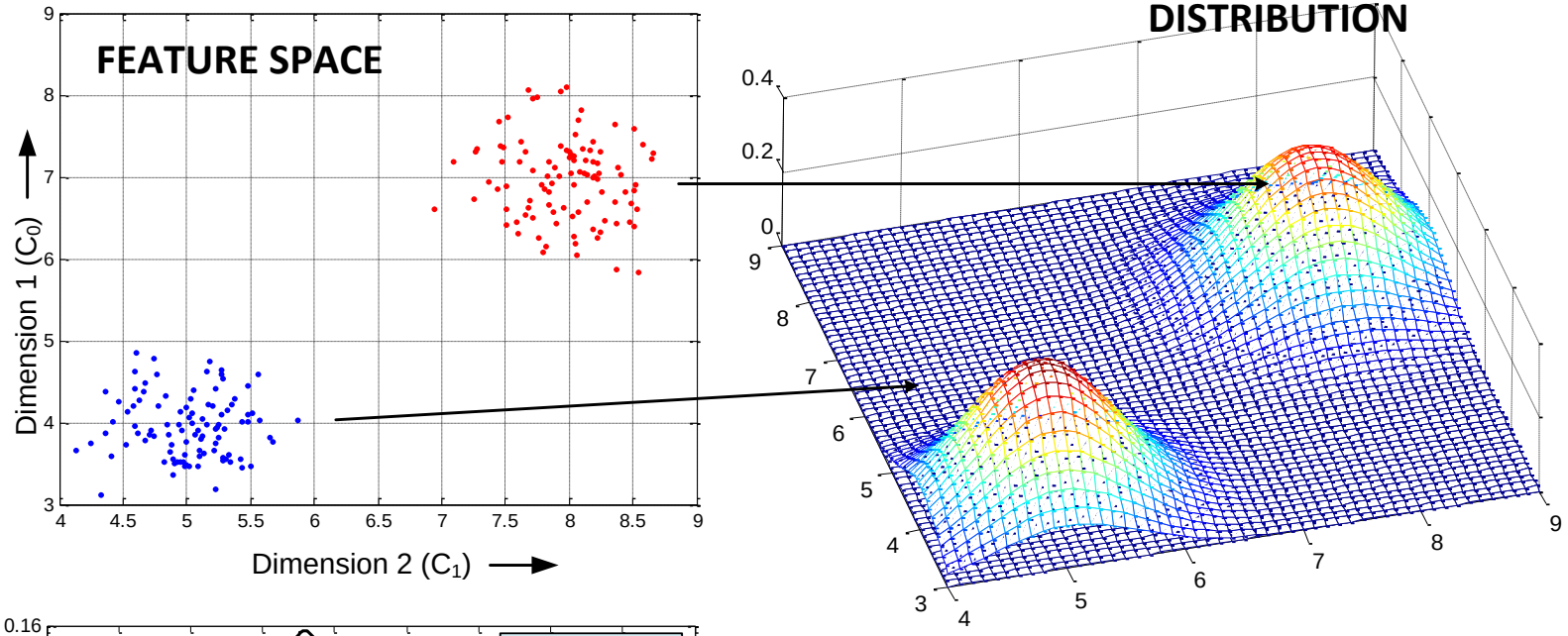


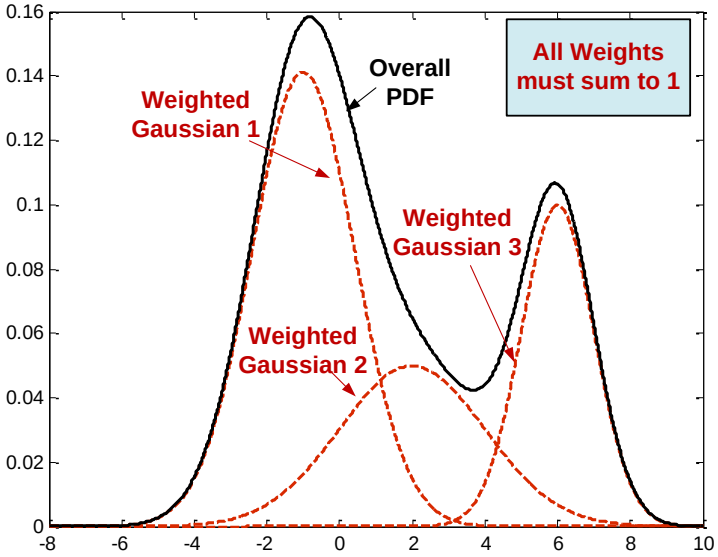
- Cepstral Mean Subtraction (CMS)
- RelAtive SpecTrAl (RASTA)
- Feature Warping
- Feature Mapping

- Nuisance Attribute Projection (NAP)
- Joint Factor Analysis (JFA)
- i-vectors
- Within Class Covariance Normalisation (WCCN)
- Linear Discriminant Analysis (LDA)
- Probabilistic Linear Discriminant Analysis (PLDA)

- Zero-normalisation (Z-norm)
- Test-normalisation (T-Norm)

Speaker Modelling

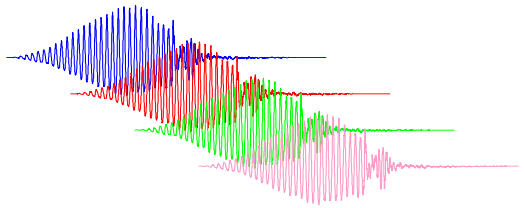




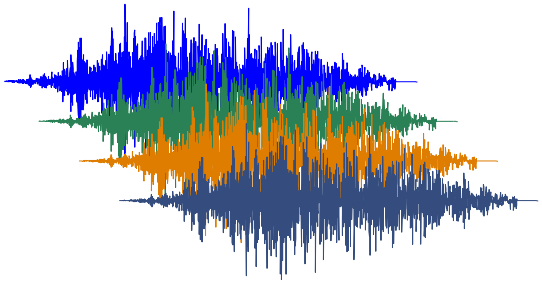
- ✓ **Probability density function approximated by 3-component Gaussian mixture models**
- ✓ **Each Gaussian mixture consist of a mean (μ), covariance (Σ) and weight (*w*)**

Database for creating UBM (example)

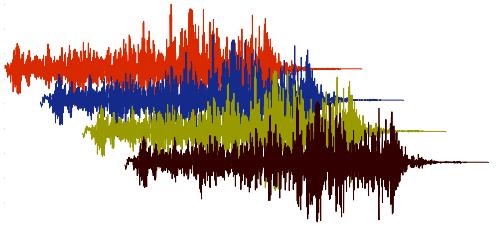
- Training set
  - 56 male speakers (each speaker consists of 2 minutes of active speech) for creating the UBM

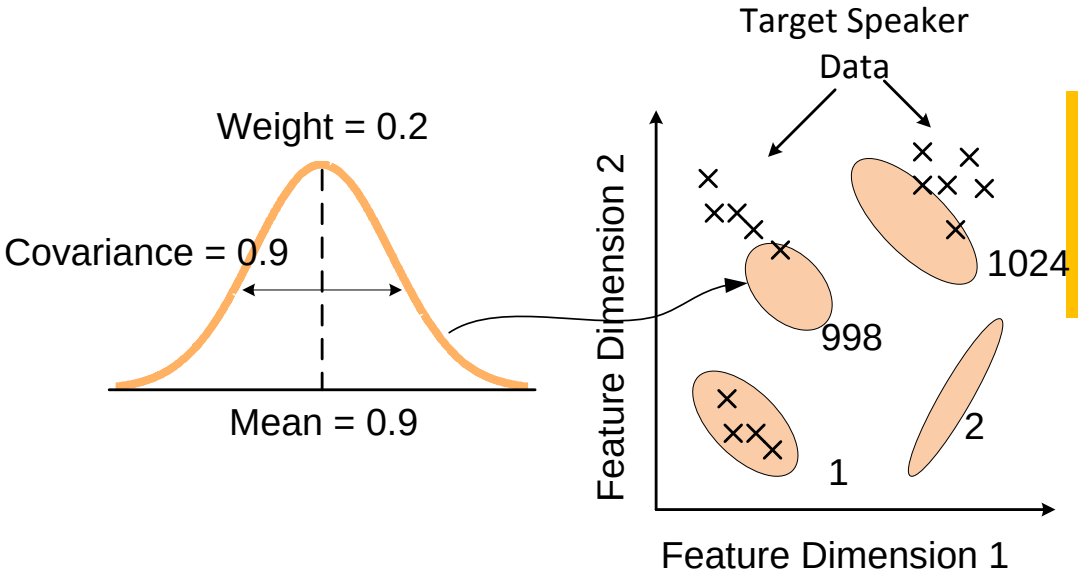


- Target set
  - 20 male speakers (each speaker consists of 2 minutes of active speech) for speaker-specific model



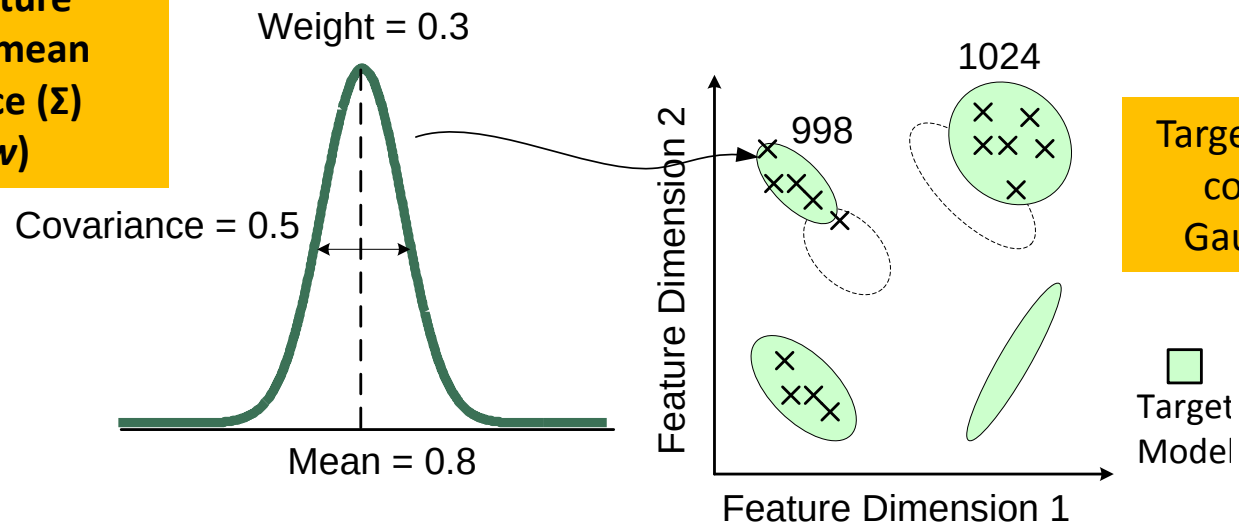
- Test set
  - 250 male utterances (each speaker has many test utterances) with the known identity





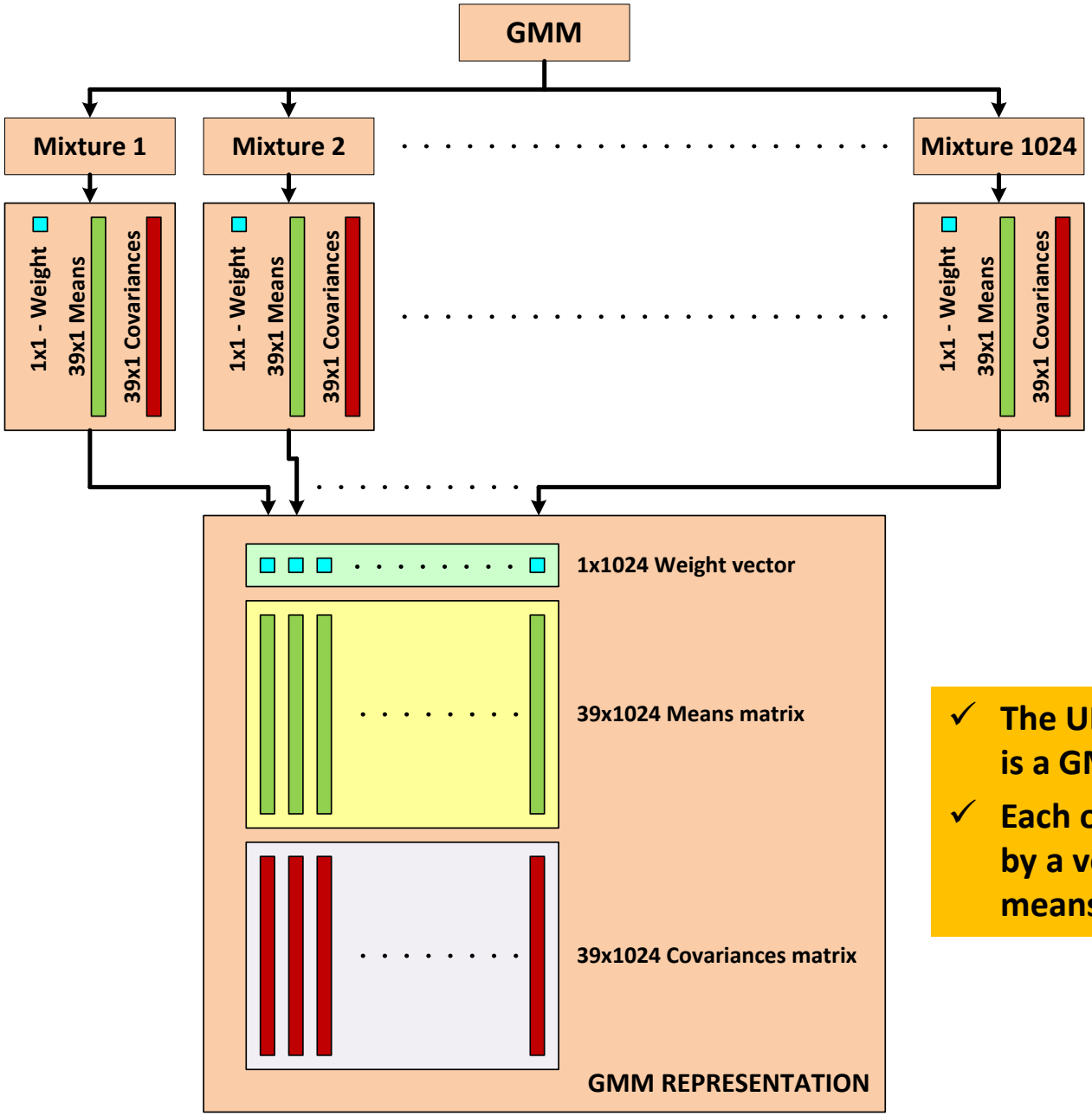
Universal Background Model (UBM) consists of 1024 Gaussian mixtures

- ✓ **Gaussian mixture consists of a mean (μ), covariance (Σ) and weight (*w*)**



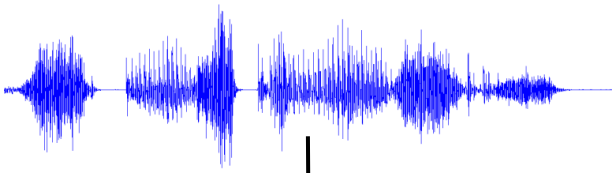
Target speaker model consists of 1024 Gaussian mixtures

Representing GMMs



- ✓ **The UBM and each speaker model is a GMM**
- ✓ **Each of them will be represented by a vector of weights, a matrix of means and a matrix of covariances**

Decision Making



Feature Extraction

Universal Background Models

Generic Male

Generic Female

Determine level of Match

Likelihood of Generic Male

Determine level of Match

Likelihood of John

Speaker Models

Speaker 1 Model

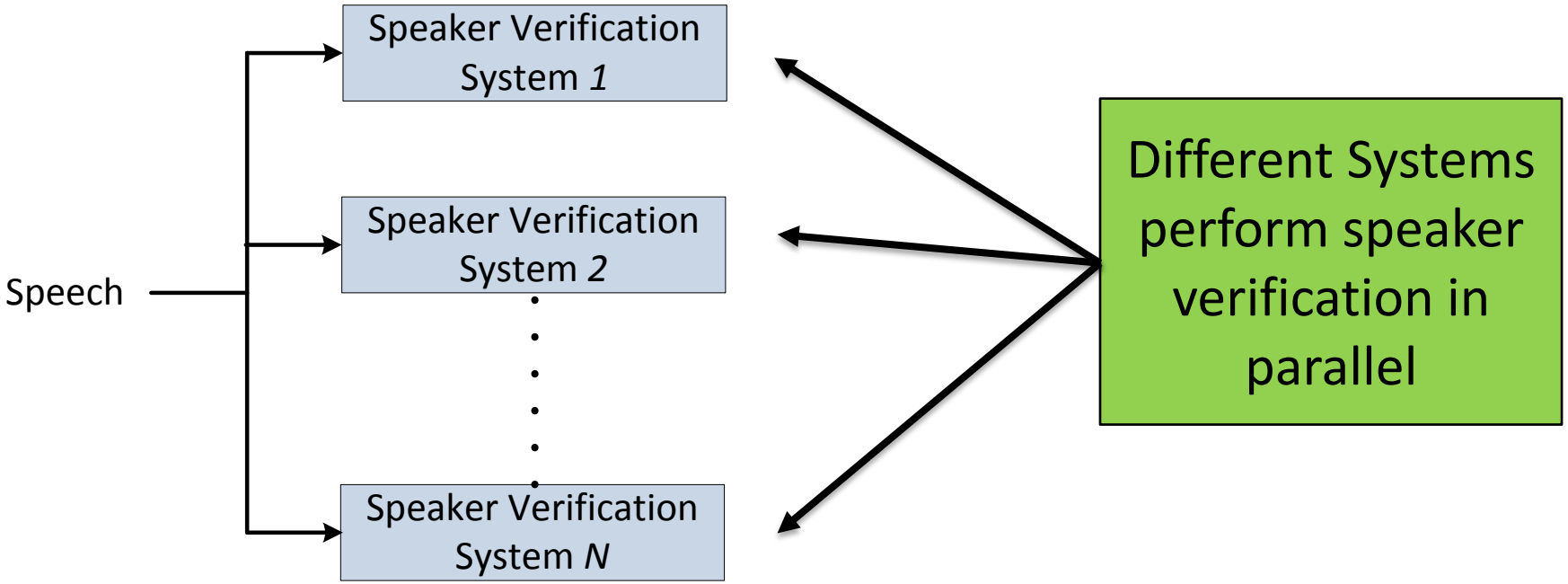
John's Model

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

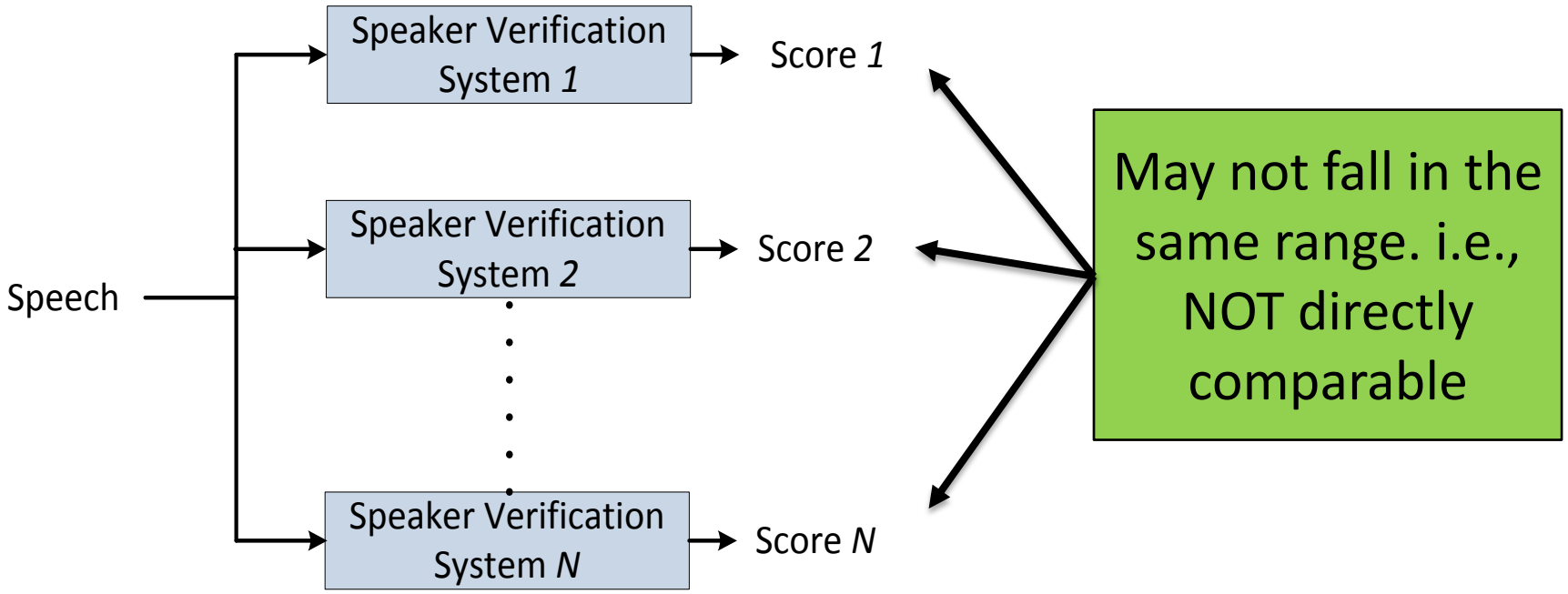
$L \lessgtr \theta$

Reject/Accept

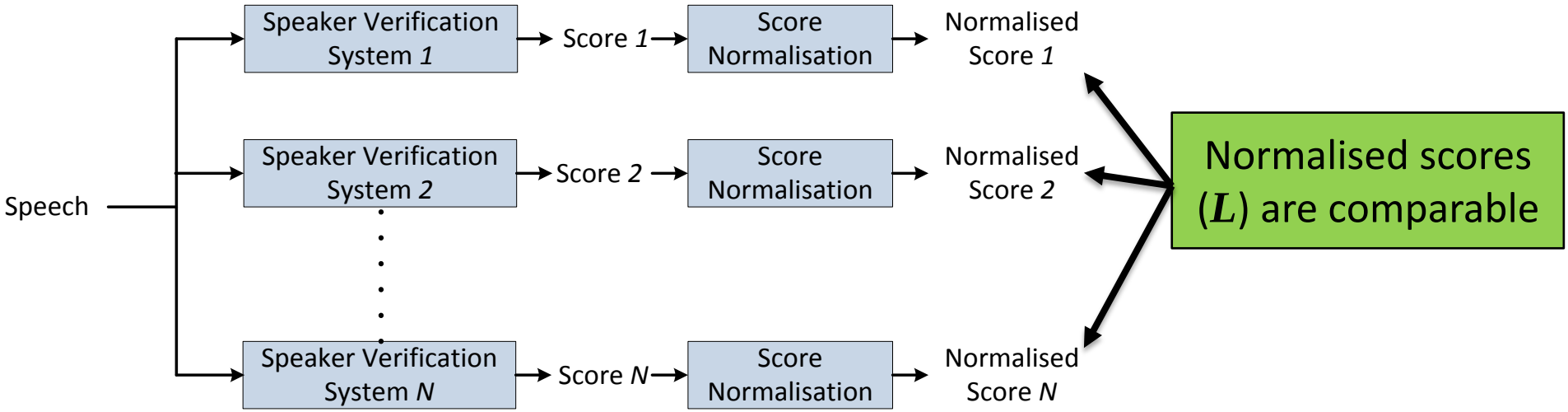
Score Normalisation



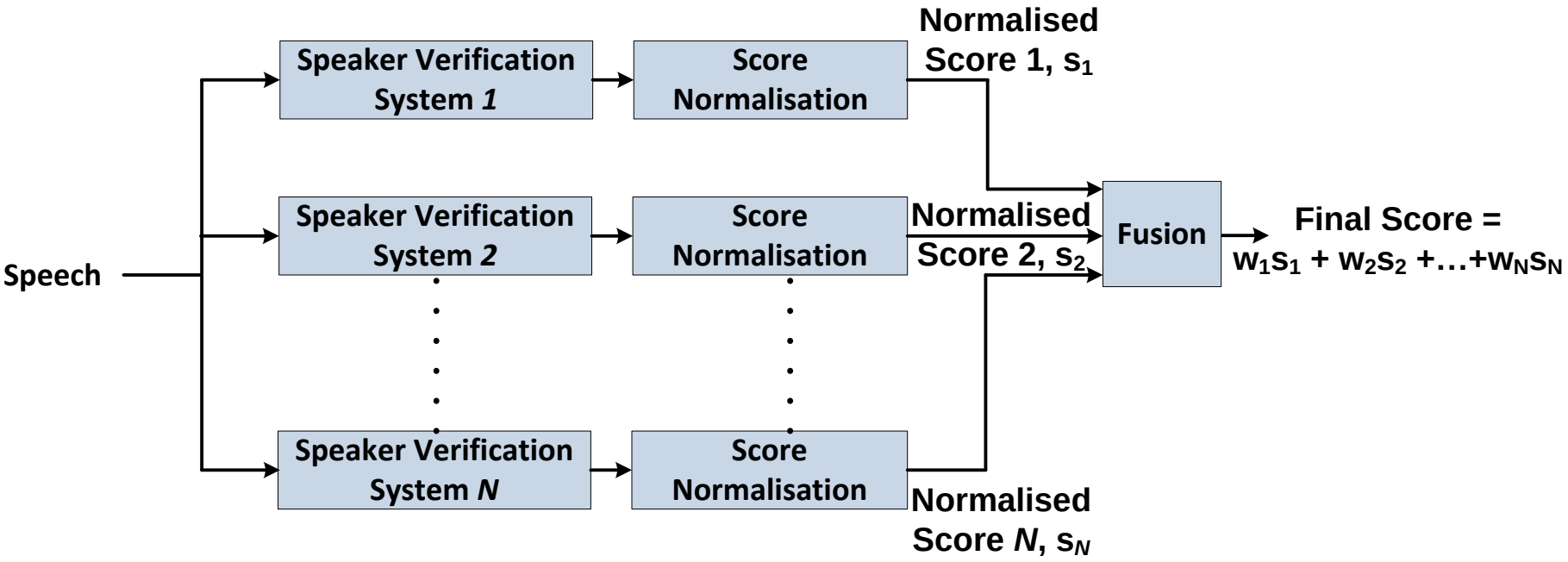
Score Normalisation



Score Normalisation



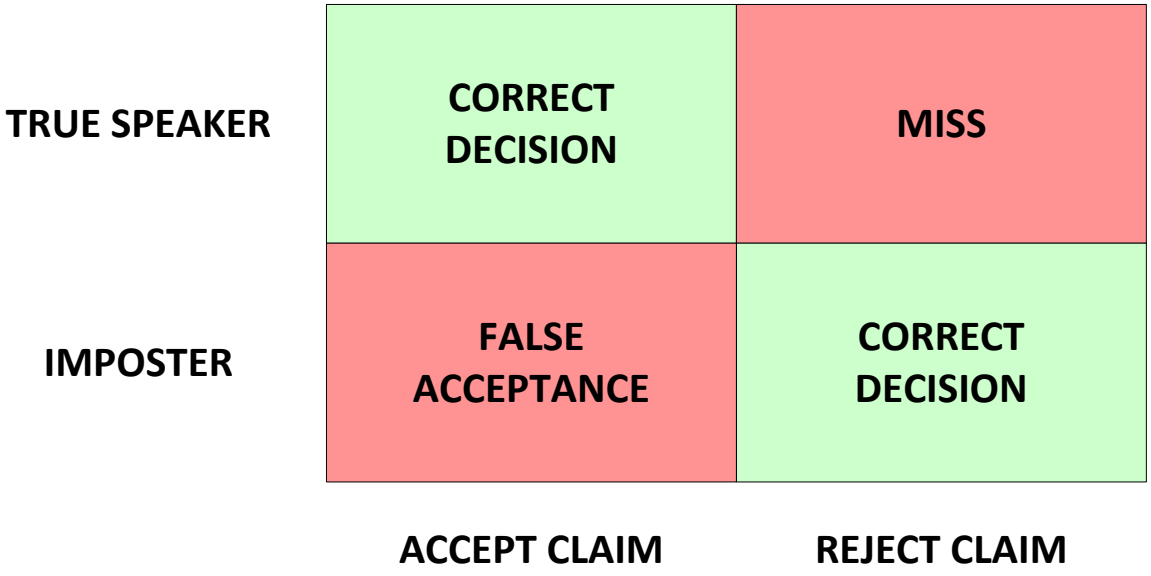
Fusion



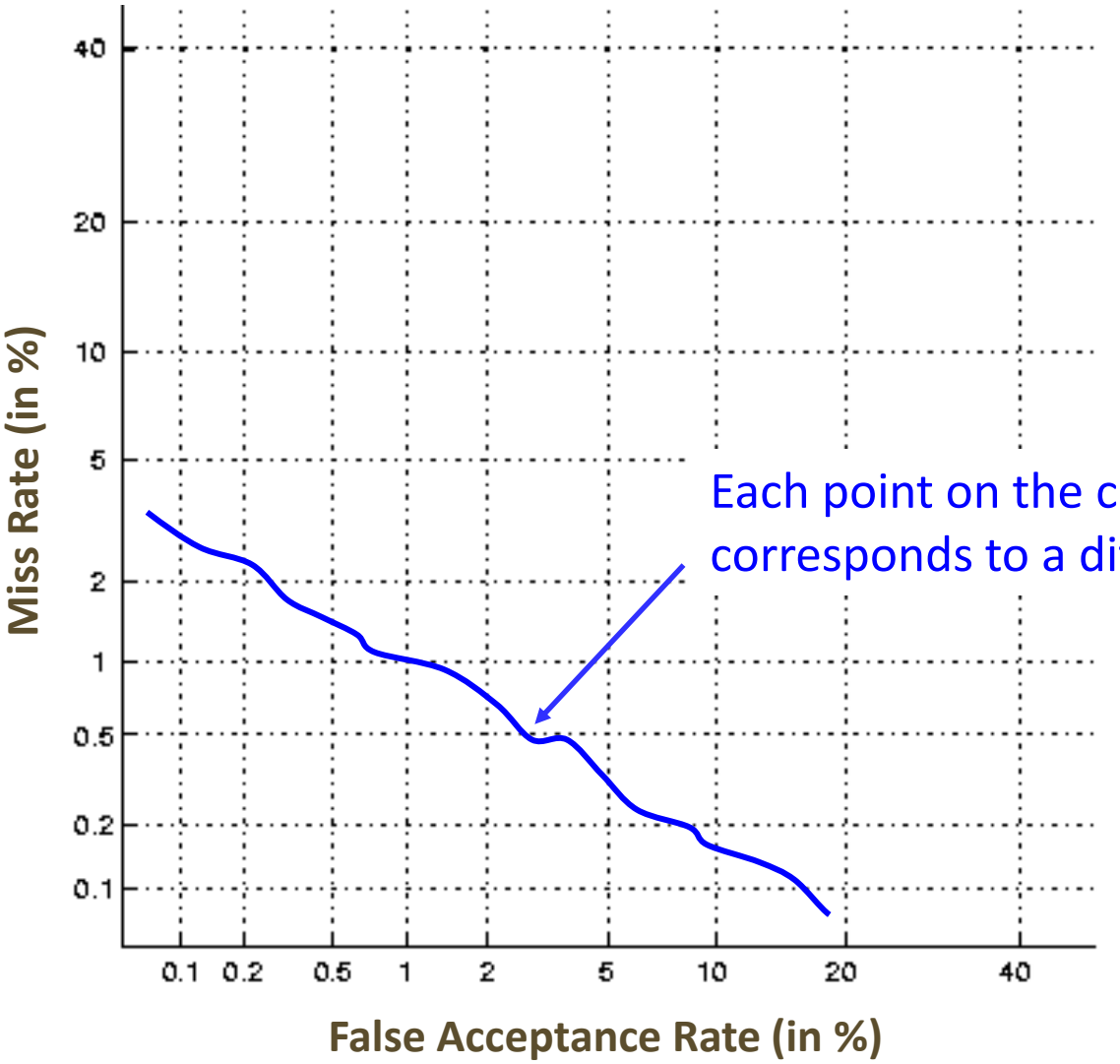
Final score will be a weighted sum of score from each system

Performance measure

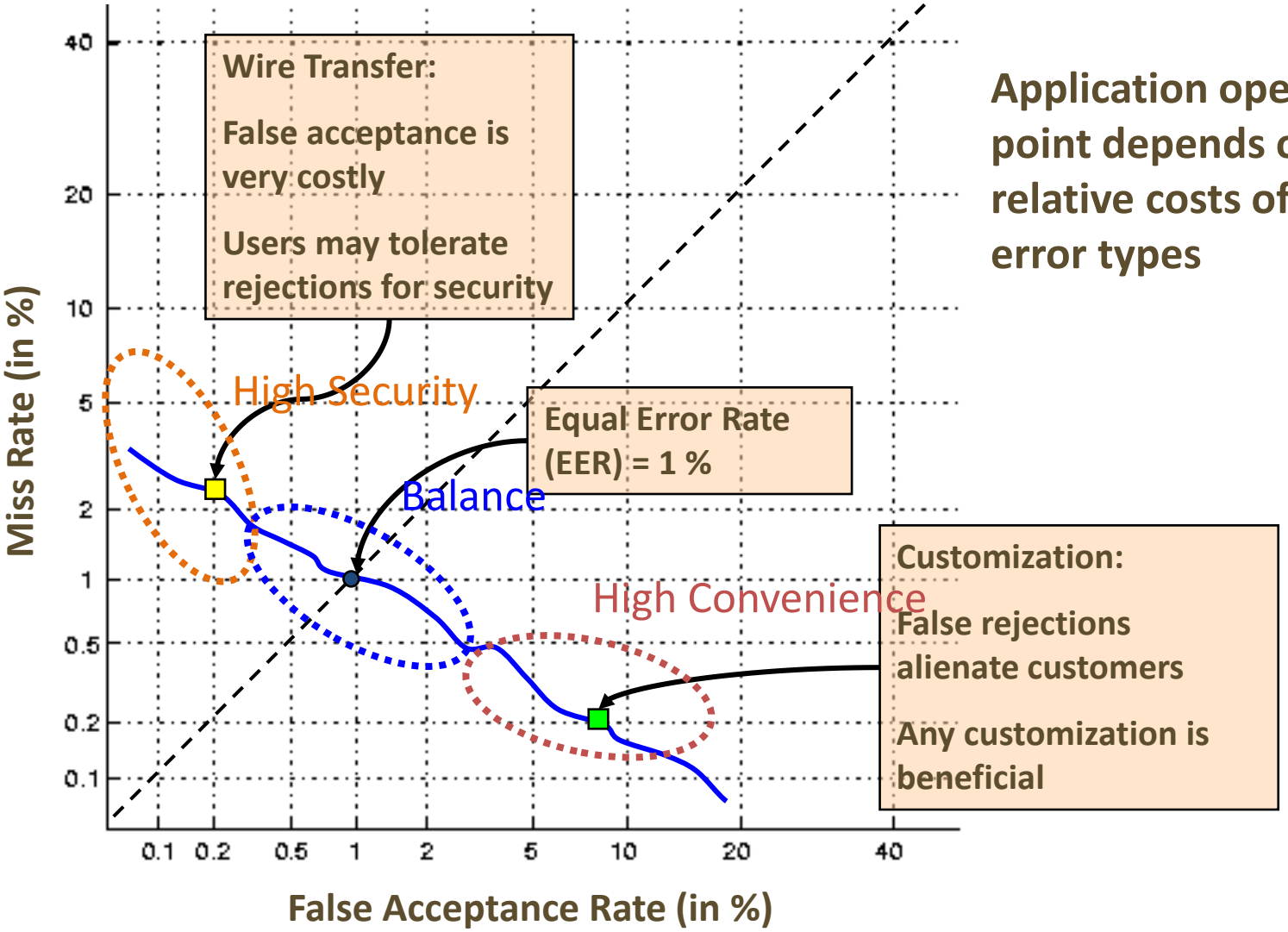
- Types of error:
  - Misses: valid identity is rejected
    - Probability of miss: ratio of the number of falsely rejected speaker tests to the total number of correct speaker trials.
  - False alarms: invalid identity is accepted
    - Probability of false alarm: ratio of the number of falsely accepted speaker tests to the total number of impostor trials



Performance measure - Detection error trade-off (DET) curve



Performance measure - Detection error trade-off (DET) curve



Application operating point depends on relative costs of the two error types

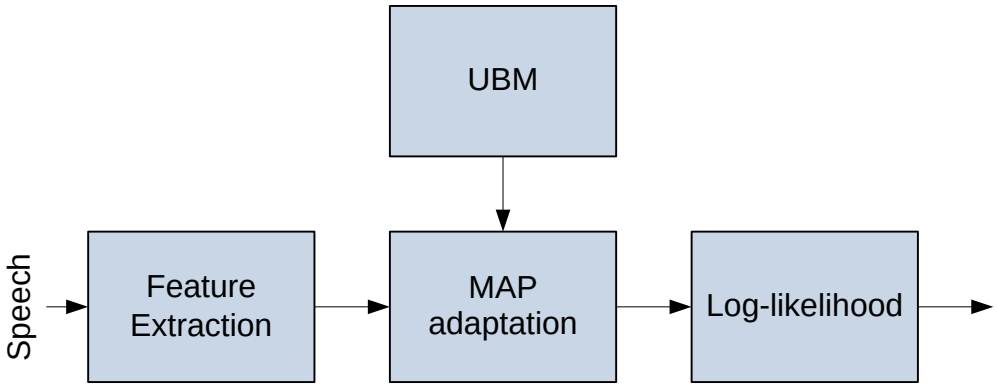
NIST Speaker Recognition Evaluation (SRE)

- Ongoing text independent speaker recognition evaluations conducted by NIST (http://www.itl.nist.gov/iad/mig/tests/spk/)
  - driving force in advancing the state-of-the-art
  - Conditions for different amounts of data
    - 10 sec.
    - 3-5 minutes
    - 8 minutes
    - Separate channel and summed channel conditions
  - English-speakers, non-English speakers, multilingual speakers

NIST SRE Trends

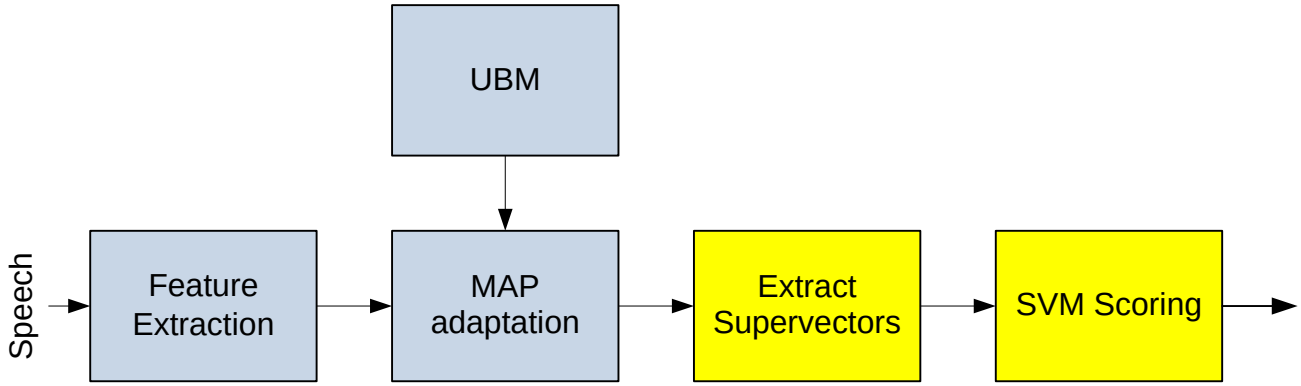
- 1996 – First SRE in current series
- 2000 – AHUMADA Spanish data, first non-English speech
- 2001 – Cellular data, Automatic Speech Recognition (ASR) transcripts provided
- 2005 – Multiple languages with bilingual speakers, room mic recordings, cross-channel trials
- 2008 – Interview data
- 2010 – High and low vocal effort, aging, HASR (Human-Assisted Speaker Recognition) Evaluation
- 2012 – Broad range of test conditions, with added noise and reverberation, target speakers defined beforehand

Basic System



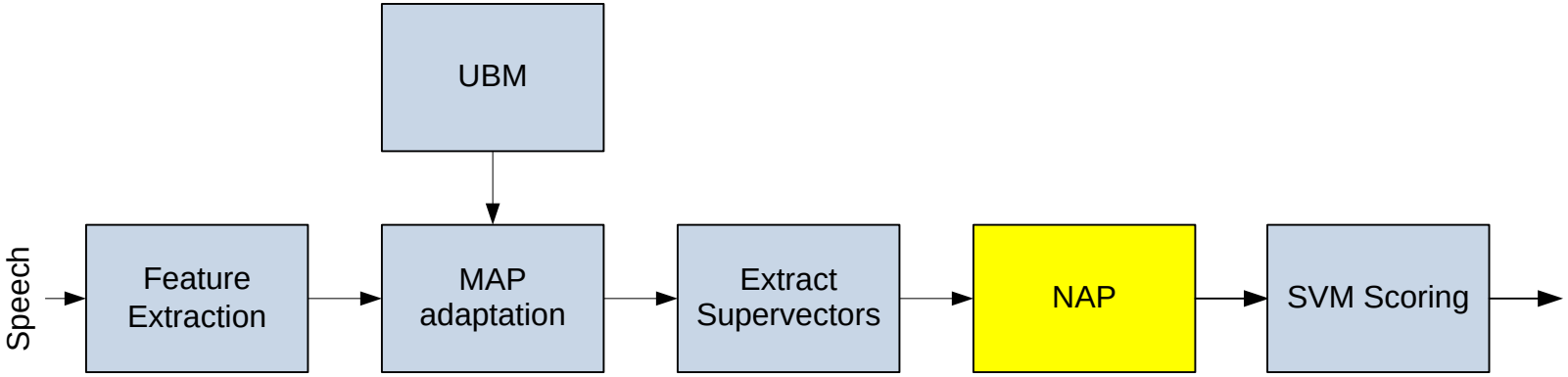
Trends

- In 2004's: Classification



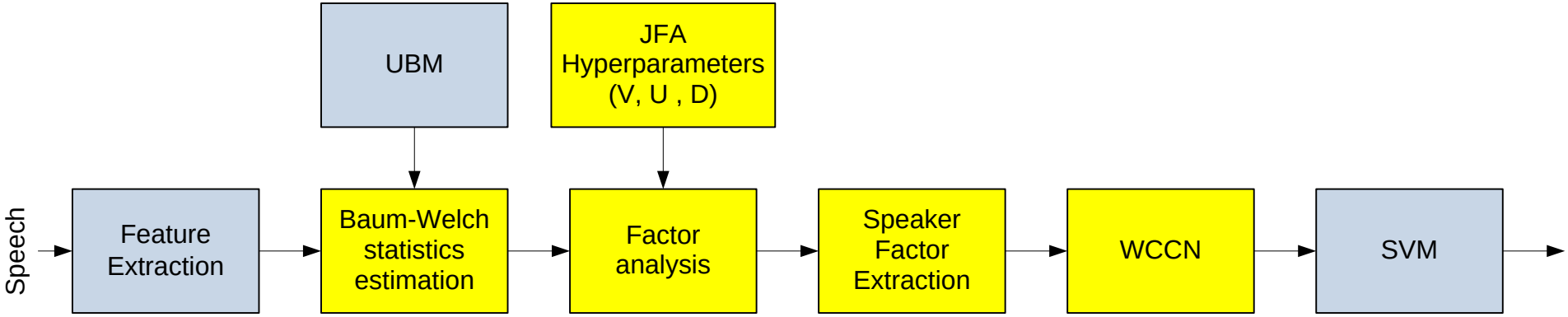
Trends

- In 2005's: Channel compensation - NAP



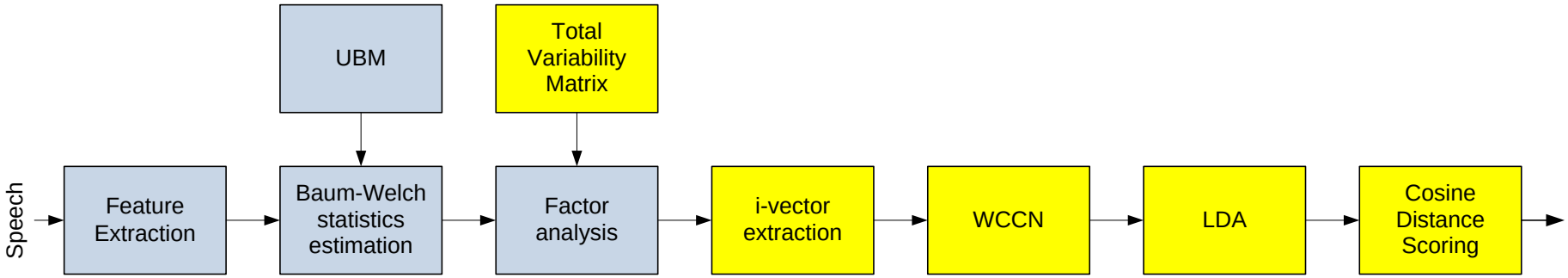
Trends

- In 2007's: Channel compensation - JFA



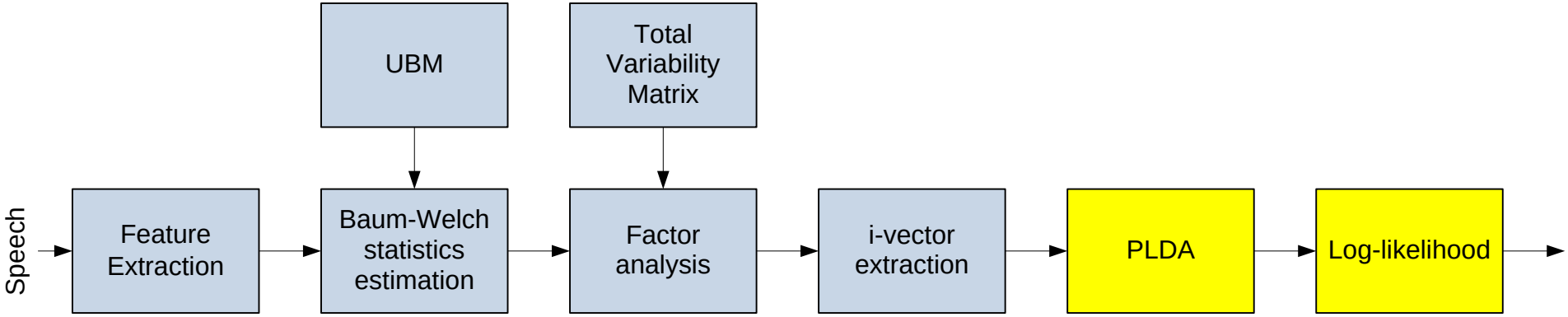
Trends

- In 2009's: Channel compensation – i-vector



Trends

- In 2009's: Channel compensation – PLDA




Q&A