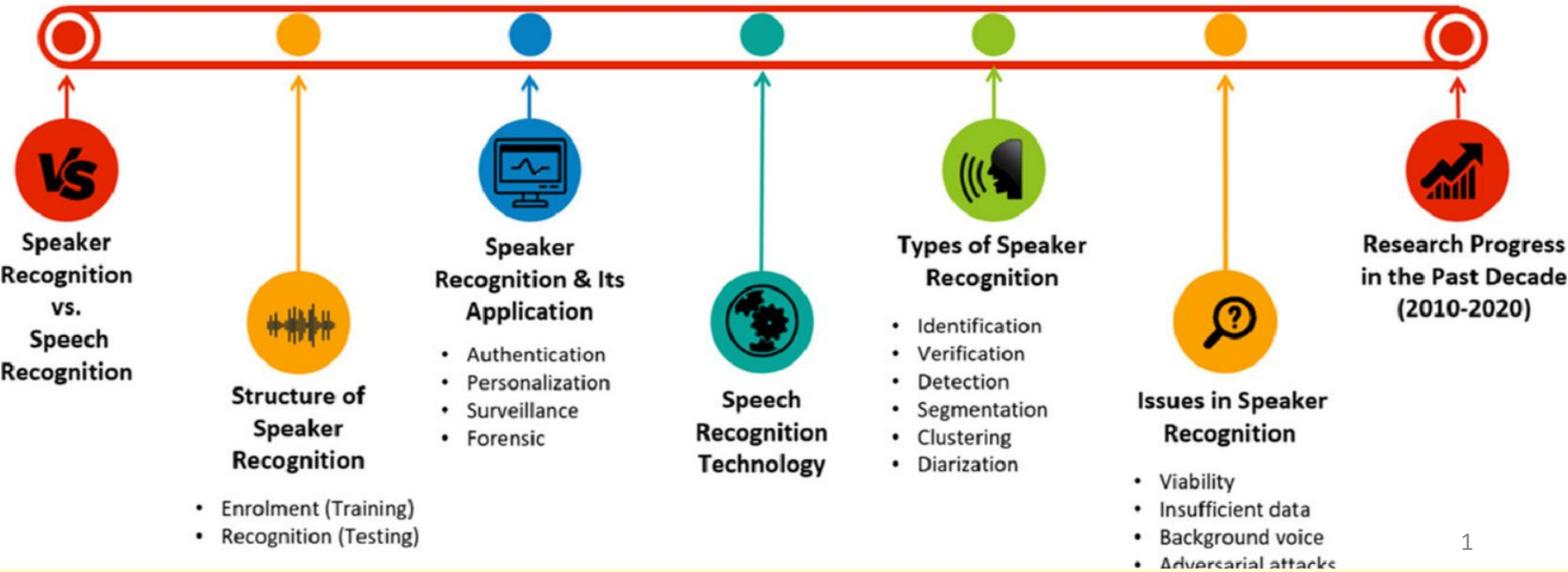
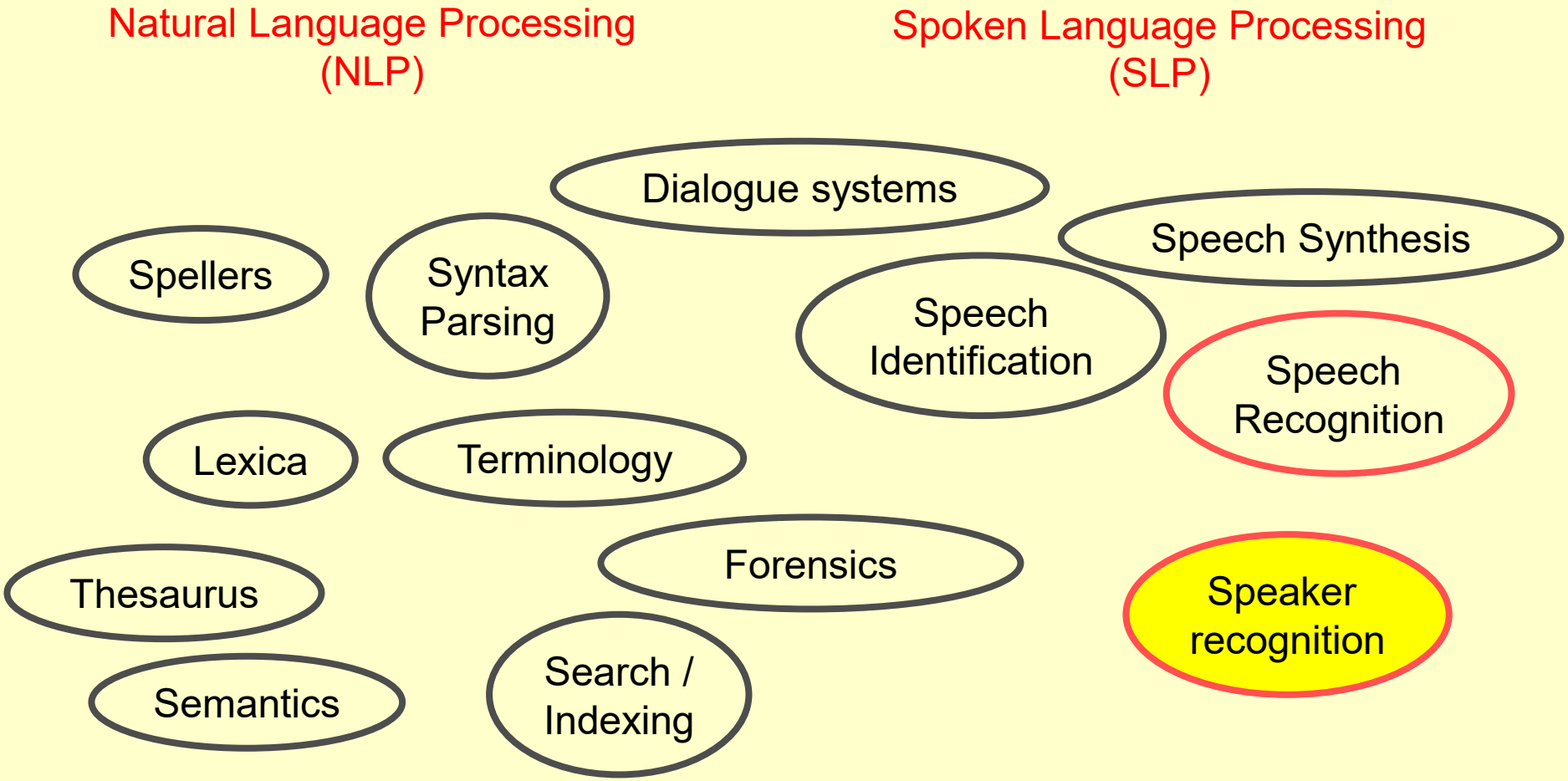
C14

RECUNOASTEREA VORBITORULUI

overview

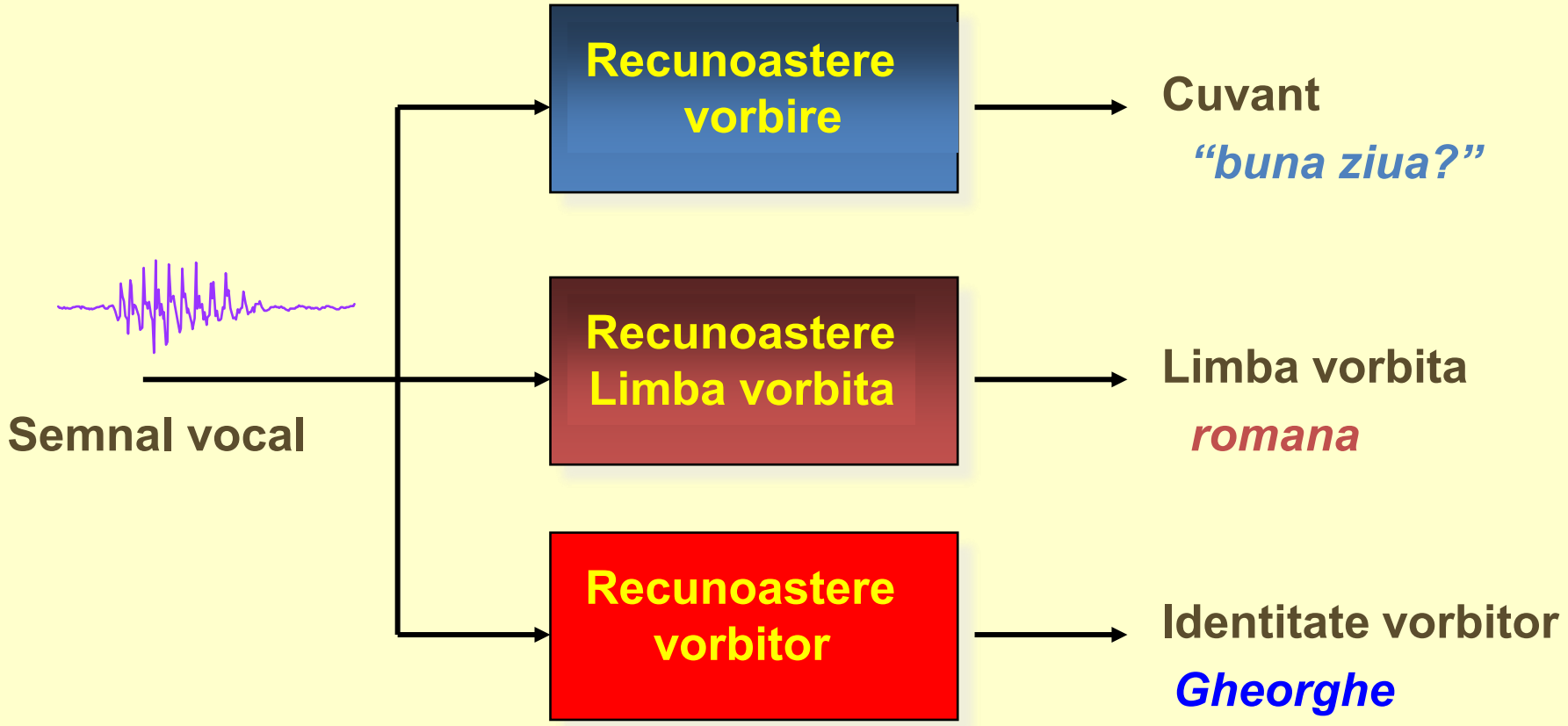


Taxonomia prelucrarii SV



Extragerea informatiilor din vorbire

Obiectiv: extragerea automata a informatiilor transmise de SV

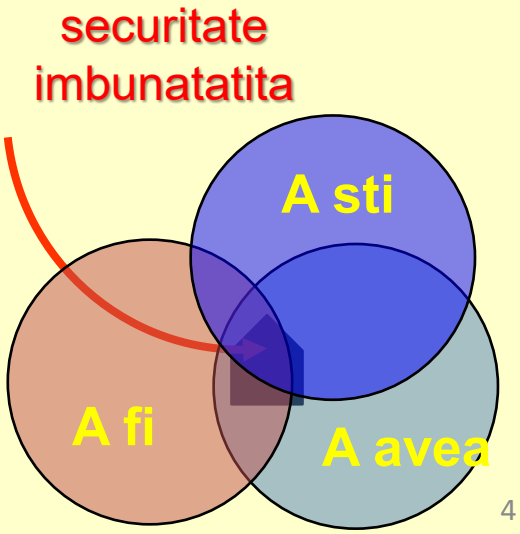


Vorbirea caracteristica biometrica

- Biometria: un atribut sau semnal uman generat ptr. autentificarea identitatii unei persoane
- Vorbirea este o caract. biometrica populara
  - ✓ Semnalul este produs natural
  - ✓ Nu necesita dispozitiv/senzor specializat
  - ✓ omniprezent: telefoane si microfoane la PC

- Biometria vocala + alte forme de securitate

- ceva ce aveti – de exemplu, insigna
- ceva ce stiti - de exemplu, parola
- ceva ce esti – de exemplu, vocea



- Recunoașterea vorbitorului a devenit intre cele mai populare metode în domeniul de identificare biometrică, deoarece vocea este cel mai comun semnal și cel mai simplu de achiziționat.

- Cu larga aplicarea a mașinilor de inteligență artificială, cercetători au descoperit că, comunicarea vocală este cea mai bună modalitate de a comunica între oameni-mașini.

- În domeniul controlului accesului, serviciilor de automatizare a tranzacțiilor prin telefon și domeniul de diarizare a vorbitorului, RV se aplică extensiv

- Metodele de recunoaștere a vorbitorilor au fost studiate și dezvoltate de-a lungul a cinci decenii

- Metodele de recunoaștere a vorbitorului au evoluat *de la compararea spectrogramelor auditive umane* si potrivirea simplă a modelului, la alinierea dinamică a timpului (DTW), la abordari mai moderne, precum recunoașterea folosind modelele statistice și la cea mai populară din ultimii ani, învățarea profundă (deep learning)

"Verificatori biometrici"

Caracteristica biometrica	EER	FAR	FRR	Parametrii de Test (conditii)	Test
Fata	-	1 %	10 %	Iluminare diferita interior/exterior	FRVT (2002)
Amprenta Digitala	-	1 %	0.1 %	Date furnizate de US Government	FpVTE (2003)
	2%	2 %	2 %	Rotatii sau distorsiuni exagerate ale pielii	FVC (2004)
Forma Mainii	1%	2 %	0.1 %	Cu inele si pozitii improprii	(2005)
Iris	0.01%	0.0001 %	0.2 %	Cele mai bune conditii	NIST (2005)
Voce	6%	2 %	10 %	Independent de text, multilingual	NIST (2004)

Caracteristicile vorbitorului

Vorbirea contine :

- informaţii **lingvistice,** care reprezintă mesajul sec, independent de cine îl transmite
- informaţii **legate de vorbitor**, care dau indicii despre identitatea celui care vorbeşte
- informaţii **afective**, legate de starea emoţională a vorbitorului (emoţie, stress, sănătate etc.)

- Vorbirea este rezultatul unei secvente complexe de transformãri produse la câteva nivele diferite: semantic, lingvistic, articulator si acustic.

- Variatiile în vorbire legate de vorbitor sunt cauzate de :

Variatii inter-vorbitor

- diferente anatomice - se datoresc formei si marimii tractului vocal
- diferente în deprinderile verbale (habit verbal) - modul în care vorbitorii au învatat sã foloseascã mecanismul vorbirii

Variatiile intra-vorbitor - datorate diferentelor între rostirile ale aceluias vorbitor

- viteza de vorbire
- starea emotionalã
- stress
- Sãnãtate
- varsta

Analiza variantei fonemice

- cercetările lui Matsumoto [Mat89] indică faptul că informaţia fonemică este semnificativ mai importantă decât cea datorată vorbitorului sau cea datorată corelaţiei dintre ele
- caracteristicile vorbitorului sunt transmise printr-un segment de vorbire prin informtia dependenta si cea independenta de foneme
- Considerând factorii datoraţi vorbitorului şi cei fonemici ca un vector caracteristic, $\mathbf{x_{psi}}$, extras din segmentul **"i"** de vorbire al fonemei "**p**" rostite de vorbitorul "**s**", poate fi exprimat astfel :
-

$$x_{psi} = \mu + a_p + \beta_s + \gamma_{ps} + e_{psi}$$

unde :

μ - este vectorul medie pe toţi vectorii observaţi

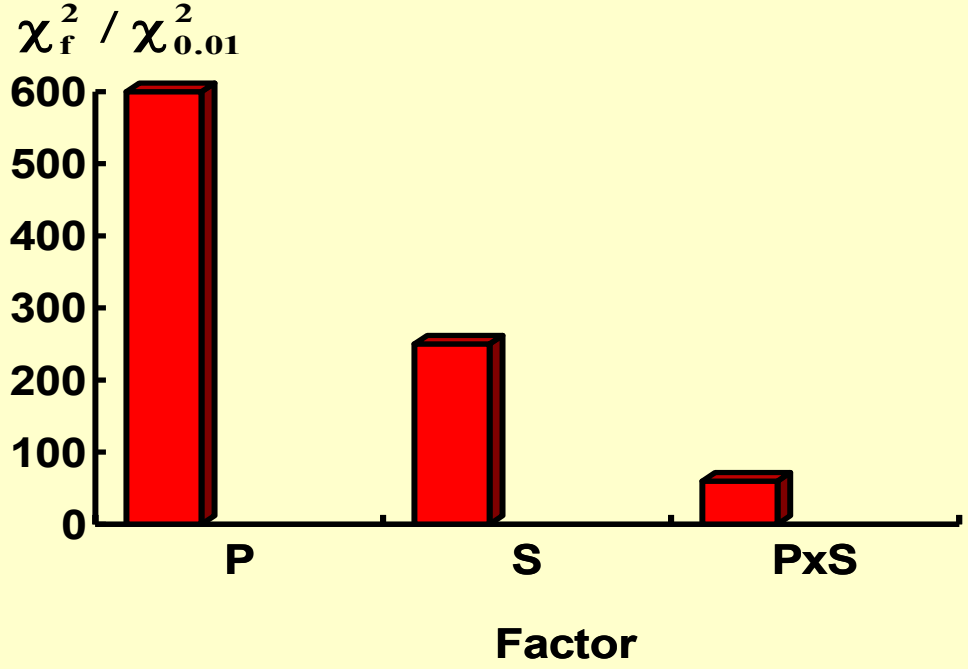
β_s - este factorul principal al vorbitorului constând în informaţia personalã independentã de foneme

a_p - este factorul principal fonemic

γ_{ps} - este factorul de interacţiune între foneme şi vorbitor care conţine informaţia personalã dependentã de fonemã

e_{psi} - termenul rezidual care implicã variaţiile datorate emoţiei, stării de sănătate etc.

- semnificaţia statisticã a fiecărui factor a fost testatã pe baza statisticã $\chi 2$



Analiza variantei factorilor S(vorbitor), P(fonemic) si SxP (interactiunea lor)

Din diagrama rezulta :

- **factorul fonemic** este foarte important (dominant) ceea ce sugerează că acesta poate corupe informatia specifică vorbitorului mai ales la recunoasterea independentă de text a vorbitorului
- **factorul fonemic dependent de vorbitor** γ_{ps} desi nu este asa de mare ca factorul principal al vorbitorului are o valoare semnificativă fiind de 60 de ori mai mare decât nivelul de semnificantă de 1%.

Caracteristici individuale

Informatiile individuale specifice vorbitorului sunt reprezentate de :

- calitatea vocii
- înaltime
- intensitate
- viteza
- intonatia
- accent
- vocabular

- **Proprietãtile parametrilor folosiți la recunoasterea vorbitorului**

Ideal ar fi ca parametrii vocali sã îndeplineascã urmãtoarele conditii :

- sã reprezinte eficient informatia dependentã de vorbitor
- sã fie usor de mãsurat
- sã fie stabili în timp
- sã aparã natural si frecvent în vorbire
- sã se modifice putin în medii diferite
- sã nu se preteze la imitare

✓ Atributele dezirabile ptr. caracteristicile ASR (Wolf) – practice, robuste si sigure

$$F = \frac{\text{variatia} - \text{medie} - \text{intervorbitor}}{\text{variatia} - \text{medie} - \text{intravorbitor}}$$

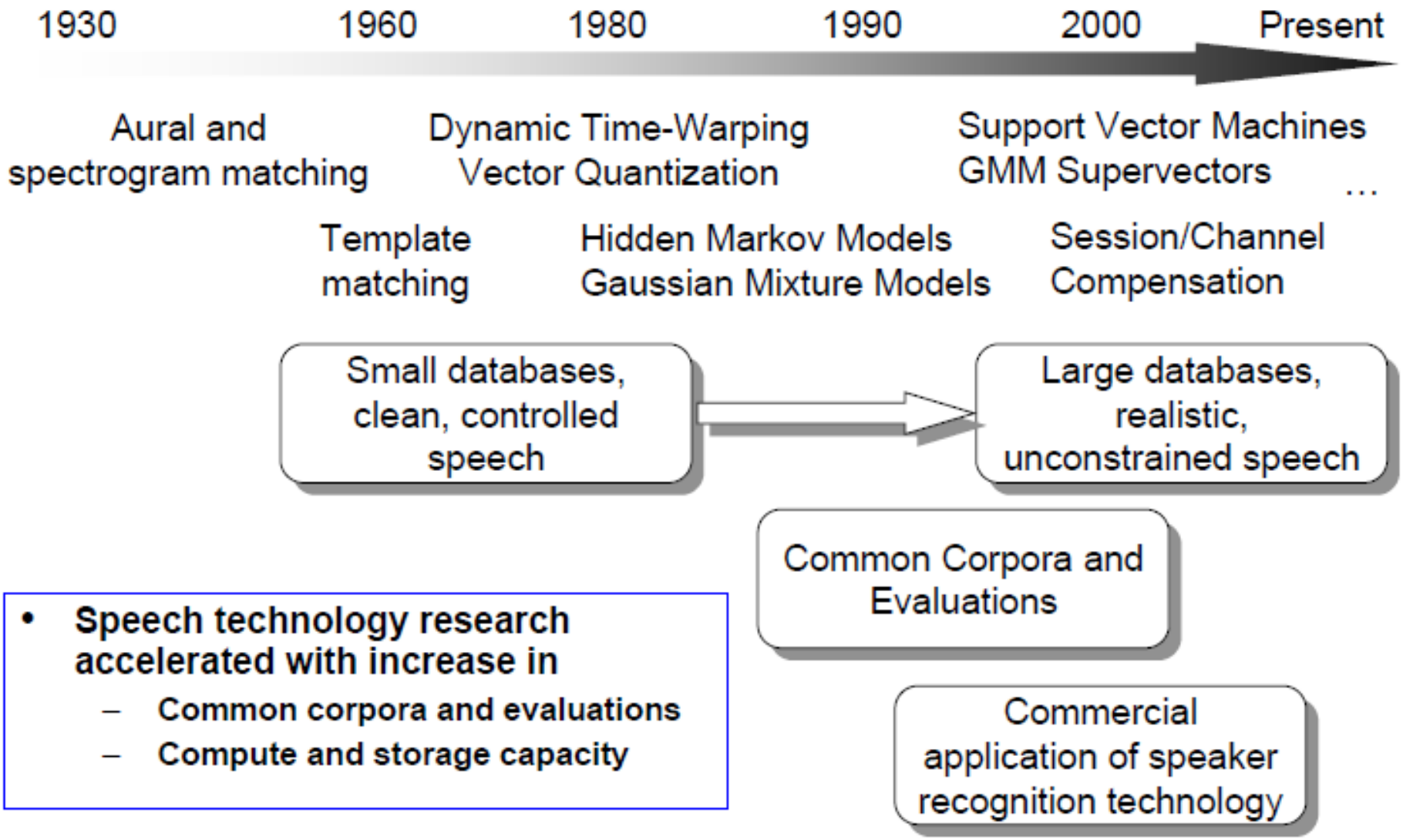
Obiectivele clasificarii si recunoasterii vorbitorilor

- identificarea sexului vorbitorului
- identificarea vârstei
- identificãrii stãrii de sãnãtate
- identificarea dispozitiei vorbitorului (stresat, vesel, calm, supãrat)
- identificarea accentului (provenienta socialã a vorbitorului)
- identificarea limbii vorbite
- identificarea unei anumite persoane este uzual numita ca ***recunoasterea vorbitorului***

- ***identificarea vorbitorului*** constã în gãsirea la ce clasã/vorbitor apartine cel mai probabil rostirea curentã/de test
- ***verificarea vorbitorului*** are ca scop validarea sau invalidarea ipotezei cã rostirea apartine vorbitorului sau clasei care o revendicã

- **s*egmentarea si diarizarea* –** are ca scop gasirea/etichetarea segmentelor care apartin aceluiasi vorbitor

Evolutia recunoasterii vorbitorului



- **Speech technology research accelerated with increase in**
  - **Common corpora and evaluations**
  - **Compute and storage capacity**

Evolutia recunoasterii vorbitorului

- Recunoașterea vorbitorului poate fi vazută în patru etape

Prima etapă a fost din anii 1960-70, care sa ***concentrat cercetarea extragerea caracteristicilor și tehnica de potrivire a șabloanelor/modelelor***

- 1962, Kesta de la Bell LABS a propus metoda spectrogramei pentru recunoașterea vorbitorului
- 1969, Luck a propus tehnica Cepstrum-ului
- 1976, Atal și colab. au folosit coeficienții (LPCC), care au îmbunătățit precizia de recunoaștere a vorbitorului.
- D.p.v. a modelului, potrivirea modelelor a fost adoptată în principal în anii 1960.
- În anii 1970, alinierea dinamică a timpului (DTW) și tehnologia de cuantizare vectorială (VQ) a devenit principala tendinta

A doua etapă a fost din anii 1980-90, ***modelele statistice*** ale vocii încep să fie aplicate si recunoașterii vorbitorului

- În ceea ce privește extragerea caracteristicilor, Davis propune ***parametrii Mel-Frequency cepstrum (MFCC)*** pentru recunoașterea vorbitorului, care devine caracteristica principală în anii următori
- În ceea ce privește modelele, abordarea clasica a fost împărțită în două tipuri.

✓ Primele tipuri sunt bazate pe ***VQ și DTW,*** care sunt denumite modele bazate pe șablon.

✓ Al doilea tip sunt modele stochastice care se bazează: pe ***mixturi gaussiene (GMM) sau Hidden Markov Model (HMM)***

- Majoritatea sistemelor de RV de ultimă generație au adoptat ***caracteristicile MFCC și au utilizat modelele de mixturi gaussiene (GMM)*** pentru modelarea vorbitorilor.
- Mixturile gaussiene s-au dovedit de succes la recunoașterea vorbitorului

A treia etapă (2000-2010)

- Metodele de recunoaştere a vorbitorului sunt bazate pe mixturile gaussiene;
- GMM au fost cele mai des folosite (propuse de Reynolds) includ clasica adaptare maximă a posteriori (MAP) a modelului universal de baza (UBM) a parametrilor (GMM-UBM) și folosesc clasificarea (SVM) a supervectorilor GMM (GMM-SVM)

- În faza de instruire, adaptarea MAP oferă o modalitate de încorporare a informațiilor anterioare, prin adaptarea parametrilor GMM din UBM. Cadrul este disponibil în abordarea problemelor generate de antrenarea cu date putine

- SVM folosește o funcție neliniară pentru a mapa datele intr-un spațiu dimensional superior (posibil infinit) și apoi găsește cel mai bun hiperplan care separă cele două clase din acest spațiu (SVM este practic un clasificator pentru 2 clase)

A patra etapă a început la începutul anilor 2000 (2010- prezent)

- Invăţarea profunda (DL) promovează dezvoltarea RV
- In această etapă, dezvoltarea tehnologiei de RV este determinata de nevoile comerciale, în timp ce deep learning, big-data si unitatile (GPU) promovează de asemenea dezvoltarea RV
- Diferite tipuri de rețele neuronale profunde (DNN) au fost propuse pentru metodele de RV
- La extragerea caracteristicilor la nivel de cadru, cercetătorii utilizează DNN pentru a extrage caracteristicile Bottlenect (BN), d-vector, i-vector și x-vector
- La nivel de model, cercetarea se concentrează pe diferite (DNN) pentru modelarea caracteristicilor acustice. Deciziile au fost făcute utilizand distanța dintre vectorul caracteristic țintă și vectorul caracteristic de testare.
- Vorbitorii care interactioneaza cu sistemul sunt adesea necunoscuți în timpul antrenarii sistemului, acest lucru reprezintă o mare provocare pentru sistemele de RV.

- Indiferent dacă sunt extrasi i-vectori sau vectorul caracteristic din sistemul DNN este format din trei module:

- Primul este ***modulul de antrenare*** care calculează reprezentarea vorbitorului
- Al doilea este ***modulul de înrolare,*** care estimează modelul vorbitorului
- Ultimul este ***modulul de evaluare,*** care are o funcție de pierdere, adecvată pentru optimizare.

- O nouă abordare a fost propusa , în care toate modulele pot fi legate împreună.
- În comparație cu metodele prezente, o astfel de metodă end-to-end (E2E) modelează enunțurile și le estimeaza direct, rezultand modele mai bune și mai compacte.
- Această abordare are ca rezultat adesea sisteme mai simplificate care au nevoie de mai puține concepte și euristici

Taxonomia sistemelor de recunoasterea vorbitorului

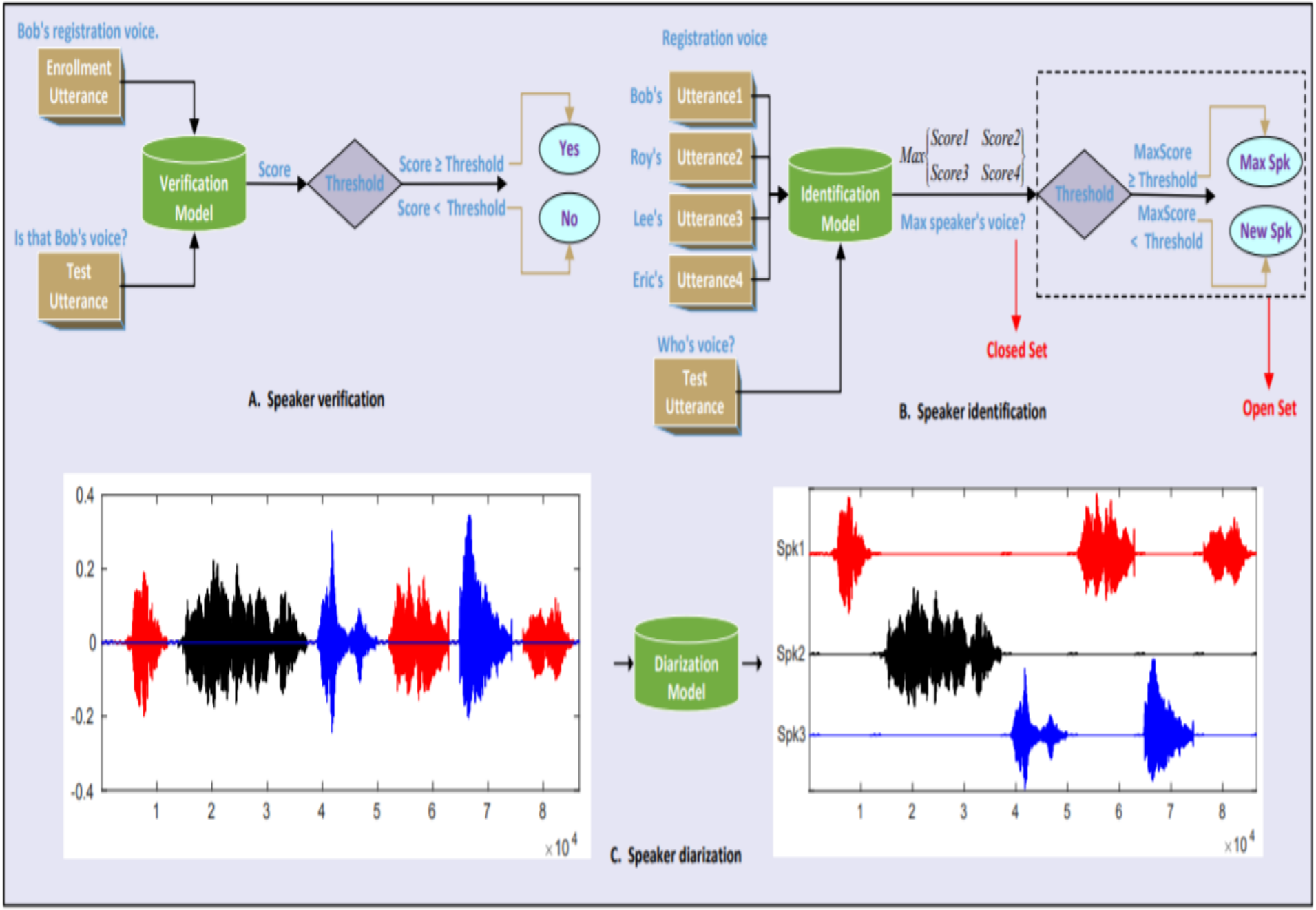
- verificarea vorbitorului
- identificarea vorbitorului

- sistemele pot fi împãrtite dupã gradul de ***dependenta de text*** :

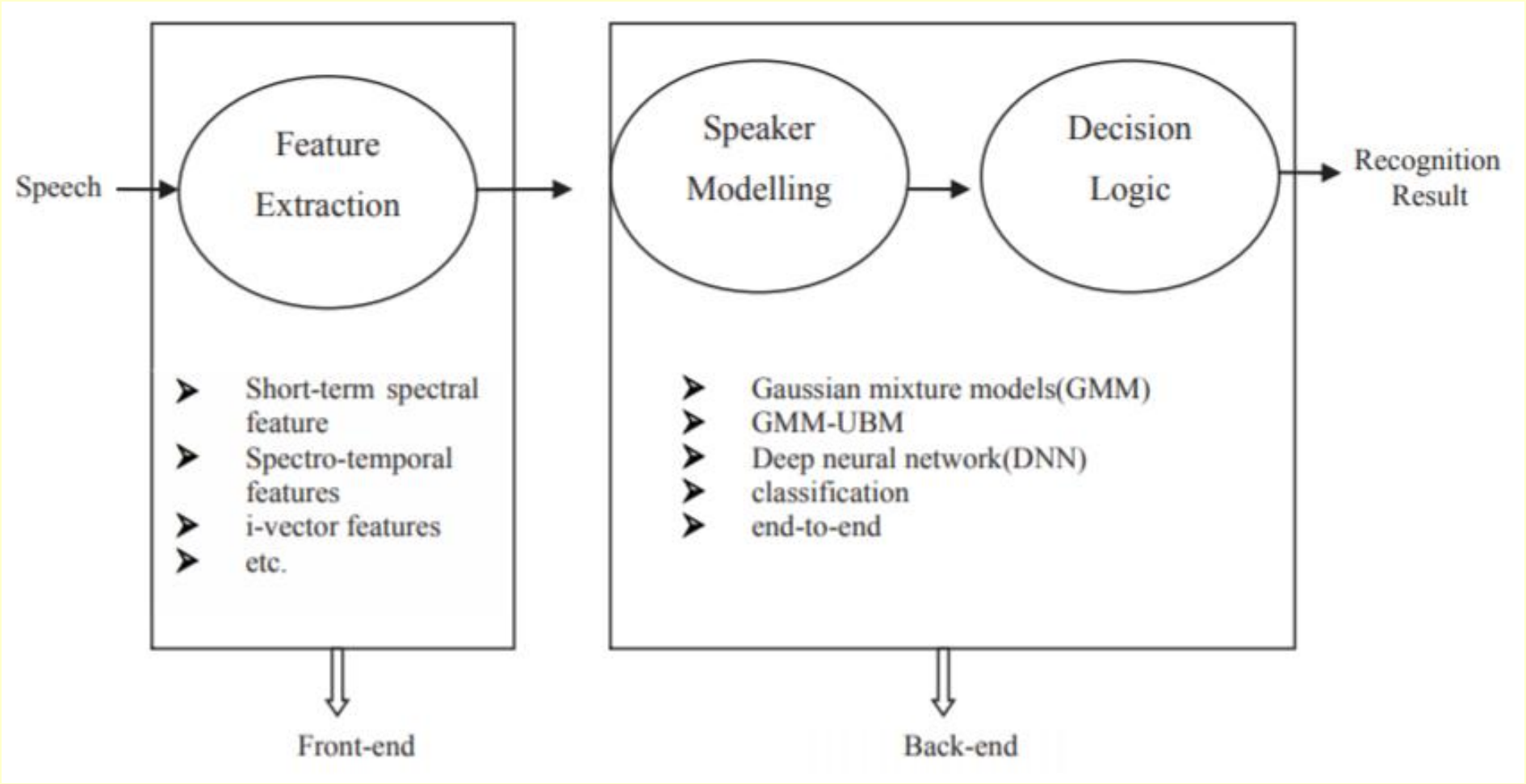
- **dependente de text** - **parole individuale**
  - **parole comune (**rigide)

- **independente de text**
  - **cu vocabular fix** (se folosesc aceleasi cuvinte într-o ordine aleatoare)
  - **dependente de un eveniment** (cautã un anumit eveniment lingvistic)
  - **vocabular fãrã restrictii** (independentã de text fãrã restrictii)

- **Segmentare si clustering (Diarizare)**



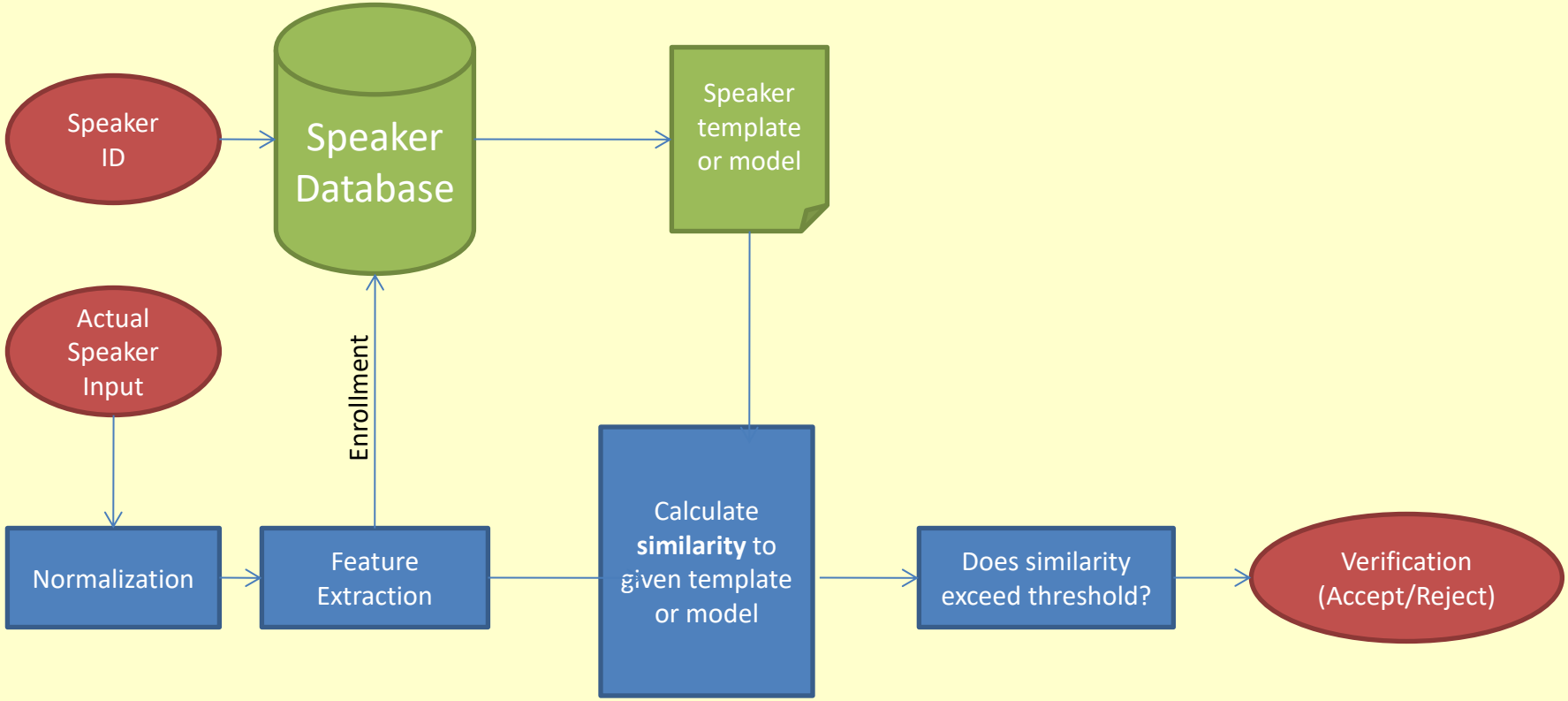
Diagrame de verificare, identificare și de diarizare a vorbitorilor



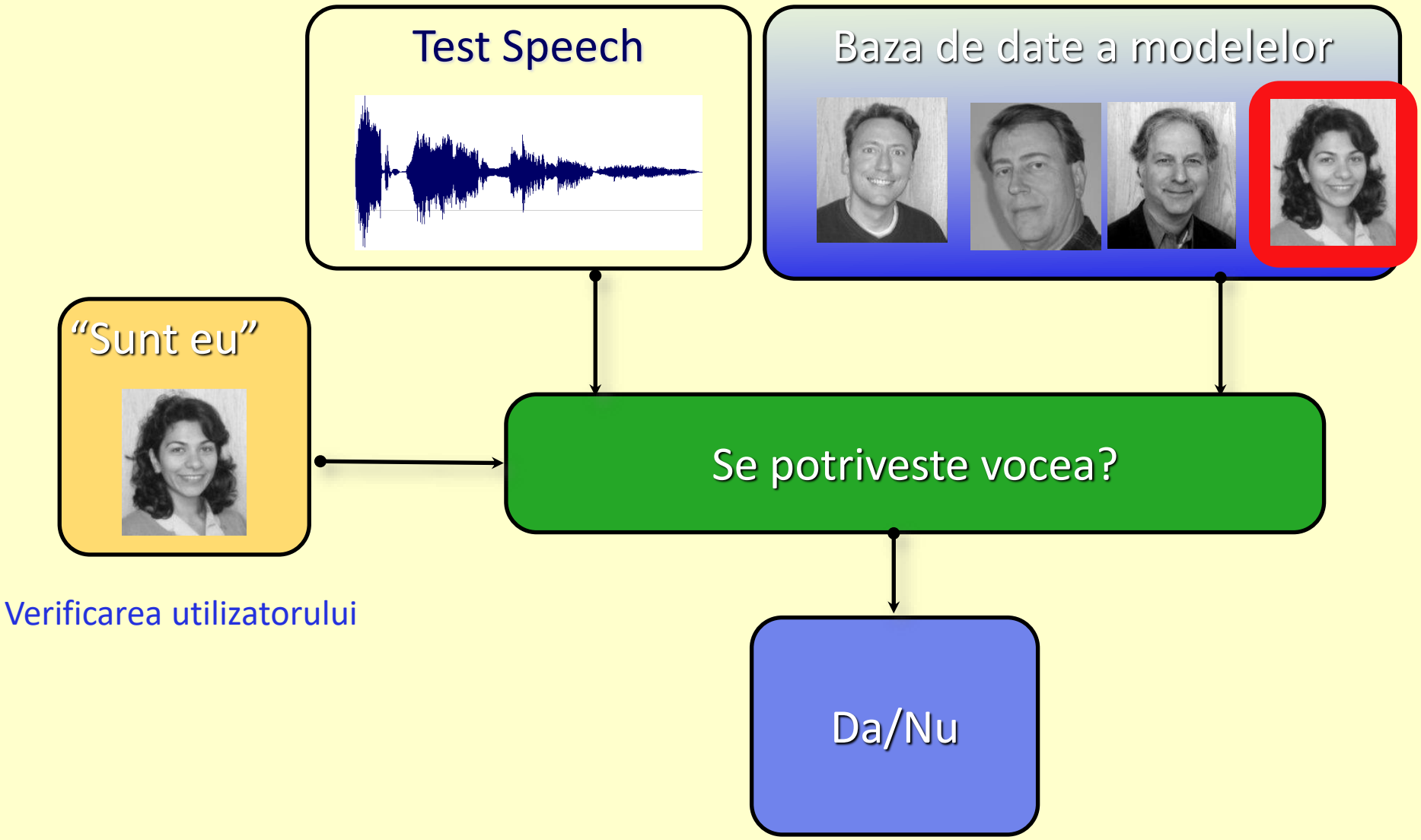
Sistem tipic de RV

Liu J, Philip Chen CL, Li T, Zuo Y, He P (2019) An overview of speaker recognition. Trends Computer Sci Inf Technol 4(1): 001-012

Verificarea vorbitorului

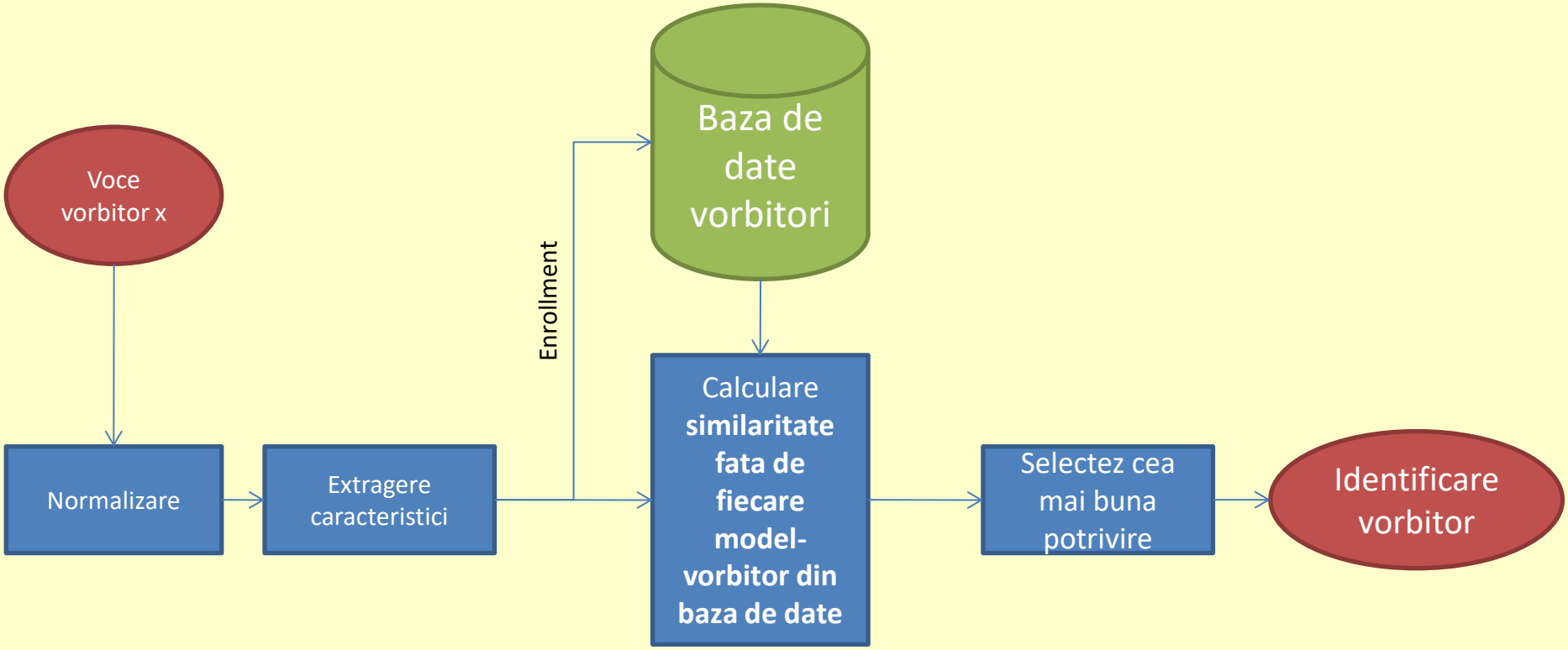


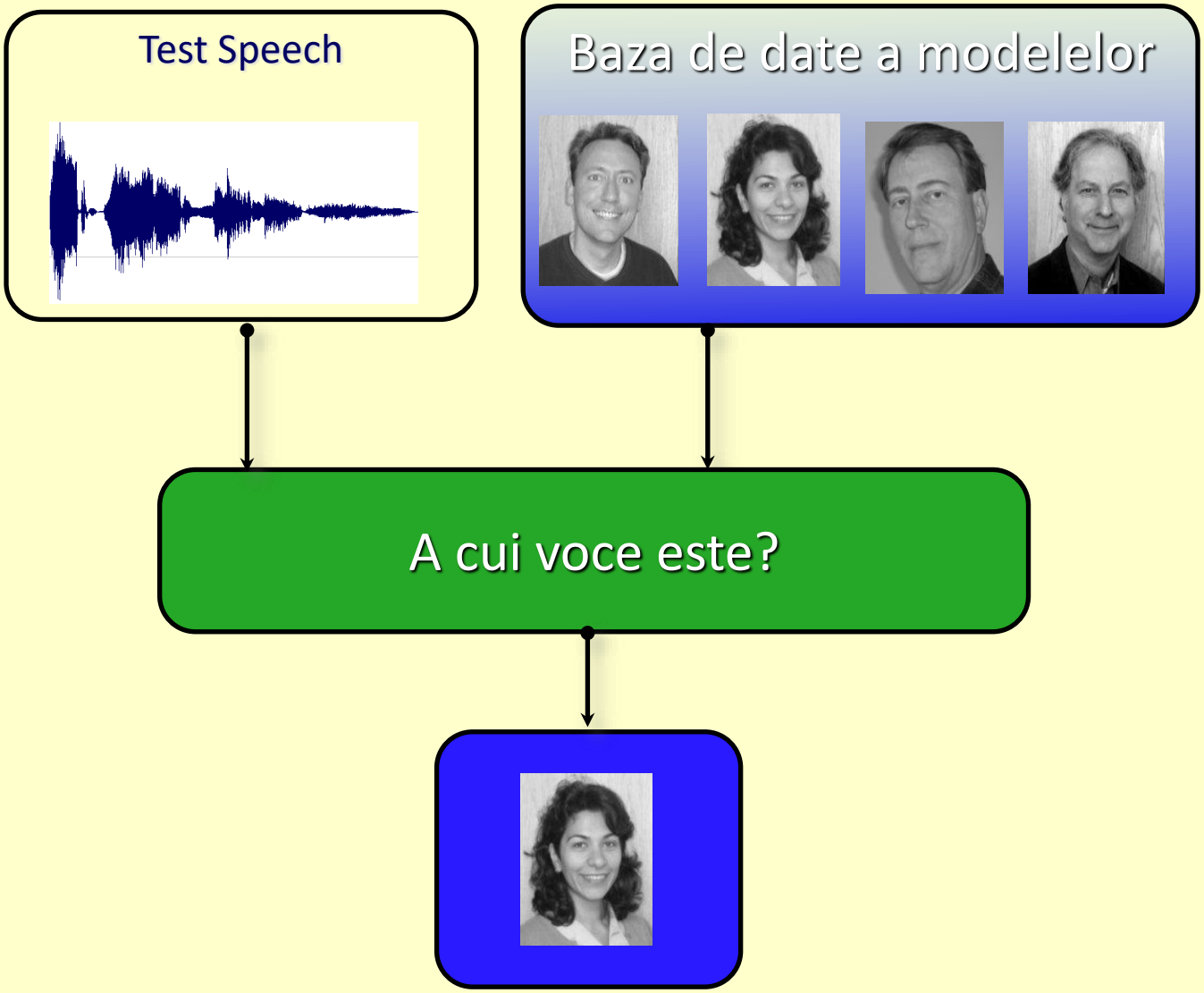
Verificare/Autentificare/Detectie



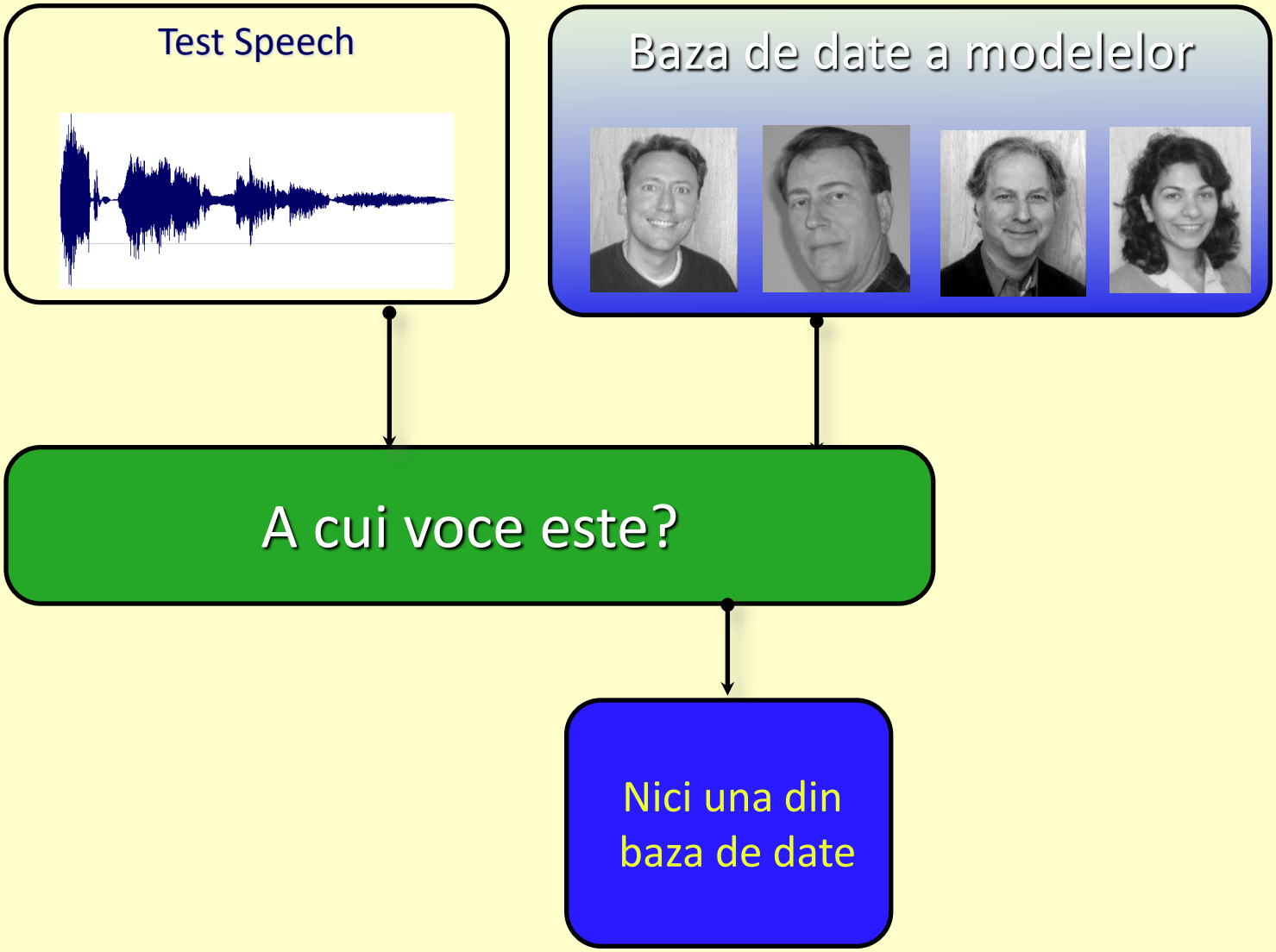
Verificarea utilizatorului

Identificarea vorbitorului



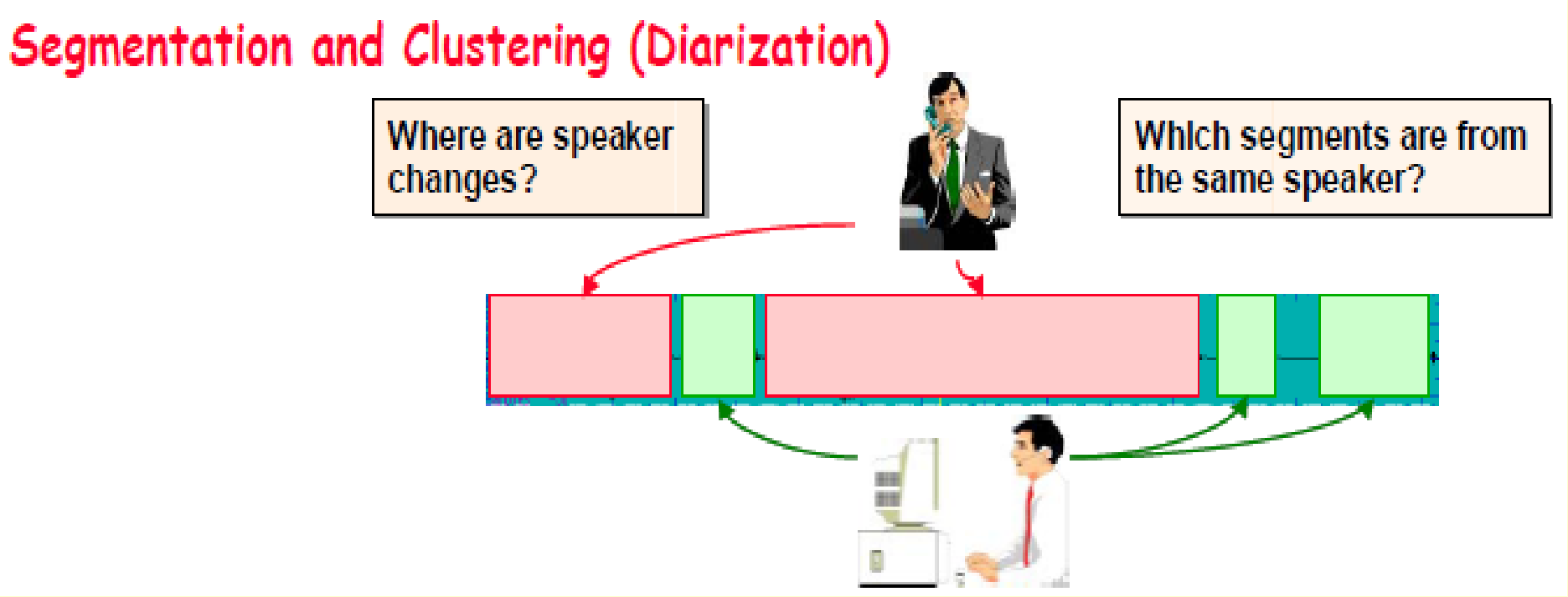


Identificare vorbitor cu set inchis



Identificare vorbitor cu set deschis

Segmentarea si clustering (Diarizare)



- Sunt disponibile Toolbox-uri open-source ptr. diarizarea vorbitorilor (***AudioSeg, ShoUT, DiarTK, ALIZÉE, LIUM*** etc.), cea mai veche este ***CMU toolkit***

Performantele sistemelor de verificare a vorbitorilor

- Exista mai multi factori de luat in considerare la conceperea unui mod de evaluare al unui sistem de verificare a vorbitorilor (VV)

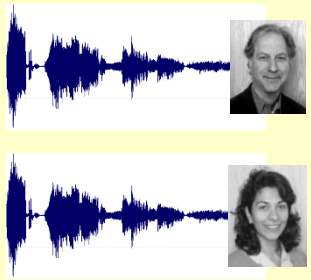
Calitatea vorbirii	- Caracteristica canalului si a microfonului - Nivel si tip de zgomot - Variabilitatea intre inrolare si verificare
Modalitatea de a vorbi	- Fixa/promptata/fraze selectate de utilizatori - Free text
Durata vorbirii	- Durata si numarul de sedinte de inrolare si verificare
Numarul de vorbitori	- Dimensiunea bazei de date

- Mai important : datele de evaluare si proiectare trebuie sa corespunda domeniului de interes al aplicatiei vizate

Faza de inrolare si test

Model ptr.
fiecare vorbitor

Faza de inrolare



Extragere caracteristici

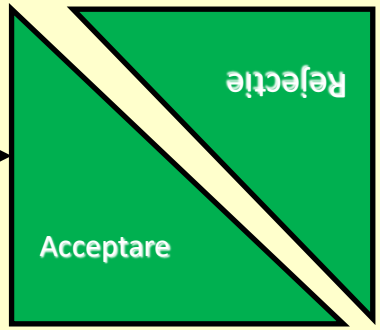
Antrenare Model

Rostiri de antrenare pentru fiecare vorbitor

Faza de recunoastere (Verificare)

Extragere caracteristici

Decizia de Verificare



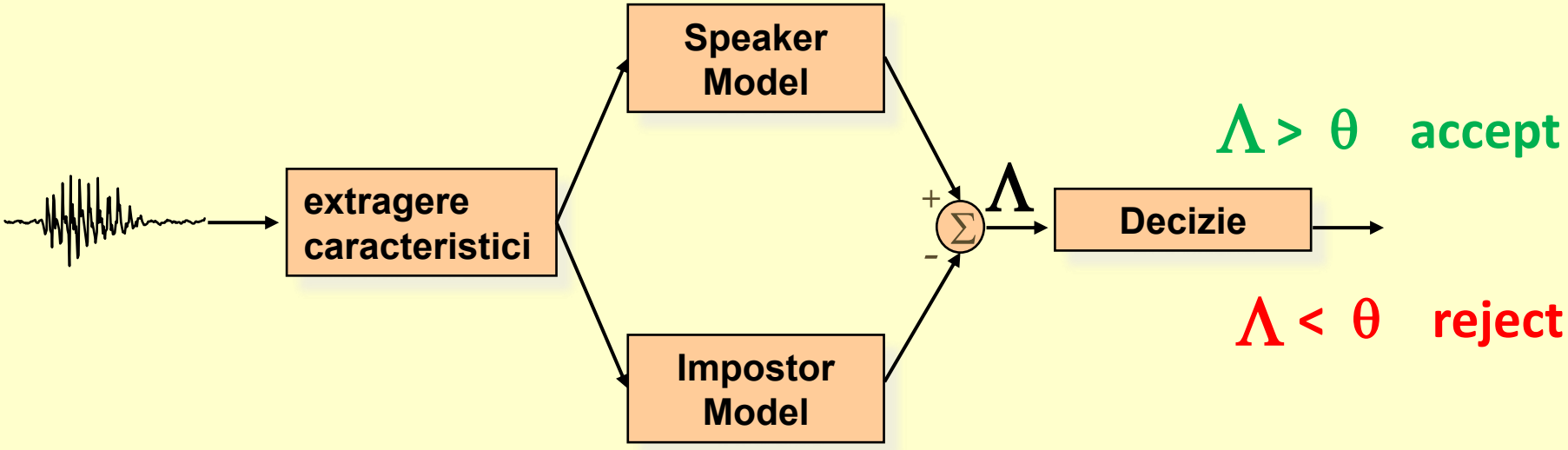
"Sunt eu!"

Luarea deciziei la verificare

- Abordarea ***deciziei de verificare*** are radacini in teoria detectiei semnalelor
- **Sunt 2** ipoteze **de test:**

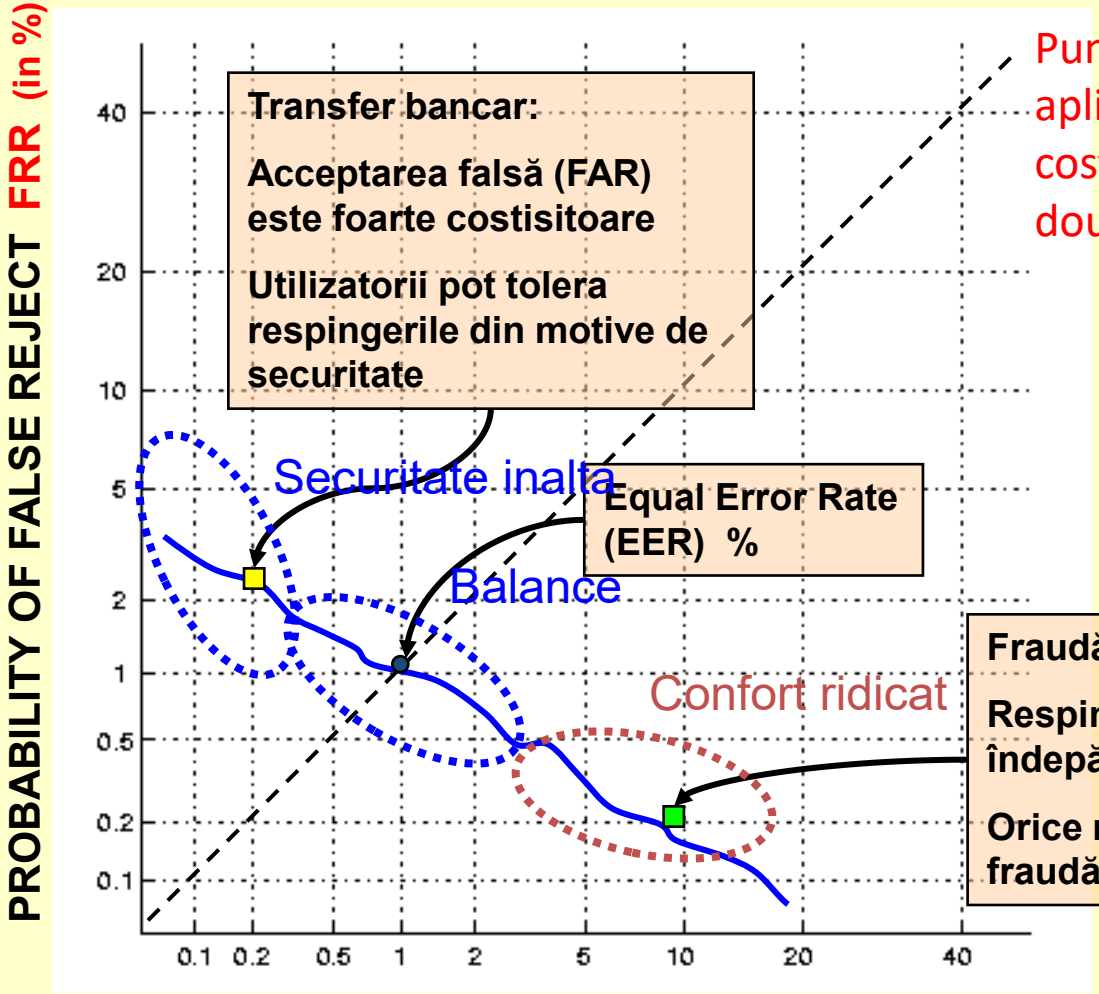
  H0: : vorbitorul este un impostor
  H1: : vorbitorul este utilizatorul revendicat
- Statistica calculata pe enuntul de test « S » ca raport de asemanare (**likelihood ratio – raportul de probabilitate)** :

$$\Lambda = \log \frac{\text{Probabilitatea ca } \mathbf{S} \text{ provine de la modelul vorbitorului}}{\text{Probabilitatea ca } \mathbf{S} \text{ nu provine de la modelul vorbitorului}}$$



Performantele Verificarii

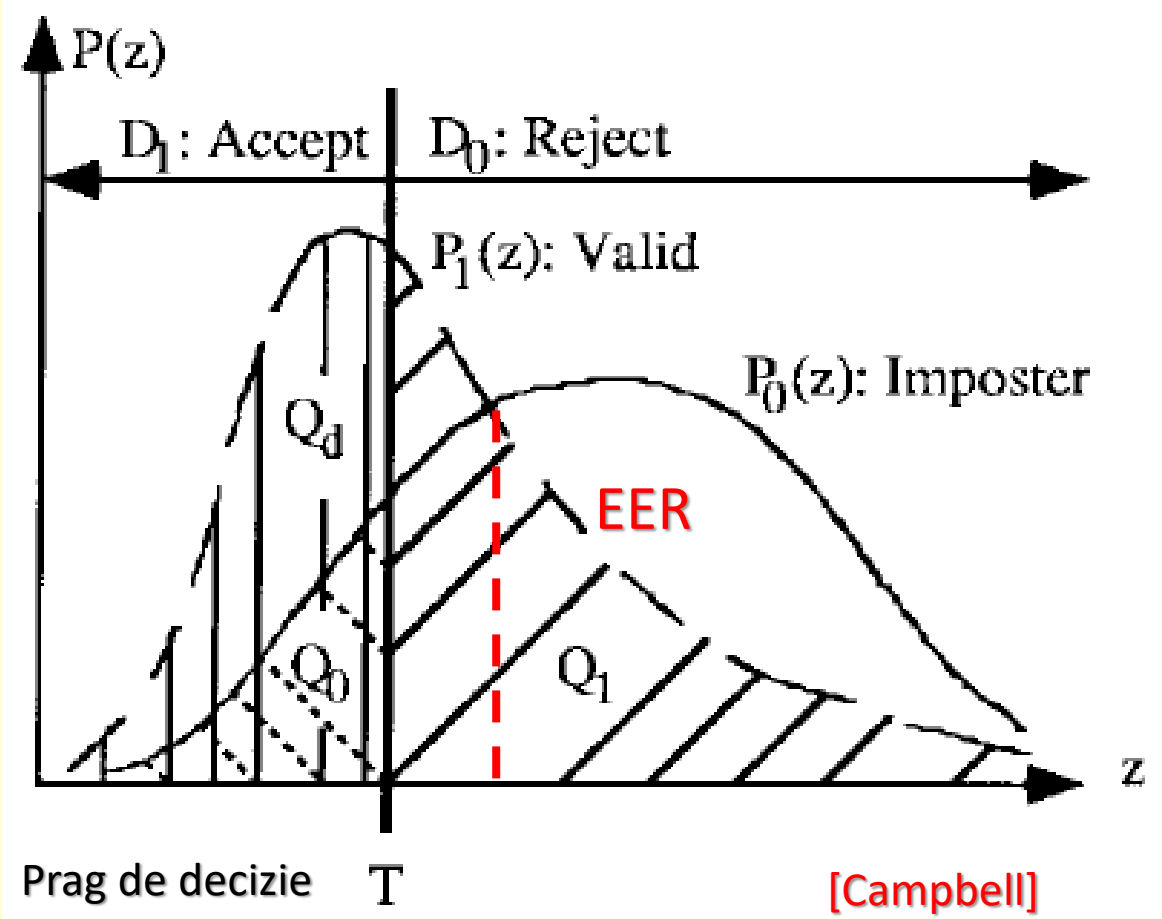
Evaluarea sistemelor de verificare a vorbitorului



Punctul de operare al aplicației depinde de costurile relative ale celor două tipuri de erori

Ipoteza de testare – luare a deciziei

Densitatile valide & impostor



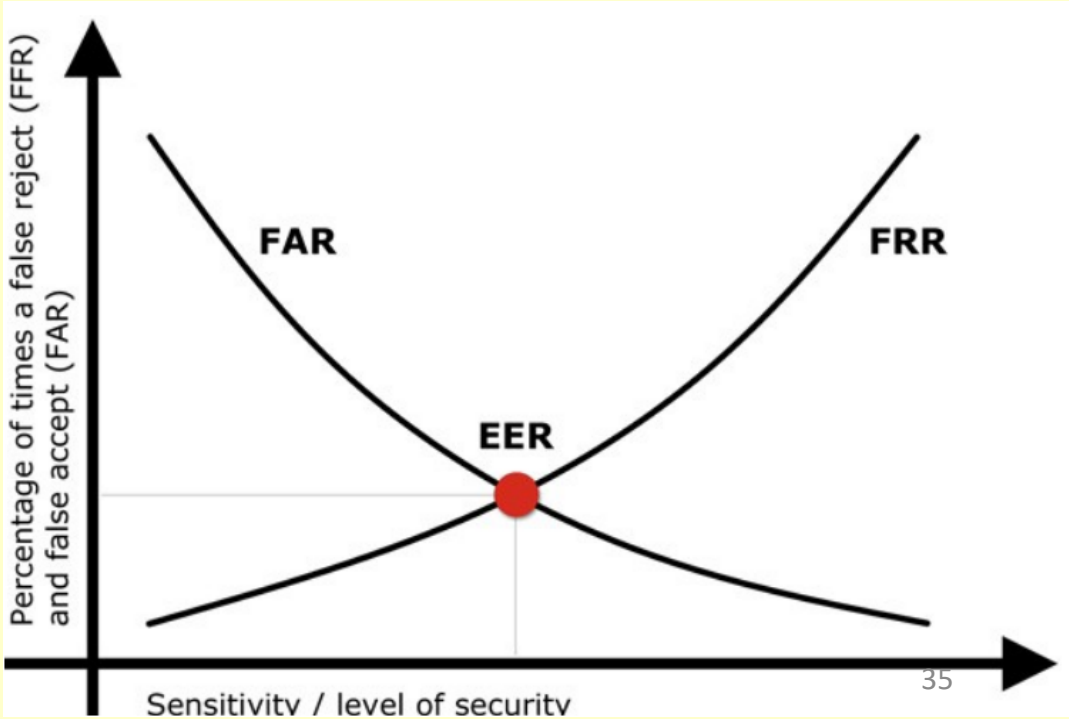
[Campbell]

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

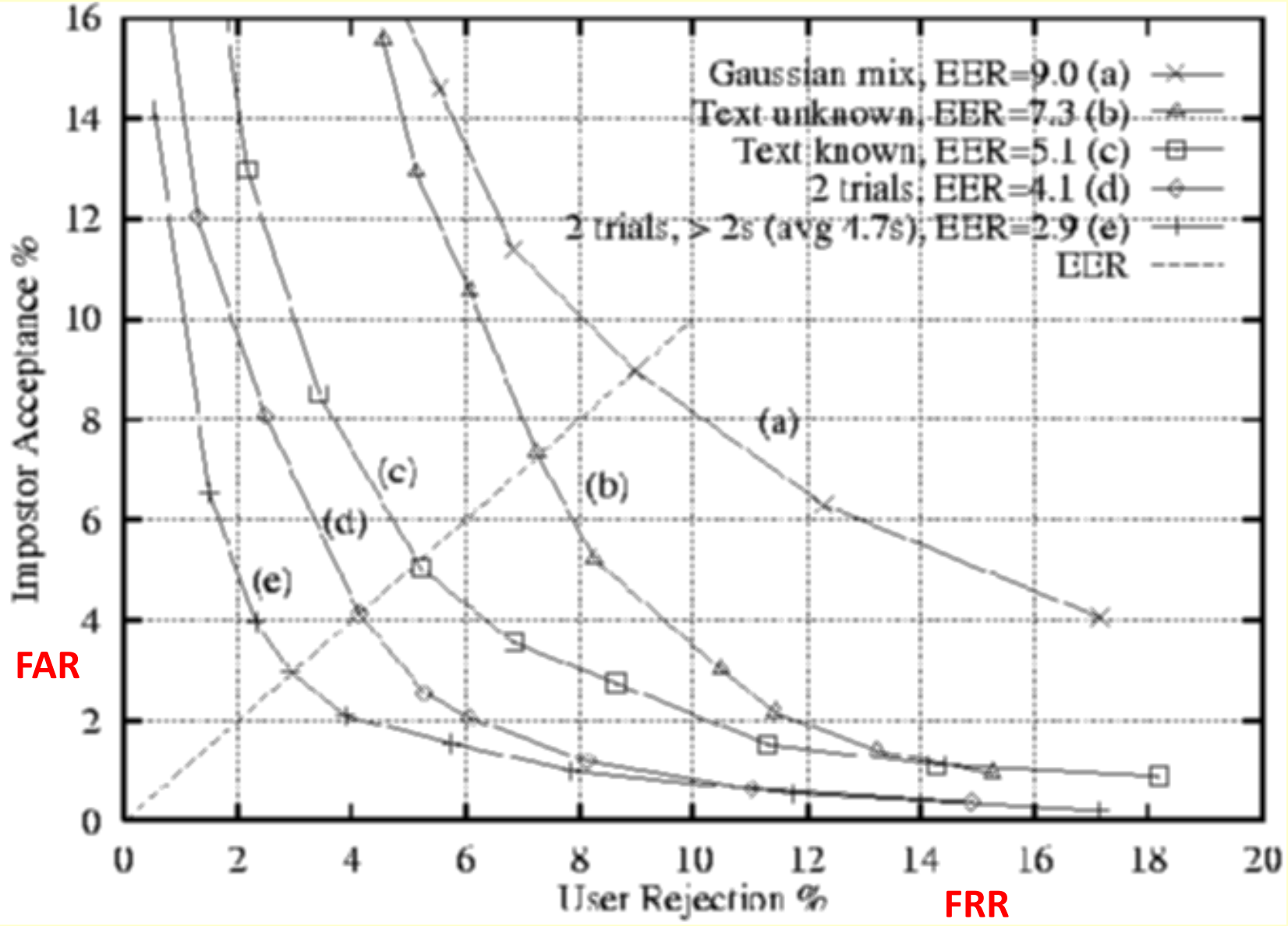
Probabilitatile performantei	Decizia D	Ipoteza I	Numele probabilitatii	Rezultatul deciziei
Q0	1	0	Semnificaţie (marimea testului)	FAR (alarma) falsa acceptare
Q1	0	1		FRR Falsa rejectie
Qd=1-Q1	1	1	Puterea testului	Acceptare corecta
1-Q0	0	0		Rejectie corecta

Precizia Sistemului

– FAR (Rata de acceptare falsă)
– FRR (Rata de respingere falsă)
– EER (Rata de egală eroare) (FAR~FRR)
– ROC
(Caracteristicile de funcționare ale receptorului)

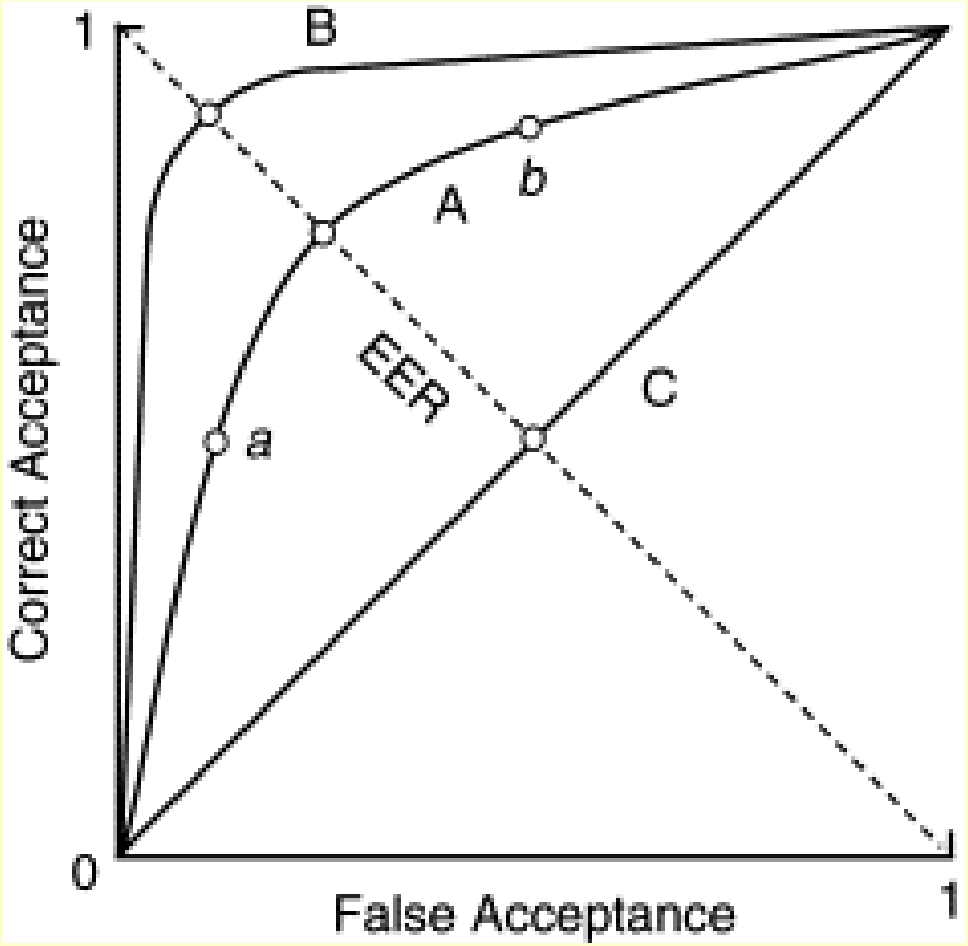


Evaluarea Performantelor



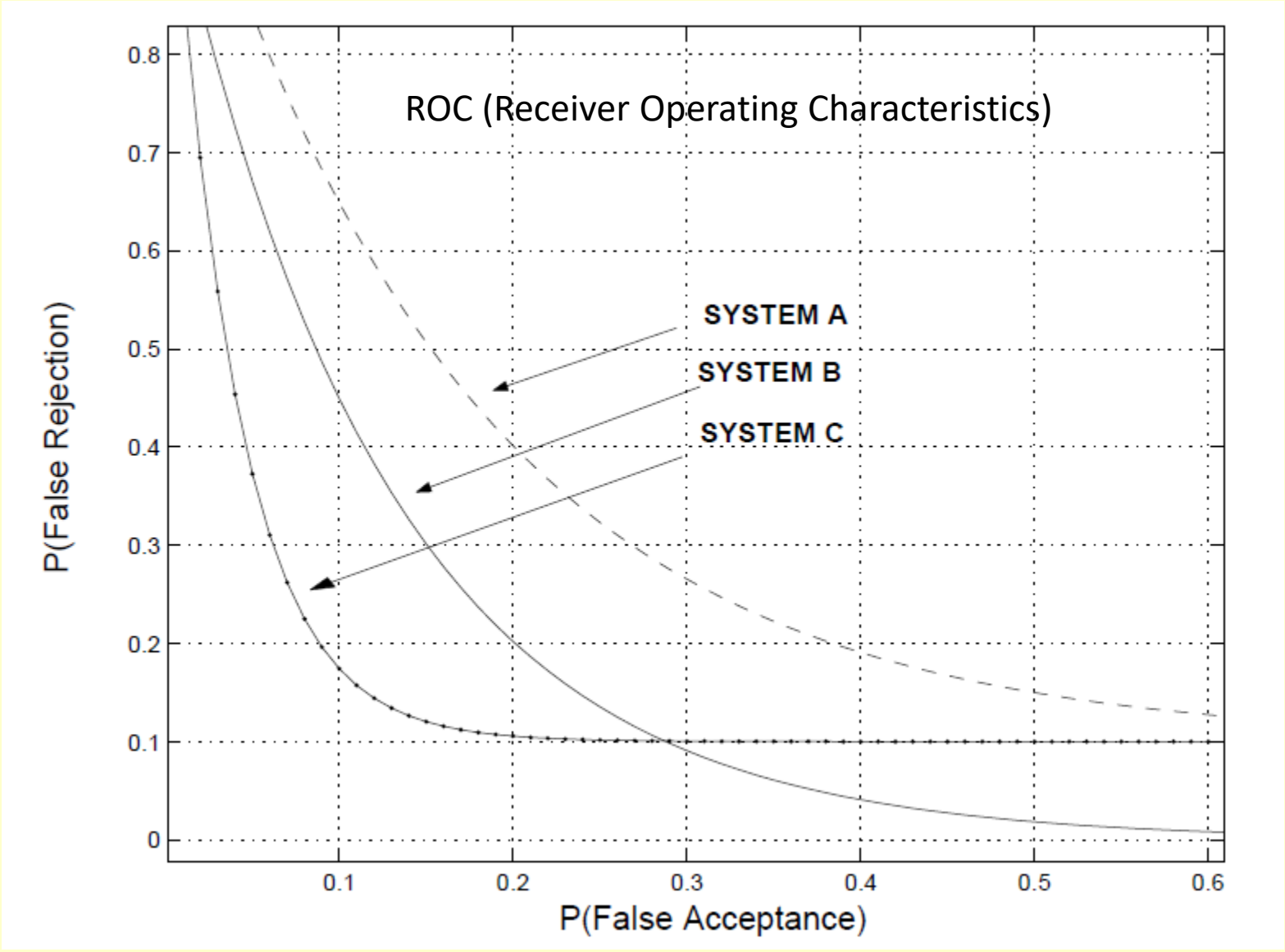
ROC (Receiver Operating Characteristics)

- Plotare diferite puncte de operare pentru diferite (valori FRR vs. FAR) denumit și DET (compromisul de detecție a erorilor)



Curbe ROC (Receiver operating characteristic);
ex. de performanță a trei sisteme de recunoaștere a vorbitorilor: A, B și C.

- Figura prezintă curbele A, B și C pentru trei sisteme. Este clar că performanța curbei B este sistematic mai bună decât cea a curbei A, iar curba C corespunde cazului limită de performanță pur aleatorie. Poziția a din figură corespunde cazului în care se utilizează un criteriu de decizie strict, iar poziția b corespunde unui caz în care se utilizează un criteriu lax

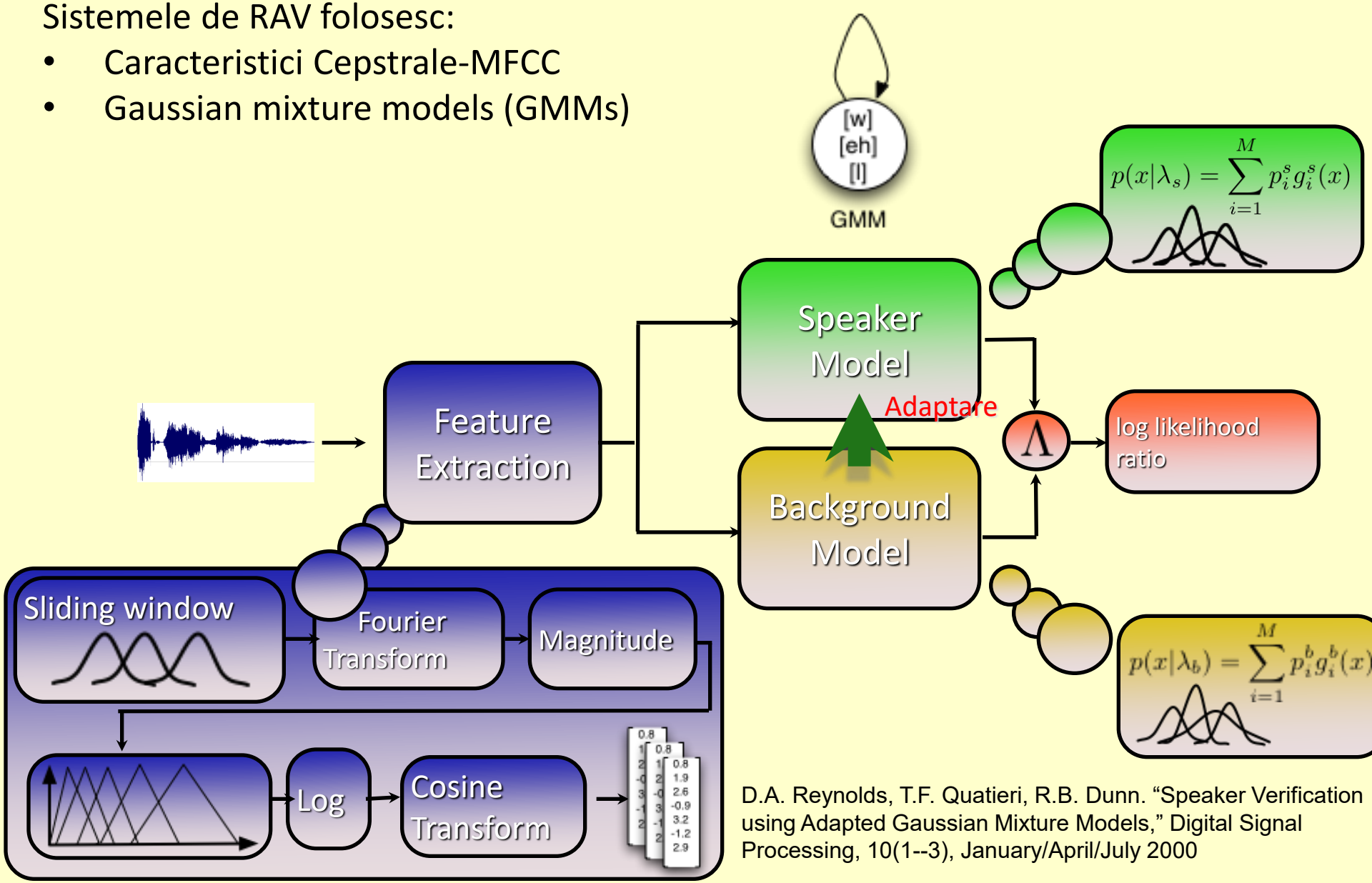


Tema. (a) Estimați EER-urile diferitelor sisteme. Ce puteti concluziona?
(b) Aplicația X utilizata face parte dintr-un sistem de inalta securitate de acces a persoanelor la documentele secrete. Având în vedere trei furnizori A; B; C, care sistem il recomandati? De ce?

Abordare RAV pe baza Spectrala

Sistemele de RAV folosesc:

- Caracteristici Cepstrale-MFCC
- Gaussian mixture models (GMMs)



D.A. Reynolds, T.F. Quatieri, R.B. Dunn. “Speaker Verification using Adapted Gaussian Mixture Models,” Digital Signal Processing, 10(1--3), January/April/July 2000

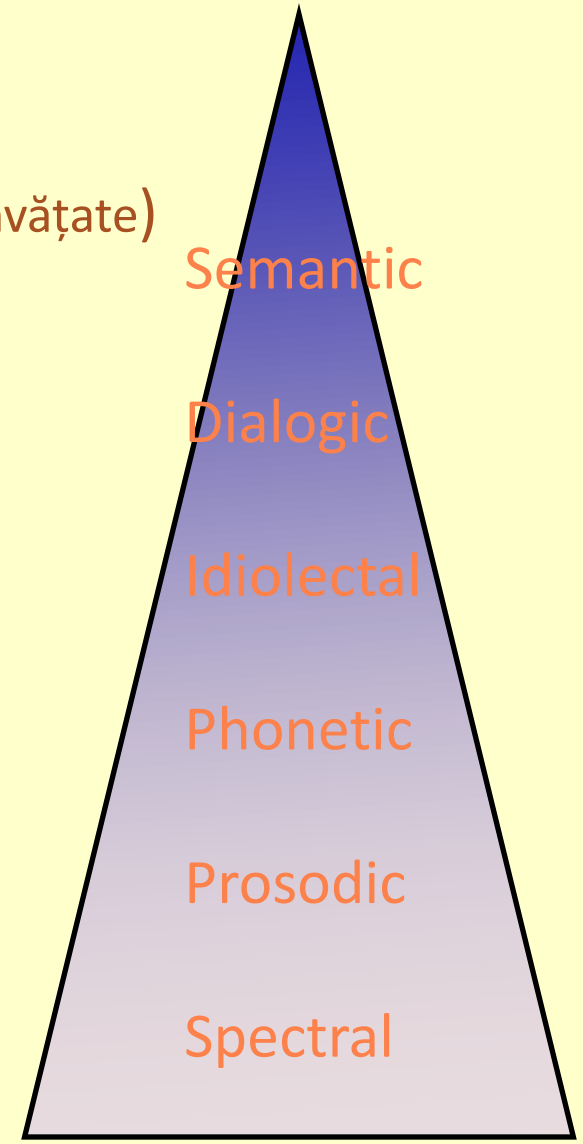
Caracteristici: Nivele de Informatie

Ierarhia cheilor perceptuale

semantics, idiolects, pronunciations, idiosyncrasies	socio-economic status, education, place of birth
prosody, rhythm, speed intonation, volume modulation	personality type, parental influence
acoustic aspects of speech, nasal, deep, breathy, rough	anatomical structure of vocal apparatus

Chei de nivel inalt
(comportamente învățate)

Chei de nivel jos
(Caracteristici fizice)

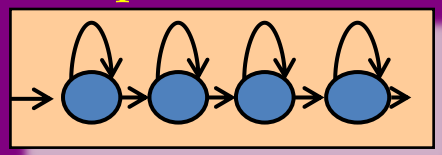


Modele vorbitor – HMM/GMM

- Formele modelelor HMM depind de aplicatie

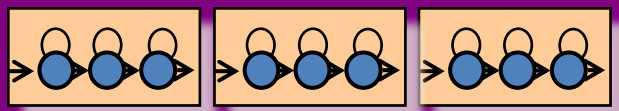
Parola Fixa → **Modele cuvant**

"Open sesame"



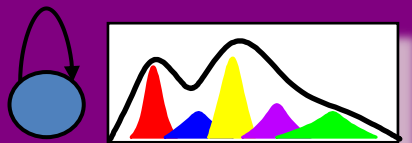
Parole Promptate → **Modele Foneme**

/s/ /i/ /x/



Text-independent → HMM cu o stare

General speech



Arhitectura Generală a unui Sistem Deep Learning

- Sistemele moderne de RV au abandonat modelele statistice tradiționale (cum ar fi GMM-UBM sau i-vectors) în favoarea *"Speaker Embeddings"* (vectori de caracteristici de dimensiune fixă extrași cu rețele neuronale).

Componentele Fluxului:

- **Front-end (Procesarea semnalului):** Transformarea formei de undă brute în reprezentări timp-frecvență (Spectrograme, Mel-Filterbanks sau MFCC). Recent, modele precum **SincNet** permit procesarea directă a semnalului brut (raw waveform).
- **Encoder Network (Speaker Embedding):** O rețea DNN care mapează un segment audio de lungime variabilă într-un vector de dimensiune fixă (ex: 256/ 512 unități).
- **Pooling Layer:** Deoarece intrările audio au durate diferite, un strat de mediere (Global Average Pooling) sau **Statistics Pooling** (medie și deviație standard) transformă datele într-un format fix.
- **Back-end (Scoring):** Compararea a doi vectori (cel de test și cel de referință) folosind **Cosine Similarity** sau **Probabilistic Linear Discriminant Analysis (PLDA)**.

Evoluția arhitecturilor Deep NN

A. d-vectors (Prima generație)

- Introduși de Google în 2014, d-vectorii utilizează rețele DNN simple (Feed-forward) antrenate la nivel de cadru (frame-level). Reprezentarea finală este media ultimului strat ascuns pentru toate cadrele unui enunț.

B. x-vectors (Standardul industrial)

- Dezvoltați la Universitatea Johns Hopkins (Snyder et al., 2018), **x-vectors** utilizează o arhitectură **Time-Delay Neural Network (TDNN)**. Aceștia pot captura contextul temporal pe termen lung mai eficient decât DNN-urile standard. Stratul de *Statistics Pooling* colectează media și varianța, oferind informații despre variația vocii.

C. ECAPA-TDNN (State-of-the-art)- Emphasized Channel Attention, Propagation, and Aggregation TDNN

este în prezent una dintre cele mai performante arhitecturi.

Introduce:

- **Squeeze-and-Excitation (SE) blocks:** Pentru a acorda atenție canalelor celor mai relevante.
- **Multi-scale feature aggregation:** Combină caracteristici din straturi diferite pentru o rezoluție mai bună.

D. ResNet și Transformers

- Recent, arhitecturile de computer vision (**ResNet-34**, **ResNet-50**) au fost adaptate pentru RV, tratând spectrogramele ca imagini.
- De asemenea, **Conformers** (care combină CNN cu Transformers) sunt utilizați pentru a capta atât dependențele locale, cât și pe cele globale ale semnalului vocal.

Funcții de Loss /obiective de antrenare

- Succesul RV depinde critic de modul în care rețeaua învață să "împingă" vorbitorii diferiți cât mai departe în spațiul vectorial și să-i "apropie" pe cei identici.
- Triplet Loss: Antrenează rețeaua folosind un triplet (Anchor, Positive, Negative). Obiectivul este ca distanța dintre Anchor și Positive să fie mai mică decât cea dintre Anchor și Negative.
- Angular Margin Losses (AAM-Softmax / ArcFace): - Acestea sunt cele mai populare în prezent. În loc de distanța euclidiană, ele *optimizează unghiul dintre vectori pe o hipersferă*, adăugând o "margine" de penalizare pentru a face modelul mai robust.

Seturi de date și evaluare

Dataset-uri Publice

•**VoxCeleb 1 & 2:** Conține peste 1 mil. de enunțuri de la peste 6.000 de celebrități extrase din videoclipuri YouTube. Este standardul de facto pentru antrenarea sistemelor de RV
•**LibriSpeech:** Utilizat adesea pentru pre-antrenare sau sarcini hibride de vorbire/vorbitor.

Metrici de Performanță

1.Equal Error Rate (EER): Punctul în care rata de acceptare falsă (FAR) este egală cu rata de respingere falsă (FRR). Cu cât este mai mic, cu atât sistemul este mai bun.
2.Minimum Detection Cost Function (minDCF): O metrică ce ponderază diferit erorile de tip fals pozitiv și fals negativ în funcție de aplicația finală.

Provocări și Direcții Viitoare

- **Robustețea la zgomot:** Vocea în medii zgomotoase (restaurante, trafic) rămâne o problemă. Se utilizează tehnici de *data augmentation* (adăugarea de zgomot artificial, reverberație).

- **Anti-Spoofing (Biometrie):** Detectarea atacurilor prin prezentare (înregistrări redate, sinteză vocală sau Deepfakes). Competiția **ASVspoof** este lider în această direcție.

- **Diarizarea Vorbitorului:** Procesul de a segmenta un fișier audio în funcție de "cine când vorbește" (Who spoke when), esențial în transcrierea ședințelor.

Comparații Modele

Model	EER (%)	Complexitate	Robust la Zgomot
GMM-UBM	12	Medie	Slab
i-vector + PLDA	4	Înaltă	Bun
x-vector (DNN)	2.5	Înaltă	Excelent

BAZE DE DATE VOCALE PENTRU RECUNOASTEREA VORBITORILOR

1. **NYNEX** (base de données de ligne terrestre)

Ex . Dire : 355-087-3567 (3x)

333-444-5678 (3X)

446-586-7632 (3X)

Carl vit dans une chambre belle.(1X)

2. **NYNEX cellulaire**

- nom de famille (X5) de locuteur
- chiffres isolés
- Commandes (dial , efface ..)
- mot de passe vocal (3x 10 chiffres)

3. **King92 -ITT** - 51 locuteurs masculins (téléphone / microphone)

4. **YOHO - ITT** - authentification du locuteur dépendant du texte (186 locuteurs 156M + 30F)

5. **Switchboard** - 2340 conversations téléphoniques TI ~ 6min. - 26CD

6. **SPIDE** => Switchboard - trois dispositifs différents -2 CD

Aplicatii ale recunoasterii vorbitorului

Speaker Verification

Aplicatii:

- Controlul accesului
- Verificarea identității prin telefon
- Bancomat automat
- Resetarea parolei
- Servicii bancare: identitate pentru conturi noi etc.
- Protecție împotriva furtului (mașini...)

Speaker Identification

Aplicatii:

- Criminalistică
- Munca in politie
- Setări automate de utilizator
- Clasificarea vorbitorului
- Publicitate

Referințe

1.Snyder, D., et al. (2018). "X-vectors: Robust DNN Embeddings for Speaker Recognition." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

2.Desplanques, B., et al. (2020). "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification." *Interspeech 2020*.

3.Chung, J. S., et al. (2018). "VoxCeleb2: Deep Speaker Recognition." *Interspeech 2018*.

4.Variani, E., et al. (2014). "Deep neural networks for small footprint text-dependent speaker verification." *ICASSP 2014* (Originile d-vector).

5.Deng, J., et al. (2019). "ArcFace: Additive Angular Margin Loss for Deep Face Recognition." (Aplicat ulterior cu succes masiv în SR).

Source	Org	Features	Method	Input	Text	Pop	Error
Atal 1974 [1]	AT&T	Cepstrum	Pattern Match	Lab	Dependent	10	i: 2%@0.5s v: 2%@1s
Markel and Davis 1979 [34]	STI	LP	Long Term Statistics	Lab	Independent	17	i: 2%@39s
Furui 1981 [16]	AT&T	Normalized Cepstrum	Pattern Match	Tele-phone	Dependent	10	v: 0.2%@3s
Schwartz, et al. 1982 [56]	BBN	LAR	Nonparametric pdf	Tele-phone	Independent	21	i: 2.5%@2s
Li and Wrench 1983 [31]	ITT	LP, Cepstrum	Pattern Match	Lab	Independent	11	i: 21%@3s i: 4%@10s
Doddington 1985 [12]	TI	Filter-bank	DTW	Lab	Dependent	200	v: 0.8%@6s
Soong, et al. 1985 [57]	AT&T	LP	VQ (size 64) Likelihood Ratio Distortion	Tele-phone	10 isolated digits	100	i: 5%@1.5s i: 1.5%@3.5s
Higgins and Wohlford 1986 [23]	ITT	Cepstrum	DTW Likelihood Scoring	Lab	Independent	11	v: 10%@2.5s v: 4.5%@10s
Attili, et al. 1988 [3]	RPI	Cepstrum, LP, Autocorr	Projected Long Term Statistics	Lab	Dependent	90	v: 1%@3s
Higgins, et al. 1991 [22]	ITT	LAR, LP-Cepstrum	DTW Likelihood Scoring	Office	Dependent	186	v: 1.7%@10s
Tishby 1991 [60]	AT&T	LP	HMM (AR mix)	Tele-phone	10 isolated digits	100	v: 2.8%@1.5s v: 0.8%@3.5s
Reynolds 1995 [48]; Reynolds and Carlson 1995 [49]	MIT-LL	Mel-Cepstrum	HMM (GMM)	Office	Dependent	138	i: 0.8%@10s v: 0.12%@10s
Che and Lin 1995 [9]	Rutgers	Cepstrum	HMM	Office	Dependent	138	i: 0.56%@2.5s i: 0.14%@10s v: 0.62%@2.5s
Colombi, et al. 1996 [10]	AFIT	Cep, Eng dCep, ddCep	HMM monophone	Office	Dependent	138	i: 0.22%@10s v: 0.28%@10s
Reynolds 1996 [50]	MIT-LL	Mel-Cepstrum, Mel-dCepstrum	HMM (GMM)	Tele-phone	Independent	416	v: 11%/16%@3s v: 6%/8%@10s v: 3%/5%@30s matched/mis-matched handset