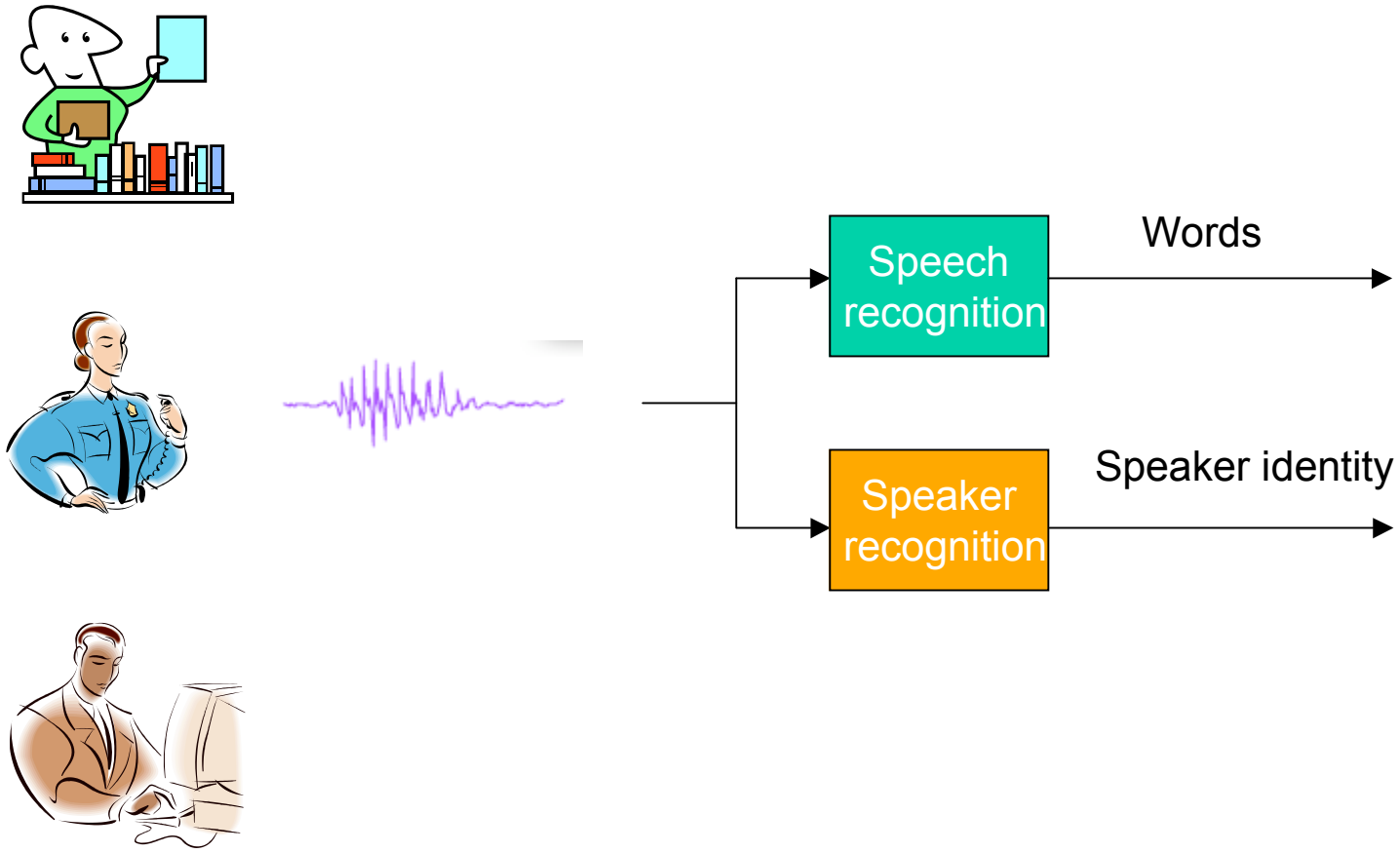
Automatic speaker verification

Outline

- Introduction
- Speaker identification vs verification
- Speaker verification overview
- The parts of a speaker verification system
- Evaluation of speaker verification performance

Speaker vs speech recognition



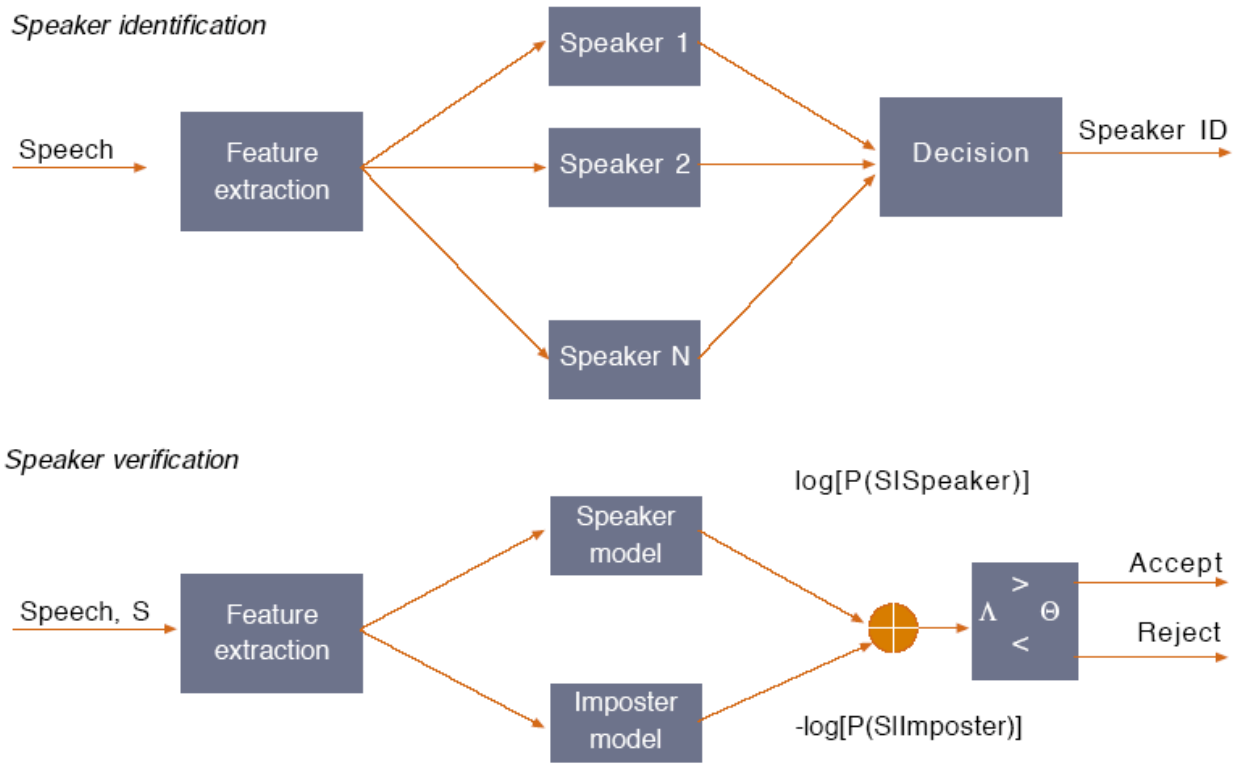
Speaker identification

- Determine the speaker identity
- Selection between a set of known voices
- The user does not claim an identity
- Closed set identification
  - Assume that all speakers are known to the system
- Open set identification
  - Possibility that speaker is not among the speakers known to the system

Speaker verification

- Synonyms: authentication, detection
- User claims an identity
- System task: Accept or reject identity claim
- The voice can come from outside the set of known speakers
  - All speakers known: closed set
- Imposter: All voices but the true identity

Identification vs verification



Speech modalities

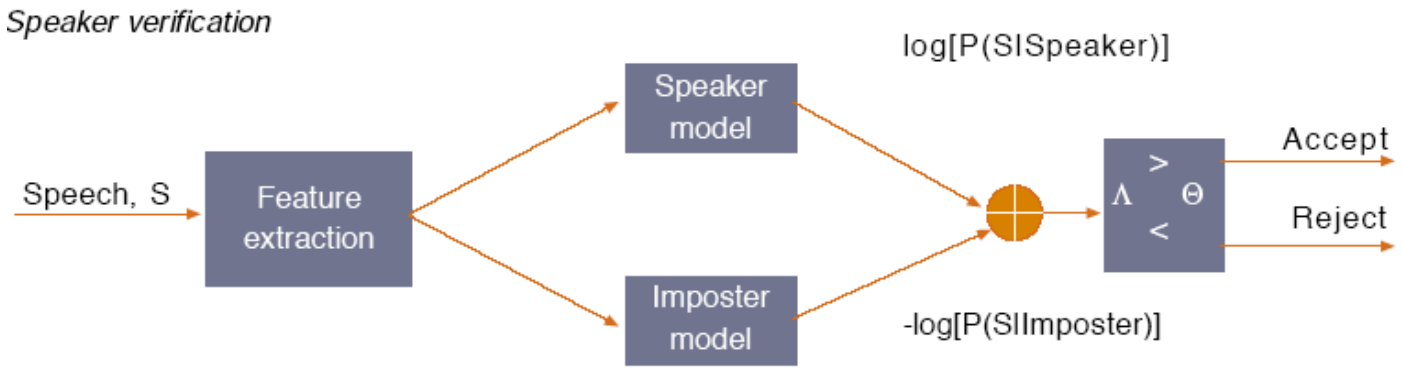
- Text-dependent speaker verification
  - The word sequence is (assumed) known to the system
    - E.g. prompted phrase, password phrase, fixed phrase
  - Knowing the text can improve system performance
  - Prompting may reduce risk of imposters using voice recordings
- Text-independent speaker verification
  - The uttered word sequence is unknown to the system
    - E.g. user initiated phrase, conversational speech
  - Less restrictions on user
  - Increased risk of imposters
  - ASR can be used for transcription

Speech for identification

- Speech is easily produced
- It does not require advanced input devices
- Can be applied using telephones, PCs
- Can be supplied with
  - Password phrase
  - Personal knowledge

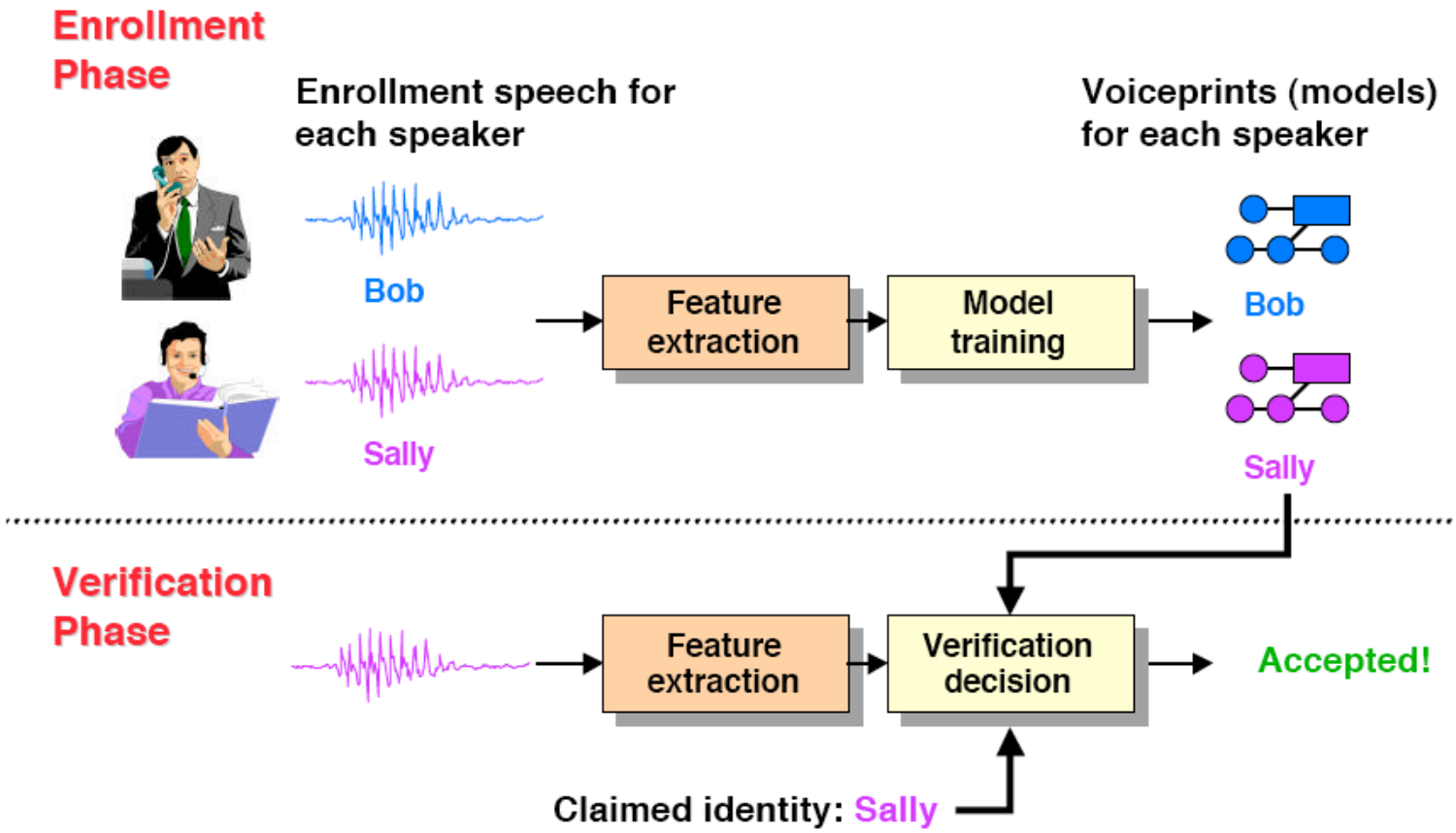
  to improve security

Speaker verification



- Which features?
- How to model the speaker
- How to model the imposters
- How to make the decision to minimize probability of error

Enrollment and verification



(From Reynods&Heck)

Features

- Desirable attributes (Wolf, ’72)
  - Occur naturally and frequently in speech
  - Easily measurable
  - Not change over time or be dependent on health condition
  - Be unaffected be environmental and transmission noise
  - Not be subject to mimicry
- A tall order
- Interestingly, the same, or similar features used as in speech recognition
  - Low-level, acoustic cues
  - Spectrum based
  - Typically: MFCCs

Models

- Hidden Markov models are predominantly used
- Type of HMM depends on application
  - Fixed phrase: Word or phrase models
  - Prompted phrase: Utterance models bulit from phone models
  - Text-independent: Single-state HMM - Gaussian Mixture model (GMM)

Text-independent speaker verification

- The imposter model is built using speech from all speakers
- GMM with high number of mixture components
- The speaker model is built using speaker adaptation
  - Relatively small amount of speech
  - MAP adaptation from the imposter model

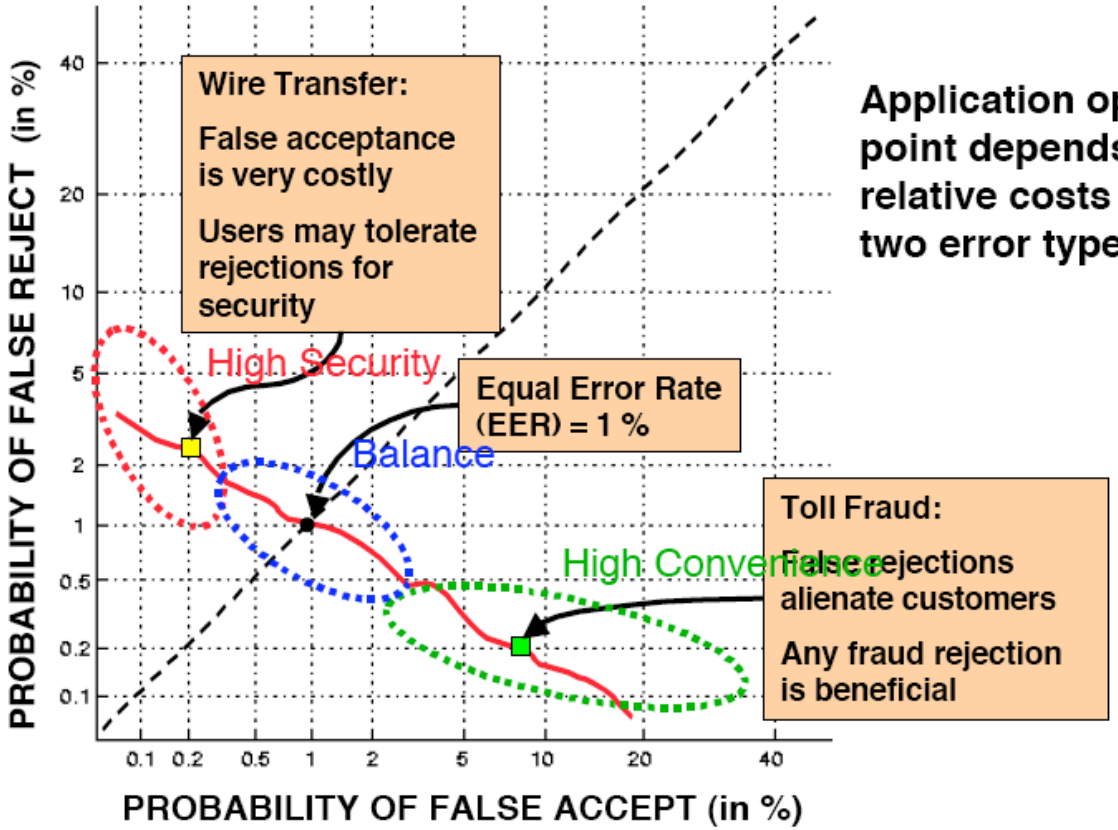
Decision

- The decision is a 2-class hypothesis test
  - H0: The speaker is an imposter
  - H1: The speaker is the claimed speaker
- Based on the log likelihood ratio
  - $\Lambda = \text{Log}\{(p(\text{speech} \mid \text{speaker model})/p(\text{speech} \mid \text{imposter model})\}$
- Desicion by threshold
  - $\Lambda < \Theta$ reject identity claim
  - $\Lambda > \Theta$ accept identity claim

Evaluation of performance

- System limitations
  - Speech quality (noise, variability, microphone...))
  - Speech modaluty (text dependent/independent...)
  - Speech duration (more speech -> more reliable)
  - Number and type of speakers
- Performance measure:
  - Two types of errors:
    - Accept imposter
    - Reject true speaker
  - Single number
    - Equal errror rate (EER)
  - Better picture: Detection Error Tradeoff (DET-curve)

DET-curve



(Reynolds&Heck)

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

Importance of the error types depend on application!