Speaker Recognition and Verification

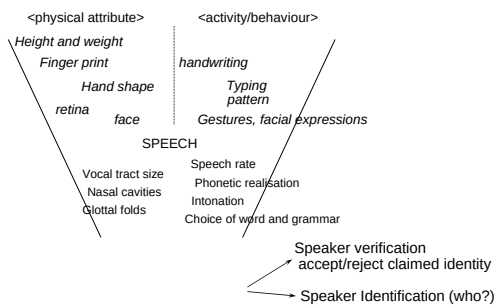
Mats Blomberg

Speech, Music and Hearing
KTH

Person identification

- Methods rely on
  - Something you **possess**
    - (E.g. key, magnetic card)
  - Something you **know**
    - (PIN-code)
  - Something you **are**
    - (physical attributes, behaviour: biometrics)

Biometric identification features



Verification / Identification

- Speaker verification
  - The claimed identity is verified by voice
  - Binary decision: “accept or reject?”, “true customer or impostor?”
  - The performance is independent of the number of registered users
- Speaker identification
  - Choose 1 of N: “Who is the speaker?”
    - Closed set: The utterance is known to come from the N trained speakers
    - Open set: The utterance may be spoken by persons outside N
  - The performance decreases with increased number of identities

Advantages/problems with speaker verification

+ Speech is natural
+ Simple to record, non-obtrusive
+ In many applications, speech may already be used for other purposes
  + Low extra cost if the application already uses speech recognition
+ Not 100% security, but
  - That’s the case for other techniques as well
  - Can be combined with other methods
  - Makes it less worthwhile for organised crime
  - Deterrent effect

– Large variability for a speaker at different occasions
  – Behaviour
  – Different microphones or microphone positions
  – Physical and mental condition
– Speech recognition problems

Application examples

- Telecommunication
  – Bank services, also complementary to manual methods
  – Credit cards
  – Information access by phone
  – Telephone call charging
- On-site
  – Entrance control
  – Authorisation
  – Home incarceration (large in USA)
- Crime investigation
  – Objective automatic techniques
- Speaker tracking
  – Find the intervals during a conversation when a certain person is speaking
  – E.g. during telephone conversation and in radio and TV

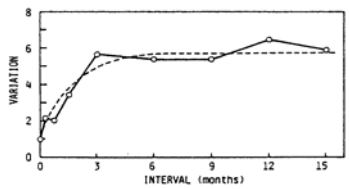
Text dependence

- SV systems have varying requirements on what the user should say
  - Fixed password
    - Highest text dependence
  - User specific password
  - Limited vocabulary
    - E.g. digits
  - The system presents the text to be spoken
    - Text-prompted
    - combination of speaker and text verification
    - Prevents playback of recorded speech
  - Any word sequence is allowed
    - Text-independent

Decreasing text dependence

Voice characteristics vary with time

Variability within one speaker



Acoustic variation among identical utterances as a function of the duration of the recordings. Average for nine male speakers. (Furui, 1986).

Same or different analysis as in speech recognition?

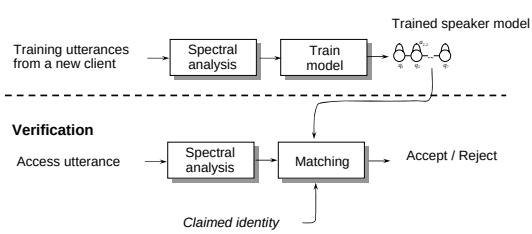
- SPEECH recognition should be SPEAKER independent
  - Should extract phonetic information but not speaker information
- SPEAKER recognition should be SPEECH independent
  - Should extract speaker information but not speech information
- This suggests that the optimal acoustic features are different between speech and speaker recognition
- However, experiments have shown that the best SPEECH representation is at the same time one of the best SPEAKER representations
- Why? Maybe the optimal representation contains both SPEECH and SPEAKER information

Modelling techniques

- HMM
  - Text-dependent systems
  - The state sequence represents allowed utterances
- GMM (Gaussian Mixture Models)
  - Text-independent systems
  - Single-state HMM with large number of Gaussian mixture components (~ 1000) representing any utterance by the speaker
  - Sequential information is not used
- Combined GMM + HMM systems

Two phases in speaker verification

Registration (training, enrolment)



Probabilistic decision criterion

- Bayes decision theory
  - The ratio between the probability scores of a client and an anti-client model is compared with a decision threshold

If $\frac{\text{P(The client sounds like this)}}{\text{P(Anybody could sound like this)}} > R$

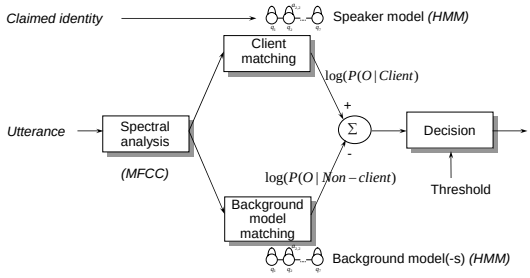
Then accept, else reject

$$\frac{P(O|\theta_C)}{P(O|\theta_{\overline{C}})} \geq R$$

O: utterance
θ_C: client C's model

The threshold R can be adjusted for
Required balance between errors,
Minimum total error
Minimum error cost

Standard system

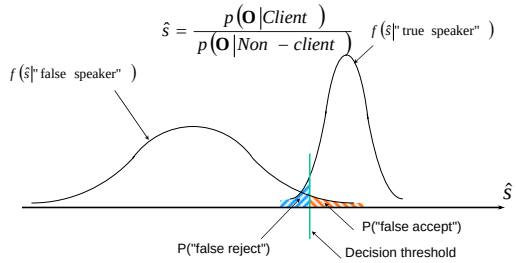


Two types of errors

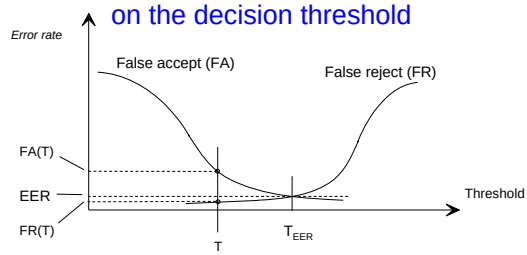
Claimed identity:	True	False
Accept	OK	False Accept (FA)
Reject	False Reject (FR)	OK

Decision:

Score distribution for true and false speaker identities



The error balance depends on the decision threshold

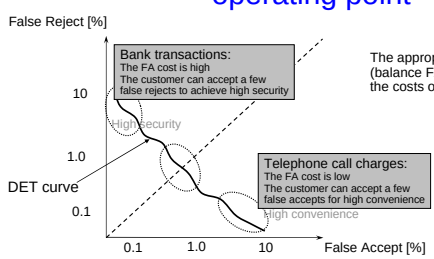


EER: Equal Error Rate, $EER = FA(T_{EER}) = FR(T_{EER})$
at an *a posteriori* determined threshold

Performance measures

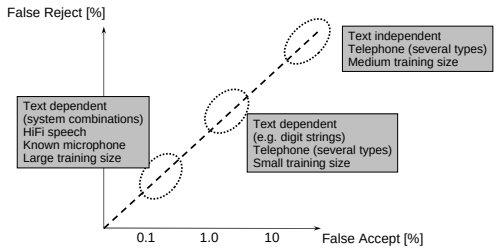
- False Rejection rate (FR)
  - FR = (Nbr *false reject utterances*) / (Nbr true ID attempts)
- False Acceptance rate (FA)
  - FA = (Nbr *false accept utterances*) / (Nbr impostor attempts)
- Half Total Error Rate (HTER)
  - HTER = (FR + FA) /2
- Equal Error Rate (EER)
  - EER = FR = FA at an a posteriori determined threshold
  - Well defined measure, but cannot be selected in practice
- Detection Error Trade-off (DET)
  - Exhibits FR and FA at different thresholds
  - Similar to “Receiver Operating Characteristics” (ROC)

Application-dependent operating point



The appropriate operating point (balance FA/FR) depends on the costs of each error type

Performance in different applications



Security aspects

- Performance is measured using casual impostors
- What is the immunity against real impostor attempts?
  - Imitations? Recordings? “Personal” speech synthesis?
- The security of conventional systems can be raised by combination with voice
  - E.g. protection if credit card + PIN code is stolen
- Preventive effect by
  - Recordings can be saved for later manual control

User aspects

- As little training as possible, preferrably nothing
  - The speaker’s variability cannot be measured
- Speaker verification should simplify for the user, preferrably transparent
- Door guard or warning bell?
- What balance FA / FR?
  - Depends on the security demands and the costs
  - True clients should not be disturbed

The CTT project PER (Prototype Entrance Receptionist)

- Visually detects the presence of a person at the TMH entrance
- Identifies personnel using speaker verification and unlocks the gate
  - Say your name and a prompted digit sequence
  - Animated talking face
- Combined HMM and GMM system
  - Comparable performance with commercial system
- In practical use since 1998

PER at the TMH entrance

Subject: the creator Håkan Melin

Summary

- Speaker verification useful today in certain applications
- Can be combined with other methods to increase security
- User aspects have to be taken into account