

Universitatea "Politehnica" din Timișoara
Facultatea de Automatică și Calculatoare

CONTRIBUȚII LA
RECUNOAȘTEREA AUTOMATĂ
A VORBIRII CONTINUE
ÎN LIMBA ROMÂNĂ

TEZĂ DE DOCTORAT

ing. Marian Boldea

Timișoara, 2003

Sumar

Această teză este bazată pe o serie de cercetări în direcția recunoașterii automate a vorbirii în limba română, cercetări în care principala problemă abordată a fost cea a modelării acustice folosind unități sublexicale pentru recunoașterea vorbirii continue cu vocabulare mari și foarte mari.

Pentru început, teza își precizează cadrul teoretic printr-o trecere în revistă a unor metode de bază în analiza și recunoașterea vorbirii.

Este prezentată apoi proiectarea și colectarea primei baze de date fonetice în limba română, destinată în primul rând cercetărilor în direcția recunoașterii 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 recunoașterii automate a vorbirii și în alte domenii.

În final sunt evaluate două posibile seturi de unități sublexicale de modelare acustică, evaluare desfășurată prin experimente de recunoaștere dependentă și independentă de vocabular a unor semnale din baza de date.

Cercetările au fost finanțate parțial, prin opt granturi și contracte, de Academia Română, Comisia Europeană, Consiliul Național al Cercetării Științifice din Învățământul Superior (fost Consiliul Național al Cercetării Științifice Universitare) și fostul Minister al Cercetării și Tehnologiei.

CUPRINS

Sumar	3
Cuprins	5
Mulțumiri	9
1 Introducere	11
1.1 Scurtă istorie	13
1.2 Stadiul actual	14
1.3 Obiectivele cercetărilor	16
1.4 Organizarea lucrării	16
2 Analiza semnalului vocal	17
2.1 Reprezentări ale semnalului vocal	18
2.2 Prelucrări în domeniul timp	20
2.2.1 Cadrarea și ferestruirea	20
2.2.2 Energia și puterea	21
2.2.3 Autocorelația	22
2.2.4 Preaccentuarea	23
2.3 Analiza prin predicție liniară	25
2.3.1 Metoda autocorelației	28
2.4 Analiza Fourier	29
2.5 Legături timp–frecvență	30
2.6 Analiza homomorfică	33
2.6.1 Cepstrul real	36
2.7 Metode perceptuale	38
2.7.1 Analiza melodică	39
2.7.2 Caracteristicile dinamice	40
2.8 Concluzii	41

3	Recunoașterea automată a vorbirii	43
3.1	Evaluarea performanțelor	44
3.1.1	Compararea prin programare dinamică	46
3.2	Distanțe acustice	47
3.3	Transformări ale spațiului acustic	49
3.4	Metode statistice de recunoaștere a vorbirii	50
3.5	Modelele Markov ascunse	52
3.5.1	Antrenarea MMA discrete	54
3.6	Modelarea lingvistică	56
3.7	Modelarea acustică	58
3.7.1	Antrenarea MMA cu mixturi gaussiene	59
3.8	Reprezentarea integrată a cunoștințelor	60
3.9	Algoritmi de căutare	61
3.10	Concluzii	63
4	Baza de date fonetice	65
4.1	Considerații de proiectare	66
4.2	Alegerea unităților de modelare	67
4.3	Materialele de înregistrat	69
4.3.1	Pasajele	70
4.3.2	Propozițiile	72
4.4	Vorbitorii	73
4.5	Analize statistice	74
4.6	Organizarea bazei de date	76
4.7	Realizarea înregistrărilor	77
4.8	Datele colectate	79
4.9	Calitatea înregistrărilor	80
4.10	Concluzii	82
5	Etichetarea semnalelor vocale	83
5.1	Alegerea nivelului de etichetare	85
5.2	Automatizarea etichetării	86
5.2.1	Evaluarea etichetării automate	87
5.3	Sistemul de etichetare	88
5.3.1	Etichetarea manuală	88
5.3.2	Transcrierea fonetică	89
5.3.3	Extragerea caracteristicilor	90
5.3.4	Modelele acustice	90
5.3.5	Segmentarea automată	92
5.3.6	Verificarea etichetării	93
5.4	Criteriile de decizie	93
5.4.1	Fenomenele specifice vorbirii fluente	95
5.4.2	Vocalele și semivocalele	95
5.4.3	Consoanele plozive	96
5.4.4	Consoanele fricative și africcate	97

5.4.5	Consoanele sonante	97
5.4.6	Problema /I/	98
5.5	Rezultate și comentarii	98
5.6	Concluzii	102
6	Experimente de modelare acustică	103
6.1	Experimente dependente de vorbitor	104
6.1.1	Decodarea lingvistică	104
6.1.2	Rezultate și comentarii	105
6.2	Experimente independente de vorbitor	106
6.3	Recunoașterea unităților de modelare	106
6.3.1	Vorbitori și date	106
6.3.2	Alternativele de modelare	108
6.3.3	Extragerea caracteristicilor	108
6.3.4	Modelele acustice	109
6.3.5	Rezultate și comentarii	109
6.4	Recunoașterea cuvintelor	114
6.4.1	Vorbitori și date	114
6.4.2	Dicționarele	115
6.4.3	Modelele acustice	115
6.4.4	Rezultate și comentarii	117
6.5	Concluzii	122
7	Încheiere	123
7.1	Contribuții	124
7.2	Continuări	125
A	Detalii ale dicționarelor	127
B	Exemple de recunoaștere	129
	Bibliografie	143

Mulțumiri

În primul rând, mulțumesc domnului profesor Crișan Strugaru, conducătorul meu de doctorat, pentru îndrumările și sprijinul moral și material pe care mi le-a acordat în permanență de-a lungul activităților de cercetare pe care această teză le valorifică.

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

De asemenea, mulțumesc profesorului Renato De Mori, cu sprijinul și sub îndrumarea căruia am efectuat un stagiul de cercetare de șase luni pe probleme ale recunoașterii automate a vorbirii la Laboratorul de Informatică al Universității din Avignon.

Esențiale au fost mijloacele financiare cu care cercetările au fost susținute de către Academia Română, Comisia Europeană, Consiliul Național al Cercetării Științifice din Învățământul Superior (fost Consiliul Național al Cercetării Științifice Universitare) și fostul Minister al Cercetării și 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 câțiva foști studenți: Alin Doroga, Tiberiu Dumitrescu, Maria Pescaru, Cosmin Munteanu.

Foarte importante au fost de asemenea sprijinul domnului profesor Nicolae Robu pentru realizarea unor înregistrări de calitate în studioul TeleUniversității, ca și înțelegerea colegilor de acolo, care au suportat timp de aproximativ doi ani problemele astfel apărute.

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

La fel de important a fost și ajutorul vorbitorilor înregistrați: deși numele fiecăruia există notat undeva, sunt totuși prea mulți (o sută) pentru a fi enumerați aici.

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

CAPITOLUL 1

Introducere

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

În ansamblul aplicațiilor 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 adecvați prelucrărilor ulterioare [200], [232], [59];
- **îmbunătățirea calității** prin reducerea efectelor zgomotelor și distorsiunilor, necesară atât pentru facilitarea comunicării între oameni, cât și ca o etapă preliminară altor prelucrări sau aplicații mai complexe [59];
- **codarea semnalului vocal**, având ca scop obținerea unei reprezentări cât mai compacte a acestuia în vederea stocării sau transmiterii, simultan cu păstrarea unei cât mai bune calități 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 recunoaștere corespunzător, sau punerii în legătură cu un vorbitor al aceleiași limbi [165];

- **recunoașterea vorbitorului**, putând urmări **identificarea** acestuia dintre mai multe persoane, de exemplu în expertize criminalistice, sau **verificarea** identității pe care acesta o pretinde, cu aplicații în controlul accesului [62], [179], [92];
- **recunoașterea automată a vorbirii**, având ca scop determinarea cu o cât mai mare exactitate a șirului de cuvinte pronunțat de un vorbitor, și utilă în aplicații de tip comandă-și-control, dictare automată etc. [194], [115], [58];
- **înțelegerea vorbirii**, care are ca obiectiv nu o determinare exactă a cuvintelor pronunțate, ci a semnificației acestora, astfel încât să se poată efectua în continuare o serie de acțiuni de răspuns corecte din punctul de vedere al aplicației [1], [156].

Deși 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 interacțiunii acestor sisteme cu utilizatorii, imperios necesară în perspectiva unei societăți informaționale, recunoașterea, înțelegerea și sinteza automată a vorbirii sunt esențiale, iar teza de față prezintă o serie de cercetări în direcția recunoașterii automate a vorbirii continue care, împreună cu altele urmărind înțelegerea [24], sinteza ei automată din text [65], [185], și integrarea tuturor acestora în sisteme de dialog [164], [163] vizează dezvoltarea de interfețe vocale om-mașină în limba română [28].

Problemele recunoașterii automate a vorbirii

Dificultatea recunoașterii automate a vorbirii este influențată de multe variabile, cele mai importante fiind:

- **tipul pronunției**: continuă sau discretă (cu pauze între cuvinte);
- **stilul de vorbire**: citită, spontană sau semispontană (ca răspuns la o cerere);
- **numărul de vorbitori** ale căror pronunții trebuie recunoscute, în funcție de acesta sistemele de recunoaștere 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 recunoaște și care poate fi **mică** ($N \times 10$), **medie** ($N \times 100$), **mare** ($N \times 1000$) sau **foarte mare** ($N \times 10000$);
- **condițiile 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, deși începute aproape imediat după inventarea calculatoarelor digitale, cercetările asupra recunoașterii automate a vorbirii să ajungă doar în zilele noastre la stadiul în care, pentru câteva limbi, există disponibile comercial sisteme de recunoaștere a vorbirii continue cu vocabulare foarte mari. Pentru o mai bună imagine globală asupra dificultăților care au trebuit depășite, 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 recunoașterii 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] și au urmărit recunoașterea unor sunete sau a unor cuvinte izolate folosind **analiza spectrală** cu filtre analogice și idei din **fonetica acustică**.

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

De-abia în anii '70 s-au pus fundamente teoretice solide pentru tratarea problemei, bazate pe recunoașterea formelor, inteligența artificială și teoria comunicației. Au fost introduse **distanțe spectrale** [110] cu interpretări psihofiziologice clare în deformarea dinamică a timpului, iar metoda ca atare a fost extinsă de la recunoașterea cuvintelor izolate la cea a cuvintelor conectate [212]. Abordări inspirate de **inteligența artificială**, cunoscute și ca **bazate pe cunoștințe (knowledge based)**, au fost materializate în **sisteme expert** în care **surse de cunoștințe** (fonetice, fonologice, lexicale, sintactice, semantice și pragmatice), încorporând cunoștințe ale unor experți umani, cooperează pentru recunoașterea pronunțiilor din semnalele prelucrate [257], [141]. Însă cea mai importantă a fost abordarea recunoașterii vorbirii ca problemă de **teoria comunicației** [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**, și **modelele Markov ascunse** – **MMA** (în literatura de limbă engleză, **hidden Markov models** – **HMM**), care stau la baza tuturor sistemelor moderne de recunoaștere automată a vorbirii.

Cele mai importante evoluții au avut loc însă în anii '80. Așa cum am menționat deja, există o gamă largă de metode pentru abordarea recunoașterii automate a vorbirii, dar mulțimea factorilor care influențează semnalul vocal, determinând caracterul său cvasialéator, face ca problema să fie una extrem de complexă, care nu admite soluții analitice, așa 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 recunoaștere a cuvintelor conectate prin deformarea dinamică a timpului [166], [38] și a utilizării modelelor Markov ascunse pentru recunoașterea cuvintelor izolate [199], [197] sau conectate [198], [201] cu vocabulare cât mai mari [101], [10] și 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 menționate construcția unor baze de date vocale de dimensiuni suficiente pentru utilizarea metodelor statistice și compararea diferitelor abordări în condiții identice [132], [192], noi variante de modelare statistică folosind MMA [11], [209], [106], utilizarea unor unități acustice sublexicale [207], [255] și modelarea lor dependentă de context [218] etc.

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

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

Pe plan național, interesul pentru cercetările asupra prelucrării automate a vorbirii s-a manifestat în special în cadrul catedrelor de electronică, automatică și calculatoare din unele centre de învățământ superior: București [63], Iași [235], Timișoara [75]. Ca în multe alte țări, un interes aparte pentru aceste cercetări au manifestat și instituțiile militare de învățământ superior și cercetare [173], [232], [154].

Preocupările în direcția recunoașterii automate a vorbirii au existat încă din anii '60 ([172], citat în [63]), dar în condițiile izolării de comunitatea științifică internațională și a lipsei unei baze materiale corespunzătoare, ele au continuat doar sporadic, grație entuziasmului unui număr redus de cercetători, și au fost limitate la recunoașterea 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], rețelele neuronale [236], [98], metodele hibride combinând modelele Markov ascunse cu rețelele neuronale sau tehnicile fuzzy [248]. Cu toate acestea, ele au rămas în continuare limitate la nivelul sunetelor sau al cuvintelor izolate, fără a aborda recunoașterea vorbirii continue.

1.2 Stadiul actual

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

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

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

Inițial, aceste două direcții 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 și pentru alte limbi, în general dintre cele cu suficient potențial comercial.

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

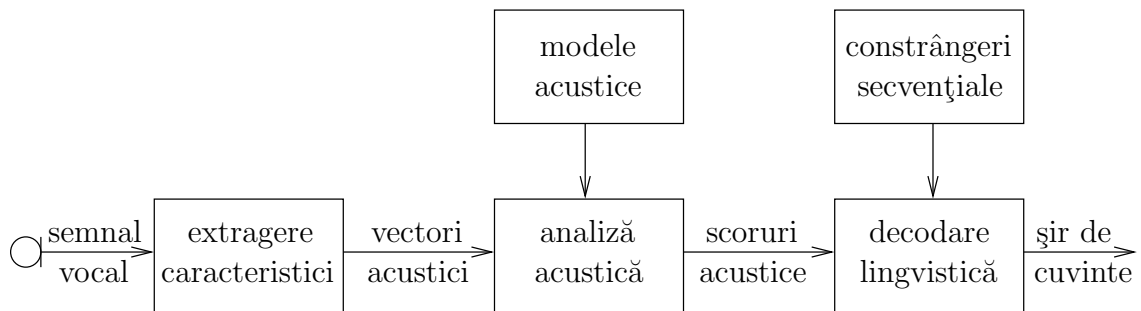


Figura 1.1: Schema bloc tipică a unui sistem de recunoaștere automată a vorbirii

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

După obținerea șirului de vectori acustici, **recunoașterea** propriu-zisă presupune rezolvarea a două probleme, tratate în blocurile următoare:

- **analiza acustică** a porțiunii de semnal corespunzătoare unui vector în raport cu niște **modele acustice**, concretizată în calculul unor **scoruri acustice**: funcție de metoda de recunoaștere utilizată și tipul modelelor acustice, scorurile pot fi **distanțe** între vectorul curent și cei din modele, sau **probabilități** condiționate ca vectorul să fi apărut în urma pronunției unităților modelate;
- **decodarea lingvistică** prin alinierea vectorilor acustici cu modelele asociate unui șir de cuvinte folosind algoritmi de **căutare**; aceasta urmărește obținerea unei estimări optime a șirului de cuvinte prin pronunțarea căruia s-a obținut semnalul, și poate include **constrângeri secvențiale**: la nivelul cuvintelor, acestea pot fi impuse prin **dicționare de pronunții** ale cuvintelor din vocabularul sistemului în termenii eventualelor unități de modelare acustică sublexicală, iar la nivelul întregii pronunții, prin **gramatici** sau **modele lingvistice statistice** ale aplicației.

În cadrul oferit de această schemă-bloc, cercetările continuă în prezent în diferite direcții: metode robuste de extragere a caracteristicilor; modelare acustică cu acuratețe sporită, independentă de vocabular, și utilizând cât mai eficient datele disponibile; adaptarea modelelor acustice și lingvistice; modelarea cuvintelor noi și a variațiilor de pronunție; reducerea complexității algoritmilor de căutare etc.

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

1.3 Obiectivele cercetărilor

Ținând cont de nivelul mondial și național, **obiectivul principal** al cercetărilor a fost **recunoașterea vorbirii continue în limba română, independentă de vorbitor, cu vocabulare în jurul a 1000 cuvinte**. Deoarece aceasta a fost (din câte cunoaștem) **prima abordare a recunoașterii automate a vorbirii continue în limba română**, acest obiectiv avea avantajele de a fi atât realist cât și semnificativ în condițiile date.

Pentru recunoașterea vorbirii continue cu vocabulare mari și foarte mari, de mii sau zeci de mii de cuvinte, esențială este **problema unităților de modelare acustică**: dacă pentru vocabulare mici, de zeci de cuvinte, fiecare cuvânt poate fi modelat separat, utilizând un număr de pronunții ale sale, odată cu creșterea mărimii vocabularelor crește și cantitatea de date necesare antrenării de modele ale cuvintelor, aceasta putând deveni prohibitivă pentru vocabulare mari și foarte mari. Soluția constă în utilizarea unor unități 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 considerații, au fost stabilite ca **obiective intermediare** în vederea atingerii obiectivului principal enunțat mai sus:

- **proiectarea și 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 și evaluarea performanțelor la acest nivel;
- **studiul unor unități de modelare acustică sublexicală** prin experimente de recunoaștere 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) și a unora utilizate în recunoașterea automată a vorbirii (capitolul 3).

În continuare sunt descrise cercetările care stau la baza tezei: proiectarea și colectarea primei baze de date fonetice în limba română, permițând cercetări asupra recunoașterii vorbirii continue, independentă de vorbitor, cu vocabulare în jurul a 1000 de cuvinte (capitolul 4); etichetarea la nivel fonetic a celei mai mari părți 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 unități de modelare acustică pentru recunoașterea independentă de vorbitor a vorbirii continue în limba română, desfășurate prin experimente folosind subseturi din baza de date (capitolul 6).

În încheiere, capitolul 7 sintetizează contribuțiile lucrării și trasează câteva direcții în care cercetările ar putea fi continuate.

CAPITOLUL 2

Analiza semnalului vocal

Nestaționaritatea semnalului vocal, cunoscută din literatura de specialitate [200], [232] și ilustrată și prin exemplele din acest capitol, impune definiții ale parametrilor săi bazate pe proprietăți 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 recunoașterii 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 prezentați în acest capitol au aplicații mergând de la delimitarea porțiunilor de semnal corespunzătoare vorbirii (energia și puterea) până la recunoașterea ei propriu-zisă.

Deși prelucrările în domeniul timp asigură extragerea unor caracteristici importante ale semnalelor vocale, se consideră că pentru recunoașterea automată a vorbirii esențiale sunt caracteristicile lor spectrale, cunoscută fiind capacitatea auzului uman de a distinge sunete după conținutul lor în componente de diferite frecvențe și 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 predicție liniară [153], sau cu luarea în considerație a unor particularități funcționale ale aparatului auditiv uman. Între aceste metode, fundamentală este analiza Fourier pe termen scurt, a cărei prezentare oferă ocazia și pentru discutarea unor legături între domeniile timp și frecvență.

Problema esențială a analizei spectrale a vorbirii pentru recunoașterea ei automată este separarea informației minime necesară în acest scop de variațiile ne semnificative, dar nu este încă foarte clar cum se poate realiza această separare. O primă posibilitate este sugerată de interpretarea spectrală a predicției liniare. O altă posibilitate este analiza homomorfică sau cepstrală, care poate fi utilizată pentru a reține doar informația referitoare la aspectul general al spectrului, determinat de forma tractului vocal, cu eliminarea detaliilor corespunzătoare excitației acestuia.

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

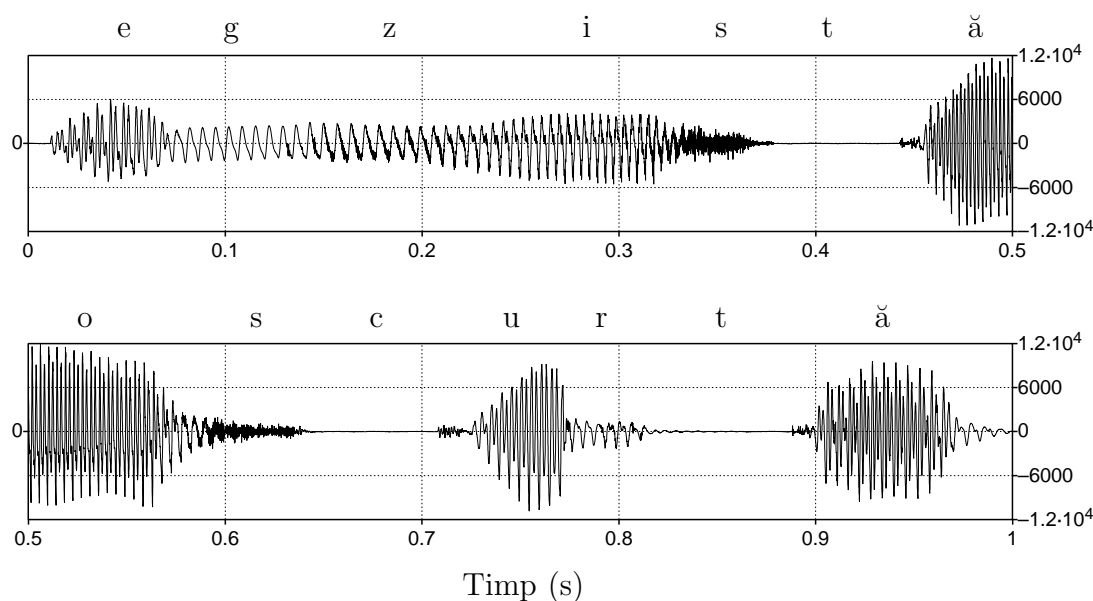


Figura 2.1: Exemplu de formă de undă a unui semnal vocal (cuvintele "există o scurtă", cu indicarea sunetelor efectiv pronunțate și a pozițiilor lor)

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, eșantionată la 16 KHz cu o rezoluție de 16 biți, a pronunției de către un bărbat a cuvintelor "există o scurtă", iar figura 2.1 prezintă forma ei de undă și sunetele efectiv pronunțate. Se observă că aspectul formei de undă variază foarte mult funcție de sunet, de la amplitudini foarte mici în timpul "t" și "c" și aspect aleator pentru "s", până la amplitudini maxime și aspect cvasiperiodic pentru vocale, de unde necesitatea analizei pe termen scurt.

O reprezentare care evidențiază simultan evoluția amplitudinii și a caracteristicilor spectrale ale unui semnal este **spectrograma** lui, care poate fi obținută prin diferite metode de estimare spectrală. În spectrogramă, intensitatea componentei de o anumită frecvență 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 nuanțele mai închise indică intensități mai mari.

Pentru exemplificare, figura 2.2 prezintă două spectrograme ale semnalului din figura 2.1. Funcție de "termenul scurt" de analiză Δt (v. secțiunile 2.2.1 și 2.5), invers proporțional cu rezoluția în timp și direct proporțional cu cea în frecvență, spectrogramele pot fi de **bandă largă**, cu rezoluție redusă în frecvență și mare în timp, respectiv de **bandă îngustă**, cu rezoluție mare în frecvență dar redusă în timp.

În cazul spectrogramei de **bandă îngustă** din figura 2.2, obținută analizând porțiuni de câte 30 ms, caracterul cvasiperiodic al semnalului pe durata vocalelor și a unor consoane se manifestă ca **striațiuni orizontale**, corespunzătoare armonicilor frecvenței fundamentale, F_0 , inversa perioadei fundamentale, T_0 , ambele variabile în timp. În cazul

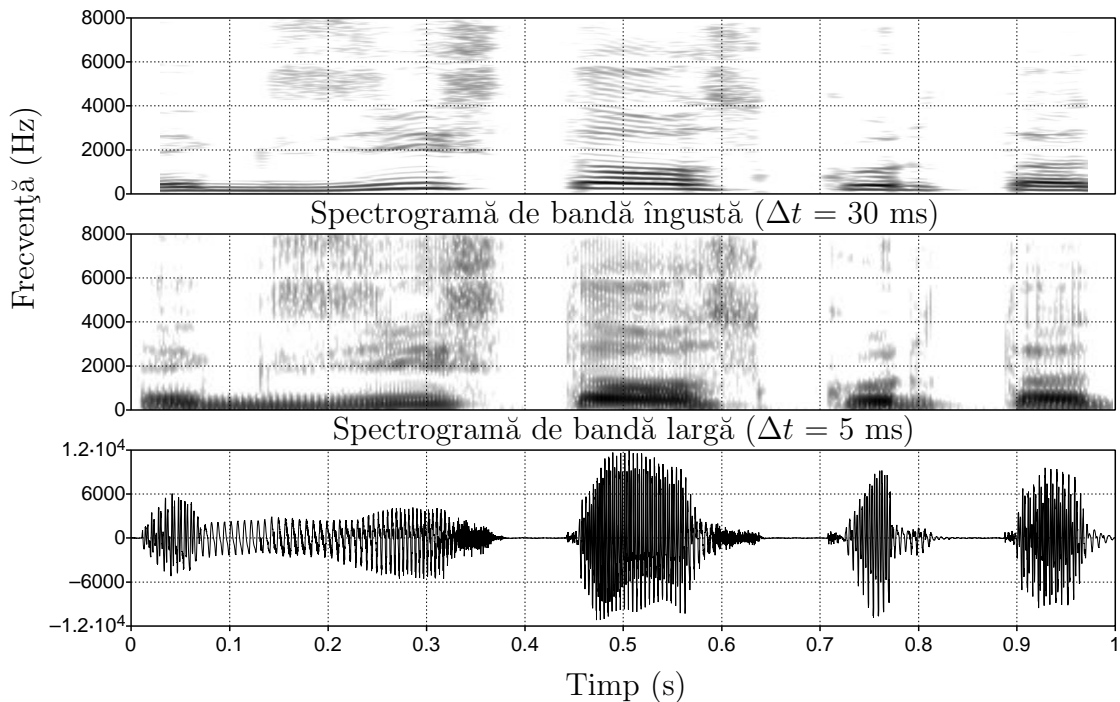


Figura 2.2: Spectrograme de bandă îngustă și largă ale semnalului din figura 2.1

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

Indiferent de tipul spectrogramei, se constată existența unor zone de intensitate sporită, situate la diferite frecvențe pe durata diferitor sunete. Aceste zone corespund așa-numiților **formanți**, care sunt rezonanțe ale unor cavități din tractul vocal (faringele, cavitatea bucală, cavitatea nazală) prin care oscilațiile de presiune acustică se propagă spre exterior pe durata producerii sunetelor vorbirii. Pentru același semnal din figura 2.1, figura 2.3 prezintă rezultatul unei estimări automate a valorilor frecvențelor primilor cinci formanți, reprezentate prin puncte suprapuse peste spectrograma de bandă largă. Frecvențele formanților, notate în literatură, în ordinea lor crescătoare, cu $F_1, F_2, F_3 \dots$,

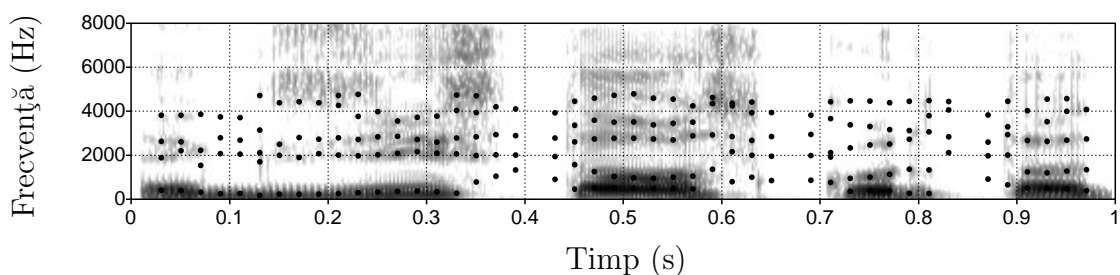


Figura 2.3: Spectrograma de bandă largă și formanții semnalului din figura 2.1

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 recunoașterii 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 și ferestruirea

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

Această operațiune este supusă unor constrângeri, uneori contradictorii:

- unele intervale trebuie să fie **suficient de scurte** pentru a evidenția 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 eșantioanele 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 eșantioanele semnalului, fiecare eșantion 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 secvență numerică de lungime finită.

Pentru atenuarea efectului porțiunilor partajate cu cadrele adiacente asupra valorilor parametrilor calculate pe baza unui cadru, eșantioanele dintr-un cadru pot fi ponderate funcție de poziția lor (pondere minimă la margini, maximă la centru) prin multiplicarea valorilor lor $s(n)$ cu o **ferastră** $w(n)$, operațiune cunoscută drept **ferestruire**:

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

După cum vom vedea în continuare, pe lângă motivul intuitiv al ponderării efectelor cadrelor adiacente, există și alte motive pentru aplicarea ferestruirii, în unele cazuri nu numai semnalului vocal, ci și 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

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

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

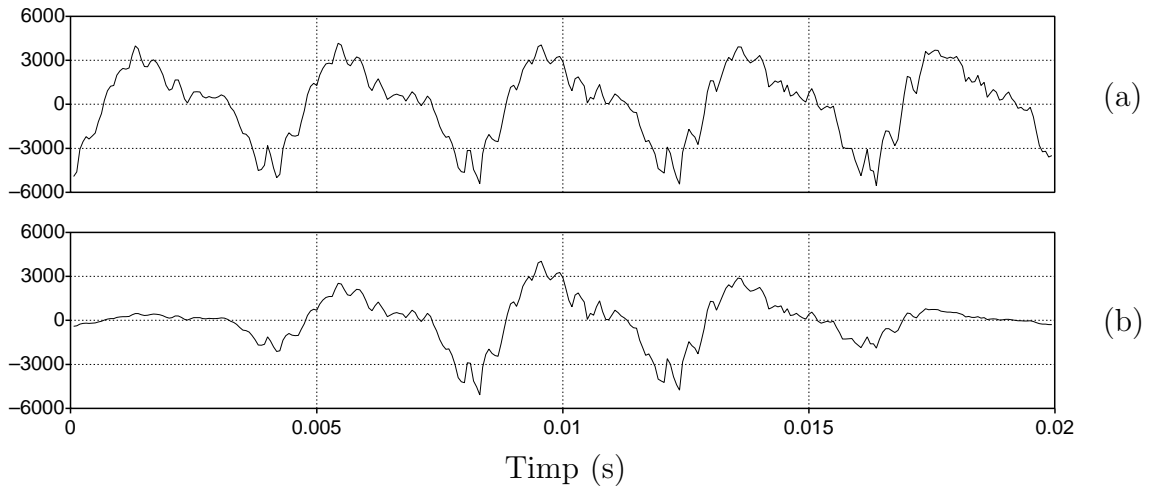


Figura 2.4: Un cadru de 20 ms dintr-un semnal vocal (a) și rezultatul multiplicării lui cu o fereastră Hamming (b)

2.2.2 Energia și puterea

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

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

Variante ale acestei definiții aplică o fereastră fie eșantioanelor semnalului [213]

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

fie pătratelor lor [180]

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

iar prin împărțirea la N a estimărilor energiei se obțin estimări ale puterii medii P .

Energia și puterea evidențiază 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 distincții 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 și logaritmică estimarea puterii medii a semnalului din figura 2.1 folosind cadre de 5 ms.

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

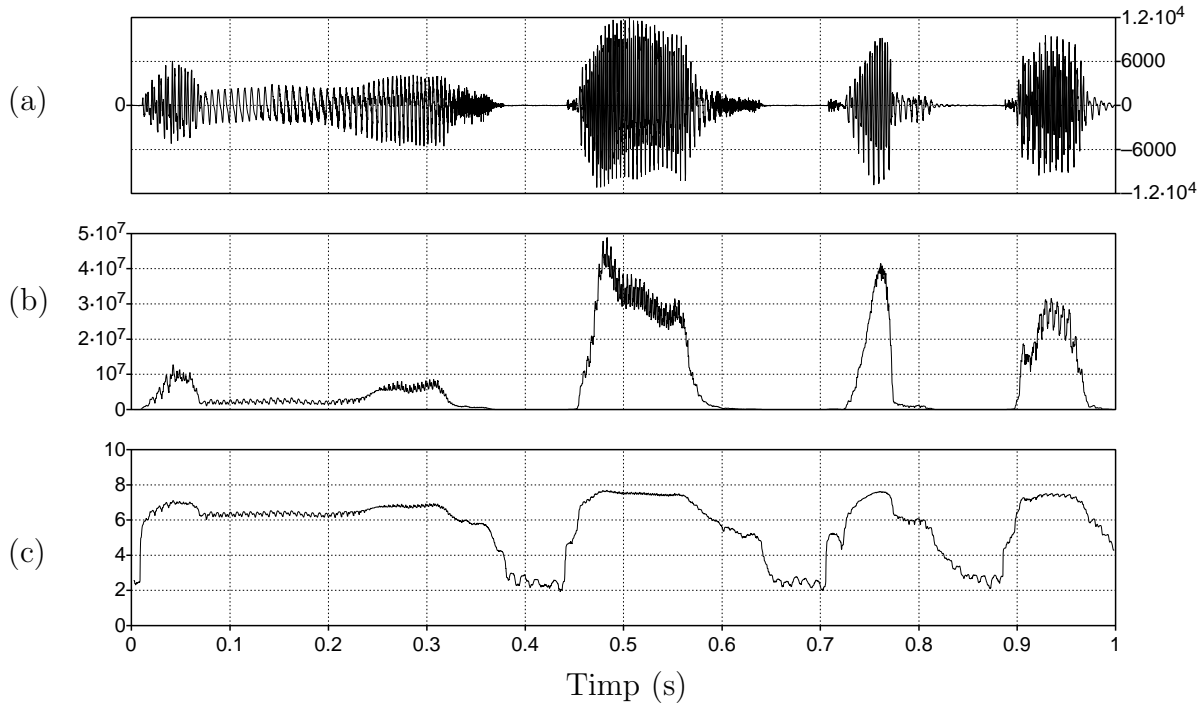


Figura 2.5: Estimare a puterii medii a semnalului din figura 2.1 (a) reprezentată pe scară liniară (b) și logaritmică (c)

2.2.3 Autocorelația

Pentru un semnal numeric, funcția de autocorelație este prin definiție [56] așteptarea statistică a produsului semnalului cu o replică a sa deplasată cu m eșantioane

$$R(m) = E[s(n)s(n+m)] \quad (2.6)$$

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

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

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

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

După cum ușor se poate demonstra și matematic, autocorelația pune în evidență periodicitatea semnalului. Ca exemplu, în figura 2.6 sunt date estimări ale funcției de autocorelație pentru o porțiune 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 periodicității imperfecte a semnalului) mai lent în cazul estimării nedeplasate decât în cel al estimării deplasate, când intervine și

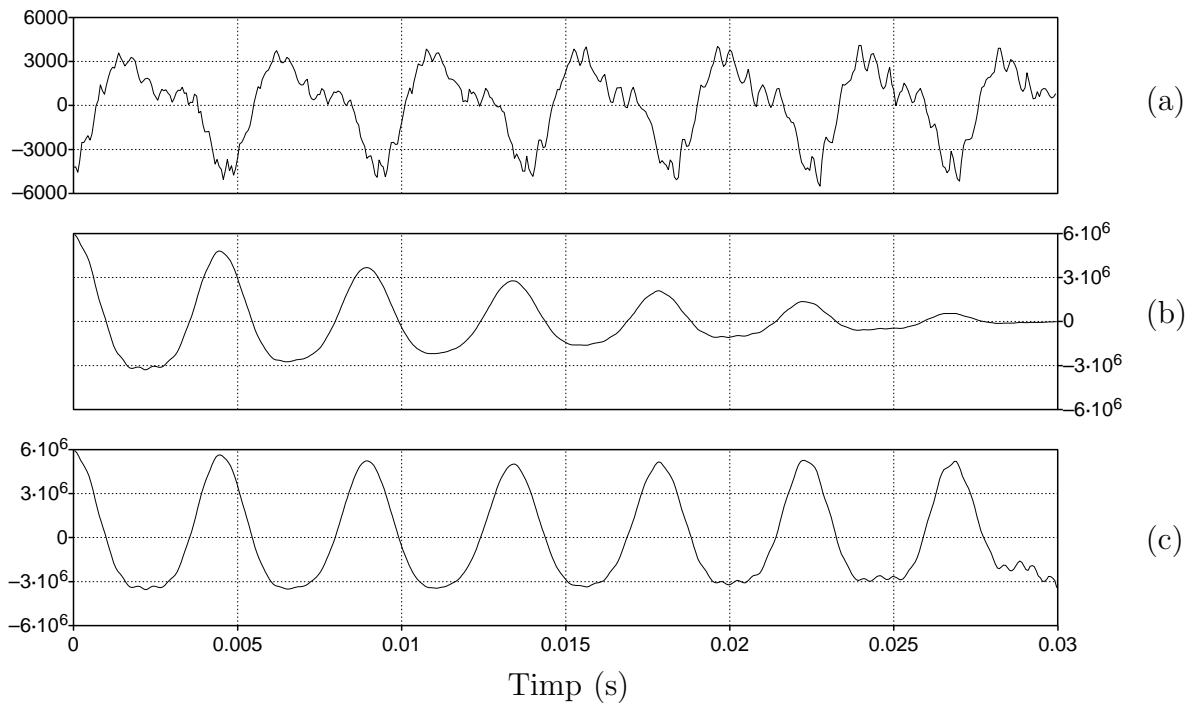


Figura 2.6: Autocorelația pentru o porțiune de semnal vocal sonor (a): estimare deplasată (b) și nedeplasată (c)

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

Remarcând și că puterea medie este acoperită de autocorelație 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 și a altor prelucrări ale semnalului vocal, însă importanța ei va putea fi pe deplin evidențiată doar după prezentarea analizei prin predicție liniară (secțiunea 2.3), utilizată în foarte multe aplicații.

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 eșantionarea și cuantizarea semnalului, fie abia după cadrarea sau ferestruirea lui, constă din trecerea lui printr-un filtru numeric cu ecuația

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

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

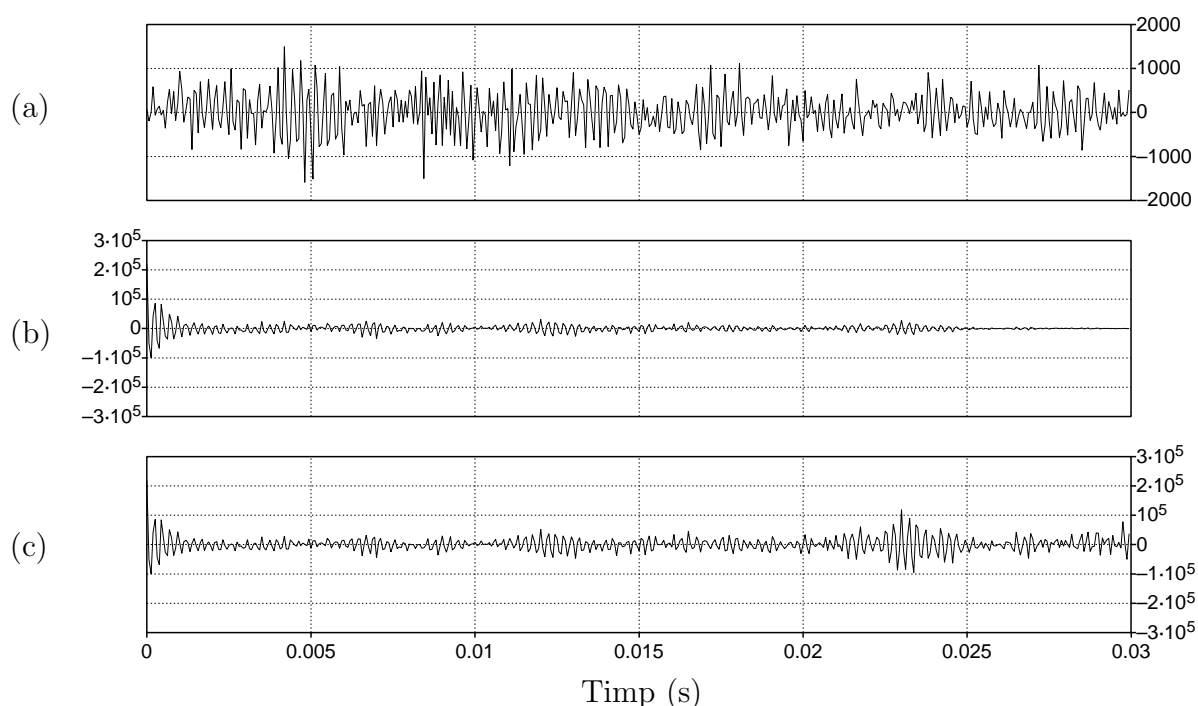


Figura 2.7: Autocorelația pentru o porțiune de semnal vocal nesonor (a): estimare deplasată (b) și nedepusată (c)

al radiației la ieșirea tractului vocal, preaccentuarea asigură o compensare a căderii de circa 12 dB/octavă a spectrului excitației sonore. În acest fel, pe durata sunetelor sonore anvelopa spectrală a semnalului preaccentuat va aproxima răspunsul în frecvență al tractului vocal, care prezintă interes din punct de vedere al recunoașterii automate a vorbirii datorită corelației dintre conținutul lingvistic al unui semnal vocal și evoluția 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 frecvențele înalte vor fi reprezentate cu niveluri de gri apropiate de cele ale frecvențelor joase (a se compara cu spectrograma de bandă largă din figura 2.2).

Deoarece excitația sonoră și 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 soluție simplă a acestei probleme, care are și avantajul de a asigura o compensare maximă a căderii spectrale, este aceea a utilizării unui coeficient de preaccentuare adaptiv [153]

$$\alpha = R(1)/R(0) \quad (2.11)$$

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

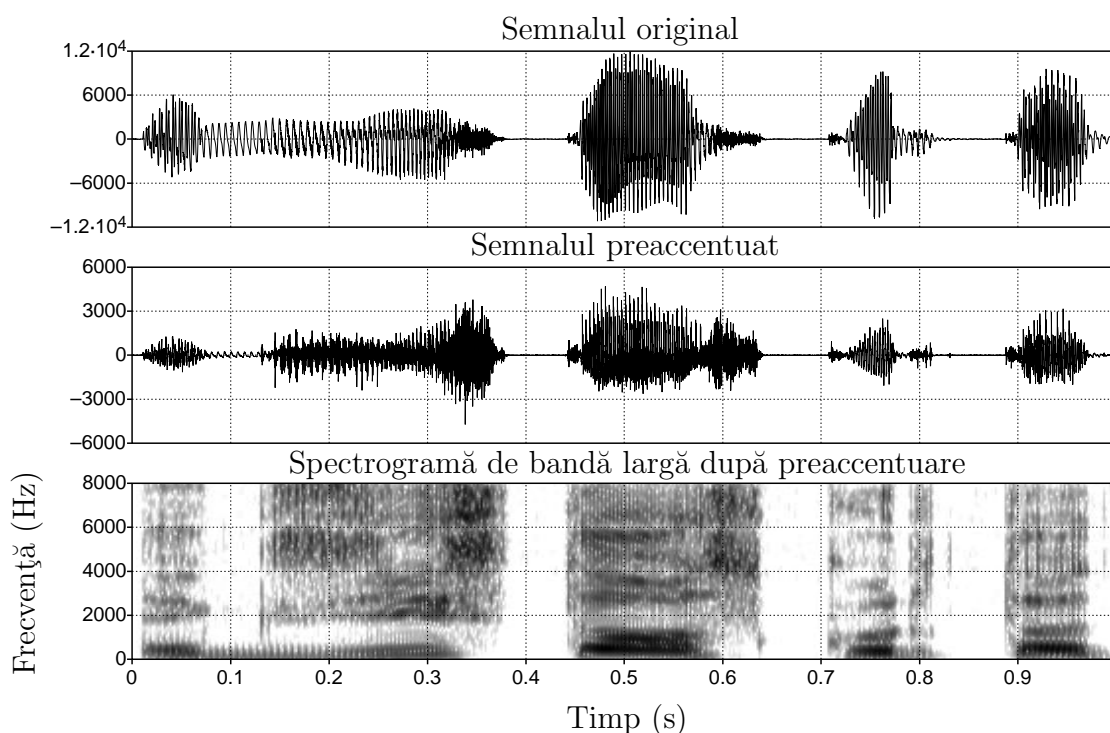


Figura 2.8: Efectul preaccentuării semnalului din figura 2.1 cu $\alpha = 0,96$

2.3 Analiza prin predicție liniară

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

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

unde P este **ordinul de predicție** și a_i sunt **coeficienții de predicție**. Diferența dintre valoarea astfel prezisă și cea efectivă constituie **eroarea de predicție**

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

a cărei transformată z este

$$E(z) = S(z)A(z) \quad (2.14)$$

unde

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

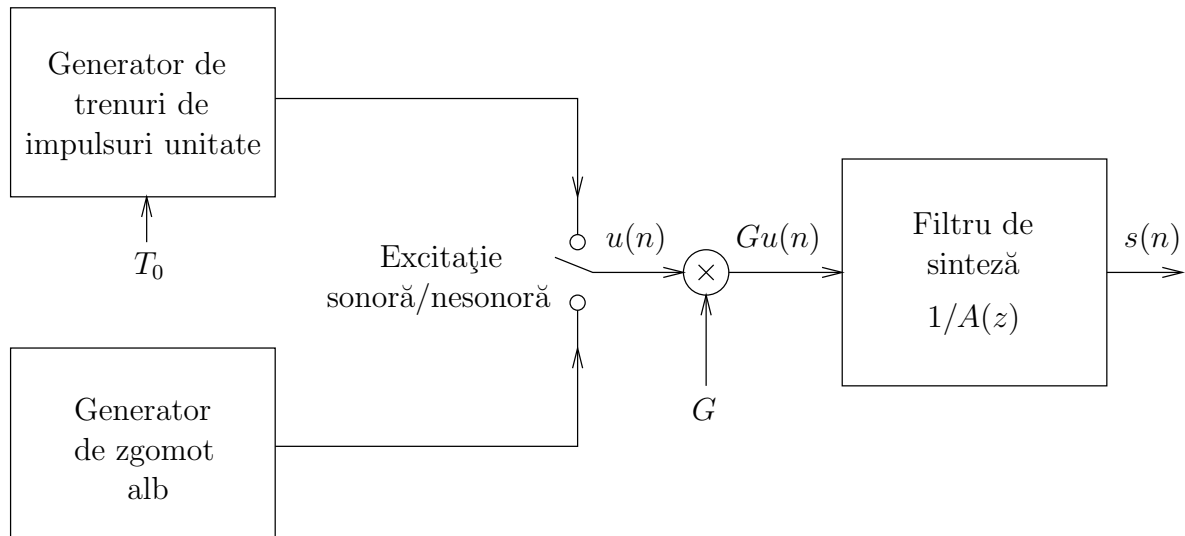


Figura 2.9: Modelul producerii vorbirii bazat pe predicție liniară

astfel că se poate scrie

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

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

Pentru porțiunile nesonore, problema compensării poliilor care modelează impulsurile glotale nu se mai pune, în schimb zerourile datorate fenomenelor de absorbție selectivă (antiformanți), ca și cele corespunzătoare radiației și preaccentuării, pot fi înlocuite printr-un număr de poli. În consecință, chiar dacă metoda nu mai pare "naturală", analiza prin predicție liniară poate fi folosită și în cazul sunetelor nesonore.

Se obține astfel un model bazat pe predicție 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 și radiației.

Numărul de coeficienți de predicție folosiți depinde de lărgimea de bandă a semnalului vocal analizat și de precizia dorită a modelării: fiecare formant impune utilizarea unei

perechi de poli complex conjugați pentru modelarea sa, iar alți câțiva 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 aștepta la un număr de formanți egal cu lărgimea de bandă în KHz a semnalului, ceea ce ar impune utilizarea unui **număr de poli**

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

unde F_s este frecvența de eșantionare a semnalului în KHz. Practic, nu există mai mult de cinci formanți semnificativi, astfel că de cele mai multe ori este suficient $P = 12 \dots 14$.

Coeficienții de predicție a_i corespunzători unui cadru pot fi calculați prin proceduri de optimizare care minimizează **eroarea pătratică de predicție**

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

Ținând cont de ecuația (2.13), eroarea pătratică de predicție 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 obțin pentru

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

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

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

unde

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

care poate fi scris în formă matricială

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

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) și (2.22), iar modul de alegere a lor duce la două metode fundamentale de analiză prin predicție liniară:

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

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

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

- **metoda autocorelației**, 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 (secțiunea 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

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

Diferența esențială între aceste metode constă în absența respectiv prezența ferestruirii și face ca fiecare să aibă proprietăți specifice ce o recomandă în anumite aplicații.

Aceste două metode, împreună cu altele și aplicații ale analizei prin predicție liniară, sunt tratate pe larg în literatură [153], [200], [59], dar datorită eficienței computaționale și stabilității soluțiilor furnizate, în recunoașterea automată a vorbirii se utilizează aproape în exclusivitate metoda autocorelației.

2.3.1 Metoda autocorelației

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

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

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

$$\sum_{k=1}^P R(|i-k|) a_k = R(i), \quad i = 1 \dots P \quad (2.27)$$

sau în formă matricială

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

Matricea $P \times P$ a valorilor autocorelației R din sistemul (2.28) este simetrică și cu elemente identice pe diagonale (Toeplitz), proprietate folosită pentru a obține 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 și ca **metoda Levinson-Durbin**) pentru rezolvarea ecuației $Ax = b$ cu A matrice Toeplitz pozitiv definită și b un vector oarecare, care ține cont de relațiile existente între elementele A și b în cazul metodei autocorelației.

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

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

și calculează seturi optime de coeficienți pentru predictorii de ordin i din ce în ce mai mare:

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

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

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

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

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

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

Pentru o modelare a semnalului folosind un predictor de ordin P , vor fi reținuți parametrii rezultați din ultima iterație a acestui algoritm. O reprezentare mai compactă a coeficienților de predicție este posibilă utilizând coeficienții $k_i \in [-1, 1]$, din care ei pot fi calculați, cunoscuți drept **coeficienți de reflexie** datorită unor legături care pot fi făcute cu modelarea tractului vocal prin tuburi acustice de secțiune constantă.

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

2.4 Analiza Fourier

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

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

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

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

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

unde limitele de sumare și integrare reflectă utilizarea timpului discret și a frecvenței normalizate (între $-1/2$ și $1/2$) 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 informația de fază, inutilă pentru recunoașterea vorbirii, se obține **spectrul de amplitudine** $|S(\omega)|$ al semnalului.

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

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

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

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

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

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

2.5 Legături timp–frecvență

Anunțate încă de la începutul capitolului, legăturile dintre domeniile timp și frecvență prezentate în continuare au fost alese pentru relevanța lor în analiza vorbirii în general și în contextul prezentei lucrări în special, fără a epuiza conexiunile existente între cele două domenii. Ele se referă la corelațiile în domeniul frecvență ale operațiunii de ferestruire din domeniul timp, precum și la posibilitatea de estimare a spectrului unui semnal pe baza analizei lui prin predicție liniară.

Efecte spectrale ale ferestruirii

După cum se cunoaște din proprietățile transformatelor Fourier [147], [56], înmulțirea a două semnale în domeniul timp are ca echivalent în domeniul frecvență convoluția

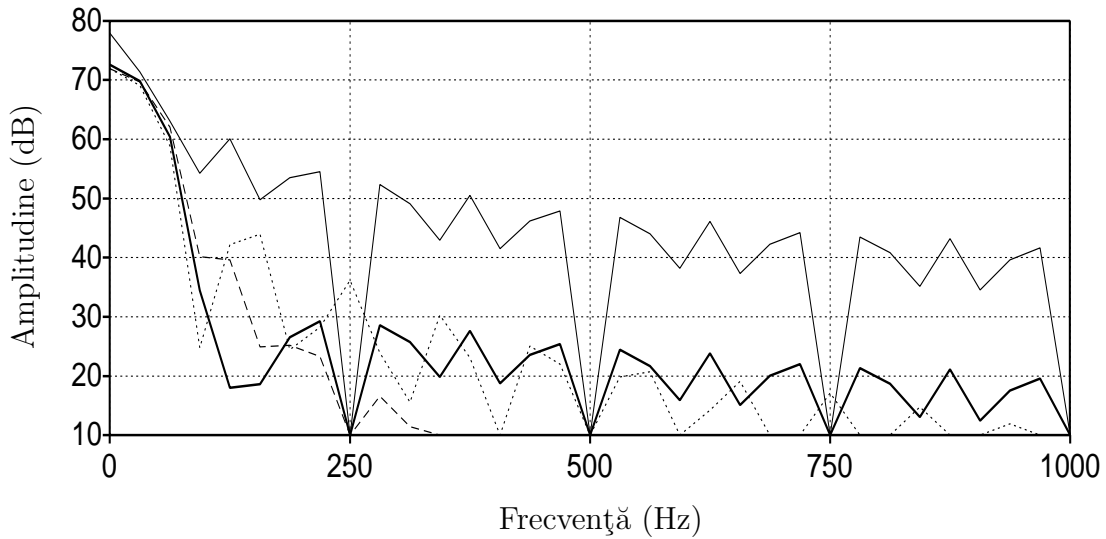


Figura 2.10: Spectre ale unor ferestre cu lungimea de 20 ms la frecvența de eșantionare de 16 KHz: Hamming (linie groasă), Hanning (întreruptă), triunghiulară (punctată), rectangulară (continuuă)

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

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

Porțiunile inferioare ale spectrelor unor ferestre dintre cele mai utilizate, exemplificate pentru o lungime de 20 ms și o frecvență de eșantionare $F_s = 16$ KHz, sunt prezentate în figura 2.10 (părțile lipsă continuă conform tendințelor sugerate): după cum se observă, toate spectrele au o componentă dominantă de joasă frecvență, denumită în literatura de specialitate **lob principal**, însoțită de **lobi secundari** de amplitudini din ce în ce mai mici spre frecvența maximă posibilă $F_s/2$.

Relația (2.39) implică faptul că **lărgimea de bandă** a lobului principal și **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 rezoluție bună în jurul unei anumite frecvențe Ω , iar atenuarea lobilor secundari previne influența componentelor de la frecvențe mai îndepărtate.

După cum se observă din figura 2.10, fereastra rectangulară, corespunzătoare simplei cadrări a semnalului, are cele mai slabe performanțe datorită lobului principal foarte larg și 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 condiții, 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 și rezoluția în frecvență (lărgimea de bandă a lobului principal)

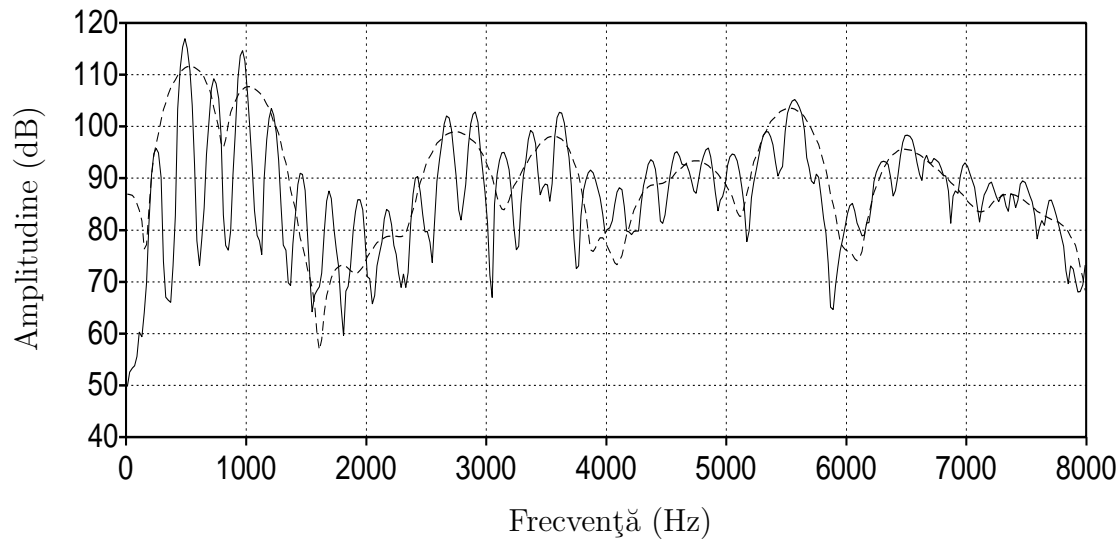


Figura 2.11: Estimări prin analiză Fourier ale spectrului semnalului din figura 2.1 la $t = 0,5$ s, realizate utilizând ferestre Hamming de 25 ms (linie continuă) și 5 ms (linie întreruptă)

ale unei ferestre sunt în relație inversă, menționată deja în secțiunea 2.1 și ilustrată în figurile 2.2 și 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 evidență în spectru prin maxime repetate, corespunzătoare armonicilor frecvenței fundamentale. Examinând structură periodică a spectrului, din faptul că armonica a patra este de aproape 1000 Hz putem chiar estima valorile frecvenței 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 predicției liniare

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

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

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

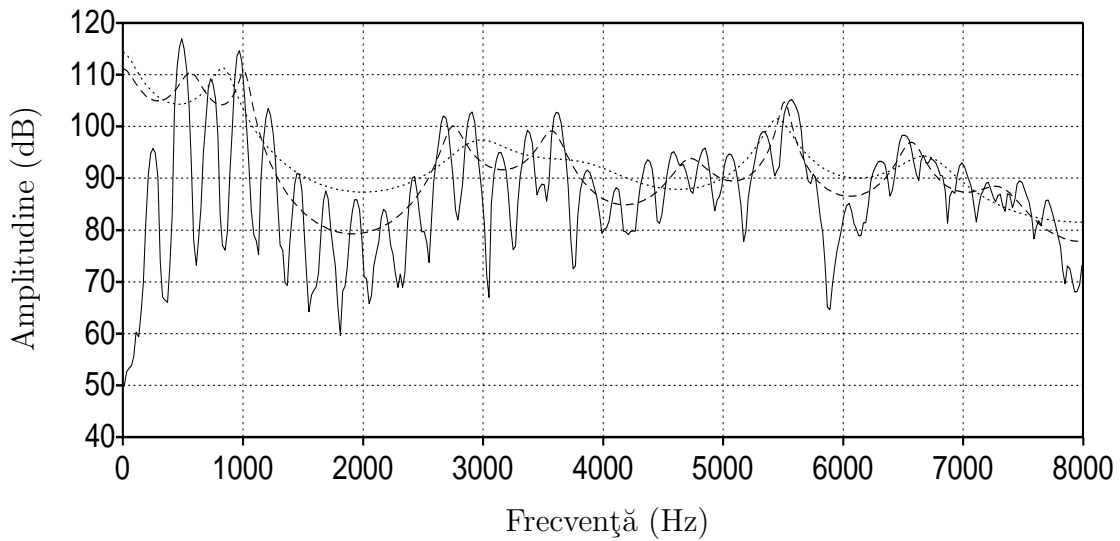


Figura 2.12: Estimări ale spectrului semnalului din figura 2.1 la $t = 0,5$ s, realizate cu o fereastră Hamming de 25 ms, prin analiză Fourier (linie continuă) și predicție liniară cu 20 (linie întreruptă) respectiv 12 poli (linie punctată)

un **spectru normalizat**

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

care pe o scară logaritmică are valoarea medie nulă

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

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

Ca și în cazul analizei Fourier de bandă largă, spectrul obținut este o aproximare, de astă dată în principal a formanților, iar numărul polilor folosiți determină precizia acestora. Pentru exemplificare, figura 2.12 prezintă trei estimări ale spectrului unei porțiuni de semnal vocal sonor prin analiză Fourier de bandă îngustă și predicție liniară cu 20 respectiv 12 poli. A doua variantă corespunde utilizării formulei (2.17) și asigură o foarte bună (poate chiar prea bună) aproximare a formanților, în timp ce a treia ”contopește” perechi de poli din cea anterioară.

2.6 Analiza homomorfică

Analiza Fourier de bandă îngustă a porțiunilor sonore ale semnalelor vocale (figurile 2.2, 2.11 și 2.12) pune în evidență prezența unor componente periodice ale spectrelor lor, corespunzătoare armonicilor frecvenței fundamentale, datorate excitației sonore care stă la baza acestor porțiuni, și cunoscute sub numele de **structură spectrală fină**.

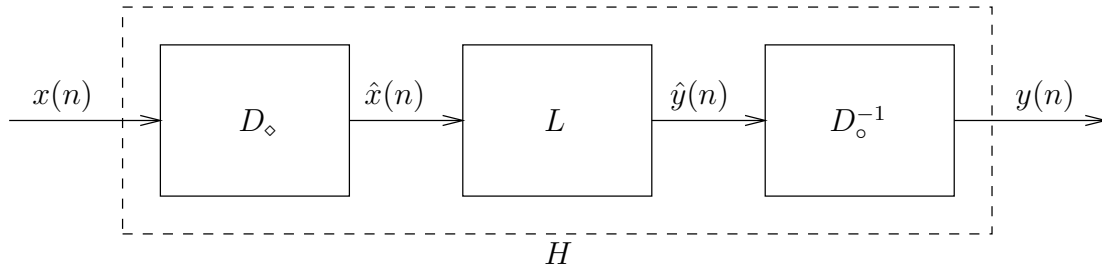


Figura 2.13: Reprezentarea canonică a unui sistem homomorfic

Această structură fină este eliminată prin netezirea asigurată de analiza Fourier de bandă largă (secțiunea 2.5, figura 2.11), care reține în principal informații despre aspectul general al spectrului, ceea ce pentru recunoașterea automată a vorbirii este avantajos deoarece variațiile frecvenței fundamentale pe durata pronunțiilor unui cuvânt (excepție 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) și frecvență ridicată, astfel că prezintă interes găsirea unei metode apte să asigure această **netezire spectrală** din cadre mai lungi (de ordinul zecilor de milisecunde) și cu frecvență mai scăzută (tipic 100 pe secundă).

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

$$L[x_1(n) + x_2(n)] = L[x_1(n)] + L[x_2(n)] \quad (2.43)$$

$$L[cx(n)] = cL[x(n)] \quad (2.44)$$

unde L este o transformare liniară.

În cazul unui sistem homomorfic, există operatori pentru combinarea intrărilor, \diamond , combinarea intrărilor cu scalari, \bullet , combinarea ieșirilor, \oplus , și combinarea ieșirilor cu scalari, \odot , iar **principiul generalizat al superpoziției** este

$$H[x_1(n) \diamond x_2(n)] = H[x_1(n)] \oplus H[x_2(n)] \quad (2.45)$$

$$H[c \bullet x(n)] = c \odot H[x(n)] \quad (2.46)$$

unde H este o transformare homomorfică.

Rezultatele referitoare la sisteme liniare pot fi utilizate și pentru cele homomorifice 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 operațiunile \diamond și \bullet) și stabilește un homomorfism între spațiul vectorial de intrare și unul intermediar, în care \diamond și \bullet sunt homomorfe cu adunarea și înmulțirea cu scalari

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

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

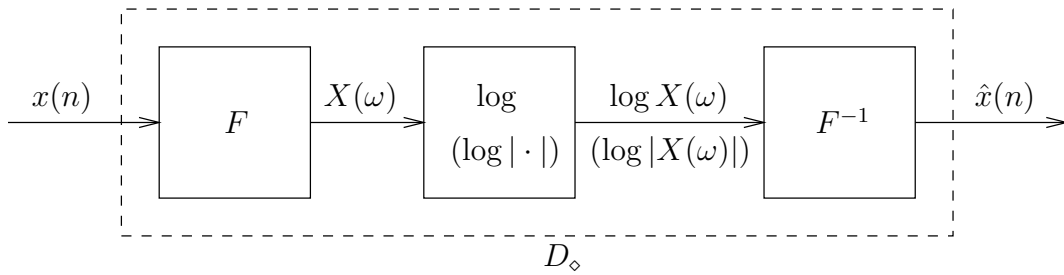


Figura 2.14: Schema bloc a unui sistem caracteristic de intrare pentru prelucrări homomorifice (cepstrale) ale semnalului vocal

Al doilea este un sistem liniar, iar ultimul este **sistemul caracteristic de ieșire** și asigură revenirea la spațiul vectorial al ieșirilor

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

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

Semnalul vocal poate fi modelat ca rezultat al convoluției în timp, respectiv înmulțirii în frecvență, dintre excitația 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 condiții, pentru prelucrarea homomorfică a semnalului vocal, sistemul caracteristic de intrare trebuie să asigure transformarea convoluției $e(n) * v(n)$ în suma transformatelor, $\hat{e}(n) + \hat{v}(n)$. Ținând cont de corespondențele

$$F[e(n) * v(n)] = F[e(n)] F[v(n)] = E(\omega) V(\omega) \quad (2.51)$$

$$\log[E(\omega) V(\omega)] = \log E(\omega) + \log V(\omega) \quad (2.52)$$

unde $*$ este operatorul de convoluție iar F transformata Fourier directă, precum și 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

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

Transformata Fourier inversă F^{-1} realizează o trecere din domeniul frecvență într-un nou domeniu de tip timp, diferit însă de cel al semnalului original, iar pentru descrierea lui au fost introduși termeni obținuți prin anagramarea unora din domeniul frecvență. Astfel, domeniul ca atare este numit **domeniul cvefrență** (anagramă din frecvență, cf. engl. quefreny, anagramă din frequency); $\hat{x}(n)$ este **cepstrul** (anagramă din spectrul, pronunțată kepsstrul) semnalului $x(n)$, având aceeași frecvență de eșantionare ca și acesta; iar prelucrarea semnalului ca atare a fost denumită **analiză cepstrală**.

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

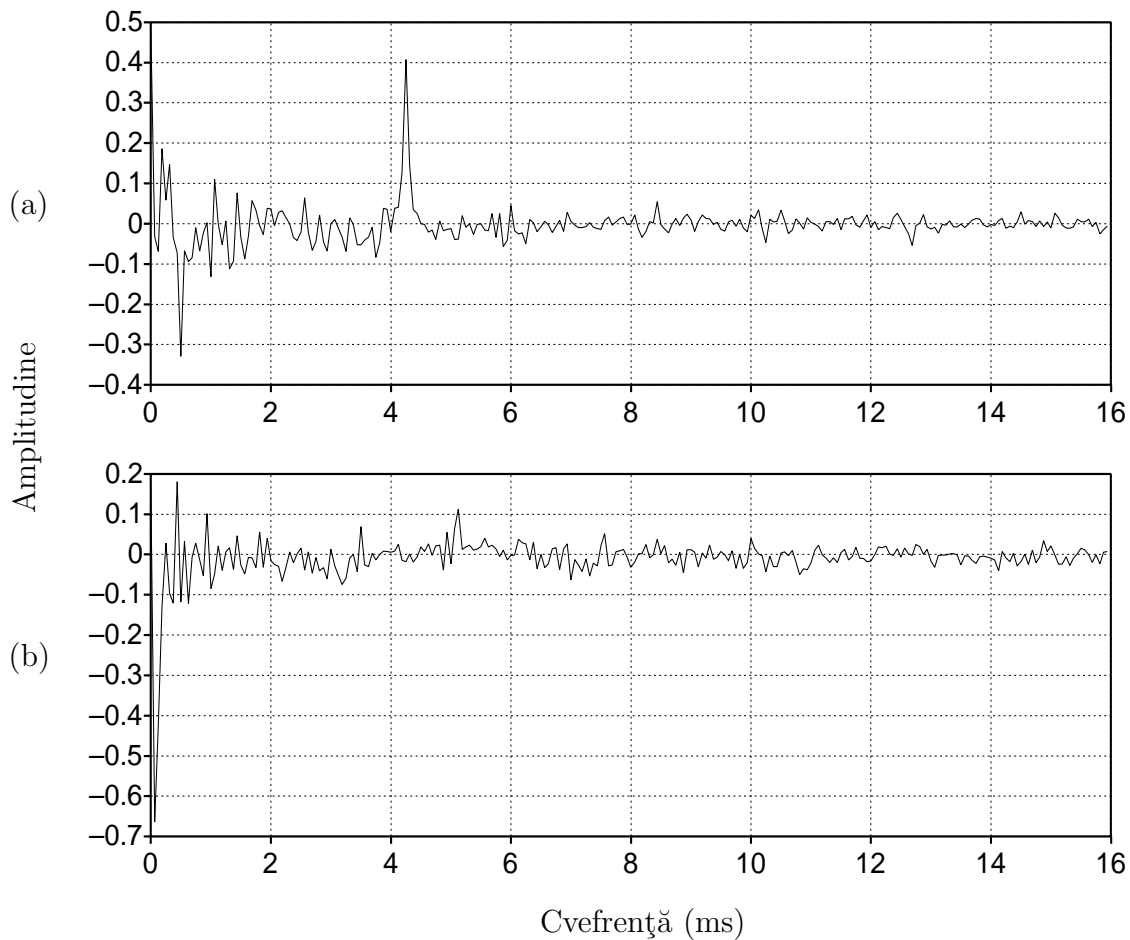


Figura 2.15: Estimări cu o fereastră Hamming de 25 ms ale cepstrului semnalului din figura 2.1 la (a) $t = 0,5$ ms (sonor) (b) $t = 0,6$ ms (nesonor)

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 ecuația (2.53) este denumit **cepstru complex**. Utilizarea logaritmului complex ridică însă o serie de probleme de ordin teoretic [175], iar informația de fază nu are importanță pentru recunoașterea vorbirii, astfel încât cel mai frecvent este utilizat **cepstrul real**, singurul considerat în continuare, obținut prin neglijarea fazei. Pentru un cadru de lungime N dintr-un semnal vocal $s(n)$, cepstrul real este o secvență de numere reale, cunoscute sub numele de **coeficienți cepstrali**, de aceeași lungime

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

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

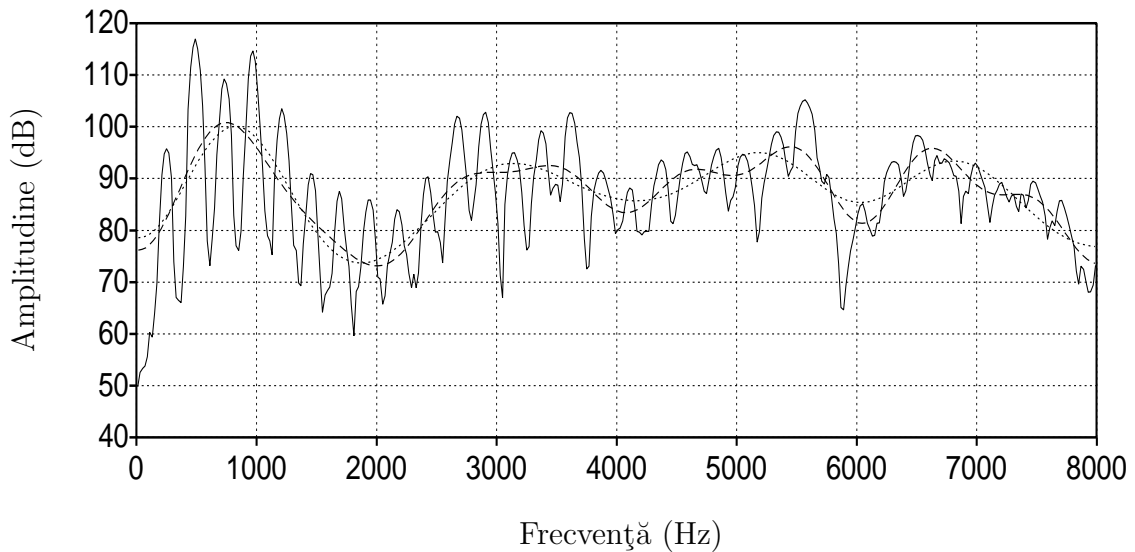


Figura 2.16: Estimări ale spectrului semnalului din figura 2.1 la $t = 0,5$ s, realizate cu o fereastră Hamming de 25 ms prin analiză Fourier de bandă îngustă (cu linie continuă) și netezite prin trunchiere cepstrală la 1 ms (cu linie întreruptă) respectiv 0,75 ms (cu linie punctată)

prezintă de asemeni simetrie pară, iar figura 2.15(a) prezintă aspectul porțiunii sale inferioare pentru semnalul pe baza căruia s-au realizat estimările spectrale din figurile 2.11 și 2.12. Se observă prezența unui maxim, corespunzător excitației sonore, la cvefrența de 4,25 ms, care concordă cu estimarea $T_0 \approx 4,2$ ms din secțiunea 2.5. Această evidențiere în cepstru a excitației sonore este motivul pentru care una din cele mai frecvente aplicații ale sale este estimarea valorilor perioadei și frecvenței fundamentale.

Pentru porțiunile 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 și în algoritmi de decizie asupra caracterului sonor sau nesonor al semnalului.

Ținând cont de ecuațiile (2.53) și (2.54), de precizările făcute pe marginea lor și de exemplele prezentate, valorile cepstrului de la cvefrențe joase pot fi puse în corespondență cu componenta lent variabilă a spectrului, determinată de tractul vocal, $V(\omega)$, iar cele de la cvefrențe superioare – cu structura spectrală fină $E(\omega)$, datorată excitației.

Această interpretare a cepstrului stă la baza eliminării structurii spectrale fine și 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 cvefrență. Dacă această operațiune 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 și netezite prin trunchiere cepstrală ale semnalului analizat și în figurile 2.11 și 2.12. Deși sunt folosiți numai primii 17 respectiv 13 coeficienți cepstrali (corespunzători cvefrențelor superioare de 1 și 0,75 ms), spectrele estimate prin trunchiere cepstrală prezintă o netezire

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

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

În încheiere, mai menționăm că predicția liniară de ordin P permite și calculul unei estimări a cepstrului, numită **cepstru de predicție liniară**, pe baza relațiilor

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

2.7 Metode perceptuale

Una din căile posibile pentru îmbunătățirea performanțelor sistemelor de recunoaștere automată a vorbirii este includerea în analiza semnalului vocal a unor prelucrări care emulează proprietăți 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 și funcțiile sistemului auditiv pe baza rezultatelor unor studii de anatomie, fiziologie și psihofiziologie, și **metode perceptuale**, care includ proprietățile acestuia fără considerarea structurilor anatomice și 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] și 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] și auditorii [222] și mergând până la implementări hardware [250], [134].

Deși în mod teoretic metodele auditorii pot contribui la apropierea performanțelor sistemelor de recunoaștere automată a vorbirii de cele umane, iar unele experimente susțin superioritatea lor [149], [145], practic această superioritate nu se manifestă [53] sau se manifestă numai în condiții de mediu dificile [119], [222]. Această situație este explicabilă prin insuficiența cunoașterii a mecanismelor pe care aceste metode încearcă să le modeleze, ceea ce împreună cu costul computațional ridicat face ca ele să nu aibă o utilitate clară în recunoașterea automată a vorbirii.

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

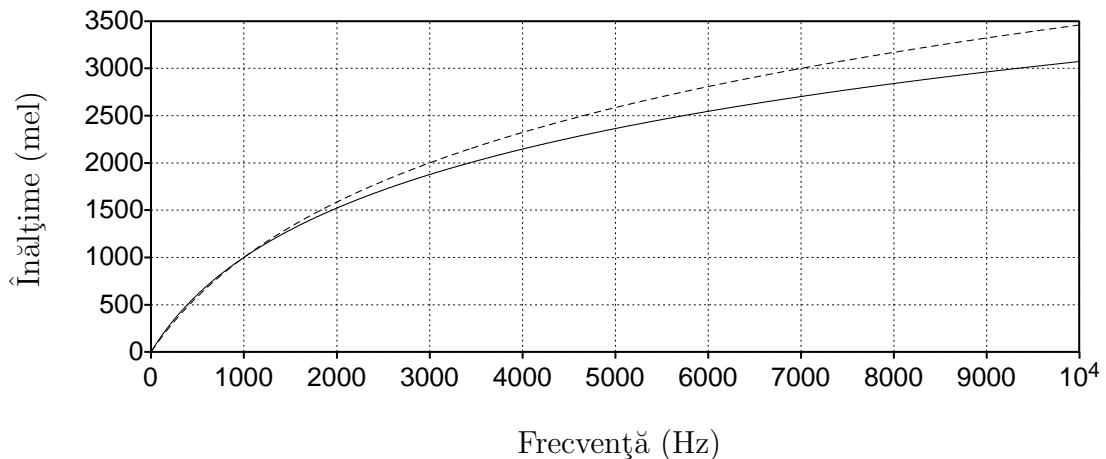


Figura 2.17: Corespondența dintre înălțime și frecvență conform ecuațiilor (2.56) (linie continuă) și (2.57) (linie întreruptă)

2.7.1 Analiza melodică

Înălțimea este o proprietate subiectivă a sunetelor [228], [173], [208], dependentă în principal de frecvența lor, pe baza căreia ele sunt clasificate de la grave la acute. Relația cantitativă dintre frecvențele și înălțimile tonurilor pure a fost studiată experimental, iar pentru măsurarea înălțimii s-a introdus o **scară melodică** cu unitatea de măsură **mel**: prin definiție [208], un ton pur de 1000 Hz cu un nivel de 40 dB are înălțimea 1000 mel, iar frecvențele tonurilor pure percepute ca având înălțimi de n ori mai mari sau mai mici sunt puse în corespondență cu aceste înălțimi.

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

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

respectiv Fant [69]

$$m = 1000 \ln(1 + f/1000) / \ln 2 \quad (2.57)$$

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

După cum s-a arătat în secțiunea 2.6.1, analiza cepstrală oferă o reprezentare foarte compactă, sub forma coeficienților cepstrali inferiori, a informației spectrale asociată tractului vocal, reținută în spectrul netezit prin trunchiere cepstrală. **Cepstrul melodic** [57] este obținut prin modificarea analizei cepstrale, cu luarea în considerație a percepției neliniare cu frecvența a înălțimii sunetelor.

Integrarea în analiza cepstrală a fenomenului percepției neliniare cu frecvența a înălțimii se face estimând un **spectru melodic** după transformarea Fourier directă

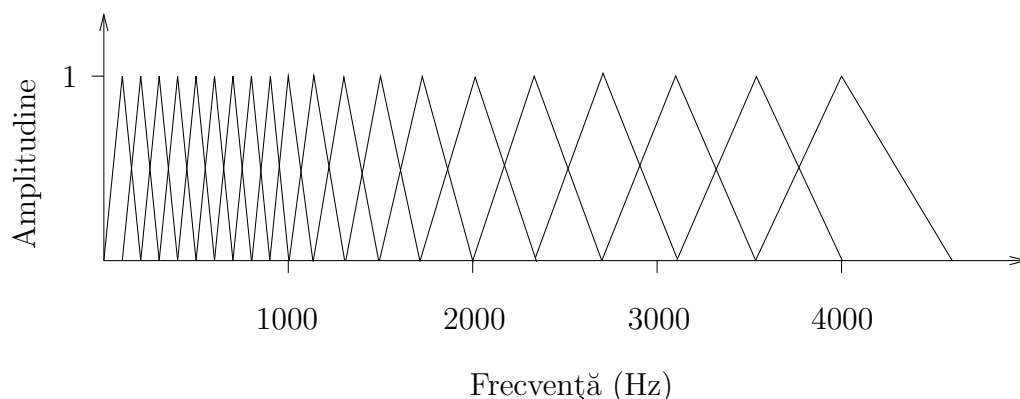


Figura 2.18: Exemple de ferestre utilizate în analiza melodică [57]

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

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

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 și exponențial crescătoare, corespunzător uneia uniformă pe scara melodică, peste această frecvență, ca exemplu putând fi dat chiar setul de ferestre utilizat în [57] (figura 2.18), dar se pot utiliza și ferestre rectangulare adiacente, uneori cu rezultate superioare [27].

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

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

unde L este numărul de coeficienți cepstrali melodici care se dorește să fie calculați.

2.7.2 Caracteristicile dinamice

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

Inițial, au fost utilizate aproximări ale valorilor derivatelor temporale ale coeficienților cepstrali și log-energiei, calculate cu formula

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

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 ecuația (2.60) realizează o netezire prea mare, și care are și avantajul simplității, este

$$d(t) = c(t+L) - c(t-L) \quad (2.61)$$

Aceste caracteristici dinamice sunt în prezent cunoscute ca delta cepstru respectiv delta log-energie, pe scurt coeficienți 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 coeficienți delta-delta ($\Delta\Delta$). Sub o formă sau alta, aceste caracteristici dinamice au fost folosite în numeroase sisteme de recunoaștere automată a vorbirii, în care eficiența lor a fost în mod repetat verificată.

Pe lângă acestea, au existat și există și alte încercări de utilizare a informațiilor despre variațiile spectrale: prin definirea de noi caracteristici (de exemplu funcții de variație 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 aplicație 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 predicție liniară, analiza cepstrală. Pentru o mai bună înțelegere a particularităților lor, prezentarea acestor metode a fost precedată de descrierea altor prelucrări și parametri (cadrare, ferestruire, energie, putere, autocorelație, preaccentuare) care intervin pe parcursul lor, iar de-a lungul întregului capitol au fost evidențiate multiplele legături existente între ele.

Una din posibilitățile de creștere a performanței sistemelor de recunoaștere 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 considerații asupra acestei posibilități, cu detalierea a două metode care o materializează: analiza melodică și caracteristicile dinamice.

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

vocabular mic de cuvinte pronunțate 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 calitățile coeficienților cepstrali în recunoașterea automată a vorbirii și au arătat că performanțele cele mai bune se obțin prin utilizarea cepstrului melodic.

CAPITOLUL 3

Recunoașterea automată a vorbirii

Indiferent de abordările folosite în încercările de a rezolva diferite probleme ale recunoașterii automate a vorbirii, esențială pentru aprecierea și compararea acestora este **evaluarea performanțelor** sistemelor în care ele sunt implementate. Din acest motiv, prezentarea unor aspecte ale evaluării performanțelor sistemelor de recunoaștere automată a vorbirii precede trecerea în revistă în acest capitol a unora dintre cele mai importante metode folosite în aceste sisteme și utilizate și 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 recunoașterii automate a vorbirii pot fi grupate în vectori acustici dintr-un spațiu acustic. Astfel, o primă problemă care apare este cea a evaluării unor **distanțe acustice**, care pot fi folosite împreună cu metode de recunoaștere a formelor [64], [76], [243] pentru a încerca o estimare locală, la nivel de cadre, a identității lingvistice a semnalelor vocale.

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

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

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

Pe lângă modelarea acustică și cea lingvistică, recunoașterea automată a vorbirii mai necesită și **reprezentarea integrată a cunoștințelor** acumulate de modele într-un spațiu unic de căutare a soluțiilor. În acest spațiu, estimarea șirului de cuvinte pronunțat se poate face prin aplicarea unor **algoritmi de căutare** având ca obiectiv maximizarea probabilității unui ipotetic șir de cuvinte în raport cu modelele.

3.1 Evaluarea performanțelor

Sistemele de recunoaștere automată a vorbirii pot face erori, ca și oamenii, în ceea ce privește șirul de cuvinte pe care îl estimează a fi fost pronunțat (figura 1.1), iar obiectivul lor fiind tocmai estimarea corectă a acestui șir, principalele metrice pentru evaluarea performanțelor lor se definesc pe baza erorilor apărute în acest proces.

Performanțele pot fi influențate de foarte mulți factori, astfel încât orice valori ale metricilor trebuie însoțite de precizarea condițiilor în care au fost obținute, condiții determinate de caracteristicile aplicației și ale sistemului de recunoaștere evaluat. Câteva dintre aceste condiții sunt:

- **mărimea vocabularului**, dată de numărul de cuvinte distincte din dicționarul sistemului evaluat și pe care acesta le poate recunoaște: într-o aplicație necesitând un vocabular redus, în care apar puține cuvinte între care sistemul trebuie să aleagă la un moment dat, erorile pot apare mai rar decât într-o aplicație cu un vocabular mai mare, între cuvintele căruia pot exista mai multe posibilități de confuzie;
- **tipul vocabularului**: funcție de luarea sau nu în calcul a posibilității ca în pronunții să apară și cuvinte din afara lui, putem avea un vocabular **deschis**, în care un singur cuvânt special va corespunde tuturor cuvintelor necunoscute care pot apare în pronunții, sau un vocabular **închis**;
- **cuvintele din vocabular**: chiar dacă o aplicație utilizează un vocabular mic și închis, prezența în el și utilizarea în proporție mare a unor cuvinte ușor de confundat din punct de vedere acustic va duce la creșterea frecvenței erorilor;
- **tipul pronunțiilor**: un sistem pentru recunoașterea pronunțiilor **discrete**, a unor cuvinte rostite cu pauze între ele, va putea obține performanțe superioare unuia pentru recunoașterea vorbirii **continue**, fără pauze între cuvinte, care trebuie să realizeze și delimitarea, nu numai recunoașterea cuvintelor;
- **stilul vorbirii**: sunt mai ușor de recunoscut semnalele vocale obținute prin **citire**, respectând restricțiile impuse de un text, decât cele rezultate dintr-o **conversație** 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 aplicației**: în cazul unei aplicații în care secvența cuvintelor este rigidă, iar fiecare cuvânt are un număr redus de succesori posibili, erorile vor fi mai puțin 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 performanțe mai bune decât unul cu modele **independente** de vorbitor, iar performanțele acestuia din urmă pot crește prin **adaptare** la vorbitor a modelelor;
- **condițiile acustice** ale mediului în care sunt rostite pronunțiile: zgomotele, reverberațiile etc. vor duce la scăderea calității semnalului și a performanțelor;

- **caracteristicile canalului de comunicație** dintre utilizator și sistem: un canal de comunicație cu o bandă mai îngustă (de exemplu de tip telefonic) sau afectat de zgomote, distorsiuni etc. poate duce la scăderea performanțelor;
- **vorbitorii** folosiți pentru realizarea evaluării, caracteristicile vocii și ale vorbirii fiind dependente de multe variabile biologice (sex, vârstă) și sociale (grad de educație, ocupație, mediu social) care pot afecta performanțele fie prin intermediul semnalului vocal, fie prin modul de utilizare a sistemului în aplicație.

Erorile din funcționarea unui sistem de recunoaștere automată a vorbirii pot apare datorită imperfecțiunii modelelor acustice sau lingvistice folosite (**erori de modelare**) sau eliminării greșite din procesul de căutare a unor cuvinte corecte (**erori de căutare**). Funcție de modul în care se manifestă la nivelul șirului de cuvinte estimat, aceste erori se pot clasifica în:

- **substituții** ale unui cuvânt pronunțat cu un altul, datorate asemănării acustice dintre ele în condițiile unei modelări acustice imperfecte, sau absenței cuvântului corect dintre alternativele permise la un moment dat;
- **insertii** ale unor cuvinte nepronunțate: acestea sunt de obicei cuvinte scurte, cel mai adesea monosilabice (prepoziții, conjuncții, pronume), asemănătoare din punct de vedere acustic cu porțiuni ale unor cuvinte mai lungi;
- **omisiuni** (în engl. deletions) ale unor cuvinte pronunțate.

Evaluarea performanțelor unui sistem de recunoaștere automată a vorbirii presupune existența unor transcrieri de referință ale semnalelor vocale folosite în acest scop, cu care să fie comparate șirurile corespunzătoare de cuvinte estimate de sistem, și a unor metode de determinare a apariției și tipului erorilor. Odată detectate și clasificate erorile, se pot calcula valorile unor metrice de performanță:

- frecvența recunoașterilor corecte sau corectitudinea

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

- frecvența substituțiilor

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

- frecvența insertiilor

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

- frecvența omisiunilor

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

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

Fiecare din metricile anterioare oferă o imagine doar asupra unui singur aspect din funcționarea unui sistem, astfel încât pentru o caracterizare globală a performanțelor au mai fost introduse două metrici care țin cont de toate tipurile de erori posibile:

- frecvența erorilor

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

- acuratețea

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

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 șirurile estimate de sistem prin comparații cu transcrierile de referință corespunzătoare. În mod tradițional, șirurile estimate sunt considerate ipoteze formulate de sistem, iar rezultatele comparațiilor sunt prezentate ca alinieri ale ipotezelor cu transcrierile de referință, însoțite de tipul fiecărei erori detectate – recunoașterile corecte nu sunt marcate:

Referință:	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 referință corespunzătoare se poate face calculând printr-un algoritm de programare dinamică distanța de editare dintre ele, definită ca fiind costul total minim al operațiunilor de editare (substituție, inserție sau ștergere/omisiune), fiecare cu un cost specific, prin care unul din cele două șiruri de cuvinte este transformat în celălalt [249].

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

Ținând cont de acest principiu, evaluarea distanței de editare $D(I, R)$ dintre ipoteza I de N_i cuvinte și referința asociată R de N_r cuvinte se poate face recursiv: notând cu C_s , C_i și C_o costul unei substituții, inserții respectiv omisiuni, distanța parțială $D(I_{1,i}, R_{1,r})$ dintre prefixele $I_{1,i}$, $R_{1,r}$ de lungimi i și r ale I respectiv R poate fi scrisă

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

Algoritmul 3.1 Compararea prin programare dinamică a unei ipoteze cu referința

```

1:  $N_i \leftarrow$  lungimea ipotezei  $I$ ,  $N_r \leftarrow$  lungimea referinței  $R$ 
2: alocă  $D[N_i, N_r]$ ,  $E[N_i, N_r]$ ,  $e[N_i + N_r - 1]$ 
3:  $D[i, r] \leftarrow \infty$ ,  $i = 1 \dots N_i$ ,  $r = 1 \dots N_r$ 
4: for  $i = 1 \dots N_i$  do
5:   for  $r = 1 \dots N_r$  do
6:      $D[i, r] \leftarrow$  valoarea conform ecuației (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  $i > 0$  și  $r > 0$  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 distanțe D de dimensiune $N_i \times N_r$, în care fiecare linie corespunde unui cuvânt recunoscut și fiecare coloană – unuia de referință, iar distanța finală se obține în $D[N_i, N_r]$.

Alinierea necesită refacerea șirului de evaluări care au minimizat distanțele parțiale calculate succesiv conform ecuației (3.7). Aceasta impune păstrarea evaluărilor într-o a doua matrice E de dimensiune $N_i \times N_r$, astfel încât odată calculată $D[N_i, N_r]$, șirul de evaluări care au dus la obținerea 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 secvența $e[n]$ de evaluări (corect, substituit, inserat, omis) care au dus la minimizarea distanțelor parțiale și finală. Odată această secvență disponibilă, calculul diferitelor matrice de performanță și alinierea celor două șiruri de cuvinte sunt imediate.

Costurile asociate erorilor (C_s , C_i și C_o) în ecuația (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 recunoaștere automată a vorbirii prin aliniere la nivel de cuvânt [181], [72], [260] se alege de obicei valori ale acestora

$$C_i + C_o > C_s > C_i = C_o \quad (3.8)$$

astfel încât o substituție să fie preferată unei perechi inserție+omisiune sau invers.

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

3.2 Distanțe acustice

Data fiind natura nestaționară 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 recunoașterii formelor [64], [76], [243], acești parametri pot fi grupați în vectori caracteristici, elemente ale unui spațiu multidimensional al caracteristicilor. În cazul recunoașterii vorbirii, vectorii caracteristici sunt cunoscuți drept **vectori acustici**, elemente ale unui **spațiu acustic**.

Evaluarea distanțelor dintre vectorii caracteristici este o problemă fundamentală în recunoașterea formelor, complicată în cazul distanțelor acustice de cerința ca ele să aibă pe cât posibil și o interpretare perceptuală: distanța dintre doi vectori acustici ar trebui să aibă valori mari dacă porțiunile de semnal din care provin sunt diferite din punct de vedere lingvistic, și valori mici, ideal nule, în cazul identității lor lingvistice.

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

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

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

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

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

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

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

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

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

sau, în cazul utilizării trunchierii cepstrale:

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

Aceasta înseamnă că distanța euclidiană dintre vectorii cepstrali satisface cerințele perceptuale enunțate anterior, ceea ce poate fi una din explicațiile succesului analizei cepstrale ca metodă de extragere a caracteristicilor pentru recunoașterea vorbirii.

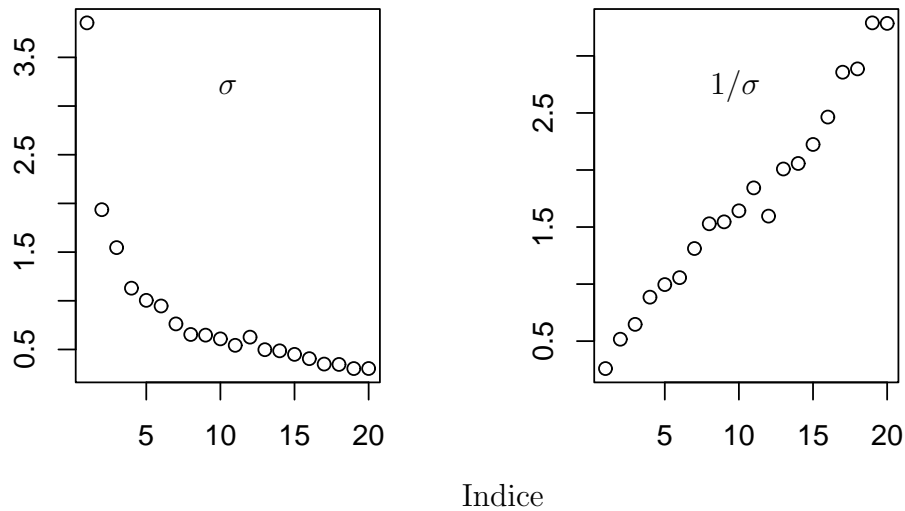


Figura 3.1: Abaterile standard și inversele lor pentru primii 20 de coeficienți cepstrali melodici, estimate din 9802 cadre (98 secunde) de semnal vocal.

3.3 Transformări ale spațiului acustic

Una dintre problemele utilizării distanțelor în recunoașterea formelor [64], [76], [243] este cauzată de diferențele dintre dispersiile componentelor vectorilor caracteristici. Ca exemplu, prezentăm în figura 3.1 valorile abaterilor standard și ale inverselor lor pentru primii 20 de coeficienți cepstrali melodici, valori estimate din 9802 cadre (98 secunde) de semnal vocal obținut prin citirea de către 20 de vorbitori uniform distribuiți pe sexe și grupe de vârstă a unei propoziții în care apar toate unitățile din tabelul 4.1.

Se observă că valorile $1/\sigma$ au o variație cvasiliniară cu indicii coeficienților, deci valorile σ sunt invers proporționale cu indicii coeficienților cepstrali. În mod corespunzător, coeficienții cepstrali vor avea ponderi invers proporționale cu indicii lor în distanțele euclidiene dintre vectorii acustici cepstrali.

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

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

unde x_i, x_j și μ_i, μ_j sunt două componente ale vectorilor, respectiv valorile lor medii. Aceasta indică posibilitatea corelației liniare a celor două componente, care poate duce la evaluări eronate ale distanțelor prin luarea repetată în calcul a unei aceleiași variații.

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

$$\mathbf{C} = \Phi \Lambda \Phi^T \quad (3.15)$$

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

$$\mathcal{E}\{\Lambda^{-1/2}\Phi^T[\mathbf{x} - \boldsymbol{\mu}_x][\mathbf{x} - \boldsymbol{\mu}_x]^T\Phi\Lambda^{-1/2}\} = \mathbf{I} \quad (3.16)$$

rezultă că transformarea căutată este

$$\mathbf{y} = \Lambda^{-1/2}\Phi^T\mathbf{x} \quad (3.17)$$

Utilizarea transformării (3.17) este costisitoare din punct de vedere computațional, iar reducerea acestui cost poate fi făcută ținând cont de particularitățile caracteristicilor folosite. În cazul coeficienților cepstrali, o asemenea particularitate este corelația lor redusă datorată transformării ortogonale prin care sunt obținuți.

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

$$w_i = 1 + \frac{L}{2} \sin \frac{i\pi}{L} \quad (3.18)$$

utilizate actualmente în majoritatea sistemelor de recunoaștere automată a vorbirii.

3.4 Metode statistice de recunoaștere a vorbirii

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

Ca și 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 variabilități a fost propusă și implementată în sistemul DRAGON [16], [15] dezvoltat la CMU în același program al ARPA. Metoda folosea un model teoretic general bazat pe funcții 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) și a căror prezentare va fi detaliată în restul capitoului.

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

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

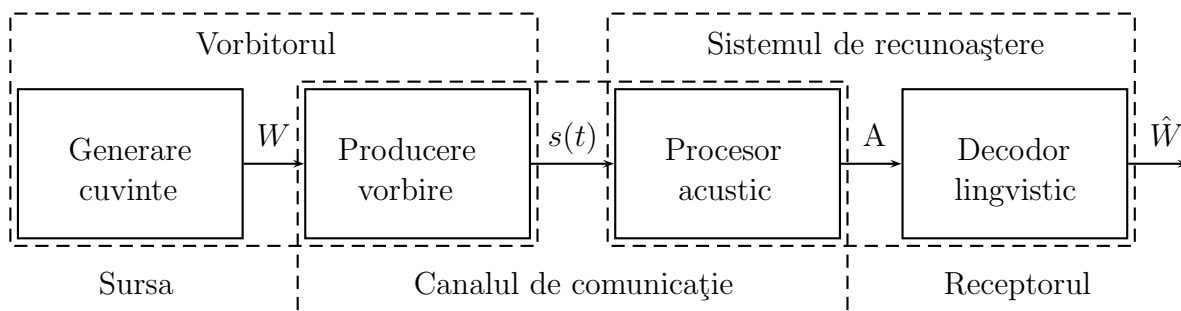


Figura 3.2: Recunoașterea vorbirii din perspectiva teoriei comunicației

având ca rezultat generarea unui șir de cuvinte W ; șirul 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 recunoaștere automată a vorbirii formează canalul de comunicație; iar receptorul este decodorul lingvistic al sistemului de recunoaștere, care generează o estimare \hat{W} a șirului de cuvinte pronunțat.

Pentru găsirea \hat{W} , decodorul lingvistic utilizează șirul A de date acustice¹ obținut de procesorul acustic din semnalul vocal $s(t)$. Dacă pentru un șir de cuvinte oarecare

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

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

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

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

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

și neglijând $P(A)$, aceeași pentru toate șirurile de cuvinte W

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

Ecuția (3.22) pune în evidență cele două probleme esențiale pentru recunoașterea 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 cunoștințe de pe nivelurile superioare (sintactic, semantic, pragmatic) din metodele bazate pe cunoștințe;

¹De-a lungul timpului au existat unele variațiuni, dar în esență A este un șir 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 obținute în urma pronunțării șirului de cuvinte W ; în cadrul metodelor bazate pe cunoștințe, aceasta era realizată prin intermediul surselor de pe nivelurile inferioare (fonetic, fonologic, lexical).

Datorită generalității menționate anterior, modelele Markov ascunse pot fi utilizate și pentru modelarea lingvistică și 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 și a evoluțiilor globale ale caracteristicilor unor procese aleatoare nestaționare, dar care pot fi considerate staționare pe porțiuni, prin **funcții de probabilitate** a valorilor acestor caracteristici. În modelele Markov ascunse, tranzițiile între stări se fac conform unor **probabilități de tranziție**, iar producerea/observarea caracteristicilor și funcțiile de probabilitate care modelează distribuțiile valorilor lor pot fi asociate stărilor sau tranzițiilor, în mod similar automatelor Moore respectiv Mealy. În continuare vom considera doar cazul funcțiilor de probabilitate asociate stărilor.

Un model Markov ascuns este definit prin:

- mulțimea stărilor $\mathcal{S} = \{s_i, i = 1 \dots N\}$;
- matricea de tranziție $\mathcal{A} = [a_{ij}]$, unde $a_{ij} = P[s(t+1) = s_j | s(t) = s_i], i, j = 1 \dots N$ sunt probabilitățile tranzițiilor între stări;
- mulțimea probabilităților inițiale ale stărilor $\Pi = \{\pi_i = P[s(0) = s_i], i = 1 \dots N\}$;
- mulțimea valorilor caracteristicilor procesului modelat \mathcal{Y} ;
- $\mathcal{B} = \{b_j(y) | b_j(y = o_t) = P[o_t | s(t) = s_j], j = 1 \dots N\}$ – mulțimea funcțiilor de probabilitate a valorilor caracteristicilor observate/produse în fiecare stare, unde prin o_t am notat valoarea observată (observația, pe scurt) la momentul t .

În această definiție, mulțimea stărilor împreună cu mulțimea probabilităților inițiale și matricea de tranziție corespund unui lanț/proces/model Markov. Stările lui nu sunt însă observabile direct, ci doar prin intermediul observațiilor generate conform funcțiilor de probabilitate asociate fiecărei stări, de unde numele de model Markov ascuns.

Modelele Markov ascunse au fost folosite inițial cu succes în recunoașterea automată a vorbirii, apoi în multe alte aplicații în care procesul studiat poate fi modelat secvențial: prelucrarea limbajului natural [152], recunoașterea scrisului de mână [48], recunoașterea feței [168], decodarea genomului uman [128] etc. Acest lucru se datorează existenței unor algoritmi eficienți pentru rezolvarea a trei **probleme fundamentale**:

- **evaluarea** probabilității $P(O|M)$ ca un șir de observații $O = o_1 o_2 \dots o_T$ să apară în urma unei realizări a procesului modelat de modelul M ;

- **decodarea** secvenței de stări $S = s(1)s(2) \dots s(T)$ parcursă prin model pentru producerea șirului de observații O ;
- **estimarea** parametrilor (probabilitățile inițiale π_i și de tranziție a_{ij} și funcțiile de probabilitate b_j) sau **antrenarea** MMA pe baza unor date corespunzătoare.

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

$$\alpha_i(t) = P[o_1 \dots o_t, s(t) = s_i | M] \quad (3.23)$$

se pot calcula recursiv:

$$\alpha_i(1) = \pi_i b_i(o_1), \quad i = 1 \dots N \quad (3.24)$$

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

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

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

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

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

dar recursia are loc în sens invers:

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

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

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

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

A doua dintre problemele enunțate, decodarea secvenței de stări parcurse printr-un model M pentru producerea unui șir de observații O , are soluții dependente de criteriul de optim utilizat. Un prim criteriu ar putea fi probabilitatea stării s_i la momentul t date fiind O și M

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

care poate fi scrisă funcție de $\alpha_i(t)$ și $\beta_i(t)$

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

astfel încât se poate face estimarea

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

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

Ultima dintre probleme, cea a antrenării MMA, are soluții dependente de tipul mulțimii \mathcal{Y} a valorilor caracteristicilor procesului modelat. În cazul în care \mathcal{Y} este o mulțime discretă, funcțiile $b_j(y)$ sunt distribuții de probabilitate, iar MMA rezultate sunt cunoscute ca **MMA discrete**. Dacă \mathcal{Y} este continuă, $b_j(y)$ sunt cel mai adesea densități parametriche de probabilitate, rezultând **MMA continue**. Alternativa este reprezentată de metodele neparametrice de estimare a probabilităților: rețele neuronale [36], [205], [161], metoda celor mai apropiați k vecini [138], mașini cu vectori suport (în engl. Support Vector Machines – SVM) [78]. Deși acestea sunt în fapt doar variante de estimare a probabilităților, sistemele rezultate sunt considerate **sisteme hibride**.

3.5.1 Antrenarea MMA discrete

Antrenarea MMA presupune estimarea valorilor parametrilor acestora: întrucât nu se cunosc soluții analitice ale problemei, toate metodele de antrenament se bazează pe optimizarea iterativă a parametrilor folosind date de antrenament corespunzătoare și diferite criterii de optim. Dintre criterii, cel mai frecvent folosit este cel al plauzibilității maxime (în engl. maximum likelihood), conform căruia estimarea de maximă plauzibilitate a parametrilor unui model M este cea care maximizează probabilitatea condiționată 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 șirurilor de observații $O = o_1 o_2 \dots o_T$, astfel încât datele de antrenament sunt incomplete: chiar dacă se cunosc valorile observațiilor, nu se știe căror stări le sunt asociate. Estimarea de maximă plauzibilitate a parametrilor din date incomplete se poate face prin algoritmul de maximizare a așteptării (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 așteptărilor (speranțelor matematice) pentru numerele de apariții ale unor evenimente, urmat de reestimarea parametrilor folosind valorile acestor așteptări.

Dacă parametrii modelului M sunt reestimați pe baza a R realizări ale procesului modelat, care au produs R secvențe de observații $O_r = o_{r,1} o_{r,2} \dots o_{r,T_r}$, $r = 1 \dots R$, calculul așteptărilor se face sumând probabilitățile corespunzătoare pentru toate secvențele și la toate momentele relevante.

Probabilitățile inițiale reestimate vor fi astfel

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

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

Reestimarea probabilităților de tranziție pe baza numerelor așteptate de tranziții între stări necesită calculul probabilităților tranzițiilor din starea s_i la momentul t în starea s_j la momentul $t + 1$ date fiind șirul de observații O și modelul M

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

Probabilitățile conjugate $P[s(t) = s_i, s(t+1) = s_j, O | M]$ pot fi calculate ca produse ale probabilităților ca modelul: 1) să se afle în starea s_i la momentul t după ce a generat prefixul $o_1 \dots o_t$; 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 $o_{t+2} \dots o_T$. Ținând cont de expresiile acestor probabilități, obținem

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

iar **probabilitățile de tranziție** reestimate vor fi

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

MMA discrete au o mulțime finită a observațiilor, $\mathcal{Y} = \{y_k, k = 1 \dots K\}$, astfel încât **funcțiile de probabilitate** $b_j(y)$ sunt distribuții de probabilitate a căror reestimare se poate face prin simpla numărare a aparițiilor valorilor y_k

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

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 convergență. Aplicarea lui este însă posibilă doar după proiectarea MMA și inițializarea parametrilor. Proiectarea MMA presupune alegerea numărului de stări N , a stărilor inițiale ($\pi_i \neq 0$) și tranzițiilor permise ($a_{ij} \neq 0$), care împreună definesc topologia unui model, iar în cazul MMA discrete trebuie stabilită și mulțimea \mathcal{Y} a valorilor caracteristicilor procesului modelat.

Deoarece algoritmul Baum-Welch realizează numai o optimizare locală a parametrilor, inițializarea lor poate fi esențială pentru performanțele obținute ulterior. Experimental, s-a constatat că probabilitățile inițiale π_i și de tranziție a_{ij} pot fi inițializate cu valori uniforme, arbitrar alese sau aleatoare fără a afecta performanțele, dar că poate fi necesară inițializarea din datele de antrenament a funcțiilor de probabilitate $b_j(y)$.

Algoritmul 3.2 Algoritmul Baum-Welch pentru antrenarea MMA discrete

```

1:  $Spi[N] \leftarrow 0$                                 {sumă ecuația (3.34)}
2:  $Nra[N, N], Nua[N] \leftarrow 0$                  {numărător/numitor ecuația (3.37)}
3:  $Nrb[N, K], Nub[N] \leftarrow 0$                  {numărător/numitor ecuația (3.38)}
4: repeat
5:   for toate secvențele  $O_r, r = 1 \dots R$  do
6:     calculează  $\alpha_{i,r}(t)$  conform ecuațiilor (3.24) și (3.25)
7:     calculează  $\beta_{i,r}(t)$  conform ecuațiilor (3.28) și (3.29)
8:     calculează  $P(O_r|M)$  conform ecuației (3.26)
9:     calculează  $\gamma_{i,r}(t)$  conform ecuației (3.32)
10:    calculează  $\xi_{ij,r}(t)$  conform ecuației (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 ecuațiilor (3.34), (3.37) și (3.38)
14: until atingerea convergenței sau a unui număr de iterații

```

3.6 Modelarea lingvistică

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

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

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

în care fiecare factor poate fi estimat pe baza unui corpus de texte. Presupunând că recunoașterea se face la nivelul propozițiilor, la o lungime a propoziției de L cuvinte și o mărime a vocabularului de V cuvinte, pentru utilizarea ecuației (3.39) ar trebui estimate

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

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

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

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

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

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

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

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

Chiar și în urma aplicării acestor metode, există încă posibilitatea ca unele evenimente (secvențe 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) probabilităților 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 și combinare.

Perplexitatea

Pentru a putea compara performanțele sistemelor de recunoaștere a vorbirii trebuie precizate (și) modelele lingvistice utilizate, iar aprecierea acestora impune cuantificarea complexității aplicației din punctul lor de vedere. Prima măsură a acestei complexități a fost așa-numitul factor de ramificare (branching factor), definit ca numărul maxim de cuvinte care pot apare după cuvântul curent. Acesta nu ținea însă cont de variațiile numărului de cuvinte care pot urma celui curent, nici de frecvențele cuvintelor, astfel încât pentru a lua în calcul aceste aspecte a fost introdusă perplexitatea.

Deși din punct de vedere teoretic poate fi pusă în legătură cu entropia sursei care generează cuvintele, în practică perplexitatea unui model lingvistic față de o aplicație se evaluează pe un corpus de texte de test, considerat reprezentativ pentru acea aplicație. Prin definiție, pentru o secvență de N cuvinte perplexitatea este

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

și 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

Așa cum am menționat deja, principala problemă în recunoașterea vorbirii este marea ei variabilitate: dacă asupra celei de la nivelul cuvintelor ne-am putut face o idee destul de exactă în secțiunea 3.6, cea de la nivelul acustic este mult mai greu de intuit și cuantificat. Și chiar dacă ea poate fi parțial redusă prin cuantizare vectorială [96], [151], [87], care înlocuiește spațiul multidimensional al vectorilor acustici cu o mulțime finită de vectori prototip, posibilitățile acestora de combinare sunt mult mai multe decât cele ale cuvintelor: de exemplu, la o frecvență de 100 vectori acustici/secundă și 256 de vectori prototip, vor exista $256^{100} \approx 10^{80}$ combinații posibile într-o singură secundă. Din acest motiv, modelarea acustică este esențială pentru orice sistem de recunoaștere a vorbirii.

Primele cercetări asupra recunoașterii vorbirii folosind MMA au apelat frecvent la cuantizare vectorială și MMA discrete pentru a reduce cerințele computaționale. Au fost obținute astfel rezultate care au mers de la recunoașterea independentă de vorbitor a unui număr mic de cuvinte pronunțate izolat [199] până la recunoașterea dependentă de vorbitor a unui număr mare (5000) [114] sau foarte mare (20000) [10] de cuvinte izolate, și au culminat la sfârșitul anilor '80 cu recunoașterea 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 densităților de probabilitate utilizate, așa încât potențialul avantaj nu se putea realiza datorită neoptimizării acestor parametri. Abia după stabilirea relațiilor de reestimare prin algoritmul Baum-Welch a parametrilor unei largi clase de densități 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 densitățile parametrice de probabilitate, cele mai performante s-au dovedit așa-numitele mixturi gaussiene, cunoscute și ca modele cu mixturi gaussiene (în engl. Gaussian Mixture Models – GMM), de forma

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

cu $\mathcal{N}_{\boldsymbol{\mu}_k, \mathbf{C}_k}$ densități normale multivariate de medii $\boldsymbol{\mu}_k$ și matrice de covarianță \mathbf{C}_k

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

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

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

normale simple, iar timpul de calcul se reduce în mod corespunzător la $O(D)$. În plus, scade în aceeași proporție și numărul parametrilor MMA care trebuie estimați.

Succesul mixturilor gaussiene are explicații multiple. Pe de o parte, alofonele unui fonem sunt pronunțate folosind configurații apropiate ale tractului vocal: cum vectorii acustici rezultați depind de aceste configurații, intuitiv ne așteptăm, și studii experimentale o confirmă [263], ca pentru un același vorbitor sau vorbitori de același sex ei să fie apropiați în spațiul acustic. Mai mult, ținând cont de numărul mare de factori de care vectorii acustici depind, ne putem aștepta, și același studii o confirmă, ca densitățile gaussiene să fie adecvate pentru modelarea distribuției lor.

Pe de altă parte, densitățile gaussiene au o mulțime de proprietăți 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 densități parametrice, de exemplu mixturi laplaciene [169], cât și metode neparametrice de estimare a densităților de probabilitate: rețele neuronale [36], [205], [161], metoda celor mai apropiați k vecini [138] sau mașini cu vectori suport (în engl. Support Vector Machines – SVM) [78]. Deși inițial promițătoare, metodele neparametrice fie s-au dovedit dificil de scalat la vocabulare foarte mari sau de adaptat la schimbările de mediu sau ale vorbitorilor (cazul rețelelor neuronale), fie au încă de trecut aceste teste, așa că pentru moment mixturile gaussiene rămân cea mai bună opțiune în modelarea acustică.

3.7.1 Antrenarea MMA cu mixturi gaussiene

Înlocuirea distribuțiilor de probabilitate din MMA discrete cu densități parametrice în MMA continue nu afectează probabilitățile inițiale și de tranziție, așa încât formulele de reestimare a lor pentru MMA discrete (secțiunea 3.5.1) rămân valabile și în acest caz. Modificările constau în introducerea unor noi formule pentru reestimarea parametrilor mixturilor: ponderile densităților w_k , vectorii medii $\boldsymbol{\mu}_k$ și matricele de covarianță \mathbf{C}_k .

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

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

Utilizând R secvențe de observații $\mathbf{O}_r = \mathbf{o}_{r,1} \mathbf{o}_{r,2} \dots \mathbf{o}_{r,T_r}$, $r = 1 \dots R$, pentru a antrena modelul M , formulele de reestimare a valorilor parametrilor mixturilor sunt

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

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

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

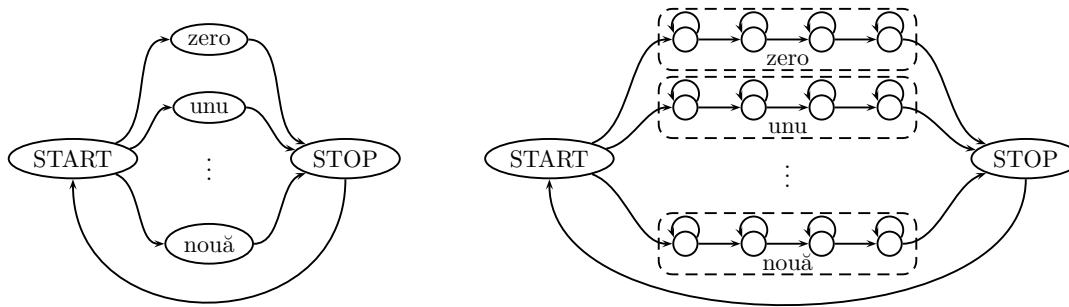


Figura 3.3: Exemplu de integrare a modelului lingvistic cu cele acustice

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 probabilităților densităților componente ale fiecărei mixturi, $\gamma_{jk,r}(t)$; reestimarea parametrilor mixturilor conform formulelor (3.47–3.49).

3.8 Reprezentarea integrată a cunoștințelor

Căutarea șirului de cuvinte prin pronunțarea căruia este cel mai probabil să fi rezultat un șir de observații acustice necesită integrarea cunoștințelor încorporate de modelele lingvistice și acustice într-un spațiu de căutare unic, în care se poate evalua produsul $P(W) \cdot P(A|W)$ din ecuația (3.22). Modelele Markov ascunse oferă un cadru pentru realizarea acestei integrări, permițând construcția ierarhică a unui MMA global care modelează atât generarea cuvintelor cât și producerea vorbirii. Pentru exemplificare considerăm o aplicație simplă constând în recunoașterea cifrelor zecimale (figura 3.3).

Pentru recunoașterea 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 rețea de recunoaștere sau rețea integrată, care include acum atât modelul lingvistic cât și modelele acustice.

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

Pentru vocabulare mari și foarte mari, modelarea acustică utilizează MMA ale unor unități sublexicale (silabe, semisilabe, foneme etc.), iar integrarea modelului lingvistic cu cele acustice presupune utilizarea unui model lexical, uzual un simplu dicționar, pentru obținerea pronunțiilor cuvintelor, eventual cu variante, în termenii unităților sublexicale. Pronunțiile la rândul lor sunt reprezentabile prin modele Markov ale căror stări pot fi

înlocuite cu modelele acustice ale unităților sublexicale componente.

Simpla înlocuire a pronunțiilor din dicționar în nodurile modelului lingvistic duce la o reprezentare costisitoare din punct de vedere computațional, în special în cazul vocabulelor mari și foarte mari, astfel încât pentru creșterea vitezei și reducerea necesarului de memorie dicționarul poate fi reprezentat printr-o structură arborescentă [12], [169], eventual comprimată [131]. O asemenea structură arborescentă poate fi folosită și pentru factorizarea probabilităților cuvintelor din modelul lingvistic statistic la nivelul unităților sublexicale [231], [5], [178]. În cazul unui model lingvistic de tip n -gram, acesta poate fi în întregime integrat cu dicționarul și transformat într-o mulțime de subarbori [70].

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

3.9 Algoritmi de căutare

Odată cu integrarea cunoștințelor acumulate de modelele lingvistice și acustice într-o rețea de recunoaștere, găsirea șirului de cuvinte cel mai probabil să fi fost pronunțat, date fiind observațiile acustice, se poate face prin căutarea în rețea a căii pentru care produsul $P(W) \cdot P(A|W)$ este maxim. În acest produs, probabilitățile lingvistice pot fi aplicate la tranzițiile între cuvinte, sau pe durata cuvintelor, dacă dicționarul a fost reprezentat sub formă de arbore și probabilitățile lingvistice au fost factorizate [231]. Datorită valorilor foarte mici ale probabilităților, înmulțirile lor repetate pot ajunge rapid să cauzeze depășiri inferioare (underflow), astfel încât pentru evitarea acestei probleme se utilizează de obicei un sistem logaritm de reprezentare [39], [41].

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

Altă problemă a sistemelor de recunoaștere automată a vorbirii este cea a echilibrului între omisiuni și inserții, pentru asigurarea căruia se utilizează o penalizare de tranziție τ , deasemeni determinată experimental, aplicată la tranzițiile între cuvinte pentru a reduce frecvența inserțiilor. Folosind o reprezentare logaritmă a probabilităților și cele două variabile de control λ și τ , ecuația (3.22) trebuie rescrisă în forma

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

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

Algoritmii fundamentali de căutare a căii corespunzătoare celui mai probabil șir de cuvinte sunt algoritmul Viterbi [247] și algoritmul A^* [210], mai cunoscut în domeniul recunoașterii 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 același moment în baza unei strategii de căutare în lățime (breadth first), iar scorurile căilor pot fi comparate deoarece toate corespund aceleiași porțiuni a semnalului; ca urmare, algoritmul Viterbi este admisibil, adică garantează găsirea căii cu scorul maxim – care poate să coincidă sau nu cu șirul de cuvinte corect.

Pentru șirul de observații $O = o_1 o_2 \dots o_T$, algoritmul Viterbi găsește calea (secvența de stări) $Q = q_1 q_2 \dots q_T$ optimă prin MMA integrat M (rețeaua de recunoaștere) prin calculul pentru fiecare moment t și fiecare stare s_i a probabilității maxime ca modelul să fi generat observațiile $o_1 \dots o_t$ și să se afle în starea s_i la momentul t

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

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

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

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

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

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

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

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

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

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

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

Deși admisibil, algoritmul Viterbi se poate dovedi prohibitiv din punct de vedere al timpului de calcul necesar căutării soluției optime. Pentru creșterea vitezei de obținere a unei soluții, el a fost modificat prin reducerea (pruning) spațiului de căutare doar la un fascicol (beam) de căi printre care este cel mai probabil să fie localizată și 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) față de starea cea mai probabilă la acel moment, a unui număr maxim permis de stări componente [231] sau a unor combinații ale acestor criterii.

Algoritmul A^* sau de decodare cu stivă este un algoritm asincron, de tip *best first*, cu o parcurgere arborescentă a spațiului de căutare, diferită de cea de tip grilaj (trellis) din cazul algoritmului Viterbi, ceea ce conduce la căi parțiale de lungimi diferite, între ale căror scoruri comparațiile directe nu au sens. Comparațiile 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 funcție de evaluare

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

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

Caracterul euristic al funcției h face ca algoritmul A^* să nu fie întotdeauna admisibil. Împreună cu dificultățile estimării propriu-zise, acest lucru a făcut ca algoritmul A^* să fie mai puțin utilizat în sistemele de recunoaștere 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 aștepta la o creștere a frecvenței utilizării lui.

În forma de bază, atât algoritmul Viterbi cât și A^* asigură găsirea unei căi optime unice, dar în practică, datorită erorilor de modelare sau căutare, aceasta poate fi diferită de cea corectă, corespunzătoare șirului de cuvinte pronunțat. Unele dintre erorile de acest tip pot fi corectate de aplicația în care este integrat sistemul de recunoaștere dacă acesta generează nu doar șirul 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 propoziții, și algoritmi în mai mulți pași [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 spații de căutare reduse pentru pași succesivi, cât și a rezultatelor finale.

3.10 Concluzii

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

Utilizarea parametrilor semnalului vocal pentru recunoașterea automată a vorbirii presupune gruparea lor în vectori dintr-un spațiu acustic. Recunoașterea ca atare poate fi precedată de o serie de operațiuni în acest spațiu: calculul unor distanțe acustice semnificative din punct de vedere perceptual; transformări ale spațiului acustic în scopul reducerii anizotropiei lui; cuantizarea vectorială care, utilizând o mulțime de vectori prototip, permite comprimarea și reprezentarea discretă a vectorilor acustici.

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

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

Găsirea șirului de cuvinte cel mai probabil să fi fost pronunțat dat fiind semnalul de recunoscut necesită construcția unei reprezentări integrate a modelelor lingvistice și acustice și evaluarea posibilelor soluții prin utilizarea unor algoritmi de căutare în spațiul soluțiilor. Modelele Markov ascunse permit rezolvarea elegantă a ambelor probleme prin construcția ierarhică a unor modele înglobându-le pe cele lingvistice și acustice, cunoscute sub numele de rețele de recunoaștere sau rețele integrate, în care șirul cel mai probabil de cuvinte este găsit cel mai adesea folosind algoritmul Viterbi.

CAPITOLUL 4

Baza de date fonetice

Utilizarea metodelor statistice de recunoaștere automată a vorbirii, prezentate în capitolul 3, impune existența unor date pe baza cărora să poată fi construite modelele implicate de ecuația (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 aplicații [15], [146], și arhive de texte private [113].

Dezvoltarea cercetărilor asupra recunoașterii automate a vorbirii, utilizând multiple abordări ale problemelor ei, a făcut necesară evaluarea nu doar a performanțelor în sine ale sistemelor de recunoaștere, ci și a semnificației statistice [37], [158] a diferențelor dintre aceste performanțe. Drept urmare, în anii '80 a început colectarea și publicarea unor baze de date vocale de uz general [45] sau specifice unor aplicații. Acestea s-au dovedit esențiale pentru dezvoltarea și testarea unor sisteme de recunoaștere, precum și pentru evaluarea și compararea acestor sisteme utilizând seturi de date standard.

Probabil cele mai cunoscute exemple de baze de date specifice unor aplicații sunt TIDIGITS [140], colectată la Texas Instruments (TI) pentru studiul recunoașterii șirurilor de cifre, și Resource Management (RM) [192], dezvoltată în cadrul programului DARPA de recunoaștere automată a vorbirii și având în vedere o aplicație de conducere a resurselor militare navale (nave de război) ale Statelor Unite. Disponibilitatea acestora din urmă a fost esențială în demonstrarea fezabilității recunoașterii independente de vorbitor a vorbirii continue cu vocabulare mari [135] și, în ultimă instanță, impunerea metodelor statistice de recunoaștere automată a vorbirii în raport cu alte abordări.

Succesul metodelor statistice, bazate pe estimarea automată din date de antrenament a parametrilor sistemelor de recunoaștere, a pus în evidență importanța fundamentală a datelor și a condus la construirea unor noi baze de date vocale, de dimensiuni din ce în ce mai mari, vizând aplicații diverse și încercând să acopere porțiuni cât mai extinse din variabilitatea semnalului vocal. Cele mai importante, prin contribuția 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 și al fostului Consiliu Național al Cercetării Științifice Universitare – CNCSU (devenit din 1999 Consiliul Național al Cercetării Științifice din Învățământul Superior – CNCSIS) prin grantul 56/1995.

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

O categorie aparte a bazelor de date vocale, destul de restrânsă datorită dificultăților implicate de construcția lor, este constituită de așa-numitele baze de date fonetice, prima și cea mai cunoscută și utilizată dintre ele fiind TIMIT [81], construită în cooperare de Texas Instruments (TI) [71] și Massachusetts Institute of Technology (MIT) [132], cu unele contribuții de la Stanford Research Institute [51]. Elementele defnitorii ale unei asemenea baze de date sunt conținutul controlat prin proiectarea corespunzătoare a materialelor înregistrate și selectarea vorbitorilor, o calitate deosebită a înregistrărilor, precum și adnotarea conținutului cu informații fonetice și fonologice. Datorită acestor caracteristici, o bază de date fonetice constituie o resursă esențială nu numai pentru cercetările în direcția recunoașterii automate a vorbirii, ci și pentru cele din alte domenii ale prelucrării automate a vorbirii, precum și o sursă de cunoștințe fundamentale, de fonetică și fonologie a limbii în care au fost pronunțate materialele înregistrate.

Date fiind inexistența unei baze de date corespunzătoare pentru cercetările asupra recunoașterii automate independentă de vorbitor a vorbirii continue în limba română și insuficiența constatată a cunoștințelor de fonetică acustică și fonologie a limbii române (insuficiență confirmată ulterior chiar și de lingviști în literatura de specialitate [234]), construcția 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 și colectarea ei [32], [34].

4.1 Considerații de proiectare

Disponibilitatea unor baze de date vocale corespunzătoare reprezintă o precondiție pentru multe cercetări fundamentale sau aplicative din diverse domenii ale științei și 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 cerințele. Uneori, aceste cerințe pot fi satisfăcute prin utilizarea unor baze de date deja existente sau a unor subseturi convenabil alese ale acestora.

Există însă și numeroase situații în care se impune colectarea unor noi baze de date, una dintre acestea fiind și extinderea la noi limbi sau dialecte a cercetărilor asupra recunoașterii automate a vorbirii: deși au fost încercate diferite metode de utilizare a datelor dintr-o limbă pentru recunoașterea pronunțiilor unei alte limbi [253], [162], unele vizând limba română [155], recunoașterea multi- și croslinguală [215] sau independentă de limbă [44], performanțele obținute se îmbunătățesc odată cu creșterea cantității datelor din noua limbă folosite pentru construirea sau adaptarea modelelor acustice.

Proiectarea și colectarea unei noi baze de date devine cu atât mai necesară atunci când cunoștințele fundamentale de fonetica și fonologia noii limbi, necesare și pentru recunoașterea automată, sunt insuficiente: din păcate, aceasta este și situația 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 influențate, în consecință, de necesitatea de a facilita:

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

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

- validarea sau obținerea unor noi cunoștințe de fonetică și fonologie a limbii române, a căror insuficiență a fost resimțită încă din această fază a cercetărilor și confirmată în parte de rezultate prezentate în capitolele următoare.

Din marea varietate de tipuri de pronunțare, condiții de mediu etc. posibile, această bază de date a fost limitată la înregistrări de calitate, într-un mediu afectat cât mai puțin de zgomote și reverberații, ale unor pronunții rezultate cu preponderență din citirea în varianta standard (literară) a limbii române a unor texte pregătite în mod special pentru a asigura un conținut al înregistrărilor corespunzător scopurilor propuse.

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

4.2 Alegerea unităților de modelare

Recunoașterea automată a vorbirii se poate realiza, după cum s-a menționat, folosind modele acustice ale cuvintelor sau ale unor unități sublexicale – silabe, semisilabe, sunete etc. Alegerea unităților de modelare are o influență decisivă asupra acurateții cu care modelele acustice reprezintă variabilitatea semnalului vocal, în special cea cauzată de coarticulație, constând în modificarea caracteristicilor sunetelor vorbirii funcție de cele adiacente și datorată mișcărilor anticipatorii și inerțiale ale articulatorilor. Exemple de manifestări ale coarticulației pot fi urmărite în figura 5.1: coarticulația este cea mai vizibilă în cazul sunetului [l], cele trei apariții ale lui având fiecare caracteristici spectrale diferite, dar poate fi observată și în cazul altor sunete – de exemplu [j] sau [a].

Utilizând cuvinte ca unități de modelare acustică, o bună parte din variabilitatea datorată coarticulației – cea corespunzătoare interacțiunii dintre sunete în interiorul cuvintelor – va fi inclusă în modelele rezultate, dar va rămâne totuși neacoperită cea cauzată de interacțiunile dintre cuvinte, localizată la extremitățile lor. Aceasta ar putea fi la rândul ei reprezentată prin modele dependente de context ale cuvintelor, dar odată cu creșterea mărimii vocabularului va crește și numărul modelelor. În plus, va crește proporțional și cantitatea de date necesare pentru antrenarea lor.

Un alt dezavantaj, poate cel mai important, al cuvintelor ca unități de modelare acustică, este acela al lipsei lor de generalitate: extinderea vocabularului unui sistem de recunoaștere 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 aplicații simple, cu vocabulare închise de zeci sau maximum sute de cuvinte.

Dintre unitățile de modelare sublexicale, silabele și unitățile derivate sunt adesea considerate unități naturale din punct de vedere al coarticulației și al posibilităților de a reprezenta variabilitatea asociată ei. În practică, ele se dovedesc însă dificil de utilizat datorită problemelor care pot apare la extinderea vocabularului, precum și a celor de

Tabelul 4.1: Unități fonetice de modelare acustică – simboluri și exemple

Simboluri		Exemple	Simboluri		Exemple	Simboluri		Exemple
IPA	ASCII		IPA	ASCII		IPA	ASCII	
i	i	vîn	ɔ	O	coate	k	k	cap
î	I	ani	p	p	păr	g	g	gât
e	e	el	b	b	ban	ts	T	țap
î	y	în	t	t	tip	tʃ	C	ce
ə	@	că	d	d	dar	dʒ	G	ger
a	a	an	f	f	foc	h	h	han
u	u	ud	v	v	vin	m	m	mic
o	o	om	s	s	stop	n	n	nas
j	j	iar	z	z	zi	l	l	lac
ɛ	E	deal	ʃ	S	și	r	r	râu
w	w	nou	ʒ	J	jok	#	-	pauză

acoperire a pronunțiilor de recunoscut, care fac necesară combinarea lor cu modele de tip fonemic. În plus, unele experimente [35] au arătat că performanțele astfel obținute 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 variabilității din vorbirea spontană, iar experimente de recunoaștere [79] au confirmat unele avantaje ale silabelor ca unități de modelare acustică. Dar chiar când sistemele evaluate au utilizat modele ale silabelor, ele au inclus și modele dependente de context de tip fonemic pentru a putea acoperi toate pronunțiile de recunoscut.

Pentru a obține modele acustice cât mai generale folosind o cantitate minimă de date pentru antrenarea lor, recunoașterea automată a vorbirii continue cu vocabulare mari și foarte mari este în general bazată pe unități de modelare de tip fonetic sau, prin legarea parametrilor la nivelul stărilor, chiar subfonetic [108]. În acest caz, variabilitatea datorată coarticulației este acoperită prin modelarea dependentă de context, o aceeași unitate putând fi reprezentată prin mai multe modele, diferențiate funcție de unitățile care o preced și/sau urmează. Se obțin astfel modele cu un grad mare de generalitate, ușor de reutilizat în cazul modificării sau extinderii vocabularului și care pot valorifica cu eficiență maximă cantitatea de date disponibile pentru antrenarea lor.

Ținând cont de aspectele menționate, proiectarea bazei de date s-a făcut având în vedere modelarea acustică prin unități sublexicale de tip fonetic. Din păcate, cunoștințele disponibile de fonetică acustică și fonologie a limbii române, pe baza cărora ar fi trebuit definit un set de asemenea unități, s-au dovedit insuficiente, existând diverse poziții ale lingviștilor în ceea ce privește fonemele limbii române [193], [244], uneori contradictorii [190] sau chiar greșite [206]. Analizând seturile de foneme identificate sau acceptate de diverși autori și luând în considerare diferențele dintre ele, am definit un set de unități care, direct sau prin combinații, acoperă aceste variante (tabelul 4.1).

Tabelul 4.1 cuprinde atât simboluri din alfabetul fonetic internațional (International Phonetic Alphabet – IPA), pentru a facilita comparațiile cu literatura lingvistică, cât și

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

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

- /I/ apare doar în poziție postconsonantică finală, în cuvinte ca *ani*, *azi*, *mari* etc.: unele sisteme fonologice ale limbii române standard îl consideră un alofon – o variantă pozițională, scurtă, asilabică și (uneori) devocalizată, a lui /i/, /j/ sau /e/; altele neagă chiar existența 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/ și /O/ sunt uneori interpretate ca alofone fie ale vocalelor /e/ respectiv /o/, fie ale semivocalelor/semiconsoanelor /j/ respectiv /w/; în mod corespunzător, va diferi mulțimea acceptată a diftongilor și triftongilor – unii dintre ei subiect, la rândul lor, al unor discuții asupra naturii lor mono- sau multifonemice.

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

Unitățile 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 recunoașterea automată a vorbirii continue în limba română, înregistrările din baza de date trebuie să conțină un număr suficient de mare de apariții ale unor unități de modelare acustică pentru a permite estimarea cu acuratețe a modelelor lor. În plus, pentru a asigura și independența de vorbitor a modelelor, aceste apariții trebuie să provină din pronunții ale cât mai multor vorbitori.

Pe lângă modelarea acustică, la construcția unei baze de date pentru cercetări asupra recunoașterii automate a vorbirii trebuie avută în vedere și 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 și pentru construcția de modele lingvistice de tip n -gram. Adesea, un asemenea corpus este obținut, datorită caracteristicilor textelor conținute, din arhive în format electronic ale unor ziare. Această metodă a fost folosită, de exemplu, pentru construcția 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 obținerii și prelucrării unui corpus de tipul menționat, precum și unele aspecte (pre-normalizare și standarde) vizate de programul COPERNICUS al Comisiei Europene, în cadrul căruia a început construcția ei [203], [204], pentru această bază de date s-a ales varianta compatibilității cu baza de date EUROM [46], dezvoltată anterior în proiectele ESPRIT SAM și SAM-A.

Compatibilitatea se referă atât la materialele conținute și vorbitorii înregistrați, cât și la utilizarea aceluiași organizări și formate ale fișierelor constituite [90].

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

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

Pentru fiecare limbă, (circa) 60 de vorbitori, (cât mai) egal distribuiți pe sexe, au înregistrat 3-5 pasaje, 0-5 propoziții de completare și numerele, iar câțiva – logatomi.

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

În jurul acestui nucleu au mai fost incluse:

- patru propoziții fonemice compacte, citite de către toți vorbitorii și utile pentru inițializarea modelelor acustice;
- propoziții individuale, specifice fiecărui vorbitor, selectate dintr-un corpus printr-un algoritm de tip greedy având ca obiectiv maximizarea numărului de difoni așteptați să apară din citirea pasajelor și a propozițiilor de completare și individuale;
- materiale care să permită dezvoltarea unor aplicații simple și studiul diferențelor între stilul citit și cel semispontan de vorbire: alfabetul limbii române (citat) și unele informații furnizate semispontan de vorbitori (numele, pe cuvinte și litere; seria și numărul buletinului de identitate; numărul de telefon; data nașterii; adresa).

Prezentăm în continuare unele detalii legate de materialele principale, utilizate pentru antrenarea și evaluarea sistemelor de recunoaștere a vorbirii – pasaje și propoziții.

4.3.1 Pasaje

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

Algoritmul 4.1 Algoritmul de grupare a pasajelor

```

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

```

ale fonemelor, traducerea și adaptarea au inclus modificări, în special ale unor toponime, pentru a crește frecvențele de apariție ale fonemelor rare.

Pentru a obține o structură cât mai ordonată a înregistrărilor și o distribuție cât mai echilibrată a fonemelor în pronunțiile vorbitorilor, cele 40 de pasaje au fost grupate în 10 clustere de câte 4 pasaje folosind un algoritm special conceput (algoritmul 4.1) care urmărește gruparea a P pasaje în K clustere de câte P/K pasaje în care fonemele să fie, în limita posibilităților, cât mai uniform distribuite.

În cursul producerii vorbirii, fonemele sunt realizate cu frecvențe diferite (în pasaje, de exemplu, numerele totale de apariții ale fonemelor, $N_p[f]$, variază între 15 și 1104), iar construcția unor modele acustice de o calitate acceptabilă impune ca fiecare fonem să aibă cel puțin un anumit număr minim de realizări. Din această cauză, fonemele rare trebuie urmărite cu prioritate maximă, iar numerele minime de apariții ale fonemelor în clustere (liniile 12, 14 și 18) sunt asigurate în ordinea crescătoare a numerelor totale de apariții ale fonemelor (linia 8): pentru a obține și o distribuție pe clustere cât mai uniformă, numărul de apariții ale fonemului f în clusterul c , $N_c[f]$, poate fi cu cel mult $\Delta[f]$ mai mic decât numărul său mediu de apariții 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 $\Delta[f]$, dependentă de fonem pentru a ține cont de inegalitatea frecvențelor de apariție, a fost calculată în implementarea practică a algoritmului cu formula

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

gruparea efectivă a pasajelor fiind realizată prin utilizarea constantei $a = 73$. În măsura în care acest lucru a fost posibil, aceeași abatere maximă $\Delta[f]$ a fost utilizată și pentru a limita depășirea numărului mediu de apariții ale unui fonem într-un cluster.

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

4.3.2 Propozițiile

Pe lângă necesitățile modelării acustice și compatibilitatea cu EUROM, pregătirea materialelor de înregistrat a luat în calcul și aspecte legate de etichetarea bazei de date. Astfel, având în vedere automatizarea etichetării folosind modele Markov ascunse, au fost create patru propoziții fonemic compacte citite de către toți vorbitorii înregistrați, propoziții așa-zise de inițializare: din citirea fiecăreia se așteaptă să se obțină minimum o realizare a fiecărui fonem, iar prin etichetarea manuală a înregistrărilor (secțiunea 5.3.1) – materialul pentru inițializarea unor modele acustice ale fonemelor.

O a doua categorie de propoziții este a celor de completare: în urma grupării pasajelor, câteva foneme erau încă slab reprezentate în unele clustere – de exemplu, fonemele /G/ și /h/ apăreau doar de câte 15 ori în total, și în multe clustere doar o singură dată. Din acest motiv au fost create, folosind același editor de transcrieri fonemice utilizat și pentru pasaje, 26 propoziții de completare care să asigure pentru toate fonemele un număr minim de apariții 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 tranșe de înregistrări ale celor 10 clustere, fiecare cluster a fost extins cu 2 sau 3 propoziții de completare, care să asigure oricărui fonem minimum 7 apariții/cluster.

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

Tabelul 4.3: Distribuția 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 și peste 50 de ani.

Cluster	Vorbitori									
0	MA_1	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	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

Existența celor 10 clustere extinse a făcut posibilă planificarea înregistrărilor în grupe de 20 de vorbitori egal distribuiți pe sexe, iar compatibilitatea cu EUROM [46] a impus înregistrarea a minimum 60 de vorbitori. Pentru a obține însă cât mai multe date pentru antrenarea, testarea și evaluarea sistemelor de recunoaștere, am anticipat o extindere până la 100 de vorbitori, cu o posibilă fază intermediară la 80 de vorbitori. Ansamblul vorbitorilor constituie, în terminologia EUROM, așa-numita mulțime ”Many Talker”, prescurtat MT, toți vorbitorii trebuind să înregistreze într-o singură sesiune:

- materiale care asigură compatibilitatea cu EUROM: câte un cluster extins (4 pasaje și 2 sau 3 propoziții de completare asociate) și cele 26 numere între 0 și 9999;
- materiale adiționale: cele 4 propoziții de inițializare, 3...7 propoziții individuale, alfabetul și informațiile personale.

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

În tabelul 4.3, primele șase coloane de vorbitori asigură compatibilitatea cu EUROM, iar următoarele două – faza intermediară de 80 vorbitori. Zece vorbitori, unul din fiecare sex și grupă de vârstă (*evidențiați* în tabel), constituie mulțimea ”Few Talker”, prescurtat FT. Ei au fost planificați să înregistreze în plus logatomii de tip CVC, iar în patru sesiuni suplimentare, decalate la câte cel puțin două săptămâni – patru noi clustere extinse și numerele. Doi vorbitori din mulțimea FT (*subliniați* în tabel), unul din fiecare sex,

Tabelul 4.4: Proprietăți ale clusterelor de pasaje extinse cu propoziții de completare: entropiile fonemelor și numerele de cuvinte distincte și 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 mulțimea ”Very Few Talker”, prescurtat VT, ei înregistrând în prima sesiune și logatomii în cinci contexte de câte două cuvinte, precum și 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 și a numerelor de cuvinte distincte și totale, numere în limitele a 8% respectiv 12% față de medie (sub două abateri standard). Constatarea este susținută și de valorile entropiei relative (divergență informațională [129] sau Kullback-Leibler) dintre distribuțiile lor fonemice, valori cuprinse între 0,02 și 0,05 biți, cu o medie de 0,032 biți și o abatere standard de 0,0075 biți.

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

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

Tabelul 4.5: Numerele de apariții ale fonemelor în transcrierile pasajelor și propozițiilor de completare, respectiv așteptate în pronunțiile pasajelor și propozițiilor de completare și 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: propozițiile de inițializare, folosite doar pentru inițializarea modelelor; numerele și logatomii CVC, utile doar pentru testare, evaluare și diagnoză; alfabetul și informațiile personale, cu pronunții greu sau imposibil de anticipat (în plus, pronunțiile 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 și la nivelul organizării și formatelor fișierelor componente prin înregistrarea ei folosind același pachet software utilizat și pentru EUROM, denumit EUROPEC [262], [90]. Acesta utilizează, gestionează și generează diverse fișiere, majoritatea de tip text:

- un fișier de descriere a vorbitorilor, completat la momentul înregistrării cu date de identificare și caracteristici ale acestora care pot avea legătură cu modul în care vorbesc (înălțimea, greutatea, limba maternă, educația etc.);
- un fișier de descriere a materialelor, precizând pentru fiecare codul de identificare și tipul (tabelul 4.6), protocolul folosit pentru înregistrare etc.;
- fișiere corpus, incluzând materialele pentru înregistrări: în cazul celor cu conținut predeterminat (pasaje, propoziții, numere, logatomi, alfabet), textul acestora, iar în cazul celor semispontane, un text generic indicând informația solicitată – nume, adresă etc.; textele au fost formate conform cerințelor EUROPEC înainte de realizarea înregistrărilor, iar pe durata lor au fost afișate pentru citire și stocate, împreună cu alte informații, în fișierele de adnotări ortografice ale semnalelor;
- protocoale de înregistrare, referite în fișierul de descriere a materialelor și specifice diferitor tipuri de materiale: acestea sunt interpretate de EUROPEC pe durata înregistrărilor, asigurând secvența dorită de operațiuni;
- fișiere de semnal, singurele de tip binar, obținute prin înregistrarea vorbitorilor în timpul citirii textelor sau al pronunțării informațiilor solicitate;
- fișiere de adnotări ortografice corespunzătoare celor de semnal, generate odată cu ele și cuprinzând fiecare: numele fișierului de semnal asociat și ale fișierelor corpus și de descriere a protocolului folosite la înregistrarea lui; date despre vorbitorul înregistrat (sex, vârstă, limbă maternă); caracteristicile înregistrării (frecvența de eșantionare, numărul și caracteristicile eșantioanelor) și localizarea în ea a porțiunilor corespunzătoare unor propoziții/părți din textele citite etc.;
- alte fișiere cu descrieri ale configurațiilor de lucru, condițiilor de înregistrare etc.

Pentru fiecare înregistrare, EUROPEC generează un fișier de semnal, unul de adnotări ortografice și unul conținând configurația folosită. Pentru identificarea vorbitorului și a materialului, numele acestor fișiere 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 fișierului. Extensia `ext` este CFG pentru fișierele

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

Materiale	Coduri de identificare	Tip
Pasaje	[O-R][0-9]	P
Propoziții de completare	F[0-9]	S
Propoziții individuale	[AB][A-Z], C[A-H], [DE][A-T]	D
Propoziții de inițializare	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 și număr B.I.	SI	I
Data naștere	SB	B
Telefon	ST	T

de configurație și de forma TRS, TRO pentru fișierele de Semnal respectiv adnotări Ortografice, unde T este o literă indicând tipul materialului (tabelul 4.6).

4.7 Realizarea înregistrărilor

Utilizarea EUROPEC pentru realizarea înregistrărilor a necesitat: configurarea unei stații de lucru SESAM [46] pentru rularea lui prin instalarea unei plăci de achiziție și prelucrare a semnalelor de tip OROS AU21 [177] într-un calculator compatibil PC; modificarea unor fonturi pentru caracterele diacritice românești; traducerea și adaptarea mesajelor EUROPEC în limba română; formatarea materialelor de înregistrat în fișiere corpus conforme cu cerințele EUROPEC; scrierea, testarea și depanarea protocoalelor de înregistrare specifice diferitelor tipuri de materiale. Fișierele astfel rezultate au fost organizate într-o structură unitară (secțiunea 4.6) utilizând fișiere de configurare și de descriere a materialelor și condițiilor de înregistrare. În sfârșit, au fost dezvoltate fișiere de comenzi pentru simplificarea menținerii acestei structuri și a operării EUROPEC.

Pentru a obține o calitate cât mai bună a înregistrărilor prin minimizarea zgomotului și distorsiunilor, acestea au fost realizate într-o cameră izolată și tratată fonic, în care vorbitorii citeau textele alocate sau pronunțau informațiile cerute sub supravegherea unor operatori plasați, împreună cu echipamentele, într-o cameră alăturată (figura 4.1). Inițial se intenționa ca vorbitorii să citească instrucțiunile și textele de pe ecranul unui monitor, așa cum erau ele afișate 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 și tratată fonic a fost înlocuit cu un interfon și listinguri ale textelor de citit. Pe lângă eliminarea zgomotului monitorului,

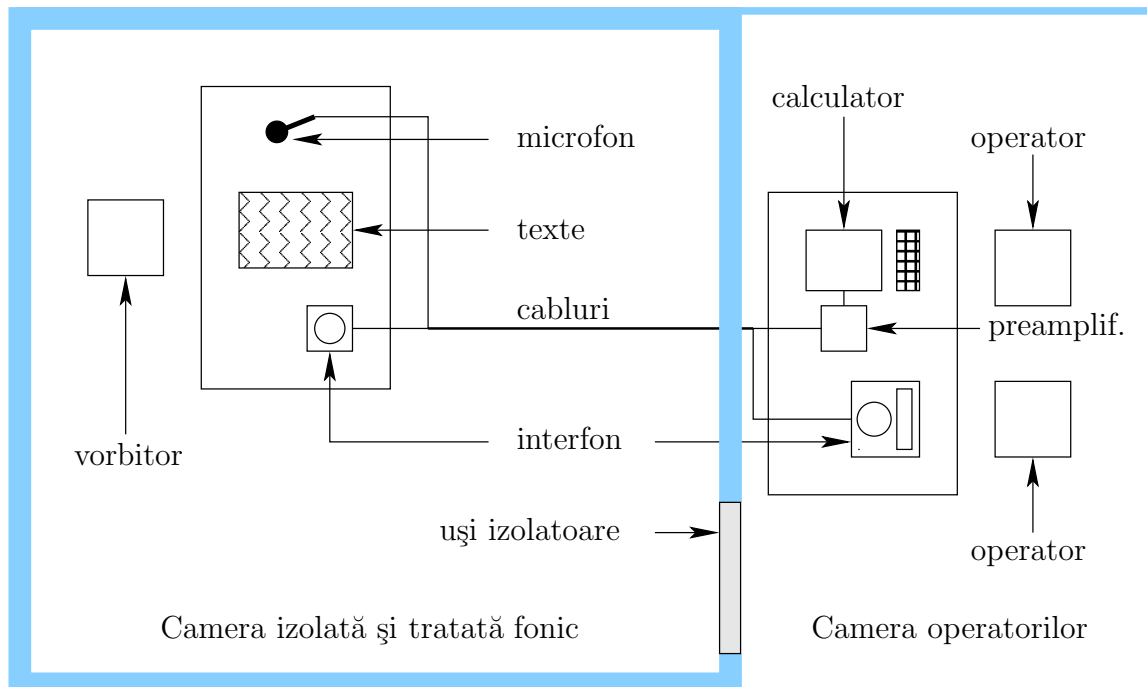


Figura 4.1: Schița camerelor folosite pentru înregistrări

această soluție are și două avantaje suplimentare: pe de o parte, facilitează detectarea de către operatori a zgomotelor sau altor probleme de pe durata înregistrărilor, permițând întreruperea și reluarea lor imediată, fără a mai aștepta o verificare auditivă ulterioară; pe de altă parte, reduce efectele nedorite care pot fi resimțite de persoane plasate într-o cameră izolată și tratată fonic, efecte mergând de la dificultăți 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 potențială a fost cea a ordinii de înregistrare a diferitor tipuri de materiale – citite sau semispontane. Pentru a evita pe cât posibil influența unui stil de vorbire asupra celuilalt, a fost stabilită următoarea ordine de înregistrare:

- informațiile personale (nume, adresă etc.) solicitate de către operatori și pronunțate semisontan de către vorbitori;
- alfabetul limbii române, citit de fiecare vorbitor în felul cu care era obișnuit;
- pasajele, citite cât mai fluent, fără a marca în mod special pauzele dintre propoziții;
- propozițiile de completare și individuale, citite cu pauze suficient de lungi între ele pentru a permite separarea vorbirii și a pauzelor cu algoritmi din EUROPEC;
- propozițiile de inițializare, citite ca și cele anterioare, dar vorbitorii au fost instruiți în mod special să le pronunțe cât mai clar și mai apropiat de varianta standard;

- numerele, citite ca și propozițiile, cu pauze între ele;
- logatomii de tip CVC, citiți izolați doar de vorbitorii din mulțimile restrânse FT și VT, și în contexte de câte două cuvinte doar de către cei din mulțimea VT;
- cuvintele-contexte CVC, citite izolat doar de vorbitorii din mulțimea VT.

Înregistrările au fost realizate folosind un microfon cu electret, omnidirecțional, de tip SONY ECM-44B, plasat la aproximativ 25 cm de gura vorbitorului și la un unghi de circa 30 grade față de direcția lui înainte. Pentru evitarea zgomotelor electrice și adaptarea la placa de achiziție și prelucrare a semnalelor din calculator, microfonul a fost conectat la aceasta prin cabluri ecranate și un preamplificator cu un câștig fix de circa 20 dB (figura 4.1). Semnalul a fost eșantionat la 20 KHz și cuantizat pe 16 biți.

Înainte sesiunii de înregistrări (a primei sesiuni, în cazul vorbitorilor din mulțimile restrânse FT și VT), fiecare vorbitor a furnizat o serie de informații personale, unele introduse și în fișierul de descriere a vorbitorilor, și a fost instruit asupra modului de lucru într-o scurtă sesiune de antrenament folosind informațiile personale și alfabetul.

Pe parcursul înregistrărilor, operatorii au urmărit permanent apariția unor zgomote sau erori majore de citire (omisiuni, inserții sau substituții de cuvinte, bâlbâieli etc.), cu întreruperea și reluarea imediată a înregistrărilor afectate. Înregistrările au fost refăcute imediat și în cazul în care asemenea probleme erau constatate prin ascultarea fișierelor de semnal. Cu excepția propozițiilor de inițializare, în cazul cărora vorbitorii au fost solicitați să le citească într-un mod cât mai clar și mai apropiat de varianta standard, variațiile de pronunție nu au constituit motive de refacere a înregistrărilor.

Pentru evitarea zgomotelor transmise în interiorul camerei izolate fonic prin pereții clădirii, înregistrările au fost efectuate în zile nelucrătoare (sâmbăta și duminica), ceea ce a făcut dificilă găsirea unor vorbitori dispuși să ia parte la înregistrări. Vorbitorii au fost recrutați urmărind satisfacerea criteriilor stabilite de sex și vârstă (secțiunea 4.4), precum și capacitatea de a citi în mod fluent materialele de înregistrat. În aceste condiții, 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 fișiere de semnal însumând peste 1,3 gigaocteți de date și cuprinzând peste 9 ore și 41 minute de înregistrări, distribuite pe diferite categorii de materiale conform tabelului 4.7.

Tabelul 4.7: Cantități de date colectate, pe categorii

	Pasaaje	Propoziții			Numere	CVC	Alfabet	Info
		Comp.	Ind.	Iniț.				
Fișiere	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"
Propoziții	2758	364	558	400	–	–	–	–

Dintre acestea, pasajele și propozițiile 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 și evaluarea sistemelor de recunoaștere automată a vorbirii, au o durată totală de 5 ore și 24 minute și cuprind 4080 de propoziții.

4.9 Calitatea înregistrărilor

Printre caracteristicile bazelor de date fonetice, enumerate la începutul capitolului, se numără și calitatea deosebită a înregistrărilor, apreciată prin elemente subiective (lipsa zgomotelor, corectitudinea pronunțiilor etc.) și 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 posibilității apariției zgomotelor. Alte zgomote acustice puteau fi produse chiar de vorbitorii însăși în timpul înregistrărilor: minimizarea lor a fost asigurată prin ascultarea de către operatori a înregistrărilor prin interfon și din fișiere în timpul și după efectuarea lor, urmată de reluare sau refacere când era cazul. Zgomotele electrice au fost minimizezate prin ecranarea legăturii microfon-preamplificator.

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

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

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

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

unde E_s și E_z sunt energia semnalului util respectiv zgomotului. Dat fiind caracterul nestaționar 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 staționar 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 distribuției valorilor energiei cadrelor de semnal (figura 4.2), care are două maxime: unul corespunzător pauzelor și fazelor de închidere ale sunetelor plozive nesonore, la valori mici, și unul corespunzător vorbirii – la valori mari.

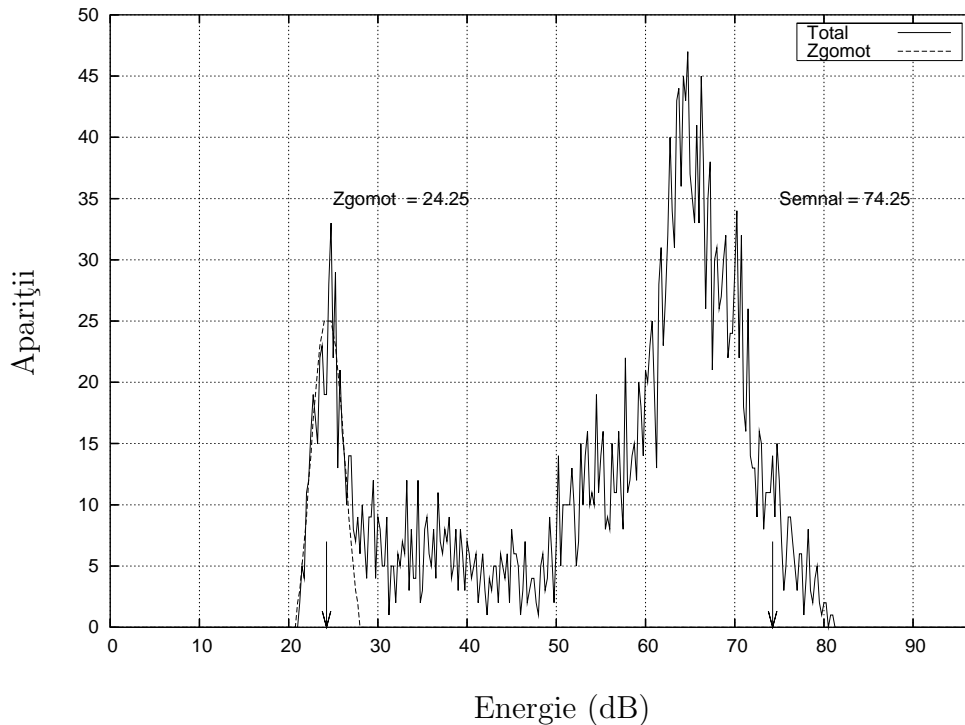


Figura 4.2: Exemplu de distribuție a energiei cadrelor dintr-o înregistrare

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% și respectiv 85%. În aceste condiții, valoarea raportului semnal/zgomot este calculată ca diferența dintre cele două valori (în decibeli).

Metode mai elaborate țin cont de distribuția 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 distribuții 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 aceeași metodă a procentelor complementare din paragraful de mai sus.

Această metodă a fost utilizată și 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 și valorile extreme de 33,75 și 57,75 dB.

Pentru comparație, analiza prin aceeași 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) și valori extreme de 27 și 69,75 dB. Ținând cont de faptul că înregistrările din TIMIT au fost făcute cu un microfon de mică distanță [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

Recunoașterea vorbirii prin metode statistice cere cantități considerabile de date pentru estimarea modelelor acustice și lingvistice pe care se bazează, iar acest capitol a prezentat proiectarea și colectarea unei baze de date cuprinzând semnale vocale în limba română, utilizabilă pentru antrenarea de modele acustice și testarea și evaluarea unor sisteme de recunoaștere 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 importanță mult mai mare din punctul de vedere al metodologiei cercetării, existența și utilizarea ei permițând efectuarea în condiții controlate a unor experimente de recunoaștere a vorbirii și comparații semnificative între rezultatele lor.

Date fiind inexistența cercetărilor anterioare asupra recunoașterii vorbirii continue în limba română și insuficiența celor de fonetică și fonologie a limbii române, constatate prin consultarea literaturii și a specialiștilor din aceste domenii, proiectarea și colectarea bazei de date au pus un accent deosebit pe calitatea înregistrărilor, în lipsa căreia unele investigații 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 unități fonetice de modelare acustică considerat acoperitor în raport cu sistemele fonologice ale limbii române identificate în literatura lingvistică, iar conținutul unei părți semnificative a înregistrărilor a fost planificat pentru a permite antrenarea de modele ale acestor unități.

Pentru a putea antrena și testa modele acustice independente de vorbitor, populația înregistrată a inclus un număr considerabil de vorbitori (100), iar pentru o cât mai bună acoperire a variabilității datorate lor, aceștia au fost selectați pentru a obține o distribuție uniformă după două variabile biologice controlabile – sexul și grupa de vârstă.

În sfârșit, pentru ca informația lingvistică inclusă în semnalele vocale înregistrate să fie cât mai puțin afectată de zgomote, distorsiuni, reverberații, o atenție deosebită a fost acordată condițiilor de realizare a înregistrărilor, iar analiza lor prin prisma câtorva criterii obiective de evaluare a calității acustice indică atingerea acestui obiectiv.

Datorită conținutului lingvistic controlat încă din faza de proiectare al unei părți semnificative a ei și calității deosebite a înregistrărilor, această bază de date va fi utilă nu doar pentru cercetările asupra recunoașterii automate a vorbirii continue, ci și în alte cercetări aplicative, precum și pentru cercetările fundamentale de fonetică acustică și fonologie a limbii române, a căror insuficiență a fost menționată.

CAPITOLUL 5

Etichetarea semnalelor vocale

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

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

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

Etichetarea semnalului vocal este deci o abstractizare dependentă de anumite puncte de vedere analitice și 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 același 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 și aplicative este realizată în secțiunea 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, și al fostului Minister al Cercetării și Tehnologiei prin contractul de grant 3019GR/1997-98.

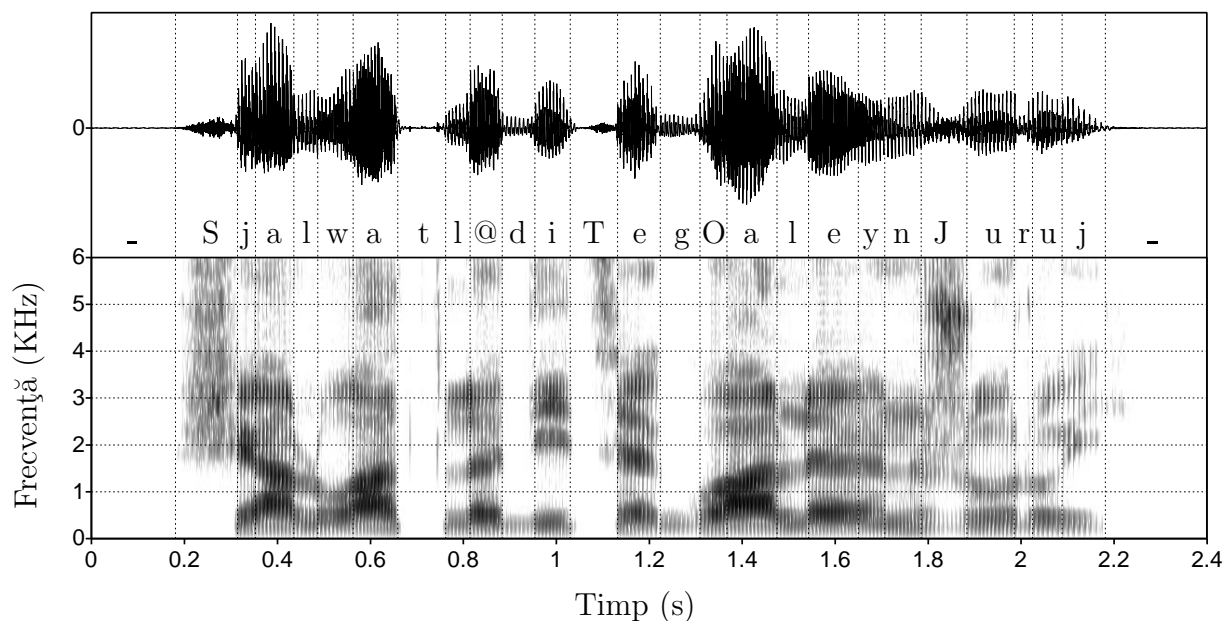


Figura 5.1: Exemplu de etichetare a propoziției ”Și-a luat lădițe goale în juru-i” folosind simbolurile ASCII din tabelul 4.1. Urmăriți cum se schimbă aspectul spectrogramei pentru diferite apariții ale sunetelor [j], [a] și [l], urmare a fenomenului de coarticulație.

Metoda de etichetare cea mai simplă consta în utilizarea unor programe adecvate [74], [107], [225] pentru vizualizarea formelor de undă și a spectrogramelor semnalelor și ascultarea lor selectivă, urmate de atribuirea manuală de poziții temporale și identități anumitor evenimente (lingvistice sau de altă natură). Deși aparent cea mai fiabilă, fiind bazată pe expertiza umană, această metodă este totuși susceptibilă de lipsa consistenței pozițiilor temporale și a identităților evenimentelor evidențiate în cadrul semnalelor, dată fiind existența unor cazuri ambigue care necesită decizii subiective ce vor varia cu experiența personală a experților implicați în etichetarea manuală.

Dar poate chiar mai important decât problema inconsistenței 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 construcției primelor baze de date fonetice, scop în care au fost propuse și experimentate diferite metode [142], [266], [68], [144], [40], iar secțiunea 5.2 face o trecere în revistă a celor mai semnificative aspecte ale problemei, precum și a metodelor de evaluare a rezultatelor etichetării automate.

Datorită utilității modelelor Markov ascunse și pentru cercetările ulterioare asupra recunoașterii automate a vorbirii, pentru etichetarea bazei de date a fost aleasă varianta unui sistem bazat pe MMA, a cărui construcție este descrisă în secțiunea 5.3.

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

În final sunt prezentate și comentate rezultatele etichetării automate, evaluate prin compararea etichetelor generate automat cu cele verificate și 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 și/sau lingvistice și corelațiile care se pot stabili între ele. Pentru cercetările noastre, dar și pentru multe altele, cele mai interesante sunt nivelurile care pot reprezenta legăturile dintre semnalele vocale și interpretările lor lingvistice în termenii unor unități fonetico-fonologice segmentale [18]:

- nivelul **fizic** cuprinde etichetele definite cu referință în exclusivitate la evenimentele fizice dintr-o pronunție, fiind în mod clar cel mai susceptibil de a fi divizat în straturi corespunzătoare diferitor metode de achiziție și prelucrare a semnalului;
- nivelul **fonetico-acustic** include etichete ce descriu evenimente omogene din punct de vedere acustic folosind termeni fonetici (închidere, eliberare, aspirație, fricțiune, sonoritate, nazalizare etc.), fără referiri la funcțiile lor lingvistice sau distinctive, sau la legături cu evenimente fizice; cu toate acestea, deciziile asupra delimitării lor temporale și a setului de simboluri utilizate ca identificatori adesea necesită sau sunt facilitate de informații asupra rolului lor în termeni fonologici;
- nivelul **fonetic restrâns** grupează etichete care caracterizează calitatea fonetică a sunetelor vorbirii folosind simboluri din alfabetul fonetic internațional (IPA) sau altele echivalente, reprezentabile pe calculator – SAMPA [251], Worldbet [104] etc.; în acest caz, impresia perceptuală a persoanei care realizează etichetarea este esențială pentru stabilirea delimitărilor temporale și a identităților sunetelor;
- nivelul **fonemic** este cel mai abstract dintre nivelurile prezentate aici, simbolurile folosite corespunzând fonemelor limbii în care a fost rostită o pronunție, așa cum apar ele în formele standard ale cuvintelor din cadrul pronunției, fără luarea în considerație a fenomenelor specifice vorbirii fluente (secțiunea 5.4.1); drept urmare, simbolurile nu vor putea fi puse întotdeauna în corespondență cu evenimente din semnal, astfel încât acest nivel nu este folosit pentru etichetarea efectivă; el este însă indispensabil ca mediator între semnal și vocabular, fiind prezent în dicționarele de pronunții pe care le folosesc sistemele de recunoaștere și sinteză a vorbirii;
- nivelul **fonetic extins**: adesea denumit fonemic, datorită faptului că utilizează de asemeni simboluri corespunzătoare fonemelor limbii în care a fost rostită pronunția adnotată, acest nivel, spre deosebire de cel anterior, ia în considerație fenomenele specifice vorbirii fluente și le reflectă ca atare, având astfel un grad de abstractizare intermediar între cel al nivelului fonetic restrâns și cel al nivelului fonemic.

Etichetarea avută în vedere aici trebuind să asigure legătura dintre caracteristicile anumitor porțiuni ale semnalului și 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 opțiune practică din motivele deja menționate, astfel încât în final am optat pentru nivelul fonetic extins.

Această opțiune are motivații multiple: nivelul fonetic extins este cel mai economic, reprezentând un maximum de informație fonetică cu un set minimal de simboluri; având în comun cu nivelul fonemic setul de simboluri folosit, are și avantajul că facilitează antrenarea și evaluarea la nivel de fonem a sistemelor de recunoașterea automată a vorbirii bazate pe modelarea acustică a unităților sublexicale de tip fonemic. În plus, acest nivel este mai fiabil decât nivelul fonetic restrâns în ceea ce privește consistența între diferiți experți a simbolurilor folosite pentru etichetarea aceluiași semnale, și aproximativ la fel de fiabil ca acesta în privința delimitărilor temporale [67].

5.2 Automatizarea etichetării

Așa după cum am precizat, etichetarea semnalului vocal în general, și cea la nivelul fonetic extins în particular, presupune două acțiuni:

- **segmentarea** semnalului prin identificarea momentelor de timp la care începe respectiv se termină o porțiune 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 folosește cel de etichetare. Preferăm însă folosirea sensului extins al termenului de etichetare, cel de segmentare și identificare, deoarece respectăm astfel definiția etichetării dată la începutul capitoului. În plus, fonemele dintr-o pronunție pot fi identificate și prin transcrierea ei, ceea ce nu înseamnă că a fost realizată o etichetare, deși a avut loc o adnotare a semnalului.

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

- **eliminarea** consta în transcrierea de către experți umani a fiecărei pronunții ce se dorește 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 rețea a unor posibile variante de pronunție [144], [126], [254]; o asemenea reprezentare poate fi obținută 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 pronunție specificate de reguli fonologice.

În principiu, multe dintre înregistrările din baza de date fiind însoțite 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 îmbogățită cu variante de pronunție prin aplicarea unor reguli fonologice, indisponibile la acel moment pentru limba română.

Ținând cont de această situație, pentru a asigura o calitate maximă a etichetării am ales varianta eliminării problemei identificării prin transcrierea manuală a semnalelor de etichetat [266] în termenii unităților fonetice din tabelul 4.1. În ceea ce privește segmentarea, am optat pentru realizarea ei folosind modele Markov ascunse datorită posibilității de a le utiliza în continuare și pentru recunoașterea automată a vorbirii.

5.2.1 Evaluarea etichetării automate

Ca și în cazul evaluării sistemelor de recunoaștere a vorbirii (secțiunea 3.1), și ieșirile sistemelor de etichetare automată pot fi privite ca ipoteze referitoare la identitățile și limitele sunetelor pronunțate în fișierele de semnal cărora le sunt asociate.

Spre deosebire de sistemele de recunoaștere, în cazul sistemelor de etichetare, pe lângă identitățile sunetelor, care pot fi comparate prin același algoritm de programare dinamică (secțiunea 3.1.1), în cazul în care identitățile coincid trebuie luate în considerație și diferențele dintre limitele de referință ale sunetelor și cele generate automat.

În plus, dacă referințele folosite pentru evaluarea sistemelor de recunoaștere sunt în general simplu de obținut prin transcrierea ortografică a fișierelor de semnal, nu aceeași este situația referințelor pentru evaluarea sistemelor de etichetare: așa cum am menționat deja, funcție de nivelul ales pentru etichetare și experiența experților implicați, între etichetările realizate de diferiți experți pot apare diferențe [67]. Ca atare, și rezultatele evaluării pot diferi funcție de referințele utilizate, o soluție propusă pentru atenuarea acestei probleme fiind cea a comparațiilor cu etichetări manuale multiple [252].

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

- un grup de metrice caracterizând performanțele de identificare – aceleași folosite și pentru evaluarea performanțelor sistemelor de recunoaștere automată a vorbirii (secțiunea 3.1); funcție de modul de tratare a problemei identificării (secțiunea 5.2), acestea vor reprezenta măsuri fie ale diferențelor de transcriere între experți umani, fie ale performanțelor combinate ale algoritmilor de traducere grafeme-foneme, de aplicare a eventualelor reguli fonologice și de etichetare propriu-zisă;
- metrice care descriu performanțele segmentării semnalelor vocale, obținute pe baza diferențelor dintre limitele stabilite automat și cele de referință; aceste diferențe sunt calculate doar dacă identitățile segmentelor comparate coincid, iar metricele derivate includ atât caracteristici globale ale distribuțiilor lor (media, mediana, abaterea standard etc. ale diferențelor sau valorilor lor absolute) cât și unele legate de anumite obiective de performanță, de tipul ”% diferențe între anumite limite.”

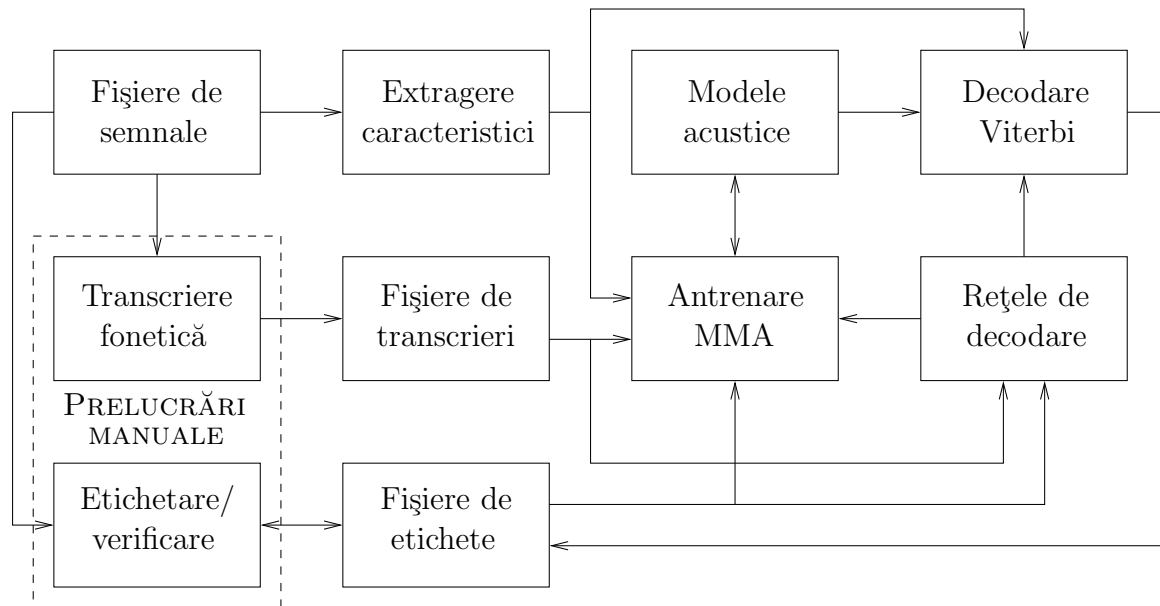


Figura 5.2: Schema bloc a sistemului de etichetare

5.3 Sistemul de etichetare

Așa cum am precizat în secțiunea 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 (secțiunea 3.9) pentru decodarea unor MMA compuse din modele acustice ale segmentelor conform transcrierilor semnalelor în termenii acestor unități, cu condiția existenței prealabile a modelelor acustice. În cazul nostru, inexistența unor modele ale unităților fonetice din tabelul 4.1 a făcut imposibilă această abordare, iar soluția a constat într-un sistem dezvoltat și utilizat în două etape (figura 5.2).

Într-o primă etapă, de dezvoltare a sistemului, a avut loc antrenarea unor modele acustice pentru unitățile din tabelul 4.1, simultan cu segmentarea automată a semnalelor utilizate în acest scop și 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 inițializare a funcțiilor de probabilitate ale modelelor acustice, construcția acestora a fost precedată de etichetarea manuală (figura 5.3) a celor 400 de propoziții de inițializare (secțiunea 4.3.2), cu o durată de circa 33'22", pe baza vizualizării formelor de undă și spectrogramelor de bandă largă ale semnalelor și a ascultării lor

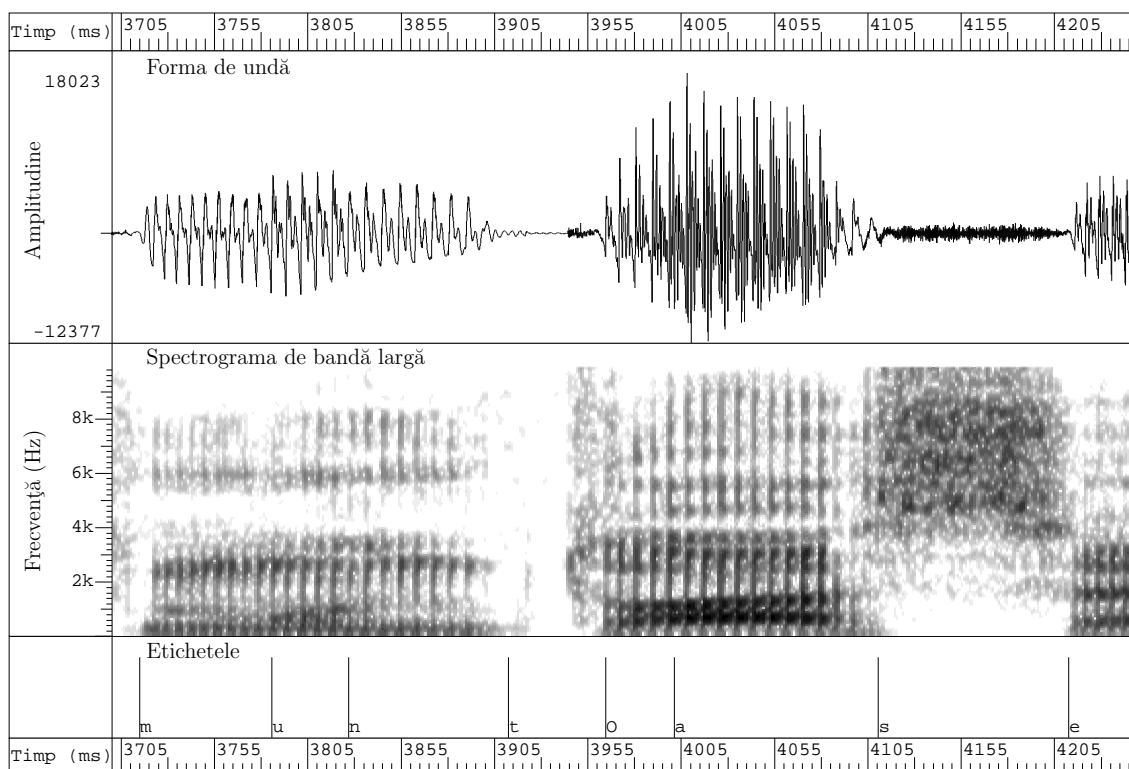


Figura 5.3: Exemplu de realizare manuală a etichetării și verificării

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

Etichetarea manuală a fost făcută la nivel fonetic extins, ținând cont de fenomenele specifice vorbirii fluente (secțiunea 5.4.1), de corelațiile generale ce se pot stabili între proprietățile acustice și cele articulatorii ale sunetelor vorbirii [69], [186], precum și de experiența acumulată în cadrul altor cercetări [266], [81], [17], [133], [55]. Dat fiind însă numărul mic de propoziții 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 secțiunea 5.4.

5.3.2 Transcrierea fonetică

Pe lângă cele 400 de propoziții de inițializare etichetate manual, în prima etapă au mai fost utilizate pasajele și propozițiile de completare înregistrate de către toți vorbitorii în prima sesiune (2230 de propoziții cu o durată totală de circa 2h56'12"), dar singura lor prelucrare manuală a fost transcrierea fonetică în termenii unităților din tabelul 4.1, ținând cont de fenomenele specifice vorbirii fluente (secțiunea 5.4.1).

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

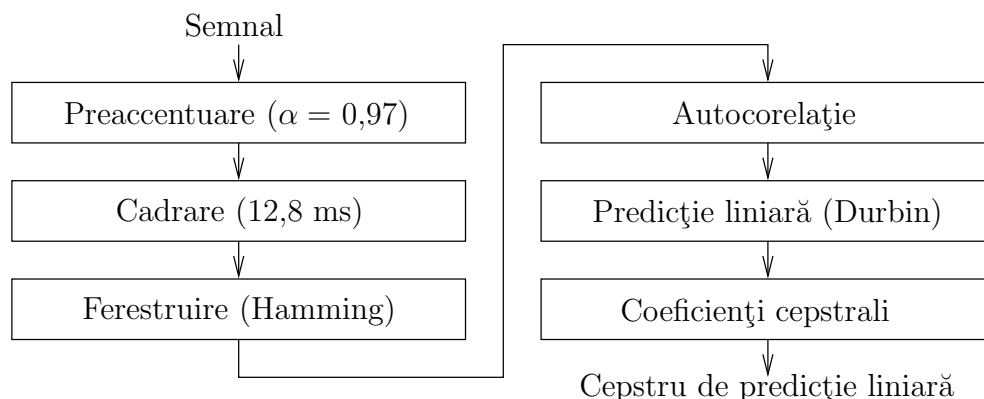


Figura 5.4: Analiza cepstrală prin predicție liniară

În etapa a doua au fost prelucrate alte 1450 de propoziții cu o durată totală de circa 1h54'26": pasajele și propozițiile de completare înregistrate în câte patru sesiuni suplimentare de cei 10 vorbitori din mulțimea FT (892 de propoziții cu o durată totală de circa 1h11'23") și cele 558 de propoziții individuale (circa 43'3"). Pentru pasaje și propozițiile de completare a fost utilizată aceeași adaptare a transcrierilor-prototip, iar propozițiile 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 și mai frecvente decât cele utilizate în mod curent pentru recunoașterea vorbirii, cu o durată de 12,8 ms (256 eșantioane) și o deplasare de 5 ms între cadre (figura 5.4).

După preaccentuarea semnalelor (secțiunea 2.2.4) cu un coeficient $\alpha = 0,97$ și cadrare, fiecare cadru a fost ferestruit cu o fereastră Hamming (secțiunea 2.2.1). În continuare, din fiecare cadru au fost calculate log-energia (secțiunea 2.2.2) și, prin metoda autocorelației (secțiunea 2.3.1), un predictor liniar de ordinul 12. Pe baza ecuației (2.55), coeficienții de predicție liniară au fost convertiți în coeficienții cepstrali $c_{1..12}$.

Coeficienții cepstrali au fost supuși unei liftrări conform ecuației (3.18) cu lungimea $L = 12$ și, împreună cu log-energia, unui proces de derivare peste intervale de 40 ms (ecuația 2.60 cu $L = 4$), aproximat prin diferențe simple la capetele fișierelor. Fișierele de semnal au fost astfel transformate în secvențe de vectori acustici 26-dimensional, incluzând fiecare 12 coeficienți cepstrali, log-energia și coeficienții lor Δ (secțiunea 2.7.2).

5.3.4 Modelele acustice

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

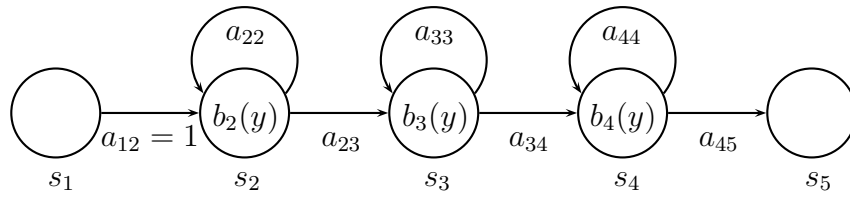


Figura 5.5: Structura MMA ale unităților fonetice

una de tip stânga-dreapta (figura 5.5) cu trei stări emițătoare și două stări (inițială și finală) utilizate doar pentru concatenarea lor în modele ale pasajelor și propozițiilor.

Corespunzător vectorilor acustici 26-dimensionalii extrași din semnale, funcțiile de probabilitate $b_j(y)$, $j = 2 \dots 4$, au fost mixturi de densități gaussiene 26-dimensionale cu matrice de covarianță diagonale (secțiunea 3.7). Stărilor inițială și finală, s_1 și s_5 , ale fiecărui model, destinate doar interconectării lor, nu le-au fost asociate funcții de probabilitate, iar stările emițătoare $s_2 \dots s_4$ au avut rolul de a modela porțiunile inițială, medie și respectiv finală ale fiecărei unități. Ținând cont de deplasarea de 5 ms a cadrelor, modelele corespund astfel unei durate minime de 15 ms a unei unități fonetice.

Datorită introducerii în MMA a stărilor inițiale și finale neemițătoare, reestimarea probabilităților de tranziție (ecuația 3.37) trebuie modificată în cazul acestor stări: în mod evident, datorită structurii alese, probabilitatea $a_{12} = 1$, iar formula de reestimare a probabilităților de tranziție în ultima stare devine

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

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

Inițializarea

Câte 200 de propoziții de inițializare înregistrate de vorbitorii de același sex și etichetate manual au fost folosite pentru inițializarea modelelor acustice ale unităților fonetice. Pentru fiecare model, toate aparițiile unității asociate au fost divizate în trei segmente egale, corespunzătoare celor trei stări emițătoare $s_2 \dots s_4$, rezultând astfel o divizare în submulțimi disjuncte a vectorilor acustici extrași din aceste propoziții.

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

$\boldsymbol{\mu}_{jk}$, matrice de covarianță diagonală \mathbf{C}_{jk} și pondere w_{jk} egale cu cele ale grupului

$$\boldsymbol{\mu}_{jk} = \frac{\sum_{i=1}^{N_{jk}} \mathbf{o}_{ijk}}{N_{jk}} \quad (5.2)$$

$$\mathbf{C}_{jk}[l] = \frac{\sum_{i=1}^{N_{jk}} (\mathbf{o}_{ijk}[l] - \boldsymbol{\mu}_{jk}[l])^2}{N_{jk}}, \quad l = 1 \dots 26 \quad (5.3)$$

$$w_{jk} = \frac{N_{jk}}{\sum_{k=1}^K N_{jk}} \quad (5.4)$$

unde \mathbf{o}_{ijk} , $i = 1 \dots N_{jk}$ sunt cei N_{jk} vectori acustici din grupul k corespunzător stării s_j , iar $\mathbf{C}_{jk}[l]$ – componentele de pe diagonala matricei sale de covarianță.

Folosind, pe lângă densitățile gaussiene astfel inițializate, probabilități de tranziție atribuite manual, datele au fost resegmentate prin algoritmul Viterbi (secțiunea 3.9) iar parametrii gaussianelor au fost reactualizați conform metodei de mai sus. După câteva iterații ale acestui proces, inițializarea a fost încheiată cu algoritmul Baum-Welch (secțiunile 3.5.1 și 3.7.1), care a permis și reestimarea probabilităților de tranziție.

Reestimarea concatenată

Inițializarea modelelor acustice s-a făcut separat pentru fiecare unitate fonetică, neținând cont că realizările lor nu apar separat, ci doar interconectate. Pentru a lua în calcul și acest aspect s-a utilizat algoritmul Baum-Welch concatenat [135], [136], care operează asupra unor modele compuse ale propozițiilor sau pasajelor folosite pentru antrenament, formate prin concatenarea de instanțe ale modelelor unităților fonetice în conformitate cu rețele de decodare obținute din transcrierile sau etichetele asociate.

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

În plus, în cazul unei structuri de tipul celei din figura 5.5 a modelelor unităților fonetice, concatenarea lor duce la modificarea formulei de reestimare a probabilităților de tranziție în stările finale: cu excepția ultimei instanțe dintr-un model compus, pentru care se utilizează formula (5.1), probabilitățile reestimate ale tranzițiilor finale devin

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

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

5.3.5 Segmentarea automată

După inițializarea și reestimarea concatenată a modelelor unităților fonetice, fișierele de semnal transcrise fonetic au fost segmentate prin decodarea Viterbi a modelelor lor compuse, având ca observații vectorii acustici extrași din acele fișiere.

În etapa de dezvoltare a sistemului, procesul de antrenare a modelelor și segmentare a semnalelor a fost reluat de câteva ori, folosind însă pentru inițializarea modelelor fișierele de semnal transcrise fonetic și 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 șase iterații de segmentare/reantrenare.

5.3.6 Verificarea etichetării

Segmentările semnalelor și alinierea la acestea ale transcrierilor lor fonetice produse în mod automat de sistemul de etichetare sunt în multe cazuri foarte bune. Există însă și destul de multe cazuri, inevitabile pentru orice sistem automat, în care ele suferă de erori de poziționare, sau cazuri, mai puțin frecvente, în care transcrierile fonetice au fost din start afectate de erori. Ținând cont de existența acestor erori, pentru a asigura o calitate cât mai bună a etichetării și a face posibilă o evaluare a sistemului de etichetare, a fost necesară verificarea și, 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 și identificarea segmentelor. Deoarece aceste criterii au fost definite și rafinate în mod iterativ pe parcursul cercetărilor legate de etichetarea semnalului vocal, pentru a asigura consistența globală a etichetării, verificarea a inclus și rezultatele etichetării manuale a propozițiilor de inițializare.

5.4 Criteriile de decizie

Datorită coarticulației sunetelor vorbirii, un fonem poate fi semnalat printr-o serie de indicii distribuite în mai multe segmente de semnal vocal, și invers, proprietățile 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 ușoară, 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 conținute în acel segment [18].

Indiciile fonemice putând fi distribuite de-a lungul mai multor segmente succesive, etichetarea și verificarea manuală au acordat o pondere redusă percepției 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 și al spectrogramelor), care pot fi cel mai adesea corelate cu mișcările 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 rezoluții mai reduse în timp datorită cadrării semnalelor în cursul generării lor.

Corelațiile dintre proprietățile articulatorii și cele acustice ale sunetelor vorbirii [69], [186] sunt suficient de puternice pentru ca experți umani să poată reconstitui conținutul unor pronunții pe baza examinării spectrogramelor lor [265], [157], ele putând fi cu atât mai mult folosite pentru decizii asupra etichetării, când semnalele pot fi și ascultate.

Tabelul 5.1: Clasificarea consoanelor limbii române după criteriile manierei și 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 și 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 și maniera de articulare (tabelul 5.1), iar în cel al vocalelor – prin locul de articulare și gradul de deschidere a cavității bucale (tabelul 5.2). În plus, consoanele pot fi distinse și după sonoritate – în tabelul 5.1, aceasta diferențiază perechile de sunete cu aceleași locuri de articulare din primele trei linii, cele sonore fiind *evidențiate*.

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

Maniera de articulare, sonoritatea și 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 proprietățile acustice ale semnalelor, fiind din această cauză baza cea mai potrivită pentru deciziile asupra segmentării [69].

Locul de articulare, relevant și pentru vocale, are drept corelate acustice valorile frecvențelor formanților, determinate de volumele cavităților componente ale tractului vocal: F_1 este determinată în principal de volumul faringelui, iar F_2 – de cel al cavității bucale. Având o evoluție în general lentă în timp (v. figurile 2.3 și 5.1), corelatele locului de articulare sunt utile în special pentru deciziile asupra identității segmentelor.

Sonoritatea este indicată cu maximum de fiabilitate de prezența în spectrogramă a primului formant, cauzat de rezonanța faringelui, deoarece vibrațiile coardelor vocale, specifice sunetelor sonore, sunt singura sursă posibilă de excitație a faringelui.

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

5.4.1 Fenomenele specifice vorbirii fluente

Pe durata etichetării manuale, a transcrierii fonetice și verificării etichetării, opțiunea pentru nivelul fonetic extins (secțiunea 5.1) a condus la luarea în considerație a unor fenomene specifice vorbirii fluente: asimilarea, eliziunea și epenteza.

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

Eliziunea, sau omiterea segmentului corespunzător unui fonem, apare în vorbirea fluentă chiar și la viteze moderate. În cazul eliziunii complete, chiar dacă un fonem apare în pronunția "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 pronunțat, chiar dacă la examinarea semnalului nu putem localiza un segment specific: în asemenea cazuri au fost etichetate componentele acustice efectiv apărute – de exemplu, nazalizarea unei vocale percepută ca realizare a unei consoane nazale (secțiunea 5.4.5).

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

5.4.2 Vocalele și semivocalele

Datorită producerii lor fără obstrucții sau constricții ale tractului vocal, vocalele sunt caracterizate prin structură formantică și, de obicei, sonoritate. Caracteristicile articulatorii și cele acustice sunt legate în acest caz prin două corelații majore: pe de o parte, cea a locului de articulare cu caracterul grav sau acut al vocalei, indicat de valoarea frecvenței 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 și caracterul difuz sau compact al spectrului, indicat de diferența $F_2 - F_1$ – mare (spectru difuz) pentru vocalele închise, mică (spectru compact) pentru cele deschise. Valorile frecvențelor formanților trebuie însă interpretate relativ la ansamblul unei pronunții, ținând cont de evoluțiile lor în timp și variațiile dintre vorbitori.

Deoarece F_2 este determinată de rezonanța cavității 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 tranziției sale.

Secvențele 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 tranziției formantice, în caz contrar se utilizează indiciile furnizate de schimbările de amplitudine ale formei de undă și variațiile de energie din spectrogramă. În lipsa altor indicii, o secvență 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 africte sunt de obicei clare datorită manierei de articulare a acestora, iar cele față de consoanele sonante corespund creșterii intensității formanților și amplitudinii și/sau complexității formei de undă.

5.4.3 Consoanele plozive

Pe durata producerii acestor consoane pot apare trei evenimente acustice distincte: liniștea, corespunzătoare închiderii articulatorilor, explozia, produsă la eliberarea lor, și aspirația de după explozie. Explozia este formată din impulsuri depășind amplitudinea oricărui zgomot de fricțiune ulterior, corespunzător aspirației, iar în cazul în care aceasta din urmă lipsește, poate fi identificată ca un impuls în forma de undă, de amplitudine mult mai mică decât a sunetului sonor următor. Aspirația este semnalul aperiodic de după explozie, al cărui sfârșit este marcat de începutul vibrațiilor coardelor vocale pentru următorul sunet sonor. Limitele consoanelor plozive corespund deci începutului închiderii articulatorilor și sfârșitului exploziei sau eventualei aspirații.

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

După o pauză, limita inițială a unei plozive sonore este poziționată la începutul vibrațiilor coardelor vocale. Dacă vibrațiile 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 și/sau spectrogramă, iar limita inițială a consoanei plozive asociate este plasată la acest moment. În lipsa acestor indicii, ca și în cazul anterior, limita inițială 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 și plozivă fiind în acest caz plasată pe baza spectrogramei, în punctul de dispariție a formanților.

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 creștere marcată a energiei la frecvențe de peste 500 Hz.

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

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

5.4.4 Consoanele fricative și africte

Datorită intensității zgomotului de înaltă frecvență care le caracterizează, /s/, /z/, /S/ și /J/ sunt cel mai ușor de identificat dintre consoanele fricative, începutul lor putând fi determinat pe baza creșterii 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 formanților în raport cu sunetele alăturate și indicii ale fricțiunii 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 diferențelor de intensitate și frecvență inferioară a zgomotului din spectrogramă, provocate de schimbarea locului de articulare, iar în cazul identității lor – prin divizarea în părți egale a segmentului fricativ.

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

Deoarece o consoană africată este realizată asemănător închiderii unei plozive, urmată de o fricțiune homorganică înlocuind explozia, criteriile din acest caz sunt similare celor folosite pentru plozive și fricative: limita inițială 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 frecvența de cca. 500 Hz. Amplitudinea semnalului variază de obicei în aceste puncte, ca și cea a oscilațiilor de înaltă frecvență, corespunzătoare formanților, suprapuse peste frecvența fundamentală, astfel încât pe durata nazalelor forma de undă este mai simplă, iar formanții sunt mai slabi și discontinui în raport cu cei ai vocalelor din jur.

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

Eliziunea nazalelor apare frecvent în vorbirea fluentă, caz în care ele pot fi totuși 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 și scăderii energiei la frecvențe înalte, dar formantul al doilea este considerabil mai puternic decât în cazul nazalelor. Când există tranziții formantice, delimitările se fac la mijlocul acestora, ca și în cazul (semi)vocalelor.

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

5.4.6 Problema /I/

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

Analiza datelor etichetate arată însă că acesta are o frecvență de realizare foarte redusă, sub 20% din cea așteptată în urma proiectării bazei de date, iar examinarea semnalelor și 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 neclaritățile legate de existența realizărilor lui fizice și proprietățile lor acustice și 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 corespondență i-au fost alocate scurte porțiuni de semnal între fonemele adiacente.

5.5 Rezultate și comentarii

Pentru dezvoltarea sistemului de etichetare descris în secțiunea 5.3 au fost etichetate manual cele 400 de propoziții de inițializare, iar pasajele și propozițiile de completare și cele individuale au fost etichetate semiautomat, pe parcursul dezvoltării sistemului (pasajele și propozițiile de completare înregistrate de toți vorbitorii în prima sesiune) și prin utilizarea lui (propozițiile individuale și pasajele și propozițiile de completare înregistrate de vorbitorii din mulțimea FT în sesiunile suplimentare).

Au fost astfel etichetate toate cele 4080 de propoziții, iar pentru asigurarea calității și a consistenței, toate fișierele de etichete au fost verificate și corectate manual conform criteriilor din secțiunea 5.4. Aceleași criterii au fost folosite și pentru a eticheta manual logatomii CVC, alfabetul și informațiile semispontane, care nu au fost prelucrate folosind sistemul de etichetare datorită simplității și/sau caracteristicilor acustice diferite de cele ale propozițiilor utilizate la construcția acestuia. Nu au fost etichetate numerele.

Dat fiind modul de dezvoltare și utilizare a sistemului de etichetare, evaluarea lui a fost făcută doar pe baza analizei etichetării celor 558 propoziții individuale pentru a obține o cât mai bună estimare a capacității lui de generalizare: aceste propoziții 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 și propozițiilor de completare utilizate pentru antrenarea modelelor acustice ale sistemului de etichetare.

Evaluarea a fost făcută conform metodologiei din secțiunea 5.2.1 prin compararea fișierelor de etichete generate automat plecând de la transcrierile fonetice manuale cu cele verificate și corectate manual. Metricele caracterizând performanțele de identificare reprezintă în acest caz o măsură a corectitudinii transcrierii fonetice manuale și au, cum era de așteptat, valori foarte bune: transcrierile inițiale au cuprins 27553 de etichete, iar fișierele corectate – 27580, deci o diferență sub 0,1%. Alinierea lor (secțiunea 3.1.1) a identificat 37 substituții (sub 0,14%), 66 omisiuni (sub 0,24%) și 39 inserții (sub 0,15%), corespunzătoare unei corectitudini de peste 99,62% și unei acurateți de peste 99,48%.

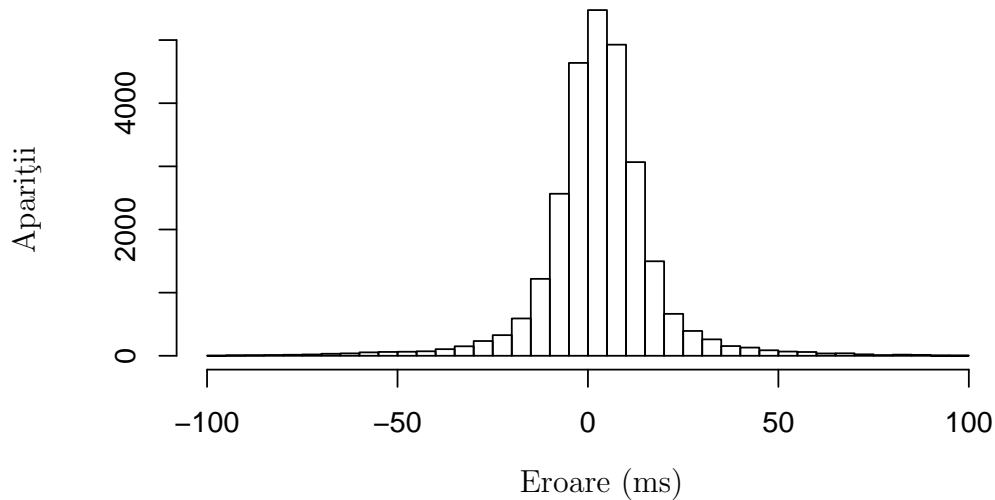


Figura 5.6: Histograma erorilor de segmentare

Analiza erorilor arată că substituțiile 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 și în foarte mică măsură altor confuzii la tastare (/s/-/S/).

Evaluarea performanțelor 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 diferențe între limitele generate de sistemul de etichetare și cele stabilite în urma verificării și corectării manuale, au fost cuprinse între -275 și $226,002$ ms, cu media de $3,027$ ms, mediana de $3,102$ ms și 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 și din histogramă, pe ansamblu sistemul a avut o ușoară tendință de întârziere a segmentării.

Pentru o apreciere mai exactă a performanțelor 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 metrici 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 și 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

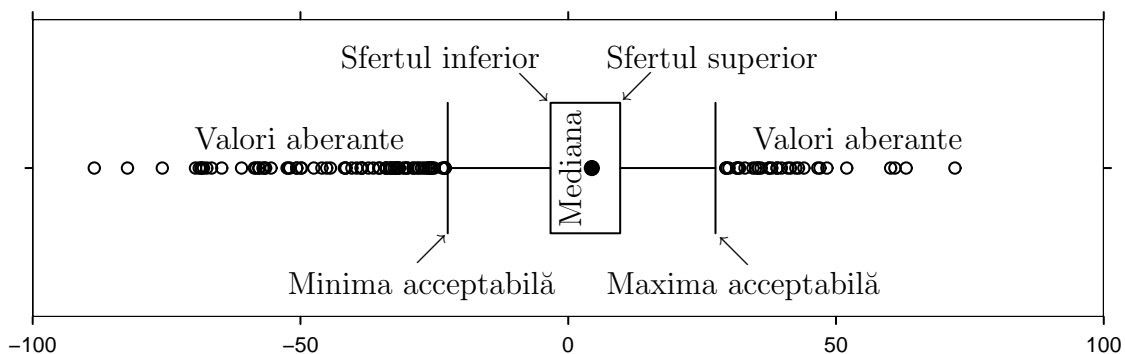
Tabelul 5.4: Clasele de sunete folosite în analiza erorilor de segmentare

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 distribuțiilor erorilor, sunt reprezentate funcție de clasele sunetelor adiacente în figura 5.8.

Explicația 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 și 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 aceeași clasă, cele mai afectate fiind, în ordine, clasele P (consoanele plozive și pauza), V (vocalele și semivocalele /E/ și /O/) și S (consoanele sonante și semivocalele /j/ și /w/). Mecanismele de producere a erorilor diferă însă între clase: în cazul clasei P, principala cauză, indicată și de deplasarea distribuției erorilor, este dificultatea de a detecta începutul unei noi închideri în lipsa unei explozii asociate celei anterioare; în cazul claselor V și S, problemele își au originea în modificările lente ale caracteristicilor spectrale ale sunetelor din aceste clase, fără evenimente ușor detectabile, iar distribuțiile sunt aproape simetrice.

**Figura 5.7:** Reprezentarea prin *boxplot* a unei distribuții

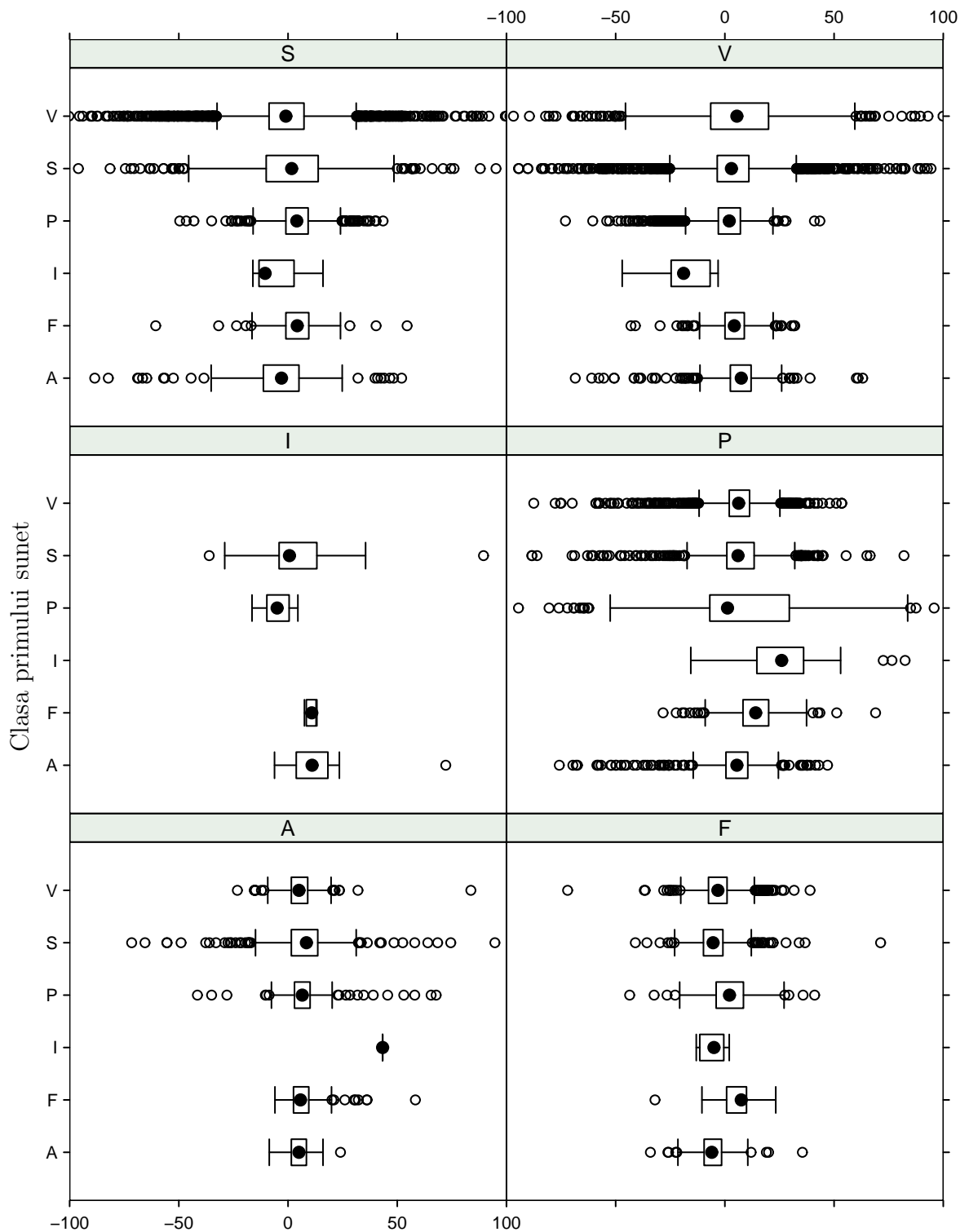


Figura 5.8: Distribuțiile erorilor de segmentare funcție de clasele sunetelor adiacente (clasa celui de-al doilea sunet este titlul unuia dintre cele șase panouri)

5.6 Concluzii

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

Din punctul de vedere al cercetărilor în direcția recunoașterii automate a vorbirii, etichetarea corespunzătoare a unei asemenea baze de date permite localizarea aparițiilor unor unități de modelare acustică și utilizarea lor pentru antrenarea modelelor acustice, și asigură transcrierile de referință necesare evaluării rezultatelor în experimentele de recunoaștere desfășurate la nivelul acestor unități 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 și al utilizării lor în cursul recunoașterii, ca și al etichetării în sine, pentru etichetarea bazei de date proiectate și colectate în cursul acestor cercetări a fost ales nivelul fonetic extins. Acesta utilizează simboluri ale fonemelor unei limbi (în cazul nostru, ale unităților 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 soluții, iar pentru etichetarea bazei noastre de date am ales varianta unui sistem bazat pe modele Markov ascunse dependente de sex ale unităților fonetice de modelare acustică.

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

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

Dezvoltarea și utilizarea sistemului de etichetare și 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 și corectate a dovedit bunele performanțe de segmentare ale acestuia – peste 80% din limitele între segmente evaluate au fost la maximum 15 ms de pozițiile de referință, iar peste 95% – la maximum 35 ms.

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

CAPITOLUL 6

Experimente de modelare acustică

Așa cum am menționat în secțiunea 3.7, modelarea acustică joacă un rol esențial în sistemele automate de recunoaștere a vorbirii, de modul în care aceasta reușește să acopere variabilitatea semnalelor vocale depinzând performanțele lor, iar experimentele descrise în acest capitol constituie prima abordare a problemelor modelării acustice folosind unități sublexicale pentru recunoașterea 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, și pentru a cărei tratare alegerea unităților de modelare acustică poate fi determinantă, este coarticulația sunetelor vorbirii, menționată deja în secțiunea 4.2 și exemplificată în figura 5.1.

Printre sursele extralingvistice se numără mediile în care sunt produse semnalele și canalele de comunicație prin care ele sunt transmise de la vorbitori la sistemele de recunoaștere: în cazul acestora, soluția constă în utilizarea unor metode de extragere a caracteristicilor capabile să reducă efectele nedorite ale mediilor sau canalelor și 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 comunicație intermediare.

Alte surse de variabilitate țin de vorbitori și cuprind caracteristicile lor biologice (sex, vârstă, stare fiziologică etc.), sociale (educație, ocupație etc.) și lingvistice (eventualul dialect vorbit, particularități de pronunție etc.) În aceste cercetări am încercat acoperirea acestui tip de variabilitate prin selectarea și 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 și mai ușor de apreciat – sexul și grupa de vârstă.

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

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

Problemele modelării acustice au fost apoi studiate prin experimente de recunoaștere independentă de vorbitor desfășurate atât la nivelul unităților de modelare, cât și la cel lexical (al cuvintelor), dependent și independent de vocabular.

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

6.1 Experimente dependente de vorbitor

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

Pentru construcția sistemului au fost folosite circa 15 minute de semnal vocal colectat de la un singur vorbitor, obținut prin citirea celor 40 de pasaje și a propozițiilor de inițializare folosite pentru înregistrarea bazei de date (aproximativ 200 de propoziții).

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

Modelele acustice au avut aceeași structură ca și 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 unități de modelare. Funcțiile de probabilitate ale stărilor emițătoare au fost mixturi gaussiene 24-dimensionale cu matrice de covarianță diagonală.

Antrenarea modelelor acustice a fost realizată prin aceeași procedură folosită și pentru etichetare (secțiunea 5.3.4), pentru inițializare fiind utilizate cele patru propoziții de inițializare etichetate manual, iar pentru reestimarea Baum-Welch concatenată – cele 40 de pasaje, transcrise fonetic. Datorită cantității 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 recunoaștere automată a vorbirii continue în limba română au utilizat în procesul de decodare lingvistică toate constrângerile secvențiale dintr-un sistem de recunoaștere tipic (figura 1.1) – dicționare de pronunții ale cuvintelor, respectiv un model statistic și o gramatică a pronunțiilor de recunoscut.

Dicționarele de pronunții și 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 pronunții distincte, incluzând două corespunzătoare pauzelor din și dintre propoziții. Transcrierea lor în termenii unităților de modelare acustică a rezultat într-un prim dicționar de pronunții, care a inclus în transcrieri și fenomene de fonetică sintactică – de exemplu, ”ce-ar” transcris /Car/ etc.

O a doua variantă a dicționarului de pronunții a fost obținută prin eliminarea /I/ din transcrierile fonetice ale cuvintelor conform secțiunii 6.1.2.

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

$$P(w_2|w_1) = N(w_1w_2)/N(w_1) \quad (6.1)$$

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

Dicționarele, modelele lingvistice și cele acustice au fost utilizate într-un algoritm Viterbi cu reducerea spațiului de căutare (pruning – secțiunea 3.9), implementat folosind liste de stări ale MMA active la un moment dat [136]. Pentru a reduce viteza de creștere a spațiului de căutare, tranzițiile între cuvinte au fost restricționate în mod euristic până la atingerea unor durate ale cuvintelor corespunzătoare unei medii de 60 ms a duratelor unităților sublexicale componente [31].

6.1.2 Rezultate și comentarii

În aceste prime experimente, aceleași date au fost folosite și pentru antrenarea MMA și pentru testarea sistemului de recunoaștere. Deși din punct de vedere metodologic o asemenea abordare (testare pe datele de antrenament [64]) este contraindicată, neputând evidenția capacitățile 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 totuși și la un prim rezultat semnificativ din punct de vedere al modelării acustice.

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

Datorită acestei situații, utilizarea setului de unități de modelare din tabelul 4.1 a avut ca urmare obținerea 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 tendința de a determina probabilități importante ale multor cuvinte scurte având pronunții terminate în /I/ (îmi, își, îți etc.), care erau inserate în mod frecvent și determinau erori de recunoaștere foarte numeroase.

Această problemă a fost rezolvată prin eliminarea lui /I/ din setul de unități de modelare acustică, urmată de modificarea corespunzătoare a dicționarului de pronunții și a transcrierilor semnalelor și construcția unor noi modele acustice. Aceasta a dus la dispariția fenomenului menționat mai sus și a făcut posibile **primele demonstrații ale recunoașterii automate a vorbirii continue în limba română** (în iunie 1997).

6.2 Experimente independente de vorbitor

Odată un prim sistem de recunoaștere disponibil, cercetările au continuat în direcția recunoașterii independente de vorbitor. Trecerea de la recunoașterea dependentă la cea independentă de vorbitor s-a făcut prin selectarea din baza de date a unor submulțimi de vorbitori cu distribuții similare pe sexe și grupe de vârstă: submulțimi de câte 60 de vorbitori pentru antrenarea modelelor acustice, respectiv 20 de vorbitori pentru testarea și evaluarea sistemelor de recunoaștere bazate pe aceste modele.

Dintre materialele folosite pentru înregistrarea bazei de date (secțiunea 4.3), cele 40 de pasaje și propozițiile de completare asociate au fost citite de toți vorbitorii, ceea ce a permis evaluarea modelelor acustice în mod dependent de vocabular. Existența în baza de date și a unor semnale obținute prin citirea de propoziții individuale, specifice fiecărui vorbitor, a făcut posibilă și evaluarea în mod independent de vocabular. Evaluările au fost făcute atât la nivelul unităților de modelare, cât și la cel lexical, al cuvintelor.

Primele experimente de recunoaștere independentă de vorbitor au fost efectuate la nivelul sublexical, al unităților de modelare acustică, și au urmărit evaluarea dependentă și independentă de vocabular a celor două seturi alternative de unități 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, obținut prin eliminarea fonemului /I/.

Deși recunoașterea și evaluarea la nivel sublexical pot fi utile pentru a îmbunătăți performanțele recunoașterii la nivel lexical [82], recunoașterea trebuie în cele din urmă efectuată la nivelul cuvintelor. Restul experimentelor au fost în consecință dedicate recunoașterii la nivel lexical, evaluată dependent și independent de vocabular.

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

6.3 Recunoașterea unităților de modelare

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

6.3.1 Vorbitori și date

În aceste experimente au fost folosiți 80 de vorbitori – 60 de antrenament și 20 de test – uniform distribuiți pe sexe și grupe de vârstă (tabelul 6.1).

Datele de antrenament au inclus pasajele și propozițiile de completare înregistrate de vorbitorii de antrenament, iar pentru teste au fost folosite mulțimi de date colectate

Tabelul 6.1: Mulțimile de vorbitori folosite în experimentele de recunoaștere a unităților 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 recunoaștere a unităților de modelare acustică

Texte	Număr propoziții	Număr cuvinte	Lungime medie a propozițiilor	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 propoziții individuale specifice vorbitorilor de test (4-5 propoziții/vorbitor).

Caracteristicile textelor folosite pentru înregistrarea datelor sunt prezentate în tabelul 6.2, iar caracteristicile mulțimilor de date de antrenament și de test – în tabelul 6.3. Față 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 propozițiile 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 mulțimea datelor de antrenament.

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

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

Date	Număr propoziții	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 și a literaturii lingvistice în privința sunetelor limbii române a condus inițial la conturarea a două posibile seturi alternative de unități de modelare acustică.

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

A doua alternativă consta în eliminarea lui /I/ din setul inițial de unități de modelare, deja testată în sistemul dependent de vorbitor. Aceasta s-a dovedit singura fezabilă în condițiile existente deoarece frecvențele reduse de apariție ale secvențelor 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

Fișierele de semnal vocal din mulțimile de date de antrenament și de test au fost supuse unei analize cepstrale melodice (secțiunea 2.7.1) conform figurii 6.1: după preaccenuarea cu un coeficient $\alpha = 0,97$, eşantioanele semnalelor au fost grupate în cadre cu lungimea de 25,6 ms (512 eşantioane) distanțate la 10 ms, fereștruite cu o fereastră Hamming și 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 obținuți 12 coeficienți cepstrali melodici.

Coeficienții cepstrali melodici au fost liftrați cu un liftru de tip sinus ridicat cu o lungime $L = 22$ (ecuația 3.18) și împreună cu o estimare a log-energiei fiecărui cadru și coeficienții lor Δ (secțiunea 2.7.2) au format vectori acustici 26-dimensionalii folosiți efectiv la antrenarea modelelor acustice și recunoaștere.

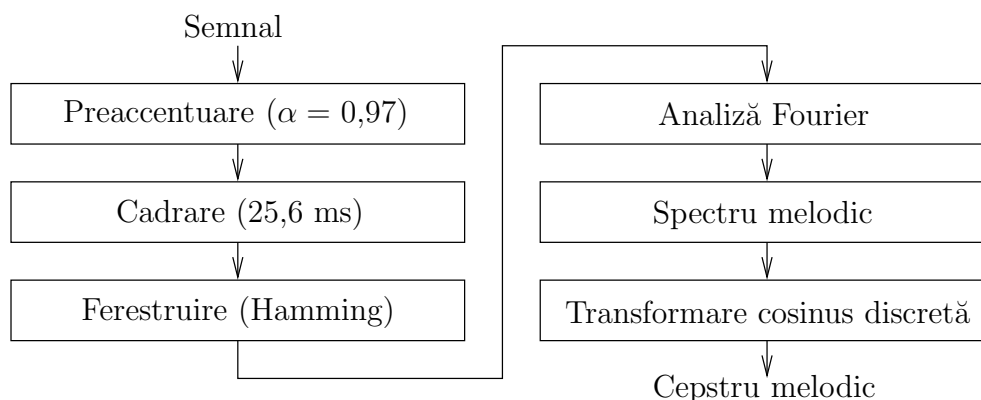


Figura 6.1: Analiza cepstrală melodică

6.3.4 Modelele acustice

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

Antrenarea MMA a fost făcută conform procedurii din secțiunea 5.3.4; pentru a evita însă unele instabilități ale algoritmului de grupare folosit, în faza de inițializare vectorii asociați unei stări emițătoare au fost utilizați pentru estimarea unei singure gaussiene. După fiecare reestimare concatenată, numărul de gaussiene al fiecărei stări emițătoare a fost incrementat prin divizarea gaussienei de pondere maximă și înlocuirea ei cu altele două de ponderi egale cu jumătate din ponderea celei inițiale. Vectorii medii ai noilor gaussiene au fost obținuți prin deplasări ale vectorului mediu al gaussienei inițiale 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 și comentarii

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

Analiza rezultatelor evidențiază faptul că procedura de antrenare utilizată conduce la modele cu o foarte bună capacitate de generalizare: frecvența recunoașterilor corecte evaluată independent de vocabular este superioară celei evaluate dependent de vocabular atât în cazul setului de bază de unități de modelare, cât și al celui redus (figura 6.2), iar acuratețea evaluată independent de vocabular este foarte apropiată de cea evaluată dependent de vocabular (figura 6.3).

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

Tabelul 6.4: Rezultatele experimentelor de recunoaștere a unităților 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 recunoaștere a unităților 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

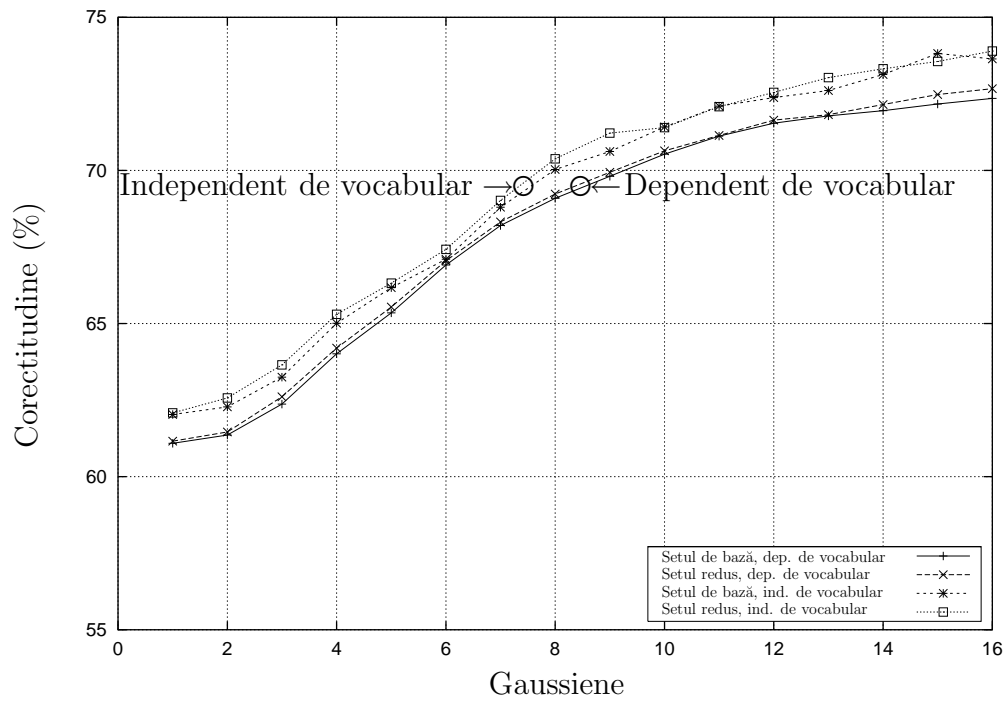


Figura 6.2: Evoluția corectitudinii recunoașterii unităților de modelare acustică funcție de numărul de gaussiene/stare emițătoare

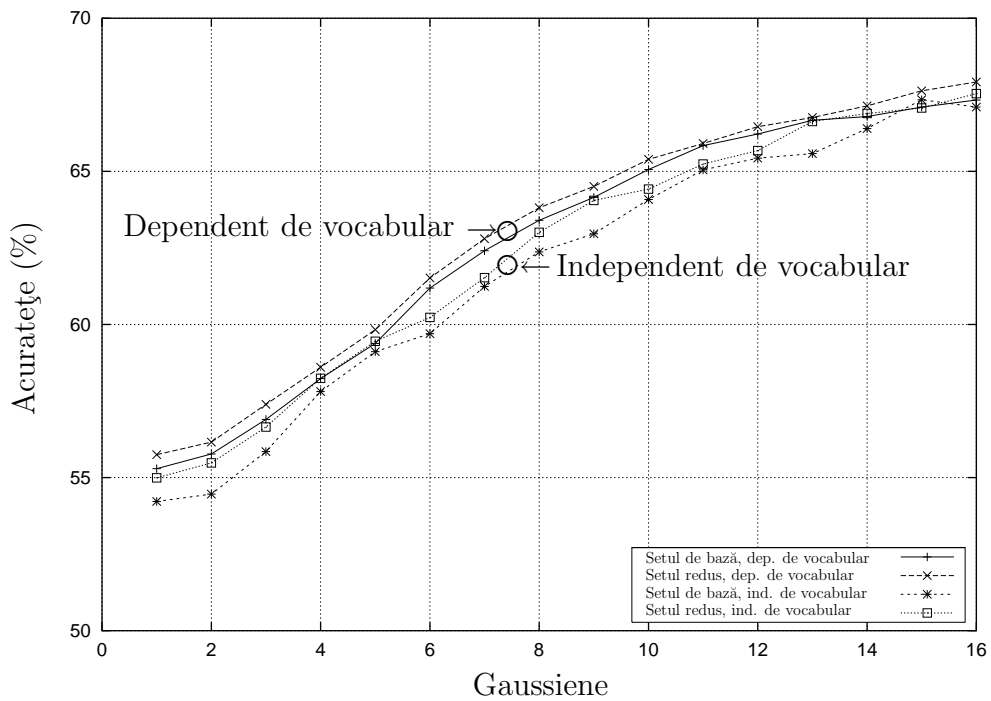


Figura 6.3: Evoluția acurateții recunoașterii unităților de modelare acustică funcție de numărul de gaussiene/stare emițătoare

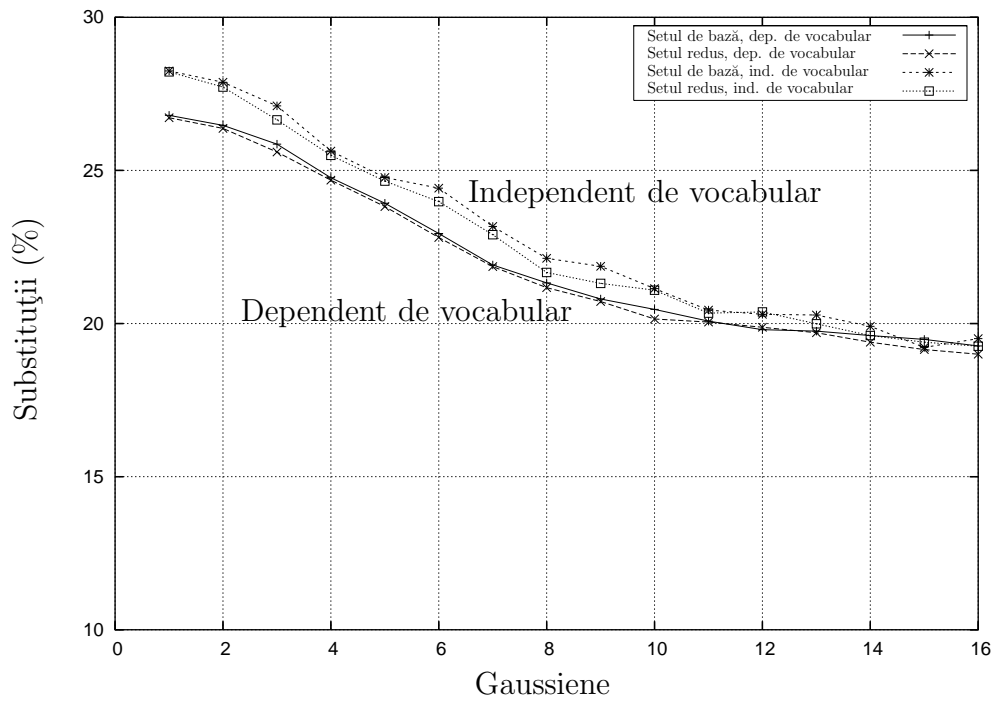


Figura 6.4: Evoluția frecvenței substituțiilor unităților de modelare acustică funcție de numărul de gaussiene/stare emițătoare

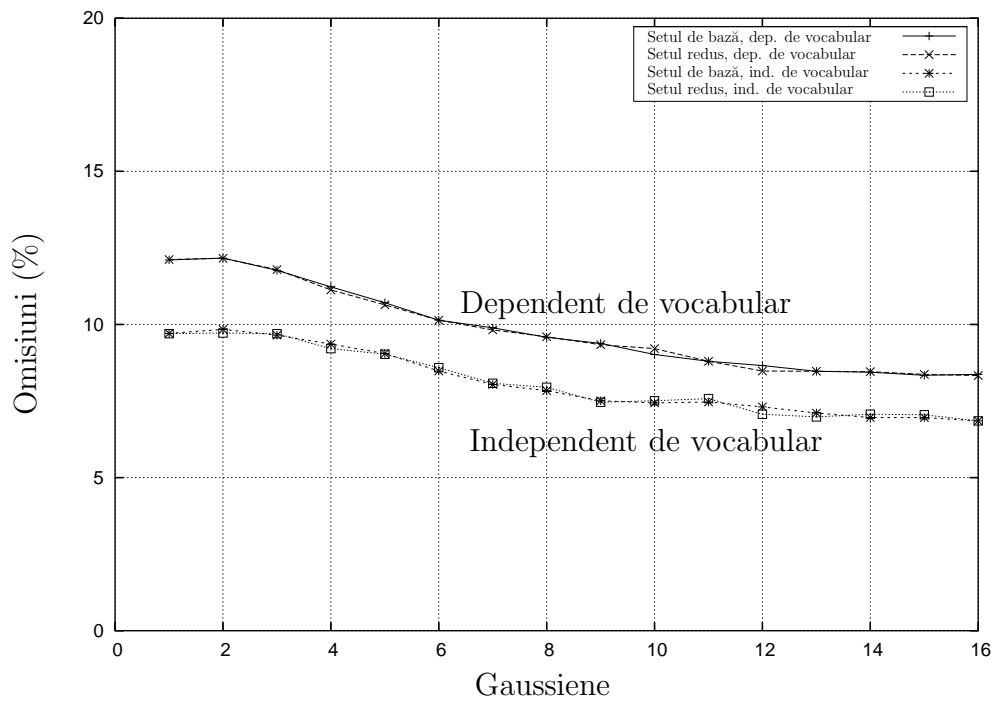


Figura 6.5: Evoluția frecvenței omisiunilor unităților de modelare acustică funcție de numărul de gaussiene/stare emițătoare

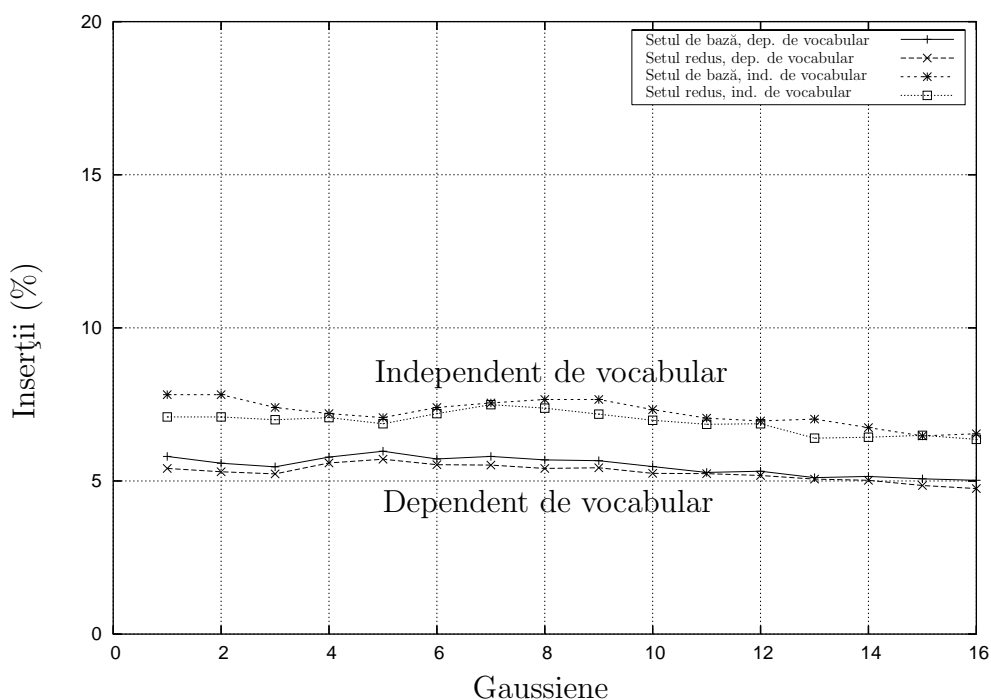


Figura 6.6: Evoluția frecvenței inserțiilor unităților de modelare acustică funcție de numărul de gaussiene/stare emițătoare

după cum se observă, setul redus (fără /I/) a condus la rezultate ușor superioare atât din punctul de vedere al corectitudinii, cât și al acurateții recunoașterii, indiferent de dependența sau independența de vocabular. Studiind și tipurile de erori – substituții (figura 6.4), omisiuni (figura 6.5) și inserții (figura 6.6) – se observă că diferențele între rezultate sunt datorate nu atât modificării setului de unități de modelare, cât mai ales dependenței sau independenței de vocabular.

Influența setului de unități de modelare asupra omisiunilor (figura 6.5) este practic neglijabilă, în acest caz diferența fiind dată de dependența/independența de vocabular. În schimb, ea este sesizabilă în cazul inserțiilor (figura 6.6) și substituțiilor (figura 6.4), care au fost în general reduse prin utilizarea setului alternativ de unități de modelare. Reducerea inserțiilor este sistematică, indiferent de dependența sau independența de vocabular, iar cea a substituțiilor se manifestă în special independent de vocabular. În plus, valorile frecvenței substituțiilor converg odată cu creșterea numărului de gaussiene.

Comparând rezultatele obținute cu unele [85] raportate în condiții apropiate (35 unități de modelare cu 16 densități gaussiene 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 mulți vorbitori (60 față de 43) și mai mult semnal (aproape 106 minute față de cca. 50), consider că aceste rezultate sunt normale.

Această opinie este susținută și de rezultatele obținute în condiții comparabile pentru limba spaniolă [35]: cu 25 unități de modelare și 3 gaussiene pe stare, dependent de

Tabelul 6.6: Vorbitorii folosiți în experimentele de recunoaștere 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 față de $C = 62,6\%$ pentru româna, iar independent de vocabular – $C = 61,4\%$ pentru spaniolă față de $C = 63,6\%$ pentru română.

6.4 Recunoașterea cuvintelor

În experimentele de recunoaștere a cuvintelor descrise în continuare, ca și în cele de recunoaștere a unităților de modelare, a fost utilizată o gramatică simplă de tip buclă (figura 3.3) pentru a obține informații referitoare strict la modelarea acustică.

Pentru aceste experimente au fost utilizate și două pachete publice de programe pentru construcția și evaluarea sistemelor de recunoaștere a vorbirii: unul dezvoltat la Universitatea statului Mississippi [176], și HTK (HMM Toolkit) [260], dezvoltat la Universitatea Cambridge [258] și 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 maturității și al documentației superioare.

6.4.1 Vorbitori și date

Experimentele de recunoaștere a cuvintelor au utilizat tot o submulțime de 60 de vorbitori de antrenament și una de 20 de vorbitori de test (tabelul 6.6), dar schimbate față de experimentele de recunoaștere a unităților de modelare pentru a maximiza cantitatea de semnale disponibile pentru testele independente de vocabular.

Datele folosite pentru antrenarea modelelor acustice și 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 propoziții, însoțită de eliminarea pauzelor dintre ele.

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

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

Date	Număr propoziții	Număr cuvinte	Cuvinte distincte	Cuvinte specifice	Durăță
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 creșterea atât a dimensiunii vocabularului aferent de la 575 la 784 cuvinte, cât și a numărului de cuvinte specifice (de la 412 la 585) și a ponderii lor în acest vocabular (de la 71,7% la 74,6%). A rezultat de asemeni și o ușoară creștere (de la 9,65 la 9,7 cuvinte) a lungimii medii a propozițiilor individuale utilizate pentru testele independente de vocabular.

6.4.2 Dicționarele

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

Corespunzător celor două seturi de unități de modelare acustică evaluate (de bază și redus), fiecare dintre aceste dicționare a avut la rândul lui două variante, în care pronunțiile cuvintelor au fost precizate în termenii unităților din setul respectiv.

Au fost adăugate pronunții alternative pentru a ține cont de posibila apariție a unor pauze între cuvinte (v. și secțiunea 6.4.3) și de variantele de pronunție ale cuvintelor (de exemplu, "optsprezece" poate fi pronunțat /optsprezeCe/ sau /opSpe/).

Deoarece gramatica de tip buclă utilizată în aceste experimente nu permite distincții între cuvintele cu pronunții identice (homofone), prin examinarea dicționarelor au fost identificate perechile de cuvinte homofone, ale căror substituții nu au fost considerate erori. Listele acestor cuvinte, împreună cu pronunțiile lor, sunt incluse în anexa A.

6.4.3 Modelele acustice

Pentru recunoașterea cuvintelor au fost folosite tot modele acustice cu o structură de tip stânga-dreapta cu două stări conectoare și trei emițătoare (figura 5.5). Deoarece analiza semnalelor a fost făcută ca și pentru experimentele de recunoaștere a unităților de modelare (secțiunea 6.3.3), funcțiile de probabilitate asociate stărilor emițătoare au fost mixturi de densități gaussiene 26-dimensionale cu matrice de covarianță diagonale.

Procedura de antrenare a fost însă modificată pentru a ține cont de dicționarele de pronunții ca intermediare între nivelul unităților de modelare și cel lexical. Astfel, după inițializarea și prima reestimare concatenată a modelelor cu o gaussiană pe stare emițătoare folosind datele de antrenament etichetate, modelul pauzelor a fost înlocuit cu

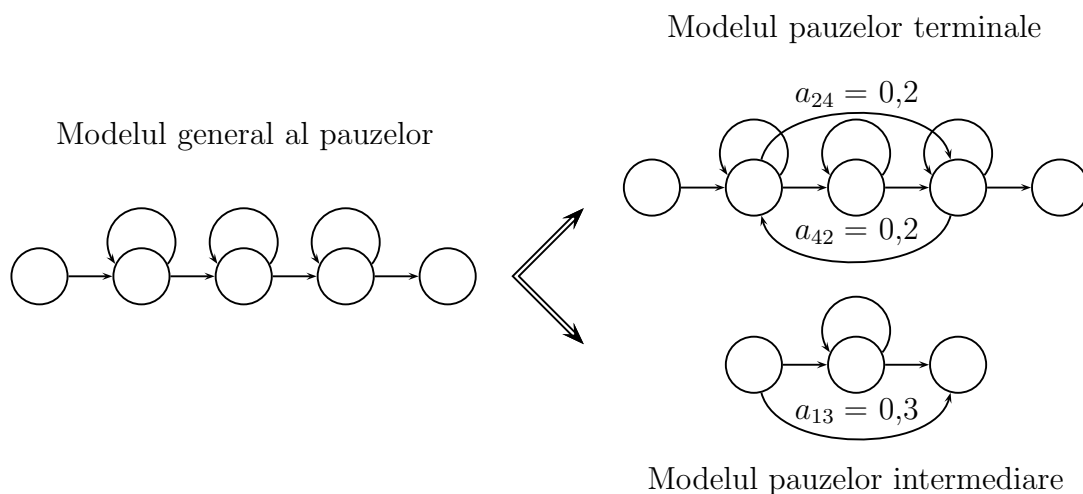


Figura 6.7: Înlocuirea modelului general al pauzelor cu modele ale pauzelor terminale (de la începutul și sfârșitul propozițiilor) și intermediare (dintre cuvinte)

alte două obținute prin editarea lui (figura 6.7): unul pentru pauzele terminale (inițială și finală) și unul pentru cele intermediare, care pot apare opțional între cuvinte [260].

Modelul pauzelor terminale a fost obținut prin copierea modelului general al pauzelor, iar cel al pauzelor intermediare a inclus o singură stare emițătoare pentru a putea modela și pauze foarte scurte între cuvinte. Inițial, această stare avea funcția de probabilitate a stării de mijloc din modelul general al pauzelor, iar în continuare parametrii acesteia au fost partajați cu cei ai stării de mijloc din modelul pauzelor terminale.

Caracterul opțional al pauzelor intermediare a fost marcat în modelul lor printr-o tranziție directă din starea inițială în cea finală. Alte tranziții au fost adăugate în modelul pauzelor terminale pentru a permite traversări mai rapide sau multiple ale acestuia. Tranzițiilor noi adăugate le-au fost atribuite probabilități arbitrare, iar probabilitățile celorlalte tranziții au fost scalate pentru a respecta restricțiile stochastice specifice.

Luarea în considerație a dicționarelor de pronunții a început prin utilizarea pentru următoarele reestimări concatenate nu a etichetelor semnalelor, ci a transcrierilor lor fonetice obținute din cele ortografice prin înlocuirea fiecărui cuvânt cu o pronunție a lui incluzând pauza finală opțională dintre cuvinte. Aceste transcrieri au fost folosite pentru încă două reestimări Baum-Welch concatenate ale modelelor.

Detectarea aparițiilor efective ale pauzelor opționale dintre cuvinte a fost realizată prin decodarea Viterbi a datelor de antrenament folosind rețele de recunoaștere obținute prin înlocuirea fiecărui cuvânt din transcrierile lor ortografice cu pronunțiile de bază și cele incluzând pauzele opționale finale, conectate în paralel. A rezultat astfel un nou set de transcrieri fonetice, care au luat în calcul atât pronunțiile din dicționare cât și pauzele dintre cuvinte și care au fost folosite pentru restul reestimărilor concatenate.

După încă două reestimări concatenate a început mărirea numărului de gaussiene pe stare emițătoare (secțiunea 6.3.4). Pentru a ține însă cont de diferențele care apar între

pronunțiile din dicționar și 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 și comentarii

Experimentele de recunoaștere a cuvintelor au fost efectuate utilizând setul de bază și cel redus de unități de modelare și MMA cu până la 16 gaussiene pe stare emițătoare. Decodarea a fost realizată folosind implementarea cu liste înlănțuite [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 tranziție și a pragului de reducere (secțiunea 3.9).

Urmărind maximizarea simultană a corectitudinii și acurateții recunoașterii și evitarea erorilor de căutare, au fost determinate experimental penalizarea de tranziție de -25 și pragul de reducere de -600 pentru testele dependente de vocabular, respectiv -30 și -900 pentru cele independente de vocabular: se observă că în cazul independenței de vocabular au fost necesare valori mai mari în valoare absolută, corespunzătoare unei tendințe mai puternice spre inserții, respectiv unui spațiu mai extins al soluțiilor.

Experimentele finale, desfășurate folosind acești parametri, au fost evaluate conform metodologiei din secțiunea 3.1, iar rezultatele sunt prezentate numeric în tabelele 6.8 și 6.9 și grafic în figurile 6.8–6.12 – atenție la diferențele de scară! Pentru a ilustra unele aspecte discutate în continuare, exemple de recunoaștere sunt incluse în anexa B.

După cum era de așteptat, dat fiind numărul mai redus de cuvinte din acest caz (784 față de 1043 – v. tabelul 6.7), performanțele independente de vocabular sunt superioare celor dependente de vocabular. Diferența dintre cele două condiții de test este ușor observabilă în toate reprezentările grafice ale evoluțiilor metricilor de performanță cu numărul de gaussiene/stare emițătoare, cu excepția frecvenței inserțiilor, în cazul căreia diferența este greu sesizabilă datorită alegerii valorilor penalizării de tranziție.

Se observă de asemeni apariția unor abateri pronunțate de la tendințele de ameliorare ale performanțelor în cazurile modelării cu 8 și 14 gaussiene/stare emițătoare, abateri manifestate la nivelul tuturor metricilor cu excepția frecvenței omisiunilor și al căror mecanism nu a fost încă elucidat: dat fiind faptul că abaterile, deși mai puțin pronunțate, apar și în cazul testelor dependente de vocabular, ar putea fi vorba de particularități ale datelor de antrenament care îngreuiază antrenarea unor modele în aceste cazuri, sau de fenomene care țin de dinamica procesului de antrenare a modelelor acustice.

Examinând rezultatele și din punctul de vedere al setului de unități de modelare folosit, constatăm că setul de bază a condus la rezultate ușor superioare în cazul testelor independente de vocabular cu peste 8 gaussiene/stare și al celor dependente de vocabular cu mai puțin de 8 gaussiene/stare, iar cel redus – în cazul celor dependente de vocabular cu peste 8 gaussiene/stare. Studiul ipotezelor decodate și al aliniilor lor cu transcrierile de referință, exemplificate în anexa B, arată însă că aceste diferențe nu pot fi atribuite direct noilor cuvinte homofone introduse prin utilizarea setului redus (anexa A).

Evaluarea la nivelul de semnificație $p = 0,05$ a semnificației statistice [37] a diferențelor dintre performanțele obținute folosind setul de bază și cel redus de unități de modelare acustică, realizată prin compararea distribuțiilor numerelor de erori din zonele eronate corespondente statistic independente [91], [182], arată că aceste diferențe sunt în general

Tabelul 6.8: Rezultatele experimentelor de recunoaștere a cuvintelor folosind setul de bază de unități 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 recunoaștere a cuvintelor folosind setul redus de unități 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

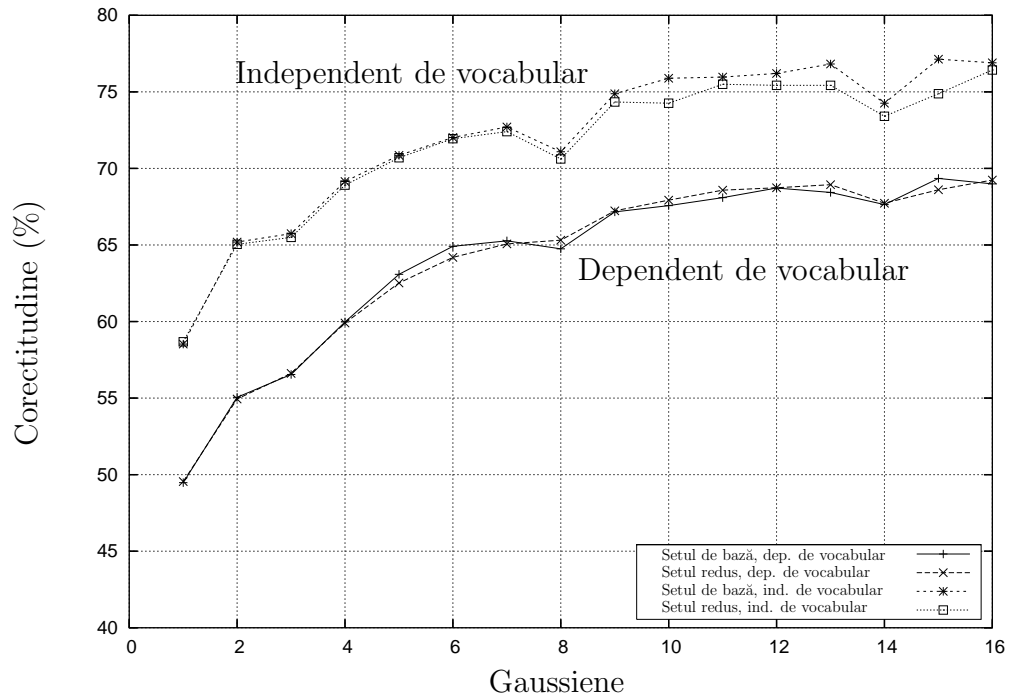


Figura 6.8: Evoluția corectitudinii recunoașterii cuvintelor funcție de numărul de gaussiene/stare emițătoare

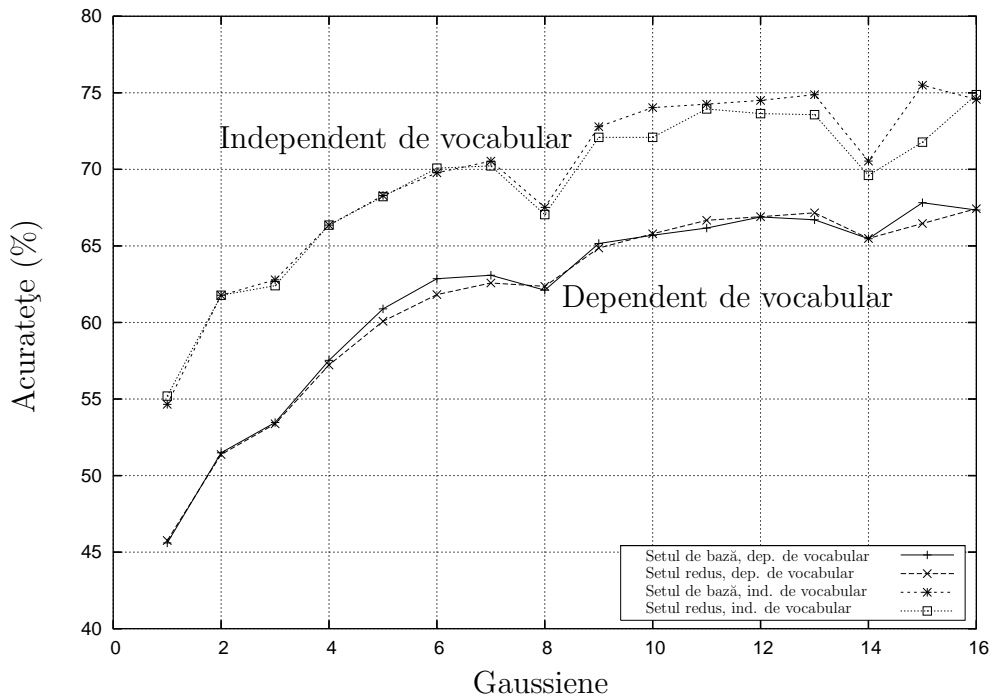


Figura 6.9: Evoluția acurateții recunoașterii cuvintelor funcție de numărul de gaussiene/stare emițătoare

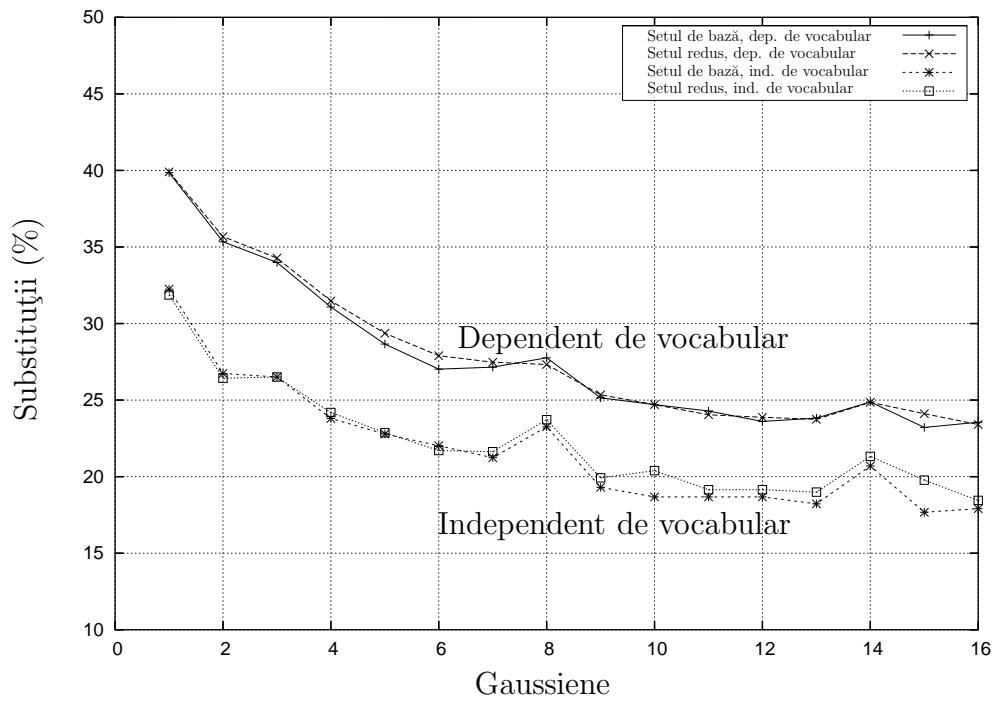


Figura 6.10: Evoluția frecvenței substituțiilor cuvintelor funcție de numărul de gaussiene/stare emițătoare

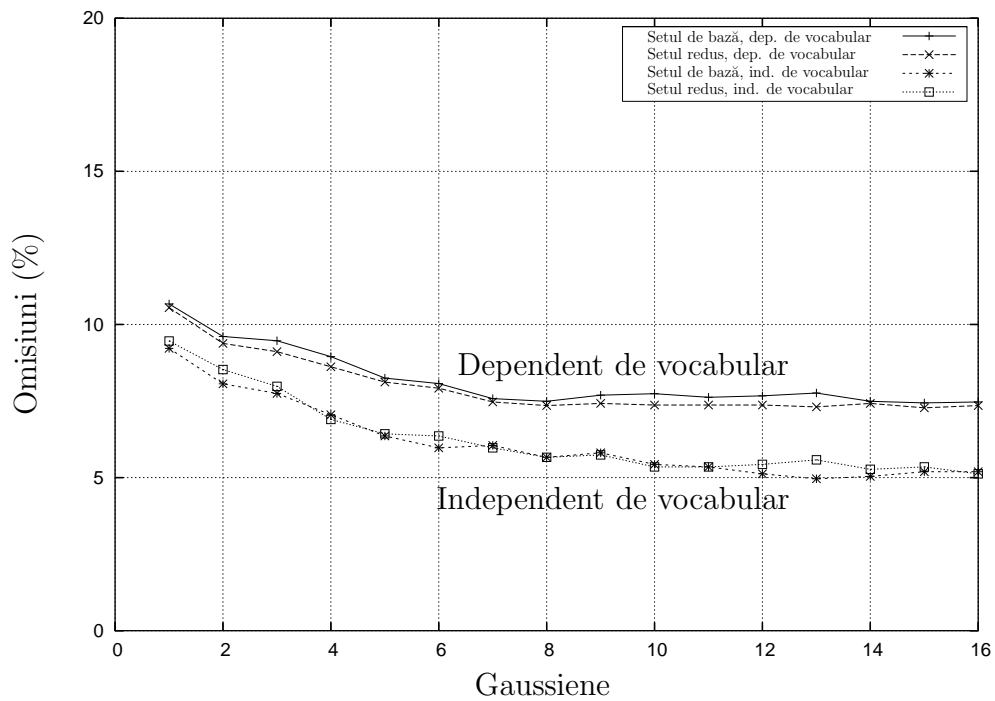


Figura 6.11: Evoluția frecvenței omisiunilor cuvintelor funcție de numărul de gaussiene/stare emițătoare

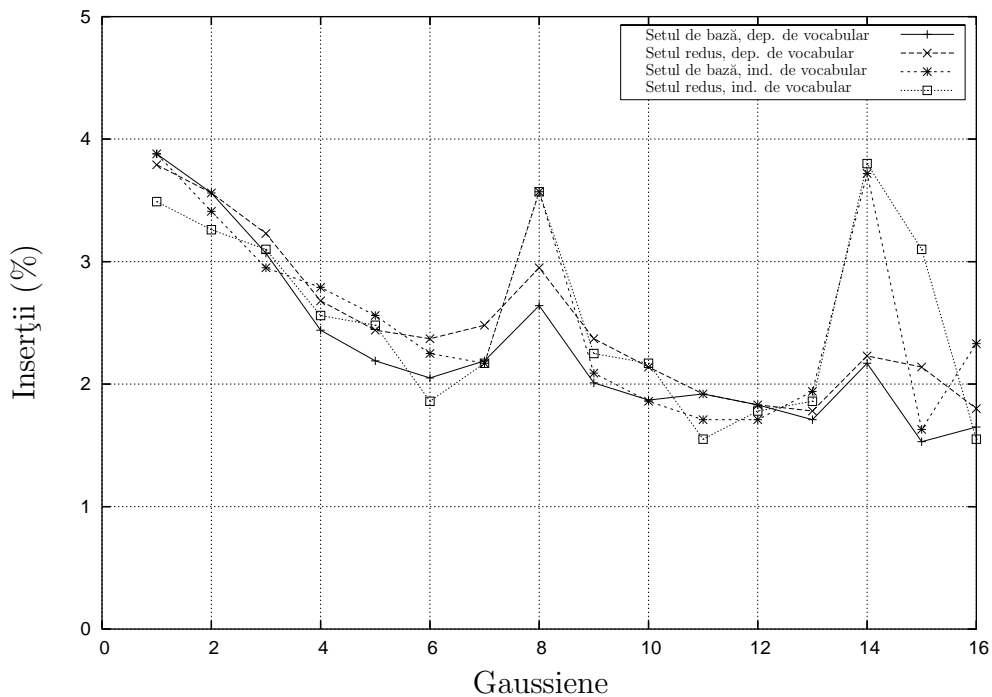


Figura 6.12: Evoluția frecvenței inserțiilor cuvintelor funcție de numărul de gaussiene/stare emițătoare

nesemnificative, singurele excepții, favorabile setului de bază, apărând în cazul testelor dependente de vocabular cu 5, 6 și 15 gaussiene/stare emițătoare și al celor independente de vocabular cu 10 și 15 gaussiene/stare emițătoare (tabelul 6.10).

Semnificația statistică redusă a diferențelor poate fi însă considerată normală dacă ținem cont și de frecvența redusă de realizare a fonemului /I/, observată pe durata etichetării semnalelor (capitolul 5) și confirmată și de analiza datelor de antrenament: din cele 774 apariții ale fonemului /I/, așteptate pe baza pronunțiilor din dicționar, doar 168 au fost efectiv realizate, corespunzător unei frecvențe de realizare de cca. 21,7%.

Această situație ar putea fi explicată prin faptul că informația lingvistică presupusă a fi transmisă prin intermediul fonemului /I/ este în realitate partajată între diferitele niveluri ale comunicării verbale (acustic, sintactic, semantic, pragmatic etc.), așa încât mesajele pot fi înțelese corect chiar și atunci când acest fonem nu este realizat.

Tabelul 6.10: Semnificația statistică la nivelul $p = 0,05$ a diferențelor dintre performanțele în recunoașterea cuvintelor folosind setul de bază și cel redus funcție de numărul de gaussiene și dependența 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 unități sublexicale ales în faza inițială a cercetărilor și folosit pentru proiectarea bazei de date construite pe parcursul lor, dar elementele noi apărute pe durata etichetării semnalelor și a experimentelor au indicat existența a două posibile seturi alternative de unități de modelare – unul incluzând consoane palatalizate și altul redus, obținut prin eliminarea /I/ – reflectând în fapt neclaritățile din literatura lingvistică referitoare la fonemele limbii române. Drept urmare, un al doilea obiectiv al experimentelor a fost compararea celor trei posibile seturi de unități de modelare, însă analiza datelor a arătat că acestea sunt suficiente doar pentru comparații între setul de bază și cel redus.

Prima comparație a fost făcută pe durata experimentelor dependente de vorbitor, când utilizarea setului de bază a fost însoțită de erori atribuite problemelor de modelare a fonemului /I/, diminuate prin eliminarea lui și utilizarea setului redus.

Compararea sistematică a celor două seturi, simultan cu evaluarea performanțelor la care conduc, a fost făcută printr-o serie de experimente independente de vorbitor. În plus, prin teste independente de vocabular, de recunoaștere a unor vocabulare semnificativ diferite de cel al datelor de antrenament, a fost evaluată și capacitatea de generalizare a modelelor acustice care să permită utilizarea lor în sisteme cu vocabulare flexibile.

Experimentele independente de vorbitor au fost desfășurate la nivelurile unităților de modelare și cuvintelor: la nivelul unităților de modelare, rezultatele au demonstrat o bună capacitate de generalizare a modelelor în raport cu schimbarea vocabularului, iar utilizarea setului redus a condus la performanțe superioare.

Comparațiile dintre cele două seturi de unități pe baza rezultatelor în recunoașterea cuvintelor arată diferențe în favoare setului de bază, manifestate în special independent de vocabular și la o modelare destul de detaliată (peste opt gaussiene/stare emițătoare), însă aceste diferențe au o semnificație statistică redusă.

Neclaritățile din literatura lingvistică, pe baza căreia a fost făcută alegerea inițială a unităților de modelare, au continuat deci și pe parcursul acestor experimente: înlocuirea setului de bază cu cel redus a condus fie la ameliorarea rezultatelor (cazul experimentelor dependente de vorbitor și al recunoașterii unităților de modelare), fie la reduceri ale performanțelor, în general nesemnificative statistic (cazul recunoașterii cuvintelor).

Cu toate neclaritățile încă persistente, experimentele au marcat primii pași spre soluționarea problemelor modelării acustice sublexicale pentru recunoașterea vorbirii continue cu vocabulare mari și foarte mari în limba română, iar rezultatele ca atare constituie un prim punct de referință pentru cercetările viitoare din domeniu.

CAPITOLUL 7

Încheiere

Această teză a descris primele cercetări (din câte cunoaștem) vizând recunoașterea automată a vorbirii continue în limba română, pentru care obiectivul principal ales inițial a fost recunoașterea 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 condițiile date, având în același timp avantajul de a fi semnificativ atât din punctul de vedere al dificultății problemei, cât și din cel al aplicabilității practice a rezultatelor – de exemplu, în sisteme de dialog vocal om-mașină orientate pe domenii bine definite, pentru care o asemenea dimensiune a vocabularului poate fi suficientă.

Metodele statistice de recunoaștere a vorbirii, care s-au impus pe plan mondial și au fost folosite și în aceste cercetări, presupun utilizarea unor modele acustice pentru a descrie evoluția în timp a proprietăților spectrale ale semnalelor vocale și a unor modele lingvistice ale succesiunii cuvintelor, iar aceste cercetări au fost axate pe problemele modelării acustice, fundamentală în raport cu cea lingvistică.

Pentru recunoașterea unui număr redus de cuvinte, modelele acustice pot fi construite separat pentru fiecare cuvânt, însă recunoașterea vorbirii continue cu vocabulare mari și foarte mari, de ordinul miilor sau zecilor de mii de cuvinte, impune utilizarea de modele acustice ale unor unități sublexicale (silabe, foneme etc.), în număr mult mai redus decât cel al cuvintelor din vocabular, antrenabile în consecință pe baza unor cantități mult mai reduse de semnale vocale, și utilizabile și în cazul schimbării vocabularului.

În principiu, cea mai economică și în același timp cea mai generală modelare acustică sublexicală ar putea fi realizată utilizând setul de foneme ale unei limbi, care sunt prin definiție categoriile abstracte ale sunetelor ei distinctive. În practică, diferențele dintre proprietățile spectrale ale alofonelor (variantele poziționale de realizare ale fonemelor), cauzate de fenomenul de coarticulație, pot conduce la utilizarea unor seturi alofonice de unități de modelare acustică sublexicală, diferite de cel al fonemelor.

În cazul limbii române, utilizarea directă a unui set de foneme sau alofone ca unități de modelare acustică sublexicală nu este posibilă din cel puțin două motive: pe de o parte, literatura lingvistică menționează mai multe seturi de foneme; pe de alta, insuficiența studiilor de fonetică și fonologie a limbii române, care ar fi permis clarificarea problemei setului de foneme și ar fi fost utile și pentru eventuale decizii de modelare alofonică.

În consecință, principala problemă investigată în această teză a fost cea a unităților de modelare acustică sublexicală, investigație din care au rezultat și contribuțiile ei.

7.1 Contribuții

Un prim grup de contribuții ale tezei sunt cele legate de proiectarea și colectarea bazei de date fonetice, prezentată în capitolul 4. Această bază de date este în sine o contribuție esențială, cea mai importantă pe termen lung: în primul rând prin cadrul pe care îl creează pentru experimente controlate de recunoaștere a vorbirii și comparații între rezultatele lor, necesare în cursul cercetărilor, și abia apoi prin materialul pe care îl furnizează pentru antrenarea modelelor acustice, testarea și evaluarea sistemelor. În plus, datorită proiectării corespunzătoare a conținutului lingvistic al unei părți importante a materialelor înregistrate și asigurarea unei foarte bune calități acustice a înregistrărilor, baza de date poate fi folosită și în cercetările fundamentale de fonetică și fonologie, a căror insuficiență a fost semnalată, precum și pentru alte cercetări aplicative. Contribuțiile legate de proiectarea bazei de date constau în:

- analiza literaturii lingvistice și definirea unui set de unități fonetice de modelare acustică (tabelul 4.1) astfel încât direct sau prin combinații acestea să acopere toate elementele diferitelor seturi de foneme ale limbii române din literatură, elemente considerate potențial 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 obținerea unor grupuri de materiale cât mai apropiate între ele din punctul de vedere al distribuțiilor unităților fonetice de modelare acustică;
- o metodologie sistematică de proiectare, bazată pe gruparea materialelor folosind algoritmul menționat și distribuirea lor pe vorbitori conform unui experiment cu o structură bloc aleatoare, utilizând ca variabile de blocare criteriile de selecție uniformă a vorbitorilor – în cazul nostru, sexul și grupa de vârstă.

Al doilea grup de contribuții este legat de etichetarea semnalelor vocale din baza de date, descrisă în capitolul 5. În cadrul cercetărilor asupra recunoașterii automate a vorbirii, etichetarea facilitează antrenarea modelelor acustice și asigură referințele pentru evaluarea experimentelor de recunoaștere la nivelul unităților de modelare. Datorită însă accesului rapid pe care îl permite la realizările acestor unități și informațiilor de durată pe care le include, ea este utilă pentru multe alte cercetări, iar împreună cu celelalte caracteristici ale bazei de date (controlul conținutului lingvistic și al vorbitorilor înregistrați și calitatea înregistrărilor) îi conferă acesteia caracterul de bază de date fonetice. Contribuțiile din acest grup includ:

- dezvoltarea unui sistem de etichetare a semnalelor vocale care combină transcrierea lor fonetică manuală și 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 identităților și limitelor segmentelor de semnal, utilizate pe durata etichetării manuale și/sau a verificării și corectării etichetelor generate automat folosind sistemul dezvoltat anterior;
- etichetarea, verificarea și corectarea etichetării majorității înregistrărilor.

Ultimul grup de contribuții este legat de utilizarea efectivă în modelarea acustică a unităților sublexicale, prezentată în capitolul 6. Experimentele de recunoaștere a vorbirii continue descrise în acest capitol au fost efectuate folosind subseturi ale bazei de date, iar obiectivul lor era inițial validarea setului de unități definit în etapa de proiectare a acesteia. Pe durata etichetării și a experimentelor a apărut însă problema reducerii setului de unități, motiv pentru care un al doilea obiectiv urmărit a fost compararea setului de bază cu cel redus. Pentru evaluarea capacității de generalizare a modelelor la schimbarea vocabularului, majoritatea experimentelor au urmărit și contrastul dintre dependența și independența de vocabular. Contribuțiile din acest grup cuprind:

- dezvoltarea unor sisteme automate de recunoaștere a vorbirii continue în limba română, dependente și independente de vorbitor, la nivel lexical și sublexical, care au fost utilizate pentru efectuarea experimentelor;
- studiul și evaluarea seturilor de bază și redus de unități fonetice de modelare acustică sublexicală pentru recunoașterea automată a vorbirii continue în limba română.

Experimentele nu au reușit să clarifice diferențele dintre setul de bază și cel redus, această clarificare rămânând în sarcina continuărilor acestor cercetări.

7.2 Continuări

Deși această teză cuprinde rezultatele câtorva ani de cercetări, ea nu a rezolvat toate problemele modelării acustice sublexicale pentru recunoașterea vorbirii continue în limba română, iar din experiența și rezultatele de până acum reies și câteva posibile continuări.

Datele colectate și utilizate în cursul acestor prime cercetări asupra recunoașterii automate a vorbirii continue în limba română au fost în mod necesar limitate cantitativ, accentul fiind pus pe calitatea lor acustică, astfel încât informația lingvistică din semnalul vocal să fie cât mai puțin afectată de zgomote, reverberații etc. Datorită acestei limitări, unele posibile alternative de modelare nu au putut fi abordate (secțiunea 6.3.2).

Chiar și așa, rezultatele obținute sugerează posibilitatea ca setul de unități fonetice de modelare acustică ales inițial (tabelul 4.1) să nu fie cel optim. O posibilă alternativă, care pe baza rezultatelor de până acum nu poate fi respinsă, este renunțarea la fonemul /I/: așa cum rezultă din analiza datelor, acesta are o frecvență de realizare redusă, ceea ce ar putea indica faptul că informația 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 și lingvistice. Datorită limitării menționate a datelor, încercările de utilizare în acest scop a unor modele statistice de tip bigram sau a unor gramatici de tip perechi-de-cuvinte (secțiunea 6.1.1) estimate pe baza

lor a dus la rezultate irelevante: date fiind perplexitățile foarte reduse ale acestor modele lingvistice, chiar cu modelele acustice cele mai simple, cu o gaussiană/stare emițătoare, frecvența recunoașterilor corecte și acuratețea s-au situat peste 95%. Deși în aparență foarte bune, asemenea rezultate sunt inutile deoarece nu permit comparații semnificative între diferite condiții testate, și în consecință nici nu au fost prezentate aici.

Studiul simultan al problemelor modelării acustice și lingvistice presupune însă un efort preliminar pentru colectarea coordonată a textelor și semnalelor vocale necesare, de exemplu plecând de la surse de texte disponibile pe Internet [23].

Date fiind dimensiunile considerabile ale unui asemenea efort, direcția de continuare a cercetărilor cea mai accesibilă pentru moment este cea a modelării acustice dependente de context, care pe lângă certe ameliorări ale performanțelor ar putea aduce contribuții și la clarificarea problemei setului optim de unități de modelare.

Din experiența acumulată se degajă și unele probleme metodologice, legate de analiza și diagnoza antrenării modelelor și recunoașterii și evaluarea performanțelor. Studiul procesului de antrenare, de exemplu pentru a putea explica abaterile de la tendința de ameliorare a performanțelor din cazurile recunoașterii cuvintelor cu MMA având 8 și 14 gaussiene/stare emițătoare (secțiunea 6.4.4), ar putea fi facilitat prin analiza și vizualizarea interactivă a evoluției modelelor și a datelor de antrenament.

Examinând exemplele de recunoaștere din anexa B, se constată că în numeroase cazuri erorile apar grupate și nu sunt independente între ele: datorită constrângerilor reduse impuse în cazul lor de dicționarele de pronunții, cuvintele scurte sunt cele mai frecvent omise sau inserate, iar omisiunile și inserțiile la rândul lor sunt adesea însoțite de substituții. Analiza unor asemenea situații ținând cont și de pronunțiile cuvintelor și similaritatea lor arată că de fapt pe lângă substituții, omisiuni și inserții există și erori de divizare sau contopire a cuvintelor. Identificarea acestora și modificarea corespunzătoare a procesului și metricilor de evaluare a performanțelor s-ar putea face recurgând la alinierea fonetico-fonologică a ipotezelor cu referințele [188], [187], [72].

În sfârșit, diagnoza recunoașterii în sine necesită recurgerea la metode de analiză a zonelor eronate, cu posibilitatea de clasificare și stabilire automată a cauzelor lor [47].

ANEXA A

Detalii ale dicționarelor

Utilizând setul de bază de unități de modelare acustică, au fost identificate următoarele perechi de cuvinte homofone:

- în datele pentru teste dependente de vocabular

Cuvânt	Homofon	Pronunție
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	Pronunție
ce-i	cei	C e j
s-a	sa	s a

Prin înlocuirea setului de bază cu cel redus de unități de modelare s-au mai adăugat câteva perechi de homofone:

- în datele pentru teste dependente de vocabular

Cuvânt	Homofon	Pronunție
bețivan	bețivani	b e T i v a n

- în datele pentru teste independente de vocabular

Cuvânt	Homofon	Pronunție
a-și	aș	a S
bun	buni	b u n
m-ar	mari	m a r

ANEXA B

Exemple de recunoaștere

În continuare sunt prezentate rezultatele recunoașterii independente de vocabular a cuvintelor folosind modele acustice cu 16 densități gaussiene/stare. Fiecare dintre cele 133 propoziții individuale folosite este identificată prin codul vorbitorului, cel al materialului citit și numărul ei de ordine în acest material – de exemplu, FA-AZ1 este prima dintre propozițiile înregistrate de vorbitorul FA prin citirea materialului AZ.

Pentru fiecare propoziție este prezentată transcrierea de referință aliniată cu ipotezele obținute folosind setul de bază și cel redus de unități de modelare acustică sublexicală. Nu au fost considerate erori substituțiile unor cuvinte homofone ale celor de referință. În rest, substituțiile și inserțiile sunt *evidențiate*, iar omisiunile – marcate prin ****.

Ref. FA-AZ1: nu se teme să-și mărturisească slăbiciune după slăbiciune

Setul de bază: nu *să* teme să-și mărturisească slăbiciune după slăbiciune

Setul redus: nu *să* teme să-și mărturisească slăbiciune după slăbiciune

Ref. FA-AZ2: mai văzusem așa ceva în niște planșe reprezentând anatomia subtilă

Setul de bază: mai văzusem așa ceva în niște planșe reprezentând anatomia subtilă

Setul redus: mai văzusem așa ceva *îmi* niște planșe reprezentând anatomia subtilă
conform yogăi
conform yogăi
conform yogăi

Ref. FA-AZ3: și zise că n-avea nici el vreo lețcaie

Setul de bază: și zise că n-avea nici *și-a* vreo lețcaie

Setul redus: și zise că n-avea nici *și-a* vreo lețcaie

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 preț decât să cunosc cartea

Setul de bază: *urma* astăzi *m-am gâfâit m-a* bun mai *vă* preț decât să cunosc cartea

Setul redus: *urma* astăzi *m-am gâfâit m-a* bun mai *vă* preț 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: circumstanțele existenței noastre sunt socotite de noi

Setul de bază: circumstanțele *insist în ți* *noastră* sunt socotite ** *numai*

Setul redus: circumstanțele *insist îmi ți* *noastră să-mi* socotite ** *numai*
toți
toți
toți

Ref. FB-AB1: o gravă acuzație era aceea de subminare morală a statului

Setul de bază: o gravă acuzație era ***** *cere* subminare morală a statului

Setul redus: o gravă acuzație era ***** *cere* subminare morală a statului

Ref. FB-AB2: câți dintre dâșii vorbeau măcar la nivelul unor discuții de salon

Setul de bază: *că stâncă* dâșii vorbeau măcar la nivelul unor discuții de salon

Setul redus: *că stâncă* dâșii vorbeau măcar la nivelul unor discuții de salon

Ref. FB-AB3: ghemul timpului nu s-a desfășurat în întregime nădăjduiesc

Setul de bază: ghemul timpului nu sa desfășurat ** întregime nădăjduiesc

Setul redus: ghemul timpului nu sa desfășurat ** întregime *văd întins*

Ref. FB-AB4: cu cât înaintez le văd mai mari și goana lor mi se pare mai

Setul de bază: cu *către* înaintez ** *** *revăd m-am a-și* goana lor mi *să* pare mai

Setul redus: cu *către* înaintez ** *** *revăd m-am a-și* goana lor mi *să* pare mai
ciudată
ciudată
ciudată

Ref. FB-AB5: se opri acolo unde îndeobște n-aveai voie s-o faci

Setul de bază: se opri acolo unde îndeobște *m-a pe* voie s-o faci

Setul redus: se opri acolo unde îndeobște *m-a te* voie s-o faci

Ref. FB-AB6: pentru a-și desăvârși penitența poetul refuza să i se mai

Setul de bază: *pe haz* desăvârși penitența poetul refuza *să-i* să **** ** mai

Setul redus: *pe crezi* desăvârși penitența poetul refuza ** *să-i* se mai
încălzească odaia
încălzească odaia
încălzească odaia

Ref. FC-AT1: statuia închipuia un subiect cu care sculptorii erau mai puțin hârșâiți

Setul de bază: statuia închipuia *au* subiect cu care sculptorii erau *m-a* puțin hârșâiți

Setul redus: statuia închipuia *au* subiect cu care sculptorii erau *m-a* puțin hârșâiți

Ref. FC-AT2: intenția spectacolului oferit de el era să-și cruțe soața și s-o rupă de

Setul de bază: intenția spectacolului oferit *doi ele* era să-și cruțe soața și s-o rupă de

Setul redus: intenția spectacolului oferit *doi ele* era să-și cruțe soața și s-o rupă de
soarta lui neagră
soarta lui neagră
soarta lui neagră

Ref. FC-AT3: nicăieri nu se înnegresc de dorința răzbunării

Setul de bază: nicăieri nu se înnegresc de dorința răzbunării

Setul redus: nicăieri nu se înnegresc de dorința răzbunării

Ref. FC-AT4: tudor mă urmărise abia stăpânindu-se să nu izbucnească-n răs

Setul de bază: tudor *m-a* urmărise abia stăpânindu-se ** *seama* izbucnească-n răs

Setul redus: tudor *m-a* urmărise abia stăpânindu-se să nu izbucnească-n răs

Ref. FC-AT5: la școală a izbucnit un adevărat scandal

Setul de bază: *era* școală *pare* izbucnit un adevărat scandal

Setul redus: la școală *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 șoptește că putea fi și mai rău

Setul de bază: orgoliul *în* șoptește că putea fi și *m-ar ou*

Setul redus: orgoliul *în* șoptește că putea fi *și-mi aer ou*

Ref. FD-AW2: ziaristul se schimonosi încercând să-și ascundă răsul

Setul de bază: ziaristul se schimonosi *pe* încercând să-și ascundă răsul

Setul redus: ziaristul se schimonosi încercând să-și ascundă răsul

Ref. FD-AW3: m-am ridicat într-o rână am înșfăcat sticla de apă de lângă pat

Setul de bază: m-am ridicat într-o rână *apară-n* înșfăcat sticla de apă de lângă pat

Setul redus: m-am ridicat într-o rână am înșfăcat sticla de apă de lângă pat

și am sorbit

pe și am sorbit

și am sorbit

Ref. FD-AW4: de astfel de vorbe se și temea duhovnicul de la putna

Setul de bază: de astfel de vorbe se și temea duhovnicul *de-l a* putna

Setul redus: de astfel de vorbe se și temea duhovnicul *de-l a* putna

Ref. FD-AW5: congestionat și grețos intenționează să se arate amenințător

Setul de bază: congestionat și grețos intenționează să *să* arate amenințător

Setul redus: congestionat și grețos intenționează să *să* arate amenințător

Ref. FD-AW6: dacă ieșeam împreună să-mi cumpere ciorapi sau o cămașă

Setul de bază: *dată* ieșeam împreună *ar* să-mi cumpere ciorapi *** *său* cămașă

Setul redus: dacă ieșeam împreună *ar* să-mi cumpere ciorapi *** *său* cămașă

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 toți cei de acolo se grăbeau să-și dezică viețile

Setul de bază: aproape toți ce-i de acolo se grăbeau să-și dezică viețile

Setul redus: aproape toți ce-i de acolo se grăbeau să-și dezică viețile

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 înmulțiseră ca niciodată urticariile de tot soiul

Setul de bază: *de* sigur că se înmulțiseră ca niciodată urticariile de *toți* soiul

Setul redus: *îi* sigur ** *câți* înmulțiseră ca niciodată urticariile de *toți* soiul

Ref. FE-BB4: tot ce ne aduce imaginația e fals

Setul de bază: tot ce ne aduce imaginația *i* fals

Setul redus: tot ce ne aduce imaginația *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 înveșmântat în negru așezat pe un scaun

Setul de bază: îl revăd acolo înveșmântat *e* negru așezat pe un scaun

Setul redus: îl revăd acolo înveșmântat *e* negru așezat *te* un scaun

Ref. FF-AG1: sfârșitul cărții de față sosea grabnic

Setul de bază: sfârșitul cărții de față sosea *dat mic*

Setul redus: sfârșitul cărții de față sosea *dat mic*

Ref. FF-AG2: în redactare se bizuia pe niște note luate chiar după ce se

Setul de bază: *până* redactare se bizuia pe niște note luate chiar după ce se

Setul redus: *până* redactare se bizuia pe niște note luate chiar după ce se

despărțise de gazda sa

despărțise de gazda sa

despărțise de gazda sa

Ref. FF-AG3: sperțul ocazionă întunecarea nădejdilor

Setul de bază: sperțul ocazionă întunecarea nădejdilor

Setul redus: sperțul 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 dorința de a istorisi mereu aceleași întâmplări cu un haz reînnoit

Setul de bază: *ea* dorința de * istorisi mereu aceleași întâmplări ** *cum* haz reînnoit

Setul redus: *ea* dorința de * istorisi mereu aceleași întâmplări ** *cum* haz reînnoit

Ref. FG-BI2: tata s-ar fi înfuriat cu siguranță și poate mă pocnea

Setul de bază: tata *să* fi *vrea* cu siguranță și poate *m-a* pocnea

Setul redus: tata *să* *fiu fu ea* cu siguranță și poate *m-a* pocnea

Ref. FG-BI3: se pomenise preluând pe neașteptate cea mai dificilă misiune

Setul de bază: se pomenise preluând *te* neașteptate cea mai dificilă misiune

Setul redus: se pomenise preluând *te* neașteptate cea mai dificilă misiune

Ref. FG-BI4: schiță o figură de gimnastică să se dezmoștească după trezire

Setul de bază: schiță *au* figură de gimnastică să se dezmoștească după trezire

Setul redus: schiță *au* figură de gimnastică să se dezmoștească 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
înțeleagă jocul
înțeleagă jocul
înțeleagă jocul

Ref. FG-BI6: impulsurile pe care le resimțam îmi depășeau puterile de stăpânire de

Setul de bază: impulsurile pe care le resimțam îmi depășeau puterile de stăpânire de

Setul redus: impulsurile *te* care le resimțam îmi depășeau 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 aștept
 Setul de bază: mă pofti să-l aștept
 Setul redus: mă pofti să-l aștept

Ref. FH-AY4: sunt sigur că-i era greu cu un nătăfleață ca mine
 Setul de bază: **** *sus ion* era greu cu ** nătăfleață ca mine
 Setul redus: **** *sus ion* era greu cu ** nătăfleață ca mine

Ref. FH-AY5: pleoapele ți le ștergeai întruna de stropii de ulei țâșniți până la ele
 Setul de bază: pleoapele ți *de* ștergeai întruna de stropii de ulei țâșniți până *lui* ele
 Setul redus: pleoapele ți *de* ștergeai întruna de stropii de ulei țâșniți până *lui* ele

Ref. FH-AY6: din când în când își freca pleoapele până le roșea
 Setul de bază: din când ** *încât* își freca *pleoape* până le roșea
 Setul redus: din când ** *încât* își freca *pleoape* până le roșea

Ref. FH-AY7: îngrija în chilia sa vreo optsprezece canari
 Setul de bază: îngrija *îmi* chilia sa *văd* optsprezece canari
 Setul redus: îngrija 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 așa de îndușmănit
 Setul de bază: ** *ne* mai *vine se* primească altul care ** *nu-i* așa de îndușmănit
 Setul redus: ** *ne* mai *vine se* primească altul care ** *nu-i* așa de îndușmănit
 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 ** ** *înaintearea* bisericii
 Setul redus: stătea ** spatele tejghelei cu lumânări ** ** *înaintearea* bisericii

umbroase
 umbroase
 umbroase

Ref. FI-BZ5: purta veșminte albe întreșesute cu fir de aur
 Setul de bază: purta veșminte albe întreșesute cu fir de aur
 Setul redus: purta veșminte albe întreșesute 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 internațional acesta angajându-l
 Setul de bază: acestea l-au impus *îi* for internațional acesta angajându-l
 Setul redus: acestea l-au impus *îi* for internațional acesta angajându-l

Ref. FJ-BC2: era de dorit să ne doboare în număr cât mai mare
 Setul de bază: *ea* de *doi* *ți* ne doboare ** număr *câți* *mă* mare
 Setul redus: *ea* de *doi* *se* ne doboare ** număr *câți* *mă* mare

Ref. FJ-BC3: iarăși simțea nevoia să se ducă la toaletă
 Setul de bază: *iară* simțea *ne vă* să se ducă la toaletă
 Setul redus: iarăși simțea *ne* *vă sus* ducă la toaletă

Ref. FJ-BC4: poate că ar trebui să vă adresați altcuiva
 Setul de bază: poate ** *ca* trebui să vă adresați altcuiva
 Setul redus: poate ** *ca* trebui să vă adresați 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 național ales a-i fi înmănat însăși reginei
 Setul de bază: a lucrat *în* costum național **** *alese* fi înmănat însăși reginei
 Setul redus: a lucrat *îmi* costum național **** *alese* fi înmănat însăși reginei

Ref. FJ-BC7: vasele de porțelan se zdrobeau de parchet rămânând fără viață
 Setul de bază: vasele de porțelan se zdrobeau de parchet *pe* rămânând fără viață
 Setul redus: vasele de porțelan se zdrobeau de parchet rămânând fără viață

Ref. MA-AX1: făcusem această alegere după matură chibzuință
 Setul de bază: făcusem această alegere *vă* matură chibzuință
 Setul redus: făcusem această alegere *vă* matură chibzuință

Ref. MA-AX2: când îl zăreai îți dădeai cu părerea că ar fi fost un om
 Setul de bază: când *e* zăreai îți dădeai cu părerea ** ** *cărții* fost ** *m-am*
 Setul redus: când *e* zăreai îți dădeai cu părerea ** ** *cărții* fost ** *m-am*
 cumsecade
 cumsecade
 cumsecade

Ref. MA-AX3: exasperat a apucat să-i spună că nu părăsește cabinetul fără
 Setul de bază: exasperat * apucat *se* spună că nu părăsește cabinetul fără
 Setul redus: exasperat * apucat *se* spună ** *cum* părăsește 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 și blândețe
 Setul de bază: a *trebui se* insist mult cu răbdare și blândețe
 Setul redus: * *atrag uiți* insist mult cu răbdare și blândețe

Ref. MA-AX6: iară eu beau vin și țuică
 Setul de bază: iară eu beau *vine* și țuică
 Setul redus: iară eu beau vin și țuică

Ref. MA-AX7: mi-a venit să pufnesc în răs
 Setul de bază: **** **** *devenise* pufnesc ** *ales*
 Setul redus: **** **** *devenise* pufnesc ** *ales*

Ref. MB-AC1: apariția unui profesor însoțitor le cenzurează instinctele dezlănțuite
 Setul de bază: apariția unui profesor însoțitor le cenzurează *e stâncă* dezlănțuite
 Setul redus: apariția unui profesor însoțitor le cenzurează *e stâncă* dezlănțuite

Ref. MB-AC2: în cele din urmă m-a încredințat că dacă se elibera postul urma
 Setul de bază: *un cele ghemul om a* încredințat că dacă se elibera postul urma
 Setul redus: *un cele ghemul om a* încredințat 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 stăpîinii s-o slăbească pentru a o putea manipula în
 Setul de bază: ** *mi-a* vor stăpîinii s-o slăbească **** pe *într-o* putea manipula în
 Setul redus: ** *mi-a* vor stăpîinii s-o slăbească **** *te într-o* putea manipula în
 interesul lor
 interesul lor
 interesul lor

Ref. MB-AC4: există și alte slujbe acolo

Setul de bază: există și alte slujbe acolo

Setul redus: există și alte slujbe acolo

Ref. MB-AC5: și astfel odiseea luă sfârșit

Setul de bază: și *noastră* odiseea luă sfârșit

Setul redus: și *noastră* odiseea *nu* sfârșit

Ref. MB-AC6: murise cam pe la optsprezece ani și-mi devenise model

Setul de bază: murise cam ** *până* optsprezece *al* și-mi devenise model

Setul redus: murise cam ** *părea* optsprezece ani și-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 simțeam nițel neliniștit de zâmbetul său enigmatic

Setul de bază: doar eu mă *simțea* nițel neliniștit de zâmbetul *s-o* enigmatic

Setul redus: doar eu mă *simțea* nițel neliniștit de zâmbetul *s-o* enigmatic

Ref. MC-AA4: și-a luat lădițe goale în juru-i

Setul de bază: și-a luat lădițe *boală* în juru-i

Setul redus: și-a luat lădițe *boală* în juru-i

Ref. MC-AA5: ieși pe șleaul ce ducea drept spre drumul de la sulița la botoșani

Setul de bază: ieși pe șleaul ce ducea drept spre drumul *de-l a* sulița la botoșani

Setul redus: ieși pe șleaul ce ducea drept spre drumul *de-l a* sulița la botoșani

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

Setul de bază: hotărâsem să mă inspir din numele său *un* *alegere*

Setul redus: hotărâsem *să-mi îi* inspir din numele *** *sub* *alegere*

pseudonimului meu

pseudonimului *i* meu

pseudonimului *i* meu

Ref. MD-AK2: își luă măsurile de prevedere esențial era să salveze pentru

Setul de bază: își *o* măsurile de prevedere *pe* esențial era să salveze pentru

Setul redus: își *l-au* măsurile de prevedere esențial era să salveze pentru

viitorime comoara de documente

viitorime comoara de documente

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 aceeași
 Setul de bază: *i rău bun bec* *circulă* bucurii cu care *să-l* rugau *un* aceeași
 Setul redus: *i rău bun bec* *circulă* bucurii cu care *să-l* rugau *un acești*
 obște parohială
 obște *planul iară*
 obște *planul iară*

Ref. MD-AK5: ion suferă de o boală cronică și mortală
 Setul de bază: ion suferă ** * *doboare* cronică și mortală
 Setul redus: ion suferă ** * *doboare* cronică și 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 securității
 Setul de bază: numai că respectivul student francez era în stipendiul *îmi* securității
 Setul redus: numai că respectivul student francez era în stipendiul securității

Ref. ME-BN2: văzusem o spinare sumețindu-se de sub fiarele unui pat
 Setul de bază: văzusem o spinare *tălpi* sumețindu-se de *și-mi* fiarele unui pat
 Setul redus: văzusem o spinare sumețindu-se de *și-mi* fiarele unui pat

Ref. ME-BN3: acum ni se adresă cu o temă vrednică de inteligența noastră
 Setul de bază: acum ** se adresă cu o temă vrednică de inteligența noastră
 Setul redus: acum ** se adresă cu o temă vrednică de inteligența noastră

Ref. ME-BN4: șleahta nemaifiind un avanpost al societății a dispărut
 Setul de bază: șleahta *pe* nemaifiind *putna* avanpost *ar* societății *te* dispărut
 Setul redus: șleahta nemaifiind *una* avanpost *ar* societății *trei* dispărut

Ref. ME-BN5: starețul părintele arsenie mi se înfățișă ca o siluetă fără
 Setul de bază: starețul *îmi* părintele arsenie *plin* se înfățișă ** *planul* siluetă fără
 Setul redus: starețul părintele arsenie mi se înfățișă *că aur* siluetă fără
 pereche
 pereche
 pereche

Ref. ME-BN6: am constatat seninul bunătații ce-i emana dintre pleoape
 Setul de bază: am constatat seninul bunătații ce-i emana dintre pleoape
 Setul redus: am constatat seninul bunătații ce-i emana dintre pleoape

Ref. ME-BN7: am socotit visul de mai sus ca o șoaptă încurajatoare
 Setul de bază: *ani* socotit *i* visul de *m-ar și uiți pe au* șoaptă încurajatoare
 Setul redus: *ani* socotit *i* visul de *m-ar* sus ** *rău* șoaptă încurajatoare
 a subconștientului
aș subconștientului
a-și subconștientului

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 obținut
 Setul de bază: guvernul ** *număr a s-o* înlocuiască după ** *ceva* fi obținut
 Setul redus: guvernul ** *număr a s-o* înlocuiască după ** *ceva* fi obținut
 recunoașterea
 recunoașterea
 recunoașterea

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 belșugul dragostei de frați
 Setul de bază: *o* trăit-o ca *pare* cu belșugul dragostei de frați
 Setul redus: *o* trăit-o ca *pare* cu belșugul dragostei de frați

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 lovească*
 Setul redus: caut un pamflet ** *cele* apărut în evenimentul ** *lumea* cu
 doi trei ani
 doi trei *ar*
 doi trei *ar*

Ref. MG-AH3: atrage atenția asupra greutăților întâmpinate în viață
 Setul de bază: atrage atenția *stau* greutăților întâmpinate *lui vrea te*
 Setul redus: atrage atenția *stau* greutăților întâmpinate *lui* viață

Ref. MG-AH4: ghemuit în patru labe se sprijini anevoie de bara de fier și-și scoase
 Setul de bază: ghemuit în *patul* labe se sprijini anevoie de bara de fier *și* scoase
 Setul redus: ghemuit în *patul* labe se sprijini anevoie de bara de fier și-și 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 ținut seama de dorința lui se risca provocarea
 Setul de bază: dacă ** *** *râs oficial* seama de dorința lui se risca provocarea
 Setul redus: dacă ** *** *râs oficial* seama de dorința 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 pușca
 Setul de bază: anghel * ***** *astăzi* mâinile soldatului pușca
 Setul redus: anghel * ***** *astăzi* mâinile soldatului pușca

Ref. MH-AR4: în interiorul cărții se desfăcea o planșă ce reprezenta trupul
 Setul de bază: *părintele o* cărții se desfăcea * *planșe* ce reprezenta trupul
 Setul redus: *părintele o* cărții se desfăcea * *planșe* ce reprezenta trupul
 omului
 ou vor
 ou vor

Ref. MH-AR5: trebuia să te uiți la vârful picioarelor sau în bec
 Setul de bază: trebuia *se* te uiți la vârful picioarelor sau *până* bec
 Setul redus: trebuia *se* te uiți la vârful picioarelor sau în bec

Ref. MH-AR6: își lipește din nou uităturile în același timp grele și pehlivane
 Setul de bază: își lipește din *** uităturile *până* același timp grele și pehlivane
 Setul redus: își lipește din *** uităturile în același timp grele și 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 știam cât erau de urâți acești ortodocși
 Setul de bază: m-am îngrijorat deoarece știam *pripit* erau de urâți acești ortodocși
 Setul redus: m-am îngrijorat deoarece știam *te te* erau de urâți acești ortodocși
 care nu renunțau
 care nu renunțau
 care nu renunțau

Ref. MI-BA3: mă zguduie fiori de emoție cu mult timp înainte de întâlnire
 Setul de bază: mă zguduie fiori de emoție cu **** timp înainte de întâlnire
 Setul redus: mă zguduie fiori de emoție cu **** timp înainte de întâlnire

Ref. MI-BA4: muștrări și sfaturi legate de frecvența lui se îngrămădeau la gura mea
 Setul de bază: muștrări și sfaturi legate de frecvența lui se îngrămădeau la gura *mă*
 Setul redus: muștrări și sfaturi legate de frecvența lui se îngrămădeau la gura *mă*

Ref. MI-BA5: am insistat continuând a crede în cinstea sa
 Setul de bază: am insistat *puțin 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 șchioapă
 împiedicată *optsprezece apă*
 împiedicată șchioapă

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*
 acuzația
acuzație
acuzație

Ref. MJ-BE2: pleoapele îi acoperă parțial ochii mari albaștri
Setul de bază: ***** *pleoape* acoperă parțial ochii mari albaștri
Setul redus: ***** *pleoape* acoperă parțial ochii *mare* albaștri

Ref. MJ-BE3: stătea singur pe stâncă îmbrățișând cerul cu brațele
Setul de bază: stătea singur pe stâncă îmbrățișând cerul cu *gura fără*
Setul redus: stătea singur *te* stâncă îmbrățișând cerul cu *dată*

Ref. MJ-BE4: dacă nu m-ar fi stăpânit amortizarea vodcii aş fi făcut multe prostii
Setul de bază: dacă ** *număr* fi stăpânit amortizarea vodcii aş ** făcut multe prostii
Setul redus: dacă ** *număr* fi stăpânit amortizarea vodcii a-și *fu cu* multe prostii

Ref. MJ-BE5: am luat o gură de apă să-mi potolesc arșița gâtlejului uscat
Setul de bază: ** **** *arată gura* de apă să-mi potolesc arșița gâtlejului uscat
Setul redus: ** *arată un gura* de apă să-mi potolesc arșița gâtlejului uscat

Ref. MJ-BE6: au fost transferați împreună cu un grup de delincvenți de drept comun
Setul de bază: *o* fost transferați *întruna* cu ** **** de delincvenți de drept comun
Setul redus: *o* fost transferați *întruna* cu ** **** de delincvenți 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 și 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 și 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 și 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 și 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 și 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 și 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 ș.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 și 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 și 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 și 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 și 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 și A.J. Fourcin. Levels of labelling. *Computer Speech and Language*, 6(1):1–14, ianuarie 1992.
- [19] L.E. Baum și 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 și N. Weiss. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *Annals of Mathematical Statistics*, 41(1):164–71, februarie 1970.
- [21] R. Bellman și S. Dreyfus. *Programarea dinamică aplicată*. Editura Tehnică, București, 1967.
- [22] L.L. Beranek. *Acoustic Measurements*. Wiley, New York, 1949.
- [23] D. Bohuş și **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 Timișoara*, vol. 40(54) din *Seria Automatică și Calculatoare*, pag. 135–38. Timișoara, 1995.
- [26] M. Boldea. Speaker Independent Isolated Word Recognition Experiments. În *Buletinul Ştiinţific al Universităţii Tehnice din Timișoara*, vol. 40(54) din *Seria Automatică și Calculatoare*, pag. 129–34. Timișoara, 1995.
- [27] M. Boldea. A Comparison of Speech Processing Methods in Speaker Independent Isolated Word Recognition. În *Buletinul Ştiinţific al Universităţii "Politehnica" din Timișoara*, vol. 41(55) din *Seria Automatică și Calculatoare*, pag. 164–70. Timișoara, 1996.
- [28] M. Boldea. Speech Technology Research at Computer Science Department, "Politehnica" University of Timișoara. 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 Timișoara, 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, Timișoara, 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 și E. Vives. Study of Subword Units for Spanish Speech Recognition. În *Proceedings EUROSPEECH'95*, pag. 1607–10, Madrid, septembrie 1995.
- [36] H. Bourlard și 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 și J.S. Hunter. *Statistics for Experimenters*. John Wiley & Sons, New York, 1978.
- [38] J.S. Bridle, M.D. Brown și 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 și M. Omologo. Automatic segmentation and labeling of speech based on Hidden Markov Models. *Speech Communication*, 12(4):357–70, august 1993.
- [41] F. Brugnara și 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 și 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 pronunțate izolat în vederea recunoașterii automate. În M. Drăgănescu și C. Burileanu (editori), *Analiza și sinteza semnalului vocal*, pag. 36–131. Editura Academiei Române, București, 1986.
- [44] W. Byrne, P. Byerlein ș.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 și 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 ș.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 și 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 și 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 ș.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 și M. Weintraub. Studies for an Adaptive Recognition Lexicon. În *Proceedings of the DARPA Speech Recognition Workshop*, San Diego, California, 1987.
- [52] M. Constantinescu și D. Cristescu. Sistem de analiză și recunoaștere automată a vorbirii. În M. Drăgănescu și C. Burileanu (editori), *Analiza și sinteza semnalului vocal*, pag. 210–20. Editura Academiei Române, București, 1986.
- [53] P. Cosi, D. Falavigna, G.A. Mian și M. Omologo. A Comparison between Mel-scale Cepstrum and Auditory Model Representation for Noisy Speech Recognition. În L. Torres, E. Masgrau și 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 și L. Rabiner. On the Applications of Multimedia Processing to Telecommunications. *Proceedings of the IEEE*, 86(5):755–824, mai 1998.
- [55] K. Croot și 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 și 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 și J.H.L. Hansen. *Discrete-Time Processing of Speech Signals*. Macmillan Publishing Company, New York, 1993.
- [60] A.P. Dempster, N.M. Laird și 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 Suppressive 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 și C. Burileanu (editori), *Analiza și sinteza semnalului vocal*, pag. 9–16. Editura Academiei Române, București, 1986.
- [64] R.O. Duda și P.E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.
- [65] T. Dumitrescu. Elemente de sinteza vorbirii în limba română. Dizertație de studii aprofundate, Departamentul de Calculatoare, Universitatea "Politehnica" din Timișoara, 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 și M. Omologo. A DTW-based Approach to the Automatic Labeling of Speech According to the Phonetic Transcription. În L. Torres, E. Masgrau și 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 ediție, 1970.
- [70] M. Federico, M. Cettolo, F. Brugnara și G. Antoniol. Language modelling for efficient beam-search. *Computer Speech and Language*, 9(4):353–79, 1995.
- [71] W.M. Fisher, G.R. Doddington și K.M. Goudie-Marshall. 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 și 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. Aplicații în industrie și telecomunicații*. Editura Mirton, Timișoara, 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 și 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 și 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 și M. Adda-Decker. Speaker-independent continuous speech dictation. *Speech Communication*, 15(1):21–37, 1994.
- [83] J.L. Gauvain și 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 și 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 și L. Lamel. Speaker-Independent Phone Recognition Using BREF. În *Proceedings DARPA Workshop on Automatic Speech Recognition*, 1992.
- [86] J.L. Gauvin, L.F. Lamel și 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 și K. Zeger. Vector Quantization Techniques in Speech Coding. În S. Furui și 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 și M.M. Sondhi (editori), *Advances in Speech Signal Processing*, cap. 15, pag. 453–85. Marcel Dekker, New York, 1992.
- [90] D. Gibbon, R. Moore și R. Winski (editori). *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, Berlin, 1997.
- [91] L. Gillick și 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 și 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. Tufiș și P. Andersen (editori), *Recent Advances in Romanian Language Technology*, pag. 178–87. Editura Academiei Române, București, 1997.
- [94] J.J. Godfrey, E.C. Holliman și 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 și M. Zirra. Neural Network Vowel Recognition in Romanian Language. În *Proceedings Second International Conference on Technical Informatics – CONTI'96*, pag. 165–72, Timișoara, 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** și C. Bărbulescu. Recunoașterea vocalelor cu ajutorul unui microsistem de calcul. În *Buletinul sesiunii științifice pentru tineret "Tehnic 2000"*, pag. 274–77, Timișoara, aprilie 1984.
- [101] V.N. Gupta, M. Lennig și 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 și 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 și 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 și M.A. Jack. *Hidden Markov Models for Speech Recognition*. Edinburgh University Press, 1990.
- [106] X.D. Huang și 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. Ioniță, C. Burileanu și M. Ioniță. 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 și 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 și 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 și R.L. Mercer. Interpolated Estimation of Markov Source Parameters from Sparse Data. În E.S. Gelsema și L.N. Kanal (editori), *Pattern Recognition in Practice*, pag. 381–97. North-Holland, 1980.
- [118] F. Jelinek, R.L. Mercer și S. Roukos. Principles of Lexical Language Modeling for Speech Recognition. În S. Furui și 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 și 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 și 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 și 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 și 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 și 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 și 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 și 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 și A.K. Foldvik. The multifarious r-sound. În *Proceedings International Conference on Spoken Language Processing*, pag. 1259–62, 1992.
- [131] R. Lacouture și R. De Mori. Lexical Tree Compression. În *Proceedings EURO-SPEECH'91*, vol. 1, pag. 581–84, 1991.
- [132] L.F. Lamel, R.H. Kassel și S. Seneff. Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus. În *Proceedings DARPA Speech Recognition Workshop*, 1986.

- [133] T. Lander și S.T. Metzler. The CSLU Labeling Guide. Center for Spoken Language Understanding, Oregon Graduate Institute, februarie 1994.
- [134] J. Lazzaro ș.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 și F. Alleva. Continuous Speech Recognition. În S. Furui și M.M. Sondhi (editori), *Advances in Speech Signal Processing*, cap. 20, pag. 623–50. Marcel Dekker, New York, 1992.
- [137] K.F. Lee și 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. Leggetter și 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 și 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 și 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 și 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 și 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 și 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 și 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 și H. Gish. Vector Quantization in Speech Coding. *Proceedings of the IEEE*, 73(11):1551–88, noiembrie 1985.
- [152] C.D. Manning și H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.
- [153] J.D. Markel și A.H. Gray. *Linear Prediction of Speech*. Springer, Berlin, 1976.
- [154] V. Marâi și Gh. Marâi. *Comanda vocală a sistemelor tehnice*. Editura Militară, București, 1991.
- [155] G. Micca, A. Frasca și 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 și J. Mariani. *Stochastically-based semantic analysis*. 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 ediție, 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 și 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 și A. Verma. On Deriving a Phoneme Model for a New Language. În *Proceedings International Conference on Spoken Language Processing*, 2000.

- [163] C. Munteanu și **M. Boldea**. A Description Language for Dialog Modeling and Management. În *Proceedings Fourth International Conference on Technical Informatics – CONTI'2000*, Timișoara, octombrie 2000.
- [164] C. Munteanu și **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 și R.A. Cole. Reviewing Automatic Language Identification. *IEEE Signal Processing Magazine*, (5):33–41, octombrie 1994.
- [166] C. Myers și 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 și 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 și 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 și Șt. Gavăt. Aparat pentru recunoașterea automată a vocalelor. *Automatica și Electronica*, (6), 1963.
- [173] E. Oancea. *Analiza și sinteza vorbirii*. Editura Militară, București, 1976.
- [174] A.V. Oppenheim și R.W. Schafer. Homomorphic Analysis of Speech. *IEEE Transactions on Audio and Electroacoustics*, 16(6):118–23, iunie 1968.
- [175] A.V. Oppenheim și R.W. Schafer. *Digital Signal Processing*. Prentice-Hall International, Londra, 1975.
- [176] M. Ordowski, N. Deshmukh, A. Ganapathiraju, J. Hamaker și 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, Franța. *Documentații ale plăcii OROS AU21*, 1995.
- [178] S. Ortmans, H. Ney și 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 și 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 și 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 și 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. Dizertație de studii aprofundate, Departamentul de Calculatoare, Universitatea "Politehnica" din Timișoara, iulie 1996.
- [186] J.M. Pickett. *The Sounds of Speech Communication*. University Park Press, Baltimore, 1980.
- [187] J. Picone, G.R. Doddington și 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 și 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, Timișoara, 1995.
- [191] W.H. Press, S.A. Teukolsky, W.T. Vetterling și B.P. Flannery. *Numerical Recipes in C*. Cambridge University Press, a doua ediție, 1992.
- [192] P. Price, W.M. Fisher, J. Bernstein și 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. Pușcariu. *Limba română: Rostirea*. Editura Academiei, București, 1959.

- [194] L. Rabiner și 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 și 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 și 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 și 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 și R.W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.
- [201] L.R. Rabiner, J.G. Wilpon și 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 și 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 și 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, București, 1968.

- [207] A.E. Rosenberg, L.R. Rabiner, J.G. Wilpon și 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 și 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 și P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 1995.
- [211] I.Gh. Șabac. *Matematici speciale*, vol. 2. Editura Didactică și Pedagogică, București, 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 și L.R. Rabiner. Digital Representations of Speech Signals. *Proceedings of the IEEE*, 63(4):662–77, 1975.
- [214] M.S. Schmidt și 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 și A. Waibel. Language Independent and Language Adaptive Large Vocabulary Speech Recognition. În *Proceedings International Conference on Spoken Language Processing*, 1998.
- [216] R. Schwartz și 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 și 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 și 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 și 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 și 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 și F. Itakura. Spectrum Distance Measures for Speech Recognition. În S. Furui și M.M. Sondhi (editori), *Advances in Speech Signal Processing*, cap. 14, pag. 419–52. Marcel Dekker, New York, 1992.
- [225] K. Sjölander și J. Beskow. WaveSurfer – an Open Source Speech Tool. În *Proceedings International Conference on Spoken Language Processing*, Beijing, China, 2000.
- [226] F.K. Soong și 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 și Pedagogică, București, 1968.
- [229] H.J.M. Steeneken și 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 și 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 și F. Cetină. *Prelucrarea numerică a semnalului vocal*. Editura Militară, București, 1984.
- [233] R.D. Stuart. *Introducere în analiza Fourier cu aplicații în tehnică*. Editura Tehnică, București, 1971.
- [234] A. Tătaru. *Limba română: Specificul pronunțării în contrast cu germana și engleza*. Editura Dacia, Cluj-Napoca, 1997.

- [235] H.N. Theodorescu, L. Buchholtzer și C. Poșă. *Comunicarea orală om-mașină*. Editura Tehnică, București, 1986.
- [236] G. Todorean, M. Costeiu și M. Giurgiu. *Rețele 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 și T. Jurca. Isolated word recognition system. *Buletinul științific și tehnic al Institutului Politehnic "Traian Vuia" Timișoara*, 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 și 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 și 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 și J. Hirschberg (editori). *Progress in Speech Synthesis*. Springer, New York, 1997.
- [243] R. Vancea, Șt. Holban și D. Ciubotariu. *Recunoașterea formelor – Aplicații*. Editura Academiei, București, 1989.
- [244] E. Vasiliu. *Fonologia limbii române*. Editura Științifică, București, 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. Șabac ș.a. Statistical and Hybrid Methods for Speech Recognition in Romanian. *International Journal of Speech Technology*, 5:259–68, septembrie 2002.
- [249] R.A. Wagner și 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 și 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 și 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 și 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 și D.T. Talkin. The Aligner: Text-to-Speech Alignment Using Markov Models. În J.P.H. Van Santen, R.W. Sproat, J.P. Olive și J. Hirschberg (editori), *Progress in Speech Synthesis*, cap. 25, pag. 313–23. Springer, New York, 1997.
- [255] J.G. Wilpon, B.H. Juang și 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 și 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 și 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 și P. Woodland. *The HTK Book Version 3.0*. Cambridge University, 2000.
- [261] S.J. Young, N.H. Russel și 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 și J.F. Serignat. Europec software v4.1 user's guide. Raport SAM-ICP-045, Institute de la Communication Parlée, Grenoble, Franța, 1991.

- [263] Y. Zhang, R. Togneri și 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 și 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 și 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.