Raport de activitate 2003

Vasile Apopei, Doina Jitca
Institutul de Informatica Teoretica – Academia Romana Filiala Iasi

Cercetari asupra metodelor de segmentare si adnotare a corpusurilor de semnal vocal

I. Modele ale productiei si recunosterii vocale. Modele pentru segmentarea corpusurilor de semnal vocal

Pentru a dezvolta metode de segmentare a corpusurilor de semnal vocal, este importanta întelegerea fenomenelor si modelelor de productie si perceptie umana a semnalului vocal, precum si cele mai recente abordari în sinteza si recunoasterea vorbirii.

I.1 Modele pentru productia vocala

Modelele de productie vocala prezentate în literatura de specialitate, urmaresc doua aspecte: modelarea tractului vocal, modelerea variabilitatii în timp a caracteristicilor sunetelor.

I.1.1 Modelarea tractului vocal

Cel mai cunoscut model al productiei vocale este cel numit "modelul sursa-filtru", descris de Fant în 1960. Conform acestui model, componentele ansamblului de productie vocala este compus din trei parti: sursa de excitatie, echivalenta generatorului undei vocale, un tub cu cavitati de rezonanta, echivalent unui filtru, în procesarea de semnal, si tractului vocal superior, în mecanismul de productie vocala, si un filtru care simuleaza fenomenul radiatiei bucale. Parametrii referitori la rezonantele tubului (în principal polii filtrului) pot fi folositi pentru a caracteriza din punct de vedere spectral portiuni ale fluxului sonor vocal.

Pentru modelarea tuturor sunetelor vorbirii, se folosesc doua tipuri de surse de excitatie:

- ? Generator de impulsuri; forma implusurilor este asemanatoare celor glotale naturale, cu o panta de crestere mai mica decât cea de scadere, si se desfasoara pe durata deschiderii glotale. Acestea sunt separate de portiuni pe care semnalul ramâne la zero si corespund intervalelor de închidere glotala. Acest tip de excitatie se produce în cazul vocalelor, consoanelor sonante (nazale, lichide) si a celorlalte consoane sonore. Trenurile de impulsuri au o panta spectrala de -12 dB/octava.
- ? Generator de zgomot aleator echivalent celui de fricatie produs la trecerea aerului prin portiunile constrictive, în cazul consoanelor fricative sau a celor cu portiuni fricative.

Identitatea fiecarui sunet este data de pozitiile particulare ale articulatorilor mobili care formeaza cavitatile rezonante ale tractului vocal superior. Acesta actioneaza ca un filtru asupra semnalului de excitatie. În stadiul final al modelarii sursa-filtru, se simuleaza fenomenul radiatiei bucale, care se concretizeaza în cresterea pantei spectrale cu +6 dB/octava. În sinteza formantica a sunetelor sonore se combina efectul radiatiei bucale cu cel produs de spectrul de excitatie si se foloseste un filtru care realizeaza, pe ansamblu, o panta spectrala de –6 dB/octava.

I.1.2 Modelarea semanlului vocal în domeniul timp

Prin modelerea variabilitatii caracteristicilor sunetelor, se urmareste reprezentarea în timp a urmatoarelor trasaturi: durata si co-articularea sunetelor, evolutia în timp a principalilor parametrii ai modelului.

a. Modele pentru durata sunetelor

La nivel vorbitorului, durata sunetelor este în corelatie cu rapiditatea vorbirii. Rapiditatea vorbirii este limitata de inertia articulatorilor iar durata sunetelor variaza functie de mobilitatea articulatorilor implicati în producerea lor (miscarea buzelor si a limbii). Duratele medii ale fonemelor variaza între 20 msec pentru consoanele plosive sonore pâna la 150 msec pentru diftongi, cu o durata medie a fonemelor de 70 msec. La vocale, durata variaza, functie de context, între valori aflate într-un raport de 1/8 si depinde de silaba în care se afla. Kanedera si Hermansky [Kanedera97] în studiul lor au pus în evidenta faptul ca modulatia perceptuala cea mai importanta a vorbirii (modificarile cele mai importante în semnalul vocal) este realizata în jurul valori de 4-5 Hz, sau 200-250 msec cât este aproximativ durata unei silabe [Greenberg96].

Klatt [Klatt76] face remarca ca în recunoasterea vorbirii informatia de durata este folosita de fiintele umane pentru a distinge:

- vocalele lungi de cele scurte,
- consoanele sonore de perechile lor surde,
- silaba de final fraza de cea neaflata în pozitie final fraza,
- vocala din silaba accentuata de cea din silaba neaccentuata.

Daca se iau în considerare multitudinea factorilor care influienteaza duratele fonemelor si perceptia, rezulta modele relativ complexe. Modelul propus de Klatt stabileste 7 factori care influenteaza structura duratelor dintr-o propozitie, si 8 reguli care tin cont de acesti factori [Klatt76].

Un model mai simplu propus de van Santen [Santen92] este capabil sa modeleze 86% din situatiile de variatie a duratelor vocalelor cuprinse într-un corpus segmentat manual. Acest model necesita urmatorii parametrii pentru controlul duratei vocalelor:

- durata intrinseca a vocalei,
- prezenta accentului de propozitie
- prezenta accentului de cuvânt,
- consoana dinaintea vocalei,
- consoana dupa vocala,
- pozitia în cadrul cuvântului
- pozitia în cadrul rostirii.

In modelul statistic dezvoltat de Chung, pentru modelarea duratei la nivel de fonem si cuvânt este folosita o structura cu arbori (ANGIE framework). Antrenarea arborelui este facuta be baza unui corpus de date. Informatia de durata cuprinsa în arbore poate fi folosita pentru a testa diferite ipoteze asupra unor posibile cuvinte si a favoriza pe acelea care au o mai buna apropiere cu modelul [Chung97]. Folosirea acestui model, în cadrul unui proces de recunoastere a condus la o scadere cu 8% a erorilor de recunoastere în vorbirea continua si cu 22% a erorilor de recunoastere a cuvintelor.

b. Modele pentru co-articularea sunetelor

În ceea ce priveste co-articularea sunetelor vecine, aceasta se evidentiaza prin tranzitii formantice lente de la unul la celalalt care fac dificila stabilirea granitelor acestora.

În modelul dezvoltat de *Öhman* [Öhman66], modificarea formei tractului vocal la tranzitiile de tipul vocala-consoana-vocala a fost descrisa prin relatia (I.1)

$$s(x,t)=v(x)+k(t) * [c(x) - v(x)] * w_c(x) \qquad \text{(I.1)}$$

unde: $s(x,t)$ este forma tractului vocal în pozitia x si la momentul de timp t,
$v(x)$ este forma tractului vocal corespunzatore vocalei respective,
$c(x)$ este forma tractului vocal corespunzatoare consoanei,
$k(t)$ este un termen de interpolare între 0 si 1
$w_c(x)$ este un terman care descrie marimea extrinderii co-articulatiei.

Autorul recunoaste ca e greu de descris cu acest model co-articularea între consoane, cum ar fi în cazul tranzitiei CVC, consoana-vocala-consoana.

În modelul bazat pe teoria punctului de articulare [Delattre55], consoanele au asignate valori fixe pentru formanti, corespunzatoare punctului de articulare, care pot sa nu fie vizibili în semnalul vocal. Acesti formanti virtuali sunt interpolati cu formantii propriu-zisi care apar la vocale, pentru a genera modificari formantice dependente de context. *Klatt* a modificat teoria punctului de articulare pentru ca valorile "formantilor" la consoane sa depinda si de tipul vocalei [Klatt87]. Folosind aceasta metoda el a atins o inteligibilitate a consoanelor de 95% în sinteza silabelor CVC, comparabila cu inteligibilitatea de 99% pentru silabele CVC rostite natural. Klatt nu a evaluat modelul pe extensii ale co-articulatei mai mari de 6 foneme.

În modelul propus de *Löfqvist*, raportat de Cohen si Massaro [Cohen93], segmentele de speech îsi suprapun functiile de dominanta care controleaza articulatorii, existând câte o functie de dominanta pe câte un articulator. Acestea pot diferi în ceea ce priveste offset-ul de timp, durata si marime, dând o mai mare sau mai mica pondere articulatorilor asociati cu segmentul de vorbire dat.

Desi aceste modele pentru co-articulatii sunt folosite cu succes în modelarea reprezentarii miscarii articulatorilor în timpul vorbirii, ele nu pot fi folosite în sistemele de recunoastere a vorbirii.

I. 2. Modele ale mecanismului de perceptie unana a vorbirii

I.2.1 Teoria motorie a perceptiei vorbirii

Teoria motorie a perceptiei vorbirii este una din cele mai cunoscute si mai des folosite. Într-o versiune noua [Liberman85] stabileste drept obiective ale perceptiei vorbirii, gesturilor fonetice intentionate ale vorbitorului necesare pentru articularea sunetelor. La nivelul creierului, acestora le corespund comenzile motoare pentru miscarea articulatorilor, care constituie comenzi invariante în raport cu un anumit sunet. Cu alte cuvinte, ceea ce percepem sunt gesturi care corespund miscari articulatorilor efectuate de un vorbitor. Tot în aceasta teorie se sustine ideea existentei în creier a unui "modul specializat" care transforma semnalul acustic în gesturi fonetice intentionate ale articulatorilor (figura 1)

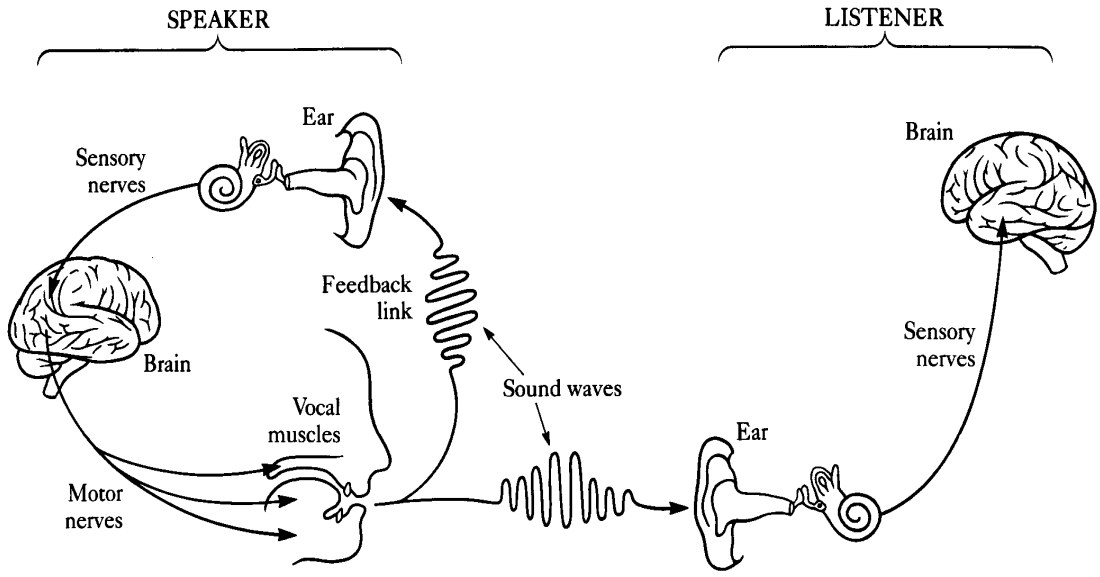

Figura 1. Modelul lui Liberman de perceptie a vorbirii

Dupa Liberman, acest modul poate lucra dupa metoda de analiza prin sinteza, în care modelul mental al sintetizatorului este folosi pentru a genera diferite proprietati acustice. Parametrii gesturilor acustice de la intrare sintetizatorului sunt modificate pâna când eroarea dintre proprietatile acustice sintetizate si proprietatile observate sunt minimizate. Iesirea acestui modul este reprezentata prin gesturile articulatorilor. Liberman si Mattingly afirma ca acest model al perceptiei este computational si indirect (usor invesabil).

Una din criticile aduse modelului este formulata de Cole care arata, în urma unui studiu [Cole80], ca recunoasterea sunetelor se poate face si de pe spectrograma, prin 'citirea' unei forme vizuale a vorbirii fara implicarea unui modul biologic specializat. Persoanele care efectueaza citirea spectrogramelor nu fac referiri la miscarile articulatorii. Acest studiu combate afirmatiile conform carora semnalul acustic este prea complex pentru a fi mapat în categorii fonetice si aparatul perceptiei auditive necesita un stadiu intermediar, de transformare a informatiei în gesturi articulatorii.

Un alt studiu, cel al lui Lane, combate ideea transformarii semnalului acustic în parametri de miscare a articulatorilor. S-au folosit stimuli de tip CV carora li s-au inversat frecventele formantice pe axa frecventelor, obtinându-se sunete nonverbale. Auditorii au fost antrenati sa învete sa le recunoasa si acest lucru a reusit, desi sunetele neprovenind din rostiri verbale nu pot fi puse în legatura cu acei presupusi parametri interni referitori la comanda miscarii articulatorilor [Lane65].

Un alt argument îl aduce Ladefoged prin ideea ca aceleasi sunete, pronuntate diferit din punct de vedere al articularii, sunt percepute ca reprezentând acelasi sunet, desi conform teoriei motoare ar trebui sa fie reprezentate intern diferit [Ladefoged93].

I.2.2 Modelul perceptiei vorbirii bazat pe trasaturi multiple

Ronald Cole si Brian Scott propun un model al perceptiei vocale bazat pe un set de trasaturi independente de context si un alt set de trasaturi dependente de context, relative la tranzitii. Ambele seturi de trasaturi sunt implicate în recunoasterea silabelor [Cole74]. Pentru integrarea silabele în cuvinte si fraze modelul foloseste proprietati referitoare la anvelopa semnalului.

Pentru unele foneme (/s/, /z/, /s/, /j/, /c/(ci), /g/(gi)), trasaturile invariante sunt suficiente pentru a le identifica în mod unic. În alte cazuri (/f/, /v/, /m/, /n/), trebuie puse în conjunctie cu cele referitoare la tranzitii. În cazul consoanelor oclusive, distinctia între cele sonore si cele surde se face pe baza trasaturilor independente de context, în timp ce locul articularii le implica atât pe acestea cât si pe cele tranzitionale. În plus fata de aceste doua tipuri de trasaturi, se pot folosi proprietatile de anvelopa ale undei în timp, pentru recunoasterea informatiei prozodice sau fonetice.

Avantajele acestui model sunt urmatoarele:

- ? efortul computational este mai mic decât în cazul teoriei motoare, care are un grad mare de complexitate;
- ? exista o mapare directa de la trasaturile acustice la cele fonetice;
- ? unda glotala este luata în considerare sub ambele aspecte, al trasaturilor invariante si al celor dependente de context.

Aspectele acestui model, carora le-au fost aduse critici, se refera la faptul ca amplitudinea undei nu este o trasatura folosita de aparatul uman de perceptie, caci receptorii senzoriali transforma informatia auditiva din domeniul timp, în domeniul frecventa. Dar, uneori informatiile din domeniul timp folosite de Cole si Scott (amplitudinea, 'pitch'-ul, durata) pot fi extrase din reprezentarea spectrala variabila în timp.

I. 2.3 Trasaturi invariante pentru perceptia consoanelor oclusive

Stevens si Blumstein în cercetarile lor asupra perceptiei vocale [Stevens78] au stabilit trasatura invarianta care sa identifice locul de articulare a consoanei oclusive, pornind de la analiza perceptuala a undei vocale produsa de sintetiza grupurilor consoana-vocala. Aceasta trasatura se refera la *forma spectrului consoanei efectuat pe cadre din bara de explozie si în zona deschiderii vocale*. Pe acest studiu s-au bazat Cole si Scott la stabilirea modelului cu trasaturi multiple, în specificarea indiciului care poate fi folosit pentru identificarea punctului de articulare a consoanelor oclusive. Cercetari asupra principalilor indicatii în determinarea articulatiei pentru consoanele plosive au fost efectuate si de Jackson [Jackson2001]. El a pus în evidenta urmatoarele indicii:

- pentru consoanele surde /p, t, k/ în realizarile tipice avem:
  - ? o crestere brusca (burst) a amplitudini semnalului vocal corespunzatoare momentului exploziei (îndepartarea ocluziei);
  - ? o scurta portiune fricativa;
  - ? în final o portiune cu zgomot de aspiratie înainte si în la începutul fonemului urmator.
- consoanele sonore /b,d, g/ în realizarile tipice avem:
  - ? o crestere brusca (burst) a amplitudini semnalului vocal corespunzatoare momentului exploziei (îndepartarea ocluziei);
  - ? o scurta portiune fricativa;
  - ? în final o portiune cu zgomot de aspiratie mai mic decât la consoanele surde;
  - ? o sonorizare mai mare decât la consoanele surde (amplitudine mai mare a semnalului vocal)

I.2.4 Modelul Fletcher-Allen

Harvey Fletcher si colegii sai a studiat mecanismul perceptii umane la Bell. Labs. Un rezultat al cercetarilor sale [Allen94] a fost masurarea recunoasterii corecte a silabelor de forma CVC calculând o rata de recunoastere a fonemelor componente (relatia I.2):

$$S = c_1 * v * c_2 \qquad \text{(I.2)}$$

unde: S este probabilitatea de recunoastere a silabei;
c_1, v, c_2 probabilitatile de recunoastere corecta a consoanelor si a vocalei dintre acestea.

Aceasta formula are o implicatii importante deoarece oameni percep fiecare fonem individual, mai degraba ca unitate de intrare într-o silaba. În plus, Fletcher a observat ca fiintele umane proceseaza benzile de frecventa independent si ca eroarea globala de recunoastere în mai multe benzi este data de multiplicarea erorilor în fiecare din acestea.

Allen a interpretat aceste rezultate astfel: fiintele umane efectueaza recunoasteri partiale în benzi de frecventa individuale si aceste rezultate partiale fuzioneaza pentru a produce o estimare a fonemului. Numarul benzilor de frecventa trebuie sa fie între 10 si 30. Allen noteaza, de asemenea, ca “reprezentarea neurala la nivelul creierului, a intensitatii sunetului este transformata într-o masura a recunoasterii partiale ... noi nu trebuie sa consideram ca aceste transformari sunt triviale”.

Bazat pe descoperirile lui Fletcher, Allen [Allen94] propune un model cascada al mecanismului de perceptie, în care energia semnalul acustic este mai întâi împartita într-un numar de benzi de frecventa intens suprapuse cu ajutorul unui bank de filtre cochlear(Figura I.2). Iesirile acestor benzi sunt folosite pentru a extrage trasaturile sunetului pentru al clasifica la nivel

de fonem. Dupa clasificarea la nivelul fonemului se face recunoastere silabei, pe care se bazeaza apoi recunoasterea la nivel de cuvânt. Allen noteaza ca în acest model simplificat nu exista feedback între diferitele nivele de recunoastere.

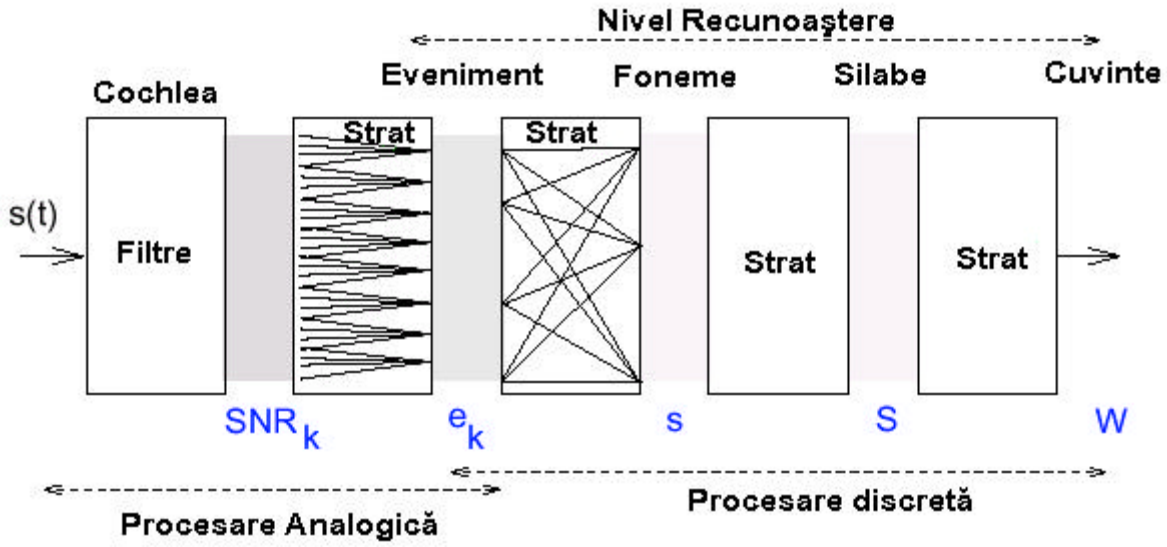

Figura I.2 Model cascada al mecanismului de perceptie propus de Allen

I.2.5 Analiza scenelor auditive

Analiza scenelor auditive (ASA) este un model teoretic al perceptiei umane în care se foloseste o procesare *bottom-up* si *top-down* pentru a determina ce parti din semnalul de vorbire apartin unui singur eveniment acustic [Bregman90]. Modelul este construit pornind de la analiza mediilor sonore complexe ce implica existenta mai multor sunete simultan. De asemenea, dupa izolarea unor componente de interes din fiecare portiune de semnal vocal, pentru integrarea lor într-un '*streams*', se foloseste criteriul similaritatii în ceea ce priveste frecventa de 'pitch', sau alte aspecte [Cooke93]. Cu ajutorul acestui model Cooke si Brown au fost capabili sa detecteze si sa extraga anumite sunete 'acoperite', cum ar fi cele acoperite de sunetul unei sirene.

I.2.6 Modelul TRACE

Modelul a fost dezvoltat de James McClelland si Jeffrey Elman în 1986 [McClelland86]. Acesta are trei nivele: trasatura, fonem si cuvânt (figura I.3). La nivelul trasaturilor sunt folosite urmatoarele caracteristici ale sunetelor:

- caracterul consonantic;
- caracterul vocalic;
- caractertul difuz;
- înaltimea sunetelor (frecventa de pitch);
- grad de sonorizare;
- puterea spectrala;
- amplitudinea *burst*-urilor zgomotului.

Fiecare din aceste elemente poate lua una din 9 valori.

Fiecare nivel este realizat din retele neurale artificiale construite prin conectarea unui numar de unitati de procesare simple prin legaturi excitatoare si inhibitoare. Procesul de recunoastere se realizeaza în urma modificarilor continui ale starilor celulelor, functie de cea a celulelor cu care sunt conectate. Fiecare celula reprezinta o ipoteza despre intrare, iar activarea ei este proportionala cu gradul de adevar al ipotezei asupra intrarii. Conexiunile dintre celule exprima relatia între ipoteze. Între celulele de pe acelasi nivel, corespunzatoare unor ipoteze ce se exclud, exista conexiuni inhibitorii. Conexiunile între straturi sunt bidirectionale, ceea ce permite procesare *bottom-up* si *top-down* simultan.

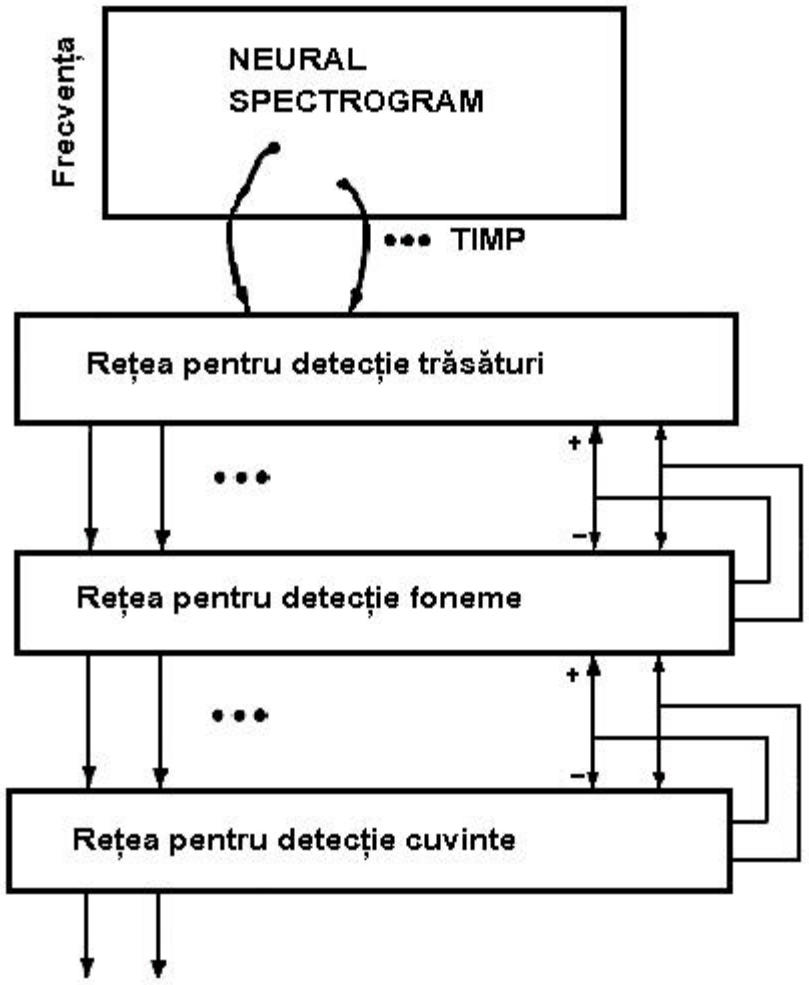

Figura I.3 Modelul TRACE dezvoltat de McLelland & Elman în 1986

Modelul TRACE este capabil sa simuleze un numar de efecte perceptibile auditiv, ceea ce îi confera un suport psihologic pentru validitatea acestuia. Aceste efecte sunt de natura lexicala *top-down*, ca, de exemplu:

- ? un raspuns mai rapid la cuvintele recunoscute decât la cele nerecunoscute,
- ? perceptia fonemelor drept categorii distincte manifestate prin stabilizarea retelei în loc de schimbari continui ale starii.

Autorii au stabilit o lista cu 11 aspecte de similaritate între functionarea modelului si fenomenul perceptiei umane, dar recunosc ca are un numar egal de neconcordante, multe din ele fiind generate de simplificarile impuse de un efort computational acceptabil.

Modelul nu se bazeaza pe retele neuronale artificilale, date fiind conexiunile bidirectionale si lipsei unui formalism de antrenare, ceea ce le apropie de cele biologice.

I.2.7 Model de perceptie a vorbirii bazat pe logica fuzzy

Modelul de perceptie a vorbirii bazat pe logica fuzzy (FLMP) nu este un model complet al recunoasterii vorbirii, pentru ca nu specifica toti pasii, de la intrarea semnalului pâna la cuvântul recunoscut. Modelul are în atentie integrarea trasaturilor pentru a ajunge la rezultate de clasificare în concordanta cu performantele umane.

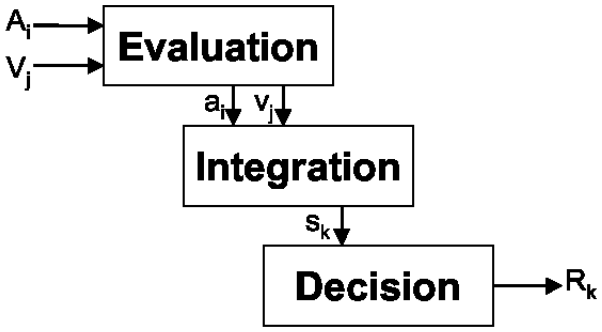

Figura I.4 Schema de procesare în trei nivele a FLMP dezvoltat de Massaro [Massaro98]

Functionarea modelului (figura I.4) se desfasoara în trei pasi (nivele): evaluarea trasaturilor, integrarea trasaturilor si clasificarea formelor.

În stadiul de evaluare a trasaturilor, semnalul vocal este analizat si anumite trasaturi sunt extrase. Spre exemplu, poate fi o trasatura numita labiala, pentru a indica locul articularii. Valorile acestor trasaturi sunt continui si ele reprezinta gradul de încredere ca segmentul de vorbire curent are respectiva trasatura. De exemplu trasatura ‘labial’ poate avea valoarea 0.80, indicând o încredere puternica în faptul ca segmentul de vorbire reprezinta un sunet labial. Folosirea valorilor continui pentru fiecare trasatura este acceptata de multe din studiile legate de perceptia vorbirii [oden78] [oden91].

Al doilea nivel consta în compararea cu formele prototip, în care intrarea este comparata cu descrierea prototip a fiecarui fonem. De exemplu fonemul /b/ poate avea trasaturile ‘labial’ si ‘sonor’ în descrierea prototip. Gradul de potrivire a semnalului de intrare cu fonemele prototip este specificat, prin intermediul unei functii de *matching*, în termeni de valori de încredere fuzzy. Spre exemplu, functia de *matching* pentru /b/ poate fi specificata astfel (relatia I.3):

$$B_s = L_s * V_s \qquad \text{(I.3)}$$

unde: B_s este gradul de potrivire al intrarii *s*, cu fonemul prototip */b/*
L_s este gradul de încredere ca sunetul sa fie labial
V_s este gradul de încredere ca sunetul de la intrare sa fie sonor.

Pe acest nivel pentru fonemul de la intrare se calculeaza câte o functie de *matching* pentru fiecare fonem prototip.

În al treilea stadiu se efectueaza clasificarea formelor. Probabilitatea de identificare a fiecarui fonem este calculata folosind modelul lui Luce [Luce59]. Pentru exemplul cu fonemul /b/, probabilitatea ca sunetul de la intrare sa fie /b/ este data de formula (I.4):

$$p(/b/|s) = B_s/(P_s + B_s + D_s) \qquad \text{(I.4)}$$

unde: P_s , B_s , D_s sunt probabilitatile ca sunetul sa apartina claselor */b/, /p/* si respectiv, */d/*.
p(/b/|s) este probabilitatea ca segmentul de voce *s* sa fie */b/*

Massaro si Friedman folosesc FLMP-ul pentru a modela rezultatele unor activitati de clasificare realizate în perceptia umana. Ei considera ca acest model asigura un echivalent sau o mai buna reprezentare a datelor prin comparatie cu un alte modele de tip *information-integration*, cum ar fi cel aditiv, cu integrare lineara, bazat pe retea cu doua straturi [Massaro90]. Modelul FLMP este matematic echivalent cu integrarea Bayes-iana daca valorile de adevar fuzzy sunt interpretate ca probabilitati, desi FLMP a fost dezvoltat pe baza unor studii psihologice fara referiri la reguli Bayes.

I.3. Modele computationale pentru recunoasterea vorbirii

Exista un numar mare de modele pentru sisteme de recunoastere a vorbirii, fiecare cu diferite perspective de abordare. Cele mai multe modele pot fi, în general, clasificate în doua categorii:

- bazate pe segment – extragerea trasaturilor se face pe segmente de rostire mai mari de 20msec
- bazate pe cadru – extragerea trasaturilor se face pe segmente de rostire mai mici de 20msec

În continuare vom face referiri la cele mai importante dintre aceste sisteme.

I.3.1 Sisteme de recunoastere a vorbirii bazate pe segment

I. 3.1.1 Sistemul SUMMIT

Sistemul SUMMIT a fost dezvoltat de Victor Zue de la MIT în 1980 iar în variante ulterioare îmbunatatite, de catre Jim Glass. Caracteristic acestui sistem este faptul ca mai întâi împarte semnalul în segmente si apoi clasifica din punt de vedere fonetic fiecare segment. Procedura generala de recunoastere în sistemul SUMMIT este urmatoarea:

1. Granitele acustice sunt determinate pe baza unei multimi de modificari spectrale. Într-o implementare mai particulara a sistemului SUMMIT [Chang97] granitele sunt plasate automat la fiecare 10msec, transformându-l efectiv dintr-un sistem bazat pe segment într-unul bazat pe cadru, dar aceasta implementare nu este folosita în mod curent deoarece necesita un timp mare de calcul.
2. O retea de segmente (dendrograma) este creata prin una din urmatoarele metode:
   - Unind segmentele mici în segmente mai mari în acord cu similaritatile lor spectrale. Aceasta este o metoda traditionala folosita în SUMMIT [Glass88], care necesita putine resurse computationale.
   - Segmentarea prin recunoastere, folosind o procedura de recunoastere prin care sunt clasificate fiecare segment sau zona, marcate fie ca foneme, fie ca portiuni tranzitorii (co-articultii). Dupa aceasta clasificare, este facuta o cautare *Viterbi* “forword-pass”, care este urmata de o cautare înapoi de tip *A**. Cautarea *A** produce un numar de alternative de segmentare fonetica care reprezinta rezultatul într-o dendograma. Aceasta metoda are un cost computational mai mare dar are performante de recunoastere mai bune.
3. Pe baza dendogramelor create în pasul 2, se efectueaza clasificarea fonetica a tuturor segmentelor, folosind urmatoarele doua metode:

? Prima metoda efectueaza recunoasterea independent de context a fiecarui segment din dendodrama. În aceasta metoda sunt între N+1 si 2*N categorii, dintre care un numar de N categorii corespund celor N foneme posibile, iar restul de N categorii sunt folosite pentru a modela segmentele neincluse în segmentarea cu ipoteze numite "ne-modelabile" sau "aproape de a fi modelate" [Chang97].

? A doua metoda efectueaza recunoasterea dependenta de context a fiecarei granite de segment din dendograma [Glass96]. Categoriile dependente de context pot fi granite fonetice sau granite interne unui fonem, si ar putea fi în numar de ($N + N^2$). În practica, numai 750 de categorii sunt folosite.

Aceste clasificatoare sunt antrenate cu aceleasi trasaturi spectrale care sunt comune si sistemelor bazate pe HMM-uri, iar clasificarea este facuta folosind o combinatie de gaussiene.

4. Cautarea continua cu un "bigram" *Viterbi forward* si, pentru cele mai bune N ipoteze, o cautare de tip *n-gram A** cu trecere înapoi. Daca ambele recunoasteri (cea independenta de context si cea a zonelor de granita) sunt efectuate în pasul 3, atunci probabilitatea finala a secventei de cuvinte este calculata prin înmultirea probabilitatilor fiecarui segment si a zonelor de granita dintre acestea.

Performantele celor mai recente sisteme SUMMIT sunt de 72% pe clasificarea fonemelor din baza TIMIT. Rezultatele la nivel de fonem sunt printre cele mai bune raportate. Un sistem bazat pe HMM raporteza 69.1% [Lamel93]procent de recunoastere iar unul bazat pe retele neurale 73.4% [Robinson94].

I.3.1.2 Sistemul FEATURE

Sistemul FEATURE a fost dezvoltat de Ronald Cole, Richard Stern la Carnegie-Mellon în anii 1980. Autorii proiectului si-au propus sa realizeze recunoasterea automata a vorbirii astfel încât sa realizeze distictia între sunete precum /p^h/(p-cu zgomot aspiratie) si /b/. Când sistemul FEATURE a fost dezvoltat, sistemele de recunoastere bazate pe cadre si potrivire cu prototipul aveau o rata de recunoastere de numai 60% pe "E-set"-urile care apar la rostirile în limba engleza a unor cifre si litere: B, C, D, E, G. P, T, V, Y si 3. Sistemul original a fost proiectat pentru recunoasterea, independenta de vorbitor, a literelor izolate A.....Z din limba engleza [Cole90] (figura I.5).

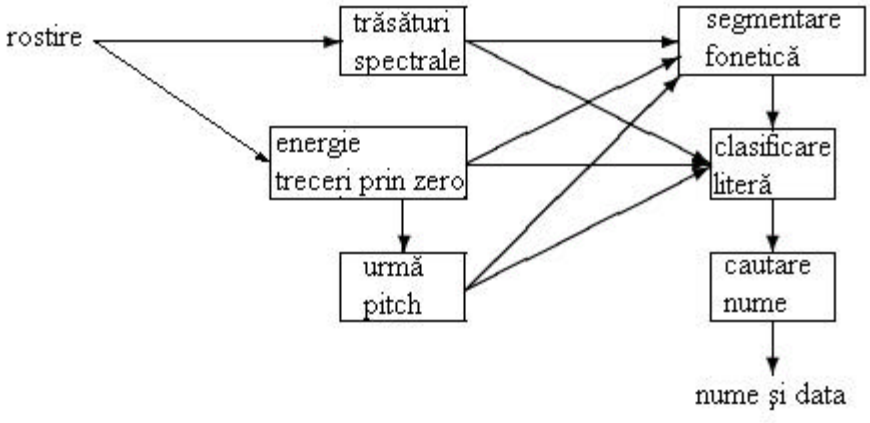

Figura I.5. Structura sistemului FEATURE dezvolatat de Roland Cole

În procesul de recunoastere al sistemului FEATURE sunt implicati urmatorii pasi:

1. *Procesarea de semnal*: Data fiind o rostire corespunzatoare unei singure litere, rutinele de procesare de semnal sunt folosite pentru extragerea informatiilor generale despre semnal, cum ar fi proprietatile spectrale, frecventa fundamentala, numarul de treceri prin zero si energia în diferite benzi de frecventa.
2. *Segmentarea*: Patru puncte sunt localizate în rostire: începutul rostirii, *onset*-ul vocalei, *offset*-ul vocalei, si sfârsitul rostirii.
3. *Extragerea trasaturilor*: Aproape 50 de trasaturi diferite sunt extrase folosind informatia determinata la pasii 1 si 2. Aceste trasaturi includ:
   - ? primii trei formanti a regiunilor corespunzatoare vocalelor;
   - ? traiectoriile formantilor;
   - ? maximul si minimul frecventelor fiecarui formant ;
   - ? durata sunetului aperiodic dinainte si de dupa vocala.

   Aceste trasaturi au fost selectate folosind inspectia vizuala a diferitelor reprezentari ale semnalului.
4. *Clasificarea*: Pentru determinarea probabilitatilor fiecareia din cele 26 de litere a fost folosita o metoda cu arbori de decizie. Fiecare nod în arbore reprezinta un grup de litere si nodurile de pe un nivel mai jos contin subseturi disjuncte ale nodurilor de pe nivele aflate mai sus. Nodurile ‘frunza’ contin litere individuale. La fiecare nod care nu este ‘frunza’, probabilitatea unei rostiri de a apartine acelui nod este determinata folosind o distributie Gaussiana a vectorilor de trasaturi. Probabilitatile sunt calculate pentru toate nodurile care nu sunt ‘frunze’, în cadrul arborelui si probabilitatea finala a unei litere date rezulta din înmultirea probabilitatilor fiecarui nod care conduce la nodul ‘frunza’. Pentru a reduce dimensiunea spatiului de decizie, la fiecare nod sunt folosite numai trasaturile relevante. De asemenea ste facuta presupunerea ca seturile de trasaturi folosite pentru clasificare la fiecare nod sunt independente.
5. *Adaptare*: Pentru o mai buna potrivire cu valorile trasaturilor, distributiile de probabilitate Gaussiene pot fi ajustate pentru fiecare vorbitor. Distributia de probabilitate este reactualizata dupa recunoasterea fiecarei rostiri si dupa confirmarea din partea utilizatorului ca recunoasterea s-a facut corect (adaptare supervizata).

Fara a folosi adaptarea pentru fiecare vorbitor, sistemul FEATURE are 89% rata de recunoastere pe litere izolate si 83% pe “E-set”-uri. Sistemul FEATURE a fost modificat de Cole, Fanty si altii pentru a folosi retele neuronale pentru clasificare si pentru a recunoaste litere rostite continuu (rostite cu sau fara pauza). Rezultatele acestui nou sistem numit “EAR” sunt 96% rata de recunoastere pe rostire la microfon, si 89% pe semnal telefonic [Cole90] . Acestea au fost cele mai bune rezultate timp de 6 ani, pâna când au aparut sistemele sofisticate cu HMM care au atins 97,3% pe semnal foarte bun si 91,7% pe cel de pe linie telefonica analogica în 1996 [Loizou96].

I. 3.2 Sisteme de recunoastere a vorbirii bazate pe cadre

I. 3.2.1 Modele bazate pe HMM

Cele mai multe metode de recunoastere automata a vorbirii se bazeaza pe modele Markov. Aceste modele au fost folosite în recunoasterea vorbirii din anul 1975 [Bengio99]. Modelul ofera un cadru matematic elegant care permite antrenarea pe corpus-uri de date, pentru unitati de vorbire cum ar fi cuvinte, foneme sau sunete dependente de context.

HMM-urile au câstigat în competitia cu alte modele statistice de recunoastere a vorbirii din urmatoarele ratiuni:

- ? din punct de vedere formal modelul matematic este bine fundamentat;
- ? problema de recunoastere a vorbirii devine o problema de recunoastere statistica a formelor
- ? performantele sistemelor bazate pe HMM sunt adesea superioare celor bazate pe cunostinte.

Modelul HMM pentru recunoastere de vorbire este cel cu stari independente legate prin arce corespunzatoare tranzitiei starilor (figura I.7). Pentru fiecare cuvânt, în etapa de antrenare se creaza câte un model (o succesiune de stari).

Fiecare stare este asociata cu o anumita unitate lingvistica (în mod obisnuit o unitate fonetica), si în orice moment de timp sistemul se poate gasi doar într-o singura stare. La fiecare 10 msec se efectueaza o tranzitie spre o noua stare. În timpul recunoasterii, sistemul estimeaza:

- ? probabilitatile de a se afla în fiecare din stari la timpul *t* pe baza probabilitatilor de a se afla în fiecare din stari la timpul *(t-1)*;
- ? probabilitatile de tranzitie de la starile anterioare la starile curente;
- ? probabilitatea starii curente de a fi asociata cu semnalul la timpul *t* (numita probabilitate de observatie).

Probabilitatea fiecarui cuvânt la momentul t_w este data de probabilitatea de a fi în stare finala asociata cu acel cuvânt la momentul t_w; o simpla comparatie a probabilitatilor starilor de final de cuvânt produce cel mai asemanator cuvânt, date fiind semnalul de intrare si configuratia HMM-ului.

În general, semnalul de intrare la timpul *t* este reprezentat la intrarea HMM (figura I.6) de o informatie în domeniul spectral, pe o fereastra mica de timp (mai mica de 16 msec). Probabilitatea ca un cuvânt w_i sa fie rezultatul observatiei *O* este data de relatia (I.5)

$$P(O \mid w_i) \approx \max_X \left[a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) * a_{x(t)x(t-1)} \right] \qquad \text{(I.5)}$$

unde: o_t reprezinta vectorul de trasaturi (figura I.6) la momentul *t*

$b_{x(t)}(o_t)$ reprezinta probabilitatea ca vectorul de trasaturi sa se afle în fiecare din starile *x(t)*

X este secventa de stari (1,2,3,4,5,6 în figura I.7) în ordinea în care genereaza trasaturile o_t

$a_{x(t)x(t-1)}$ reprezinta probabilitati de tranzitie de la starea *x(t-1)* la starea *x(t)*

Estimarea probabilitatilor de observatie este efectuata folosind o combinatie de Gaussiene sau cuantizarea vectoriala (VQ). Cautarea prin starile HMM, pentru a o gasi cel mai asemanator cuvânt, este efectuata folosind diversi algoritmi de cautare din programare dinamica (Viterbi, Baum-Welch).

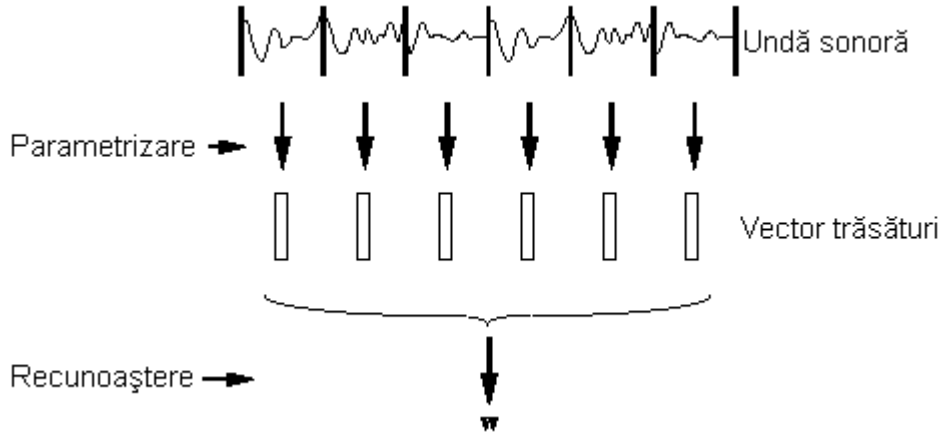

Figura I.6 Modul de parametrizare a semnalului pentru recunoasterea unui cuvânt

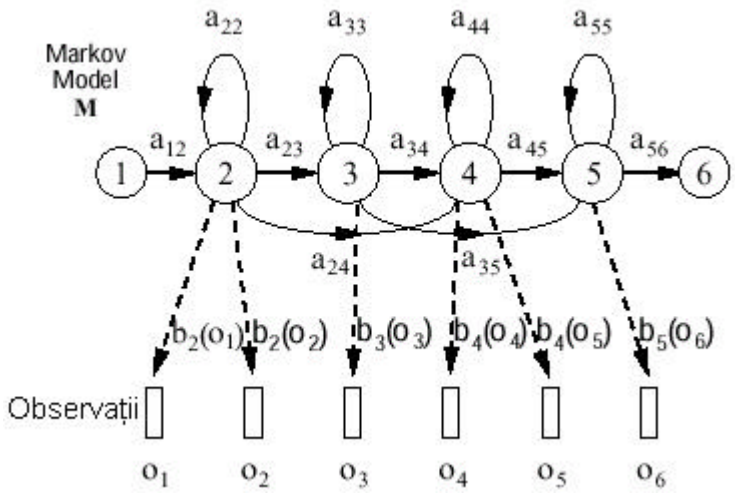

Figura I.7 Structura unui lant Markov

I.3.2.2 Modele hibride HMM/ANN

Sistemele de recunoastere hibride au fost dezvoltate în centrele de cercetare de la Cambridge University [Robinson94], ICSI [Bourland94] si OGI [Hosom98]. Principala diferenta între modelul standard si cel hibrid consta în estimarea probabilitatilor de observatie. În sistemele standard, aceste probabilitati sunt estimate folosind combinatii de Gaussiene. În cele hibride, probabilitatile de observatie sunt estimate folosind retele neuronale. Acestea au câteva avantaje fata de modelele cu combinatii de Gaussiene pentru ca au mai bune proprietati discriminative, nu impun ca trasaturile de intrare sa fie necorelate si nu cer ca datele sa se potriveasa cu modele Gaussiene [Bourland92].

Dezavantajele modelelor hibride sunt urmatoarele:

? durata mare de timp necesara pentru antrenarea clasificatorului. Aceasta crestere a timpului de antrenare este uneori compensata de scaderea timpului necesar pentru estimarea probabilitatilor de observatie în timpul recunoasterii.

? pentru a modifica proprietatile modelelor, trebuie sa se faca o noua antrenare a retelei. La combinatiile de Gaussiene, modelele fonetice pot fi ajustate individual dupa antrenare.
? antrenarea ANN necesita o buna estimare a locatiilor fiecarui fonem. Modelul HMM standard poate fi antrenat fara astfel de informatie desi performantele acestuia pot beneficia în urma transcrierii manuale prealabile.

I.3.2.3 Sisteme de recunoastere bazate pe silaba

Acest sistem a fost dezvoltat de Su-Lin Wu, Steven Greenberg si colegi lor de la International Computer Science Institute (ICSI) pornind de la realitatile psihoacustice si anume faptul ca silaba este importanta în procesul de perceptie al vorbirii [Wu97]. Intrarea la aceste sisteme este o “spectrograma modulata” care reprezinta variatia în timp a continutul spectral al semnalului vocal, obtinut prin filtrarea trece jos sub 10 Hz [Greenberg97]. Aceste benzi pun în evidenta evolutia semnalului de-a lungul silabei, suprimând modificarile cele mai lente si cele mai rapide. Spectrograma astfel modulata este aplicata la intrarea unei retele neuronale într-o fereastra de 185 msec. La iesirea retelei se obtin ponderi pentru un numar de 124 categorii de semisilabe. Folosind o cautare Viterbi pentru iesirile retelei se gaseste cea care corespunde cel mai mult cu o anumita secventa.

Performantele sistemului bazat pe silaba (90,2% cuvinte corect recunoscute pentru semnal de pe linia telefonica) nu sunt atât de bune ca ale sistemelor de recunoastere bazate pe fonem (93,2%) dar combinarea acestora dau o performanta mai buna (94,5%) decât în folosirea lor separata. Aceste rezultate indica faptul ca tipurile de erori facute de fiecare sistem sunt într-o oarecare masura independente si ca cercetarile continuate în acest domeniu sunt fructuoase.