Overview of Automatic Speaker Recognition

JHU 2008 Workshop Summer School

Douglas A. Reynolds, PhD
Senior Member of Technical Staff
M.I.T. Lincoln Laboratory

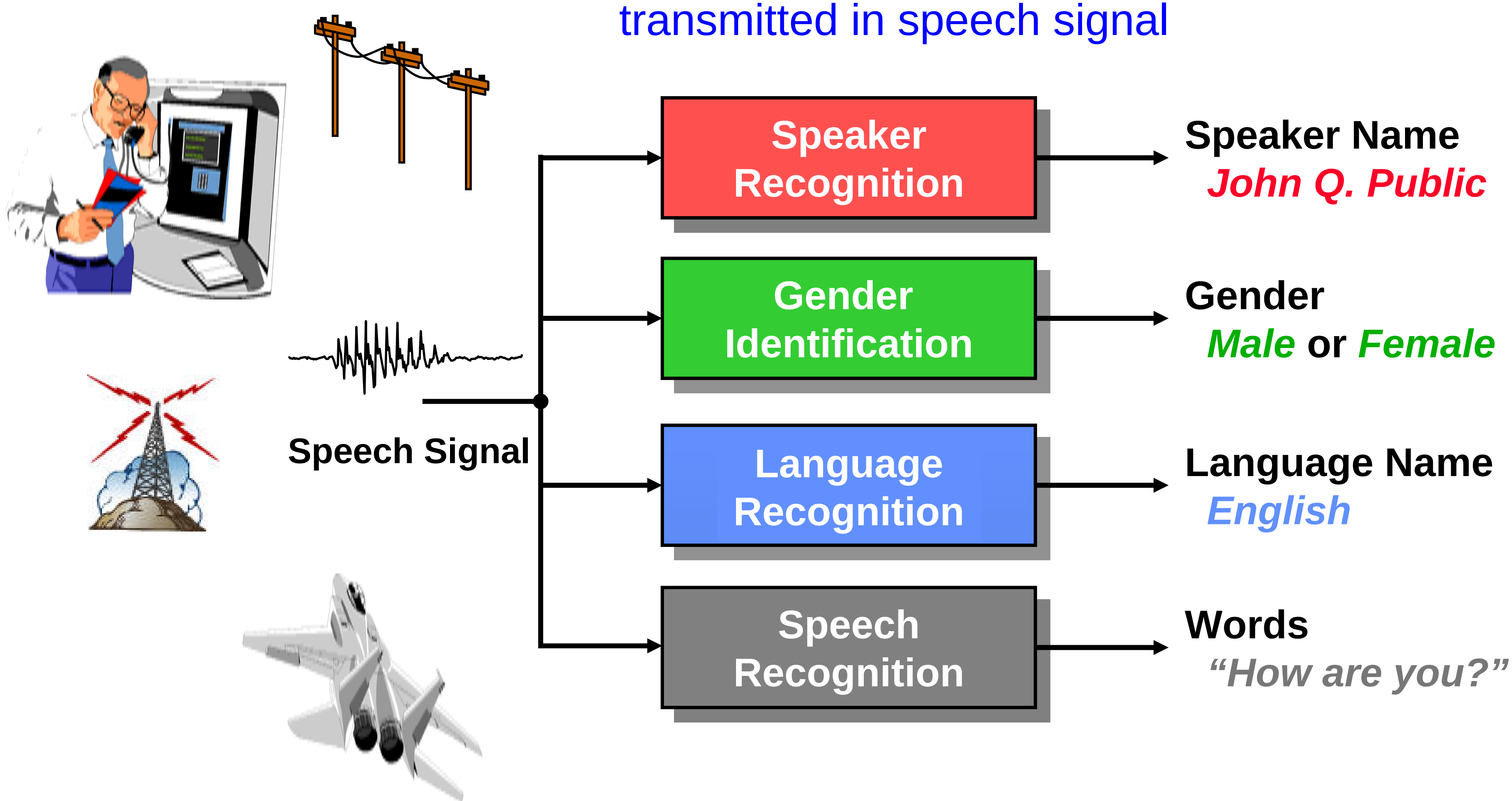
This work was sponsored by the Department of Defense under Air Force contract FA8721-05-C-0002 . Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government. These slides may not be further distributed without permission of MITLL.

Outline

- **Background and Theory**
  - Terminology
  - Components of recognition systems
    - Features and models
- **Evaluation and Performance**
  - Evaluation metrics and design
  - Performance survey
- **Following talk**
  - Factor analysis and discriminative training

Speech Processing Technologies:
Extracting Information from Speech

Goal: Automatically extract information transmitted in speech signal



Speaker Recognition Applications

Access Control
Physical facilities

Computer networks and websites

Transaction Authentication
Telephone banking

Remote credit card purchases

Law Enforcement
Forensics

Home parole

Speech Data Management
Voice mail browsing

Speech skimming

Personalization
Intelligent answering machine

Voice-web / device customization

Speech Modalities

Application dictates different speech modalities:

Text-dependent	Text-independent

Text-dependent

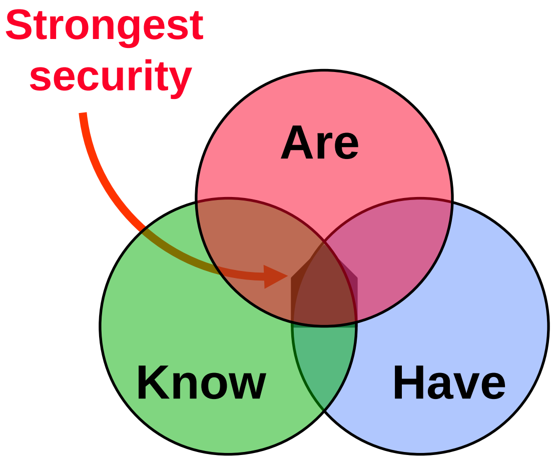
- Recognition system knows text spoken by person
- Examples: fixed phrase, prompted phrase
- Used for applications with strong control over user input
- Knowledge of spoken text can improve system performance

Text-independent

- Recognition system does not know text spoken by person
- Examples: User selected phrase, conversational speech
- Used for applications with less control over user input
- More flexible system but also more difficult problem
- Speech recognition can provide knowledge of spoken text

Voice Biometric

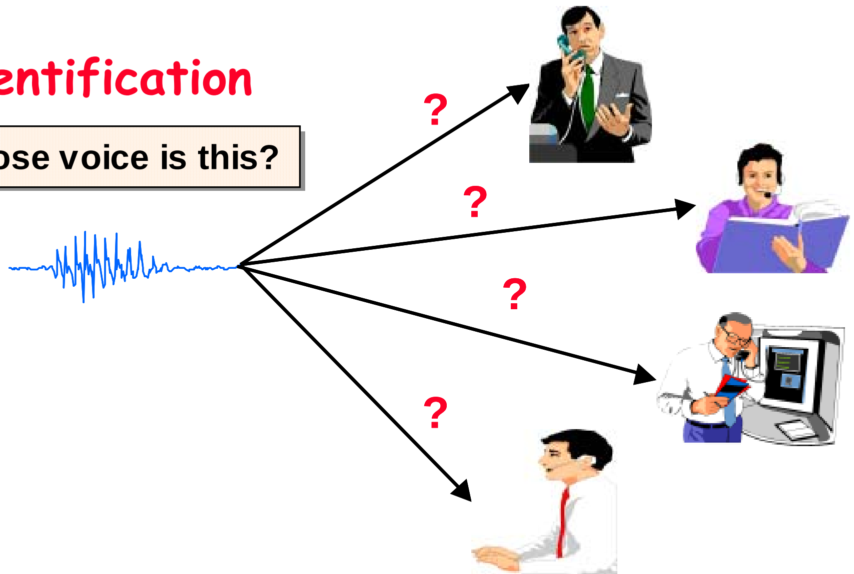
- **Speaker verification is often referred to as a voice biometric**
- **Biometric:** a human generated signal or attribute for authenticating a person's identity
- **Voice is a popular biometric:**
  - natural signal to produce
  - does not require a specialized input device
  - ubiquitous: telephones and microphone equipped PC
- **Voice biometric can be combined with other forms of security**
  - Something you have - e.g., badge
  - Something you know - e.g., password
  - Something you are - e.g., voice



Speaker Recognition Tasks

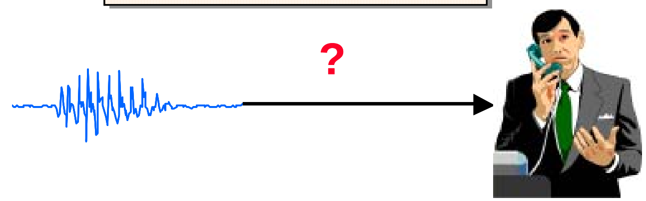
Identification

Whose voice is this?



Verification/Authentication/Detection

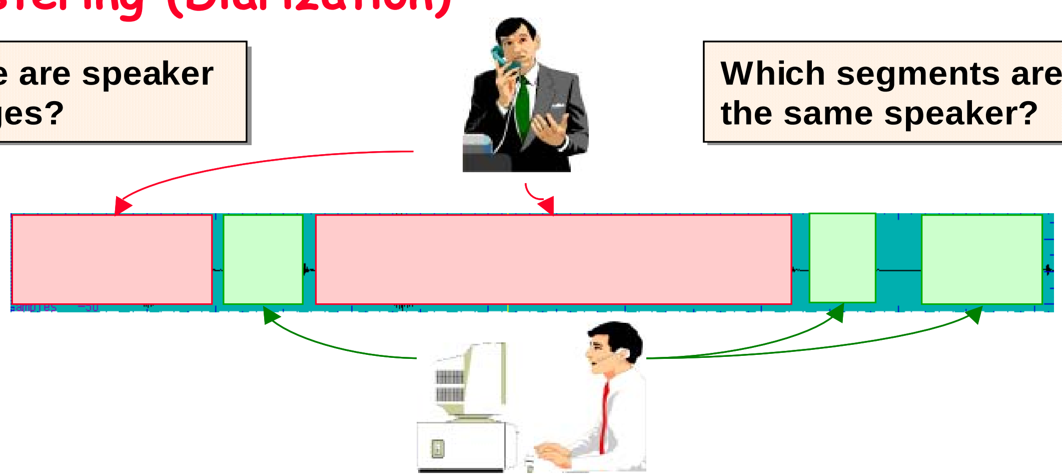
Is this Bob's voice?



Segmentation and Clustering (Diarization)

Where are speaker changes?

Which segments are from the same speaker?



Likelihood Ratio Test

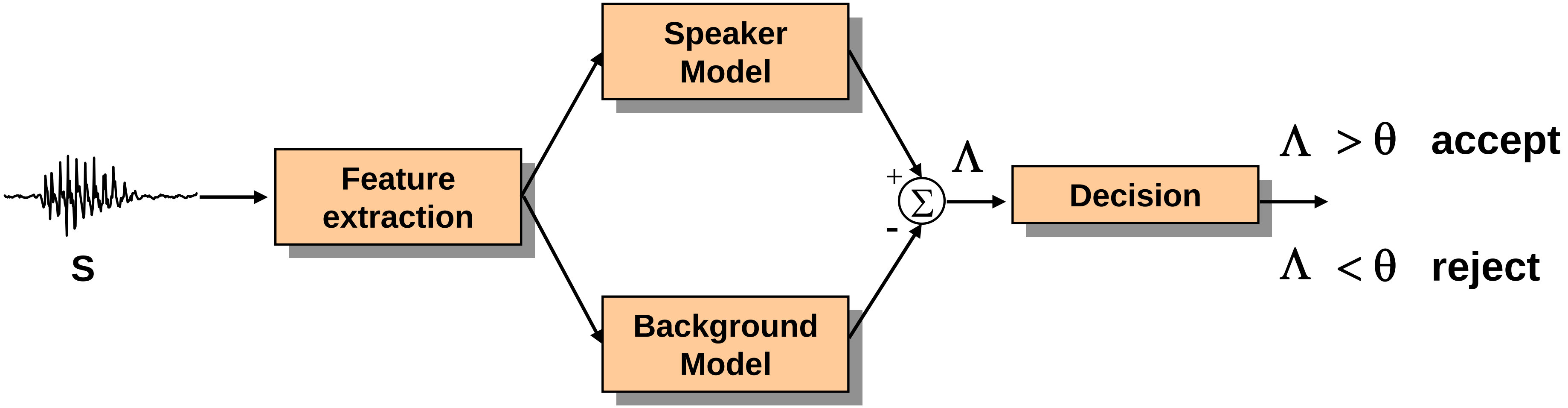
Speaker detection decision approaches have roots in signal detection theory

- **2-class Hypothesis test:**

  H0: the speaker is <u>not</u> the target speaker
  H1: the speaker is the target speaker.

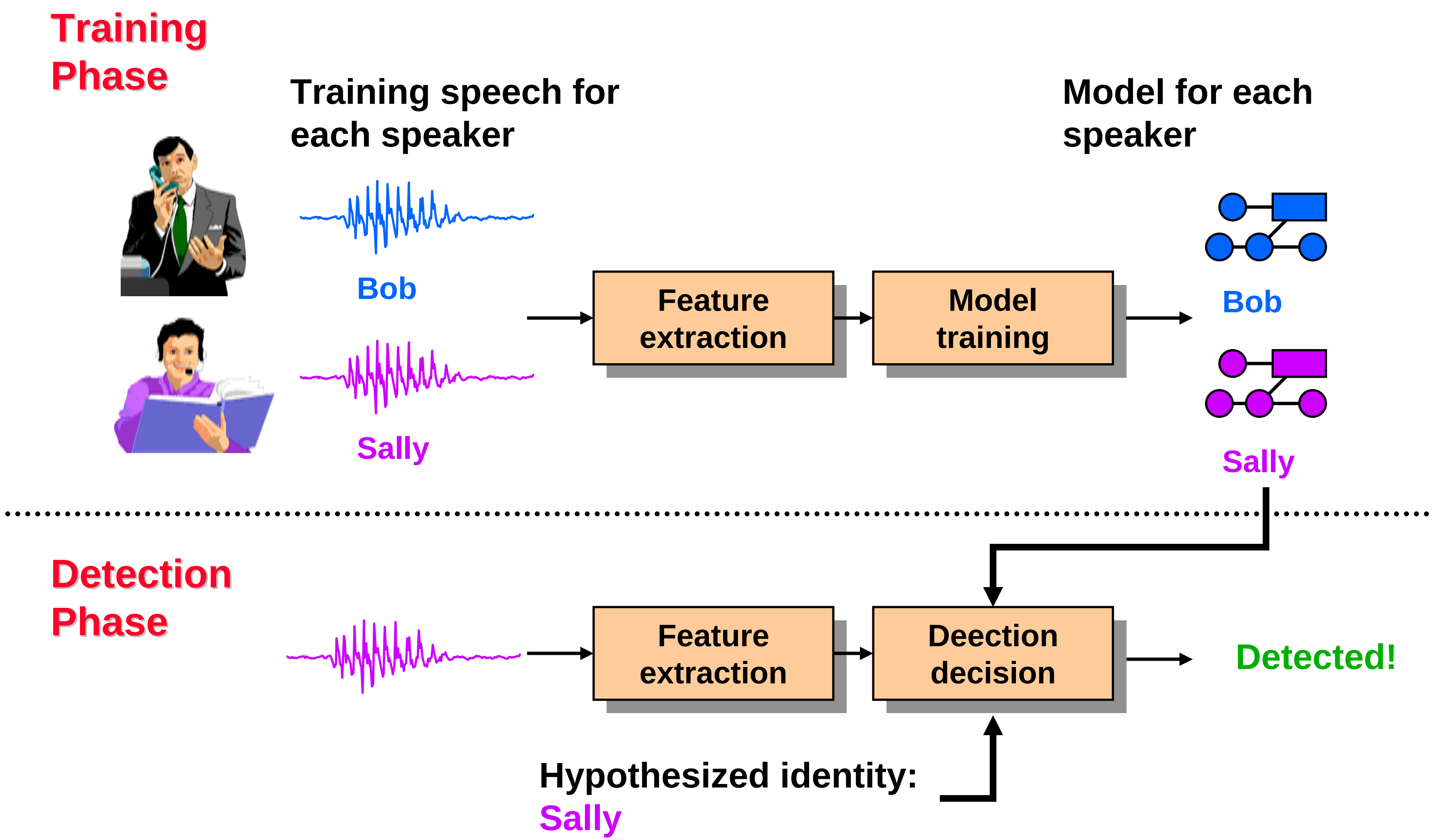
- **Statistic computed on test utterance S as likelihood ratio:**

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



Phases of Speaker Detection System

Two distinct phases to any speaker detection system



Features for Speaker Recognition

- **Humans use several levels of perceptual cues for speaker recognition**

High-level cues (learned traits)

Low-level cues (physical traits)

Hierarchy of Perceptual Cues

Semantics, idiolect, pronunciations, idiosyncrasies	**Socio-economic status, education, place of birth**
Prosodics, rhythm, speed intonation, volume modulation	**Personality type, parental influence**
Acoustic aspect of speech, nasal, deep, breathy, rough	**Anatomical structure of vocal apparatus**

Difficult to automatically extract

Easy to automatically extract

- **There are no exclusive speaker identity cues**
- **This workshop will primarily focus on acoustic cues**

Features for Speaker Recognition

- Desirable attributes of features for an automatic system (Wolf '72)

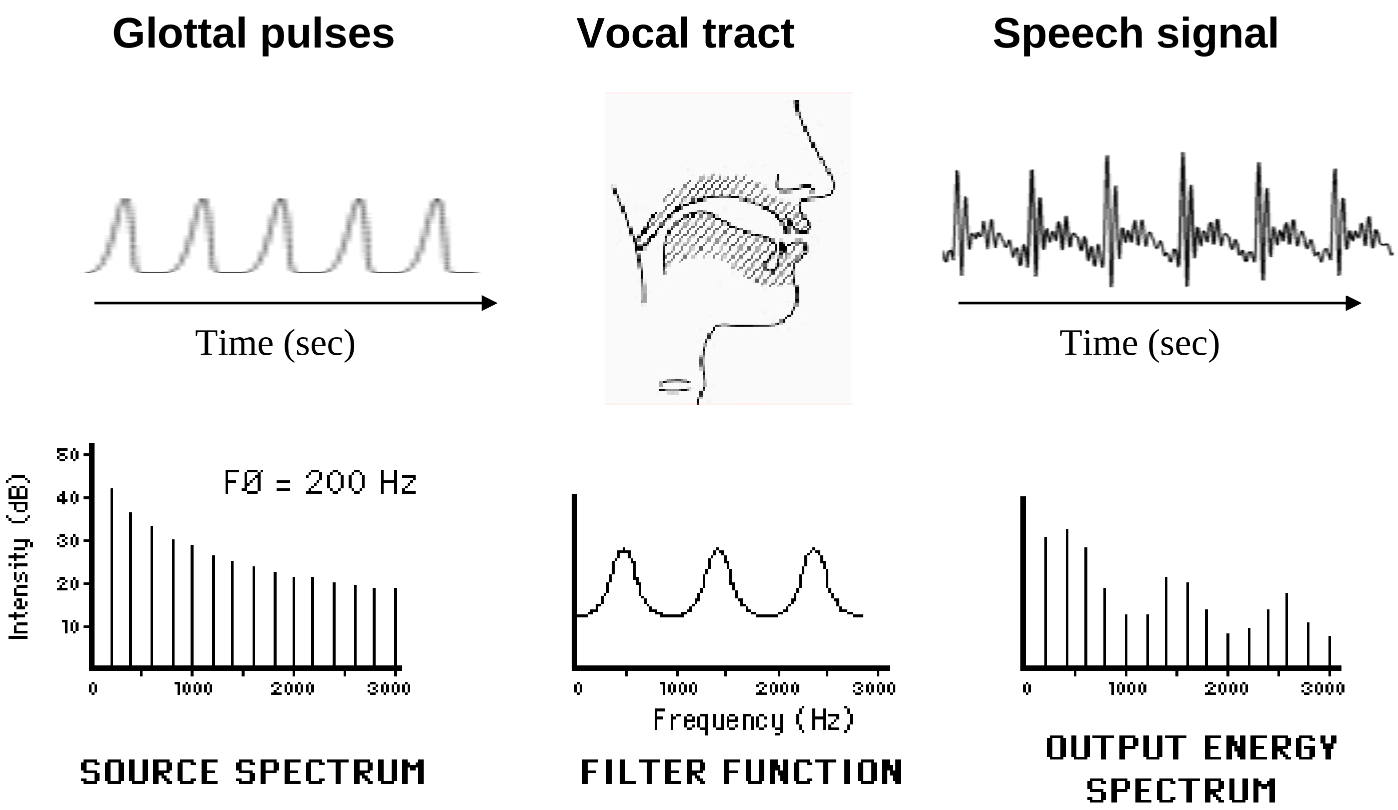
Practical	• Occur naturally and frequently in speech • Easily measurable
Robust	• Not change over time or be affected by speaker's health • Not be affected by reasonable background noise nor depend on specific transmission characteristics
Secure	• Not be subject to mimicry

- No feature has all these attributes

- Features derived from spectrum of speech have proven to be the most effective in automatic systems

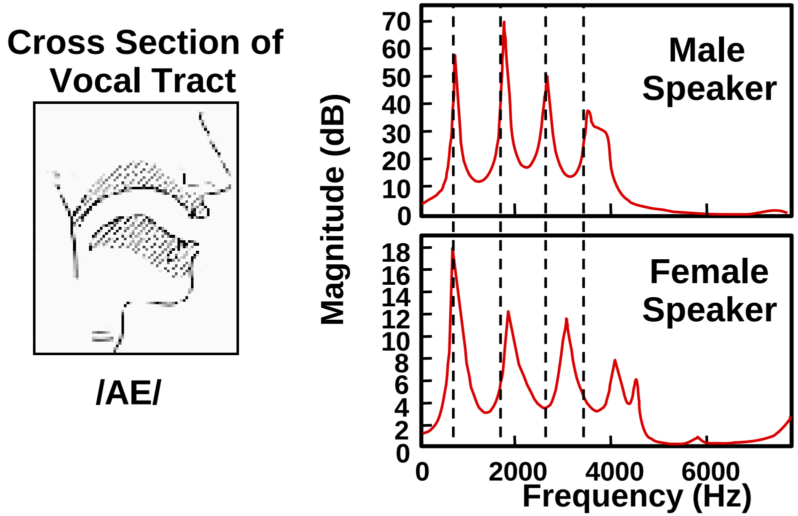
Speech Production

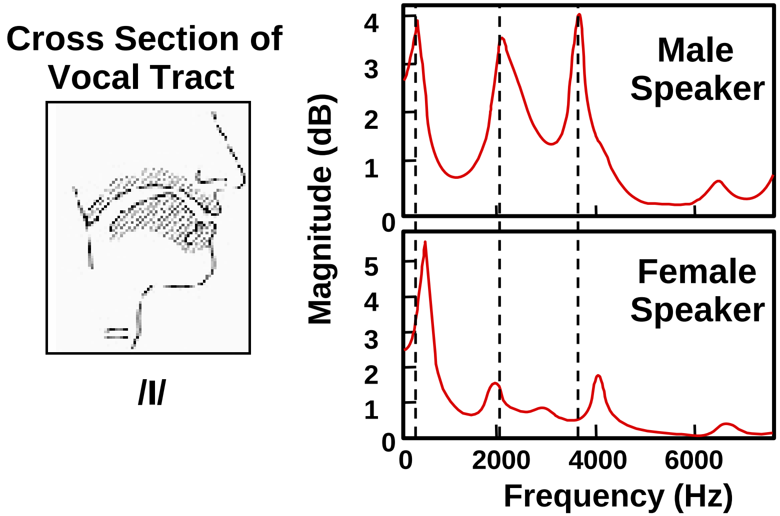
- **Speech production model: source-filter interaction**
  - Anatomical structure (vocal tract/glottis) conveyed in speech spectrum



Vocal Tract Configurations

- Different speakers will have different spectra for similar sounds

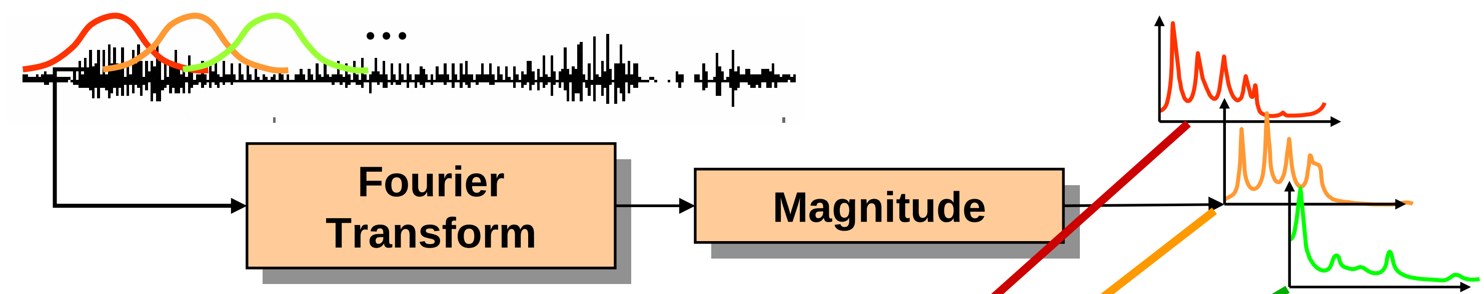




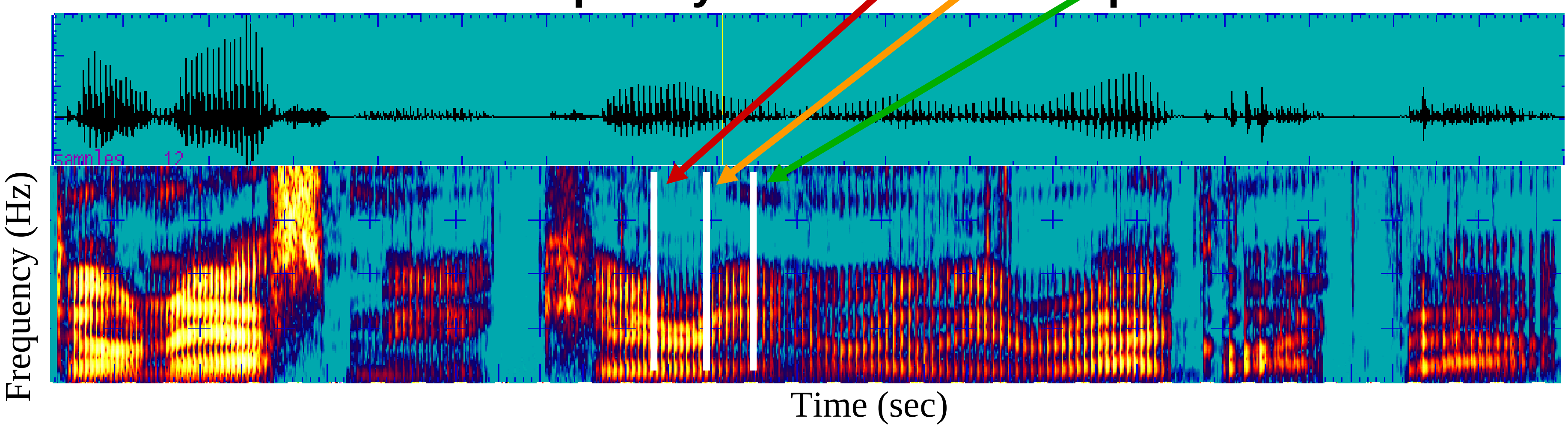
- Differences are in location and magnitude of peaks in spectrum
  – Peaks are known as formants and represent resonances of vocal cavity
- The spectrum captures the formant location and, to some extent, pitch without explicit formant or pitch tracking

Spectral Analysis

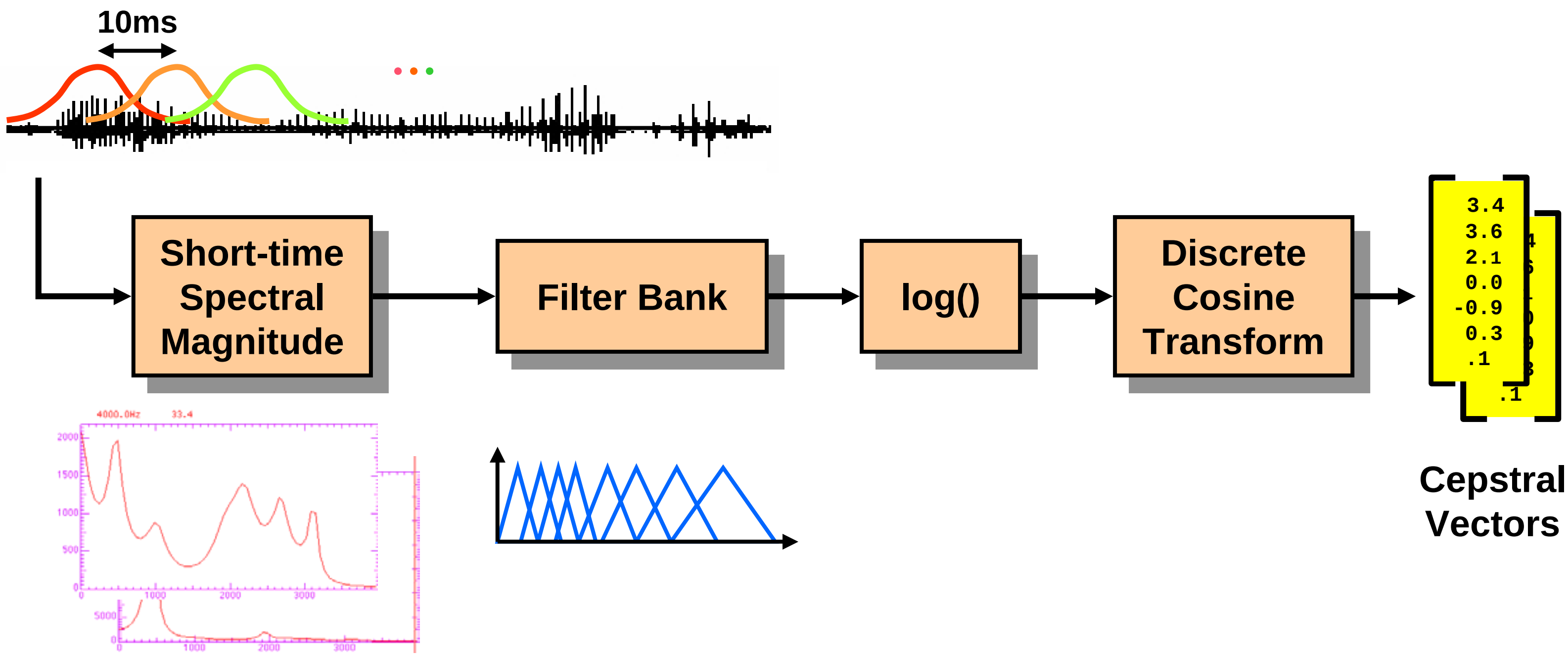
- **Speech is a continuous evolution of the vocal tract**
  - Need to extract time series of spectra
  - Use a sliding window - 20 ms window, 10 ms shift



- **Produces time-frequency evolution of the spectrum**



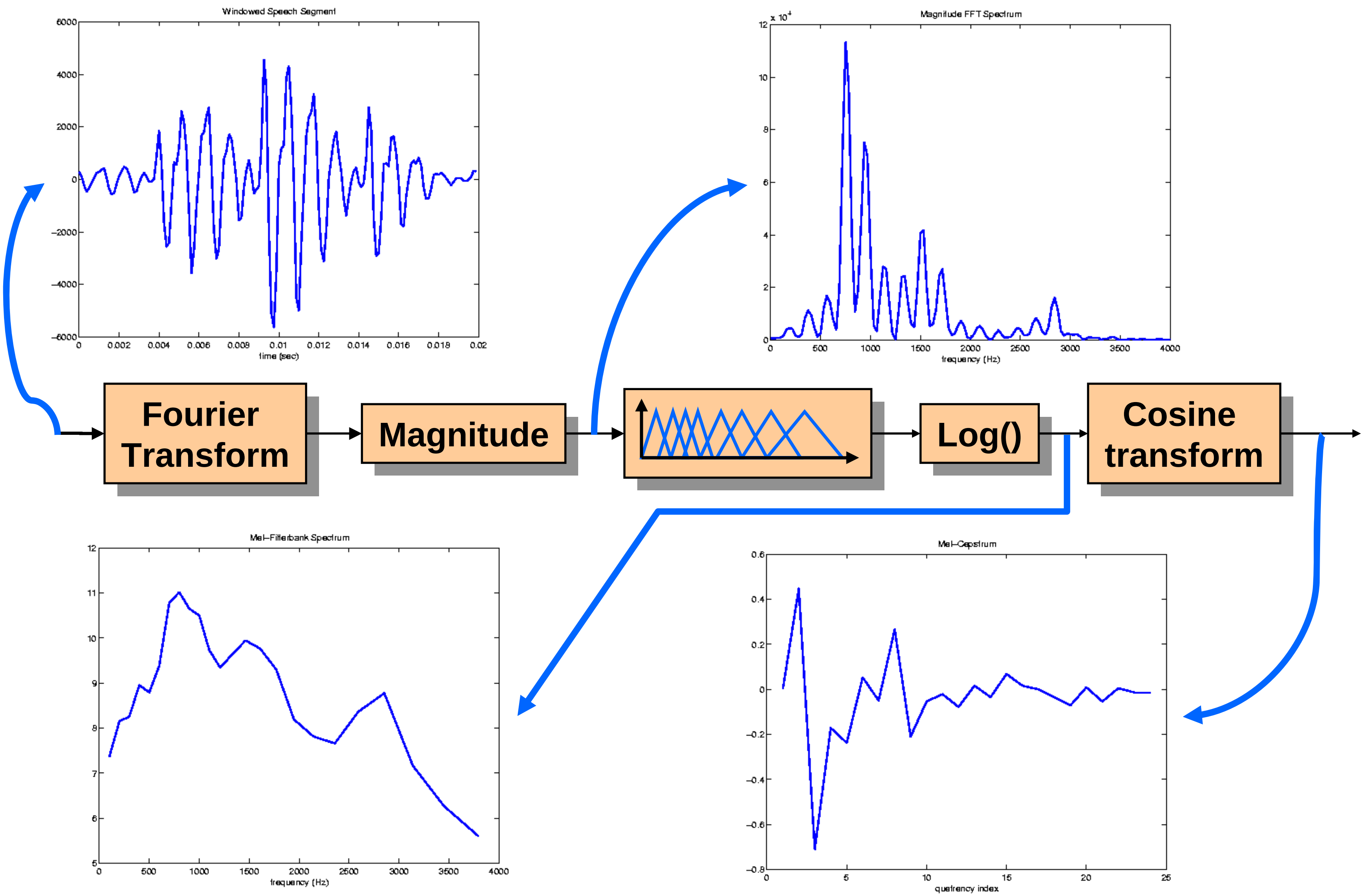
Spectral Features



- **Difference (delta) cepstra are often appended to vector**
- **Typical feature vector dimension: 25-49**
- **Additional front-end processing**
  - **Speech activity detection**
  - **Compensation for channel variability**

    blind deconvolution, mean and variance norm, etc.

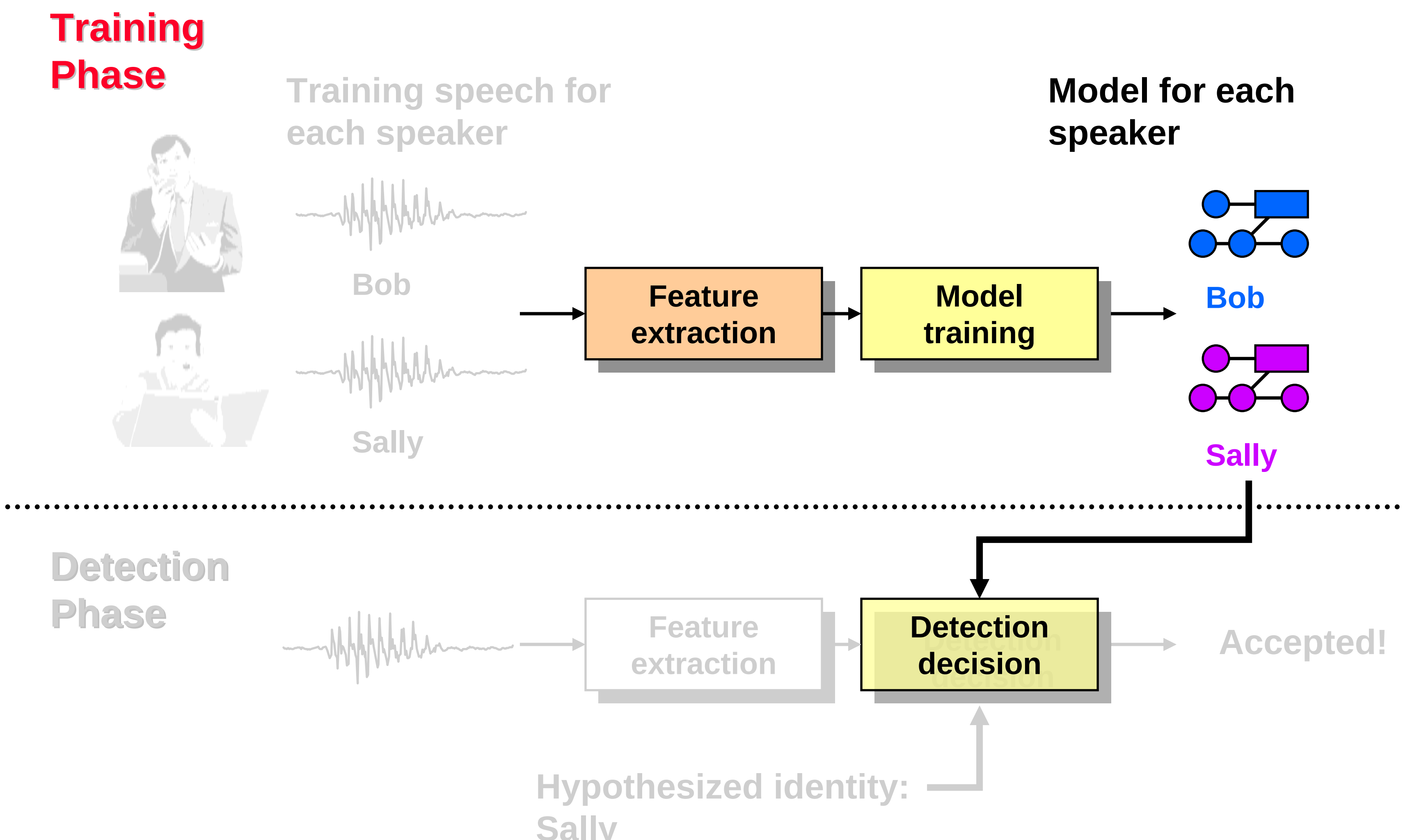
Spectral Features



Phases of Speaker Detection System

Speaker Models

Two distinct phases to any speaker Detection system



Speaker Models

- **Speaker models are used to represent the speaker-specific information conveyed in the feature vectors**

- **Desirable attributes of a speaker model**
  - **Theoretical underpinning**
  - **Generalizable to new data**
  - **Parsimonious representation (size and computation)**

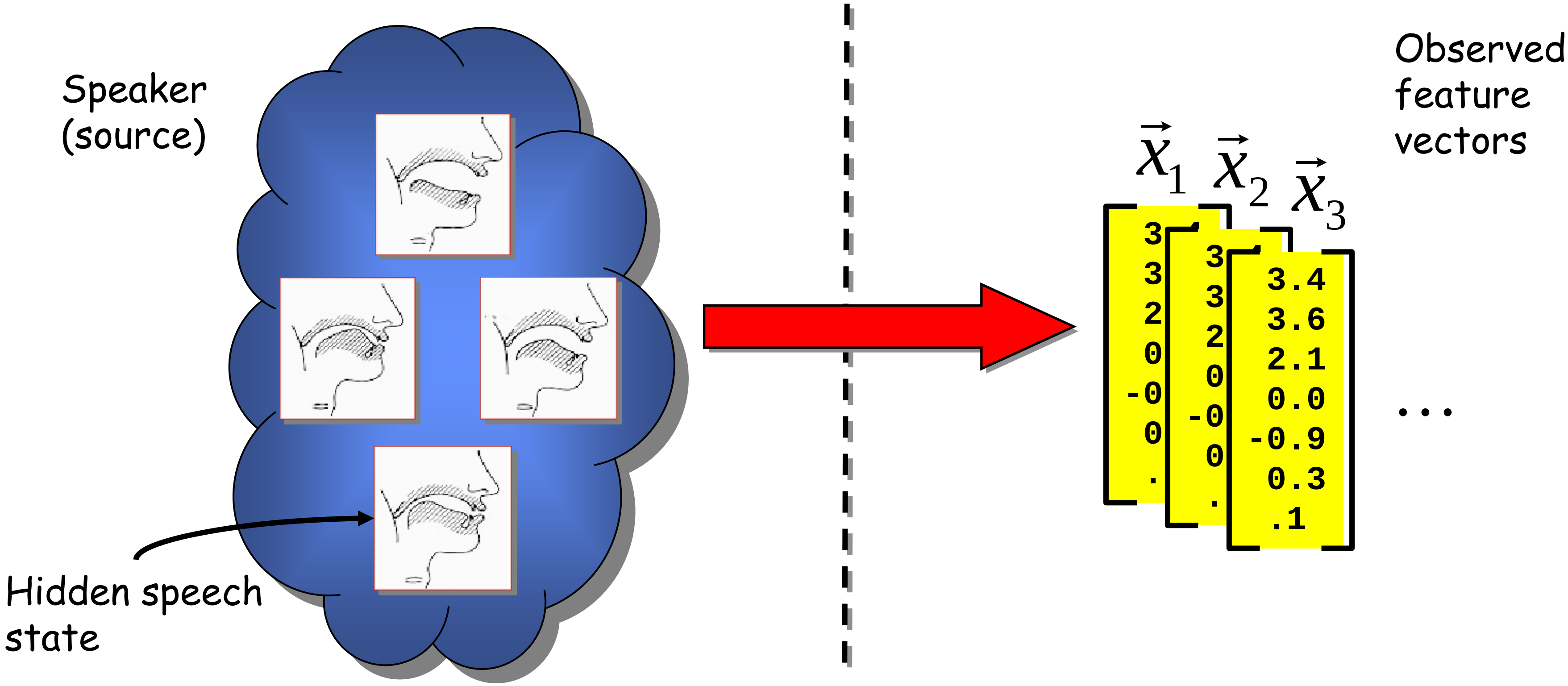
- **Many different modeling techniques have been applied to speaker recognition problems**
  - **Generative, discriminative, parametric, non-parametric, etc.**
  - **We will focus on two popular and successful approaches**

GMM-UBM – Gaussian Mixture Models adapted from a Universal Background Model

SVM-GSV – Support Vector Machines using GMM SuperVectors

Gaussian Mixture Model

- **Treat speaker as a hidden random source generating observed feature vectors**
  - **Source has "states" corresponding to different speech sounds**

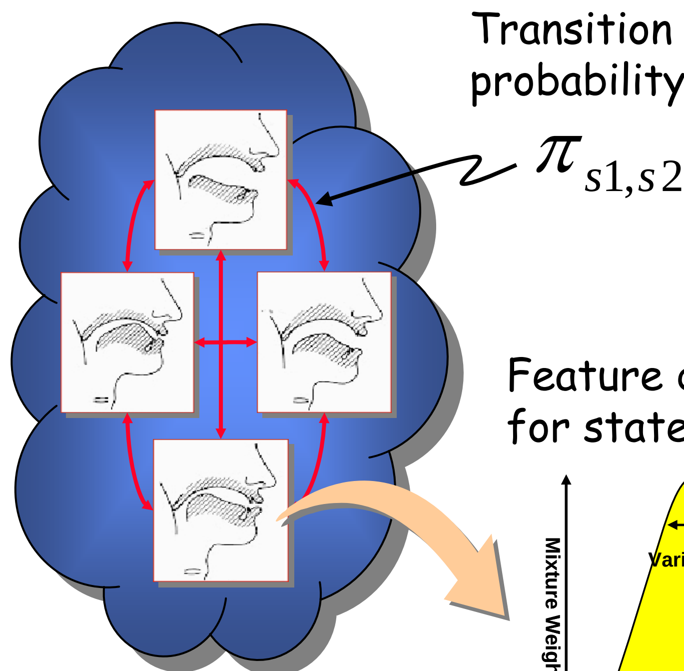


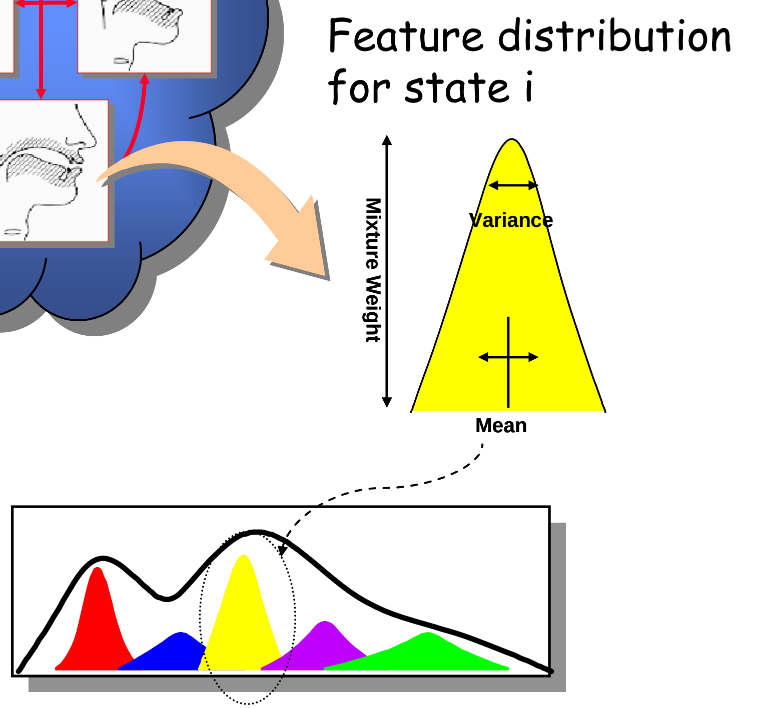
Gaussian Mixture Model

- **Hypothesize feature vectors generated from each state follow a Gaussian distribution**
  - **Total pdf is a Hidden Markov Model**
- **Transition between states based on modality of speech**
  - **Text-dependent case will have ordered states**
  - **Text-independent case will allow all transitions**
- **For text-independent case, pdf is a Gaussian Mixture Model**

$$p(\vec{x} \mid \lambda_s) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \qquad \lambda_s = (p_i, \vec{\mu}_i, \Sigma_i)$$

- **Parameters can be estimated from training speech using Expectation Maximization (EM) algorithm**

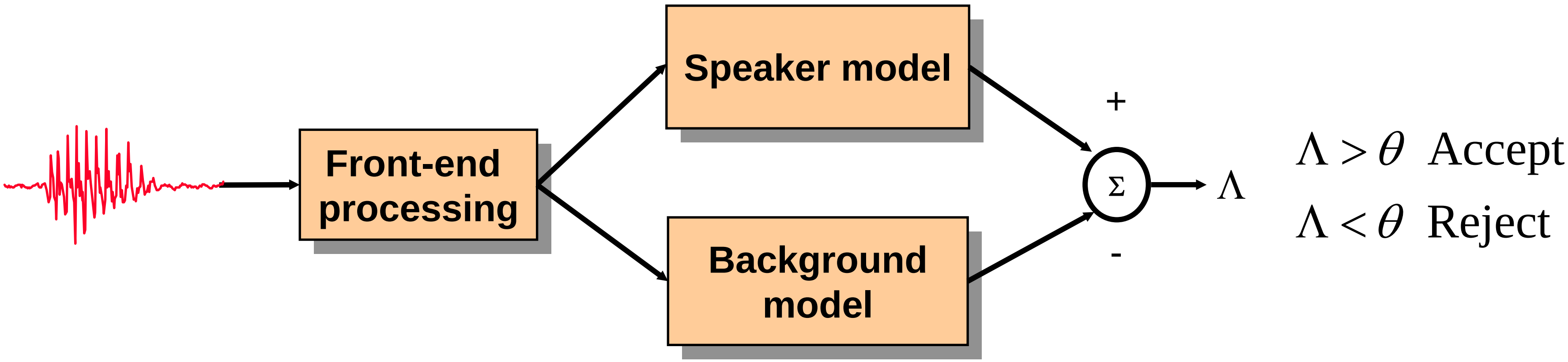




Gaussian Mixture Model
Background Model

- **We now can use the GMM to compute a log-likelihood ratio score**

$$LLR = \Lambda = \log p(S \mid H1) - \log p(S \mid H0)$$



- **The H1 likelihood is computed using the claimed speaker GMM**
- **But we also need an alternative model for H0 likelihood**

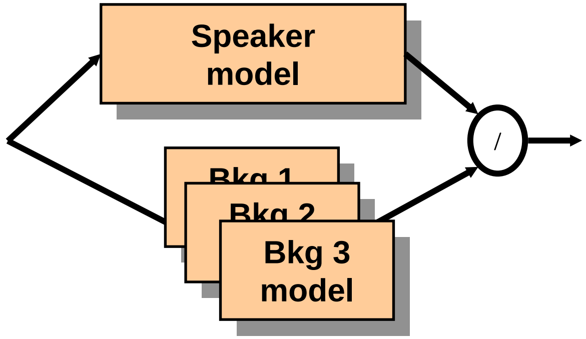
Background Modeling

- There are two main approaches for creating an alternative model for the likelihood ratio test

Cohorts/Likelihood Sets/Background Sets (Higgins, DSPJ91)

– Use a collection of other speaker models

– The likelihood of the alternative is some function, such as average, of the individual impostor model likelihoods



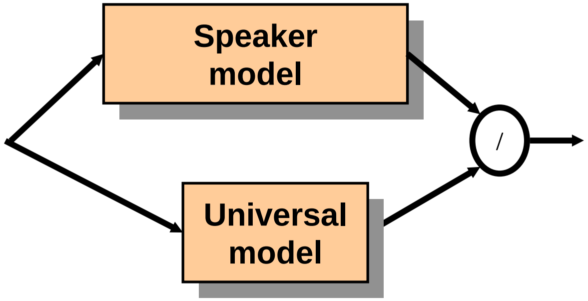
$$p(S \mid H0) = f\,(p(S \mid Bkg(b),\ b = 1,\ldots,B)$$

General/World/Universal Background Model (Carey, ICASSP91)

– Use a single speaker-independent model

– Trained on speech from a large number of speakers to represent general speech patterns

– Often MAP adaptation used to derive speaker model (Reynolds 96)

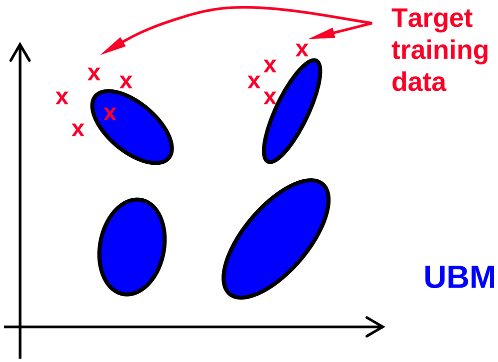


$$p(S \mid H0) = p(S \mid UBM)$$

GMM-UBM

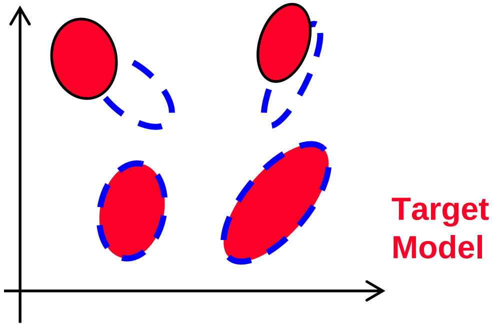
Relevance MAP from UBM

- **The target speaker model is derived from the UBM using unsupervised Bayesian adaptation**
  - **Probabilistically align target training data into UBM mixture states**
  - **Update mixture weights, means and variances based on the number of occurrences in mixtures**



- **Based on development experiments, only means are adapted**

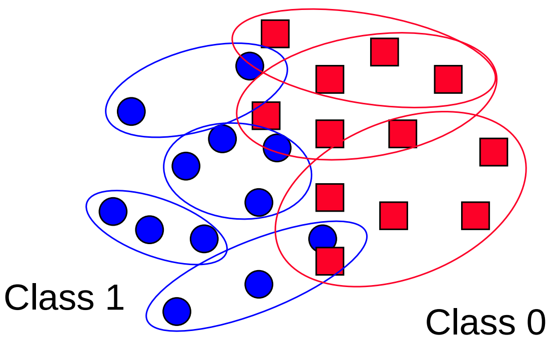
$$\mu_{\mathbf{tgt}} = \gamma\, \mu_{\mathbf{trn}} + (1-\gamma)\, \mu_{\mathbf{ubm}}$$
$$\gamma = n / (n+r)$$



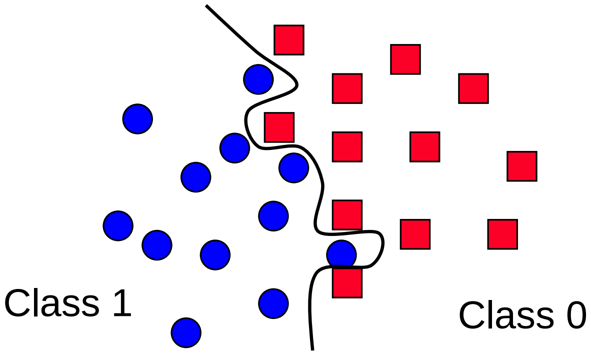
- **Adaptation only updates parameters representing acoustic events seen in target training data**
  - **Unseen events in testing do not count as evidence for or against target**
- **Other adaptation techniques can be applied**
  - **MLLR, Eigen-voices**

Generative vs. Discriminative Models
Support Vector Machines

- **The GMM-UBM is considered a *generative* model**
  - **The model is focused on representing the total distribution of the speaker data**
  - **Parameters estimated with Maximum likelihood or Maximum A-Posteriori criteria**
  - **Competition with other models comes through likelihood ratio**



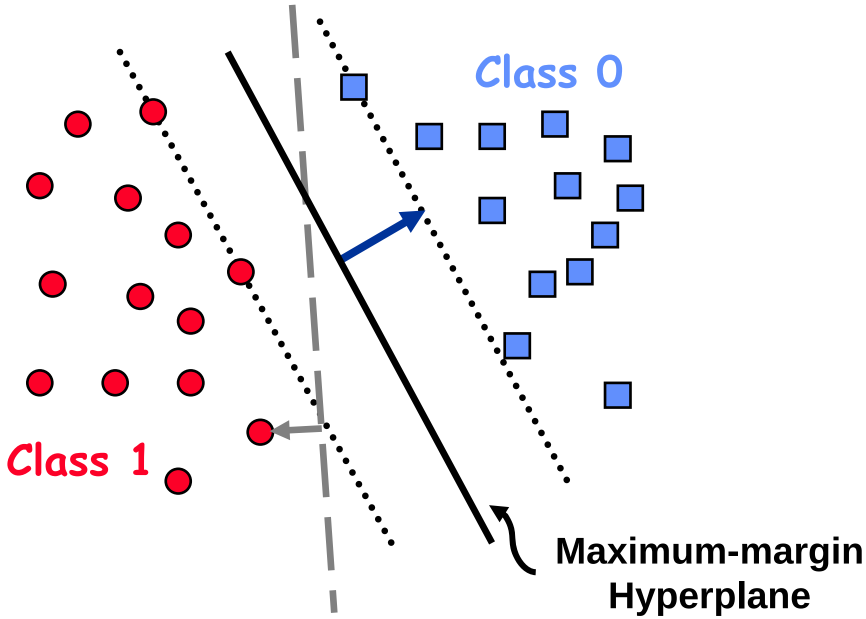
- **Support Vector Machines (SVMs) are an example of *discriminative* models**
  - **The model is focused on representing the boundary between competing speaker data**
  - **Parameters are estimated with a maximum margin (separation boundary) criteria**
  - **Competition with other classes directly optimized in model training**



Support Vector Machine
Maximum Margin Hyperplane

- **Margin:** ***Distance from the separating hyperplane to the nearest training sample***

- **Classifier that uses a maximum-margin separating hyperplane boundary provides good generalization**
  - **Minimizes expected classification error on unseen test samples**
  - **Only one hyperplane maximizes margin**



- **Can map non-separable data to higher dimensional space where a hyperplane can be found**
  - **x → b(x)**
  - **Define kernel (distance) in high-dimensional space**

$$K(\mathbf{x},\mathbf{y}) = \mathbf{b}(\mathbf{x})^t\mathbf{b}(\mathbf{y})$$

Support Vector Machine

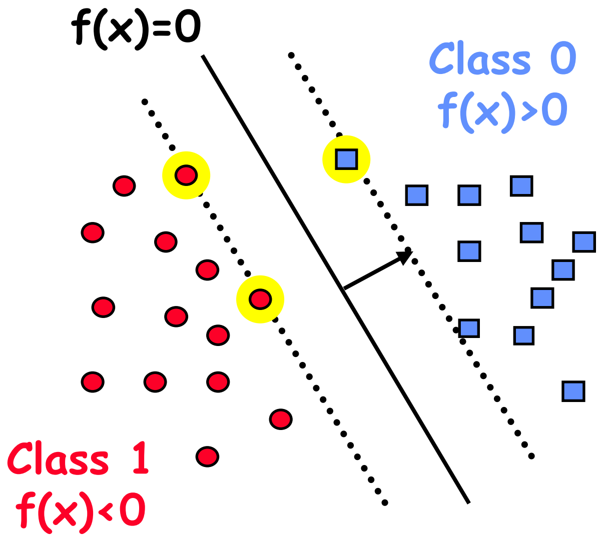
Support Vectors

- **SVM discriminant function**

$$f(\mathbf{x}) = \sum_i \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + c$$

where

$$\alpha_i = \text{ weights}$$
$$y_i = \pm 1 \text{ (class labels)}$$
$$K(\bullet, \bullet) = \text{ kernel function}$$
$$\mathbf{x}_i = \text{support vectors}$$



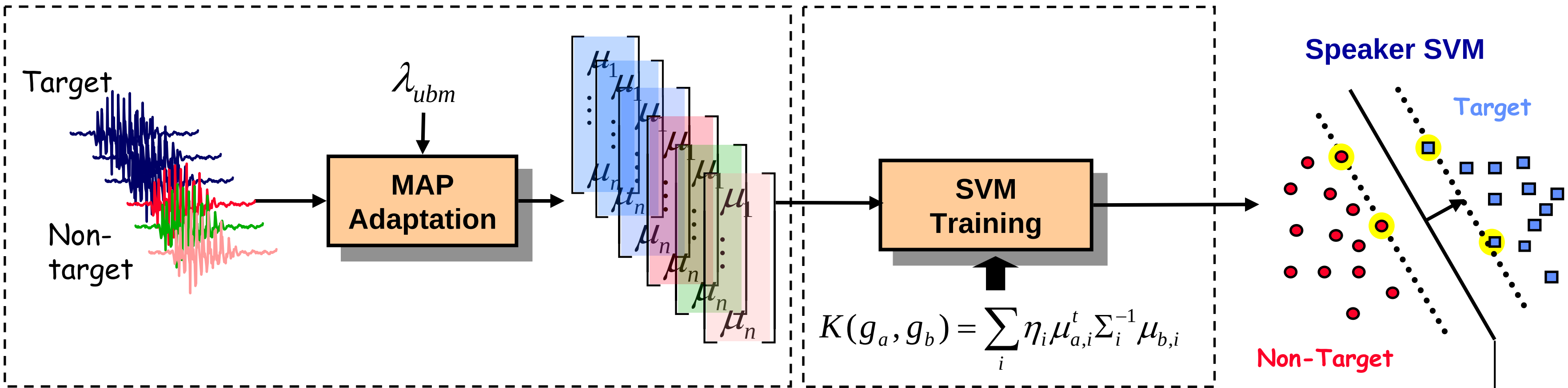
- **Number of training samples retained as support vectors is often small**
- **Projection into high-dimensional space can be explicit (b(x)) or implicit in kernel**
- **With explicit projection, scoring is a single dot-product**

$$f(\mathbf{x}) = \sum_i \alpha_i y_i b(\mathbf{x})^t b(\mathbf{x}_i) + c = b(\mathbf{x})^t \left[\sum_i \alpha_i y_i b(\mathbf{x}_i) \right] + c = b(\mathbf{x})^t \mathbf{m} + c$$

SVM using GMM Super Vectors

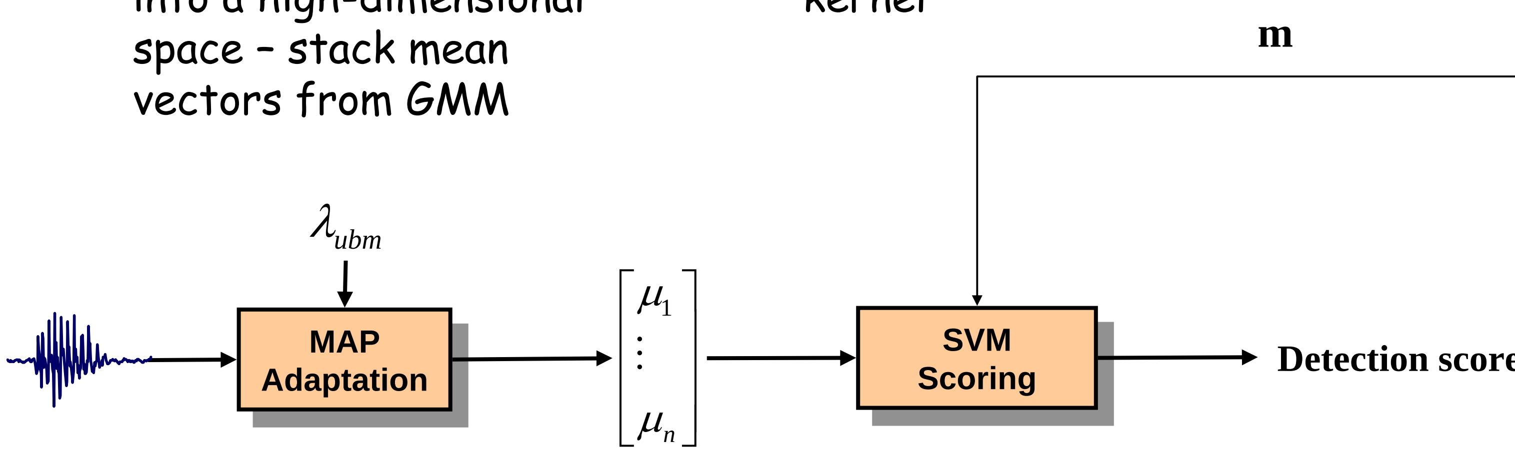
SVM-GSV

- **The SVM-GSV is a merging of GMM-UBM and SVM classifiers**



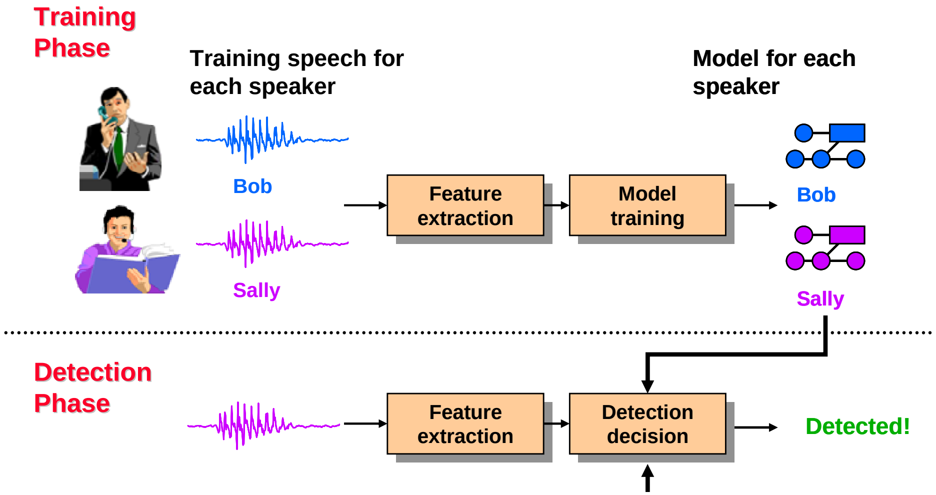
Project each utterance into a high-dimensional space – stack mean vectors from GMM

Use a KL-based kernel



Speaker Detection Systems

Feature/Models Recap



- **Cepstral Features:**
  - Capture salient speaker information from speech signal
  - Short-time spectral features convey information about vocal apparatus

- **GMM-UBM models:**
  - GMMs model the distribution of feature vectors (generative)
  - Roughly capture underlying sound classes in speech
  - Likelihood ratio formed with a UBM
  - MAP adaptation from UBM used to derive speaker models

- **SVM-GSV models:**
  - SVMs model the boundary between classes (discriminative)
  - GMM-UBM stacked mean vectors form SuperVector
  - SVM learns speaker-dependent likelihood ratio

Speaker Detection Systems
Channel/Session Effects

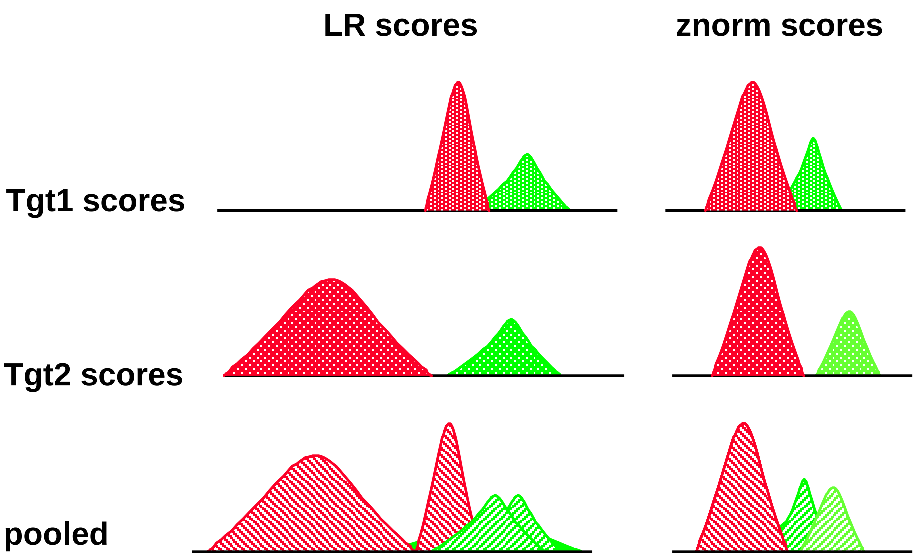
The largest challenge to practical use of speaker detection systems is channel/session variability

- **Variability** refers to changes in channel effects between training and successive detection attempts
- **Channel/session effects** encompasses several factors
  - The microphones
    - Carbon-button, electret, hands-free, array, etc
  - The acoustic environment
    - Office, car, airport, etc.
  - The transmission channel
    - Landline, cellular, VoIP, etc.
- Anything which affects the spectrum can cause problems
  - Speaker and channel effects are bound together in spectrum and hence features used in speaker verifiers
- Channel/session compensation occurs at several levels
  - Features: blind-deconvolution
  - Models: Eigen-channels
  - Scores: Z-norm, T-norm

Z Score Normalization
Znorm

- **Target model LR scores have different biases and scales for test data**
- **Znorm attempts to remove these bias and scale differences from the LR scores**
  - **Estimate mean and standard-deviation of non-target, same-sex utterances from data similar to test data**
  - **During testing normalize LR score**
  - **Align each model’s non-target scores to N(0,1)**

$$Z_{Tgt}(x) = \frac{\Lambda_{Tgt}(x) - \mu_{Tgt}}{\sigma_{Tgt}}$$

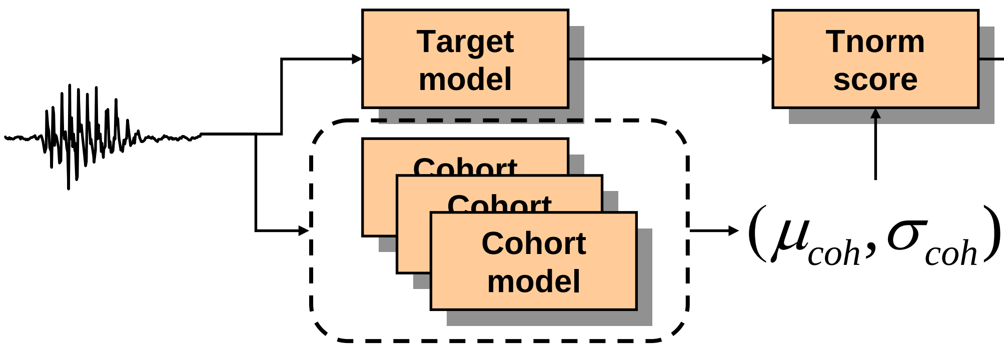


MIT Lincoln Laboratory

Test Score Normalization

Tnorm

- **Introduced in 1999 by Ensigma (DSP Journal January 2000)**
- **Estimates bias and scale parameters for score normalization using fixed “cohort” set of speaker models**
  - **Normalizes target score relative to a non-target model ensemble**
  - **Similar to standard cohort normalization except for standard deviation scaling**

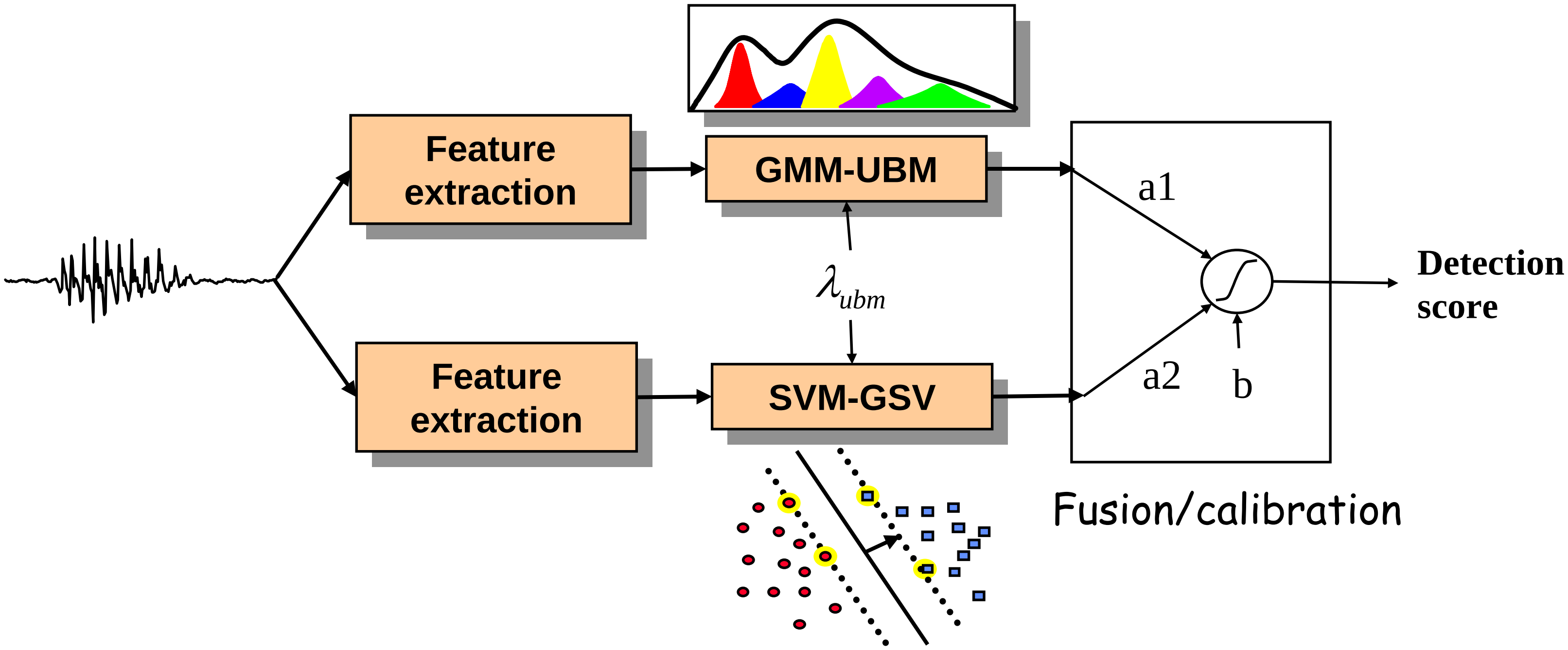


$$T_{tgt}(u) = \frac{\Lambda_{tgt}(u) - \mu_{coh}}{\sigma_{coh}}$$

- **Used cohorts of same gender as target**
- **Can be used in conjunction with Znorm**
  - **ZTnorm or TZnorm depending on order**

Speaker Recognition Systems
Score Fusion

- **Scores from different types of features/models can be combined using a simple fuser**
  - **Requires scores from some development data to train fuser**
- **A generalized linear regression fuser works well**
- **An added benefit is we can get calibrated scores**
  - **E.g. [0-1] posterior probability estimates**



Outline

- Background and Theory
  - Terminology
  - Components of recognition systems
    - Features and models

- **Evaluation and Performance**
  - **Evaluation metrics and design**
  - **Performance survey**

Evaluation Metrics

- In speaker detection, there are two types of errors that can occur

  Miss: incorrectly reject a target trial

  Also known as a false reject or Type-I error

  False Alarm: incorrectly accept a non-target trial

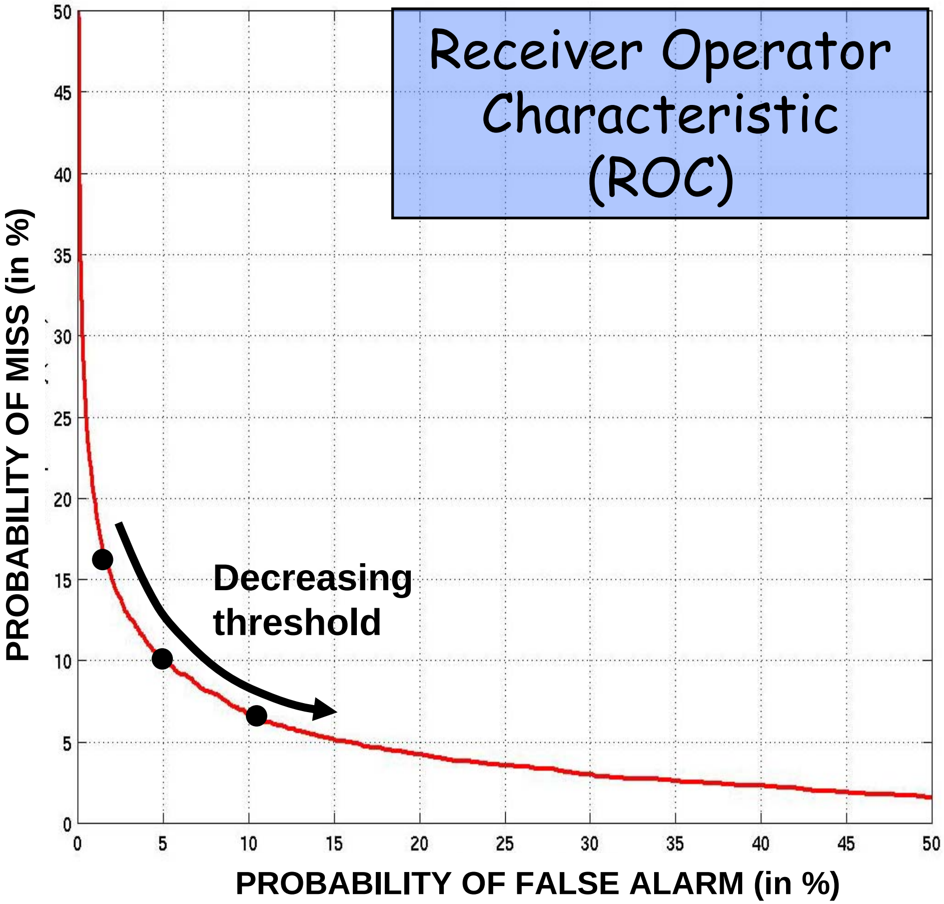
  Also known as a false accept or Type-II error

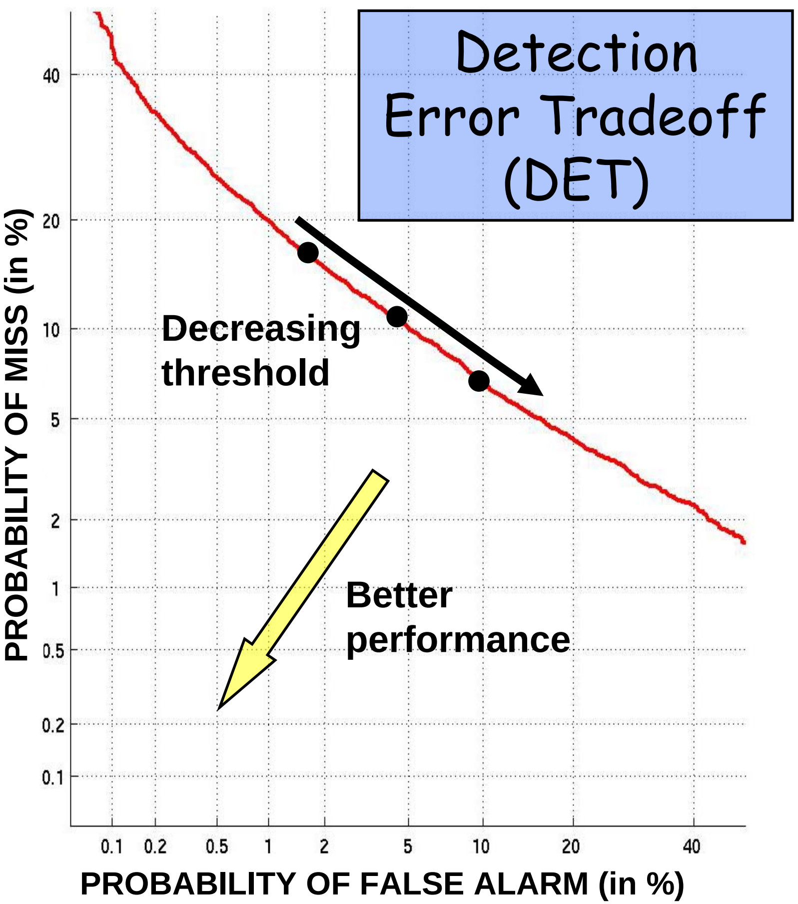
- The performance of a detection system is a measure of the trade-off between these two errors
  - The tradeoff is usually controlled by adjustment of the decision threshold
- In an evaluation, N_{target} target trials (test speaker = model speaker) and $N_{non\text{-}target}$ non-target trials (test speaker != model speaker) are conducted and error probabilities are estimated at threshold θ

$$\Pr(\text{miss} \mid \theta) = \frac{\#\ \text{target trial scores} < \theta}{\mathrm{N}_{\text{target}}} \qquad \Pr(\text{false alarm} \mid \theta) = \frac{\#\ \text{non-target trial scores} > \theta}{\mathrm{N}_{\text{non-target}}}$$

Evaluation Metrics
ROC and DET Curves

Plot of Pr(miss) vs. Pr(fa) shows system performance
DET plots Pr(miss) and Pr(fa) on normal deviate scale

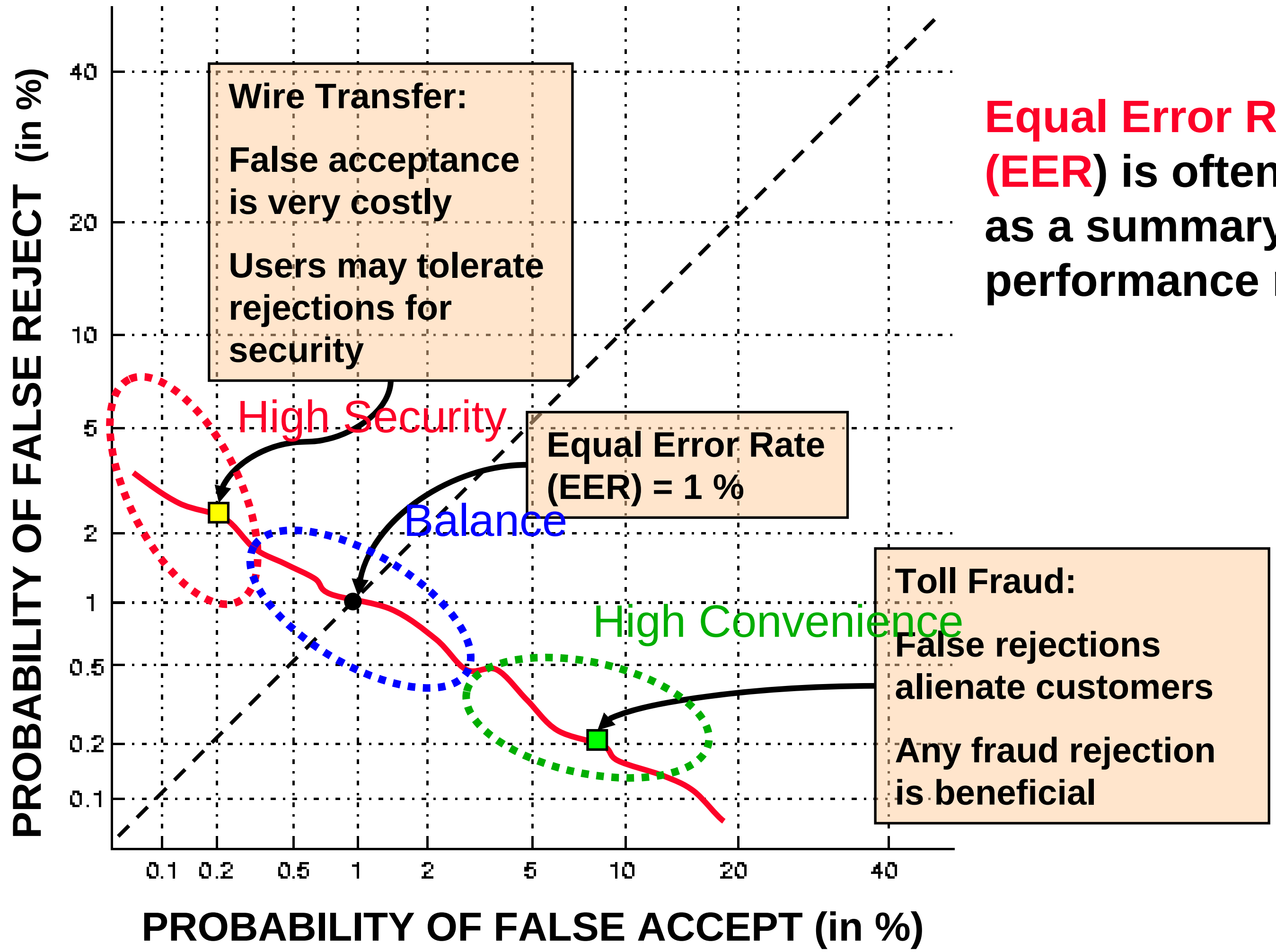




Evaluation Metrics

DET Curve

Application operating point depends on relative costs of the two errors



Equal Error Rate (EER) is often quoted as a summary performance measure

Evaluation Metrics
Decision Cost Function

- **In addition to EER, a decision cost function (DCF) is also used to measure performance**

$$\boxed{\mathrm{DCF}(\theta) = \mathrm{C(miss)Pr(tgt)\,Pr(miss} \mid \theta) + \mathrm{C(fa)Pr(non)Pr(fa} \mid \theta)}$$

C(miss) = cost of a miss

Pr(tgt) = prior probability of target trial

C(fa) = cost of a false alarm

Pr(non) = 1-Pr(tgt) = prior probability of non-target trial

- **For application specific costs and priors, we can compare systems based on value of DCF**

Evaluation Design

Data Selection Factors

- **Performance numbers are only meaningful when evaluation conditions are known**

Speech quality	– Channel and microphone characteristics – Ambient noise level and type – Variability between enrollment and verification speech
Speech modality	– Fixed/prompted/user-selected phrases – Free text
Speech duration	– Duration and number of sessions of enrollment and verification speech
Speaker population	– Size and composition – Experience

The evaluation data and design should match the application domain of interest

Evaluation Design

NIST Speaker Recognition Evaluations

- **Annual NIST evaluations of speaker verification technology (since 1995)**
- **Aim:** Provide a common paradigm for comparing technologies
- **Focus:** Conversational telephone & microphone speech (text-independent)

Technology Consumers

Application domain / parameters

Data Provider

Linguistic Data Consortium

Evaluation Coordinator

NIST Spoken Natural Language Processing

Comparison of technologies on common task

Technology Developers

MIT Lincoln Laboratory

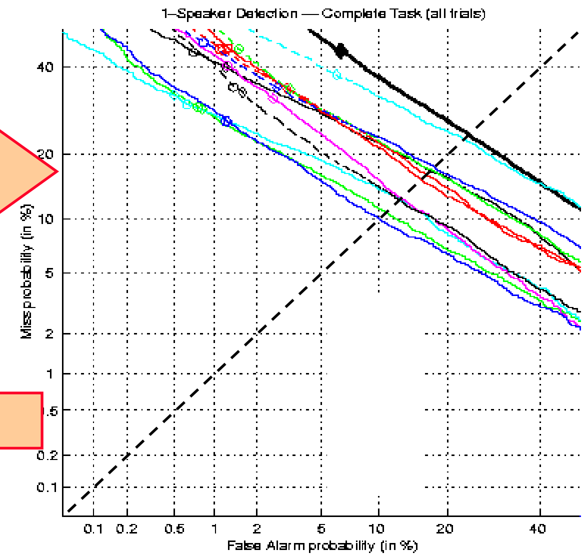
Institute for Infocomm Research A*STAR

Brno University of Technology, Faculty of Information Technology

Laboratoire d'Informatique, Université d'Avignon

Evaluate

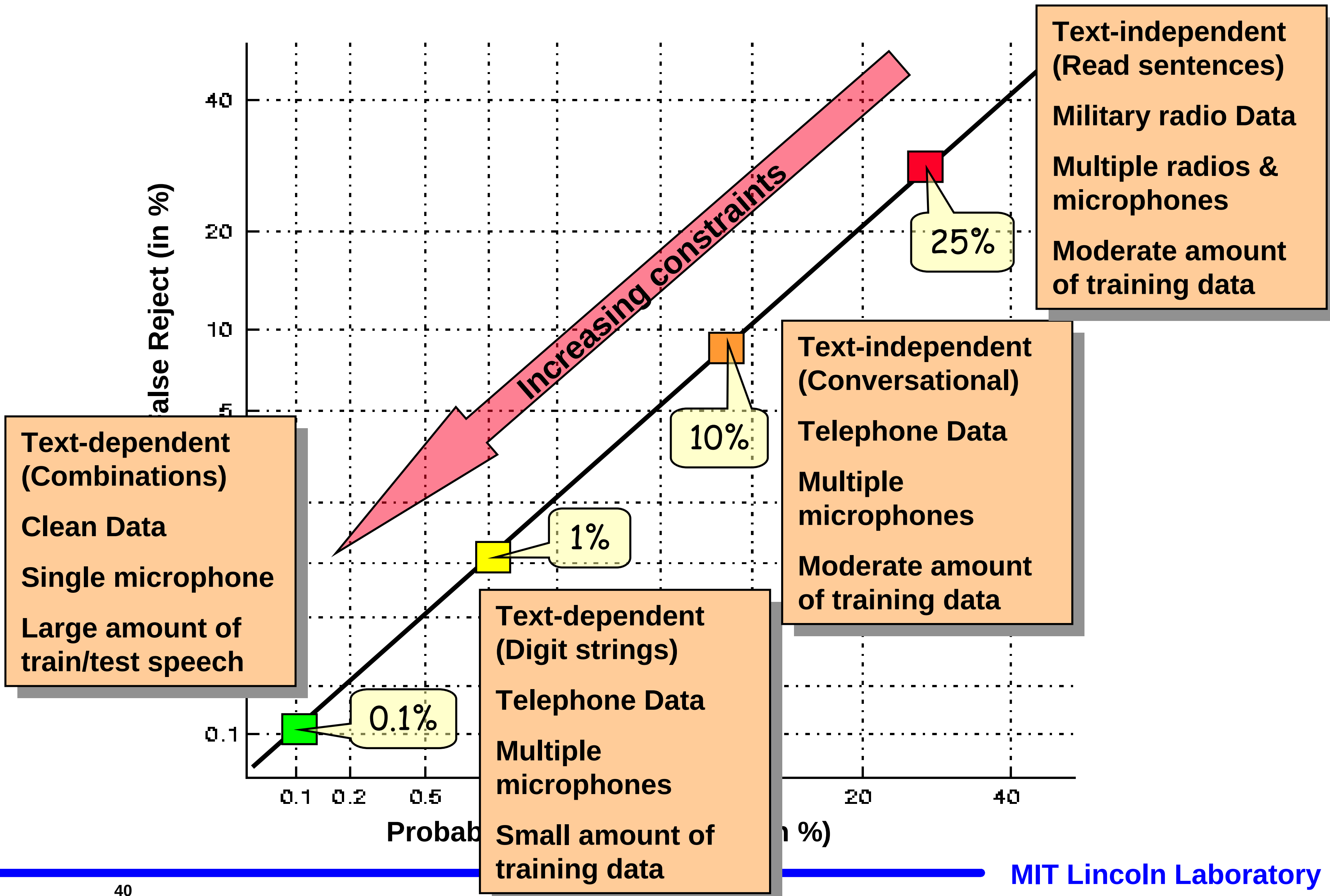
Improve



http://www.nist.gov/speech/tests/spk/index.htm

Performance Survey

Range of Performance



Performance
NIST SRE 2008

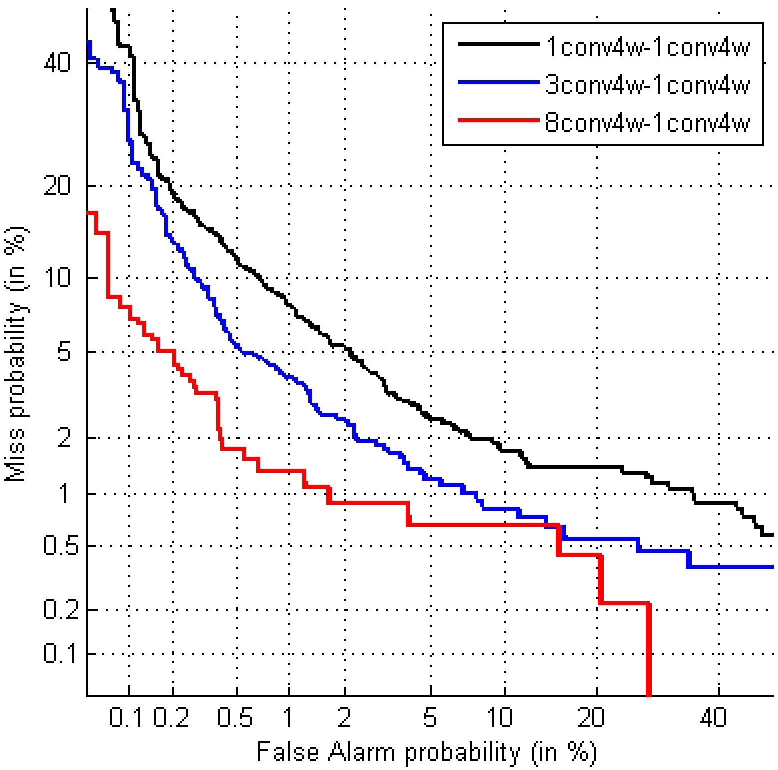
- **Large number of conditions broadly covering**
  - Language: English, non-English (33 languages represented)
  - Channels: Telephone, microphones (various locations)
  - Sessions: Multiple 2.5 minute training telephone calls
  - Duration: Train and test with 10 sec of speech
  - Mutli-speakers: More than one speaker in train/test data
- **46 sites participated employing > 100 systems for all conditions**
  - Many variations and different system fusions
  - GMM-UBM, SVM-GSV and channel compensation common components over almost all systems
- **Workshop will focus on 1-2 conditions from SRE08**
  - Telephone 1-8 conversation train, 1 conversation test
  - Cross-microphone train/test

SRE08

Telephone Train/Test

- **Results are representative of most GMM-UBM and SVM-GSV systems**
- **Language and channel/session can have large effects**

English train/test

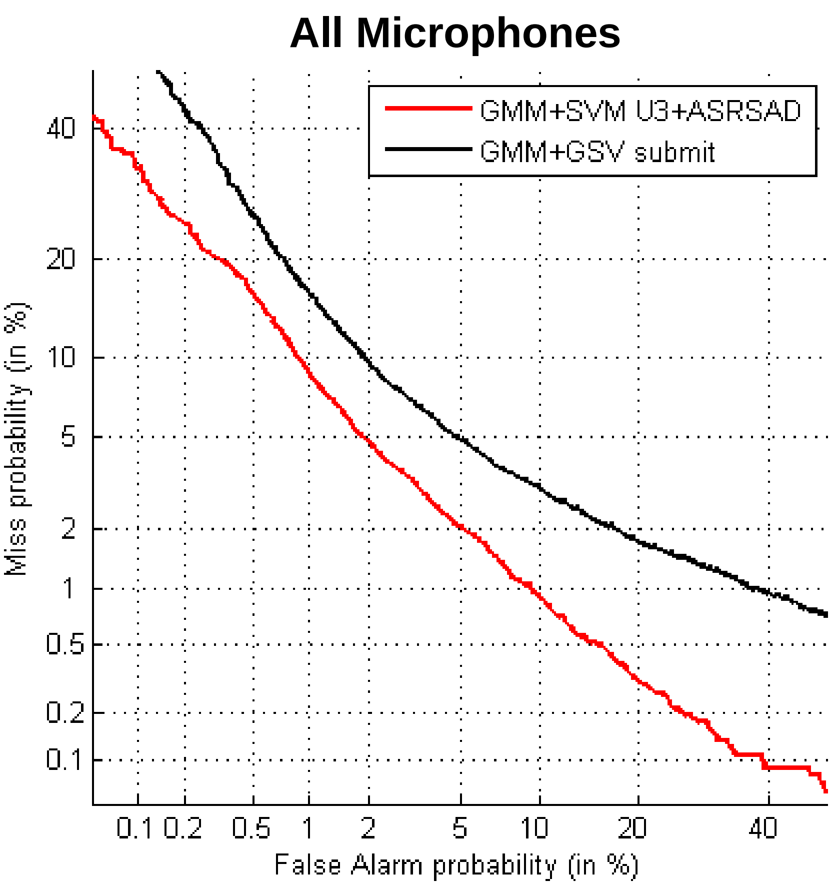


GMM +GSV	US-ENG		ENG		ALL	
	EER	DCF	EER	DCF	EER	DCF
1c/1c	3.7	1.6	3.6	1.7	6.0	2.9
8c/1c	1.5	0.43	1.3	0.57	2.4	1.3

SRE08

Interview Microphones Train/Test

- **Analysis found improvements with better speech activity detection and channel compensation**



ALL GMM+ GSV	3U+ASRSAD		Submit	
	EER	DCF	EER	DCF
Intmic/ intmic	3.3	1.9	5.0	2.6

Summary

- **This talk provided a broad overview of speaker recognition technology conveying**
  - **An understanding of the major concepts behind modern speaker recognition systems**
    - **Feature and models**
  - **The identification of key elements in evaluating performance of a speaker recognition system**
  - **An indication of the range of expected performance**

- **The following talk will focus on new and powerful techniques used with speaker recognition systems to improve robustness and accuracy**