Automatic Speaker Recognition
Brief Introduction

Alberto Abad

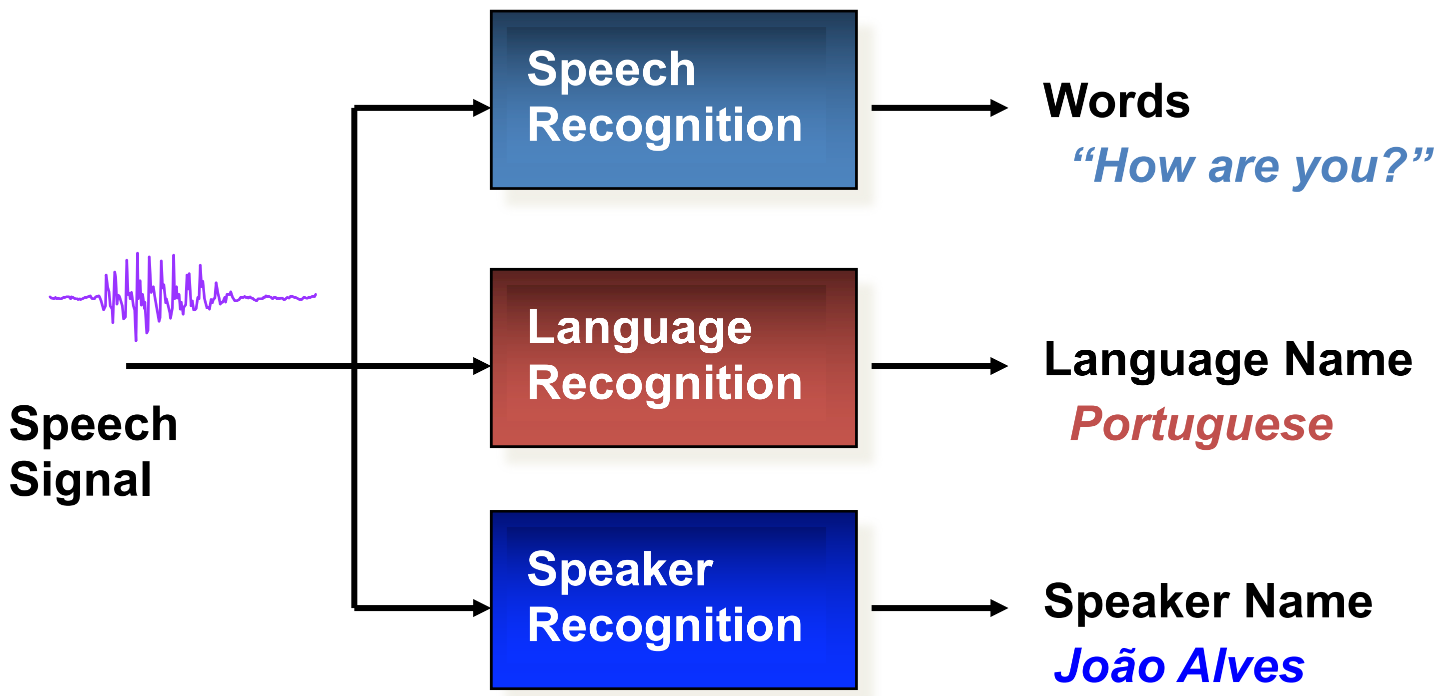
IST/INESC-ID Lisboa, Portugal

alberto.abad@tecnico.ulisboa.pt

Speech Processing - **IST** Lisboa, April 2020

Speech processing

Example of classical applications



Speech processing: Speech coding, Speech enhancement, Audio segmentation, Text-to-speech synthesis, Automatic speech recognition, Speaker and language identification

Text processing: Morphological analysis, Syntactic analysis, Semantic analysis, Discourse analysis, Named entity extraction, NL Generation, Information retrieval, Summarization, Question answering, Machine translation, Text analytics

Spoken language processing Speech understanding, Speech synthesis from concepts, Spoken/multimodal dialog systems, Classification of multimedia documents, Summarization of spoken documents, Question answering on multimedia documents, Rich Transcription of multimedia documents, Speech-to-speech machine translation, Speech analytics

Image after [1]

Voice biometrics

- Biometric authentication paradigm:

What you posses/know
Token, password, etc.

What you are (physiological)
DNA, fingerprints, iris, …

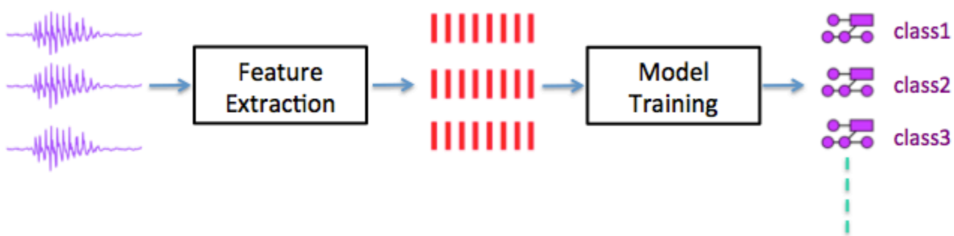
What you produce (behavioural)
Voice, signature, …

- Speech/voice is one form of biometric that carries lots of personal (identity) information:
  - Gender, age, accent, region, social class, illnesses (cold), style of speaking, mood, etc.
- Some advantages/particularities of voice:
  - It allows for remote authentication
  - Non intrusiveness
  - Low cost and wide availability
  - Ease of transmission, small storage space

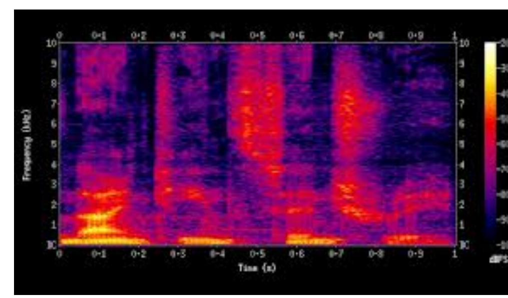
Voice biometrics
Preliminary considerations

- Voice biometrics can be seen as a common pattern classification problem, but with the particularities of **SPEECH** pattern classification problems:
  - Most important one is the time nature of input (and in some cases also output)

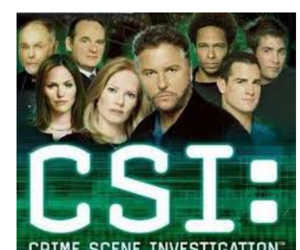
- Learning/Training phase



- Classification phase



- Some extra cautions (before going into detail):
  - Wrong idea → graphical representation of speech based on spectrogram is as reliable as a fingerprint or DNA
  - False premise → All voices are unique (and discernable)



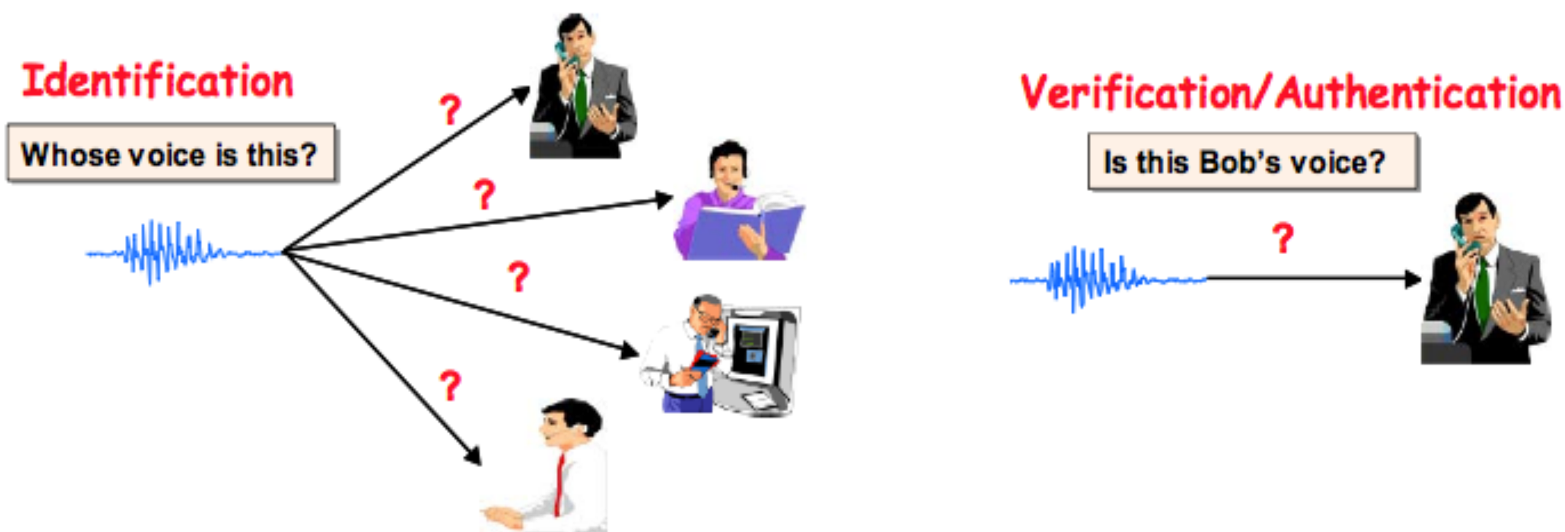
Outline

- Automatic Speaker recognition
  - Intro
  - Classical approaches:
    - Features
    - Models
  - The problem of inter-session variability
  - Advanced topics

- Evaluation and performance of speaker verification
  - Evaluation measures
  - SRE evaluation challenges
    - NIST SRE
    - NIST HASR

Introduction to SR
Speaker recognition Tasks

Identification vs Verification



– Closed-set vs open-set **identification** (the *unknown* option)

Slide after [1]

Introduction to SR
Speech modalilties

Application dictates different speech modalities:

- **Text-dependent recognition**
  - Highly constrained 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**
  - Unconstrained 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

Slide after [1]

Introduction to SR

Speaker recognition applications

Access Control
Physical facilities
Computer networks and websites

Transaction Authentication
Telephone banking
Remote credit card purchases

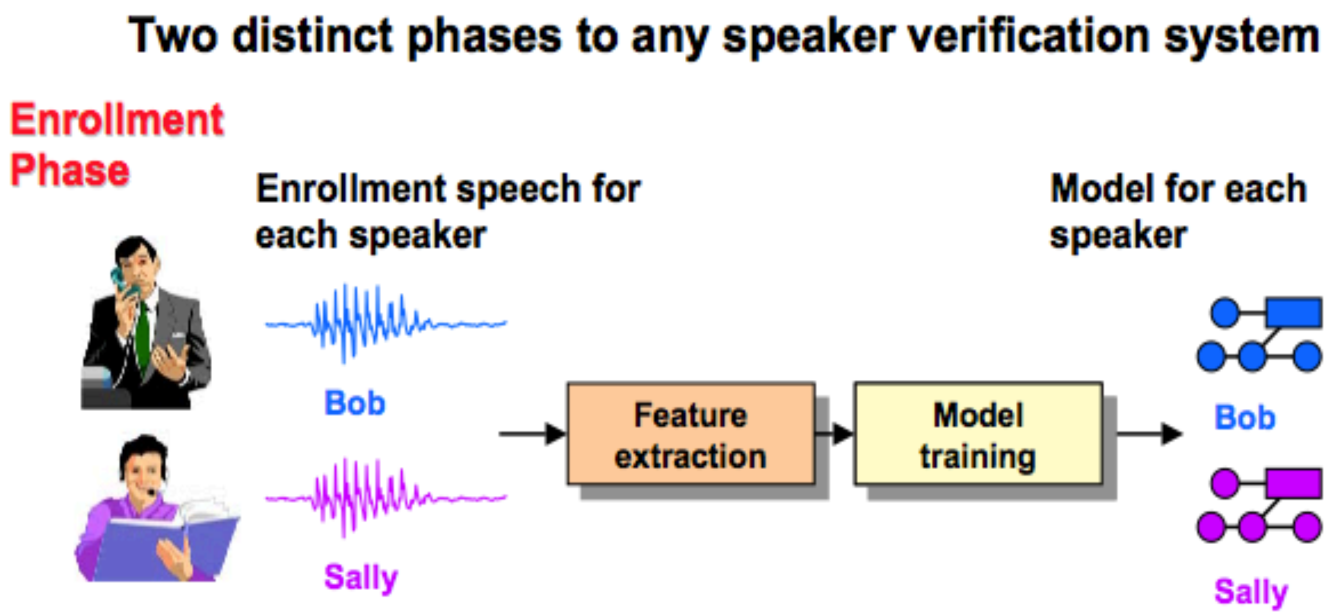
Law Enforcement
Forensics
Home parole

Speech Data Management
Voice mail browsing
Speech skimming

Personalization
Intelligent answering machine
Voice-web / device customization

Slide after [1]

Speaker Recognition



Slide after [1]

Speaker Recognition: Features (I)

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

High-level cues (learned traits)

Difficult to automatically extract

Hierarchy of Perceptual Cues

Semantics, diction, idiolect, pronunciation, 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**

Low-level cues (physical traits)

Easy to automatically extract

- **There are no exclusive speaker identity cues**
- **Low-level acoustic cues most common for automatic systems**

Slide after [1]

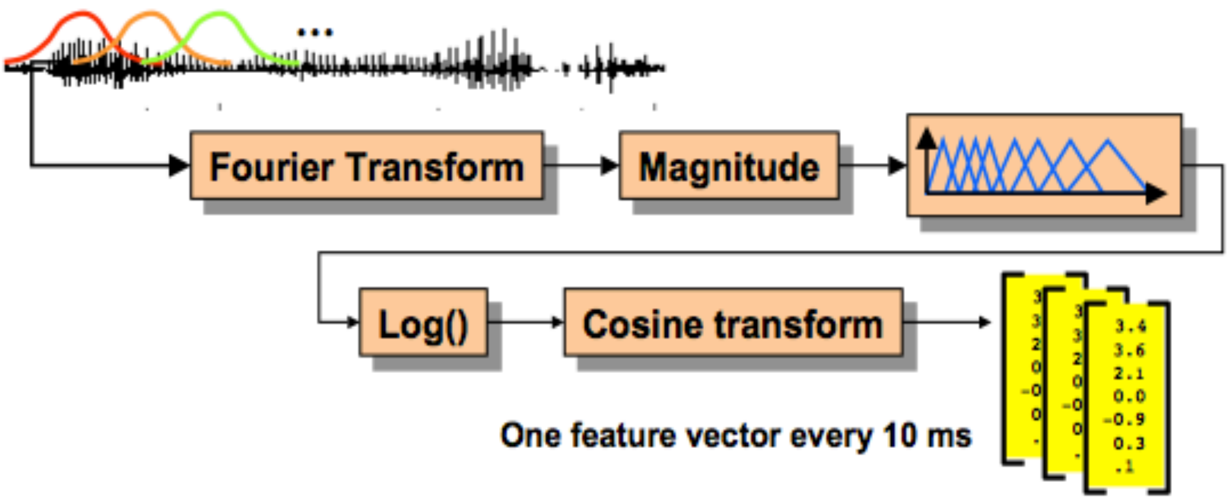
Speaker Recognition: Features (II)

- Desirable attributes of features for automatic methods:
  - **Practical**
    - Occure naturally and frequently in speech
    - Easy to measure
  - **Robust**
    - Not change over time or affected by speakers' health
    - Not (very) affected by noise and channel
  - **Secure**
    - Not be subject to mimicry
- In practice,
  - No feature has all these attributes
  - Features derived from spectrum speech are the most successful

Speaker Recognition: Features (III)

MFCC (Mel-frequency cepstral coefficients)

- Primary feature used in speaker recognition systems are cepstral feature vectors
- Some form of blind deconvolution is used to remove stationary channel effects
- Time differential cepstra (delta cepstra) are usually appended to cepstral features
- Typically 24-40 dimensional feature vectors are used



Slide after [1]

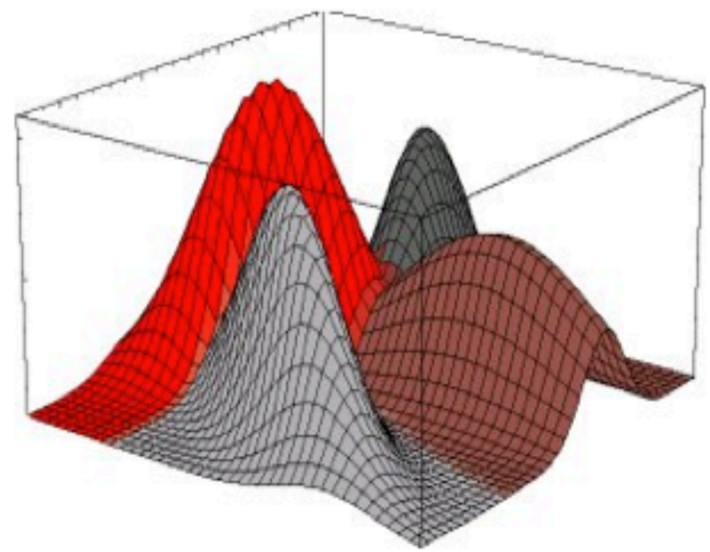
Speaker Recognition: Models

- **Speaker models** are used to represent the specific-speaker information in the feature vectors
- Several **different** modelling techniques have been applied:
  - Template matching (DTW for text-dependent)
  - Nearest neighbour
  - Neural networks
  - Hidden Markov Models
    - Single state HMM → **GMM**
  - Support vector machines
- Models provide some sort of score, reliability measure or **likelihood** for the target speakers

SR models: GMM (I)

- **A GMM is a weighted sum of Gaussian distributions**

$$p(\vec{x} \mid \lambda_s) = \sum_{i=1}^{M} p_i b_i(\vec{x})$$



$$\lambda_s = (p_i, \vec{\mu}_i, \Sigma_i)$$

p_i = mixture weight (Gaussian prior proability)

$\vec{\mu}_i$ = mixture mean vector

Σ_i = mixture covariance matrix

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp(-\tfrac{1}{2}(\vec{x} - \vec{\mu}_i)' \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i))$$

Slide after [1]

SR models: GMM (II)

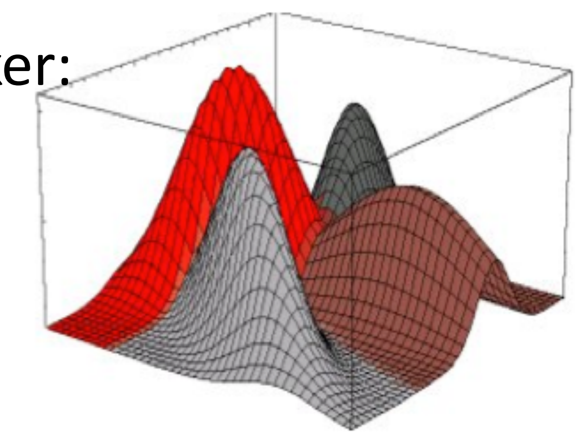
- In order to use GMMs we need:
  1. A method to estimate the model parameters using the training/enrolment data → **EM algorithm**
  2. Compute the (log-)**likelihood** of a sequence of features given a GMM

$$\log p(\vec{x}_1,\ldots,\vec{x}_N \mid \lambda) = \sum_{n=1}^{N} \log p(\vec{x}_n \mid \lambda)$$

$$= \sum_{n=1}^{N} \log\left(\sum_{i=1}^{\mathrm{M}} p_i b_i(\vec{x}_n) \right)$$

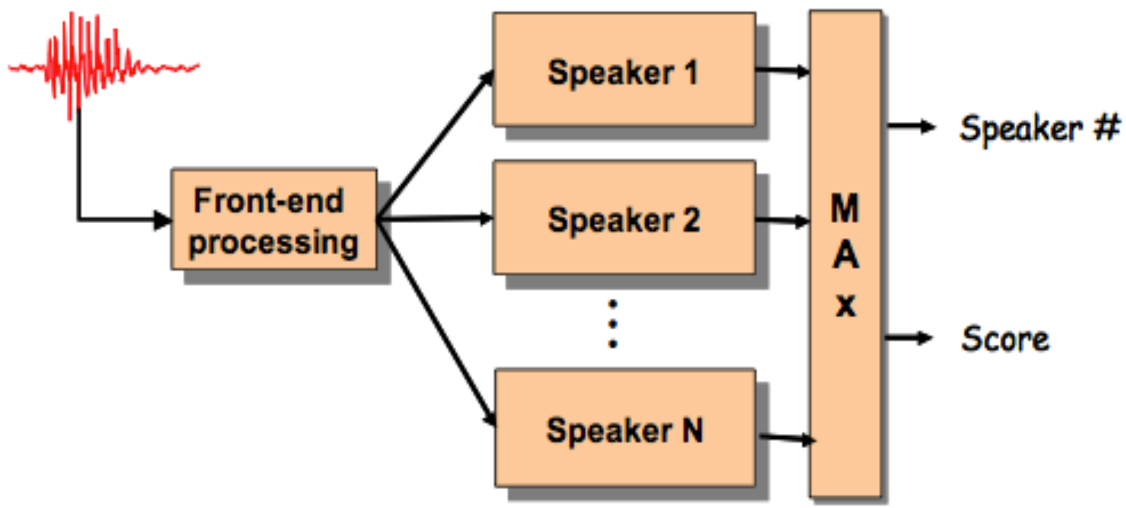
SR models: GMM-ML

- Conventional **GMM-ML** approach:
  - Use cepstral features as front-end
  - In **train** phase:
    - Train a GMM model per target speaker:
      - Apply EM algorithm for ML estimation
  - In **test** phase:
    - Compute log-likelihoods for scoring:
      - Speaker ID → MAX(LL)
      - Speaker Verification → log-likelihood compared to a threshold or impostor model

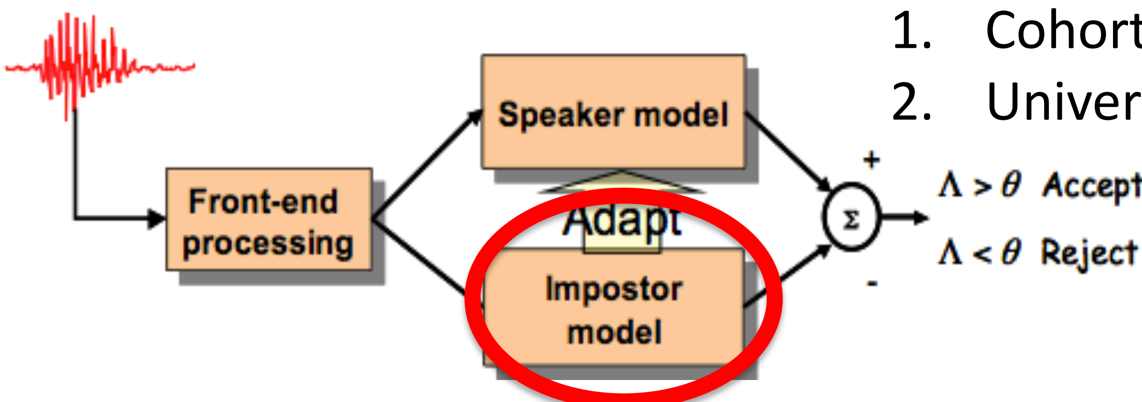


SR models: Impostor model

Identification



Verification



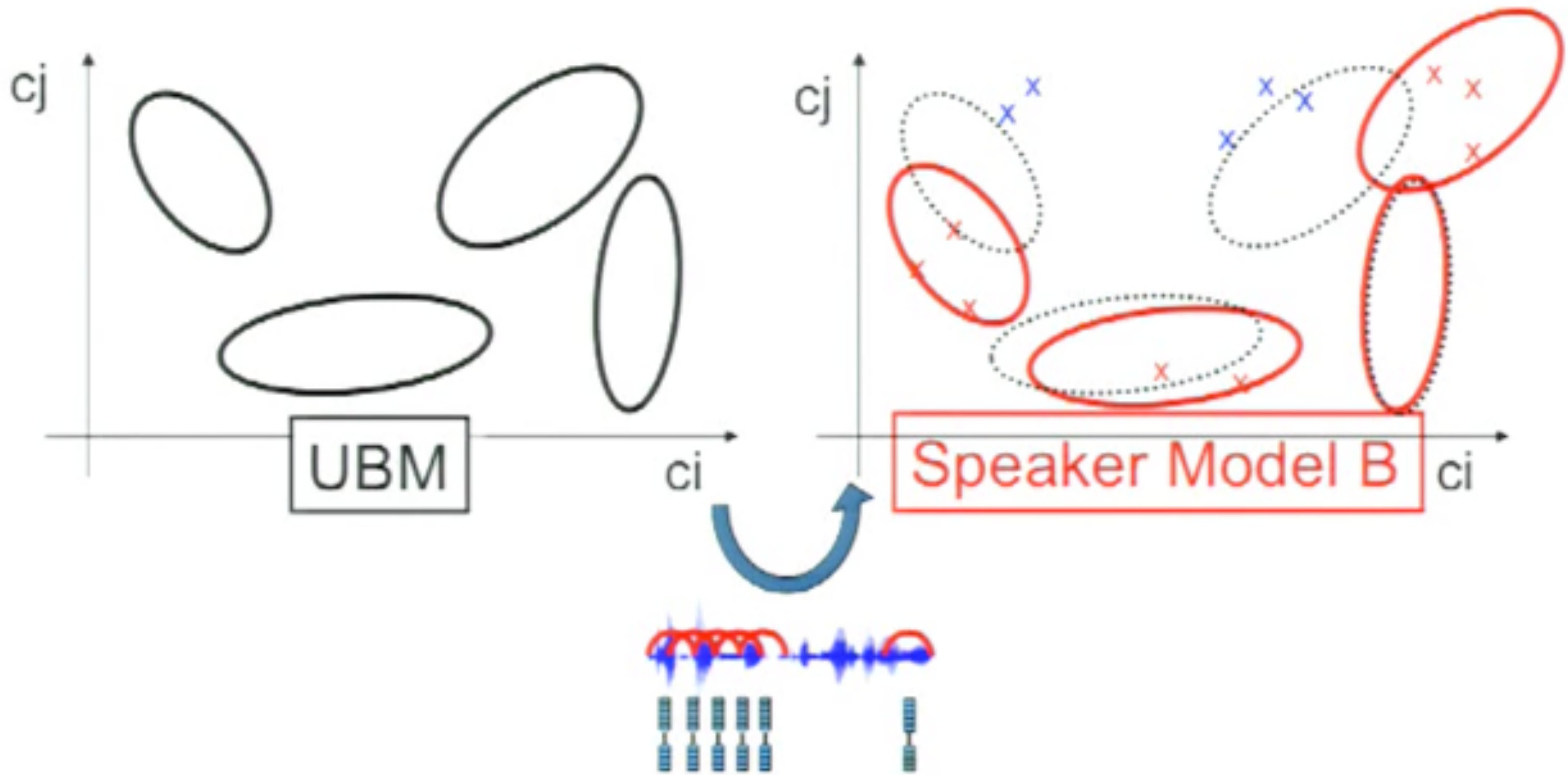
- Impostor model approaches:
  1. Cohort of impostors
  2. Universal model

Slide after [1]

SR models: GMM-UBM (I)

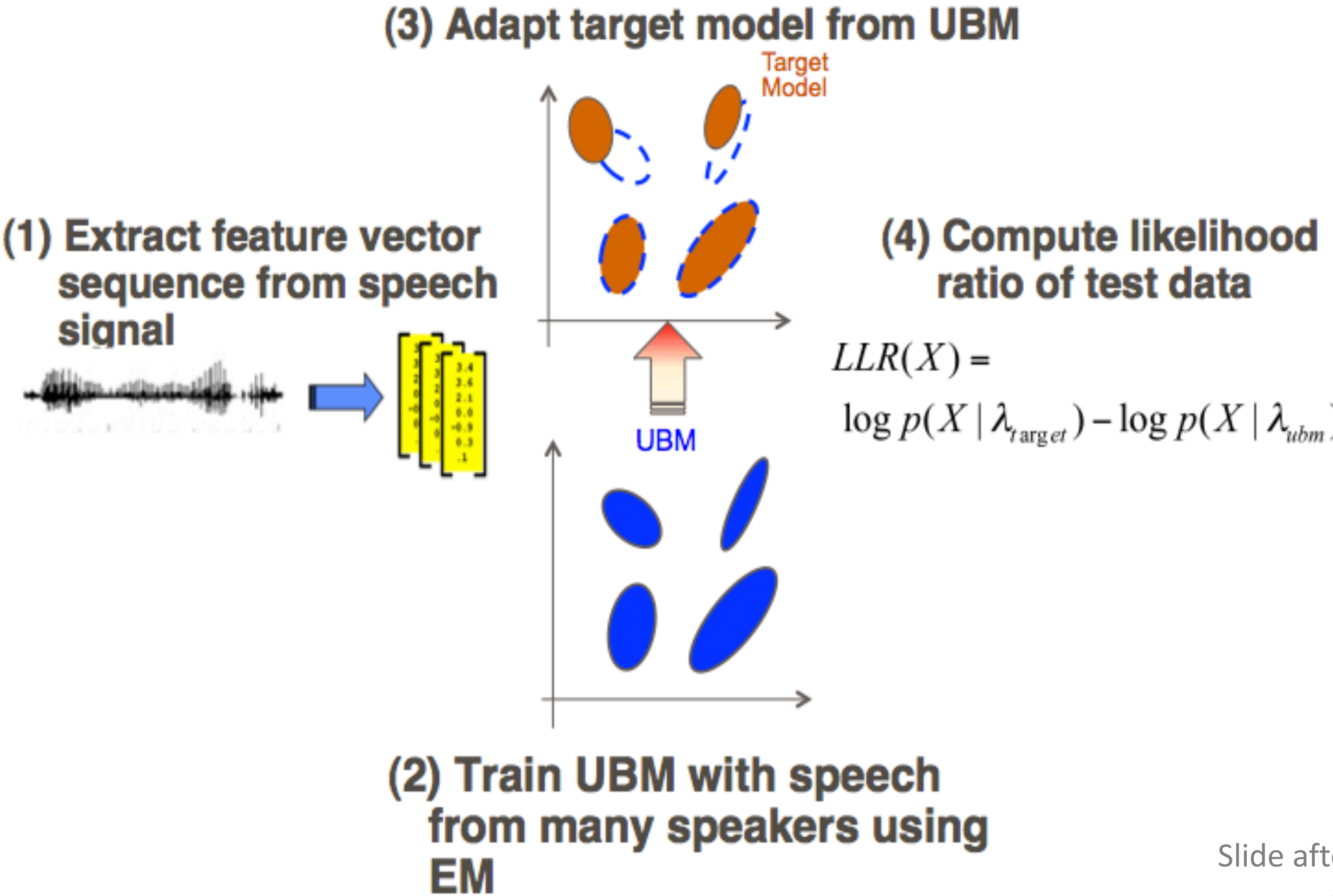
- **GMM-UBM** approach [Reynolds2000]:
  - Use cepstral features as feature extraction
  - In **train** phase:
    - Estimate the parameters of an UBM (Universal Background Model) with data from different speakers, channels, noise conditions, etc…
    - Adapt the UBM to each one of the target speakers:
      - Use MAP adaptation (usually only-means)
  - In **test** phase is like in previous GMM-ML approach.
  - **Advantages**
    - Needs less data,
    - permits updating only seen events,
    - keeps correspondence between means, allows fast scoring (top-M)

SR models: GMM-UBM (II)



Slide after [3]

SR models: GMM-UBM (III)



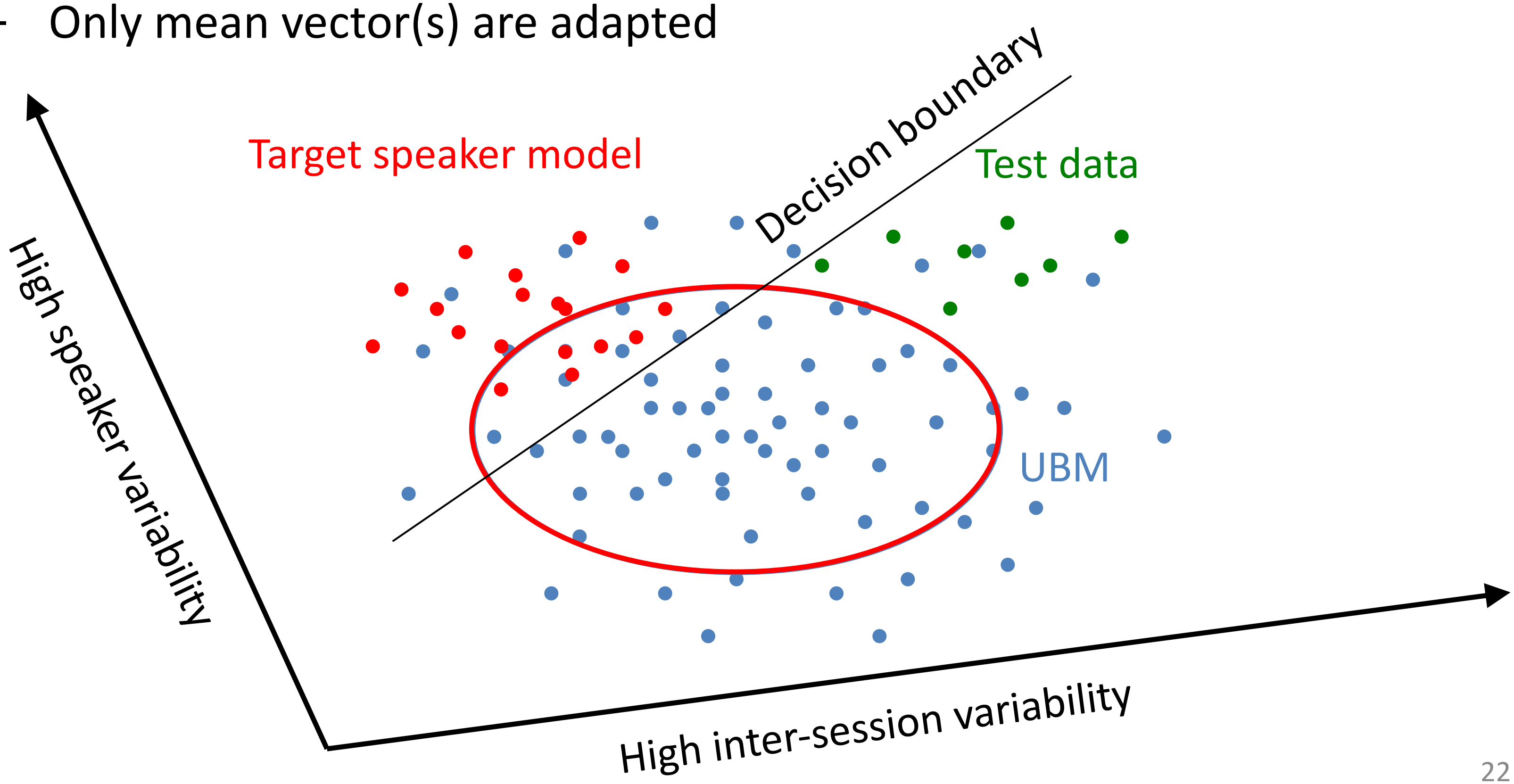
Slide after [1]

Inter-session variability (I)

- Variability refers to changes in channel effects (and other) between training and successive detection attempts
- Session variability 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.
  - The differences in speaker voice
    - Aging, mood, spoken language, etc.

Inter-session variability (II)

- Relevance MAP adaptation example (GMM-UBM):
  - 2D features
  - Single Gaussian model
  - Only mean vector(s) are adapted



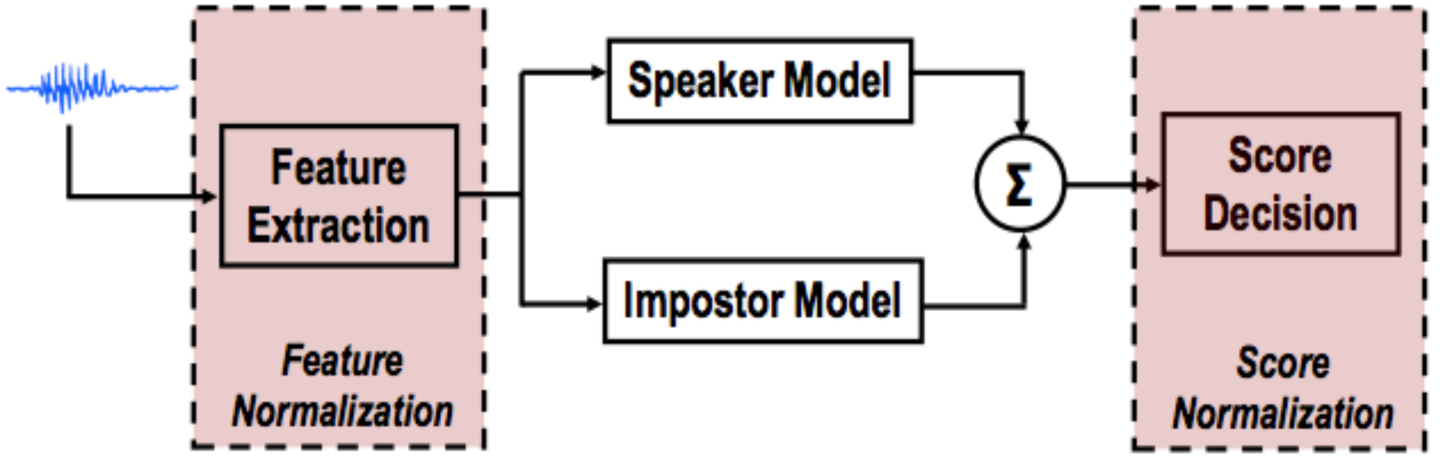
Inter-session variability (III)

- The largest challenge to practical use of speaker recognition systems is channel/session variability
- Most of the research during the last decade focused on developing more robust systems to session variability:
  - Feature level
    - Normalization, robust speech enhancement, alternative features (high-level)
  - Model level
    - More robust models (GMM-SVM), compensation at high dimensional space (NAP), factor analysis and explicit channel modeling
  - Score level
    - Score normalization (T-norm, Z-norm, etc.)
  - Back-end level
    - Calibration, fusion, etc.

Advanced Topics

Feature extraction (I)

- Channel can be (partially) compensated at the feature level



- Typical ways of increasing feature robustness are:
  - Use of **VAD** (Voice Activity Detector)
  - Apply **speech enhancement** methods:
    - RASTA processing, Wiener filtering, etc.
  - Feature **normalizations**:
    - CM(V)N, Feature warping, etc.

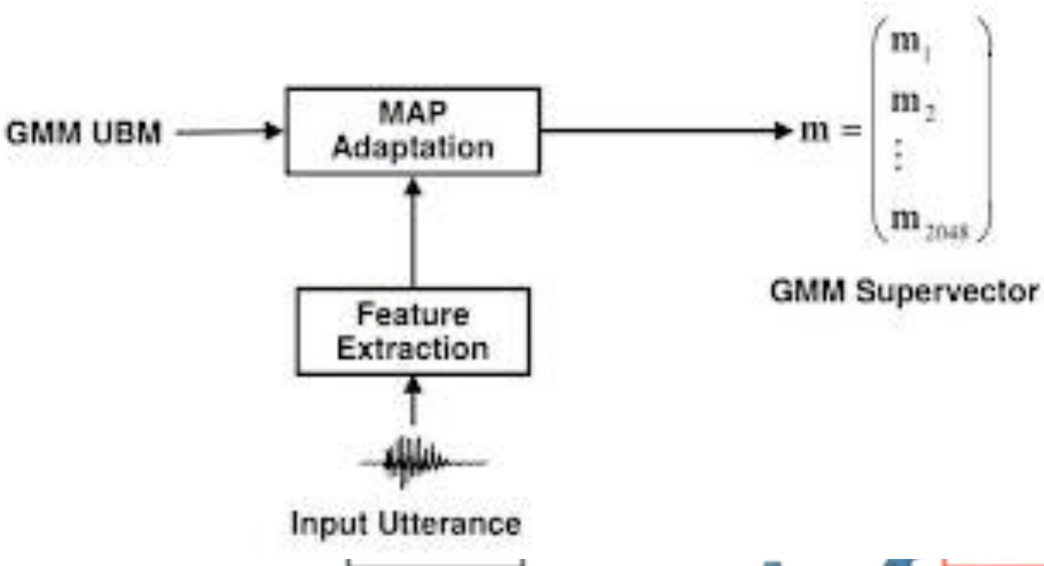
Advanced Topics
Feature extraction (II)

High-level Features:

- Extract and apply all levels of information from the speech signal conveying speaker identity
  - Acoustic: Use spectral features conveying vocal tract information
  - Prosodic: Use features derived from prosody (pitch, energy tracks) to characterize speaker-specific prosodic patterns
  - Phonetic: Use phone sequences to characterize speaker- specific pronunciations and speaking patterns
  - Idiolect: Use word sequences to characterize speaker- specific use of word patterns
  - Linguistic: Use linguistic patterns to characterize speaker- specific conversation style
- Combine them (ensemble of different systems), usually at the score level
  - Feature level combination is also possible
    - Feature selection; Feature dimensionality reduction (PCA)

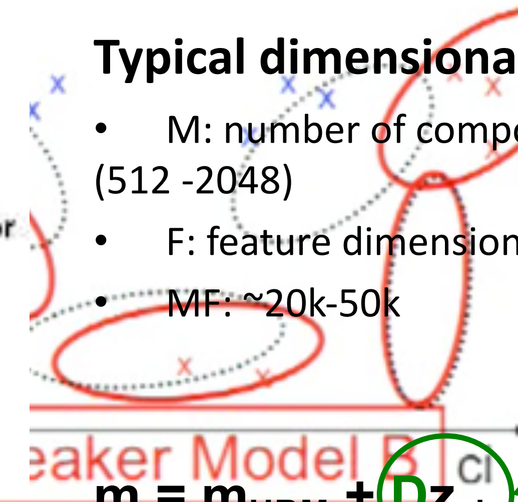
Improved modelling approaches

GMM-UBM: The supervector concept



Typical dimensionality:

- M: number of components (512 -2048)
- F: feature dimensions (20-60)
- MF: ~20k-50k



MAP

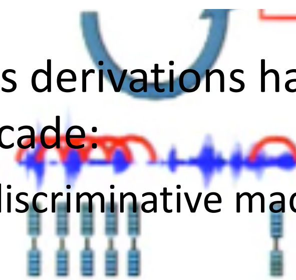
$$\mathbf{m} = \mathbf{m}_{UBM} + \mathbf{D}\mathbf{z}_{sh}$$

D = Full rank diagonal matrix (relevance MAP)

$\mathbf{z}_{sh}$ = Full rank vector

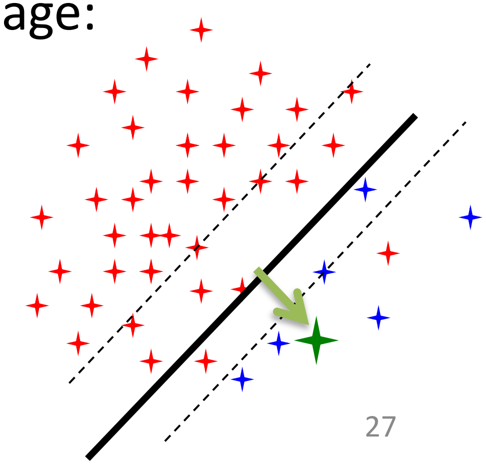
- The supervector concept and its derivations has had a **huge impact** in in the last decade:

1. As a kind of feature extraction for discriminative machine learning methods → GMM-SVM
2. As a tool for Factor Analysis derivation and session variability explicit modelling → JFA & i-vectors



Improved modelling approaches
GMM – SVM (I)

- The Gaussian/GMM super vector (GSV) in one of the recent most successful approaches for SR:
  - In SR comparable (even better) to standard GMM-UBM system with t-norm.
- GSV technique combines both GMM with Support Vector Machines (SVM):
  - GMM-UBM is efficient well-known technique in SR and LR.
  - SVM have proven to be a novel effective method for SR and LR (introduce discriminative training).
- **Main idea** Use a vector of the stacked means of GMM-UBM adapted models (super vectors) to characterize the speaker/language:
  - SVMs perform a nonlinear mapping from an high-dimensional input space.
  - More efficient/faster and improved modelling (discriminative).



Improved modelling approaches
GMM – SVM (II)

How does it work in practice?

- Super vector extraction...
  - Train 1 GMM (MAP adapted) for each train and test segment.
  - Use always the same UBM for adaptation (to keep sorting).
  - Stacked means need to be normalized (it does not work well without normalization).
- SVM model training...
  - Train "1 vs ALL" classifiers for each target class:
    - Target class super-vectors are positive samples for the SVM training.
    - A large set of background super-vectors are the negative samples for SVM training.
  - Careful needed due to unbalanced data sets (in SR it is usual to have only 1 positive supervector).
- SVM classifying...
  - Each test supervector is classified/scored with each target classifier to obtain speaker/language scores.

Improved modelling approaches
NAP for GMM-SVM (I)

Introduction to NAP

The SVM nuisance attribute projection (NAP) method works by removing subspaces that cause variability in the kernel, constructing a new kernel:

$$K(m^a, m^b) = b(m^a)^T b(m^b)$$

$$\Downarrow$$

$$K(m^a, m^b) = [Pb(m^a)]^T [Pb(m^b)] = b(m^a)^T Pb(m^b) = b(m^a)^T (I - vv^T) b(m^b)$$

$b(m^k)$ is the normalized super vector: $b(m_n^k) = \sqrt{\lambda_n}\Sigma_n^{-1/2} m_n^k$

P is a projection matrix with **v** the variability directions

Objective Find **P** according to variability compensation criteria desired.

Improved modelling approaches

NAP for GMM-SVM (II)

HOWTO in simple words/steps

1. Form the matrix **M** → differences of the SV with respect to its class SV mean.

2. Find the variability directions **v** → The normalized eigenvectors of MM^t.

3. Find the projection matrix $\mathbf{P} = \mathbf{I} - \mathbf{v}\mathbf{v}^t$ → Select the most important variability directions (the ones corresponding to larger eigenvalues).

4. Apply **P** to the training SV set → train new 1vsALL SVM classifiers.

5. Apply **P** to the test SV before SVM classification → Obtain target scores.

This compensation method may be applied to any general high-dimensionality SVM based classification task.

Improved modelling approaches

Factor Analysis (I)

- Factor Analysis (FA) is a method for investigating if a number of variables are linearly related to a small number of unobservable factors. Example:

Student no.	Grade in: Finance, Y_1	Marketing, Y_2	Policy, Y_3
1	3	6	5
2	7	3	3
3	10	9	8
4	3	9	7
5	10	6	5

$$Y_1 = \beta_{10} + \beta_{11} F_1 + \beta_{12} F_2 + e_1$$
$$Y_2 = \beta_{20} + \beta_{21} F_1 + \beta_{22} F_2 + e_2$$
$$Y_3 = \beta_{30} + \beta_{31} F_1 + \beta_{32} F_2 + e_3$$

- FA propose solutions for estimating the loading factors matrix and also for estimating the (low-dimensionality) factors.

Improved modelling approaches
Factor Analysis (II)

GMM-UBM (MAP) → $\mathbf{m} = \mathbf{m}_{UBM} + \mathbf{D}\mathbf{z}_{sh}$

- **D** diagonal full-rank
- $\mathbf{z}_{sh}$: speaker (and more) component

Eigenvoices → $\mathbf{m}_s = \mathbf{m}_{UBM} + \mathbf{V}\mathbf{y}$

(inspired in FA approach for image processing, *eigenfaces*)

- **V** speaker variability subspace (low-rank)
- **y** speaker (loading) factors for every speaker

Eigenchannels → $\mathbf{m} = \mathbf{m}_s + \mathbf{U}\mathbf{x}$

(the speaker and and session components can be linearly decomposed)

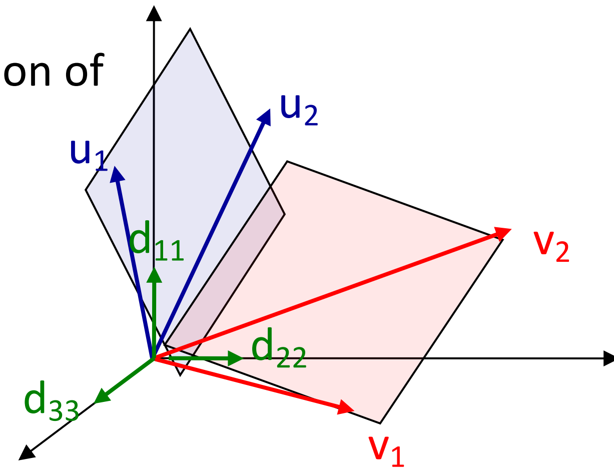
- **U** session/variability sub-space (low-rank)
- **x** channel (loading) factors for every utterance

Slide after [3]

Improved modelling approaches
Factor Analysis (III)

Current most successful FA based methods for SR:

1. Joint Factor Analysis (**JFA**) [Kenny2005]
   - Represent speaker mean supervectors as a combination of low-dimensional speaker and channel factors
   - "Put together" eigenvoices, eigenchannels and MAP



$$\mathbf{m} = \mathbf{s} + \mathbf{c} = \mathbf{m}_{UBM} + \mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{z} + \mathbf{U}\mathbf{x}$$

2. Total variability (or **i-vector**) [Dehak2009]
   - Represent speaker means depending on total variability

$$\mathbf{m} = \mathbf{m}_{UBM} + \mathbf{T}\mathbf{w}$$

   - **w** are called i-vectors (~400-600 dimensions)
     - They contain all speaker and channel variability (can be compensated later)
     - It is used as a low-dimensional representation (on top of them other models can be trained)
     - Cosine scoring after compensation methods like LDA or WCCN (for simple SR)
   - **i-vector + PLDA** scoring is the **current?** de facto standard

Advanced Topics

Score normalization (I)

- Scores normalization contributes to compensate variability of inter-speaker and inter-session in decision making.
- Normalize the log likelihood ratio score with mean and standard deviations

$$LLR(\chi_{test}, S)_{norm} = \frac{LLR(\chi_{test}, S) - \mu}{\sigma}$$

- Most common approaches:
  - Zero Normalization (Z-norm)
  - Test Normalization (T-norm)

Advanced Topics

Score normalization (II)

- Z-norm
  - Compensate inter-speaker variability
  - Estimate mean and variance from a set of log likelihood ratio score target model against impostor utterances
  - Normalize based on speaker model
- T-norm
  - Compensate inter-session variability
  - Estimate mean and variance from a set of log likelihood ratio score impostor model against test utterances
  - Normalize based on test utterances

Advanced Topics

Fusion and Calibration

- In SR the goal is to produce verification scores that favor (or disfavor) two speech samples belong to the same (different) speaker (FRONT-END)
- Calibration is a fundamental problem in Speaker Verification (BACK-END)
  - Its objective is to transform the scores (or set thresholds) so that task-specific thresholds can be applied to take decisions that minimize the cost function
    - Usually, the objective is to produce well-calibrated log-likelihood ratios
    - If well-calibrated log-likelihood ratios, decision thresholds are theoretically defined
  - Most successful systems, in addition to calibration, they fuse several sub-systems.
- There are may approaches for Fusion&Calibration:
  - The Focal/Bosaris toolkits based on Linear Logistic Regression:

$$\hat{s}_t = \beta + \sum_{i=1}^{N} \alpha_i \cdot s_t(i)$$

$$\hat{s}_t \approx \log \frac{P(\hat{s}_t | H_{\mathrm{target}})}{P(\hat{s}_t | H_{\mathrm{non-target}})}$$

https://sites.google.com/site/bosaristoolkit/

Break

Speaker Recognition Recap

- Main challenges:
  - Enrolment usually a single (short) utterance
  - Session variability → Compensation
- Classical approach to SR [<2000]
  - 1 GMM trained per speaker on top of MFCC features
- Major modeling improvements:
  - GMM-MAP [~2000]: A Universal Background Model used as seed to adapt to speaker characteristics
  - GSV-SVM [~2004]: Adapted Gaussian means are concatenated to obtain a super(large) vector + SVM → Difficult channel compensation
  - FA [>2005]: Super-vector variability lays in a low-dimensional space:
    - JFA [~2006]: Model specifically speaker and channel subspaces
    - i-vector [2009-]: Model a single low-dimension space → do channel compensation later in this low-dimension space

Advanced Topics

It was happening in 2017...

- i-vectors **extremely** successful:
  - Recent efforts (2015) of making of i-vectors more than a de-facto standard http://www.voicebiometry.org
  - Recent SR evaluations do not rely (directly) on speech samples
    - The 2013-2014 SR i-vector Machine Learning Challenge: https://ivectorchallenge.nist.gov/evaluations/1

- Deep learning has **also** arrived to SR:
  - As a replacement of GMM-UBM in i-vectors
  - As features based on DNN

- Emergence of related tasks:
  - ASVspoof: Automatic Speaker Verification Spoofing and Countermeasures Challenge: http://www.spoofingchallenge.org
  - Privacy issues!?

Advanced Topics

… in 2018, welcome x-Vectors (bye bye i-vectors)!!

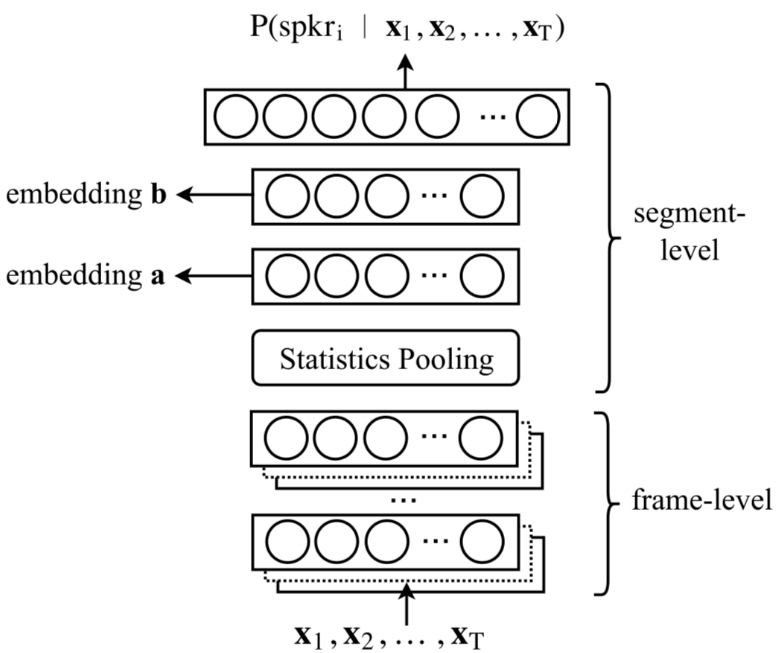


Figure 1: *Diagram of the DNN. Segment-level embeddings (e.g.,* **a** *or* **b***) can be extracted from any layer of the network after the statistics pooling layer.*

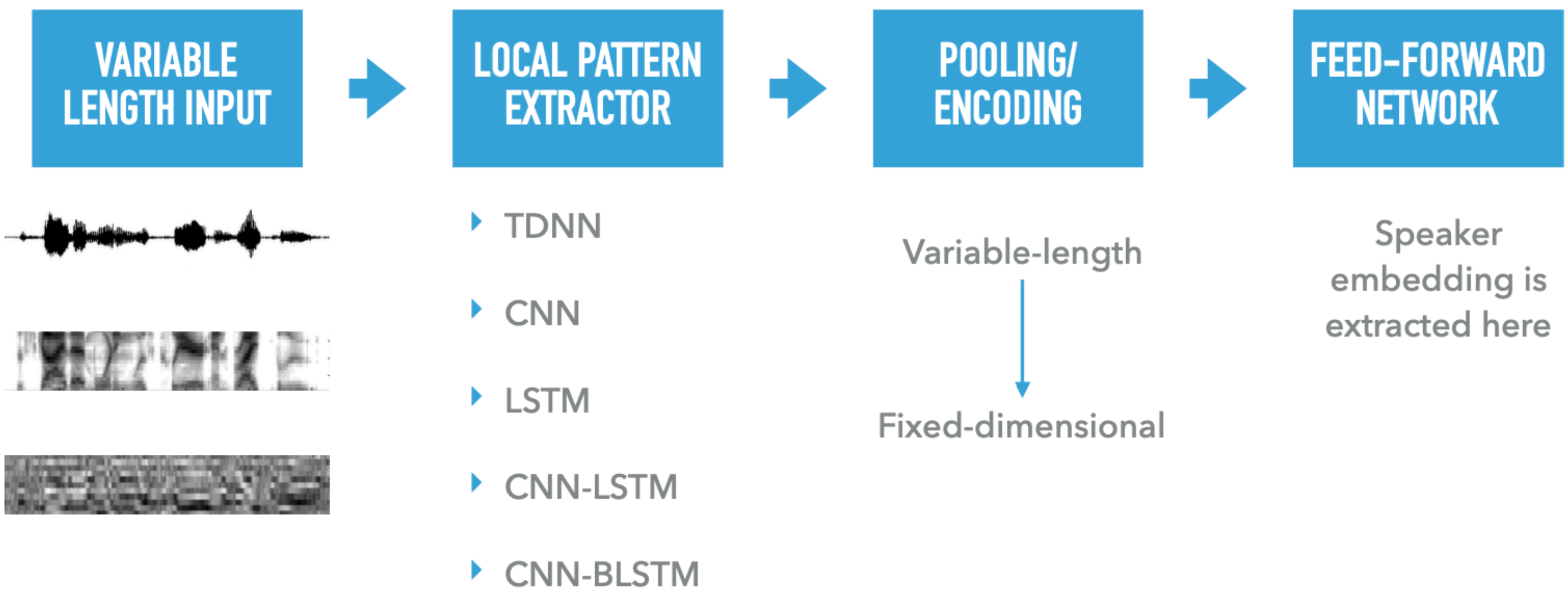
			SITW Core			SRE16 Cantonese		
			EER(%)	DCF10^{-2}	DCF10^{-3}	EER(%)	DCF10^{-2}	DCF10^{-3}
4.1	Original systems	i-vector (acoustic)	9.29	0.621	0.785	9.23	0.568	0.741
		i-vector (BNF)	**9.10**	**0.558**	**0.719**	9.68	0.574	0.765
		x-vector	9.40	0.632	0.790	**8.00**	**0.491**	**0.697**
4.2	PLDA aug.	i-vector (acoustic)	8.64	0.588	0.755	8.92	0.544	0.717
		i-vector (BNF)	8.00	**0.514**	**0.689**	8.82	0.532	0.726
		x-vector	**7.56**	0.586	0.746	**7.45**	**0.463**	**0.669**
4.3	Extractor aug.	i-vector (acoustic)	8.89	0.626	0.790	9.20	0.575	0.748
		i-vector (BNF)	7.27	**0.533**	0.730	8.89	0.569	0.777
		x-vector	**7.19**	0.535	**0.719**	**6.29**	**0.428**	**0.626**
4.4	PLDA and extractor aug.	i-vector (acoustic)	8.04	0.578	0.752	8.95	0.555	0.720
		i-vector (BNF)	6.49	0.492	0.690	8.29	0.534	0.749
		x-vector	**6.00**	**0.488**	**0.677**	**5.86**	**0.410**	**0.593**
4.5	Incl. VoxCeleb	i-vector (acoustic)	7.45	0.552	0.723	9.23	0.557	0.742
		i-vector (BNF)	6.09	0.472	0.660	8.12	0.523	0.751
		x-vector	**4.16**	**0.393**	**0.606**	**5.71**	**0.399**	**0.569**

Table 2. Results using data augmentation in various systems. "Extractor" refers to either the UBM/**T** or the embedding DNN. For each experiment, the best results are **boldface**.

Snyder, David, et al. "X-vectors: Robust DNN embeddings for speaker recognition." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

Advanced Topics

After that, DNN architectures for speaker embedding



Outline

- Automatic Speaker recognition
  - Intro
  - Classical approaches:
    - Features
    - Models
  - The problem of inter-session variability
  - Advanced topics

- Evaluation and performance of speaker verification
  - Evaluation measures
  - SRE evaluation challenges
    - NIST SRE
    - NIST HASR

Evaluation measures
Trial definition

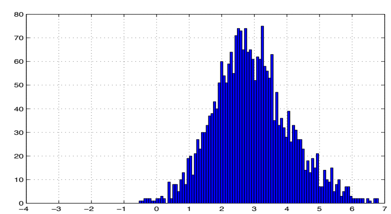
- Speaker verification tasks usually consist of a set of verification trials.

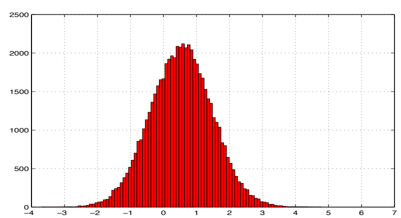
- **Test trials**: given a test segment, determine whether a given speaker is actually speaking
  - Target trials → The speaker is speaking in the test segment
  - Non-target/Impostor trials → The speaker is NOT speaking in the test segment

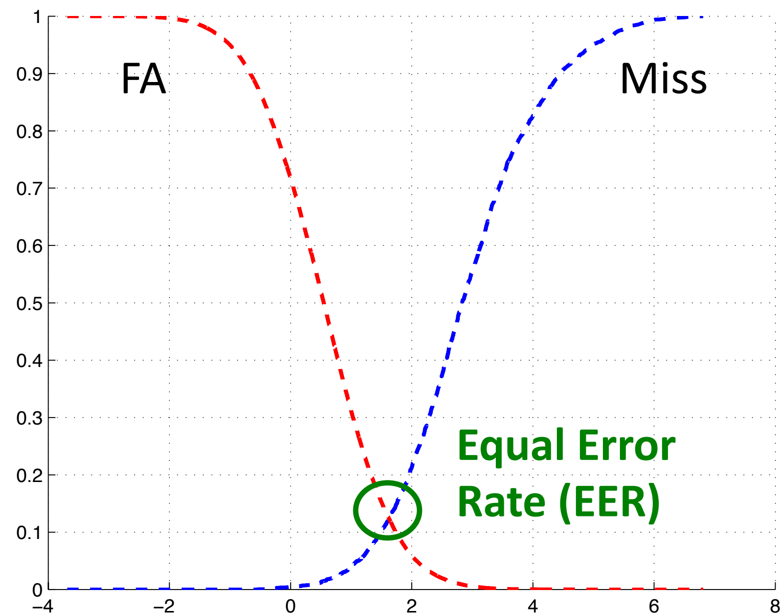
- Each trial (usually) requires two outputs:
  - Actual decision → True/false
  - Likelihood score → Confidence in decision

Evaluation measures
Decision errors

- Two types of actual decision errors:
  - Missed detections ($P_{miss|target}$): Percentage of target trials rejected incorrectly
  - False Alarms ($P_{fa|impostor}$): Percentage of impostor trials accepted incorrectly

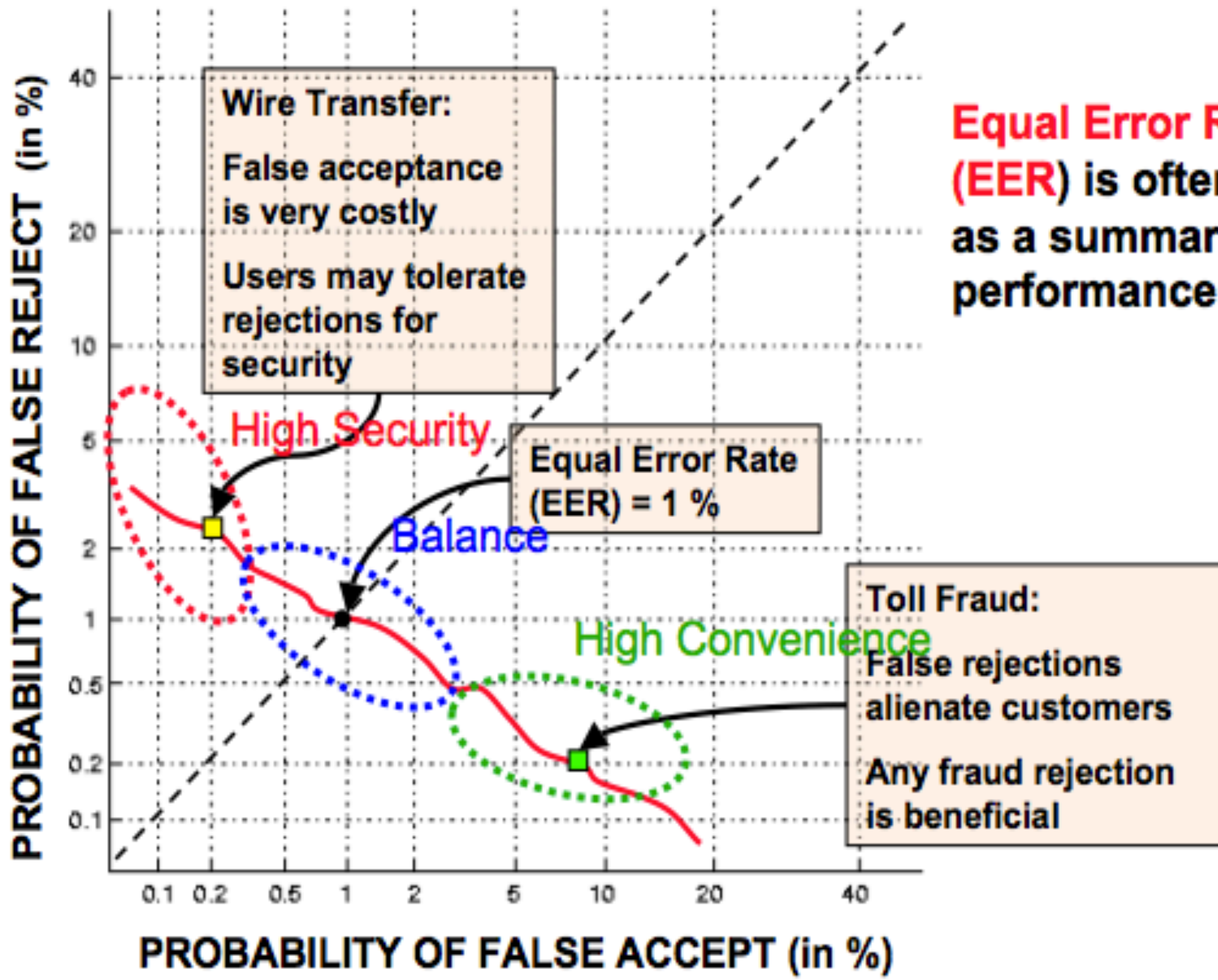






Evaluation measures
DET curve

- DET plots P_{miss} vs P_{FA} for every threshold (like ROC curves):
  - Axis follow normal distribution scale



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

Slide after [1]

Evaluation measures

Cost function

$$\boxed{\begin{aligned} C_{Det} = \; & C_{Miss} \times P_{Miss|Target} \times P_{Target} \\ & + C_{FalseAlarm} \times P_{FalseAlarm|NonTarget} \times (1-P_{Target}) \end{aligned}}$$

- C_{Det} depends on application Costs and Priors

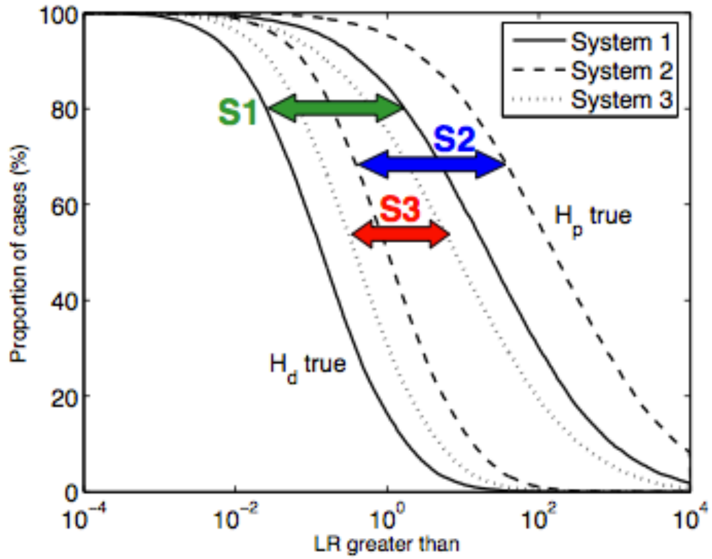
- If scores are well-calibrated likelihood ratios, the minimum expected cost Bayes decision threshold is:

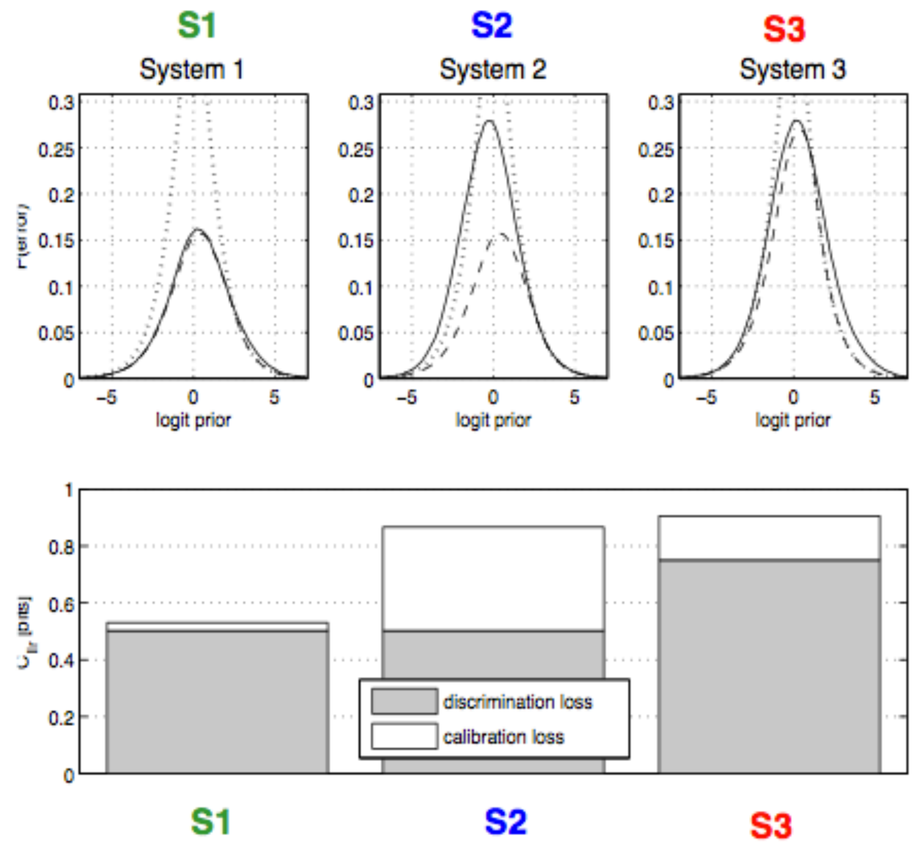
  $\boxed{LR \geq TH_{Bayes}}$ $\quad TH_{Bayes} = C_{Miss}/C_{FalseAlarm} \times (1-P_{target})/P_{target}$

  - The decisions are influenced by the cost of the errors we make

- CALIBRATION LOSS: C_{Det} – $minC_{Det}$
  - C_{Det} computed from decisions, $minC_{Det}$ from scores
  - Measure the goodness of the threshold selection

Evaluation Measures
Importance of calibration





- S1 and S2 systems have the same DET curve!!
- Not only discrimination is important, calibration is a must

Slide after [2]

NIST SRE

- Speech Group at the (US) National Institute of Standards and Technology
- (Bi-)Annual evaluations of speaker verification technology (since 1996)
  - Aim: Provide a common paradigm for comparing technologies
  - Provides: evaluation plan, common test sets, standard metrics, etc.

http://www.itl.nist.gov/iad/mig/tests/sre/

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

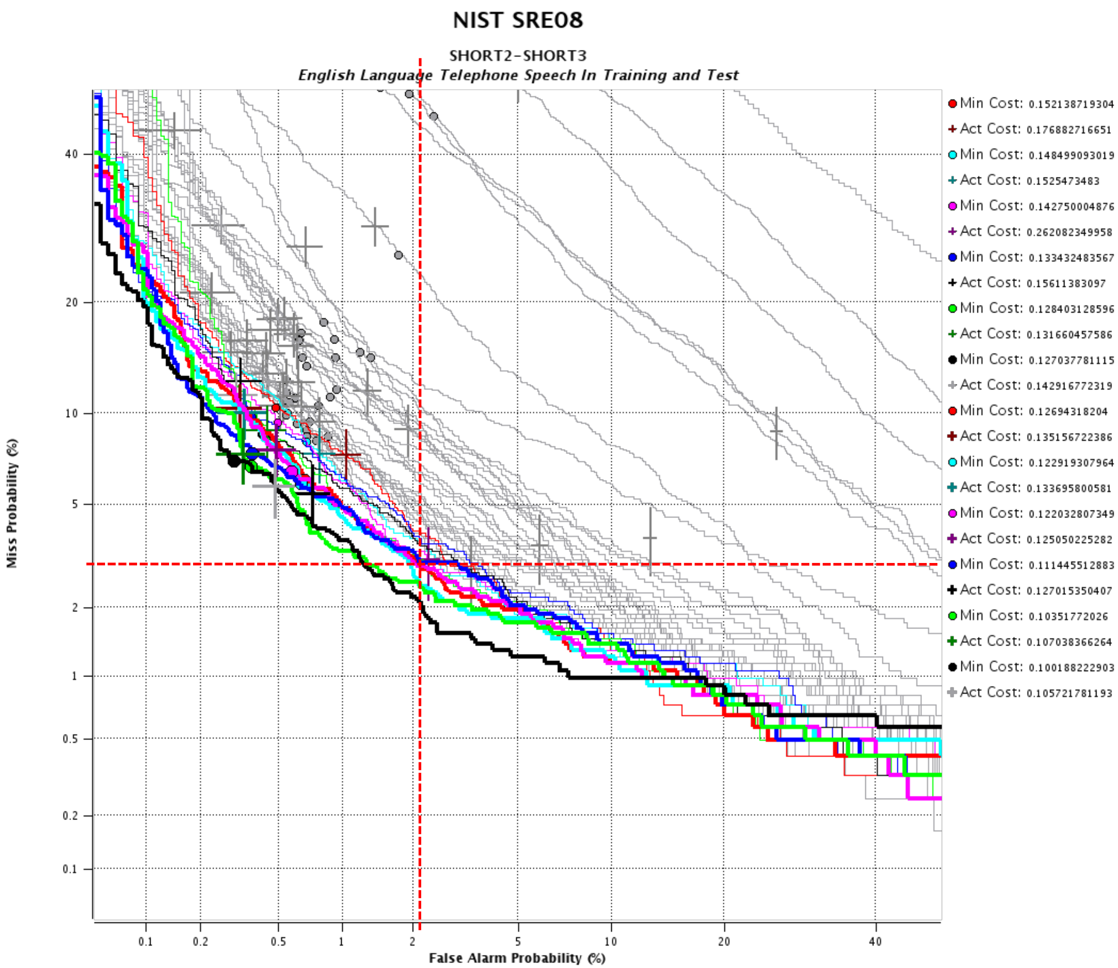
NIST SRE

- **Task** focused initially on conversational telephone speech with novelties every year:
  - Up to 2005: 2 side conversations with telephone mikes
  - In 2005: Recording of alternate microphones (same conversations)
  - In 2008: New interview speech (different style)
  - In 2010: Focus on new operation point (very-low FA) & vocal effort
  - In 2012: Included additive & environmental noise and submitted scores LLRs
  - *In 2013-2014: NIST SRE i-vector challenge*
  - In 2016: Fixed/common training data
  - In 2018: CTS + VoIP + audio from video
  - In 2019 (planned for December): Similar to 2018 + audio-visual condition (amateur videos)
- Mandatory/core vs optional conditions:
  - Core conditions usually involve 1-side conversation (~5 minutes) for enrolment and one 1-side conversation test utterances:
    - Can be telephone, microphone, interview, same/different language, etc
  - Alternate conditions usually involve shorter segments (10-secs), multiple enrolment utterances (~8) or summed channels test utterances
  - The amount of trials has been increased during the years in all conditions, eg.:
    - Mandatory/core condition in SRE2008 ~100 000
    - Mandatory/core condition in SRE2012 ~ 1 000 000 (extended 100 000 000)
- The amount of data and computational processing load involved in these evaluations is HUGE

NIST SRE Results

NIST SRE 2008 core condition

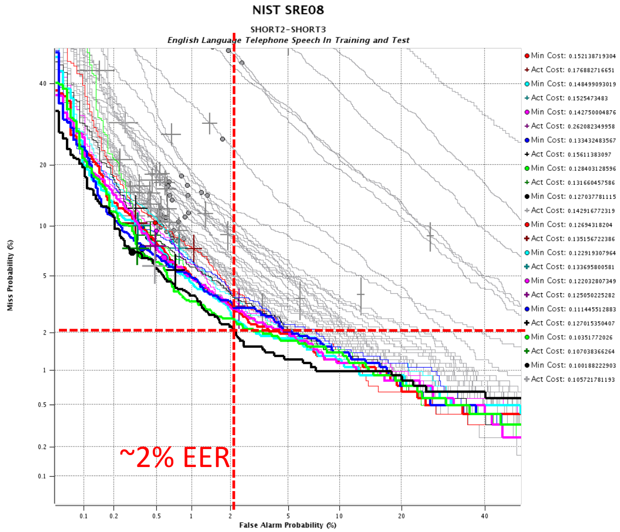
Tel-tel + only English sub-conditions



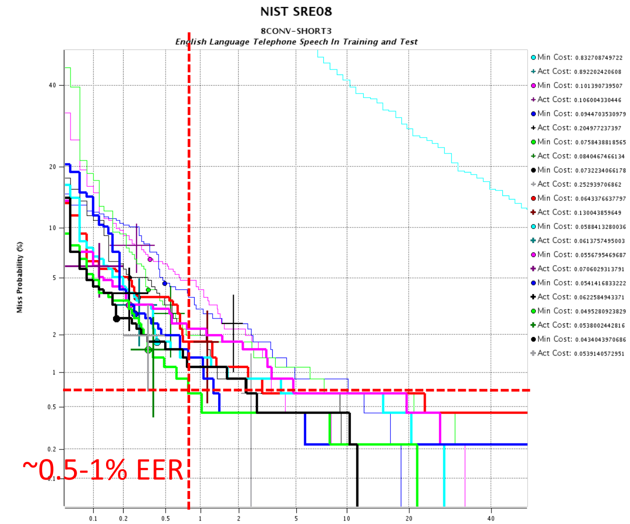
NIST SRE Results

NIST SRE 2008: Importance of speech length

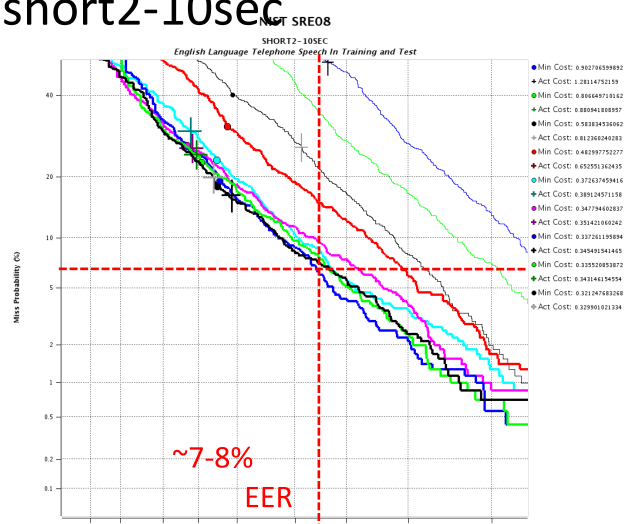
short2-short3



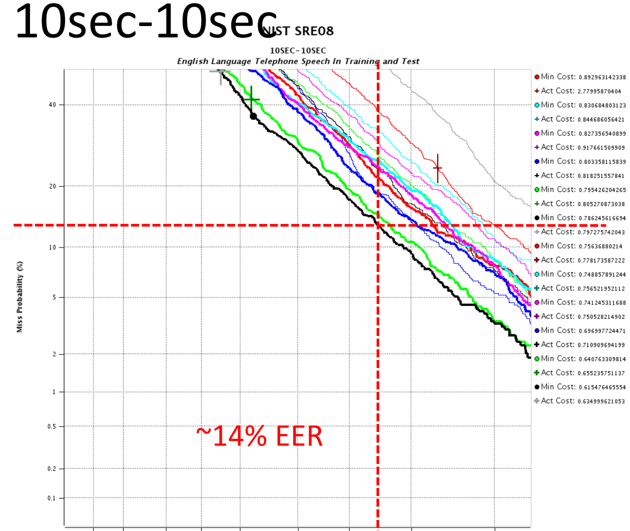
8conv-short3



short2-10sec



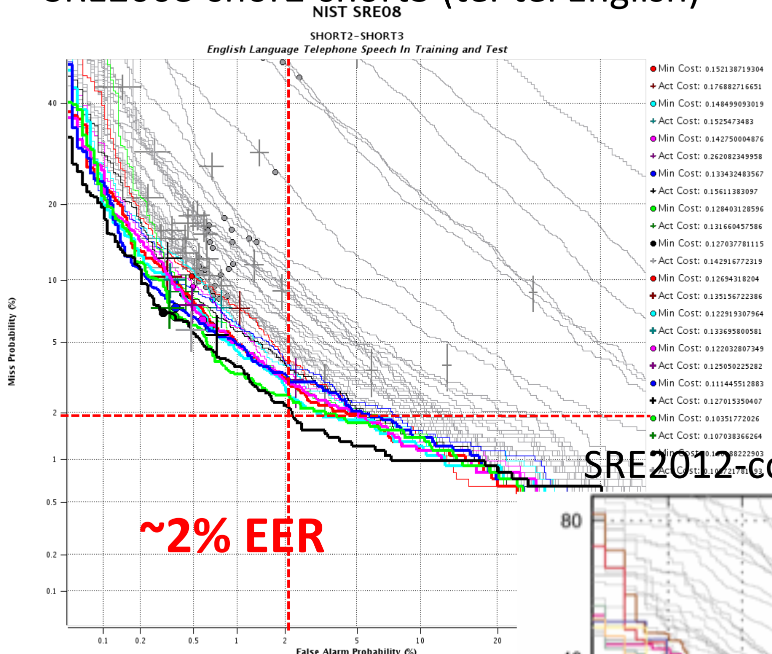
10sec-10sec



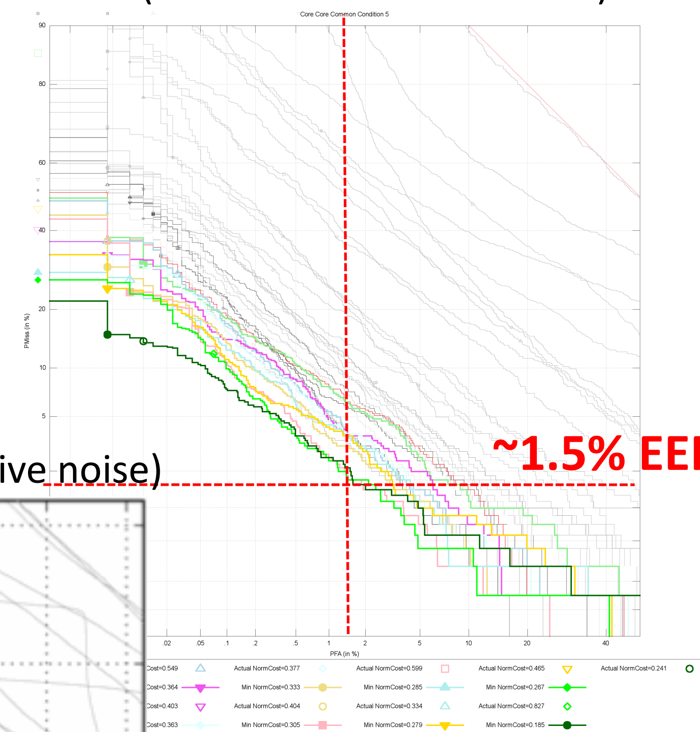
NIST SRE Results

Recent years evolution

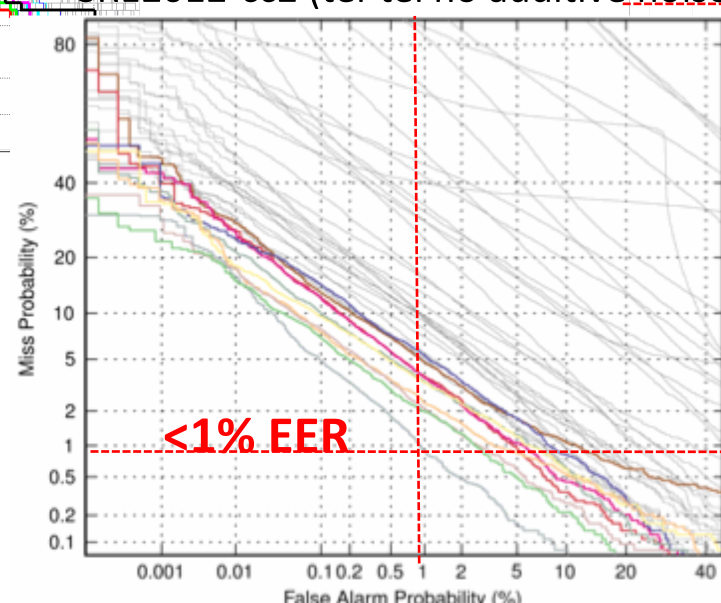
SRE2008-shor2-short3 (tel-tel English)



SRE2010-cc5 (tel-tel normal vocal effort)



SRE2012-cc2 (tel-tel no additive noise)



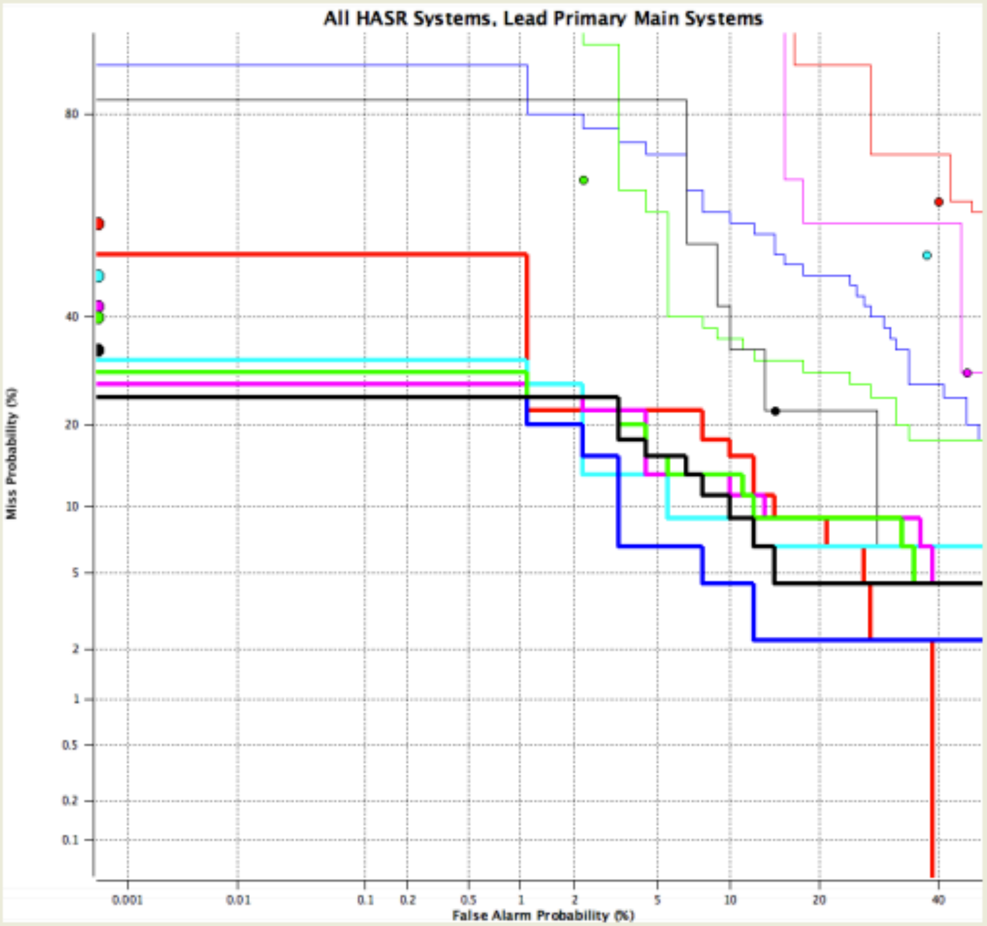
NIST HASR 2010

- A pilot test with difficult cross-channel trials of the NIST SRE 2010
  - HASR1 – 15 trials
  - HASR2 – 150 trials
- Trials to be processed separately and independently
  - Automated email used to submit each trial's output before next trial was accessible
  - Unlimited listening (in whatever order) permitted for training and test data
- Human listeners could be one person or a panel
- A decision and a likelihood score were required for each trial
- Decisions could be made from:
  - A combination of automatic processing and human expertise, or
  - Solely based on human listening
- Scoring
  - Count number of Misses and False

NIST HASR 2010: HASR1 Results

Site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Misses	FAs	Total
System 1	t	f	f	f	f	f	t	f	f	t	f	f	f	t	f	2	-	2
System 2	t	t	f	f	t	f	t	t	f	t	f	f	t	f	t	1	3	4
System 3	t	t	f	f	t	t	f	f	f	t	t	f	f	t	f	2	3	5
System 4	t	t	f	f	t	t	f	f	f	t	t	f	f	t	t	1	3	4
System 5	t	t	f	f	t	f	t	t	f	t	f	f	t	f	t	1	3	4
System 6	t	f	t	t	f	t	f	f	t	f	t	f	f	t	f	4	5	9
System 7	f	t	f	t	f	f	f	t	f	f	f	f	f	t	f	5	3	8
System 8	f	t	t	t	f	t	f	t	t	t	t	f	f	t	f	4	7	11
System 9	t	t	f	t	t	f	f	f	t	t	t	t	t	t	f	2	6	8
System 10	t	t	f	t	t	f	f	f	t	t	t	t	t	t	f	2	6	8
System 11	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	-	9	9
System 12	f	f	t	f	t	t	t	t	t	t	t	t	f	t	t	1	6	7
System 13	f	t	t	f	t	t	t	f	t	t	t	t	t	t	f	2	7	9
System 14	f	t	t	f	t	t	t	f	t	t	t	t	t	t	f	2	7	9
System 15	t	f	f	f	f	f	t	f	f	t	t	f	f	t	f	2	1	3
System 16	f	t	f	f	f	f	t	f	f	t	t	f	f	t	f	3	2	5
System 17	t	t	t	t	f	t	f	f	f	t	t	f	f	t	f	3	5	8
System 18	t	t	t	t	t	t	f	f	t	t	t	t	t	f	t	2	8	10
System 19	f	f	f	f	t	f	f	t	f	t	t	f	f	t	t	2	2	4
System 20	f	f	f	f	f	t	f	f	f	t	f	f	f	f	f	5	1	6
KEY	**T**	**F**	**F**	**F**	**T**	**F**	**T**	**F**	**F**	**T**	**F**	**F**	**F**	**T**	**T**	-	-	-
Number of Errors	*8*	*14*	*8*	*8*	*8*	*11*	*11*	*7*	*9*	*2*	*15*	*7*	*8*	*4*	*13*	*46*	*87*	***133***

NIST HASR 2010: HASR2 Results



- 135 HASR2 trials
- Six HASR systems (thin lines)
  one system = decision only
- Six Automatic systems (thick lines)

Summary

- Speech contains lots of identity information → It can be used for biometric authentication:
  - Identifying speakers is difficult (even “human assisted”)
- Classical systems for speaker authentication are based on:
  - MFCC features → Most common features in any speech application
  - GMM-UBM → GMM speaker models based on MAP UBM adaptation
- Session variability is the most challenging limitation:
  - Most successful current modelling approach → i-VECTORS
  - Length of enrolment and test utterances is also an important problem
- Steady and consistent improvements in the last :
  - Actually, very good results are obtained in the order of <1%EER
  - In a pilot experience for human-assisted SR, automatic methods performed better than assisted ones (**caution!** very difficult trials)
- Some relevant topics not discussed today:
  - Privacy issues?
  - Spoofing attacks/fooling the system?

References

- These are some presentations that inspired and were used for this course:

[1] Douglas A. Reynolds, "Overview of Automatic Speaker Recognition" http://www.fit.vutbr.cz/study/courses/SRE/public/prednasky/2009-10/07_spkid_doug/sid_tutorial.pdf

[2] Joaquin González-Rodríguez, "An Overview of the NIST Series of Speaker Recognition Evaluations and Technologies" http://tv.uvigo.es/matterhorn/20021

[3] Javier González-Domínguez, "Session Variability Compensation in Speaker Recognition" http://tv.uvigo.es/matterhorn/20022

- Recommended reading:

[Reynolds2000] Douglas A. Reynolds, Thomas F. Quatieri, Robert B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing 10(1-3): 19-41, 2000

[Kenny2005] Patrick Kenny, "Joint factor analysis of speaker and session variability : Theory and algorithms", Technical report CRIM-06/08-13 Montreal, CRIM, 2005

[Campell2006] W. M. Campbell, D. E. Sturim, D. A. Reynolds, A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in Proc ICASSP, Tolouse, May 2006

[Dehak2009] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification", In Proc Interspeech 2009, Brighton, UK, September 2009

Publicly available tools

- General Speech Recognition toolkits:
  - HTK - http://htk.eng.cam.ac.uk
  - KALDI - http://kaldi.sourceforge.net
- Specific Speaker Recognition toolkits:
  - ALIZE http://mistral.univ-avignon.fr/index_en.html
  - SPEAR https://pypi.python.org/pypi/bob.bio.spear
- Other useful tools:
  - I-vectors: http://www.voicebiometry.org + KALDI
  - Fusion & Calibration
    - Focal: https://sites.google.com/site/nikobrummer/focal
    - Bosaris: https://sites.google.com/site/bosaristoolkit/
  - DET plotting: http://www.nist.gov/itl/iad/mig/upload/DETware_v2-1-tar.gz