L16: Speaker recognition

Introduction

Measurement of speaker characteristics

Construction of speaker models

Decision and performance

Applications

[This lecture is based on Rosenberg et al., 2008, in Benesty et al., (Eds)]

Introduction

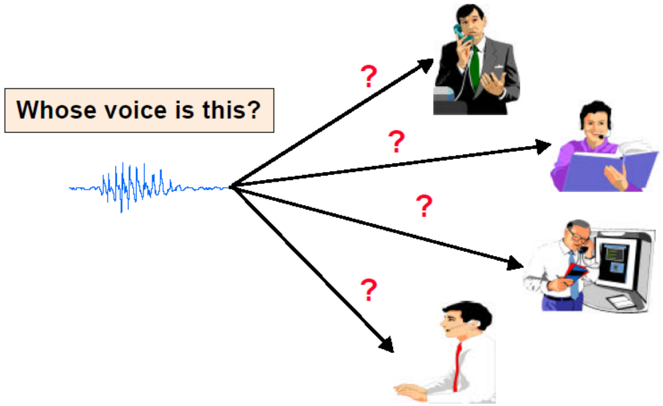
Speaker identification vs. verification

- Speaker identification
  - The goal is to match a voice sample from an unknown speaker to one of several of labeled speaker models
  - No identity is claimed by the user
  - Open-set identification: it is possible that the unknown speaker is not in the set of speaker models
    - If no satisfactory match is found, a *no-match* decision is provided
  - Closed-set : the unknown speaker is one of the known speakers
  - Speaker may be cooperative or uncooperative
  - Performance degrades as the number of comparisons increases

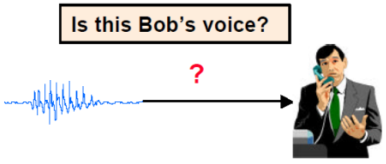
Introduction

- Speaker verification
  - User makes a claim as to his/her identity, and the goal is to determine the authenticity of the claim
  - In this case, the voice samples are compared only with the speaker model of the claimed identity
  - Can be thought of as a special case of open-set identification (one vs. all)
  - Speaker is generally assumed to be cooperative
  - Because only one comparison is made, performance is independent of the size of the speaker population

Speaker identification

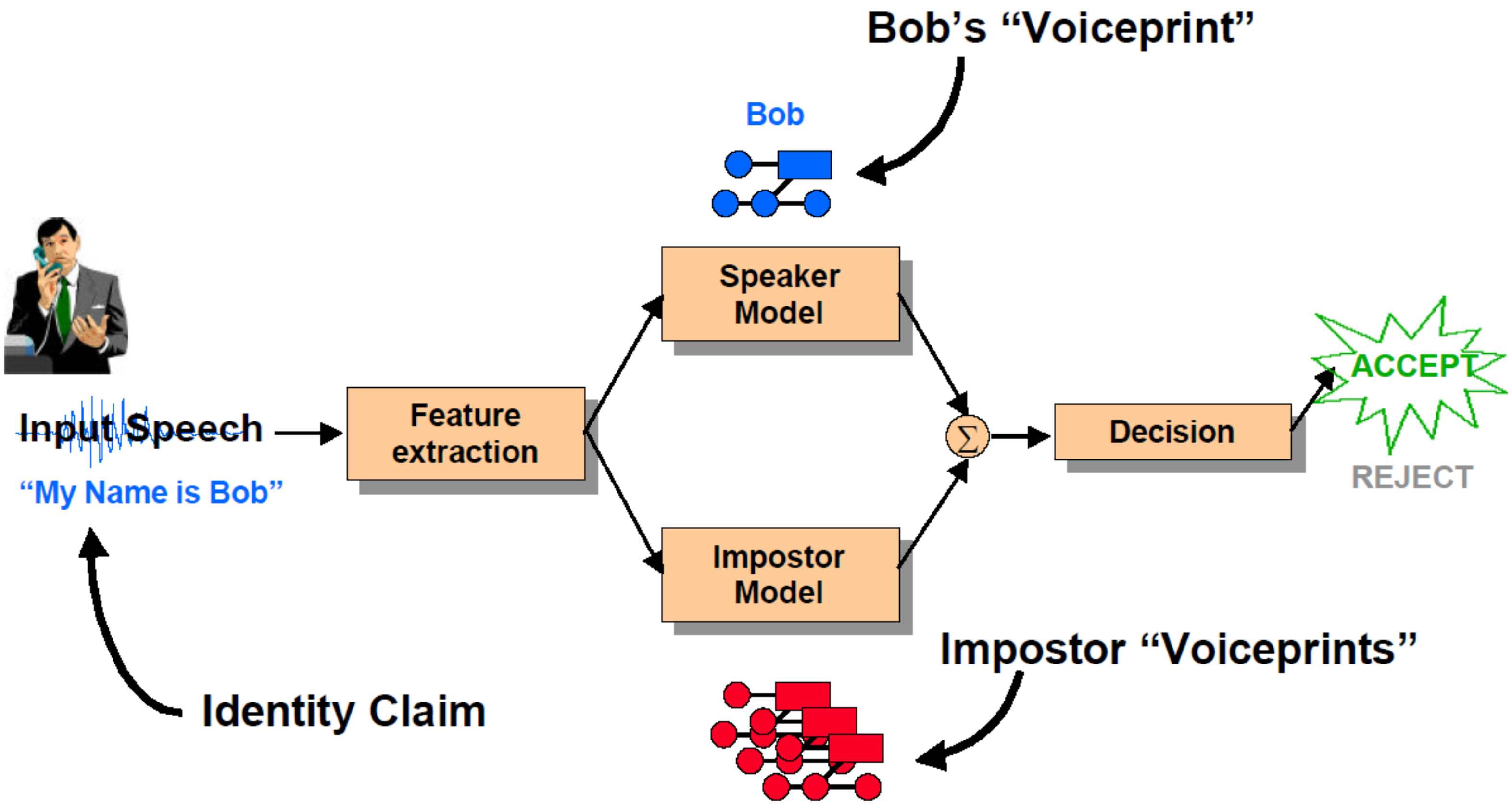


Speaker verification



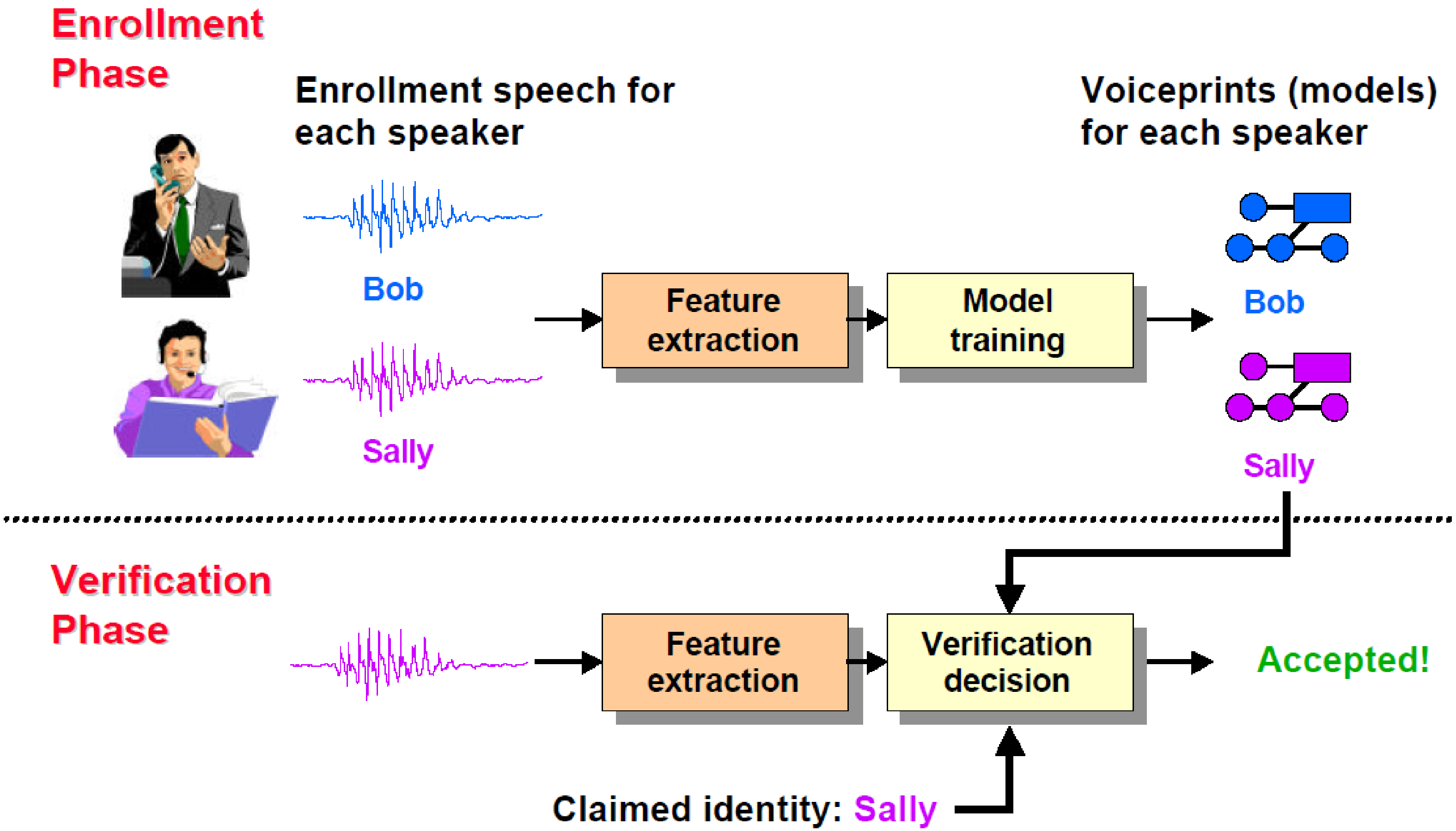
http://www.ll.mit.edu/mission/communications/ist/publications/aaas00-dar-pres.pdf

Components of a speaker verification system



From http://www.ll.mit.edu/mission/communications/ist/publications/aaas00-dar-pres.pdf

Two distinct phases to any speaker verification system



From http://www.ll.mit.edu/mission/communications/ist/publications/aaas00-dar-pres.pdf

Text-dependent vs. text-independent

- Text-dependent recognition
  - Recognition system knows the text spoken by the person, either fixed passwords or prompted phrases
  - These systems assume that the speaker is cooperative
  - Suited for security applications
    - To prevent impostors from playing back recorded passwords from authorized speakers, random prompted phrases can be used
- Text-independent recognition
  - Recognition system does not know text spoken by person, which could be user-selected phrases or conversational speech
  - Unsuited for security applications (e.g., impostor playing back a recording from an authorized speaker)
  - Suited for identification of uncooperative speakers
  - More flexible system but also more difficult problem

Measurement of speaker characteristics

Types of speaker characteristics

- Low-level features
  - Associated with the periphery in the brain's perception of speech
    - Segmental: formants are relatively hard to track reliably, so one generally uses short-term spectral measurements (e.g., LPC, filter-bank analysis)
    - Supra-segmental: Pitch periodicity is easy to extract, but also requires a prior voiced/unvoiced detector
    - Long term averages of these measures may be used if one does not need to resolve detailed individual differences
- High-level features
  - Associated with more central locations in the perception mechanism
    - Perception of words and their meaning
    - Syntax and prosody
    - Dialect and idiolect (variety of a language unique to a person)
  - These features are relatively harder to extract than low-level features

Low-level features

- Short-time spectra, generally MFCCs
    - Isn't this counterintuitive?
        - Speech recognition should be speaker independent, whereas speaker recognition should be speech independent
        - This would suggest that the optimal acoustic features would be different,
    - However, the best speech representation turns out to be also a good speaker representation (!) … *perhaps the optimal representation contains both speech and speaker information?*
- Cepstral mean subtraction
    - Subtracts the cepstral average over a sufficiently long speech recording
    - Removes convolutional distortions in slowly varying channels
- Dynamic information
    - Derivatives (Δ) and second derivatives (Δ^2) of the above features are also useful (both for speech and for speaker recognition)
- Pitch and energy <u>averages</u>
    - Robust pitch extraction is hard and pitch has large intra-speaker variation

Linguistic measurements

- Can only be used with long recordings (i.e., indexing broadcast, passive surveillance), not with conventional text-dependent systems
- Word usage
  - Vocabulary choices, word frequencies, part-of-speech frequencies
  - Spontaneous speech, such as fillers and hesitations
  - Susceptible to errors introduced by LVCSR systems
- Phone sequences and lattices
  - Models of phone sequences output by ASR using phonotactic grammars can be used to represent speaker characteristics
  - However, lexical constraints generally used to improve ASR may prevent extraction of phone sequences that are unique to a speaker
- Other linguistic features
  - Pronunciation modeling of carefully chosen words
  - Pitch and energy contours, duration of phones and pauses

Construction of speaker models

Speaker recognition models can be divided into two classes

- Non-parametric models
  - These models make few structural assumptions about the data
  - Effective when there is sufficient enrollment data to be matched to the test data
  - Models are based on techniques such as
    - Template matching (DTW)
    - Nearest-neighbors models
- Parametric models
  - Offer a parsimonious representation of structural constraints
  - Can make effective use of enrollment data if constraints are chosen properly
  - Models are based on techniques such as
    - Vector quantization,
    - Gaussian mixture models,
    - Hidden Markov models, and
    - Support vector machines (will not be discussed here)

Non-parametric models

- Template matching
  - The simplest form of speaker modeling; rarely used in real applications today
  - Appropriate for fixed-password speaker verification systems
  - Enrollment data consists of a small number of repetitions of the password
  - Test data is compared against each of the enrollment utterances and the identity claim is accepted if the distance is below a threshold
  - Feature vectors for test and enrollment data are aligned with DTW
- Nearest-neighbors modeling
  - It can be shown that, given enrollment data from a speaker X, the local density (likelihood) for test utterance y is (see CSCE 666 lecture notes)

$$p_{nn}(y;X) = \frac{1}{V[d_{nn}(y,X)]} = \frac{1}{V\left[\min_{x_j \in X} \left\| y - x_j \right\| \right]}$$

    - where $V[r] \sim r^D$ is the volume of a *D*-dimensional hyper-sphere of radius *r*

- Taking logs and removing constant terms, we can define a similarity measure between Y and X as

$$s_{nn}(Y;X) = -\sum\nolimits_{y_j \in Y} \ln[d_{nn}(y,X)\ ]$$

  - and the speaker with greatest $s_{nn}(Y;X)$ is identified

- It has been shown that the following measure provides significantly better results than $s_{nn}(Y;X)$

$$\begin{aligned} s'_{nn}(Y;X) = & \frac{1}{N_y} \sum_{y_j \in \mathrm{Y}} \min_{x_i \in X} \left\| y_j - x_i \right\|^2 \\ & + \frac{1}{N_x} \sum_{x_j \in X} \min_{y_i \in Y} \left\| y_i - x_j \right\|^2 \\ & - \frac{1}{N_y} \sum_{y_j \in \mathrm{Y}} \min_{y_i \in Y; j \neq i} \left\| y_i - y_j \right\|^2 \\ & - \frac{1}{N_x} \sum_{x_j \in X} \min_{x_i \in X; j \neq i} \left\| x_i - x_j \right\|^2 \end{aligned}$$

Parametric models

- Vector quantization
  - Generally based on *k-means*, which we presented in an earlier lecture
    - Since *k* is unknown, an iterative technique based on the Linde-Buzo-Gray (LBG) algorithm is generally used
    - LBG: start with $k = 1$, choose the cluster with largest variance and partition into two by adding a small perturbation to their means $(\mu \pm \epsilon)$, and repeat
  - Once VQ models are available for the target speaker, evaluate sum-squared-error measure D to determine authenticity of the claim

$$D = \sum_{j=1}^{J} \sum_{x_i \to \mu_j} (x_i - \mu_j)$$

    - where μ_j is the sample mean of test vectors assigned to the j-th cluster
  - VQ may be used for text-dependent and text-independent systems
  - Temporal aspects may be included by clustering sequences of feature vectors
  - While VQ is still useful, it has been superseded by more advanced models such as GMMs and HMMs

- Gaussian mixture models
  - GMMs can be thought of as a generalization of k-means where each cluster is allowed to have its own covariance matrix
    - As we saw in an earlier lecture, model parameters (mean, covariance, mixing coefficients) are learned with the EM algorithm
  - Given trained model λ, test utterance scores are obtained as the average log-likelihood given by

$$s(Y|\lambda) = \frac{1}{T}\sum_{t=1}^{T} \log[p(y_t|\lambda)]$$

  - When used for speaker verification, the final decision is based on a likelihood ratio test of the form

$$\frac{p(Y|\lambda)}{p(Y|\lambda_{BG})}$$

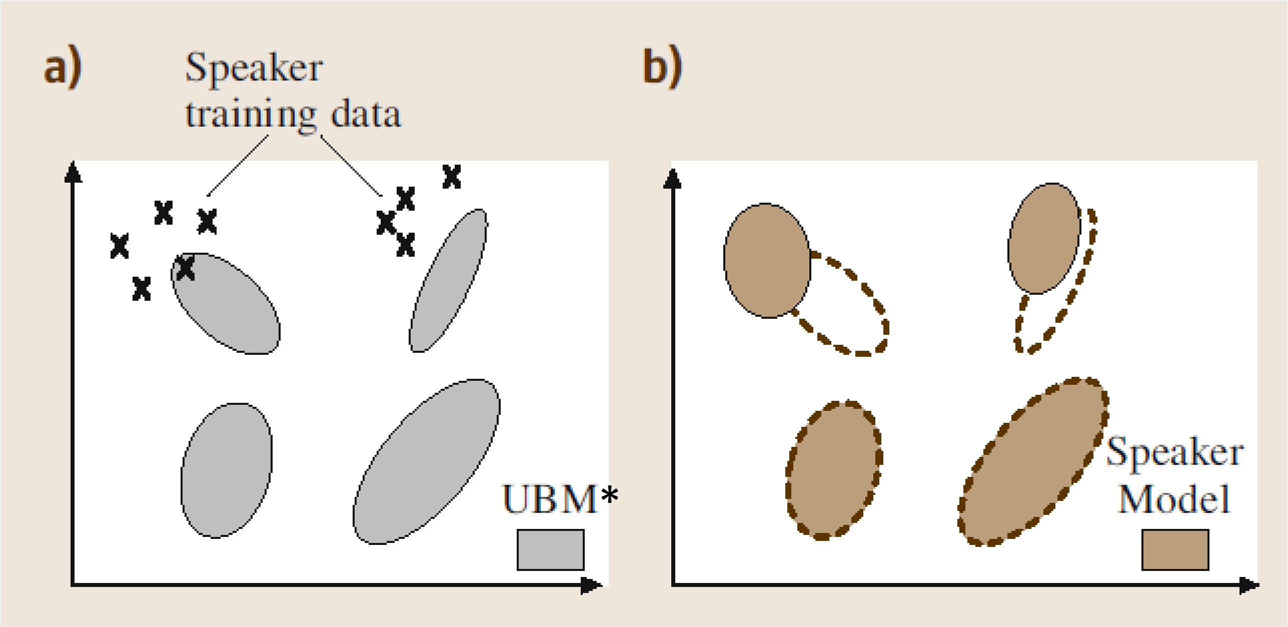
    - where λ_{BG}represents a background model trained on a large independent speech database
    - As we will see, the target speaker model λ can also be obtained by adapting λ_{BG}, which tends to give more robust results
  - GMMs are suitable for text-independent speaker recognition but do not model the temporal aspects of speech

- Hidden Markov Models
  - For text-dependent systems, HMMs have been shown to be very effective
    - HMMs may be trained at the phone, word or sentence level, depending on the password vocabulary (e.g., digit sequences are commonly used)
  - HMMs are generally trained using maximum likelihood (Baum-Welch)
    - Discriminative training techniques may be used if examples from competing speakers are available (e.g., closed-set identification)
  - For text-independent systems, ergodic HMMs may be used
    - Unlike the left-right HMMs generally used in ASR, ergodic HMMs allow all possible transitions between states
      - In this way emission probabilities will tend to represent different spectral characteristics (associated with different phones), whereas transition probabilities allow some modeling of temporal information
    - Experimental comparison of GMMs and ergodic HMMs, however, show that the addition of the transition probabilities in HMMs has little effect on performance

Adaptation

- In most speaker recognition scenarios, the speech data available for enrollment is too limited to train models
  - In fixed-password speaker authentication systems, the enrollment data may be recorded in a single call
    - As a result, enrollment and test conditions may be mismatched: different telephone handsets and networks (landline vs. cellular), background noises
  - In text-independent models, additional problems may result from mismatches in linguistic content
- For these reasons, adaptation techniques may be used to build models for specific target speakers
  - When used in fixed-password systems, model adaptation can reduce error rates significantly

Adapting a hypothesized speaker model (for GMMs)



[Reynolds & Campbell, 2008, in Benesty et al., (Eds)]

*UBM: universal background model

Decision and performance

Decision rules

- The previous models provide a score $s(Y|\lambda)$ that measures the match between a given test utterance Y and a speaker model λ
  - Identification systems produce a set of scores, one for each target speaker
    - In this case, the decision is to choose the speaker $\hat{S}$ with maximum score

$$\hat{S} = \arg\max_{j} s(Y|\lambda_j)$$

  - Verification systems output only one score, that of the claimed speaker
    - Here, a verification decision is obtained by comparing the score against a predetermined threshold

$$s(Y|\lambda_i) \geq \theta \Rightarrow Y \in \lambda_i$$

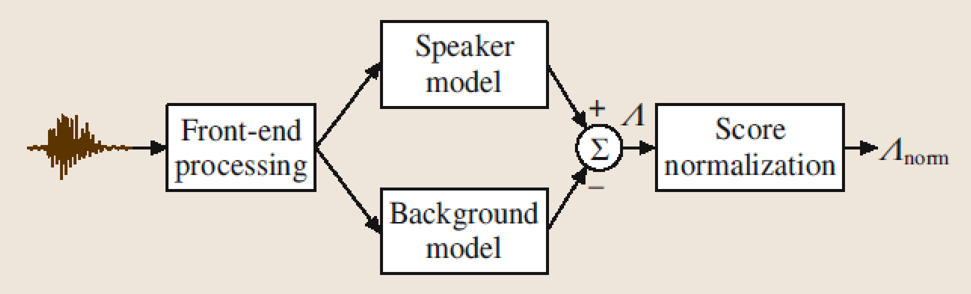
  - Open-set identification relies on two steps
    - a closed-step identification to find the most likely speaker, and
    - a verification step to test whether the match is good enough

Threshold setting and score normalization

- When the score is obtained in a probabilistic framework, one may employ Bayesian decision theory to determine the threshold θ
  - Given false acceptance c_{fa} and false rejection c_{fr} rates and the prior probability of an impostor p_{imp}, the optimal threshold $\theta^*$is

$$\theta^* = \frac{c_{fa}}{c_{fr}} \frac{p_{imp}}{1 - p_{imp}}$$

- In practice, however, the score $s(Y|\lambda)$ does not behave as theory predicts due to modeling errors
  - To address this issue, various forms of normalization have been proposed over the years, such as Z-norm, H-norm, T-norm, etc.

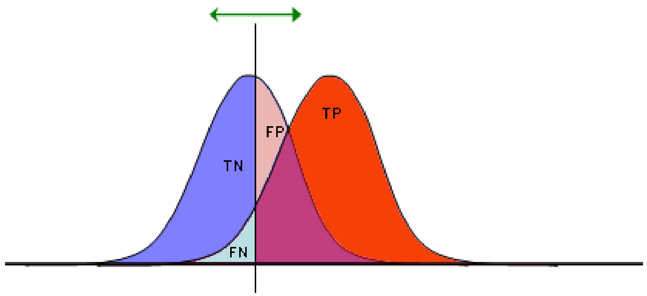


[Reynolds & Campbell, 2008, in Benesty et al., (Eds)]

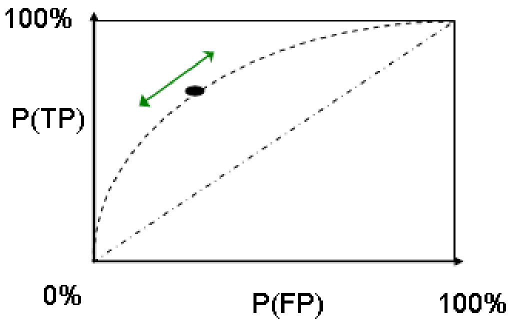
Errors and DET

- SID systems are evaluated based on the probability of misclassification
- Verification systems, in contrast, are evaluated based on two types of errors: false acceptance errors, and false rejection errors
  - The probability of these two errors (p_{fa}, p_{fr}) varies in opposite directions when the decision threshold θ is varied
    - The tradeoff between the two types of errors is often displayed as a curve known as the receiver operating characteristic (ROC) in decision theory
- Detection error threshold (DET)
  - In speaker verification, the two errors are converted to normal deviates $(\mu = 0; \sigma = 1)$ and plotted in log scale, and the curve is known as a DET
    - The DET highlights differences between systems more clearly
    - If the two errors are Gaussian with $\sigma = 1$ the curve is linear with slope -1, which helps rank systems based on how close their DET is to the ideal

Generating ROC curves

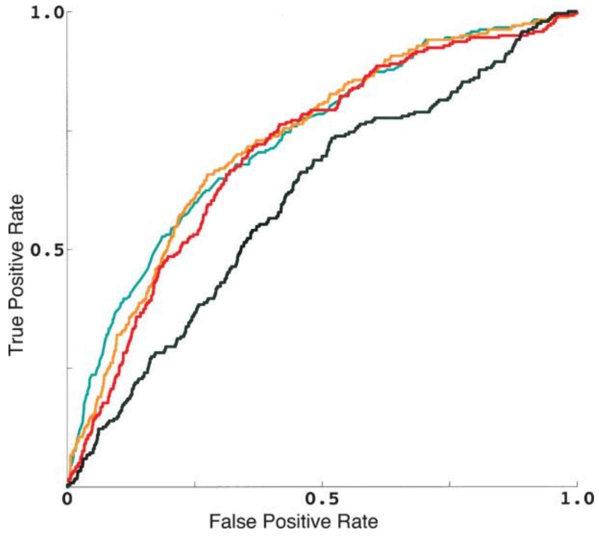


TP	FP
FN	TN
1	1



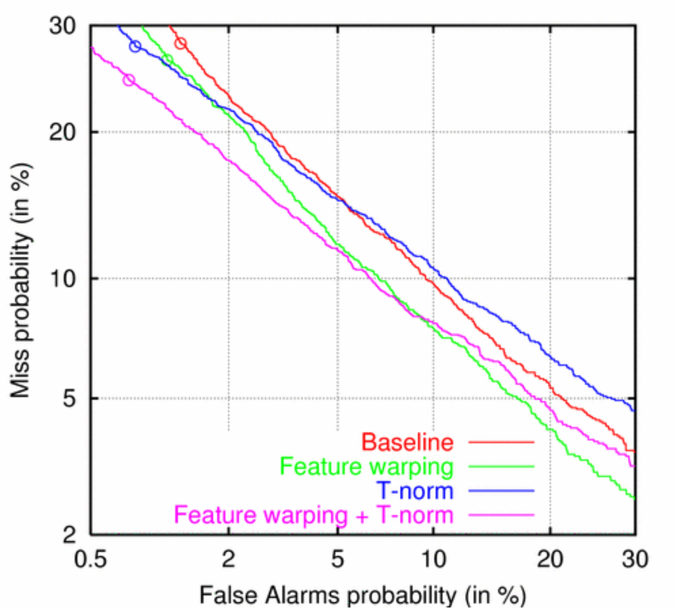
http://en.wikipedia.org/wiki/Receiver_operating_characteristic

ROC



http://genome.cshlp.org/content/18/2/206/F4.expansion

DET



http://www.limsi.fr/RS2003GB/CHM2003GB/TLP2003/TLP9/modelechmgb.html

Selecting a detection threshold

- The DET shows how the system behaves over a range of thresholds, but does not indicate which threshold should be used
- Two criteria are commonly used to select an operating point
- Equal error rate (EER)
  - The threshold at which the two errors are equal $p_{fa} = p_{fr}$
- Detection cost function (DCF)
  - The threshold that minimizes the expected risk based on the prior probability of impostors and the relative cost of the two types of errors

$$C = p_{imp} c_{fa} p_{fa} + \left(1 - p_{imp}\right) c_{fr} p_{fr}$$

Applications

Transaction authentication

- Toll fraud prevention, telephone credit card purchases, telephone brokerage (e.g., stock trading)

Access control

- Physical facilities, computers and data networks

Monitoring

- Remote time and attendance logging, home parole verification, prison telephone usage

Information retrieval

- Customer information for call centers, audio indexing (speech skimming device), speaker diarisation

Forensics

- Voice sample matching

From http://www.ll.mit.edu/mission/communications/ist/publications/aaas00-dar-pres.pdf