'olitehnica" din Timişoara
Automatică şi Calculatoare

CONTRIBUŢII LA RECUNOAŞTEREA AUTOMATĂ A VORBIRII CONTINUE ÎN LIMBA ROMÂNĂ

TEZĂ DE DOCTORAT

ing. Marian Boldea

Timişoara, 2003

Universitatea "Politehnica" din Timişoara
Facultatea de Automatică şi Calculatoare

CONTRIBUŢII LA RECUNOAŞTEREA AUTOMATĂ A VORBIRII CONTINUE ÎN LIMBA ROMÂNĂ

TEZĂ DE DOCTORAT

ing. Marian Boldea

Timişoara, 2003

Sumar

Această teză este bazată pe o serie de cercetări în direcţia recunoaşterii automate a vorbirii în limba română, cercetări în care principala problemă abordată a fost cea a modelării acustice folosind unităţi sublexicale pentru recunoaşterea vorbirii continue cu vocabulare mari şi foarte mari.

Pentru început, teza îşi precizează cadrul teoretic printr-o trecere în revistă a unor metode de bază în analiza şi recunoaşterea vorbirii.

Este prezentată apoi proiectarea şi colectarea primei baze de date fonetice în limba română, destinată în primul rând cercetărilor în direcţia recunoaşterii automate independentă de vorbitor a vorbirii continue.

Urmează etichetarea acestei baze de date la nivel fonetic utilizând un sistem dezvoltat în acest scop, etichetare menită să faciliteze folosirea ei în cercetările asupra recunoaşterii automate a vorbirii şi în alte domenii.

În final sunt evaluate două posibile seturi de unităţi sublexicale de modelare acustică, evaluare desfăşurată prin experimente de recunoaştere dependentă şi independentă de vocabular a unor semnale din baza de date.

Cercetările au fost finanţate parţial, prin opt granturi şi contracte, de Academia Română, Comisia Europeană, Consiliul Naţional al Cercetării Ştiinţifice din Învăţământul Superior (fost Consiliul Naţional al Cercetării Ştiinţifice Universitare) şi fostul Minister al Cercetării şi Tehnologiei.

CUPRINS

Mulţumiri

În primul rând, mulţumesc domnului profesor Crişan Strugaru, conducătorul meu de doctorat, pentru îndrumările şi sprijinul moral şi material pe care mi le-a acordat în permanenţă de-a lungul activităţilor de cercetare pe care această teză le valorifică.

Apoi, pentru rolul decisiv pe care încrederea şi ajutorul lor l-au avut în conturarea conţinutului tezei, câtorva colegi din proiectul european BABEL: dr. Lori Lamel de la LIMSI-CNRS, Orsay-Paris, profesorul Peter Roach de la Universitatea din Reading şi profesorul William Barry de la Universitatea din Saarbruecken.

De asemeni, mulţumesc profesorului Renato De Mori, cu sprijinul şi sub îndrumarea căruia am efectuat un stagiu de cercetare de şase luni pe probleme ale recunoaşterii automate a vorbirii la Laboratorul de Informatică al Universităţii din Avignon.

Esenţiale au fost mijloacele financiare cu care cercetările au fost susţinute de către Academia Română, Comisia Europeană, Consiliul Naţional al Cercetării Ştiinţifice din Învăţământul Superior (fost Consiliul Naţional al Cercetării Ştiinţifice Universitare) şi fostul Minister al Cercetării şi Tehnologiei.

Cu toate acestea, cele mai multe dintre rezultatele prezentate nu ar fi fost posibile fără ajutorul unei echipe din care de-a lungul anilor au făcut parte câţiva foşti studenţi: Alin Doroga, Tiberiu Dumitrescu, Maria Pescaru, Cosmin Munteanu.

Foarte importante au fost de asemeni sprijinul domnului profesor Nicolae Robu pentru realizarea unor înregistrări de calitate în studioul TeleUniversităţii, ca şi înţelegerea colegilor de acolo, care au suportat timp de aproximativ doi ani problemele astfel apărute.

Calitatea înregistrărilor a fost asigurată şi prin ajutorul domnului Constantin Nanasi, proiectantul şi constructorul preamplificatorului liniar utilizat pentru realizarea lor.

La fel de important a fost şi ajutorul vorbitorilor înregistraţi: deşi numele fiecăruia există notat undeva, sunt totuşi prea mulţi (o sută) pentru a fi enumeraţi aici.

În sfârşit, mulţumiri amestecate cu scuze tuturor celor pe care nu i-am numit dar care, într-un fel sau altul, mai mult sau mai puţin, m-au sprijinit sau stimulat.

CAPITOLUL 1

Introducere

Vorbirea constituind un element distinctiv al speciei şi forma cea mai naturală de comunicare pentru fiinţele umane, este normal ca studiul ei să prezinte interes pentru o mare varietate de discipline, începând cu anatomia şi fiziologia, trecând prin cele lingvistice (fonetică, fonologie) şi terminând cu cele tehnice, interesate de prelucrarea vorbirii în diverse scopuri aplicative. În condiţiile în care actualmente are loc o adevărată fuziune între sistemele de calcul şi cele de comunicaţii, cu efecte revoluţionare asupra modului în care oamenii comunică [196], [73], [54], prelucrarea automată a vorbirii capătă o importanţă deosebită, iar disciplinele tehnice, în calitate de solicitante şi beneficiare ale unor rezultate din domeniile fundamentale, devin motorul cercetărilor asupra vorbirii.

În ansamblul aplicaţiilor tehnice bazate pe prelucrarea automată a vorbirii putem distinge ca domenii de bază:

- **analiza semnalului vocal**, care are un rol fundamental în raport cu toate celelalte domenii, urmărind caracterizarea lui prin extragerea unor parametri adecvaţi prelucrărilor ulterioare [200], [232], [59];
- **îmbunătăţirea calităţii** prin reducerea efectelor zgomotelor şi distorsiunilor, necesară atât pentru facilitarea comunicării între oameni, cât şi ca o etapă preliminară altor prelucrări sau aplicaţii mai complexe [59];
- **codarea semnalului vocal**, având ca scop obţinerea unei reprezentări cât mai compacte a acestuia în vederea stocării sau transmiterii, simultan cu păstrarea unei cât mai bune calităţi a semnalului refăcut pe baza acestei reprezentări [227], [111];
- **sinteza semnalului vocal**, aflată în strânsă legătură cu codarea: în sens restrâns, permite refacerea dintr-o reprezentare codată a unui semnal; în sens mai larg, ea utilizează reprezentări textuale sau conceptuale ale unor mesaje pentru transpunerea lor în formă sonoră [200], [2], [66], [242];
- **identificarea limbii** vorbite de o persoană, care poate servi pentru sinteza unor mesaje de răspuns adecvate, activării unui sistem de recunoaştere corespunzător, sau punerii în legătură cu un vorbitor al aceleiaşi limbi [165];

- **recunoaşterea vorbitorului**, putând urmări **identificarea** acestuia dintre mai multe persoane, de exemplu în expertize criminalistice, sau **verificarea** identităţii pe care acesta o pretinde, cu aplicaţii în controlul accesului [62], [179], [92];

- **recunoaşterea automată a vorbirii**, având ca scop determinarea cu o cât mai mare exactitate a şirului de cuvinte pronunţat de un vorbitor, şi utilă în aplicaţii de tip comandă-şi-control, dictare automată etc. [194], [115], [58];

- **înţelegerea vorbirii**, care are ca obiectiv nu o determinare exactă a cuvintelor pronunţate, ci a semnificaţiei acestora, astfel încât să se poată efectua în continuare o serie de acţiuni de răspuns corecte din punctul de vedere al aplicaţiei [1], [156].

Deşi într-o formă sau alta toate domeniile enumerate se bazează pe utilizarea unor sisteme de calcul de o complexitate mai mare sau mai mică, din punctul de vedere al facilitării interacţiunii acestor sisteme cu utilizatorii, imperios necesară în perspectiva unei societăţi informaţionale, recunoaşterea, înţelegerea şi sinteza automată a vorbirii sunt esenţiale, iar teza de faţă prezintă o serie de cercetări în direcţia recunoaşterii automate a vorbirii continue care, împreună cu altele urmărind înţelegerea [24], sinteza ei automată din text [65], [185], şi integrarea tuturor acestora în sisteme de dialog [164], [163] vizează dezvoltarea de interfeţe vocale om-maşină în limba română [28].

Problemele recunoaşterii automate a vorbirii

Dificultatea recunoaşterii automate a vorbirii este influenţată de multe variabile, cele mai importante fiind:

- **tipul pronunţiei**: continuă sau discretă (cu pauze între cuvinte);

- **stilul de vorbire**: citită, spontană sau semispontană (ca răspuns la o cerere);

- **numărul de vorbitori** ale căror pronunţii trebuie recunoscute, în funcţie de acesta sistemele de recunoaştere automată a vorbirii putând fi clasificate în **dependente** de, **independente** de, sau **adaptive** la vorbitor ;

- **dimensiunea vocabularului**, dată de numărul de cuvinte pe care sistemul le poate recunoaşte şi care poate fi **mică** (Nx10), **medie** (Nx100), **mare** (Nx1000) sau **foarte mare** (Nx10000);

- **condiţiile de mediu** care se reflectă asupra caracteristicilor semnalului: curat, zgomotos, distorsionat etc.

Aceste variabile contribuie în moduri specifice la dificultatea problemei, iar împreună au făcut ca, deşi începute aproape imediat după inventarea calculatoarelor digitale, cercetările asupra recunoaşterii automate a vorbirii să ajungă doar în zilele noastre la stadiul în care, pentru câteva limbi, există disponibile comercial sisteme de recunoaştere a vorbirii continue cu vocabulare foarte mari. Pentru o mai bună imagine globală asupra dificultăţilor care au trebuit depăşite, să amintim că timpul necesar ajungerii oamenilor pe Lună a fost de câteva ori mai scurt...

1.1 Scurtă istorie

Pe plan mondial, cercetările asupra recunoaşterii automate a vorbirii au început aproape imediat după inventarea calculatoarelor digitale (1947, cf. [189]), dar primele cercetări atestate prin articole publicate au avut loc în anii '50 [189], [194] şi au urmărit recunoaşterea unor sunete sau a unor cuvinte izolate folosind **analiza spectrală** cu filtre analogice şi idei din **fonetica acustică**.

Anii '60 au adus recunoaşterea cuvintelor izolate, bazată pe existenţa unor tipare ale acestora, prin **deformarea dinamică a timpului** (în engleză **dynamic time warping**) ([246], citat de exemplu în [245]), şi s-a abordat recunoaşterea vorbirii continue, dar rezultatele cu cel mai mare impact ulterior au apărut printre metodele de analiză a semnalului vocal: **analiza homomorfică** [174] şi cea prin **predicţie liniară** [7].

De-abia în anii '70 s-au pus fundamente teoretice solide pentru tratarea problemei, bazate pe recunoaşterea formelor, inteligenţa artificială şi teoria comunicaţiei. Au fost introduse **distanţe spectrale** [110] cu interpretări psihofiziologice clare în deformarea dinamică a timpului, iar metoda ca atare a fost extinsă de la recunoaşterea cuvintelor izolate la cea a cuvintelor conectate [212]. Abordări inspirate de **inteligenţa artificială**, cunoscute şi ca **bazate pe cunoştinţe** (**knowledge based**), au fost materializate în **sisteme expert** în care **surse de cunoştinţe** (fonetice, fonologice, lexicale, sintactice, semantice şi pragmatice), încorporând cunoştinţe ale unor experţi umani, cooperează pentru recunoaşterea pronunţiilor din semnalele prelucrate [257], [141]. Însă cea mai importantă a fost abordarea recunoaşterii vorbirii ca problemă de **teoria comunicaţiei** [116], ceea ce a permis utilizarea unor **metode statistice** [15], [113], cu un fundament matematic riguros, pentru tratarea ei. Instrumentele emblematice ale acestei abordări sunt **modelele lingvistice** de tip **n-gram**, şi **modelele Markov ascunse – MMA** (în literatura de limbă engleză, **hidden Markov models – HMM**), care stau la baza tuturor sistemelor moderne de recunoaştere automată a vorbirii.

Cele mai importante evoluţii au avut loc însă în anii '80. Aşa cum am menţionat deja, există o gamă largă de metode pentru abordarea recunoaşterii automate a vorbirii, dar mulţimea factorilor care influenţează semnalul vocal, determinând caracterul său cvasialeator, face ca problema să fie una extrem de complexă, care nu admite soluţii analitice, aşa încât orice nouă idee, oricât de bine justificată din punct de vedere teoretic, nu poate fi validată decât prin experimente cât se poate de cuprinzătoare.

Au avut astfel loc cercetări asupra unor noi algoritmi de recunoaştere a cuvintelor conectate prin deformarea dinamică a timpului [166], [38] şi a utilizării modelelor Markov ascunse pentru recunoaşterea cuvintelor izolate [199], [197] sau conectate [198], [201] cu vocabulare cât mai mari [101], [10] şi a vorbirii continue dependentă [50] de, adaptivă [219] la, sau independentă [137] de vorbitor. Atingerea acestor obiective a presupus însă cercetări vizând mult mai multe domenii conexe, ca exemple putând fi menţionate construcţia unor baze de date vocale de dimensiuni suficiente pentru utilizarea metodelor statistice şi compararea diferitelor abordări în condiţii identice [132], [192], noi variante de modelare statistică folosind MMA [11], [209], [106], utilizarea unor unităţi acustice sublexicale [207], [255] şi modelarea lor dependentă de context [218] etc.

Principalul rezultat a fost tranşarea disputelor între diferitele abordări în favoarea metodelor statistice, a căror maturizare a dus în anii '90 la apariţia pentru câteva limbi a

unor sisteme comerciale rulând pe calculatoare uzuale, dar folosind microfoane speciale, de mică distanţă, capabile să recunoască vorbirea continuă cu vocabulare mari şi foarte mari şi cu o frecvenţă a erorilor suficient de redusă pentru a fi utilizabile în aplicaţii.

Pe plan naţional, interesul pentru cercetările asupra prelucrării automate a vorbirii s-a manifestat în special în cadrul catedrelor de electronică, automatică şi calculatoare din unele centre de învăţământ superior: Bucureşti [63], Iaşi [235], Timişoara [75]. Ca în multe alte ţări, un interes aparte pentru aceste cercetări au manifestat şi instituţiile militare de învăţământ superior şi cercetare [173], [232], [154].

Preocupările în direcţia recunoaşterii automate a vorbirii au existat încă din anii '60 ([172], citat în [63]), dar în condiţiile izolării de comunitatea ştiinţifică internaţională şi a lipsei unei baze materiale corespunzătoare, ele au continuat doar sporadic, graţie entuziasmului unui număr redus de cercetători, şi au fost limitate la recunoaşterea de sunete [100] sau cuvinte izolate [43], [52], [238].

După 1989, aceste cercetări au cunoscut o relativă dezvoltare prin adoptarea unor metode moderne: cuantizarea vectorială [26], [93], deformarea dinamică a timpului [93], [109], modelele Markov ascunse [26], [93], reţelele neuronale [236], [98], metodele hibride combinând modelele Markov ascunse cu reţelele neuronale sau tehnicile fuzzy [248]. Cu toate acestea, ele au rămas în continuare limitate la nivelul sunetelor sau al cuvintelor izolate, fără a aborda recunoaşterea vorbirii continue.

1.2 Stadiul actual

Aşa cum am precizat, cercetările desfăşurate pe parcursul a aproape jumătate de secol au atins în anii '90 stadiul valorificării comerciale, incluzând două direcţii principale.

Prima vizează sistemele de recunoaştere automată a vorbirii continue cu vocabulare foarte mari, de zeci de mii de cuvinte, având ca aplicaţie tipică dictarea de documente. Dintre aceste sisteme putem menţiona NaturallySpeaking, primul disponibil comercial, în 1997, produs de firma Dragon Systems din S.U.A., şi ViaVoice, produs de IBM. Asemenea sisteme sunt livrate cu modele acustice independente de vorbitor şi modele lingvistice generale sau specializate pentru un anumit domeniu, dar adaptabile la particularităţile vorbitorilor care le utilizează şi ale documentelor dictate.

A doua direcţie este cea a sistemelor de dialog om-calculator, dedicate pentru anumite aplicaţii, cel mai frecvent servicii în domenii bine definite (un exemplu tipic fiind cel al informaţiilor, rezervărilor şi vânzărilor de bilete prin telefon pentru mijloace publice de transport). Datorită numărului mare de utilizatori potenţiali şi a necesităţii de a furniza cât mai rapid serviciul cerut, acestea utilizează modele acustice independente de vorbitor, dificil de adaptat în condiţiile unor interacţiuni de scurtă durată, dar au avantajul unor vocabulare limitate de aplicaţii, de ordinul sutelor sau miilor de cuvinte.

Iniţial, aceste două direcţii vizau cu precădere limba engleză, asupra căreia au fost concentrate eforturi de cercetare deosebite în S.U.A., în cadrul unor programe speciale ale Defense Advanced Research Projects Agency (DARPA), dar ulterior au apărut sisteme similare şi pentru alte limbi, în general dintre cele cu suficient potenţial comercial.

Indiferent însă de caracteristicile şi domeniul lor de utilizare, sistemele contemporane de recunoaştere automată a vorbirii pot fi descrise prin schema bloc din figura 1.1.

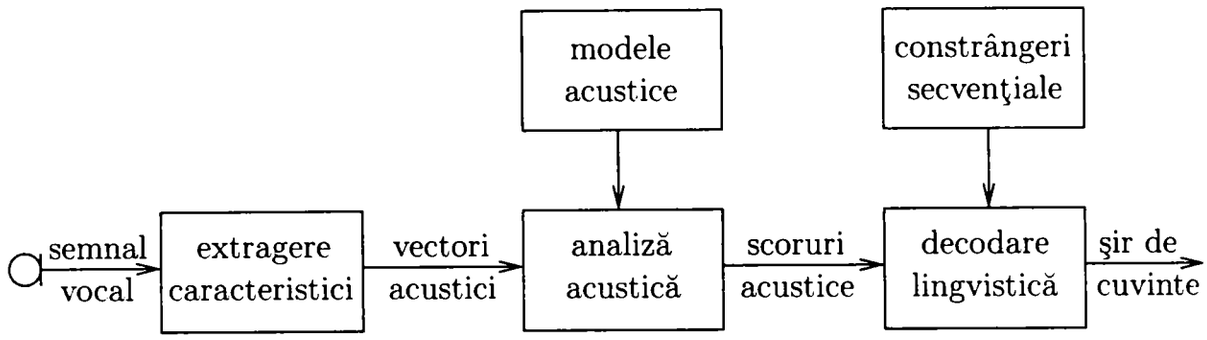

Figura 1.1: Schema bloc tipică a unui sistem de recunoaştere automată a vorbirii

Primul bloc, de **extragere a caracteristicilor**, transformă semnalul vocal rezultat dintr-o pronunţie într-un şir de **vectori acustici** care îl caracterizează pe intervale scurte, în conformitate cu natura sa nestaţionară.

După obţinerea şirului de vectori acustici, **recunoaşterea** propriu-zisă presupune rezolvarea a două probleme, tratate în blocurile următoare:

- **analiza acustică** a porţiunii de semnal corespunzătoare unui vector în raport cu nişte **modele acustice**, concretizată în calculul unor **scoruri acustice**: funcţie de metoda de recunoaştere utilizată şi tipul modelelor acustice, scorurile pot fi **distanţe** între vectorul curent şi cei din modele, sau **probabilităţi** condiţionate ca vectorul să fi apărut în urma pronunţiei unităţilor modelate;

- **decodarea lingvistică** prin alinierea vectorilor acustici cu modelele asociate unui şir de cuvinte folosind algoritmi de **căutare**; aceasta urmăreşte obţinerea unei estimări optime a şirului de cuvinte prin pronunţarea căruia s-a obţinut semnalul, şi poate include **constrângeri secvenţiale**: la nivelul cuvintelor, acestea pot fi impuse prin **dicţionare de pronunţii** ale cuvintelor din vocabularul sistemului în termenii eventualelor unităţi de modelare acustică sublexicală, iar la nivelul întregii pronunţii, prin **gramatici** sau **modele lingvistice statistice** ale aplicaţiei.

În cadrul oferit de această schemă-bloc, cercetările continuă în prezent în diferite direcţii: metode robuste de extragere a caracteristicilor; modelare acustică cu acurateţe sporită, independentă de vocabular, şi utilizând cât mai eficient datele disponibile; adaptarea modelelor acustice şi lingvistice; modelarea cuvintelor noi şi a variaţiilor de pronunţie; reducerea complexităţii algoritmilor de căutare etc.

O direcţie aparte a cercetărilor, în care se înscriu şi cele descrise în această teză, urmăreşte extinderea aplicării metodelor existente la noi limbi şi/sau dialecte, fie prin construcţia şi utilizarea unor noi resurse lingvistice adecvate (baze de date vocale, pentru modelarea acustică, şi arhive de texte în format electronic, pentru cea lingvistică), fie prin adaptarea unor modele acustice din alte limbi.

1.3 Obiectivele cercetărilor

Ţinând cont de nivelul mondial şi naţional, **obiectivul principal** al cercetărilor a fost **recunoaşterea vorbirii continue în limba română, independentă de vorbitor, cu vocabulare în jurul a 1000 cuvinte**. Deoarece aceasta a fost (din câte cunoaştem) **prima abordare a recunoaşterii automate a vorbirii continue în limba română**, acest obiectiv avea avantajele de a fi atât realist cât şi semnificativ în condiţiile date.

Pentru recunoaşterea vorbirii continue cu vocabulare mari şi foarte mari, de mii sau zeci de mii de cuvinte, esenţială este **problema unităţilor de modelare acustică**: daca pentru vocabulare mici, de zeci de cuvinte, fiecare cuvânt poate fi modelat separat, utilizând un număr de pronunţii ale sale, odată cu creşterea mărimii vocabularelor creşte şi cantitatea de date necesare antrenării de modele ale cuvintelor, aceasta putând deveni prohibitivă pentru vocabulare mari şi foarte mari. Soluţia consta în utilizarea unor unitaţi de modelare acustică sublexicală (silabe, semisilabe, sunete etc.), în număr mult mai mic decât cel al cuvintelor din vocabular, astfel încât cantitatea de date necesare antrenării modelelor să fie mult mai redusă. În plus, devine astfel posibilă utilizarea unor vocabulare flexibile, incluzând cuvinte inexistente în datele de antrenament.

Plecând de la aceste consideraţii, au fost stabilite ca **obiective intermediare** în vederea atingerii obiectivului principal enunţat mai sus:

- **proiectarea şi colectarea unei baze de date vocale** corespunzătoare, care să poată fi utilizată pentru a antrena modele acustice sublexicale;
- **etichetarea bazei de date** pentru a facilita antrenarea unor modele acustice sublexicale şi evaluarea performanţelor la acest nivel;
- **studiul unor unităţi de modelare acustică sublexicală** prin experimente de recunoaştere independentă de vorbitor a vorbirii continue.

1.4 Organizarea lucrării

Pentru început, teza realizează o sinteză a cadrului teoretic în care se plasează prin trecerea în revistă a unor metode de analiză a semnalului vocal (capitolul 2) şi a unora utilizate în recunoaşterea automată a vorbirii (capitolul 3).

În continuare sunt descrise cercetările care stau la baza tezei: proiectarea şi colectarea primei baze de date fonetice în limba română, permiţând cercetări asupra recunoaşterii vorbirii continue, independentă de vorbitor, cu vocabulare în jurul a 1000 de cuvinte (capitolul 4); etichetarea la nivel fonetic a celei mai mari părţi din această bază de date, realizată folosind un sistem automat, dezvoltat ad-hoc, de aliniere a semnalelor cu transcrierile lor fonetice (capitolul 5); studii asupra unor unităţi de modelare acustică pentru recunoaşterea independentă de vorbitor a vorbirii continue în limba română, desfăşurate prin experimente folosind subseturi din baza de date (capitolul 6).

În încheiere, capitolul 7 sintetizează contribuţiile lucrării şi trasează câteva direcţii în care cercetările ar putea fi continuate.

CAPITOLUL 2

Analiza semnalului vocal

Nestaţionaritatea semnalului vocal, cunoscută din literatura de specialitate [200], [232] şi ilustrată şi prin exemplele din acest capitol, impune definiţii ale parametrilor săi bazate pe proprietăţi locale ale acestuia. Estimarea valorilor acestor parametri se poate face prin metode de **analiză pe termen scurt**, iar în continuare le vom trece în revistă pe cele mai importante din punctul de vedere al recunoaşterii automate a vorbirii.

Prelucrările în domeniul timp [200], [180], [59] operează direct asupra eşantioanelor semnalului, iar unele au ca rezultate valori ale unor parametri ai acestuia. Parametrii prezentaţi în acest capitol au aplicaţii mergând de la delimitarea porţiunilor de semnal corespunzătoare vorbirii (energia şi puterea) până la recunoaşterea ei propriu-zisă.

Deşi prelucrările în domeniul timp asigură extragerea unor caracteristici importante ale semnalelor vocale, se consideră că pentru recunoaşterea automată a vorbirii esenţiale sunt caracteristicile lor spectrale, cunoscută fiind capacitatea auzului uman de a distinge sunete după conţinutul lor în componenete de diferite frecvenţe şi amplitudini [160], [184], [180], [3]. Ele pot fi extrase folosind tehnici clasice din prelucrarea semnalelor (filtrare, analiză Fourier [175]), pot fi derivate din analiza prin predicţie liniară [153], sau cu luarea în consideraţie a unor particularităţi funcţionale ale aparatului auditiv uman. Între aceste metode, fundamentală este analiza Fourier pe termen scurt, a cărei prezentare oferă ocazia şi pentru discutarea unor legături între domeniile timp şi frecvenţă.

Problema esenţială a analizei spectrale a vorbirii pentru recunoaşterea ei automată este separarea informaţiei minime necesară în acest scop de variaţiile nesemnificative, dar nu este încă foarte clar cum se poate realiza această separare. O primă posibilitate este sugerată de interpretarea spectrală a predicţiei liniare. O altă posibilitate este analiza homomorfică sau cepstrală, care poate fi utilizată pentru a reţine doar informaţia referitoare la aspectul general al spectrului, determinat de forma tractului vocal, cu eliminarea detaliilor corespunzătoare excitaţiei acestuia.

Includerea în cadrul analizei semnalului vocal a unor prelucrări motivate pe baza unor particularităţi funcţionale ale aparatului auditiv uman este ilustrată în finalul capitolului prin analiza melodică şi caracteristicile dinamice.

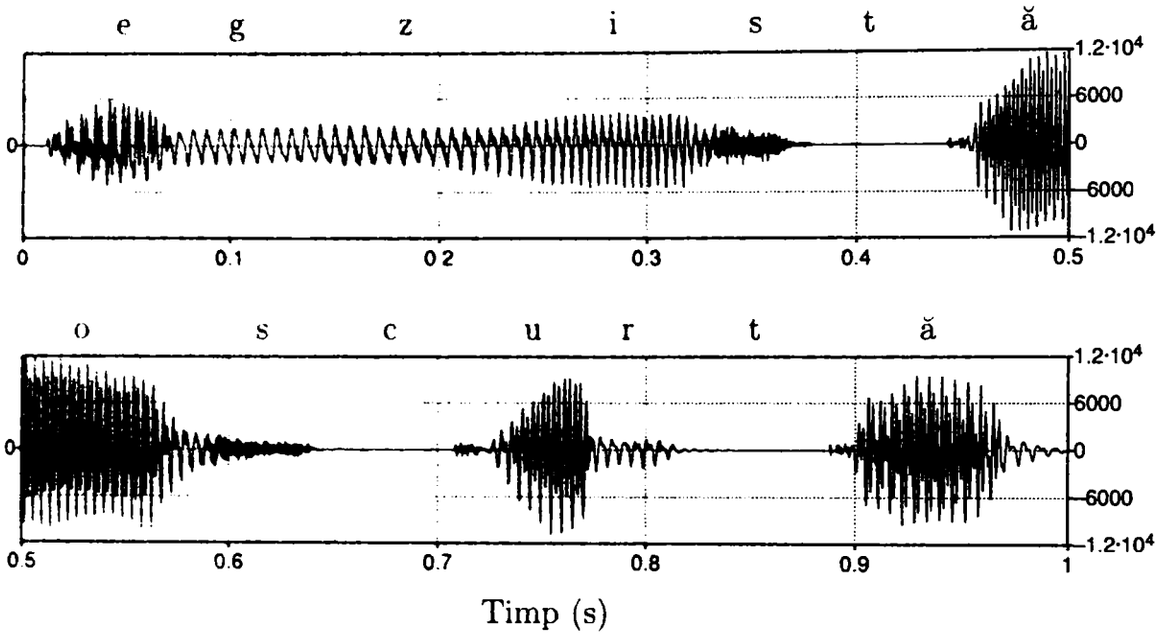

Figura 2.1: Exemplu de formă de undă a unui semnal vocal (cuvintele "există o scurtă", cu indicarea sunetelor efectiv pronunţate şi a poziţiilor lor)

2.1 Reprezentări ale semnalului vocal

Reprezentarea primară a unui semnal vocal este **forma de undă**: în continuare vom utiliza pentru diferite exemplificări înregistrarea numerică, eşantionată la 16 KHz cu o rezoluţie de 16 biţi, a pronunţiei de către un bărbat a cuvintelor "există o scurtă", iar figura 2.1 prezintă forma ei de undă şi sunetele efectiv pronunţate. Se observă că aspectul formei de undă variază foarte mult funcţie de sunet, de la amplitudini foarte mici în timpul "t" şi "c" şi aspect aleator pentru "s", până la amplitudini maxime şi aspect cvasiperiodic pentru vocale, de unde necesitatea analizei pe termen scurt.

O reprezentare care evidenţiază simultan evoluţia amplitudinii şi a caracteristicilor spectrale ale unui semnal este **spectrograma** lui, care poate fi obţinută prin diferite metode de estimare spectrală. În spectrogramă, intensitatea componentei de o anumită frecvenţă a semnalului la un moment dat este indicată prin nivelul de gri sau culoarea punctului de coordonate corespunzătoare. În continuare se vor utiliza spectrograme cu niveluri de gri, în care nuanţele mai închise indică intensităţi mai mari.

Pentru exemplificare, figura 2.2 prezintă două spectrograme ale semnalului din figura 2.1. Funcţie de "termenul scurt" de analiză Δt (v. secţiunile 2.2.1 şi 2.5), invers proporţional cu rezoluţia în timp şi direct proporţional cu cea în frecvenţă, spectrogramele pot fi de **bandă largă**, cu rezoluţie redusă în frecvenţă şi mare în timp, respectiv de **bandă îngustă**, cu rezoluţie mare în frecvenţă dar redusă în timp.

În cazul spectrogramei de **bandă îngustă** din figura 2.2, obţinută analizând porţiuni de câte 30 ms, caracterul cvasiperiodic al semnalului pe durata vocalelor şi a unor consoane se manifestă ca **striaţiuni orizontale**, corespunzătoare armonicelor frecvenţei fundamentale, F_0, inversa perioadei fundamentale, T_0, ambele variabile în timp. În cazul

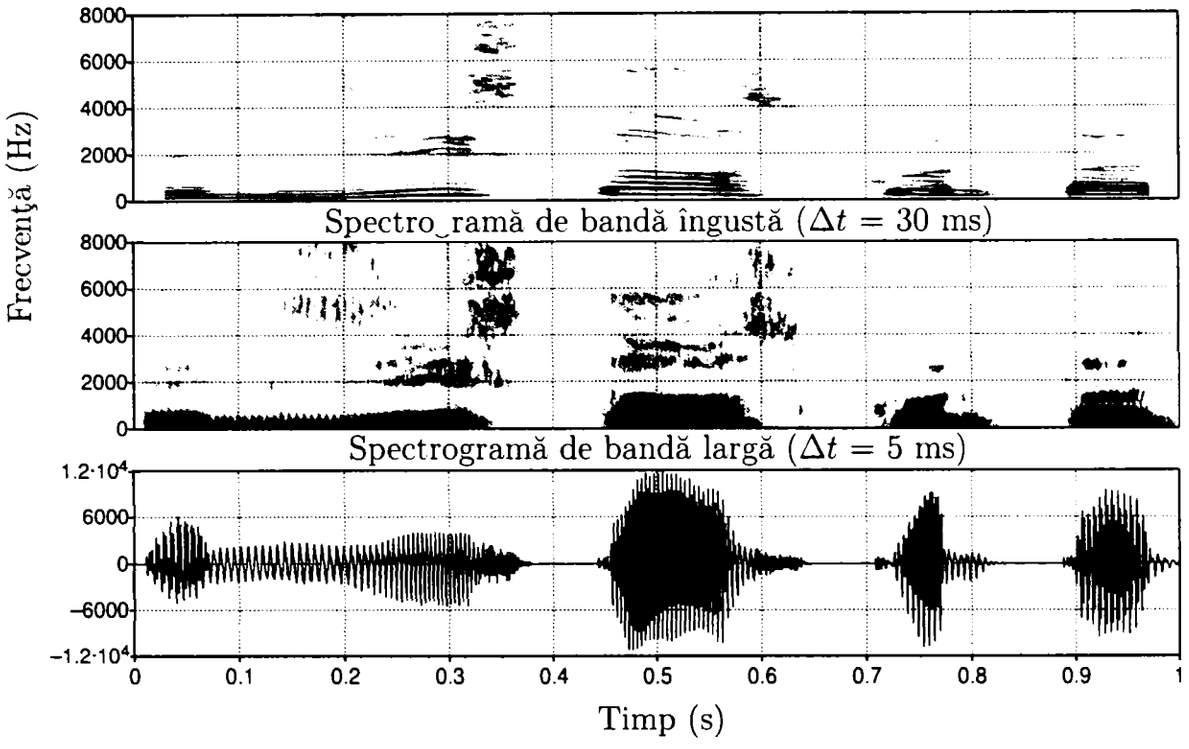

Figura 2.2: Spectrograme de bandă îngustă şi largă ale semnalului din figura 2.1

spectrogramei de **bandă largă**, cu intervalul de analiză de 5 ms, cvasiperiodicitatea se manifestă sub forma unor **striaţiuni verticale**, corespunzătoare variaţiilor energiei semnalului din intervalul analizat cu poziţia acestui interval.

Indiferent de tipul spectrogramei, se constată existenţa unor zone de intensitate sporită, situate la diferite frecvenţe pe durata diferitor sunete. Aceste zone corespund aşa-numiţilor **formanţi**, care sunt rezonanţe ale unor cavităţi din tractul vocal (faringele, cavitatea bucală, cavitatea nazală) prin care oscilaţiile de presiune acustică se propagă spre exterior pe durata producerii sunetelor vorbirii. Pentru acelaşi semnal din figura 2.1, figura 2.3 prezintă rezultatul unei estimări automate a valorilor frecvenţelor primilor cinci formanţi, reprezentate prin puncte suprapuse peste spectrograma de bandă largă. Frecvenţele formanţilor, notate în literatură, în ordinea lor crescătoare, cu $F_1, F_2, F_3 \ldots$,

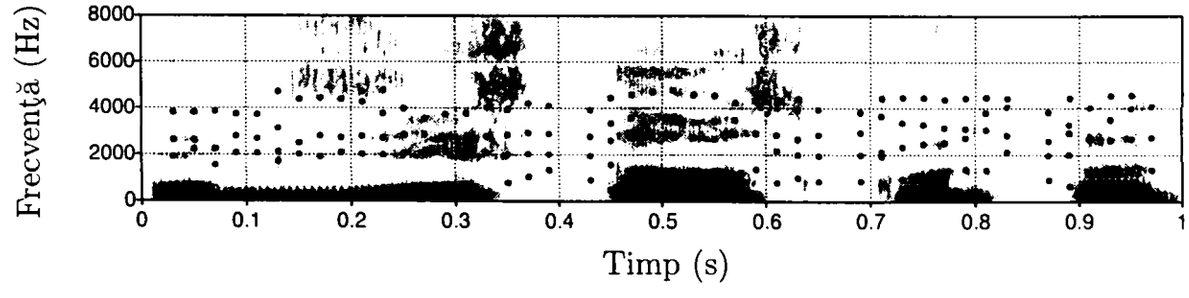

Figura 2.3: Spectrograma de bandă largă şi formanţii semnalului din figura 2.1

ar putea fi o reprezentare foarte compactă a sunetelor vorbirii, iar în fonetică ele sunt folosite pentru a caracteriza din punct de vedere acustic vocalele unei limbi, dar în scopul recunoaşterii automate a vorbirii ele sunt insuficiente din multe motive, cel mai evident fiind acela că există zone ale semnalelor pentru care nu pot fi estimate.

2.2 Prelucrări în domeniul timp

2.2.1 Cadrarea şi ferestruirea

Presupunând că semnalul vocal de prelucrat a fost deja trecut în forma numerică $s(n)$ printr-un proces de eşantionare şi cuantizare, primul pas în aplicarea oricărei metode de analiză pe termen scurt a acestuia este **selectarea intervalelor** de semnal de analizat.

Această operaţiune este supusă unor constrângeri, uneori contradictorii:

- unele intervale trebuie să fie **suficient de scurte** pentru a evidenţia evenimente acustice semnificative cu o durată foarte scurtă, cum ar fi faza de eliberare a aerului din cursul producerii sunetelor plozive, de tipul consoanelor "t" sau "c";
- în alte cazuri, ele trebuie să fie **suficient de lungi** pentru a putea realiza estimarea valorilor unor parametri (de exemplu T_0, F_0) pe baza lor;
- intervalele trebuie să acopere toate eşantioanele semnalului.

Aceste constrângeri ar putea fi satisfăcute utilizând **segmente** de lungime variabilă, însă implementarea acestei idei este dificilă, astfel că în practică se utilizează cel mai adesea **cadre** de lungime fixă N, cu o deplasare $d \leq N$ între două cadre succesive. Acestea includ toate eşantioanele semnalului, fiecare eşantion făcând parte din unul sau mai multe cadre, după cum acestea sunt disjuncte ($d = N$) sau nu ($d < N$), iar fiecare cadru poate fi considerat o secvenţă numerică de lungime finită.

Pentru atenuarea efectului porţiunilor partajate cu cadrele adiacente asupra valorilor parametrilor calculate pe baza unui cadru, eşantioanele dintr-un cadru pot fi ponderate funcţie de poziţia lor (pondere minimă la margini, maximă la centru) prin multiplicarea valorilor lor $s(n)$ cu o **fereastră** $w(n)$, operaţiune cunoscută drept **ferestruire**:

$$s_w(n) = \begin{cases} s(n)w(n), & 0 \leq n \leq N-1 \\ 0, & \text{altfel} \end{cases} \tag{2.1}$$

După cum vom vedea în continuare, pe lângă motivul intuitiv al ponderării efectelor cadrelor adiacente, există şi alte motive pentru aplicarea ferestruirii, în unele cazuri nu numai semnalului vocal, ci şi unor mărimi derivate din acesta. Cel mai important dintre aceste motive este efectul de netezire al ferestruirii, netezire datorată faptului că ferestrele uzuale reprezintă răspunsuri la impuls ale unor filtre trece jos [175].

Dintre ferestrele utilizate în prelucrarea semnalelor, cea mai frecvent folosită pentru prelucrarea automată a semnalului vocal este fereastra de tip Hamming

$$w(n) = 0{,}54 - 0{,}46\cos(2\pi\frac{n}{N-1}), \quad 0 \leq n \leq N-1 \tag{2.2}$$

din motive care vor fi prezentate în secţiunea 2.5, iar un exemplu de multiplicare a unui cadru lung de 20 ms cu o asemenea fereastră este dat în figura 2.4.

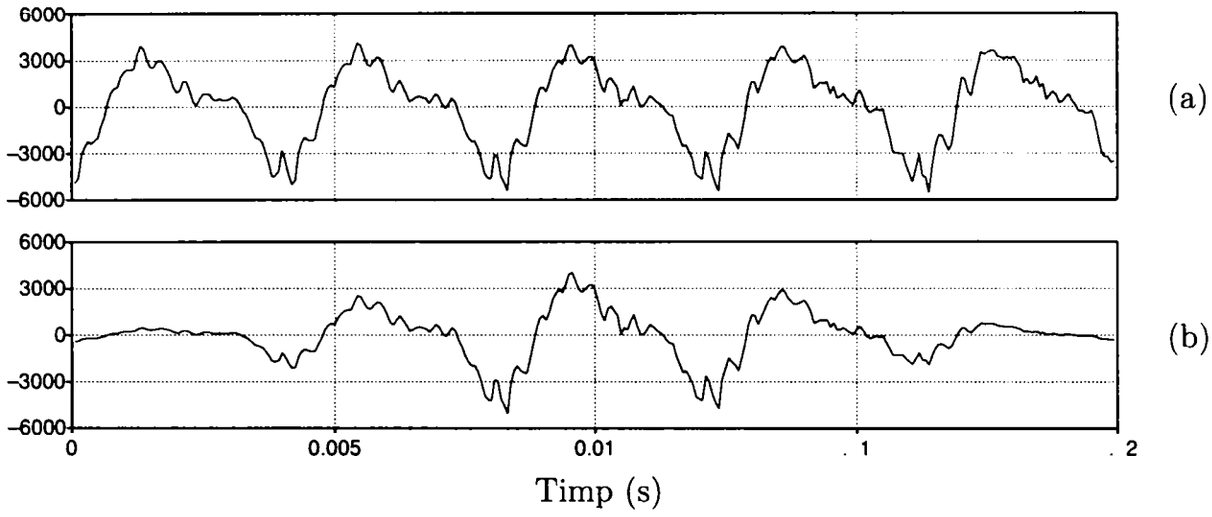

Figura 2.4: Un cadru de 20 ms dintr-un semnal vocal (a) şi rezultatul multiplicării lui cu o fereastră Hamming (b)

2.2.2 Energia şi puterea

Un prim parametru al semnalului vocal este energia, definită pentru un cadru ca

$$E = \sum_{n=0}^{N-1} s^2(n) \tag{2.3}$$

Variante ale acestei definiţii aplică o ferestruire fie eşantioanelor semnalului [213]

$$E_1 = \sum_{n=0}^{N-1} s_w^2(n) \tag{2.4}$$

fie pătratelor lor [180]

$$E_2 = \sum_{n=0}^{N-1} s^2(n)w(n) \tag{2.5}$$

iar prin împărţirea la N a estimărilor energiei se obţin estimări ale puterii medii P.

Energia şi puterea evidenţiază sunetele sonore, vocalele în primul rând, pe durata cărora ating valori maxime. Datorită ridicării la pătrat, gama lor dinamică este prea mare pentru ca reprezentările lor pe o scară liniară să permită distincţii ale sunetelor cu amplitudini reduse, astfel încât de obicei ele sunt reprezentate pe o **scară logaritmică** (în decibeli). Pentru exemplificare, figura 2.5 prezintă pe scară liniară şi logaritmică estimarea puterii medii a semnalului din figura 2.1 folosind cadre de 5 ms.

Reprezentările pe scară logaritmică ale energiei şi puterii pot fi privite ca reprezentări pe scară liniară ale logaritmilor lor, cunoscuţi drept **log-energie** respectiv **log-putere**. Aceştia includ dependenţa logaritmică, în conformitate cu legea Weber-Fechner [228], dintre **intensitatea** obiectivă şi **tăria** subiectivă a unui sunet.

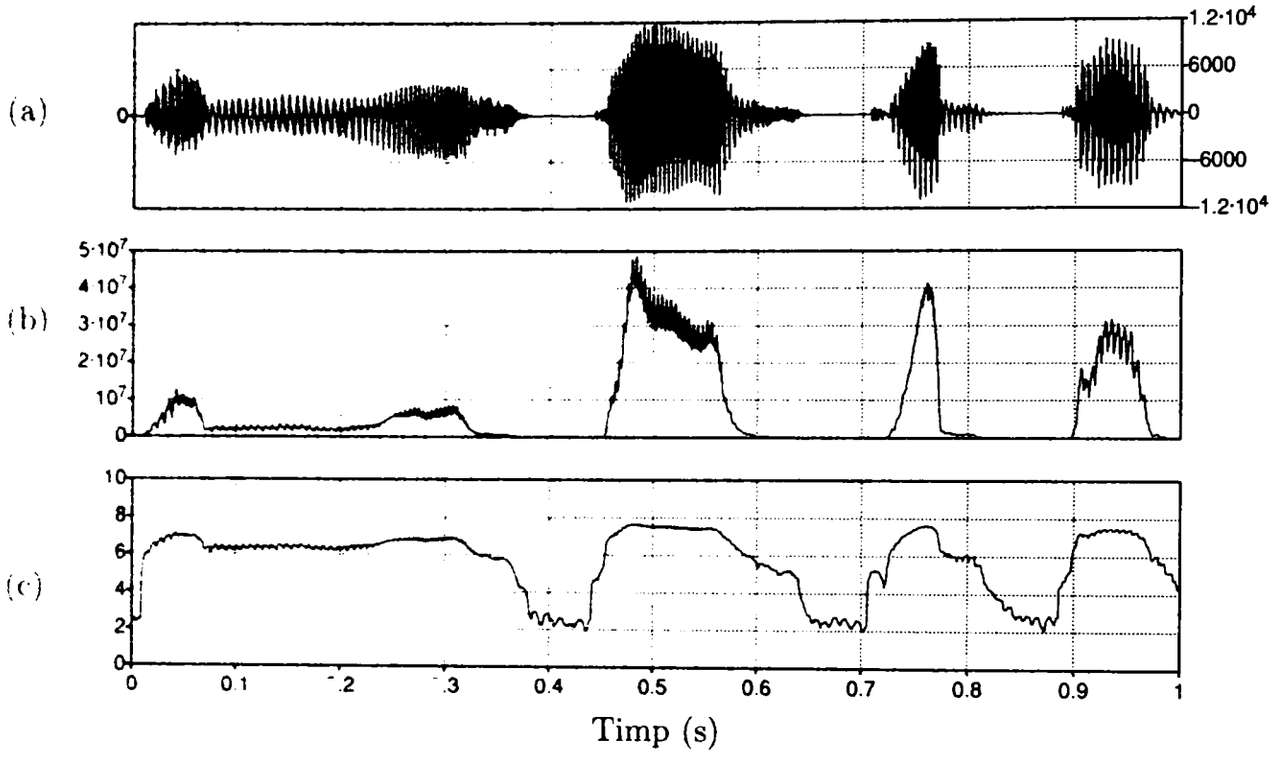

Figura 2.5: Estimare a puterii medii a semnalului din figura 2.1 (a) reprezentată pe scară liniară (b) şi logaritmică (c)

2.2.3 Autocorelaţia

Pentru un semnal numeric, funcţia de autocorelaţie este prin definiţie [56] aşteptarea statistică a produsului semnalului cu o replică a sa deplasată cu m eşantioane

$$R(m) = E[s(n)s(n+m)] \tag{2.6}$$

Pentru un cadru, valoarea ei se poate estima "natural", dar **deplasat**

$$R(m) = \frac{1}{N} \sum_{n=0}^{N-|m|-1} s(n)s(n+|m|) \tag{2.7}$$

sau **nedeplasat**, prin mediere numai peste produsele efectiv folosite

$$R(m) = \frac{1}{N-|m|} \sum_{n=0}^{N-|m|-1} s(n)s(n+|m|) \tag{2.8}$$

După cum uşor se poate demonstra şi matematic, autocorelaţia pune în evidenţă periodicitatea semnalului. Ca exemplu, în figura 2.6 sunt date estimări ale funcţiei de autocorelaţie pentru o porţiune sonoră dintr-un semnal vocal: se observă că intervalele dintre maximele ei succesive au durate aproximativ egale cu perioada fundamentală a semnalului, iar maximele scad (datorită periodicităţii imperfecte a semnalului) mai lent în cazul estimării nedeplasate decât în cel al estimării deplasate, când intervine şi

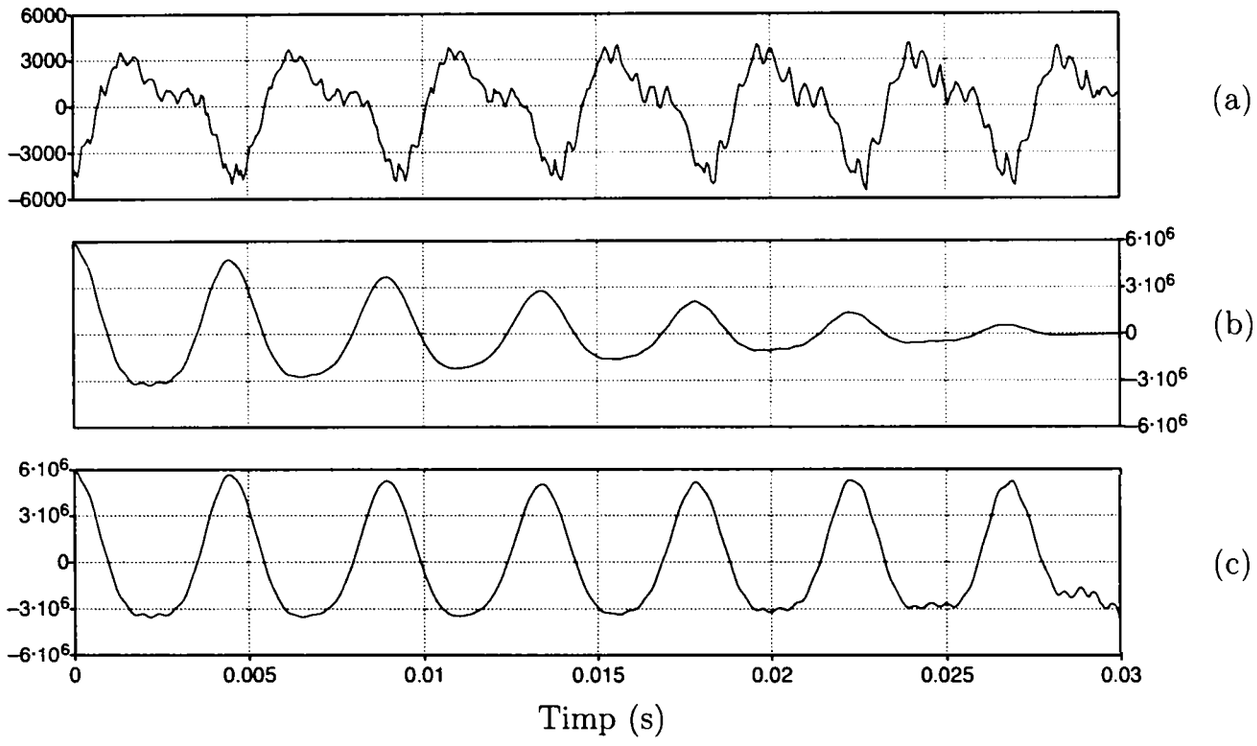

Figura 2.6: Autocorelaţia pentru o porţiune de semnal vocal sonor (a): estimare deplasată (b) şi nedeplastă (c)

scăderea numărului de produse luate efectiv în calcul. În cazul zonelor nesonore ale semnalului, autocorelaţia are maxime mult mai reduse şi aperiodice (figura 2.7).

Remarcând şi că puterea medie este acoperită de autocorelaţie ca un caz particular

$$P = R(0) \tag{2.9}$$

apare în mod natural ideea utilizării ei pentru realizarea clasificării sonor/nesonor, a estimării perioadei fundamentale T_0 şi a altor prelucrări ale semnalului vocal, însă importanţa ei va putea fi pe deplin evidenţiată doar după prezentarea analizei prin predicţie liniară (secţiunea 2.3), utilizată în foarte multe aplicaţii.

2.2.4 Preaccentuarea

O altă prelucrare simplă a semnalului vocal în domeniul timp [200], [232], [59], care poate fi aplicată cu rezultate practic echivalente [153] fie imediat după eşantionarea şi cuantizarea semnalului, fie abia după cadrarea sau ferestruirea lui, constă din trecerea lui printr-un filtru numeric cu ecuaţia

$$y(n) = x(n) - \alpha x(n-1)\,, \quad \alpha \approx 1\,,\, \alpha \leq 1 \tag{2.10}$$

Ea are ca efect accentuarea frecvenţelor înalte cu aproximativ 6 dB/octavă şi din această cauză este cunoscută sub numele de preaccentuare. Conjugată cu efectul similar

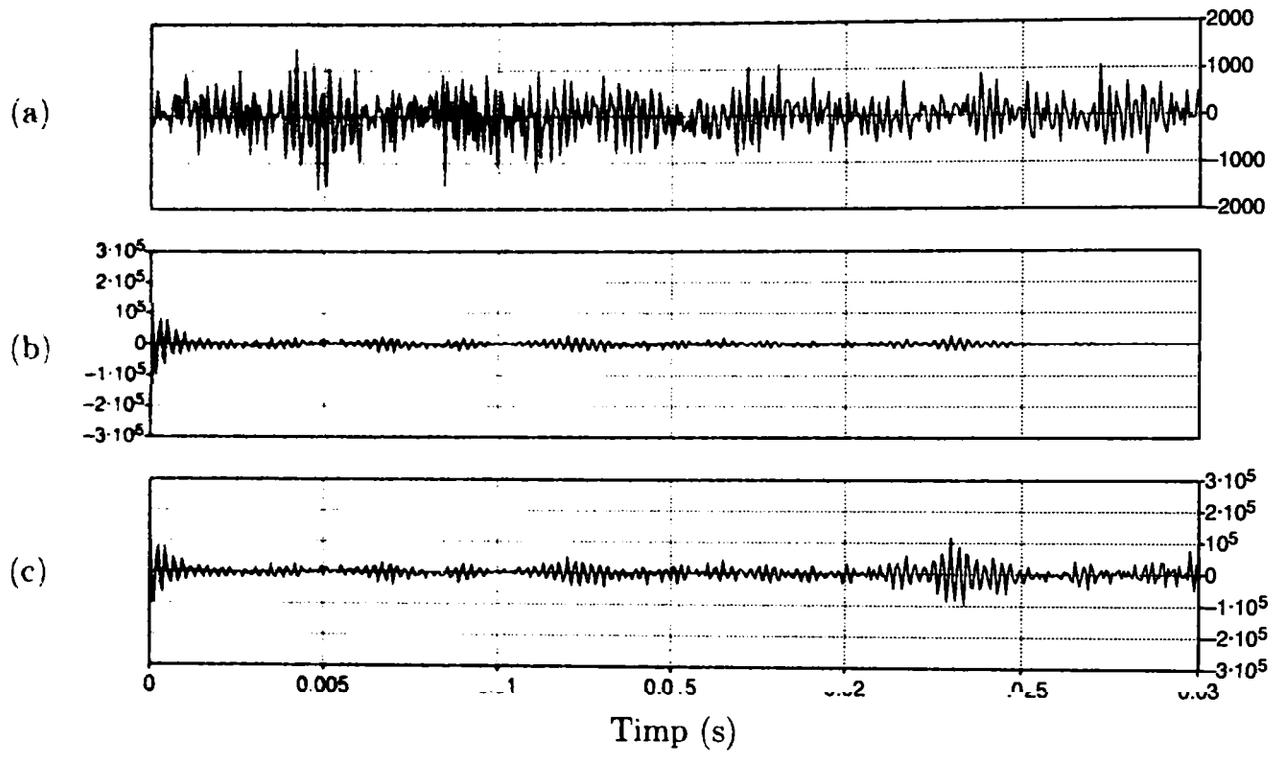

Figura 2.7: Autocorelaţia pentru o porţiune de semnal vocal nesonor (a): estimare deplasată (b) şi nedeplasată (c)

al radiaţiei la ieşirea tractului vocal, preaccentuarea asigură o compensare a căderii de circa 12 dB/octavă a spectrului excitaţiei sonore. În acest fel, pe durata sunetelor sonore anvelopa spectrală a semnalului preaccentuat va aproxima răspunsul în frecvenţă al tractului vocal, care prezintă interes din punct de vedere al recunoaşterii automate a vorbirii datorită corelaţiei dintre conţinutul lingvistic al unui semnal vocal şi evoluţia formei tractului vocal pe parcursul producerii lui.

Efectul preaccentuării cu un **coeficient de preaccentuare** $\alpha = 0{,}96$ este ilustrat în figura 2.8: după cum se observă, sunetele nesonore (cei doi "s") sunt amplificate, iar cele sonore sunt atenuate, astfel încât gama dinamică a semnalului este comprimată, iar în spectrogramă frecvenţele înalte vor fi reprezentate cu niveluri de gri apropiate de cele ale frecvenţelor joase (a se compara cu spectrograma de bandă largă din figura 2.2).

Deoarece excitaţia sonoră şi căderea spectrală asociată ei nu apar pe durata sunetelor nesonore, preaccentuarea ar trebui aplicată doar sunetelor sonore, ceea ce presupune determinarea în prealabil a caracterului sonor sau nesonor al semnalului; o soluţie simplă a acestei probleme, care are şi avantajul de a asigura o compensare maximă a căderii spectrale, este aceea a utilizării unui coeficient de preaccentuare adaptiv [153]

$$\alpha = R(1)/R(0) \tag{2.11}$$

apropiat de 1 în porţiunile sonore ale semnalului, respectiv 0 în cele nesonore. În practică, cel mai frecvent se utilizează valori fixe apropiate de 1, $\alpha \in (0.9 \ldots 1)$, alese uneori de forma $1 - 2^{-n}$ în vederea implementării simple în virgulă fixă.

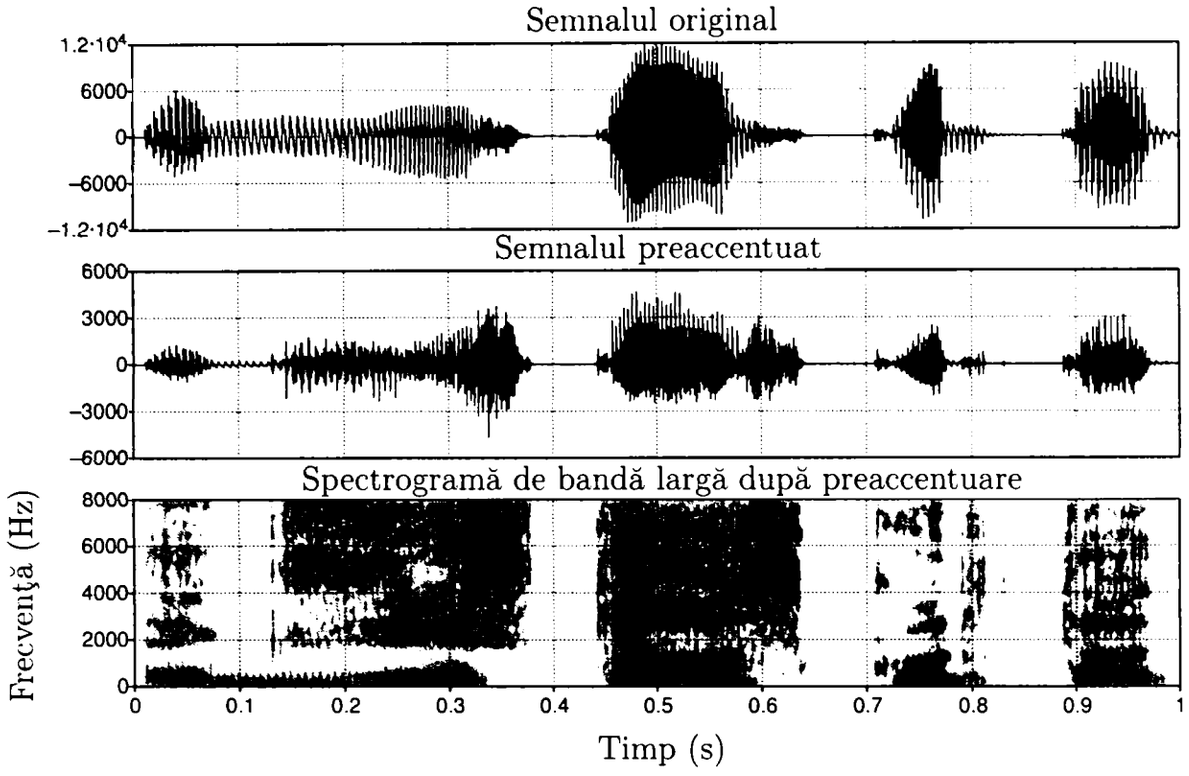

Figura 2.8: Efectul preaccentuării semnalului din figura 2.1 cu $\alpha = 0{,}96$

2.3 Analiza prin predicţie liniară

Prelucrare care oferă o reprezentare a semnalului vocal sub forma unor valori ale parametrilor unui model cu structură fixă, analiza prin predicţie liniară [153], [200], [59] poate fi obţinută prin mai multe abordări ale modelării semnalului vocal [153], fiecare bazată pe diferite ipoteze şi aproximări. Toate abordările conduc la optimizarea parametrilor unui **predictor liniar** care calculează o estimare $\hat{s}(n)$ a valorii unui eşantion $s(n)$ al semnalului ca o combinaţie liniară a unui număr de eşantioane anterioare

$$\hat{s}(n) = \sum_{i=1}^{P} a_i s(n-i) \tag{2.12}$$

unde P este **ordinul de predicţie** şi a_i sunt **coeficienţii de predicţie**. Diferenţa dintre valoarea astfel prezisă şi cea efectivă constituie **eroarea de predicţie**

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{i=1}^{P} a_i s(n-i) \tag{2.13}$$

a cărei transformată z este

$$E(z) = S(z)A(z) \tag{2.14}$$

unde

$$A(z) = 1 - \sum_{i=1}^{P} a_i z^{-i} \tag{2.15}$$

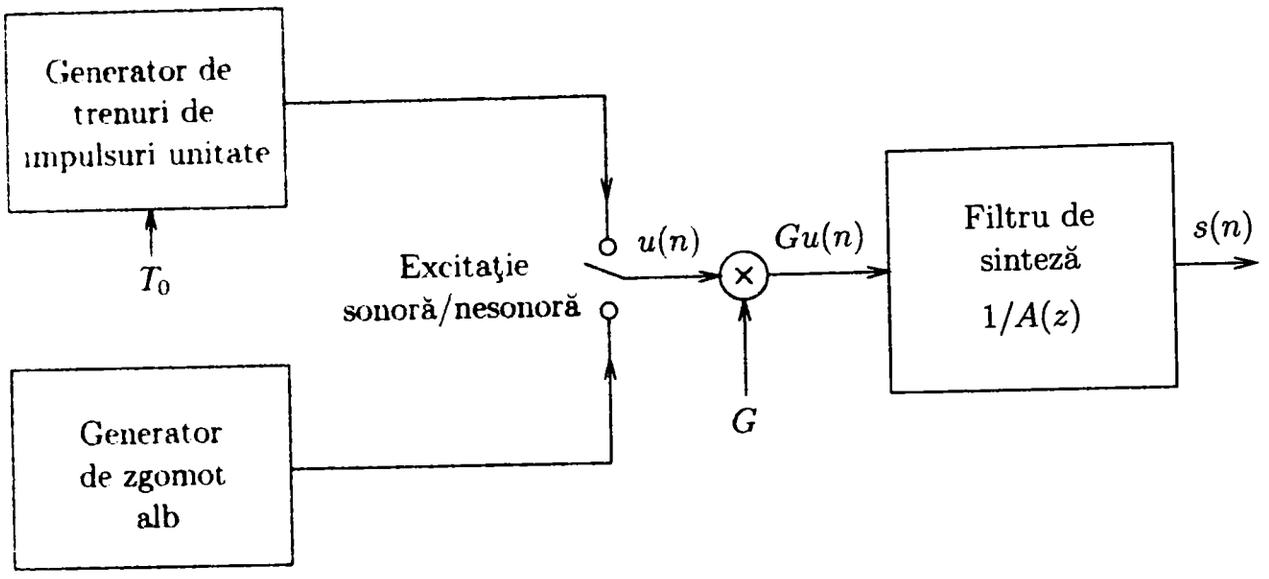

Figura 2.9: Modelul producerii vorbirii bazat pe predicţie liniară

astfel că se poate scrie

$$S(z) = E(z)\frac{1}{A(z)} \tag{2.16}$$

Dat fiind modelul discret al producerii vorbirii [200], [59], [29], predicţia liniară pare "naturală" doar pentru porţiunile sonore ale semnalului vocal, cele nesonore, rezultate din excitaţii cu caracter aleator ale tractului vocal, fiind prin definiţie impredictibile. Pe porţiunile sonore, presupunând semnalul preaccentuat şi cu aproximaţia că zerourile datorate radiaţiei şi preaccentuării compensează polii care modelează impulsurile glotale, din modelul discret al producerii vorbirii va mai rămâne de determinat modelul tractului vocal. Identificând $E(z)$, $S(z)$ şi $A(z)$ cu componente ale modelului discret al producerii vorbirii, ecuaţiile (2.14) şi (2.16) definesc un **model de analiză** respectiv **sinteză** a semnalului vocal preaccentuat: prima permite extragerea excitaţiei modelului tractului vocal din semnal, a doua sinteza semnalului din excitaţie. Din această cauză, $A(z)$ este denumit **filtru de analiză** sau **filtru invers**, iar $1/A(z)$ – **filtru de sinteză**.

Pentru porţiunile nesonore, problema compensării polilor care modelează impulsurile glotale nu se mai pune, în schimb zerourile datorate fenomenelor de absorbţie selectivă (antiformanţi), ca şi cele corespunzătoare radiaţiei şi preaccentuării, pot fi înlocuite printr-un număr de poli. În consecinţă, chiar dacă metoda nu mai pare "naturală", analiza prin predicţie liniară poate fi folosită şi în cazul sunetelor nesonore.

Se obţine astfel un model bazat pe predicţie liniară al producerii vorbirii (figura 2.9), variantă a modelului discret al producerii vorbirii, simplificat prin înglobarea în filtrul de sinteză a efectelor impulsului glotal, tractului vocal şi radiaţiei.

Numărul de coeficienţi de predicţie folosiţi depinde de lărgimea de bandă a semnalului vocal analizat şi de precizia dorită a modelării: fiecare formant impune utilizarea unei

perechi de poli complex conjugaţi pentru modelarea sa, iar alţi câţiva poli (2... 4) sunt necesari pentru suplinirea zerourilor în cazul sunetelor nesonore. Pe baza teoriei acustice a producerii vorbirii [200], pentru o lungime uzuală a tractului vocal de cca. 17 cm ne putem aştepta la un număr de formanţi egal cu lărgimea de bandă în KHz a semnalului, ceea ce ar impune utilizarea unui **număr de poli**

$$P = F_s + 2 \ldots 4 \tag{2.17}$$

unde F_s este frecvenţa de eşantionare a semnalului în KHz. Practic, nu există mai mult de cinci formanţi semnificativi, astfel că de cele mai multe ori este suficient $P = 12 \ldots 14$.

Coeficienţii de predicţie a_i corespunzători unui cadru pot fi calculaţi prin proceduri de optimizare care minimizează **eroarea pătratică de predicţie**

$$\mathcal{E} = \sum_n e^2(n) \tag{2.18}$$

Ţinând cont de ecuaţia (2.13), eroarea pătratică de predicţie devine

$$\mathcal{E} = \sum_n [s(n) - \sum_{i=1}^{P} a_i\, s(n-i)]^2 \tag{2.19}$$

iar valorile optime ale a_i se obţin pentru

$$\frac{\partial \mathcal{E}}{\partial a_i} = 0\,, \qquad i = 1 \ldots P \tag{2.20}$$

sau, înlocuind (2.19) în (2.20), din sistemul de ecuaţii liniare

$$\sum_{k=1}^{P} \phi(i,k)\, a_k = \phi(i,0)\,, \qquad i = 1 \ldots P \tag{2.21}$$

unde

$$\phi(i,k) = \sum_n s(n-i)\, s(n-k)\,, \quad i = 1 \ldots P\,, \quad k = 0 \ldots P \tag{2.22}$$

care poate fi scris în formă matricială

$$\begin{bmatrix} \phi(1,1) & \phi(1,2) & \cdots & \phi(1,P) \\ \phi(2,1) & \phi(2,2) & \cdots & \phi(2,P) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(P,1) & \phi(P,2) & \cdots & \phi(P,P) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} \phi(1,0) \\ \phi(2,0) \\ \vdots \\ \phi(P,0) \end{bmatrix} \tag{2.23}$$

Având în vedere că în (2.22) apar indici de forma $n-k$, $n-i$, limitele de sumare au fost în mod deliberat omise în (2.18), (2.19) şi (2.22), iar modul de alegere a lor duce la două metode fundamentale de analiză prin predicţie liniară:

- **metoda covarianţei**, în care se sumează doar **erorile din cadrul analizat**, astfel încât $\phi(i,k)$ devine:

$$\phi(i,k) = \sum_{n=0}^{N-1} s(n-i)\, s(n-k)\,, \quad i = 1 \ldots P\,, \quad k = 0 \ldots P \tag{2.24}$$

  iar ferestruirea nu este permisă pentru a nu afecta eşantioanele utilizate în calcule dar necuprinse în cadru; numele este datorat similitudinii formale a matricei $[\phi(i,k)]$ cu matricea de covarianţă a unei variabile aleatoare P-dimensionale;

- **metoda autocorelaţiei**, în care se sumează **toate erorile nenule**, ceea ce impune înlocuirea eşantioanelor din afara cadrului cu valori nule pentru ca erorile să fie cauzate doar de cele din cadru; pentru a reduce erorile la capetele cadrului, această înlocuire se face prin ferestruire (secţiunea 2.2.1) cu o fereastră care atenuează progresiv eşantioanele marginale (figura 2.4), astfel că limitele efective de sumare sunt cele ale valorilor nenule ale semnalului ferestruit $s_w(n)$, iar formal

$$\phi(i,k) = \sum_{n=0}^{N+P-1} s_w(n-i)\, s_w(n-k)\,, \quad i = 1 \ldots P\,, \quad k = 0 \ldots P \tag{2.25}$$

Diferenţa esenţială între aceste metode constă în absenţa respectiv prezenţa ferestruirii şi face ca fiecare să aibă proprietăţi specifice ce o recomandă în anumite aplicaţii.

Aceste două metode, împreună cu altele şi aplicaţii ale analizei prin predicţie liniară, sunt tratate pe larg în literatură [153], [200], [59], dar datorită eficienţei computaţionale şi stabilităţii soluţiilor furnizate, în recunoaşterea automată a vorbirii se utilizează aproape în exclusivitate metoda autocorelaţiei.

2.3.1 Metoda autocorelaţiei

Dacă în ecuaţia (2.25) sunt luate în calcul numai produsele care pot fi nenule, expresia $\phi(i,k)$ devine

$$\varphi(i,k) = \sum_{n=0}^{N-|i-k|-1} s_w(n)\, s_w(n+|i-k|) = N\, R(|i-k|) \tag{2.26}$$

unde R este estimatorul deplasat al autocorelaţiei $s_w(n)$ (secţiunea 2.2.3, ecuaţia 2.7), de unde numele metodei. Lungimea cadrului N fiind factor comun în $\phi(i,k)$, sistemul (2.21) poate fi rescris

$$\sum_{k=1}^{P} R(|i-k|)\, a_k = R(i)\,, \qquad i = 1 \ldots P \tag{2.27}$$

sau în formă matricială

$$\begin{bmatrix} R(0) & R(1) & \cdots & R(P-1) \\ R(1) & R(0) & \cdots & R(P-2) \\ \vdots & \vdots & & \vdots \\ R(P-1) & R(P-2) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(P) \end{bmatrix} \tag{2.28}$$

Matricea $P \times P$ a valorilor autocorelaţiei R din sistemul (2.28) este simetrică şi cu elemente identice pe diagonale (Toeplitz), proprietate folosită pentru a obţine proceduri recursive de rezolvare a sistemului. Cea mai eficientă este **metoda lui Durbin**: aceasta este o particularizare a unui algoritm propus de Levinson (motiv pentru care mai este cunoscută şi ca **metoda Levinson-Durbin**) pentru rezolvarea ecuaţiei $Ax = b$ cu A matrice Toeplitz pozitiv definită şi b un vector oarecare, care ţine cont de relaţiile existente între elementele A şi b în cazul metodei autocorelaţiei.

Algoritmul recursiv al lui Durbin pleacă de la lipsa oricărei predicţii ($P = 0$), pentru care **eroarea medie pătratică de predicţie** ε este

$$\varepsilon^{(0)} = R(0) \tag{2.29}$$

şi calculează seturi optime de coeficienţi pentru predictori de ordin din ce în ce mare:

$$k_i = \frac{R(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j)}{\varepsilon^{(i-1)}}, \qquad i = 1 \ldots P \tag{2.30}$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i\, a_{i-j}^{(i-1)}, \quad j = 1 \ldots i-1 \tag{2.31}$$

$$a_i^{(i)} = k_i \tag{2.32}$$

$$\varepsilon^{(i)} = (1 - k_i^2)\, \varepsilon^{(i-1)} \tag{2.33}$$

$$G^{(i)} = \sqrt{\varepsilon^{(i)}} \tag{2.34}$$

unde mărimile indexate superior cu (i) corespund predictorului de ordin i, iar sumarea din ecuaţia (2.30) este omisă pentru $i = 1$.

Pentru o modelare a semnalului folosind un predictor de ordin P, vor fi reţinuţi parametrii rezultaţi din ultima iteraţie a acestui algoritm. O reprezentare mai compactă a coeficienţilor de predicţie este posibilă utilizând coeficienţii $k_i \in [-1, 1]$, din care ei pot fi calculaţi, cunoscuţi drept **coeficienţi de reflexie** datorită unor legături care pot fi făcute cu modelarea tractului vocal prin tuburi acustice de secţiune constantă.

O reprezentare completă a semnalului vocal folosind analiza prin predicţie liniară trebuie să includă pentru fiecare cadru, pe lângă coeficienţii de predicţie a_i şi câştigul G, clasificarea sonor/nesonor şi valoarea perioadei fundamentale T_0. Asemenea reprezentări stau la baza unor algoritmi de codare a semnalului vocal [227], dar în recunoaşterea automată a vorbirii singurii parametri specifici analizei prin predicţie liniară cu o utilitate validată experimental sunt coeficienţii de predicţie.

2.4 Analiza Fourier

Ca şi predicţia liniară, analiza Fourier [233], [211] este o metodă de aproximare a semnalelor, de data aceasta prin componente cu diferite frecvenţe, amplitudini şi faze. Semnalele continue periodice sunt reprezentate prin serii Fourier, teoretic infinite, de armonice ale unei frecvenţe fundamentale unice, iar cele aperiodice – prin integrale şi transformate Fourier, extinse la toate frecvenţele.

Pentru un semnal numeric $s(n)$, amplitudinile şi fazele componentelor sale pot fi calculate utilizând **transformata Fourier** (TF)

$$S(\omega) = F[s(n)] = \sum_{n=-\infty}^{\infty} s(n)\, e^{-j\omega n} \tag{2.35}$$

din care semnalul poate fi refăcut prin **transformata Fourier inversă** (TFI)

$$s(n) = F^{-1}[S(\omega)] = \frac{1}{2\pi} \int_{-\pi}^{\pi} S(\omega)\, e^{j\omega n}\, d\omega \tag{2.36}$$

unde limitele de sumare şi integrare reflectă utilizarea timpului discret şi a frecvenţei normalizate (între $-1/2$ şi $1/2$) în reprezentarea semnalelor numerice.

Valorile transformatei Fourier $S(\omega)$ sunt în cazul cel mai general complexe, definind o **anvelopă spectrală complexă**, iar dacă se neglijează informaţia de fază, inutilă pentru recunoaşterea vorbirii, se obţine **spectrul de amplitudine** $|S(\omega)|$ al semnalului.

Ca şi în cazul energiei şi puterii, gama dinamică a spectrului de amplitudine este prea mare pentru o reprezentare grafică pe o scară liniară, astfel încât şi pentru reprezentarea lui se utilizează o scară logaritmică. Similar, reprezentarea pe o scară logaritmică a spectrului poate fi privită şi ca o reprezentare pe o scară liniară a logaritmului spectrului, sau **log-spectrul**, semnalului analizat, care include dependenţa logaritmică, conform legii Weber-Fechner [228], dintre intensitatea obiectivă şi tăria subiectivă a unui sunet.

Sub forma dată, transformata Fourier este o funcţie de argument continuu, calculată pe baza unui număr infinit de valori, motive pentru care ea este imposibil de utilizat în practică. Pentru o secvenţă $s(n)$, $0 \leq n \leq N-1$, cum este cazul oricărui semnal numeric real şi cu atât mai mult al unui cadru de semnal vocal, spectrul poate fi eşantionat în N puncte echidistante folosind **transformata Fourier discretă** (TFD)

$$S(k) = \begin{cases} \sum_{n=0}^{N-1} s(n)\, e^{-j\,(2\pi/N)\,k\,n}, & 0 \leq k \leq N-1 \\ 0, & \text{altfel} \end{cases} \tag{2.37}$$

Secvenţa $s(n)$ poate fi refăcută prin **transformata Fourier discretă inversă** (TFDI)

$$s(n) = \begin{cases} \frac{1}{N} \sum_{k=0}^{N-1} S(k)\, e^{j\,(2\pi/N)\,k\,n}, & 0 \leq n \leq N-1 \\ 0, & \text{altfel} \end{cases} \tag{2.38}$$

Transformata Fourier rapidă (TFR) este denumirea generică dată unei clase de algoritmi [175], [202], [147] pentru calculul eficient al TFD şi TFDI în anumite condiţii referitoare la lungimea N a secvenţei analizate (de obicei, $N = 2^m$), iar existenţa a numeroase implementări (ex. [191]) ale unor asemenea algoritmi face ca în prezent ei să fie folosiţi foarte frecvent. În cazul unei lungimi N care nu satiface condiţiile, ele pot fi îndeplinite prin adăugarea unui număr convenabil de eşantioane nule după ferestruire.

2.5 Legături timp–frecvenţă

Anunţate încă de la începutul capitolului, legăturile dintre domeniile timp şi frecvenţă prezentate în continuare au fost alese pentru relevanţa lor în analiza vorbirii în general şi în contextul prezentei lucrări în special, fără a epuiza conexiunile existente între cele două domenii. Ele se referă la corelaţiile în domeniul frecvenţă ale operaţiunii de ferestruire din domeniul timp, precum şi la posibilitatea de estimare a spectrului unui semnal pe baza analizei lui prin predicţie liniară.

Efecte spectrale ale ferestruirii

După cum se cunoaşte din proprietăţile transformatelor Fourier [147], [56], înmulţirea a două semnale în domeniul timp are ca echivalent în domeniul frecvenţă convoluţia

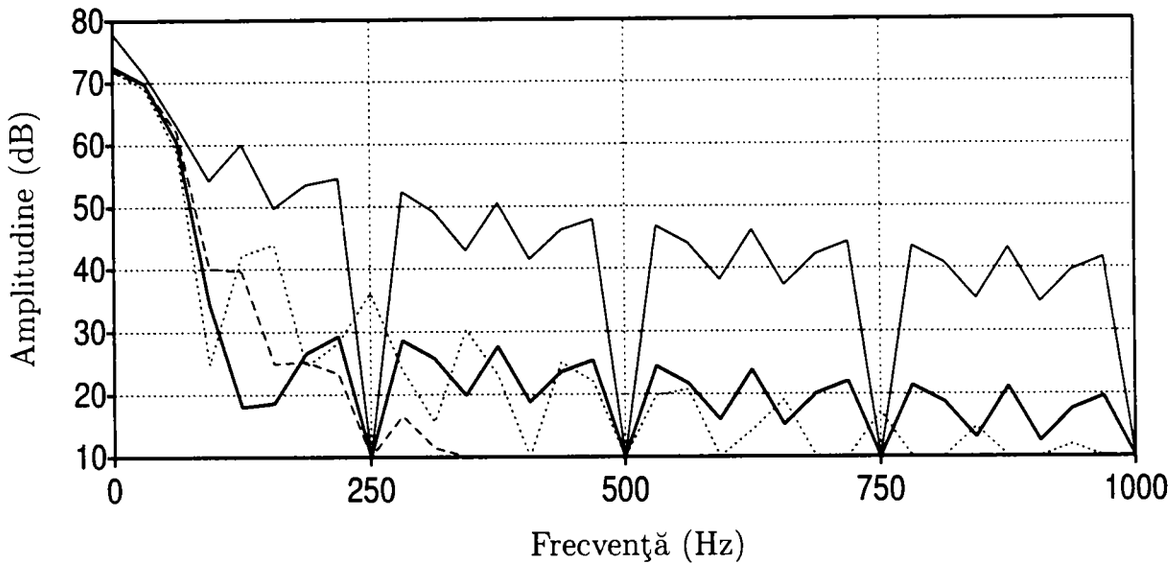

Figura 2.10: Spectre ale unor ferestre cu lungimea de 20 ms la frecvenţa de eşantionare de 16 KHz: Hamming (linie groasă), Hanning (întreruptă), triunghiulară (punctată), rectangulară (continuă)

spectrelor lor, astfel încât, cadrarea şi ferestruirea semnalului vocal fiind înmulţiri ale lui cu ferestre rectangulare sau de alt tip, ele vor determina modificări spectrale:

$$s_w(n) = s(n)\, w(n) \Leftrightarrow S_w(\Omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S(\Omega - \omega)\, W(\omega)\, d\omega \tag{2.39}$$

Porţiunile inferioare ale spectrelor unor ferestre dintre cele mai utilizate, exemplificate pentru o lungime de 20 ms şi o frecvenţă de eşantionare $F_s = 16$ KHz, sunt prezentate în figura 2.10 (părţile lipsă continuă conform tendinţelor sugerate): după cum se observă, toate spectrele au o componentă dominantă de joasă frecvenţă, denumită în literatura de specialitate **lob principal**, însoţită de **lobi secundari** de amplitudini din ce în ce mai mici spre frecvenţa maximă posibilă $F_s/2$.

Relaţia (2.39) implică faptul că **lărgimea de bandă** a lobului principal şi **atenuarea** celor secundari în raport cu acesta sunt cele două elemente care determină calitatea unei ferestre din punct de vedere al analizei spectrale: o bandă îngustă a lobului principal asigură o rezoluţie bună în jurul unei anumite frecvenţe Ω, iar atenuarea lobilor secundari previne influenţa componentelor de la frecvenţe mai îndepărtate.

După cum se observă din figura 2.10, fereastra rectangulară, corespunzătoare simplei cadrări a semnalului, are cele mai slabe performanţe datorită lobului principal foarte larg şi atenuării slabe a lobilor secundari; cea triunghiulară are lobul principal cel mai îngust, dar atenuarea lobilor secundari nu este foarte bună; fereastra Hanning asigură o atenuare foarte puternică a lobilor secundari dar are un lob principal destul de larg. În aceste condiţii, fereastra Hamming asigură cel mai bun compromis, motiv pentru care ea este utilizată cel mai frecvent în prelucrarea automată a semnalului vocal.

Lungimea în timp şi rezoluţia în frecvenţă (lărgimea de bandă a lobului principal)

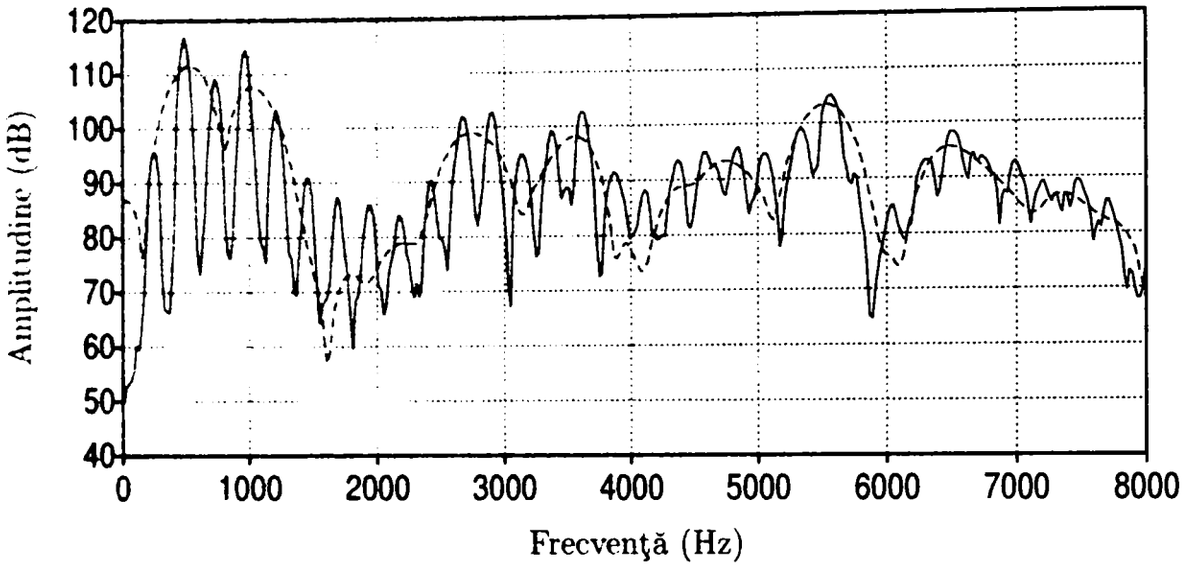

Figura 2.11: Estimări prin analiză Fourier ale spectrului semnalului din figura 2.1 la t = 0,5 s, realizate utilizând ferestre Hamming de 25 ms (linie continuă) şi 5 ms (linie întreruptă)

ale unei ferestre sunt în relaţie inversă, menţionată deja în secţiunea 2.1 şi ilustrată în figurile 2.2 şi 2.11. În figura 2.11, fereastra de 25 ms, corespunzătoare unei analize de bandă îngustă (a lobului principal), cuprinde câteva perioade fundamentale, iar această periodicitate este pusă în evidenţă în spectru prin maxime repetate, corespunzătoare armonicilor frecvenţei fundamentale. Examinând structură periodică a spectrului, din faptul că armonica a patra este de aproape 1000 Hz putem chiar estima valorile frecvenţei fundamentale $F_0 \approx 240$ Hz, respectiv perioadei fundamentale $T_0 \approx 4{,}2$ ms.

Fereastra de 5 ms include o singură perioadă fundamentală din semnalul analizat, iar periodicitatea acestuia nu se poate manifesta în spectrul estimat, care datorită lărgimii mai mari a lobului principal din spectrul ferestrei este o variantă netezită a celui anterior.

Interpretarea spectrală a predicţiei liniare

Analiza prin predicţie liniară [153] modelează spectrul sursei de excitaţie a tractului vocal, răspunsul lui în frecvenţă, şi efectele radiaţiei sonore, înglobate în semnalul vocal, printr-un singur filtru având numai poli, $G/A(z)$, de fază minimă, şi al cărui spectru de amplitudine îl aproximează pe cel al semnalului. În aceste condiţii, estimarea spectrului de amplitudine al unui cadru de semnal se poate face pe baza analizei prin predicţie liniară prin evaluarea modelului numai poli rezultat pe cercul unitate din planul z

$$S(\omega) = \frac{G}{A(z)|_{z=e^{j\omega}}} = \frac{G}{A(e^{j\omega})} \tag{2.40}$$

Estimarea se poate face şi numai pe baza filtrului de sinteză $1/A(z)$, rezultatul fiind

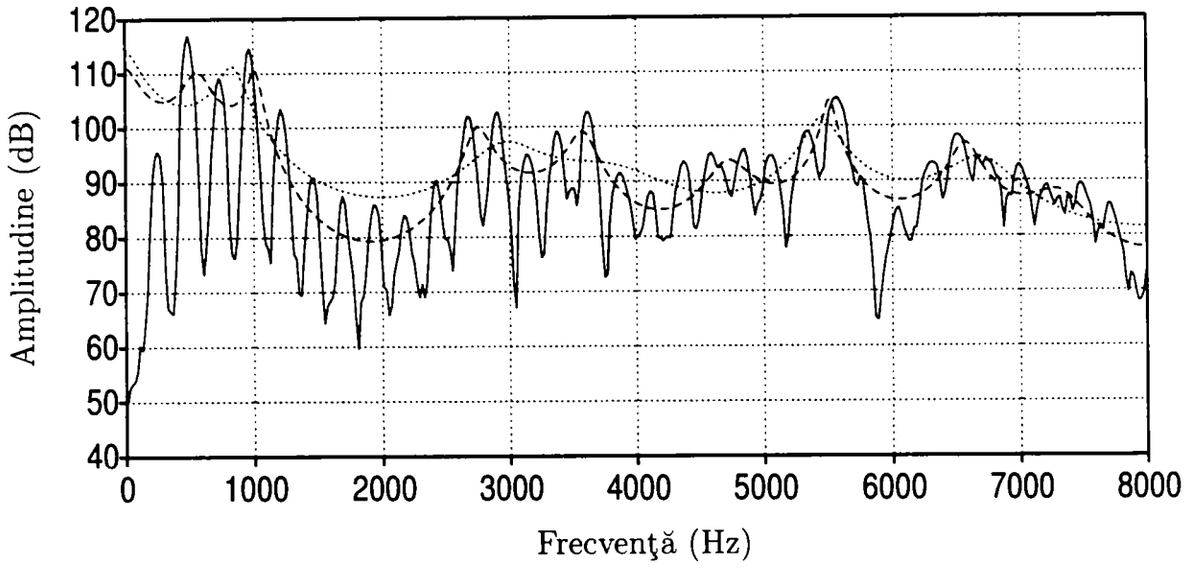

Figura 2.12: Estimări ale spectrului semnalului din figura 2.1 la t = 0,5 s, realizate cu o fereastră Hamming de 25 ms, prin analiză Fourier (linie continuă) şi predicţie liniară cu 20 (linie întreruptă) respectiv 12 poli (linie punctată)

un **spectru normalizat**

$$S_N(\omega) = \frac{1}{A(z)|_{z=e^{j\omega}}} = \frac{1}{A(e^{j\omega})} \tag{2.41}$$

care pe o scară logaritmică are valoarea medie nulă

$$\int_{-\pi}^{\pi} \log |A(e^{j\omega})|\, d\omega = 0 \tag{2.42}$$

ceea ce face posibile comparaţii spectrale între semnale cu amplitudini diferite.

Ca şi în cazul analizei Fourier de bandă largă, spectrul obţinut este o aproximare, de astă dată în principal a formanţilor, iar numărul polilor folosiţi determină precizia acesteia. Pentru exemplificare, figura 2.12 prezintă trei estimări ale spectrului unei porţiuni de semnal vocal sonor prin analiză Fourier de bandă îngustă şi predicţie liniară cu 20 respectiv 12 poli. A doua variantă corespunde utilizării formulei (2.17) şi asigură o foarte bună (poate chiar prea bună) aproximare a formanţilor, în timp ce a treia "contopeşte" perechi de poli din cea anterioară.

2.6 Analiza homomorfică

Analiza Fourier de bandă îngustă a porţiunilor sonore ale semnalelor vocale (figurile 2.2, 2.11 şi 2.12) pune în evidenţă prezenţa unor componente periodice ale spectrelor lor, corespunzătoare armonicilor frecvenţei fundamentale, datorate excitaţiei sonore care stă la baza acestor porţiuni, şi cunoscute sub numele de **structură spectrală fină**.

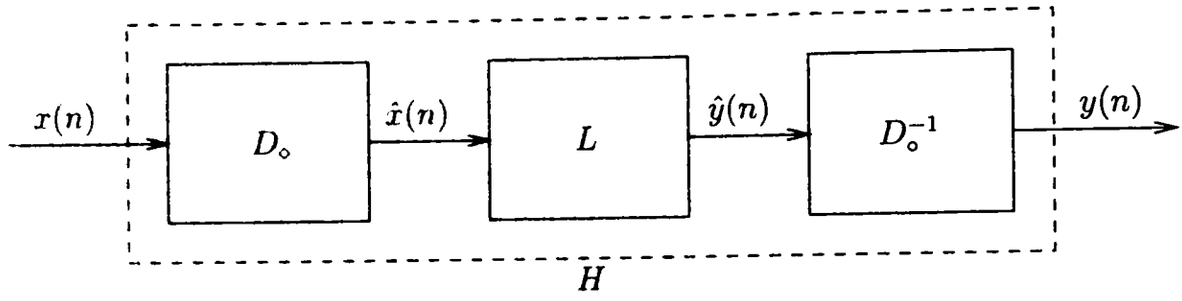

Figura 2.13: Reprezentarea canonică a unui sistem homomorfic

Această structură fină este eliminată prin netezirea asigurată de analiza Fourier de bandă largă (secţiunea 2.5, figura 2.11), care reţine în principal informaţii despre aspectul general al spectrului, ceea ce pentru recunoaşterea automată a vorbirii este avantajos deoarece variaţiile frecvenţei fundamentale pe durata pronunţiilor unui cuvânt (excepţie fac limbile tonale, de tipul chinezei) sunt irelevante din acest punct de vedere. Analiza Fourier de bandă largă are însă dezavantajul utilizării unor cadre cu lungime scurtă (de ordinul milisecundelor) şi frecvenţă ridicată, astfel că prezintă interes găsirea unei metode apte să asigure această **netezire spectrală** din cadre mai lungi (de ordinul zecilor de milisecunde) şi cu frecvenţă mai scăzută (tipic 100 pe secundă).

Un cadru teoretic cuprinzător pentru separarea şi eliminarea structurii spectrale fine este oferit de **sistemele homomorfice** [175], care extind principiul superpoziţiei din sistemele liniare

$$L[x_1(n) + x_2(n)] = L[x_1(n)] + L[x_2(n)] \tag{2.43}$$

$$L[c\,x(n)] = c\,L[x(n)] \tag{2.44}$$

unde L este o transformare liniară.

În cazul unui sistem homomorfic, există operatori pentru combinarea intrărilor, $\diamond$, combinarea intrărilor cu scalari, $\bullet$, combinarea ieşirilor, $\oplus$, şi combinarea ieşirilor cu scalari, $\odot$, iar **principiul generalizat al superpoziţiei** este

$$H[x_1(n) \diamond x_2(n)] = H[x_1(n)] \oplus H[x_2(n)] \tag{2.45}$$

$$H[c \bullet x(n)] = c \odot H[x(n)] \tag{2.46}$$

unde H este o transformare homomorfică.

Rezultatele referitoare la sisteme liniare pot fi utilizate şi pentru cele homomorfice recurgând la reprezentări canonice [175], orice sistem homomorfic fiind decompozabil în trei subsisteme, fiecare homomorfic la rândul lui (figura 2.13).

Primul subsistem este denumit **sistem caracteristic de intrare** (determinat de operaţiunile $\diamond$ şi $\bullet$) şi stabileşte un homomorfism între spaţiul vectorial de intrare şi unul intermediar, în care $\diamond$ şi $\bullet$ sunt homomorfe cu adunarea şi înmulţirea cu scalari

$$D_\diamond[x_1(n) \diamond x_2(n)] = \hat{x}_1(n) + \hat{x}_2(n) \tag{2.47}$$

$$D_\diamond[c \bullet x(n)] = c\,\hat{x}(n) \tag{2.48}$$

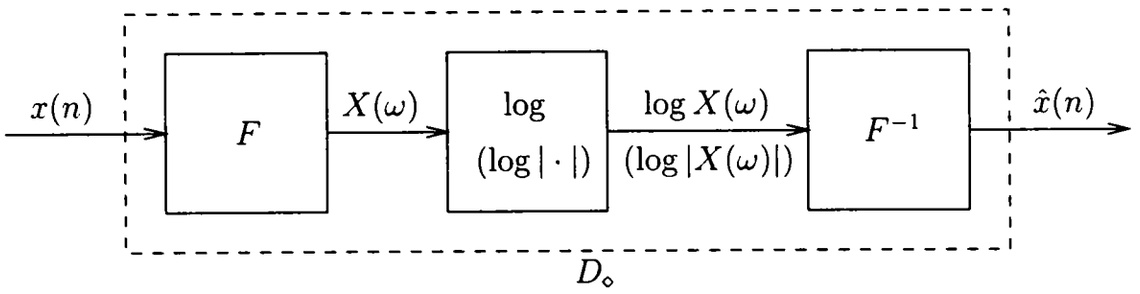

Figura 2.14: Schema bloc a unui sistem caracteristic de intrare pentru prelucrări homomorfice (cepstrale) ale semnalului vocal

Al doilea este un sistem liniar, iar ultimul este **sistemul caracteristic de ieşire** şi asigură revenirea la spaţiul vectorial al ieşirilor

$$D_{\circ}^{-1}[\hat{y}_1(n) + \hat{y}_2(n)] = y_1(n) \oplus y_2(n) \tag{2.49}$$

$$D_{\circ}^{-1}[c\,\hat{y}(n)] = c \odot y(n) \tag{2.50}$$

Semnalul vocal poate fi modelat ca rezultat al convoluţiei în timp, respectiv înmulţirii în frecvenţă, dintre excitaţia tractului vocal $e(n)$, manifestată sub forma structurii spectrale fine $E(\omega)$, cu răspunsul la impuls al tractului, $v(n)$, având corespondent spectrul netezit $V(\omega)$. În aceste condiţii, pentru prelucrarea homomorfică a semnalului vocal, sistemul caracteristic de intrare trebuie să asigure transformarea convoluţiei $e(n) * v(n)$ în suma transformatelor, $\hat{e}(n) + \hat{v}(n)$. Ţinând cont de corespondenţele

$$F\,[e(n) * v(n)] = F\,[e(n)]\,F\,[v(n)] = E(\omega)\,V(\omega) \tag{2.51}$$

$$\log\,[E(\omega)\,V(\omega)] = \log E(\omega) + \log V(\omega) \tag{2.52}$$

unde $*$ este operatorul de convoluţie iar F transformata Fourier directă, precum şi de utilitatea interpretării rezultatului pe care o permite, o transformare foarte utilizată pentru **sistemul caracteristic de intrare**, corespunzătoare schemei bloc din figura 2.14, este

$$\hat{x}(n) = D_{\circ}\,[x(n)] = F^{-1}\{\log F\,[x(n)]\} \tag{2.53}$$

Transformata Fourier inversă F^{-1} realizează o trecere din domeniul frecvenţă într-un nou domeniu de tip timp, diferit însă de cel al semnalului original, iar pentru descrierea lui au fost introduşi termeni obţinuţi prin anagramarea unora din domeniul frecvenţă. Astfel, domeniul ca atare este numit **domeniul cvefrenţă** (anagramă din frecvenţă, cf. engl. quefrency, anagramă din frequency); $\hat{x}(n)$ este **cepstrul** (anagramă din spectrul, pronunţată kepstrul) semnalului $x(n)$, având aceeaşi frecvenţă de eşantionare ca şi acesta; iar prelucrarea semnalului ca atare a fost denumită **analiză cepstrală**.

Importanţa analizei homomorfice sau cepstrale a semnalului vocal este dată, după cum vom vedea în continuare, de posibilităţile oferite pentru separarea componentelor spectrale corespunzătoare excitaţiei, $E(\omega)$, respectiv tractului vocal, $V(\omega)$.

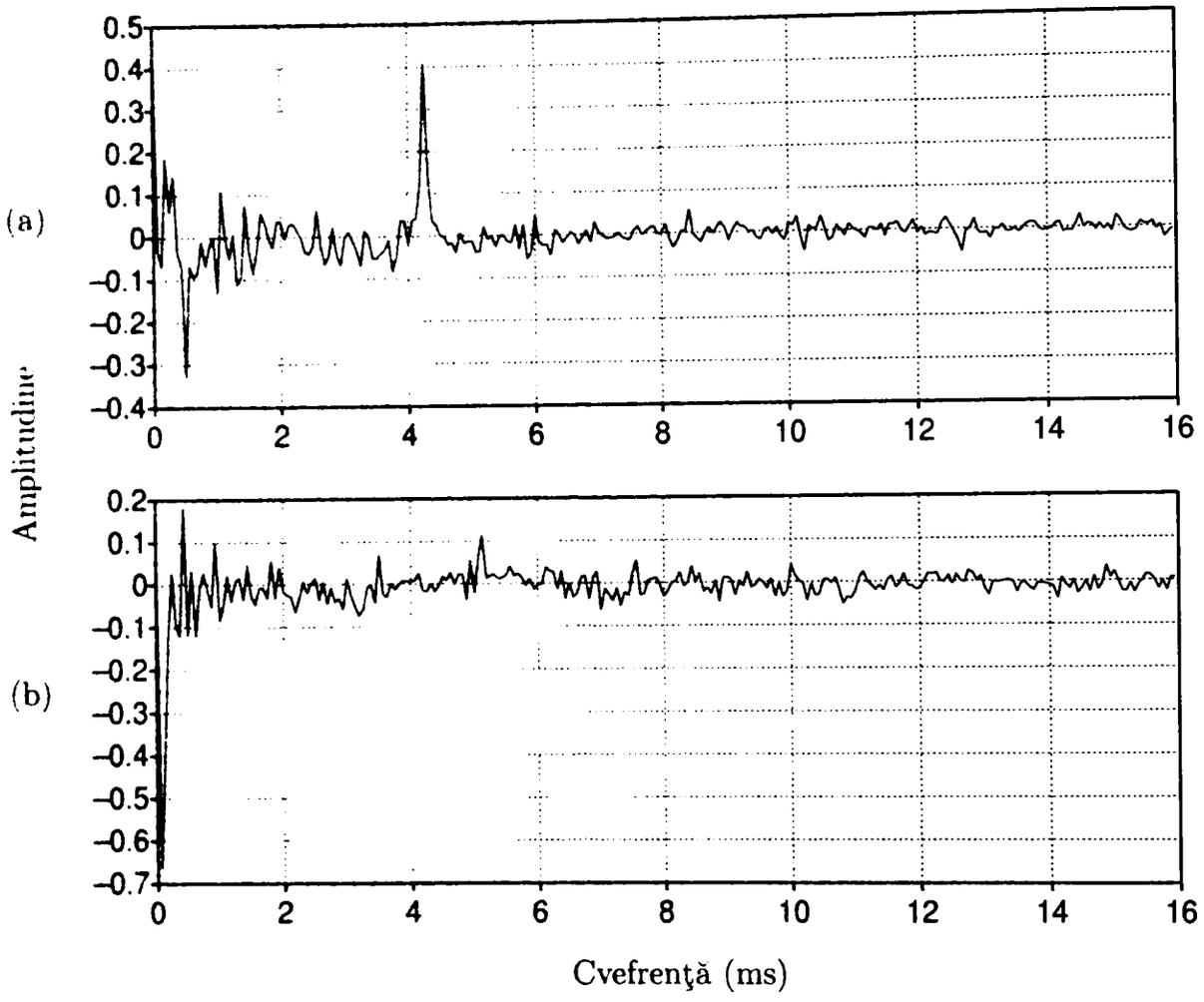

Figura 2.15: Estimări cu o fereastră Hamming de 25 ms ale cepstrului semnalului din figura 2.1 la (a) t = 0,5 ms (sonor) (b) t = 0,6 ms (nesonor)

2.6.1 Cepstrul real

Transformata Fourier a unui semnal real $F[x(n)]$ are în cazul cel mai general valori complexe, astfel încât $\hat{x}(n)$ din ecuaţia (2.53) este denumit **cepstru complex**. Utilizarea logaritmului complex ridică însă o serie de probleme de ordin teoretic [175], iar informaţia de fază nu are importanţă pentru recunoaşterea vorbirii, astfel încât cel mai frecvent este utilizat **cepstrul real**, singurul considerat în continuare, obţinut prin neglijarea fazei. Pentru un cadru de lungime N dintr-un semnal vocal $s(n)$, cepstrul real este o secvenţă de numere reale, cunoscute sub numele de **coeficienţi cepstrali**, de aceeaşi lungime

$$c_l = c(l) = \frac{1}{N}\sum_{k=0}^{N-1} \log|S(k)|\, e^{j(2\pi/N)kl}, \quad l = 0 \ldots N-1 \tag{2.54}$$

Spectrul de amplitudine $|S(k)|$ fiind o secvenţă reală cu paritate pară, cepstrul real

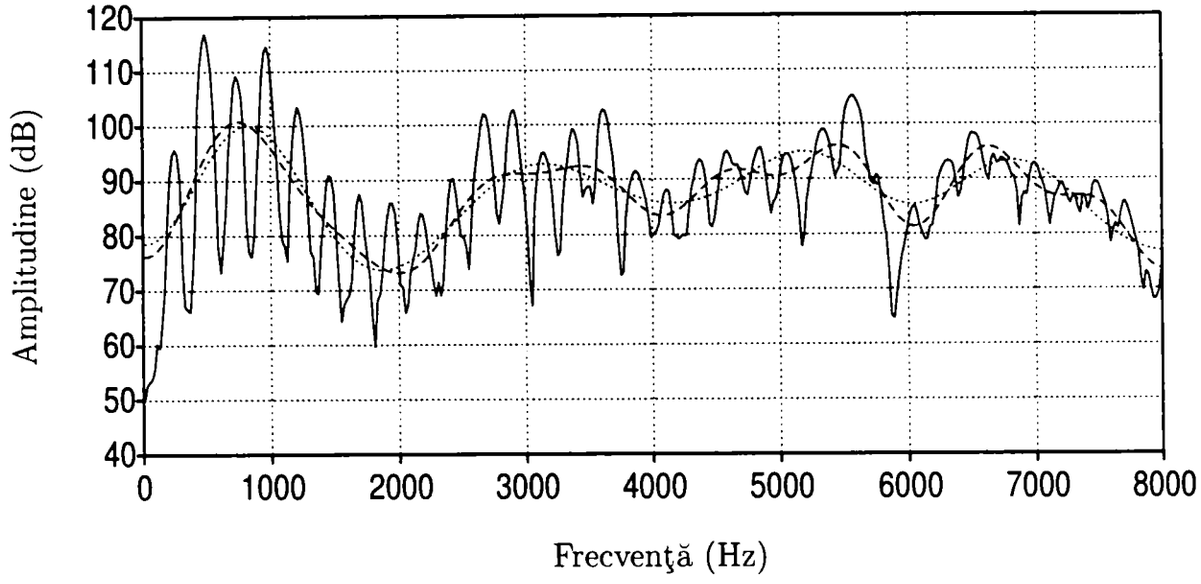

Figura 2.16: Estimări ale spectrului semnalului din figura 2.1 la t = 0,5 s, realizate cu o fereastră Hamming de 25 ms prin analiză Fourier de bandă îngustă (cu linie continuă) şi netezite prin trunchiere cepstrală la 1 ms (cu linie intreruptă) respectiv 0,75 ms (cu linie punctată)

prezintă de asemeni simetrie pară, iar figura 2.15(a) prezintă aspectul porţiunii sale inferioare pentru semnalul pe baza căruia s-au realizat estimările spectrale din figurile 2.11 şi 2.12. Se observă prezenţa unui maxim, corespunzător excitaţiei sonore, la cvefrenţa de 4,25 ms, care concordă cu estimarea $T_0 \approx 4{,}2$ ms din secţiunea 2.5. Această evidenţiere în cepstru a excitaţiei sonore este motivul pentru care una din cele mai frecvente aplicaţii ale sale este estimarea valorilor perioadei şi frecvenţei fundamentale.

Pentru porţiunile nesonore ale semnalelor vocale (figura 2.15(b)), acest maxim nu este prezent, sau dacă apare are valori mult mai mici, ceea ce oferă posibilitatea utilizării cepstrului şi în algoritmi de decizie asupra caracterului sonor sau nesonor al semnalului.

Ţinând cont de ecuaţiile (2.53) şi (2.54), de precizările făcute pe marginea lor şi de exemplele prezentate, valorile cepstrului de la cvefrenţe joase pot fi puse în corespondenţă cu componenta lent variabilă a spectrului, determinată de tractul vocal, $V(\omega)$, iar cele de la cvefrenţe superioare – cu structura spectrală fină $E(\omega)$, datorată excitaţiei.

Această interpretare a cepstrului stă la baza eliminării structurii spectrale fine şi netezirii spectrale prin **trunchiere cepstrală**, care este un caz de **liftrare** (anagramă din filtrare) a cepstrului [122] constând din anularea valorilor lui peste o anumită cvefrenţă. Dacă această operaţiune este urmată de o transformare Fourier directă a cepstrului trunchiat, rezultatul va fi **log-spectrul netezit** al semnalului, cu netezirea dependentă de trunchiere. Ca exemple, figura 2.16 prezintă estimări spectrale de bandă îngustă şi netezite prin trunchiere cepstrală ale semnalului analizat şi în figurile 2.11 şi 2.12. Deşi sunt folosiţi numai primii 17 respectiv 13 coeficienţi cepstrali (corespunzători cvefrenţelor superioare de 1 şi 0,75 ms), spectrele estimate prin trunchiere cepstrală prezintă o netezire

mai bună decât cea rezultată din predicţie liniară (figura 2.12) sau analiză Fourier de bandă largă (figura 2.11). Aceasta este una din cauzele pentru care coeficienţii cepstrali inferiori (de indice mic, $l = 1 \ldots L$, cu $L = 10 \ldots 20$) formează setul de caracteristici ale semnalului vocal cel mai utilizat actualmente în recunoaşterea automată a vorbirii.

După cum se observă şi din ecuaţiile (2.53) şi (2.54), $c_0 = c(0)$ este o măsură a energiei semnalului din cadrul analizat, iar renunţarea la el permite comparaţii spectrale între cadre de semnal cu energii diferite, astfel încât el nu prezintă interes deosebit din punct de vedere al recunoaşterii automate a vorbirii.

În încheiere, mai menţionăm că predicţia liniară de ordin P permite şi calculul unei estimări a cepstrului, numită **cepstru de predicţie liniară**, pe baza relaţiilor

$$c_l = \begin{cases} \log(G), & l = 0 \\ a_l + \sum_{k=1}^{l-1} (k/l)\, c_k\, a_{l-k}, & 1 \leq l \leq P \\ \sum_{k=1}^{l-1} (k/l)\, c_k\, a_{l-k}, & l > P \end{cases} \tag{2.55}$$

2.7 Metode perceptuale

Una din căile posibile pentru îmbunătăţirea performanţelor sistemelor de recunoaştere automată a vorbirii este includerea în analiza semnalului vocal a unor prelucrări care emulează proprietăţi ale sistemului auditiv uman [3], [180], [184], [160], [80]. După modul în care se realizează această emulare, putem distinge **metode auditorii**, care modelează elementele structurale şi funcţiile sistemului auditiv pe baza rezultatelor unor studii de anatomie, fiziologie şi psihofiziologie, şi **metode perceptuale**, care includ proprietăţile acestuia fără considerarea structurilor anatomice şi mecanismelor fiziologice implicate.

Dintre metodele auditorii pot fi considerate clasice, prin prisma volumului de cercetări deja acumulate în jurul lor, modelul cochlear al lui Lyon [148], [150], modelul auditor al lui Seneff [220] şi modelul auditor EIH (Ensemble Interval Histogram) al lui Ghitza [88], [89], dar cercetările în acest domeniu sunt încă departe de a fi epuizate, exemple în acest sens putând fi găsite începând cu domeniile modelării cochleare [124], [61] şi auditorii [222] şi mergând până la implementări hardware [250], [134].

Deşi în mod teoretic metodele auditorii pot contribui la apropierea performanţelor sistemelor de recunoaştere automată a vorbirii de cele umane, iar unele experimente susţin superioritatea lor [149], [145], practic această superioritate nu se manifestă [53] sau se manifestă numai în condiţii de mediu dificile [119], [222]. Această situaţie este explicabilă prin insuficienta cunoaştere a mecanismelor pe care aceste metode încearcă să le modeleze, ceea ce împreună cu costul computaţional ridicat face ca ele să nu aibă o utilitate clară în recunoaşterea automată a vorbirii.

Prin contrast, metodele perceptuale nu intră în detaliile modelării aparatului auditiv uman, evitând astfel dificultăţile asociate, ci se rezumă la considerarea proprietăţilor sale funcţionale puse în evidenţă prin experimente psihoacustice. Printre aceste metode se găsesc unele dintre cele mai eficiente în recunoşterea automată a vorbirii, cu o utilitate consacrată, iar în restul acestui capitol vor fi prezentate două dintre ele.

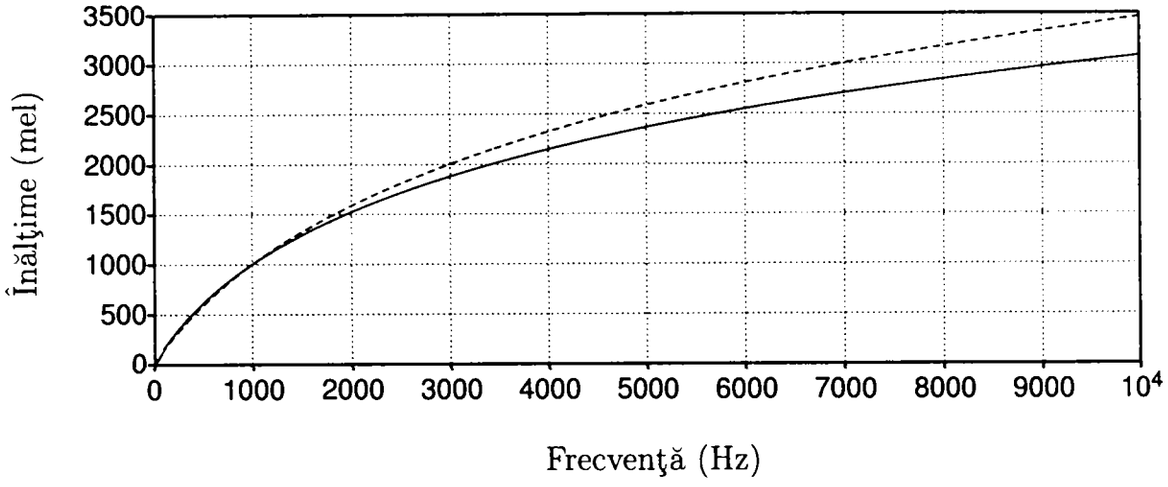

Figura 2.17: Corespondenţa dintre înălţime şi frecvenţă conform ecuaţiilor (2.56) (linie continuă) şi (2.57) (linie întreruptă)

2.7.1 Analiza melodică

Înălţimea este o proprietate subiectivă a sunetelor [228], [173], [208], dependentă în principal de frecvenţa lor, pe baza căreia ele sunt clasificate de la grave la acute. Relaţia cantitativă dintre frecvenţele şi înălţimile tonurilor pure a fost studiată experimental, iar pentru măsurarea înălţimii s-a introdus o **scară melodică** cu unitatea de măsură **mel**: prin definiţie [208], un ton pur de 1000 Hz cu un nivel de 40 dB are înălţimea 1000 mel, iar frecvenţele tonurilor pure percepute ca având înălţimi de n ori mai mari sau mai mici sunt puse în corespondenţă cu aceste înălţimi.

Rezultatele acestor experimente variază de la subiect la subiect, dar pe baza unui număr suficient de mare de subiecţi au putut fi elaborate formulări analitice ale relaţiei dintre frecvenţa f (în Hz) şi înălţimea pe scara melodică m (în mel) ale unui ton pur, două dintre ele fiind datorate lui Beranek ([22], citat în [239])

$$m = 2595 \log_{10}(1 + f/700) = 1127 \ln(1 + f/700) \tag{2.56}$$

respectiv Fant [69]

$$m = 1000 \ln(1 + f/1000)/\ln 2 \tag{2.57}$$

Reprezentările grafice asociate ambelor formule (figura 2.17) evidenţiază echivalenţa lor calitativă, dar în practică prima este cea mai frecvent utilizată în prezent, deşi există încă multe alte formule posibile [240].

După cum s-a arătat în secţiunea 2.6.1, analiza cepstrală oferă o reprezentare foarte compactă, sub forma coeficienţilor cepstrali inferiori, a informaţiei spectrale asociată tractului vocal, reţinută în spectrul netezit prin trunchiere cepstrală. **Cepstrul melodic** [57] este obţinut prin modificarea analizei cepstrale, cu luarea în consideraţie a percepţiei neliniare cu frecvenţa a înălţimii sunetelor.

Integrarea în analiza cepstrală a fenomenului percepţiei neliniare cu frecvenţa a înălţimii se face estimând un **spectru melodic** după transformarea Fourier directă

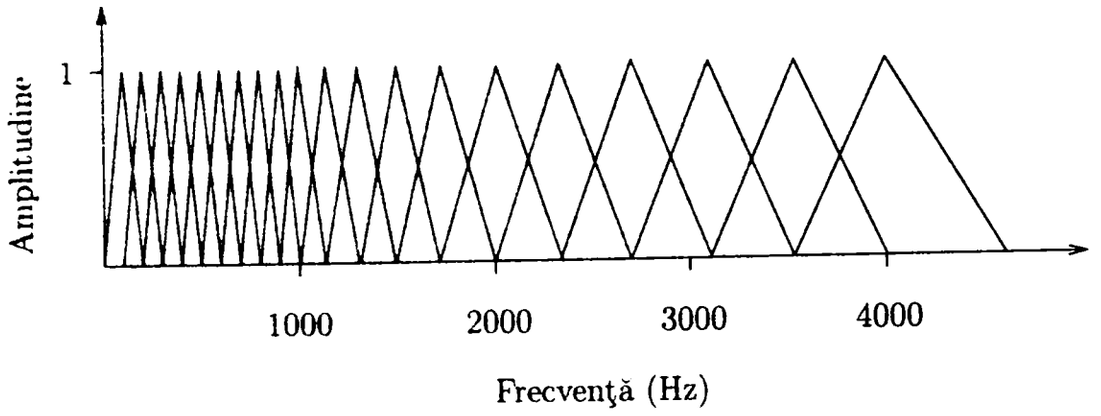

Figura 2.18: Exemple de ferestre utilizate în analiza melodică [57]

(figura 2.14) prin sumarea ponderată a componentelor spectrale din benzi de frecvenţă corespunzătoare unui număr M de filtre trece bandă. Prin această sumare sunt atenuate şi efectele structurii spectrale fine, iar pentru un cadru de semnal vocal de lungime N ea se poate realiza cu o relaţie de forma

$$Y(m) = \sum_{k=0}^{N/2-1} |S(k)|\, w_m(k)\,, \quad m = 1 \ldots M \tag{2.58}$$

unde w_m sunt ferestre care ponderează diferitele componente spectrale $S(k)$ pentru a simula caracteristicile dorite ale filtrelor trece bandă. În practică cele mai folosite sunt ferestrele triunghiulare de lărgime uniformă până la 1000 Hz şi exponenţial crescătoare, corespunzător uneia uniformă pe scara melodică, peste această frecvenţă, ca exemplu putând fi dat chiar setul de ferestre utilizat în [57] (figura 2.18), dar se pot utiliza şi ferestre rectangulare adiacente, uneori cu rezultate superioare [27].

Odată spectrul melodic disponibil, cepstrul melodic se poate calcula ca şi cel real, prin aplicarea transformatei Fourier inverse, sau se poate ţine seama de paritatea spectrului melodic şi utiliza o transformată cosinus discretă

$$c_l = \sum_{m=1}^{M} [\log Y(m)] \cos[l\,(m - 0{,}5)\,\pi/M]\,, \quad l = 1 \ldots L \tag{2.59}$$

unde L este numărul de coeficienţi cepstrali melodici care se doreşte să fie calculaţi.

2.7.2 Caracteristicile dinamice

Experimente perceptuale au arătat că informaţia fonetică este localizată mai ales în zonele cu variaţie spectrală mare ale semnalelor vocale [77], iar această constatare a condus la utilizarea acestor variaţii în recunoaşterea automată a vorbirii.

Iniţial, au fost utilizate aproximări ale valorilor derivatelor temporale ale coeficienţilor cepstrali şi log-energiei, calculate cu formula

$$d(t) = \frac{\sum_{k=-L}^{L} kc(t+k)}{\sum_{k=-L}^{L} k^2} \tag{2.60}$$

în care $c(t)$ este valoarea caracteristicii statice (coeficient cepstral sau log-energie) la momentul t, iar $2L$ este lungimea în cadre a unei ferestre, centrată la momentul t, în care sunt cuprinse valorile caracteristicii statice pe baza cărora se calculează $d(t)$.

Altă formulă [223], introdusă deoarece se considera că ecuaţia (2.60) realizează o netezire prea mare, si care are şi avantajul simplităţii, este

$$d(t) = c(t+L) - c(t-L) \tag{2.61}$$

Aceste caracteristici dinamice sunt în prezent cunoscute ca delta cepstru respectiv delta log-energie, pe scurt coeficienţi delta (Δ), iar pe baza lor se pot calcula, în mod asemănător, aproximări ale derivatelor de ordinul doi (cele de ordin superior nu s-au dovedit utile) ale caracteristicilor statice, cunoscute drept coeficienţi delta-delta ($\Delta\Delta$). Sub o formă sau alta, aceste caracteristici dinamice au fost folosite în numeroase sisteme de recunoaştere automată a vorbirii, în care eficienţa lor a fost în mod repetat verificată.

Pe lângă acestea, au existat şi există şi alte încercări de utilizare a informaţiilor despre variaţiile spectrale: prin definirea de noi caracteristici (de exemplu funcţii de variaţie spectrală [42]), prin includerea unor etape de derivare temporală a spectrului în procesul de analiză (de exemplu în prelucrările RASTA [102]), sau chiar prin integrarea unor caracteristici dinamice generalizate în structura modelelor Markov ascunse [49].

2.8 Concluzii

După cum s-a arătat la începutul capitolului 1, analiza semnalului vocal are un rol fundamental în orice aplicaţie bazată pe prelucrarea automată a vorbirii, iar în acest capitol s-a urmărit identificarea metodelor de analiză cele mai relevante pentru obiectivele propuse ale cercetărilor care stau la baza acestei teze.

Au fost astfel trecute în revistă principalele metode ce permit descrierea spectrală a semnalului vocal: analiza Fourier, analiza prin predicţie liniară, analiza cepstrală. Pentru o mai bună înţelegere a particularităţilor lor, prezentarea acestor metode a fost precedată de descrierea altor prelucrări şi parametri (cadrare, ferestruire, energie, putere, autocorelaţie, preaccentuare) care intervin pe parcursul lor, iar de-a lungul întregului capitol au fost evidenţiate multiplele legături existente între ele.

Una din posibilităţile de creştere a performanţei sistemelor de recunoaştere automată a vorbirii fiind introducerea în analiza semnalului vocal a unor prelucrări inspirate de caracteristicile aparatului auditiv uman, în finalul capitolului au fost prezentate câteva consideraţii asupra acestei posibilităţi, cu detalierea a două metode care o materializează: analiza melodică şi caracteristicile dinamice.

Înainte de a fi utilizate pentru cercetările din această teză, majoritatea metodelor au fost evaluate experimental [27] într-o aplicaţie simplă, de recunoaştere automată a unui

vocabular mic de cuvinte pronunţate izolat [26], folosind o bază de date vocale de mici dimensiuni [25] colectată ad-hoc în etapa de fundamentare a cercetărilor. Aceste evaluări au confirmat calităţile coeficienţilor cepstrali în recunoaşterea automată a vorbirii şi au arătat că performanţele cele mai bune se obţin prin utilizarea cepstrului melodic.

CAPITOLUL 3

Recunoaşterea automată a vorbirii

Indiferent de abordările folosite în încercările de a rezolva diferite probleme ale recunoaşterii automate a vorbirii, esenţială pentru aprecierea şi compararea acestora este **evaluarea performanţelor** sistemelor în care ele sunt implementate. Din acest motiv, prezentarea unor aspecte ale evaluării performanţelor sistemelor de recunoaştere automată a vorbirii precede trecerea în revistă în acest capitol a unora dintre cele mai importante metode folosite în aceste sisteme şi utilizate şi în cercetările proprii.

Majoritatea metodelor de analiză a semnalului vocal descrise în capitolul 2 estimează valori ale unor parametri locali ai acestuia, valori care în scopul recunoaşterii automate a vorbirii pot fi grupate în vectori acustici dintr-un spaţiu acustic. Astfel, o primă problemă care apare este cea a evaluării unor **distanţe acustice**, care pot fi folosite împreună cu metode de recunoaştere a formelor [64], [76], [243] pentru a încerca o estimare locală, la nivel de cadre, a identităţii lingvistice a semnalelor vocale.

Datorită inegalităţii dintre gamele în care componentele vectorilor acustici pot lua valori, distanţele acustice acordă ponderi inegale acestor componente, ceea ce duce la o anizotropie a spaţiului acustic. Această anizotropie poate fi redusă prin utilizarea unor **transformări ale spaţiului acustic**, bazate pe distribuţia vectorilor acustici.

Tratarea matematică riguroasă a majorităţii problemelor recunoaşterii automate a vorbirii a devenit posibilă prin considerarea ei ca o problemă de teoria comunicaţiei [116], ceea ce a creat premisele utilizării sistematice a unor **metode statistice**, capabile să acopere variabilitatea semnalelor vocale. În prezent, instrumentele matematice esenţiale în recunoaşterea vorbirii sunt **modelele Markov ascunse (MMA)**.

Recunoaşterea vorbirii prin metode statistice se bazează pe **modelarea acustică** a cuvintelor şi **modelarea lingvistică** a succesiunii lor. Problema principală în această teză fiind modelarea acustică pentru recunoaşterea automată a vorbirii continue în limba română (secţiunea 1.3), în continuare vor fi detaliate în special aspecte legate de aceasta.

Pe lângă modelarea acustică şi cea lingvistică, recunoaşterea automată a vorbirii mai necesită şi **reprezentarea integrată a cunoştinţelor** acumulate de modele într-un spaţiu unic de căutare a soluţiilor. În acest spaţiu, estimarea şirului de cuvinte pronunţat se poate face prin aplicarea unor **algoritmi de căutare** având ca obiectiv maximizarea probabilităţii unui ipotetic şir de cuvinte în raport cu modelele.

3.1 Evaluarea performanţelor

Sistemele de recunoaştere automată a vorbirii pot face erori, ca şi oamenii, în ceea ce priveşte şirul de cuvinte pe care îl estimează a fi fost pronunţat (figura 1.1), iar obiectivul lor fiind tocmai estimarea corectă a acestui şir, principalele metrice pentru evaluarea performanţelor lor se definesc pe baza erorilor apărute în acest proces.

Performanţele pot fi influenţate de foarte mulţi factori, astfel încât orice valori ale metricelor trebuie însoţite de precizarea condiţiilor în care au fost obţinute, condiţii determinate de caracteristicile aplicaţiei şi ale sistemului de recunoaştere evaluat. Câteva dintre aceste condiţii sunt:

- **mărimea vocabularului**, dată de numărul de cuvinte distincte din dicţionarul sistemului evaluat şi pe care acesta le poate recunoaşte: într-o aplicaţie necesitând un vocabular redus, în care apar puţine cuvinte între care sistemul trebuie să aleagă la un moment dat, erorile pot apare mai rar decât într-o aplicaţie cu un vocabular mai mare, între cuvintele căruia pot exista mai multe posibilităţi de confuzie;

- **tipul vocabularului**: funcţie de luarea sau nu în calcul a posibilităţii ca în pronunţii să apară şi cuvinte din afara lui, putem avea un vocabular **deschis**, în care un singur cuvânt special va corespunde tuturor cuvintelor necunoscute care pot apare în pronunţii, sau un vocabular **închis**;

- **cuvintele din vocabular**: chiar dacă o aplicaţie utilizează un vocabular mic şi închis, prezenţa în el şi utilizarea în proporţie mare a unor cuvinte uşor de confundat din punct de vedere acustic va duce la creşterea frecvenţei erorilor;

- **tipul pronunţiilor**: un sistem pentru recunoaşterea pronunţiilor **discrete**, a unor cuvinte rostite cu pauze între ele, va putea obţine performanţe superioare unuia pentru recunoaşterea vorbirii **continue**, fără pauze între cuvinte, care trebuie să realizeze şi delimitarea, nu numai recunoaşterea cuvintelor;

- **stilul vorbirii**: sunt mai uşor de recunoscut semnalele vocale obţinute prin **citire**, respectând restricţiile impuse de un text, decât cele rezultate dintr-o **conversaţie** spontană, în care regulile gramaticale pot să nu fie respectate, apar ezitări, reluări, vorbitorii se pot întrerupe reciproc, se pot suprapune etc.;

- **gramatica sau modelul lingvistic al aplicaţiei**: în cazul unei aplicaţii în care secvenţa cuvintelor este rigidă, iar fiecare cuvânt are un număr redus de succesori posibili, erorile vor fi mai puţin frecvente decât în cazul uneia în care cuvintele se pot succeda mai liber, iar succesorii unui cuvânt pot fi în număr mai mare;

- **tipul modelelor acustice**: un sistem cu modele **dependente** de vorbitor poate avea performanţe mai bune decât unul cu modele **independente** de vorbitor, iar performanţele acestuia din urmă pot creşte prin **adaptare** la vorbitor a modelelor;

- **condiţiile acustice** ale mediului în care sunt rostite pronunţiile: zgomotele, reverberaţiile etc. vor duce la scăderea calităţii semnalului şi a performanţelor;

- **caracteristicile canalului de comunicaţie** dintre utilizator şi sistem: un canal de comunicaţie cu o bandă mai îngustă (de exemplu de tip telefonic) sau afectat de zgomote, distorsiuni etc. poate duce la scăderea performanţelor;

- **vorbitorii** folosiţi pentru realizarea evaluării, caracteristicile vocii şi ale vorbirii fiind dependente de multe variabile biologice (sex, vârstă) şi sociale (grad de educaţie, ocupaţie, mediu social) care pot afecta performanţele fie prin intermediul semnalului vocal, fie prin modul de utilizare a sistemului în aplicaţie.

Erorile din funcţionarea unui sistem de recunoaştere automată a vorbirii pot apare datorită imperfecţiunii modelelor acustice sau lingvistice folosite (**erori de modelare**) sau eliminării greşite din procesul de căutare a unor cuvinte corecte (**erori de căutare**). Funcţie de modul în care se manifestă la nivelul şirului de cuvinte estimat, aceste erori se pot clasifica în:

- **substituţii** ale unui cuvânt pronunţat cu un altul, datorate asemănării acustice dintre ele în condiţiile unei modelări acustice imperfecte, sau absenţei cuvântului corect dintre alternativele permise la un moment dat;

- **inserţii** ale unor cuvinte nepronunţate: acestea sunt de obicei cuvinte scurte, cel mai adesea monosilabice (prepoziţii, conjuncţii, pronume), asemănătoare din punct de vedere acustic cu porţiuni ale unor cuvinte mai lungi;

- **omisiuni** (în engl. deletions) ale unor cuvinte pronunţate.

Evaluarea performanţelor unui sistem de recunoaştere automată a vorbirii presupune existenţa unor transcrieri de referinţă ale semnalelor vocale folosite în acest scop, cu care să fie comparate şirurile corespunzătoare de cuvinte estimate de sistem, şi a unor metode de determinare a apariţiei şi tipului erorilor. Odată detectate şi clasificate erorile, se pot calcula valorile unor metrice de performanţă:

- frecvenţa recunoaşterilor corecte sau corectitudinea

$$C = \frac{N_C}{N} \cdot 100 \quad [\%] \tag{3.1}$$

- frecvenţa substituţiilor

$$S = \frac{N_S}{N} \cdot 100 \quad [\%] \tag{3.2}$$

- frecvenţa inserţiilor

$$I = \frac{N_I}{N} \cdot 100 \quad [\%] \tag{3.3}$$

- frecvenţa omisiunilor

$$O = \frac{N_O}{N} \cdot 100 \quad [\%] \tag{3.4}$$

unde N, N_C, N_S, N_I şi N_O sunt numărul de cuvinte de referinţă, corect recunoscute, substituite, inserate şi respectiv omise.

Fiecare din metricele anterioare oferă o imagine doar asupra unui singur aspect din funcţionarea unui sistem, astfel încât pentru o caracterizare globală a performanţelor au mai fost introduse două metrice care ţin cont de toate tipurile de erori posibile:

- frecvenţa erorilor

$$E = \frac{N_S + N_I + N_O}{N} \cdot 100 \quad [\%] \tag{3.5}$$

- acurateţea

$$A = 100 - E = (1 - \frac{N_S + N_I + N_O}{N}) \cdot 100 \quad [\%] \tag{3.6}$$

3.1.1 Compararea prin programare dinamică

Determinarea cuvintelor recunoscute corect, substituite, inserate sau omise presupune marcarea cu unul din aceste atribute a fiecărui cuvânt din şirurile estimate de sistem prin comparaţii cu transcrierile de referinţă corespunzătoare. În mod tradiţional, şirurile estimate sunt considerate ipoteze formulate de sistem, iar rezultatele comparaţiilor sunt prezentate ca alinieri ale ipotezelor cu transcrierile de referinţă, însoţite de tipul fiecărei erori detectate – recunoaşterile corecte nu sunt marcate:

Referinţă:	cuv_ref_1	cuv_ref_2	cuv_ref_3		...	cuv_ref_{N_r}
Ipoteză:		cuv_ip_1	cuv_ip_2	cuv_ip_3	...	cuv_ip_{N_i}
Evaluare:	O	S		I		

Compararea unei ipoteze cu transcrierea de referinţă corespunzătoare se poate face calculând printr-un algoritm de programare dinamică distanţa de editare dintre ele, definită ca fiind costul total minim al operaţiunilor de editare (substituţie, inserţie sau ştergere/omisiune), fiecare cu un cost specific, prin care unul din cele două şiruri de cuvinte este transformat în celălalt [249].

Programarea dinamică consta în sinteza dinamică a unor strategii sau programe (de unde şi numele) global optime din decizii succesive optime la fiecare moment. Ea se bazează pe **principiul de optimizare** enunţat de Bellman [21]: "O strategie optimă are proprietatea că, oricare ar fi starea iniţială şi decizia iniţială, deciziile rămase trebuie să constituie o strategie optimă în raport cu starea care rezultă din prima decizie."

Ţinând cont de acest principiu, evaluarea distanţei de editare $D(I, R)$ dintre ipoteza I de N_i cuvinte şi referinţa asociată R de N_r cuvinte se poate face recursiv: notând cu C_s, C_i şi C_o costul unei substituţii, inserţii respectiv omisiuni, distanţa parţială $D(I_{1,i}, R_{1,r})$ dintre prefixele $I_{1,i}$, $R_{1,r}$ de lungimi i şi r ale I respectiv R poate fi scrisă

$$D(I_{1,i}, R_{1,r}) = \min \begin{cases} D(I_{1,i-1}, R_{1,r-1}), & I_i \text{ corect} \\ D(I_{1,i-1}, R_{1,r-1}) + C_s, & I_i \text{ substituit lui } R_r \\ D(I_{1,i-1}, R_{1,r}) + C_i, & I_i \text{ inserat} \\ D(I_{1,i}, R_{1,r-1}) + C_o, & R_r \text{ omis} \end{cases} \tag{3.7}$$

Algoritmul 3.1 Compararea prin programare dinamică a unei ipoteze cu referinţa

1: $N_i \leftarrow$ lungimea ipotezei I, $N_r \leftarrow$ lungimea referinţei R
2: *alocă* $D[N_i, N_r]$, $E[N_i, N_r]$, $e[N_i + N_r - 1]$
3: $D[i, r] \leftarrow \infty$, $i = 1 \ldots N_i$, $r = 1 \ldots N_r$
4: **for** $i = 1 \ldots N_i$ **do**
5: **for** $r = 1 \ldots N_r$ **do**
6: $D[i, r] \leftarrow$ valoarea conform ecuaţiei (3.7)
7: $E[i, r] \leftarrow$ evaluarea care a minimizat $D[i, r]$
8: **end for**
9: **end for**
10: $n = 0$ {refacere în ordine inversă}
11: **while** $i > 0$ şi $r > 0$ **do**
12: $e[n + 1] = E[i, r], n = n + 1, i = i - 1$ **if** $e[n] \neq$ omis, $r = r - 1$ **if** $e[n] \neq$ inserat
13: **end while**

Această recursie se poate implementa într-o matrice de distanţe D de dimensiune $N_i \times N_r$, în care fiecare linie corespunde unui cuvânt recunoscut şi fiecare coloană - unuia de referinţă, iar distanţa finală se obţine în $D[N_i, N_r]$.

Alinierea necesită refacerea şirului de evaluări care au minimizat distanţele parţiale calculate succesiv conform ecuaţiei (3.7). Aceasta impune păstrarea evaluărilor într-o a doua matrice E de dimensiune $N_i \times N_r$, astfel încât odată calculată $D[N_i, N_r]$, şirul de evaluări care au dus la obţinerea ei să poată fi refăcut prin parcurgere în sens invers.

Întregul proces de comparare este descris de algoritmul 3.1: în liniile 1–3 se pregătesc variabilele necesare, liniile 4–9 corespund calculului $D(I, R)$, iar în liniile 10–13 se reface secvenţa $e[n]$ de evaluări (corect, substituit, inserat, omis) care au dus la minimizarea distanţelor parţiale şi finală. Odată această secvenţă disponibilă, calculul diferitelor metrice de performanţă şi alinierea celor două şiruri de cuvinte sunt imediate.

Costurile asociate erorilor (C_s, C_i şi C_o) în ecuaţia (3.7) determină în bună măsură comportamentul algoritmului 3.1. Cea mai simplă posibilitate este de a le atribui valori egale, dar pentru evaluarea sistemelor de recunoaştere automată a vorbirii prin aliniere la nivel de cuvânt [181], [72], [260] se aleg de obicei valori ale acestora

$$C_i + C_o > C_s > C_i = C_o \tag{3.8}$$

astfel încât o substituţie să fie preferată unei perechi inserţie+omisiune sau invers.

Dat fiind criteriul de minimizare urmărit, o asemenea alegere face ca algoritmul 3.1 să favorizeze inserţiile şi omisiunile, care vor fi plasate în poziţii premergătoare substituţiilor chiar şi atunci când nu este cazul. Ameliorarea acestei situaţii se poate obţine prin coborârea la nivel sublexical în calculul $D(I, R)$ şi utilizarea drept costuri a unor distanţe motivate fonetic sau fonologic între unităţi acustice sublexicale [188], [187], [182], [72].

3.2 Distanţe acustice

Dată fiind natura nestaţionară a semnalului vocal, analiza lui se face, după cum am văzut în capitolul 2, la nivelul unor cadre cu lungimi de ordinul $n \times 10$ ms, fiecare descris

prin valorile anumitor parametri. În termenii generali ai recunoaşterii formelor [64], [76], [243], aceşti parametri pot fi grupaţi în vectori caracteristici, elemente ale unui spaţiu multidimensional al caracteristicilor. În cazul recunoaşterii vorbirii, vectorii caracteristici sunt cunoscuţi drept **vectori acustici**, elemente ale unui **spaţiu acustic**.

Evaluarea distanţelor dintre vectorii caracteristici este o problemă fundamentală în recunoaşterea formelor, complicată în cazul distanţelor acustice de cerinţa ca ele să aibă pe cât posibil şi o interpretare perceptuală: distanţa dintre doi vectori acustici ar trebui să aibă valori mari dacă porţiunile de semnal din care provin sunt diferite din punct de vedere lingvistic, şi valori mici, ideal nule, în cazul identităţii lor lingvistice.

De-a lungul timpului au fost propuse şi studiate multe distanţe acustice [224], [194], toate bazate pe faptul că informaţia spectrală este esenţială pentru identificarea sunetelor vorbirii [160], [184], [180], [3]. Distanţa acustică dintre două cadre ar putea fi deci evaluată în primă instanţă printr-o metrică Minkowsky de ordin p:

$$l_p(X,Y) = \sqrt[p]{\sum_{k=0}^{N-1} |X(k) - Y(k)|^p} \tag{3.9}$$

unde X, Y sunt spectrele discrete ale celor două cadre.

Ţinând cont că tăria percepută subiectiv a unui sunet, conform legii Weber-Fechner, este proporţională cu logaritmul intensităţii lui obiective, iar recunoaşterea de către oameni a vorbirii nu depinde de faza diferitelor componente spectrale ale acesteia, o distanţă mai potrivită din punct de vedere perceptual este una în care spectrul este înlocuit cu log-spectrul de amplitudine:

$$d_p(X,Y) = \sqrt[p]{\sum_{k=0}^{N-1} |log|X(k)| - log|Y(k)||^p} \tag{3.10}$$

Conform secţiunii 2.6.1, ecuaţia (2.54), log-spectrul de amplitudine este transformata Fourier a cepstrului real:

$$log|S(k)| = \sum_{l=0}^{N-1} c(l) e^{-j(2\pi/N)kl} \tag{3.11}$$

astfel încât pentru $p = 2$ ecuaţia (3.10) devine:

$$d_2(X,Y) = \sqrt{\sum_{l=0}^{N-1} |c_X(l) - c_Y(l)|^2} \tag{3.12}$$

sau, în cazul utilizării trunchierii cepstrale:

$$d_2(X,Y) \approx \sqrt{\sum_{l=0}^{L} |c_X(l) - c_Y(l)|^2} \tag{3.13}$$

Aceasta înseamnă că distanţa euclidiană dintre vectorii cepstrali satisface cerinţele perceptuale enunţate anterior, ceea ce poate fi una din explicaţiile succesului analizei cepstrale ca metodă de extragere a caracteristicilor pentru recunoaşterea vorbirii.

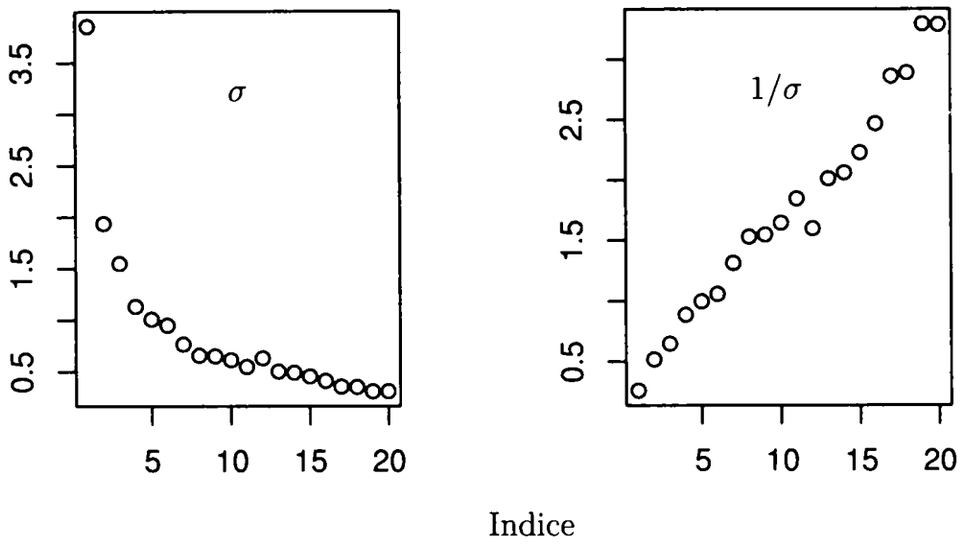

Figura 3.1: Abaterile standard şi inversele lor pentru primii 20 de coeficienţi cepstrali melodici, estimate din 9802 cadre (98 secunde) de semnal vocal.

3.3 Transformări ale spaţiului acustic

Una dintre problemele utilizării distanţelor în recunoaşterea formelor [64], [76], [243] este cauzată de diferenţele dintre dispersiile componentelor vectorilor caracteristici. Ca exemplu, prezentăm în figura 3.1 valorile abaterilor standard şi ale inverselor lor pentru primii 20 de coeficienţi cepstrali melodici, valori estimate din 9802 cadre (98 secunde) de semnal vocal obţinut prin citirea de către 20 de vorbitori uniform distribuiţi pe sexe şi grupe de vârstă a unei propoziţii în care apar toate unităţile din tabelul 4.1.

Se observă că valorile $1/\sigma$ au o variaţie cvasiliniară cu indicii coeficienţilor, deci valorile σ sunt invers proporţionale cu indicii coeficienţilor cepstrali. În mod corespunzător, coeficienţii cepstrali vor avea ponderi invers proporţionale cu indicii lor în distanţele euclidiene dintre vectorii acustici cepstrali.

O altă problemă poate apare dacă două sau mai multe componente ale vectorilor caracteristici au covarianţa nenulă

$$c_{ij} = \mathcal{E}\{(x_i - \mu_i)(x_j - \mu_j)\} \neq 0 \tag{3.14}$$

unde x_i, x_j şi μ_i, μ_j sunt două componente ale vectorilor, respectiv valorile lor medii. Aceasta indică posibilitatea corelaţiei liniare a celor două componente, care poate duce la evaluări eronate ale distanţelor prin luarea repetată în calcul a unei aceleiaşi variaţii.

În cazul general, soluţia acestor probleme este decorelarea caracteristicilor şi normalizarea dispersiilor lor prin aplicarea unei transformări liniare. Aceasta poate fi obţinută plecând de la matricea de covarianţă $\boldsymbol{C} = [c_{ij}]$, care fiind simetrică ($c_{ij} = c_{ji}$) poate fi scrisă

$$\boldsymbol{C} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T \tag{3.15}$$

unde Φ este matricea vectorilor ei proprii normaţi, iar Λ matricea diagonală a valorilor ei proprii. Deoarece

$$\mathcal{E}\{\Lambda^{-1/2}\Phi^T[\boldsymbol{x}-\boldsymbol{\mu_x}][\boldsymbol{x}-\boldsymbol{\mu_x}]^T\Phi\Lambda^{-1/2}\} = \boldsymbol{I} \tag{3.16}$$

rezultă că transformarea căutată este

$$\boldsymbol{y} = \Lambda^{-1/2}\Phi^T\boldsymbol{x} \tag{3.17}$$

Utilizarea transformării (3.17) este costisitoare din punct de vedere computaţional, iar reducerea acestui cost poate fi făcută ţinând cont de particularităţile caracteristicilor folosite. În cazul coeficienţilor cepstrali, o asemenea particularitate este corelaţia lor redusă datorată transformării ortogonale prin care sunt obţinuţi.

Renunţând la decorelare, normalizarea dispersiilor coeficienţilor cepstrali poate fi aproximată mult mai simplu prin înmulţirea lor cu ponderi constante w_i care pot fi [237] a) inversele dispersiilor: $w_i = 1/\sigma_i$; b) indicii coeficienţilor cepstrali: $w_i = i$. Ambele variante pot fi privite fie ca metode de introducere a unor distanţe ponderate, fie ca operaţiuni de liftrare, iar un studiu efectuat din această ultimă perspectivă [122] a condus la ponderi de forma

$$w_i = 1 + \frac{L}{2}\sin\frac{i\pi}{L} \tag{3.18}$$

utilizate actualmente în majoritatea sistemelor de recunoaştere automată a vorbirii.

3.4 Metode statistice de recunoaştere a vorbirii

Pe lângă deformarea dinamică a timpului, la începutul anilor '70 se mai utilizau în cercetările asupra recunoaşterii automate a vorbirii metode bazate pe cunoştinţe, implementate în sisteme expert, cele mai cunoscute fiind cele dezvoltate în cadrul unui program ARPA [141], [257], [127]. Pentru transcrierea semnalelor vocale, acestea foloseau informaţii furnizate de aşa-numite surse de cunoştinţe (fonetice, fonologice, lexicale, sintactice, semantice şi pragmatice), în fapt seturi de reguli formulate de experţi umani.

Ca şi tiparele în cazul deformării dinamice a timpului, regulile s-au dovedit incapabile să reprezinte variabilitatea semnalelor vocale. O metodă care s-a dovedit adecvată pentru reprezentarea acestei variabilităţi a fost propusă şi implementată în sistemul DRAGON [16], [15] dezvoltat la CMU în acelaşi program al ARPA. Metoda folosea un model teoretic general bazat pe funcţii de probabilitate ale unor procese Markov, care în timp au ajuns să fie cunoscute sub numele de modele Markov ascunse – MMA (în engl. Hidden Markov Models – HMM) şi a căror prezentare va fi detaliată în restul capitolului.

Aceeaşi metodă a fost propusă, în mod independent, de cercetători de la IBM [116], [13], [113] plecând de la tratarea recunoaşterii automate a vorbirii ca o problemă de teoria comunicaţiei, ceea ce a oferit un cadru teoretic cuprinzător pentru tratarea matematică riguroasă a diferitelor ei aspecte, nu doar a metodei de reprezentare a cunoştinţelor.

Această abordare utilizează o particularizare a modelului general al unui sistem de comunicaţie [221], prezentată în figura 3.2: sursa de informaţie din modelul general devine o abstractizare a proceselor cognitive premergătoare formulării unui mesaj şi

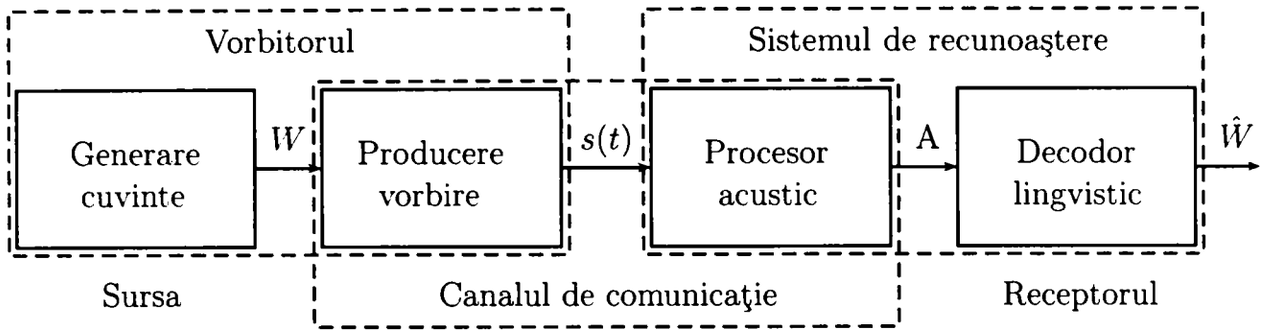

Figura 3.2: Recunoaşterea vorbirii din perspectiva teoriei comunicaţiei

având ca rezultat generarea unui şir de cuvinte W; şirul de cuvinte W este transformat în semnalul vocal $s(t)$ de către organele de producere a vorbirii; acestea împreună cu procesorul acustic din sistemul de recunoaştere automată a vorbirii formează canalul de comunicaţie; iar receptorul este decodorul lingvistic al sistemului de recunoaştere, care generează o estimare $\hat{W}$ a şirului de cuvinte pronunţat.

Pentru găsirea $\hat{W}$, decodorul lingvistic utilizează şirul A de date acustice[1] obţinut de procesorul acustic din semnalul vocal $s(t)$. Dacă pentru un şir de cuvinte oarecare

$$W = w_1 w_2 \dots w_n \tag{3.19}$$

probabilitatea ca el să fi fost pronunţat, date fiind datele acustice A, este $P(W|A)$, pentru minimizarea frecvenţei erorilor sistemul va alege

$$\hat{W} = \arg\max_W P(W|A) \tag{3.20}$$

sau, exprimând $P(W|A)$ conform formulei lui Bayes

$$P(W|A) = \frac{P(W) \cdot P(A|W)}{P(A)} \tag{3.21}$$

şi neglijând $P(A)$, aceeaşi pentru toate şirurile de cuvinte W

$$\hat{W} = \arg\max_W [P(W) \cdot P(A|W)] \tag{3.22}$$

Ecuaţia (3.22) pune în evidenţă cele două probleme esenţiale pentru recunoaşterea automată a vorbirii prin metode statistice:

- modelarea generării cuvintelor, cunoscută drept **modelare lingvistică** (în engl. language modeling), astfel ca utilizând modelul lingvistic (language model) rezultat să se poată estima $P(W)$; aceasta corespunde surselor de cunoştinţe de pe nivelurile superioare (sintactic, semantic, pragmatic) din metodele bazate pe cunoştinţe;

[1] De-a lungul timpului au existat unele variaţiuni, dar în esenţă A este un şir de vectori acustici.

- modelarea producerii vorbirii, cunoscută drept **modelare acustică**, având ca obiectiv construirea unor modele acustice pe baza cărora să fie calculată probabilitatea $P(A|W)$ ca datele acustice A să fi fost obţinute în urma pronunţării şirului de cuvinte W; în cadrul metodelor bazate pe cunoştinţe, aceasta era realizată prin intermediul surselor de pe nivelurile inferioare (fonetic, fonologic, lexical).

Datorită generalităţii menţionate anterior, modelele Markov ascunse pot fi utilizate şi pentru modelarea lingvistică şi pentru cea acustică, astfel încât o scurtă prezentare a lor precede discutarea celor două probleme.

3.5 Modelele Markov ascunse

Modelele Markov ascunse – MMA (în engl. Hidden Markov Models – HMM) [195], [105], [59], [115] sunt automate finite stochastice folosite pentru descrierea statisticilor locale şi a evoluţiilor globale ale caracteristicilor unor procese aleatoare nestaţionare, dar care pot fi considerate staţionare pe porţiuni, prin **funcţii de probabilitate** a valorilor acestor caracteristici. În modelele Markov ascunse, tranziţiile între stări se fac conform unor **probabilităţi de tranziţie**, iar producerea/observarea caracteristicilor şi funcţiile de probabilitate care modelează distribuţiile valorilor lor pot fi asociate stărilor sau tranziţiilor, în mod similar automatelor Moore respectiv Mealy. În continuare vom considera doar cazul funcţiilor de probabilitate asociate stărilor.

Un model Markov ascuns este definit prin:

- mulţimea stărilor $\mathcal{S} = \{s_i, i = 1 \ldots N\}$;
- matricea de tranziţie $\mathcal{A} = [a_{ij}]$, unde $a_{ij} = P[s(t+1) = s_j | s(t) = s_i], i, j = 1 \ldots N$ sunt probabilităţile tranziţiilor între stări;
- mulţimea probabilităţilor iniţiale ale stărilor $\Pi = \{\pi_i = P[s(0) = s_i], i = 1 \ldots N\}$;
- mulţimea valorilor caracteristicilor procesului modelat $\mathcal{Y}$;
- $\mathcal{B} = \{b_j(y) | b_j(y = o_t) = P[o_t | s(t) = s_j], j = 1 \ldots N\}$ – mulţimea funcţiilor de probabilitate a valorilor caracteristicilor observate/produse în fiecare stare, unde prin o_t am notat valoarea observată (observaţia, pe scurt) la momentul t.

În această definiţie, mulţimea stărilor împreună cu mulţimea probabilităţilor iniţiale şi matricea de tranziţie corespund unui lanţ/proces/model Markov. Stările lui nu sunt însă observabile direct, ci doar prin intermediul observaţiilor generate conform funcţiilor de probabilitate asociate fiecărei stări, de unde numele de model Markov ascuns.

Modelele Markov ascunse au fost folosite iniţial cu succes în recunoaşterea automată a vorbirii, apoi în multe alte aplicaţii în care procesul studiat poate fi modelat secvenţial: prelucrarea limbajului natural [152], recunoaşterea scrisului de mână [48], recunoaşterea feţei [168], decodarea genomului uman [128] etc. Acest lucru se datorează existenţei unor algoritmi eficienţi pentru rezolvarea a trei **probleme fundamentale**:

- **evaluarea** probabilităţii $P(O|M)$ ca un şir de observaţii $O = o_1 o_2 \ldots o_T$ să apară în urma unei realizări a procesului modelat de modelul M;

- **decodarea** secvenţei de stări $S = s(1)s(2)\ldots s(T)$ parcursă prin model pentru producerea şirului de observaţii O;
- **estimarea** parametrilor (probabilităţile iniţiale π_i şi de tranziţie a_{ij} şi funcţiile de probabilitate b_j) sau **antrenarea** MMA pe baza unor date corespunzătoare.

În toate cazurile, recursia joacă un rol esenţial pentru calculul eficient al diferitelor mărimi care intervin. Astfel, probabilităţile $\alpha_i(t)$ ca modelul M să se afle în starea s_i la momentul t după ce a produs prefixul $o_1 \ldots o_t$ din şirul de observaţii $O = o_1 o_2 \ldots o_T$

$$\alpha_i(t) = P[o_1 \ldots o_t, s(t) = s_i | M] \tag{3.23}$$

se pot calcula recursiv:

$$\alpha_i(1) = \pi_i b_i(o_1), \quad i = 1 \ldots N \tag{3.24}$$

$$\alpha_j(t) = \left[\sum_{i=1}^{N} \alpha_i(t-1) a_{ij}\right] b_j(o_t), \quad j = 1 \ldots N, \quad t = 2 \ldots T \tag{3.25}$$

Deoarece recursia se face în sensul direct al timpului t, $\alpha_i(t)$ sunt numite probabilităţi "înainte", iar pe baza lor se poate rezolva prima dintre problemele de mai sus:

$$P(O|M) = \sum_{s_i \text{ finală}} \alpha_i(T) \tag{3.26}$$

Similar se pot calcula probabilităţile "înapoi" $\beta_i(t)$ ca modelul M, aflat în starea s_i la momentul t, să producă în continuare sufixul $o_{t+1} \ldots o_T$

$$\beta_i(t) = P[s(t) = s_i, o_{t+1} \ldots o_T | M] \tag{3.27}$$

dar recursia are loc în sens invers:

$$\beta_i(T) = \begin{cases} 1 & \text{pentru } s_i \text{ stare finală} \\ 0 & \text{altfel} \end{cases} \tag{3.28}$$

$$\beta_i(t) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_j(t+1), \quad i = 1 \ldots N, \quad t = T-1 \ldots 1 \tag{3.29}$$

unde iniţializarea $\beta_i(T)$ s-a făcut astfel încât ecuaţia (3.26) să devină o particularizare a cazului general

$$P(O|M) = \sum_{i=1}^{N} \alpha_i(t) \beta_i(t), \quad t = 1 \ldots T \tag{3.30}$$

A doua dintre problemele enunţate, decodarea secvenţei de stări parcurse printr-un model M pentru producerea unui şir de observaţii O, are soluţii dependente de criteriul de optim utilizat. Un prim criteriu ar putea fi probabilitatea stării s_i la momentul t date fiind O şi M

$$\gamma_i(t) = P[s(t) = s_i | O, M] = \frac{P[s(t) = s_i, O|M]}{P(O|M)} \tag{3.31}$$

care poate fi scrisă funcţie de $\alpha_i(t)$ şi $\beta_i(t)$

$$\gamma_i(t) = \frac{\alpha_i(t)\beta_i(t)}{P(O|M)} \tag{3.32}$$

astfel încât se poate face estimarea

$$s(t) = \arg\max_{s_i \in S} \gamma_i(t), \quad t = 1 \ldots T \tag{3.33}$$

Dacă modelul conţine stări între care nu există tranziţii, secvenţa rezultată poate să nu formeze o cale continuă prin model, fiind deci invalidă. Eliminarea acestei probleme se poate face prin impunerea condiţiei suplimentare ca între stările succesive să existe tranziţii, iar rezultatul este algoritmul Viterbi [247] care va fi prezentat în secţiunea 3.9.

Ultima dintre probleme, cea a antrenării MMA, are soluţii dependente de tipul mulţimii $\mathcal{Y}$ a valorilor caracteristicilor procesului modelat. În cazul în care $\mathcal{Y}$ este o mulţime discretă, funcţiile $b_j(y)$ sunt distribuţii de probabilitate, iar MMA rezultate sunt cunoscute ca **MMA discrete**. Dacă $\mathcal{Y}$ este continuă, $b_j(y)$ sunt cel mai adesea densităţi parametrice de probabilitate, rezultând **MMA continue**. Alternativa este reprezentată de metodele neparametrice de estimare a probabilităţilor: reţele neuronale [36], [205], [161], metoda celor mai apropiaţi k vecini [138], maşini cu vectori suport (în engl. Support Vector Machines - SVM) [78]. Deşi acestea sunt în fapt doar variante de estimare a probabilităţilor, sistemele rezultate sunt considerate **sisteme hibride**.

3.5.1 Antrenarea MMA discrete

Antrenarea MMA presupune estimarea valorilor parametrilor acestora: întrucât nu se cunosc soluţii analitice ale problemei, toate metodele de antrenament se bazează pe optimizarea iterativă a parametrilor folosind date de antrenament corespunzătoare şi diferite criterii de optim. Dintre criterii, cel mai frecvent folosit este cel al plauzibilităţii maxime (în engl. maximum likelihood), conform căruia estimarea de maximă plauzibilitate a parametrilor unui model M este cea care maximizează probabilitatea condiţionată ca datele de antrenament să fi fost generate de M, date fiind valorile parametrilor lui.

Problema estimării parametrilor MMA este complicată de "ascunderea" stărilor s_i în spatele şirurilor de observaţii $O = o_1 o_2 \ldots o_T$, astfel încât datele de antrenament sunt incomplete: chiar dacă se cunosc valorile observaţiilor, nu se ştie căror stări le sunt asociate. Estimarea de maximă plauzibilitate a parametrilor din date incomplete se poate face prin algoritmul de maximizare a aşteptării (Expectation-Maximization) [60], [159], a cărui particularizare la MMA este algoritmul Baum-Welch [19], [20]. Acesta realizează o optimizare iterativă a parametrilor unui model M prin calculul pe baza lor a aşteptărilor (speranţelor matematice) pentru numerele de apariţii ale unor evenimente, urmat de reestimarea parametrilor folosind valorile acestor aşteptări.

Dacă parametrii modelului M sunt reestimaţi pe baza a R realizări ale procesului modelat, care au produs R secvenţe de observaţii $O_r = o_{r,1} o_{r,2} \ldots o_{r,T_r}, r = 1 \ldots R$, calculul aşteptărilor se face sumând probabilităţile corespunzătoare pentru toate secvenţele şi la toate momentele relevante.

Probabilităţile iniţiale reestimate vor fi astfel

$$\hat{\pi}_i = \frac{1}{R}\sum_{r=1}^{R}\gamma_{i,r}(1)\,, \quad i = 1\ldots N \tag{3.34}$$

unde $\gamma_{i,r}(1)$ este probabilitatea ca modelul să se afle în starea s_i la momentul $t = 1$ de pe parcursul generării secvenţei O_r.

Reestimarea probabilităţilor de tranziţie pe baza numerelor aşteptate de tranziţii între stări necesită calculul probabilităţilor tranziţiilor din starea s_i la momentul t în starea s_j la momentul $t+1$ date fiind şirul de observaţii O şi modelul M

$$\xi_{ij}(t) = P[s(t) = s_i, s(t+1) = s_j | O, M] = \frac{P[s(t) = s_i, s(t+1) = s_j, O|M]}{P(O|M)} \tag{3.35}$$

Probabilităţile conjugate $P[s(t) = s_i, s(t+1) = s_j, O|M]$ pot fi calculate ca produse ale probabilităţilor ca modelul: 1) să se afle în starea s_i la momentul t după ce a generat prefixul $o_1 \ldots o_t$; 2) să treacă din starea s_i în starea s_j; 3) în starea s_j să genereze o_{t+1}; 4) din starea s_j la momentul $t+1$ să genereze în continuare $o_{t+2} \ldots o_T$. Ţinând cont de expresiile acestor probabilităţi, obţinem

$$\xi_{ij}(t) = \frac{\alpha_i(t) a_{ij} b_j(o_{t+1}) \beta_j(t+1)}{P(O|M)}\,, \quad i,j = 1\ldots N\,, \quad t = 1\ldots T-1 \tag{3.36}$$

iar **probabilităţile de tranziţie** reestimate vor fi

$$\hat{a}_{ij} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r-1}\xi_{ij,r}(t)}{\sum_{r=1}^{R}\sum_{t=1}^{T_r-1}\gamma_{i,r}(t)}\,, \quad i,j = 1\ldots N \tag{3.37}$$

MMA discrete au o mulţime finită a observaţiilor, $\mathcal{Y} = \{y_k, k = 1\ldots K\}$, astfel încât **funcţiile de probabilitate** $b_j(y)$ sunt distribuţii de probabilitate a căror reestimare se poate face prin simpla numărare a apariţiilor valorilor y_k

$$\hat{b}_j(y_k) = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_{j,r}(t)\cdot\delta(o_{r,t}, y_k)}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_{j,r}(t)}\,, \quad j = 1\ldots N\,, \quad k = 1\ldots K \tag{3.38}$$

unde δ este simbolul lui Kronecker: dacă $x = y$, $\delta(x,y) = 1$, altfel $\delta(x,y) = 0$.

Algoritmul Baum-Welch (algoritmul 3.2) integrează toate aceste calcule, repetate de un anumit număr de ori sau până la atingerea unui criteriu de convergenţă. Aplicarea lui este însă posibilă doar după proiectarea MMA şi iniţializarea parametrilor. Proiectarea MMA presupune alegerea numărului de stări N, a stărilor iniţiale ($\pi_i \neq 0$) şi tranziţiilor permise ($a_{ij} \neq 0$), care împreună definesc topologia unui model, iar în cazul MMA discrete trebuie stabilită şi mulţimea $\mathcal{Y}$ a valorilor caracteristicilor procesului modelat.

Deoarece algoritmul Baum-Welch realizează numai o optimizare locală a parametrilor, iniţializarea lor poate fi esenţială pentru performanţele obţinute ulterior. Experimental, s-a constatat că probabilităţile iniţiale π_i şi de tranziţie a_{ij} pot fi iniţializate cu valori uniforme, arbitrar alese sau aleatoare fără a afecta performanţele, dar că poate fi necesară iniţializarea din datele de antrenament a funcţiilor de probabilitate $b_j(y)$.

Algoritmul 3.2 Algoritmul Baum-Welch pentru antrenarea MMA discrete

1: $Spi[N] \leftarrow 0$ {sumă ecuaţia (3.34)}
2: $Nra[N,N], Nua[N] \leftarrow 0$ {numărător/numitor ecuaţia (3.37)}
3: $Nrb[N,K], Nub[N] \leftarrow 0$ {numărător/numitor ecuaţia (3.38)}
4: **repeat**
5: **for** toate secvenţele $O_r, r = 1 \ldots R$ **do**
6: *calculează* $\alpha_{i,r}(t)$ conform ecuaţiilor (3.24) şi (3.25)
7: *calculează* $\beta_{i,r}(t)$ conform ecuaţiilor (3.28) şi (3.29)
8: *calculează* $P(O_r|M)$ conform ecuaţiei (3.26)
9: *calculează* $\gamma_{i,r}(t)$ conform ecuaţiei (3.32)
10: *calculează* $\xi_{ij,r}(t)$ conform ecuaţiei (3.36)
11: *actualizează* $Spi[i], Nra[i,j], Nua[i], Nrb[j,k], Nub[j]$
12: **end for**
13: *reestimează* $\pi_i, a_{ij}, b_j(y_k)$ conform ecuaţiilor (3.34), (3.37) şi (3.38)
14: **until** atingerea convergenţei sau a unui număr de iteraţii

3.6 Modelarea lingvistică

Procesul de generare a cuvintelor poate fi modelat în mod determinist prin gramatici [1], [123]. Datorită rigidităţii lor, vizibilă de exemplu în cazul limbajelor formale, acestea nu pot însă acoperi variabilitatea mesajelor decât în cazul unor aplicaţii foarte simple, pentru care ele pot fi definite iar vorbitorii pot fi antrenaţi să le respecte.

Alternativa este reprezentată de modelarea statistică [14], [118], [115], [152]: pentru aceasta, probabilitatea unui şir de cuvinte $P(W)$ poate fi descompusă într-un produs de probabilităţi condiţionate

$$P(W) = \prod_{i=1}^{L} P(w_i|w_{i-1} \ldots w_1) \tag{3.39}$$

în care fiecare factor poate fi estimat pe baza unui corpus de texte. Presupunând că recunoaşterea se face la nivelul propoziţiilor, la o lungime a propoziţiei de L cuvinte şi o mărime a vocabularului de V cuvinte, pentru utilizarea ecuaţiei (3.39) ar trebui estimate

$$N_p = \sum_{l=1}^{L} V \times V^{l-1} = \sum_{l=1}^{L} V^l = \frac{V^{L+1} - V}{V - 1} \tag{3.40}$$

probabilităţi, ceea ce cu excepţia unor aplicaţii foarte simple este imposibil din punct de vedere practic. De exemplu, pentru $L = 5$ şi $V = 1000$, $N_p \approx 10^{15}$, astfel că utilizând o reprezentare în virgulă flotantă simplă precizie, numai memoria necesară păstrării acestor probabilităţi ar fi de ordinul 10^6 Go, fără să mai considerăm şi spaţiul de disc necesar păstrării textelor din care sunt estimate. În plus, foarte multe dintre condiţiile $w_{i-1} \ldots w_1$ din ecuaţia (3.39) nu vor apare în aceste texte.

Ameliorarea unora dintre aceste probleme se face prin limitarea numărului de cuvinte din condiţiile $w_{i-1} \ldots w_1$ şi/sau gruparea lor în categorii: au rezultat astfel modelele lingvistice statistice de tip n-**gram** şi/sau utilizând **clase de cuvinte**. În cazul unui

model lingvistic statistic de tip n-gram, probabilitatea unui cuvânt este condiţionată doar de maximum $n-1$ cuvinte anterioare

$$P(W) = \prod_{i=1}^{L} P(w_i|w_{i-1}\dots w_{i-n+1}) \tag{3.41}$$

iar dacă sunt definite şi clase de cuvinte $C_k = \{w_{k,1},\dots w_{k,N_k}\}, k = 1\dots K$

$$P(W) = \prod_{i=1}^{L} P[w_i|C(w_i)] \cdot P[C(w_i)|C(w_{i-1})\dots C(w_{i-n+1})] \tag{3.42}$$

unde $C(w)$ este clasa din care face parte cuvântul w. Ambele ecuaţii pot fi interpretate ca rezultate ale modelării Markov a producerii cuvintelor: în cazul ecuaţiei (3.41), printr-un proces Markov cu stări definite de condiţiile $w_{i-1}\dots w_{i-n+1}$, iar în cel al ecuaţiei (3.42) printr-un MMA cu stările identificate de secvenţele de clase $C(w_{i-1})\dots C(w_{i-n+1})$.

Chiar şi în urma aplicării acestor metode, există încă posibilitatea ca unele evenimente (secvenţe de cuvinte sau clase) să nu apară în textele de antrenament. Pentru rezolvarea problemei au fost introduse diferite metode: netezirea modelelor prin reducerea (în engl. discounting) probabilităţilor unor evenimente apărute urmată de redistribuirea masei de probabilitate astfel eliberată celor care nu au apărut [167], [125], [256], [170]; combinarea mai multor modele prin interpolare [117] sau repliere (backing-off) [125] etc.

Obiectul acestei teze nefiind modelarea lingvistică, nu intrăm în detalii ale acestor probleme, ci ne rezumăm să precizăm că în practică modelele cele mai utilizate sunt cele de tip bigram sau trigram cu diferite variante de netezire şi combinare.

Perplexitatea

Pentru a putea compara performanţele sistemelor de recunoaştere a vorbirii trebuie precizate (şi) modelele lingvistice utilizate, iar aprecierea acestora impune cuantificarea complexităţii aplicaţiei din punctul lor de vedere. Prima măsură a acestei complexităţi a fost aşa-numitul factor de ramificare (branching factor), definit ca numărul maxim de cuvinte care pot apare după cuvântul curent. Acesta nu ţinea însă cont de variaţiile numărului de cuvinte care pot urma celui curent, nici de frecvenţele cuvintelor, astfel încât pentru a lua în calcul aceste aspecte a fost introdusă perplexitatea.

Deşi din punct de vedere teoretic poate fi pusă în legătură cu entropia sursei care generează cuvintele, în practică perplexitatea unui model lingvistic faţă de o aplicaţie se evaluează pe un corpus de texte de test, considerat reprezentativ pentru acea aplicaţie. Prin definiţie, pentru o secvenţă de N cuvinte perplexitatea este

$$PP = \frac{1}{\sqrt[N]{P(w_1\dots w_N)}} \tag{3.43}$$

şi poate fi interpretată, într-un mod compatibil cu factorul de ramificare anterior, ca media geometrică a numărului de cuvinte care urmează după cel curent, fiind deci de dorit să fie cât mai mică.

3.7 Modelarea acustică

Aşa cum am menţionat deja, principala problemă în recunoaşterea vorbirii este marea ei variabilitate: dacă asupra celei de la nivelul cuvintelor ne-am putut face o idee destul de exactă în secţiunea 3.6, cea de la nivelul acustic este mult mai greu de intuit şi cuantificat. Şi chiar dacă ea poate fi parţial redusă prin cuantizare vectorială [96], [151], [87], care înlocuieşte spaţiul multidimensional al vectorilor acustici cu o mulţime finită de vectori prototip, posibilităţile acestora de combinare sunt mult mai multe decât cele ale cuvintelor: de exemplu, la o frecvenţă de 100 vectori acustici/secundă şi 256 de vectori prototip, vor exista $256^{100} \approx 10^{80}$ combinaţii posibile într-o singură secundă. Din acest motiv, modelarea acustică este esenţială pentru orice sistem de recunoaştere a vorbirii.

Primele cercetări asupra recunoaşterii vorbirii folosind MMA au apelat frecvent la cuantizare vectorială şi MMA discrete pentru a reduce cerinţele computaţionale. Au fost obţinute astfel rezultate care au mers de la recunoaşterea independentă de vorbitor a unui număr mic de cuvinte pronunţate izolat [199] până la recunoaşterea dependentă de vorbitor a unui număr mare (5000) [114] sau foarte mare (20000) [10] de cuvinte izolate, şi au culminat la sfârşitul anilor '80 cu recunoaşterea independentă de vorbitor a vorbirii continue cu un vocabular de 1000 cuvinte [135].

Până în prima jumătate a anilor '80, modelarea acustică prin MMA continue [15], [113], care avea teoretic avantajul eliminării erorilor de cuantizare, a fost făcută folosind metode neoptimale de estimare a parametrilor densităţilor de probabilitate utilizate, aşa încât potenţialul avantaj nu se putea realiza datorită neoptimizării acestor parametri. Abia după stabilirea relaţiilor de reestimare prin algoritmul Baum-Welch a parametrilor unei largi clase de densităţi de probabilitate [143], [120], [121] a devenit posibilă utilizarea de o manieră optimă a MMA continue, iar în anii '90 ele s-au impus prin superioritatea demonstrată în mod constant în evaluările anuale DARPA [259].

Dintre densităţile parametrice de probabilitate, cele mai performante s-au dovedit aşa-numitele mixturi gaussiene, cunoscute şi ca modele cu mixturi gaussiene (în engl. Gaussian Mixture Models – GMM), de forma

$$b(\boldsymbol{y}) = \sum_{k=1}^{K} w_k \mathcal{N}_{\boldsymbol{\mu}_k, \boldsymbol{C}_k}(\boldsymbol{y}), \quad \sum_{k=1}^{K} w_k = 1 \tag{3.44}$$

cu $\mathcal{N}_{\boldsymbol{\mu}_k, \boldsymbol{C}_k}$ densităţi normale multivariate de medii $\boldsymbol{\mu}_k$ şi matrice de covarianţă $\boldsymbol{C}_k$

$$\mathcal{N}_{\boldsymbol{\mu}_k, \boldsymbol{C}_k}(\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{C}_k|}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_k)^T \boldsymbol{C}_k^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_k)} \tag{3.45}$$

unde $\boldsymbol{y}$ este un vector acustic de dimensiune D.

Utilizarea unor matrice de covarianţă complete duce la un timp de calcul $O(D^2)$ necesar pentru evaluarea probabilităţilor observaţiilor, cu valori tipice ale lui D în jur de 30. Pentru creşterea vitezei calculelor, cele mai multe sisteme care utilizează vectori acustici cepstrali se bazează pe corelaţia redusă a acestora (v. secţiunea 3.3) şi recurg la matrice de covarianţă diagonale, cu elemente nule în afara diagonalelor: $c_{ij} = 0, i \neq j$. În acest fel, evaluarea unei densităţi normale multivariate se reduce la cea a D densităţi

normale simple, iar timpul de calcul se reduce în mod corespunzător la $O(D)$. În plus, scade în aceeaşi proporţie şi numărul parametrilor MMA care trebuie estimaţi.

Succesul mixturilor gaussiene are explicaţii multiple. Pe de o parte, alofonele unui fonem sunt pronunţate folosind configuraţii apropiate ale tractului vocal: cum vectorii acustici rezultaţi depind de aceste configuraţii, intuitiv ne aşteptăm, şi studii experimentale o confirmă [263], ca pentru un acelaşi vorbitor sau vorbitori de acelaşi sex ei să fie apropiaţi în spaţiul acustic. Mai mult, ţinând cont de numărul mare de factori de care vectorii acustici depind, ne putem aştepta, şi acelaşi studii o confirmă, ca densităţile gaussiene să fie adecvate pentru modelarea distribuţiei lor.

Pe de altă parte, densităţile gaussiene au o mulţime de proprietăţi matematice care le fac instrumentul ideal pentru modelarea incertitudinii, oferind totodată posibilitatea de adaptare a lor la schimbări datorate mediului sau vorbitorilor [83], [84], [139].

Alternativele la mixturile gaussiene încercate de-a lungul timpului au inclus atât densităţi parametrice, de exemplu mixturi laplaciene [169], cât şi metode neparametrice de estimare a densităţilor de probabilitate: reţele neuronale [36], [205], [161], metoda celor mai apropiaţi k vecini [138] sau maşini cu vectori suport (în engl. Support Vector Machines – SVM) [78]. Deşi iniţial promiţătoare, metodele neparametrice fie s-au dovedit dificil de scalat la vocabulare foarte mari sau de adaptat la schimbările de mediu sau ale vorbitorilor (cazul reţelelor neuronale), fie au încă de trecut aceste teste, aşa că pentru moment mixturile gaussiene rămân cea mai bună opţiune în modelarea acustică.

3.7.1 Antrenarea MMA cu mixturi gaussiene

Înlocuirea distribuţiilor de probabilitate din MMA discrete cu densităţi parametrice în MMA continue nu afectează probabilităţile iniţiale şi de tranziţie, aşa încât formulele de reestimare a lor pentru MMA discrete (secţiunea 3.5.1) rămân valabile şi în acest caz. Modificările constau în introducerea unor noi formule pentru reestimarea parametrilor mixturilor: ponderile densităţilor w_k, vectorii medii $\boldsymbol{\mu}_k$ şi matricele de covarianţă $\boldsymbol{C}_k$.

Densităţile gaussiene multivariate componente ale mixturilor reprezintă un nou nivel "ascuns" al modelelor, sub cel al stărilor, aşa încât pentru estimarea lor probabilitatea $\gamma_j(t)$ de ocupare a stării s_j la momentul t trebuie divizată în probabilităţile ca densităţile componente să fi emis observaţia o_t în timp ce modelul se afla în această stare

$$\gamma_{jk}(t) = \gamma_j(t) \frac{w_{jk} \mathcal{N}_{\boldsymbol{\mu}_{jk}, \boldsymbol{C}_{jk}}(o_t)}{\sum_{k=1}^{K} w_{jk} \mathcal{N}_{\boldsymbol{\mu}_{jk}, \boldsymbol{C}_{jk}}(o_t)}, \quad k = 1 \ldots K \tag{3.46}$$

Utilizând R secvenţe de observaţii $\boldsymbol{O}_r = \boldsymbol{o}_{r,1}\boldsymbol{o}_{r,2}\ldots\boldsymbol{o}_{r,T_r}$, $r = 1 \ldots R$, pentru a antrena modelul M, formulele de reestimare a valorilor parametrilor mixturilor sunt

$$\hat{\boldsymbol{\mu}}_{jk} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{jk,r}(t) \cdot \boldsymbol{o}_{r,t}}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{jk,r}(t)}, \quad k = 1 \ldots K \tag{3.47}$$

$$\hat{\boldsymbol{C}}_{jk} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{jk,r}(t) \cdot (\boldsymbol{o}_{r,t} - \boldsymbol{\mu}_{jk})(\boldsymbol{o}_{r,t} - \boldsymbol{\mu}_{jk})^T}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{jk,r}(t)}, \quad k = 1 \ldots K \tag{3.48}$$

$$\hat{w}_{jk} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{jk,r}(t)}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{j,r}(t)} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{jk,r}(t)}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \sum_{k=1}^{K} \gamma_{jk,r}(t)}, \quad k = 1 \ldots K \tag{3.49}$$

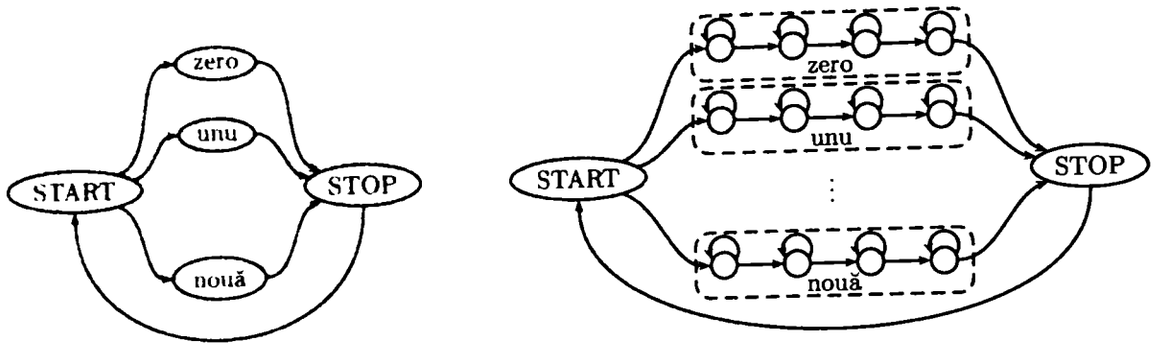

Figura 3.3: Exemplu de integrare a modelului lingvistic cu cele acustice

Modificările care trebuie aduse algoritmului 3.2 pentru antrenarea MMA cu mixturi gaussiene includ: introducerea unor noi variabile pentru numărătorul respectiv numitorul fiecăreia din aceste formule; calculul probabilităţilor densităţilor componente ale fiecărei mixturi, $\gamma_{jk,r}(t)$; reestimarea parametrilor mixturilor conform formulelor (3.47–3.49).

3.8 Reprezentarea integrată a cunoştinţelor

Căutarea şirului de cuvinte prin pronunţarea căruia este cel mai probabil să fi rezultat un şir de observaţii acustice necesită integrarea cunoştinţelor incorporate de modelele lingvistice şi acustice într-un spaţiu de căutare unic, în care se poate evalua produsul $P(W) \cdot P(A|W)$ din ecuaţia (3.22). Modelele Markov ascunse oferă un cadru pentru realizarea acestei integrări, permiţând construcţia ierarhică a unui MMA global care modelează atât generarea cuvintelor cât şi producerea vorbirii. Pentru exemplificare considerăm o aplicaţie simplă constând în recunoaşterea cifrelor zecimale (figura 3.3).

Pentru recunoaşterea cifrelor zecimale, modelul lingvistic poate fi o gramatică simplă de tipul buclă de cuvinte, reprezentabilă prin diagrama de sintaxă din stânga figurii 3.3, în care o cifră poate fi urmată de oricare alta. Presupunând că modelarea acustică este realizată prin MMA ale cuvintelor, integrarea acestora cu modelul lingvistic se poate face prin inserare în nodurile corespunzătoare ale diagramei de sintaxă, rezultatul fiind MMA extins din dreapta figurii 3.3, denumit frecvent reţea de recunoaştere sau reţea integrată, care include acum atât modelul lingvistic cât şi modelele acustice.

Dacă gramatica nu specifică probabilităţile cuvintelor, tranziţiile din starea START sunt implicit considerate echiprobabile, astfel încât probabilităţile lor pot fi omise din calcule, nerealizând nici o diferenţiere a cuvintelor. Dar aceeaşi buclă de cuvinte poate reprezenta un model lingvistic de tip unigram, caz în care probabilităţile cuvintelor trebuie atribuite tranziţiilor corespunzătoare din starea START sau în starea STOP.

Pentru vocabulare mari şi foarte mari, modelarea acustică utilizează MMA ale unor unităţi sublexicale (silabe, semisilabe, foneme etc.), iar integrarea modelului lingvistic cu cele acustice presupune utilizarea unui model lexical, uzual un simplu dicţionar, pentru obţinerea pronunţiilor cuvintelor, eventual cu variante, în termenii unităţilor sublexicale. Pronunţiile la rândul lor sunt reprezentabile prin modele Markov ale căror stări pot fi

înlocuite cu modelele acustice ale unităţilor sublexicale componente.

Simpla înlocuire a pronunţiilor din dicţionar în nodurile modelului lingvistic duce la o reprezentare costisitoare din punct de vedere computaţional, în special în cazul vocabularelor mari şi foarte mari, astfel încât pentru creşterea vitezei şi reducerea necesarului de memorie dicţionarul poate fi reprezentat printr-o structură arborescentă [12], [169], eventual comprimată [131]. O asemenea structură arborescentă poate fi folosită şi pentru factorizarea probabilităţilor cuvintelor din modelul lingvistic statistic la nivelul unităţilor sublexicale [231], [5], [178]. În cazul unui model lingvistic de tip n-gram, acesta poate fi în întregime integrat cu dicţionarul şi transformat într-o mulţime de subarbori [70].

Dimensiunea unei reţele de recunoaştere este dependentă în primul rând de modelul lingvistic, motiv pentru care integrarea cunoştinţelor în reţele statice este fezabilă doar pentru modele lingvistice simple (unigram, bigram) şi/sau cu vocabulare de dimensiuni moderate. Pentru vocabulare mari şi foarte mari şi modele mai complexe, de tip trigram sau de mai mare întindere, integrarea se face în mod dinamic [171], [241], [264], prin construcţia efectivă a spaţiului de căutare doar în zona celor mai plauzibile ipoteze.

3.9 Algoritmi de căutare

Odată cu integrarea cunoştinţelor acumulate de modelele lingvistice şi acustice într-o reţea de recunoaştere, găsirea şirului de cuvinte cel mai probabil să fi fost pronunţat, date fiind observaţiile acustice, se poate face prin căutarea în reţea a căii pentru care produsul $P(W) \cdot P(A|W)$ este maxim. În acest produs, probabilităţile lingvistice pot fi aplicate la tranziţiile între cuvinte, sau pe durata cuvintelor, dacă dicţionarul a fost reprezentat sub formă de arbore şi probabilităţile lingvistice au fost factorizate [231]. Datorită valorilor foarte mici ale probabilităţilor, înmulţirile lor repetate pot ajunge rapid să cauzeze depăşiri inferioare (underflow), astfel încât pentru evitarea acestei probleme se utilizează de obicei un sistem logaritmic de reprezentare [39], [41].

Una din slăbiciunile MMA, datorată proprietăţii Markov, este imposibilitatea lor de a ţine cont de corelaţia observaţiilor succesive, ceea ce face ca probabilităţile acustice să fie subevaluate. Pentru corectare, probabilităţile lingvistice sunt de obicei ridicate la o putere supraunitară $\lambda > 1$, determinată experimental şi cunoscută sub diferite nume, pe care în continuare o vom denumi pondere lingvistică (language weight) [136].

Altă problemă a sistemelor de recunoaştere automată a vorbirii este cea a echilibrului între omisiuni şi inserţii, pentru asigurarea căruia se utilizează o penalizare de tranziţie τ, deasemeni determinată experimental, aplicată la tranziţiile între cuvinte pentru a reduce frecvenţa inserţiilor. Folosind o reprezentare logaritmică a probabilităţilor şi cele două variabile de control λ şi τ, ecuaţia (3.22) trebuie rescrisă în forma

$$\hat{W} = \arg\max_{W}[|W|\tau + \lambda \log P(W) + \log P(A|W)] \tag{3.50}$$

unde $|W|$ este lungimea şirului de cuvinte W, astfel încât în loc de probabilitatea unui şir de cuvinte este mai corect să vorbim despre scorul lui.

Algoritmii fundamentali de căutare a căii corespunzătoare celui mai probabil şir de cuvinte sunt algoritmul Viterbi [247] şi algoritmul A^* [210], mai cunoscut în domeniul recunoaşterii automate a vorbirii ca algoritmul de decodare cu stivă (stack decoding)

[112], [115]. Algoritmul Viterbi este unul de tip sincron, care evaluează toate căile posibile până la un acelaşi moment în baza unei startegii de căutare în lăţime (breadth first), iar scorurile căilor pot fi comparate deoarece toate corespund aceleiaşi porţiuni a semnalului; ca urmare, algoritmul Viterbi este admisibil, adică garantează găsirea căii cu scorul maxim - care poate să coincidă sau nu cu şirul de cuvinte corect.

Pentru şirul de observaţii $O = o_1 o_2 \dots o_T$, algoritmul Viterbi găseşte calea (secvenţa de stări) $Q = q_1 q_2 \dots q_T$ optimă prin MMA integrat M (reţeaua de recunoaştere) prin calculul pentru fiecare moment t şi fiecare stare s_i a probabilităţii maxime ca modelul să fi generat observaţiile $o_1 \dots o_t$ şi să se afle în starea s_i la momentul t

$$\delta_i(t) = \max_{q_1 \dots q_{t-1}} P[o_1 \dots o_t, q_1 \dots q_t, q_t = s_i | M] \tag{3.51}$$

simultan cu memorarea stării anterioare care a asigurat maximizarea, $\psi_i(t)$. Ca şi până acum, calculele se pot realiza recursiv

$$\delta_i(1) = \pi_i b_i(o_1), \quad i = 1 \dots N \tag{3.52}$$

$$\psi_i(1) = 0, \quad i = 1 \dots N \tag{3.53}$$

$$\delta_j(t) = \max_i [\delta_i(t-1) a_{ij}] b_j(o_t), \quad j = 1 \dots N, \quad t = 2 \dots T \tag{3.54}$$

$$\psi_j(t) = \arg\max_i [\delta_i(t-1) a_{ij}], \quad j = 1 \dots N, \quad t = 2 \dots T \tag{3.55}$$

iar calea optimă este obţinută în final prin parcurgerea în sens invers a matricei $\psi_i(t)$:

$$q_T^* = \arg \max_{s_i \text{ finală}} \delta_i(T), \quad q_t^* = \psi_{q_{t+1}^*}(t+1), \quad t = T-1 \dots 1 \tag{3.56}$$

Pe lângă decodarea lingvistică, algoritmul Viterbi permite şi obţinerea unei aproximări a probabilităţii $P(O|M)$, cunoscută ca aproximarea Viterbi, prin cea a căii optime

$$P^*(O|M) = \max_{s_i \text{ finală}} \delta_i(T) \tag{3.57}$$

iar pe baza corespondenţei dintre stările $q_1^* q_2^* \dots q_T^*$ şi observaţiile $o_1 o_2 \dots o_T$, parametrii modelelor acustice pot fi reestimaţi prin aşa-numita metodă Viterbi de antrenare a MMA, care utilizează doar evenimentele asociate cu parcurgerea căii de probabilitate maximă.

Deşi admisibil, algoritmul Viterbi se poate dovedi prohibitiv din punct de vedere al timpului de calcul necesar căutării soluţiei optime. Pentru creşterea vitezei de obţinere a unei soluţii, el a fost modificat prin reducerea (pruning) spaţiului de căutare doar la un fascicol (beam) de căi printre care este cel mai probabil să fie localizată şi cea optimă [146], [171], ceea ce a dus la un algoritm de căutare neadmisibil. Determinarea fascicolului considerat la un moment dat se poate face pe baza unui prag de reducere (pruning threshold) faţă de starea cea mai probabilă la acel moment, a unui număr maxim permis de stări componente [231] sau a unor combinaţii ale acestor criterii.

Algoritmul A^* sau de decodare cu stivă este un algoritm asincron, de tip *best first*, cu o parcurgere arborescentă a spaţiului de căutare, diferită de cea de tip grilaj (trellis) din cazul algoritmului Viterbi, ceea ce conduce la căi parţiale de lungimi diferite, între ale căror scoruri comparaţiile directe nu au sens. Comparaţiile se fac doar la nivelul unor

căi complete, $Q = q_1 q_2 \dots q_T$, pe baza unor estimări ale scorurilor lor date de o funcţie de evaluare

$$f(Q) = g(Q_{1,t}) + h(Q_{t+1,T}) \tag{3.58}$$

unde $g(Q_{1,t})$ este scorul exact al căii parţiale $Q_{1,t} = q_1 \dots q_t$, iar $h(Q_{t+1,T})$ este o funcţie euristică realizând o estimare a scorului restului căii $Q_{t+1,T} = q_{t+1} \dots q_T$.

Caracterul euristic al funcţiei h face ca algoritmul A^* să nu fie întotdeauna admisibil. Împreună cu dificultăţile estimării propriu-zise, acest lucru a făcut ca algoritmul A^* să fie mai puţin utilizat în sistemele de recunoaştere a vorbirii. Oferind o decuplare ideală a modelului lingvistic de cele acustice [8], el are însă o serie de avantaje în cazul utilizării unor modele lingvistice de tip n-gram de mai mare întindere ($n > 3$), astfel încât pe viitor ne putem aştepta la o creştere a frecvenţei utilizării lui.

În forma de bază, atât algoritmul Viterbi cât şi A^* asigură găsirea unei căi optime unice, dar în practică, datorită erorilor de modelare sau căutare, aceasta poate fi diferită de cea corectă, corespunzătoare şirului de cuvinte pronunţat. Unele dintre erorile de acest tip pot fi corectate de aplicaţia în care este integrat sistemul de recunoaştere dacă acesta generează nu doar şirul de cuvinte corespunzător căii optime, ci o structură de date mai complexă. Au apărut astfel algoritmi de tip N-Best [217], [216], generând o listă a celor mai probabile N propoziţii, şi algoritmi în mai mulţi paşi [9], [226], eventual cu generarea unor structuri de date mai complicate, de tip latice sau graf de cuvinte [4], [103]. Aceste structuri de date asigură reprezentarea atât a unor spaţii de căutare reduse pentru paşi succesivi, cât şi a rezultatelor finale.

3.10 Concluzii

Pentru recunoaşterea automată a vorbirii au fost încercate de-a lungul timpului diferite metode, iar acest capitol le-a prezentat pe cele care, în urma performanţelor demonstrate prin evaluarea sistemelor în care au fost implementate, se numără printre cele mai frecvent utilizate în sistemele contemporane de recunoaştere a vorbirii şi au fost utilizate şi în cercetările proprii. Pentru a fi clar modul în care se poate face evaluarea sistemelor, primele au fost descrise chiar metodele şi metricele utilizate în acest scop.

Utilizarea parametrilor semnalului vocal pentru recunoaşterea automată a vorbirii presupune gruparea lor în vectori dintr-un spaţiu acustic. Recunoaşterea ca atare poate fi precedată de o serie de operaţiuni în acest spaţiu: calculul unor distanţe acustice semnificative din punct de vedere perceptual; transformări ale spaţiului acustic în scopul reducerii anizotropiei lui; cuantizarea vectorială care, utilizând o mulţime de vectori prototip, permite comprimarea şi reprezentarea discretă a vectorilor acustici.

Dintre diferitele metode încercate pentru recunoaşterea propriu-zisă, doar două au rezistat până în prezent: deformarea dinamică a timpului şi modelele Markov ascunse. Deformarea dinamică a timpului este încă utilizată pentru recunoaşterea dependentă de vorbitor a unor vocabulare mici de cuvinte rostite izolat, având avantajul generării simple a tiparelor cu care sunt comparate ulterior pronunţiile de recunoscut. Această simplitate nu permite însă reprezentarea cu suficientă precizie a marii variabilităţi a semnalelor vocale, care poate fi acoperită doar apelând la metode statistice.

Recunoaşterea automată a vorbirii prin metode statistice presupune modelarea ei lingvistică şi acustică şi integrarea modelelor rezultate într-un spaţiu unic al soluţiilor. Atingerea acestor obiective poate fi realizată prin utilizarea modelelor Markov ascunse, care constituie fundamentul sistemelor moderne de recunoaştere a vorbirii, iar acest capitol a inclus o vedere de ansamblu asupra lor, cu detalierea aspectelor esenţiale, alte amănunte legate de utilizarea lor urmând a fi prezentate în restul tezei.

Găsirea şirului de cuvinte cel mai probabil să fi fost pronunţat dat fiind semnalul de recunoscut necesită construcţia unei reprezentări integrate a modelelor lingvistice şi acustice şi evaluarea posibilelor soluţii prin utilizarea unor algoritmi de căutare în spaţiul soluţiilor. Modelele Markov ascunse permit rezolvarea elegantă a ambelor probleme prin construcţia ierarhică a unor modele înglobându-le pe cele lingvistice şi acustice, cunoscute sub numele de reţele de recunoaştere sau reţele integrate, în care şirul cel mai probabil de cuvinte este găsit cel mai adesea folosind algoritmul Viterbi.

CAPITOLUL 4

Baza de date fonetice

Utilizarea metodelor statistice de recunoaştere automată a vorbirii, prezentate în capitolul 3, impune existenţa unor date pe baza cărora să poată fi construite modelele implicate de ecuaţia (3.22): semnale vocale pentru modelele acustice, respectiv texte pentru modelele lingvistice. În cursul primelor cercetări au fost utilizate semnale vocale colectate ad-hoc, uzual specifice unor aplicaţii [15], [146], şi arhive de texte private [113].

Dezvoltarea cercetărilor asupra recunoaşterii automate a vorbirii, utilizând multiple abordări ale problemelor ei, a făcut necesară evaluarea nu doar a performanţelor în sine ale sistemelor de recunoaştere, ci şi a semnificaţiei statistice [37], [158] a diferenţelor dintre aceste performanţe. Drept urmare, în anii '80 a început colectarea şi publicarea unor baze de date vocale de uz general [45] sau specifice unor aplicaţii. Acestea s-au dovedit esenţiale pentru dezvoltarea şi testarea unor sisteme de recunoaştere, precum şi pentru evaluarea şi compararea acestor sisteme utilizând seturi de date standard.

Probabil cele mai cunoscute exemple de baze de date specifice unor aplicaţii sunt TIDIGITS [140], colectată la Texas Instruments (TI) pentru studiul recunoaşterii şirurilor de cifre, şi Resource Management (RM) [192], dezvoltată în cadrul programului DARPA de recunoaştere automată a vorbirii şi având în vedere o aplicaţie de conducere a resurselor militare navale (nave de război) ale Statelor Unite. Disponibilitatea acesteia din urmă a fost esenţială în demonstrarea fezabilităţii recunoaşterii independente de vorbitor a vorbirii continue cu vocabulare mari [135] şi, în ultimă instanţă, impunerea metodelor statistice de recunoaştere automată a vorbirii în raport cu alte abordări.

Succesul metodelor statistice, bazate pe estimarea automată din date de antrenament a parametrilor sistemelor de recunoaştere, a pus în evidenţă importanţa fundamentală a datelor şi a condus la construirea unor noi baze de date vocale, de dimensiuni din ce în ce mai mari, vizând aplicaţii diverse şi încercând să acopere porţiuni cât mai extinse din variabilitatea semnalului vocal. Cele mai importante, prin contribuţia la atingerea nivelului actual, rămân cele din cadrul programului amintit al DARPA, program care

Cercetări realizate cu sprijinul Comisiei Europene prin contractul COPERNICUS 1304/1994 şi al fostului Consiliu Naţional al Cercetării Ştiinţifice Universitare - CNCSU (devenit din 1999 Consiliul Naţional al Cercetării Ştiinţifice din Învăţământul Superior - CNCSIS) prin grantul 56/1995.

de-a lungul timpului a inclus sisteme de informaţii despre traficul aerian [99], dictare automată [183], transcrierea convorbirilor telefonice [94] şi a emisiunilor radio-TV [95].

O categorie aparte a bazelor de date vocale, destul de restrânsă datorită dificultăţilor implicate de construcţia lor, este constituită de aşa-numitele baze de date fonetice, prima şi cea mai cunoscută şi utilizată dintre ele fiind TIMIT [81], construită în cooperare de Texas Instruments (TI) [71] şi Massachusetts Institute of Technology (MIT) [132], cu unele contribuţii de la Stanford Research Institute [51]. Elementele definitorii ale unei asemenea baze de date sunt conţinutul controlat prin proiectarea corespunzătoare a materialelor înregistrate şi selectarea vorbitorilor, o calitate deosebită a înregistrărilor, precum şi adnotarea conţinutului cu informaţii fonetice şi fonologice. Datorită acestor caracteristici, o bază de date fonetice constituie o resursă esenţială nu numai pentru cercetările în direcţia recunoaşterii automate a vorbirii, ci şi pentru cele din alte domenii ale prelucrării automate a vorbirii, precum şi o sursă de cunoştinţe fundamentale, de fonetică şi fonologie a limbii în care au fost pronunţate materialele înregistrate.

Date fiind inexistenţa unei baze de date corespunzătoare pentru cercetările asupra recunoaşterii automate independentă de vorbitor a vorbirii continue în limba română şi insuficienţa constatată a cunoştinţelor de fonetică acustică şi fonologie a limbii române (insuficienţă confirmată ulterior chiar şi de lingvişti în literatura de specialitate [234]), construcţia unei asemenea baze de date s-a impus ca o primă etapă a cercetărilor, iar în acest capitol vor fi prezentate detalii legate de proiectarea şi colectarea ei [32], [34].

4.1 Consideraţii de proiectare

Disponibilitatea unor baze de date vocale corespunzătoare reprezintă o precondiţie pentru multe cercetări fundamentale sau aplicative din diverse domenii ale ştiinţei şi tehnologiei vorbirii, iar amploarea pe care aceste cercetări o pot lua face imposibilă din punct de vedere practic colectarea, pentru o anumită limbă, a unei baze de date care să le satisfacă simultan cerinţele. Uneori, aceste cerinţe pot fi satisfăcute prin utilizarea unor baze de date deja existente sau a unor subseturi convenabil alese ale acestora.

Există însă şi numeroase situaţii în care se impune colectarea unor noi baze de date, una dintre acestea fiind şi extinderea la noi limbi sau dialecte a cercetărilor asupra recunoaşterii automate a vorbirii: deşi au fost încercate diferite metode de utilizare a datelor dintr-o limbă pentru recunoaşterea pronunţiilor unei alte limbi [253], [162], unele vizând limba română [155], recunoaşterea multi- şi croslinguală [215] sau independentă de limbă [44], performanţele obţinute se îmbunătăţesc odată cu creşterea cantităţii datelor din noua limbă folosite pentru construirea sau adaptarea modelelor acustice.

Proiectarea şi colectarea unei noi baze de date devine cu atât mai necesară atunci când cunoştinţele fundamentale de fonetica şi fonologia noii limbi, necesare şi pentru recunoaşterea automată, sunt insuficiente: din păcate, aceasta este şi situaţia limbii române, în cazul căreia aceste lipsuri se manifestă în literatura de specialitate fie prin lipsa abordării unor subiecte, fie prin tratarea lor contradictorie sau chiar eronată.

Deciziile de proiectare a bazei de date descrise aici au fost influenţate, în consecinţă, de necesitatea de a facilita:

- modelarea acustică a semnalului vocal, în vederea atingerii obiectivului principal

stabilit al acestor cercetări (secţiunea 1.3) – recunoaşterea vorbirii continue în limba română, independentă de vorbitor, cu vocabulare în jurul a 1000 de cuvinte;

- validarea sau obţinerea unor noi cunoştinţe de fonetică şi fonologie a limbii române, a căror insuficienţă a fost resimţită încă din această fază a cercetărilor şi confirmată în parte de rezultate prezentate în capitolele următoare.

Din marea varietate de tipuri de pronunţare, condiţii de mediu etc. posibile, această bază de date a fost limitată la înregistrări de calitate, într-un mediu afectat cât mai puţin de zgomote şi reverberaţii, ale unor pronunţii rezultate cu preponderenţă din citirea în varianta standard (literară) a limbii române a unor texte pregătite în mod special pentru a asigura un conţinut al înregistrărilor corespunzător scopurilor propuse.

Dat fiind obiectivul principal al cercetărilor, proiectarea bazei de date, descrisă în următoarele trei secţiuni, a cuprins: alegerea unor unităţi de modelare acustică adecvate; pregătirea materialelor de înregistrat; specificarea caracteristicilor vorbitorilor şi alocarea materialelor pe care le vor înregistra.

4.2 Alegerea unităţilor de modelare

Recunoaşterea automată a vorbirii se poate realiza, după cum s-a menţionat, folosind modele acustice ale cuvintelor sau ale unor unităţi sublexicale – silabe, semisilabe, sunete etc. Alegerea unităţilor de modelare are o influenţă decisivă asupra acurateţii cu care modelele acustice reprezintă variabilitatea semnalului vocal, în special cea cauzată de coarticulaţie, constând în modificarea caracteristicilor sunetelor vorbirii funcţie de cele adiacente şi datorată mişcărilor anticipatorii şi inerţiale ale articulatorilor. Exemple de manifestări ale coarticulaţiei pot fi urmărite în figura 5.1: coarticulaţia este cea mai vizibilă în cazul sunetului [l], cele trei apariţii ale lui având fiecare caracteristici spectrale diferite, dar poate fi observată şi în cazul altor sunete – de exemplu [j] sau [a].

Utilizând cuvinte ca unităţi de modelare acustică, o bună parte din variabilitatea datorată coarticulaţiei – cea corespunzătoare interacţiunii dintre sunete în interiorul cuvintelor – va fi inclusă în modelele rezultate, dar va rămâne totuşi neacoperită cea cauzată de interacţiunile dintre cuvinte, localizată la extremităţile lor. Aceasta ar putea fi la rândul ei reprezentată prin modele dependente de context ale cuvintelor, dar odată cu creşterea mărimii vocabularului va creşte şi numărul modelelor. În plus, va creşte proporţional şi cantitatea de date necesare pentru antrenarea lor.

Un alt dezavantaj, poate cel mai important, al cuvintelor ca unităţi de modelare acustică, este acela al lipsei lor de generalitate: extinderea vocabularului unui sistem de recunoaştere bazat pe modele ale cuvintelor necesită date suplimentare, constând din înregistrări ale cuvintelor nou introduse în vocabular, pentru antrenarea unor noi modele acustice. Din aceste motive, modele acustice ale cuvintelor sunt utilizate doar în sisteme pentru aplicaţii simple, cu vocabulare închise de zeci sau maximum sute de cuvinte.

Dintre unităţile de modelare sublexicale, silabele şi unităţile derivate sunt adesea considerate unităţi naturale din punct de vedere al coarticulaţiei şi al posibilităţilor de a reprezenta variabilitatea asociată ei. În practică, ele se dovedesc însă dificil de utilizat datorită problemelor care pot apare la extinderea vocabularului, precum şi a celor de

Tabelul 4.1: Unităţi fonetice de modelare acustică – simboluri şi exemple

Simboluri		Exemple	Simboluri		Exemple	Simboluri		Exemple
IPA	ASCII		IPA	ASCII		IPA	ASCII	
i	i	v*i*n	o̯	O	c*o*ate	k	k	*c*ap
i̯	I	am*i*	p	p	*p*ăr	g	g	*g*ât
e	e	*e*l	b	b	*b*an	ts	T	*ţ*ap
ɨ	y	*î*n	t	t	*t*ip	tʃ	C	*c*e
ə	@	c*ă*	d	d	*d*ar	dʒ	G	*g*er
ɑ	a	*a*n	f	f	*f*oc	h	h	*h*an
u	u	*u*d	v	v	*v*in	m	m	*m*ic
o	o	*o*m	s	s	*s*top	n	n	*n*as
j	j	*i*ar	z	z	*z*i	l	l	*l*ac
e̯	E	d*e*al	ʃ	S	*ş*i	r	r	*r*âu
w	w	n*o*u	ʒ	J	*j*ok	#	-	*pauză*

acoperire a pronunţiilor de recunoscut, care fac necesară combinarea lor cu modele de tip fonemic. În plus, unele experimente [35] au arătat că performanţele astfel obţinute sunt inferioare celor ale sistemelor bazate pe modele dependente de context de tip fonemic. Ulterior, studii teoretice [97] au arătat că silabele ar putea fi cea mai bună unitate de modelare a variabilităţii din vorbirea spontană, iar experimente de recunoaştere [79] au confirmat unele avantaje ale silabelor ca unităţi de modelare acustică. Dar chiar când sistemele evaluate au utilizat modele ale silabelor, ele au inclus şi modele dependente de context de tip fonemic pentru a putea acoperi toate pronunţiile de recunoscut.

Pentru a obţine modele acustice cât mai generale folosind o cantitate minimă de date pentru antrenarea lor, recunoaşterea automată a vorbirii continue cu vocabulare mari şi foarte mari este în general bazată pe unităţi de modelare de tip fonetic sau, prin legarea parametrilor la nivelul stărilor, chiar subfonetic [108]. În acest caz, variabilitatea datorată coarticulaţiei este acoperită prin modelarea dependentă de context, o aceeaşi unitate putând fi reprezentată prin mai multe modele, diferenţiate funcţie de unităţile care o preced şi/sau urmează. Se obţin astfel modele cu un grad mare de generalitate, uşor de reutilizat în cazul modificării sau extinderii vocabularului şi care pot valorifica cu eficienţă maximă cantitatea de date disponibile pentru antrenarea lor.

Ţinând cont de aspectele menţionate, proiectarea bazei de date s-a făcut având în vedere modelarea acustică prin unităţi sublexicale de tip fonetic. Din păcate, cunoştinţele disponibile de fonetică acustică şi fonologie a limbii române, pe baza cărora ar fi trebuit definit un set de asemenea unităţi, s-au dovedit insuficiente, existând diverse poziţii ale lingviştilor în ceea ce priveşte fonemele limbii române [193], [244], uneori contradictorii [190] sau chiar greşite [206]. Analizând seturile de foneme identificate sau acceptate de diverşi autori şi luând în considerare diferenţele dintre ele, am definit un set de unităţi care, direct sau prin combinaţii, acoperă aceste variante (tabelul 4.1).

Tabelul 4.1 cuprinde atât simboluri din alfabetul fonetic internaţional (International Phonetic Alphabet - IPA), pentru a facilita comparaţiile cu literatura lingvistică, cât şi

simboluri ASCII, pentru adnotarea fonetică pe calculator a semnalelor vocale, alese cât mai aproape de simbolurile IPA sau cât mai sugestive şi utilizate şi în continuare.

Trei unităţi sunt implicate într-o bună parte din diferenţele dintre seturile de foneme ale limbii române identificate şi/sau acceptate în literatura lingvistică:

- /I/ apare doar în poziţie postconsonantică finală, în cuvinte ca ani, azi, mari etc.: unele sisteme fonologice ale limbii române standard îl consideră un alofon – o variantă poziţională, scurtă, asilabică şi (uneori) devocalizată, a lui /i/, /j/ sau /e/; altele neagă chiar existenţa separată a realizărilor lui fizice, reducându-l la un rol diacritic, de marcaj al palatalizării consoanei precedente – în acest caz, sistemul fonologic include cu statut de foneme o serie de consoane palatalizate;

- /E/ şi /O/ sunt uneori interpretate ca alofone fie ale vocalelor /e/ respectiv /o/, fie ale semivocalelor/semiconsoanelor /j/ respectiv /w/; în mod corespunzător, va diferi mulţimea acceptată a diftongilor şi triftongilor – unii dintre ei subiect, la rândul lor, al unor discuţii asupra naturii lor mono- sau multifonemice.

Alte diferenţe între sisteme fonologice (de ex. statutul de foneme sau alofone pentru consoanele [k] şi [g] palatalizate) au fost considerate neglijabile din punctul de vedere al cercetărilor noastre, putând fi acoperite prin modelare dependentă de context.

Unităţile din tablelul 4.1 vor fi denumite în continuare foneme (sg. fonem), notate /x/, când va fi vorba de categorii abstracte de sunete distinctive ale limbii române, respectiv sunete, notate [x], pentru a indica realizări fizice ale acestor categorii.

4.3 Materialele de înregistrat

Scopul principal al cercetărilor fiind recunoaşterea automată a vorbirii continue în limba română, înregistrările din baza de date trebuie să conţină un număr suficient de mare de apariţii ale unor unităţi de modelare acustică pentru a permite estimarea cu acurateţe a modelelor lor. În plus, pentru a asigura şi independenţa de vorbitor a modelelor, aceste apariţii trebuie să provină din pronunţii ale cât mai multor vorbitori.

Pe lângă modelarea acustică, la construcţia unei baze de date pentru cercetări asupra recunoaşterii automate a vorbirii trebuie avută în vedere şi modelarea lingvistică. Într-o fază mai avansată a cercetărilor, aceasta se poate realiza prin extragerea textelor ce vor fi citite pentru înregistrare dintr-un corpus de texte de mai mari dimensiuni, utilizabil şi pentru construcţia de modele lingvistice de tip n-gram. Adesea, un asemenea corpus este obţinut, datorită caracteristicilor textelor conţinute, din arhive în format electronic ale unor ziare. Această metodă a fost folosită, de exemplu, pentru construcţia bazei de date WSJ, plecând de la arhive ale *Wall Street Journal*, în cadrul programului DARPA [183], sau a bazei de date franceze BREF [86], utilizând texte din *Le Monde*.

Dat fiind stadiul incipient al cercetărilor noastre, dificultatea obţinerii şi prelucrării unui corpus de tipul menţionat, precum şi unele aspecte (pre-normalizare şi standarde) vizate de programul COPERNICUS al Comisiei Europene, în cadrul căruia a început construcţia ei [203], [204], pentru această bază de date s-a ales varianta compatibilităţii cu baza de date EUROM [46], dezvoltată anterior în proiectele ESPRIT SAM şi SAM-A.

Compatibilitatea se referă atât la materialele conţinute şi vorbitorii înregistraţi, cât şi la utilizarea aceloraşi organizări şi formate ale fişierelor constituente [90].

EUROM a fost proiectată cu un conţinut comparabil pentru cele 11 limbi oficiale (la acea dată) din ţările Uniunii Europene:

- 40 de **pasaje** de câte 5 propoziţii legate tematic, cu teme comune în toate limbile, utile pentru antrenarea, testarea şi evaluarea sistemelor de recunoaştere;
- un număr de **propoziţii** de completare, asociate pasajelor, compuse în mod special pentru a compensa variaţiile frecvenţelor fonemelor în pasaje;
- 100 de **numere** întregi între 0 şi 9999, selectate pentru a acoperi principalele lor constrângeri fonotactice, utile pentru testarea şi evaluarea unor sisteme;
- **logatomi** de forma CVC (consoană-vocală-consoană), izolaţi şi în cinci contexte de câte două cuvinte, pentru diagnoza sistemelor [229], [230].

Pentru fiecare limbă, (circa) 60 de vorbitori, (cât mai) egal distribuiţi pe sexe, au înregistrat 3-5 pasaje, 0-5 propoziţii de completare şi numerele, iar câţiva – logatomi.

Compatibilitatea cu EUROM este asigurată printr-un nucleu obţinut plecând de la materialele din aceasta: 40 de pasaje; 26 propoziţii de completare; un set de 26 numere între 0 şi 9999, acoperind constrângerile fonotactice ale numerelor din această gamă în limba română şi minimizat plecând de la diagramele lor de sintaxă; logatomi de forma CVC şi cinci perechi de cuvinte-contexte, proiectate pentru limba română conform aceloraşi principii utilizate şi în cazul EUROM [230].

În jurul acestui nucleu au mai fost incluse:

- patru propoziţii fonemic compacte, citite de către toţi vorbitorii şi utile pentru iniţializarea modelelor acustice;
- propoziţii individuale, specifice fiecărui vorbitor, selectate dintr-un corpus printr-un algoritm de tip greedy având ca obiectiv maximizarea numărului de difoni aşteptaţi să apară din citirea pasajelor şi a propoziţiilor de completare şi individuale;
- materiale care să permită dezvoltarea unor aplicaţii simple şi studiul diferenţelor între stilul citit şi cel semispontan de vorbire: alfabetul limbii române (citit) şi unele informaţii furnizate semispontan de vorbitori (numele, pe cuvinte şi litere; seria şi numărul buletinului de identitate; numărul de telefon; data naşterii; adresa).

Prezentăm în continuare unele detalii legate de materialele principale, utilizate pentru antrenarea şi evaluarea sistemelor de recunoaştere a vorbirii – pasajele şi propoziţiile.

4.3.1 Pasajele

Criteriul temelor lor comune în toate limbile din EUROM a impus traducerea şi adaptarea pasajelor dintr-o versiune iniţială în limba engleză. Realizate ca transcrieri fonemice folosind un editor cu facilităţi de calcul şi afişare în timp real a unor statistici

Algoritmul 4.1 Algoritmul de grupare a pasajelor

```
 1: P← numărul de pasaje, K ← numărul de clustere    {K divizor al lui P}
 2: FON ← listă foneme, PAS ← listă pasaje, CLUS ← listă clustere (vide)
 3: for f ∈ FON, p ∈ PAS, c ∈ CLUS do
 4:   N_p[f] ← numărul de apariţii ale f în pasajul p
 5:   N_c[f] ← 0     {clustere vide}
 6:   N[f] ← numărul total de apariţii ale f
 7: end for
 8: ordonează FON în ordinea crescătoare a numărului de apariţii ale fonemelor
 9: for f ∈ FON do
10:   ordonează CLUS în ordinea descrescătoare a numărului de apariţii ale f
11:   ordonează PAS în ordinea crescătoare a numărului de apariţii ale f
12:   while ∃ c ∈ CLUS a.î. N_c[f] < N[f]/K - Δ[f] do
13:     for c ∈ CLUS do
14:       if N_c[f] < N[f]/K - Δ[f], c incomplet şi PAS nevidă then
15:         adaugă la c primul pasaj p din PAS
16:         N_c[f] ← N_c[f] + N_p[f]
17:         şterge primul pasaj din PAS
18:       else if N_c[f] < N[f]/K - Δ[f] şi c complet sau PAS vidă then
19:         abandonează încercarea de grupare
20:       end if
21:     end for
22:   end while
23: end for
24: for p ∈ PAS do
25:   adaugă p la primul cluster incomplet
26: end for
```

ale fonemelor, traducerea şi adaptarea au inclus modificări, în special ale unor toponime, pentru a creşte frecvenţele de apariţie ale fonemelor rare.

Pentru a obţine o structură cât mai ordonată a înregistrărilor şi o distribuţie cât mai echilibrată a fonemelor în pronunţiile vorbitorilor, cele 40 de pasaje au fost grupate în 10 clustere de câte 4 pasaje folosind un algoritm special conceput (algoritmul 4.1) care urmăreşte gruparea a P pasaje în K clustere de câte P/K pasaje în care fonemele să fie, în limita posibilităţilor, cât mai uniform distribuite.

În cursul producerii vorbirii, fonemele sunt realizate cu frecvenţe diferite (în pasaje, de exemplu, numerele totale de apariţii ale fonemelor, N_p[f], variază între 15 şi 1104), iar construcţia unor modele acustice de o calitate acceptabilă impune ca fiecare fonem să aibă cel puţin un anumit număr minim de realizări. Din această cauză, fonemele rare trebuie urmărite cu prioritate maximă, iar numerele minime de apariţii ale fonemelor în clustere (liniile 12, 14 şi 18) sunt asigurate în ordinea crescătoare a numerelor totale de apariţii ale fonemelor (linia 8): pentru a obţine şi o distribuţie pe clustere cât mai uniformă, numărul de apariţii ale fonemului f în clusterul c, N_c[f], poate fi cu cel mult Δ[f] mai mic decât numărul său mediu de apariţii într-un cluster, N[f]/K. Abaterea maximă

Tabelul 4.2: Gruparea pasajelor în clustere

Cluster	0	1	2	3	4	5	6	7	8	9
Pasaje	P6	O7	O1	O0	O9	P2	O6	O5	O4	O8
	Q0	P4	P7	O2	P8	Q4	P1	Q2	P9	P0
	Q3	Q5	Q1	O3	Q8	R1	P3	Q7	Q6	R4
	R0	R2	R7	Q9	R3	R9	P5	R6	R5	R8

admisibilă $\Delta[\mathtt{f}]$, dependentă de fonem pentru a ţine cont de inegalitatea frecvenţelor de apariţie, a fost calculată în implementarea practică a algoritmului cu formula

$$\Delta[\mathtt{f}] = a \cdot \frac{\mathtt{N[f]} - \min_{\mathtt{f}}(\mathtt{N[f]})}{\max_{\mathtt{f}}(\mathtt{N[f]}) - \min_{\mathtt{f}}(\mathtt{N[f]})} \tag{4.1}$$

gruparea efectivă a pasajelor fiind realizată prin utilizarea constantei $a = 73$. În măsura în care acest lucru a fost posibil, aceeaşi abatere maximă $\Delta[\mathtt{f}]$ a fost utilizată şi pentru a limita depăşirea numărului mediu de apariţii ale unui fonem într-un cluster.

Pasajele au fost identificate printr-un cod format dintr-o literă (O, P, Q sau R) şi o cifră zecimală (0...9), iar gruparea lor în clustere este prezentată în tabelul 4.2.

4.3.2 Propoziţiile

Pe lângă necesităţile modelării acustice şi compatibilitatea cu EUROM, pregătirea materialelor de înregistrat a luat în calcul şi aspecte legate de etichetarea bazei de date. Astfel, având în vedere automatizarea etichetării folosind modele Markov ascunse, au fost create patru propoziţii fonemic compacte citite de către toţi vorbitorii înregistraţi, propoziţii aşa-zise de iniţializare: din citirea fiecăreia se aşteaptă să se obţină minimum o realizare a fiecărui fonem, iar prin etichetarea manuală a înregistrărilor (secţiunea 5.3.1) - materialul pentru iniţializarea unor modele acustice ale fonemelor.

O a doua categorie de propoziţii este a celor de completare: în urma grupării pasajelor, câteva foneme erau încă slab reprezentate în unele clustere – de exemplu, fonemele /G/ şi /h/ apăreau doar de câte 15 ori în total, şi în multe clustere doar o singură dată. Din acest motiv au fost create, folosind acelaşi editor de transcrieri fonemice utilizat şi pentru pasaje, 26 propoziţii de completare care să asigure pentru toate fonemele un număr minim de apariţii în fiecare cluster. Acest număr minim a fost ales având din nou în vedere automatizarea etichetării: studii anterioare [214] au indicat că etichetarea automată folosind MMA ale fonemelor necesită pentru antrenarea unui MMA un număr minim de cca. 70 realizări ale fonemului asociat; ca urmare, având în vedere etichetarea în tranşe de înregistrări ale celor 10 clustere, fiecare cluster a fost extins cu 2 sau 3 propoziţii de completare, care să asigure oricărui fonem minimum 7 apariţii/cluster.

Ultima categorie de propoziţii, incluse pentru o mai mare varietate a materialelor înregistrate, sunt cele individuale, specifice unui anumit vorbitor. Ele au fost obţinute plecând de la texte literare, din care într-o primă fază au fost extrase propoziţii de dimensiuni convenabile pentru înregistrări. Printr-un algoritm de tip greedy urmărind maximizarea numărului de difoni, din corpus au fost apoi alese 558 propoziţii.

Tabelul 4.3: Distribuţia vorbitorilor pe clustere extinse alocate pentru citire. Prima din cele două litere care formează codul unui vorbitor arată sexul (F,G - feminin, M,N - masculin), iar indicii arată grupele de vârstă, numerotate în ordinea: sub 20 de ani; 20-29; 30-39; 40-49; 50 şi peste 50 de ani.

Cluster	Vorbitori									
0	MA_1	$\underline{FF_2}$	MK_3	FP_4	MU_5	GA_1	NF_2	GK_3	NP_4	GU_5
1	FA_1	MF_2	FK_3	MP_4	FU_5	NA_1	GF_2	NK_3	GP_4	NU_5
2	MB_2	FG_3	ML_4	FQ_5	MV_1	GB_2	NG_3	GL_4	NQ_5	GV_1
3	FB_2	MG_3	FL_4	MQ_5	FV_1	NB_2	GG_3	NL_4	GQ_5	NV_1
4	$\underline{MC_3}$	FH_4	MM_5	FR_1	MX_2	GC_3	NH_4	GM_5	NR_1	GX_2
5	FC_3	MH_4	FM_5	MR_1	FX_2	NC_3	GH_4	NM_5	GR_1	NX_2
6	MD_4	FI_5	MN_1	FS_2	MY_3	GD_4	NI_5	GN_1	NS_2	GY_3
7	FD_4	MI_5	FN_1	MS_2	FY_3	ND_4	GI_5	NN_1	GS_2	NY_3
8	ME_5	FJ_1	MO_2	FT_3	MZ_4	GE_5	NJ_1	GO_2	NT_3	GZ_4
9	FE_5	MJ_1	FO_2	MT_3	FZ_4	NE_5	GJ_1	NO_2	GT_3	NZ_4

4.4 Vorbitorii

Existenţa celor 10 clustere extinse a făcut posibilă planificarea înregistrărilor în grupe de 20 de vorbitori egal distribuiţi pe sexe, iar compatibilitatea cu EUROM [46] a impus înregistrarea a minimum 60 de vorbitori. Pentru a obţine însă cât mai multe date pentru antrenarea, testarea şi evaluarea sistemelor de recunoaştere, am anticipat o extindere până la 100 de vorbitori, cu o posibilă fază intermediară la 80 de vorbitori. Ansamblul vorbitorilor constituie, în terminologia EUROM, aşa-numita mulţime "Many Talker", prescurtat MT, toţi vorbitorii trebuind să înregistreze într-o singură sesiune:

- materiale care asigură compatibilitatea cu EUROM: câte un cluster extins (4 pasaje şi 2 sau 3 propoziţii de completare asociate) şi cele 26 numere între 0 şi 9999;
- materiale adiţionale: cele 4 propoziţii de iniţializare, 3...7 propoziţii individuale, alfabetul şi informaţiile personale.

Dată fiind repetarea înregistrărilor celor 10 clustere extinse, pe lângă reprezentarea egală a celor două sexe am mai urmărit şi distribuţia uniformă a vorbitorilor în cinci grupe de vârstă - sub 20, 20-29, 30-39, 40-49, 50 şi peste 50 ani. Vorbitorii au fost astfel repartizaţi pe clustere într-o structură bloc aleatoare (randomized block design) [37], [158] având ca variabile de blocare sexul şi grupa de vârstă (tabelul 4.3).

În tabelul 4.3, primele şase coloane de vorbitori asigură compatibilitatea cu EUROM, iar următoarele două - faza intermediară de 80 vorbitori. Zece vorbitori, unul din fiecare sex şi grupă de vârstă (*evidenţiaţi* în tabel), constituie mulţimea "Few Talker", prescurtat FT. Ei au fost planificaţi să înregistreze în plus logatomii de tip CVC, iar în patru sesiuni suplimentare, decalate la câte cel puţin două săptămâni - patru noi clustere extinse şi numerele. Doi vorbitori din mulţimea FT (subliniaţi în tabel), unul din fiecare sex,

Tabelul 4.4: Proprietăţi ale clusterelor de pasaje extinse cu propoziţii de completare: entropiile fonemelor şi numerele de cuvinte distincte şi totale

Cluster	Entropie	Cuvinte distincte	Total cuvinte
0	4,51	177	272
1	4,54	170	244
2	4,49	182	256
3	4,49	178	267
4	4,48	161	219
5	4,51	174	238
6	4.51	171	238
7	4,55	177	246
8	4.53	166	239
9	4,55	188	259
Toate	4,53	1160	2478

formează mulţimea "Very Few Talker", prescurtat VT, ei înregistrând în prima sesiune şi logatomii în cinci contexte de câte două cuvinte, precum şi aceste cuvinte izolate.

4.5 Analize statistice

Analiza clusterelor extinse (tabelul 4.4) a arătat că ele au caracteristici comparabile atât prin prisma valorilor entropiei fonemelor, cât şi a numerelor de cuvinte distincte şi totale, numere în limitele a 8% respectiv 12% faţă de medie (sub două abateri standard). Constatarea este susţinută şi de valorile entropiei relative (divergenţă informaţională [129] sau Kullback-Leibler) dintre distribuţiile lor fonemice, valori cuprinse între 0,02 şi 0,05 biţi, cu o medie de 0,032 biţi şi o abatere standard de 0,0075 biţi.

Se constată de asemeni că numărul de 1160 cuvinte distincte din clusterele extinse permite abordarea obiectivului principal al cercetărilor (secţiunea 1.3) – recunoaşterea vorbirii continue, independentă de vorbitor, cu vocabulare în jurul a 1000 de cuvinte.

Pentru a verifica satisfacerea criteriului numărului minim de apariţii ale unui fonem (secţiunea 4.3.2) şi a estima cantitatea de date disponibilă pentru antrenarea modelelor acustice, a fost realizată şi o statistică a fonemelor (tabelul 4.5) în transcrierile fonemice ale pasajelor şi propoziţiilor de completare, respectiv a numerelor de pronunţii ale lor aşteptate să apară in înregistrările de pasaje şi propoziţii de completare şi individuale ale 60, 80 şi 100 vorbitori. În tabelul 4.5, fonemele sunt în general în ordinea crescătoare a numerelor (aşteptate) de apariţii. Câteva excepţii apar în cazul transcrierilor pentru perechile de foneme /z/-/b/, /p/-/m/ şi /t/-/n/, ale căror numere de apariţii sunt în ordine inversă: diferenţele dintre numere sunt însă mici, iar textele literare din care provin propoziţiile individuale, luate în calcul doar în cazul ultimelor trei coloane, pot fi considerate mai reprezentative pentru limba română decât propoziţiile de completare şi pasajele, astfel încât aceste câteva excepţii pot fi neglijate.

Tabelul 4.5: Numerele de apariţii ale fonemelor în transcrierile pasajelor şi propoziţiilor de completare, respectiv aşteptate în pronunţiile pasajelor şi propoziţiilor de completare şi individuale de către 60, 80, 100 vorbitori

Fonem	Transcrieri	60 vorbitori	80 vorbitori	100 vorbitori
h	70	457	605	751
G	72	474	630	785
J	72	492	650	804
O	76	556	732	904
w	81	598	792	970
g	98	742	968	1183
z	*112*	813	1068	1316
b	*108*	850	1108	1349
E	119	969	1262	1542
f	129	1015	1326	1601
T	132	1025	1348	1642
v	147	1097	1442	1782
I	177	1252	1653	2039
C	181	1324	1739	2145
S	195	1424	1853	2273
y	222	1798	2349	2852
j	293	2227	2935	3613
p	*363*	2759	3591	4396
m	*362*	2887	3780	4643
o	377	2912	3835	4727
d	384	2987	3897	4777
k	437	3361	4392	5395
l	456	3613	4728	5790
s	478	3757	4919	6017
@	508	3870	5073	6198
u	598	4777	6231	7601
i	677	5286	6912	8437
t	*711*	5443	7141	8750
n	*707*	5513	7224	8812
r	803	6054	7950	9756
a	1226	9213	12079	14823
e	1233	9330	12232	15034

Analizele nu au inclus: propoziţiile de iniţializare, folosite doar pentru iniţializarea modelelor: numerele şi logatomii CVC, utile doar pentru testare, evaluare şi diagnoză; alfabetul şi informaţiile personale, cu pronunţii greu sau imposibil de anticipat (în plus, pronunţiile semispontane pot avea caracteristici acustice diferite de ale celor citite).

4.6 Organizarea bazei de date

Compatibilitatea cu EUROM a bazei de date a fost asigurată şi la nivelul organizării şi formatelor fişierelor componente prin înregistrarea ei folosind acelaşi pachet software utilizat şi pentru EUROM, denumit EUROPEC [262], [90]. Acesta utilizează, gestionează şi generează diverse fişiere, majoritatea de tip text:

- un fişier de descriere a vorbitorilor, completat la momentul înregistrării cu date de identificare şi caracteristici ale acestora care pot avea legătură cu modul în care vorbesc (înălţimea, greutatea, limba maternă, educaţia etc.);
- un fişier de descriere a materialelor, precizând pentru fiecare codul de identificare şi tipul (tabelul 4.6), protocolul folosit pentru înregistrare etc.;
- fişiere corpus, incluzând materialele pentru înregistrări: în cazul celor cu conţinut predeterminat (pasaje, propoziţii, numere, logatomi, alfabet), textul acestora, iar în cazul celor semispontane, un text generic indicând informaţia solicitată – nume, adresă etc.; textele au fost formatate conform cerinţelor EUROPEC înainte de realizarea înregistrărilor, iar pe durata lor au fost afişate pentru citire şi stocate, împreună cu alte informaţii, în fişierele de adnotări ortografice ale semnalelor;
- protocoale de înregistrare, referite în fişierul de descriere a materialelor şi specifice diferitor tipuri de materiale: acestea sunt interpretate de EUROPEC pe durata înregistrărilor, asigurând secvenţa dorită de operaţiuni;
- fişiere de semnal, singurele de tip binar, obţinute prin înregistrarea vorbitorilor în timpul citirii textelor sau al pronunţării informaţiilor solicitate;
- fişiere de adnotări ortografice corespunzătoare celor de semnal, generate odată cu ele şi cuprinzând fiecare: numele fişierului de semnal asociat şi ale fişierelor corpus şi de descriere a protocolului folosite la înregistrarea lui; date despre vorbitorul înregistrat (sex, vârstă, limbă maternă); caracteristicile înregistrării (frecvenţa de eşantionare, numărul şi caracteristicile eşantioanelor) şi localizarea în ea a porţiunilor corespunzătoare unor propoziţii/părţi din textele citite etc.;
- alte fişiere cu descrieri ale configuraţiilor de lucru, condiţiilor de înregistrare etc.

Pentru fiecare înregistrare, EUROPEC generează un fişier de semnal, unul de adnotări ortografice şi unul conţinând configuraţia folosită. Pentru identificarea vorbitorului şi a materialului, numele acestor fişiere includ codurile lor, fiind de forma `vvmmnnnn.ext`, unde `vv` este codul vorbitorului (tabelul 4.3), `mm` – codul materialului (tabelul 4.6), iar `nnnn` – un număr de ordine al fişierului. Extensia `ext` este CFG pentru fişierele

Tabelul 4.6: Codurile de identificare (două caractere, sub formă de expresii regulate) şi tipurile materialelor înregistrate

Materiale	Coduri de identificare	Tip
Pasaje	[O-R][0-9]	P
Propoziţii de completare	F[0-9]	S
Propoziţii individuale	[AB][A-Z], C[A-H], [DE][A-T]	D
Propoziţii de iniţializare	I0	0
Numere	N0	N
Logatomi CVC, contexte	S[1-3], [A-C][1-5], Z1	C
Alfabet	SZ	Z
Nume	SN	N
Adresă	SA	A
Serie şi număr B.I.	SI	I
Dată naştere	SB	B
Telefon	ST	T

de configuraţie şi de forma TRS, TRO pentru fişierele de *S*emnal respectiv adnotări *O*rtografice, unde T este o literă indicând tipul materialului (tabelul 4.6).

4.7 Realizarea înregistrărilor

Utilizarea EUROPEC pentru realizarea înregistrărilor a necesitat: configurarea unei staţii de lucru SESAM [46] pentru rularea lui prin instalarea unei plăci de achiziţie şi prelucrare a semnalelor de tip OROS AU21 [177] într-un calculator compatibil PC; modificarea unor fonturi pentru caracterele diacritice româneşti; traducerea şi adaptarea mesajelor EUROPEC în limba română; formatarea materialelor de înregistrat în fişiere corpus conforme cu cerinţele EUROPEC; scrierea, testarea şi depanarea protocoalelor de înregistrare specifice diferitelor tipuri de materiale. Fişierele astfel rezultate au fost organizate într-o structură unitară (secţiunea 4.6) utilizând fişiere de configurare şi de descriere a materialelor şi condiţiilor de înregistrare. În sfârşit, au fost dezvoltate fişiere de comenzi pentru simplificarea menţinerii acestei structuri şi a operării EUROPEC.

Pentru a obţine o calitate cât mai bună a înregistrărilor prin minimizarea zgomotelor şi distorsiunilor, acestea au fost realizate într-o cameră izolată şi tratată fonic, în care vorbitorii citeau textele alocate sau pronunţau informaţiile cerute sub supravegherea unor operatori plasaţi, împreună cu echipamentele, într-o cameră alăturată (figura 4.1). Iniţial se intenţiona ca vorbitorii să citească instrucţiunile şi textele de pe ecranul unui monitor, aşa cum erau ele afişate de EUROPEC, dar experimente preliminare au arătat că bobinele de deflexie ale monitoarelor constituie o sursă semnificativă de zgomot.

Deoarece la momentul realizării înregistrărilor monitoarele cu cristale lichide erau practic inaccesibile, monitorul din camera izolată şi tratată fonic a fost înlocuit cu un interfon şi listinguri ale textelor de citit. Pe lângă eliminarea zgomotului monitorului,

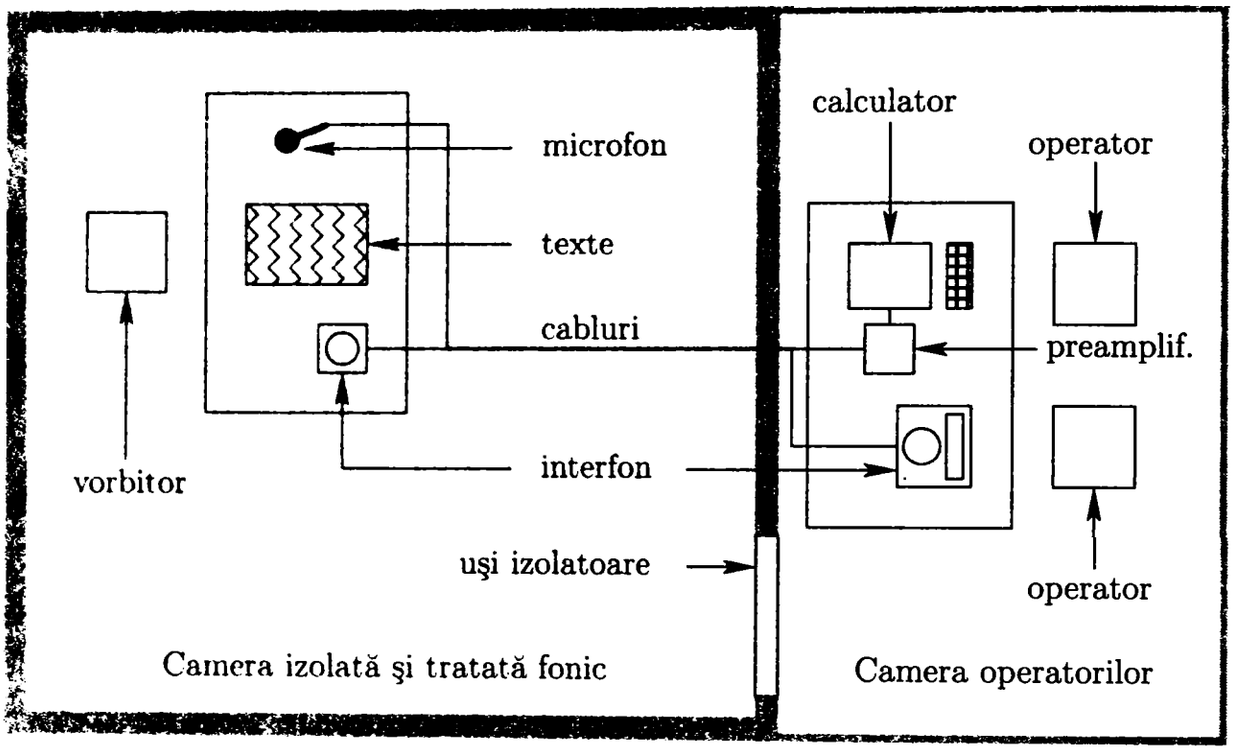

Figura 4.1: Schiţa camerelor folosite pentru înregistrări

această soluţie are şi două avantaje suplimentare: pe de o parte, facilitează detectarea de către operatori a zgomotelor sau altor probleme de pe durata înregistrărilor, permiţând întreruperea şi reluarea lor imediată, fără a mai aştepta o verificare auditivă ulterioară; pe de altă parte, reduce efectele nedorite care pot fi resimţite de persoane plasate într-o cameră izolată şi tratată fonic, efecte mergând de la dificultăţi în controlul volumului vocii până la claustrofobie. Interfonul era controlat de un operator astfel încât în timpul înregistrărilor să fie activă doar comunicarea dinspre vorbitor spre operatori.

O altă problemă potenţială a fost cea a ordinii de înregistrare a diferitor tipuri de materiale – citite sau semispontane. Pentru a evita pe cât posibil influenţa unui stil de vorbire asupra celuilalt, a fost stabilită următoarea ordine de înregistrare:

- informaţiile personale (nume, adresă etc.) solicitate de către operatori şi pronunţate semispontan de către vorbitori;
- alfabetul limbii române, citit de fiecare vorbitor în felul cu care era obişnuit;
- pasajele, citite cât mai fluent, fără a marca în mod special pauzele dintre propoziţii;
- propoziţiile de completare şi individuale, citite cu pauze suficient de lungi între ele pentru a permite separarea vorbirii şi a pauzelor cu algoritmii din EUROPEC;
- propoziţiile de iniţializare, citite ca şi cele anterioare, dar vorbitorii au fost instruiţi în mod special să le pronunţe cât mai clar şi mai apropiat de varianta standard;

- numerele, citite ca şi propoziţiile, cu pauze între ele;
- logatomii de tip CVC, citiţi izolaţi doar de vorbitorii din mulţimile restrânse FT şi VT, şi în contexte de câte două cuvinte doar de către cei din mulţimea VT;
- cuvintele-contexte CVC, citite izolat doar de vorbitorii din mulţimea VT.

Înregistrările au fost realizate folosind un microfon cu electret, omnidirecţional, de tip SONY ECM-44B, plasat la aproximativ 25 cm de gura vorbitorului şi la un unghi de circa 30 grade faţă de direcţia lui înainte. Pentru evitarea zgomotelor electrice şi adaptarea la placa de achiziţie şi prelucrare a semnalelor din calculator, microfonul a fost conectat la aceasta prin cabluri ecranate şi un preamplificator cu un câştig fix de circa 20 dB (figura 4.1). Semnalul a fost eşantionat la 20 KHz şi cuantizat pe 16 biţi.

Înaintea sesiunii de înregistrări (a primei sesiuni, în cazul vorbitorilor din mulţimile restrânse FT şi VT), fiecare vorbitor a furnizat o serie de informaţii personale, unele introduse şi în fişierul de descriere a vorbitorilor, şi a fost instruit asupra modului de lucru într-o scurtă sesiune de antrenament folosind informaţiile personale şi alfabetul.

Pe parcursul înregistrărilor, operatorii au urmărit permanent apariţia unor zgomote sau erori majore de citire (omisiuni, inserţii sau substituţii de cuvinte, bâlbâieli etc.), cu întreruperea şi reluarea imediată a înregistrărilor afectate. Înregistrările au fost refăcute imediat şi în cazul în care asemenea probleme erau constatate prin ascultarea fişierelor de semnal. Cu excepţia propoziţiilor de iniţializare, în cazul cărora vorbitorii au fost solicitaţi să le citească într-un mod cât mai clar şi mai apropiat de varianta standard, variaţiile de pronunţie nu au constituit motive de refacere a înregistrărilor.

Pentru evitarea zgomotelor transmise în interiorul camerei izolate fonic prin pereţii clădirii, înregistrările au fost efectuate în zile nelucrătoare (sâmbăta şi duminica), ceea ce a făcut dificilă găsirea unor vorbitori dispuşi să ia parte la înregistrări. Vorbitorii au fost recrutaţi urmărind satisfacerea criteriilor stabilite de sex şi vârstă (secţiunea 4.4), precum şi capacitatea de a citi în mod fluent materialele de înregistrat. În aceste condiţii, înregistrarea a 100 de vorbitori a durat peste un an (martie 1996 – iunie 1997).

4.8 Datele colectate

Din înregistrarea celor 100 de vorbitori au rezultat peste 1700 fişiere de semnal însumând peste 1,3 gigaocteţi de date şi cuprinzând peste 9 ore şi 41 minute de înregistrări, distribuite pe diferite categorii de materiale conform tabelului 4.7.

Tabelul 4.7: Cantităţi de date colectate, pe categorii

	Pasaje	Propoziţii			Numere	CVC	Alfabet	Info
		Comp.	Ind.	Iniţ.				
Fişiere	560	140	100	100	140	62	100	500
Durată	3h39'12"	28'23"	43'3"	33'22"	2h21'29"	19'8"	43'29"	52'45"
Propoziţii	2758	364	558	400	–	–	–	–

Dintre acestea, pasajele şi propoziţiile de diverse tipuri, care sunt materialele cele mai importante din punctul de vedere al obiectivului principal al cercetărilor, fiind cele mai potrivite pentru antrenarea, testarea şi evaluarea sistemelor de recunoaştere automată a vorbirii, au o durată totală de 5 ore şi 24 minute şi cuprind 4080 de propoziţii.

4.9 Calitatea înregistrărilor

Printre caracteristicile bazelor de date fonetice, enumerate la începutul capitolului, se numără şi calitatea deosebită a înregistrărilor, apreciată prin elemente subiective (lipsa zgomotelor, corectitudinea pronunţiilor etc.) şi măsuri cantitative. Aceste aspecte au fost avute în vedere pe toată durata sau chiar dinaintea începerii înregistrărilor.

Prin utilizarea pentru înregistrări a unei camere izolate fonic s-a urmărit atenuarea zgomotelor acustice propagate din exterior, iar prin efectuarea lor în zile nelucrătoare – reducerea la minimum a posibilităţii apariţiei zgomotelor. Alte zgomote acustice puteau fi produse chiar de vorbitorii însăşi în timpul înregistrărilor: minimizarea lor a fost asigurată prin ascultarea de către operatori a înregistrărilor prin interfon şi din fişiere în timpul şi după efectuarea lor, urmată de reluare sau refacere când era cazul. Zgomotele electrice au fost minimizate prin ecranarea legăturii microfon-preamplificator.

Pe lângă zgomote, semnalul putea fi afectat şi de distorsiuni cauzate de reflexiile şi reverberaţiile incintei acustice folosite pentru înregistrări: acestea au fost evitate graţie tratării fonice a camerei utilizate. O altă categorie de distorsiuni posibile ale semnalului erau cele datorate neliniarităţilor sau saturării lanţului de înregistrare. Pentru evitarea neliniarităţilor au fost utilizate un microfon cu o caracteristică cât mai liniară şi un preamplificator liniar, iar saturaţia a fost supravegheată cu ajutorul EUROPEC.

Eficienţa măsurilor de asigurare a calităţii a fost evaluată obiectiv prin estimarea şi verificarea unor caracteristici ale semnalelor rezultate. Astfel, componenta continuă, estimată ca medie a tuturor eşantioanelor dintr-un fişier, a fost cuprinsă între $-0{,}08$ şi $0{,}07$, cu o medie practic nulă şi o abatere standard sub 0,01, fiind deci nesemnificativă în raport cu amplitudinile posibile ale semnalelor, cuprinse între -32768 şi 32767. La rândul ei, saturaţia a apărut doar într-un fişier, sub forma a patru secvenţe de 3-4 eşantioane cu valori minime (-32768), toate în limitele unui singur segment de circa 20 ms.

O altă măsură a calităţii înregistrărilor este raportul semnal/zgomot, definit ca [59]

$$RSZ = 10 \log_{10} \frac{E_s}{E_z} = 10 \left(\log_{10} E_s - \log_{10} E_z\right) \quad [\text{dB}] \tag{4.2}$$

unde E_s şi E_z sunt energia semnalului util respectiv zgomotului. Dat fiind caracterul nestaţionar al semnalului vocal, valorile raportului semnal/zgomot sunt în acest caz variabile în timp, astfel încât nu poate fi estimată decât o valoare medie. În plus, zgomotul este dificil de separat de semnal: în principiu, un zgomot staţionar ar putea fi estimat din analiza pauzelor de vorbire, dar delimitarea acestora este o problemă în sine.

Unele metode de estimare a valorii raportului semnal/zgomot evită identificarea pauzelor prin utilizarea distribuţiei valorilor energiei cadrelor de semnal (figura 4.2), care are două maxime: unul corespunzător pauzelor şi fazelor de închidere ale sunetelor plozive nesonore, la valori mici, şi unul corespunzător vorbirii – la valori mari.

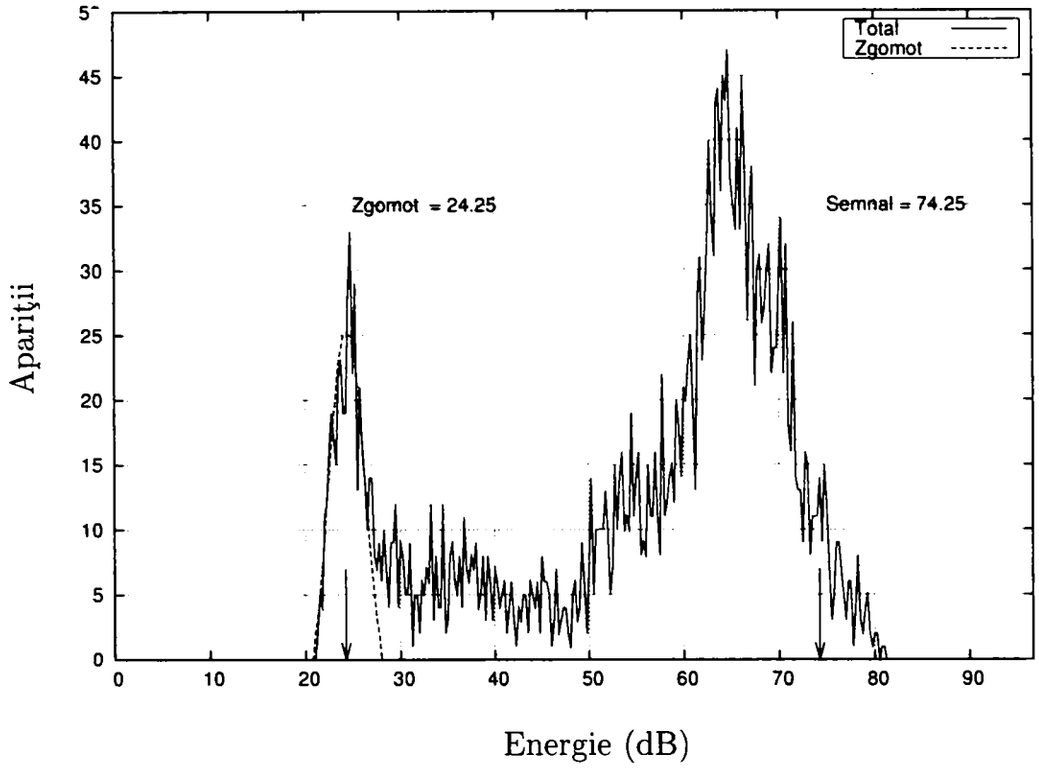

Figura 4.2: Exemplu de distribuţie a energiei cadrelor dintr-o înregistrare

O metodă simplă este ca energiei zgomotului să-i fie atribuită valoarea sub care se află un anumit procent din valori, iar celei a semnalului – valoarea sub care se încadrează un procent complementar. Aceste procente, fixe, sunt alese în mod empiric, în mod curent fiind utilizate valorile de 15% şi respectiv 85%. În aceste condiţii, valoarea raportului semnal/zgomot este calculată ca diferenţa dintre cele două valori (în decibeli).

Metode mai elaborate ţin cont de distribuţia efectivă a valorilor energiei pentru a estima modele ale zgomotului. O asemenea metodă, folosită de National Institute of Science and Technology al Statelor Unite, estimează energia zgomotului ca media unei distribuţii de tip cosinus ridicat a valorilor sale (cu linie întreruptă în figura 4.2). Odată energia zgomotului estimată, valoarea raportului semnal/zgomot este calculată prin aceeaşi metodă a procentelor complementare din paragraful de mai sus.

Această metodă a fost utilizată şi pentru verificarea înregistrărilor din această bază de date, iar rezultatele au confirmat calitatea lor: raportul semnal/zgomot are media de 48,41 dB, abaterea standard de 2,91 dB şi valorile extreme de 33,75 şi 57,75 dB.

Pentru comparaţie, analiza prin aceeaşi metodă a bazei de date TIMIT a condus la o valoare medie a raportului semnal/zgomot de 53,71 dB (cu 5,3 dB mai mare), cu o abatere standard de 5,19 dB (de 1,78 ori mai mare) şi valori extreme de 27 şi 69,75 dB. Ţinând cont de faptul că înregistrările din TIMIT au fost făcute cu un microfon de mică distanţă [81], pe când ale noastre – cu unul plasat la cca. 25 cm, putem considera că baza noastră de date este de o calitate comparabilă, ba chiar mai bună din punctul de vedere al dispersiei valorilor raportului semnal/zgomot.

4.10 Concluzii

Recunoaşterea vorbirii prin metode statistice cere cantităţi considerabile de date pentru estimarea modelelor acustice şi lingvistice pe care se bazează, iar acest capitol a prezentat proiectarea şi colectarea unei baze de date cuprinzând semnale vocale în limba română, utilizabilă pentru antrenarea de modele acustice şi testarea şi evaluarea unor sisteme de recunoaştere a vorbirii continue de tipul celor vizate prin aceste cercetări.

Pe lângă utilitatea directă pentru antrenarea modelelor acustice, această bază de date are însă o importanţă mult mai mare din punctul de vedere al metodologiei cercetării, existenţa şi utilizarea ei permiţând efectuarea în condiţii controlate a unor experimente de recunoaştere a vorbirii şi comparaţii semnificative între rezultatele lor.

Date fiind inexistenţa cercetărilor anterioare asupra recunoaşterii vorbirii continue în limba română şi insuficienţa celor de fonetică şi fonologie a limbii române, constatate prin consultarea literaturii şi a specialiştilor din aceste domenii, proiectarea şi colectarea bazei de date au pus un accent deosebit pe calitatea înregistrărilor, în lipsa căreia unele investigaţii necesare în această fază a cercetărilor ar fi dificile sau chiar imposibile.

Astfel, în etapa de proiectare a bazei de date a fost ales un set de unităţi fonetice de modelare acustică considerat acoperitor în raport cu sistemele fonologice ale limbii române identificate în literatura lingvistică, iar conţinutul unei părţi semnificative a înregistrărilor a fost planificat pentru a permite antrenarea de modele ale acestor unităţi.

Pentru a putea antrena şi testa modele acustice independente de vorbitor, populaţia înregistrată a inclus un număr considerabil de vorbitori (100), iar pentru o cât mai bună acoperire a variabilităţii datorate lor, aceştia au fost selectaţi pentru a obţine o distribuţie uniformă după două variabile biologice controlabile – sexul şi grupa de vârstă.

În sfârşit, pentru ca informaţia lingvistică inclusă în semnalele vocale înregistrate să fie cât mai puţin afectată de zgomote, distorsiuni, reverberaţii, o atenţie deosebită a fost acordată condiţiilor de realizare a înregistrărilor, iar analiza lor prin prisma câtorva criterii obiective de evaluare a calităţii acustice indică atingerea acestui obiectiv.

Datorită conţinutului lingvistic controlat încă din faza de proiectare al unei părţi semnificative a ei şi calităţii deosebite a înregistrărilor, această bază de date va fi utilă nu doar pentru cercetările asupra recunoaşterii automate a vorbirii continue, ci şi în alte cercetări aplicative, precum şi pentru cercetările fundamentale de fonetică acustică şi fonologie a limbii române, a căror insuficienţă a fost menţionată.

CAPITOLUL 5

Etichetarea semnalelor vocale

Valoarea unei baze de date de tipul celei descrise în capitolul 4 creşte dacă, pe lângă fişierele de semnal, ea include şi informaţii suplimentare referitoare la conţinutul lor. Deoarece înregistrările din această bază de date au fost făcute prin citirea unor texte pregătite în prealabil sau prin solicitarea anumitor informaţii, iar programul utilizat pentru realizarea acestor înregistrări a fost astfel implementat, fiecare fişier de semnal are asociat unul de adnotări ortografice (secţiunea 4.6), cuprinzând printre altele fie textul citit, fie descrierea informaţiei solicitate. Deşi aceste informaţii sunt fără îndoială utile, utilitatea semnalelor poate fi în continuare crescută dacă ele sunt etichetate.

Prin **etichetarea semnalului vocal** vom înţelege definirea unor evenimente din cadrul acestuia, evenimente identificate prin coordonate (limite) temporale şi simboluri (etichete) alese dintr-o mulţime finită şi definite în termeni acustici, fiziologici, fonetici sau aparţinând unor niveluri lingvistice superioare [18]. Dată fiind utilitatea ei pentru cercetările preconizate asupra recunoaşterii automate a unităţilor sublexicale de modelare acustică, ca şi pentru alte cercetări fundamentale şi aplicative, etichetarea bazei de date [33], [34] a fost un obiectiv (secţiunea 1.3) urmărit încă din faza de proiectare a ei.

Un exemplu de etichetare a unei propoziţii din baza de date în termenii unităţilor fonetice din tabelul 4.1 este prezentat în figura 5.1: localizările realizărilor unităţilor fonetice sunt indicate prin limite temporale (liniile punctate verticale), iar identităţile lor - prin plasarea între aceste limite a simbolurilor ASCII corespunzătoare.

Etichetarea semnalului vocal este deci o abstractizare dependentă de anumite puncte de vedere analitice şi teoretice, care conduc la diferite **niveluri de etichetare**, fiecare nivel putând fi la rândul lui format din **straturi** grupând multiple aspecte ce pot fi plasate pe acelaşi nivel. O prezentare a unora dintre nivelurile de etichetare cele mai relevante din punctul de vedere al utilizării bazelor de date vocale în cercetările fundamentale şi aplicative este realizată în secţiunea 5.1, împreună cu o justificare a nivelului ales pentru etichetarea semnalelor din baza noastră de date.

Cercetări realizate cu sprijinul Comisiei Europene prin contractul COPERNICUS 1304/1994, al fostului CNCSU (din 1999 CNCSIS) prin grantul 354/1996, al Academiei Române prin grantul 136/1997, şi al fostului Minister al Cercetării şi Tehnologiei prin contractul de grant 3019GR/1997-98.

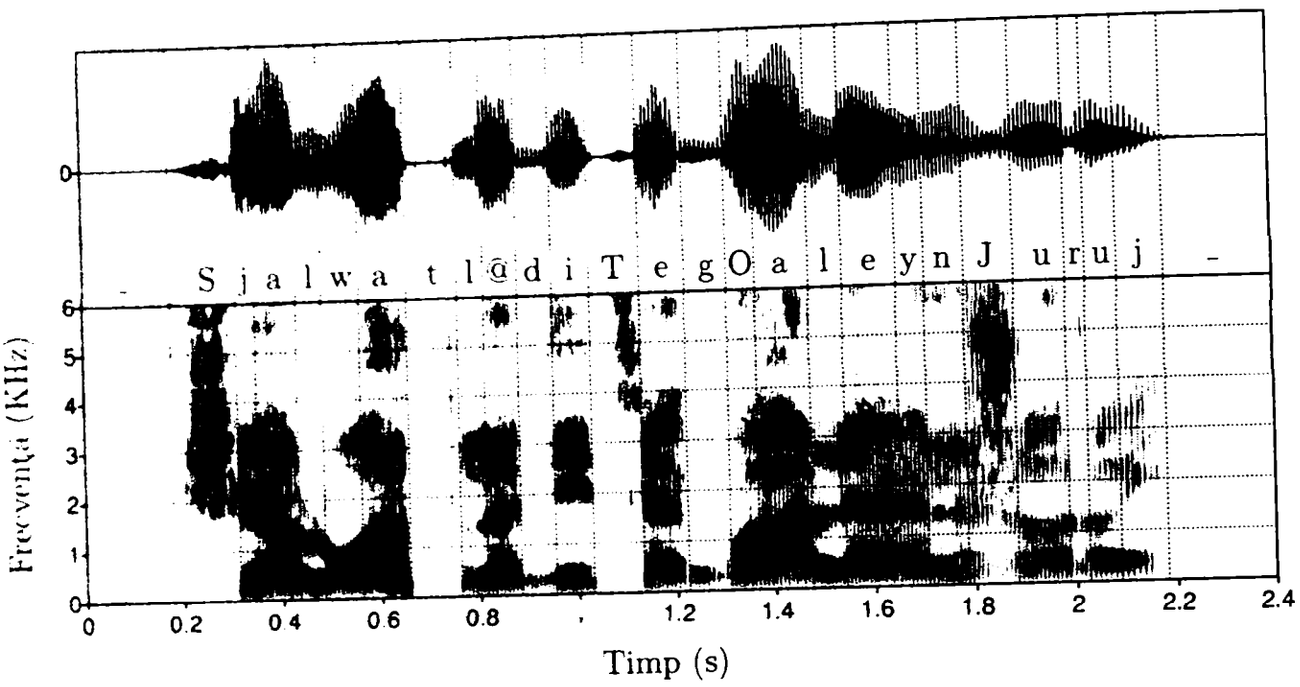

Figura 5.1: Exemplu de etichetare a propoziţiei "Şi-a luat lădiţe goale în juru-i" folosind simbolurile ASCII din tabelul 4.1. Urmăriţi cum se schimbă aspectul spectrogramei pentru diferite apariţii ale sunetelor [j], [a] şi [l], urmare a fenomenului de coarticulaţie.

Metoda de etichetare cea mai simplă consta în utilizarea unor programe adecvate [74], [107], [225] pentru vizualizarea formelor de undă şi a spectrogramelor semnalelor şi ascultarea lor selectivă, urmate de atribuirea manuală de poziţii temporale şi identităţi anumitor evenimente (lingvistice sau de altă natură). Deşi aparent cea mai fiabilă, fiind bazată pe expertiza umană, această metodă este totuşi susceptibilă de lipsa consistenţei poziţiilor temporale şi a identităţilor evenimentelor evidenţiate în cadrul semnalelor, dată fiind existenţa unor cazuri ambigue care necesită decizii subiective ce vor varia cu experienţa personală a experţilor implicaţi în etichetarea manuală.

Dar poate chiar mai important decât problema inconsistenţei etichetării este faptul că volumul de muncă solicitat pentru etichetarea manuală este foarte mare, de ordinul sutelor de ori durata semnalelor. Din acest motiv, automatizarea etichetării semnalelor vocale a fost urmărită încă din fazele preliminare ale construcţiei primelor baze de date fonetice, scop în care au fost propuse şi experimentate diferite metode [142], [266], [68], [144], [40], iar secţiunea 5.2 face o trecere în revistă a celor mai semnificative aspecte ale problemei, precum şi a metodelor de evaluare a rezultatelor etichetării automate.

Datorită utilităţii modelelor Markov ascunse şi pentru cercetările ulterioare asupra recunoaşterii automate a vorbirii, pentru etichetarea bazei de date a fost aleasă varianta unui sistem bazat pe MMA, a cărui construcţie este descrisă în secţiunea 5.3.

Pentru a asigura calitatea etichetării şi a permite evaluarea sistemului de etichetare, rezultatele etichetării au fost verificate şi, acolo unde a fost cazul, corectate manual pe baza unor criterii de decizie prezentate în secţiunea 5.4.

În final sunt prezentate şi comentate rezultatele etichetării automate, evaluate prin compararea etichetelor generate automat cu cele verificate şi corectate manual.

5.1 Alegerea nivelului de etichetare

Etichetarea semnalului vocal se poate face la diferite niveluri, luând sau nu în calcul anumite fenomene fizice şi/sau lingvistice şi corelaţiile care se pot stabili între ele. Pentru cercetările noastre, dar şi pentru multe altele, cele mai interesante sunt nivelurile care pot reprezenta legăturile dintre semnalele vocale şi interpretările lor lingvistice în termenii unor unităţi fonetico-fonologice segmentale [18]:

- nivelul **fizic** cuprinde etichetele definite cu referinţă în exclusivitate la evenimentele fizice dintr-o pronunţie, fiind în mod clar cel mai susceptibil de a fi divizat în straturi corespunzătoare diferitor metode de achiziţie şi prelucrare a semnalului;

- nivelul **fonetico-acustic** include etichete ce descriu evenimente omogene din punct de vedere acustic folosind termeni fonetici (închidere, eliberare, aspiraţie, fricţiune, sonoritate, nazalizare etc.), fără referiri la funcţiile lor lingvistice sau distinctive, sau la legături cu evenimente fizice; cu toate acestea, deciziile asupra delimitării lor temporale şi a setului de simboluri utilizate ca identificatori adesea necesită sau sunt facilitate de informaţii asupra rolului lor în termeni fonologici;

- nivelul **fonetic restrâns** grupează etichete care caracterizează calitatea fonetică a sunetelor vorbirii folosind simboluri din alfabetul fonetic internaţional (IPA) sau altele echivalente, reprezentabile pe calculator - SAMPA [251], Worldbet [104] etc.; în acest caz, impresia perceptuală a persoanei care realizează etichetarea este esenţială pentru stabilirea delimitărilor temporale şi a identităţilor sunetelor;

- nivelul **fonemic** este cel mai abstract dintre nivelurile prezentate aici, simbolurile folosite corespunzând fonemelor limbii în care a fost rostită o pronunţie, aşa cum apar ele în formele standard ale cuvintelor din cadrul pronunţiei, fără luarea în consideraţie a fenomenelor specifice vorbirii fluente (secţiunea 5.4.1); drept urmare, simbolurile nu vor putea fi puse întotdeauna în corespondenţă cu evenimente din semnal, astfel încât acest nivel nu este folosit pentru etichetarea efectivă; el este însă indispensabil ca mediator între semnal şi vocabular, fiind prezent în dicţionarele de pronunţii pe care le folosesc sistemele de recunoaştere şi sinteză a vorbirii;

- nivelul **fonetic extins**: adesea denumit fonemic, datorită faptului că utilizează de asemeni simboluri corespunzătoare fonemelor limbii în care a fost rostită pronunţia adnotată, acest nivel, spre deosebire de cel anterior, ia în consideraţie fenomenele specifice vorbirii fluente şi le reflectă ca atare, având astfel un grad de abstractizare intermediar între cel al nivelului fonetic restrâns şi cel al nivelului fonemic.

Etichetarea avută în vedere aici trebuind să asigure legătura dintre caracteristicile anumitor porţiuni ale semnalului şi categorii lingvistice asociabile lor, doar ultimele trei dintre aceste niveluri, care includ simboluri motivate lingvistic, prezintă interes în

continuare. Dintre acestea, nivelul fonemic este eliminat ca opţiune practică din motivele deja menţionate, astfel încât în final am optat pentru nivelul fonetic extins.

Această opţiune are motivaţii multiple: nivelul fonetic extins este cel mai economic, reprezentând un maximum de informaţie fonetică cu un set minimal de simboluri; având în comun cu nivelul fonemic setul de simboluri folosit, are şi avantajul că facilitează antrenarea şi evaluarea la nivel de fonem a sistemelor de recunoaşterea automată a vorbirii bazate pe modelarea acustică a unităţilor sublexicale de tip fonemic. În plus, acest nivel este mai fiabil decât nivelul fonetic restrâns în ceea ce priveşte consistenţa între diferiţi experţi a simbolurilor folosite pentru etichetarea aceloraşi semnale, şi aproximativ la fel de fiabil ca acesta în privinţa delimitărilor temporale [67].

5.2 Automatizarea etichetării

Aşa după cum am precizat, etichetarea semnalului vocal în general, şi cea la nivelul fonetic extins în particular, presupune două acţiuni:

- **segmentarea** semnalului prin identificarea momentelor de timp la care începe respectiv se termină o porţiune din semnal ce poate fi asociată unui fonem;
- **identificarea** fonemului corespunzător unui segment.

În literatura de specialitate, termenul de etichetare cu sensul dat de noi este adesea înlocuit cu cel de adnotare, iar în locul celui de identificare se foloseşte cel de etichetare. Preferăm însă folosirea sensului extins al termenului de etichetare, cel de segmentare şi identificare, deoarece respectăm astfel definiţia etichetării dată la începutul capitolului. În plus, fonemele dintr-o pronunţie pot fi identificate şi prin transcrierea ei, ceea ce nu înseamnă că a fost realizată o etichetare, deşi a avut loc o adnotare a semnalului.

Automatizarea totală a etichetării ar presupune deci atât segmentarea semnalului, cât şi identificarea (recunoaşterea) fonemului reprezentat de fiecare segment. Evident, astfel pusă, problema este mai dificilă chiar şi decât recunoaşterea vorbirii şi nu are soluţii cunoscute. În practică, automatizarea etichetării poate fi facilitată simplificând sau eliminând **problema identificării**:

- **eliminarea** consta în transcrierea de către experţi umani a fiecărei pronunţii ce se doreşte a fi etichetată în termenii setului de simboluri ales, singura incertitudine care rămâne fiind cea cauzată de subiectivismul inerent procedurii [266];
- **simplificarea** ei se poate face prin reducerea numărului de alternative dintre care trebuie făcută identificarea unui anumit segment prin utilizarea unei reprezentări de tip reţea a unor posibile variante de pronunţie [144], [126], [254]; o asemenea reprezentare poate fi obţinută plecând de la o transcriere ortografică (existentă dacă semnalul a rezultat din citirea unor texte) prin prelucrare cu componenta de traducere grafeme-foneme dintr-un sistem de conversie text-vorbire, urmată de adăugarea unor variante de pronunţie specificate de reguli fonologice.

În principiu, multe dintre înregistrările din baza de date fiind însoţite de transcrieri ortografice, am fi putut încerca o automatizare completă a etichetării apelând la a doua variantă de eliminare a problemei identificării prin utilizarea unui sistem de conversie text-vorbire deja existent [185], [65]. Pentru o etichetare de calitate însă, reprezentarea fonemică astfel generată ar fi trebuit îmbogăţită cu variante de pronunţie prin aplicarea unor reguli fonologice, indisponibile la acel moment pentru limba română.

Ţinând cont de această situaţie, pentru a asigura o calitate maximă a etichetării am ales varianta eliminării problemei identificării prin transcrierea manuală a semnalelor de etichetat [266] în termenii unităţilor fonetice din tabelul 4.1. În ceea ce priveşte segmentarea, am optat pentru realizarea ei folosind modele Markov ascunse datorită posibilităţii de a le utiliza în continuare şi pentru recunoaşterea automată a vorbirii.

5.2.1 Evaluarea etichetării automate

Ca şi în cazul evaluării sistemelor de recunoaştere a vorbirii (secţiunea 3.1), şi ieşirile sistemelor de etichetare automată pot fi privite ca ipoteze referitoare la identităţile şi limitele sunetelor pronunţate în fişierele de semnal cărora le sunt asociate.

Spre deosebire de sistemele de recunoaştere, în cazul sistemelor de etichetare, pe lângă identităţile sunetelor, care pot fi comparate prin acelaşi algoritm de programare dinamică (secţiunea 3.1.1), în cazul în care identităţile coincid trebuie luate în consideraţie şi diferenţele dintre limitele de referinţă ale sunetelor şi cele generate automat.

În plus, dacă referinţele folosite pentru evaluarea sistemelor de recunoaştere sunt în general simplu de obţinut prin transcrierea ortografică a fişierelor de semnal, nu aceeaşi este situaţia referinţelor pentru evaluarea sistemelor de etichetare: aşa cum am menţionat deja, funcţie de nivelul ales pentru etichetare şi experienţa experţilor implicaţi, între etichetările realizate de diferiţi experţi pot apare diferenţe [67]. Ca atare, şi rezultatele evaluării pot diferi funcţie de referinţele utilizate, o soluţie propusă pentru atenuarea acestei probleme fiind cea a comparaţiilor cu etichetări manuale multiple [252].

Corespunzător celor două probleme – segmentarea şi identificarea – implicate de etichetare, evaluarea sistemelor de etichetare poate fi realizată prin intermediul a două categorii de metrice:

- un grup de metrice caracterizând performanţele de identificare – aceleaşi folosite şi pentru evaluarea performanţelor sistemelor de recunoaştere automată a vorbirii (secţiunea 3.1); funcţie de modul de tratare a problemei identificării (secţiunea 5.2), acestea vor reprezenta măsuri fie ale diferenţelor de transcriere între experţi umani, fie ale performanţelor combinate ale algoritmilor de traducere grafeme-foneme, de aplicare a eventualelor reguli fonologice şi de etichetare propriu-zisă;

- metrice care descriu performanţele segmentării semnalelor vocale, obţinute pe baza diferenţelor dintre limitele stabilite automat şi cele de referinţă; aceste diferenţe sunt calculate doar dacă identităţile segmentelor comparate coincid, iar metricele derivate includ atât caracteristici globale ale distribuţiilor lor (media, mediana, abaterea standard etc. ale diferenţelor sau valorilor lor absolute) cât şi unele legate de anumite obiective de performanţă, de tipul "% diferenţe între anumite limite."

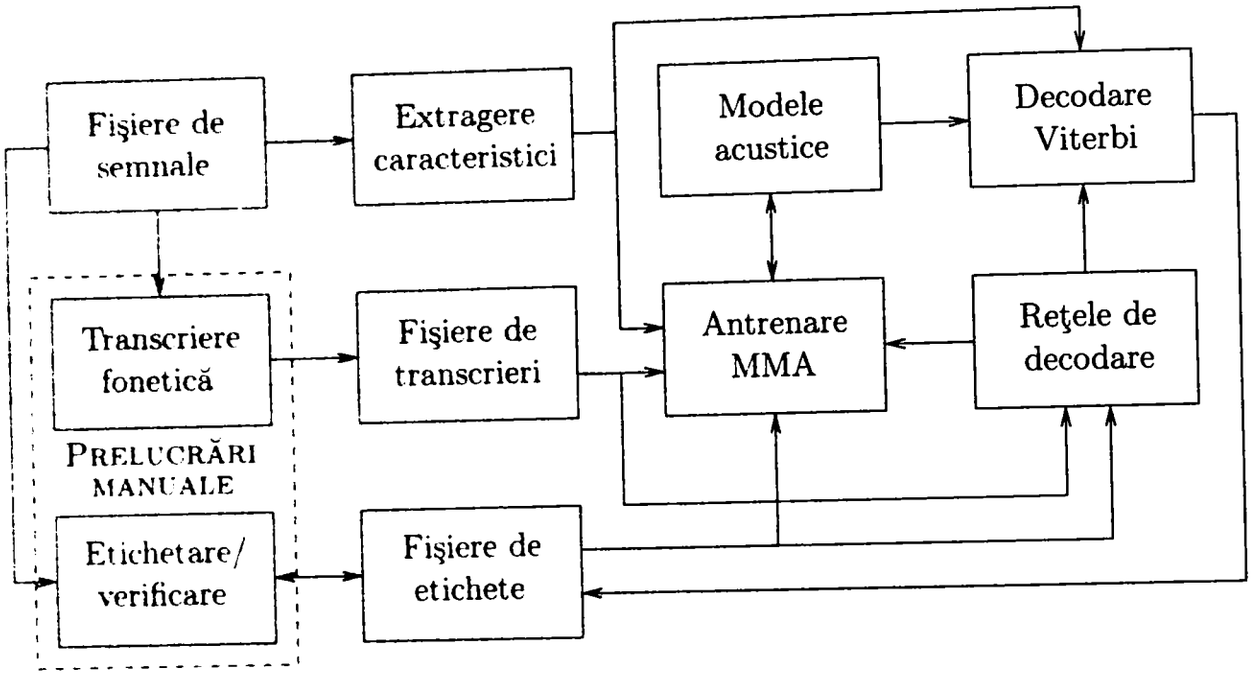

Figura 5.2: Schema bloc a sistemului de etichetare

5.3 Sistemul de etichetare

Aşa cum am precizat în secţiunea 5.2, pentru etichetarea semnalelor din baza de date am ales varianta unui sistem de etichetare în care problema identificării a fost eliminată prin transcrierea manuală a semnalelor, iar segmentarea a fost automatizată prin utilizarea modelelor Markov ascunse.

Automatizarea segmentării semnalelor vocale ar putea fi realizată foarte simplu prin utilizarea algoritmului Viterbi (secţiunea 3.9) pentru decodarea unor MMA compuse din modele acustice ale segmentelor conform transcrierilor semnalelor în termenii acestor unităţi, cu condiţia existenţei prealabile a modelelor acustice. În cazul nostru, inexistenţa unor modele ale unităţilor fonetice din tabelul 4.1 a făcut imposibilă această abordare, iar soluţia a constat într-un sistem dezvoltat şi utilizat în două etape (figura 5.2).

Într-o primă etapă, de dezvoltare a sistemului, a avut loc antrenarea unor modele acustice pentru unităţile din tabelul 4.1, simultan cu segmentarea automată a semnalelor utilizate în acest scop şi alinierea la ele a transcrierilor lor fonetice. În etapa a doua au fost prelucrate alte semnale, neutilizate în prima, folosind modelele astfel antrenate.

5.3.1 Etichetarea manuală

Pentru cât mai buna iniţializare a funcţiilor de probabilitate ale modelelor acustice, construcţia acestora a fost precedată de etichetarea manuală (figura 5.3) a celor 400 de propoziţii de iniţializare (secţiunea 4.3.2), cu o durată de circa 33'22", pe baza vizualizării formelor de undă şi spectrogramelor de bandă largă ale semnalelor şi a ascultării lor

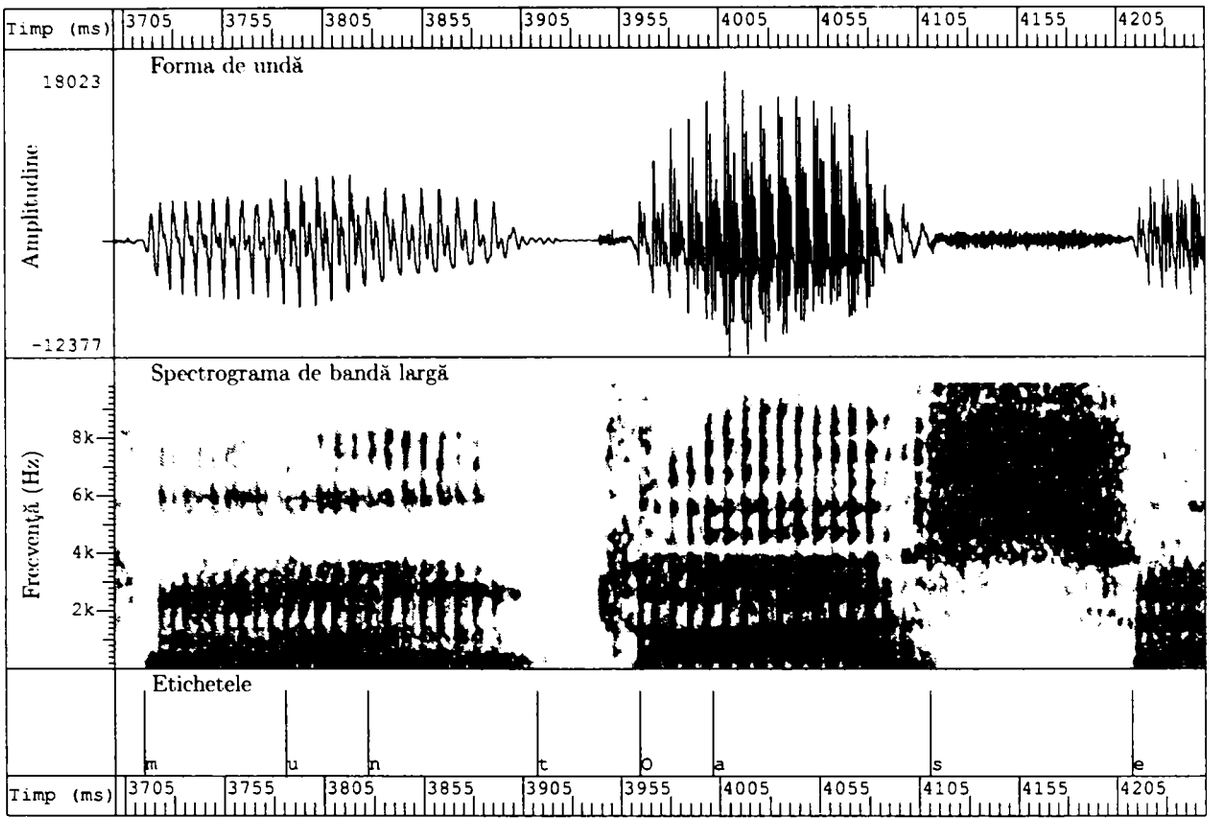

Figura 5.3: Exemplu de realizare manuală a etichetării şi verificării

selective folosind pachetul de programe SFS (Speech Filing System) [107].

Etichetarea manuală a fost făcută la nivel fonetic extins, ţinând cont de fenomenele specifice vorbirii fluente (secţiunea 5.4.1), de corelaţiile generale ce se pot stabili între proprietăţile acustice şi cele articulatorii ale sunetelor vorbirii [69], [186], precum şi de experienţa acumulată în cadrul altor cercetări [266], [81], [17], [133], [55]. Dat fiind însă numărul mic de propoziţii distincte implicate, variabilitatea problemelor apărute a fost limitată, iar unele dintre criteriile de decizie folosite pentru a le trata au fost reevaluate pe parcurs, astfel încât prezentarea lor unitară va fi făcută în secţiunea 5.4.

5.3.2 Transcrierea fonetică

Pe lângă cele 400 de propoziţii de iniţializare etichetate manual, în prima etapă au mai fost utilizate pasajele şi propoziţiile de completare înregistrate de către toţi vorbitorii în prima sesiune (2230 de propoziţi cu o durată totală de circa 2h56'12"), dar singura lor prelucrare manuală a fost transcrierea fonetică în termenii unităţilor din tabelul 4.1, ţinând cont de fenomenele specifice vorbirii fluente (secţiunea 5.4.1).

Deoarece fiecare pasaj sau propoziţie de completare au fost repetate de câte zece ori, pentru accelerarea transcrierii s-a plecat de la transcrieri-prototip, obţinute din cele ortografice şi adaptate pe baza ascultării semnalelor şi a examinării formelor lor de undă.

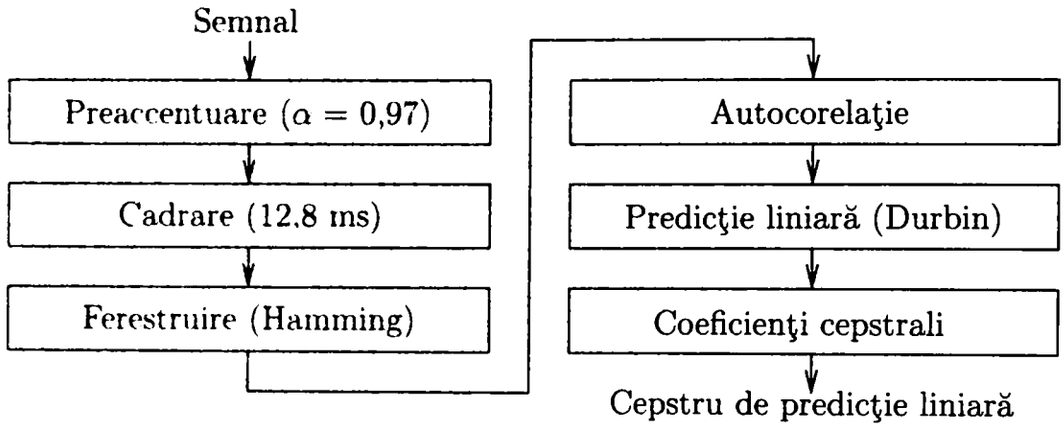

Figura 5.4: Analiza cepstrală prin predicţie liniară

În etapa a doua au fost prelucrate alte 1450 de propoziţii cu o durată totală de circa 1h54'26": pasajele şi propoziţiile de completare înregistrate în câte patru sesiuni suplimentare de cei 10 vorbitori din mulţimea FT (892 de propoziţii cu o durată totală de circa 1h11'23") şi cele 558 de propoziţii individuale (circa 43'3"). Pentru pasaje şi propoziţiile de completare a fost utilizată aceeaşi adaptare a transcrierilor-prototip, iar propoziţiile individuale, dată fiind unicitatea înregistrărilor lor, au fost transcrise separat.

5.3.3 Extragerea caracteristicilor

Deoarece este de dorit ca etichetarea să fie realizată cu o precizie temporală cât mai bună, extragerea caracteristicilor a fost făcută la nivelul unor cadre mai scurte şi mai frecvente decât cele utilizate în mod curent pentru recunoaşterea vorbirii, cu o durată de 12,8 ms (256 eşantioane) şi o deplasare de 5 ms între cadre (figura 5.4).

După preaccentuarea semnalelor (secţiunea 2.2.4) cu un coeficient $\alpha = 0{,}97$ şi cadrare, fiecare cadru a fost ferestruit cu o fereastră Hamming (secţiunea 2.2.1). În continuare, din fiecare cadru au fost calculate log-energia (secţiunea 2.2.2) şi, prin metoda autocorelaţiei (secţiunea 2.3.1), un predictor liniar de ordinul 12. Pe baza ecuaţiei (2.55), coeficienţii de predicţie liniară au fost convertiţi în coeficienţii cepstrali $c_{1\ldots12}$.

Coeficienţii cepstrali au fost supuşi unei liftrări conform ecuaţiei (3.18) cu lungimea $L = 12$ şi, împreună cu log-energia, unui proces de derivare peste intervale de 40 ms (ecuaţia 2.60 cu $L = 4$), aproximat prin diferenţe simple la capetele fişierelor. Fişierele de semnal au fost astfel transformate în secvenţe de vectori acustici 26-dimensionali, incluzând fiecare 12 coeficienţi cepstrali, log-energia şi coeficienţii lor Δ (secţiunea 2.7.2).

5.3.4 Modelele acustice

În vederea segmentării automate a semnalelor transcrise au fost construite şi utilizate modele Markov ascunse ale unităţilor fonetice din tabelul 4.1. Structura modelelor a fost

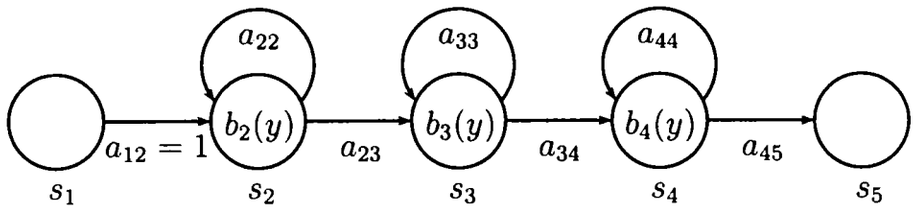

Figura 5.5: Structura MMA ale unităţilor fonetice

una de tip stânga-dreapta (figura 5.5) cu trei stări emiţătoare şi două stări (iniţială şi finală) utilizate doar pentru concatenarea lor în modele ale pasajelor şi propoziţiilor.

Corespunzător vectorilor acustici 26-dimensionali extraşi din semnale, funcţiile de probabilitate $b_j(y), j = 2 \ldots 4$, au fost mixturi de densităţi gaussiene 26-dimensionale cu matrice de covarianţă diagonale (secţiunea 3.7). Stărilor iniţială şi finală, s_1 şi s_5, ale fiecărui model, destinate doar interconectării lor, nu le-au fost asociate funcţii de probabilitate, iar stările emiţătoare $s_2 \ldots s_4$ au avut rolul de a modela porţiunile iniţială, medie şi respectiv finală ale fiecărei unităţi. Ţinând cont de deplasarea de 5 ms a cadrelor, modelele corespund astfel unei durate minime de 15 ms a unei unităţi fonetice.

Datorită introducerii în MMA a stărilor iniţiale şi finale neemiţătoare, reestimarea probabilităţilor de tranziţie (ecuaţia 3.37) trebuie modificată în cazul acestor stări: în mod evident, datorită structurii alese, probabilitatea $a_{12} = 1$, iar formula de reestimare a probabilităţilor de tranziţie în ultima stare devine

$$\hat{a}_{45} = \frac{\sum_{r=1}^{R} \gamma_{4,r}(T_r)}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{4,r}(t)} \tag{5.1}$$

Pentru o calitate cât mai bună a segmentării, modelele acustice au fost construite în două variante, una pentru fiecare sex, dar procesul de construcţie a fost identic şi a constat din două faze: una de iniţializare, folosind propoziţiile etichetate manual, şi una de reestimare utilizând semnalele transcrise fonetic în prima etapă, de dezvoltare a sistemului de etichetare.

Iniţializarea

Câte 200 de propoziţii de iniţializare înregistrate de vorbitorii de acelaşi sex şi etichetate manual au fost folosite pentru iniţializarea modelelor acustice ale unităţilor fonetice. Pentru fiecare model, toate apariţiile unităţii asociate au fost divizate în trei segmente egale, corespunzătoare celor trei stări emiţătoare $s_2 \ldots s_4$, rezultând astfel o divizare în submulţimi disjuncte a vectorilor acustici extraşi din aceste propoziţii.

Pe baza vectorilor acustici corespunzători unei stări s_j au fost iniţializaţi parametrii densităţilor sale gaussiene: pentru a obţine K gaussiene, vectorii au fost grupaţi folosind un algoritm de tip K-medii [6], iar fiecărui grup k i-a fost asociată o gaussiană de medie

μ_{jk}, matrice de covarianţă diagonală C_{jk} şi pondere w_{jk} egale cu cele ale grupului

$$\mu_{jk} = \frac{\sum_{i=1}^{N_{jk}} o_{ijk}}{N_{jk}} \tag{5.2}$$

$$C_{jk}[l] = \frac{\sum_{i=1}^{N_{jk}} (o_{ijk}[l] - \mu_{jk}[l])^2}{N_{jk}}, \quad l = 1 \ldots 26 \tag{5.3}$$

$$w_{jk} = \frac{N_{jk}}{\sum_{k=1}^{K} N_{jk}} \tag{5.4}$$

unde $o_{ijk}, i = 1 \ldots N_{jk}$ sunt cei N_{jk} vectori acustici din grupul k corespunzător stării s_j, iar $C_{jk}[l]$ - componentele de pe diagonala matricei sale de covarianţă.

Folosind, pe lângă densităţile gaussiene astfel iniţializate, probabilităţi de tranziţie atribuite manual, datele au fost resegmentate prin algoritmul Viterbi (secţiunea 3.9) iar parametrii gaussienelor au fost reactualizaţi conform metodei de mai sus. După câteva iteraţii ale acestui proces, iniţializarea a fost încheiată cu algoritmul Baum-Welch (secţiunile 3.5.1 şi 3.7.1), care a permis şi reestimarea probabilităţilor de tranziţie.

Reestimarea concatenată

Iniţializarea modelelor acustice s-a făcut separat pentru fiecare unitate fonetică, neţinând cont că realizările lor nu apar separat, ci doar interconectate. Pentru a lua în calcul şi acest aspect s-a utilizat algoritmul Baum-Welch concatenat [135], [136], care operează asupra unor modele compuse ale propoziţiilor sau pasajelor folosite pentru antrenament, formate prin concatenarea de instanţe ale modelelor unităţilor fonetice în conformitate cu reţele de decodare obţinute din transcrierile sau etichetele asociate.

Pentru reestimarea parametrilor modelului unei unităţi fonetice, algoritmul Baum-Welch concatenat utilizează tot formulele din varianta de bază, dar sumele sunt calculate peste toate instanţele lui din modelele compuse ale înregistrărilor prelucrate.

În plus, în cazul unei structuri de tipul celei din figura 5.5 a modelelor unităţilor fonetice, concatenarea lor duce la modificarea formulei de reestimare a probabilităţilor de tranziţie în stările finale: cu excepţia ultimei instanţe dintr-un model compus, pentru care se utilizează formula (5.1), probabilităţile reestimate ale tranziţiilor finale devin

$$\hat{a}_{45} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r - 1} \gamma_{4,r} a_{45}}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{4,r}} \tag{5.5}$$

Formulele de reestimare, cu sumele de la numitori şi numărători calculate conform celor de mai sus, sunt aplicate în paralel pentru toate modelele unităţilor fonetice, astfel încât parametrii acestora sunt reestimaţi simultan.

5.3.5 Segmentarea automată

După iniţializarea şi reestimarea concatenată a modelelor unităţilor fonetice, fişierele de semnal transcrise fonetic au fost segmentate prin decodarea Viterbi a modelelor lor compuse, având ca observaţii vectorii acustici extraşi din acele fişiere.

În etapa de dezvoltare a sistemului, procesul de antrenare a modelelor şi segmentare a semnalelor a fost reluat de câteva ori, folosind însă pentru iniţializarea modelelor fişierele de semnal transcrise fonetic şi etichetele lor generate automat în pasul anterior.

Etichetarea finală a fost realizată utilizând modele acustice cu mixturi de patru gaussiene pe stare, rezultate în urma a şase iteraţii de segmentare/reantrenare.

5.3.6 Verificarea etichetării

Segmentările semnalelor şi alinierile la acestea ale transcrierilor lor fonetice produse în mod automat de sistemul de etichetare sunt în multe cazuri foarte bune. Există însă şi destul de multe cazuri, inevitabile pentru orice sistem automat, în care ele suferă de erori de poziţionare, sau cazuri, mai puţin frecvente, în care transcrierile fonetice au fost din start afectate de erori. Ţinând cont de existenţa acestor erori, pentru a asigura o calitate cât mai bună a etichetării şi a face posibilă o evaluare a sistemului de etichetare, a fost necesară verificarea şi, acolo unde a fost cazul, corectarea manuală a etichetelor generate automat, realizate într-un mod asemănător etichetării manuale (figura 5.3).

Verificarea etichetării presupune însă definirea unor criterii de decizie asupra celor două procese componente – segmentarea semnalului vocal şi identificarea segmentelor. Deoarece aceste criterii au fost definite şi rafinate în mod iterativ pe parcursul cercetărilor legate de etichetarea semnalului vocal, pentru a asigura consistenţa globală a etichetării, verificarea a inclus şi rezultatele etichetării manuale a propoziţiilor de iniţializare.

5.4 Criteriile de decizie

Datorită coarticulaţiei sunetelor vorbirii, un fonem poate fi semnalat printr-o serie de indicii distribuite în mai multe segmente de semnal vocal, şi invers, proprietăţile unui segment pot fi determinate de mai multe foneme succesive [69], [265]. Aceasta face ca localizarea realizărilor fonemelor, urmărită prin etichetarea la nivelul fonetic extins, să nu fie întotdeauna uşoară, iar în unele cazuri nici măcar posibilă, astfel încât desemnarea unui segment ca realizare a unui fonem se face într-o oarecare măsură arbitrar, pe baza indiciilor fonemice principale considerate a fi conţinute în acel segment [18].

Indiciile fonemice putând fi distribuite de-a lungul mai multor segmente succesive, etichetarea şi verificarea manuală au acordat o pondere redusă percepţiei auditorii a semnalelor în deciziile asupra delimitării segmentelor asociate fonemelor percepute ca realizate. Delimitarea a fost astfel bazată în primul rând pe caracteristicile acustice ale semnalelor (aspectul formelor de undă şi al spectrogramelor), care pot fi cel mai adesea corelate cu mişcările articulatorii efectuate pentru realizarea fonemelor.

La rândul lor, formelor de undă, ca reprezentări primare ale semnalelor vocale, le-a fost acordată uneori o pondere superioară spectrogramelor, care au dezavantajul unei rezoluţii mai reduse în timp datorită cadrării semnalelor în cursul generării lor.

Corelaţiile dintre proprietăţile articulatorii şi cele acustice ale sunetelor vorbirii [69], [186] sunt suficient de puternice pentru ca experţi umani să poată reconstitui conţinutul unor pronunţii pe baza examinării spectrogramelor lor [265], [157], ele putând fi cu atât mai mult folosite pentru decizii asupra etichetării, când semnalele pot fi şi ascultate.

Tabelul 5.1: Clasificarea consoanelor limbii române după criteriile manierei şi locului de articulare (simboluri ASCII conform tabelului 4.1)

Maniera de articulare	**Locul de articulare**					
	Bilabiale	Labio-dentale	Dentale	Postal-veolare	Palatale	Velare
Plozive	p *b*		t *d*			k *g*
Fricative		f *v*	s *z*	S *J*		h
Africate			T	C *G*		
Nazale	m		n			
Laterale			l			
Vibrante			r			
Semivocale	w O				j E	

Tabelul 5.2: Clasificarea vocalelor limbii române după gradul de deschidere şi locul de articulare (simboluri ASCII conform tabelului 4.1)

Gradul de deschidere	**Locul de articulare**		
	Anterioare	Centrale	Posterioare
Închise	i	y	u
Medii	e	@	o
Deschise		a	

Caracterizarea articulatorie a sunetelor limbii române [234] poate fi făcută în cazul consoanelor prin locul şi maniera de articulare (tabelul 5.1), iar în cel al vocalelor – prin locul de articulare şi gradul de deschidere a cavităţii bucale (tabelul 5.2). În plus, consoanele pot fi distinse şi după sonoritate – în tabelul 5.1, aceasta diferenţiază perechile de sunete cu aceleaşi locuri de articulare din primele trei linii, cele sonore fiind *evidenţiate.*

Distincţia sonor/nesonor nu este relevantă pentru consoanele sonante (nazale, laterale şi vibrante) şi semivocale, care sunt toate sonore. Semivocalele au fost incluse împreună cu consoanele datorită diferenţelor fonetice (sunt sunete tranzitorii) şi fonologice (nu pot forma nucleul unei silabe) faţă de vocale. Distincţia în cadrul perechilor de semivocale cu acelaşi loc de articulare se poate face însă similar vocalelor, după gradul de deschidere.

Maniera de articulare, sonoritatea şi locul de articulare pot fi ierarhizate în această ordine din punctul de vedere al discriminării consoanelor [186]. Dintre acestea, maniera de articulare este cea mai strâns corelată cu proprietăţile acustice ale semnalelor, fiind din această cauză baza cea mai potrivită pentru deciziile asupra segmentării [69].

Locul de articulare, relevant şi pentru vocale, are drept corelate acustice valorile frecvenţelor formanţilor, determinate de volumele cavităţilor componente ale tractului vocal: F_1 este determinată în principal de volumul faringelui, iar F_2 – de cel al cavităţii bucale. Având o evoluţie în general lentă în timp (v. figurile 2.3 şi 5.1), corelatele locului de articulare sunt utile în special pentru deciziile asupra identităţii segmentelor.

Sonoritatea este indicată cu maximum de fiabilitate de prezenţa în spectrogramă a primului formant, cauzat de rezonanţa faringelui, deoarece vibraţiile coardelor vocale, specifice sunetelor sonore, sunt singura sursă posibilă de excitaţie a faringelui.

Criteriile de decizie, grupate în continuare pe categorii, au fost formulate plecând de la obiectivele cercetărilor proprii şi caracteristicile generale ale sunetelor vorbirii [69], [186] şi ţinând cont de rezultatele altor proiecte similare [266], [81], [17], [133], [55] şi experienţa acumulată pe parcursul acestor cercetări. În măsura în care acest lucru este posibil, ele sunt definite cu referire la proprietăţile acustice şi articulatorii ale sunetelor limbii române, dar există şi cazuri ambigue în care lipsa unor repere clare a impus introducerea unor reguli pentru rezolvarea ambiguităţilor şi asigurarea consistenţei.

5.4.1 Fenomenele specifice vorbirii fluente

Pe durata etichetării manuale, a transcrierii fonetice şi verificării etichetării, opţiunea pentru nivelul fonetic extins (secţiunea 5.1) a condus la luarea în consideraţie a unor fenomene specifice vorbirii fluente: asimilarea, eliziunea şi epenteza.

Asimilarea, constând în transformarea unor sunete sub influenţa celor adiacente (de exemplu, "exemplu" pronunţat [egzemplu]), a fost marcată prin atribuirea identităţilor corespunzătoare realizărilor efective ale segmentelor acustice implicate.

Eliziunea, sau omiterea segmentului corespunzător unui fonem, apare în vorbirea fluentă chiar şi la viteze moderate. În cazul eliziunii complete, chiar dacă un fonem apare în pronunţia "standard" a unui cuvânt, el nu a fost etichetat, neexistând un segment cu care să poată fi asociat. Uneori, contrastul fonemic este realizat prin intermediul altor trăsături distinctive, iar fonemul respectiv poate fi perceput ca pronunţat, chiar dacă la examinarea semnalului nu putem localiza un segment speific: în asemenea cazuri au fost etichetate componentele acustice efectiv apărute – de exemplu, nazalizarea unei vocale percepută ca realizare a unei consoane nazale (secţiunea 5.4.5).

Epenteza (introducerea unor segmente acustice suplimentare faţă de pronunţiile "standard") se manifestă în special sub forma unor semivocale (de exemplu, /j/ la începutul cuvintelor "ei", "ele"), dar şi ca pauze în interiorul cuvintelor, datorate lipsei de sincronizare a mişcărilor articulatorii, şi a fost marcată ca atare în etichetare.

5.4.2 Vocalele şi semivocalele

Datorită producerii lor fără obstrucţii sau constricţii ale tractului vocal, vocalele sunt caracterizate prin structură formantică şi, de obicei, sonoritate. Caracteristicile articulatorii şi cele acustice sunt legate în acest caz prin două corelaţii majore: pe de o parte, cea a locului de articulare cu caracterul grav sau acut al vocalei, indicat de valoarea frecvenţei formantului al doilea, F_2 – înaltă (acută) pentru vocalele anterioare, respectiv joasă (gravă) pentru cele posterioare; pe de altă parte, cea dintre gradul de deschidere şi caracterul difuz sau compact al spectrului, indicat de diferenţa $F_2 - F_1$ – mare (spectru difuz) pentru vocalele închise, mică (spectru compact) pentru cele deschise. Valorile frecvenţelor formanţilor trebuie însă interpretate relativ la ansamblul unei pronunţii, ţinând cont de evoluţiile lor în timp şi variaţiile dintre vorbitori.

Deoarece F_2 este determinată de rezonanţa cavităţii bucale, al cărui volum poate varia cel mai mult de-a lungul tractului vocal, formantul al doilea are cea mai mare dinamică, iar în lipsa altor indicii, delimitarea se poate face la mijlocul tranziţiei sale.

Secvenţele vocală-vocală sau semivocală-vocală sunt în general cele mai dificil de delimitat: dacă cele două componente nu sunt similare din punctul de vedere al locului de articulare, delimitarea se face la mijlocul tranziţiei formantice, în caz contrar se utilizează indiciile furnizate de schimbările de amplitudine ale formei de undă şi variaţiile de energie din spectrogramă. În lipsa altor indicii, o secvenţă semivocală-vocală a fost în general divizată în raportul o treime din durată – semivocala, două treimi – vocala.

Limitele în raport cu consoanele plozive, fricative sau africate sunt de obicei clare datorită manierei de articulare a acestora, iar cele faţă de consoanele sonante corespund creşterii intensităţii formanţilor şi amplitudinii şi/sau complexităţii formei de undă.

5.4.3 Consoanele plozive

Pe durata producerii acestor consoane pot apare trei evenimente acustice distincte: liniştea, corespunzătoare închiderii articulatorilor, explozia, produsă la eliberarea lor, şi aspiraţia de după explozie. Explozia este formată din impulsuri depăşind amplitudinea oricărui zgomot de fricţiune ulterior, corespunzător aspiraţiei, iar în cazul în care aceasta din urmă lipseşte, poate fi identificată ca un impuls în forma de undă, de amplitudine mult mai mică decât a sunetului sonor următor. Aspiraţia este semnalul aperiodic de după explozie, al cărui sfârşit este marcat de începutul vibraţiilor coardelor vocale pentru următorul sunet sonor. Limitele consoanelor plozive corespund deci începutului închiderii articulatorilor şi sfârşitului exploziei sau eventualei aspiraţii.

Când o plozivă este precedată de un sunet fricativ sau sonor, începutul închiderii coincide cu scăderea energiei sunetului anterior la toate frecvenţele, sau peste cca. 500 Hz, dacă sonoritatea continuă pe durata închiderii şi este vizibilă în spectrogramă.

După o pauză, limita iniţială a unei plozive sonore este poziţionată la începutul vibraţiilor coardelor vocale. Dacă vibraţiile nu sunt vizibile nici în forma de undă, nici în spectrogramă, această limită este plasată cu cca. 50 ms înaintea exploziei.

Dacă după o pauză apare o închidere nesonoră, începutul ei este de obicei marcat printr-un mic impuls în forma de undă şi/sau spectrogramă, iar limita iniţială a consoanei plozive asociate este plasată la acest moment. În lipsa acestor indicii, ca şi în cazul anterior, limita iniţială este plasată cu cca. 50 ms înaintea exploziei.

Dacă o plozivă nesonoră este precedată de o vocală, sonoritatea este adesea prelungită pe durata închiderii, limita dintre vocală şi plozivă fiind în acest caz plasată pe baza spectrogramei, în punctul de dispariţie a formanţilor.

O consoană plozivă urmată de una fricativă sau nazală nu are de obicei o explozie identificabilă, caz în care ea include numai o închidere, iar limita ei finală este plasată acolo unde apare o creştere marcată a energiei la frecvenţe de peste 500 Hz.

Limita dreaptă a unei consoane plozive finale este plasată la sfârşitul exploziei, iar în lipsa acestei - la cca. 50 ms de la începutul consoanei.

Într-o secvenţă de două consoane plozive, dacă nu există indicii ale exploziei primei consoane, închiderea este împărţită în mod egal între cele două.

5.4.4 Consoanele fricative şi africate

Datorită intensităţii zgomotului de înaltă frecvenţă care le caracterizează, /s/, /z/, /S/ şi /J/ sunt cel mai uşor de identificat dintre consoanele fricative, începutul lor putând fi determinat pe baza creşterii energiei acestui zgomot, vizibilă în spectrogramă.

Celelalte consoane fricative sunt semnalate de asemeni prin zgomot în spectrogramă, dar vizibilitatea acestuia poate fi foarte redusă; ele sunt însă caracterizate de scăderi ale energiei formanţilor în raport cu sunetele alăturate şi indicii ale fricţiunii care permit delimitarea lor. Dacă nici unele dintre aceste indicii nu sunt vizibile, limitele vor fi plasate prin excludere, acolo unde nu există în mod clar altceva.

Fricativele adiacente sunt separate pe baza diferenţelor de intensitate şi frecvenţă inferioară a zgomotului din spectrogramă, provocate de schimbarea locului de articulare, iar în cazul identităţii lor – prin divizarea în părţi egale a segmentului fricativ.

Dacă o fricativă este însoţită de o închidere produsă prin apropierea articulatorilor mai mult decât necesarul pentru fricţiune, închiderea este inclusă în fricativă, iar dacă între un sunet sonor şi o fricativă apare o scurtă pauză, aceasta este atribuită fricativei.

Deoarece o consoană africată este realizată asemănător închiderii unei plozive, urmată de o fricţiune homorganică înlocuind explozia, criteriile din acest caz sunt similare celor folosite pentru plozive şi fricative: limita iniţială este stabilită folosind regulile pentru închiderile plozivelor, iar cea finală – pe baza regulilor pentru fricative.

5.4.5 Consoanele sonante

Limitele închiderii orale de pe durata unei consoane nazale sunt marcate prin scăderea, relativă la segmentele adiacente, a energiei din spectrogramă peste frecvenţa de cca. 500 Hz. Amplitudinea semnalului variază de obicei în aceste puncte, ca şi cea a oscilaţiilor de înaltă frecvenţă, corespunzătoare formanţilor, suprapuse peste frecvenţa fundamentală, astfel încât pe durata nazalelor forma de undă este mai simplă, iar formanţii sunt mai slabi şi discontinui în raport cu cei ai vocalelor din jur.

Nazalele adiacente cu locuri de articulare diferite sunt distinse prin deplasări ale formanţilor, iar cele îngemănate pentru care nici forma de undă nici spectrograma nu oferă indicii asupra unei posibile delimitări se segmentează în părţi egale.

Eliziunea nazalelor apare frecvent în vorbirea fluentă, caz în care ele pot fi totuşi indicate prin nazalizarea vocalei precedente, iar dacă pot fi percepute, ele sunt marcate prin alocarea câtorva perioade fundamentale din aceste zone de nazalizare.

Laterala /l/ poate fi delimitată de vocale similar nazalelor, pe baza schimbărilor de amplitudine ale formei de undă şi scăderii energiei la frecvenţe înalte, dar formantul al doilea este considerabil mai puternic decât în cazul nazalelor. Când există tranziţii formantice, delimitările se fac la mijlocul acestora, ca şi în cazul (semi)vocalelor.

Vibranta /r/ are cea mai mare variabilitate a articulării şi, în mod corespunzător, a caracteristicilor acustice [130]. În cazul lui /r/ dental ("normal") apar scăderi şi chiar întreruperi ale fluxului de aer, care duc la scăderi ale amplitudinii formei de undă şi energiei din spectrogramă, iar delimitarea se face la ultima respectiv prima oscilaţie a coardelor vocale anterioară sau următoare acestor scăderi. Varianta "graseiată" poate fi delimitată pe baza reducerii amplitudinii formei de undă şi a energiei formanţilor.

5.4.6 Problema /I/

Comparând tabelul 4.1 cu tabelele 5.1 şi 5.2, se observă că /I/ nu apare printre sunetele descrise în acestea din urmă, literatura lingvistică mărginindu-se (atunci când îl menţionează) să-l descrie ca un alofon nesilabic al unei vocale sau semivocale (v. secţiunea 4.2), fără a detalia caracteristicile lui acustice sau articulatorii, şi mai ales fără dovezi experimentale convingătoare, ci mai mult citând părerile unor nume prestigioase.

Analiza datelor etichetate arată însă că acesta are o frecvenţă de realizare foarte redusă, sub 20% din cea aşteptată în urma proiectării bazei de date, iar examinarea semnalelor şi a spectrogramelor lor arată o foarte mare variabilitate a caracteristicilor lui acustice, mergând până la lipsa unor segmente care i-ar putea fi asociate.

Cu toate neclarităţile legate de existenţa realizărilor lui fizice şi proprietăţile lor acustice şi articulatorii, fonemul /I/ a fost inclus în transcrierile fonetice atunci când a fost perceput, iar în lipsa unor segmente care să-i poată fi puse în corespondenţă i-au fost alocate scurte porţiuni de semnal între fonemele adiacente.

5.5 Rezultate şi comentarii

Pentru dezvoltarea sistemului de etichetare descris în secţiunea 5.3 au fost etichetate manual cele 400 de propoziţii de iniţializare, iar pasajele şi propoziţiile de completare şi cele individuale au fost etichetate semiautomat, pe parcursul dezvoltării sistemului (pasajele şi propoziţiile de completare înregistrate de toţi vorbitorii în prima sesiune) şi prin utilizarea lui (propoziţiile individuale şi pasajele şi propoziţiile de completare înregistrate de vorbitorii din mulţimea FT în sesiunile suplimentare).

Au fost astfel etichetate toate cele 4080 de propoziţii, iar pentru asigurarea calităţii şi a consistenţei, toate fişierele de etichete au fost verificate şi corectate manual conform criteriilor din secţiunea 5.4. Aceleaşi criterii au fost folosite şi pentru a eticheta manual logatomii CVC, alfabetul şi informaţiile semispontane, care nu au fost prelucrate folosind sistemul de etichetare datorită simplităţii şi/sau caracteristicilor acustice diferite de cele ale propoziţiilor utilizate la construcţia acestuia. Nu au fost etichetate numerele.

Dat fiind modul de dezvoltare şi utilizare a sistemului de etichetare, evaluarea lui a fost făcută doar pe baza analizei etichetării celor 558 propoziţii individuale pentru a obţine o cât mai bună estimare a capacităţii lui de generalizare: aceste propoziţii nu au fost folosite pentru dezvoltarea sistemului, iar vocabularul lor, de peste 2500 de cuvinte, este mult diferit de cel de 1160 de cuvinte al pasajelor şi propoziţiilor de completare utilizate pentru antrenarea modelelor acustice ale sistemului de etichetare.

Evaluarea a fost făcută conform metodologiei din secţiunea 5.2.1 prin compararea fişierelor de etichete generate automat plecând de la transcrierile fonetice manuale cu cele verificate şi corectate manual. Metricele caracterizând performanţele de identificare reprezintă în acest caz o măsură a corectitudinii transcrierii fonetice manuale şi au, cum era de aşteptat, valori foarte bune: transcrierile iniţiale au cuprins 27553 de etichete, iar fişierele corectate - 27580, deci o diferenţă sub 0,1%. Alinierea lor (secţiunea 3.1.1) a identificat 37 substituţii (sub 0,14%), 66 omisiuni (sub 0,24%) şi 39 inserţii (sub 0,15%), corespunzătoare unei corectitudini de peste 99,62% şi unei acurateţi de peste 99,48%.

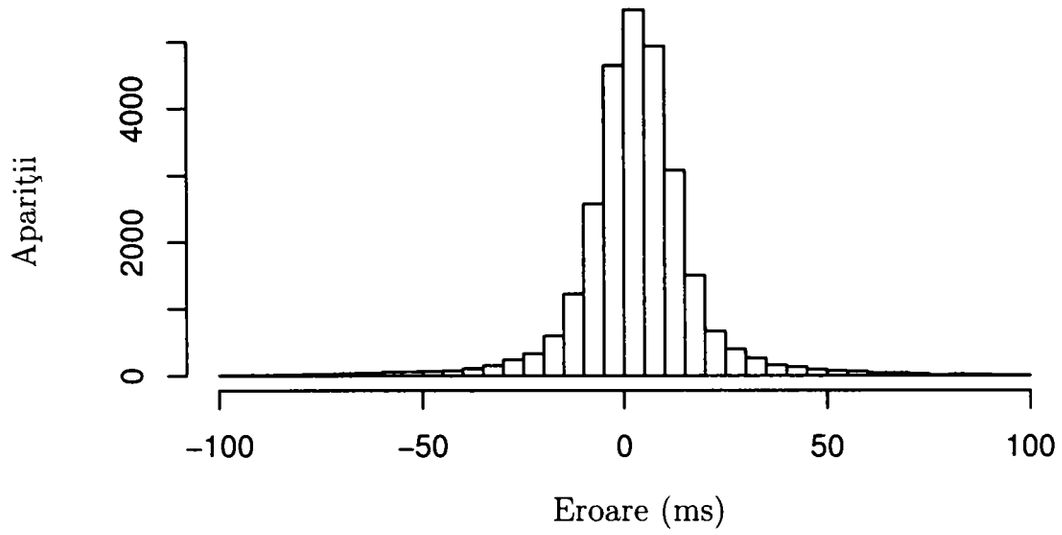

Figura 5.6: Histograma erorilor de segmentare

Analiza erorilor arată că substituţiile cele mai frecvente au fost datorate confuziei /j/-/J/ la tastare (de 7 ori), /I/ a fost fonemul cel mai frecvent inserat (de 10 ori), iar pauza /_/ – cea mai frecvent omisă. În rest, erorile pot fi atribuite incertitudinii inerente procesului de transcriere şi în foarte mică măsură altor confuzii la tastare (/s/-/S/).

Evaluarea performanţelor de segmentare ale sistemului a fost bazată pe analiza a 27241 de cazuri în care ambele etichete din jurul unei limite între segmente au fost corecte. Erorile de segmentare, calculate ca diferenţe între limitele generate de sistemul de etichetare şi cele stabilite în urma verificării şi corectării manuale, au fost cuprinse între −275 şi 226,002 ms, cu media de 3,027 ms, mediana de 3,102 ms şi abaterea standard de 17,278 ms. Erorile au fost mai mari de 100 ms în valoare absolută doar în 91 (0,33%) dintre cele 27241 de cazuri, iar figura 5.6 prezintă histograma lor între aceste limite: după cum se observă din cifrele de mai sus şi din histogramă, pe ansamblu sistemul a avut o uşoară tendinţă de întârziere a segmentării.

Pentru o apreciere mai exactă a performanţelor de segmentare ale sistemelor de etichetare, o metrică frecvent folosită este procentajul erorilor care se încadrează între anumite limite. În cazul sistemului prezentat aici, valorile acestei metrice pentru câteva limite sunt prezentate în tabelul 5.3: se observă că peste 80% din delimitările automate ale segmentelor au fost la maximum 15 ms de cele considerate corecte în urma verificării şi corectării manuale, iar peste 95% – la maximum 35 ms.

Tabelul 5.3: Procentajele erorilor de segmentare între anumite limite

Limite	±5 ms	±10 ms	±15 ms	±20 ms	±25 ms	±30 ms	±35 ms
%	37,68	64,87	80,51	88,09	91,70	93,98	95,50

Tabelul 5.4: Clasele de sunete folosite în analiza erorilor de segmentare

Clasa	Sunetele
A	T C G
F	f v s z S J h
I	I
P	p b t d k g _
S	m n l r j w
V	i e E y @ a u o O

O analiză mai detaliată a erorilor de segmentare a fost făcută prin considerarea câtorva clase de sunete pentru care au fost observate probleme similare (tabelul 5.4), iar rezultatele acestei analize, sub forma unor *boxplot*-uri ale distribuţiilor erorilor, sunt reprezentate funcţie de clasele sunetelor adiacente în figura 5.8.

Explicaţia reprezentărilor prin *boxplot*-uri din figura 5.8 poate fi urmărită în detaliu în figura 5.7: sfertul inferior este valoarea sub care sunt plasate cele mai mici 25% din valori; cel superior – valoarea peste care se află cele mai mari 25% din valori; minima şi maxima acceptabilă sunt la maximum 1,5 intervale între sferturi sub, respectiv peste sfertul adiacent; iar valorile aberante, din afara lor (*outliers*), sunt reprezentate individual.

Revenind la figura 5.8, se observă, pe baza intervalelor între sferturi, că erorile cele mai mari apar în cazul delimitării unor sunete din aceeaşi clasă, cele mai afectate fiind, în ordine, clasele P (consoanele plozive şi pauza), V (vocalele şi semivocalele /E/ şi /O/) şi S (consoanele sonante şi semivocalele /j/ şi /w/). Mecanismele de producere a erorilor diferă însă între clase: în cazul clasei P, principala cauză, indicată şi de deplasarea distribuţiei erorilor, este dificultatea de a detecta începutul unei noi închideri în lipsa unei explozii asociate celei anterioare; în cazul claselor V şi S, problemele îşi au originea în modificările lente ale caracteristicilor spectrale ale sunetelor din aceste clase, fără evenimente uşor detectabile, iar distribuţiile sunt aproape simetrice.

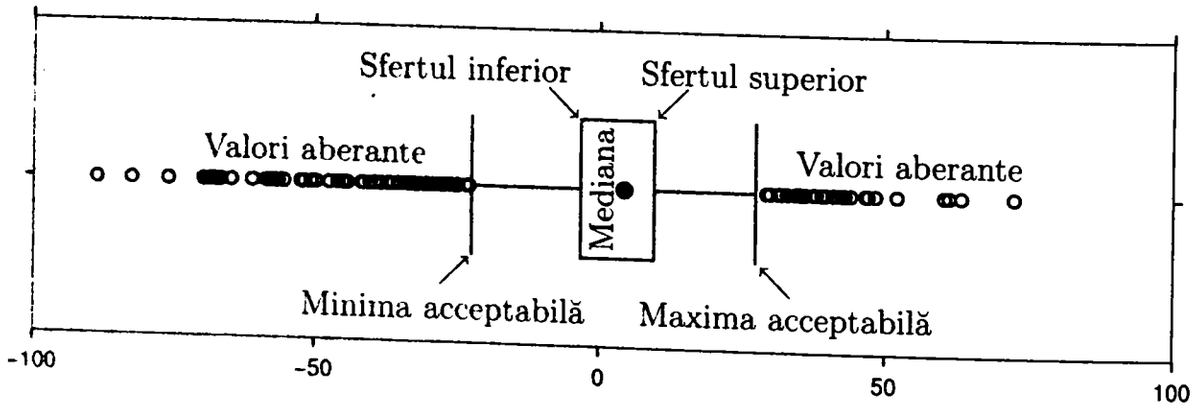

Figura 5.7: Reprezentarea prin *boxplot* a unei distribuţii

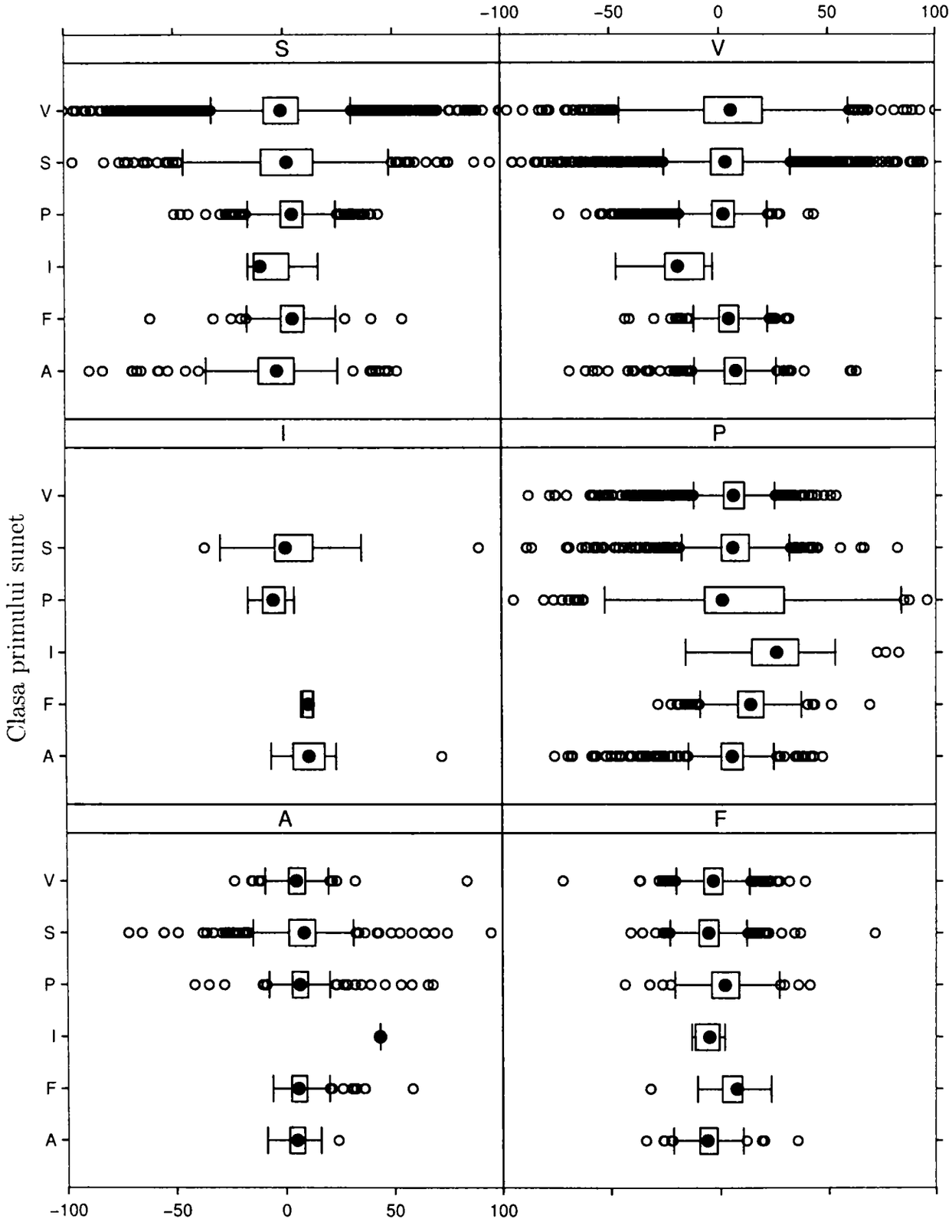

Figura 5.8: Distribuţiile erorilor de segmentare funcţie de clasele sunetelor adiacente (clasa celui de-al doilea sunet este titlul unuia dintre cele şase panouri)

5.6 Concluzii

Etichetarea semnalelor vocale, constând în definirea unor evenimente identificate prin coordonate temporale şi etichete, asigură baza mecanismelor de indexare şi regăsire rapidă a acestor evenimente în cadrul unei baze de date vocale.

Din punctul de vedere al cercetărilor în direcţia recunoaşterii automate a vorbirii, etichetarea corespunzătoare a unei asemenea baze de date permite localizarea aparţiiilor unor unităţi de modelare acustică şi utilizarea lor pentru antrenarea modelelor acustice, şi asigură transcrierile de referinţă necesare evaluării rezultatelor în experimentele de recunoaştere desfăşurate la nivelul acestor unităţi de modelare.

Etichetarea poate fi făcută la diferite niveluri, posibil divizate în straturi, iar datorită avantajelor pe care le prezintă din punctul de vedere al antrenării modelelor acustice şi al utilizării lor în cursul recunoaşterii, ca şi al etichetării în sine, pentru etichetarea bazei de date proiectate şi colectate în cursul acestor cercetări a fost ales nivelul fonetic extins. Acesta utilizează simboluri ale fonemelor unei limbi (în cazul nostru, ale unităţilor fonetice de modelare acustică) pentru a identifica realizările lor efective.

Efectuată manual, etichetarea poate fi foarte mare consumatoare de timp, astfel încât pentru automatizarea ei au fost încercate de-a lungul timpului diferite soluţii, iar pentru etichetarea bazei noastre de date am ales varianta unui sistem bazat pe modele Markov ascunse dependente de sex ale unităţilor fonetice de modelare acustică.

Acest sistem nu elimină însă total intervenţiile manuale, cum nu o face nici un alt sistem de etichetare, ele rămânând necesare fie pentru a eticheta materialele folosite în procesul de iniţializare a modelelor, fie pentru a transcrie fonetic înregistrările ce urmează a fi prelucrate, fie pentru a verifica şi corecta etichetele generate automat.

Pentru verificarea şi corectarea etichetelor generate automat, ca şi pentru etichetarea manuală, este necesară existenţa unui cadru de referinţă al acestor procese, concretizat prin criterii de decizie asupra delimitării şi identificării evenimentelor etichetate. Un asemenea set de criterii, bazat în general pe corelaţii între caracteristici articulatorii şi acustice ale sunetelor, a fost formulat şi folosit şi pentru aceste cercetări.

Dezvoltarea şi utilizarea sistemului de etichetare şi a criteriilor de decizie au permis etichetarea aproape totală a bazei de date descrise anterior, iar evaluarea sistemului prin compararea unor etichete generate de el cu cele verificate şi corectate a dovedit bunele performanţe de segmentare ale acestuia – peste 80% din limitele între segmente evaluate au fost la maximum 15 ms de poziţiile de referinţă, iar peste 95% – la maximum 35 ms.

Deşi efectuată în primul rând în scopul cercetărilor asupra recunoaşterii automate a vorbirii continue, sperăm ca etichetarea bazei de date să contribuie şi la dezvoltarea altor domenii – de exemplu, pe baza informaţiilor temporale incluse s-ar putea construi modele de ritm şi durată a sunetelor pentru conversia text-vorbire.

CAPITOLUL 6

Experimente de modelare acustică

Aşa cum am menţionat în secţiunea 3.7, modelarea acustică joacă un rol esenţial în sistemele automate de recunoaştere a vorbirii, de modul în care aceasta reuşeşte să acopere variabilitatea semnalelor vocale depinzând performanţele lor, iar experimentele descrise în acest capitol constituie prima abordare a problemelor modelării acustice folosind unităţi sublexicale pentru recunoaşterea vorbirii continue în limba română.

Variabilitatea semnalelor vocale are surse multiple, dintre care unele lingvistice, altele extralingvistice. O sursă de variabilitate de natură intrinsec lingvistică, şi pentru a cărei tratare alegerea unităţilor de modelare acustică poate fi determinantă, este coarticulaţia sunetelor vorbirii, menţionată deja în secţiunea 4.2 şi exemplificată în figura 5.1.

Printre sursele extralingvistice se numără mediile în care sunt produse semnalele şi canalele de comunicaţie prin care ele sunt transmise de la vorbitori la sistemele de recunoaştere: în cazul acestora, soluţia consta în utilizarea unor metode de extragere a caracteristicilor capabile să reducă efectele nedorite ale mediilor sau canalelor şi să ofere reprezentări spectrale robuste ale semnalelor. Acest tip de variabilitate nu a fost prezent în aceste cercetări deoarece semnalele folosite au fost colectate într-un mediu controlat prin înregistrare directă pe calculator, fără canale de comunicaţie intermediare.

Alte surse de variabilitate ţin de vorbitori şi cuprind caracteristicile lor biologice (sex, vârstă, stare fiziologică etc.), sociale (educaţie, ocupaţie etc.) şi lingvistice (eventualul dialect vorbit, particularităţi de pronunţie etc.) În aceste cercetări am încercat acoperirea acestui tip de variabilitate prin selectarea şi înregistrarea în baza de date a unui număr semnificativ de vorbitori, urmărind în mod riguros două criterii biologice, pe care le-am considerat cele mai importante şi mai uşor de apreciat – sexul şi grupa de vârstă.

Studiile asupra modelării acustice descrise în acest capitol au fost realizate folosind subseturi din baza de date proiectată, colectată şi etichetată anterior în acest scop. Pentru început a fost dezvoltat un sistem de recunoaştere automată a vorbirii continue dependent de vorbitor, care a permis o primă punere în evidenţă a unor probleme legate de alegerea unităţilor fonetice de modelare acustică.

Cercetări realizate cu sprijinul fostului CNCSU (devenit din 1999 CNCSIS) prin granturile 56/1995, 355/1996 şi 281/1998, şi al CNCSIS prin grantul 567/1999.

Problemele modelării acustice au fost apoi studiate prin experimente de recunoaştere independentă de vorbitor desfăşurate atât la nivelul unităţilor de modelare, cât şi la cel lexical (al cuvintelor), dependent şi independent de vocabular.

Rezultatele acestor studii, evaluate conform metodologiei descrise în secţiunea 3.1, sunt din câte cunoaştem primele referitoare la modelarea acustică pentru recunoaşterea automată a vorbirii continue în limba română, putând fi deci considerate un punct de referinţă pentru eventuale alte cercetări viitoare în acest domeniu.

6.1 Experimente dependente de vorbitor

Prima etapă a studiului modelării acustice sublexicale pentru recunoaşterea automată a vorbirii continue în limba română a constat în dezvoltarea unui sistem de recunoaştere dependent de vorbitor [31]. Acest sistem a fost bazat pe acelaşi set de unităţi fonetice de modelare acustică folosit şi pentru proiectarea bazei de date (tabelul 4.1).

Pentru construcţia sistemului au fost folosite circa 15 minute de semnal vocal colectat de la un singur vorbitor, obţinut prin citirea celor 40 de pasaje şi a propoziţiilor de iniţializare folosite pentru înregistrarea bazei de date (aproximativ 200 de propoziţii).

În vederea antrenării modelelor acustice şi a recunoaşterii, semnalul a fost supus unei analize cepstrale prin predicţie liniară similară celei folosite la etichetare (secţiunea 5.3.3), dar din cadre cu o lungime (25,6 ms - 512 eşantioane) şi o deplasare (10 ms) tipice pentru sistemele de recunoaştere a vorbirii. Log-energia nu a fost însă inclusă printre caracteristicile extrase, astfel încât vectorii acustici rezultaţi au fost 24-dimensionali.

Modelele acustice au avut aceeaşi structură ca şi în cazul etichetării (figura 5.5), dar datorită deplasării de 10 ms între cadre acestea au corespuns acum unei durate minime de 30 ms a unei unităţi de modelare. Funcţiile de probabilitate ale stărilor emiţătoare au fost mixturi gaussiene 24-dimensionale cu matrice de covarianţă diagonale.

Antrenarea modelelor acustice a fost realizată prin aceeaşi procedură folosită şi pentru etichetare (secţiunea 5.3.4), pentru iniţializare fiind utilizate cele patru propoziţii de iniţializare etichetate manual, iar pentru reestimarea Baum-Welch concatenată - cele 40 de pasaje, transcrise fonetic. Datorită cantităţii mici de date folosite, numărul de gaussiene dintr-o stare a fost limitat la $K = 5$.

6.1.1 Decodarea lingvistică

Aceste prime experimente de recunoaştere automată a vorbirii continue în limba română au utilizat în procesul de decodare lingvistică toate constrângerile secvenţiale dintr-un sistem de recunoaştere tipic (figura 1.1) - dicţionare de pronunţii ale cuvintelor, respectiv un model statistic şi o gramatică a pronunţiilor de recunoscut.

Dicţionarele de pronunţii şi modelele lingvistice folosite în aceste experimente au fost bazate în exclusivitate pe textele celor 40 de pasaje. O primă analiză a textelor a rezultat în identificarea unui vocabular de 1041 cuvinte cu pronunţii distincte, incluzând două corespunzătoare pauzelor din şi dintre propoziţii. Transcrierea lor în termenii unităţilor de modelare acustică a rezultat într-un prim dicţionar de pronunţii, care a inclus în transcrieri şi fenomene de fonetică sintactică - de exemplu, "ce-ar" transcris /Car/ etc.

O a doua variantă a dicţionarului de pronunţii a fost obţinută prin eliminarea /I/ din transcrierile fonetice ale cuvintelor conform secţiunii 6.1.2.

Pentru a simplifica o primă implementare a unui algoritm de decodare prin eliminarea necesităţii de a gestiona şi informaţii despre forma ortografică a cuvintelor, modelele lingvistice au fost bazate pe transcrierile fonetice ale pasajelor şi au inclus un model statistic de tip bigram, în care probabilitatea $P(w_2|w_1)$ de apariţie a unui cuvânt w_2 dat fiind predecesorul lui w_1 a fost estimată prin frecvenţa relativă

$$P(w_2|w_1) = N(w_1w_2)/N(w_1) \tag{6.1}$$

şi o gramatică de tip perechi-de-cuvinte (cf. engl. word-pair), în care toate cuvintele care urmează unui cuvânt au fost considerate echiprobabile.

Dicţionarele, modelele lingvistice şi cele acustice au fost utilizate într-un algoritm Viterbi cu reducerea spaţiului de căutare (pruning – secţiunea 3.9), implementat folosind liste de stări ale MMA active la un moment dat [136]. Pentru a reduce viteza de creştere a spaţiului de căutare, tranziţiile între cuvinte au fost restricţionate în mod euristic până la atingerea unor durate ale cuvintelor corespunzătoare unei medii de 60 ms a duratelor unităţilor sublexicale componente [31].

6.1.2 Rezultate şi comentarii

În aceste prime experimente, aceleaşi date au fost folosite şi pentru antrenarea MMA şi pentru testarea sistemului de recunoaştere. Deşi din punct de vedere metodologic o asemenea abordare (testare pe datele de antrenament [64]) este contraindicată, neputând evidenţia capacităţile de generalizare ale modelelor, ea este avantajoasă din punct de vedere al punerii la punct a algoritmilor. În cazul particular discutat aici, ea a condus totuşi şi la un prim rezultat semnificativ din punct de vedere al modelării acustice.

Pe durata etichetării bazei de date (capitolul 5) a devenit discutabilă existenţa fonemului /I/, postulată de lingvişti în sistemul fonologic al limbii române cu acceptarea cea mai largă la acest moment [244], [234]: deşi el ar trebui să aibă asociate segmente de semnal disticte, specifice, adesea asemenea segmente sunt imposibil de identificat.

Datorită acestei situaţii, utilizarea setului de unităţi de modelare din tabelul 4.1 a avut ca urmare obţinerea unui model acustic pentru /I/ caracterizat prin lipsă de specificitate – în lipsa segmentelor de semnal corespunzătoare lui /I/, modelul pentru acesta a fost construit prin colectarea unor segmente (minimum 30 ms) din sunetele adiacente. Rezultatul a fost că acest model avea tendinţa de a determina probabilităţi importante ale multor cuvinte scurte având pronunţii terminate în /I/ (îmi, îşi, îţi etc.), care erau inserate în mod frecvent şi determinau erori de recunoaştere foarte numeroase.

Această problemă a fost rezolvată prin eliminarea lui /I/ din setul de unităţi de modelare acustică, urmată de modificarea corespunzătoare a dicţionarului de pronunţii şi a transcrierilor semnalelor şi construcţia unor noi modele acustice. Aceasta a dus la dispariţia fenomenului menţionat mai sus şi a făcut posibile **primele demonstraţii ale recunoaşterii automate a vorbirii continue în limba română** (în iunie 1997).

6.2 Experimente independente de vorbitor

Odată un prim sistem de recunoaştere disponibil, cercetările au continuat în direcţia recunoaşterii independente de vorbitor. Trecerea de la recunoaşterea dependentă la cea independentă de vorbitor s-a făcut prin selectarea din baza de date a unor submulţimi de vorbitori cu distribuţii similare pe sexe şi grupe de vârstă: submulţimi de câte 60 de vorbitori pentru antrenarea modelelor acustice, respectiv 20 de vorbitori pentru testarea şi evaluarea sistemelor de recunoaştere bazate pe aceste modele.

Dintre materialele folosite pentru înregistrarea bazei de date (secţiunea 4.3), cele 40 de pasaje şi propoziţiile de completare asociate au fost citite de toţi vorbitorii, ceea ce a permis evaluarea modelelor acustice în mod dependent de vocabular. Existenţa în baza de date şi a unor semnale obţinute prin citirea de propoziţii individuale, specifice fiecărui vorbitor, a făcut posibilă şi evaluarea în mod independent de vocabular. Evaluările au fost făcute atât la nivelul unităţilor de modelare, cât şi la cel lexical, al cuvintelor.

Primele experimente de recunoaştere independentă de vorbitor au fost efectuate la nivelul sublexical, al unităţilor de modelare acustică, şi au urmărit evaluarea dependentă şi independentă de vocabular a celor două seturi alternative de unităţi de modelare acustică, conturate ca urmare a dezvoltării sistemului dependent de vorbitor: fonemele limbii române cu cea mai largă acceptare la acest moment (tabelul 4.1), considerat set de bază, respectiv un set redus, obţinut prin eliminarea fonemului /I/.

Deşi recunoaşterea şi evaluarea la nivel sublexical pot fi utile pentru a îmbunătăţi performanţele recunoaşterii la nivel lexical [82], recunoaşterea trebuie în cele din urmă efectuată la nivelul cuvintelor. Restul experimentelor au fost în consecinţă dedicate recunoaşterii la nivel lexical, evaluată dependent şi independent de vocabular.

Pentru a obţine informaţii referitoare strict la modelarea acustică, experimentele au fost efectuate utilizând gramatici deterministe de tip buclă (figura 3.3), în care toate unităţile de modelare respectiv cuvintele au fost considerate echiprobabile.

6.3 Recunoaşterea unităţilor de modelare

Experienţa acumulată până la acest moment, pe durata etichetării bazei de date (capitolul 5) şi a experimentelor de recunoaştere dependentă de vorbitor (secţiunea 6.1), indica posibila inadecvare a setului de unităţi fonetice folosit pentru proiectarea bazei de date (tabelul 4.1) în vederea modelării acustice sublexicale pentru recunoaşterea automată a vorbirii continue în limba română. Pentru clarificarea acestei probleme, primele experimente de recunoaştere independentă de vorbitor au fost efectuate la nivelul unităţilor de modelare acustică [30] şi au urmărit evaluarea performanţelor asigurate de posibile seturi alternative de unităţi de modelare.

6.3.1 Vorbitori şi date

În aceste experimente au fost folosiţi 80 de vorbitori – 60 de antrenament şi 20 de test – uniform distribuiţi pe sexe şi grupe de vârstă (tabelul 6.1).

Datele de antrenament au inclus pasajele şi propoziţiile de completare înregistrate de vorbitorii de antrenament, iar pentru teste au fost folosite mulţimi de date colectate

Tabelul 6.1: Mulţimile de vorbitori folosite în experimentele de recunoaştere a unităţilor de modelare (v. tabelul 4.3 pentru interpretarea codificării)

Antrenament						Test	
MA_1	FF_2	MK_3	FP_4	MU_5	GA_1	NF_2	GK_3
FA_1	MF_2	FK_3	MP_4	FU_5	NA_1	GF_2	NK_3
MB_2	FG_3	ML_4	FQ_5	MV_1	GB_2	NG_3	GL_4
FB_2	MG_3	FL_4	MQ_5	FV_1	NB_2	GG_3	NL_4
MC_3	FH_4	MM_5	FR_1	MX_2	GC_3	NH_4	GM_5
FC_3	MH_4	FM_5	MR_1	FX_2	NC_3	GH_4	NM_5
MD_4	FI_5	MN_1	FS_2	MY_3	GD_4	NI_5	GN_1
FD_4	MI_5	FN_1	MS_2	FY_3	ND_4	GI_5	NN_1
ME_5	FJ_1	MO_2	FT_3	MZ_4	GE_5	NJ_1	GO_2
FE_5	MJ_1	FO_2	MT_3	FZ_4	NE_5	GJ_1	NO_2

Tabelul 6.2: Caracteristici ale textelor citite pentru înregistrarea datelor folosite în experimentele de recunoaştere a unităţilor de modelare acustică

Texte	Număr propoziţii	Număr cuvinte	Lungime medie a propoziţiilor	Cuvinte distincte
Pasaje	197	2217	11,3 cuvinte	1043
Completare	26	263	10,1 cuvinte	174
Individuale	91	878	9,65 cuvinte	575

de la vorbitorii de test:

- pentru teste dependente de vocabular (DV), înregistrările pasajelor făcute de către cei 20 de vorbitori de test;
- pentru teste independente de vocabular (IV), cele 91 de propoziţii individuale specifice vorbitorilor de test (4-5 propoziţii/vorbitor).

Caracteristicile textelor folosite pentru înregistrarea datelor sunt prezentate în tabelul 6.2, iar caracteristicile mulţimilor de date de antrenament şi de test – în tabelul 6.3. Faţă de experimentele dependente de vorbitor, numărul de cuvinte distincte din pasaje a crescut la 1043 datorită considerării separate a componentelor câtorva cuvinte compuse.

Din tabele se poate observa că propoziţiile de completare au adăugat doar 116 cuvinte distincte la cele 1043 din pasajele utilizate pentru teste dependente de vocabular, adică 10% din cele 1159 cuvinte distincte din mulţimea datelor de antrenament.

Se observă de asemeni că în propoziţiile individuale, utilizate în testele independente de vocabular, aproape 72% dintre cuvintele distincte sunt specifice acestora: deoarece vorbirea este foarte greu separabilă în submulţimi cu vocabulare disjuncte, am considerat aceste valori ca fiind acceptabile pentru cele două condiţii de test – dependent (DV) respectiv independent de vocabular (IV).

Tabelul 6.3: Caracteristici ale datelor utilizate pentru antrenarea modelelor acustice şi teste dependente (DV) respectiv independente de vocabular (IV) în experimentele de recunoaştere a unităţilor de modelare acustică

Date	Număr propoziţii	Număr cuvinte	Cuvinte distincte	Cuvinte specifice	Durată
Antrenare	1338	14880	1159	1159 (100%)	1h45'50"
Teste DV	394	4434	1043	0 (0%)	30'37"
Teste IV	91	878	575	412 (71,7%)	7'13"

6.3.2 Alternativele de modelare

Analiza problemelor cauzate de utilizarea lui /I/ ca unitate de modelare acustică şi a literaturii lingvistice în privinţa sunetelor limbii române a condus iniţial la conturarea a două posibile seturi alternative de unităţi de modelare acustică.

Prima alternativă ar fi constat în înlocuirea secvenţelor de tipul consoană-/I/ prin variante palatalizate ale consoanelor, corespunzător teoriilor lingvistice care postulează existenţa consoanelor palatalizate în limba română. O analiză a etichetării datelor de antrenament a arătat însă că frecvenţele de apariţie ale secvenţelor consoană-/I/ sunt foarte reduse în raport cu frecvenţele consoanelor de bază – în general sub 1%, singurul caz în care acest prag a fost depăşit fiind cel al consoanei /r/ (1,56%).

A doua alternativă consta în eliminarea lui /I/ din setul iniţial de unităţi de modelare, deja testată în sistemul dependent de vorbitor. Aceasta s-a dovedit singura fezabilă în condiţiile existente deoarece frecvenţele reduse de apariţie ale secvenţelor consoană-/I/ constatate, conjugate cu dimensiunile bazei de date colectate, nu permit antrenarea de modele ale consoanelor palatalizate, presupuse de prima alternativă.

6.3.3 Extragerea caracteristicilor

Fişierele de semnal vocal din mulţimile de date de antrenament şi de test au fost supuse unei analize cepstrale melodice (secţiunea 2.7.1) conform figurii 6.1: după preaccentuarea cu un coeficient $\alpha = 0{,}97$, eşantioanele semnalelor au fost grupate în cadre cu lungimea de 25,6 ms (512 eşantioane) distanţate la 10 ms, ferestruite cu o fereastră Hamming şi prelucrate folosind un algoritm de transformare Fourier rapidă. Spectrul de amplitudine rezultat a fost transformat într-unul melodic prin sumări ponderate ale componentelor sale, corespunzătoare unui bloc de 30 filtre triunghiulare uniform distribuite pe scara melodică, iar printr-o transformare cosinus discretă a spectrului melodic au fost obţinuţi 12 coeficienţi cepstrali melodici.

Coeficienţii cepstrali melodici au fost liftraţi cu un liftru de tip sinus ridicat cu o lungime $L = 22$ (ecuaţia 3.18) şi împreună cu o estimare a log-energiei fiecărui cadru şi coeficienţii lor Δ (secţiunea 2.7.2) au format vectori acustici 26-dimensionali folosiţi efectiv la antrenarea modelelor acustice şi recunoaştere.

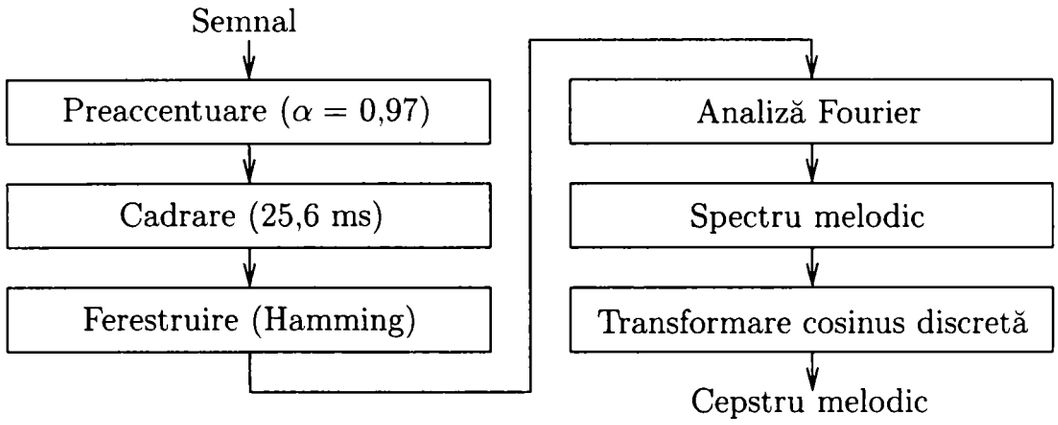

Figura 6.1: Analiza cepstrală melodică

6.3.4 Modelele acustice

Modelarea acustică a fost realizată folosind tot MMA de tipul stânga-dreapta cu două stări conectoare şi trei stări emiţătoare (figura 5.5), având ca funcţii de probabilitate $b(y)$ mixturi gaussiene cu matrice de covarianţă diagonale 26-dimensionale.

Antrenarea MMA a fost făcută conform procedurii din secţiunea 5.3.4; pentru a evita însă unele instabilităţi ale algoritmului de grupare folosit, în faza de iniţializare vectorii asociaţi unei stări emiţătoare au fost utilizaţi pentru estimarea unei singure gaussiene. După fiecare reestimare concatenată, numărul de gaussiene al fiecărei stări emiţătoare a fost incrementat prin divizarea gaussienei de pondere maximă şi înlocuirea ei cu altele două de ponderi egale cu jumătate din ponderea celei iniţiale. Vectorii medii ai noilor gaussiene au fost obţinuţi prin deplasări ale vectorului mediu al gaussienei iniţiale cu $\pm 20\%$ din abaterea standard corespunzătoare. După fiecare mărire a numărului de gaussiene, modelele au fost reestimate folosind algoritmul Baum-Welch concatenat.

6.3.5 Rezultate şi comentarii

Experimentele de recunoaştere a unităţilor sublexicale de modelare acustică din setul de bază (tabelul 4.1) şi cel redus (fără /I/) s-au desfăşurat folosind MMA cu până la 16 gaussiene/stare emiţătoare şi au fost evaluate conform metodologiei prezentate în secţiunea 3.1, iar rezultatele sunt prezentate sub formă numerică în tabelele 6.4 respectiv 6.5 şi sub formă grafică în figurile 6.2–6.6.

Analiza rezultatelor evidenţiază faptul că procedura de antrenare utilizată conduce la modele cu o foarte bună capacitate de generalizare: frecvenţa recunoaşterilor corecte evaluată independent de vocabular este superioară celei evaluate dependent de vocabular atât în cazul setului de bază de unităţi de modelare, cât şi al celui redus (figura 6.2), iar acurateţea evaluată independent de vocabular este foarte apropiată de cea evaluată dependent de vocabular (figura 6.3).

Dar aspectul cel mai interesant se referă la unităţile de modelare acustică folosite:

Tabelul 6.4: Rezultatele experimentelor de recunoaştere a unităţilor de modelare acustică folosind setul de bază (tabelul 4.1)

Număr de gaussiene	Dependent de vocabular					Independent de vocabular				
	C	A	S	O	I	C	A	S	O	I
1	61,09	55,29	26,80	12,11	5,80	62,03	54,22	28,25	9,71	7,82
2	61,36	55,77	26,47	12,17	5,58	62,28	54,46	27,88	9,84	7,82
3	62,37	56,90	25,86	11,77	5,46	63,25	55,85	27,11	9,65	7,40
4	64,01	58,23	24,76	11,23	5,78	65,01	57,81	25,63	9,36	7,20
5	65,35	59,38	23,93	10,71	5,97	66,17	59,11	24,77	9,05	7,07
6	66,91	61,19	22,94	10,14	5,72	67,10	59,70	24,42	8,48	7,40
7	68,20	62,41	21,91	9,89	5,80	68,80	61,24	23,17	8,04	7,55
8	69,09	63,40	21,33	9,58	5,69	70,03	62,37	22,13	7,84	7,66
9	69,81	64,15	20,80	9,38	5,66	70,62	62,96	21,87	7,51	7,66
10	70,53	65,06	20,46	9,02	5,47	71,42	64,08	21,14	7,44	7,33
11	71,12	65,84	20,08	8,80	5,28	72,10	65,05	20,44	7,47	7,05
12	71,54	66,22	19,80	8,66	5,32	72,38	65,43	20,30	7,31	6,96
13	71,78	66,67	19,75	8,47	5,11	72,61	65,58	20,28	7,11	7,02
14	71,95	66,79	19,61	8,44	5,15	73,13	66,40	19,91	6,96	6,74
15	72,17	67,10	19,48	8,34	5,07	73,82	67,34	19,22	6,96	6,47
16	72,35	67,33	19,27	8,37	5,02	73,64	67,10	19,51	6,85	6,54

Tabelul 6.5: Rezultatele experimentelor de recunoaştere a unităţilor de modelare acustică folosind setul redus (fără /I/)

Număr de gaussiene	Dependent de vocabular					Independent de vocabular				
	C	A	S	O	I	C	A	S	O	I
1	61,16	55,75	26,72	12,12	5,41	62,08	54,99	28,22	9,70	7,09
2	61,46	56,16	26,37	12,16	5,30	62,57	55,48	27,72	9,72	7,09
3	62,61	57,39	25,60	11,79	5,23	63,65	56,65	26,65	9,70	7,00
4	64,20	58,60	24,68	11,13	5,59	65,30	58,24	25,49	9,21	7,07
5	65,54	59,83	23,82	10,64	5,71	66,32	59,45	24,65	9,03	6,87
6	67,05	61,52	22,81	10,14	5,53	67,42	60,23	23,98	8,59	7,20
7	68,32	62,80	21,86	9,82	5,52	69,02	61,53	22,90	8,08	7,49
8	69,23	63,81	21,17	9,60	5,41	70,38	63,01	21,67	7,95	7,38
9	69,94	64,51	20,72	9,34	5,43	71,22	64,05	21,31	7,46	7,18
10	70,64	65,39	20,15	9,21	5,25	71,40	64,42	21,09	7,51	6,98
11	71,14	65,91	20,05	8,80	5,24	72,08	65,24	20,34	7,58	6,85
12	71,64	66,46	19,88	8,48	5,18	72,55	65,68	20,38	7,07	6,87
13	71,82	66,76	19,70	8,47	5,06	73,03	66,63	19,99	6,98	6,40
14	72,15	67,14	19,39	8,46	5,02	73,32	66,89	19,61	7,07	6,43
15	72,48	67,63	19,15	8,37	4,85	73,56	67,07	19,39	7,05	6,49
16	72,67	67,92	19,00	8,33	4,75	73,90	67,54	19,26	6,85	6,36

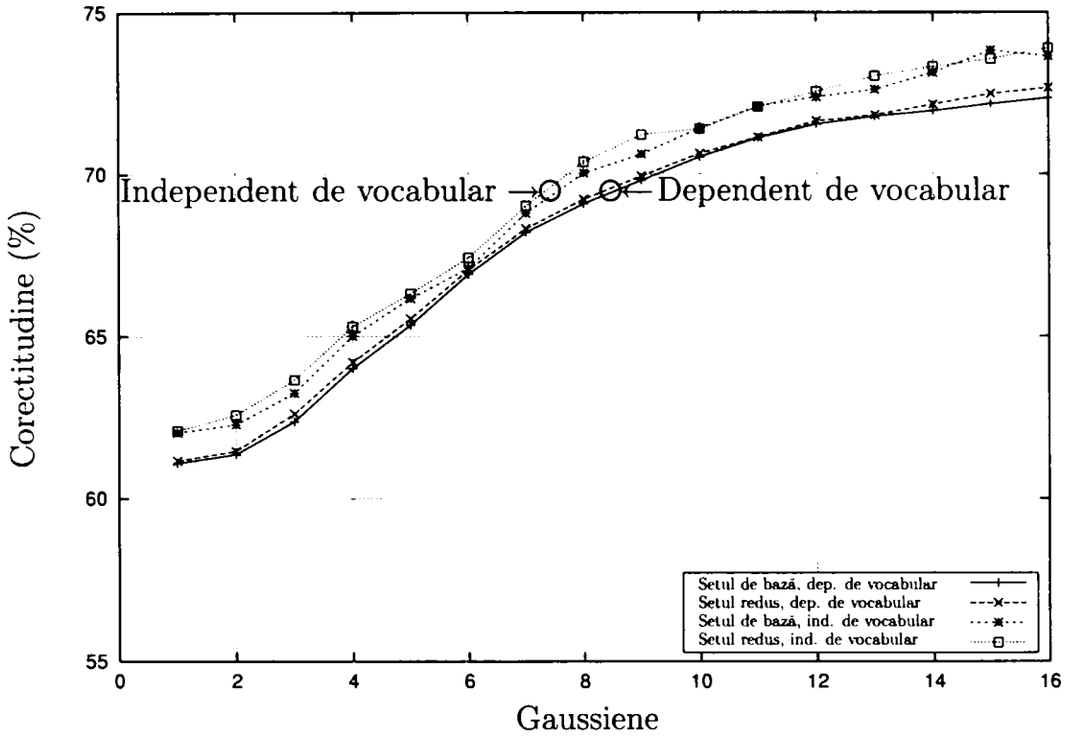

Figura 6.2: Evoluţia corectitudinii recunoaşterii unităţilor de modelare acustică funcţie de numărul de gaussiene/stare emiţătoare

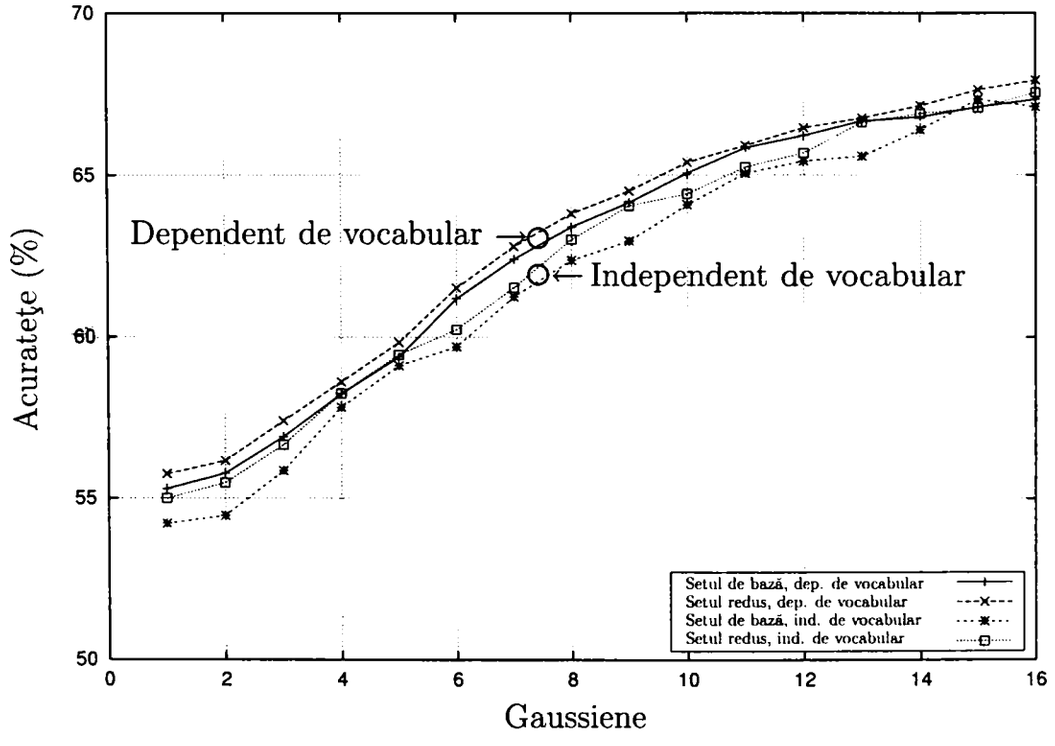

Figura 6.3: Evoluţia acurateţii recunoaşterii unităţilor de modelare acustică funcţie de numărul de gaussiene/stare emiţătoare

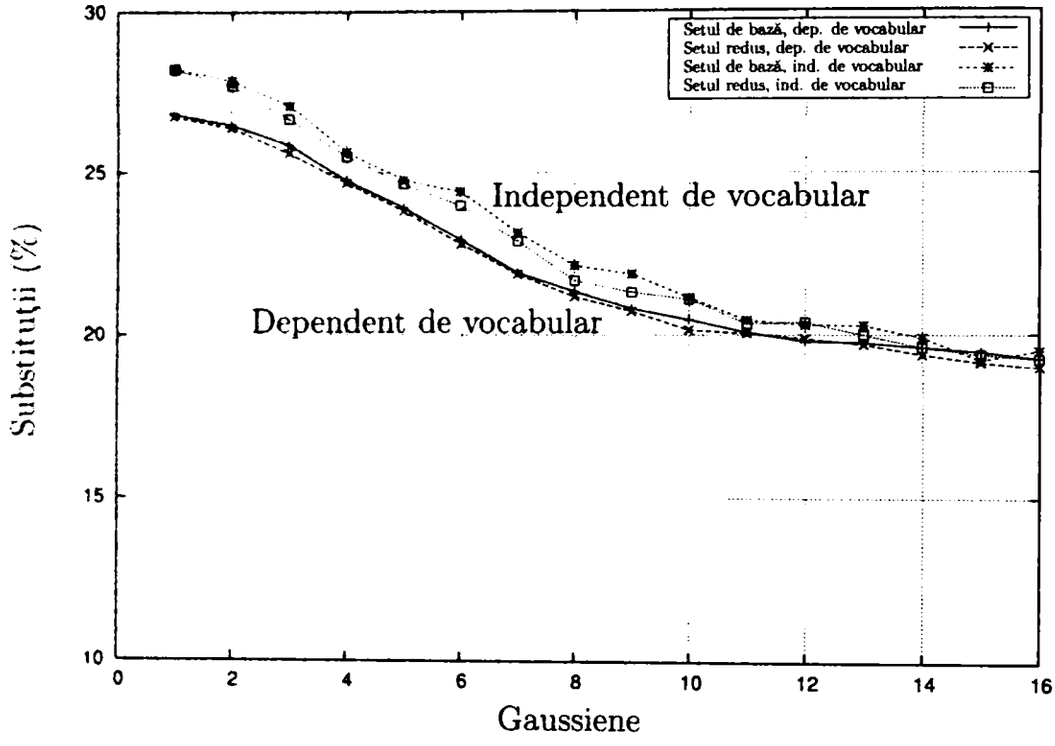

Figura 6.4: Evoluţia frecvenţei substituţiilor unităţilor de modelare acustică funcţie de numărul de gaussiene/stare emiţătoare

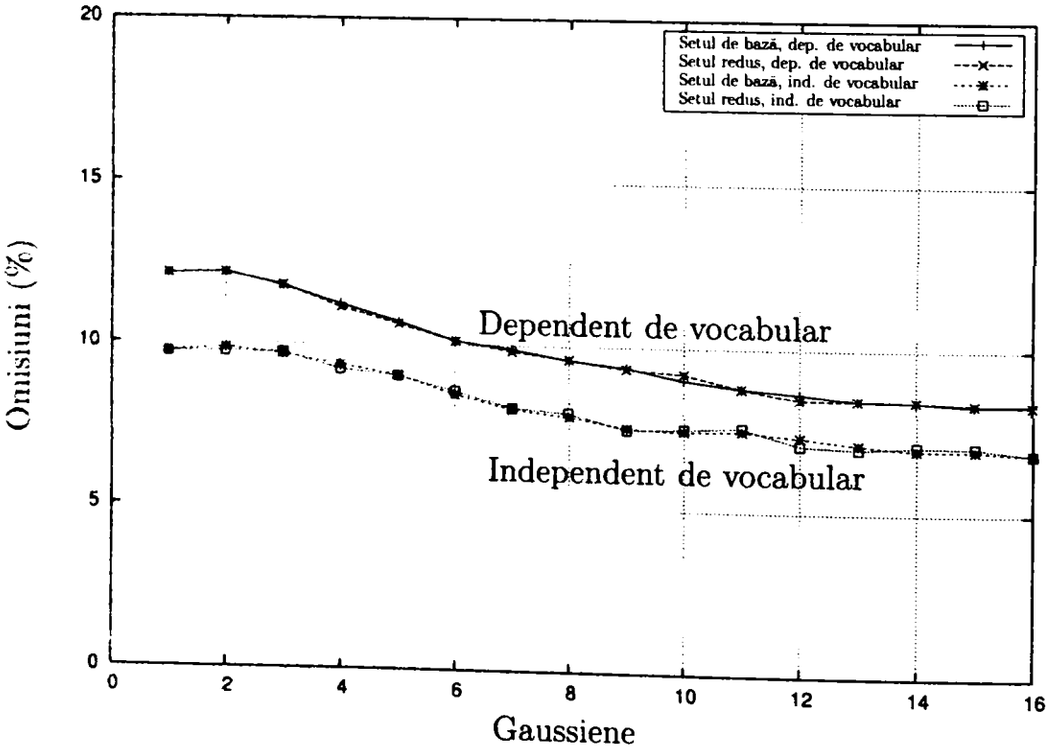

Figura 6.5: Evoluţia frecvenţei omisiunilor unităţilor de modelare acustică funcţie de numărul de gaussiene/stare emiţătoare

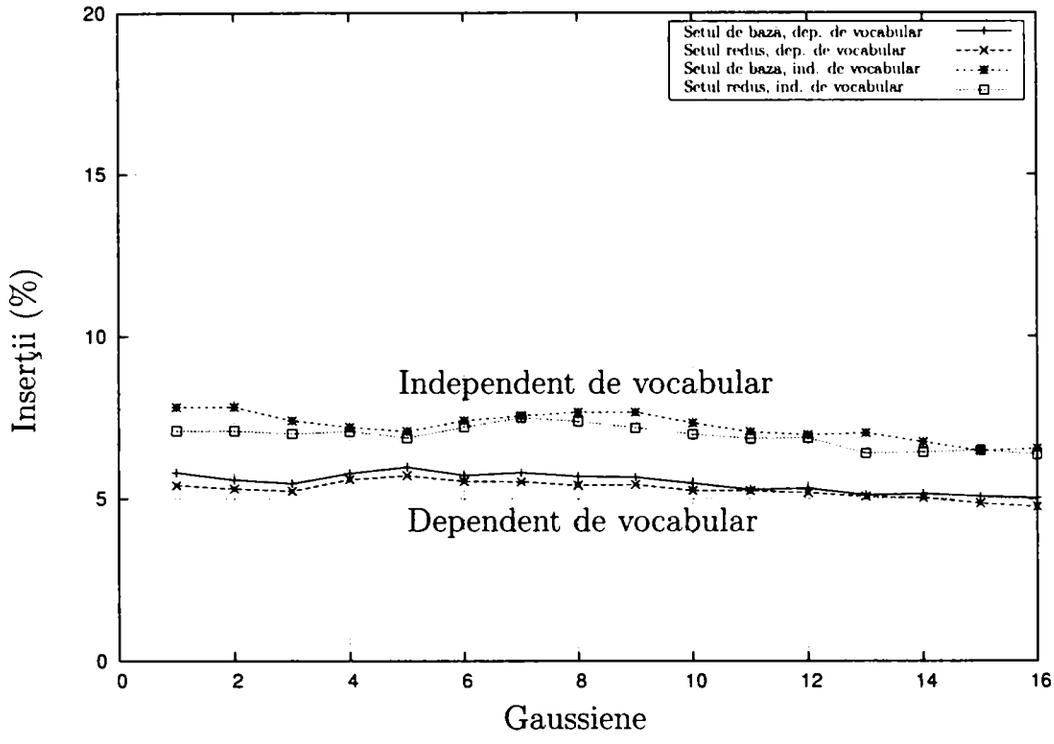

Figura 6.6: Evoluţia frecvenţei inserţiilor unităţilor de modelare acustică funcţie de numărul de gaussiene/stare emiţătoare

după cum se observă, setul redus (fără /I/) a condus la rezultate uşor superioare atât din punctul de vedere al corectitudinii, cât şi al acurateţii recunoaşterii, indiferent de dependenţa sau independenţa de vocabular. Studiind şi tipurile de erori – substituţii (figura 6.4), omisiuni (figura 6.5) şi inserţii (figura 6.6) – se observă că diferenţele între rezultate sunt datorate nu atât modificării setului de unităţi de modelare, cât mai ales dependenţei sau independenţei de vocabular.

Influenţa setului de unităţi de modelare asupra omisiunilor (figura 6.5) este practic neglijabilă, în acest caz diferenţa fiind dată de dependenţa/independenţa de vocabular. În schimb, ea este sesizabilă în cazul inserţiilor (figura 6.6) şi substituţiilor (figura 6.4), care au fost în general reduse prin utilizarea setului alternativ de unităţi de modelare. Reducerea inserţiilor este sistematică, indiferent de dependenţa sau independenţa de vocabular, iar cea a substituţiilor se manifestă în special independent de vocabular. În plus, valorile frecvenţei substituţiilor converg odată cu creşterea numărului de gaussiene.

Comparând rezultatele obţinute cu unele [85] raportate în condiţii apropiate (35 unităţi de modelare cu 16 densităţi gausiene pe stare) pentru o altă limbă romanică, franceza (C = 62,4%, A = 59,2%, S = 25,4%, O = 12,2%, I = 3,2%), s-ar putea spune că ele sunt mult mai bune decât acestea din urmă. Luând însă în calcul că modelele folosite aici au fost antrenate cu mai mulţi vorbitori (60 faţă de 43) şi mai mult semnal (aproape 106 minute faţă de cca. 50), consider că aceste rezultate sunt normale.

Această opinie este susţinută şi de rezultatele obţinute în condiţii comparabile pentru limba spaniolă [35]: cu 25 unităţi de modelare şi 3 gaussiene pe stare, dependent de

Tabelul 6.6: Vorbitorii folosiţi în experimentele de recunoaştere a cuvintelor (v. tabelul 4.3 pentru interpretarea codificării)

Antrenament						Test	
MU_5	GA_1	NF_2	GK_3	NP_4	GU_5	MA_1	FF_2
FU_5	NA_1	GF_2	NK_3	GP_4	NU_5	FA_1	MF_2
MV_1	GB_2	NG_3	GL_4	NQ_5	GV_1	MB_2	FG_3
FV_1	NB_2	GG_3	NL_4	GQ_5	NV_1	FB_2	MG_3
MX_2	GC_3	NH_4	GM_5	NR_1	GX_2	MC_3	FH_4
FX_2	NC_3	GH_4	NM_5	GR_1	NX_2	FC_3	MH_4
MY_3	GD_4	NI_5	GN_1	NS_2	GY_3	MD_4	FI_5
FY_3	ND_4	GI_5	NN_1	GS_2	NY_3	FD_4	MI_5
MZ_4	GE_5	NJ_1	GO_2	NT_3	GZ_4	ME_5	FJ_1
FZ_4	NE_5	GJ_1	NO_2	GT_3	NZ_4	FE_5	MJ_1

vocabular, C = 63,4% pentru spaniolă faţă de C = 62,6% pentru româna, iar independent de vocabular - C = 61,4% pentru spaniolă faţă de C = 63,6% pentru română.

6.4 Recunoaşterea cuvintelor

În experimentele de recunoaştere a cuvintelor descrise în continuare, ca şi în cele de recunoaştere a unităţilor de modelare, a fost utilizată o gramatică simplă de tip buclă (figura 3.3) pentru a obţine informaţii referitoare strict la modelarea acustică.

Pentru aceste experimente au fost utilizate şi două pachete publice de programe pentru construcţia şi evaluarea sistemelor de recunoaştere a vorbirii: unul dezvoltat la Universitatea statului Mississippi [176], şi HTK (HMM Toolkit) [260], dezvoltat la Universitatea Cambridge [258] şi firma Entropic Research Laboratories Inc., făcut public în anul 2000, după preluarea acestei firme de către compania Microsoft.

În continuare vor fi prezentate doar experimentele bazate pe al doilea dintre cele două pachete, care are avantajul maturităţii şi al documentaţiei superioare.

6.4.1 Vorbitori şi date

Experimentele de recunoaştere a cuvintelor au utilizat tot o submulţime de 60 de vorbitori de antrenament şi una de 20 de vorbitori de test (tabelul 6.6), dar schimbate faţă de experimentele de recunoaştere a unităţilor de modelare pentru a maximiza cantitatea de semnale disponibile pentru testele independente de vocabular.

Datele folosite pentru antrenarea modelelor acustice şi teste dependente de vocabular (tabelul 6.7) păstrează cele mai multe din caracteristicile celor anterioare (tabelul 6.3), diferite fiind doar duratele: pe de o parte datorită schimbării vorbitorilor, pe de alta - datorită segmentării pasajelor în propoziţii, însoţită de eliminarea pauzelor dintre ele.

Setul de date pentru teste independente de vocabular a fost extins de la 91 la 133 de propoziţii individuale (6-7 propoziţii/vorbitor), având acum o durată de aproape 10

Tabelul 6.7: Caracteristici ale datelor utilizate pentru antrenarea modelelor acustice şi teste dependente (DV) respectiv independente de vocabular (IV) în experimentele de recunoaştere a cuvintelor

Date	Număr propoziţii	Număr cuvinte	Cuvinte distincte	Cuvinte specifice	Durată
Antrenare	1338	14880	1159	1159 (100%)	1h37'11"
Teste DV	394	4434	1043	0 (0%)	27'38"
Teste IV	133	1290	784	585 (74,6%)	9'53"

minute. Această extindere a fost urmată de creşterea atât a dimensiunii vocabularului aferent de la 575 la 784 cuvinte, cât şi a numărului de cuvinte specifice (de la 412 la 585) şi a ponderii lor în acest vocabular (de la 71,7% la 74,6%). A rezultat de asemeni şi o uşoară creştere (de la 9,65 la 9,7 cuvinte) a lungimii medii a propoziţiilor individuale utilizate pentru testele independente de vocabular.

6.4.2 Dicţionarele

Cuvintele distincte din cele trei mulţimi de date (pentru antrenament şi pentru teste dependente şi independente de vocabular – tabelul 6.7) au fost grupate în două dicţionare de pronunţii: unul utilizat pe durata antrenării modelelor şi a testelor dependente de vocabular, şi unul pentru testele independente de vocabular.

Corespunzător celor două seturi de unităţi de modelare acustică evaluate (de bază şi redus), fiecare dintre aceste dicţionare a avut la rândul lui două variante, în care pronunţiile cuvintelor au fost precizate în termenii unităţilor din setul respectiv.

Au fost adăugate pronunţii alternative pentru a ţine cont de posibila apariţie a unor pauze între cuvinte (v. şi secţiunea 6.4.3) şi de variantele de pronunţie ale cuvintelor (de exemplu, "optsprezece" poate fi pronunţat /optsprezeCe/ sau /opSpe/).

Deoarece gramatica de tip buclă utilizată în aceste experimente nu permite distincţii între cuvintele cu pronunţii identice (homofone), prin examinarea dicţionarelor au fost identificate perechile de cuvinte homofone, ale căror substituţii nu au fost considerate erori. Listele acestor cuvinte, împreună cu pronunţiile lor, sunt incluse în anexa A.

6.4.3 Modelele acustice

Pentru recunoaşterea cuvintelor au fost folosite tot modele acustice cu o structură de tip stânga-dreapta cu două stări conectoare şi trei emiţătoare (figura 5.5). Deoarece analiza semnalelor a fost făcută ca şi pentru experimentele de recunoaştere a unităţilor de modelare (secţiunea 6.3.3), funcţiile de probabilitate asociate stărilor emiţătoare au fost mixturi de densităţi gaussiene 26-dimensionale cu matrice de covarianţă diagonale.

Procedura de antrenare a fost însă modificată pentru a ţine cont de dicţionarele de pronunţii ca intermediare între nivelul unităţilor de modelare şi cel lexical. Astfel, după iniţializarea şi prima reestimare concatenată a modelelor cu o gaussiană pe stare emiţătoare folosind datele de antrenament etichetate, modelul pauzelor a fost înlocuit cu

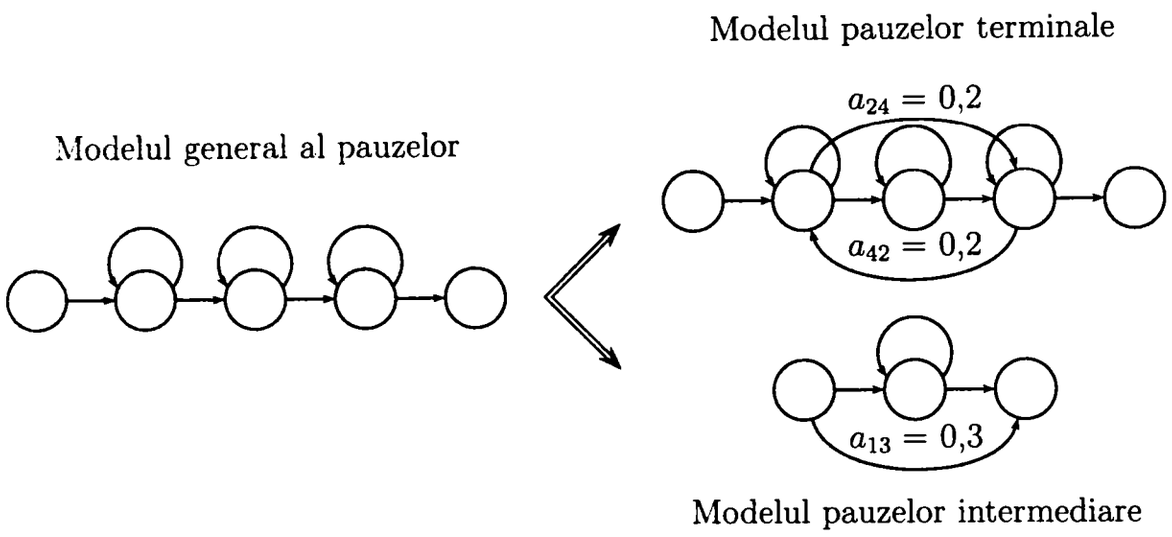

Figura 6.7: Înlocuirea modelului general al pauzelor cu modele ale pauzelor terminale (de la începutul şi sfârşitul propoziţiilor) şi intermediare (dintre cuvinte)

alte două obţinute prin editarea lui (figura 6.7): unul pentru pauzele terminale (iniţială şi finală) şi unul pentru cele intermediare, care pot apare opţional între cuvinte [260].

Modelul pauzelor terminale a fost obţinut prin copierea modelului general al pauzelor, iar cel al pauzelor intermediare a inclus o singură stare emiţătoare pentru a putea modela şi pauze foarte scurte între cuvinte. Iniţial, această stare avea funcţia de probabilitate a stării de mijloc din modelul general al pauzelor, iar în continuare parametrii acesteia au fost partajaţi cu cei ai stării de mijloc din modelul pauzelor terminale.

Caracterul opţional al pauzelor intermediare a fost marcat în modelul lor printr-o tranziţie directă din starea iniţială în cea finală. Alte tranziţii au fost adăugate în modelul pauzelor terminale pentru a permite traversări mai rapide sau multiple ale acestuia. Tranziţiilor noi adăugate le-au fost atribuite probabilităţi arbitrare, iar probabilităţile celorlalte tranziţii au fost scalate pentru a respecta restricţiile stochastice specifice.

Luarea în consideraţie a dicţionarelor de pronunţii a început prin utilizarea pentru următoarele reestimări concatenate nu a etichetelor semnalelor, ci a transcrierilor lor fonetice obţinute din cele ortografice prin înlocuirea fiecărui cuvânt cu o pronunţie a lui incluzând pauza finală opţională dintre cuvinte. Aceste transcrieri au fost folosite pentru încă două reestimări Baum-Welch concatenate ale modelelor.

Detectarea apariţiilor efective ale pauzelor opţionale dintre cuvinte a fost realizată prin decodarea Viterbi a datelor de antrenament folosind reţele de recunoaştere obţinute prin înlocuirea fiecărui cuvânt din transcrierile lor ortografice cu pronunţiile de bază şi cele incluzând pauzele opţionale finale, conectate în paralel. A rezultat astfel un nou set de transcrieri fonetice, care au luat în calcul atât pronunţiile din dicţionare cât şi pauzele dintre cuvinte şi care au fost folosite pentru restul reestimărilor concatenate.

După încă două reestimări concatenate a început mărirea numărului de gaussiene pe stare emiţătoare (secţiunea 6.3.4). Pentru a ţine însă cont de diferenţele care apar între

pronunţiile din dicţionar şi cele efective ale cuvintelor, fiecare incrementare a numărului de gaussiene a fost urmată nu de una, ci de două reestimări Baum-Welch concatenate.

6.4.4 Rezultate şi comentarii

Experimentele de recunoaştere a cuvintelor au fost efectuate utilizând setul de bază şi cel redus de unităţi de modelare şi MMA cu până la 16 gaussiene pe stare emiţătoare. Decodarea a fost realizată folosind implementarea cu liste înlănţuite [261] a algoritmului Viterbi cu reducere, inclusă în pachetul HTK, iar experimentele finale, ale căror rezultate sunt prezentate aici, au fost precedate de unele preliminare, mult mai cuprinzătoare, pentru alegerea penalizării de tranziţie şi a pragului de reducere (secţiunea 3.9).

Urmărind maximizarea simultană a corectitudinii şi acurateţii recunoaşterii şi evitarea erorilor de căutare, au fost determinate experimental penalizarea de tranziţie de -25 şi pragul de reducere de -600 pentru testele dependente de vocabular, respectiv -30 şi -900 pentru cele independente de vocabular: se observă că în cazul independenţei de vocabular au fost necesare valori mai mari în valoare absolută, corespunzătoare unei tendinţe mai puternice spre inserţii, respectiv unui spaţiu mai extins al soluţiilor.

Experimentele finale, desfăşurate folosind aceşti parametri, au fost evaluate conform metodologiei din secţiunea 3.1, iar rezultatele sunt prezentate numeric în tabelele 6.8 şi 6.9 şi grafic în figurile 6.8–6.12 – atenţie la diferenţele de scară! Pentru a ilustra unele aspecte discutate în continuare, exemple de recunoaştere sunt incluse în anexa B.

După cum era de aşteptat, dat fiind numărul mai redus de cuvinte din acest caz (784 faţă de 1043 – v. tabelul 6.7), performanţele independente de vocabular sunt superioare celor dependente de vocabular. Diferenţa dintre cele două condiţii de test este uşor observabilă în toate reprezentările grafice ale evoluţiilor metricelor de performanţă cu numărul de gaussiene/stare emiţătoare, cu excepţia frecvenţei inserţiilor, în cazul căreia diferenţa este greu sesizabilă datorită alegerii valorilor penalizării de tranziţie.

Se observă de asemeni apariţia unor abateri pronunţate de la tendinţele de ameliorare ale performanţelor în cazurile modelării cu 8 şi 14 gaussiene/stare emiţătoare, abateri manifestate la nivelul tuturor metricelor cu excepţia frecvenţei omisiunilor şi al căror mecanism nu a fost încă elucidat: dat fiind faptul că abaterile, deşi mai puţin pronunţate, apar şi în cazul testelor dependente de vocabular, ar putea fi vorba de particularităţi ale datelor de antrenament care îngreuiază antrenarea unor modele în aceste cazuri, sau de fenomene care ţin de dinamica procesului de antrenare a modelelor acustice.

Examinând rezultatele şi din punctul de vedere al setului de unităţi de modelare folosit, constatăm că setul de bază a condus la rezultate uşor superioare în cazul testelor independente de vocabular cu peste 8 gaussiene/stare şi al celor dependente de vocabular cu mai puţin de 8 gaussiene/stare, iar cel redus – în cazul celor dependente de vocabular cu peste 8 gaussiene/stare. Studiul ipotezelor decodate şi al alinierilor lor cu transcrierile de referinţă, exemplificate în anexa B, arată însă că aceste diferenţe nu pot fi atribuite direct noilor cuvinte homofone introduse prin utilizarea setului redus (anexa A).

Evaluarea la nivelul de semnificaţie $p = 0{,}05$ a semnificaţiei statistice [37] a diferenţelor dintre performanţele obţinute folosind setul de bază şi cel redus de unităţi de modelare acustică, realizată prin compararea distribuţiilor numerelor de erori din zonele eronate corespondente statistic independente [91], [182], arată că aceste diferenţe sunt în general

Tabelul 6.8: Rezultatele experimentelor de recunoaştere a cuvintelor folosind setul de bază de unităţi de modelare

Număr de gaussiene	Dependent de vocabular					Independent de vocabular				
	C	A	S	O	I	C	A	S	O	I
1	49,48	45,60	39,85	10,67	3,88	58,53	54,65	32,25	9,22	3,88
2	55,05	51,49	35,34	9,61	3,56	65,19	61,78	26,74	8,06	3,41
3	56,54	53,47	33,99	9,47	3,07	65,74	62,79	26,51	7,75	2,95
4	59,97	57,53	31,08	8,95	2,44	69,15	66,36	23,80	7,05	2,79
5	63,08	60,89	28,66	8,25	2,19	70,85	68,29	22,79	6,36	2,56
6	64,91	62,86	27,02	8,07	2,05	72,02	69,77	22,02	5,97	2,25
7	65,27	63,08	27,15	7,58	2,19	72,71	70,54	21,24	6,05	2,17
8	64,75	62,11	27,76	7,49	2,64	71,09	67,52	23,26	5,66	3,57
9	67,16	65,16	25,15	7,69	2,01	74,88	72,79	19,30	5,81	2,09
10	67,57	65,70	24,70	7,74	1,87	75,89	74,03	18,68	5,43	1,86
11	68,09	66,17	24,29	7,62	1,92	75,97	74,26	18,68	5,35	1,71
12	68,72	66,89	23,61	7,67	1,83	76,20	74,50	18,68	5,12	1,71
13	68,43	66,71	23,82	7,76	1,71	76,82	74,88	18,22	4,96	1,94
14	67,64	65,47	24,88	7,49	2,17	74,26	70,54	20,70	5,04	3,72
15	69,35	67,82	23,21	7,44	1,53	77,13	75,50	17,67	5,19	1,63
16	68,99	67,34	23,55	7,47	1,65	76,90	74,57	17,91	5,19	2,33

Tabelul 6.9: Rezultatele experimentelor de recunoaştere a cuvintelor folosind setul redus de unităţi de modelare

Număr de gaussiene	Dependent de vocabular					Independent de vocabular				
	C	A	S	O	I	C	A	S	O	I
1	49,55	45,76	39,90	10,55	3,79	58,68	55,19	31,86	9,46	3,49
2	54,94	51,38	35,68	9,38	3,56	65,04	61,78	26,43	8,53	3,26
3	56,61	53,38	34,28	9,11	3,23	65,50	62,40	26,51	7,98	3,10
4	59,90	57,22	31,48	8,62	2,68	68,91	66,36	24,19	6,90	2,56
5	62,52	60,08	29,36	8,12	2,44	70,70	68,22	22,87	6,43	2,48
6	64,19	61,82	27,90	7,92	2,37	71,94	70,08	21,71	6,36	1,86
7	65,07	62,58	27,47	7,47	2,48	72,40	70,23	21,63	5,97	2,17
8	65,31	62,36	27,33	7,35	2,95	70,62	67,05	23,72	5,66	3,57
9	67,23	64,86	25,35	7,42	2,37	74,34	72,09	19,92	5,74	2,25
10	67,93	65,79	24,70	7,37	2,14	74,26	72,09	20,39	5,35	2,17
11	68,58	66,67	24,04	7,37	1,92	75,50	73,95	19,15	5,35	1,55
12	68,74	66,91	23,88	7,37	1,83	75,43	73,64	19,15	5,43	1,78
13	68,94	67,16	23,75	7,31	1,78	75,43	73,57	18,99	5,58	1,86
14	67,73	65,49	24,85	7,42	2,23	73,41	69,61	21,32	5,27	3,80
15	68,61	66,46	24,11	7,28	2,14	74,88	71,78	19,77	5,35	3,10
16	69,24	67,43	23,41	7,35	1,80	76,43	74,88	18,45	5,12	1,55

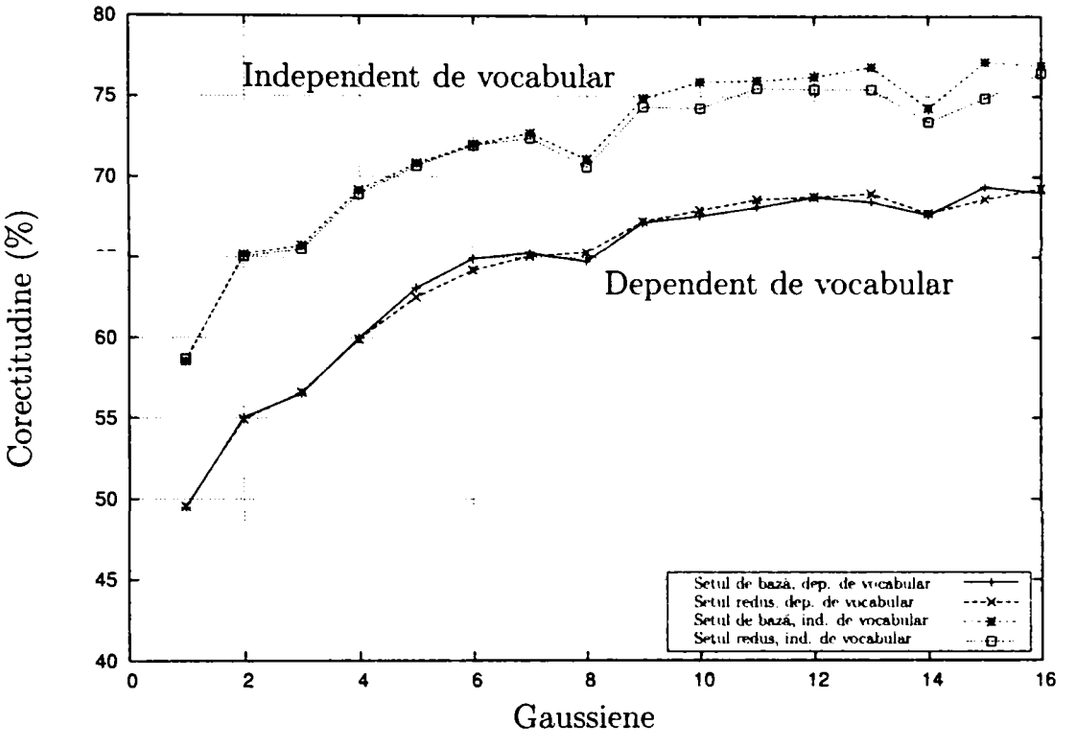

Figura 6.8: Evoluţia corectitudinii recunoaşterii cuvintelor funcţie de numărul de gaussiene/stare emiţătoare

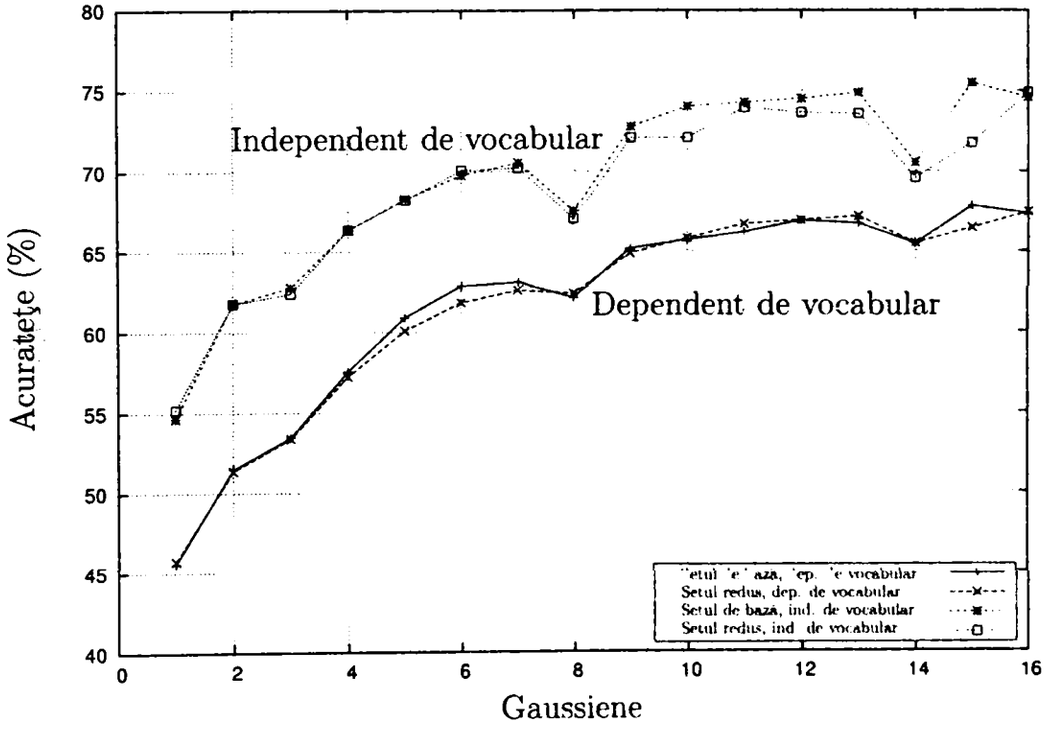

Figura 6.9: Evoluţia acurateţii recunoaşterii cuvintelor funcţie de numărul de gaussiene/stare emiţătoare

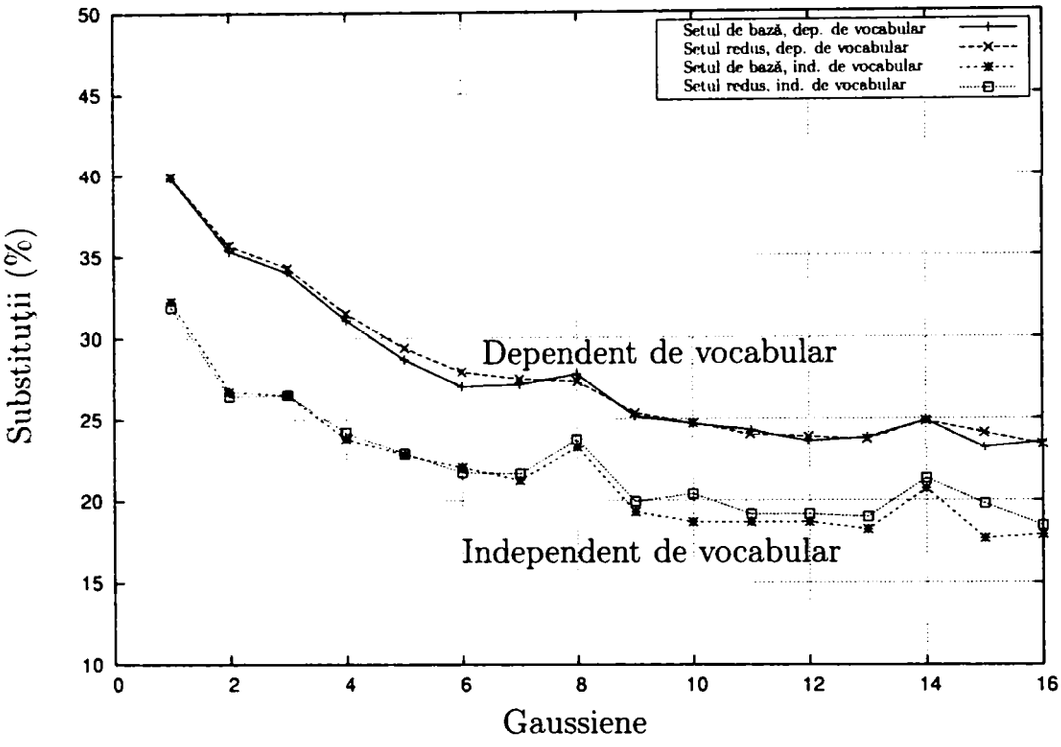

Figura 6.10: Evoluţia frecvenţei substituţiilor cuvintelor funcţie de numărul de gaussiene/stare emiţătoare

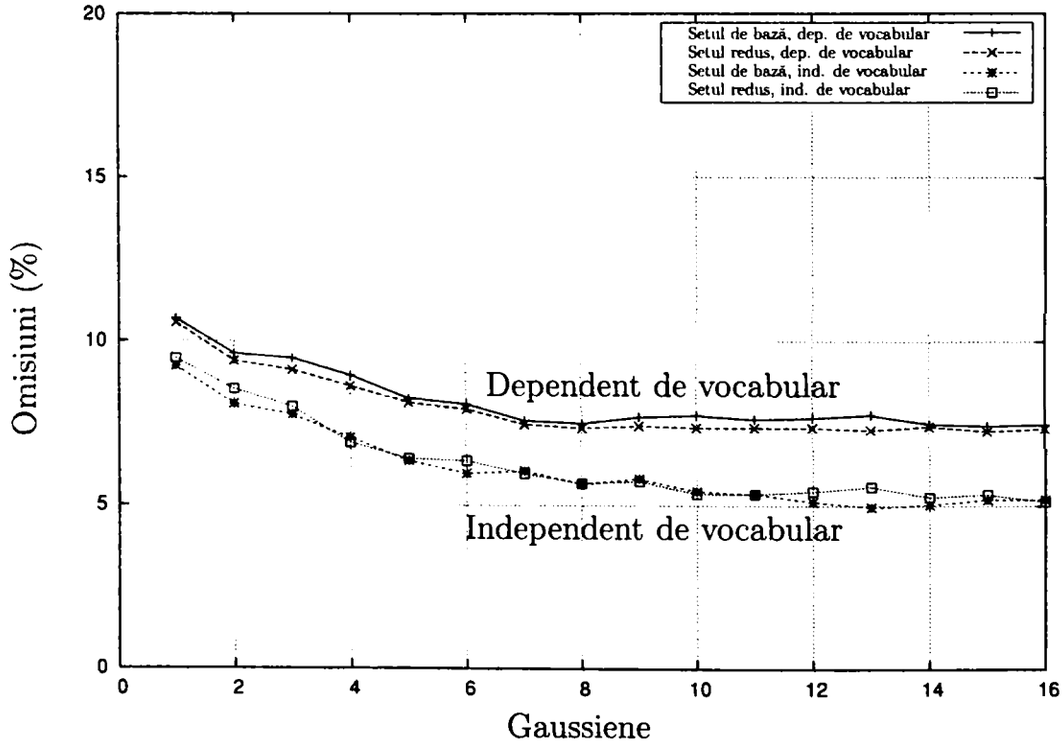

Figura 6.11: Evoluţia frecvenţei omisiunilor cuvintelor funcţie de numărul de gaussiene/stare emiţătoare

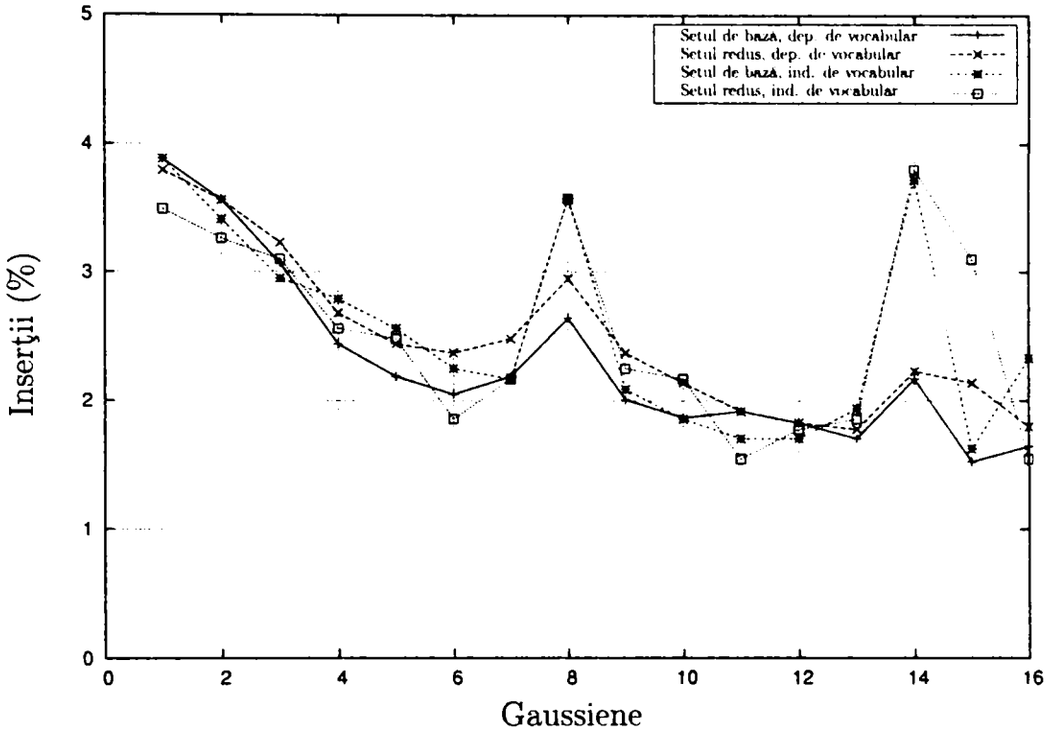

Figura 6.12: Evoluţia frecvenţei inserţiilor cuvintelor funcţie de numărul de gaussiene/stare emiţătoare

nesemnificative, singurele excepţii, favorabile setului de bază, apărând în cazul testelor dependente de vocabular cu 5, 6 şi 15 gaussiene/stare emiţătoare şi al celor independente de vocabular cu 10 şi 15 gaussiene/stare emiţătoare (tabelul 6.10).

Semnificaţia statistică redusă a diferenţelor poate fi însă considerată normală dacă ţinem cont şi de frecvenţa redusă de realizare a fonemului /I/, observată pe durata etichetării semnalelor (capitolul 5) şi confirmată şi de analiza datelor de antrenament: din cele 774 apariţii ale fonemului /I/, aşteptate pe baza pronunţiilor din dicţionar, doar 168 au fost efectiv realizate, corespunzător unei frecvenţe de realizare de cca. 21,7%.

Această situaţie ar putea fi explicată prin faptul că informaţia lingvistică presupusă a fi transmisă prin intermediul fonemului /I/ este în realitate partajată între diferitele niveluri ale comunicării verbale (acustic, sintactic, semantic, pragmatic etc.), aşa încât mesajele pot fi înţelese corect chiar şi atunci când acest fonem nu este realizat.

Tabelul 6.10: Semnificaţia statistică la nivelul $p = 0{,}05$ a diferenţelor dintre performanţele în recunoaşterea cuvintelor folosind setul de bază şi cel redus funcţie de numărul de gaussiene şi dependenţa de vocabular

Gaussiene	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Teste DV	–	–	–	–	+	+	–	–	–	–	–	–	–	–	+	–
Teste IV	–	–	–	–	–	–	–	–	–	+	–	–	–	–	+	–

6.5 Concluzii

Experimentele descrise în acest capitol au urmărit în primul rând validarea setului de unităţi sublexicale ales în faza iniţială a cercetărilor şi folosit pentru proiectarea bazei de date construite pe parcursul lor, dar elementele noi apărute pe durata etichetării semnalelor şi a experimentelor au indicat existenţa a două posibile seturi alternative de unităţi de modelare - unul incluzând consoane palatalizate şi altul redus, obţinut prin eliminarea /I/ - reflectând în fapt neclarităţile din literatura lingvistică referitoare la fonemele limbii române. Drept urmare, un al doilea obiectiv al experimentelor a fost compararea celor trei posibile seturi de unităţi de modelare, însă analiza datelor a arătat că acestea sunt suficiente doar pentru comparaţii între setul de bază şi cel redus.

Prima comparaţie a fost făcută pe durata experimentelor dependente de vorbitor, când utilizarea setului de bază a fost însoţită de erori atribuite problemelor de modelare a fonemului /I/, diminuate prin eliminarea lui şi utilizarea setului redus.

Compararea sistematică a celor două seturi, simultan cu evaluarea performanţelor la care conduc, a fost făcută printr-o serie de experimente independente de vorbitor. În plus, prin teste independente de vocabular, de recunoaştere a unor vocabulare semnificativ diferite de cel al datelor de antrenament, a fost evaluată şi capacitatea de generalizare a modelelor acustice care să permită utilizarea lor în sisteme cu vocabulare flexibile.

Experimentele independente de vorbitor au fost desfăşurate la nivelurile unităţilor de modelare şi cuvintelor: la nivelul unităţilor de modelare, rezultatele au demonstrat o bună capacitate de generalizare a modelelor în raport cu schimbarea vocabularului, iar utilizarea setului redus a condus la performanţe superioare.

Comparaţiile dintre cele două seturi de unităţi pe baza rezultatelor în recunoaşterea cuvintelor arată diferenţe în favoare setului de bază, manifestate în special independent de vocabular şi la o modelare destul de detaliată (peste opt gaussiene/stare emiţătoare), însă aceste diferenţe au o semnificaţie statistică redusă.

Neclarităţile din literatura lingvistică, pe baza căreia a fost făcută alegerea iniţială a unităţilor de modelare, au continuat deci şi pe parcursul acestor experimente: înlocuirea setului de bază cu cel redus a condus fie la ameliorarea rezultatelor (cazul experimentelor dependente de vorbitor şi al recunoaşterii unităţilor de modelare), fie la reduceri ale performanţelor, în general nesemnificative statistic (cazul recunoaşterii cuvintelor).

Cu toate neclarităţile încă persistente, experimentele au marcat primii paşi spre soluţionarea problemelor modelării acustice sublexicale pentru recunoaşterea vorbirii continue cu vocabulare mari şi foarte mari în limba română, iar rezultatele ca atare constituie un prim punct de referinţă pentru cercetările viitoare din domeniu.

CAPITOLUL 7

Încheiere

Această teză a descris primele cercetări (din câte cunoaştem) vizând recunoaşterea automată a vorbirii continue în limba română, pentru care obiectivul principal ales iniţial a fost recunoaşterea independentă de vorbitor a unor vocabulare în jurul a 1000 de cuvinte: dat fiind stadiul incipient al cercetărilor, acest obiectiv a fost considerat realist în condiţiile date, având în acelaşi timp avantajul de a fi semnificativ atât din punctul de vedere al dificultăţii problemei, cât şi din cel al aplicabilităţii practice a rezultatelor – de exemplu, în sisteme de dialog vocal om-maşină orientate pe domenii bine definite, pentru care o asemenea dimensiune a vocabularului poate fi suficientă.

Metodele statistice de recunoaştere a vorbirii, care s-au impus pe plan mondial şi au fost folosite şi în aceste cercetări, presupun utilizarea unor modele acustice pentru a descrie evoluţia în timp a proprietăţilor spectrale ale semnalelor vocale şi a unor modele lingvistice ale succesiunii cuvintelor, iar aceste cercetări au fost axate pe problemele modelării acustice, fundamentală în raport cu cea lingvistică.

Pentru recunoaşterea unui număr redus de cuvinte, modelele acustice pot fi construite separat pentru fiecare cuvânt, însă recunoaşterea vorbirii continue cu vocabulare mari şi foarte mari, de ordinul miilor sau zecilor de mii de cuvinte, impune utilizarea de modele acustice ale unor unităţi sublexicale (silabe, foneme etc.), în număr mult mai redus decât cel al cuvintelor din vocabular, antrenabile în consecinţă pe baza unor cantităţi mult mai reduse de semnale vocale, şi utilizabile şi în cazul schimbării vocabularului.

În principiu, cea mai economică şi în acelaşi timp cea mai generală modelare acustică sublexicală ar putea fi realizată utilizând setul de foneme ale unei limbi, care sunt prin definiţie categoriile abstracte ale sunetelor ei distinctive. În practică, diferenţele dintre proprietăţile spectrale ale alofonelor (variantele poziţionale de realizare ale fonemelor), cauzate de fenomenul de coarticulaţie, pot conduce la utilizarea unor seturi alofonice de unităţi de modelare acustică sublexicală, diferite de cel al fonemelor.

În cazul limbii române, utilizarea directă a unui set de foneme sau alofone ca unităţi de modelare acustică sublexicală nu este posibilă din cel puţin două motive: pe de o parte, literatura lingvistică menţionează mai multe seturi de foneme; pe de alta, insuficienţa studiilor de fonetică şi fonologie a limbii române, care ar fi permis clarificarea problemei setului de foneme şi ar fi fost utile şi pentru eventuale decizii de modelare alofonică.

În consecinţă, principala problemă investigată în această teză a fost cea a unităţilor de modelare acustică sublexicală, investigaţie din care au rezultat şi contribuţiile ei.

7.1 Contribuţii

Un prim grup de contribuţii ale tezei sunt cele legate de proiectarea şi colectarea bazei de date fonetice, prezentată în capitolul 4. Această bază de date este în sine o contribuţie esenţială, cea mai importantă pe termen lung: în primul rând prin cadrul pe care îl creează pentru experimente controlate de recunoaştere a vorbirii şi comparaţii între rezultatele lor, necesare în cursul cercetărilor, şi abia apoi prin materialul pe care îl furnizează pentru antrenarea modelelor acustice, testarea şi evaluarea sistemelor. În plus, datorită proiectării corespunzătoare a conţinutului lingvistic al unei părţi importante a materialelor înregistrate şi asigurarea unei foarte bune calităţi acustice a înregistrărilor, baza de date poate fi folosită şi în cercetările fundamentale de fonetică şi fonologie, a căror insuficienţă a fost semnalată, precum şi pentru alte cercetări aplicative. Contribuţiile legate de proiectarea bazei de date constau în:

- analiza literaturii lingvistice şi definirea unui set de unităţi fonetice de modelare acustică (tabelul 4.1) astfel încât direct sau prin combinaţii acestea să acopere toate elementele diferitelor seturi de foneme ale limbii române din literatură, elemente considerate potenţial semnificative pentru modelarea acustică sublexicală;

- un algoritm de grupare a pasajelor (algoritmul 4.1), în cazul cel mai general al unor materiale de citit în vederea înregistrării, care în limitele specifice limbajului natural asigură obţinerea unor grupuri de materiale cât mai apropiate între ele din punctul de vedere al distribuţiilor unităţilor fonetice de modelare acustică;

- o metodologie sistematică de proiectare, bazată pe gruparea materialelor folosind algoritmul menţionat şi distribuirea lor pe vorbitori conform unui experiment cu o structură bloc aleatoare, utilizând ca variabile de blocare criteriile de selecţie uniformă a vorbitorilor – în cazul nostru, sexul şi grupa de vârstă.

Al doilea grup de contribuţii este legat de etichetarea semnalelor vocale din baza de date, descrisă în capitolul 5. În cadrul cercetărilor asupra recunoaşterii automate a vorbirii, etichetarea facilitează antrenarea modelelor acustice şi asigură referinţele pentru evaluarea experimentelor de recunoaştere la nivelul unităţilor de modelare. Datorită însă accesului rapid pe care îl permite la realizările acestor unităţi şi informaţiilor de durată pe care le include, ea este utilă pentru multe alte cercetări, iar împreună cu celelalte caracteristici ale bazei de date (controlul conţinutului lingvistic şi al vorbitorilor înregistraţi şi calitatea înregistrărilor) îi conferă acesteia caracterul de bază de date fonetice. Contribuţiile din acest grup includ:

- dezvoltarea unui sistem de etichetare a semnalelor vocale care combină transcrierea lor fonetică manuală şi alinierea automată a transcrierilor fonetice cu semnalele asociate, folosit pentru a realiza etichetarea bazei de date la nivel fonetic extins;

- elaborarea unor criterii de decizie asupra identităţilor şi limitelor segmentelor de semnal, utilizate pe durata etichetării manuale şi/sau a verificării şi corectării etichetelor generate automat folosind sistemul dezvoltat anterior;

- etichetarea, verificarea şi corectarea etichetării majorităţii înregistrărilor.

Ultimul grup de contribuţii este legat de utilizarea efectivă în modelarea acustică a unităţilor sublexicale, prezentată în capitolul 6. Experimentele de recunoaştere a vorbirii continue descrise în acest capitol au fost efectuate folosind subseturi ale bazei de date, iar obiectivul lor era iniţial validarea setului de unităţi definit în etapa de proiectare a acesteia. Pe durata etichetării şi a experimentelor a apărut însă problema reducerii setului de unităţi, motiv pentru care un al doilea obiectiv urmărit a fost compararea setului de bază cu cel redus. Pentru evaluarea capacităţii de generalizare a modelelor la schimbarea vocabularului, majoritatea experimentelor au urmărit şi contrastul dintre dependenţa şi independenţa de vocabular. Contribuţiile din acest grup cuprind:

- dezvoltarea unor sisteme automate de recunoaştere a vorbirii continue în limba română, dependente şi independente de vorbitor, la nivel lexical şi sublexical, care au fost utilizate pentru efectuarea experimentelor;

- studiul şi evaluarea seturilor de bază şi redus de unităţi fonetice de modelare acustică sublexicală pentru recunoaşterea automată a vorbirii continue în limba română.

Experimentele nu au reuşit să clarifice diferenţele dintre setul de bază şi cel redus, această clarificare rămânând în sarcina continuărilor acestor cercetări.

7.2 Continuări

Deşi această teză cuprinde rezultatele câtorva ani de cercetări, ea nu a rezolvat toate problemele modelării acustice sublexicale pentru recunoaşterea vorbirii continue în limba română, iar din experienţa şi rezultatele de până acum reies şi câteva posibile continuări.

Datele colectate şi utilizate în cursul acestor prime cercetări asupra recunoaşterii automate a vorbirii continue în limba română au fost în mod necesar limitate cantitativ, accentul fiind pus pe calitatea lor acustică, astfel încât informaţia lingvistică din semnalul vocal să fie cât mai puţin afectată de zgomote, reverberaţii etc. Datorită acestei limitări, unele posibile alternative de modelare nu au putut fi abordate (secţiunea 6.3.2).

Chiar şi aşa, rezultatele obţinute sugerează posibilitatea ca setul de unităţi fonetice de modelare acustică ales iniţial (tabelul 4.1) să nu fie cel optim. O posibilă alternativă, care pe baza rezultatelor de până acum nu poate fi respinsă, este renunţarea la fonemul /I/: aşa cum rezultă din analiza datelor, acesta are o frecvenţa de realizare redusă, ceea ce ar putea indica faptul că informaţia lingvistică presupusă a fi transmisă prin intermediul lui este de fapt partajată între mai multe niveluri ale comunicării verbale.

Ca atare, clarificarea problemelor legate de acest aspect ar putea avea loc doar prin studiul simultan al problemelor modelării acustice şi lingvistice. Datorită limitării menţionate a datelor, încercările de utilizare în acest scop a unor modele statistice de tip bigram sau a unor gramatici de tip perechi-de-cuvinte (secţiunea 6.1.1) estimate pe baza

lor a dus la rezultate irelevante: date fiind perplexităţile foarte reduse ale acestor modele lingvistice, chiar cu modelele acustice cele mai simple, cu o gaussiană/stare emiţătoare, frecvenţa recunoaşterilor corecte şi acurateţea s-au situat peste 95%. Deşi în aparenţă foarte bune, asemenea rezultate sunt inutile deoarece nu permit comparaţii semnificative între diferite condiţii testate, şi în consecinţă nici nu au fost prezentate aici.

Studiul simultan al problemelor modelării acustice şi lingvistice presupune însă un efort preliminar pentru colectarea coordonată a textelor şi semnalelor vocale necesare, de exemplu plecând de la surse de texte disponibile pe Internet [23].

Date fiind dimensiunile considerabile ale unui asemenea efort, direcţia de continuare a cercetărilor cea mai accesibilă pentru moment este cea a modelării acustice dependente de context, care pe lângă certe ameliorări ale performanţelor ar putea aduce contribuţii şi la clarificarea problemei setului optim de unităţi de modelare.

Din experienţa acumulată se degajă şi unele probleme metodologice, legate de analiza şi diagnoza antrenării modelelor şi recunoaşterii şi evaluarea performanţelor. Studiul procesului de antrenare, de exemplu pentru a putea explica abaterile de la tendinţa de ameliorare a performanţelor din cazurile recunoaşterii cuvintelor cu MMA având 8 şi 14 gaussiene/stare emiţătoare (secţiunea 6.4.4), ar putea fi facilitat prin analiza şi vizualizarea interactivă a evoluţiei modelelor şi a datelor de antrenament.

Examinând exemplele de recunoaştere din anexa B, se constată că în numeroase cazuri erorile apar grupate şi nu sunt independente între ele: datorită constrângerilor reduse impuse în cazul lor de dicţionarele de pronunţii, cuvintele scurte sunt cele mai frecvent omise sau inserate, iar omisiunile şi inserţiile la rândul lor sunt adesea însoţite de substituţii. Analiza unor asemenea situaţii ţinând cont şi de pronunţiile cuvintelor şi similaritatea lor arată că de fapt pe lângă substituţii, omisiuni şi inserţii există şi erori de divizare sau contopire a cuvintelor. Identificarea acestora şi modificarea corespunzătoare a procesului şi metricelor de evaluare a performanţelor s-ar putea face recurgând la alinierea fonetico-fonologică a ipotezelor cu referinţele [188], [187], [72].

În sfârşit, diagnoza recunoaşterii în sine necesită recurgerea la metode de analiză a zonelor eronate, cu posibilitatea de clasificare şi stabilire automată a cauzelor lor [47].

ANEXA A

Detalii ale dicţionarelor

Utilizând setul de bază de unităţi de modelare acustică, au fost identificate următoarele perechi de cuvinte homofone:

- în datele pentru teste dependente de vocabular

Cuvânt	Homofon	Pronunţie
ce-i	cei	C e j
de-a	dea	d E a
ea	ia	j a
s-a	sa	s a
s-ar	sar	s a r
s-au	sau	s a w

- în datele pentru teste independente de vocabular

Cuvânt	Homofon	Pronunţie
ce-i	cei	C e j
s-a	sa	s a

Prin înlocuirea setului de bază cu cel redus de unităţi de modelare s-au mai adăugat câteva perechi de homofone:

- în datele pentru teste dependente de vocabular

Cuvânt	Homofon	Pronunţie
beţivan	beţivani	b e T i v a n

- în datele pentru teste independente de vocabular

Cuvânt	Homofon	Pronunţie
a-şi	aş	a S
bun	buni	b u n
m-ar	mari	m a r

ANEXA B

Exemple de recunoaştere

În continuare sunt prezentate rezultatele recunoaşterii independente de vocabular a cuvintelor folosind modele acustice cu 16 densităţi gaussiene/stare. Fiecare dintre cele 133 propoziţii individuale folosite este identificată prin codul vorbitorului, cel al materialului citit şi numărul ei de ordine în acest material – de exemplu, FA-AZ1 este prima dintre propoziţiile înregistrate de vorbitorul FA prin citirea materialului AZ.

Pentru fiecare propoziţie este prezentată transcrierea de referinţă aliniată cu ipotezele obţinute folosind setul de bază şi cel redus de unităţi de modelare acustică sublexicală. Nu au fost considerate erori substituţiile unor cuvinte homofone ale celor de referinţă. În rest, substituţiile şi inserţiile sunt *evidenţiate*, iar omisiunile – marcate prin ****.

Ref. FA-AZ1: nu se teme să-şi mărturisească slăbiciune după slăbiciune
Setul de bază: nu *să* teme să-şi mărturisească slăbiciune după slăbiciune
Setul redus: nu *să* teme să-şi mărturisească slăbiciune după slăbiciune

Ref. FA-AZ2: mai văzusem aşa ceva în nişte planşe reprezentând anatomia subtilă
Setul de bază: mai văzusem aşa ceva în nişte planşe reprezentând anatomia subtilă
Setul redus: mai văzusem aşa ceva *îmi* nişte planşe reprezentând anatomia subtilă
conform yogăi
conform yogăi
conform yogăi

Ref. FA-AZ3: şi zise că n-avea nici el vreo leţcaie
Setul de bază: şi zise că n-avea nici *şi-a* vreo leţcaie
Setul redus: şi zise că n-avea nici *şi-a* vreo leţcaie

Ref. FA-AZ4: respira gâfâit pe marginea propriei sale gropi
Setul de bază: respira gâfâit pe marginea propriei sale gropi
Setul redus: respira gâfâit *te* marginea *tot mi* sale gropi

Ref. FA-AZ5: până astăzi n-am râvnit la bun mai de preţ decât să cunosc cartea
Setul de bază: *urma* astăzi *m-am gâfâit m-a* bun mai *vă* preţ decât să cunosc cartea
Setul redus: *urma* astăzi *m-am gâfâit m-a* bun mai *vă* preţ decât să cunosc cartea

Ref. FA-AZ5: evreul era foarte cumsecade
Setul de bază: *te vreo ea* foarte cumsecade
Setul redus: *te vreo ea* foarte cumsecade

Ref. FA-AZ7: circumstanţele existenţei noastre sunt socotite de noi
Setul de bază: circumstanţele *insist în ţi* *noastră* sunt socotite ** *numai*
Setul redus: circumstanţele *insist îmi ţi* *noastră să-mi* socotite ** *numai*
toţi
toţi
toţi

Ref. FB-AB1: o gravă acuzaţie era aceea de subminare morală a statului
Setul de bază: o gravă acuzaţie era ***** *cere* subminare morală a statului
Setul redus: o gravă acuzaţie era ***** *cere* subminare morală a statului

Ref. FB-AB2: câţi dintre dânşii vorbeau măcar la nivelul unor discuţii de salon
Setul de bază: *că stâncă* dânşii vorbeau măcar la nivelul unor discuţii de salon
Setul redus: *că stâncă* dânşii vorbeau măcar la nivelul unor discuţii de salon

Ref. FB-AB3: ghemul timpului nu s-a desfăşurat în întregime nădăjduiesc
Setul de bază: ghemul timpului nu sa desfăşurat ** întregime nădăjduiesc
Setul redus: ghemul timpului nu sa desfăşurat ** întregime *văd întins*

Ref. FB-AB4: cu cât înaintez le văd mai mari şi goana lor mi se pare mai
Setul de bază: cu *către* înaintez ** *** *revăd m-am a-şi* goana lor mi *să* pare mai
Setul redus: cu *către* înaintez ** *** *revăd m-am a-şi* goana lor mi *să* pare mai
ciudată
ciudată
ciudată

Ref. FB-AB5: se opri acolo unde îndeobşte n-aveai voie s-o faci
Setul de bază: se opri acolo unde îndeobşte *m-a pe* voie s-o faci
Setul redus: se opri acolo unde îndeobşte *m-a te* voie s-o faci

Ref. FB-AB6: pentru a-şi desăvârşi penitenţa poetul refuza să i se mai
Setul de bază: *pe haz* desăvârşi penitenţa poetul refuza *să-i* să **** ** mai
Setul redus: *pe crezi* desăvârşi penitenţa poetul refuza ** *să-i* se mai
încălzească odaia
încălzească odaia
încălzească odaia

Ref. FC-AT1: statuia închipuia un subiect cu care sculptorii erau mai puţin hârşâiţi
Setul de bază: statuia închipuia *au* subiect cu care sculptorii erau *m-a* puţin hârşâiţi
Setul redus: statuia închipuia *au* subiect cu care sculptorii erau *m-a* puţin hârşâiţi

Ref. FC-AT2: intenţia spectacolului oferit de el era să-şi cruţe soaţa şi s-o rupă de
Setul de bază: intenţia spectacolului oferit *doi ele* era să-şi cruţe soaţa şi s-o rupă de
Setul redus: intenţia spectacolului oferit *doi ele* era să-şi cruţe soaţa şi s-o rupă de
soarta lui neagră
soarta lui neagră
soarta lui neagră

Ref. FC-AT3: nicăieri nu se înnegresc de dorinţa răzbunării
Setul de bază: nicăieri nu se înnegresc de dorinţa răzbunării
Setul redus: nicăieri nu se înnegresc de dorinţa răzbunării

Ref. FC-AT4: tudor mă urmărise abia stăpânindu-se să nu izbucnească-n râs
Setul de bază: tudor *m-a* urmărise abia stăpânindu-se ** *seama* izbucnească-n râs
Setul redus: tudor *m-a* urmărise abia stăpânindu-se să nu izbucnească-n râs

Ref. FC-AT5: la şcoală a izbucnit un adevărat scandal
Setul de bază: *era* şcoală *pare* izbucnit un adevărat scandal
Setul redus: la şcoală *ar* izbucnit un adevărat scandal

Ref. FC-AT6: mi-am făcut un program riguros de autoanaliză
Setul de bază: mi-am făcut un program riguros de autoanaliză
Setul redus: mi-am făcut un program riguros de autoanaliză

Ref. FC-AT7: cel dintâi fu expediat grabnic să întâmpine un oaspete oficial
Setul de bază: *ce dintre i* fu expediat grabnic să întâmpine *au* oaspete oficial
Setul redus: *ce dintre i* fu expediat grabnic să întâmpine *au* oaspete oficial

Ref. FD-AW1: orgoliul îmi şopteşte că putea fi şi mai rău
Setul de bază: orgoliul *în* şopteşte că putea fi şi *m-ar ou*
Setul redus: orgoliul *în* şopteşte că putea fi *şi-mi aer ou*

Ref. FD-AW2: ziaristul se schimonosi încercând să-şi ascundă râsul
Setul de bază: ziaristul se schimonosi *pe* încercând să-şi ascundă râsul
Setul redus: ziaristul se schimonosi încercând să-şi ascundă râsul

Ref. FD-AW3: m-am ridicat într-o rână am înşfăcat sticla de apă de lângă pat
Setul de bază: m-am ridicat într-o rână *apară-n* înşfăcat sticla de apă de lângă pat
Setul redus: m-am ridicat într-o rână am înşfăcat sticla de apă de lângă pat
şi am sorbit
pe şi am sorbit
şi am sorbit

Ref. FD-AW4: de astfel de vorbe se şi temea duhovnicul de la putna
Setul de bază: de astfel de vorbe se şi temea duhovnicul *de-l a* putna
Setul redus: de astfel de vorbe se şi temea duhovnicul *de-l a* putna

Ref. FD-AW5: congestionat şi greţos intenţionă să se arate ameninţător
Setul de bază: congestionat şi greţos intenţionă să *să* arate ameninţător
Setul redus: congestionat şi greţos intenţionă să *să* arate ameninţător

Ref. FD-AW6: dacă ieşeam împreună să-mi cumpere ciorapi sau o cămaşă
Setul de bază: *dată* ieşeam împreună *ar* să-mi cumpere ciorapi *** *său* cămaşă
Setul redus: dacă ieşeam împreună *ar* să-mi cumpere ciorapi *** *său* cămaşă
profitam s-o atrag într-o biserică
profitam *** *soarta* într-o biserică
profitam *** *soarta* într-o biserică

Ref. FD-AW7: o afirm cu mâhnire
Setul de bază: *ou* afirm cu mâhnire
Setul redus: *ou* afirm cu mâhnire

Ref. FE-BB1: aproape toţi cei de acolo se grăbeau să-şi dezică vieţile
Setul de bază: aproape toţi ce-i de acolo se grăbeau să-şi dezică vieţile
Setul redus: aproape toţi ce-i de acolo se grăbeau să-şi dezică vieţile

Ref. FE-BB2: până atunci nu avusesem prilejul să stau în preajma unui ins cu un
Setul de bază: până atunci nu avusesem prilejul *se* stau în preajma unui ins ** *cum*
Setul redus: *în* atunci nu avusesem prilejul *se* stau în preajma unui ins ** *cum*
atare handicap
atare handicap
atare handicap

Ref. FE-BB3: e sigur că se înmulţiseră ca niciodată urticariile de tot soiul
Setul de bază: *de* sigur că se înmulţiseră ca niciodată urticariile de *toţi* soiul
Setul redus: *îi* sigur ** *câţi* înmulţiseră ca niciodată urticariile de *toţi* soiul

Ref. FE-BB4: tot ce ne aduce imaginaţia e fals
Setul de bază: tot ce ne aduce imaginaţia *i* fals
Setul redus: tot ce ne aduce imaginaţia *i* fals

Ref. FE-BB5: eroul îi cere diavolului să-i arate lumea spirituală de care este
Setul de bază: eroul *i* cere diavolului *se* arate lumea spirituală de *care-i* este
Setul redus: eroul *i* cere diavolului *se* arate lumea spirituală de *care-i* este
înfometat
înfometat
înfometat

Ref. FE-BB6: îi vine să se rupă de această femeie care-i cere să se jertfească
Setul de bază: îi vine *să-l* se rupă de această femeie care-i cere *se* se jertfească
Setul redus: îi vine *să-l* se rupă de această femeie care-i cere *se* se jertfească
pentru ea
pentru ea
pentru ea

Ref. FE-BB7: îl revăd acolo înveşmântat în negru aşezat pe un scaun
Setul de bază: îl revăd acolo înveşmântat *e* negru aşezat pe un scaun
Setul redus: îl revăd acolo înveşmântat *e* negru aşezat *te* un scaun

Ref. FF-AG1: sfârşitul cărţii de faţă sosea grabnic
Setul de bază: sfârşitul cărţii de faţă sosea *dat mic*
Setul redus: sfârşitul cărţii de faţă sosea *dat mic*

Ref. FF-AG2: în redactare se bizuia pe nişte note luate chiar după ce se
Setul de bază: *până* redactare se bizuia pe nişte note luate chiar după ce se
Setul redus: *până* redactare se bizuia pe nişte note luate chiar după ce se
despărţise de gazda sa
despărţise de gazda sa
despărţise de gazda sa

Ref. FF-AG3: şperţul ocazionă întunecarea nădejdilor
Setul de bază: şperţul ocazionă întunecarea nădejdilor
Setul redus: şperţul ocazionă întunecarea nădejdilor

Ref. FF-AG4: scrierea ce o închei aici a bâjbâit în căutarea unui răspuns al
Setul de bază: scrierea ce *om ce-i* aici a bâjbâit în căutarea *mi* răspuns *ar*
Setul redus: scrierea ce *om ce-i* aici a bâjbâit în căutarea *mi* răspuns *ar*
acestei dileme
acestei dileme
acestei dileme

Ref. FF-AG5: perdeaua se umflase de o pală de aer
Setul de bază: perdeaua *să* umflase de o pală de *a-i*
Setul redus: perdeaua *să* umflase de o pală de *a-i*

Ref. FF-AG6: doar că nu ne-au aruncat cu lovituri de cizmă pe trepte-n jos
Setul de bază: **** **** *dacă lumea* aruncat cu lovituri de cizmă pe trepte-n jos
Setul redus: **** *dacă* nu ne-au aruncat cu lovituri de cizmă *te* trepte-n jos

Ref. FG-BI1: era dorinţa de a istorisi mereu aceleaşi întâmplări cu un haz reînnoit
Setul de bază: *ea* dorinţa de * istorisi mereu aceleaşi întâmplări ** *cum* haz reînnoit
Setul redus: *ea* dorinţa de * istorisi mereu aceleaşi întâmplări ** *cum* haz reînnoit

Ref. FG-BI2: tata s-ar fi înfuriat cu siguranţă şi poate mă pocnea
Setul de bază: tata *să* fi *vrea* cu siguranţă şi poate *m-a* pocnea
Setul redus: tata *să fiu fu ea* cu siguranţă şi poate *m-a* pocnea

Ref. FG-BI3: se pomenise preluând pe neaşteptate cea mai dificilă misiune
Setul de bază: se pomenise preluând *te* neaşteptate cea mai dificilă misiune
Setul redus: se pomenise preluând *te* neaşteptate cea mai dificilă misiune

Ref. FG-BI4: schiţă o figură de gimnastică să se dezmorţească după trezire
Setul de bază: schiţă *au* figură de gimnastică să se dezmorţească după trezire
Setul redus: schiţă *au* figură de gimnastică să se dezmorţească după trezire

Ref. FG-BI5: rămânea doar să înjghebeze atât de bine planul încât nimeni să nu-i
Setul de bază: rămânea doar să înjghebeze atât de bine planul încât *nu-i* să nu-i
Setul redus: rămânea *ba* să înjghebeze atât de bine planul încât *nu-i* să nu-i
înţeleagă jocul
înţeleagă jocul
înţeleagă jocul

Ref. FG-BI6: impulsurile pe care le resimţam îmi depăşeau puterile de stăpânire de
Setul de bază: impulsurile pe care le resimţam îmi depăşeau puterile de stăpânire de
Setul redus: impulsurile *te* care le resimţam îmi depăşeau puterile de stăpânire de
sine
sine
sine

Ref. FG-BI7: din prima după-amiază am dat fuga la mănăstirea locului
Setul de bază: din prima după-amiază *al* dat fuga la mănăstirea locului
Setul redus: din prima după-amiază *al* dat fuga la mănăstirea locului

Ref. FH-AY1: eu am rămas mirată fiindcă tata nu părea un familist
Setul de bază: *i* *om* rămas mirată fiindcă tata nu părea *om* familist
Setul redus: *i* *om* rămas mirată fiindcă tata nu părea *om* familist

Ref. FH-AY2: peste câtva timp m-am pomenit victimă a unui sumar interogatoriu
Setul de bază: peste câtva timp m-am pomenit victimă *au mi* sumar interogatoriu
Setul redus: peste câtva timp m-am pomenit victimă *au mi* sumar interogatoriu

Ref. FH-AY3: mă pofti să-l aştept
Setul de bază: mă pofti să-l aştept
Setul redus: mă pofti să-l aştept

Ref. FH-AY4: sunt sigur că-i era greu cu un nătăfleaţă ca mine
Setul de bază: **** *sus* *ion* era greu cu ** nătăfleaţă ca mine
Setul redus: **** *sus* *ion* era greu cu ** nătăfleaţă ca mine

Ref. FH-AY5: pleoapele ţi le ştergeai întruna de stropii de ulei ţâşniţi până la ele
Setul de bază: pleoapele ţi *de* ştergeai întruna de stropii de ulei ţâşniţi până *lui* ele
Setul redus: pleoapele ţi *de* ştergeai întruna de stropii de ulei ţâşniţi până *lui* ele

Ref. FH-AY6: din când în când îşi freca pleoapele până le roşea
Setul de bază: din când ** *încât* îşi freca *pleoape* până le roşea
Setul redus: din când ** *încât* îşi freca *pleoape* până le roşea

Ref. FH-AY7: îngrijea în chilia sa vreo optsprezece canari
Setul de bază: îngrijea *îmi* chilia sa *văd* optsprezece canari
Setul redus: îngrijea în chilia *sau* *văd* optsprezece canari

Ref. FI-BZ1: gazda mea se dovedea prea indulgentă cu ifosele copilului ce eram
Setul de bază: *un* gazda mea se dovedea prea indulgentă cu ifosele copilului *cel* *al*
Setul redus: *un* gazda mea se dovedea prea indulgentă cu ifosele copilului *cel* *al*

Ref. FI-BZ2: a cunoscut de mic foamea batjocura
Setul de bază: a cunoscut de mic foamea batjocura
Setul redus: a cunoscut de mic foamea batjocura

Ref. FI-BZ3: nu e mai bine să primească altul care nu e aşa de înduşmănit
Setul de bază: ** *ne* mai *vine* *se* primească altul care ** *nu-i* aşa de înduşmănit
Setul redus: ** *ne* mai *vine* *se* primească altul care ** *nu-i* aşa de înduşmănit
ca mine
ca mine
ca mine

Ref. FI-BZ4: stătea în spatele tejghelei cu lumânări de la intrarea bisericii
Setul de bază: stătea ** spatele tejghelei cu lumânări ** ** *înaintarea* bisericii
Setul redus: stătea ** spatele tejghelei cu lumânări ** ** *înaintarea* bisericii
umbroase
umbroase
umbroase

Ref. FI-BZ5: purta veşminte albe întreţesute cu fir de aur
Setul de bază: purta veşminte albe întreţesute cu fir de aur
Setul redus: purta veşminte albe întreţesute cu fir de aur

Ref. FI-BZ6: eram plin de nădejdi
Setul de bază: *te aur* plin *unde* nădejdi
Setul redus: *te aur* plin de nădejdi

Ref. FI-BZ7: tudor mă întrerupse pentru prima dată în cursul acelei după-amieze
Setul de bază: tudor ** întrerupse pentru prima dată în cursul acelei după-amieze
Setul redus: tudor *m-a* întrerupse pentru prima dată în cursul acelei după-amieze

Ref. FJ-BC1: acestea l-au impus unui for internaţional acesta angajându-l
Setul de bază: acestea l-au impus *îi* for internaţional acesta angajându-l
Setul redus: acestea l-au impus *îi* for internaţional acesta angajându-l

Ref. FJ-BC2: era de dorit să ne doboare în număr cât mai mare
Setul de bază: *ea* de *doi* *ţi* ne doboare ** număr *câţi mă* mare
Setul redus: *ea* de *doi* *se* ne doboare ** număr *câţi mă* mare

Ref. FJ-BC3: iarăşi simţea nevoia să se ducă la toaletă
Setul de bază: *iară* simţea *ne vă* să se ducă la toaletă
Setul redus: iarăşi simţea *ne* *vă sus* ducă la toaletă

Ref. FJ-BC4: poate că ar trebui să vă adresaţi altcuiva
Setul de bază: poate ** *ca* trebui să vă adresaţi altcuiva
Setul redus: poate ** *ca* trebui să vă adresaţi altcuiva

Ref. FJ-BC5: vine cu ciomagul subsuoară
Setul de bază: vine cu ciomagul subsuoară
Setul redus: vine cu ciomagul subsuoară

Ref. FJ-BC6: a lucrat un costum naţional ales a-i fi înmânat însăşi reginei
Setul de bază: a lucrat *în* costum naţional **** *alese* fi înmânat însăşi reginei
Setul redus: a lucrat *îmi* costum naţional **** *alese* fi înmânat însăşi reginei

Ref. FJ-BC7: vasele de porţelan se zdrobeau de parchet rămânând fără viaţă
Setul de bază: vasele de porţelan se zdrobeau de parchet *pe* rămânând fără viaţă
Setul redus: vasele de porţelan se zdrobeau de parchet rămânând fără viaţă

Ref. MA-AX1: făcusem această alegere după matură chibzuinţă
Setul de bază: făcusem această alegere *vă* matură chibzuinţă
Setul redus: făcusem această alegere *vă* matură chibzuinţă

Ref. MA-AX2: când îl zăreai îţi dădeai cu părerea că ar fi fost un om
Setul de bază: când *e* zăreai îţi dădeai cu părerea ** ** *cărţii* fost ** *m-am*
Setul redus: când *e* zăreai îţi dădeai cu părerea ** ** *cărţii* fost ** *m-am*
cumsecade
cumsecade
cumsecade

Ref. MA-AX3: exasperat a apucat să-i spună că nu părăseşte cabinetul fără
Setul de bază: exasperat * apucat *se* spună că nu părăseşte cabinetul fără
Setul redus: exasperat * apucat *se* spună ** *cum* părăseşte cabinetul fără
aprobarea râvnită
aprobarea râvnită
aprobarea râvnită

Ref. MA-AX4: colonelul tăcea respectuos
Setul de bază: colonelul tăcea respectuos
Setul redus: colonelul tăcea respectuos

Ref. MA-AX5: a trebuit să insist mult cu răbdare şi blândeţe
Setul de bază: a *trebui se* insist mult cu răbdare şi blândeţe
Setul redus: * *atrag uiţi* insist mult cu răbdare şi blândeţe

Ref. MA-AX6: iară eu beau vin şi ţuică
Setul de bază: iară eu beau *vine* şi ţuică
Setul redus: iară eu beau vin şi ţuică

Ref. MA-AX7: mi-a venit să pufnesc în râs
Setul de bază: **** **** *devenise* pufnesc ** *ales*
Setul redus: **** **** *devenise* pufnesc ** *ales*

Ref. MB-AC1: apariţia unui profesor însoţitor le cenzurează instinctele dezlănţuite
Setul de bază: apariţia unui profesor însoţitor le cenzurează *e stâncă* dezlănţuite
Setul redus: apariţia unui profesor însoţitor le cenzurează *e stâncă* dezlănţuite

Ref. MB-AC2: în cele din urmă m-a încredinţat că dacă se elibera postul urma
Setul de bază: *un* cele *ghemul om a* încredinţat că dacă se elibera postul urma
Setul redus: *un* cele *ghemul om a* încredinţat că dacă se elibera postul urma
să fiu întrebat de-l mai râvneam
să fiu întrebat de-l mai râvneam
să fiu întrebat de-l mai râvneam

Ref. MB-AC3: pe ea vor stăpînii s-o slăbească pentru a o putea manipula în
Setul de bază: ** *mi-a* vor stăpînii s-o slăbească ****** *pe într-o* putea manipula în
Setul redus: ** *mi-a* vor stăpînii s-o slăbească ****** *te într-o* putea manipula în
interesul lor
interesul lor
interesul lor

Ref. MB-AC4: există şi alte slujbe acolo
Setul de bază: există şi alte slujbe acolo
Setul redus: există şi alte slujbe acolo

Ref. MB-AC5: şi astfel odiseea luă sfârşit
Setul de bază: şi *noastră* odiseea luă sfârşit
Setul redus: şi *noastră* odiseea *nu* sfârşit

Ref. MB-AC6: murise cam pe la optsprezece ani şi-mi devenise model
Setul de bază: murise cam ** *până* optsprezece *al* şi-mi devenise model
Setul redus: murise cam ** *părea* optsprezece ani şi-mi devenise model

Ref. MC-AA1: a ajuns să-l vadă pe poetul voiculescu întins în patul de acasă
Setul de bază: * ajuns *al* vadă *apă* poetul voiculescu întins în patul de acasă
Setul redus: * ajuns *al* vadă *că* poetul voiculescu întins în patul de acasă

Ref. MC-AA2: crezi că pantelimon arată a pustnic
Setul de bază: crezi că pantelimon arată a pustnic
Setul redus: crezi că pantelimon arată a pustnic

Ref. MC-AA3: doar eu mă simţeam niţel neliniştit de zâmbetul său enigmatic
Setul de bază: doar eu mă *simţea* niţel neliniştit de zâmbetul *s-o* enigmatic
Setul redus: doar eu mă *simţea* niţel neliniştit de zâmbetul *s-o* enigmatic

Ref. MC-AA4: şi-a luat lădiţe goale în juru-i
Setul de bază: şi-a luat lădiţe *boală* în juru-i
Setul redus: şi-a luat lădiţe *boală* în juru-i

Ref. MC-AA5: ieşi pe şleaul ce ducea drept spre drumul de la suliţa la botoşani
Setul de bază: ieşi pe şleaul ce ducea drept spre drumul *de-l a* suliţa la botoşani
Setul redus: ieşi pe şleaul ce ducea drept spre drumul *de-l a* suliţa la botoşani

Ref. MC-AA6: anchetatorul a dat ordin unui gealat să mă lovească la tălpi
Setul de bază: anchetatorul a dat ordin unui gealat ** *seama* lovească la tălpi
Setul redus: anchetatorul a dat ordin unui gealat ** *să-mi* lovească la tălpi

Ref. MD-AK1: hotărâsem să mă inspir din numele său în alegerea
pseudonimului meu
Setul de bază: hotărâsem să mă inspir din numele său *un alegere*
pseudonimului *i* meu
Setul redus: hotărâsem *să-mi îi* inspir din numele *** *sub alegere*
pseudonimului *i* meu

Ref. MD-AK2: îşi luă măsurile de prevedere esenţial era să salveze pentru
viitorime comoara de documente
Setul de bază: îşi *o* măsurile de prevedere *pe* esenţial era să salveze pentru
viitorime comoara de documente
Setul redus: îşi *l-au* măsurile de prevedere esenţial era să salveze pentru
viitorime comoara de documente

Ref. MD-AK3: tocmai eu m-am găsit să judec proorocirea sa cu privire la viitorul
Setul de bază: tocmai *om am* găsit *se* judec proorocirea sa cu privire ** *lovituri*
Setul redus: tocmai *om am* găsit *se* judec proorocirea sa cu privire ** *lovituri*
meu
meu
meu

Ref. MD-AK4: erau buni de chefuri la bucurii cu care se rugau în aceeaşi
Setul de bază: *i rău bun bec circulă* bucurii cu care *să-l* rugau *un* aceeaşi
Setul redus: *i rău bun bec circulă* bucurii cu care *să-l* rugau *un aceşti*
obşte parohială
obşte *planul iară*
obşte *planul iară*

Ref. MD-AK5: ion suferă de o boală cronică şi mortală
Setul de bază: ion suferă ** * *doboare* cronică şi mortală
Setul redus: ion suferă ** * *doboare* cronică şi mortală

Ref. MD-AK6: am îngenuncheat
Setul de bază: am îngenuncheat
Setul redus: am îngenuncheat

Ref. ME-BN1: numai că respectivul student francez era în stipendiul securităţii
Setul de bază: numai că respectivul student francez era în stipendiul *îmi* securităţii
Setul redus: numai că respectivul student francez era în stipendiul securităţii

Ref. ME-BN2: văzusem o spinare sumeţindu-se de sub fiarele unui pat
Setul de bază: văzusem o spinare *tălpi* sumeţindu-se de *şi-mi* fiarele unui pat
Setul redus: văzusem o spinare sumeţindu-se de *şi-mi* fiarele unui pat

Ref. ME-BN3: acum ni se adresă cu o temă vrednică de inteligenţa noastră
Setul de bază: acum ** se adresă cu o temă vrednică de inteligenţa noastră
Setul redus: acum ** se adresă cu o temă vrednică de inteligenţa noastră

Ref. ME-BN4: şleahta nemaifiind un avanpost al societăţii a dispărut
Setul de bază: şleahta *pe* nemaifiind *putna* avanpost *ar* societăţii *te* dispărut
Setul redus: şleahta nemaifiind *una* avanpost *ar* societăţii *trei* dispărut

Ref. ME-BN5: stareţul părintele arsenie mi se înfăţişă ca o siluetă fără
Setul de bază: stareţul *îmi* părintele arsenie *plin* se înfăţişă ** *planul* siluetă fără
Setul redus: stareţul părintele arsenie mi se înfăţişă *că aur* siluetă fără
pereche
pereche
pereche

Ref. ME-BN6: am constatat seninul bunătăţii ce-i emana dintre pleoape
Setul de bază: am constatat seninul bunătăţii ce-i emana dintre pleoape
Setul redus: am constatat seninul bunătăţii ce-i emana dintre pleoape

Ref. ME-BN7: am socotit visul de mai sus ca o şoaptă încurajatoare
Setul de bază: *ani* socotit *i* visul de *m-ar şi uiţi pe au* şoaptă încurajatoare
Setul redus: *ani* socotit *i* visul de *m-ar* sus ** *rău* şoaptă încurajatoare
a subconştientului
aş subconştientului
a-şi subconştientului

Ref. MF-AD1: venerabilul îmi dezleagă misterul smeririi necesare tuturor
Setul de bază: venerabilul *în* dezleagă misterul smeririi necesare tuturor
Setul redus: venerabilul îmi dezleagă misterul smeririi necesare tuturor

Ref. MF-AD2: guvernul nu mai vrea să înlocuiască după ce vor fi obţinut
Setul de bază: guvernul ** *număr a s-o* înlocuiască după ** *ceva* fi obţinut
Setul redus: guvernul ** *număr a s-o* înlocuiască după ** *ceva* fi obţinut
recunoaşterea
recunoaşterea
recunoaşterea

Ref. MF-AD3: ajunsesem absolut îngrozitor de suportat de către cei care mă
Setul de bază: ajunsesem absolut îngrozitor *vă* suportat de către ** *ce-i cam*
Setul redus: ajunsesem absolut îngrozitor de suportat de către ** *ce-i cam*
acceptau în preajma lor
acceptau ** preajma lor
acceptau ** preajma lor

Ref. MF-AD4: imaturi o duceam într-un permanent duel verbal
Setul de bază: imaturi o duceam într-un permanent *de-l* verbal
Setul redus: imaturi *tot ducea* într-un permanent *de-l* verbal

Ref. MF-AD5: au trăit-o ca atare cu belşugul dragostei de fraţi
Setul de bază: *o* trăit-o ca *pare* cu belşugul dragostei de fraţi
Setul redus: *o* trăit-o ca *pare* cu belşugul dragostei de fraţi

Ref. MF-AD6: avusese o îndelungată perioadă de refacere după o boală
Setul de bază: avusese o îndelungată perioadă *vă* refacere după o *goală*
Setul redus: avusese o îndelungată perioadă *vă* refacere după *ou* boală

Ref. MG-AH1: ba mai găsesc câte una goală prin care nici nu se circulă
Setul de bază: ba mai găsesc *** *întruna* goală prin care nici *smuls e* circulă
Setul redus: ba *m-a* găsesc *** *întruna boală atunci* care nici *smuls e* circulă

Ref. MG-AH2: caut un pamflet ce mi-a apărut în evenimentul în urmă cu
Setul de bază: caut un pamflet ** *cele* apărut în evenimentul ** *pe loveaşcă*
Setul redus: caut un pamflet ** *cele* apărut în evenimentul ** *lumea* cu
doi trei ani
doi trei *ar*
doi trei *ar*

Ref. MG-AH3: atrage atenţia asupra greutăţilor întâmpinate în viaţă
Setul de bază: atrage atenţia *stau* greutăţilor întâmpinate *lui vrea te*
Setul redus: atrage atenţia *stau* greutăţilor întâmpinate *lui* viaţă

Ref. MG-AH4: ghemuit în patru labe se sprijini anevoie de bara de fier şi-şi scoase
Setul de bază: ghemuit în *patul* labe se sprijini anevoie de bara de fier *şi* scoase
Setul redus: ghemuit în *patul* labe se sprijini anevoie de bara de fier şi-şi scoase
capul la vedere
capul *a* vedere
capul *a* vedere

Ref. MG-AH5: nu coboară
Setul de bază: *l-au* coboară
Setul redus: *l-au* coboară

Ref. MG-AH6: dacă nu s-ar fi ţinut seama de dorinţa lui se risca provocarea
Setul de bază: dacă ** *** *râs oficial* seama de dorinţa lui se risca provocarea
Setul redus: dacă ** *** *râs oficial* seama de dorinţa lui se risca provocarea
unui incendiu
ulei incendiu
ulei incendiu

Ref. MH-AR1: îi jignisem prin cuvintele prea pripit alese
Setul de bază: *îl* jignisem prin cuvintele prea pripit alese
Setul redus: îi jignisem prin cuvintele prea pripit alese

Ref. MH-AR2: lucrurile nu stau întocmai cum le-am descris
Setul de bază: lucrurile nu stau întocmai *cu* le-am descris
Setul redus: lucrurile nu stau întocmai *cu* le-am descris

Ref. MH-AR3: anghel a smuls din mâinile soldatului puşca
Setul de bază: anghel * ***** *astăzi* mâinile soldatului puşca
Setul redus: anghel * ***** *astăzi* mâinile soldatului puşca

Ref. MH-AR4: în interiorul cărţii se desfăcea o planşă ce reprezenta trupul
Setul de bază: *părintele o* cărţii se desfăcea * *planşe* ce reprezenta trupul
Setul redus: *părintele o* cărţii se desfăcea * *planşe* ce reprezenta trupul
omului
ou vor
ou vor

Ref. MH-AR5: trebuia să te uiţi la vârful picioarelor sau în bec
Setul de bază: trebuia *se* te uiţi la vârful picioarelor sau *până* bec
Setul redus: trebuia *se* te uiţi la vârful picioarelor sau în bec

Ref. MH-AR6: îşi lipeşte din nou uităturile în acelaşi timp grele şi pehlivane
Setul de bază: îşi lipeşte din *** uităturile *până* acelaşi timp grele şi pehlivane
Setul redus: îşi lipeşte din *** uităturile în acelaşi timp grele şi pehlivane

Ref. MH-AR7: din micul bloc de alături se auzi un zgomot
Setul de bază: din micul **** *vorbe* alături *să* auzi *nu* zgomot
Setul redus: din micul **** *vorbe* alături *să* auzi un zgomot

Ref. MI-BA1: nu îndrăzneau să apară-n lumea oamenilor mari unde eu eram
Setul de bază: *nou* îndrăzneau să apară-n lumea oamenilor mari *închei o* eram
Setul redus: *nou* îndrăzneau *soarta vă* lumea oamenilor m-ar *închei o* eram
admis
admis
admis

Ref. MI-BA2: m-am îngrijorat deoarece ştiam cât erau de urâţi aceşti ortodocşi
Setul de bază: m-am îngrijorat deoarece ştiam *pripit* erau de urâţi aceşti ortodocşi
Setul redus: m-am îngrijorat deoarece ştiam *te te* erau de urâţi aceşti ortodocşi
care nu renunţau
care nu renunţau
care nu renunţau

Ref. MI-BA3: mă zguduie fiori de emoţie cu mult timp înainte de întâlnire
Setul de bază: mă zguduie fiori de emoţie cu **** timp înainte de întâlnire
Setul redus: mă zguduie fiori de emoţie cu **** timp înainte de întâlnire

Ref. MI-BA4: mustrări şi sfaturi legate de frecvenţa lui se îngrămădeau la gura mea
Setul de bază: mustrări şi sfaturi legate de frecvenţa lui se îngrămădeau la gura *mă*
Setul redus: mustrări şi sfaturi legate de frecvenţa lui se îngrămădeau la gura *mă*

Ref. MI-BA5: am insistat continuând a crede în cinstea sa
Setul de bază: am insistat *puţin* *una* crede în cinstea sa
Setul redus: am insistat continuând a crede în cinstea sa

Ref. MI-BA6: în urma incidentului fu convocat un consiliu profesoral
Setul de bază: în urma incidentului fu convocat un consiliu profesoral
Setul redus: în urma incidentului fu convocat un consiliu profesoral

Ref. MI-BA7: bulgării mari de pământ îmi fac înaintarea anevoioasă
Setul de bază: bulgării **** *mare* pământ *până* fac înaintarea anevoioasă *pe*
Setul redus: bulgării **** *mare* pământ *timp* fac înaintarea anevoioasă
împiedicată şchioapă
împiedicată *optsprezece apă*
împiedicată şchioapă

Ref. MJ-BE1: la toate întrebările procurorul consemnase că voiculescu nega
Setul de bază: *a pat* întrebările *ani* procurorul consemnase *cu* voiculescu *n-am*
Setul redus: *a* toate întrebările *i* procurorul consemnase *cu* voiculescu *n-am*
acuzaţia
acuzaţie
acuzaţie

Ref. MJ-BE2: pleoapele îi acoperă parţial ochii mari albaştri
Setul de bază: ******** *pleoape* acoperă parţial ochii mari albaştri
Setul redus: ******** *pleoape* acoperă parţial ochii *mare* albaştri

Ref. MJ-BE3: stătea singur pe stâncă îmbrăţişând cerul cu braţele
Setul de bază: stătea singur pe stâncă îmbrăţişând cerul cu *gura fără*
Setul redus: stătea singur *te* stâncă îmbrăţişând cerul cu *dată*

Ref. MJ-BE4: dacă nu m-ar fi stăpânit amorţirea vodcii aş fi făcut multe prostii
Setul de bază: dacă ** *număr* fi stăpânit amorţirea vodcii aş ** făcut multe prostii
Setul redus: dacă ** *număr* fi stăpânit amorţirea vodcii a-şi *fu cu* multe prostii

Ref. MJ-BE5: am luat o gură de apă să-mi potolesc arşiţa gâtlejului uscat
Setul de bază: ** **** *arată gura* de apă să-mi potolesc arşiţa gâtlejului uscat
Setul redus: ** *arată un* *gura* de apă să-mi potolesc arşiţa gâtlejului uscat

Ref. MJ-BE6: au fost transferaţi împreună cu un grup de delincvenţi de drept comun
Setul de bază: *o* fost transferaţi *întruna* cu ** **** de delincvenţi de drept comun
Setul redus: *o* fost transferaţi *întruna* cu ** **** de delincvenţi de drept comun

Ref. MJ-BE7: am mâncat un ou fiert
Setul de bază: *a* mâncat *una* ou fiert
Setul redus: *a* mâncat un ou fiert

BIBLIOGRAFIE

[1] J. Allen. *Natural Language Understanding.* Benjamin/Cummings, Redwood City, California, 1995.

[2] J. Allen, M.S. Hunnicut şi D. Klatt. *From text to speech: The MITalk system.* Cambridge University Press, 1987.

[3] J.B. Allen. How Do Humans Process and Recognize Speech? În R.P. Ramachandran şi R.J. Mammone (editori), *Modern Methods of Speech Processing*, cap. 11, pag. 251–75. Kluwer Academic Publishers, Boston, 1995.

[4] F. Alleva, X.D. Huang şi M.Y. Hwang. An Improved Search Algorithm Using Incremental Knowledge for Continuous Speech Recognition. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pag. 307–310, 1993.

[5] F. Alleva, X.D. Huang şi M.Y. Hwang. Improvements on the Pronunciation Prefix Tree Search Organization. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pag. 133–36, 1996.

[6] M.R. Anderberg. *Cluster Analysis for Applications.* Academic Press, New York, 1973.

[7] B.S. Atal şi S.L. Hanauer. Speech Analysis and Synthesis by Linear Prediction of the Speech Wave. *The Journal of the Acoustical Society of America*, 50(2):637–55, 1971.

[8] X.L. Aubert. A Brief Overview of Decoding Techniques for Large Vocabulary Continuous Speech Recognition. În *Proceedings of the ISCA ITRW ASR2000*, pag. 91–96, Paris, 2000.

[9] S. Austin, R. Schwartz şi P. Placeway. The Forward-Backward Search Algorithm. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pag. 697–700, 1991.

[10] A. Averbuch ş.a. Experiments with the Tangora 20,000 Word Speech Recognizer. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pag. 701–704, 1987.

[11] L.R. Bahl, P.F. Brown, P.V. de Souza şi R.L. Mercer. Speech recognition with continuous-parameter hidden Markov models. *Computer Speech and Language*, 2(3/4):219–34, 1987.

[12] L.R. Bahl, S.V. De Gennaro, P.S. Gopalakrishnan şi R.L. Mercer. A Fast Approximate Acoustic Match for Large Vocabulary Speech Recognition. În *Proceedings EUROSPEECH'89*, vol. 1, pag. 156–58, 1989.

[13] L.R. Bahl şi F. Jelinek. Decoding for Channels with Insertions, Deletions, and Substitutions with Applications to Speech Recognition. *IEEE Transactions on Information Theory*, 21(4):404–11, iulie 1975.

[14] L.R. Bahl, F. Jelinek şi R.L. Mercer. A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–90, martie 1983.

[15] J.K. Baker. *Stochastic Modeling as A Means of Automatic Speech Recognition.* Teză de doctorat, Carnegie Mellon University, aprilie 1975.

[16] J.K. Baker. The DRAGON System – An Overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):24–29, februarie 1975.

[17] W. Barry. Labelling criteria: Phonemic and acoustic-segment labelling. ESPRIT Project 2589 (SAM) Report, University College, Londra, octombrie 1990.

[18] W.J. Barry şi A.J. Fourcin. Levels of labelling. *Computer Speech and Language*, 6(1):1–14, ianuarie 1992.

[19] L.E. Baum şi T. Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *Annals of Mathematical Statistics*, 37(6):1554–63, decembrie 1966.

[20] L.E. Baum, T. Petrie, G. Soules şi N. Weiss. A Maximization Technique Ocurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *Annals of Mathematical Statistics*, 41(1):164–71, februarie 1970.

[21] R. Bellman şi S. Dreyfus. *Programarea dinamică aplicată.* Editura Tehnică, Bucureşti, 1967.

[22] L.L. Beranek. *Acoustic Measurements.* Wiley, New York, 1949.

[23] D. Bohuş şi **M. Boldea**. A Web-based Text Corpora Development System. În *Proceedings Second International Conference on Language Resources and Evaluation - LREC2000*, Atena, Grecia, mai 2000.

[24] D. Bohuş şi **M. Boldea**. Stochastic Speech Understanding for Human-Computer Dialogue. *Romanian Journal of Information Science and Technology*, 4(3-4):261-72, 2001.

[25] M. Boldea. A Database of Spoken Romanian Isolated Digits. În *Buletinul Ştiinţific al Universităţii Tehnice din Timişoara*, vol. 40(54) din *Seria Automatică şi Calculatoare*, pag. 135–38. Timişoara, 1995.

[26] M. Boldea. Speaker Independent Isolated Word Recognition Experiments. În *Buletinul Ştiinţific al Universităţii Tehnice din Timişoara*, vol. 40(54) din *Seria Automatică şi Calculatoare*, pag. 129–34. Timişoara, 1995.

[27] M. Boldea. A Comparison of Speech Processing Methods in Speaker Independent Isolated Word Recognition. În *Buletinul Ştiinţific al Universităţii "Politehnica" din Timişoara*, vol. 41(55) din *Seria Automatică şi Calculatoare*, pag. 164–70. Timişoara, 1996.

[28] M. Boldea. Speech Technology Research at Computer Science Department, "Politehnica" University of Timişoara. În D. Tufiş şi P. Andersen (editori), *Recent Advances in Romanian Language Technology*, pag. 174–77, Bucureşti, 1997. Editura Academiei Române.

[29] M. Boldea. Analiza semnalului vocal pentru recunoaşterea automată a vorbirii. Referat de doctorat, Departamentul de Calculatoare, Universitatea "Politehnica" din Timişoara, decembrie 1999.

[30] M. Boldea. Speaker Independent Phoneme Recognition in Romanian. În *Proceedings 12th International Conference on Control Systems and Computer Science - CSCS12*, vol. 2, pag. 7–12, Bucureşti, mai 1999.

[31] M. Boldea şi A. Doroga. Towards Automatic Recognition of Continuous Speech in Romanian. În *Proceedings Third International Conference on Technical Informatics - CONTI'98*, vol. 3, pag. 216–25, Timişoara, octombrie 1998.

[32] M. Boldea, A. Doroga, T. Dumitrescu şi M. Pescaru. Preliminaries to a Romanian Speech Database. În *Proceedings International Conference on Spoken Language Processing*, vol. 3, pag. 1934–37, Philadelphia, octombrie 1996.

[33] M. Boldea şi C. Munteanu. Labeling a Romanian Speech Database. În *Proceedings Second International Workshop "Speech and Computer" - SPECOM'97*, pag. 77–80, Cluj-Napoca, octombrie 1997.

[34] M. Boldea, C. Munteanu şi A. Doroga. Design, Collection, and Annotation of a Romanian Speech Database. În *Proceedings LREC Workshop on Speech Database Development for Central and Eastern European Languages*, pag. 43–46, Granada, Spania, mai 1998.

[35] A. Bonafonte, R. Estany şi E. Vives. Study of Subword Units for Spanish Speech Recognition. În *Proceedings EUROSPEECH'95*, pag. 1607–10, Madrid, septembrie 1995.

[36] H. Bourlard şi N. Morgan. Continuous Speech Recognition by Connectionist Statistical Methods. *IEEE Transactions on Neural Networks*, 4(6):893–909, noiembrie 1993.

[37] G.E.P. Box, W.G. Hunter şi J.S. Hunter. *Statistics for Experimenters.* John Wiley & Sons, New York, 1978.

[38] J.S. Bridle, M.D. Brown şi R.M. Chamberlain. An Algorithm for Connected Word Recognition. În J.P. Haton (editor), *Automatic Speech Analysis and Recognition*, pag. 191–204. D. Reidel Publishing Company, Dordrecht, Olanda, 1982.

[39] P.F. Brown. *The Acoustic-Modeling Problem in Automatic Speech Recognition.* Teză de doctorat, Carnegie Mellon University, mai 1987.

[40] F. Brugnara, D. Falavigna şi M. Omologo. Automatic segmentation and labeling of speech based on Hidden Markov Models. *Speech Communication*, 12(4):357–70, august 1993.

[41] F. Brugnara şi R. De Mori. Acoustic Modelling. În R. De Mori (editor), *Spoken Dialogues with Computers*, cap. 5, pag. 141–70. Academic Press, Londra, 1998.

[42] F. Brugnara, R. De Mori, D. Giuliani şi M. Omologo. Improved Connected Digit Recognition Using Spectral Variation Functions. În *Proceedings International Conference on Spoken Language Processing*, pag. 627–30, Banff, Canada, 1992.

[43] C. Burileanu. Caracterizarea unui vocabular limitat de cuvinte pronunţate izolat în vederea recunoaşterii automate. În M. Drăgănescu şi C. Burileanu (editori), *Analiza şi sinteza semnalului vocal*, pag. 36–131. Editura Academiei Române, Bucureşti, 1986.

[44] W. Byrne, P. Byerlein ş.a. Towards Language Independent Acoustic Modeling. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pag. 1029–32, 2000.

[45] R. Carre, R. Descout, M. Eskenazi, J. Mariani şi M. Rossi. The French Language Database: Defining, Planning, and Recording a Large Database. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, 1984.

[46] D. Chan, A. Fourcin ş.a. EUROM – A Spoken Language Resource for the EU. În *Proceedings EUROSPEECH'95*, vol. 1, pag. 867–70, Madrid, septembrie 1995.

[47] L.L. Chase. *Error-Responsive Feedback Mechanisms for Speech Recognizers.* Teză de doctorat, Carnegie Mellon University, aprilie 1997.

[48] M.Y. Chen, A. Kundu şi J. Zhou. Off-Line Handwritten Word Recognition Using a Hidden Markov Model Type Stochastic Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):481–96, mai 1994.

[49] R. Chengalvarayan şi L. Deng. Use of Generalized Dynamic Feature Parameters for Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 5(3):232–42, mai 1997.

[50] Y.L. Chow ş.a. BYBLOS: The BBN Continuous Speech Recognition System. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pag. 89–92, 1987.

[51] M. Cohen, G. Baldwin, J. Bernstein, H. Murveit şi M. Weintraub. Studies for an Adaptive Recognition Lexicon. În *Proceedings of the DARPA Speech Recognition Workshop*, San Diego, California, 1987.

[52] M. Constantinescu şi D. Cristescu. Sistem de analiză şi recunoaştere automată a vorbirii. În M. Drăgănescu şi C. Burileanu (editori), *Analiza şi sinteza semnalului vocal*, pag. 210–20. Editura Academiei Române, Bucureşti, 1986.

[53] P. Cosi, D. Falavigna, G.A. Mian şi M. Omologo. A Comparison between Mel-scale Cepstrum and Auditory Model Representation for Noisy Speech Recognition. În L. Torres, E. Masgrau şi M.A. Lagunas (editori), *SIGNAL PROCESSING V: Theories and Applications*, pag. 1199–1201. Elsevier Science Publishers, 1990.

[54] R.V. Cox, B.G. Haskell, Y. Lecun, B. Shahraray şi L. Rabiner. On the Applications of Multimedia Processing to Telecommunications. *Proceedings of the IEEE*, 86(5):755–824, mai 1998.

[55] K. Croot şi B. Taylor. Criteria for Acoustic-Phonetic Segmentation and Word Labelling in the Australian National Database of Spoken Language. Speech, Hearing and Language Research Centre, Macquarie University, 1995.

[56] R.I. Damper. *Introduction to Discrete-Time Signals and Systems.* Chapman & Hall, Londra, 1995.

[57] S.B. Davis şi P. Mermelstein. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–66, august 1980.

[58] R. De Mori (editor). *Spoken Dialogues with Computers.* Academic Press, Londra, 1998.

[59] J.R. Deller, J.G. Proakis şi J.H.L. Hansen. *Discrete-Time Processing of Speech Signals.* Macmillan Publishing Company, New York, 1993.

[60] A.P. Dempster, N.M. Laird şi D.B. Rubin. Maximum Likelihood from Incomplete Data via the *EM* Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[61] L. Deng. Processing of Acoustic Signals in a Cochlear Model Incorporating Laterally Coupled Supressive Elements. *Neural Networks*, 5(1):19–34, 1992.

[62] G.R. Doddington. Speaker Recognition – Identifying People by their Voices. *Proceedings of the IEEE*, 73(11):1651–64, noiembrie 1985.

[63] M. Drăgănescu. Tehnologia vorbirii. În M. Drăgănescu şi C. Burileanu (editori), *Analiza şi sinteza semnalului vocal*, pag. 9–16. Editura Academiei Române, Bucureşti, 1986.

[64] R.O. Duda şi P.E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, New York, 1973.

[65] T. Dumitrescu. Elemente de sinteza vorbirii în limba română. Dizertaţie de studii aprofundate, Departamentul de Calculatoare, Universitatea "Politehnica" din Timişoara, iulie 1996.

[66] T. Dutoit. *An Introduction to Text-to-Speech Synthesis.* Kluwer Academic Publishers, Dordrecht, Olanda, 1997.

[67] B. Eisen. Reliability of Speech Segmentation and Labelling at Different Levels of Transcription. În *Proceedings EUROSPEECH'93*, vol. 1, pag. 673–76, Berlin, 1993.

[68] D. Falavigna şi M. Omologo. A DTW-based Approach to the Automatic Labeling of Speech According to the Phonetic Transcription. În L. Torres, E. Masgrau şi M.A. Lagunas (editori), *SIGNAL PROCESSING V: Theories and Applications*, pag. 1139–42. Elsevier Science Publishers, 1990.

[69] G.C.M. Fant. Analysis and synthesis of speech processes. În B. Malmberg (editor), *Manual of phonetics*, cap. 8, pag. 173–277. North Holland, Amsterdam, a doua ediţie, 1970.

[70] M. Federico, M. Cettolo, F. Brugnara şi G. Antoniol. Language modelling for efficient beam-search. *Computer Speech and Language*, 9(4):353–79, 1995.

[71] W.M. Fisher, G.R. Doddington şi K.M. Goudie-Marshal. The DARPA Speech Recognition Research Database: Specification and Status. În *Proceedings of the DARPA Speech Recognition Workshop*, Palo Alto, California, februarie 1986.

[72] W.M. Fisher şi J.H. Fiscus. Better Alignment Procedures for Speech Recognition Evaluation. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pag. 59–62, 1993.

[73] J.L. Flanagan. Technologies for Multimedia Communications. *Proceedings of the IEEE*, 82(4):590–603, aprilie 1994.

[74] Center for Spoken Language Understanding. Speech Tools User Manual. Oregon Graduate Institute of Science and Technology, Beaverton, Oregon, august 1993.

[75] L. Fortuna. *Vorbirea artificială. Aplicaţii în industrie şi telecomunicaţii.* Editura Mirton, Timişoara, 1996.

[76] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, New York, 1972.

[77] S. Furui. Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34(1):52–59, februarie 1986.

[78] A. Ganapathiraju. *Support Vector Machines for Speech Recognition.* Teză de doctorat, Mississippi State University, ianuarie 2002.

[79] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski şi G.R. Doddington. Syllable-Based Large Vocabulary Continuous Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 9(4):358–66, mai 2001.

[80] M. Garman. *Psycholinguistics.* Cambridge University Press, 1990.

[81] J.S. Garofolo, L.F. Lamel, W.M. Fisher, D.S. Pallett şi N.L. Dahlgren. *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus.* U.S. Department of Commerce Technology Administration, National Institute of Standards and Technology, Gaithersburg, Maryland, 1993.

[82] J.L. Gauvain, L.F. Lamel, G. Adda şi M. Adda-Decker. Speaker-independent continuous speech dictation. *Speech Communication*, 15(1):21–37, 1994.

[83] J.L. Gauvain şi C.H. Lee. Bayesian learning for hidden Markov model with Gaussian mixture state observation densities. *Speech Communication*, 11:205–13, 1992.

[84] J.L. Gauvain şi C.H. Lee. Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–98, aprilie 1994.

[85] J.L. Gauvin şi L. Lamel. Speaker-Independent Phone Recognition Using BREF. În *Proceedings DARPA Workshop on Automatic Speech Recognition*, 1992.

[86] J.L. Gauvin, L.F. Lamel şi M. Eskénazi. Design Considerations and Text Selection for BREF, a large French read-speech corpus. În *Proceedings International Conference on Spoken Language Processing*, pag. 1097–1100, 1990.

[87] A. Gersho, S. Wang şi K. Zeger. Vector Quantization Techniques in Speech Coding. În S. Furui şi M.M. Sondhi (editori), *Advances in Speech Signal Processing*, cap. 2, pag. 49–84. Marcel Dekker, New York, 1992.

[88] O. Ghitza. Auditory nerve representation as a front-end for speech recognition in a noisy environment. *Computer, Speech, and Language*, 1(1):109–31, 1986.

[89] O. Ghitza. Auditory Nerve Representation as a Basis for Speech Processing. În S. Furui şi M.M. Sondhi (editori), *Advances in Speech Signal Processing*, cap. 15, pag. 453–85. Marcel Dekker, New York, 1992.

[90] D. Gibbon, R. Moore şi R. Winski (editori). *Handbook of Standards and Resources for Spoken Language Systems.* Mouton de Gruyter, Berlin, 1997.

[91] L. Gillick şi S.J. Cox. Some Statistical Issues in the Comparison of Speech Recognition Algorithms. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, pag. 532–35, 1989.

[92] H. Gish şi M. Schmidt. Text-Independent Speaker Identification. *IEEE Signal Processing Magazine*, (5):18–32, octombrie 1994.

[93] M. Giurgiu. Results on Automatic Speech Recognition in Romanian. În D. Tufiş şi P. Andersen (editori), *Recent Advances in Romanian Language Technology*, pag. 178–87. Editura Academiei Române, Bucureşti, 1997.

[94] J.J. Godfrey, E.C. Holliman şi J. McDaniel. SWITCHBOARD: Telephone Speech Corpus for Research and Development. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pag. 517–20, 1992.

[95] D. Graff. The 1996 Broadcast News Speech and Language-Model Corpus. În *Proceedings of the DARPA Speech Recognition Workshop*, Chantilly, Virginia, 1997.

[96] R.M. Gray. Vector Quantization. *IEEE Acoustics, Speech, and Signal Processing Magazine*, pag. 4–29, aprilie 1984.

[97] S. Greenberg. Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, 29:159–76, 1999.

[98] O. Grigore, I. Gavăt şi M. Zirra. Neural Network Vowel Recognition in Romanian Language. În *Proceedings Second International Conference on Technical Informatics - CONTI'96*, pag. 165–72, Timişoara, octombrie 1996.

[99] Multisite ATIS Data Collection Working Group. Multi-Site Data Collection for a Spoken Language Corpus. În *Proceedings of the DARPA Workshop on Speech and Natural Language*, pag. 7–14, Harriman, New York, februarie 1992.

[100] V. Groza, **M. Boldea** şi C. Bărbulescu. Recunoaşterea vocalelor cu ajutorul unui microsistem de calcul. În *Buletinul sesiunii ştiinţifice pentru tineret "Tehnic 2000"*, pag. 274–77, Timişoara, aprilie 1984.

[101] V.N. Gupta, M. Lennig şi P. Mermelstein. Integration of Acoustic Information in a Large Vocabulary Word Recognizer. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pag. 697–700, 1987.

[102] H. Hermansky şi N. Morgan. RASTA Processing of Speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–89, octombrie 1994.

[103] I.L. Hetherington, M.S. Philips, J.R. Glass şi V.W. Zue. *A** Word Network Search for Continuous Speech Recognition. În *Proceedings EUROSPEECH'93*, vol. 3, pag. 1533–36, Berlin, 1993.

[104] J.L. Hieronimus. Ascii Phonetic Symbols for the World's Languages: Worldbet. *Journal of the International Phonetic Association*, 1993.

[105] X.D. Huang, Y. Ariky şi M.A. Jack. *Hidden Markov Models for Speech Recognition.* Edinburgh University Press, 1990.

[106] X.D. Huang şi M.A. Jack. Semi-continuous hidden Markov models for speech signals. *Computer Speech and Language*, 3:239–51, 1989.

[107] M. Huckvale. SFS for Users. University College, Londra, martie 1996.

[108] M.Y. Hwang. *Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition.* Teză de doctorat, Carnegie Mellon University, decembrie 1993.

[109] M. Ioniţă, C. Burileanu şi M. Ioniţă. DTW Algorithm with Associated Matrix for a Password Access System. În *Proceedings Second International Workshop "Speech and Computer" – SPECOM'97*, pag. 91–96, Cluj-Napoca, octombrie 1997.

[110] F. Itakura. Minimum Prediction Residual Principle Applied to Speech Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):67–71, februarie 1975.

[111] N. Jayant, J. Johnston şi R. Safranek. Signal Compression Based on Models of Human Perception. *Proceedings of the IEEE*, 81(10):1385–1422, octombrie 1993.

[112] F. Jelinek. Fast Sequential Decoding Algorithm Using a Stack. *IBM Journal of Research and Development*, noiembrie 1969.

[113] F. Jelinek. Continuous Speech Recognition by Statistical Methods. *Proceedings of the IEEE*, 64(4):532–56, aprilie 1976.

[114] F. Jelinek. The Development of an Experimental Discrete Dictation Recognizer. *Proceedings of the IEEE*, 73(11):1616–24, noiembrie 1985.

[115] F. Jelinek. *Statistical Methods for Speech Recognition.* MIT Press, 1997.

[116] F. Jelinek, L.R. Bahl şi R.L. Mercer. Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech. *IEEE Transactions on Information Theory*, 21(3):250–56, mai 1975.

[117] F. Jelinek şi R.L. Mercer. Interpolated Estimation of Markov Source Parameters from Sparse Data. În E.S. Gelsema şi L.N. Kanal (editori), *Pattern Recognition in Practice*, pag. 381–97. North-Holland, 1980.

[118] F. Jelinek, R.L. Mercer şi S. Roukos. Principles of Lexical Language Modeling for Speech Recognition. În S. Furui şi M.M. Sondhi (editori), *Advances in Speech Signal Processing*, cap. 21, pag. 651–99. Marcel Dekker, New York, 1992.

[119] C.R. Jankowski Jr., H.D.H. Vo şi R.P. Lippmann. A Comparison of Signal Processing Front Ends for Automatic Word Recognition. *IEEE Transactions on Speech and Audio Processing*, 3(3):286–93, iulie 1995.

[120] B.H. Juang. Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains. *AT&T Technical Journal*, 64(6):1235–49, iulie-august 1985.

[121] B.H. Juang, S.E. Levinson şi M.M. Sondhi. Maximum Likelihood Estimation for Multivariate Mixture Observations of Markov Chains. *IEEE Transactions on Information Theory*, 32(2):307–309, martie 1986.

[122] B.H. Juang, L.R. Rabiner şi J.G. Wilpon. On the Use of Bandpass Liftering in Speech Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(7):947–54, iulie 1987.

[123] D. Jurafsky şi J.H. Martin. *Speech and Language Processing.* Prentice Hall, 2000.

[124] J.M. Kates. A Time-Domain Digital Cochlear Model. *IEEE Transactions on Signal Processing*, 39(12):2573–92, decembrie 1991.

[125] S.M. Katz. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3):400–401, martie 1987.

[126] A. Kipp, M.B. Wesenick şi F. Schiel. Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora. În *Proceedings International Conference on Spoken Language Processing*, vol. 1, pag. 106–109, Philadelphia, octombrie 1996.

[127] J.W. Klovstad şi L.F. Mondshein. The CASPERS Linguistic Analysis System. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):118–23, februarie 1975.

[128] A. Krogh. An introduction to hidden Markov models for biological sequences. În S. L. Salzberg, D.B. Searls şi S. Kasif (editori), *Computational Methods in Molecular Biology*, cap. 4, pag. 45–63. Elsevier, Amsterdam, 1998.

[129] S. Kullback. *Information Theory and Statistics.* John Wiley & Sons, New York, 1959.

[130] K. Kvale şi A.K. Foldvik. The multifarious r-sound. În *Proceedings International Conference on Spoken Language Processing*, pag. 1259–62, 1992.

[131] R. Lacouture şi R. De Mori. Lexical Tree Compression. În *Proceedings EUROSPEECH'91*, vol. 1, pag. 581–84, 1991.

[132] L.F. Lamel, R.H. Kassel şi S. Seneff. Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus. În *Proceedings DARPA Speech Recognition Workshop*, 1986.

[133] T. Lander şi S.T. Metzler. The CSLU Labeling Guide. Center for Spoken Language Understanding, Oregon Graduate Institute, februarie 1994.

[134] J. Lazzaro ş.a. Silicon Auditory Processors as Computer Peripherals. *IEEE Transactions on Neural Networks*, 4(3):523–28, mai 1993.

[135] K.F. Lee. *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System.* Teză de doctorat, Carnegie Mellon University, aprilie 1988.

[136] K.F. Lee şi F. Alleva. Continuous Speech Recognition. În S. Furui şi M.M. Sondhi (editori), *Advances in Speech Signal Processing*, cap. 20, pag. 623–50. Marcel Dekker, New York, 1992.

[137] K.F. Lee şi H.W. Hon. Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pag. 123–26, 1988.

[138] F. Lefèvre. *Estimation de probabilité non-paramétrique pour la reconnaissance markovienne de la parole.* Teză de doctorat, Université Pierre et Marie Curie, Paris, ianuarie 2000.

[139] C.J. Leggeter şi P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(2):171–85, aprilie 1995.

[140] R.G. Leonard. A Database for Speaker-Independent Digit Recognition. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, pag. 42.11.1–4, 1984.

[141] V.R. Lesser, R.D. Fennel, L.D. Erman şi D.R. Reddy. Organization of the Hearsay II Speech Understanding System. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):11–24, februarie 1975.

[142] H.C. Leung şi V. Zue. A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, pag. 2.7.1–4, 1984.

[143] L.A. Liporace. Maximum Likelihood Estimation for Multivariate Observations of Markov Sources. *IEEE Transactions on Information Theory*, 28(5):729–34, septembrie 1982.

[144] A. Ljolje şi M.D. Riley. Automatic Segmentation and Labeling of Speech. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pag. 473–76, 1991.

[145] E.P. Loeb şi R.F. Lyon. Experiments in Isolated Digit Recognition with a Cochlear Model. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, 1987.

[146] B.T. Lowerre. *The HARPY Speech Recognition System.* Teză de doctorat, Carnegie Mellon University, aprilie 1976.

[147] P.A. Lynn şi W. Fuerst. *Digital Signal Processing with Computer Applications.* John Wiley & Sons, New York, 1992.

[148] R.F. Lyon. A Computational Model of Filtering, Detection, and Compression in the Cochlea. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, 1982.

[149] R.F. Lyon. Speech Recognition Experiments with a Cochlear Model. În *Proceedings DARPA Speech Recognition Workshop*, Palo Alto, California, februarie 1986.

[150] R.F. Lyon şi C. Mead. An analog electronic cochlea. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(7):1119–34, iulie 1988.

[151] J. Makhoul, S. Roucos şi H. Gish. Vector Quantization in Speech Coding. *Proceedings of the IEEE*, 73(11):1551–88, noiembrie 1985.

[152] C.D. Manning şi H. Schütze. *Foundations of Statistical Natural Language Processing.* MIT Press, 1999.

[153] J.D. Markel şi A.H. Gray. *Linear Prediction of Speech.* Springer, Berlin, 1976.

[154] V. Marâi şi Gh. Marâi. *Comanda vocală a sistemelor tehnice.* Editura Militară, Bucureşti, 1991.

[155] G. Micca, A. Frasca şi M.G. Di Benedetto. Cross-lingual Interpolation of Speech Recognition Models. În *Proceedings of the Language Resources and Evaluation Conference*, vol. 3, pag. 1589–92, 2000.

[156] W. Minker, A. Waibel şi J. Mariani. *Stochastically-based semantic snalysis.* Kluwer Academic Publishers, Boston/Dordrecht/Londra, 1999.

[157] MIT. Speech spectrogram reading. Cursul 6.67s, iulie 1985.

[158] D.C. Montgomery. *Design and Analysis of Experiments.* John Wiley & Sons, New York, a treia ediţie, 1991.

[159] T.K. Moon. The Expectation-Maximization Algorithm. *IEEE Signal Processing Magazine*, 13(6):47–60, noiembrie 1996.

[160] B.C.J. Moore. *An Introduction to the Psychology of Hearing.* Academic Press, Londra, 1982.

[161] N. Morgan şi H. Bourlard. Neural Networks for Statistical Recognition of Continuous Speech. *Proceedings of the IEEE*, 83(5):742–70, mai 1995.

[162] N. Mukherjee, N. Rajput, L.V. Subramaniam şi A. Verma. On Deriving a Phoneme Model for a New Language. În *Proceedings International Conference on Spoken Language Processing*, 2000.

[163] C. Munteanu şi **M. Boldea**. A Description Language for Dialog Modeling and Management. În *Proceedings Fourth International Conference on Technical Informatics - CONTI'2000*, Timişoara, octombrie 2000.

[164] C. Munteanu şi **M. Boldea**. MDWOZ: A Wizard of Oz Environment for Dialog Systems Development. În *Proceedings Second International Conference on Language Resources and Evaluation - LREC2000*, Atena, Grecia, mai 2000.

[165] Y.K. Muthusamy, E. Barnard şi R.A. Cole. Reviewing Automatic Language Identification. *IEEE Signal Processing Magazine*, (5):33–41, octombrie 1994.

[166] C. Myers şi L.R. Rabiner. Connected Word Recognition Using a Level Building Dynamic Time Warping Algorithm. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pag. 951–55, 1981.

[167] A. Nádas. On Turing's Formula for Word Probabilities. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(6):1414–16, decembrie 1985.

[168] A.V. Nefian. *A Hidden Markov Model-Based Approach for Face Detection and Recognition.* Teză de doctorat, Georgia Institute of Technology, august 1999.

[169] H. Ney. Modeling and Search in Continuous Speech Recognition. În *Proceedings EUROSPEECH'93*, vol. 1, pag. 491–98, Berlin, 1993.

[170] H. Ney, U. Essen şi R. Knesser. On structuring probabilistic dependencies in stochastic language modeling. *Computer Speech and Language*, 8(1), ianuarie 1994.

[171] H. Ney, D. Mergel, A. Noll şi A. Paeseler. Data Driven Search Organization for Continuous Speech Recognition. *IEEE Transactions on Signal Processing*, 40(2):272–81, februarie 1992.

[172] E. Nicolau, I. Weber şi Şt. Gavăt. Aparat pentru recunoaşterea automată a vocalelor. *Automatica şi Electronica*, (6), 1963.

[173] E. Oancea. *Analiza şi sinteza vorbirii.* Editura Militară, Bucureşti, 1976.

[174] A.V. Oppenheim şi R.W. Schafer. Homomorphic Analysis of Speech. *IEEE Transactions on Audio and Electroacoustics*, 16(6):118–23, iunie 1968.

[175] A.V. Oppenheim şi R.W. Schafer. *Digital Signal Processing.* Prentice-Hall International, Londra, 1975.

[176] M. Ordowski, N. Deshmukh, A. Ganapathiraju, J. Hamaker şi J. Picone. A Public Domain Speech-to-Text System. În *Proceedings of EUROSPEECH'99*, vol. 5, pag. 2127–30, Budapesta, septembrie 1999.

[177] OROS, Meylan, Franţa. *Documentaţii ale plăcii OROS AU21*, 1995.

[178] S. Ortmans, H. Ney şi A. Eiden. Language-Model Look-Ahead for Large Vocabulary Speech Recognition. În *Proceedings International Conference on Spoken Language Processing*, vol. 4, pag. 2095–98, 1996.

[179] D. O'Shaughnessy. Speaker Recognition. *IEEE Acoustics, Speech, and Signal Processing Magazine*, (5):4–17, octombrie 1986.

[180] D. O'Shaughnessy. *Speech Communication: Human and Machine.* Addison-Wesley, 1987.

[181] D.S. Pallet. Benchmark Tests for DARPA Resource Management Database Performance Evaluations. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, pag. 536–39, 1989.

[182] D.S. Pallet, W.M. Fisher şi J.G. Fiscus. Tools for the Analysis of Benchmark Speech Recognition Tests. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, pag. 97–100, 1990.

[183] D.B. Paul şi J.M. Baker. The Design for the Wall Street Journal-based CSR Corpus. În *Proceedings of the DARPA Workshop on Speech and Natural Language*, pag. 357–62, Harriman, New York, februarie 1992.

[184] W.H. Perkins şi R.D. Kent. *Textbook of Functional Anatomy of Speech, Language, and Hearing.* Taylor & Francis, Londra, Philadelphia, 1986.

[185] M. Pescaru. Prelucrări de texte pentru sinteza automată a vorbirii în limba română. Dizertaţie de studii aprofundate, Departamentul de Calculatoare, Universitatea "Politehnica" din Timişoara, iulie 1996.

[186] J.M. Pickett. *The Sounds of Speech Communication.* University Park Press, Baltimore, 1980.

[187] J. Picone, G.R. Doddington şi D.S. Pallett. Phone-Mediated Word Alignment for Speech Recognition Evaluation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(3):559–62, martie 1990.

[188] J. Picone, K.M. Goudie-Marshall, G.R. Doddington şi W. Fisher. Automatic Text Alignment for Speech System Evaluation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34(4):780–84, august 1986.

[189] L.C.W. Pols. Real-Time Recognition of Spoken Words. *IEEE Transactions on Computers*, 20(9):972–78, septembrie 1971.

[190] H. Pârlog. *The Sound of Sounds.* Hestia Publishing House, Timişoara, 1995.

[191] W.H. Press, S.A. Teukolsky, W.T. Vetterling şi B.P. Flannery. *Numerical Recipes in C.* Cambridge University Press, a doua ediţie, 1992.

[192] P. Price, W.M. Fisher, J. Bernstein şi D.S. Pallett. The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pag. 651–54, 1988.

[193] S. Puşcariu. *Limba română: Rostirea.* Editura Academiei, Bucureşti, 1959.

[194] L. Rabiner şi B.H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

[195] L.R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–86, februarie 1989.

[196] L.R. Rabiner. Applications of Voice Processing to Telecommunications. *Proceedings of the IEEE*, 82(2):199–228, februarie 1994.

[197] L.R. Rabiner, B.H. Juang, S.E. Levinson şi M.M. Sondhi. Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities. *AT&T Technical Journal*, 64(6):1211–34, iulie-august 1985.

[198] L.R. Rabiner şi S.E. Levinson. A Speaker-Independent, Syntax-Directed, Connected Word Recognition System Based on Hidden Markov Models and Level Building. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(3):561–73, iunie 1985.

[199] L.R. Rabiner, S.E. Levinson şi M.M. Sondhi. On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent, Isolated Word Recognition . *The Bell System Technical Journal*, 62(4):1075–1105, aprilie 1983.

[200] L.R. Rabiner şi R.W. Schafer. *Digital Processing of Speech Signals.* Prentice-Hall, 1978.

[201] L.R. Rabiner, J.G. Wilpon şi F.K. Soong. High Performance Connected Digit Recognition Using Hidden Markov Models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(8):1214–25, august 1989.

[202] R.W. Ramirez. *The FFT: Fundamentals and Concepts.* Tektronix, Inc., Beaverton, Oregon, 1975.

[203] P. Roach, S. Arnfield, W. Barry, J. Baltova, **M. Boldea**, A. Fourcin, W. Gonet, R. Gubrynowicz, E. Hallum, L. Lamel, K. Marasek, A. Marchal, E. Meister şi K. Vicsi. BABEL: An Eastern European Multi-Language Database. În *Proceedings International Conference on Spoken Language Processing*, Philadelphia, 1996.

[204] P.J. Roach, S. Arnfield, W. Barry, S. Dimitrova, **M. Boldea**, A. Fourcin, W. Gonet, R. Gubrynowicz, E. Hallum, L. Lamel, K. Marasek, A. Marchal, E. Meister şi K. Vicsi. BABEL: A Database of Central and Eastern European Languages. În *Proceedings First International Conference on Language Resources and Evaluation - LREC*, vol. 1, pag. 371–74, Granada, Spania, mai 1998.

[205] A.J. Robinson. An Application of Recurrent Nets to Phone Probability Estimation. *IEEE Transactions on Neural Networks*, 5(2):298–305, martie 1994.

[206] A. Roceric-Alexandrescu. *Fonostatistica limbii române.* Editura Academiei, Bucureşti, 1968.

[207] A.E. Rosenberg, L.R. Rabiner, J.G. Wilpon şi D. Kahn. Demisyllable-Based Isolated Word Recognition System. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-31(3):713–26, iunie 1983.

[208] M. Rossi. *Électroacoustique.* Presses polytechniques romandes, Lausanne, 1986.

[209] S. Roucos şi M.O. Dunham. A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition. În *Proceedings International Conference on Acoustics, Speech. and Signal Processing*, 1987.

[210] S. Russel şi P. Norvig. *Artificial Intelligence: A Modern Approach.* Prentice Hall, 1995.

[211] I.Gh. Şabac. *Matematici speciale*, vol. 2. Editura Didactică şi Pedagogică, Bucureşti, 1965.

[212] H. Sakoe. Two-Level DP-Matching – A Dynamic Programming-Based Pattern Matching Algorithm for Connected Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(6):588–95, decembrie 1979.

[213] R.W. Schafer şi L.R. Rabiner. Digital Representations of Speech Signals. *Proceedings of the IEEE*, 63(4):662–77, 1975.

[214] M.S. Schmidt şi G.S. Watson. The Evaluation and Optimization of Automatic Speech Segmentation. În *Proceedings EUROSPEECH'91*, vol. 2, pag. 701–704, Genova, Italia, 1991.

[215] T. Schultz şi A. Waibel. Language Independent and Language Adaptive Large Vocabulary Speech Recognition. În *Proceedings International Conference on Spoken Language Processing*, 1998.

[216] R. Schwartz şi S. Austin. A Comparison of Several Approximate Algorithms for Finding Multiple (N-BEST) Sentence Hypotheses. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pag. 701–704, 1991.

[217] R. Schwartz şi Y.L. Chou. The N-Best Algorithm: An Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pag. 81–84, 1990.

[218] R. Schwartz şi Y. Chow. Improved Hidden Markov Modeling of Phonemes for Continuous Speech Recognition. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, 1984.

[219] R. Schwartz, Y.L. Chow şi F. Kubala. Rapid Speaker Adaptation using a Probabilistic Spectral Mapping. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pag. 633–36, 1987.

[220] S. Seneff. A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics*, 16:55–76, 1988.

[221] C. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–56, iulie, septembrie 1948.

[222] H. Sheikhzadeh şi L. Deng. Speech Analysis and Recognition Using Interval Statistics Generated from a Composite Auditory Model. *IEEE Transactions on Speech and Audio Processing*, 6(1):90–94, ianuarie 1998.

[223] K. Shikano. Evaluation of LPC Spectral Matching Measures for Phonetic Unit Recognition. Raport tehnic CMU-CS-86-108, Carnegie Mellon University, februarie 1986.

[224] K. Shikano şi F. Itakura. Spectrum Distance Measures for Speech Recognition. În S. Furui şi M.M. Sondhi (editori), *Advances in Speech Signal Processing*, cap. 14, pag. 419–52. Marcel Dekker, New York, 1992.

[225] K. Sjölander şi J. Beskow. WaveSurfer - an Open Source Speech Tool. În *Proceedings International Conference on Spoken Language Processing*. Beijing, China, 2000.

[226] F.K. Soong şi E.F. Huang. A Tree-Trellis Based Fast Search for Finding the N Best Sentence Hypotheses in Continuous Speech Recognition. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pag. 705–708, 1991.

[227] A.S. Spanias. Speech Coding: A Tutorial Review. *Proceedings of the IEEE*, 82(10):1541–82, octombrie 1994.

[228] D. Stanomir. *Electroacustică*. Editura Didactică şi Pedagogică, Bucureşti, 1968.

[229] H.J.M. Steeneken şi J.G. van Velden. Objective and Diagnostic Assessment of (Isolated) Word Recognizers. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, pag. 540–43, 1989.

[230] H.J.M. Steeneken şi J.G. van Velden. Recognizer assessment by means of CVC-words as available in the EUROM-1 data-base. Raport tehnic, TNO – Institute for Perception, Soesterbergh, Olanda, 1991.

[231] V. Steinbiss. Improvements in Beam Search. În *Proceedings International Conference on Spoken Language Processing*, vol. 4, pag. 2143–46, 1994.

[232] G. Stolojanu, V. Podaru şi F. Cetină. *Prelucrarea numerică a semnalului vocal*. Editura Militară, Bucureşti, 1984.

[233] R.D. Stuart. *Introducere în analiza Fourier cu aplicaţii în tehnică*. Editura Tehnică, Bucureşti, 1971.

[234] A. Tătaru. *Limba română: Specificul pronunţării în contrast cu germana şi engleza*. Editura Dacia, Cluj-Napoca, 1997.

[235] H.N. Theodorescu, L. Buchholtzer şi C. Poşa. *Comunicarea orală om-maşină.* Editura Tehnică, Bucureşti, 1986.

[236] G. Toderean, M. Costeiu şi M. Giurgiu. *Reţele neuronale artificiale.* Editura Albastră, Cluj-Napoca, 1995.

[237] Y. Tohkura. A Weighted Cepstral Distance Measure for Speech Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(10):1414–22, octombrie 1987.

[238] L. Toma şi T. Jurca. Isolated word recognition system. *Buletinul ştiinţific şi tehnic al Institutului Politehnic "Traian Vuia" Timişoara*, 1990.

[239] H. Traunmüller. Auditory scales of frequency representation. Pagină WWW, http://www.ling.su.se/staff/hartmut/bark.htm, august 1997.

[240] S. Umesh, L. Cohen şi D. Nelson. Fitting the Mel Scale. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[241] V. Valtchev, J.J. Odell, P.C. Woodland şi S.J. Young. A Dynamic Network Decoder Design for Large Vocabulary Speech Recognition. În *Proceedings International Conference on Spoken Language Processing*, vol. 2, pag. 1351–54, 1994.

[242] J.P.H. van Santen, R.W. Sproat, J.P. Olive şi J. Hirschberg (editori). *Progress in Speech Synthesis.* Springer, New York, 1997.

[243] R. Vancea, Şt. Holban şi D. Ciubotariu. *Recunoaşterea formelor - Aplicaţii.* Editura Academiei, Bucureşti, 1989.

[244] E. Vasiliu. *Fonologia limbii române.* Editura Ştiinţifică, Bucureşti, 1965.

[245] T.K. Vintsiuk. Two Approaches to Create a Dictation/Translation Machine. În *Proceedings Second International Workshop "Speech and Computer" - SPECOM'97*, pag. 1–7, Cluj-Napoca, octombrie 1997.

[246] T.K. Vintsyuk. Speech Discrimination by Dynamic Programming. *Kibernetika*, 4(1):81–88, 1968.

[247] A.J. Viterbi. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*, 13(2):260–69, aprilie 1967.

[248] Z. Vâlsan, I. Gavăt, B. Şabac ş.a. Statistical and Hybrid Methods for Speech Recognition in Romanian. *International Journal of Speech Technology*, 5:259–68, septembrie 2002.

[249] R.A. Wagner şi M.J. Fischer. The String-to-String Correction Problem. *Journal of the ACM*, 21(1):168–173, ianuarie 1974.

[250] L. Watts, D.A. Kerns, R.F. Lyon şi C.A. Mead. Improved Implementation of the Silicon Cochlea. *IEEE Journal of Solid-State Circuits*, 27(5):692–700, mai 1992.

[251] J.C. Wells. Computer-coding the IPA: a proposed extension of SAMPA. Department of Phonetics and Linguistics, University College, Londra, 1995.

[252] M.B. Wesenick şi A. Kipp. Estimating the Quality of Phonetic Transcriptions and Segmentations of Speech Signals. În *Proceedings International Conference on Spoken Language Processing*, vol. 1, pag. 129–32, Philadelphia, 1996.

[253] B. Wheatley, K. Kondo, W. Anderson şi Y. Muthusamy. An Evaluation of Cross-Language Adaptation for Rapid HMM Development in a New Language. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pag. 237–40, 1994.

[254] C.W. Wightman şi D.T. Talkin. The Aligner: Text-to-Speech Alignment Using Markov Models. În J.P.H. Van Santen, R.W. Sproat, J.P. Olive şi J. Hirschberg (editori), *Progress in Speech Synthesis*, cap. 25, pag. 313–23. Springer, New York, 1997.

[255] J.G. Wilpon, B.H. Juang şi L.R. Rabiner. An Investigation on the Use of Acoustic Sub-Word Units for Automatic Speech Recognition. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pag. 821–24, 1987.

[256] I.H. Witten şi T.C. Bell. The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression. *IEEE Transactions on Information Theory*, 37(4):1085–94, iulie 1991.

[257] W.A. Woods. Motivation and Overview of SPEECHLIS: An Experimental Prototype for Speech Understanding Research. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):2–10, februarie 1975.

[258] S.J. Young. The HTK Hidden Markov Model Toolkit: Design and Philosophy. Raport tehnic TR-152, Cambridge University Engineering Department, 1994.

[259] S.J. Young şi L.L. Chase. Speech recognition evaluation: a review of the U.S. CSR and LVCSR programmes. *Computer Speech and Language*, 12(4):263–79, octombrie 1998.

[260] S.J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev şi P. Woodland. *The HTK Book Version 3.0.* Cambridge University, 2000.

[261] S.J. Young, N.H. Russel şi J.H.S. Thornton. Token Passing: a Conceptual Model for Connected Speech Recognition Systems. Raport tehnic TR-38, Cambridge University Engineering Department, 1989.

[262] J. Zeiliger şi J.F. Serignat. Europec software v4.1 user's guide. Raport SAM-ICP-045, Institute de la Comunication Parlé, Grenoble, Franţa, 1991.

[263] Y. Zhang, R. Togneri şi M. Adler. Phoneme-Based Vector Quantization in a Discrete HMM Speech Recognizer. *IEEE Transactions on Speech and Audio Processing*, 5(1):26–32, ianuarie 1997.

[264] Q. Zhou şi W. Chou. An Approach to Continuous Speech Recognition Based on Layered Self-Adjusting Decoding Graph. În *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, pag. 1779–82, 1997.

[265] V.W. Zue. The Use of Speech Knowledge in Automatic Speech Recognition. *Proceedings of the IEEE*, 73(11):1602–15, noiembrie 1985.

[266] V.W. Zue şi S. Seneff. Transcription and Alignment of the TIMIT Database. În *Proceedings Second Symposium on Advanced Man-Machine Interface Through Spoken Language*, Oahu, Hawaii, noiembrie 1988.