Speaker Recognition

Sadaoki Furui

Tokyo Institute of Technology

Department of Computer Science

furui@cs.titech.ac.jp

Outline

- Speaker recognition methods
  - Applications of speaker recognition technology
  - Classification of speaker recognition methods
  - Structure of speaker recognition systems
  - Intra-speaker variation and evaluation of speaker recognition features
  - VQ, HMM (hidden Markov model) and GMM (Gaussian mixture model)
- Increasing robustness
  - Likelihood/distance normalization
  - Higher-level information
  - “Person authentication by voice: A need for caution”
- Multimodal speaker recognition
  - Feature fusion
  - Combination with face images
  - Combination with ear images
  - Fusion of spectral envelope and fundamental frequency information
- Future trends

Speaker recognition

- **Speaker verification**: confirm the identity claim (banking transactions, database access services, security control for confidential information)
- **Speaker identification**: determine from registered speakers (criminal investigations)

"voice key"

- **Text-dependent** methods
- **Text-independent** methods

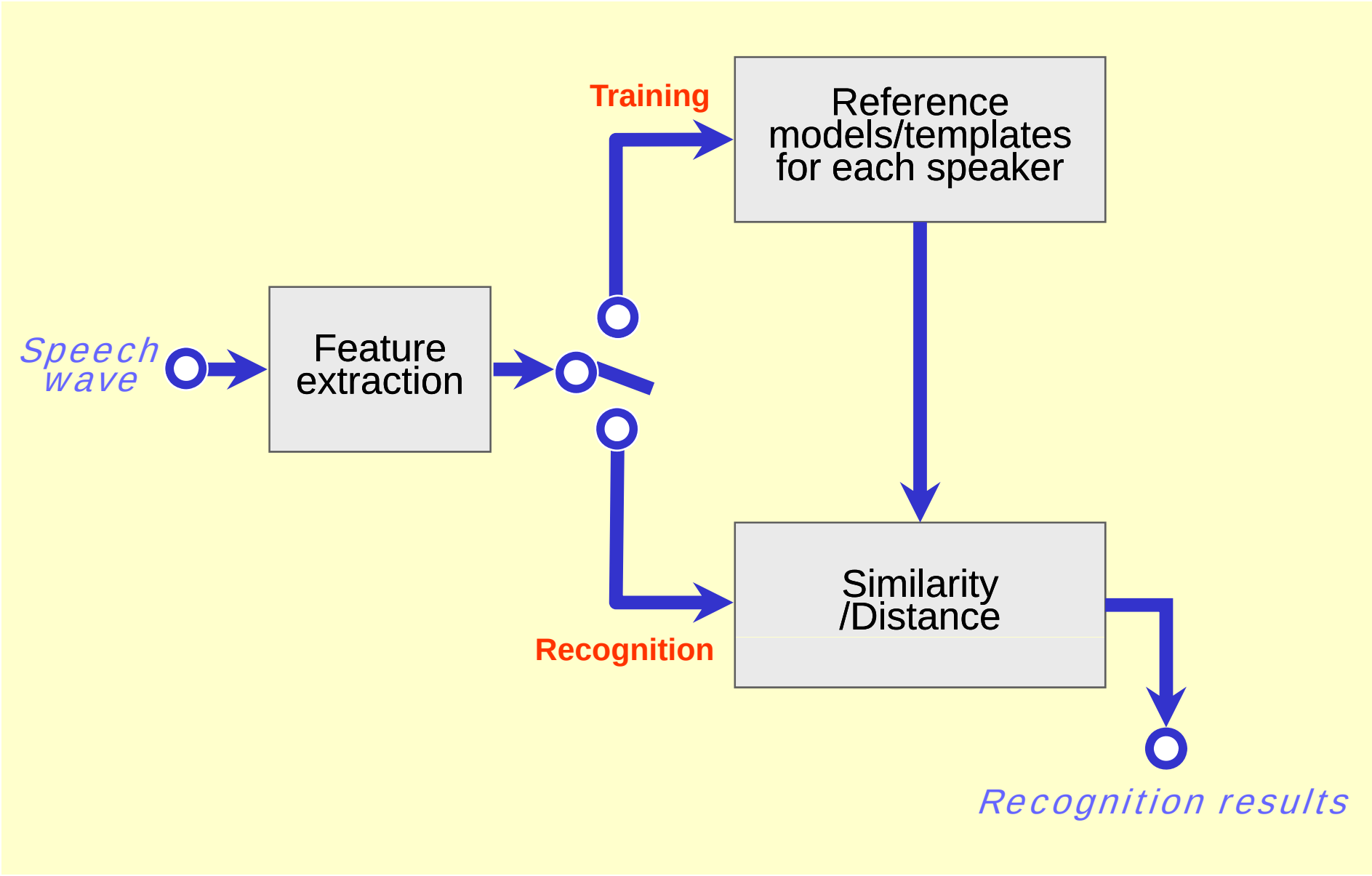
- Intersession variability (variability over time) of speech waves and spectra

⇒ Spectral/likelihood equalization (normalization)

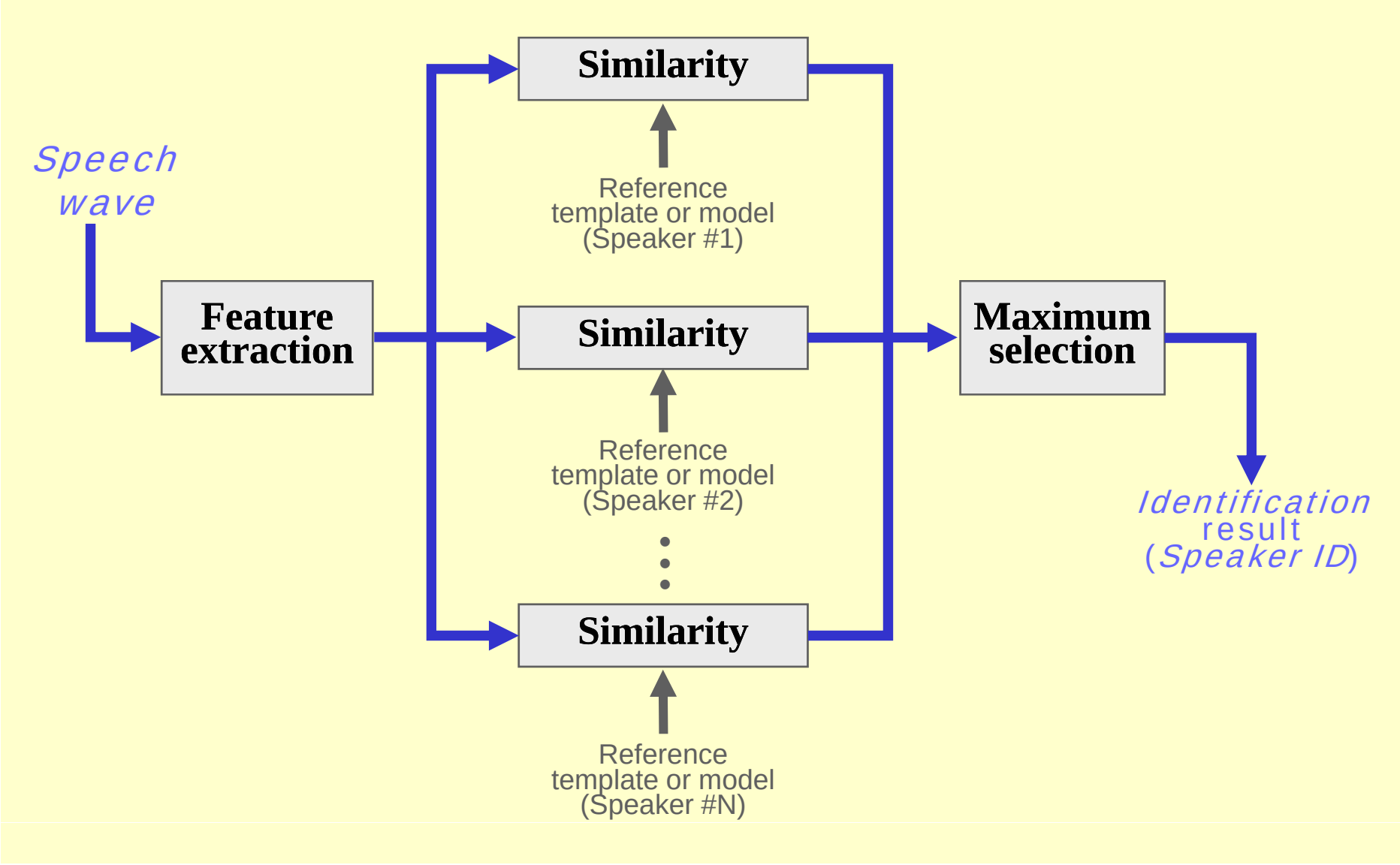
Applications of speaker recognition technology

- **Access control**: For physical facilities, computer networks, websites and automated password reset services.
- **Transaction authentication**: For telephone banking and remote electronic and mobile purchases (e- and m-commerce).
- **Law enforcement**: Home-parole monitoring, prison call monitoring and corroborating aural/spectral inspections of voice samples for forensic analysis.
- **Speech data management**: Label incoming voice mail with speaker name for browsing and/or action. Annotate recorded meetings or video with speaker labels for quick indexing and filing.
- **Personalization**: Store and retrieve personal setting/preferences for multi-user site or device. Use speaker characteristics for directed advertisement or services.

Principal structure of speaker recognition systems

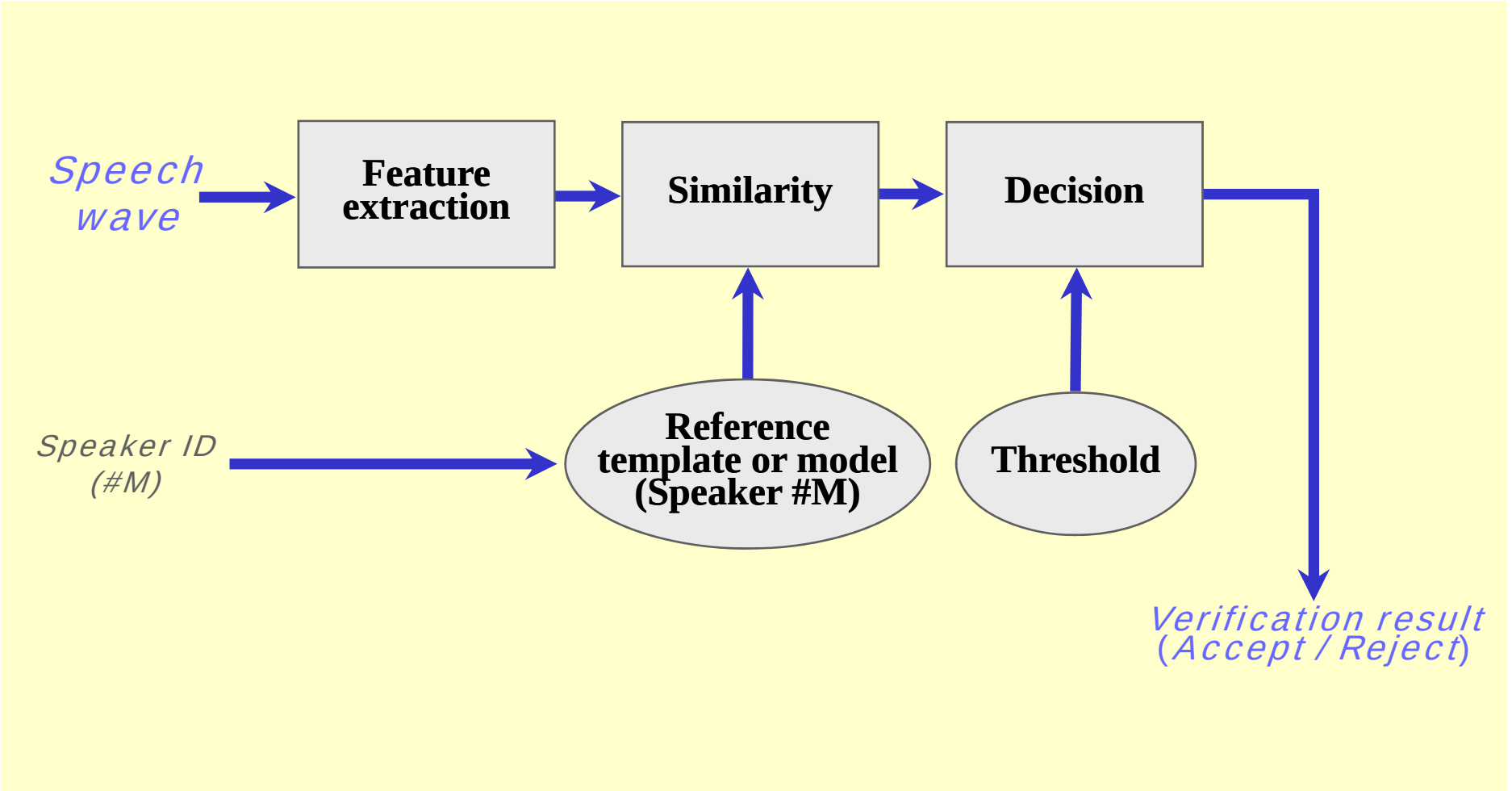


Basic structure of speaker recognition systems



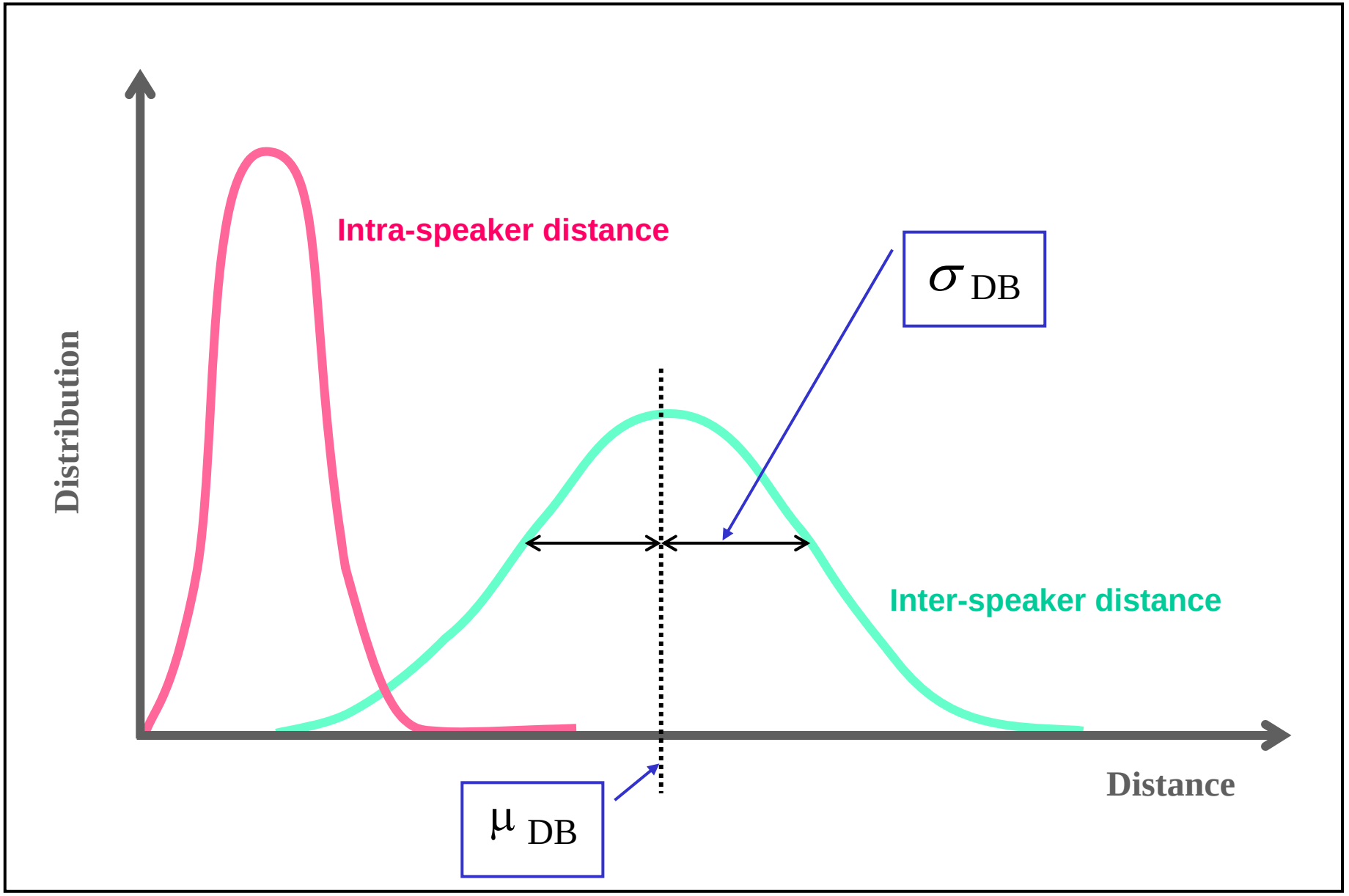
(a) Speaker identification

Basic structure of speaker recognition systems (cont.)



(b) Speaker verification

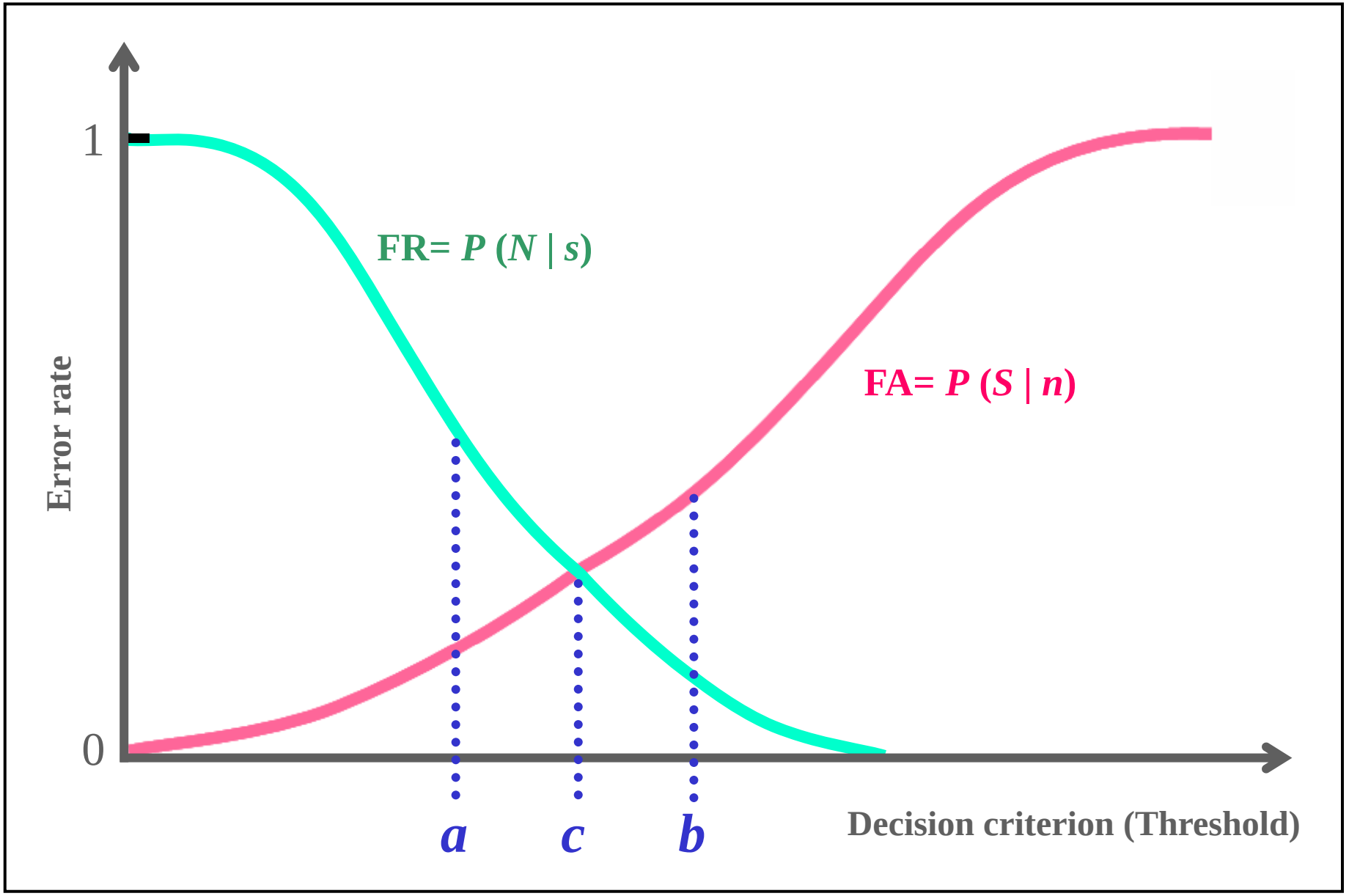
Example of typical intraspeaker and interspeaker distance distributions



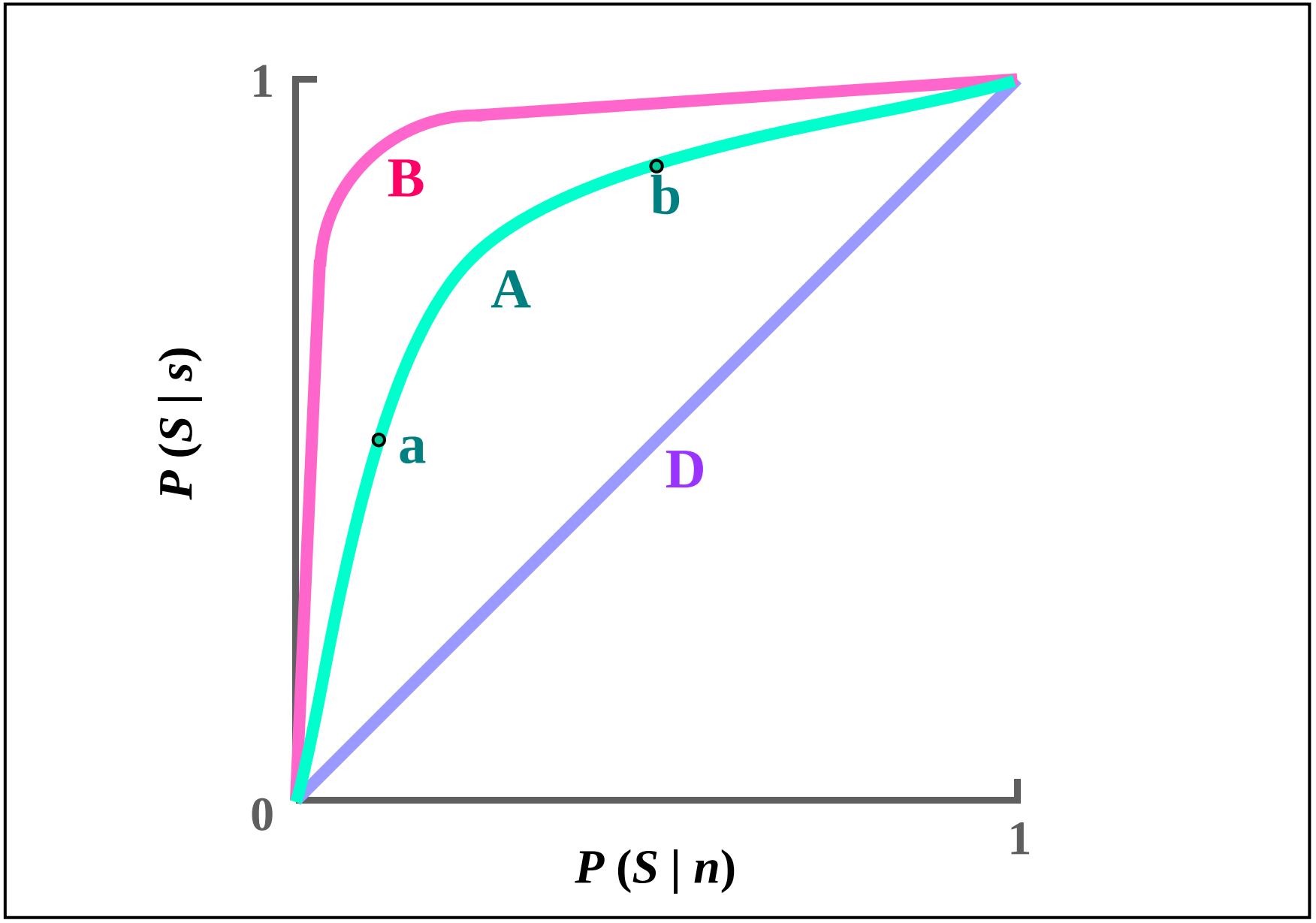
Four conditional probabilities in speaker verification

Decision condition \ Input utterance condition	s (customer)	n (impostor)
S (accept)	$P(S \mid s)$	$P(S \mid n)$
N (reject)	$P(N \mid s)$	$P(N \mid n)$

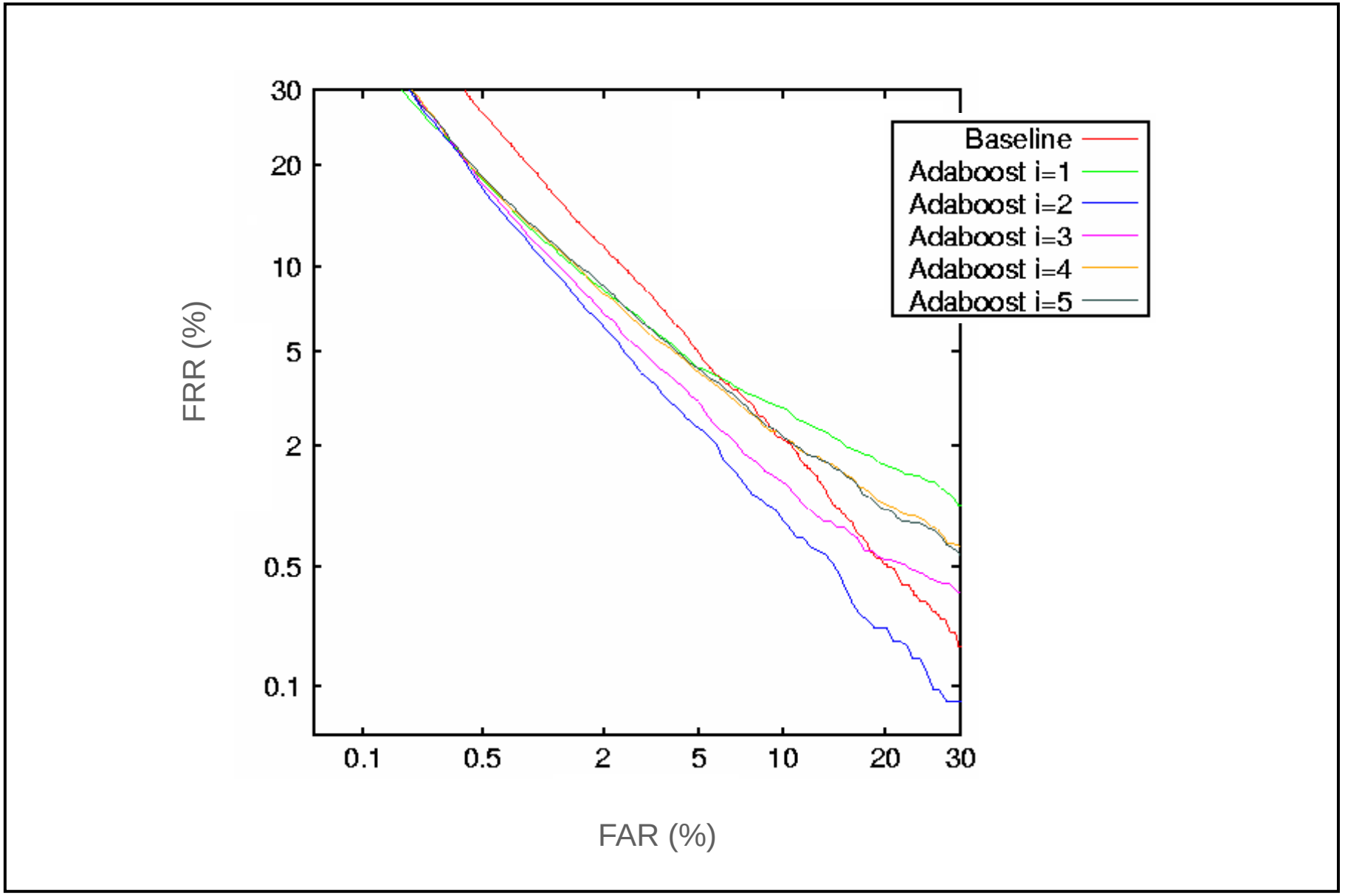
Relationship between error rate and decision criterion (threshold) in speaker verification



Receiver operating characteristic (ROC) curves; performance examples of three speaker verification systems: A, B, and D



Examples of the DET (detection error trade-off) curve
(Normal deviate scales)



NIST speaker recognition evaluation (C_{DET} cost)

$$C_{DET} = (C_{FR} \times P_{FR} \times P_C) + (C_{FA} \times P_{FA} \times P_{NonC})$$

P_{FR} : False rejection rate

P_{FA} : False acceptance rate

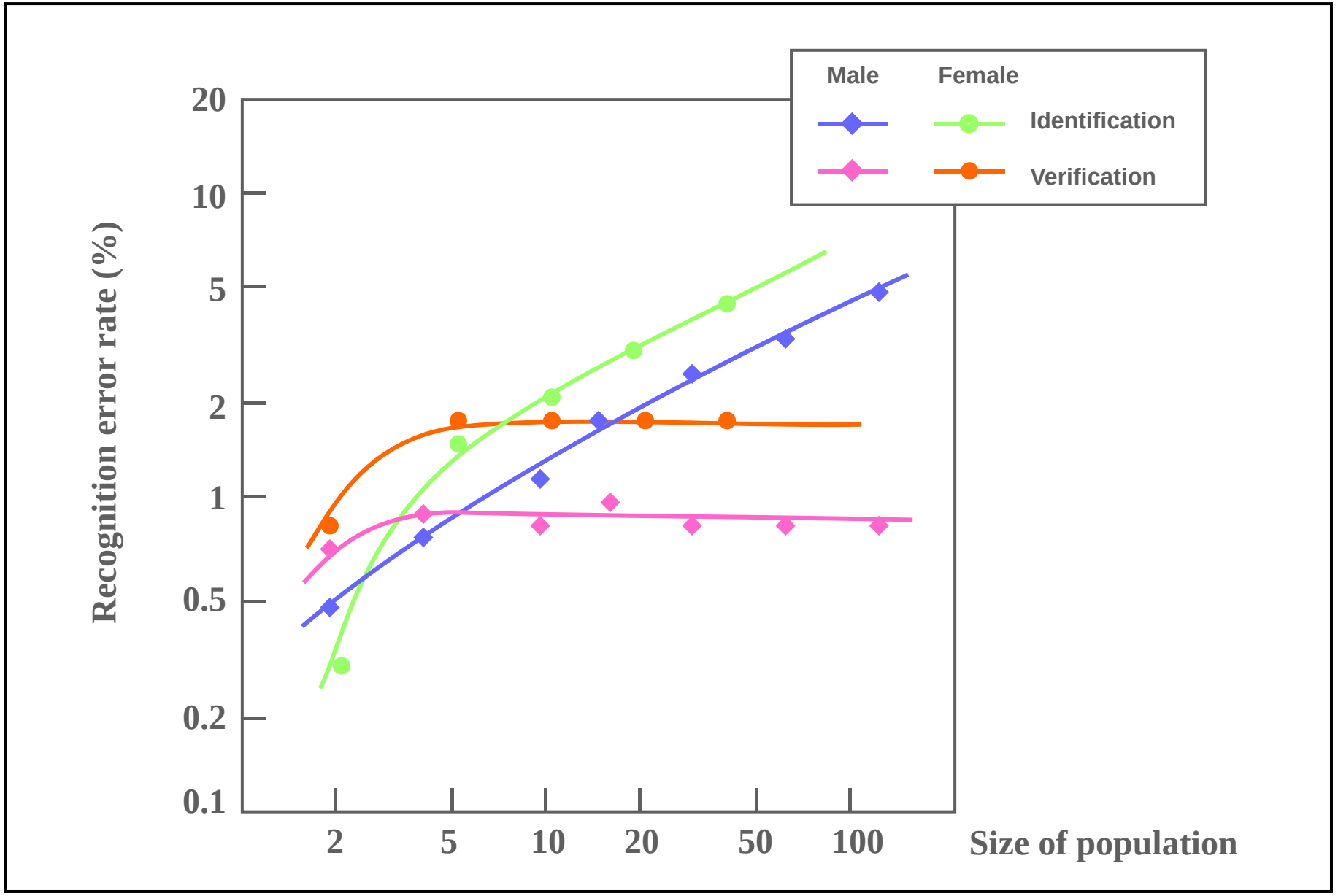
C_{FR} : Cost of false rejection (10)

C_{FA} : Cost of false acceptance (1)

P_C : A priori probability of a customer (0.01)

$P_{NonC} = 1 - P_C$

Recognition error rates as a function of population size in speaker identification and verification

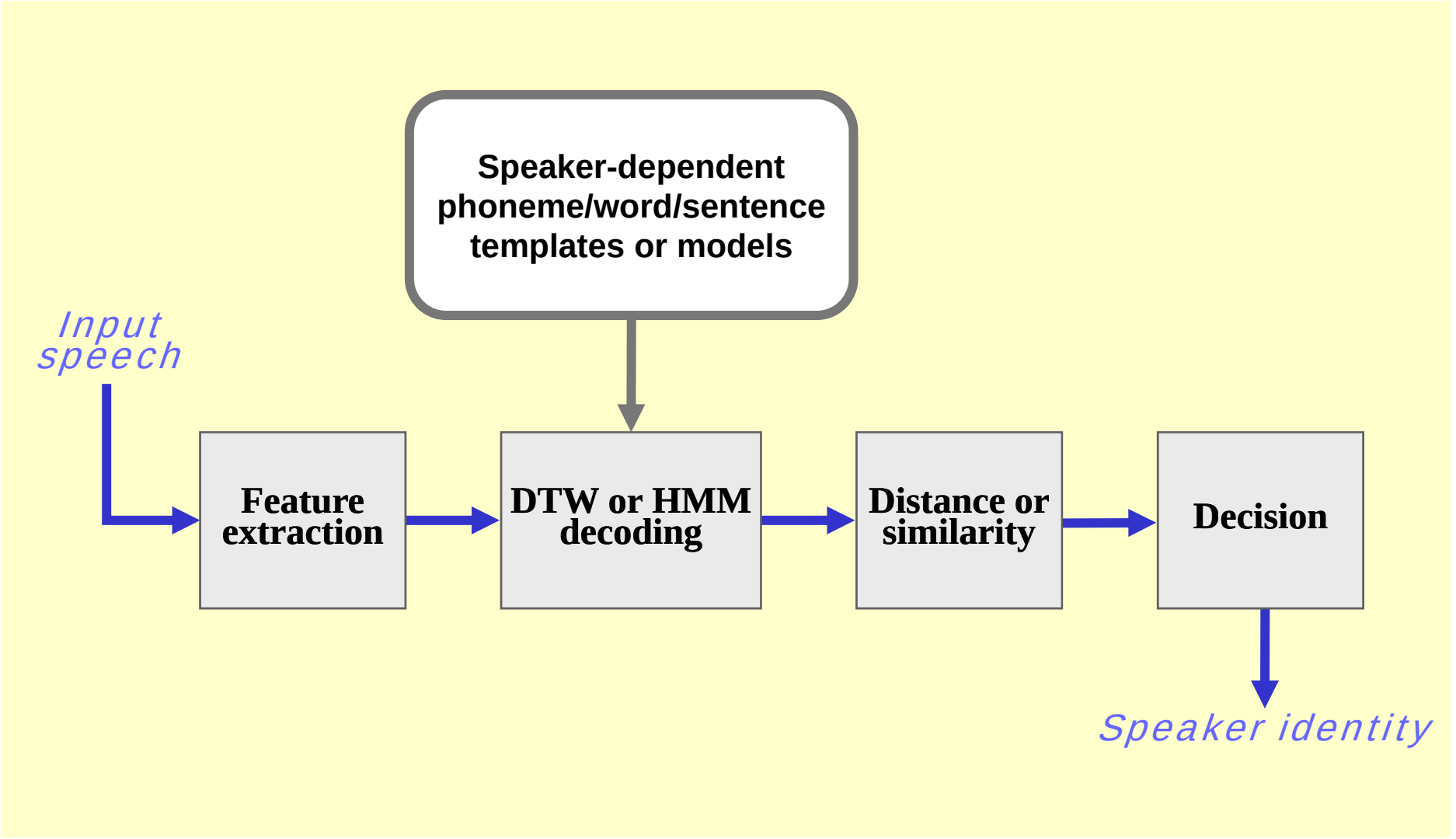


Text-dependent vs. text-independent methods

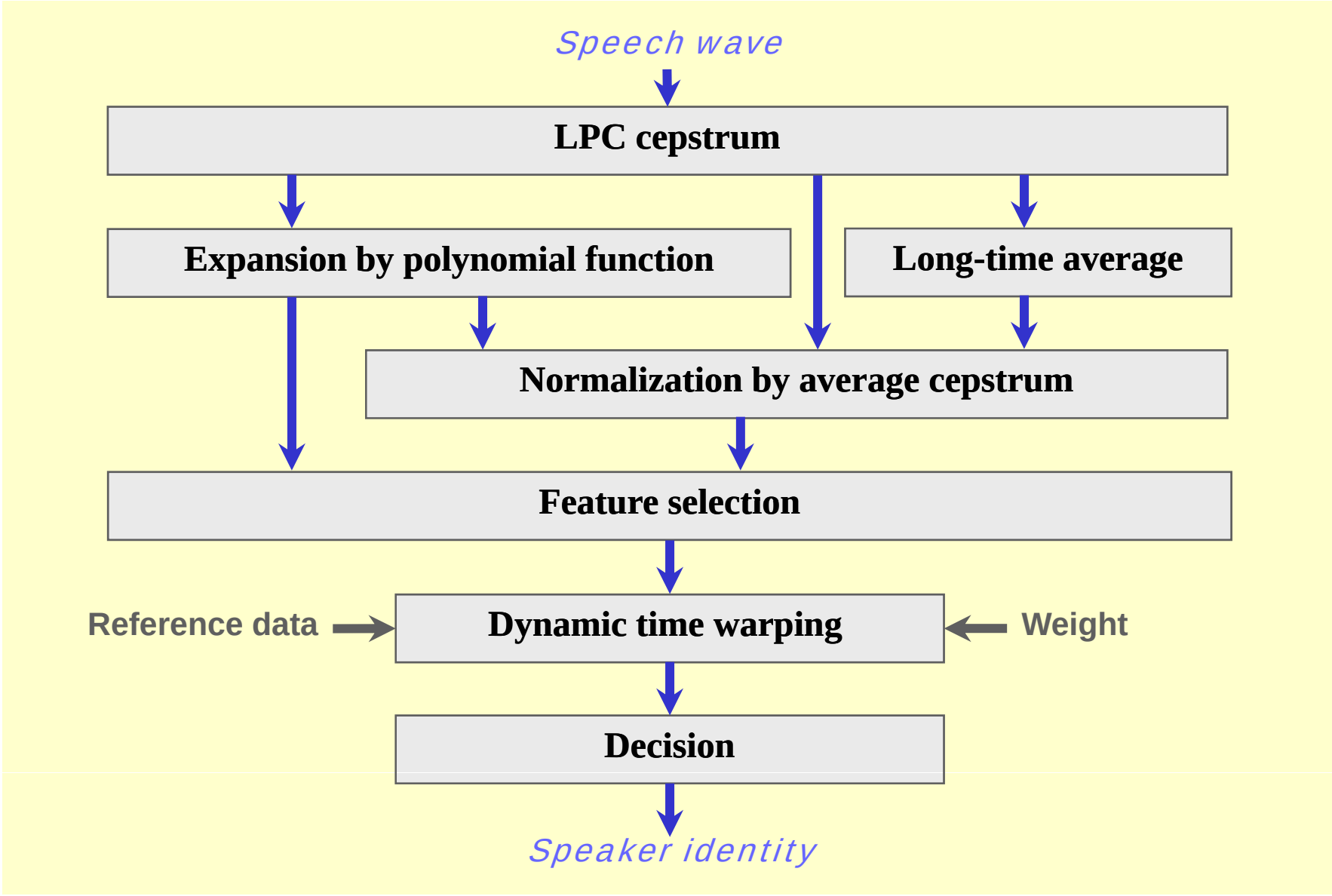
Text-dependent methods are usually based on template matching techniques. The structure of the systems is, therefore, rather simple. Since this method can directly exploit the voice individuality associated with each phoneme or syllable, it generally achieves higher recognition performance than the text-independent method.

Text-independent methods can be used in several applications in which predetermined key words cannot be used. Another advantage is that it can be done sequentially, until a desired significance level is reached, without the annoyance of repeating the key words again and again.

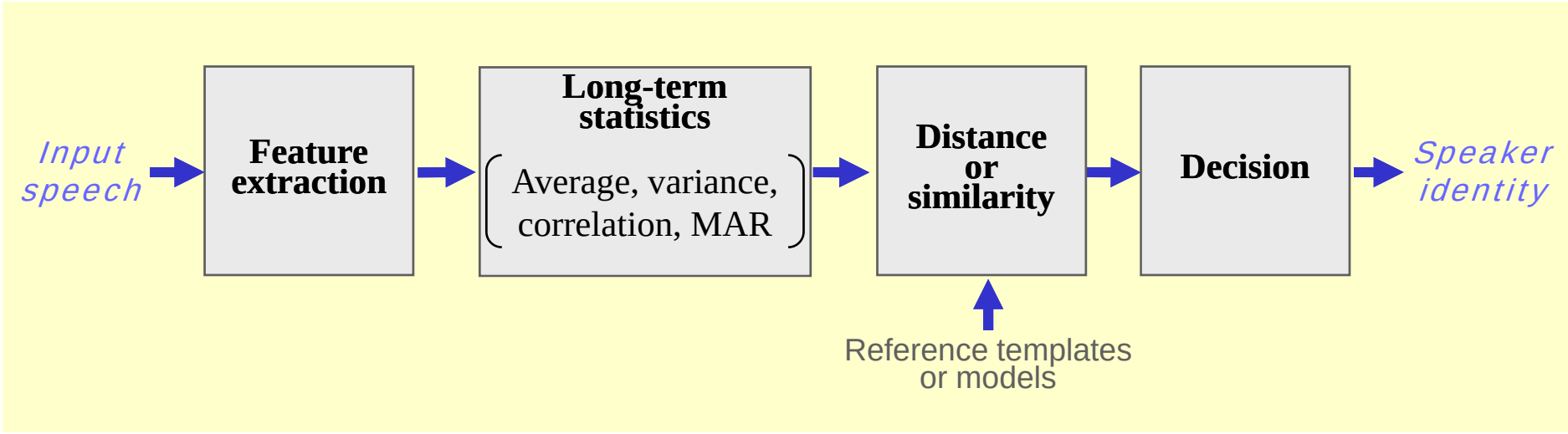
Basic structure of DTW/HMM-based text-dependent speaker recognition methods



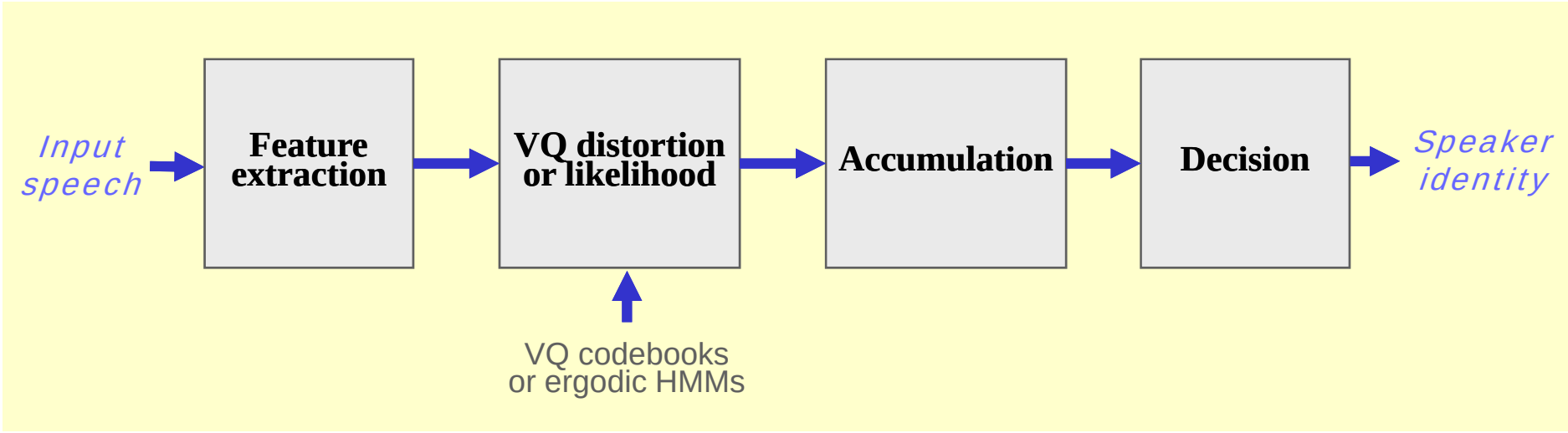
Block diagram indicating principal operation of speaker recognition method using time series of cepstral coefficients and their orthogonal polynomial coefficients



Basic structures of text-independent speaker recognition methods



(a) Long-term-statistics-based method

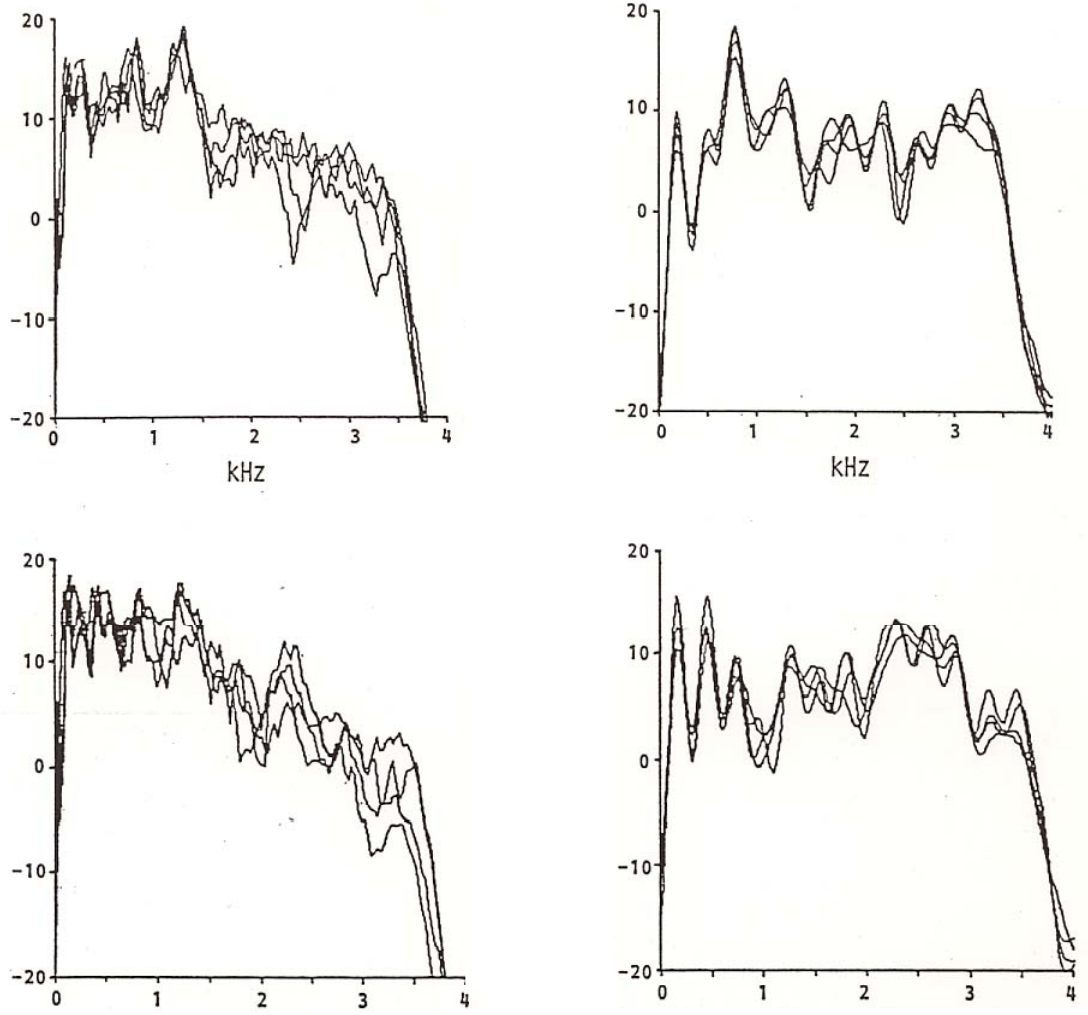


(b) VQ/HMM-based method

Variation of the long-time averaged spectrum at four sessions over eight months, and corresponding spectral envelopes derived from cepstrum coefficients weighted by the square root of inverse variances

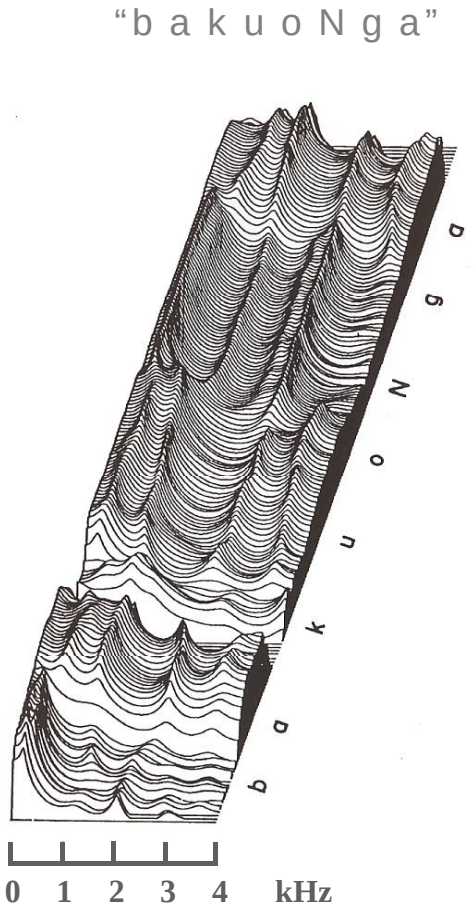
Sub. W

Sub. T

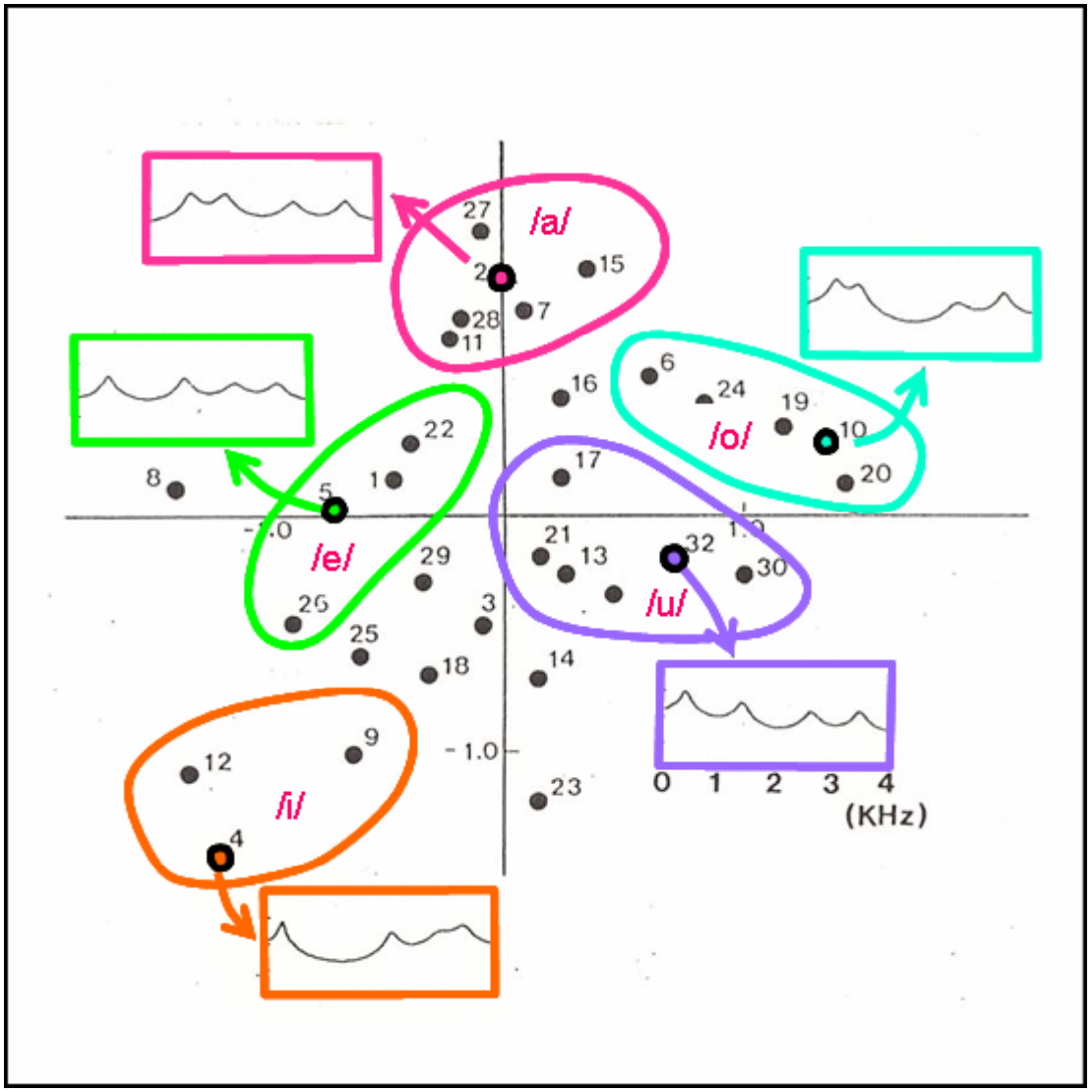


(a) Long-time averaged spectra (b) Envelopes by weighted cepstrum

Vector quantization (VQ)-based text-independent speaker recognition

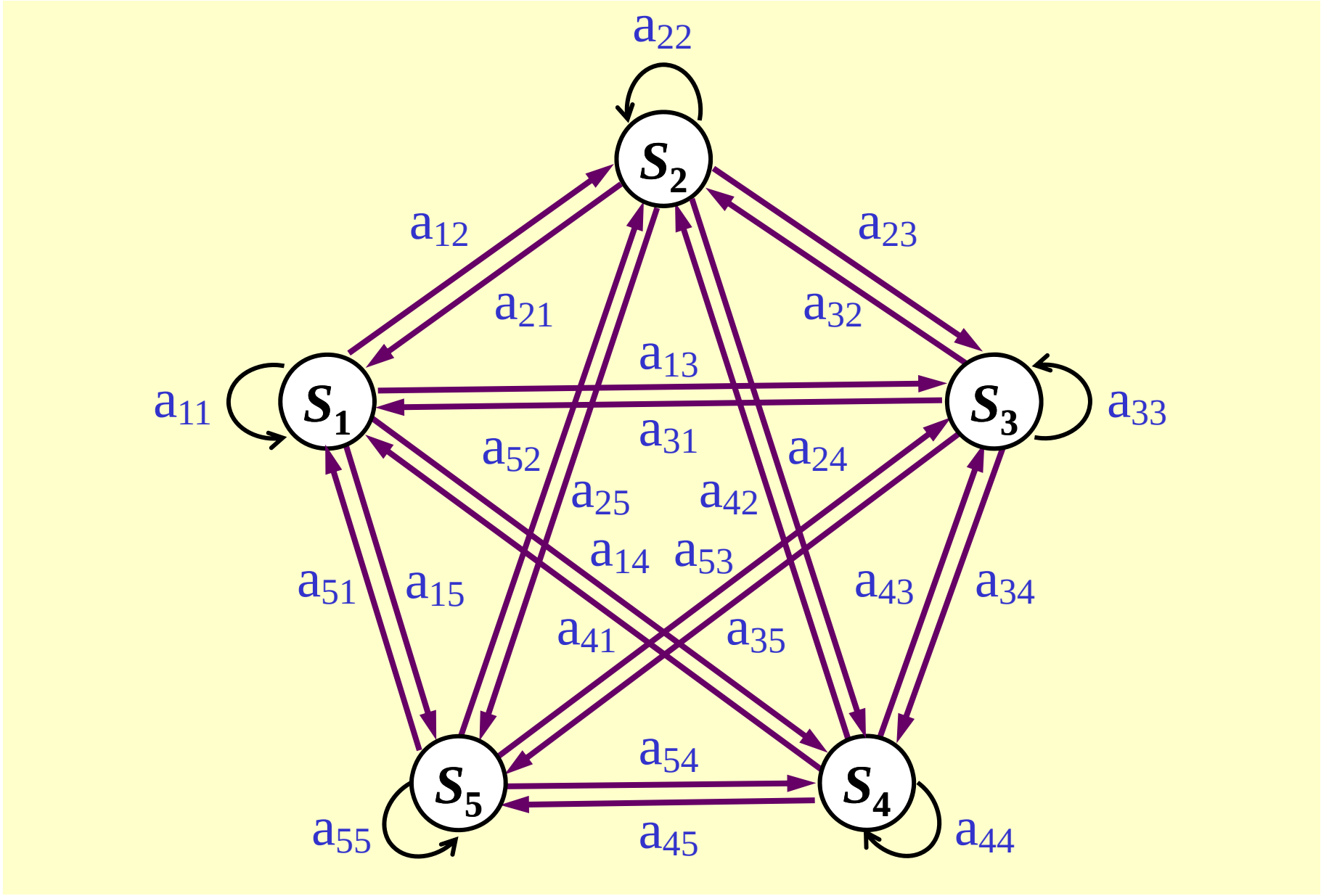


Spectral envelopes

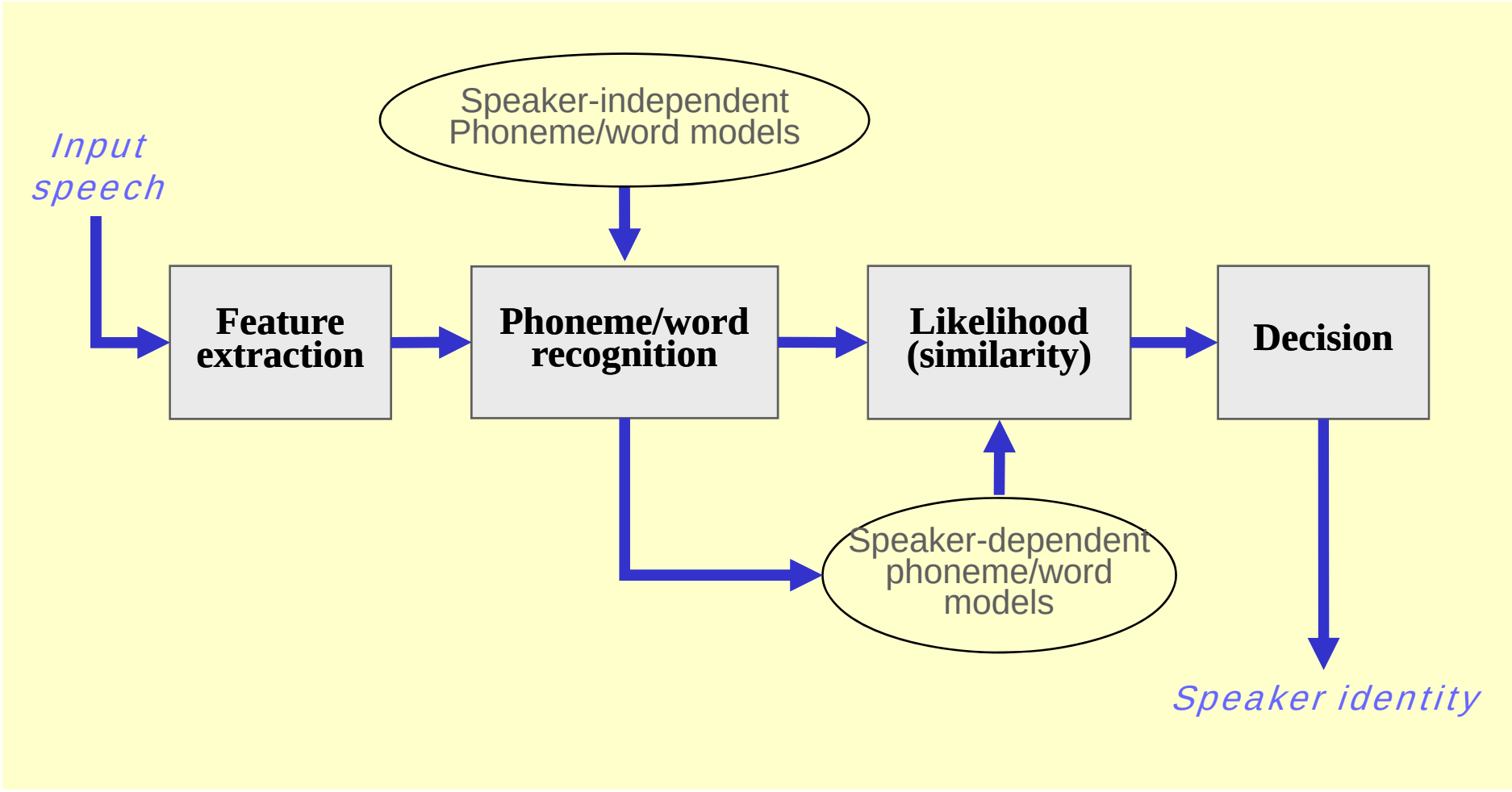


Speaker-specific codebook

A five-state ergodic HMM for text-independent speaker verification



Basic structures of text-independent speaker recognition methods (cont.)



(c) Speech-recognition-based method

Text-prompted speaker recognition method

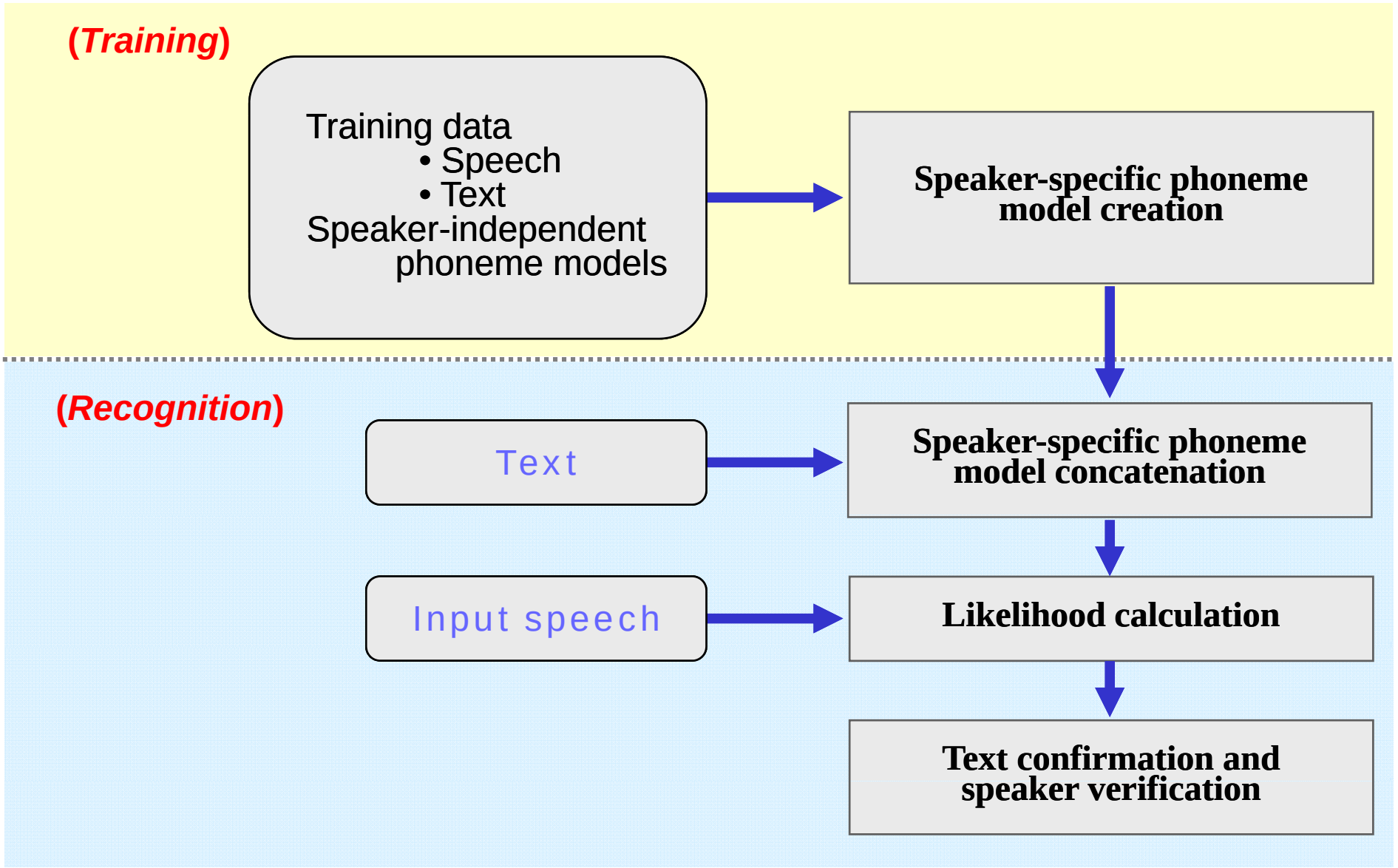
This method is facilitated by using **speaker-specific phoneme models** as basic acoustic units.

The recognition system prompts each user with a new key sentence every time the system is used, and accepts the input utterance only when it decides that the registered speaker has uttered the prompted sentence.

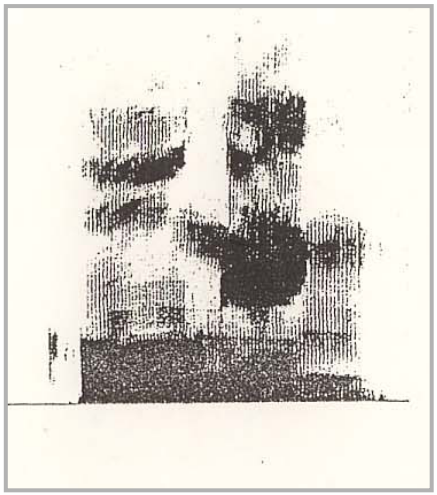
Because the vocabulary is unlimited, prospective impostors cannot know in advance what sentence they will be asked to repeat. Thus a pre-recorded voice can easily be rejected.

One of the major issues in this method is how to properly create the speaker-specific phoneme models with training utterances of a limited size for each speaker.

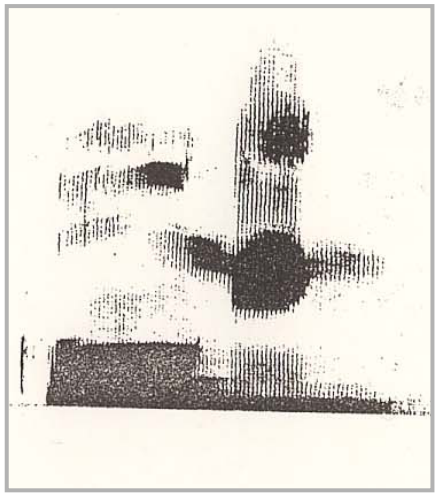
Block diagram of the text-prompted speaker recognition method



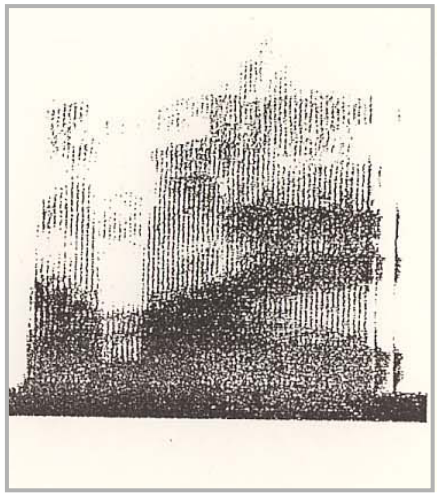
Sound spectrograms for word utterances by several speakers



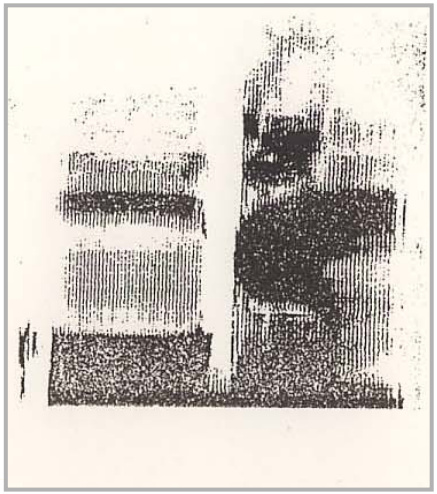
Speaker S /ko:geN/



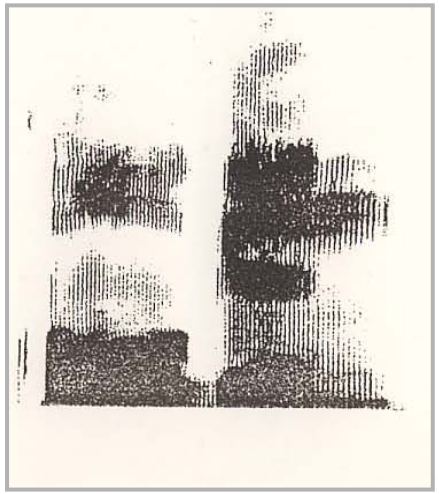
Same (2 years later)



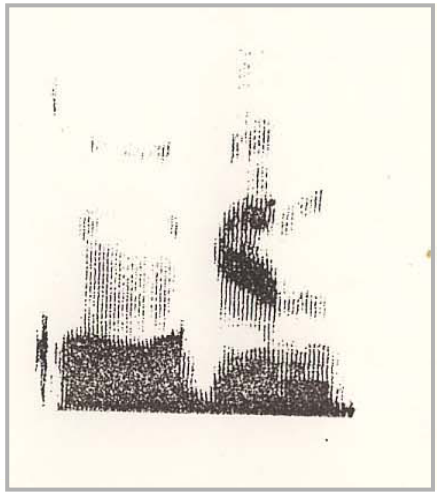
Speaker S /baNgo:/



Speaker M /ko:geN/



Speaker F /ko:geN/



Speaker U /ko:geN/

Intersession variability (variability over time)

- Speakers
- Recording and transmission conditions
- Noise

Normalization

- Parameter domain
- Distance/similarity domain

Parameter-domain normalization

Cepstral mean normalization (subtraction) (CMN, CMS)

- Linear channel effects
- Long-term spectral variation

Delta-cepstral coefficients

Distance/similarity-domain normalization

- ***Likelihood ratio***

$$\log L(X) = \log p(X|S{=}S_c) - \log p(X|S{\neq}S_c)$$

S_c : claimed speaker

- ***A posteriori probability***

$$\log L(X) = \log p(X|S{=}S_c) - \log \sum_{S \in Ref} p(X|S)$$

Ref : reference speaker

Both are almost equally effective. Difference exists in whether or not the claimed speaker is included in the speaker set for normalization.

Cohort speakers

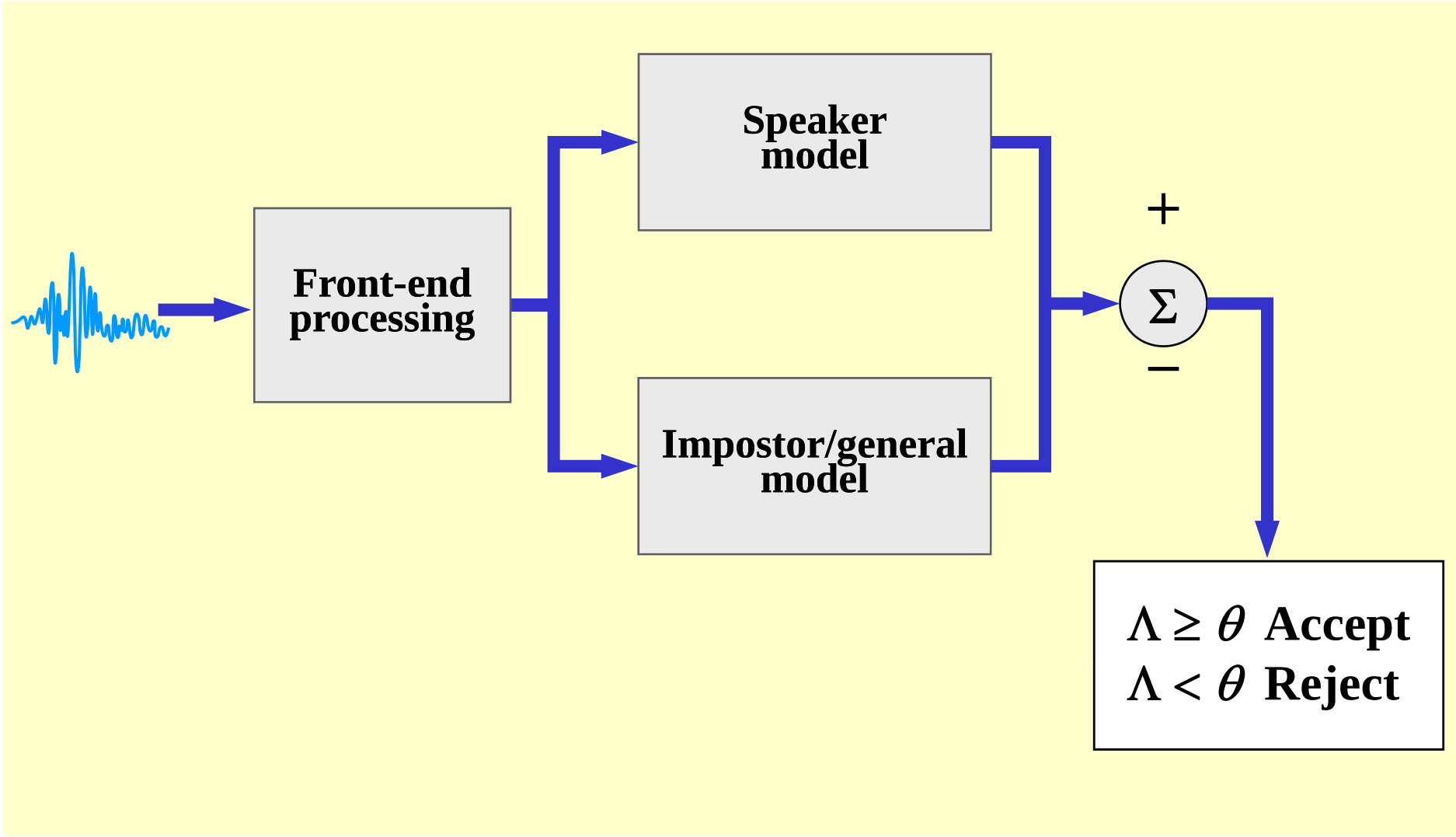
$$\log L(X) = \log p(X|S{=}S_c) - \log \sum_{S \in Cohort} p(X|S{\neq}S_c)$$

- Typical of the general population, or
- Population near the claimed speaker

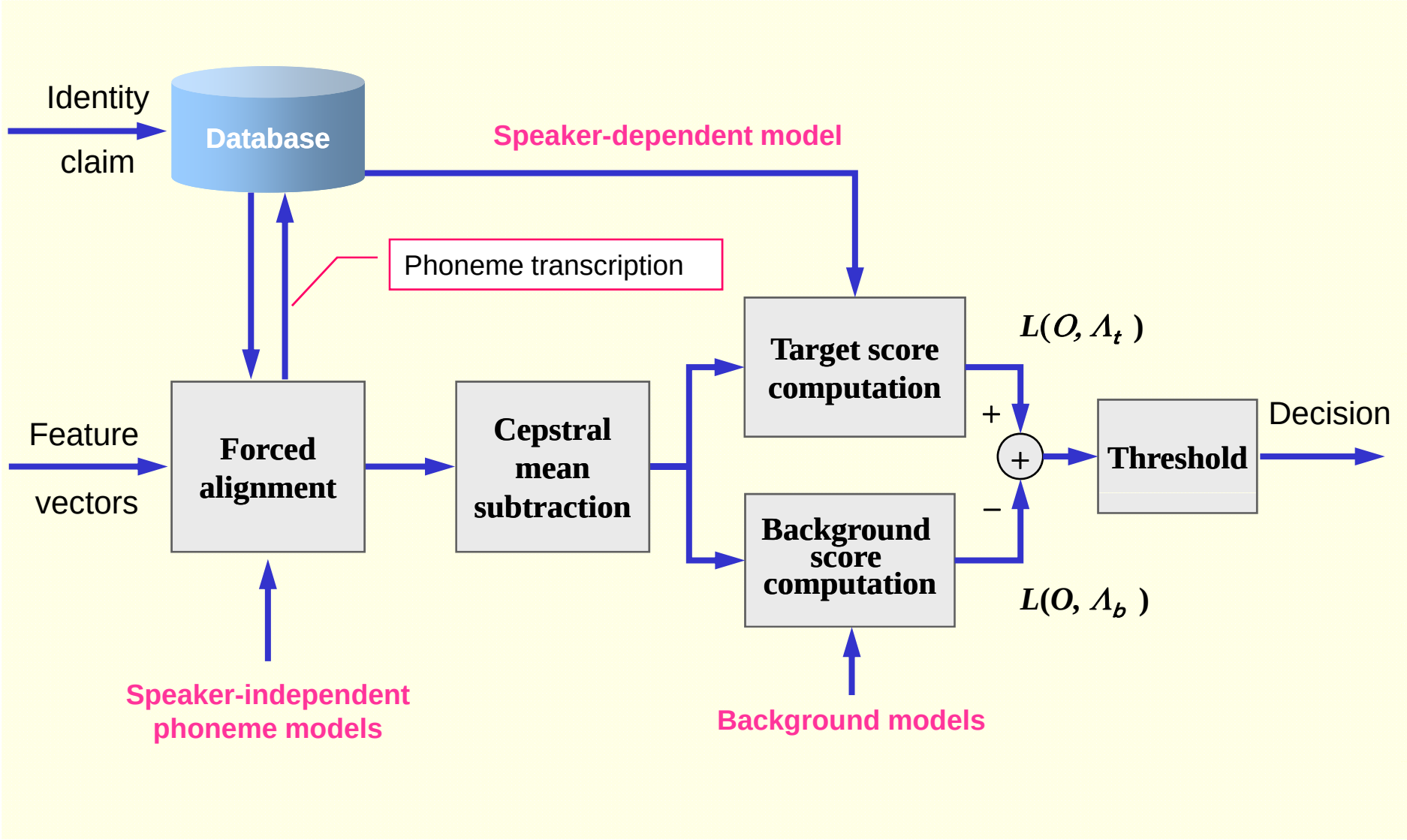
Normalization by a general/world model

- A Gaussian mixture which models the parameter distribution for free-text utterances by many speakers

Distance/similarity normalization by impostor/general model



A fixed-phrase speaker verification system



Z-Norm (Zero Normalization)

$$S_{Z-Norm} = \frac{S(X,U,I) - \mu_U}{\sigma_U}$$

$S(X,U,I)$: Score of an input utterance X obtained using a user model U and an impostor model I

μ_U, σ_U : Mean value and standard deviation of $S(X, U, I)$ for impostor utterances

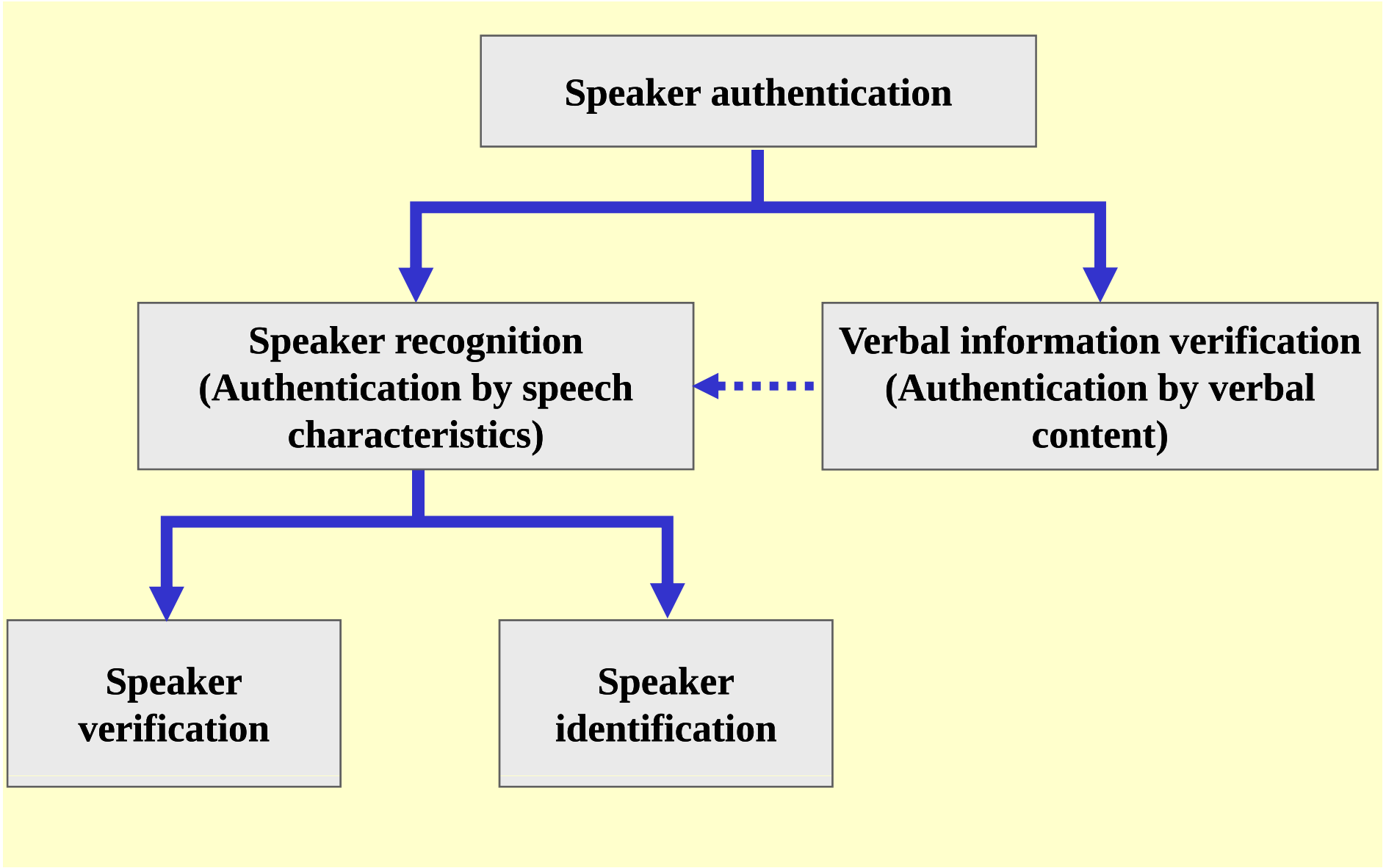
T-Norm (Test Normalization)

$$S_{T-Norm} = \frac{P(X|U) - \mu_X}{\sigma_X}$$

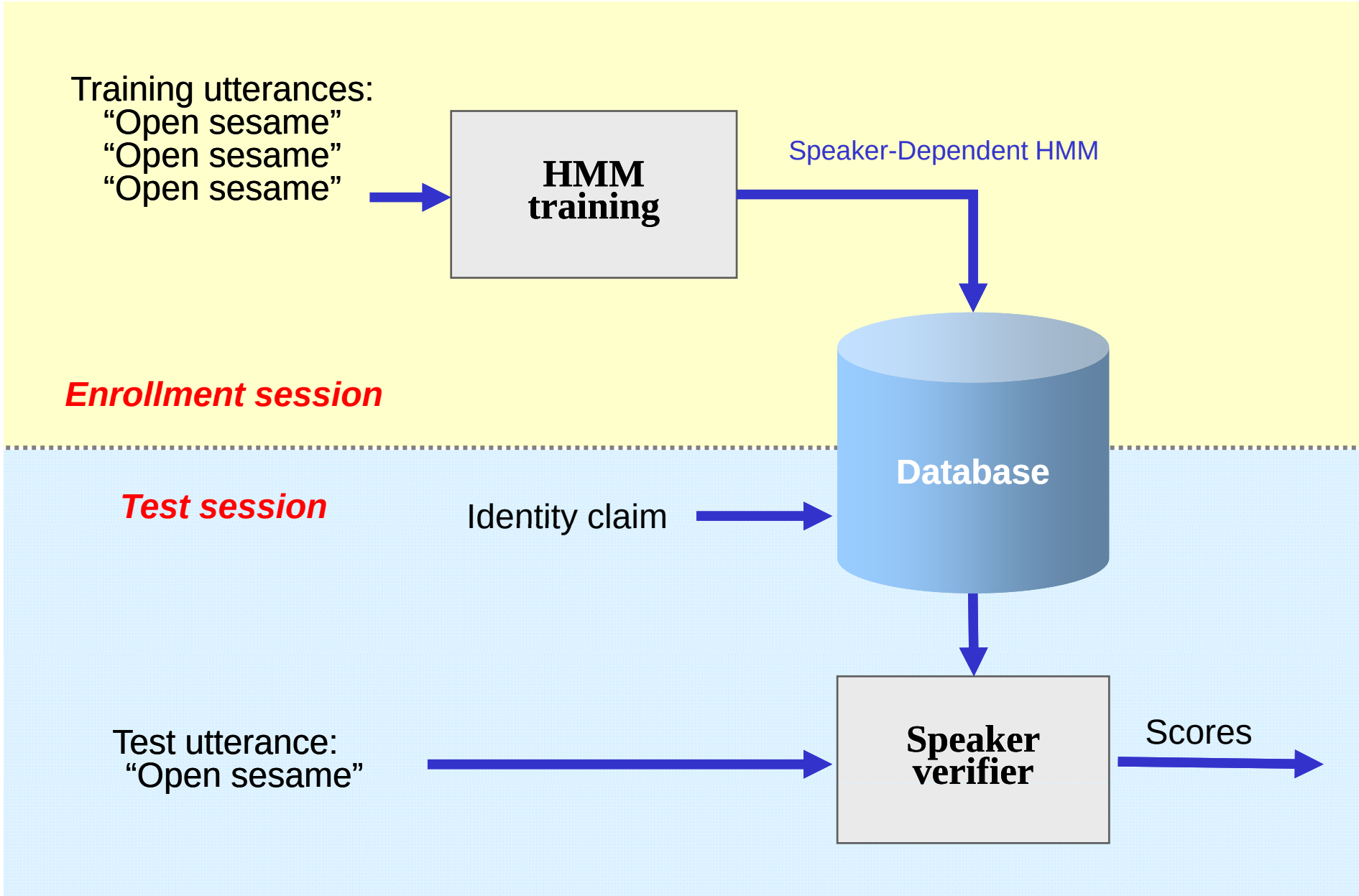
$P(X|U)$: Likelihood of an input utterance X for a user model U

μ_X, σ_X : Mean value and standard deviation of likelihood values for a cohort model

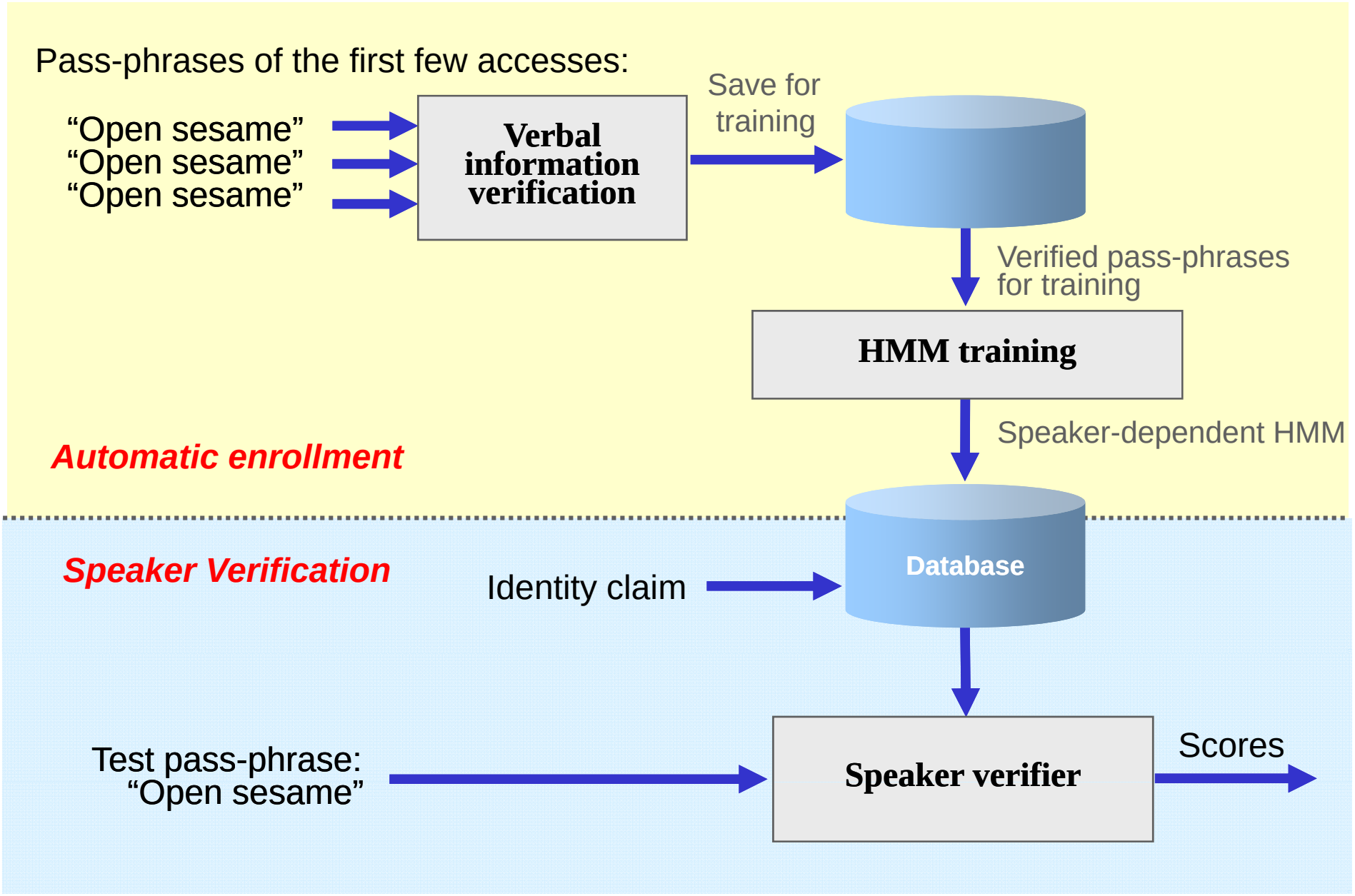
Speaker authentication approaches



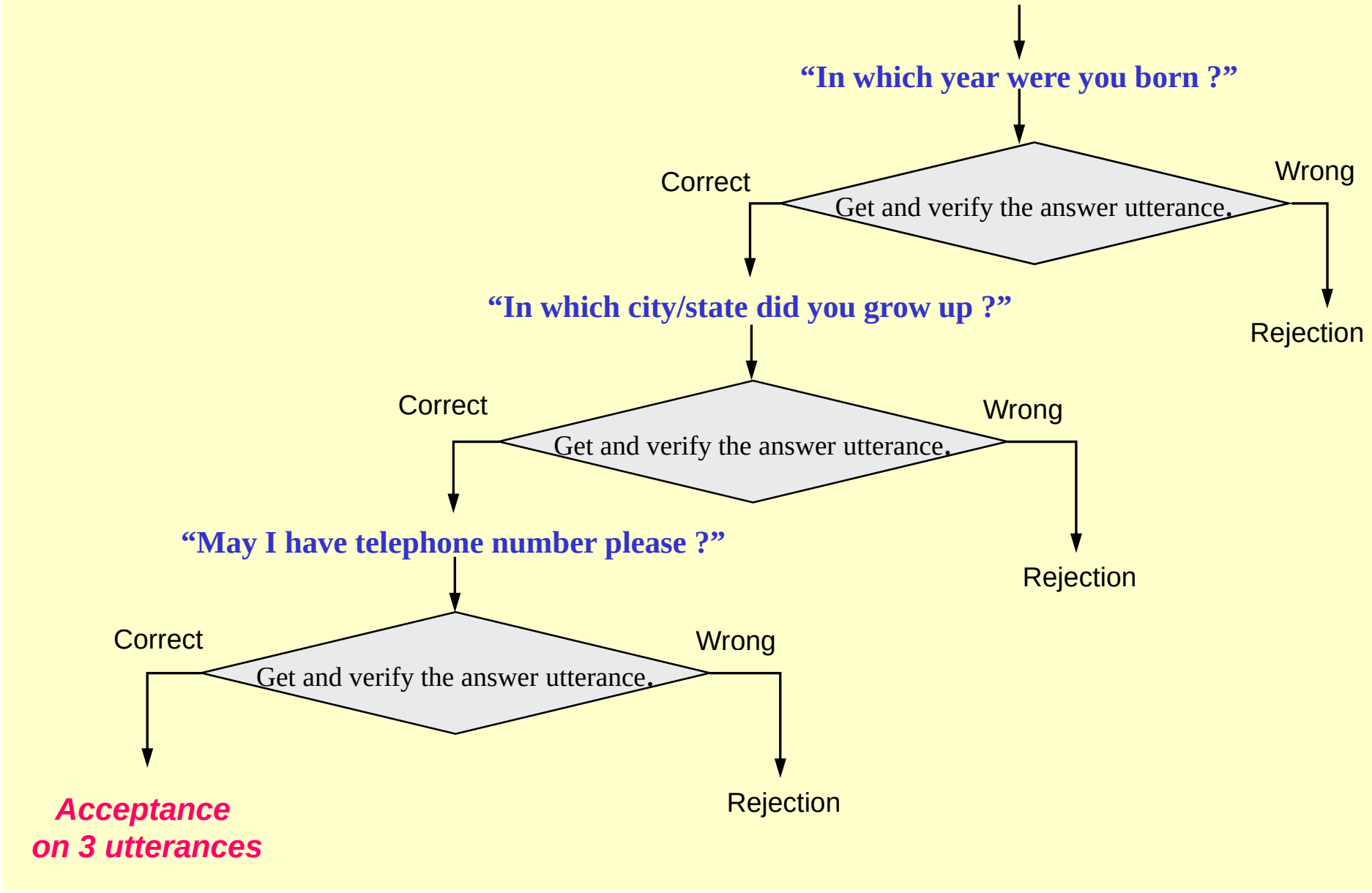
Conventional speaker verification system



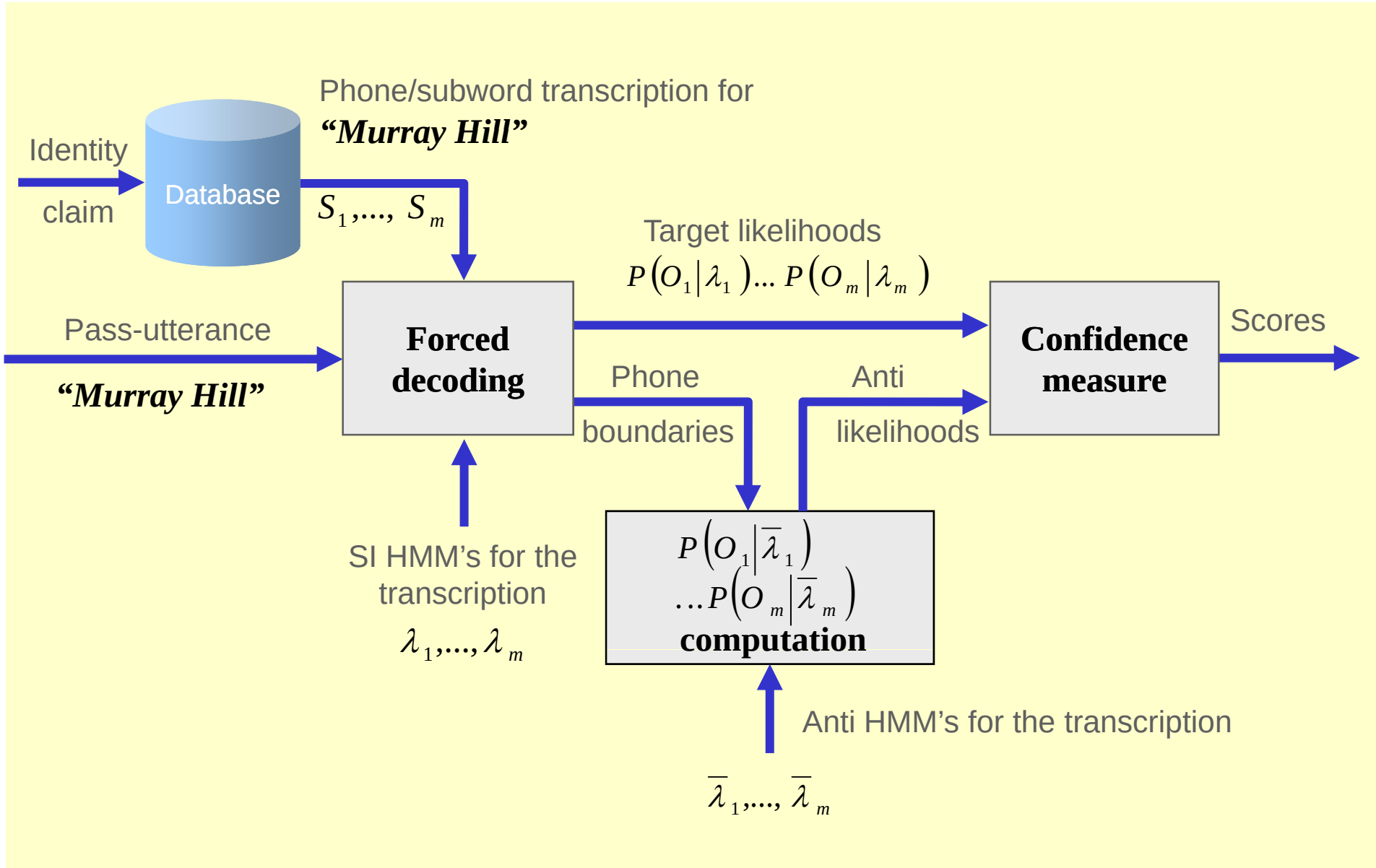
Speaker verification system including verbal information verification (VIV)



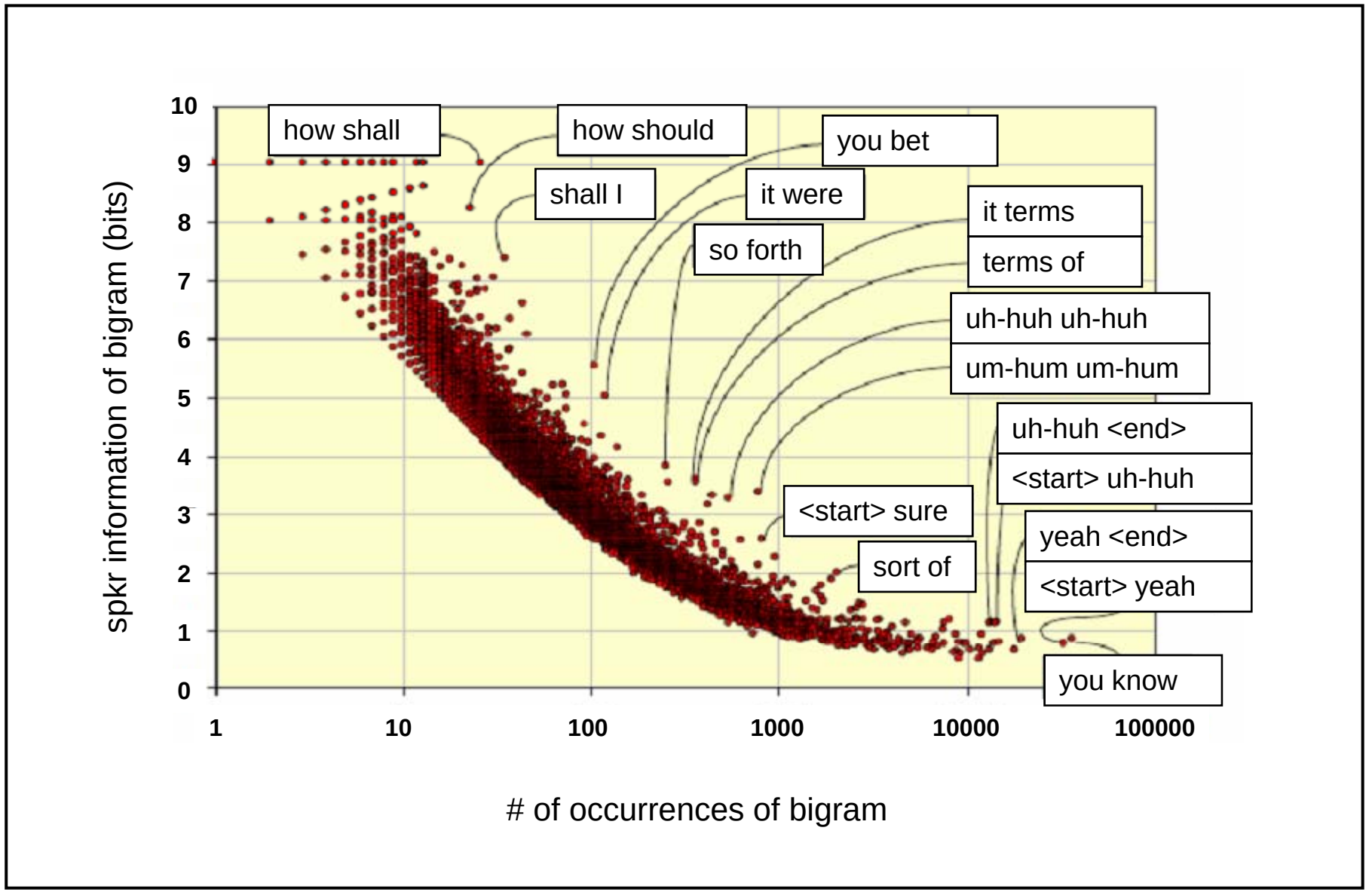
An example of verbal information verification by asking sequential questions



Utterance verification in verbal information verification (VIV)



Speaker information contained in word bigrams, tabulated over the whole SwitchBoard corpus (G. Doddington)



Speaker recognition by idiolectal differences between speakers

$$Score = \frac{\sum_{k} \left\{ N_{tokens}(k) \cdot \log \left[\Lambda_{TS}(k) / \Lambda_{BG}(k) \right] \right\}}{\sum_{k} \left\{ N_{tokens}(k) \right\}}$$

$N_{tokens}(k)$: Number of occurrences of N-gram type k in the test segment

$\Lambda_{TS}(k)$: Likelihood of N-gram type k for the test speaker model

$\Lambda_{BG}(k)$: Likelihood of N-gram type k for the background model

“Person authentication by voice: A need for caution”

AFCP and SpLC SIG of ISCA

***Proc. Eurospeech 2003*, pp. 33-36**

“At the present time, there is no scientific process that enables one to uniquely characterize a person’s voice or to identify with absolute certainty an individual from his or her voice.”

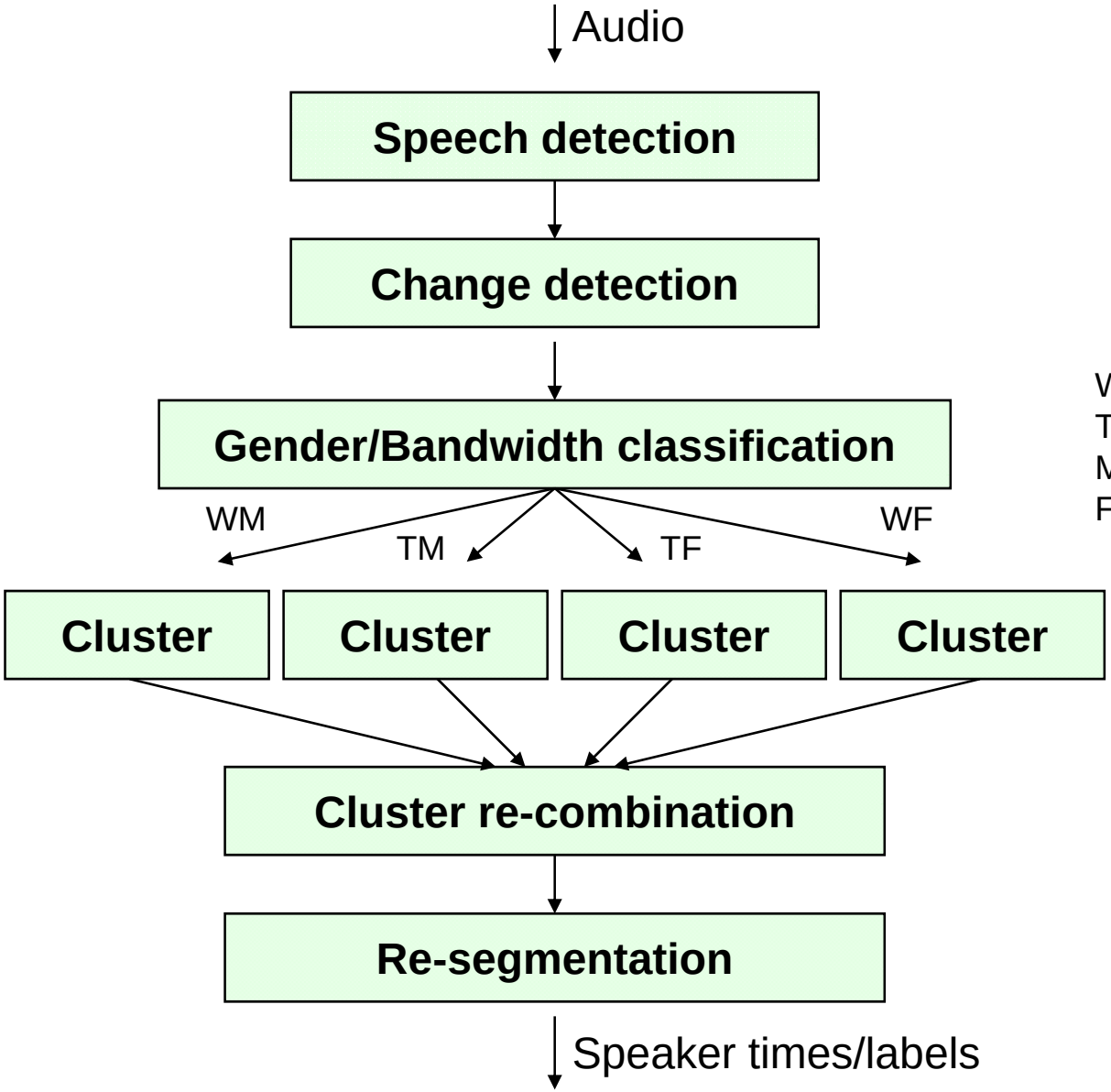
The following prerequisites are required to provide a reasonable level of performance in speaker recognition

- Speakers must not try to disguise their voice.
- The recording conditions and signal processing techniques are known or controlled.
- Speech, recorded in similar conditions as the test signal, is available to register a speaker in the system.
- Reference values for similarity measures must have been established in similar conditions as the test signal. Decision thresholds must have been calibrated from these reference values and tuned as a function of a specific application.

Applying additional constraints can result in improved performance:

- Speakers must be **willing to be recognized** and cooperate with the system.
- Potential impersonators must be prevented from using sophisticated technology to **modify or disguise** their voice.
- The use of **speech synthesis devices** is not allowed.
- The **linguistic content** of the message includes words already known to the system, so that the similarity between different voices can be calculated on the basis of similar contents.

Prototypical diarization system



W: wide bandwidth
T: narrow/telephone bandwidth
M: male
F: female