Motto:

“ Each domain has its Holy Grail, and automatic speech recognition is ours “

J. Flanagan

ASRSV – curs 9

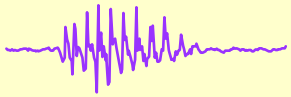
Recunoasterea Semnalului Vocal

http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/index.html

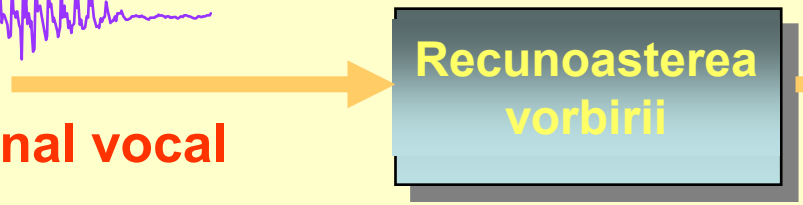
https://www.ibm.com/watson/services/speech-to-text/

https://medium.com/descript/a-brief-history-of-asr-automatic-speech-recognition-b8f338d4c0e5

Cuprins



Semnal vocal



Text

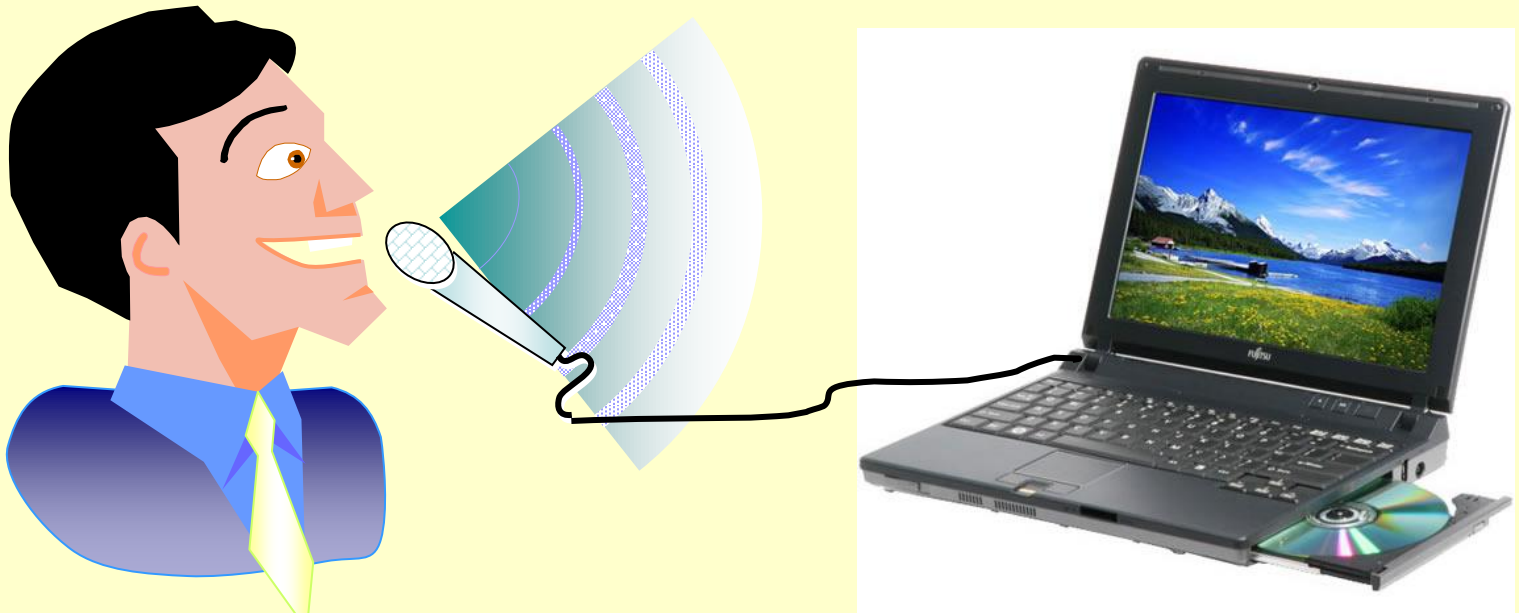
“Ce mai faci?”

1.COMUNICAREA ORALA OM-MASINA

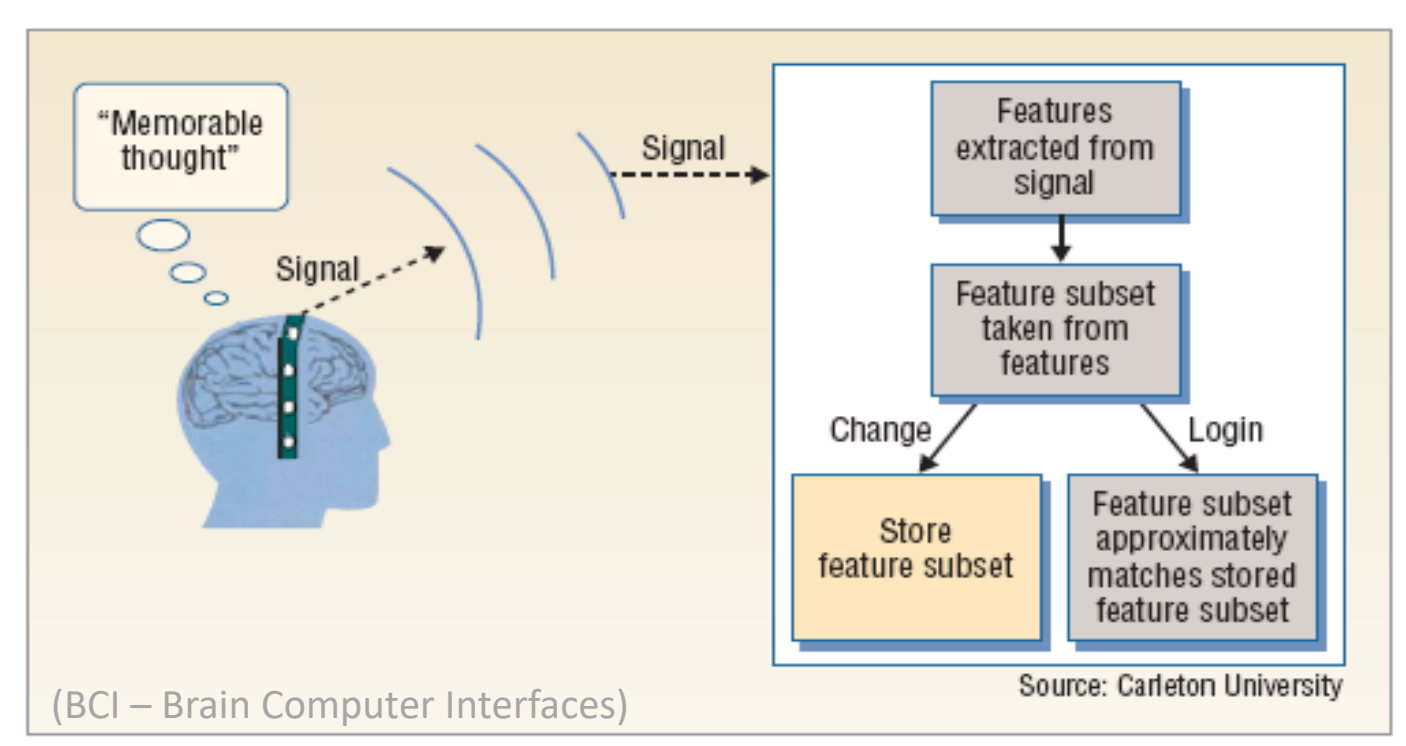
- Tehnologia care permite calculatoarelor să identifice și să proceseze vorbirea umană, transformând-o în text sau comenzi executabile (speech to text) si sa transmita mesaje vocale ca raspuns (text to speech)

Impact și Utilizare

- Accesibilitate: Pentru persoanele cu dizabilități
- Productivitate: Dictare, transcrieri automate
- Interfață naturală: Dispozitive inteligente, mașini
- Eficiență: Suport pentru clienți, analiză de interacțiuni



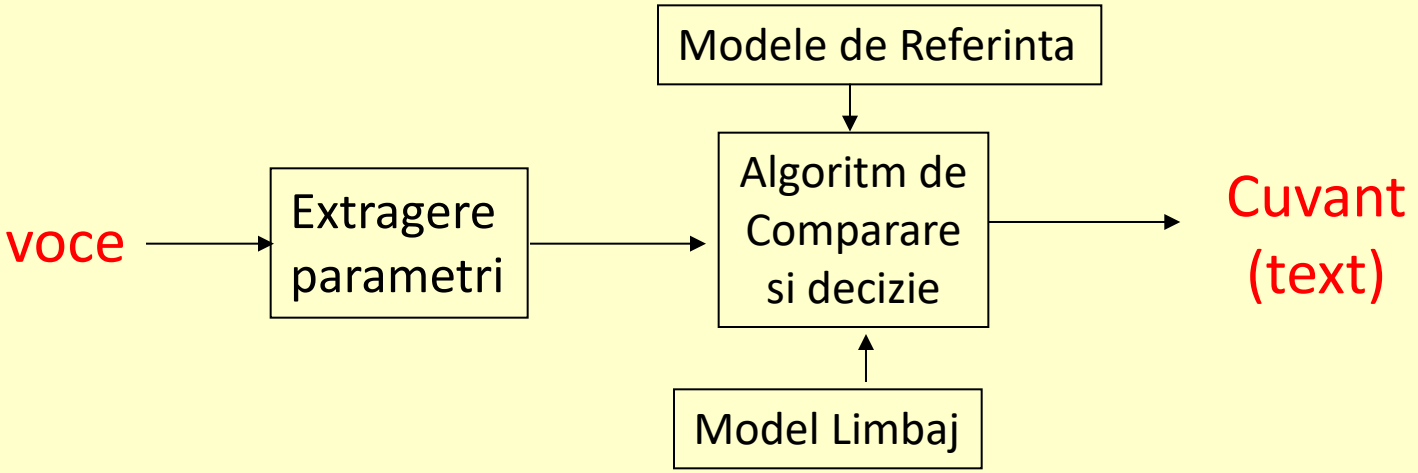
Carleton University's proposed BCI-based biometric system



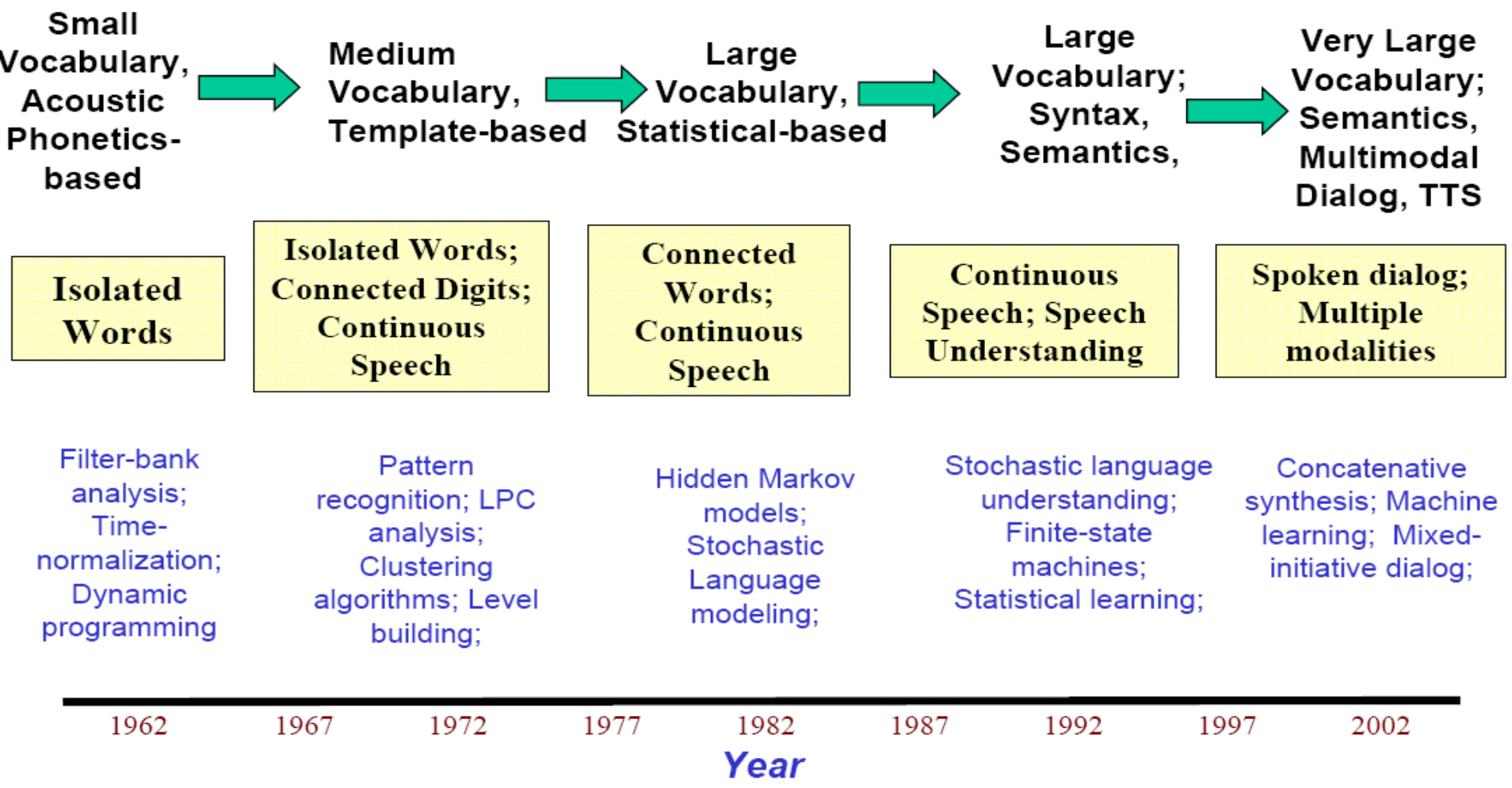
- Subjects use specific thoughts as passwords (called pass-thoughts). When someone tries to access a protected computer system or building, they think of their pass-thought. A headpiece with electrodes records the brain signals. The system extracts the signal's features for computer processing, which includes identification of the feature subset that best and most consistently represents the pass-thought.
- The biometric system then compares the subset to those recorded for authorized users.

2.Recunoasterea vorbirii

- Recunoasterea vorbirii (RV) este *baza interactiunii om-calculator* folosind vocea (speech to text)
- Un sistem ASR este o piatră de temelie a inteligenței artificiale
- Recunoasterea vorbirii poate fi in diferite contexte
  - Dependenta/ independenta de vorbitor
  - Cuvintele rostite izolat sau continuu
  - Vocabular redus sau mare
  - In mediu linistit sau zgomotos



Milestones in Speech and Multimodal Technology Research



Future of Speech Recognition Technologies

Very Large Vocabulary, Limited Tasks, *Controlled* Environment

Very Large Vocabulary, Limited Tasks, *Arbitrary* Environment

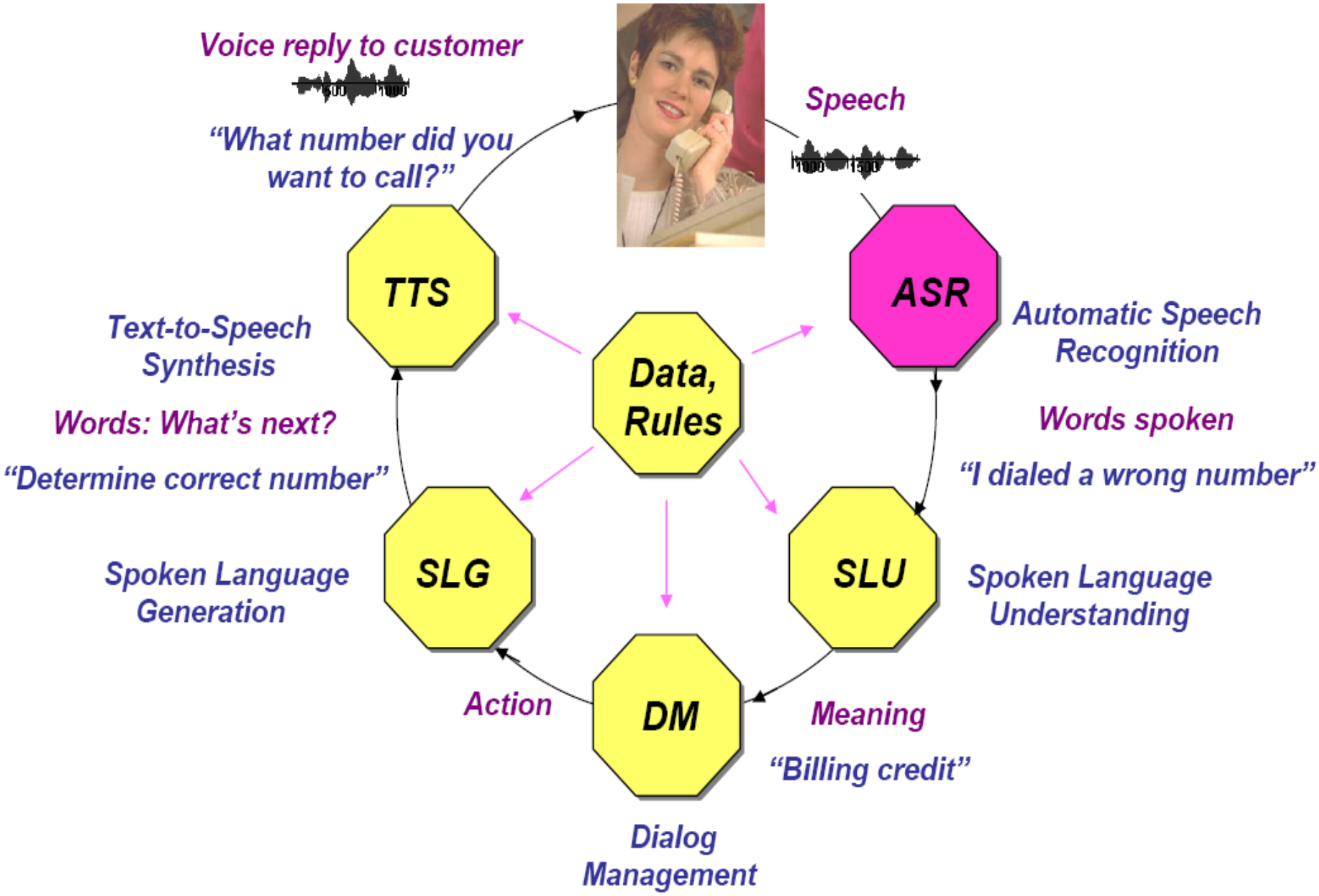
Unlimited Vocabulary, *Unlimited* Tasks, Many Languages

Dialog Systems

Robust Systems

Multilingual Systems; Multimodal Speech Enabled Devices

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


Voice reply to customer
"What number did you want to call?"
Speech
TTS
ASR
Text-to-Speech Synthesis
Automatic Speech Recognition
Data, Rules
Words: What's next?
"Determine correct number"
Words spoken
"I dialed a wrong number"
SLG
SLU
Spoken Language Generation
Spoken Language Understanding
Action
DM
Meaning
"Billing credit"
Dialog Management

1. Clasificarea sistemelor de recunoastere automata a SV

Factorii de complexitate care determina dificultatea recunoasterii pot fi clasificati astfel:

- Dependenta/independenta de vorbitor
- Cuvintele rostite izolat sau continuu
- Extensia vocabularului
- Limbajul
- Mediul ambiant

- *Dependența sau independenta de vorbitor*

Recunoasterea semnalului vocal pentru vocea unui singur vorbitor nu este o problema chiar usoara datorita variabilitatii intralocutor inerente procesului de vorbire. S-a constatat ca variabilitatea este legata de :

- variatii de debit verbal
- emotie
- stress, oboseala, raceala etc.

Recunoasterea independenta de vorbitor (multilocutor) este o problema dificila, in masura in care la **variabilitatea intralocutor** se adauga cea **interlocutor** legata de caracteristicile anatomice, a habitului lingvistic sau accentelor regionale etc.

- *Cuvintele sunt rostite izolat sau continuu*

Daca vorbitorul (vorbitorii) marcheaza cu o pauza dupa fiecare cuvant pronuntat, complexitatea problemei se reduce, pentru ca granitele intre cuvinte sunt bine definite (spre deosebire de vorbirea continua) ceea ce limiteaza mult combinatiile pentru a accede la un lexic, de exemplu.

- cuvintele pot fi considerate ca entitati globale nu ca o succesiune de entitati elementare.

- ***Dimensiunea vocabularului -*** Este mai dificil de a lucra pe un vocabular foarte extins (mii, zeci de mii de cuvinte) decat pe un vocabular restrans (cateva zeci de cuvinte). Dimensiunea vocabularului este un parametru insuficient, un grup de cuvinte foarte diferite unele de altele fiind mult mai usor de tratat decit cuvinte *apropiate din punct de vedere fonetic* (ex. moale, foale, oale, boale, toale).

- ***Limbajul -*** Daca la tratarea semnalului vocal se tine cont si de sintaxa limbajului folosit de locutor (si eventual si de semantica si pragmatica in cadrul unei aplicatii) va fi mult mai usor pentru un limbaj rigid, foarte constrins decat pentru unul suplu, natural.
  - Exista metode cantitative de studiu al acestui tip de complexitate in legatura cu dimensiunea vocabularului folosind notiuni ca: *factor de ramificatie mediu, perplexitate, vocabular echivalent* etc.

- ***Mediul ambiant -*** Independent de cele enumerate anterior, mediul acustic si conditiile de captare a sunetului constituie un factor important: prezenta zgomotului (chiar stationar) degradeaza in general foarte mult performantele sistemelor de recunoastere. Mai mult un zgomot puternic atrage dupa sine o variabilitate sporita a locutorului (efect Lombard).

MEDIU	Zgomot de reverberatie corelat cu vorbirea Zgomot de reflexie necorelat Zgomot aditiv (stationar/nestationar)
VORBITOR	Atributele vorbitorului:- dialect, sex, vârsta Modul de vorbire : - zgomot de respiratie si buze - stress - efect LOMBARD - debit verbal - nivel - frecventa fundamentala - cooperativitate
ECHIPAMENT DE INTRARE	Microfon (transmitator) Distanta de microfon Filtre Sistemul de transmisie (distorsiuni, zgomot, ecou) Echipament de inregistrare

Sistemele de recunoaştere se pot clasifica in ordinea dificultatii crescatoare astfel :

Sistemele de recunoaştere se pot clasifica in ordinea dificultatii crescatoare astfel :

- **Recunoasterea cuvintelor izolate apartinind unui vocabular limitat (<100 cuvinte) si sistemul adaptat la un vorbitor dat (sistem monolocutor)**
- **Recunoasterea cuvintelor izolate apartinind unui vocabular extins (mii de cuvinte) si sistemul adaptat la un vorbitor dat (sistem monolocutor)**
- **Recunoasterea cuvintelor izolate sistemul fiind independent de vorbitor (sistem multilocutor)**
- **Recunoasterea cuvintelor inlantuite dintr-un vocabular limitat (ex. cele 10 cifre) , dar pronuntate in orice ordine fara pauza intre ele.**
- **Recunoasterea de fraze scurte bazate pe un vocabular limitat, monolocutor.**
- **Recunoasterea de fraze scurte bazate pe un vocabular limitat , multilocutor.**
- **Recunoasterea vorbirii continue mono/multilocutor**
- **Intelegerea vorbirii este o etapa suplimentara care cere sistemului o actiune adecvata continutului mesajului vocal**

O problema foarte apropiata de recunoasterea vorbirii este ***recunoaşterea locutorului***.

Sisteme de recunoaşterea vorbirii

Recunoaşterea şi înţelegerea

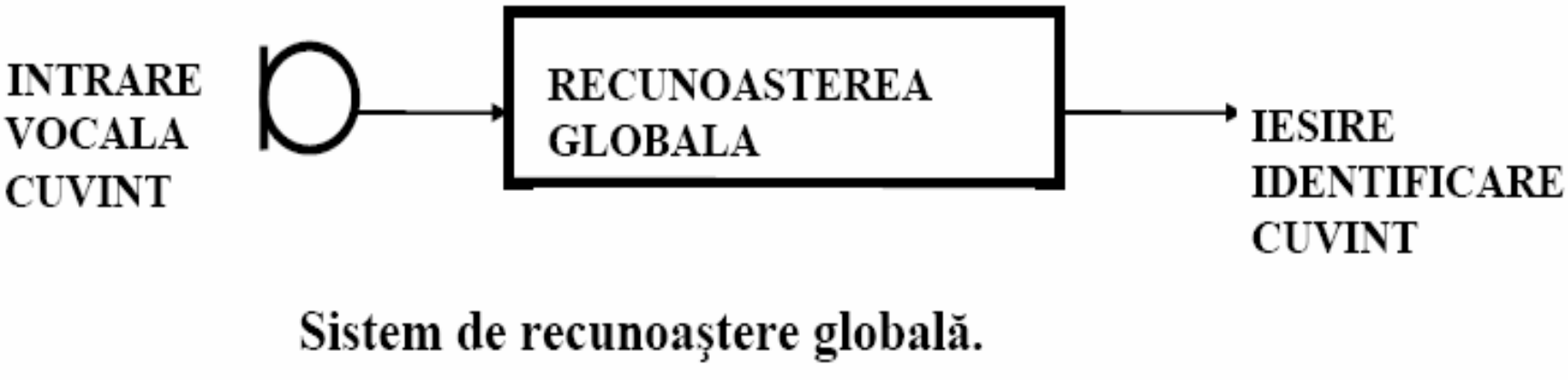
- Sunt două obiective diferite care se pot ataşa unui sistem. Recunoaşterea în sens strict conduce la o aplicatie de tip dictare automata (masina de scris fonetica), in timp ce *intelegerea cauta sa ajunga la o semnificatie* a enuntului vorbit (ca de ex. in cazul dialogului om-masina). Aceasta distinctie nu se remarca decat in cazul sistemelor destul de evoluate. Pentru un limbaj de comanda simplu, prin cuvinte izolate sau intrari de date recunoasterea = intelegere.

Sisteme auto-organizate şi bazate pe cunoştinţe

- Un sistem poate folosi metode matematice pentru a extrage date reprezentative (esantioane de invatare) structuri si parametri pe care ii va folosi pentru efectuarea recunoasterii. Este altfel spus *"auto-organizat ",* dar aceasta nu inseamna ca este total ignorant la cunostintele umane (care au stat la baza definirii unei metode matematice pe un tip specific de date), ci le foloseste implicit. (DTW, HMM)
- O alta abordare *utilizeaza explicit in sistemele de cunostinte experti umani* care au studiat vorbirea la mai multe nivele: acustic, fonetic, lingvistic *- sisteme bazate pe cunostinte.* Astfel de sisteme sunt concepute sa realizeze *prelucrari de tipul inteligenta artificiala* cooperand cu sistemele clasice de tratare a semnalului si de recunoastere a formelor.

Recunoaşterea globală şi recunoaşterea analitică

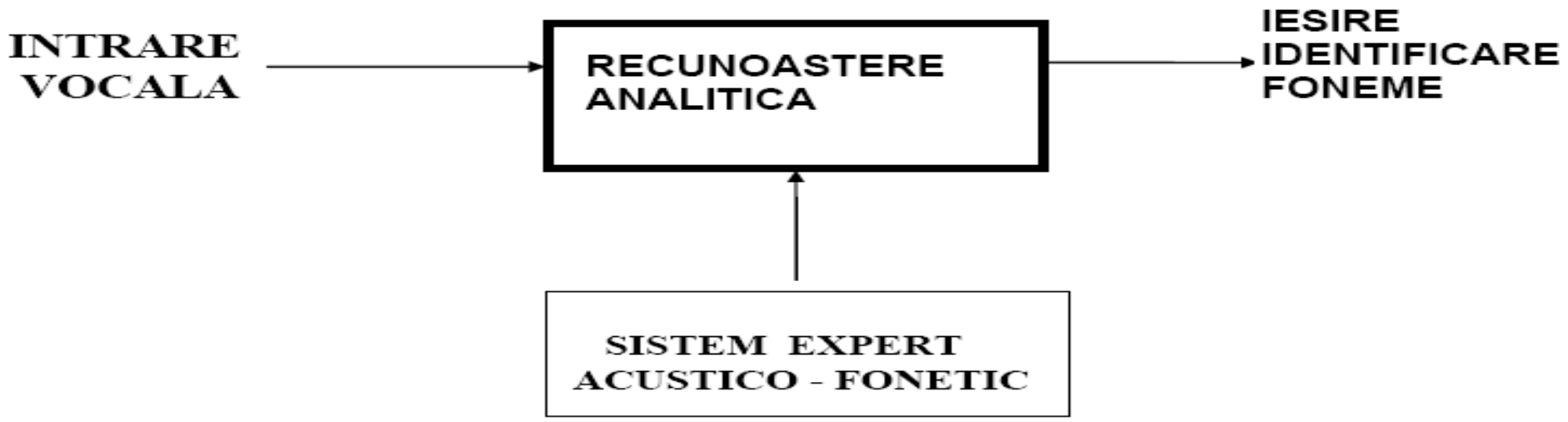
- **Recunoaşterea globală** consta in compararea entitatilor de tip cuvint (de identificat) intre ele fara a le descompune in alte subentitati, in acest caz avem recunoasterea globala.



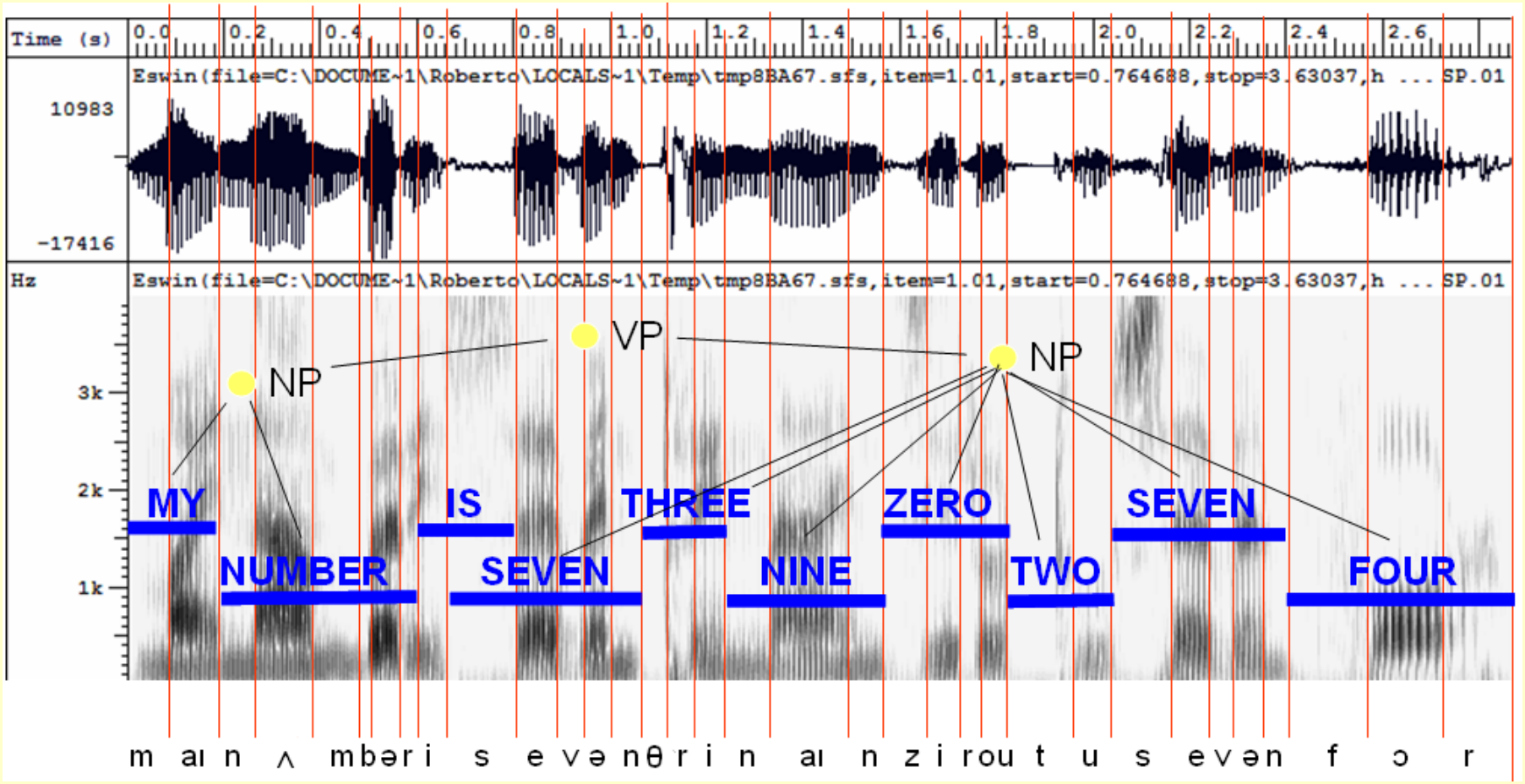
Sistem de recunoaştere globală.

• **Recunosterea analitica** consta din metode care folosesc intr-un *stadiu intermediar entitati elementare* (foneme, difoni, demisilabe) care permit descrierea entitatilor de nivel superior

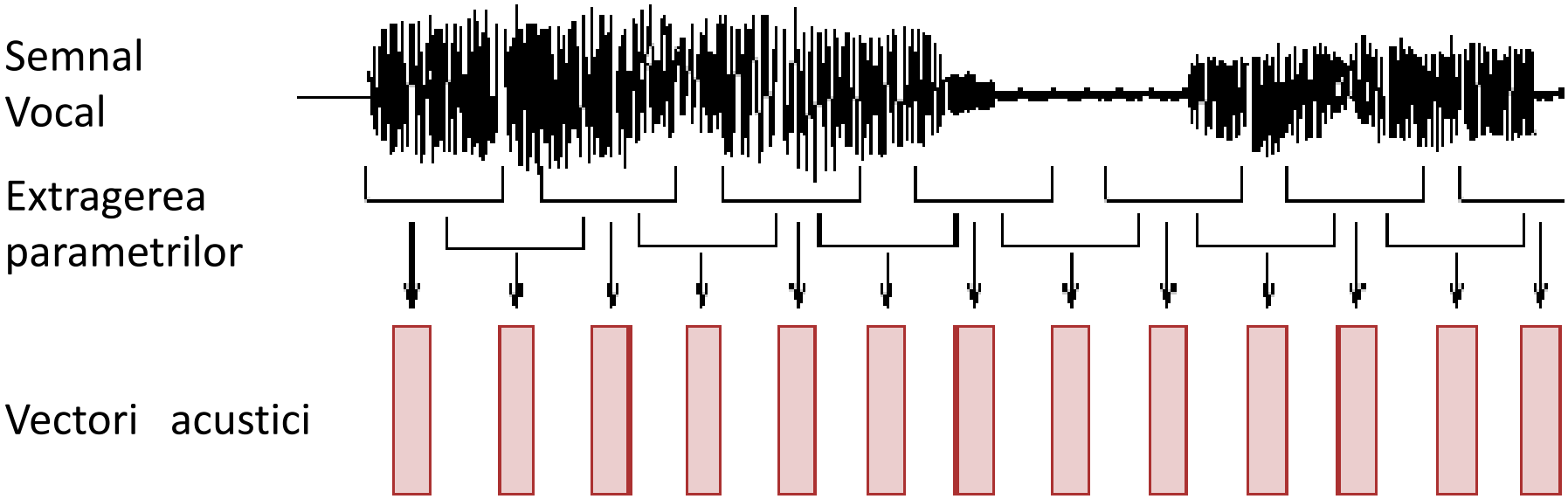
- anumite entitati elementare ale recunoasterii analitice nu au nevoie de o definitie fonetica, ele putand fi definite “ad-hoc“ (definite de exemplu prin cuantizarea vectoriala).



Sistem de recunoaştere analitică.



- Identificarea fonemelor individuale
- Identificarea cuvintelor
- Identificarea structurii și/sau semnificației propozițiilor
- Interpretarea caracteristicilor prozodice (tonul, intensitatea, durata)



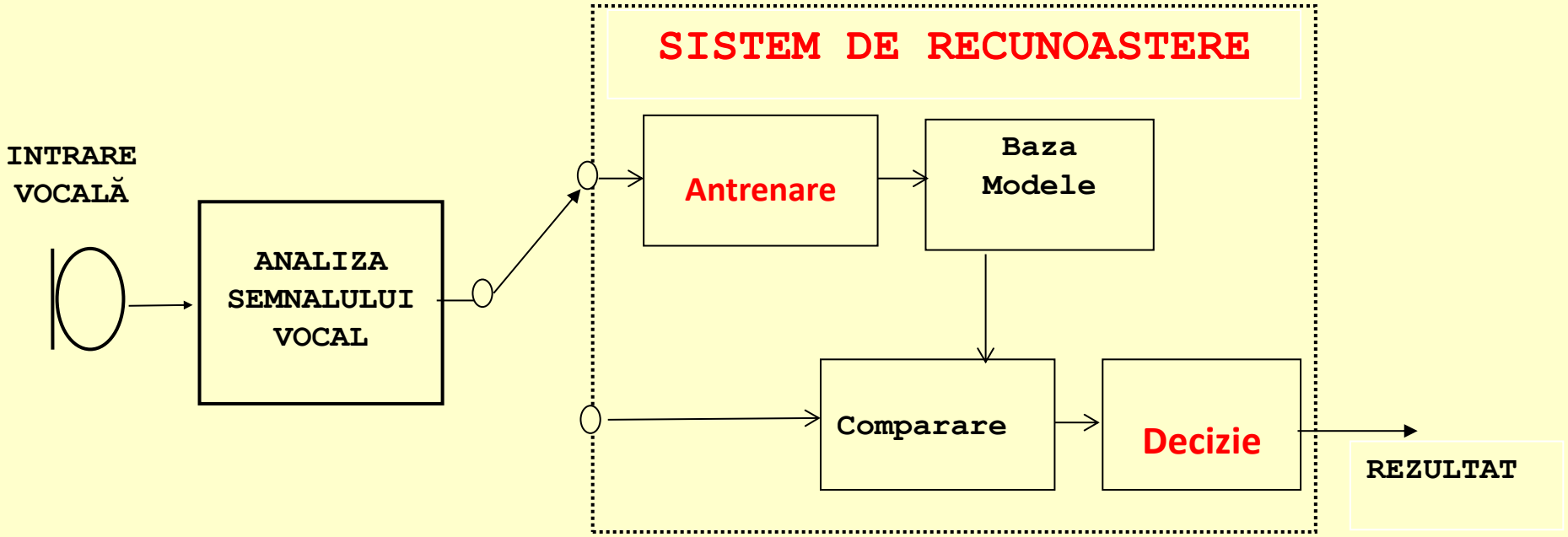
- ***Nivelul de baza pentru recunoasterea vorbirii este nivelul acustic***

- parametrii care caracterizeaza un spectru vocal se numeste ***vector acustic/spectral***

- se fixeza o metrica in spatiul marginit de un anumit tip de parametri

- distantele intre vectorii spectrali sunt mai mult sau mai putin eficiente in functie de capacitatea lor de a caracteriza spatiul acustic si de volumul de calcul care il implica estimarea lor.

- ptr. reducerea numarului de distante de calculate, adesea se imparte spatiul vectorilor acustici in clase, fiecare clasa fiind reprezentata de un vector particular numit centroid (cuantizarea vectoriala).



- antrenarea – se achiziţionează cunoştinţele şi le organizează în clase sau modele

- recunoasterea

• *Rostire necunoscuta:* ***X={x1,x2,..xl,...,xN},*** *unde* $xl=[xl1,xl2,\ldots, xld\ ]^T$

ROSTIRI CUNOSCUTE (Baza modele)

$\boldsymbol{Y1}=\{y11,y12,\ldots,y1J\}$ (1)

$\boldsymbol{Y2}=\{y21,y22,\ldots,y2J\}$ (2)

...

$\boldsymbol{YK}=\{yk1,yk2,\ldots,ykJ\}$ (M)

• Calculez $D(\boldsymbol{X},\boldsymbol{Yk})$ ptr. $k=1,\ldots M$

• *Recunoscut:* $\boldsymbol{X}=\boldsymbol{Y}_{best}$, cu $D(\boldsymbol{X},\boldsymbol{Y}_{best})\leq D(\boldsymbol{X},\boldsymbol{Y}_k)$ ptr. $k=1,\ldots M$

• OK - ptr. recunoasterea cuvintelor izolate dependente de vorbitor

- Sunt mai multe *abordări clasice* ale clasificării/deciziei în recunoastere, astfel:
  - clasificarea automată
  - discriminarea funcţională
  - metodele statistice bayesiene
  - metoda proximităţii maxime a celor k-vecini (k-NN)
  - metode stochastice (HMM)
  - metode conexioniste (reţele neuronale)
  - metode structurale.

- ***Inteligența artificială*** permite învățarea automată a relației complexe între semnalele audio și structurile lingvistice fără intervenție umană detaliată.

Principalele modele AI moderne includ:

- DNN (Deep Neural Networks) - utilizate inițial pentru clasificarea fonemelor.

- CNN (Convolutional Neural Networks) – captează modele spațiale în spectrograma audio.

- RNN / LSTM / GRU – modelează dependențele temporale lungi din vorbire.

- Transformer / Attention Models (ex: Whisper, wav2vec 2.0)** – procesează întregul semnal în paralel și oferă performanțe superioare cu latență redusă.

3. Metricã în spatiul acustic

Distanţa sau mãsura deosebirii

Într-un spatiu metric n-dimensional ***distanta dintre doi vectori a si b*** trebuie sã satisfacã urmãtoarele conditii :

$d(a,b) > 0$, $d(a,b)=0$ ddacã $a=b$ (1)

$d(a,b)=d(b,a)$ - simetrie (2)

$d(a,b) < d(a,c)+d(c,b)$ - inegalitatea triunghiului (3)

Daca simetria nu este respectata se foloseste:

$dS(a,b)= 1/2[d(a,b)+d(b,a)]$ (4)

Definirea unei distante între 2 spectre trebuie :

- sã aibã semnificatie în planul acustic
- sã se preteze formalismului matematic
- sã defineascã un spatiu de parametri judicios alesi.

- Reprezentarea parametricã a caracteristicilor unui segment de vorbire poate fi privitã ca un vector N-dimensional sau ca un punct în spaţiul N-dimensional al caracteristicilor, unde N este numãrul caracteristicilor reprezentate.

Pentru un vector cu n componente norma Holder este:

$$d_p(x,y) = \left[\sum_{k=1}^{N} |x_k - y_k|^p\right]^{1/p} = \|x - y\|_p$$

Ea defineşte o distanţă şi are diferite semnificaţii pentru valorile lui p, astfel :

p=1 - suma absolută a diferentelor coordonatelor (City block, Manhattan, L_1 norm)

$$d_1(x,y) = \sum_{i=1}^{N} |x_i - y_i|$$

Un exemplu tipic al acesteia este distanţa Hamming, care este numărul de biţi care diferă între 2 vectori binari)

p=2 distanţa medie pãtraticã (euclidianã)

$$d_2(x,y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}$$

p= ∞ distanţa maximã între componente (distanţa maximului)

$$d_\infty(x,y) = \max|x_i - y_i|_{i=1,N}$$

Daca B este o matrice definitã pozitiv de ordinul k, se poate defini o distantã ponderatã astfel:

$$d_B(x,y) = (x-y)'\ B\ (x-y)$$

- Deoarece aceastã matrice B depinde de unul din vectori, aceastã definitie nu satisface proprietatea de simetrie (2).

- Nu toate distanţele folosite în recunoaşterea automatã a vorbirii sunt metrice, de exemplu *distanţa Mahalanobis* sau distanţa covarianţei ponderate, care pentru douã modele x şi y este datã de relaţia :

$$d(x,y) = (x-y)^T W^{-1} (x-y)$$

unde W^{-1} este o matrice definitã pozitiv care ponderează diferit caracteristicile individuale ale modelului depinzând de utilitatea lor în identificarea segmentelor de vorbire în spaţiul caracteristicilor.

Distanţa euclidianã rezulta ptr. W= I, în timp ce distanţa Mahalanobis face ca W sã fie matricea de autocovarianţã corespunzãtoare vectorului de referinţã.

- avantajele teoretice ale distanţei Mahalanobis din pãcate nu sunt folosite pentru cã este dificil de estimat W, dintr-un set redus de date de antrenament

- se foloseşte de obicei distanţa euclidianã sau LPC care necesitã numai N înmulţiri pentru un vector caracteristic n-dimensional faţã N^2 de pentru distanţa Mahalanobis.

- se mai folosesc distanţele Cebîşev sau “city-block “ care reduc precizia sistemului de recunoaştere în favoarea unui volum de calcul mai redus.

$$Dc(T,R) = \sum_{n=1}^{N} |T_n - R_n|$$

unde T_n ,R_n reprezintã al n-lea parametru din modelul de test şi din cel de referinţã.

O altă distanţã înruditã, cu volum de calcul redus, este următoarea :

$$Ds(T,R) = \frac{\max_n |T_n - R_n| - \min_n |T_n - R_n|}{\sum_{n=1}^{N} T_n}$$

Distanţe LPC

Douã distanţe eficiente şi asemãnãtoare sunt folosite cu coeficienţii LPC:

- distanţa Itakura-Saito (IS)
- LLR (log likelihood ratio).

Ele prezintă avantajul folosirii modelării matematice LPC pentru ponderarea efectelor diferiţilor coeficienţi fără a fi nevoie de ortogonalitate. Forma cea mai folosită a d_{IS} între două modele reprezentate prin vectorul LPC de test T şi unul de referinţă a_i (i=1,...L) este următoarea :

$$d_{IS_i} = d_{IS}(a_i, a_T) = \frac{\sigma_i^2 a_i^T v_T a_i}{\sigma_T^2 a_T^T v_T a_T} + \log\left(\frac{\sigma_T^2}{\sigma_i^2}\right) - 1$$

unde v_T este matricea de autocorelaţie a rostirii de test, σ_T ,σ_i - parametrii de câştig pentru modelele de test şi de referinţă.

Distanţa LLR este o versiune normalizată a distanţei Itakura-Saito, LLR presupune că *amplitudinea este irelevantă* pentru recunoaştere deoarece cuvântul poate fi rostit mai tare sau mai slab. :

$$d_{LLR_i} = d_{LLR}(a_i, a_T) = \log \frac{a_i^T v_T a_i}{a_T^T v_T a_T}$$

Distanţe cepstrale

•Cepstrul unui semnal este definit ca transformata Fourier a log spectrului semnalului.

•Pentru un spectru de putere X(ω) care este simetric faţã de ω=0 şi este periodic pentru o secvenţã de eşantioane, seria Fourier a log|X(ω)| se poate exprima astfel :

$$\log X(\omega) = \sum_{n=-\infty}^{\infty} c_n e^{-jn\omega}$$

Dacã semnalul vocal este modelat cu un spectru numai poli, de fazã minimã :

$$X(\omega) \rightarrow \frac{\sigma^2}{\left|A(e^{j\omega})\right|^2}$$

- cepstrul rezultat are mai multe proprietãţi interesante. Pentru un filtru stabil numai poli log A(z) se aflã în interiorul cercului unitate şi se poate reprezenta prin dezvoltare Laurent :

$$\log[\sigma/A(z)] = \log\sigma + \sum_{m=1}^{\infty} c_n z^{-n}$$

Diferenţiind cei doi membri ai ecuaţiei în raport cu 1/z, obţinem:

$$c_n = -a_n - 1/n \sum_{k=1}^{n-1} k c_k a_{n-k} \qquad , n>0$$

a_0=1 şi a_k=0 pentru k>p.

$$d_c^2(L) = \sum_{n=1}^{L} (C_n - C_n')^2$$

Sisteme şi metode utilizate în recunoaştere

•*Abordările traditionale* de recunoaştere a vorbirii şi vorbitorului în raport cu metodele şi tehnicile folosite se pot grupa în mai multe familii :

- sisteme bazate pe spectrograme
- sisteme bazate pe metodele programării dinamice (DTW)
- sisteme care folosesc cuantizarea vectorială (VQ)
- sisteme bazate pe modele Markov ascunse (HMM)
- sisteme utilizând reţelele neuronale NN
- sisteme bazate pe metoda TESPAR FANN
- sisteme folosind metode algebrice/statistice
- combinaţii ale lor.

•*Abordari IA*

1. Metode Tradiționale (Bazate pe Componente)

Înainte de dominația modelelor end-to-end, sistemele ASR erau alcătuite din mai multe componente separate, antrenate individual:

- **Modelul Acustic:** Utilizează algoritmi precum **Modelele Markov Ascunse (HMM)** sau, mai târziu, **Rețelele Neurale (DNN/RNN)** pentru a converti caracteristicile acustice ale sunetului (ex. coeficienți MFCC) în **unități fonetice** (foneme).
- **Modelul de Pronunție (Lexicon):** Conține o listă de cuvinte și modul în care acestea sunt pronunțate (secvențe de foneme).
- **Modelul de Limbă:** Utilizează **n-grame** sau alte metode statistice/neurale pentru a prezice probabilitatea ca o anumită secvență de cuvinte să apară într-o limbă. Ajută la rezolvarea ambiguității (*"casa ta e mare"* vs. *"casata Ema re"*).
- Aceste componente erau integrate prin algoritmi de decodare, cum ar fi **algoritmul Viterbi**, pentru a găsi cea mai probabilă secvență de cuvinte.

2. Metodele End-to-End

Metodele **end-to-end** reprezintă o schimbare de paradigmă. Ele antrenează un **singur model neuronal** pentru a mapa direct semnalul audio de intrare la secvența de text de ieșire, eliminând nevoia de componente separate (model acustic, lexicon și model de limbă distincte).

Avantajele principale includ:

- **Simplitate:** arhitectură unică, mai ușor de implementat și optimizat.
- **Performanță Superioară:** Modelul optimizează direct eroarea finală (de cuvânt), iar componentele învață să lucreze împreună mai eficient.

Trei abordări majore end-to-end sunt:

A. Connectionist Temporal Classification (CTC)
B. Sequence-to-Sequence (Seq2Seq) cu mecanism de atenție (arhitectura encoder-decoder)
C. Modele bazate pe Transformer

- Se bazează exclusiv pe mecanismul de atenție (self-attention) pentru a procesa datele audio și a genera textul, permițând o *paralelizare* mult mai mare și, implicit, o antrenare mai rapidă și pe seturi de date mai mari.
- Exemple notabile sunt modelele de mari dimensiuni precum **Wav2Vec 2.0** sau arhitecturi folosite în **OpenAI Whisper**.

3. Modele preantrenate (self-supervised) - tendințe actuale

- Cea mai recentă evoluție în ASR se bazează pe **învățarea auto-supervizată** (*Self-Supervised Learning* - SSL):

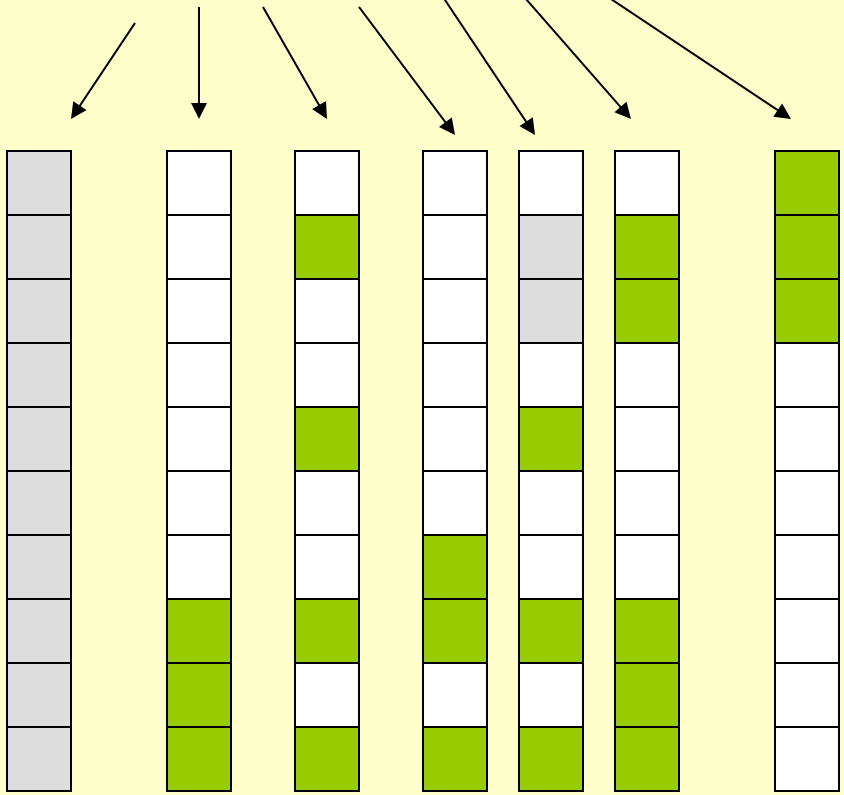
• **Pre-antrenare:** Modelele sunt antrenate pe un volum masiv de date audio *netranscrise* (fără etichete) pentru a învăța reprezentări bogate și generale ale vorbirii (ex. prezicerea unor segmente de vorbire mascate).

• **Ajustare Fină (Fine-Tuning):** Modelul pre-antrenat este apoi ajustat (antrenat în continuare) pe un set de date *transcrise* (cu etichete) mult mai mic, folosind o metodă end-to-end (precum CTC). Această abordare (*ex. Wav2Vec 2.0, HuBERT*) a îmbunătățit dramatic performanța ASR, mai ales pentru limbile cu resurse de date transcrise limitate.

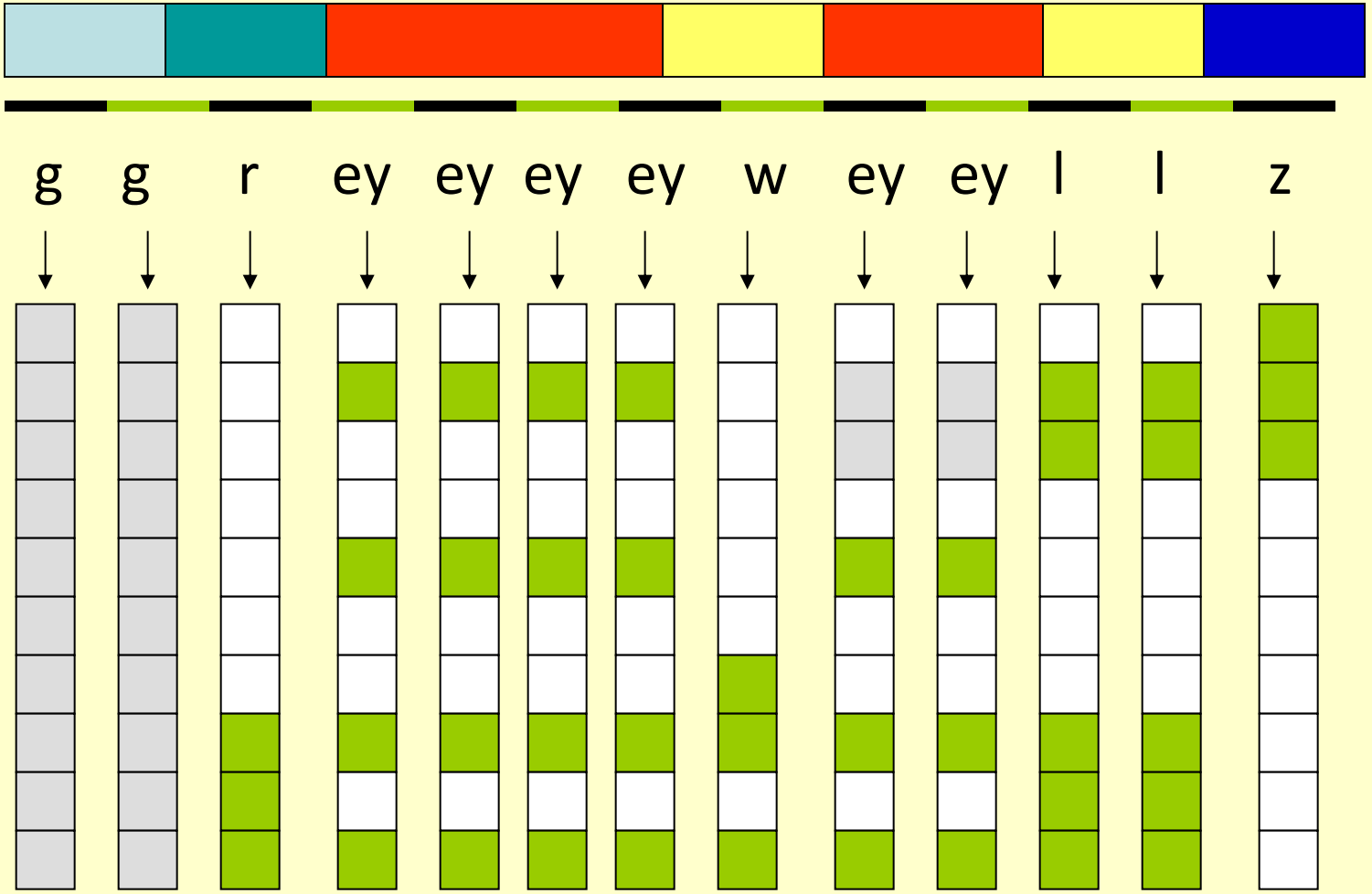
4. Metode de recunoasterea vorbirii

Cuvinte: grey whales

Foneme: g r e^{y} w e^{y} l z

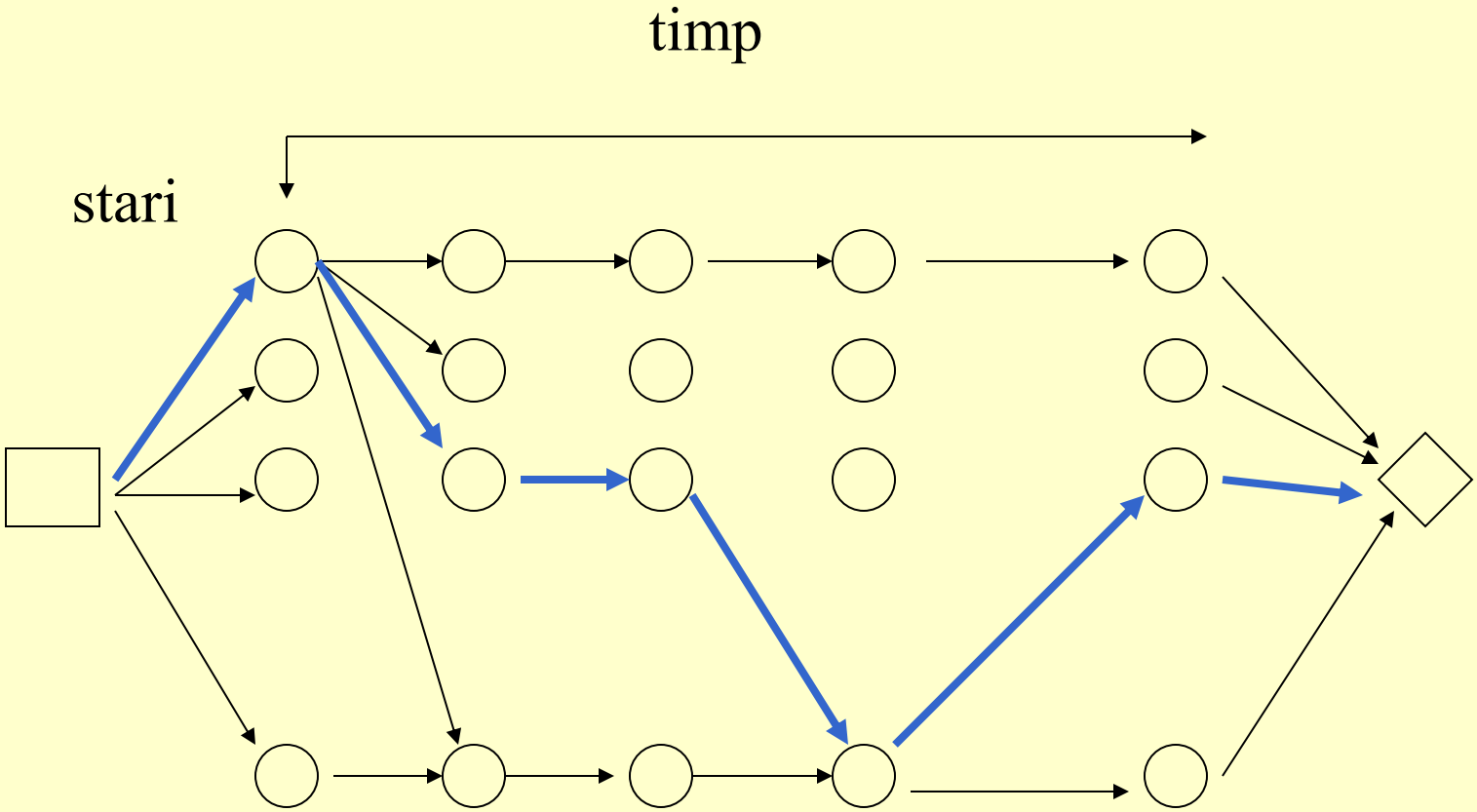


-Fiecare fonema are diferite caracteristici (de ex. Distributia de putere).



Cum se pot compara cuvintele daca exista diferente temporale sau altele?

Solutia: Programarea Dinamica



- Se ***cauta o cale*** care corespunde fonemelor cele mai probabile care au generat secventa de observatii “vectorii acustici” (extrasi din fiecare cadru de vorbire)

5. METODA ALINIERII DINAMICE (DTW)

- **Alinierea liniarã si neliniarã -** La calcularea distanţelor dintre douã rostiri se presupune cã cele douã modele au aceeaşi lungime. Variaţiile în lungime ale cuvintelor se datorează în mod special modului de rostire al vocalelor: rapid sau lent

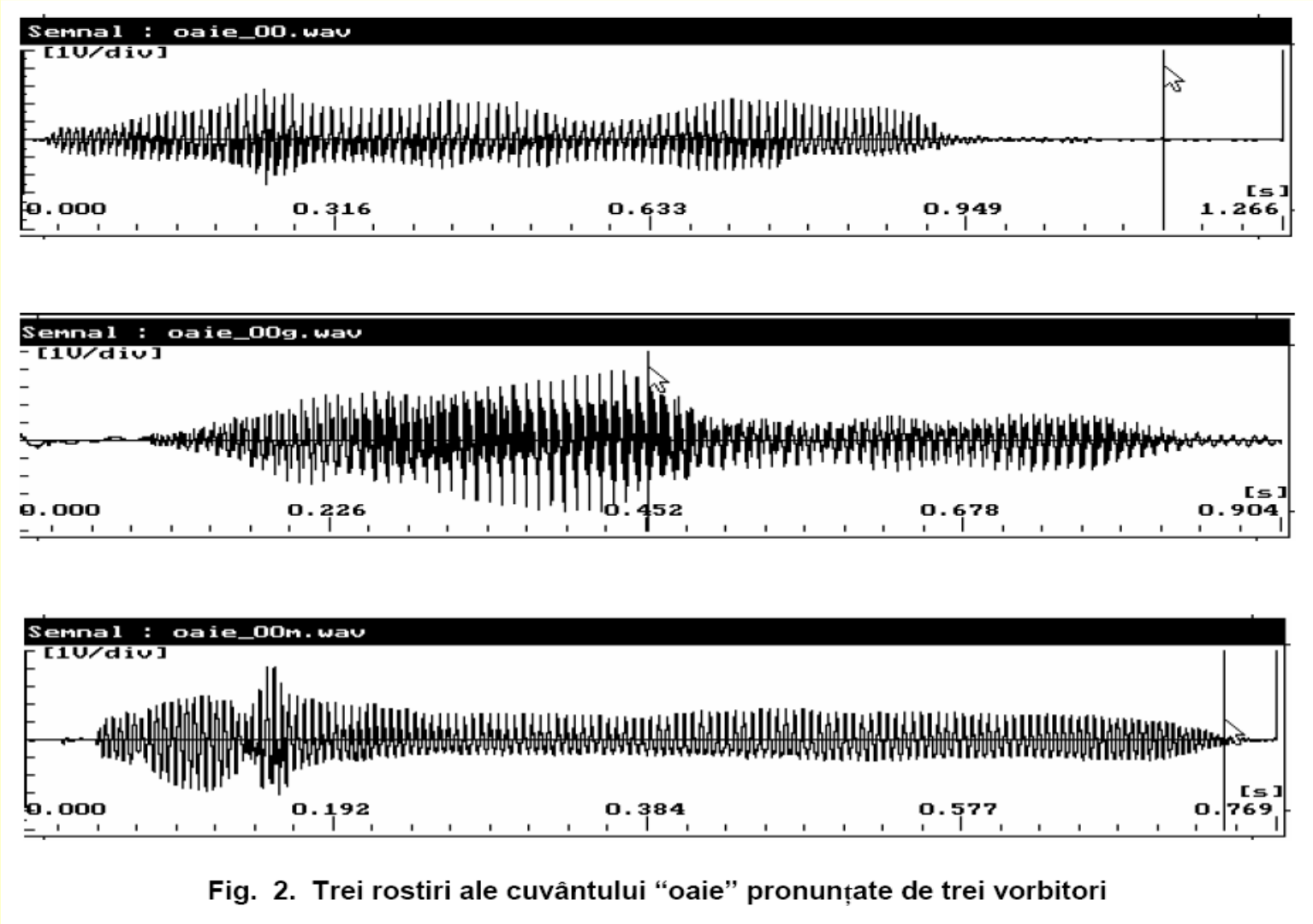


Fig. 2. Trei rostiri ale cuvântului "oaie" pronunțate de trei vorbitori

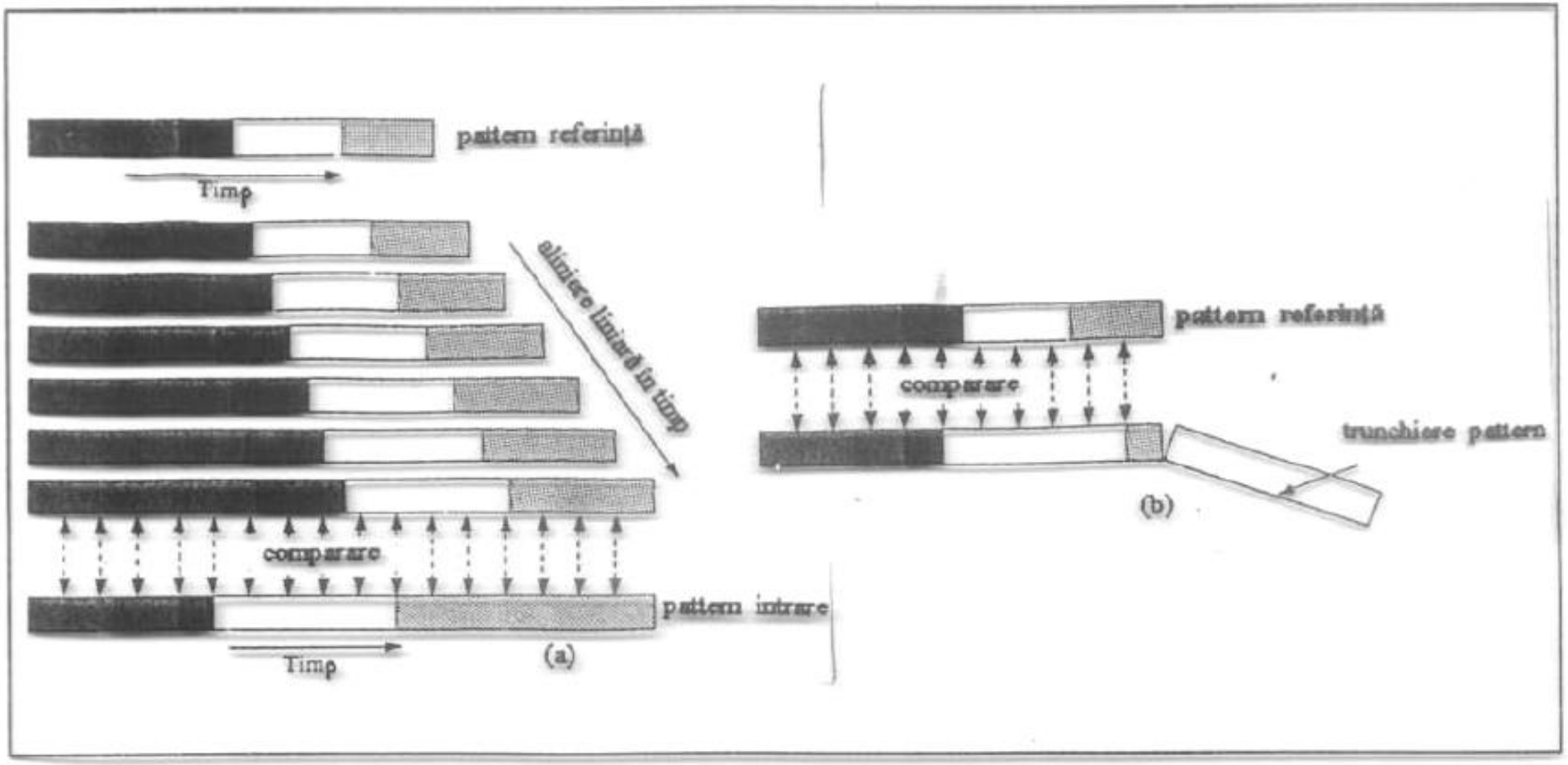


Fig. 3. Alinierea liniară în timp prin comprimare/dilatare proporțională (a) sau prin trunchiere (b)

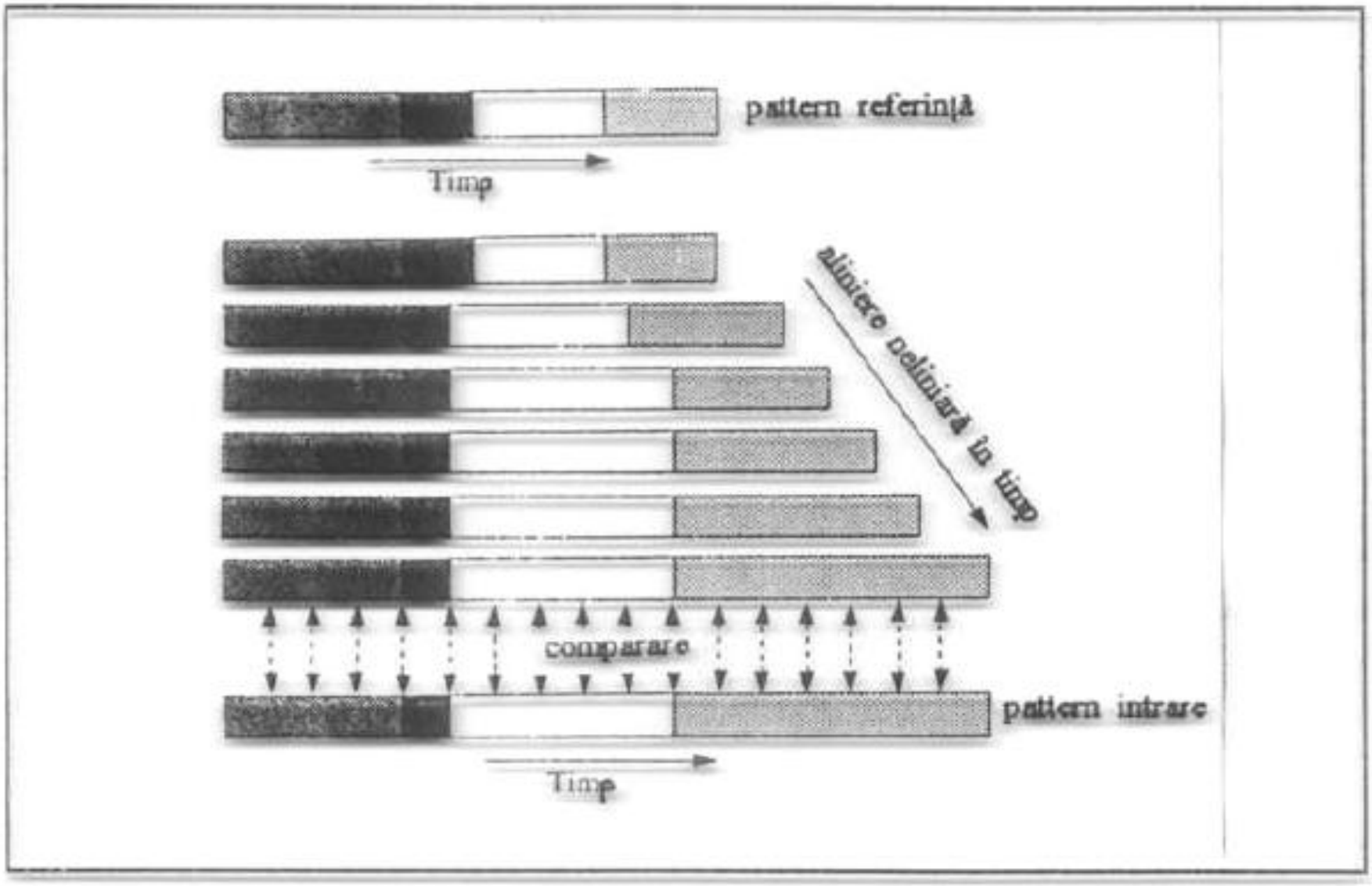


Fig. 4. Alinierea neliniarã în timp

DTW este un algoritm care ***calculează similaritatea*** dintre două secvențe de date care pot varia în timp sau viteză. În recunoașterea vorbirii, acesta compară semnalele de vorbire (de exemplu, un cuvânt pronunțat de un utilizator cu un model de referință).

De ce este necesar?

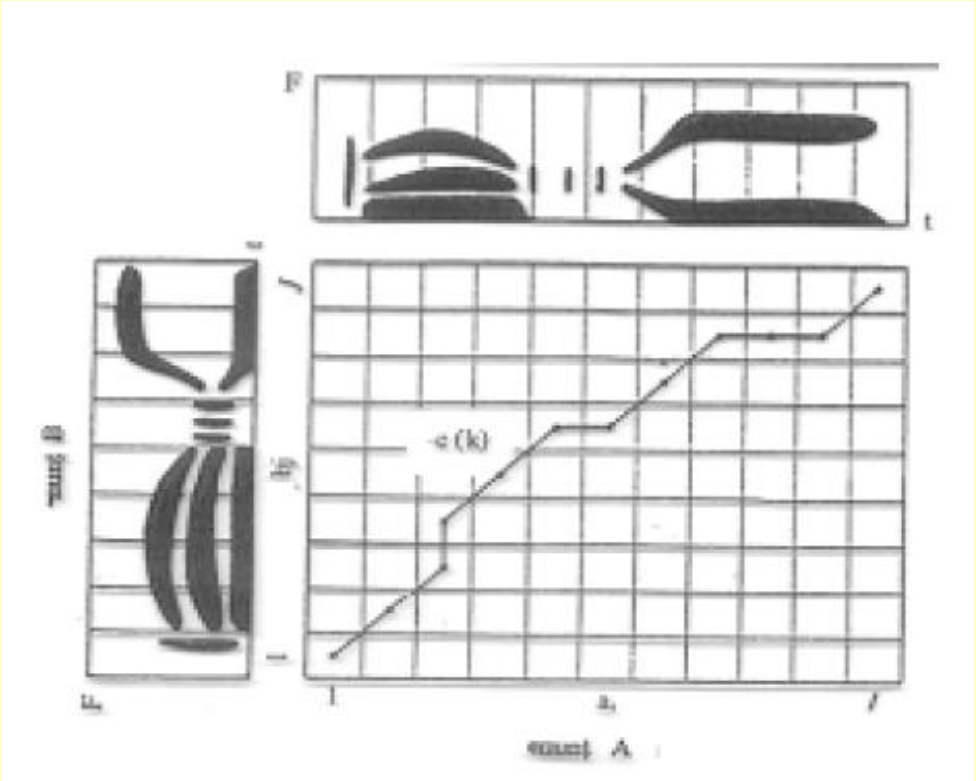
- Oamenii vorbesc cu **viteze diferite** (ex: "Hei" poate fi rostit rapid sau lent).
- Metodele simple (ex: distanța Euclidiană) eșuează dacă semnalele nu sunt sincronizate temporal.
- DTW **aliniază** nelinear cele două semnale în timp, pentru a găsi corespondența optimă.

Metoda DTW sau "potrivirea elasticã" a modelelor este o tehnica bazatã pe programarea dinamicã, des folositã la rezolvarea problemelor de cost/drum minim
- Aceastã metodã DTW, ***combinã alinierea şi calculul distanţei*** printr-o procedurã de programare dinamicã care, pentru varianta standard, presupune urmãtoarele :

- *variaţia globalã în viteza de vorbire pentru un vorbitor care rosteşte acelaşi cuvânt în diferite ocazii poate fi prelucrată prin normalizare liniarã în timp*
- *Variaţiile locale în viteza de vorbire, pentru care alinierea liniarã este inadecvatã, sunt mici şi pot fi tratate folosind ponderi de distanţã cunoscute ca şi constrângeri locale de continuitate*
- *fiecare cadru din rostirea de test contribuie în mod egal la recunoaştere*
- *se foloseşte o singurã mãsurã de distanţã aplicatã uniform pe toate cadrele*

Algoritmul de aliniere dinamică

- Algoritmul DTW comprimã sau dilatã axa timp, neliniar pentru a potrivi poziţia aceluiaşi fonem în modelul de intrare (test) şi cel de referinţã
-Acest proces poate fi realizat folosind tehnica programãrii dinamice propusã de Bellman
- Algoritmul DTW - alinierea dinamicã temporalã a vorbirii - Slutsker, Vintsyuk, Velichko şi Zaguruyko în Rusia şi în paralel de Sakoe şi Chiba în Japonia (1968).
- Aceastã tehnicã a avut impact în recunoaşterea vorbirii, fiind în prezent o metodã importantã şi des aplicatã



Prezentarea algoritmului

Se presupun două secvențe în timp, de parametrii reprezentați de vectorii :

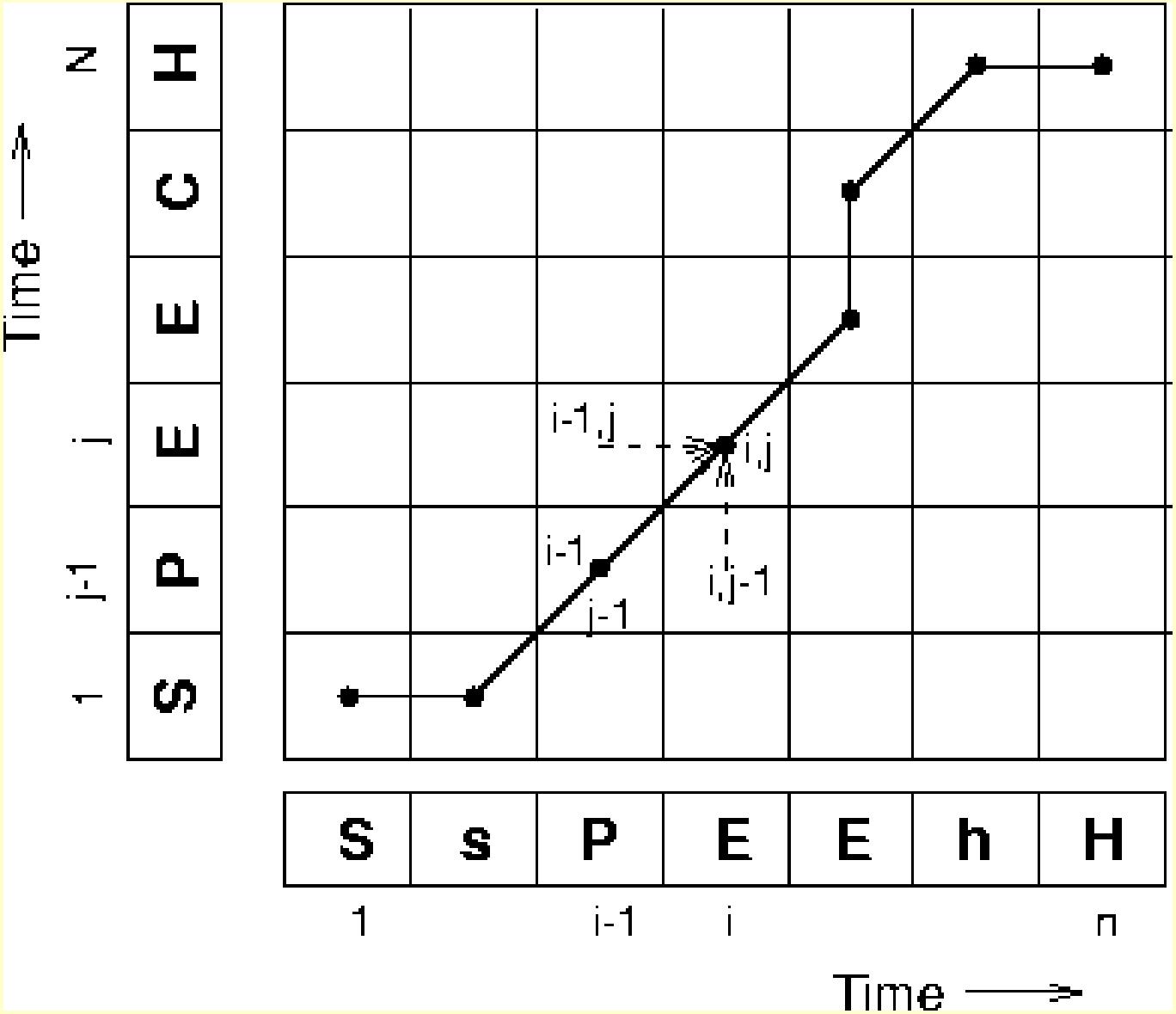
$$A= a_1, a_2,.....a_i \qquad B= b_1,b_2,.....b_j \qquad (1)$$

care trebuiesc comparați. Considerând planul împărțit de A şi B ca în figura .1., funcția de aliniere în timp indică corespondența între axele timp ale secvențelor A şi B şi poate fi reprezentată printr-o secvență de puncte (latice) în plan c =(i,j).

$$F = c_1, c_2, ...C_k.....,C_{\mathbf{K}} \quad , \quad C_{\mathbf{K}}=(i_k, j_k) \qquad (2)$$

Dacă distanța spectrală între doi vectori caracteristici a_i şi b_j se reprezintă prin: d(c) = d(i,j), atunci suma distanțelor până la sfârşitul secvenței de-a lungul lui F se poate reprezenta prin: (3)

$$D(F) = \frac{\sum_{k=1}^{K} d(c_k) W_k}{\sum_{k=1}^{K} W_k}$$


Time
N
H
C
E
E
P
S
j
j-1
1
i-1,j
i,j
i-1
i,j-1
j-1
S
s
P
E
E
h
H
1
i-1
i
n
Time

Cu cât valoarea este mai micã, cu atât potrivirea (asemãnarea) dintre A şi B este mai bunã. W_k reprezintã ponderi corelate cu F care trebuie sã respecte urmãtoarele [Fur89]:

- condiția de monotonie şi continuitate:

$$i_{k-1} \le i_k \ , \ j_{k-1} \le j_k \qquad \text{şi} \qquad 0 \le i_k - i_{k-1} \le 1, \quad 0 \le j_k - j_{k-1} \le 1 \tag{4}$$

- condiția la limite (capete):

$$i_1 = j_1 = 1, \ i_k = I, \ j_k = J \tag{5}$$

- condiția de limitare a ferestrei:

$$| i_k - j_k | \ \le r \ , \ r = \text{constant}. \tag{6}$$

care se aplicã pentru a preveni contracții/extensii exagerate.

Ecuația se simplificã, definind W_k astfel încât numitorul relației .3 sã fie constant, deci independent de drum (F). De exemplu pentru:

$$W_k = (i_k - i_{k-1}) + (j_k - j_{k-1}), \ i_0 = j_0 = 0 \tag{7}$$

W_k devine distanța 'city-block' (3.9.13) şi

$$\sum_{k=1}^{k} W_k = I + J \tag{8}$$

iar relația (3) devine :

$$D(F) = \frac{1}{I+J} \sum_{k=1}^{K} d(c_k) . W_k \tag{9}$$

Obiectivul minimizării funcției F poate fi realizat fără a examina toate posibilitățile funcției F, ținând cont de caracterul aditiv:

$$g(ck)= g(i,j)= \min_{c1,\dots Ck-1} \sum_{m=1}^{K} d(c_m)w_m = \min_{c1,\dots Ck-1} [\sum_{m=1}^{K-1} d(c_m)w_m + d(c_k)w_k] =$$

$$= \min_{Ck-1} [\min_{c1,\dots Ck-2} [\sum_{m=1}^{K-1} d(c_m)w_m] + d(c_k)w_k] =$$

$$= \min_{Ck-1} [g(c_{k-1}) + d(c_k)w_k] \qquad (10)$$

Folosind condiționările (.4), (.5) şi (.6) pentru F şi formularea de mai sus pentru W_k, relația (.10) se poate rescrie:

$$g(i,j) = \min \begin{pmatrix} g(i, j-1) + d(i, j) \\ g(i-1, j-1) + 2d(i, j) \\ g(i-1, j) + d(i, j) \end{pmatrix} \qquad (11)$$

Fie condițiile inițiale:

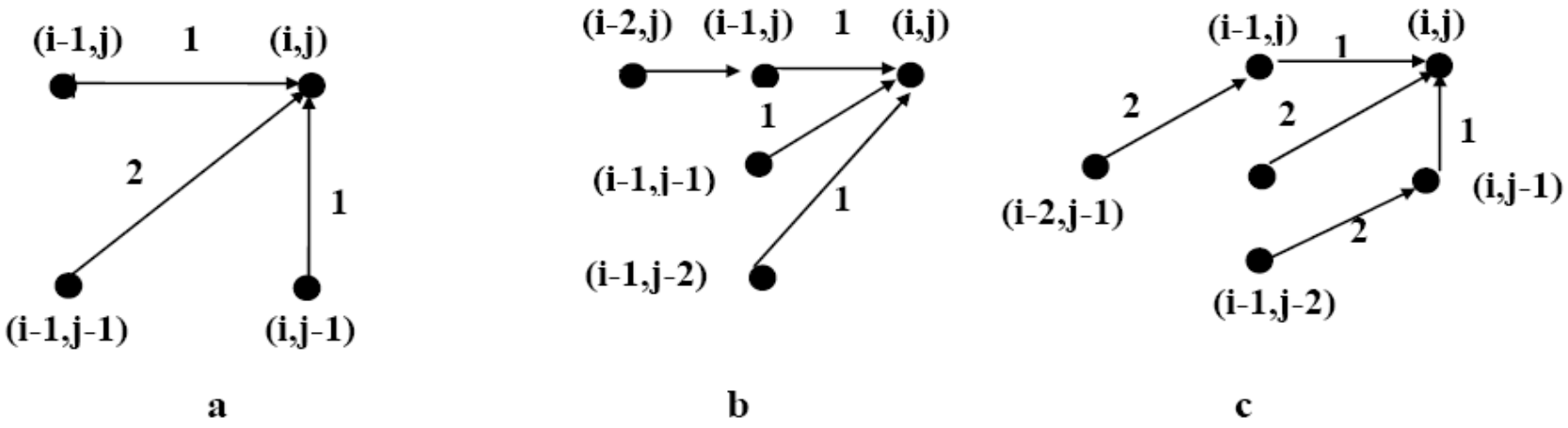
$$g(1,1)=2*d(1,1) \text{ şi } j=1 \qquad (12)$$

Plecând de la această distanță locală se calculează distanța între cele două secvențe A şi B cu ajutorul relației (.11), variind I între limitele ferestrei definite. Calculele se repetă incrementând j până j=J, iar distanța globală se obține ca:

$$D(F)=D(A,B)=g(I,J)/(I+J) \qquad (13)$$

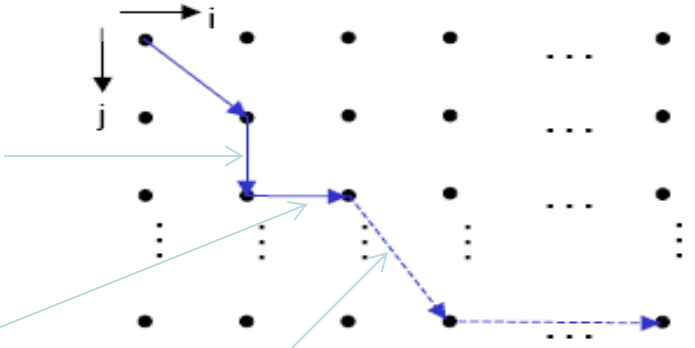
Variante ale comparației prin programare dinamică

Pentru evaluarea funcției de aliniere şi calculul distanței D(F) (.3) trebuie definiți funcția de ponderare W_k şi coeficientul de normalizare $\sum W_k$. Sunt mai multe tipuri de funcții de ponderare propuse (Sakoe, Myers) care rezultă din experimentele de recunoaştere şi care reprezintă constrângeri locale ce se impun (figura .2) .



Constrângeri locale

Weighted Directed Path: Arcs



Insertion : $\{G[i,j], G[i,j+1]\} \rightarrow w(0,1)$

Deletion : $\{G[i,j], G[i+1,j]\} \rightarrow w(1,0)$

Change or Substitution : $\{G[i,j], G[i+1,j+1]\} \rightarrow w(1,1)$

```
# define Mx   100000
        g[0][0]=0
         for (j=1; j≤ J; j ++)   g[0][j]=Mx ;
         for (I=1 ; i≤ I; i  ++)
        {
                        g [i][0]=Mx ;
                        for (j=1; j≤ J; j++)
                    {
                            d= dist (i,j) ;
                            g[i][j]=min ( g[i-1][j]+d ,
                                                g[i-1][j-1]+d+d,
                                                  g[i][j-1]+d ) ;
                    }
         }

        D=g[I][J]/I+J;
```

unde dist(i,j) este o funcţie care returnează distanţa spectralã între evenimentele:
i a rostirii A şi j din rostirea B.

Volumul de calcul cerut de algoritmul DTW

- algoritmul trebuie sã calculeze o **distanţã localã** d(i,j) şi una **cumulatã (parţialã)** g(i,j).
- pentru a compara douã cuvinte izolate de lungime I şi J se tratează toate punctele (i,j) ale reţelei de puncte [1,I] x [1,J], la recunoaşterea unui cuvânt izolat este nevoie să se analizeze MxQxL puncte, unde
  - M reprezintã dimensiunea vocabularului,
  - Q este numãrul mediu de referinţe/cuvânt ,
  - L este lungimea medie a cuvintelor.

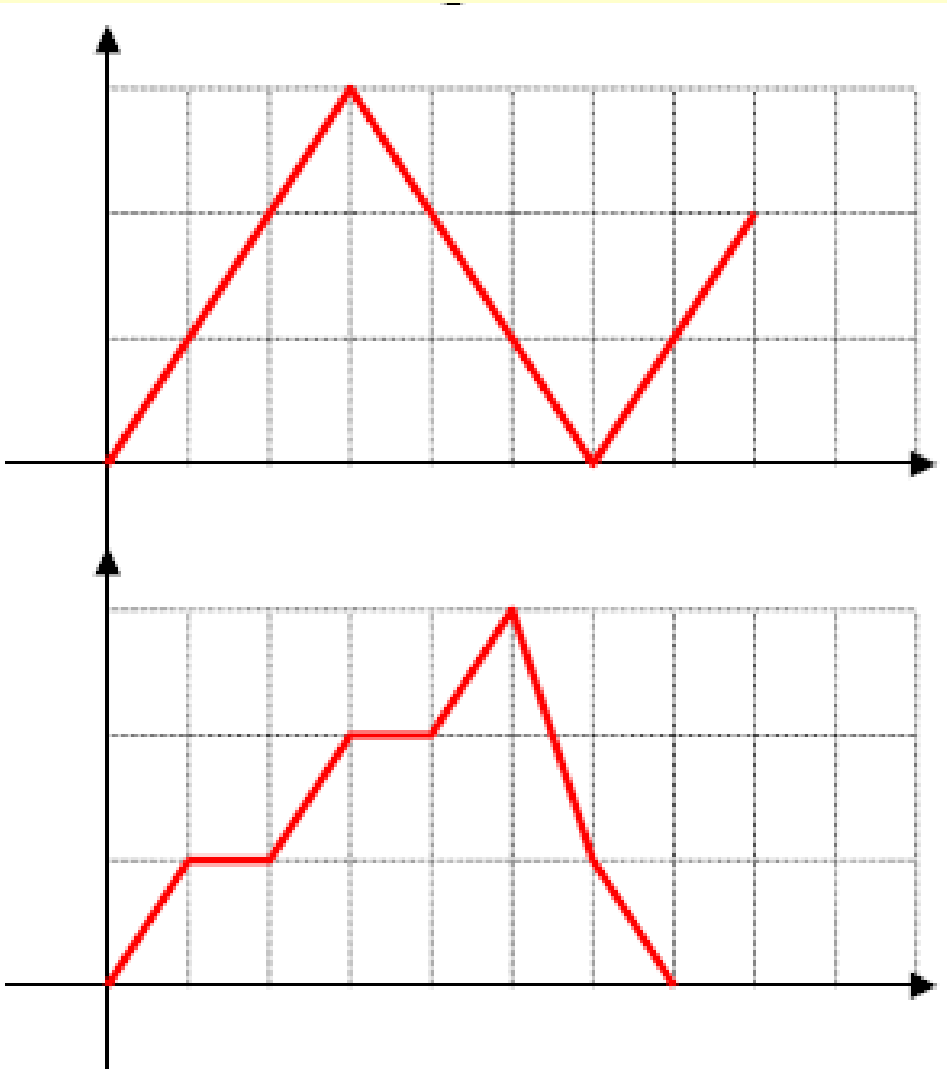
In general, numãrul real de puncte evaluate este mult mai mic din douã motive:
- din domeniul [1,I]x[1,J] interesează doar o zonã în jurul diagonalei, suprafaţa analizatã reducându-se cu un factor de 0,4 ...0,1
- compararea unui cuvânt de recunoscut cu o referinţã poate fi abandonatã dacã: indicele de distanţã D(T,R) nu se poate situa sub un prag dat.

Tema. Cautati variante modificate ale algoritmului DTW.

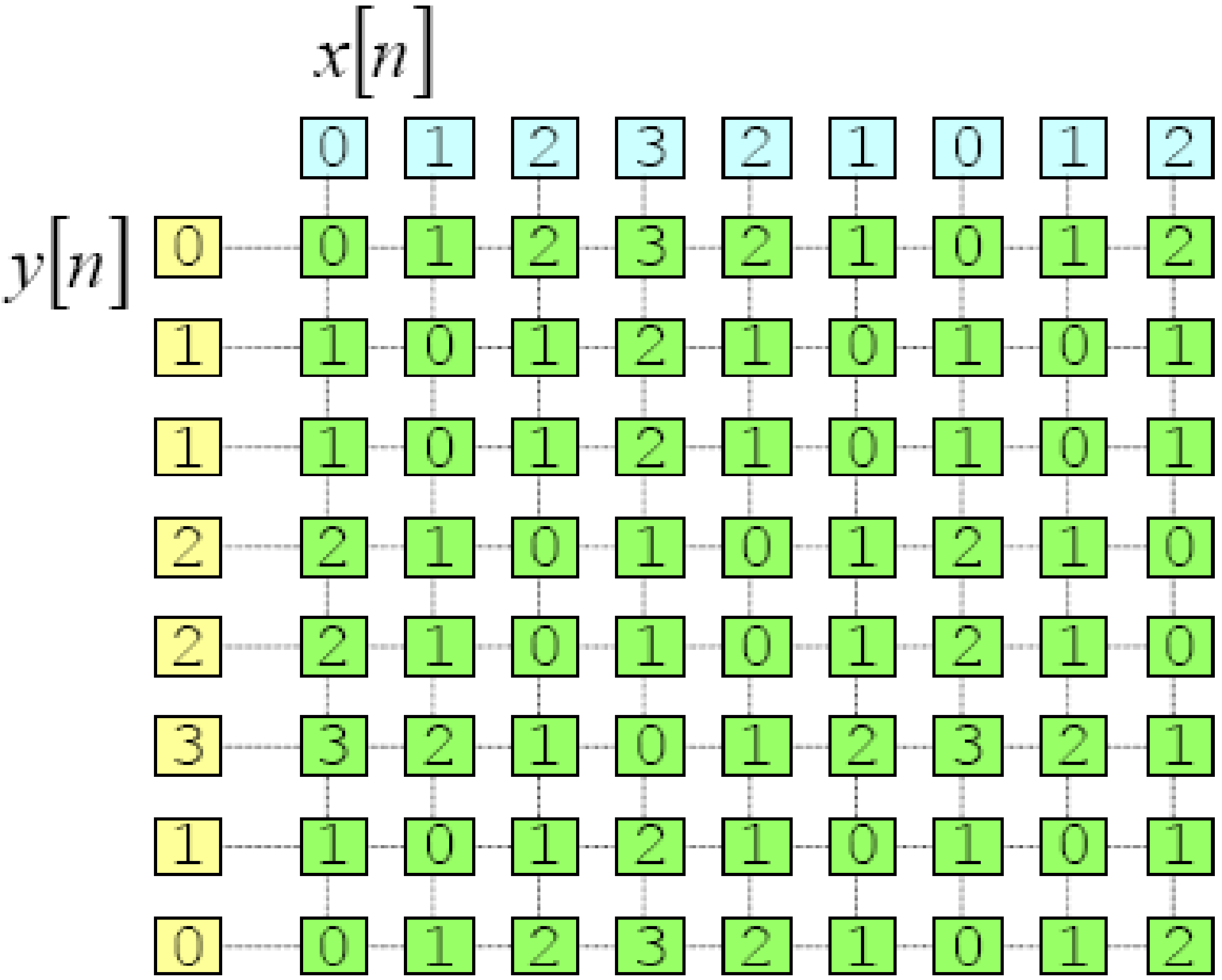
Exemplu. Găsiţi calea de aliniere şi distanta intre următoarele două secvenţe. (City-block)

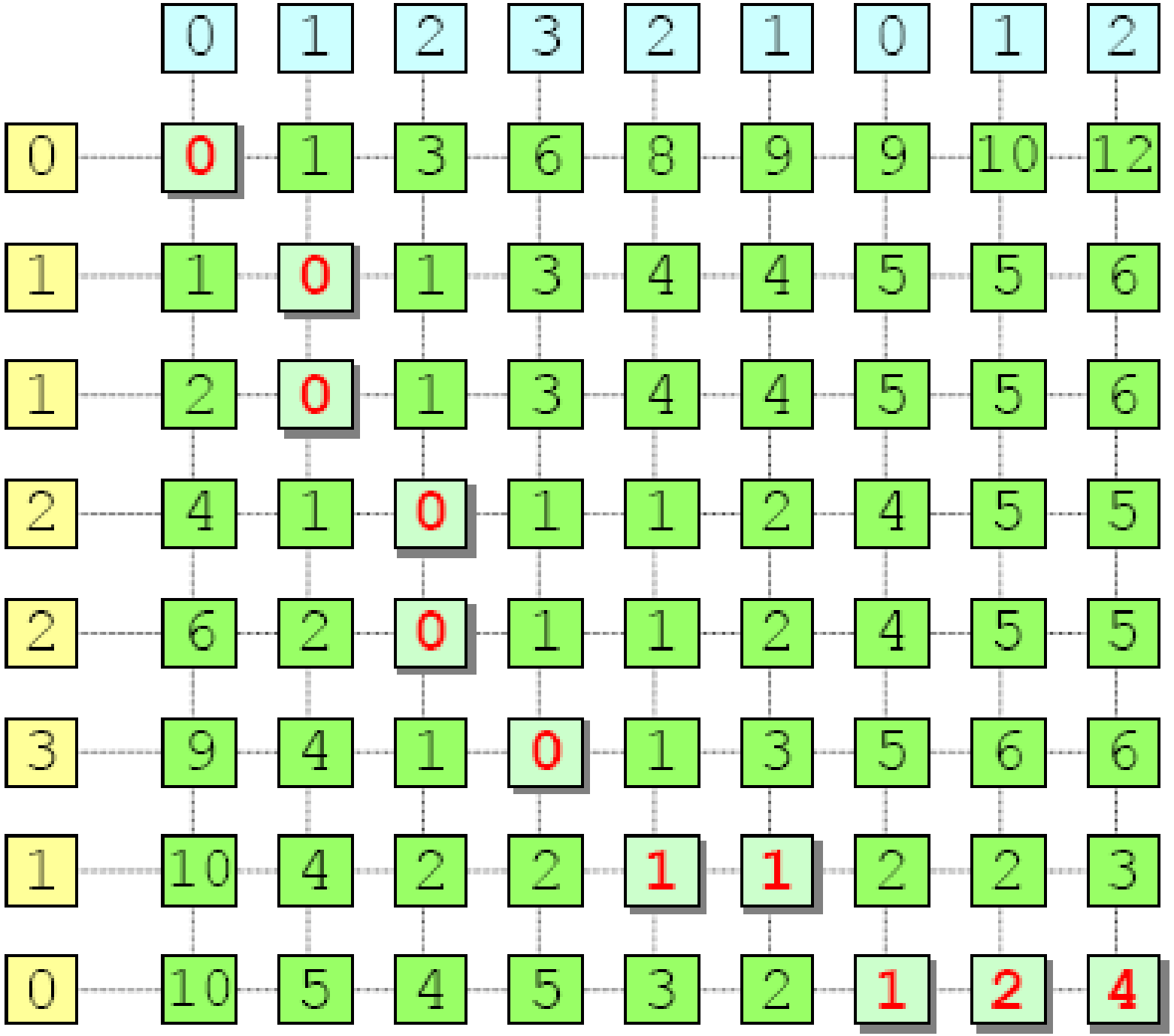
$$x[n] = \{0,1,2,3,2,1,0,1,2\}$$

$$y[n] = \{0,1,1,2,2,3,1,0\}$$



$$d(i,j) = |x[i] - y[j]|$$


x[n]
y[n]
0 1 2 3 2 1 0 1 2
0 0 1 2 3 2 1 0 1 2
1 1 0 1 2 1 0 1 0 1
1 1 0 1 2 1 0 1 0 1
2 2 1 0 1 0 1 2 1 0
2 2 1 0 1 0 1 2 1 0
3 3 2 1 0 1 2 3 2 1
1 1 0 1 2 1 0 1 0 1
0 0 1 2 3 2 1 0 1 2


0 1 2 3 2 1 0 1 2
0 0 1 3 6 8 9 9 10 12
1 1 0 1 3 4 4 5 5 6
1 2 0 1 3 4 4 5 5 6
2 4 1 0 1 1 2 4 5 5
2 6 2 0 1 1 2 4 5 5
3 9 4 1 0 1 3 5 6 6
1 10 4 2 2 1 1 2 2 3
0 10 5 4 5 3 2 1 2 4

Cum funcționează DTW? Rezumat

1. **Extragere caracteristici**: Semnalele de vorbire sunt convertite în secvențe de vectori de caracteristici (ex: coeficienți MFCC).
2. **Calcul matrice distanțe**: Se calculează o matrice de distanțe între fiecare pereche de puncte din cele două secvențe.
3. **Găsire cale optimă**: Se găsește calea cu costul minim de-alungul matricei, care "deformează" temporal una dintre secvențe pentru a se potrivi cu cealaltă.
4. **Scor similaritate**: Costul totale al căii reprezintă gradul de similitudine.

Exemplu:

- **Secvența A**: "H e e e y'' **Secvența B**: "H e y"
- DTW va alinia "e e e" din A cu un singur "e" din B, compensând diferența de durată.

Avantaje:

- Robust la variații de viteză în vorbire.
- Simplu și eficient pentru vocabular limitat.
- Folosit în sisteme "template-based" (fiecare cuvânt este stocat ca un model).

Dezavantaje:

- Computațional costisitor pentru baze de date mari.
- Sensibil la zgomot și diferențe de intonație.
- Înlocuit parțial de modelele HMM și rețelele neuronale în sistemele moderne.

Aplicații:

- Recunoaștere vorbire izolată (cuvinte sau comenzi scurte).
- Verificare vorbitor.
- Aliniere temporală în bioinformatică sau analiză gesturi.