Institutul de Informatică Teoretică
Academia Română - Filiala Iaşi
Secţia de Ştiinţa şi Tehnologia Informaţiei

Doctorand: **Apopei Vasile**

ANALIZA UNOR SISTEME NELINIARE CU APLICAŢII ÎN PRELUCRAREA SEMNALELOR

Rezumatul tezei de doctorat

Conducător ştiinţific:
Prof. dr. Horia-Nicolai Teodorescu
Membru corespondent al Academiei Române

- Iaşi 2008 –

Cuprins

Capitolul 1

Introducere în prelucrarea de semnal.

Procesarea semnalelor reprezintă un domeniu de cercetare vast, care se ocupă cu dezvoltarea de metode şi algoritmi pentru analiza, extragerea de trăsături, interpretarea, codificarea, transformarea şi manipularea semnalelor.

Din punct de vedere tehnic, semnalele pot fi definite ca fiind suportul fizic al transmiterii informației în şi între sisteme. Această definiție a semnalelor se bazează pe modelarea sistemică a lumii înconjurătoare şi a mecanismelor de transmitere a informației. Sistemele care furnizează la ieşirea lor semnale sunt văzute ca surse de semnal.

Semnalele pot proveni din surse diverse: audio, imagini, semnale biomedicale, din procese fizice sau chimice, etc. De multe ori, pentru a putea fi procesate, semnale provenite de la sisteme sunt transformate în semnale electrice cu ajutorul unor dispozitive electrice sau electronice: microfoane; camere de luat vederi; senzori sau traductori termici, optici, de presiune, de poziție, de proximitate, de accelerație şi viteză, etc. Majoritatea semnalelor provenite din lumea înconjurătoare prezintă variație continuă în timp şi, pentru procesarea acestora, se folosesc sisteme analogice.

Pentru analiza teoretică a sistemelor şi semnalelor se recurge la reprezentarea acestora prin funcții matematice. Funcțiile matematice folosite pentru aceste reprezentări depind în primul rând de timp (exemplu de reprezentare a unui semnal sinusoidal: $x(t)=a*\sin(\omega t)$). În reprezentarea matematică a semnalelor pot interveni şi alte variabile cu semnificație fizică (spațiul, temperatura, frecvența, amplitudinea, caracteristici ale sistemelor etc.).

Apariția microprocesoarelor şi progresele înregistrate de sistemele de calcul electronic au determinat apariția şi dezvoltarea după 1950 a unui nou subdomeniu de procesare a semnalelor, procesarea digitală a semnalelor (în engleză *Digital Signal Processing – DSP*). Pentru fi procesate cu ajutorul calculatoarelor, semnalele continui în timp sunt supuse unui proces de conversie în semnale digitale. Această conversie se realizează cu ajutorul unor circuite electronice numite convertoare analog-digital (în limba engleză *Analog to Digital Convertor – ADC*). Uneori rezultatele procesării digitale a semnalelor sunt reintroduse ca intrare în sisteme analogice. Pentru aceasta au fost realizate circuite speciale de conversie a semnalelor digitale în semnale analogice (*Digital to Analog Convertor – DAC*). În figura 1.1 este prezentată schema bloc a unui sistem de procesare digitală a semnalelor care este pus în legătură cu sisteme de procesare analogică a semnalelor. În cadrul procesărilor digitale timpul nu mai este o variabilă continuă, ci o variabilă discretă (în figura 1.1 notat cu *n*).

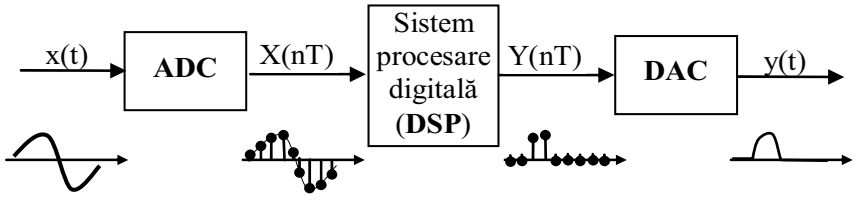

Figura 1.1 Schema bloc a unui sistem de procesare digitală a semnalelor

Legătura între timpul continuu şi cel discret se face cu ajutorul perioadei de eşantionare *T*. Prin eşantionare o parte din informația transmisă de semnal se poate pierde. Pentru a reduce pierderea de informație transmisă de semnalele analogice, perioada de eşantionare trebuie să respecte teorema Nyquist-Shannon.

În funcție de evoluția amplitudinii semnalelor în timp sau spațiu, acestea se pot clasifica în deterministe şi nedeterministe sau aleatoare. Semnalele deterministe pot fi descrise complet prin ecuații matematice liniare sau neliniare. Semnalele aleatoare sunt cele a căror evoluție în timp nu poate fi anticipată cu certitudine, ca de exemplu: semnalul vocal,

semnalul video, semnalul muzical etc. Cu cât aceste semnalele sunt mai imprevizibile (entropia semnalului este mai mare) cu atât cantitatea de informație transmisă de acestea este mai mare (Shannon 1950, Onicescu 1966).

Din multitudinea surselor şi claselor de semnal, am ales pentru cercetările prezentate în această lucrare clasa semnalelor provenite din comunicarea umană, numite în literatura de specialitate semnal vocal.

Semnalul vocal are o structură complexă şi variabilă în timp (H.N. Teodorescu ş.a 1997, Stylianou 2001) în care se pot distinge mai multe tipuri de segmente: segmente cu comportare cvasi-periodică pentru sunetele sonore; non-deterministe staționare pentru sunetele fricative sonore şi nesonore. Pentru caracterizarea variabilității în timp, semnalul vocal poate fi văzut, din punct de vedere sistemic, ca fiind ieşirea unui sistem neliniar care are la intrare mai multe semnale de excitație, cuantizate pe nivele, care furnizează informații despre: contextual fonetic, sintactic şi semantic; starea emoțională a vorbitorului; interrelația dintre vorbitori (în cazul dialogului); interrelația dintre vorbitor şi audiență ş.a. (Teodorescu 2005). În figura 1.2 este prezentată o schemă cu modul de conectare a acestor semnale la intrarea unui sistem de sinteză prozodică.

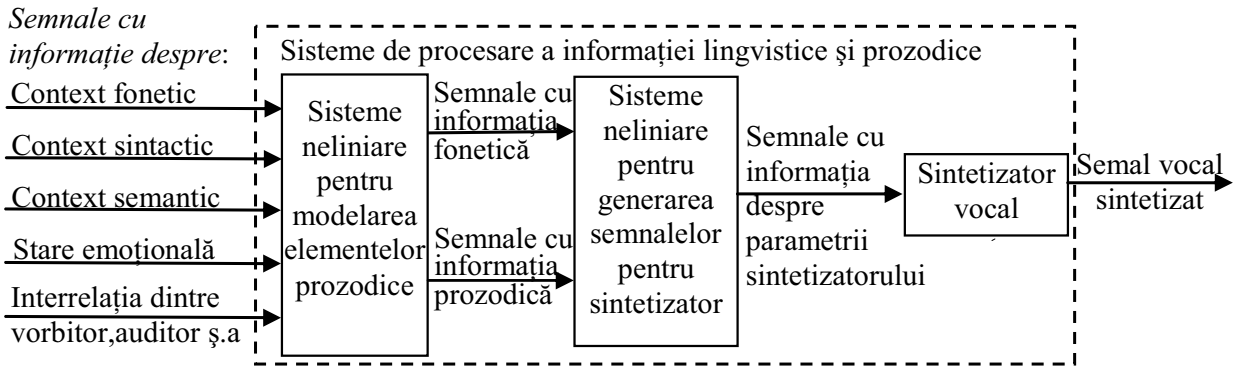

Figura 1.2 Schema bloc a unui sistem de sinteză prozodică.

Procesarea semnalelor vocale reprezintă o direcție de cercetare importantă datorită implicațiile pe care aceasta le are în domeniile medical, lingvistic, fonologic, telecomunicații, tehnologiei vorbirii.

Aplicațiile medicale care se bazează pe procesarea semnalului vocal urmăresc cu precădere evidențierea de trăsături care să diferențieze vocile patologice de cele normale (Teodorescu 1987), analiza posibilităților de recuperare a pacienților cu afecțiuni neurologice (dislexie) sau a vocilor profesionale (profesori, actori, jurişti, preoți, solişti vocali etc.).

Procesarea din punct de vedere lingvistic şi fonologic a semnalelor vocale are ca scop analiza principalelor elemente care influențează prozodia şi elaborarea de modele prozodice.

Ultimul domeniu luat în discuție şi cel mai complex prin interdisciplinaritate este domeniul tehnologiei vorbirii. Tehnologia vorbirii foloseşte rezultatele cercetărilor din domeniile procesării de semnal, procesării limbajului natural, modelării prozodice şi emoționale a vorbirii cu scopul de a dezvolta aplicații de compresie, recunoaştere şi sinteză vocală, şi înglobarea acestora în sisteme de dialog vorbit om-maşină. Aceste aplicații sunt utilizate în domeniul medical de persoanele cu handicap motor sau vizual, industrie, telefonie, transporturi ş.a.

În țara noastră problematica recunoaşterii şi sintezei vocale a fost abordată începând cu anii ’60 la Institutul de Fonetică al Academiei Române (analiza şi sinteza vocală), Universitatea Bucureşti (modelarea matematică a proceselor lingvistice) şi la Academia Militară. După 1980 au început să apară şi alte grupuri de cercetare, care au abordat această problematică (Institutul Central de Informatică, Filialele din Bucureşti şi Iaşi ale Institutului

de Tehnică de Calcul, Facultatea de Electronică de la Institutul Politehnic Bucureşti, Facultatea de Electronică de la Institutul Politehnic Iaşi, Institutul Politehnic Timişoara, Institutul Politehnic Cluj, Institutul de Medicină Iaşi).

1.1 Direcții de cercetare în domeniul procesării semnalului vocal

Până la începutul anilor ’95, cercetările din domeniul recunoaşterii şi sintezei vocale au urmărit dezvoltarea de metode, algoritmi şi soluții pentru realizarea unor tipuri de sintetizatoare vocale (vocodere, articulatorii, formantice, bazate pe modulația AM-FM sau concatenative) şi realizarea unor sisteme de recunoaştere vocală bazate pe rețele neuronale şi lanțuri Markov ascunse (HMM). Tot în această perioadă au intrat în atenția cercetătorilor teoriile fonetice şi fonologice cu implicații în modelarea aspectelor prozodice ale semnalului vocal (în mod special, intonația). Acestea au creat premisele trecerii într-o nouă etapă a sistemelor de sinteză şi recunoaştere vocală, prin realizarea de descrieri ale semnalului vocal din punct de vedere prozodic şi al stărilor emoționale (Fant 2004, Furui 2007).

Conform teoriei propuse de Ladd (1996), prozodia unei propoziții poate fi exprimată prin structuri ierarhice care realizează gruparea cuvintelor în unități intonaționale, de diferite mărimi, în funcție de proeminența relativă (*„weak”/”strong”*) a unităților intonaționale vecine. Pentru analiza proeminențelor relative dintre unitățile intonaționale se folosesc, în general, următoarele elemente prozodice: conturul frecvenței fundamentale, intensitatea şi durata sunetelor, durata pauzelor.

În ultimii ani au început să apară definiții mai complete pentru modelele prozodice. Conform acestor definiții, modelele prozodice realizează o reprezentare fonologică a vorbirii pe baza unor relații între funcțiile şi formele (elementele şi evenimentele) prozodiei (Hirst 2007, Shih 2006, Kohler 2005, Batliner 2003). În prezentarea funcțiilor prozodiei Shih (2006) ia în discuție funcțiile lexicale (accentele şi contrastele lexicale care apar între cuvinte), funcțiile intonației interogative (interogația totală, interogațiile ne-totale, interogația declarativă, interogația în ecou) şi funcțiile paralingvistice (segmente de discurs, transmiterea de stări emoționale) ale prozodiei. Kohler (2005) pune în evidență o legătură între funcțiile comunicative ale prozodiei şi formele de pe conturul intonațional pe baza unei analize a contextului semantic şi pragmatic a transmiterii mesajului de vorbitor către ascultător. Teodorescu H.N. (2005) propune completarea structurii de informații rezultată în urma analizei morfologice, sintactice şi de discurs (a textului) cu informații despre emoție, interrelația vorbitor-receptor şi starea vorbitorului .

Cu toate că, în literatura de specialitate, mulți autori susțin ideia conform căreia frecvența fundamentală (F0) este cel mai important element prozodic în stabilirea proeminențelor dintr-o rostire, există cercetări (Kochanski G. şa. 2005) care susțin faptul că intensitatea şi durata sunetelor joacă un rol mai mare în stabilirea proeminențelor, iar frecvența F0 joacă un rol minor. În opinia lor vorbitorii realizează proeminențele în primul rând prin „pattern-uri” de durată şi energie. Feraru M şi Teodorescu H.N (2008) completează lista elementelor care influențează prozodia cu energia primilor patru formanții din componența semnalului vocal (F1-F4).

La nivelul semnalului vocal, descrierile prozodice şi cele emoționale sunt realizate, cu ajutorul unor sisteme neliniare, pe baza unor parametrii extraşi din unda vocală: conturul frecvenței F0, durata şi energia segmentelor sonore, durata segmentelor nesonore şi pauzelor, timbrul vocii ş.a. Pentru a fi utilizate în aplicații, aceste descrieri sunt introduse în modele intonaționale, modele de durată, modele de energie şi respectiv, modele pentru pauze. Modelele realizează legătura între variația în timp a acestor parametrii şi structura de informații a textului asociat semnalului vocal. Ansamblul acestor modele formează împreună modelul prozodic.

Modelele prozodice au contribuit semnificativ la creşterea performanţelor sistemelor de recunoaştere vocală (Glass 2003, Batliner 2003 ş.a) şi sistemelor de conversie text-voce (Schröder 2004, Shih 2006). Ca o consecinţă a creşterii performanţelor sistemelor de recunoaştere vocală au început să apară aplicaţii pentru adnotarea automată a corpusurilor de voce (Matthew & Jain 1997, Meinedo H. & J. Neto 2003), aplicaţii care sunt folosite la dezvoltarea corpusurilor de semnal vocal. De asemenea s-a trecut la realizarea sistemelor de înţelegere a vorbirii (E. Shriberg & A. Stolcke 2004) şi a sistemelor de dialog vorbit om-maşină (Peckham 1991, Wahlster 2000, Batliner 2003).

În cadrul sistemelor de recunoaştere vocală, modelele prozodice sunt utilizate pentru a ajuta procesul de recunoaştere în situaţiile de incertitudine. Schema de principiu a unui sistem de recunoaştere vocală care foloseşte modul prozodic este ilustrată în figura 1.3. Modulul care cuprinde modelul prozodic şi modelul de limbaj primeşte ca intrare semnalele derivate extrase din unda vocală (curba de energie şi conturul frecvenţei F0) şi secvenţe ipotetice de unităţi segmentale (silabe, cuvinte, grupuri de foneme) reprezentabile sub forma unui graf, provenite de la un submodul de recunoaştere vocală iniţială. La ieşirea acestui modul se completează graful de unităţi segmentale cu informaţii probabilistice despre evenimentele prozodice (graniţe de unităţi intonaţionale, accente) asociate unităţilor segmentele de la intrare. Secvenţa de evenimente prozodice ipotetice, împreună cu secvenţele de unităţi segmentale ipotetice constituie intrare în modulul de decizie finală al sistemului de recunoaştere vocală.

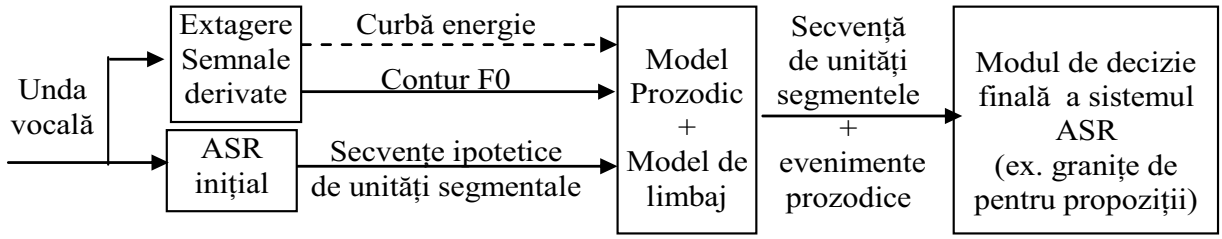

Figura 1.3 Schema unui sistem de recunoaştere vocală cu model prozodic (după Washlster 2000)

Folosirea modulului prozodic în sistemul Verbmobil (Wahlster 2000) a dus la îmbunătăţirea procentului de recunoaştere şi înţelegere a vorbirii prin diferenţierea graniţelor de unitate intonaţională şi a celor de unitate intonaţională intermediară, de graniţele de cuvânt sau graniţele agramaticale.

În cadrul sistemelor de conversie text-voce, modelele prozodice sunt utilizate în principal pentru predicţia evenimentelor prozodice (graniţe de unităţi intonaţionale, tipuri de accente sintactice şi semantice) care pot fi asociate rostirii unui text. Implementarea prozodiei în sinteza vocală a permis obţinerea de semnal vocal sintetizat cu nuanţe de conţinut semantic. Detalii despre sistemele de conversie text-voce vor fi prezentate în secţiunea 1.2 şi capitolele următoare.

1.2 Stadiul actual în domeniul sistemelor de conversie text-voce

Sistemele de conversie text-voce (în limba engleză "*Text-to-Speech*" - TtS) realizează conversia unui text în semnal vocal sintetizat. Aceste sisteme sunt rezultatul cercetărilor interdisciplinare din domeniile: sinteză vocală; procesarea limbaj natural; analiza şi descrierea parametrică a semnalului vocal din punct de vedere fonetic şi fonologic. Evaluarea performanţelor acestor sisteme s-a făcut la început numai pe baza percepţiei psiho-acustice a acurateţei, inteligibilităţi şi naturaleţei sunetelor produse (van Santen 1998), urmând ca apoi evaluarea să fie completată (Bonafonte 2006) cu informaţii despre performanţele modulului de procesare a limbajului natural şi ale modulului prozodic.

Preocupări pentru realizarea de sisteme care să producă sunete asemănătoare vorbirii,

datează încă din secolele XI-XII (G.Aurillac, A. Magnus, R. Bacon) când s-au obținut primele capete vorbitoare (în engleză *„speaking heads”*) pe baza automatelor hidraulice şi pneumatice (Wikipedia). Abia în a doua jumătate a secolului al XVIII-lea C.G. Kratzenstein şi Wolfgang von Kemplemen au realizat primele sisteme mecanice de producere a sunetelor vocalice (Teodorescu ş.a 1986). Următorul pas important în evoluția sistemelor de sinteză a semnalului vocal a fost realizarea sistemelor electronice de sinteză bazate pe codarea-decodarea semnalului vocal cu ajutorul filtrelor analogice şi rezonatori serie sau paraleli.

Următoarele etape importante în evoluția sistemelor de sinteză sunt asociate în general cu dezvoltarea tehnologiei calculatoarelor, sintetizatoarelor pe bază de reguli şi sintetizatoarele concatenative.

În țara noastră primele sisteme de sinteză vocală sunt raportate începând cu anul 1997 la Bucureşti (C. Burileanu ş.a. 1997) şi Cluj (A. Frentz ş.a. 1997) primele sisteme de conversie text-voce pentru limba română folosind sinteză concatenativă bazată pe difoni.

Prin includerea elementelor prozodice în sinteza vocală, sistemele de conversie text-voce dezvoltate în ultimi ani sunt capabile să transmită mesaje cu conținut semantic şi emoțional. Ele au în componență următoarele module (figura 1.4 (b)):

- modulul de procesare a textului (în engleză ***N**atural **L**anguage **P**rocessing*) - completează textul de intrare cu informații despre structura morfo-sintactică şi conținutul semantic a cuvintelor; realizează fonetizarea textului de intrare;
- modulul prozodic – realizează frazarea textului şi generează descrieri parametrice pentru elementele prozodice utilizate în implementare (intonația, intensitatea sunetelor, durata fonemelor şi pauzelor);
- modulul fonetic - generează semnalele pentru comanda sintetizatorului vocal pe baza informațiilor fonetice şi prozodice primite de la modulele anterioare;
- sintetizator vocal – realizează generarea unui semnal sintetizat pe baza semnalelor generate de modulul fonetic.

Proiectarea şi realizarea modulelor unui sistem de conversie text-voce implică parcurgerea mai multor etape de procesare şi analiză pe corpusuri paralele text-voce, adnotate multiplu (figura 1.4 (a)). Aceste etape vizează două direcții principale: analiza şi adnotarea corpusurilor de voce la nivel fonetic, fonologic, prozodic şi emoțional; analiza şi adnotarea corpusurilor de text la nivel morfologic, sintactic şi semantic.

Adnotările multiple obținute pe corpusurile paralele constituie intrări pentru procesul de învățare al modulul prozodic. Analiza şi adnotarea la nivel fonetic a corpusului de voce poate fi folosită şi pentru perfecționarea modulului fonetic prin îmbunătățirea descrierilor parametrice realizate pentru sunete sau grupuri de sunete.

Dacă ținem cont de interacțiunile dintre modulele sistemului de conversie text-voce şi etapele procesului de analiză, este lesne de înțeles faptul că descrierea parametrică a sunetelor este dependentă de modul de realizare a sintezei vocale (concatenative, formantice, pe baza modulației AM-FM) iar descrierea prozodică este dependentă de modelul prozodic care se implementează în sistemul de conversie text-voce. În fiecare etapă de analiză şi procesare, la nivelul semnalului vocal şi a textului, se folosesc sisteme neliniare.

Sistemele de conversie text-voce pentru limba română, realizate în ultimii ani, apelează la diferite modalități de introducere a elementelor prozodice. Sistemul dezvoltat la Universitatea „Politehnica” Bucureşti foloseşte un model prozodic bazat pe reguli lingvistice care țin cont de poziția accentului lexical, informații despre semnele de punctuație şi tipul de rostire (declarativă, interogativă, exclamativă sau imperativă) pentru stabilirea unei intonații cu patru nivele tonale de realizare a accentelor (D. Burileanu ş.a. 2004). Sistemul dezvoltat la Universitatea Tehnică din Cluj foloseşte pentru introducerea accentelor lexicale, într-un sistem de sinteză bazat pe concatenarea de silabe, un set de reguli lingvistice (Buza ş.a. 2007).

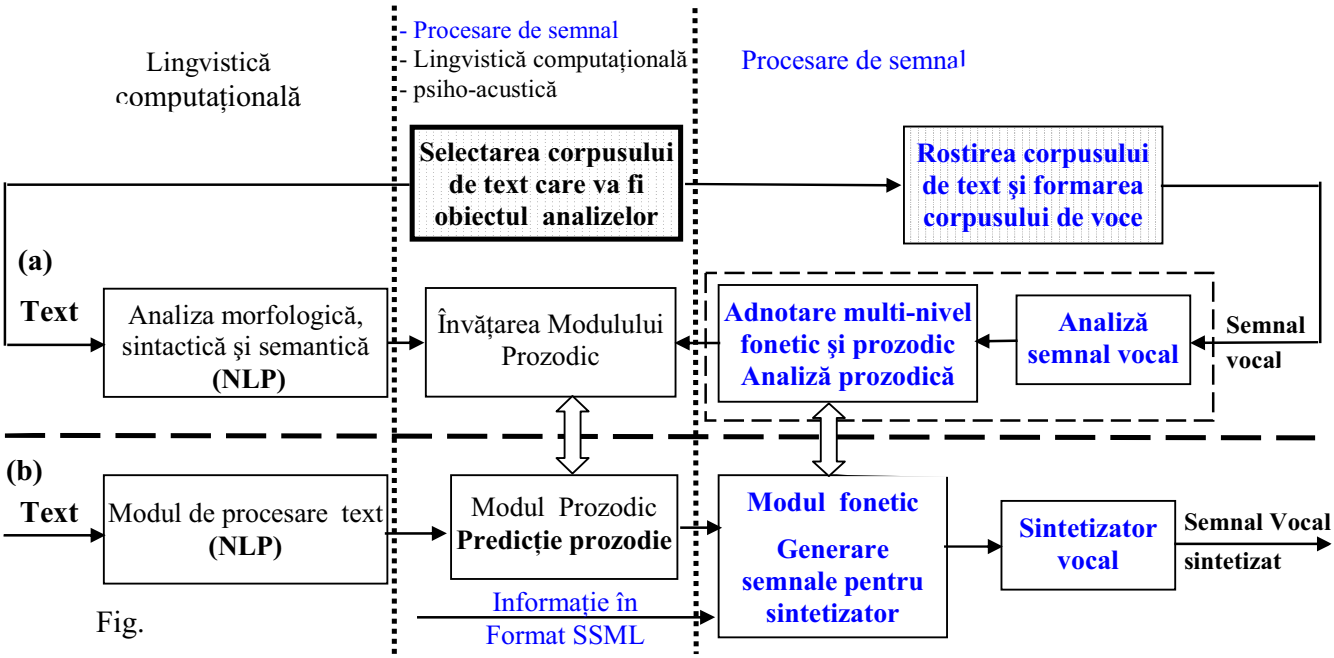

Fig. 1.4. Etape de procesare şi analiză la nivelul semnalului vocal şi textului în vederea realizării unui sistem de conversie text-voce cu modul prozodic: (a) etapele procesului de învățare a prozodiei; (b) schema bloc a unui sistem de conversie text-voce.

Introducerea elementelor de prozodie în sistemul de conversie text-voce dezvoltat în cadrul Institutului de Informatică Teoretică s-a realizat pe texte adnotate în format XML şi a cunoscut două etape de dezvoltare: a) prima abordare s-a bazat pe un model propus de H.N. Teodorescu şi s-a concretizat la nivelul adnotării VoiceXML pe împărțirea din punct de vedere intonațional a unui fragment din „Ecleziastul" (Teodorescu, Ceauşu, Apopei 2003) în grupuri de cuvinte cu pattern-uri intonaționale. Pentru delimitarea grupurilor de cuvinte s-a folosit *tag*-ul „break" cu două valori (0 şi 2) prin care se indică prezența unor pauze, iar pentru descrierea tonurilor de realizare a accentelor lexicale din cadrul acestor grupuri s-a introdus pentru cuvinte atributul „pitch" cu două valori („high"/ „low"), asociat în general cuvintelor de la începutul şi de la sfârşitul grupurilor de cuvinte. Împărțirea frazelor în grupuri de cuvinte şi nivelul tonurilor erau stabilite în funcție de anumite clase de mărci textuale şi semne de punctuație folosind *n-grame*; b) a doua abordare urmăreşte realizarea unei legături între: - analiza semantică (Teodorescu 2005, Kohler 2005) şi împărțirea din punct de vedere intonațional a rostirilor unor texte folosind teoriile fonologice şi în special teoria autosegmental-metrică (Pierrehumbert 1980, Ladd 1996); şi analiza morfologică, sintactică şi semantică a acestor texte. În acest context am dezvoltat un model fonologic ierarhizat (Apopei 2006, 2007) în care pentru marcarea evenimentele tonale de pe conturul frecvenței fundamentale, evenimente prin care se realizează accentele lexicale, am folosit în principal etichete din sistemul de adnotare a intonației ToBI. Această abordare a prozodiei a fost dezvoltată în cadrul temelor de cercetare ale Institutului de Informatică Teoretică şi a fost concepută din perspectiva realizării unei punți de legătură între cercetările din domeniul lingvisticii computaționale (Tufiş 2000,2007, Cristea 2003, 2005, Curteanu 2007, Forăscu 2006, 2008) şi cele din domeniul analizei şi sintezei vocale pentru limba română (Teodorescu H.N. 2003, 2005, 2008, Burileanu D. 2006, Grigoraş Fl. 1997,1999, Jitcă 2002, 2003).

În cadrul cercetărilor efectuate am insistat mai mult pe modelarea aspectelor legate, în special, de dinamica (variabilitatea) semnalului vocal şi de modelarea aspectele prozodice ale acestuia. Pentru aceste modelări am utilizat sisteme neliniare inteligente, în care neliniaritățile sunt introduse prin reguli, prin indicații (etichete) etc. Principalele probleme abordate sunt: descrierea parametrică a fonemelor şi co-articularea sunetelor; metode şi algoritmi de procesare a semnalului vocal pentru etapele de analiză şi adnotare a prozodiei;

implementarea rezultatelor obținute în etapele de analiză într-un sistem de conversie text-voce.

1.3 Stadiul actual în domeniul predicției intonației şi al teoriilor fonologice

Elementele prozodice studiate şi implementate în sistemele *text-to-speech* sunt derivate din caracteristici acustice ale vocii. Cele mai importante elemente prozodice luate în considerare de modelele prozodice sunt: intonația, intensitatea sunetelor, durata silabelor (fonemelor) şi a pauzelor. Intonația este o caracteristică acustică a semnalelor vocale dată în principal de variația frecvenței fundamentale F0 şi depinde de modul în care vorbitorul realizează frazarea (gruparea) şi accentuarea cuvintelor. Implementarea intonației în sinteza vocală presupune generarea automată a "melodiei" corespunzătoare rostirii unui text, pe baza unor modele intonaționale care pun în corespondență structura sintactică şi conținutul semantic al textului cu un set de evenimente intonaționale şi un set de pattern-uri la nivelul frecvenței F0.

Cercetările efectuate în domeniul predicției intonației (evenimentelor intonaționale) au pus în evidență existența următoarele ipoteze de lucru: (a) structura intonațională poate fi complet determinată pe baza structurii morfologice şi sintactice a textului (Chomsky&Halle 1968, Selkirk 1984); (b) structura intonațională reflectă mai mult conținutul semantic decât structura sintactică a textului (Selkirk 1999, Gussenhoven 1992, 2007); (c) structura intonațională şi cea sintactică reflectă conținutul semantic şi structura de informații (Steedman 1991, Huesinger 1999). În cadrul ultimelor două ipoteze de lucru, pe baza noțiunilor de focus şi conținut semantic se pune în evidență faptul că, melodia rostirii unui text este determinată în principal de conținutul semantic şi emoțional al mesajului care trebuie transmis.

Implementarea modulelor de generare a conturului frecvenței F0 în sinteza vocală se realizează în principal prin două clase de modele: modele bazate pe principiul superpoziției şi modele care descriu conturul frecvenței F0 printr-o secvență de evenimente tonale cu anumite semnificații fonetice şi/sau fonologice.

Modelele intonaționale bazate pe principiul superpoziției consideră conturul frecvenței F0 ca o rezultantă a sumării mai multor componente intonaționale. Dintre acestea cele mai importante componente (Öhman 1967, Fujisaki 1983, 2004) se referă la intonația frazei intonaționale şi intonația corespunzătoare accentului de cuvânt. Cele mai cunoscute implementări ale modelelor intonaționale bazate pe principiul superpoziției sunt cele raportate Mixdorff (1998, 2003) şi Santen (2002).

Modelele intonaționale care interpretează conturul frecvenței F0 ca o secvență de evenimente tonale s-au dezvoltat în principal din două considerente: necesitatea de adnotare prozodică a corpusurilor de voce; predicția şi generarea conturului frecvenței F0 în sinteza vocală. Adnotarea prozodică realizează descrieri fonetice şi fonologice pentru evenimentele prozodice, descrieri care să dea sens şi semnificație conturului intonațional. Descrierile realizate pentru conturul intonațional, după 1980, au la bază teoriile fonetice şi fonologice. Dintre acestea, cele mai utilizate sunt fonologia metrică şi fonologia autosegmentală.

Fonologia metrică, introdusă de Liberman (1975) pentru studiul accentului şi ritmului, realizează descrierea intonației pe baza de proeminențe relative, de tip *weak/ strong*, între unități intonaționale (fig. 1.5).

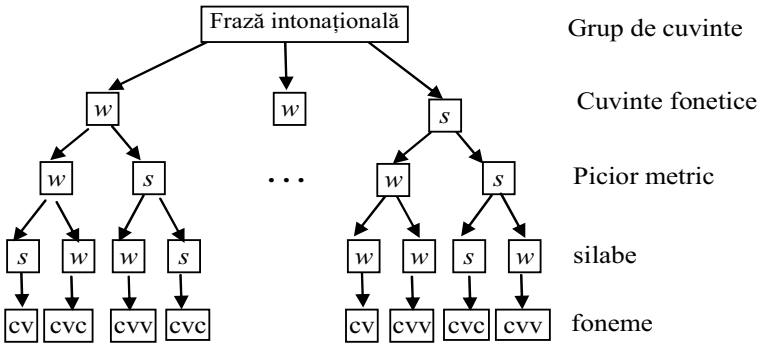

Figura 1.5. Exemplu de descriere a intonației pe baza fonologiei metrice
Notații: w=*weak*, s=*strong*, c=consoană, v=vocală

Fonologia autosegmentală, propusă de Goldsmith (1975), este derivată din fonologia generativă (Chomsky&Halle 1968) şi foloseşte o reprezentare multi-nivel pentru a pune în corespondență nivelele fonetic, silabic şi tonal (fig. 1.6).

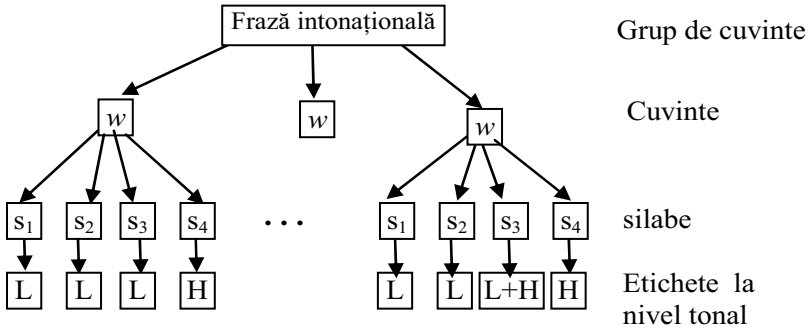

Figura 1.6. Exemplu de descriere a intonației pe baza fonologiei autosegmentale
Notații: w=*word*, s_i=silabe, L=nivel tonal scăzut, H=nivel tonal ridicat

Pe baza acestor teorii s-a dezvoltat teoria autosegmental-metrică bazată pe secvențe de tonuri a lui Pierrehumbert (1980), preluată şi dezvoltată ulterior de Ladd (1986, 1999). Descrierile fonologice ale intonației pe baza teoriei autosegmental-metrice realizează descrieri ierarhice pentru intonație. Evenimentele de pe conturul frecvenței fundamentale F0 sunt asociate unui set de descrieri fonologice. Acest set de descrieri au asociat la nivelul conturului frecvenței F0 un set de primitive de contur, reprezentative pentru intonație.

Ținând cont de caracterul complex şi interdisciplinar al problematicii analizei şi predicției elementelor prozodice şi de impactul descrierilor fonologice asupra modelării intonației, cercetările efectuate pe perioada elaborării tezei s-au axat pe stabilirea unei structuri ierarhice de adnotare a prozodiei pe corpusurilor de voce pentru limba română.

În țara noastră, predicția elemente prozodice s-a axat la început pe predicția accentelor lexicale folosind rețelele neuronale (D. Burileanu ş.a. 1999) sau reguli (Buză ş.a. 2007). În prezent, pentru predicția elementelor prozodice, din lucrările publicate, se disting două direcții: un model prozodic bazat pe reguli lingvistice care ține cont de poziția accentului lexical, informații despre semnele de punctuație şi tipul de rostire (declarativă, interogativă, exclamativă sau imperativă) pentru stabilirea unei intonații cu patru nivele tonale de realizare a accentelor (D. Burileanu ş.a. 2006); folosirea unui model prozodic ierarhic (Apopei şi Jitcă 2006, 2007) bazat pe teoriile fonologice (Pierrehumbert 1980, Ladd 1996) care porneşte în predicția intonației de la împărțirea textului în fraze intonaționale, iar la nivelul frazelor intonaționale diferențiază mai multe posibilități de realizare a accentelor lexicale (pe baza setului de etichete din sistemul de adnotare a intonației ToBI).

Capitolul 2.

Procesarea semnalului vocal

Problematica analizei şi procesării semnalului vocal reprezintă un domeniu de interes atât pentru cercetătorii din domeniul tehnologiei vorbirii, cât şi pentru cei din alte domenii, cum ar fi cel lingvistic sau medical. Pentru fiecare domeniu de cercetare, procesarea şi analiza semnalului vocal are drept scop, în esență, extragerea de informații (valorile unor trăsături acustice, fonetice sau fonologice) cu ajutorul cărora să se poată face clasificări, codificări, descrieri parametrice şi interpretări specifice. Pentru domeniul tehnologiei vorbirii trăsăturile au drept scop realizarea de descrieri parametrice în vederea recunoaşterii şi sintezei vocale; în domeniul lingvisticii computaționale se urmăreşte corelarea valorilor unor trăsături extrase din unda vocală cu structura semantică şi de discurs a textului corespunzător; în cercetările medicale, importante sunt corelările valorilor unor trăsături extrase din unda vocală cu anumite particularități, legate de starea de sănătate a subiecților care le-au rostit. Aceste trăsături extrase din unda vocală rezultă din analiza semnalului vocal în domeniul timp, în domeniul frecvență sau în spațiul stărilor (Keller ş.a.1993, Teodorescu ş.a 1997).

Valorile trăsăturilor folosite pentru analiza semnalului vocal în domeniul timp pot fi determinate direct din unda vocală (durate, amplitudini), sau derivat din aceasta cum ar fi: energia semnalului (corespunzătoare intensității sunetului definite în procesul de percepție vocală); frecvența fundamentală (conturul F0), frecvența trecerilor prin zero, valorile componentelor armonice şi aleatorii din semnalul vocal. În afara acestor trăsături care au semnificație fizică imediată, în practică, se mai foloseşte caracterizarea semnalului vocal în domeniul timp prin coeficienții de predicție lineară, care rezultă dintr-o modelare lineară a acestuia.

Trăsăturile folosite pentru analiza semnalului vocal în domeniul frecvență sunt: frecvența fundamentală; amplitudinile şi frecvențele formaților; benzile de frecvență ale componentelor de zgomot; energia în benzi de frecvență (coeficienții MFC). Analiza semnalului vocal este necesară atât pentru modelări locale la nivelul unităților segmentale care compun unda vocală (foneme, alofoni, difoni, trifoni sau silabe), cât şi pentru modelări ale componentelor legate de dinamica semnalului vocal, co-articularea sunetelor şi elementele prozodice.

Pentru modelări locale la nivelul unităților segmentale se foloseşte proprietatea de cvasi-staționaritate a semnalului vocal pe durate de 10-20 msec. Pe baza acestei aproximări se obțin componentele armonice care caracterizează fonemele şi alofonii.

Modelarea aspectelor prozodice vizează identificarea unor pattern-uri şi a unor reguli care să descrie evoluția în timp a elementelor prozodice extrase din semnalul vocal. Schema bloc a unui proces complet de analiză şi modelare a prozodiei este prezentată în figura 2.1. Pentru a putea fi folosite în aplicațiile de recunoaştere şi sinteză vocală, descrierile prozodice trebuie corelate cu: funcțiile semantice, comunicative şi pragmatice ale prozodiei (Kohler 2005, Teodorescu 2005); structura ierarhică a intonației; structura de informații (structură rezultată în urma analizei morfologice, sintactice şi semantice) a textului asociat rostirii.

În urma procesului de analiză a intonației, se obține un set de descrieri macroprozodice care corespund anumitor funcții semantice, comunicative şi pragmatice ale prozodiei, precum şi un set de descrieri microprozodice, care corespund evenimentelor tonale de pe conturul frecvenței F0.

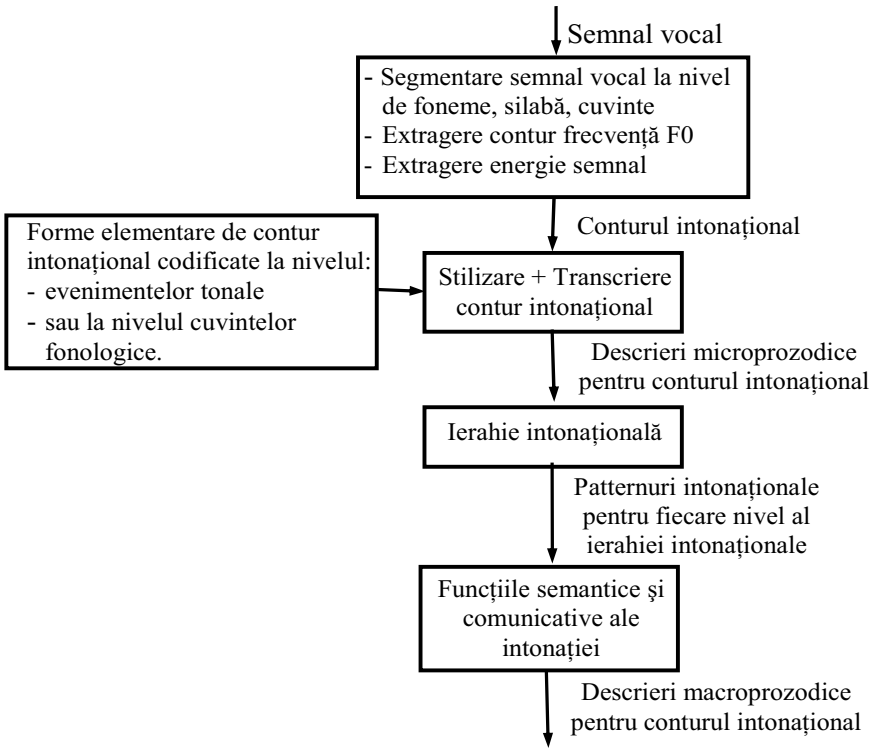

Figura 2.1. Schema bloc a procesului de analiză şi modelare a prozodiei

Un model prozodic complet trebuie să permită generarea descrierilor microprozodice pe baza unui set de descrieri macroprozodice. Descrierile macroprozodice rezultă în urma aplicării funcţiilor semantice, comunicative şi pragmatice ale intonaţiei (Kohler 2005, Teodorescu 2005), structurii de informaţii a textului de intrare (figura 2.2).

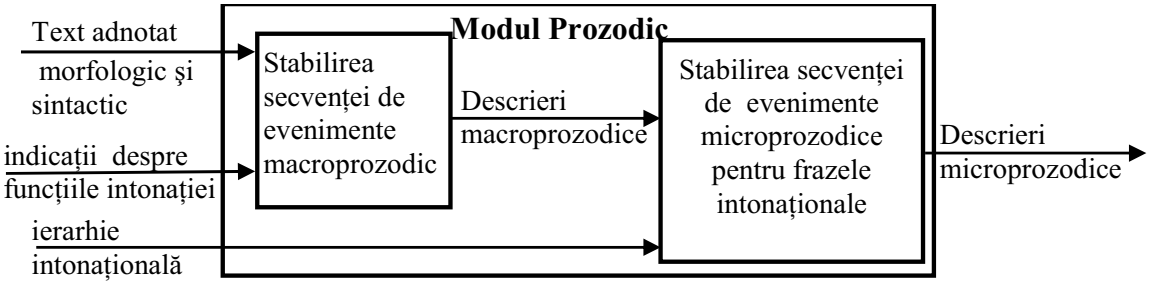

Figura 2.2. Schema bloc a unui model prozodic

Secvenţele de evenimente microprozodice sunt specifice modelului prozodic utilizat şi ierarhiei intonaţionale adoptate. Ele trebuie să ofere posibilitatea modelării la nivelul silabelor a principalelor elemente prin care se materializează prozodia (frecvenţa fundamentală F0, energia, durata silabelor şi pauzelor). Pe parcursul cercetărilor care au stat la baza elaborări prezentei teze am acordat o atenţie deosebită studiului metodelor de estimare a frecvenţei fundamentale şi analizei secvenţelor de evenimente tonale de pe conturul frecvenţei F0.

2.1 Metode de extragere a frecvenţei fundamentale. Conturul frecvenţei fundamentale

În domeniul analizei şi sintezei semnalului vocal, frecvenţa fundamentală (F0) este definită ca frecvenţa de excitaţie a coardelor vocale. Variaţia în timp a frecvenţei fundamentale pe durata unei rostiri este percepută la nivel psiho-acustic prin intonaţie sau melodia rostirii.

La nivelul semnalului vocal, frecvența F0 determină un aspect cvasi-periodic al undei vocale pe durata vocalelor şi consoanelor sonorante. În figura 2.1 semnalul vocal, corespunzător unui segment vocalic, prezintă un caracter repetitiv al unei formei de undă de perioadă T0=1/F0. Formele de undă pot fi considerate aproximativ identice doar pe durata a câtorva perioade succesive (2-4 perioade), pentru care semnalul vocal prezintă caracteristicile semnalelor staționare. Pe această aproximare se bazează majoritatea metodelor de analiză ale semnalului vocal.

Generarea semnalului vocal pe baza excitației produsă de vibrația corzilor vocale determină caracterul armonic al acestuia. Metodele de analiză în domeniul frecvență oferă informații relative la componentele frecvențiale din unda vocală. Pentru a se evidenția în spectru de frecvență armonicile superioare ale frecvenței fundamentale, cadrul de analiză trebuie să conțină 3-4 perioade T0 (datorită teoremei lui Shanon şi a faptului că fereastra de analiză nu este întotdeauna sincronă cu periodicitatea semnalului vocal).

În figura 2.2 este redat spectrul de frecvență al segmentului vocalic din figura 2.1, calculat pe baza unui algoritm de transformată Fourier de tip FFT. Algoritmii de extragere automată a frecvenței fundamentale din spectrul de frecvență se bazează în principal pe determinarea poziției peak-urilor spectrale în banda de joasă frecvență (50-1000 Hz). Primul *peak* spectral corespunde frecvenței fundamentale, iar celelalte *peak-uri* se poziționează la valori de frecvență egale cu multiplii frecvenței F0. (figura 2.2).

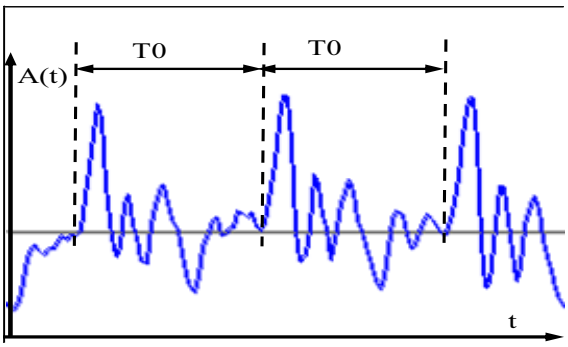

Figura 2.1. Reprezentarea în domeniul timp a undei vocale corespunzătoare unui segment vocalic

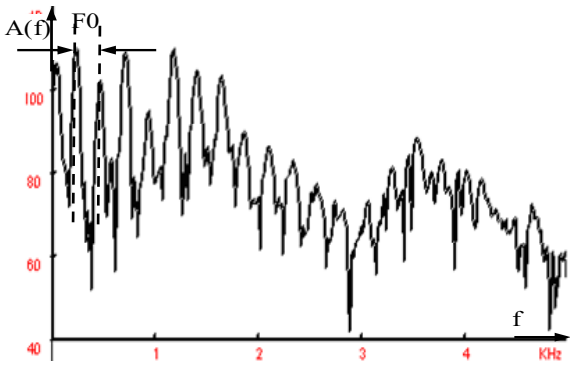

Figura 2.2. Reprezentarea în domeniul frecvența a *peak-urilor* rezultate din analiza FFT a semnalului vocal reprezentat în figura 2.1

Pornind de la aceste observații şi de la necesitatea obiectivă de a extrage melodia din semnalul vocal corespunzător unui anumite rostiri, pentru estimarea frecvenței fundamentale s-au dezvoltat metode care folosesc fie analiza în domeniul timp, fie analiza în domeniul frecvență a semnalului vocal fie analiza în timp şi frecvență.

Corectitudinea şi acuratețea determinărilor frecvenței fundamentale este influențată de următoarele cauze obiective: perioada excitației coardelor vocale se modifică permanent, fapt care generează nestaționarietatea semnalului vocal; interacțiunea dintre oscilatorii tractului vocal superior şi excitația glotală determină uneori atenuarea, până la dispariție, a unor armonici ale frecvenței fundamentale din spectrul de putere; dificultăți în stabilirea exactă chiar şi pe unda vocală în domeniul timp a începutului şi sfârşitului perioadei excitației glotale; dificultăți în a distinge automat segmentele nesonore de segmentele sonore de nivel foarte scăzut; distorsiuni ale semnalului vocal datorate diferitelor surse de zgomot (transmisii telefonice, zgomote de fond, zgomote datorate surselor de alimentare ale componentelor hardware, etc.)

Pentru compararea performanțelor detectoarelor de pitch, L.R. Rabiner (1976) propune un set de criterii, dintre care vom enumera următoarele: acuratețea în estimarea perioadei

fundamentale; acuratețea în stabilirea segmentelor sonore şi nesonore din unda vocală; robustețe la diferite surse de zgomot şi vorbitori; viteza de operare; complexitatea algoritmului.

2.5 Particularități de implementare ale unei metode de estimare a frecvenței F0 bazată pe funcția de autocorelație.

În figura 2.10 este prezentată schema bloc a algoritmului de estimare a frecvenței F0, în cadrul unei ferestre de analiză, bazat pe calculul funcției de autocorelație şi AMDF. Algoritmul a fost implementat şi testat în mediul de programare Matlab.

Semnalul din cadrul ferestrei de analiză este trecut printr-un filtru Batterworth trece jos de ordin 2, cu frecvența de trecere fixată la 700 Hz. La ieşirea filtrului se calculează energia semnalului din fereastra de analiză, se estimează funcția AMDF şi funcția de autocorelație. Autocorelația este calculată cu funcția Matlab *xcorr*.

Pentru determinarea intervalului în care se caută poziția primului minim local semnificativ din funcția *AMDF* şi poziția maximului din autocorelație m-am folosit de valorile minime şi maxime ale frecvenței fundamentale, care poate apare în cadrul undelor sonore şi de valoarea frecvenței de eşantionare a semnalului vocal. Astfel intervalul *[li, ls]* de căutare a punctelor de extrem local este dat de relațiile (2.22), (2.23).

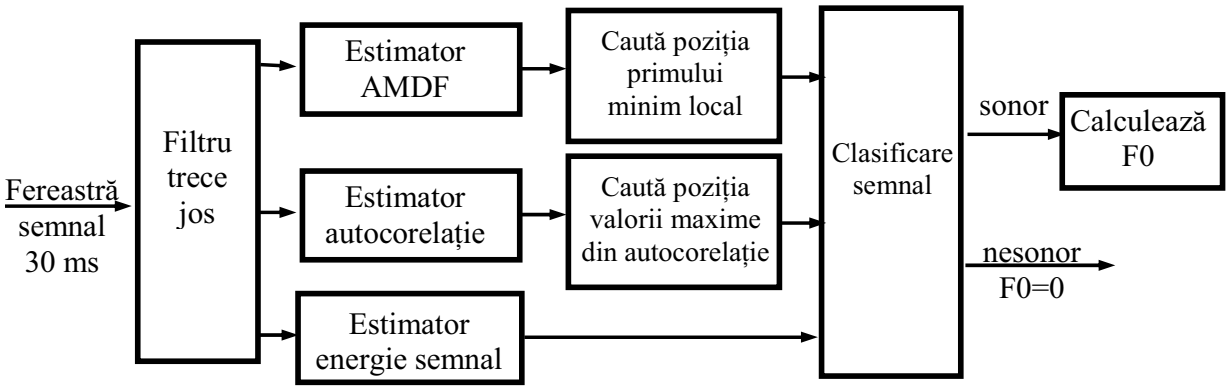

Figura 2.10. Diagrama algoritmului de estimare a valori frecvenței fundamentale pe baza funcției de autocorelație şi AMDF

$$ls=floor(fs/F0min) \quad (2.22)$$

$$li=ceil(fs/F0max) \quad (2.23)$$

unde: *fs* este frecvența de eşantionare a semnalului vocal;

F0min este valoarea minimă a frecvenței fundamentale care poate apare în cadrul rostirii;

F0max este valoarea maximă a frecvenței fundamentale care poate apare în cadrul rostirii.

Cu pozițiile punctelor semnificative de extrem local determinate şi cu valorile energiei semnalului vocal, maximul funcției AMDF şi minimul funcției de autocorelație se intră într-un clasificator pe bază de reguli care stabileşte dacă fereastra analizată corespunde unui segment vocal sonor sau nesonor. Pentru ferestrele clasificate ca fiind sonore, se calculează valoarea frecvenței F0 pe baza poziției punctului de minim din funcția AMDF.

2.6 Particularități de implementare a unei metode de estimare a frecvenței F0 în domeniul frecvență

În figura 2.11 este prezentată schema bloc a algoritmului de estimare a frecvenței F0, în cadrul unei ferestre de analiză, bazat pe calculul funcției cepstrum. Algoritmul a fost implementat şi testat în mediul de programare Visual C++.
Semnalul din cadrul ferestrei de analiză este trecut printr-un filtru Batterworth trece jos de ordin 2, cu frecvența de trecere fixată la 700 Hz, multiplicat cu o fereastră Hamming după care se valorile rezultate se depun într-un vector de dimensiune 1024. Acest vector de date este utilizat pentru estimarea unui spectru de putere în 1024 puncte spectrale folosind un algoritm de FFT. Cu valorile rezultate pentru spectrul de putere se intră într-o procedură de calcul care face estimarea cepstrumului pentru datele din fereastra analizată.

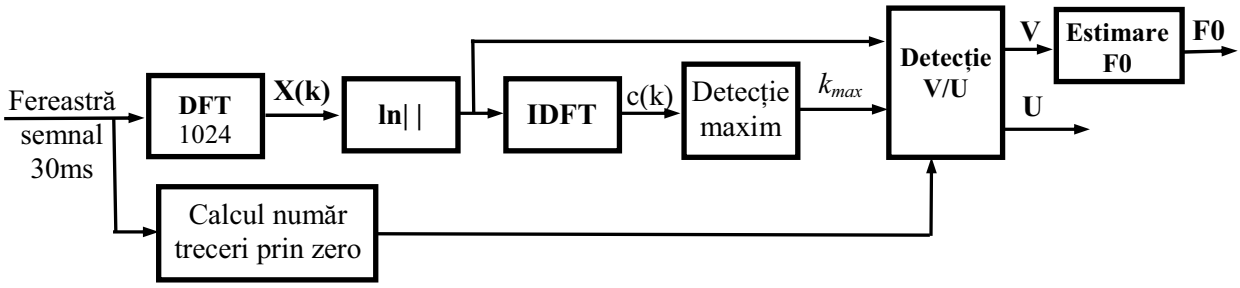

Figura 2.11. Diagrama algoritmului de estimare a valori frecvenței fundamentale pe baza funcției cepstrum şi maximelor din spectrul de putere

Pentru determinarea intervalului în care se caută poziția maximului local semnificativ pe axa quefrenței m-am folosit de relațiile (2.9), (2.10). Maximul semnificativ trebuie să aibă valoarea mai mare de 3 ori decât valoarea absolută medie determinată pentru intervalul analizat. Cu poziția maximului (k_{max}) determinată pe axa quefrenței, spectrul de putere şi frecvența trecerilor prin zero al semnalului din fereastra se intră într-un clasificator care stabileşte dacă semnalul din fereastra analizată este sonor sau nesonor. Pentru semnalul considerat sonor, cu ajutorul relației (2.24) se calculează o posibilă valoare pentru frecvența f0 .

$$f_o = f_s / k_{\max} \qquad (2.24)$$

Cu valoarea f_0 determinată pentru frecvența fundamentală, se caută în spectrul de putere posibile maxime locale care se repetă pe axa frecvenței la intervale egale cu valoarea f_0 , fie la intervale egale cu jumătate sau cu dublul valorii f_0. Valorile estimate pentru frecvența fundamentală prin cele două metode sunt trecute prin două filtre mediane cu dimensiune 3, iar decizia asupra valori finale pentru F0 se ia după următoarea regulă:

- dacă modulul diferenței dintre valorile deduse pentru frecvența F0 din spectrul de putere şi cea din cepstrum este mai mică decât jumătate din valoarea dedusă din cepstrum, atunci frecvența F0 capătă valoarea f_0;
- altfel, se acceptă valoarea care este cea mai apropiată de valoarea lui F0 determinată la pasul anterior, urmând a fi corectată eventual la iterația următoare dacă modulul diferenței scade sub jumătate din valoarea dedusă din cepstrum astfel:
  - ✔ dacă valoarea este apropiată de cea de la iterația următoare, atunci aceasta se păstrează;
  - ✔ altfel, se acceptă pentru F0, valoarea care este între cea determinată la pasul anterior şi cea de la pasul următor.

Dacă modulul diferenței dintre valorile deduse pentru frecvența F0 din spectrul de

putere şi cea din cepstrum se păstrează la valori mari pe mai mult de două fereastre de analiză, se consideră că semnalul are caracter nesonor şi valorile frecvenței F0 pentru respectivele ferestre de analiză sunt egalate cu zero.

2.7 Contribuții personale

Cercetările privind modelarea melodiei semnalului vocal impun dezvoltarea şi implementarea de algoritmi pentru estimarea cât mai corectă a frecvenței fundamentale. Aparent o problemă uşoară, abordată foarte frecvent în literatura de specialitate prin diverse metode, estimarea frecvenței fundamentale în contextul dinamicii nestaționare a semnalului vocal, rămâne o problemă destul de complicată şi generatoare de noi abordări.

Pentru a face față acestei provocări a trebuit să analizez mai multe metode de estimare a frecvenței fundamentale în domeniul timp, domeniul frecvență şi în domeniul timp-frecvență. În urma analizei efectuate am constatat că fiecare metodă reuşeşte să estimeze corect frecvența F0 în anumite condiții de zgomot şi componente armonice ale semnalului vocal.

După trecerea etapei de analiză am reuşit să implementez două metode de estimare a frecvenței F0: una *în domeniul timp* bazată pe combinarea metodei de estimare folosind *funcția de autocorelație* cu o metodă bazată pe *funcția mediei diferenței amplitudinilor (AMDF)*; ce de a doua *în domeniul frecvență* bazată prin combinarea metodei de estimare folosind *funcția cepstrum* cu o metodă de estimare a armonicilor superioare ale frecvenței F0 din spectrul de frecvență al semnalului.

Prin folosirea celor două metode de estimare a frecvenței fundamentale pe aceleaşi semnale vocale, am constatat următoarele: pe segmentele de semnal vocal sonore pe care ambele metode oferă estimări corecte pentru frecvența F0, metoda de estimare a frecvenței în domeniul timp reuşeşte să ofere rezultate care se corelează mai bine cu periodicitatea prezentă la nivelul semnalului vocal în domeniul timp; metoda de estimare a frecvenței fundamentale în domeniul timp reuşeşte să estimeze valori corecte pentru frecvența F0 pe segmente de semnal sonor de intensitate redusă, pe care metoda de estimare în domeniul frecvență estimează rezultate eronate.

Capitolul 3

Sinteza vocală

Istoricului producerii sunetelor artificiale cu ajutorul sistemelor create de om pune în evidență următoarele etape de dezvoltare: etapa capetelor vorbitoare (în engleză *„speaking heads”*) realizate cu automate hidraulice şi pneumatice; etapa sistemelor mecanice de producere a sunetelor vocalice; etapa sistemelor electronice de producere a sunetelor vocalice; etapa sistemelor de producție vocală bazate pe tehnologia calculatoarelor şi a sistemelor electronice de control automat.

Apariția sistemelor de conversie text-voce a avut un impact deosebit în coagularea rezultatelor din domenii de cercetare care până atunci s-au dezvoltat separat: realizarea de sisteme pentru sinteză vocală, procesarea semnalului vocal, procesarea limbajului natural, analiza intonației pentru diferite tipuri de propoziții (Palmer 1922).

După prezentarea tipurilor de sintetizatoare utilizate pentru sinteza vocală, în secțiunea 3.2 este prezentat sintetizatorul formantic de tip Klatt. În secțiunea 3.3 este prezentată contribuția autorului referitoare la implementarea sintetizatorului formantic pentru limba română şi realizarea unei modelări pentru variația valorilor centrale ale formanților F1 şi F2 la co-articularea sunetelor.

3.1 Sisteme pentru sinteza semnalului vocal

Sistemele de sinteză vocală (*speech synthesizer* în limba engleză) sunt parte integrantă a sistemelor de conversie text-voce. Rolul sintetizatoarelor vocale este acela de a transforma informația fonetică (secvența de foneme) şi informația prozodică în semnal vocal. În componența acestora, se pot distinge două blocuri principale: modulul de generare al semnalelor de comandă a sintetizatorului şi modulul de sinteză propriu-zisă (figura 3.1).

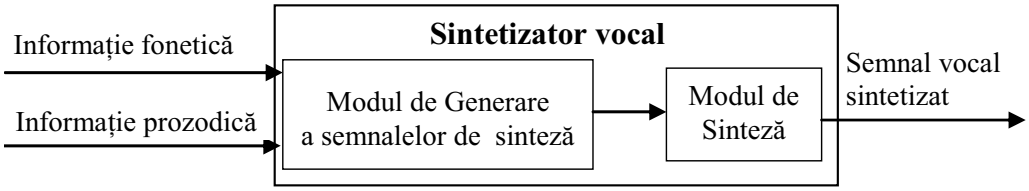

Figura 3.1 Structura generală a unui sintetizator vocal.

Modulul de generare a semnalelor de sinteză (figura 3.2) stabileşte pe baza informațiilor fonetice şi prozodice, evoluțiile temporale ale semnalelor de comanda pentru modulul de sinteza propriu-zisă.

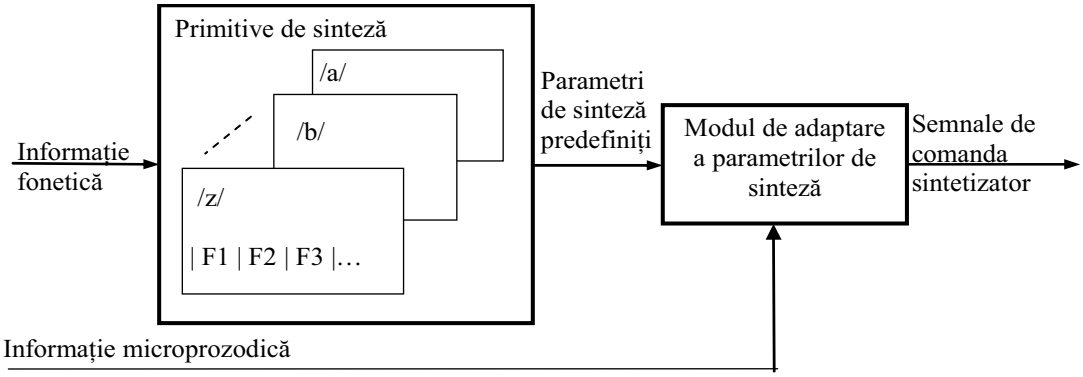

Figura 3.2 Schema bloc a modulului de generare a semnalelor de comanda pentru sintetizator.

Fiecărui fonem, în funcție de contextul fonetic în care apare, i se asociază o descriere parametrică. Descrierile parametrice poartă denumirea de primitive de sinteză sau unități acustice, iar formatul de reprezentare al acestora depinde de tipul de sintetizator. Informația prozodică de la intrarea sintetizatorului vocal se referă la durata sunetelor asociate fonemelor, la conturul frecvenței fundamentale f_0 şi la energia semnalului sintetizat.

Există două moduri de realizare a sintetizatoarelor vocale, numite în literatura de specialitate: sintetizatoare bazate pe reguli (*rules-based synthesizer* în limba engleză) şi respectiv, sintetizatoare concatenative. Sintetizatoarele bazate pe reguli au implementate în modulul de generare a semnalelor de comanda pentru sintetizator un set de legi de variație a semnalelor de control al sintezei. Numărul şi semnificația semnalele de comandă sunt determinate de modelul producției vocale pe care se bazează sintetizatorul: sintetizatoare articulatorii, bazate pe modele articulatorii; sintetizatoare formantice sau folosind modulația AM-FM (Potamianos 1997), bazate pe modele acustice în cadrul cărora tractul vocal este modelat în domeniul frecvență.

Sintetizatoarele concatenative se bazează pe înlănțuirea unor primitive de sinteză obținute prin codarea parametrică a unor segmente acustice provenite din rostiri naturale. Segmentele acustice pot fi de următoarele tipuri: morfeme (foneme şi alofoni), difoni, jumătăți de fonem, silabe, cuvinte sau fraze. Sintetizatoarele concatenative comerciale dezvoltate în ultimii ani (Loquendo, InfoVox) folosesc pentru introducerea elementelor prozodice algoritmi de selecție a unităților acustice (în engleză *unit selection*).

3.2 Prezentare generală a sintetizatorului Klatt

Modelul de sintetizator formantic propus de Klatt (1980) şi folosit la realizarea sistemului MITtalk (Allen 1987) pentru limba engleză, a fost preluat în cadrul mai multor sisteme de conversie text-voce, dintre care putem aminti sistemul comercial DECtalk (1983), sistemul de la Speech Technology Laboratory (Javkin 1989) şi sintetizatorul JSRU (Holmes 1983). În literatura de specialitate sunt cunoscute două variante ale sintetizatorului Klatt: modelul Klatt80 şi modelul Klatt88. În figura 3.7 este prezentată o variantă a modelului Klatt80.

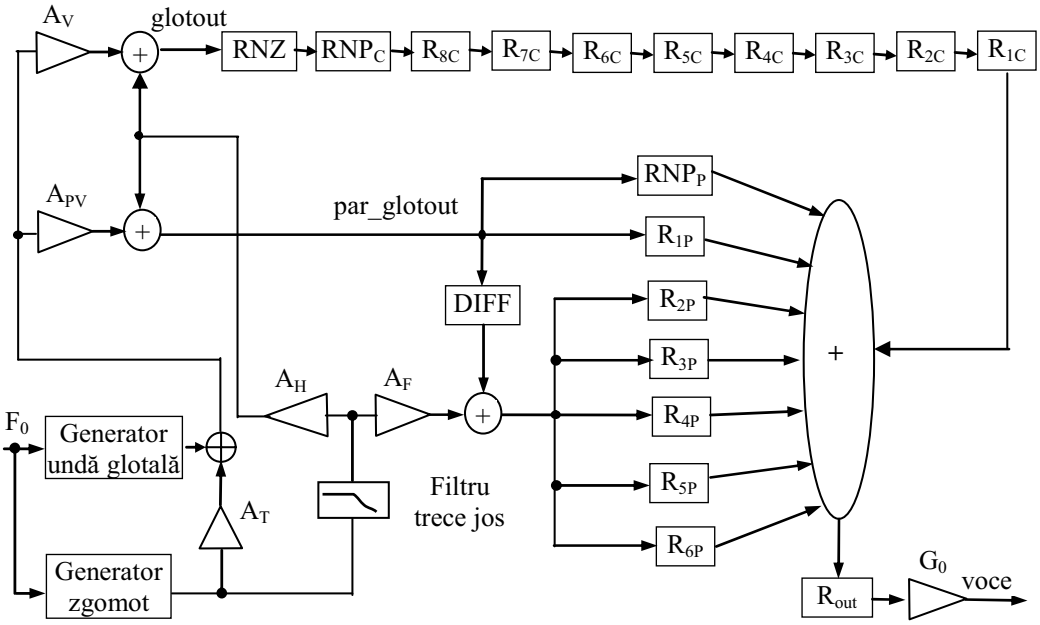

Figura 3.7. Schema bloc a sintetizatorului Klatt (Simmons 1994)

Sursa de excitaţie glotală este formată din două componente: un generator de undă glotală şi un generator de zgomot alb. Sistemul de filtre care modelează efectul tractul vocal şi cavităţilor nazale asupra undei glotale este format din rezonatorii conectaţi pe două ramuri distincte: ramura serie (R_{1C}-R_{8C}) şi ramura paralel (RNP_P, R_{1P}-R_{6P}).

Ieşirea rezonatorilor de pe ramura serie se sumează cu ieşirile rezonatorilor de pe ramura paralel pentru a forma semnalul vocal sintetizat.

3.2.1 Semnale pentru controlul sintetizatorului Klatt

Variabilitatea în timp a semnalul vocal sintetizat se obţine pe baza unor semnale care conţin informaţia despre modificarea în timp a principalelor componente din domeniul spectrului de frecvenţe al semnalului vocal (figura 3.2). Ele sunt asociate componentelor sintetizatorului Klatt (figura 3.8), care contribuie la generarea semnalului vocal sintetizat: semnale pentru controlul formei de undă a excitaţiei glotale ($S_{F0}(t)$ şi $S_{Av}(t)$); semnale pentru controlul filtrelor de pe tractul vocal şi cavitatea nazală ($S_{Tv}(t)$); semnale pentru controlul rezonatorul care modelează fenomenul radiaţiei bucale($S_{rb}(t)$).

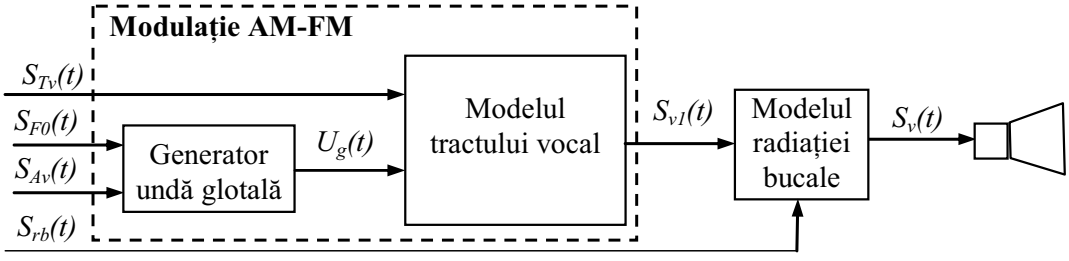

Figura 3.8. Schema semnalelor comandă a sintetizatorului Klatt

Pe baza semnalelor de intrare, ansamblul generator undă glotală, tract vocal realizează o modulație în amplitudine şi frecvență (AM-FM) pentru semnalul vocal. Rezonatorii şi antirezonatorii utilizați pentru modelarea tractului vocal şi cavității nazale, determina o creştere semnificativă a numărului semnalelor de control. Astfel, pentru un număr de o opt rezonatori serie şi şase rezonatori paraleli, cumulat cu semnalele necesare pentru controlul generatorului undei glotale, se obține un număr de 40 semnale la modelul Klat80 şi 48 semnale la modelul Klat88.

3.2.2 Generarea semnalelor de intrare pentru sintetizatorul Klatt

Numărul mare al semnalelor de intrare în sintetizator, influențat de numărul de parametrii la modelul Klatt80, a determinat dezvoltarea de metode pentru generarea acestor semnale astfel încât sunetele sintetizate să se apropie de sunetele naturale. Astfel s-au dezvoltat metode bazate pe modele articulatorii (Stevens 1991, 2002), metode bazate pe modele cu auto-organizare secvențială (Breidegard 2003) şi metode bazate pe reguli euristice. Toate metodele, de generare a semnalelor de intrare în sintetizator, urmăresc evidențierea celor mai importante aspecte pentru realizarea contrastelor fonetice pe baza unor simplificări ale vorbirii naturale.

Folosirea primitivelor de sinteză, simplifică procedura de generare a semnalelor de intrare în sintetizator. Sarcina modului de generare a acestor semnale se reduce la introducerea informațiilor prozodice (intonație, durata foneme, amplitudine) şi a efectelor datorate de co-articularea sunetelor (figura 3.2).

3.3 Sistem de sinteză vocală pentru limba română

Pornind de la o variantă a sintetizatorului Klatt (dezvoltată de Simmons 1994), începând din anul 2000, ne-am propus să realizăm în cadrul Institutului de Informatică Teoretică Iaşi un sistem integrat cu ajutorul căruia să putem studia probleme de analiză şi sinteză a semnalului vocal. În prima fază s-a proiectat o interfață grafică pentru vizualizarea şi analiza rezultatelor obținute din procesarea semnalului vocal. Ulterior am proiectat şi implementat o interfață grafică (figura 3.13) pentru vizualizarea şi modificarea semnalelor de la intrarea sintetizatorului, cu scopul de a studia efectul acestora în semnalul vocal sintetizat şi pentru îmbunătățirea descrierilor parametrice, a sunetelor limbii române, pentru sintetizatorul Klatt.

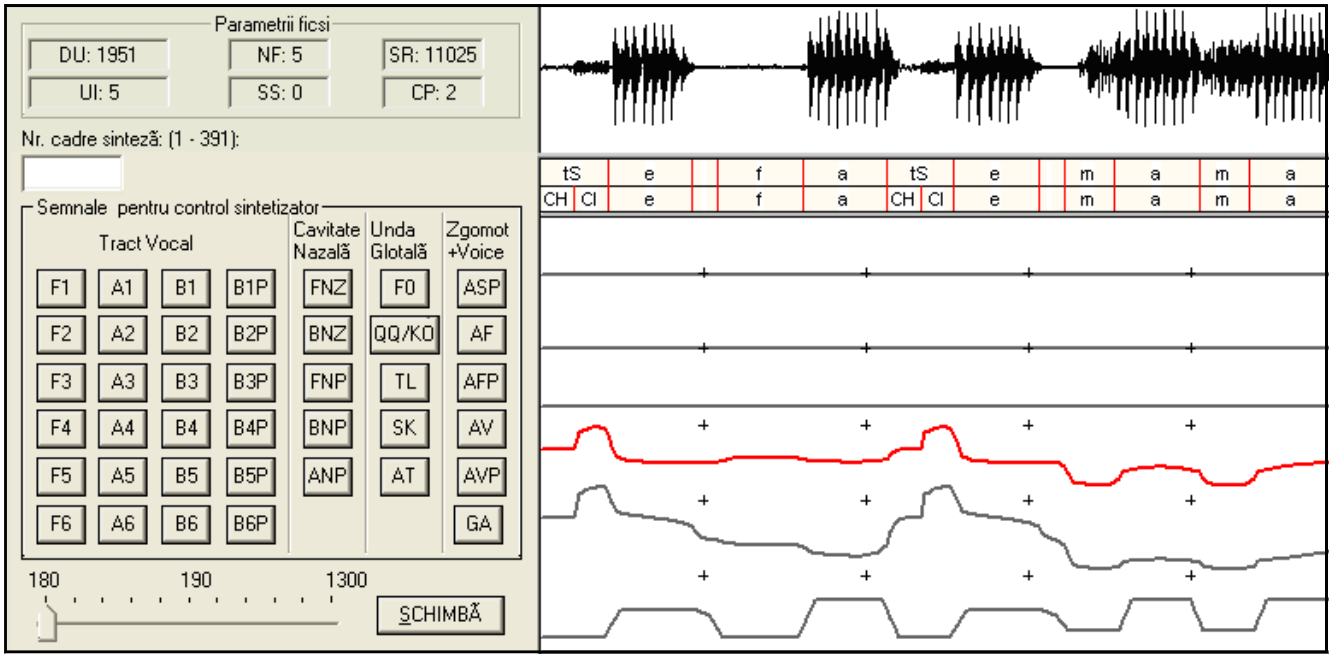

Figura 3.13. Interfață grafică pentru vizualizarea şi modificarea semnalelor de intrare în sintetizatorul Klatt

Până în anul 2003 am participat la proiectarea, implementarea şi testarea următoarele module din cadrul sistemului de analiză şi sinteză vocală pentru limba română (Jitcă 2003):

- implementarea unei interfețe grafice a utilizatorului cu sistemul de analiză şi sinteză vocală;
- implementarea unor funcții de analiză spectrală şi temporală a semnalului vocal;
- definirea parametrilor pentru primitivele de sinteză asociate fonemelor limbii române aflat la baza sintetizatorului vocal;

proiectarea, implementarea şi testarea primei variante a sistemului de conversie

- text-voce pentru limba română;
- implementarea şi testarea modulului de despărțire în silabe şi de poziționare a accentului în cadrul cuvintelor prezentat în lucrarea (Jitcă ş.a. 2002b);
- implementarea unui modulul de editare a conturului F0, pentru analiza intonației în limba română, publicat în lucrarea (Jitcă ş.a. 2002e).
- Implementarea algoritmului de adaptare la contextul fonetic a frecvenței centrale a formanților F2 şi F3 pentru fonemele /l/ şi /r/ (2002a)

După această etapă, cercetările efectuate pentru sintetizatorul formantic au vizat analiza posibilităților de modelare a co-articularii sunetelor, implementarea unui modul software pentru îmbunătățirea tranzițiilor formanților între foneme (Apopei ş.a. 2004a) şi analiza posibilităților de implementare a elementelor prozodice prin controlul conturului frecvenței F0, duratei silabelor, duratei pauzelor şi energiei fonemelor.

3.3.1 Modelarea co-articulării sunetelor

Analiza în domeniul frecvență a undelor vocale naturale evidențiază influențe ale frecventelor centrale ale formanților între sunetele (fonemele) vecine. Aceste influențe se materializează prin modificarea, între anumite limite, a valorilor de stabilitate ale formanților şi prin tranziții între valorile de stabilitate la trecerea de la un fonem la altul. În literatura de specialitate, aceste efecte naturale care apar în timpul producției vocale poartă numele de co-articularea sunetelor.

Co-articularea sunetelor se datorează efectelor de inerție care apar în mişcarea, fără eforturi deosebite din partea vorbitorului, a unor organe implicate în procesul de vorbire: buzele, vălul paltin, limba, maxilare cu sistemul de masticație şi laringele. Aceste mişcări determină efecte anticipatorii sau de influență către sunetele următoare (în engleză „*carry-over*") la nivelul evoluției frecvențelor centrale ale formanților. Majoritatea cercetărilor care abordează fenomenul co-articulării sunetelor se bazează pe caracterul articulator al mecanismului producției vocale şi leagă co-articularea de împărțirea în silabe a cuvintelor.

La o trecere mai atentă asupra modelelor care abordează problematica co-articulării sunetelor din punct de vedere al producției şi percepției vocale se poate constata că acestea se pot clasifica în patru categorii: metode care modelează co-articularea sunetelor din punct de vedere al percepției auditive fără modelarea tractului vocal (Delattre şi Liberman 1955, Klatt 1979, 1987, Wickelgren 1969); metode care corelează mişcarea organelor implicate în procesul de vorbire cu modificările frecvențelor centrale ale formanților (Öhman 1966, Stevens 1994, 2002, Carré 1999); metode (în engleză *visual speech synthesis)* care modelează co-articularea sunetelor din punct de vedere al percepției vizuale cu modelarea unor articulatori ai producției vocale şi ai mimici feței (Löfqvist 1990, Pelachaud, Badler şi Steedman 1991, Cohen & Massaro 1993/2003); metode (în engleză *audio-visual speech synthesis* sau *talking head)* care corelează modele articulatorii al producției vocale cu percepția acustică şi vizuală a vorbirii (Cohen & Massaro 2003, Beskow 2003, Fagel 2003).

Modelele din prima categorie au fost dezvoltate pe baza următoarelor teorii: *locus*

theory (Delattre 1955, Klatt 1979/1987); teoria trăsăturilor extinse (*feature-spreading,* Henke 1966); teoria alofonilor (Wickelgren 1969/1972); teoria rezistentei la co-articulare (Bladon & Al-Bamerni 1976; Hawkins 1994/2000); teoria constrângerilor articulatorii (Recanses 1987);

3.3.2 Îmbunătățirea sintezei vocale formantice prin introducerea tranzițiilor neliniare în generarea semnalelor F2 şi F3

Pornind de la posibilitățile oferite de sintetizatorul formantic de tip Klatt şi de la analiza modelelor care abordează co-articularea fonemelor, în lucrarea (Apopei 2004a) ne-am propus să realizăm o modelare, cu funcții neliniare de dominanță, a variației formanților F2 şi F3 la sintetizatorul Klatt. Pentru aceasta am redefinit noțiunea de dominanță. În varianta originală a sintetizatorului Klatt fiecare fonem are stabilită o valoare ce exprimă dominanța şi este folosită în calculul traseelor liniare de tranziție (secțiunea 3.2.4.).

La sintetizatorul Klatt valorile dominanței determină modul de realizare a tranziției liniare între valorile de stabilitate ale fonemelor vecine, mai precis, impune durata şi panta celor două segmente lineare ce o compun. Dominanța este dată de valoarea unei variabile din structura de date asociată cu definiția fiecărui fonem-element sau parte dintr-un fonem. Pe baza valorilor variabilei în cadrul setului de foneme s-au constatat 3 categorii de dominanță: dominanțe mari - corespunzătoare valorilor 17-26; dominanțe medii - corespunzătoare valorilor 10-17; dominanțe mici - corespunzătoare valorilor mai mici de 10.

Conform, teoriei co-articulatorii a sunetelor, pe durata sunetelor dominante, articulatorii îşi ating pozițiile țintă (punctul de articulare corespunzător rostirii izolate a acestora) şi se mențin pe acestea un anumit interval de timp (de stabilitate). Pe durata celor slab dominante pozițiile articulatorilor pot varia continuu dinspre sau/şi spre pozițiile fonemelor vecine dominante, dacă diferențele între pozițiile lor țintă sunt mari.

Modelarea neliniară a porțiunilor de tranziție pe care am aplicat-o sintetizatotului Klatt, realizează o variație a formanților F2 şi F3 , între două foneme vecine, mai apropiată ca evoluție de variația observabilă pe semnale naturale. Astfel pentru un caz concret al co-articulării fonemelor m/ şi /i/ (la care variațiile de frecvență ale formanților F2 şi F3 între foneme sunt de aproximativ 1000 Hz) tranzițiile au fost modificate (figura 3.15) de la forma inițială trasată cu linie punctată, la forma nouă trasată cu linie continuă. Această nouă evoluție a valorilor frecvențelor centrale ale formanților este mai apropiată de evoluția naturală şi se poate explica prin mişcarea articulatorilor.

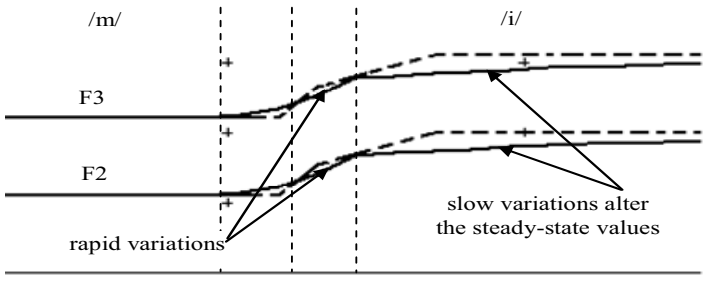

Figure 3.15 Conturul frecvențelor formantice F2 şi F3 la tranziția dintre sunetele /m/ şi /i/

În modul nou de tratare a co-articulării, pentru un fonem în zona de stabilitate se pot obține valori diferite pentru formanți în funcție de contextul fonetic. De asemenea formanții pot avea variații continue în jurul valorii de stabilitate. Cu alte cuvinte, porțiunii de stabilitate a unui formant nu îi mai corespunde o singură valoare în cadrul sintezei ci, în cazul general, o gamă de variație. În reprezentarea grafică din figura 3.17 se ilustrează modul cum se influențează fonemele adiacente în carul sintezei cuvântului 'ape'.

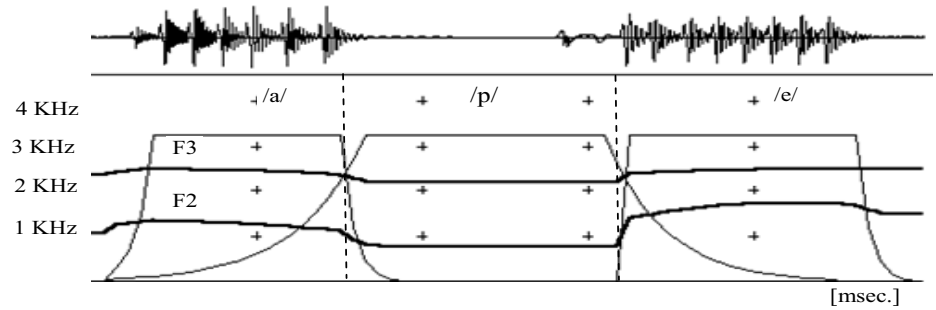

Figura 3.17 Forma de undă (sus), funcțiile de dominanță (jos, cu linie subțire) şi traseele frecvențelor formantice F2, F3 (jos, cu linie groasă) din sinteza rostirii cuvântului 'ape'

Funcțiile de dominanță corespunzătoare vocalelor (foneme cu dominanță slabă) au durate scurte de revenire la zero iar la consoana /p/ (formată din 3 segmente) fronturile dominanțelor spre şi dinspre fonemele vecine evoluează pe toată durata acestora. Fronturile funcțiilor de dominanță din stânga şi din dreapta fonemului nu sunt identice, ci se stabilesc funcție de cel de-al doilea fonem implicat în tranziție. De aceea fronturile funcției de dominanță ale vocalei /e/ nu sunt identice. Astfel explozia consoanei oclusive impune un front abrupt (în partea stângă) spre deosebire de cel din dreapta.

3.4 Contribuții personale

Contribuțiile din acest capitol sunt legate de necesitatea modelării tranzițiilor dintre foneme şi de analiza posibilităților de implementare a elementelor prozodice la sintetizatorul formantic Klatt. Analiza în domeniul frecvență a undelor vocale naturale a pus în evidență influențe ale frecventelor centrale ale formanților între sunetele (fonemele) vecine. Aceste influențe se materializează prin modificarea, între anumite limite, a valorilor de stabilitate ale formanților şi prin tranziții între valorile de stabilitate la trecerea de la un fonem la altul. În literatura de specialitate, aceste efecte naturale care apar în timpul producției vocale poartă numele de co-articularea sunetelor. Din punct de vedere al fenomenului producției vocale aceste influențe se explică cu ajutorul efectelor de inerție care apar în mişcarea, fără eforturi deosebite din partea vorbitorului, unor organe implicate în procesul de vorbire: buzele, vălul paltin, limba, maxilare cu sistemul de masticație şi laringele.

După etapa de analiză a principalelor metode şi teorii existente pentru modelarea co-articulării sunetelor, am ajuns la concluzia că pentru cazul sintetizatorului formantic este de interes găsirea unei metode de modelare a efectului co-articularii din punct de vedere al percepției auditive dar care să țină cont de fenomenul producției vocale.

Pornind de la posibilitățile oferite de sintetizatorul formantic de tip Klatt şi de la analiza modelelor care abordează co-articularea fonemelor, în lucrarea (Apopei 2004a) am propus o modelare, cu funcții neliniare de dominanță, a variației formanților F2 şi F3 la sintetizatorul Klatt. Această modelare a fost inspirată din modelul Cohen şi Massaro (1993, 2003).

Folosind această modelare a variației formanților, la tranziția dintre foneme, am reuşit să îmbunătățesc calitatea semnalelor vocale sintetizate cu ajutorul sintetizatorul formantic de tip Klatt.

Analiza elementelor componente ale unui sintetizator (figura 3.1) şi a posibilităților de control a parametrilor la sintetizatorul formantic Klatt, m-a condus la ideia de a realiza implementarea elementelor prozodice cu ajutorul unor submodule, care să fie incluse în modulul fonetic din componența sistemului de conversie text-voce .

Capitolul 4.

Analiza prozodiei. Modele prozodice

Elementele prozodice studiate şi implementate în sistemele *text-to-speech* sunt derivate din caracteristici acustice ale vocii. Cele mai importante elemente prozodice luate în considerare de modelele prozodice sunt: intonația, intensitatea sunetelor, durata silabelor (fonemelor) şi a pauzelor. Intonația este o caracteristică acustică a semnalelor vocale dată în principal de variația frecvenței fundamentale F0 şi depinde de modul în care vorbitorul realizează frazarea (gruparea) şi accentuarea cuvintelor. Implementarea intonației în sinteza vocală presupune generarea automată a "melodiei" corespunzătoare rostirii unui text, pe baza unor modele intonaționale care pun în corespondență structura sintactică şi semantică a textului cu un set de evenimente intonaționale şi un set de pattern-uri la nivelul frecvenței F0.

Modelele intonaționale realizează o reprezentare fonologică a vorbirii pe baza unor relații între funcțiile şi formele (evenimentele) prozodice şi intonaționale (Hirst 2007, Shih 2006, Batliner 2003). Evenimentele prozodice sunt reprezentate prin variații în timp ale elementelor prozodice. Relațiile stabilite în cadrul modelelor intonaționale urmăresc asocierea de evenimente prozodice pentru transmiterea unor stări emoționale şi de atitudine prin voce, dezambiguizarea componentelor verbale ale comunicației. Evenimentele prozodice se evidențiază la nivelul vorbirii prin: modul de frazare a rostirilor, proeminențele accentelor (stabilesc funcțiile cuvintelor în rostire), modul de realizare a tonurilor de graniță. Există numeroase cercetări care încearcă să asocieze evenimentele intonaționale şi prozodice cu structura sintactică, semantică şi/sau de discurs a unui text (Bachenko şi Fitzpatrick 1990, Wang şi Hirschberg 1992, Ostendorf şi Veilleux 1994, Taylor şi Black 1998, Heusinger 1999, Taylor ş.a 2006, Gussenhoven 2007, Steedman 2000, ş.a)

Modelele prozodice rezultate prin analiza semnalului vocal sunt de interes atât pentru îmbunătățirea performanțelor în sistemele de recunoaştere vocală (ASR) cât şi a celor din domeniul sintezei vocale. În prima direcție se lucrează pentru a se crea modele prozodice care să permită identificarea granițelor unităților intonaționale, constituind indicii clare de final de cuvânt sau propoziție/frază. Acestea completează modelul acustic şi de limbaj folosit în cadrul sistemelor de recunoaştere bazate pe HMM.

4.1 Modele intonaționale şi prozodice

Modele intonaționale şi prozodice dezvoltate după 1980 au la bază teoriile fonetice şi fonologice. Dintre acestea, cele mai utilizate sunt fonologia metrică a lui Liberman introdusă pentru studiul ritmului (1975), preluată de Ladd şi aplicată intonației (1983) sau cea autosegmental-metrică bazată pe secvențe de tonuri a lui Pierrehumbert (1980), Ladd 1996. În cadrul fonologiei metrice, intonația este văzută prin proeminențe relative (de tip *weak, strong*), realizate prin evenimente tonale la nivelul frecvenței F0, între grupuri de unități segmentale (foneme). Unitățile segmentale sunt grupate în silabe, silabele în cuvinte fonologice, iar cuvintele în unități din ce în ce mai mari până se ajunge la fraza intonațională. În acest mod se poate asocia rostirii unui text o anumită ierarhie intonațională (Di Cristo 2004).

Împărțirea rostirii unui text în fraze intonaționale şi stabilirea unităților intonaționale segmentale proeminente, din cadrul unei fraze intonaționale, este cunoscută în literatura de specialitate ca *frazare* sau *frazare prozodică* (*prosodic phrasing)*.

În continuare voi face o scurtă trecere în revistă a celor mai cunoscute modele intonaționale şi prozodice folosite în adnotarea corpusurilor de voce, sinteza şi recunoaşterea vocală.

4.1.1 Modele fonologice

Cu ajutorul modelele fonologice se realizează descrieri discrete ale intonației pe bază de evenimente fonologice organizate după structuri ierarhice. În cadrul acestor ierarhii, unitatea intonațională pentru care se urmăreşte analiza intonației în vederea predicției elementelor prozodice este fraza intonațională. În cadrul frazelor intonaționale, evenimentele fonologice de pe conturul frecvenței fundamentale F0 au asociate un set de etichete şi un set de primitive de contur reprezentative pentru intonație.

Dezvoltarea inventarului de evenimente fonologice se bazează pe analiza fonetică a conturului frecvenței F0 din perspectiva producției şi a percepției vocale. Cele mai cunoscute modele intonaționale din această categorie sunt: modelul propus de Pierrehumbert (1980), modelul ToBI (Silverman ş.a. 1992, Backman ş.a 1993) şi modelul propus de Ladd (1996).

4.1.2 Modele fonetice bazate pe reprezentări numerice

Modelele fonetice realizează descrierea intonației printr-un set de parametrii care variază continuu pe durata unei fraze intonaționale. Pentru a fi funcționale modelele fonetice folosesc pentru predicția parametrilor descrieri fonologice sau trăsături lingvistice. Din punct de vedere al modului în care este percepută realizarea intonației, modelele fonetice se pot clasifica în liniare şi bazate pe principiul superpoziției.

Descrierile bazate pe principul superpoziției tratează intonația ca rezultanta sumei a două componente importante: intonația la nivelul frazei intonaționale şi intonația cuvintelor. Spre deosebire de acestea, modelele fonologice consideră intonația realizată printr-o secvență de pattern-uri elementare de contur F0 care corespund unor evenimente intonaționale.

4.1.3 Modele fonetice bazate pe principiul superpoziției

Cele mai reprezentative modele intonaționale bazate pe principiul superpoziției sunt modelul Öhman (1967) şi modelul Fujisaki (1983, 2004) Aceste modele consideră, conturul frecvenței F0 ca o rezultantă a sumării mai multor componente intonaționale. Dintre acestea cele mai importante componente se referă la intonația frazei intonaționale şi intonația corespunzătoare accentului de cuvânt.

Modele bazate pe principiul superpoziției diferă între ele prin componentele intonaționale din a căror suprapunere se obține conturului intonațional. Astfel unele modele completează setul de componente ale modelului Fujisaki (Thorsen 1983,1995, Santen 2002) iar altele folosesc componente diferite (Gårding 1983, Bruce 1984).

Pentru alinierea valorilor frecvenței fundamentale pe cele trei tipuri de curbe, se foloseşte o structură cu repere de timp şi durate pentru fiecare segment şi fonem.

4.1.4 Modele prozodice bazate pe informații semantice şi fonologice

Modelul prozodic propus de Batliner (1998, 2003) pentru sistemul Verbmobil, propune ca informațiile prozodice să fie asociate unor segmente de vorbire mai mari decât fonemele, cum ar fi: silabele, cuvintele, frazele intonaționale şi segmente de vorbire fără echivalent sintactic (*whole turns*). Segmentele au asociate o serie de proprietăți ca: tăria, frecvența fundamentală, rata vorbirii, calitatea vocii, durata, ritmul ş.a. Aceste proprietăți sunt corelate cu următoarele trăsături acustice: frecvența fundamentală, energia semnalului vocal, frecvența trecerilor prin zero, ş.a.

Cele mai importante evenimente şi aspecte prozodice avute în vedere de acest model sunt: stabilirea granițelor şi a tipului acestora; stabilirea accentelor şi a tipului acestora; tipul

de propoziție (afirmativ/interogativ); starea emoțională a vorbitorului. Pentru adnotarea informațiilor prozodice, Batliner propune folosirea următoarelor clase de etichete:

- ✔ etichete pentru granițe acustico-prozodice care se stabilesc pe baza de trăsături acustice.
- ✔ etichete pentru granițe sintactico-prozodice care se stabilesc pe bază de trăsături sintactico-semantice.
- ✔ etichete pentru tipuri de accente.
- ✔ etichete pentru tipul propoziției.

Pentru etichetarea automată a corpusurilor de voce pe baza acestui model, Batliner (2003) a utilizat o rețea neuronală de tip perceptron multistrat, care foloseşte ca intrare un set de 95 de trăsături prozodice şi 30 de trăsături ale părților de vorbire. Trăsăturile prozodice au fost determinate pe ferestre de diferite lungimi (la nivel de silabă sau la nivel de cuvânt), iar la intrarea rețelei s-au luat în considerare valorile trăsăturilor de pe cinci ferestre de analiză (fereastra curentă, două ferestre anterioare, două ferestre posterioare). Părțile de vorbire au fost împărțite în 6 clase după cum urmează: AUX (cuvinte auxiliare); PAJ (particule, articole, interjecții); VERB(verbe); APN (adjective şi participii neflexionate); API (adjective şi participii flexionate); NOUN (substantive proprii şi comune).

Modelul KIM

Modelul KIM (**K**iel **I**ntonation **M**odel) a fost dezvoltat la Universitatea din Kiel pentru a furniza informații fonologice despre prozodia în limba germană (Kohler 1997) în cadrul proiectului Verbmobil. Modelul corelează informații despre următoarele elemente fonetice şi fonologice: accentul lexical; accentul propoziției; intonația; sincronizarea formei evenimentelor de pe conturul frecvenței F0 „*peaks*" şi „*valleys*" cu silabele accentuate; informații despre granițele prozodice exprimate; viteza de vorbire între granițele prozodice; tendințele de downstep sau upstep al succesiunilor *„peaks"/"valleys"* şi evenimentul de „*pitch reset*" de pe conturul frecvenței F0.

Pentru transcrierea prozodiei cu acest model Kohler (1991) a folosit sistemul de etichetare PROLAB dezvoltat pentru adnotarea corpusurilor de voce în limba germană. Kohler (1997) propune pentru generarea intonației un sistem bazat pe două nivele:

- definirea unor pattern-uri prozodice controlate fonologic printr-un număr mic de puncte semnificative de pe conturul frecvenței F0 (la nivel macroprozodic).
- conturul frecvenței F0 rezultă prin concatenarea acestor pattern-uri fonologice.

Modelul KIM şi setul de etichete PROLAB au stat la baza sistemului de conversie text-voce INFOVOX şi au fost utilizate pentru analiza şi modelarea prozodiei în limba germană. Folosind acest model Kohler (2005) pune în evidență o legătură între funcțiile comunicative ale prozodiei şi formele de pe conturul intonațional pe baza unei analize a contextului semantic şi pragmatic a transmiterii mesajului de vorbitor către ascultător.

4.2 Modelarea duratei sunetelor şi pauzelor

Durata sunetelor este corelată cu viteza de vorbire (rapiditatea vorbirii), în engleză *speech rate*. Un model al duratei sunetelor trebuie să țină cont de o limită inferioară impusă de inerția/mobilitatea articulatorilor implicați în producerea lor (mişcarea buzelor şi a limbii). Duratele medii ale fonemelor variază între 20 msec pentru consoanele plozive sonore, până la 150 msec pentru diftongi, cu o durată medie a fonemelor de 75 msec. La vocale, durata variază funcție de context între valori aflate într-un raport de 1/8 şi depinde de silaba în care se află. Kanedera ş.a (1997) au pus în evidență faptul că modulația perceptuală cea mai importantă a vorbirii (modificările cele mai importante în semnalul vocal) este

realizată în jurul valori de 4-5 Hz, sau 200-250 msec cât este aproximativ durata unei silabe (Greenberg 1996, Arai 1997). Dacă se iau în considerare multitudinea factorilor care influențează duratele şi percepția fonemelor, rezultă modele relativ complexe.

Modelul propus de Klatt pe baza relației (4.6) foloseşte 7 factori şi 11 reguli pentru modificarea duratei fonemelor dintr-o propoziție (Klatt 1979, 1987).

$$DUR = MINDUR + \frac{(INHDUR - MINDUR)* PRCNT}{100} \tag{4.6}$$

unde: *MINDUR* este durata minimă a fonemului accentuat;

INHDUR este durata intrinsecă a fonemului ;

PRCNT este procentul de micşorare sau creştere aplicat pe baza celor 11 reguli.

Santen (1997) a rescris relația (4.6), pentru modelarea duratei grupurilor CV sub forma:

$$DUR(v,c,p) = S_{1,1}(v)S_{1,2}(c)S_{1,3}(p) + S_{2,1}(v) \tag{4.7}$$

unde: $S_{1,1}(v)$ este durata netă a fonemului *INHDUR- MINDUR;*

$S_{1,2}(c)$ este o constantă care depinde de consoană precedentă;

$S_{1,3}(p)$ este o constantă care depinde de poziția în frază;

$S_{2,1}(v)$ este durata minimă a unei vocale.

Această scriere este în concordanță cu modelul "*sumă de produse*" propus de Santen (1993). Conform acestui model, durata unui fonem este dată de relația (4.8):

$$DUR(d) = \sum_{i \in K} \prod_{j \in I_i} S_{i,j}(d_j) \tag{4.8}$$

Pentru predicția duratei sunetelor s-au dezvoltat şi metode bazate pe sisteme de învățare: rețele neuronale (Campbell 1992) şi arbori de regresie (Bagshaw 1998, Strom 2002).

4.3 Modelarea intensități sunetelor

Modelele dezvoltate pentru reprezentarea tăriei (intensității) sunetelor provin din cercetările efectuate în domeniul modelării psiho-acustice a vorbirii. Aceste modele sunt folosite cu succes în domeniul recunoaşterii vocale şi analizei rostirilor emoționale. În domeniul sintezei vocale, pentru predicția tăriei sunetelor s-au dezvoltat modele bazate pe reguli (Dohalská M., ş.a. 2001), modele bazate arbori de regresie (Bagshaw 1998) metode statistice cu HMM etc.

Modelul bazat pe arbori de regresie propus de Bagshaw (1998) se bazează pe determinarea a doi parametrii asociați silabei (p = proeminența, l = lungimea). Aceşti parametrii sunt estimați, prin metode statistice, pe baza energiei semnalului vocal şi duratei silabelor extrase de pe corpusuri de semnal vocal. Pe lângă aceşti doi parametrii, pentru estimarea energiei cu relația (4.9) fonemele se împart în mai multe categorii. Categoriile de împărțire a fonemelor se stabilesc în funcție de următorii parametrii: eticheta fonemului; contextul de grup în care apare fonemul respectiv (grup consoane, grup vocale, consoană-vocală; poziția fonemului şi grupului în silabă (*onset, nucleu, coda*); poziția silabei în cadrul cuvântului (finală, nonfinală).

$$\hat{e}_i = e_{mi} + (w'_{mi} \cdot p + w''_{mi} \cdot l) \cdot \sigma_{mi} \tag{4.9}$$

unde: e_{mi} = energia medie a fonemului din categoria *i;*

σ_{mi} = deviația standard medie a variației energiei pentru fonemul din categoria *i;*

w'_{mi}, w''_{mi} = coeficienți de ponderare a parametrilor p şi l pentru fonemul din categoria *i.*

4.4 Descrierea contururilor intonaționale în limba română

În limba română, ca şi în cazul altor limbi europene, în studiile de fonologie (Gramatica Academiei Române 2005, L.Dascălu-Jinga 2001, Turculeț 1999) există doar descrieri ale contururilor intonaționale (contururi stilizate ale frecvenței F0) pentru diverse tipuri de rostiri, cum ar fi cele declarative neutrale, interogative, exclamative etc.

Scopul cercetărilor efectuate, în ultimii ani la Institutul de Informatică Teoretică, în domeniul modelării intonației a fost acela de a realiza descrieri fonologice pe baza cărora să putem trece la implementarea intonației în sinteza vocală pentru limba română.

În cadrul acestei secțiuni vom prezenta o ierarhie pentru unitățile intonaționale şi câteva exemple de etichetare a intonației pentru limba română.

4.4.1. Prezentarea ierarhiei unităților intonaționale

Pentru adnotarea contururilor intonaționale am utilizat o ierarhie intonațională (figura 4.4) care să poată fi implementată şi în format XML (Apopei ş.a. 2006b, 2006c). În cadrul acestei ierarhii, cea mai mică unitate căreia i se poate asocia un eveniment prozodic este silaba. Silabele constituie părți componente ale cuvintelor. Cuvintele sunt purtătoare ale accentelor sintactice sau lexicale. Cuvintele se grupează în unități de accentuare (AU). Unitățile de accentuare cuprind un cuvânt cu accent şi unul sau mai multe cuvinte clitice. Există situații în care unitățile de accentuare pot include pe lângă cuvântul accentuat, un alt cuvânt neclitic, dar care şi-a pierdut complet accentul în vecinătatea acestuia.

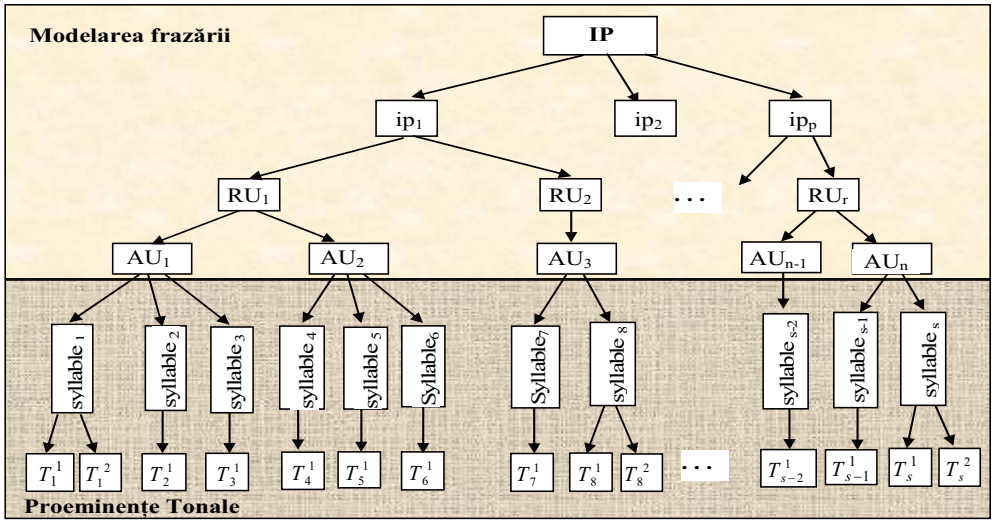

Fig. 4.4. Ierarhia unităților intonaționale utilizată pentru descrierea conturului frecvenței F0

O unitate de accentuare purtătoare de accent puternic se grupează în cadrul acestei ierarhii cu alte unități care includ cuvinte purtătoare de accente mai slabe, formând unități ritmice (Di Cristo 2004), sau grupuri de unități de accentuare. Aceste grupări sunt susținute şi de existența unor grupuri sintactico-semantice diferite la nivelul textului (grup verbal, grup nominal, grup adjectival etc.). La nivelul conturului frecvenței F0, unitățile ritmice (grupuri de unități de accentuare) delimitează segmente de contur cu pattern-uri specifice semantici intonaționale a rostirii. Una sau mai multe unități ritmice compun o frază intonațională (*intonational phrase*, în limba engleză şi notată, IP), sau o frază intonațională intermediară (*intermediate phrase*, în limba engleză şi notată, ip).

În lipsa unor definiții explicite pentru frazele intonaționale (*intonational phrase)* şi frazele intonaționale intermediare (există doar exemplificări ale acestora pe cazuri particulare de contururi F0), am caracterizat aceste unități intonaționale într-o manieră

funcțională, cu care să putem opera în analiza contururilor naturale şi apoi în sinteza vocală. Astfel, urmărind tendințele de creştere (upstepping) sau scădere (downstepping) a tonurilor țintă din cadrul accentelor de pitch sau lexicale, am identificat puncte de pe conturul F0 în care se produce o schimbare a tendințelor în curs şi plasarea evoluției tonurilor țintă pe o nouă tendință de creştere/descreştere.

Dacă aceste puncte nu coincid cu sfârşitul frazei intonaționale, atunci acestea au fost marcate ca sfârşituri frază intermediară (în sistemul ToBI, tonurile de sfârşit de frază intermediară - *phrase accent*). Identificarea finalurilor de fraze intonaționale este mai uşoară atunci când sunt însoțite de pauze scurte). Frazele intonaționale sunt urmate de pauze mai lungi după tonurile finale (în sistemul ToBI aceste sunt numite tonuri de graniță - *boundary tones*).

La nivelul rostiri propozițiilor afirmative dintr-un text am identificat următoarele secvențe de tonuri pentru unitățile ritmice (4.10):

[H L+!H*] , [H* L*+!H]* (4.10)

În această relație se observă că unitatea de accentuare de la începutul unității ritmice are un accent de pitch, de tip H*, cu un ton țintă mai înalt iar ultima unitate de accentuare are fie un accent de pitch de tip L+H* sau L*+H cu tonul țintă High mai jos decât cel al primei unități de accentuare. Unitățile ritmice pun în evidență contraste tonale locale între două unități de accentuare.

În lucrarea (Apopei ş.a 2005a) am folosit exemplul rostirii naturale a textului "*Avea sentimentul că mai fusese prin cartierul respectiv odată ...*" al cărei contur F0 este prezentat în figura 4.5. Folosind perspectiva dată de o ierarhie intonațională cu două nivele nu am putut decât să împărțim fraza intonațională care cuprinde această porțiune de frază, în cinci fraze intermediare, formate din două şi respectiv câte o unitate de accentuare (varianta 4.11). Se observă însă că în rostirea acestui text, unitățile de accentuare se grupează câte două prin asocierea unui accent de pitch ce tinde mai repede la punctul țintă high cu unul mai lent în ridicarea spre punctul propriu țintă *high*. Această succesiune de combinații de tonuri țintă formează ritmul frazei intonaționale. În consecință pentru a reda mai bine sensul melodic al frazei am împărțit fraza din exemplul de mai sus în două fraze intermediare, iar a doua subîmpărțită în două unități ritmice (varianta 4.12). În redarea celor două variante de frazare a textului cu „/" s-au separat unitățile de accentuare, cu paranteze rotunde am delimitat unitățile ritmice, cu paranteze pătrate am încadrat frazele intermediare iar cu acolade, fraza intonațională.

{[Avea sentimentul]$_{ip}$ [că mai fusese] $_{ip}$ [prin cartierul] $_{ip}$ [respectiv] $_{ip}$ [odată] $_{ip}$}$_{IP}$ (4.11)

{[(Avea/sentimentul) $_{RU}$]$_{ip}$ [(că mai fusese/ prin cartierul)$_{RU}$ (respectiv/ odată) $_{RU}$] $_{ip}$}$_{IP}$ (4.12)

În figura 4.5 conturul F0 este adnotat din perspectiva ierarhiei cu 4 nivele în maniera descrisă de varianta (4.12). În această interpretare grupând ultimele patru unități de accentuare câte două creăm posibilitatea de a le pune într-o anumită relație melodică, de a identifica un pattern melodic ce apoi să poată fi reprodus în sinteză. În cazul nostru este vorba de o combinație de două tipuri de accente, unul care ridică tonul mai repede şi celălalt mai întârziat. Cu trei nivele de unități peste nivelul unității de accentuare se poate urmări cu conturul melodic mai bine sintagmarea textului pe mai multe nivele.

În analiza contururilor frecvenței F0 am avut în vedere următoarele evenimente intonaționale: accentele de pitch, produse pe durata silabelor accentuate (în engleză „*pitch accent*"); tonurile de sfârşit ale frazelor intonaționale intermediare; tonurile graniță ale frazelor intonaționale; alte tonuri semnificative din conturul F0 (în engleză „*target ton*"), care se pot afla fie pe silaba anterioară silabei accentuate, fie pe silaba următoare. Pentru

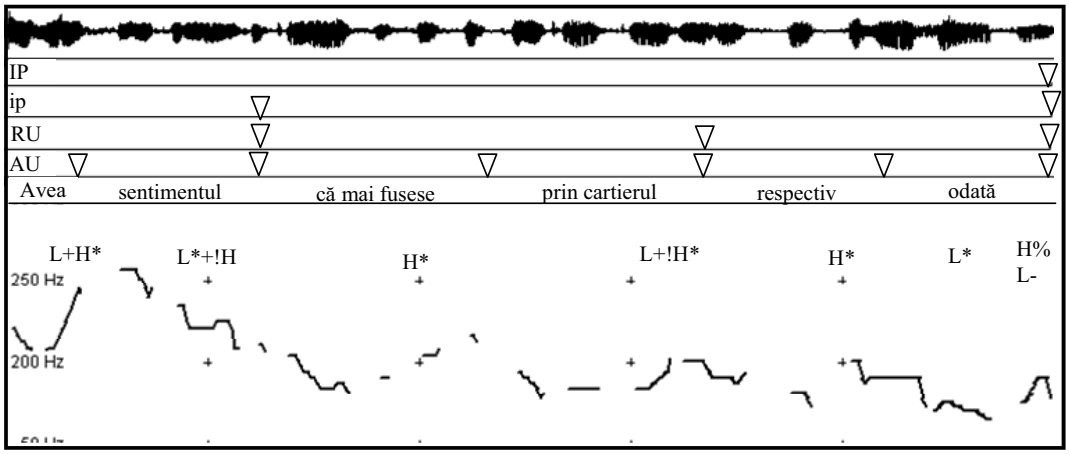

Figura 4.5. Conturul frecvenței F0 pentru rostirea textului
"*Avea sentimentul că mai fusese prin cartierul respectiv odată …*"

marcarea primelor trei tipuri de evenimente s-au folosit etichetele sistemului de adnotare ToBI (completat cu GToBI - *German ToBI*), iar pentru ultima categorie s-au adăugat două etichete, H+ şi L+, care au fost folosite şi în alte aplicații de adnotare prozodică (Baumann S. ş.a 2004).

4.4.2. Etichete pentru accentele de pitch

- **H***

H* este eticheta pentru accentul ce se formează printr-o creştere semnificativă a frecvenței F0 (peste nivelul unui simplu accent gramatical), pe durata unei silabe accentuate. Creşterea se poate realiza fie prin variație continuă începând cu vocala silabei accentuate atingând valoarea maximă între mijlocul şi sfârşitul vocalei, fie prin salt crescător al frecvenței când vocala silabei accentuate este precedată de o consoană nesonoră. Forma sub care acest tip de accent se identifică cel mai uşor este cea de vârf cu un ton central ridicat față de cele ale silabelor neaccentuate vecine, ca în cazul cuvântului /*domnilor*/ (figura 4.6).

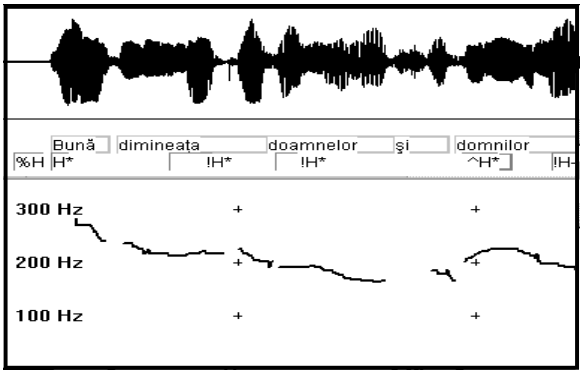

Figura 4.6. Unda vocală şi conturul frecvenței fundamentale al rostirii "Bună dimineața doamnelor şi domnilor"

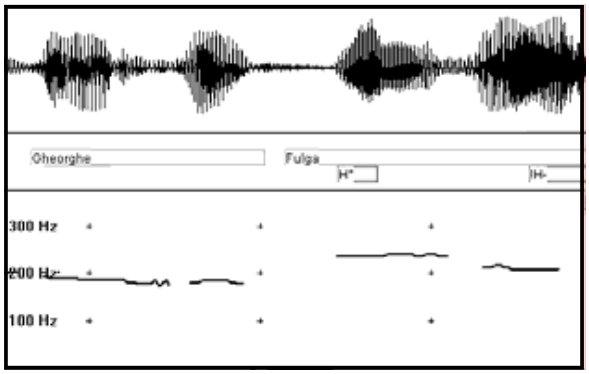

Figura 4.7. Unda vocală şi conturul frecvenței fundamentale al rostirii "Gheorghe Fulga"

Când vocala silabei accentuate este precedată sau urmată de consoane nesonore, fronturile vârfului nu apar în forma (pattern-ul) de accent, ca în cazul cuvântului /*Fulga*/ unde se atinge valoarea maximă la începutul vocalei /u/ ce se menține până la sfârşitul silabei accentuate (figura 4.7. Gama de variație a frecvenței de pitch în cadrul accentului H* este mai mare când tonul silabei neaccentuate anterioare este mai aproape de nivelul de *Low* şi mică în caz contrar (de exemplu, în cadrul unui focus larg (Ladd 1996)).

- **!H***

Când accentul de pitch de tip H* apare pe o tendință de coborâre a liniei de bază a conturului F0, acesta nu mai atinge înălțimea tonurilor „High” anterioare. Standardul ToBI prevede precedarea tonurilor „High” de caracterul “!” indiferent de tipul de etichetă în care apare. În figura 4.8 al doilea accent pe verbul “fost” are un ton țintă de nivel mai mic decât tonul High din cadrul accentului de pitch de tip H*+L asociat pronumele “tu” şi în consecință s-a adnotat cu !H*.

O altă categorie de tonuri !H* sunt cele care nu sunt însoțite de creşteri pe silaba accentuată ci formează paliere pe durata acesteia şi sunt urmate de scăderi semnificative de ton pe silaba a următoare. Este cazul accentelor „*High*” care apar înaintea finalurilor „*Low*” al frazelor intonaționale sau a celor formate pe o pantă abruptă a liniei de bază a conturului F0. Astfel de tonuri sunt cele din figura 4.6 care se formează pe cuvintele „di-mi-neá-ța” şi „‘doám-ne-lor”.

- **^H***

Când nivelul tonal al unui accent de tip H* este mai înalt decât precedentul de acelaşi tip, se creează o situație denumită în engleză “upstep”. În acest caz etichetei i se adaugă în față semnul diacritic “^”. În figura 4.6 apare o situație de “upstep” pe silaba accentuată din cuvântul “*dóm-(ni-lor)*” pe care nivelul tonului țintă „High” este mai ridicat decât pe cel al cuvântului „doamnelor”.

- **H+!H***

Acest accent este caracterizat de o cădere pe silaba accentuată dinspre un ton ridicat spre un alt ton ridicat, de nivel mai jos, care deşi se apropie de nivelul Low (jos) este mai ridicat decât acesta din urmă. Acest tip de accent se află în figura 4.9 pe verbul monosilabic “*fóst*”, pe durata căruia frecvența F0 scade dar nu până la tonul “Low” din finalul propoziției.

- **L***

Eticheta L* este folosită pentru marcarea accentului căruia îi corespunde în conturul de pitch o formă de “vale”, format dintr-un front scăzător până la un nivel de Low minim al tonului țintă, pe durata vocalei silabei accentuate. În figura 4.9 acest tip de accent apare pe silaba accentuată a cuvântului *“(întotdea)-ú-(na)*”. În figura 4.10 acest tip de accent apare pe silaba accentuată a cuvântului *“o-bráji*” şi este urmată de un ton de graniță de tip H-.

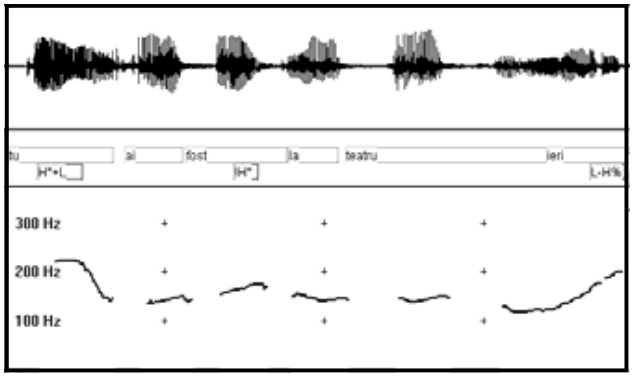

Figura 4.8 Unda vocală şi conturul frecvenței fundamentale al rostirii “Tu ai fost la teatru ieri?”

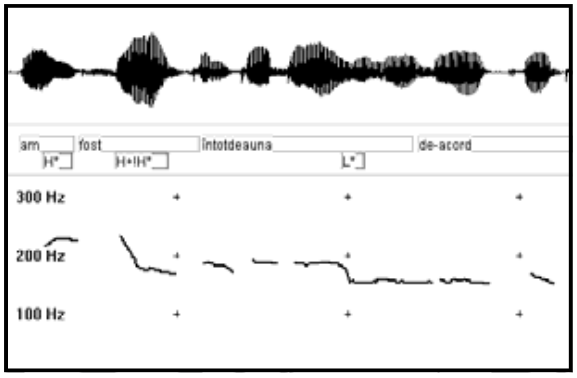

Figura 4.9 Unda vocală şi conturul frecvenței fundamentale al rostirii “Am fost întotdeauna de acord”

- **L+H***

Această etichetă corespunde unui tip de accent mai proeminent decât cel de tip H*. În (Debusmann ş.a. 2005) acesta este considerat a fi accentul cel mai des întâlnit pentru cuvintele ce formează rema într-un text, analizat din punct de vedere al teoriei discursului.

După cum se observă în figura 4.10 frecvența de pitch atinge un nivel "low" şi se menține pe consoana sonoră /n/ a silabei accentuate "*nî*" şi apoi creşte pe durata vocalei silabei accentuate până la un ton țintă „high". Este un accent bitonal format din două tonuri țintă, unul jos şi celălalt ridicat.

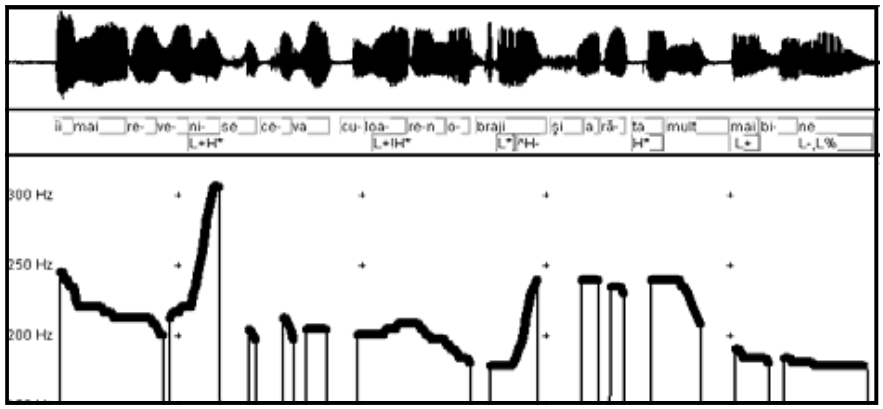

Figura 4.10. Unda vocală şi conturul frecvenței fundamentale al rostirii "Îi mai revenise ceva culoare-n obraji şi arăta mult mai bine."

- **L*+H**

Această etichetă corespunde unui tip de accent bitonal la care mai mult de 50% din durata silabei accentuate se menține la un nivel tonal scăzut "Low" după care se produce o creştere pronunțată a frecvenței F0 şi atingerea unui nivel tonal ridicat „High" (figura 4.11). Frecvența de pitch atinge un nivel "Low" şi se menține pe consoana sonoră /n/ (a cărei durată este mai mare decât durata vocalei) a silabei accentuate "*nóul*" şi apoi începe a creşte pe durata vocalei accentuate până la un ton țintă „high".

- **H+L***

Accentul de tip H+L* este tot din categoria celor bitonale şi cuprinde o variație scăzătoare de la un nivel țintă „High", pe durata vocalei silabei accentuate, spre un nivel țintă "Low". În figura 4.12 acest tip de accent apare pe cuvintele "*totuş*i" şi „*vorbească*" care sunt înaintea tonurilor de graniță ale frazelor intonaționale intermediare. În adnotările realizate acest tip de accent l-am întâlnit cu precădere pe cuvintele aflate în finalul frazelor intonaționale intermediare.

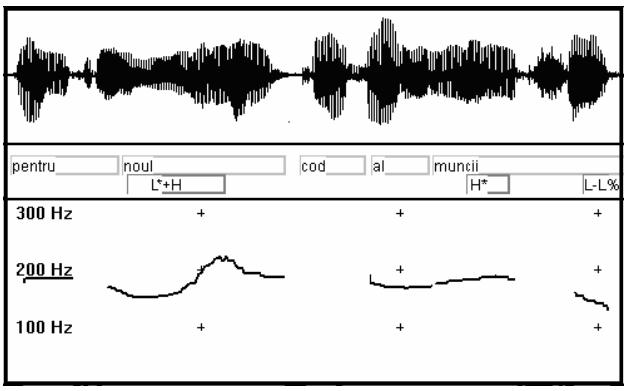

Figura 4.11. Unda vocală şi conturul frecvenței fundamentale al rostirii "..pentru noul cod al muncii"

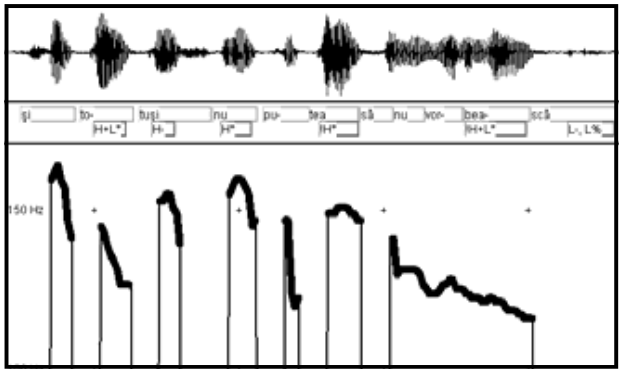

Figura 4.12 Unda vocală şi conturul frecvenței fundamentale al rostirii "Şi totuşi nu putea să nu vorbească"

4.4.3 Etichete pentru tonurile de frază intonațională intermediară

Tonurile ce formează accentele de frază împart o frază intonațională în mai multe fraze (unități) intermediare corespunzătoare sintagmelor, la nivelul textului. Sunt două feluri de accente de frază: "Low" notat (L-) şi "High" notat (H-). În figura 4.13 este redat conturul

frecvenței F0 corespunzător unui text format din două sintagme din finalul unei fraze. Prima unitate intermediară are un ton final H- de acelaşi nivel cu tonul accentului de pitch precedent H* iar cea de-a doua un ton de tip L-.

…(prognoza zilei)$_{H-}$ (cu Florinela Popa)$_{L-L\%}$"

În acest caz granițele sunt foarte clare şi nu sunt dubii în identificarea lor. Granițele sintagmelor prezintă o mare varietate de manifestare începând cu pauzele clare, însoțite de o creştere sau scădere locală de F0, până la o subtilă modificare lentă de pitch care provoacă o definiție neambiguă. Astfel, sunt divergențe de păreri despre faptul că o graniță de sintagmă este sau nu prezentă. În literatură definițiile granițelor de IP sunt vagi (Ladd 1996). O altă problemă o constituie faptul că, deşi se observă unele trăsături fonetice care să constituie graniță de frază intonațională, aceasta nu se percepe auditiv. Se greşeşte uneori datorită faptului că se încearcă împărțirea în fraze, ținând cont de constituenții sintactici, semantici şi de discurs sau se ignoră faptul că structura prozodică este mai simplă decât cea sintactică/ semantică (Ladd 1996).

4.4.4 Etichetele pentru tonurile de granițe finale ale frazelor intonaționale

Unitățile intonaționale corespunzătoare propozițiilor/frazelor se termină fie cu un ton jos ("Low") notat L%, fie cu un ton ridicat ("High"), notat H%. Deoarece un sfârşit de propoziție/frază implică şi un ton de sfârşit al ultimei sintagme, rezultă că în finalul unei propoziții/fraze se pot produce următoarele combinații de tonuri: L-L%, H-H%, H-L%, L-H% ce vor fi exemplificate pe rostiri din corpus-ul de voce în limba română.

- **L-L%**

Această este combinația de tonuri specifică finalurilor propozițiilor afirmative în care tonul L% înseamnă o cădere accentuată sub tonul de mediu de *Low* al propoziției. Secvența de tonuri apare în figura 4.13 pe finalul de contur F0.

- **H-H%**

Secvența de tonuri H-H% apare în cazul în care ultima sintagmă se termină cu un accent de frază ridicat (H-) şi propoziția/fraza are un puternic caracter ascendent care se traduce printr-o ridicare suplimentară a tonului la nivelul H%. Această secvență de tonuri se întâlneşte la propozițiile interogative totale (figura 4.14 prezintă un puternic accent imperativ)

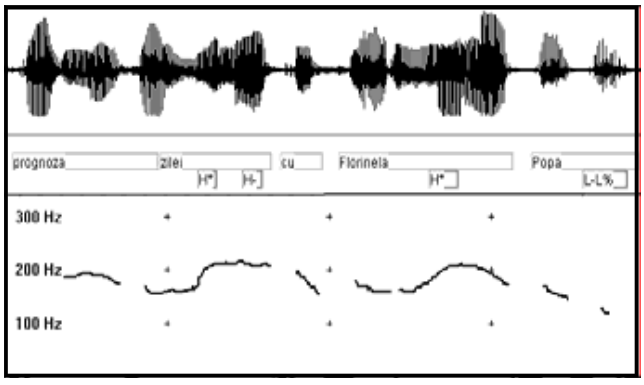

Figure 4.13 Unda vocală şi conturul frecvenței fundamentale al rostirii "..prognoza zilei cu Florinela Popa"

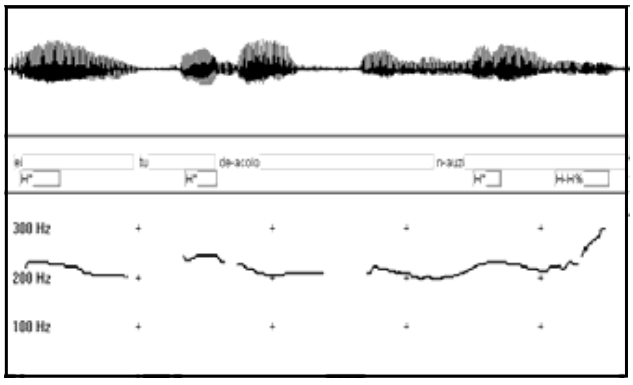

Figure 4.14. Unda vocală şi conturul frecvenței fundamentale al rotirii "Ei, tu de acolo, n-auzi?"

- **L-H%**

Această etichetă este specifică propozițiilor afirmative urmate de virgulă, când frecvența F0 se ridică de la un ton Low la care a ajuns printr-o secvență L*L- spre un ton ridicat H%. În cazul propoziției secundare *"Când venea vorba de război, ..."* din figura 4.15, pe ultima

silabă, care este şi accentuată, se realizează şi accentul L* şi cel de final sintagmă L-. Pe prelungirea silabei se formează tonul H%.

- **H-L%**

Această secvență de tonuri este specifică frazelor intonaționale cu continuare (figura 4.16).Finalul propoziției este ascendent începând cu ultima silabă accentuată (accent H*) şi apoi scade puțin la o valoare care nu este o valoare reală de *Low*, a cărei durată mai lungă formează un platou. Aceasta este o valoare în mijlocul gamei de pitch a vorbitorului. Prin creşterea duratei ultimei vocale se poate genera un platou la nivelul de pitch dat de această valoare intermediară.

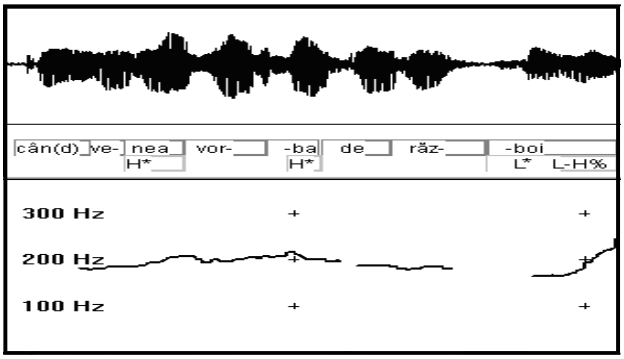

Figura 4.15. Unda vocală şi conturul frecvenței fundamentale al rostirii "Când venea vorba de război…"

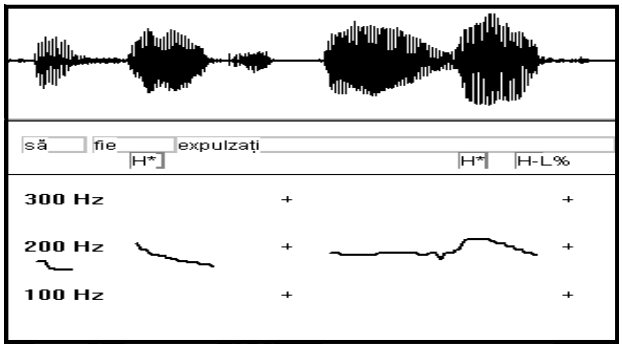

Figura 4.16. Unda vocală şi conturul frecvenței fundamentale al rostirii "…să fie expulzați."

În figura 4.16 tonul L% care urmează accentului de frază H- corespunde unui final de ştire radio, sugerând continuarea ştirii.

4.5 Adnotarea intonației pe corpusuri de voce

În această secțiune voi prezenta două exemple de etichetare a evenimentelor tonale de pe conturul F0 al semnalului vocal, în corelație cu împărțirea în subunități intonaționale, pe baza ierarhiei intonaționale propuse. În figura 4.17, rostirea propoziției "*[Winston][îşi duse/ paharul/la buze][cu o oarecare/ nerăbdare]"* este împărțită din punct de vedere intonațional în două unități **IP.** Tonul de graniță final al primei unități **IP** (frază intonațională) se formează pe ultima silabă neaccentuată a cuvântului *Winston*, fiind caracterizat de durată mare şi o variație de ton semnificativă marcată L-H%.

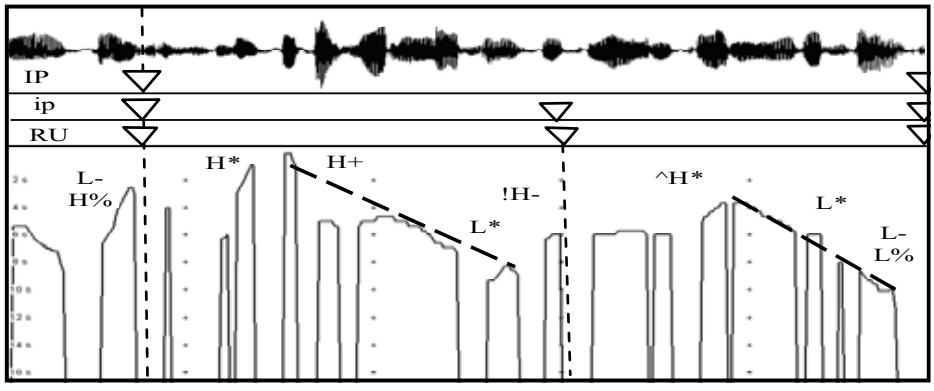

Figura 4.17 Unda vocală şi conturul F0 al rostirii
"*[Winston] [îşi duse/ paharul/la buze][cu o oarecare/ nerăbdare]"*

A doua unitate intonațională de tip **IP** se compune din două unități de tip **ip** iar în figură sunt marcate în mod corespunzător, cele două tendințe descrescătoare ale tonurilor țintă.

Unităţile de tip **ip** sunt formate din câte o unitate ritmică (**RU**). Tonul de început al fiecărei tendinţe descrescătoare se formează la finalul primelor unităţi din cadrul fiecărui **ip**. Ridicarea tonului înaintea unei noi tendinţe descrescătoare a tonurilor ţintă se numeşte în literatura de specialitate „reset F0". Prima unitate **ip** se termină printr-o uşoară ridicare de ton (accentul de frază !H-) iar cea dea doua încheie odată cu unitatea **IP** formând combinaţia de tonuri L-L%. În cadrul ultimilor două unităţi ritmice accentele puternice se formează pe verbul *duse*(H*) şi respectiv, adjectivul *oarecare*(^H*).

Rostirea al cărui contur este reprezentat în figura 4.18 reprezintă un exemplu de intonaţie ritmată generată de succesiunea unor accente puternice, de tărie apropiată, corespunzătoare fiecărui cuvânt din unităţile IP1 şi IP3. Astfel, fiecare unitate de accentuare formează singure unităţi ritmice separate.

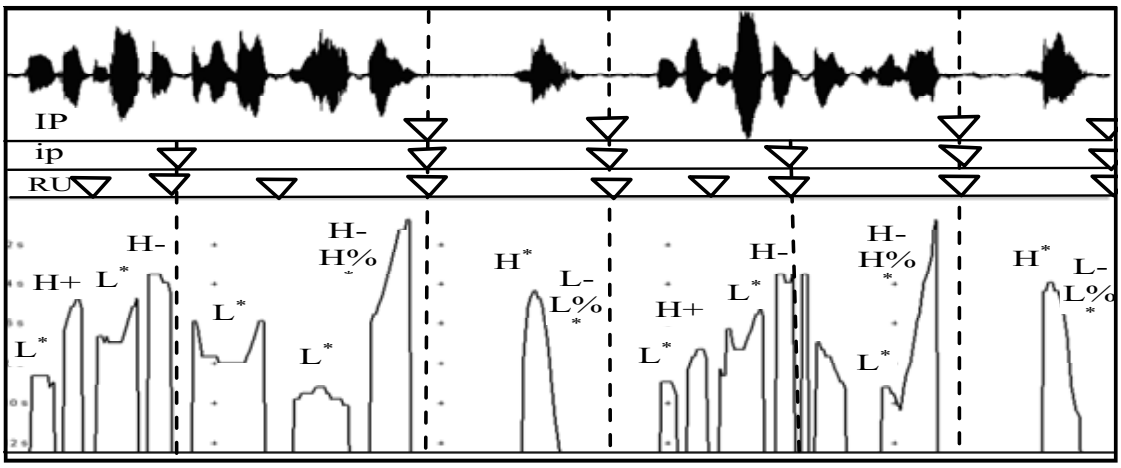

Figura 4.18 Unda vocală şi conturul F0 al rostirii
"-[(Sunteţi)(gata)][(să vă daţi/viaţa?)][-Da.] [(Sunteţi)(gata)][(să ucideţi)?][- Da.]"

Unităţile de accentuare prezintă acelaşi tip de accent (notat L*) şi este generat de tonul menţinut la nivel *Low* pe silaba accentuată, urmat de o creştere pe silaba neaccentuată, imediat următoare. Tonurile ţintă ridicate, astfel formate au fot marcate cu H- atunci când corespund unor finaluri de **ip,** fie cu etichete de tonuri ţintă H+. Unităţile de tip **ip** din cadrul unităţilor IP1 şi IP3 corespunzând unor intonaţii interogative, au o tendinţă de creştere a tonurilor ţintă, numită în engleză *upsteping*. Răspunsurile la aceste interogaţii se desfăşoară în cadrul unităţilor IP2 şi IP4, ce conţin ca evenimente, accentele de pitch de tip H* şi secvenţele finale de tip L-L%.

4.6 Contribuţii personale

Contribuţiile din acest capitol sunt legate în principal de necesitatea elaborării unui ierarhii intonaţionale care să stea la baza modelării prozodiei în limba română. Pentru a realiza acest deziderat am efectuat o analiză a principalelor modele intonaţionale care stau la baza modelelor prozodice. În urma acestei analize am constatat că pentru a fi utilizate în sinteza vocală, modelele intonaţionale trebuie corelate cu modele fonologice (care au la bază ierarhii intonaţionale), cu structurile sintactico-semantice ale textelor şi funcţiile prozodiei (Batliner 2003, Kohler 2005, Teodorescu 2005, Shih 2006, Hirst 2007 ş.a).

În urma studiului principalelor modele intonaţionale, pe baza analizei contururilor intonaţionale din limba română (Apopei ş.a. 2005b, Apopei ş.a. 2006a, Turculeţ& Apopei 2006) şi al încercărilor de a implementa aceste contururi în sinteza vocală (Apopei ş.a. 2005a) am ajuns să înţelegem legătura dintre modelele fonologice şi modelele prozodice (în particular modelele intonaţionale). Astfel am reuşit să propun o ierarhie intonaţională

(Apopei ş.a. 2006b, 2006c) cu care să putem grupa evenimentele intonaționale şi să abordăm problematica sintezei prozodice în limba română.

Această ierarhie intonațională a stat la baza dezvoltării schemei de adnotare a evenimentelor microprozodice prezentată în sectiunea 5.2.1., a cercetărilor ulterioare privind înțelegerea intonației în limba română şi a implememtării elementelor prozodice în sinteza vocală.

Capitolul 5.

Sinteza prozodică

Implementarea prozodiei în sinteza vocală presupune generarea automată a "melodiei" corespunzătoare rostiri unui text, pe baza unor modele care pun în corespondență structura de informații a textului rezultată din analiza morfologică, sintactică şi semantică sintactică cu un set de evenimente prozodice care au asociate, în principal, descrieri parametrice pentru frecvența fundamentă F0, pauze, durata şi intensitatea sunetelor.

Funcție de performanțele dorite pentru sistemul de conversie text-voce modelul prozodic poate să realizeze descrierea parametrică pentru următoarele elemente prozodice: numai a intonației la nivelul accentelor lexicale; intonația la nivelul unor grupuri de cuvinte (sintagme); intonația la nivelul propozițiilor folosind reguli lingvistice sau sisteme de învățare automată; corelarea descrierii intonației cu alte elemente prozodice (durata şi intensitatea sunetelor, tempoul şi ritmul vorbirii ş.a).

Majoritatea modelelor intonaționale dezvoltate până în prezent asociază evenimentele intonaționale de pe conturul frecvenței F0 cu forma acestora şi mai puțin cu funcția (înțelesul, semnificația) acestora în comunicare. În ultimii ani au început să apară definiții şi implementări mai complexe pentru modelele prozodice. Conform acestora, modelele prozodice realizează o reprezentare fonologică a vorbirii pe baza unor relații între funcțiile şi formele (elementele şi evenimentele) prozodiei (Hirst 2007, Shih 2006, Batliner 2003). Din categoria modelelor intonaționale care pun în legătură evenimentele intonaționale cu funcția acestora în comunicare cel mai reprezentativ este modelul PENTA (Xu 2004, 2007). Acest model se distinge de modelele tradiționale (Xu 2004a) prin următoarele elemente: face o separație clară între componentele intonației care au înțeles în comunicare (pe care le numeşte şi componente funcționale) şi primitivele de contur F0 definite prin formă; propune un mecanism pentru realizarea prin intonație a mai multor înțelesuri în comunicare; stabileşte o legătură între mecanismul de generare a conturului frecvenței F0 pe baza primitivelor de formă şi componentele funcționale ale melodiei unei rostiri.

Teodorescu H.N. (2005) propune completarea structurii de informații rezultată în urma analizei morfologice, sintactice şi de discurs (a textului) cu informații despre limbajul folosit (colocvial, oficial, artistic etc.), emoție, inter-relația vorbitor-receptor şi starea vorbitorului. Pentru predicția evenimentelor prozodice asociate unui text, Teodorescu (2005) propune folosirea unui principiu de maximizare a informației contextuale cuprinse în noua structură de informații asociată textului.

Cercetările efectuate în cadrul Institutului de Informatică Teoretică până în anul 2003 au urmărit introducerea primelor elemente prozodice în sintetizatorul dezvoltat în cadrul institutului. Acestea s-au rezumat la împărțirea cuvintelor în silabe şi stabilirea silabei accentuate (Jitcă, Apopei 2003) folosind un sistem ierarhic format din două rețele neuronale.

Modelul prozodic dezvoltat după anul 2003, în cadrul Institutului de Informatică Teoretică Iaşi a urmărit realizarea unei legături între text şi voce prin intermediul unor scheme de reprezentare asemănătoare cu cele descrise în diverse implementări realizate sub platforma *VoiceXML (http://www.w3.org/TR/voicexml)*. Pornind de la acest deziderat

introducerea elementelor de prozodie a început pe texte adnotate, în format XML, morfologic în prima etapă, morfologic şi sintactic în a doua etapă.

În prima etapă s-a pornit de la o adnotare la nivel morfologic a unui fragment din „Ecleziastul”. Pe baza unui model propus de H.N. Teodorescu s-a realizat o împărțire a acestuia din punct de vedere intonațional în grupuri de cuvinte cu pattern-uri intonaționale (Teodorescu, Ceauşu, Apopei 2003). Pentru delimitarea grupurilor de cuvinte s-a folosit *tag*-ul „break” cu două valori (0 şi 2) prin care se indică prezența unor pauze iar pentru descrierea tonurilor de realizare a accentelor lexicale din cadrul acestor grupuri s-a introdus pentru cuvinte atributul „pitch” cu două valori („high”/ „low”), asociat în general cuvintelor de la începutul şi de la sfârşitul grupurilor de cuvinte. Împărțirea frazelor în grupuri de cuvinte şi nivelul tonurilor erau stabilite în funcție de anumite clase de mărci textuale şi semne de punctuație folosind *n-grame*.

În ce-a de a doua etapă, am propus să realizăm o implementare a elementelor prozodice pe baza teoriei autosegmental-metrice. Am reuşit să propunem un model fonologic ierarhizat (Apopei şi Jitcă 2006, 2007) care să realizeze împărțirea textului în fraze intonaționale şi diviziuni ale acestora prin diferențierea mai multor moduri de realizare a accentelor lexicale (în principal pe baza setului de etichete din sistemul de adnotare a intonației ToBI). În cadrul cercetărilor efectuate pentru modelarea prozodiei am folosit rostiri ale unor fragmente din romanul “1984” al autorului George Orwell şi din corpusul de voce al Seminarului de Dialectologie al Universității „Al. I. Cuza” Iaşi.

Modelarea prozodică pe care am elaborat-o a fost dezvoltată în cadrul temelor de cercetare ale Institutului de Informatică Teoretică şi a fost gândită din perspectiva realizării unei punți de legătură între cercetările din domeniul lingvisticii computaționale (Tufiş 2000,2007, Cristea 2003, 2005, Curteanu 2007, Forăscu 2006, 2008) şi cele din domeniul analizei şi sintezei vocale pentru limba română (Teodorescu H.N. 2003, 2005, 2008, Burileanu D. 2006, Grigoraş Fl. 1997,1999, Jitcă 2002, 2003).

5.1 Structura unui sistem pentru conversia Text-Voce cu modul prozodic

Sistemele de conversie text-voce (în limba engleză “*Text-to-Speech*” - TtS) cu modul prozodic sunt rezultatul cercetărilor interdisciplinare din domeniile: procesarea semnalului vocal, lingvistica computațională, analiza şi descrierea parametrică a semnalului vocal din punct de vedere fonetic şi fonologic, psiho-acustică. Aceste sisteme au în componență următoarele module (figura 5.1):

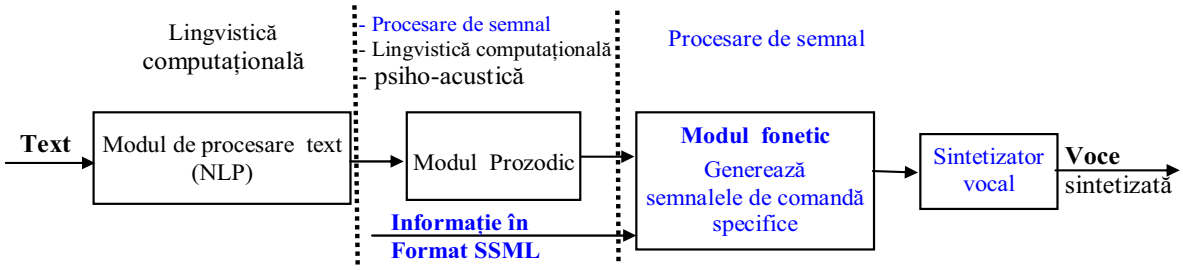

Fig. 5.1. Schema bloc a unui sistem de conversie text-voce cu modul prozodic

- modulul de procesare a textului (în engleză *Natural Language Processing* - NLP) - completează textul de intrare cu informații despre structura morfologică, sintactică şi semantică a cuvintelor;
- modulul prozodic - generează descrieri parametrice pentru elementele prozodice specifice modelului utilizat în implementare (intonația, pauzele, durata şi intensitatea fonemelor);
- modulul fonetic - generează semnalele pentru comanda sintetizatorului vocal pe baza

informațiilor fonetice şi prozodice.

- sintetizator vocal – realizează generarea unui semnal sintetizat pe baza semnalelor generate de modulul fonetic.

În partea superioară a figurii 5.1 am trecut domeniile de cercetare implicate în realizarea fiecărui modul. Introducerea elementelor prozodice în vocea sintetizată cu ajutorul modelelor prozodice necesită parcurgerea următoarelor etape (figura 5.2):

- împărțirea textului în fraze intonaționale şi stabilirea accentelor proeminente (**Frazare**);
- stabilirea secvenței de evenimente prozodice pentru frazele intonaționale sau a secvenței de forme de contur pentru unitățile de accentuare;
- asocierea de evenimentele prozodice pentru informațiile fonetice.

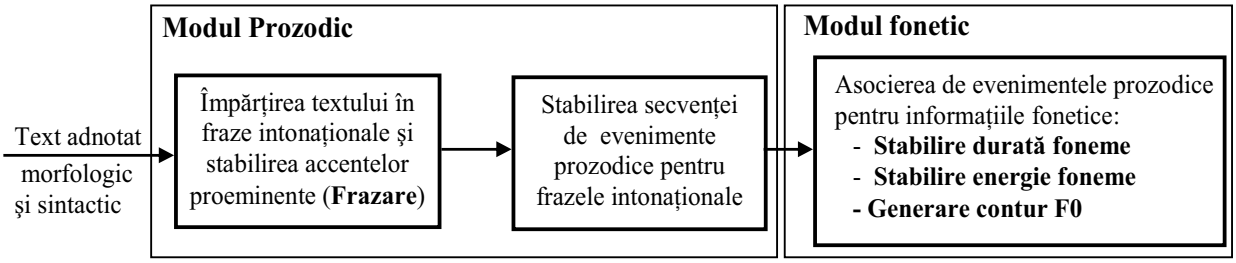

Figura 5.2. Principalele etape de procesare implicate de introducerea prozodiei în sistemele de conversie text-voce

Primele două etape sunt componente ale modulului prozodic iar ce-a de a treia aparține modulului fonetic. Modul de implementare a ultimei etape este dependent de tipul de sintetizator utilizat pentru sinteza vocală.

5.1.2 Modulul prozodic

Modulul prozodic, din sistemele de conversie text-voce de ultimă generație, este responsabil de asocierea elementelor prozodice (începând cu modul de frazare a textului şi stabilirea accentelor sintactice şi terminând cu materializarea elementelor prozodice în vocea sintetizată) pentru textul de intrare adnotat la nivel morfologic, sintactic, semantic şi de discurs.

Indiferent de modelul prozodic şi intonațional implementat în cadrul modulului prozodic prima etapă parcursă în vederea predicției de prozodie o reprezintă împărțirea textului în fraze intonaționale. Frazele intonaționale reprezintă, pentru majoritatea modelelor prozodice, cele mai mari unități prozodice pentru care se face predicția evenimentelor prozodice. Împărțirea unui text în fraze intonaționale şi stabilirea accentelor proeminente se poate realiza prin reguli (Bachenko şi Fitzpatrick 1990, Dohalská ş.a 2001), arbori de decizie (Wang şi Hirschberg 1992, Ostendorf şi Veilleux 1994), rețele neuronale (Hwang ş.a 1996) sau *n-grame* realizate cu lanțuri Markov ascunse (Taylor şi Black 1998, Taylor ş.a 2006).

Taylor şi Black (1998) au pus în evidență faptul că majoritatea frazelor intonaționale pentru limba engleză au între trei şi şase cuvinte. Frazele intonaționale sunt diferențiate (Tao 2002, Huang 1997, Schröder 2003, 2004, Xu 2004, ş.a) în general prin: structura morfologică şi sintactică; categoria şi tipul *mărcii* care delimitează finalul frazei intonaționale; poziția accentelor lexicale şi gramaticale; structura silabică; funcția (înțelesul, semnificația) acestora în comunicare. Pe baza acestor elemente, în următoarea etapă are loc asocierea secvenței de evenimente prozodice sau de forme de contur intonațional pentru unitățile de accentuare din cadrul frazelor intonaționale.

Pentru predicția secvenței de evenimente prozodice la nivelul frazelor intonaționale se folosesc seturi de reguli (Mixdorff şi Fujisaki 1995, Jilka ş.a. 1999, Becker ş.a 2006) şi/sau algoritmi de învățare. Sistemele de predicție a prozodiei bazate pe algoritmi de învățare

folosesc pentru antrenare corpusuri de voce adnotate prozodic şi corpusuri de text adnotate la nivelul morfologic, sintactic şi semantic. Cele mai cunoscute tehnici de învățare utilizate în acest domeniu se bazează pe rețele neuronale (Hwang ş.a 1996), arbori de decizie (Syrdal ş.a 1998) şi lanțuri Markov ascunse (Taylor ş.a 2000, Sun 2001, Tokuda ş.a 2002). Sistemele de predicție a prozodiei bazate pe tehnici de învățare prezintă avantajul de a adapta uşor pentru diferite tipuri de vorbire (normală, emoțională) şi pentru diferiți vorbitori.

Modelul prozodic care evidențiază cel mai bine legătura dintre funcțiile comunicative ale prozodiei şi formele de contur F0 prin care se materializează aceste funcții, a fost propus de către Xu Y.(2004, 2005, 2006) sub forma modelul PENTA (**P**arallel **EN**coding and **T**arget **A**pproximation). Pentru generarea conturului frecvenței F0 se propune procesarea paralelă a textului (**P**arallel **EN**coding) din punct de vedere al informației referitoare la funcțiile comunicative ale prozodiei (Kohler 2005, Xu 2006) şi utilizarea algoritmului „***T**arget **A**pproximation*" de aproximare a conturului melodic al frecvenței F0 pe baza unor puncte țintă (Xu ş.a 1998, Xu & Wang 2001).

La ieşirea modulului prozodic informația morfologică, sintactică şi semantică, asociată textului de la intrare, este completată cu indicații referitoare la forma evenimentelor prozodice. Această informație poate constitui intrarea modulului fonetic (în cadrul procesărilor on-line) sau poate fi salvată în fişiere cu structura VoiceXML (în cadrul procesărilor off-line) .

5.1.3 Modulul fonetic

Modulul fonetic, din cadrul sistemelor text-voce, generează semnalele de comandă specifice sintetizatorului vocal pe baza informațiilor primite de la modulele precedente. Intrarea modulul fonetic poate veni direct de la ieşirea modulului prozodic sau dintr-un fişier în format SSML (***S**peech **S**ynthesis **M**arkup **L**anguage)*. Fişierele în format SSML conțin pe lângă structura de informații a textului (adnotare morfologică, sintactică şi semantică) şi informații despre elementele prozodice care vor fi asociate de sintetizatorul vocal. În cadrul acestui modul se realizează următoarele procesări asupra informației lingvistice şi prozodice (figura 5.4):

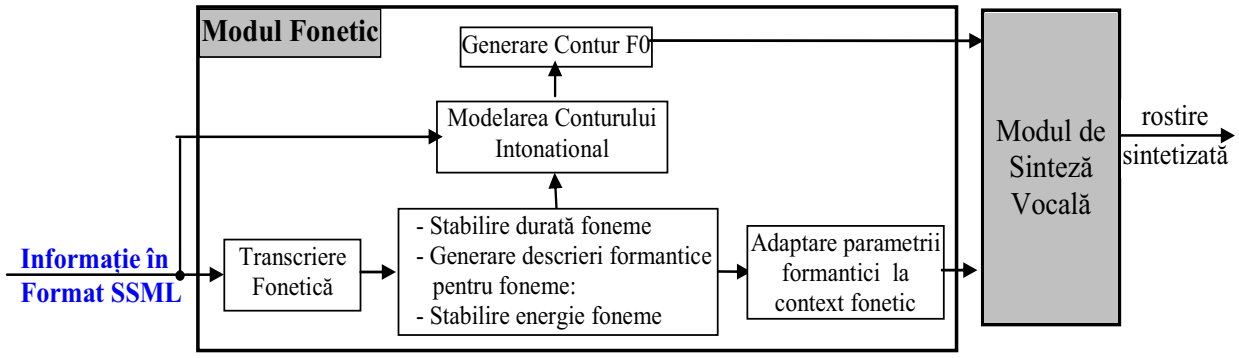

Fig. 5.4. Schema bloc a modulului fonetic

- fonetizarea automată a textului de intrare folosind diferite alfabete fonetice pentru codificarea fonemelor şi alofonilor acestora: alfabetul fonetic internațional (IPA), WorldBet (Hieronymus 1993) sau X-SAMPA (Wells 2000);
- asocierea descrierilor parametrice pentru primitivele de sinteză (difoni, foneme şi alofoni);
- generarea descrierilor parametrice pentru energia şi durata fonemelor şi alofonilor;
- rezolvă problema coarticulării dintre foneme prin adaptarea descrierilor parametrice ale primitivelor de sinteză la contextul fonetic (secțiunea 3.3 din teză);.

- generarea semnalelor de comanda pentru sintetizator;
- sincronizarea în timp a descrierilor parametrice pentru foneme cu parametrii de modificare ai frecvenței F0

În cadrul modulului fonetic cercetărilor efectuate în perioada elaborării tezei au vizat analiza posibilităților de generare a diferite forme de contur intonațional, sincrone cu desfăşurarea în timp a fonemelor, pe baza de descrieri fonologice ale conturului frecvenței F0.

5.2 Utilizarea informației prozodice în format XML

Dezvoltarea sistemelor vocale de dialog om-maşină a determinat crearea în cadrul standardului XML (Extensible Markup Language) a unor scheme de reprezentare a informației prin care să coreleze textul cu vocea. Aceste reprezentări au condus la apariția platformei *VoiceXML (http://www.w3.org/TR/voicexml)* în cadrul căreia se pot dezvolta aplicații pentru sinteza vocală, recunoaşterea vocală, pronunția de cuvinte din lexicoane, telefonie ş.a. Pentru fiecare tip de aplicație s-au dezvoltat standarde de reprezentare a informației după cum urmează: pentru recunoaşterea vocală standardele SRGS- ***S**peech **R**ecognition **G**rammar **S**pecification* şi SISR – ***S**emantic **I**nterpretation for **S**peech **R**ecognition*; pentru sinteza vocală standardul SSML - ***S**peech **S**ynthesis **M**arkup **L**anguage*; pentru pronunția de lexicoane standardul PLS - ***P**ronunciation **L**exicon **S**pecification*.

Primele reprezentări XML ale informației prozodice pentru sinteza vocală au urmărit introducerea unor indicații macro-prozodice la intrarea sintetizatoarelor vocale. Participanți la consorțiul multi-național Festival au propus standardul SABLE (Taylor ş.a. 1997, Sproat ş.a. 1998) dezvoltat pe baza standardelor STML (dezvoltat la *Bell Labs* şi Universitatea din Edinburgh) şi JAML (dezvoltat la *Sun Microsystems*). Pentru indicații micro-prozodice de finețe, în caz de nevoie, autorii standardului propun utilizarea de atribute suplimentare pentru a specifica durata unor foneme şi diferite moduri de realizare a accententelor sintactice (în format ToBI).

Ulterior, în cadrul sistemului de sinteză vocală, MARY (**M**odular **A**rchitecture for **R**esearch on speech s**Y**nthesis), dezvoltat pentru limba germană, s-a propus o reprezentare a informației prozodice la nivel micro-prozodic cu ajutorul standardelor MARYXML sau BOSXML (Schröder 2004).

Pe lângă schemele de reprezentare a informațiilor prozodice dezvoltate sub standardul W3C SSML (Walker & Hunt, 2001), firma Microsoft (2002) a dezvoltat propriul standard (SAPI) pentru marcarea indicațiilor prozodice.

5.2.1 Schemă XML de adnotare a intonației pentru limba română

În cadrul grupului nostru de cercetare, ideia introducerii elementelor de prozodie în sinteza vocală pentru limba română, prin reprezentare în format XML, a fost propusă de H.N. Teodorescu (2002) într-un grant CNCSIS. A fost realizată o schemă de adnotare a evenimentelor macroprozodice cu două *taguri* (*break* şi *pitch*). *Tag*-ul „break”, cu două valori (0 şi 2), indica prezența unor pauze pentru delimitarea, în sinteză, a grupurilor de cuvinte. *Tag*-ul „pitch”, prin valorile („high”/ „low”), indica trendul conturului intonațional şi implicit al accentelor lexicale pe durata grupului de cuvinte.

În această etapă la stabilirea schemei pentru adnotarea intonației în format XML s-au avut în vedere unitățile intonaționale din ierarhia prezentată în secțiunea 4.4, creând câte un tag pentru marcarea unităților de pe fiecare nivel. Tag-urile utilizate împreună cu atributele lor au fost prezentate în lucrarea (Apopei ş.a. 2006).

5.2.2 Studiu de caz privind asocierea evenimentele intonaționale cu atributele din formatul XML

Exemplificarea modului de asociere dintre evenimentele de pe conturul frecvenței F0 şi atributele din formatul XML este prezentat pentru intonația rostirii textului *„Avem de discutat lucruri serioase, zece minute nu-i nevoie să mai faci pe valetul”* este reprezentată în figura 5.5, prin unda vocală şi curba F0 (Raport 2006a).

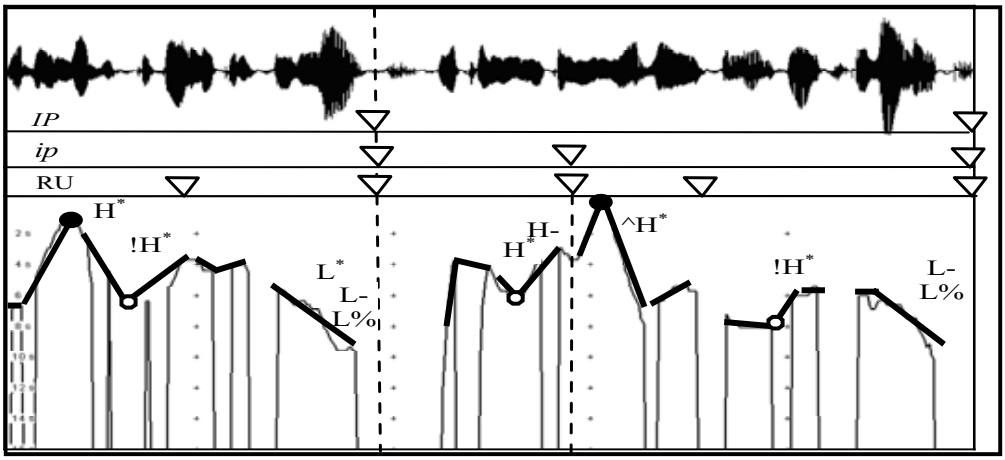

Figura 5.5 Unda vocală şi conturul F0 al rostirii textului
“*[(Avem/de discutat)(lucruri/serioase)],[(zece/minute)] [(nu-i nevoie)(să mai faci/pe valetul)]* ”

Intonația este formată din două fraze intonaționale. În cadrul primei unități **IP** se formează două unități ritmice iar în cadrul celei de a doua unității **IP** se formează două fraze intermediare **ip** ce cuprind cele trei unități ritmice. În figură sunt marcate formele aproximative ale segmentelor de contur F0 corespunzătore unităților de accentuare. Se constată forme de contur tipice pentru finalurile unităților **IP**, definit de secvențele de tonuri L-L% caracterizate de scăderea tonului pe durata întregii unități terminale.

Formele de contur F0 pe unitățile de accentuare de la începutul unităților **IP/ip** au tendințe de creştere până la atingerea tonului țintă cel mai ridicat din cadrul frazei intonaționale, de la care urmează tendința de descreştere (unitățile de accentuare al căror ton de nivel maxim sunt marcate cu cerculeț negru). Formele de contur tipice ale unităților de accentuare din interiorul unei unități IP, în care accentul este pe ultima sau penultima silabă, sunt formate din două segmente: unul descrescător înaintea silabei accentuate şi al doilea crescător începând cu silaba accentuată (unitățile marcate în figura 5.5 cu cerculeț în punctul în care începe creşterea).

Delimitările realizate în figura 5.5, corespunzătoare unităților intonaționale, au fost aplicate textului într-un fişier în format XML generând structurarea acestuia din punct de vedere intonațional (figura 5.6).

```
<IP BeginTonLabel="%L" BoundaryTonLabel="L%">
    <RU>
      <AU>
        <W TonalGroupLabel="L+H*" ToneValues="4,1">Avem</W>
      </AU>
      <AU>
        <W>de </W>
        <W TonalGroupLabel="!H*" ToneValues="3">discutat</W>
      </AU>
    </RU>
    <RU>
      <AU>
        <W>lucruri</W>
      </AU>
```

```
        <AU>
          <W TonalGroupLabel="H+L*" ToneValues="8,10">serioase</W>
        </AU>
      </RU>
</IP>
<IP BeginTonLabel="%L" BoundaryTonLabel="??">
  <ip PhraseTonLabel="H-">
      <RU>
        <AU>
          <W TonalGroupLabel="H*" ToneValues="3">zece</W>
        </AU>
        <AU>
          <W TonalGroupLabel="L+H*" ToneValues="6,3">minute</W>
        </AU>
      </RU>
  </ip>
  <ip PhraseTonLabel="L-">
      <RU>
        <AU>
          <W TonalGroupLabel="L+^H*" ToneValues="3,0">nu-i</W>
        </AU>
        <AU>
          <W>nevoie</W>
        </AU>
      </RU>
      <RU>
        <AU>
          <W>să</W>
          <W>mai</W>
          <W TonalGroupLabel="!H*" ToneValues="6">faci</W>
        </AU>
        <AU>
          <W>pe</W>
          <W TonalGroupLabel="L*" ToneValues="9">valetul</W>
        </AU>
      </RU>
  </ip>
</IP>
```

Figura 5.6 Exemplu de adnotare prozodică în fomat XML a textului
"*[(Avem/de discutat)(lucruri/serioase)][(zece/minute)] [(nu-i nevoie)(să mai faci/pe valetul)]* "
corespunzător rostirii din figura 5.5

Pentru pune în evidență mai uşor modul de asociere a informației prozodice în cele două reprezentări (grafică şi XML) am folosit următoarele convenții de notație pe text: granițele de fraze intonaționale (*IP/ip*) sunt marcate cu paranteze acoladă; granițele de unități ritmice sau grupuri de unități de accentuare sun marcate cu paranteze pătrate; separarea unităților de accentuare sa realizat cu caracterul '/'.

5.3 Forme de intonații în corelație cu sintaxa, semantica şi emoția

Cercetările efectuate în ultimul timp asupra modelelor prozodice pun în evidență tot mai mult corelarea elementelor prozodice cu structurilor sintactico-semantice ale textelor şi funcțiile prozodiei (Kohler 2005, Teodorescu 2005, Shih 2006, Hirst 2007 ş.a). În această secțiune voi prezenta o modalitate de abordare a relației dintre elementele prozodice cu structura sintactică a textelor asociate rostirilor şi semantica acestora.

5.3.1 Studiu de caz pentru intonația propozițiilor afirmative

Pentru a corela intonația neutrală a rostirii propozițiilor afirmative, cu structura sintactică şi semantică am realizat o analiză comparativă a descrierilor intonaționale pentru

rostirilor unui set de propoziții selectate din romanul "1984" al autorului George Orwell. În vederea identificării elementelor prozodice comune, fiecărei propoziție a fost rostită de către 4 vorbitori (Raport 2006a) iar descrierile intonaționale au fost realizate folosind ierarhia intonațională prezentată în secțiunea 4.4.

Analiza contururilor intonaționale a generat observații referitoare la secvențele de evenimente intonaționale de pe conturul unităților intonaționale de tip IP/ip, elementele prozodice, precum şi contextele textuale (structurile silabice) şi semantice care influențează forma conturului F0 în cadrul unităților de accentuare.

Analiza comparativă a descrierilor intonaționale a pus în evidență existența unor elemente la nivelul structurilor intonaționale şi sintactico-semantice care respectă anumite reguli. Există situații când aceste reguli nu mai sunt respectate. Cauza principală care a generat abateri de la reguli, pe rostirile analizate, a fost focalizarea diferită a unor cuvinte. Elementele prozodice care se supun unor reguli lingvistice sunt:

- realizarea finalurilor de frază intonațională prin aceeaşi combinație dintre accentul de pitch şi accentul de frază.
- formele de contur F0 ale unităților de accentuare de începutul şi sfârşitul propozițiilor/ frazelor cu anumite structuri sintactice .
- generarea accentelor de pitch prin aceleaşi tipuri de accente în cazul frazelor intonaționale cu aceiaşi structură sintactică şi semantică.

În urma corelării descrierile intonaționale cu structurilor morfologice şi sintactice ale frazelor am putut stabili următoarele reguli pentru gruparea cuvintelor în unități ale ierarhiei intonaționale:

- Grupurile verbale fac parte din aceiaşi unitate de accentuare. De exemplu grupul verbal "putea fi dat" se rosteşte cu accent H* pe silaba */tea/,* se menține la nivel *High* pe */fi/ şi* coboară formând un accent secundar de tip *H+!H** pe silaba */ dat/.*
- Cuvintele care exprimă gradele de comparație ale adjectivelor sau adverbelor intră în aceiaşi unitate de accentuare cu acestea din urmă. În cazul în care gradul de comparație este exprimat printr-un unui singur cuvânt (ca de exemplu *mai* bine) acesta se rosteşte pe trendul descrescător al unității de accentuare iar adjectivul sau adverbul pe cel crescător. În cazul mai multor cuvinte care exprimă gradul de comparație (de exemplu, *mult mai* bine) acestea se rostesc pe trendul descrescător al unității de accentuare iar adjectivul sau adverbul pe cel crescător.
- Unitățile de accentuare care alcătuiesc grupul nominal intră în aceiaşi unitate intonațională.
- Particula de negație *nu* preia accentul principal (H*) din cadrul locuțiuni verbale iar pe verb se realizează un accent secundar de tip *H+!H**.
- Particula *nici,* exprimând şi ea o negație, se rosteşte pe un ton țintă semnificativ High.
- Împărțirea în unități intonaționale păstrează structura sintactică a textului în sensul că niciodată o unitate intonațională nu va conține cuvinte care aparțin la două unități sintactice diferite.

5.3.2 Studiu de caz pentru intonația propozițiilor interogative totale

Propozițiile interogative totale (*Yes-No Question*) fac obiectul analizei comparative a realizării intonației în diferite limbi. Concluziile unei astfel de analize sunt prezentate de Ladd în lucrarea sa *Intonational Phonology* din 1996, sau de Hirst şi Di Cristo în 1998, în care sunt luate în discuție şi câteva exemple din limba română.

Studiul început în colaborare cu specialişti din domeniul lingvisticii (Turculeț 2006), referitor la modalitățile de realizare a intonației interogative totale din limba română, a avut

în intenție să concretizeze caracterizările făcute de cercetătorii lingvişti, prin descrieri pe baza evenimentelor acustice extrase din semnalul vocal. Ulterior am continuat cercetările cu scopul de a realiza descrierea prozodiei propozițiilor interogative totale cu secvențe de etichete ToBI şi durata silabelor accentuate (Apopei 2006a). Descrierile obținute în această etapă au pus în evidență faptul că pe lângă contururilor intonaționale propuse de Ladd (1996) şi cele propuse de L. Dascălu-Jinga (2001) mai există şi alte variante intonaționale. Preocupat de utilizarea descrierilor prozodice, în sinteza vocală pentru propozițiile interogative totale, am ajuns să obțin un inventar al formelor pentru conturul frecvenței F0 (Apopei 2008).

Caracterizările intonației interogative totale efectuate de L. Dascălu-Jinga se referă la cele două caracteristici principale ale intonației interogative: emfaza interogativă (cel mai proeminent accent sau cuvântul la care se referă întrebarea) şi forma conturului final (conturul melodic final care începe cu ultima silabă accentuată). Referitor la conturul terminal al unei propoziții interogative totale L. Dascălu-Jinga (2001) afirmă: este *Ascendent*, indiferent de poziția emfazei interogative în cazul cuvintelor finale oxitone (se termină cu silaba accentuată); este *Ascendent*, în cazul cuvintelor finale neoxitone, când emfaza interogativă este pe finalul propoziției (caz notat cu E); este *Ascendent-Descendent* în cazul cuvintelor finale neoxitone, când emfaza interogativă nu este pe finalul propoziției (caz notat cu NE). Referitor la emfaza interogativă, autoarea precizează faptul că aceasta se caracterizează printr-o proeminență negativă adică ton coborât şi/sau descendent.

Descrierea intonației pentru limba română, propusă de către Ladd (1996), se face pornind de la rostirile neutrale ale enunțurilor (5.1) şi (5.2) în următorii termeni : se stabileşte poziția şi tipul de ton al accentului nuclear notat cu „*" ; se stabileşte forma conturului final prin secvența de tonuri HL; se stabileşte poziția şi tipul de ton a celui de-al doilea accent proeminent în una din variante.

Ladd exprimă aceeaşi idee ca L. Dascălu-Jinga conform căreia accentul nuclear se realizează printr-o proeminență negativă, şi ca urmare acesta este marcat cu simbolul L*.

Astfel, pentru varianta intonațională NE (L. Dascălu-Jinga) în care emfaza interogativă se realizează pe verbul *"văzut"* descrierea dată de Ladd este următoarea:

Ai vă-zut a-fi-şul a-ces-ta? *(5.1)*
L H L*

Ai văzut regele? *(5.2)*
L HL*

Descrierea intonației în cel de-al doilea caz (E) în care emfaza interogativă se realizează pe final, este următoarea:

Ai vă-zut a-fi-şul a-ces-ta? *(5.3)*
H L HL*

Ai văzut regele? *(5.4)*
*H L*HL*

În încercarea de a găsi o descriere comună între cele două variante intonaționale, Ladd concluzionează următoarele:

– accentul L* se produce pe silaba accentuată din cadrul emfazei interogative
– secvența HL se produce pe ultima silabă accentuată şi următoarele neaccentuate, în cazul emfazei în poziție ne-terminală (NE)
– secvența HL se produce pe ultima silabă neaccentuată, în cazul emfazei în poziție terminală (E).

Descrierile celor doi autori diferă doar în ceea ce priveşte conturul terminal în cazul variantei intonaționale cu emfaza în poziție finală .

În studiul efectuat am intenționat să concluzionăm prin ce fel de evenimente fonologice

(accente de pitch, tonuri de graniță) se poate descrie intonația rostirilor interogative totale în limba română şi să comparăm concluziile noastre cu afirmațiilor autorilor Ladd şi L. Dascălu-Jinga. Rezultatele studiul efectuat adăugă celor descrise de autorii citați o serie de caracterizări referitoare la: împărțirea curbei melodice a unui enunț, în cazul general al rostirilor neneutrale, în unități intonaționale; descrierea conturului F0 al unei unități intonaționale prin secvențe de tonuri şi alte mărimi acustice (durată, energie); stabilirea poziției emfazei interogative atât în cazurile neutrale cât şi al celor neneutrale. Analiza intonației interogative totale în limba română s-a efectuat pe un corpus de voce construit după metodologia prezentată în lucrarea (Apopei 2008).

5.3.2.1 Prezentarea rezultatelor analizei

Contururile melodice rezultate din rostiri au fost împărțite în unități intonaționale conform ierarhiei prezentate în secțiunea 4.4.1. Astfel am reuşit să pun în evidență rostiri desfăşurate într-o singură frază intonațională şi rostiri ne-neutrale formate din două sau trei fraze intermediare. În cadrul fiecărei unități intonaționale s-au indicat accentele şi tonurile semnificative cu ajutorul etichetelor prezentate în secțiunea 4.4. Frazele intermediare se formează când în rostire apar accente sintactice în poziții diferite de cea a emfazei interogative, realizate prin accente de pitch aproximativ de aceeaşi proeminență.

Descrierea variantelor intonaționale ale rostirilor analizate s-a făcut pe baza următoarelor trăsături acustice şi fonologice: accentele şi tonurile principale din cadrul fiecărei unități intonaționale; poziția emfazei interogative; durata părților sonore din cadrul silabelor accentuate asociate accentelor principale; poziția silabelor accentuate de energie maximă.

Pentru prezentarea variantelor intonaționale am folosit aceleaşi reprezentări grafice şi simbolice ca cele utilizate în lucrarea (Apopei & Jitcă 2008). Pentru reprezentarea grafică a variantelor intonaționale am folosit o reprezentare schematică a conturului median a frecvenței F0 şi următoarele codificări: granițele de subunități intonaționale (*ip*-paranteze acoladă, *RU*-paranteze pătrate); etichete pentru tonurile principale de pe conturul frecvenței F0; poziția emfazei interogative (E); segmentele crescătoare/descrescătoare ale conturului final au fost notate cu (R) şi repectiv (F).

Pentru reprezentarea simbolică a variantelor intonaționale am folosit etichete tonale din setul prezentat în secțiune 4.4 grupate în concordanță cu ierarhia intonațională (*ip*- paranteze acoladă, *RU* - paranteze pătrate). Pentru marcarea emfazei interogative evenimentul tonal corespunzător a fost subliniat.

Varianta intonațională „V1”

La varianta intonațională V1 contorul melodic al frecvenței F0 este generat printr-o secvență de accente gramaticale (*stress sequence*) care se termină cu un ton final ascendent. Unul din cuvintele frazei intonaționale este mai proeminent prin durata şi/sau energia silabei accentuate sau printr-un *pitch accent* de tip L*. Acel cuvânt este purtătorul emfazei interogative şi determină tipul de accent de pitch de pe ultima silabă accentuată a frazei intonaționale.

Dacă emfaza interogativă este în poziție non-finală accentul de pitch de pe ultimul cuvânt este de tip H* sau L+H* şi generează un segment ascendent în conturul final al frecvenței F0. Prezența unui cuvânt non-oxiton în poziția finală determină după segmentul ascendent al conturului (datorat accentului de pitch) un segment descendent.

Dacă emfaza interogativă este pe primul cuvânt non-clitic al frazei intonaționale (fig. 5.7.a) conturul mediu al frecvenței F0 prezintă o mică creştere pe cuvintele intermediare ale frazei intonaționale. Dacă emfaza este în poziție mediană, conturul frecvenței F0 începe de

la un nivel puțin mai ridicat după care scade până la o valoare minimă pe cuvântul purtător al emfazei interogative (fig. 5.7.b). Cuvântul proeminent al emfazei interogative prezintă de cele mai multe ori pe lângă nivelul tonal scăzut şi o durată/energie mai mare pentru silaba accentuată.

Prezența emfazei interogative în poziție finală determină pe ultimul cuvânt non-clitic al frazei intonaționale un accent de pitch de tip L* sau L*+H. În acest caz ultimul cuvânt este purtător a două evenimente fonologice: emfaza interogativă determinată de tonul țintă de nivel scăzut şi segmentul ascendent al conturului final. În cazul cuvintelor finale oxitone, conturul final conține numai segmentul ascendent (fig. 5.7.c). În cazul cuvintelor finale non-oxitone conturul final poate fi unul descendent–ascendent sau unul descendent–ascendent–descendent (fig. 5.7.d).

Conturul din fig. 5.7.d l-am întâlnit pe rostiri non-neutrale cu cuvântul final

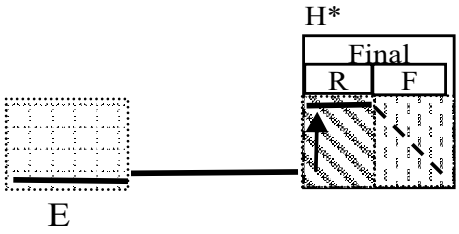

Fig. 5.7.a. Conturul schematic al frecvenței F0 al variantei intonaționale V1 cu emfaza în prima poziție

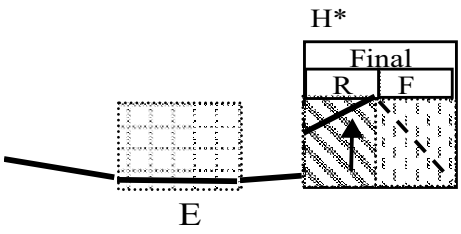

Fig. 5.7.b. Conturul schematic al frecvenței F0 al variantei intonaționale V1 cu emfaza în poziție mediană

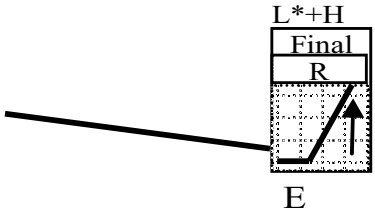

Fig. 5.7.c. Conturul schematic al frecvenței F0 al variantei intonaționale V1 cu emfaza în poziție finală şi contur final ascendent

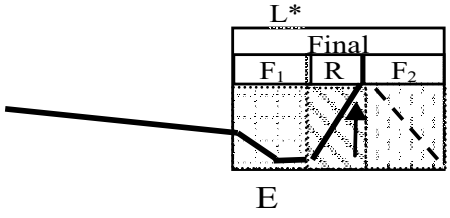

Fig. 5.7.d. Conturul schematic al frecvenței F0 al variantei intonaționale V1 cu emfaza în poziție finală şi contur final descendent-ascendent (-descendent)

proparoxiton. Acest contur corespunde cu descrierea făcută de Ladd (secvențele tonale (5.3) şi (5.4)) şi a fost întâlnită pe rostiri mai puțin neutrale decât cele realizate cu un contur ca cel din fig. 5.7.c.

Varianta intonațională „V2”

Varianta intonațională V2 se poate obține din varianta V1 prin transformarea unui accent gramatical de la începutul frazei intonaționale din simplu *stress* într-un accent de pitch de tipul H* sau L+H*. În consecință pe conturul frecvenței F0 apar două segmente ascendente (fig. 5.8). Primul segment este datorat accentului de pitch de tip H* şi produce o creştere a nivelului frecvenței F0 la un nivel intermediar. Cuvintele care se rostesc la nivel intermediar prezintă accente gramaticale de tip *stress* şi se rostesc pe un trend uşor descendent, iar unul dintre cuvinte are silaba accentuată cu durată mai mare fapt ce determină poziția emfazei interogative. Am grupat în această variantă contururile pentru frecvența F0 care au prezentat pe cuvântul din poziție finală un accent de tip H* şi emfaza interogativă în poziție non-finală. Cele cu emfaza interogativă în poziție finală le-am grupat în varianta V3.

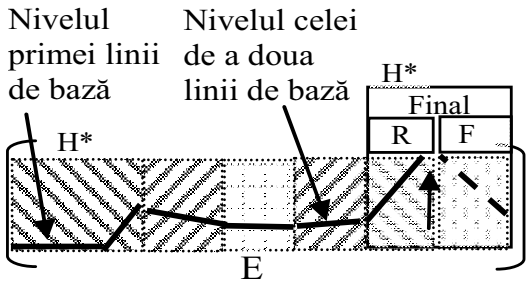

Fig. 5.8. Conturul schematic al frecvenței F0 al variantei intonaționale V2 cu emfaza în poziție non-finală

Varianta intonațională „V3”

Varianta intonațională V3 poate fi văzută ca derivând din varianta V2 la care emfaza interogativă este în poziție finală. Cele două cuvinte proeminente corespund la un accent de pitch de tip H* (uzual cuvântul *temă*) în prima parte a frazei intonaționale şi celălalt este cuvântul final pe care se află emfaza interogativă cu accent tip L* sau L*+H (uzual cuvântul *remă*).

Primul accent de pitch de tip H* sau L+H* generează prima creştere a conturului frecvenței F0. După această creştere, conturul frecvenței F0 prezintă un tendință descrescătoare, pe care se pot afla cuvinte cu accent de tip *stress,* iar pe ultimul cuvânt apare un accent de tip L* or L*+H.

În funcție de poziția relativă nivelului de „*low”* de pe ultimul cuvânt şi nivelul tonal de la începutul frazei intonaționale am pus în evidență trei sub-variante. La prima sub-variantă (fig. 5.9.a) nivelul de „*low”* de pe ultimul cuvânt este foarte apropiat de nivelul tonal de la începutul frazei intonaționale. Celelalte două sub-variante au un grad diferit de emoție. A doua sub-variantă (fig. 5.9.b) are nivelul de „*low”* de pe ultimul cuvânt mai ridicat decât nivelul tonal de la începutul frazei intonaționale şi corespunde intonațiilor intonaționale care cresc tensiunea în vorbire (interogație cu mirare, interogație cu bucurie). A treia sub-variantă (fig. 5.9.c) are nivelul de „*low”* de pe ultimul cuvânt mai coborât decât nivelul tonal de la începutul frazei intonaționale şi corespunde intonațiilor intonaționale care scad tensiunea în vorbire (interogație cu dezamăgire, interogație cu supărare).

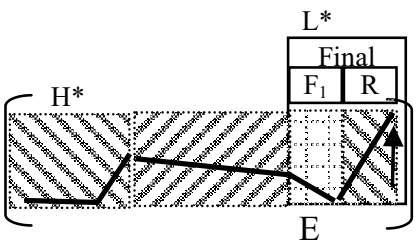

Fig. 5.9.a Conturul schematic al frecvenței F0 al variantei intonaționale V3 cu emfaza în poziție finală

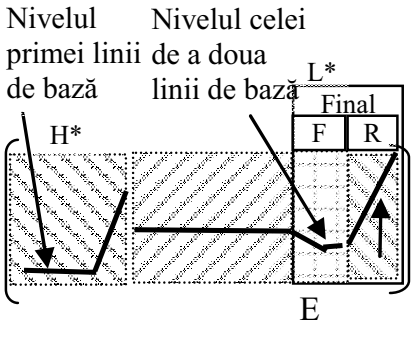

Fig. 5.9.b Conturul schematic al frecvenței F0 al variantei intonaționale V3 şi nivelul de ton pentru accentul L* mai ridicat decât nivelul de început al frazei intonaționale

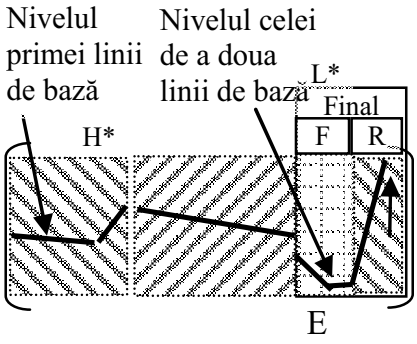

Fig. 5.9.c Conturul schematic al frecvenței F0 al variantei intonaționale V3 şi nivelul de ton pentru accentul L* mai scăzut decât nivelul de început al frazei intonaționale

Din analiza acestor sub-variante intonaționale putem afirma că sub-variantele emoționale implică o schimbare a nivelului linei de bază (nivelul de *„low”*) şi pot fi interpretate ca rezultate prin suprapunerea secvenței de tonuri de pe conturul din fig. 5.9.a pe tendință crescătoare (fig. 5.9.b) sau pe una descrescătoare (fig. 5.9.c). Gradul de emoție al rostirilor rezultate cu aceste variante depinde de gama de frecvență în care se realizează tonurile finale.

Varianta intonațională „V4”

Varianta intonațională V4 am întâlnit-o numai în rostirile considerate de noi emoționale. Conturul intonațional al unei fraze intonaționale conține o secvență de unități ritmice (uzual două unități ritmice), fiecare formată dintr-o unitate de accentuare cu intonație interogativă proprie. Cuvântul din unitatea de accentuare inițială are un accent de tip *L+H** sau secvențe de tonuri *(L*, H+)* iar unitatea de accentuare din poziție finală are un accent de tip L*, L*+H sau H+L*

Suprapunerea acestei secvențe de tonuri pe o tendință crescătoare (fig. 5.10.a) determină o creştere a nivelului emoției tensiunii în rostire. În figura 5.10.a. un accent de tip *L* H+* generează o intonație interogativă pentru prima unitate ritmică şi al doilea accent de tip *L** realizează focusul pe cuvântul final.

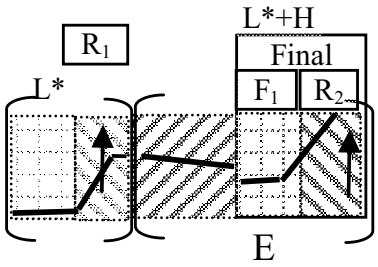

Fig. 5.10.a. Conturul schematic al frecvenței F0 al variantei intonaționale V4 şi contur final ascendent

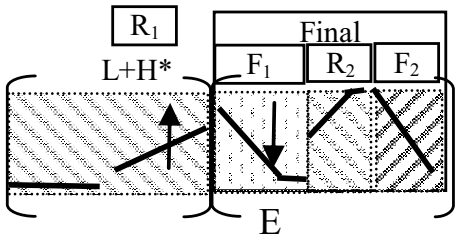

Fig. 5.10.b. Conturul schematic al frecvenței F0 al variantei intonaționale V4 şi contur final descendent-ascendent-descendent

În fig. 5.10.b este prezentată o rostire emoțională cu un cuvânt final proparoxiton. Primul accent de tip L+H* generează o intonație interogativă pentru prima unitate ritmică şi a al doilea accent de L* realizează focusul pe cuvântul final. Conturul frecvenței F0 corespunzător la ultimele două silabe neaccentuate au contur de tip ascendent-descendent cu variații mari ale frecvenței F0.

Varianta intonațională „V5”

Varianta intonațională V5 a fost întâlnită la vorbitorii din N-V Ardealului. Această intonație este caracterizată de poziționarea emfazei interogative în poziție non-finală. Conturul median al frecvenței F0 păstrează un nivel scăzut la începutul frazei intonaționale, până şi pe durata cuvântului cu emfaza interogativă, după care începe a creşte până la un nivel „*high*” pe care-l atinge pe silaba dinaintea ultimei silabe accentuate (fig. 5.11.a). Pe durata ultimei silabe accentuate conturul frecvenței F0 scade până la nivel de *„low”*. Creşterea conturului intonațional poate începe sau nu de pe silaba accentuată a cuvântului cu emfaza interogativă.

Există vorbitori care generează la finalul frazei intonaționale un mic segment crescător R_2 după descreşterea de pe silaba accentuată (fig. 5.11.b). În acest caz apare un contur final de tip descendent-ascendent iar descrierea intonației poate fi făcută cu următoarea secvențe

de tonuri: {%L stress H+ ^H* ^H+ H+L* L-!H%}.

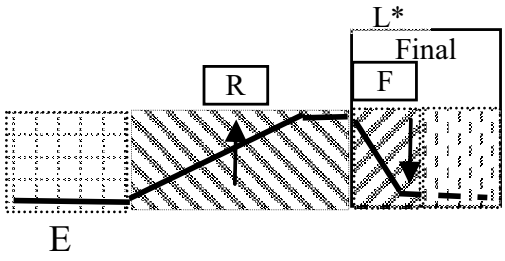

Fig. 5.11.a. Conturul schematic al frecvenței F0 al variantei intonaționale V5 şi contur final descendent

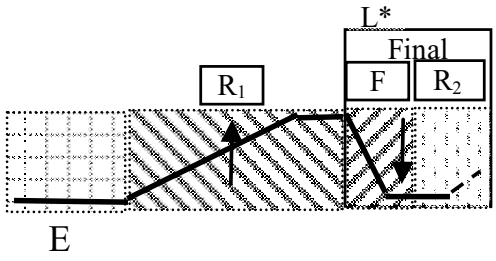

Fig. 5.11.b. Conturul schematic al frecvenței F0 al variantei intonaționale V5 şi contur final descendent-ascendent

Accentele de tip *stress* de pe tendința crescătoare (dintre tonurile H+ şi ^H+) nu sunt proeminente deoarece ele corespund unei mişcări în aceiaşi direcție a conturului frecvenței F0.

5.3.2.2 Concluzii pe baza rezultatelor analizei intonației integrațiilor totale

Analiza conturului intonațional pentru propozițiile interogative totale a dus la identificarea pentru frazele intonaționale a un număr de cinci variante intonaționale. Pe baza acestor variante intonaționale, conturul melodic al propozițiilor interogative totale poate fi realizat din mai multe fraze intonaționale separate prin tonuri de graniță şi pauze. În general, acestor segmente de pe conturul frecvenței F0 le corespund la nivelul textului diferite grupuri sintactice. Fiecare frază intonațională are propria ei emfază interogativă dar ultima este cea care le domină pe cele precedente. Variantele intonaționale prezentate de L. Dascălu-Jinga (2001) şi Ladd (1996) se regăsesc în variantele intonaționale prezentate în secțiunea 5.3.2.1.

Varianta V1 corespunde la intonația descrisă de L. Dascălu-Jinga în două cazuri diferite: cazul emfazei finale şi cel al emfazei în poziție nonfinală. Ea pune în evidență cazurile cu contur final ascendent şi respectiv, ascendent-descendent. În urma acestei analize rezultă posibilitatea de a genera emfaza finală cu accent de tip L* (cu descreşterea frecvenței) şi realizarea unui contur final de tip descendent-ascendent.

Varianta intonațională descrisă de Ladd (1996) pentru cazul accentului nuclear în poziție finală corespunde cu varianta V4 din această prezentare. Această variantă reprezintă un caz particular de contur melodic pentru intonația propozițiilor interogative totale din limba română şi anume cele caracterizate de un contur final descendent-ascendent-descendent şi care au cuvânt final proparoxiton. În mod regulat pentru intonația propozițiilor interogative totale cu emfaza finală am întâlnit varianta V3, la care conturul final este de tip descendent-ascendent.

Varianta intonațională V5 a fost identificată pe rostiri ale unor vorbitori provenind din regiunea N-V a României (Ardeal) şi se caracterizează printr-o creştere continuă a conturului median al frecvenței F0 între cuvântul purtător al emfazei interogative şi începutul ultimei silabe accentuate din fraza intonațională.

Pentru realizarea intonațiilor non-neutrale propozițiile interogative totale se divizează în mai multe fraze intonaționale iar fiecare unitate sintactică poate să devină purtătoare de accente proeminente. Pentru realizarea silabelor proeminente vorbitorii apelează la *„pattern-uri”* de durată şi energie suprapuse peste *„pattern-uri”* la nivelul frecvenței F0.

Consider că o clasificare chiar şi cu grade de încredere fuzzy a tipurilor de emoție percepute pentru o anumită intonație (o intonație poate transmite mai mute tipuri de emoție) împreună cu analiza sintactico-semantică a textului ar putea duce la generarea de intonații

interogative cât mai naturale pentru propozițiile interogative totale. Această clasificare ar ajuta la o asociere a variantelor de contur intonațional la unitățile ierarhiei intonaționale corespunzătoare textului de intrare.

5.4 Aspecte ale implementării intonației în sinteza vocală

Conturul frecvenței F0 corespunzător unei fraze intonaționale/intermediare poate fi văzut ca o succesiune de pattern-uri corespunzătoare unităților ritmice ce le compun. La rândul lor, pattern-urile unităților ritmice rezultă din secvențierea pattern-urilor unităților de accentuare componente. Acestea din urmă sunt adecvate evenimentelor tonale ce le conțin şi poziției în cadrul unităților ritmice şi frazei intonaționale. Partea semnificativă a pattern-ului unei AU este cea care generează evenimentul sau secvența de evenimente tonale. Vom numi această parte drept pattern de eveniment/evenimente, cum ar fi de exemplu, accentul de pith de tip H* sau L*, accentul de tip stress, accentele de frază intermediară şi tonurile de graniță sau o combinație a acestora (Apopei 2007, Raport 2007b).

Pattern-urile de eveniment tonal/secvență se desfăşoară în cadrul unităților de accentuare respective şi implică silaba accentuată iar uneori pe cea anterioară şi următoare acesteia. Acestea pot fi descrise folosind secvențe de etichete ToBI corespunzătoare accentelor de pitch şi unor tonuri din una din categoriile următoare: accente de frază (H-, L-), tonuri de graniță (H%, L%) şi alte tonuri semnificative din punct de vedere al pattern-ului, notate de noi H+, L+.

Modulul de generare a frecvenței F0, implementat în această etapă, se bazează pe schema bloc a modulului fonetic prezentată în figura 5.4. Intrarea modulului fonetic a fost formată dintr-un fişier XML care conține textul împărțit în silabe şi structurat în unități de intonație pe baza ierarhiei din secțiunea 4.4 (frază intonațională, frază intermediară, unitate de accentuare, grup de unități de accentuare). Cu ajutorul acestei ierarhii se poate realiza frazarea unui text de intrare precum şi stabilirea proeminențelor evenimentelor intonaționale.

Submodulul fonetic prelucrează secvența de silabe de la intrare şi generează secvența de foneme prin care se realizează rostirea sintetizată. Fiecare fonem prin atributul de durată generează o serie de timp care este folosită pentru desfăşurarea pe axa timpului a undei vocale şi a tonurilor de pe conturul frecvenței F0. Pentru stabilirea duratei şi energiei fonemelor, sistemele recente de sinteză vocală au implementate modele de durată şi energie (intensitatea) sunetelor. Secvența de descrieri parametrice, pentru fonemele prin care se materializează rostirea, este transformată în semnale de comandă pentru intrarea modulului de sinteză vocală.

În generarea conturului F0 am prevăzut efectuarea a două etape de prelucrare corespunzătoare celor două submodule al modulului fonetic: cel de modelare a conturului intonațional şi cel de generare propriu zisă. Primul submodul transformă arborele rostirii format din structura unităților intonaționale aşa cum este descrisă în fişierul XML de la intrarea modulului fonetic, într-o secvență de evenimente de contur F0 care au asociate câte un pattern al frecvenței F0 şi o poziție în spațiul (timp, frecvență), definită prin limitele maxime şi minime între care acesta se desfăşoară (figura 5.12). Sunt mapate mai întâi frazele intonaționale şi intermediare împreună cu tendințele lor de *downsteping* sau *upsteping*. Apoi în cadrul acestor limite sunt fixate pozițiile unităților ritmice şi a unităților de accentuare. Pentru fiecare unitate de accentuare am delimitat o secvență de patru regiuni (figura 5.5) cu ajutorul cărora am definit mişcările de pitch pentru evenimentele de pe conturul frecvenței F0 ce trebuie sintetizat:

- Un segment de variație a frecvenței F0 pe durata silabei anterioare celei accentuate (*segmentul I*)
- O porțiune de salt în frecvență pe porțiunea consonantică a silabei accentuate (dacă

aceasta există -*segmentul II*)

- O variație continuă a frecvenței pe durata nucleului silabei accentuate (*segmentul III*)
- O variație continuă sau un salt în frecvență pe silabele neaccentuate ce urmează celei accentuate (*segmentul IV*)

Variațiile şi salturile în frecvență pot fi şi crescătoare şi descrescătoare. Un pattern particular depinde de structura fonetică a cuvântului şi de proeminența evenimentelor şi acestea determină raporturi diferite între proiecțiile acestor segmente pe cele două axe: a timpului şi a frecvenței.

Pentru un eveniment corespunzător unui accent de pitch de tip H* urmat de un ton țintă ^H+, pattern-ul este cel din figura 5.12.a. Un accent sintactic proeminent de tip H* poate fi generat prin creşterea gamei de frecvență în cadrul segmentului III şi scăderea acesteia în cadrul segmentului II. În plus segmentul IV se transformă într-unul descrescător, pentru că pe silaba accentuată se atinge maximul tonal (figura 5.12.b). Când accentul H* este conținut de prima unitate de accentuare din cadrul unei fraze intonaționale, acesta nu este proeminent iar tonul țintă înalt care trebuie atins la începutul frazei va implica o variație crescătoare suplimentară pe silaba/silabele neaccentuate următoare din cadrul unității de accentuare. Pentru generarea pattern-ului în acest caz am stabilit o gama de variație mică pentru segmentul III (silaba accentuată) dar mai mare pentru segmentul IV (figura 5.12.c). În cazul când evenimentul de tip H* apare în cadrul ultimei unități de accentuare dintr-o frază intonațională ridicarea pe segmentele II şi III nu are amplitudine şi proeminența acestuia se realizează prin căderea bruscă pe silaba imediat următoare celei accentuate (segmentul IV), ca în figura 5.12.d.

În general în propozițiile afirmative, pattern-urile accentului de pitch crescător când acesta apare în cadrul ultimei unități de accentuare dintr-o unitate ritmică, nonfinală, este de tip L+H*, ca în figura 5.12.e. Pe durata segmentului II tonul se menține la nivel *low* sau creşte lent, urmând ca segmentul III să genereze ridicarea rapidă până la tonul țintă *high*. Pentru evenimentele de tip L* asociate cu secvența de tonuri L- H% sau accentele de frază de tip H-, pattern-ul este cel din figura 5.12.f în care tonul țintă *low* se atinge în cadrul segmentului III şi urmează ridicarea pe segmentul IV pentru tonurile finale. În afară de accentul L* cu variație descrescătoare pe silaba accentuată, acesta se poate genera prin menținerea la un ton *low* pe toată durata silabei accentuate, urmată de o creştere bruscă pe silaba imediat următoare ca în figura 5.12.h. Evenimentul L* asociat cu secvența de tonuri H-H% generează patter-nul din figura 5.12.g.

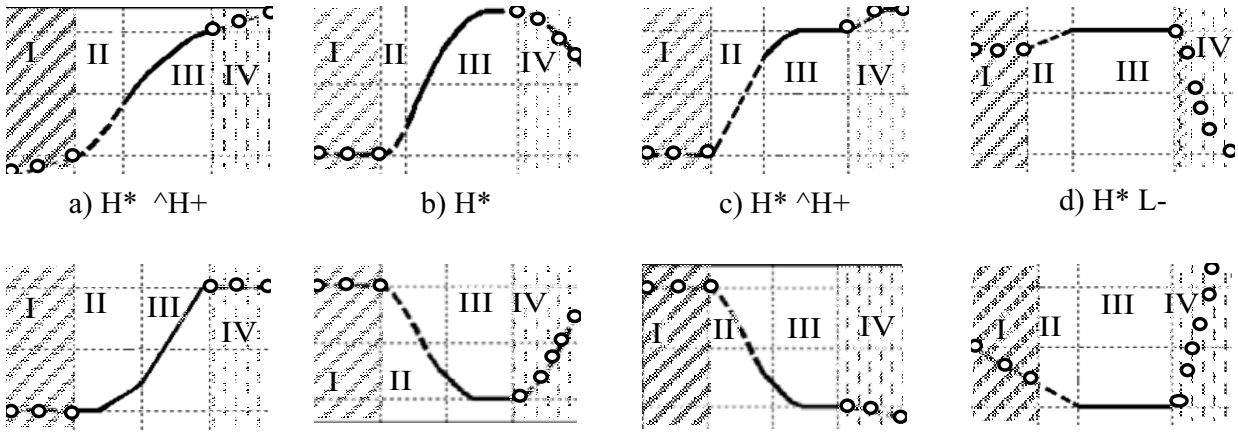

Figura 5.12 Pattern-uri pentru evenimente tonale din cadrul unităților de accentuare

Revenind la nivelul pattern-ului unei unități de accentuare, precizăm că după stabilirea pattern-ului de eveniment/secvență de evenimente, completarea pattern-ului unității de accentuare se face prin interpolare lineară cu tonurile de sfârşit şi de început a unității precedente, şi respectiv, următoare.

Într-o unitate ritmică pe lângă unitățile de accentuare cu evenimente de pitch, pot exista şi unități cu evenimente de tip simplu accent gramatical (stress). Pattern-urile evenimentelor de tip stress care apar pe un nivel aproximativ constant al conturului F0 mediu în cadrul unității ritmice, fie la nivel low, fie la nivel high este cel din figura 5.13.a, cu variație crescătoare pe silaba accentuată, respectiv cel din figura 5.13.b când se realizează cu variație descrescătoare pe silaba accentuată. Alte două pattern-uri pentru evenimentul de stress este cel care se desfăşoară pe o tendință de downsteping sau upsteping. În cazul particular al silabei accentuate în poziție de început a unității de accentuare şi în condițiile tendinței de downsteping, pattern-ul este în esență o treaptă căzătoare ca în figura 5.13.c. În cazul particular al silabei accentuate în poziție finală a unității de accentuare şi în condițiile tendinței de upsteping, pattern-ul constă într-o treaptă crescătoare ca în figura 5.13.d.

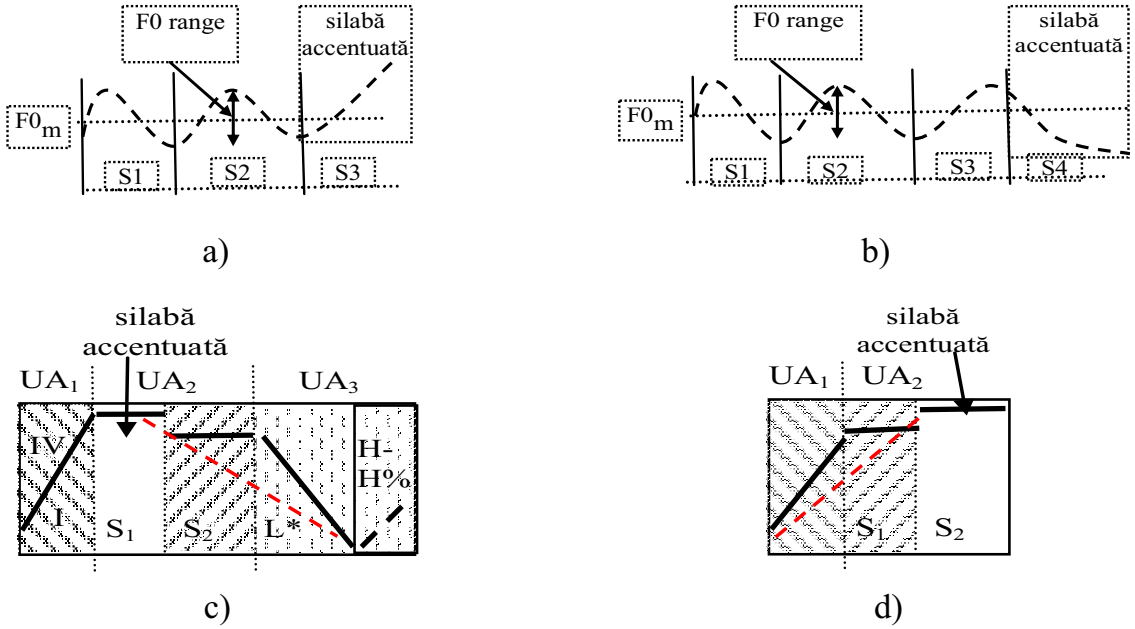

Figura 5.13 Pattern-uri pentru evenimente accente de tip stress din cadrul unităților de accentuare

Aceste pattern-uri de eveniment corespund părților semnificative din conturul F0 la nivelul unităților de accentuare. Completarea conturului la nivelul unităților de accentuare se face prin interpolare lineară între ultimul ton al pattern-ului de eveniment al unității precedente şi primul ton al pattern-ului de eveniment al unității curente. Segmentele de contur generate pentru unitățile de accentuare sunt poziționate în spațiul (timp, frecvență) de modulul de modelare a conturului intonațional prin aplicarea unor pattern-uri la nivelul unităților ritmice pe care le compun şi ținând cont de poziția unității ritmice în cadrul frazelor intonaționale/intermediare .

5.5 Generarea conturului frecvenței F0

Modulul de generare a frecvenței F0, implementat în cadrul sistemului text-voce se bazează pe schema bloc a modului fonetic prezentată în figura 5.4. În această etapă a cercetărilor, intrarea modulului fonetic este constituită dintr-un fişier XML care conține textul împărțit în silabe şi structurat în unități de intonație pe baza ierahiei din secțiunea 4.4 (frază intonațională, frază intermediară, unitate de accentuare, grup de unități de accentuare).

Submodulul fonetic prelucrează secvența de silabe de la intrare şi generează secvența de foneme prin care se realizează rostirea sintetizată. Fiecare fonem prin atributul de durată generează o serie de timp care este folosită pentru desfăşurarea pe axa timpului a undei vocale şi a tonurilor de pe conturul frecvenței F0. Pe baza secvenței de descrieri parametrice are loc generarea semnalelor de comandă de la intrarea modulului de sinteză vocală.

În generarea conturului F0 am prevăzut efectuarea a două etape de prelucrare (Apopei & Jitcă 2007) corespunzătoare celor două submodule al modulului fonetic: cel de modelare a conturului intonațional şi cel de generare propriu zisă. Primul submodul translatează arborele rostirii format din structura unităților intonaționale aşa cum este descrisă în fişierul XML de la intrarea modulului fonetic, într-o secvență de evenimente de contur F0 care au asociate câte un pattern al frecvenței F0 şi o poziție în spațiul (timp, frecvență), definită prin limitele maxime şi minime între care acesta se desfăşoară (vezi figura 5.14 în care este prezentat schematic modul în care este interpretată informația din structura prozodică). Sunt mapate mai întâi frazele intonaționale şi intermediare împreună cu tendințele lor de *downsteping* sau *upsteping*. În cadrul acestor limite sunt fixate pozițiile unităților ritmice (grupurilor de unități de accentuare) şi a unităților de accentuare.

În figura 5.14 este ilustrat modul în care este interpretată informația din structura prozodică a textului de intrare *„Sunteți gata să ucideți?”* şi apoi mapată în regiuni ale spațiului (timp, frecvență).

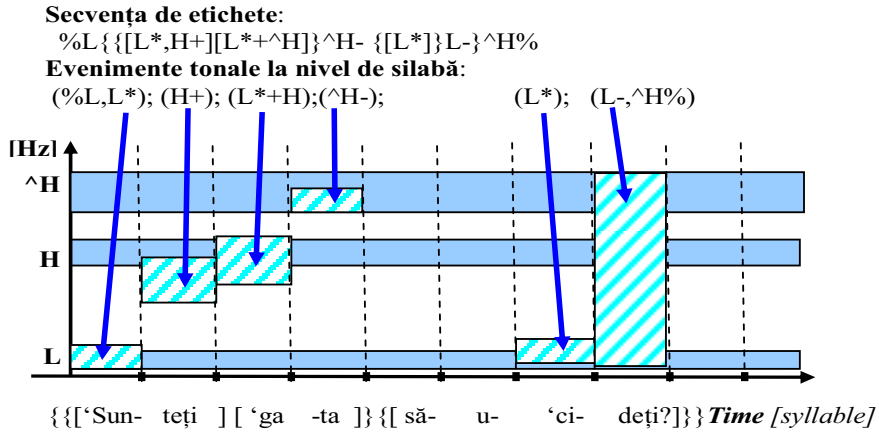

Figura 5.14 Poziționarea evenimentelor din descrierea XML în spațiul (timp, frecvență)

Rostirea de sintetizat este formată din două fraze intermediare, cea dintâi având o tendință de upsteping pe care se desfăşoara două unități ritmice, fiecare cu câte o unitate de accentuare.

Secvența de evenimente pusă în evidență în acest caz este următoarea:

- tonul de început şi punctul țintă *low* a primului accent de pitch L* (silaba */sun/*);
- ridicarea de ton de pe silaba neaccentuată */țeți/* care creează accentul L* împreună cu tonul anterior;
- accentul de pitch L*+H cu tonul țintă pe silaba accentuată /*ga*/;
- tonul de *„accent phrase”* cu care se termină prima frază intermediară (H-);
- accentul de pitch L* al ultimului cuvânt cu tonul țintă pe silaba */ci/*;
- tonurile L-H% pe ultima silabă.

Stabilirea regiunilor şi a pozițiilor acestora în timp şi frecvență este realizată pe bază de reguli euristice deduse din analiza mai multor rostiri afirmative sau interogative.

Submodulul de generare a conturului F0 selectează pentru fiecare eveniment, din secvența generată la pasul anterior, un pattern elementar de contur F0 (forme de contur de tipul celor prezentate în fig. 5.12) în acord cu: etichetele tonale din fişierul XML, poziția în cadrul unităților intonaționale şi contextul fonetic. Pattern-urile sunt poziționate şi apoi scalate astfel încât să se încadreze în regiunile delimitate prin gama de variație a frecvenței

F0 şi durata fonemelor (figura 5.15).

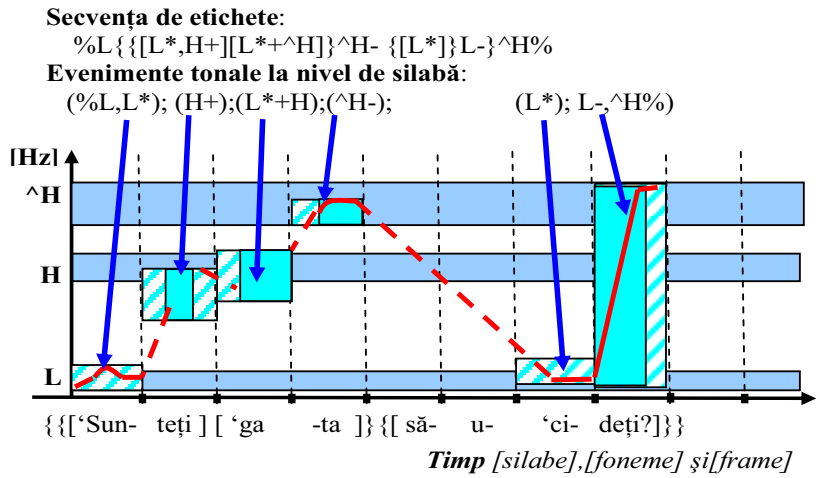

Figura 5.15 Generarea pattern-urilor pentru evenimentele semnificative de pe conturul frecvenței F0

Prin desfăşurarea pattern-urilor de evenimente tonale în interiorul regiunilor delimitate în etapa anterioară, se generează segmentele semnificative din cadrul conturului F0. În această etapă a cercetărilor, submodulul de generare a conturului F0 „leagă” formele de contur generate pentru evenimentele tonale prin segmente de interpolare liniară.

Funcționarea acestui submodul se bazează existența unui inventar de pattern-uri de evenimente cu variante pentru diferite contexte semantice, emoționale şi tip de intonație (afirmativă, exclamativă sau interogativă).

5.6 Contribuții personale

În această etapă am propus o schemă de adnotare a elementelor prozodice în format XML care să poată fi folosită atât în analiza contururilor intonaționale (Apopei ş.a. 2006) cât şi în sinteza vocală prozodică. La proiectarea acestei scheme s-au avut în vedere unitățile intonaționale din ierarhia prezentată în secțiunea 4.4, creând câte un tag pentru marcarea unităților de pe fiecare nivel. Tag-urile utilizate împreună cu atributele lor au fost prezentate în lucrarea (Apopei ş.a. 2006).

Pentru folosirea acestei scheme în adnotarea corpusurilor de voce, am prevăzut atribute pentru etichetarea fonologică a evenimentelor şi pentru informații cantitative legate de nivelul tonurilor țintă asociate evenimentele etichetate. În acest scop s-a împărțit gama de variație a frecvenței F0 în cadrul rostirii de adnotat în semitonuri şi s-a realizat o scală de măsurare a tonurilor cu baza la nivelul tonului celui mai înalt din rostire.

Următoarea contribuție constă în rezultatele unui studiu referitor la relația dintre elementele prozodice, structura sintactică şi semantică a textelor asociate rostirilor şi, funcțiile prozodiei. Acest studiu a fost realizat pe două categorii de rostiri: rostiri ale unor propoziții afirmative, şi rostiri ale unor propoziții interogative totale. Analiza contururilor intonaționale a generat observații referitoare la secvențele de evenimente intonaționale de pe conturul unităților intonaționale de tip IP/ip, funcțiile prozodiei, precum şi contextele textuale (structurile silabice) şi semantice care influențează forma conturului F0 în cadrul unităților de accentuare.

Analiza contururilor intonaționale pentru propozițiile afirmative (Raport 2006a), a condus la obținerea unor reguli lingvistice pentru forma evenimentelor prozodice, pentru modul de grupare a cuvintelor în unități unități ale ierarhiei intonaționale şi pentru stabilirea unităților de accentuare proeminente (proeminențe în sensul teoriei autosegmental-metrice).

Conturului melodic al propozițiilor afirmative poate fi realizat din una sau mai multe fraze intonaționale separate prin tonuri de graniță şi pauze.

Analiza conturului intonațional pentru propozițiile interogative totale a dus la identificarea pentru frazele intonaționale a un număr de cinci variante intonaționale. Pe baza acestor variante intonaționale, conturului melodic al propozițiilor interogative totale poate fi realizat din mai multe fraze intonaționale separate prin tonuri de graniță şi pauze. În general, acestor segmente de pe conturul frecvenței F0 le corespund la nivelul textului diferite grupuri sintactice. Fiecare frază intonațională are propria ei emfază interogativă dar ultima frază intonațională este cea le domină pe cele precedente (Apopei 2008).

O altă contribuție constă în implementarea unui modulul de generare a frecvenței F0, în cadrul sistemului text-voce dezvoltat în cadrul institutului. Acest modul realizaeză două etape de procesare a informației intonaționale (Apopei & Jitcă 2007): cel de modelare a conturului intonațional şi cel de generare propriu-zisă.

În prima etapă se translatează arborele rostirii, format din structura unităților intonaționale aşa cum este descrisă în fişierul XML de la intrarea modulului fonetic, într-o secvență de evenimente de contur F0 care au asociate câte un pattern al frecvenței F0 şi o poziție în spațiul (timp, frecvență). Sunt mapate mai întâi frazele intonaționale şi intermediare împreună cu tendințele lor de *downsteping* sau *upsteping*. În cadrul acestor limite sunt fixate pozițiile unităților ritmice (grupurilor de unități de accentuare) şi a unităților de accentuare.

În cea de a doua etapă are loc generarea conturului F0 prin selectartea pentru fiecare eveniment tonal, din secvența generată la pasul anterior, a unui pattern elementar de contur F0 (forme de contur de tipul celor prezentate în fig. 5.12). Această selecție are loc în acord cu: etichetele tonale din fişierul XML, poziția în cadrul unităților intonaționale şi contextul fonetic. Pattern-urile sunt poziționate şi apoi scalate astfel încât să se încadreze în regiunile delimitate prin gama de variație a frecvenței F0 şi durata fonemelor.

Capitolul 6

Contribuții şi direcții de cercetare viitoare

În acest capitol sunt trecute în revistă contribuțiile autorului în domeniul analizei unor sisteme neliniare cu aplicații procesarea semnalului vocal, dezvoltările şi preocupările viitoare. Contribuțiile au rezultat în urma parcurgerii planului de cercetare şi a indicațiilor conducătorului de doctorat. Contribuțiile au vizat, în general, modelarea componentei dinamice a semnalului vocal şi de modelarea aspectelor prozodice ale acestuia. Aceste modelări au fost realizate cu scopul îmbunătățirii sistemului de conversie text-voce din cadrul Institutului de Informatică Teoretică şi de a crea premisele realizării de modelări prozodice pentru alte tipuri de sintetizatoare vocale. Modelările au fost realizate cu sisteme neliniare inteligente, în care neliniaritățile au fost introduse prin reguli, prin indicații (etichete) etc. Principalele probleme abordate sunt: descrierea parametrică a fonemelor şi co-articularea sunetelor; generarea semnalelor de comandă pentru sintetizatorul Klatt; prezentarea metodelor şi algoritmilor de procesare utilizați în etapele de analiză şi adnotare a prozodiei; implementarea rezultatelor obținute în etapele de analiză într-un sistem de conversie text-voce. Rezultatele obținute au fost incluse în lucrări ştiințifice publicate în reviste de specialitate şi prezentate la conferințe şi simpozioane, în rapoarte de cercetare şi participări la proiecte de cercetare. Toate acestea s-au desfăşurat în acord cu planul de cercetare din cadrul Institutului de Informatică Teoretică Iaşi, colectivul de 'Procesare semnale', în perioada anilor 2001-2008 şi cu experiența anterioară acumulată în urma colaborări cu prof. dr. H.N. Teodorescu, CS II N.Curteanu, conf. dr. Fl. Grigoraş şi CS III

dr. D. Jitcă. Cu ultimi doi am colaborat în perioada în care dumnealor se aflau în programul de pregătire şi de elaborare a tezelor de doctorat.

6.1 Contribuții la modelarea componentelor neliniare ale semnalului vocal

Principalele contribuții ale autorului se referă la: (1) implementarea a două metode de detecție a frecvenței fundamentele F0; (2) modelarea co-articulării fonemelor cu funcții de dominanță neliniare şi îmbunătățirea tranzițiilor formanților între foneme; (3) proiectarea pe baza fonologiei autosegmental-metrice a unei ierarhii de unități intonaționale pentru modelarea fonologică a intonației din limba română; (4) proiectarea unei scheme XML pentru adnotarea prozodică a textelor de la intrarea sistemului de conversie text-voce pentru limba română dezvoltat în cadrul institutului; (5) realizarea unei analize a formelor de contur intonațional în corelație cu structura sintactică şi semantică a textelor asociate rostirilor şi funcțiile prozodiei**;** (6) implementarea unui modul software care să interpreteze fişiere XML cu structura proiectată pentru indicațiile microprozodice şi generarea conturului frecvenței F0 pe baza acestor indicații în sinteza vocală pe sintetizatorul dezvoltat în cadrul Institutului de Informatică Teoretică.

6.1.1 Implementarea de metode de estimare a frecvenței fundamentele F0

Cercetările privind modelarea melodiei semnalului vocal impun dezvoltarea şi implementarea de algoritmi pentru estimarea cât mai corectă a frecvenței fundamentale. Aparent o problemă uşoară, abordată foarte frecvent în literatura de specialitate prin diverse metode, estimarea frecvenței fundamentale în contextul dinamicii nestaționare a semnalului vocal, rămâne o problemă destul de complicată şi generatoare de noi abordări.

Pentru a face față acestei provocări a trebuit să analizez mai multe metode de estimare a frecvenței fundamentale în domeniul timp, domeniul frecvență şi în domeniul timp-frecvență. În urma analizei efectuate am constatat că fiecare metodă reuşeşte să estimeze corect frecvența F0 în anumite condiții de zgomot şi componente armonice ale semnalului vocal.

După trecerea etapei de analiză am reuşit să implementez două metode de estimare a frecvenței F0: una *în domeniul timp* bazată pe combinarea metodei de estimare folosind *funcția de autocorelație* cu o metodă bazată pe *funcția mediei diferenței amplitudinilor (AMDF)*; ce de a doua *în domeniul frecvență* bazată prin combinarea metodei de estimare folosind *funcția cepstrum* cu o metodă de estimare a armonicilor superioare ale frecvenței F0 din spectrul de frecvență al semnalului.

Prin folosirea celor două metode de estimare a frecvenței fundamentale pe aceleaşi semnale vocale, am constatat următoarele: pe segmentele de semnal vocal sonore pe care ambele metode oferă estimări corecte pentru frecvența F0, metoda de estimare a frecvenței în domeniul timp reuşeşte să ofere rezultate care se corelează mai bine cu periodicitatea prezentă la nivelul semnalului vocal în domeniul timp; metoda de estimare a frecvenței fundamentale în domeniul timp reuşeşte să estimeze valori corecte pentru frecvența F0 pe segmente de semnal sonor de intensitate redusă, pe care metoda de estimare în domeniul frecvență estimează rezultate eronate.

6.1.2 Modelarea co-articulării fonemelor cu funcții de dominanță neliniare şi îmbunătățirea tranzițiilor formanților între foneme

Următoarea problemă abordată în cadrul cercetărilor efectuate pe parcursul elaborării tezei a fost generată de necesitatea modelării tranzițiilor dintre foneme şi de analiză a posibilităților de implementare a elementelor microprozodice la sintetizatorul formantic

Klatt folosit în cadrul institutului (Capitolul 3).

Analiza în domeniul frecvență a undelor vocale naturale a pus în evidență influențe ale frecventelor centrale ale formanților între sunetele (fonemele) vecine. Aceste influențe se materializează prin modificarea, între anumite limite, a valorilor de stabilitate ale formanților şi prin tranziții între valorile de stabilitate la trecerea de la un fonem la altul. În literatura de specialitate, aceste efecte naturale care apar în timpul producției vocale poartă numele de co-articularea sunetelor. Din punct de vedere al fenomenului producției vocale aceste influențe se explică cu ajutorul efectelor de inerție care apar în mişcarea, fără eforturi deosebite din partea vorbitorului, unor organe implicate în procesul de vorbire: buzele, vălul paltin, limba, maxilare cu sistemul de masticație şi laringele.

După etapa de analiză a principalelor metode şi teorii existente pentru modelarea co-articulării sunetelor, am ajuns la concluzia că pentru cazul sintetizatorului formantic este de interes găsirea unei metode de modelare a efectului co-articularii din punct de vedere al percepției auditive dar care să țină cont de fenomenul producției vocale.

Pornind de la posibilitățile oferite de sintetizatorul formantic de tip Klatt şi de la analiza modelelor care abordează co-articularea fonemelor, în lucrarea (Apopei 2004a) am propus o modelare, cu funcții neliniare de dominanță, a variației formanților F2 şi F3 la sintetizatorul Klatt. Această modelare a fost inspirată din modelul Cohen şi Massaro (1993, 2003). Cu această modelare a variației formanților, la tranziția dintre foneme, am reuşit să îmbunătățesc calitatea semnalelor vocale sintetizate cu ajutorul sintetizatorul formantic de tip Klatt.

Analiza elementelor componente ale unui sintetizator (figura 3.1) şi a posibilităților de control a parametrilor la sintetizatorul formantic Klatt, ne-a condus la ideea de a realiza implementarea elementelor microprozodice cu ajutorul unor submodule, care să fie incluse în modulul fonetic din componența sistemului de conversie text-voce (fig. 5.1 şi fig. 5.2).

6.1.3 Proiectarea unei ierarhii de unități intonaționale pentru modelarea fonologică a intonației din limba română

Contribuțiile din capitolul 4 sunt legate în principal de necesitatea elaborării unui ierarhii intonaționale care să stea la baza modelului prozodic pentru limba română. Pentru a realiza acest deziderat am efectuat o analiză a principalelor modele intonaționale, cu aplicabilitate în sinteza şi recunoaşterea vocală şi care stau la baza realizării modelelor prozodice. În urma acestei analize am constatat că pentru a fi utilizate în sinteza vocală, modelele intonaționale trebuie corelate cu modele fonologice (care au la bază ierarhii intonaționale), cu structurile sintactico-semantice ale textelor şi funcțiile prozodiei (Batliner 2003, Kohler 2005, Teodorescu 2005, Shih 2006, Hirst 2007 ş.a).

În urma studiului principalelor modele intonaționale, pe baza analizei contururilor intonaționale din limba română (Apopei ş.a. 2005b, Apopei ş.a. 2006a, Turculeț& Apopei 2006) şi al încercărilor de a implementa aceste contururi în sinteza vocală (Apopei ş.a. 2005a) am ajuns să înțelegem legătura dintre modelele fonologice şi modelele prozodice (în particular modelele intonaționale). Astfel am reuşit să propun o ierarhie intonațională (Apopei ş.a. 2006b, 2006c) cu care să putem grupa evenimentele intonaționale şi să abordăm problematica sintezei prozodice în limba română.

Această ierarhie intonațională a stat la baza dezvoltării schemei de adnotare a evenimentelor microprozodice prezentată în secțiunea 5.2.1., a cercetărilor ulterioare privind înțelegerea intonației în limba română şi a implementării elementelor prozodice în sinteza vocală.

6.1.4 Proiectarea unei scheme XML pentru adnotarea microprozodică a textelor de la intrarea sistemelor de conversie text-voce pentru limba română

În cadrul grupului de cercetare de la Institutul de Informatică Teoretică, ideea introducerii elementelor de prozodie în sinteza vocală pentru limba română, prin reprezentare în format XML, a fost propusă de H.N. Teodorescu (2002) într-un grant CNCSIS. În acea etapă, a fost realizată o schemă de adnotare a evenimentelor macroprozodice cu două *taguri* (*break* şi *pitch*). *Tag*-ul „break", cu două valori (0 şi 2), indica prezența unor pauze pentru delimitarea, în sinteză, a grupurilor de cuvinte. *Tag*-ul „pitch", prin valorile („high"/ „low"), indica trendul conturului intonațional, şi implicit al accentelor lexicale, pe durata grupului de cuvinte.

În această etapă am propus o schemă de adnotare a elementelor prozodice în format XML care să poată fi folosită atât în analiza contururilor intonaționale (Apopei ş.a. 2006) cât şi în sinteza vocală prozodică. La proiectarea acestei scheme s-au avut în vedere unitățile intonaționale din ierarhia prezentată în secțiunea 4.4, creând câte un tag pentru marcarea unităților de pe fiecare nivel. Tag-urile utilizate împreună cu atributele lor au fost prezentate în lucrarea (Apopei ş.a. 2006).

Pentru folosirea acestei scheme în adnotarea corpusurilor de voce, am prevăzut atribute pentru etichetarea fonologică a evenimentelor şi pentru informații cantitative legate de nivelul tonurilor țintă asociate evenimentele etichetate. În acest scop s-a împărțit gama de variație a frecvenței F0 în cadrul rostirii de adnotat în semitonuri şi s-a realizat o scală de măsurare a tonurilor cu baza la nivelul tonului celui mai înalt din rostire.

6.1.5 Analiza formelor de contur intonațional în corelație cu structura sintactică şi semantică a textelor asociate rostirilor şi funcțiile prozodiei

Acest studiu a fost realizat pe două categorii de rostiri: rostiri ale unor propoziții afirmative, şi rostiri ale unor propoziții interogative totale. Analiza contururilor intonaționale a generat observații referitoare la secvențele de evenimente intonaționale de pe conturul unităților intonaționale de tip IP/ip, funcțiile prozodiei, precum şi contextele textuale (structurile silabice) şi semantice care influențează forma conturului F0 în cadrul unităților de accentuare.

Analiza contururilor intonaționale pentru propozițiile afirmative (Raport 2006a), a condus la obținerea unor reguli lingvistice pentru forma evenimentelor prozodice, pentru modul de grupare a cuvintelor în unități ale ierarhiei intonaționale şi pentru stabilirea unităților de accentuare proeminente (proeminențe în sensul teoriei autosegmental-metrice). Conturului melodic al propozițiilor afirmative poate fi realizat din una sau mai multe fraze intonaționale separate prin tonuri de graniță şi pauze.

Analiza conturului intonațional pentru propozițiile interogative totale a dus la identificarea pentru frazele intonaționale a un număr de cinci variante intonaționale. Pe baza acestor variante intonaționale, conturului melodic al propozițiilor interogative totale poate fi realizat din mai multe fraze intonaționale separate prin tonuri de graniță şi pauze. În general, acestor segmente de pe conturul frecvenței F0 le corespund la nivelul textului diferite grupuri sintactice. Fiecare frază intonațională are propria ei emfază interogativă dar ultima frază intonațională este cea le domină pe cele precedente (Apopei 2008).

6.1.6 Implementarea unui modul software pentru generarea în sinteza vocală a conturului frecvenței F0 pe baza indicațiilor microprozodice

Conturul frecvenței F0 corespunzător unei fraze intonaționale/intermediare poate fi văzut ca o succesiune de pattern-uri corespunzătoare unităților ritmice ce le compun. La

rândul lor, pattern-urile unităților ritmice rezultă din secvențierea pattern-urilor unităților de accentuare componente. Acestea din urmă sunt adecvate evenimentelor tonale ce le conțin şi poziției în cadrul unităților ritmice şi frazei intonaționale. Partea semnificativă a pattern-ului unei AU este cea care generează evenimentul sau secvența de evenimente tonale. Vom numi această parte drept pattern de eveniment/evenimente, cum ar fi de exemplu, accentul de pith de tip H* sau L*, accentul de tip stress, accentele de frază intermediară şi tonurile de graniță sau o combinație a acestora (Apopei 2007, Raport 2007b).

Modulul de generare a frecvenței F0, implementat în cadrul sistemului text-voce conține două etape de procesare (Apopei & Jitcă 2007): cel de modelare a conturului intonațional şi cel de generare propriu-zisă.

În prima etapă se translatează arborele rostirii, format din structura unităților intonaționale aşa cum este descrisă în fişierul XML de la intrarea modulului fonetic, într-o secvență de evenimente de contur F0 care au asociate câte un pattern al frecvenței F0 şi o poziție în spațiul (timp, frecvență). Sunt mapate mai întâi frazele intonaționale şi intermediare împreună cu tendințele lor de *downsteping* sau *upsteping*. În cadrul acestor limite sunt fixate pozițiile unităților ritmice (grupurilor de unități de accentuare) şi a unităților de accentuare.

În cea de a doua etapă are loc generarea conturului F0 prin selectarea pentru fiecare eveniment tonal, din secvența generată la pasul anterior, a unui pattern elementar de contur F0 (forme de contur de tipul celor prezentate în fig. 5.12). Această selecție are loc în acord cu: etichetele tonale din fişierul XML, poziția în cadrul unităților intonaționale şi contextul fonetic. Pattern-urile sunt poziționate şi apoi scalate astfel încât să se încadreze în regiunile delimitate prin gama de variație a frecvenței F0 şi durata fonemelor.

6.2 Dezvoltări şi direcţii de cercetare viitoare

Modelarea prozodiei propusă în această lucrare, pe baza unei ierarhii intonaţionale, a fost gândită din perspectiva realizării unei punţi de legătură între cercetările din domeniul lingvisticii computaţionale (Tufiş 2000,2007, Cristea 2003, 2005, Curteanu 2007, Forăscu 2006, 2008) şi cele din domeniul analizei şi sintezei vocale pentru limba română (Teodorescu H.N. 2005, 2008, Burileanu D. 2006, Grigoraş Fl. 1999, Jitcă 2002, 2003). În acest context vom continua analiza elementelor prozodice pe corpusuri paralele text-voce.

Cercetările efectuate pe descrierile conturilor intonaţionale cu ajutorul schemei XML propuse au pus în evidenţă necesitatea unei înţelegeri mai profunde a legăturilor dintre secvenţele de evenimente intonaţionale şi structurile sintactico-semantice, în corelaţie cu funcţiile comunicative ale prozodiei (Kohler 2005, Teodorescu 2005) şi realizarea de descrieri macroprozodice. Realizarea descrierilor macroprozodice ar permite completarea modelului prozodic cu elemente specifice rostirilor emoţionale. Un început al acestui demers a fost propus în (Raport 2008a), unde, conturului F0 la nivelul frazele intonaţionale/ intermediare este văzut ca o succesiune de pattern-uri corespunzătoare unităţilor ritmice (grupuri de unităţi de accentuare) ce le compun. Stabilirea unui set de funcţii comunicative (Kohler 2005, Teodorescu 2005) în corelaţie cu un set de forme de contur intonaţional la nivelul unităţilor ritmice ar facilita realizarea descrierilor macroprozodice. Trecerea de la descrierile macroprozodice la descrierile microprozodice urmând a fi realizată prin punerea în corespondenţă a formelor de contur de la nivelul unităţilor ritmice cu un set de descrieri microprozodice

Realizarea de descrieri macroprozodice ar deschide posibilitatea folosirii cercetărilor din această teză şi în domeniul sistemelor „*Spoken Human-Computer Dialogue"*(Popescu & Caelen & Burileanu 2007), cu posibile implicaţii în cadrul unor proiecte naţionale de ajutor a persoanelor cu handicap vizual.

Voi continua colaborarea cu grupul de cercetare de la *Seminarul de dialectologie şi sociolingvistică* al Facultăţii de Litere din cadrul Universităţii "Al. I. Cuza" din Iaşi, în probleme de extragere şi prelucrare statistică a unor parametri fonetici din rostiri dialectale în limba română. Colaborarea vizează, în acest moment, participarea alături de foneticieni, la programul european "*L'Atlas Multimedia Prosodique de l'Espace Roman*" ("AMPER"), pentru studiul graiurilor din Moldova şi Basarabia.

Bibliografie selectivă

1. Allen J., Hunicutt M.S., Klatt D., *From text to speech*, The MITalk System, Cambridge University Press, Cambridge, England, 1987
2. d'Alessandro C., Mertens P. (1995), *Automatic pitch contour stylization using a model of tonal perception,* Computer Speech and Language, 9 (3), pp.257-288.
3. d'Alessandro C., Castellengo M. (1994). *The pitch of short-duration vibrato tones*, Journal of the Acoustical Society of America 95, pp.1617-1630.
4. Ali S., Hirst D.J*., Analysis by Synthesis of English Intonation Patterns:Generalising from form to function*, International Congress of Phonetic Sciences, Saarbrücken, Germany, 6-10 August 2007
5. Apopei V., Jitcă D., (2008), *Intonational Variations for Romanian Yes-No Questions,* In Proceedings of the 5th European Conference on Intelligent Systems and Technologies (ECIT 2008), Iasi, July 10-12, 2008.
6. Apopei V., Jitcă D. (2007), *Module for Generating the F0 Contour Using as Input a Text Structured by Prosodic Information*, Advances in Spoken Language Technology (SpeD 2007), The Publishing House of the Romanian Academy, Eds. C. Burileanu, H.N. Teodorescu, pp.119-126.
7. Apopei V, Jitcă D, Turculeţ A. (2006a) *Intonational structures in Romanian Yes-No Questions,* Computer Science Journal of Moldavia Chişinău, vol 14, nr. 1(40), 2006, pp. 113-137
8. Apopei V., Jitcă D. (2006b), *A set of Intonational Category for Romanian Speech and Text*

Annotation, Proceedings ECIT 2006, Iaşi, september 20-23, 2006, Advances in Intelligent Systems and Technologies, ISBN 978-973-730-265-6, pp.117-124.
9. Apopei V, Jitcă D. (2006c), *Schemă XML de adnotare a intonației în cadrul corpusurilor de text*, Resurse lingvistice si instrumente pentru prelucrarea limbii române, Editura Universității ”Al.I. Cuza” Iaşi, România, pg. 9-14
10. Apopei V., Jitca D, Teodorescu H.N, (2005a) *Implementation of stress and emotion rendering rules in synthesized speech*, Trends in Speech Technology, Editura Academiei Romane, pp. 67-72, 2005
11. Apopei V., Jitca D, (2005b), *Romanian Intonational Annotation Based on Tone Sequence Model*, SASM 2005, Iasi, Romania, May 5-7, 2005
12. Apopei V., Jitcă D., Grigoraş F. (2004a), *Modeling Formantic Transitions in Klatt Speech Synthesizer*, Proceedings ECIT-2004 Conference, pp. 137-148, Iaşi, România.
13. Apopei V., Jitcă D., Grigoraş F., (2004b), *Folosirea trăsăturilor acustice în segmentarea semnalului vocal. Metode de segmentare*, SIA-2004, 24-25 Septembrie, 2004, Iasi, Zilele academice ieşene.
14. Apopei V., Jitcă D., Grigoraş F. (2003a), *Studiul trăsăturilor acustice necesare pentru evidențierea evenimentelor fonetice în vederea segmentării semnalului vocal*, Simpozionul Sistme de Inteligență Artificială SIA 2003, septembrie 2003, Iaşi, România.
15. Apopei V., Zbancioc M., (2003b), *Metode de silabificare a cuvintelor limbii române bazate pe reguli şi pe rețele neuronale - studiu comparativ*, Simpozionul Sistme de Inteligență Artificială SIA 2003, septembrie 2003, Iaşi, România.
16. Batliner A., Nöth E. (2003), *Prosody and Automatic Speech Recognition -Why not yet a Success Story and where to go from here*, Proceedings of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing, pages 357–364, Tokyo.
17. Burileanu D., Negrescu C., *Prosody Modeling for an Embedded TTS System Implementation*, Proceedings of the 14th European Signal Processing Conference EUSIPCO 2006, Florence, Italy, pp. 715-718, Sept. 4-8, 2006.
18. Burileanu D., Dervis A., *Modeling the Fundamental Frequency Contour for Text-to-Speech Synthesis in Romanian*, Proceedings of the International Conference Communications 2004, Bucharest, Vol. 1, pp. 189-192, 2004.
19. Burileanu D.,. Dan C, Sima M., Burileanu C., *A Parser-Based Text Preprocessor for Romanian Language TTS Synthesis*, Proceedings of the 6th European Conference on Speech Communication and Technology EUROSPEECH’99, Budapest, Hungary, Vol. 5, pp. 2063-2066, Sep. 1999.
20. Cristea D.: *The relationship between discourse structure and referentiality in Veins Theory*, in Wolfgang Mentzel and Cristina Vertan (eds.) Natural Language processing between Linguistic Inquiry and System Engineering, Editura Universității „Al.I.Cuza” Iaşi, iulie 2003, pag.9-22.
21. Curteanu N., Trandabăț D.M (2007), *Functional FX-bar Projections for Local and Global Text Structures. The Anatomy of Predication*, Revue Roumaine de Linguistique, Romanian Academy Editorial House, Bucharest, 2007, No. 1-2, pp. 161-194.
22. Dascălu-Jinga L., *Melodia vorbirii în limba română*, Editura Univers enciclopedic, Bucureşti, 2001.
23. Forăscu C.: *Temporal Information Processing*, In Proceedings of the 5th European Conference on Intelligent Systems and Technologies (ECIT 2008), Iasi, July 10-12, 2008.
24. Furui S., *Tokyo Institute of Technology, 50 years of progress in speech recognition technology -- Where we are, and where we should go*, International Conference on Acoustics, Speech, and Signal Processing (ICASSP), which will be held at the Hawai‘i Convention Center in Honolulu, April 15 - 20, 2007
25. Fujisaki H. (2004), *Prosody, Information, and Modeling—With emphasis on tonal features of speech*, Proc. Speech Prosody 2004 Nara, pp. 1–10.
26. Grigoras Fl. (2002), *Metode Soft-Computing în analiza şi sinteza vocală*, Ed. Artes, 2002, Iaşi.
27. Grigoras Fl., Apopei V., Jitcă D., Teodorescu H.N., (2000), *Conclusions from a Research on Soft-Computing Rule-Based Speech Synthesis for Romanian Language*, ECIT’2000-European Conference on Intelligent Technologies, Technical University “G. Asachi” Iasi, September 25-28, 2000, CD-ROM Proceedings, ISBN 973-95156-7-4.
28. Grigoras F., Teodorescu H.N., Jain L.C., Apopei V., *Fuzzy and Knowledge-based Control for Speech Synthesis*, ECC ’99, Karlsruhe, Germany, 1999

29. Grigoraş Fl., Teodorescu H.N., Apopei V., *Nonlinear Analysis and Synthesis of Speech*, Studies in Informatics and Control, Vol. 7, No. 1, March 1998, Romanian Academy Publishing House, pp. 57 – 72, 1998
30. Gussenhoven C. (2007), *Types of focus in English*, In Chungmin Gordon & Buring (eds.), Topic and Focus: Cross-linguistic Perspectives on Meaning and Intonation, Springer, pp. 83–100, 2007.
31. von Heusinger K. (1999), *Intonation and Information Structure,* Habil. thesis, University of Konstanz.
32. Hirst D.J. (2005), *Form and function in the representation of speech prosody*. Speech Communication, 46 (3-4), pp.334-347.
33. Hirst D., Cristo A.D., Espesser R. (2000), *Levels of representation and levels of analysis for intonation*, Prosody Theory and Experiment (Horne, M. , ed.), Dordrecht, The Netherlands: Kluwer.
34. Hirst D., Di Christo, A., Espesser R. (2000), *Levels of representation and levels of analysis for the description of intonation systems*, http:// aune.lpl.univ-aix.fr /~hirst/articles/2000 Hirst&al.pdf, (pp. 1-21).
35. Huang N. E.(2005), *Hilbert-Huang Transform and its Applications*, N. Huang et al. (Eds.), World Scientific Publishing.
36. *Jitcă D., Teodorescu H. N., Apopei V., Grigoraş Fl.(2003a),* An ANN-Based Method to Improve the Phonetic Transcription and Prosody Modules of a TtS System for the Romanian Language*, Proc. SPED2003, aprilie, Bucureşti, 2003, Speech Technology and Human Computer Dialogue, editor C. Burileanu, Editura Academiei Române, pp. 43-50.*
37. *Jitcă D., Apopei V., (2003b),* Conclusions on Analysis and Synthesis of Large Semivocalic Formantic Transitions*, Proceedings SCS 2003, Iaşi, pp. 177- 180, 2003.*
38. *Jitcă D., Teodorescu H.N., Apopei V., Grigoraş Fl., (2002a),* Improved Speech Synthesis Using Fuzzy Methods*, Speech Technology Journal, 5, pp. 227-235, Kluwer Academic Publishers, 2002.*
39. *Jitcă D., Apopei V., Grigoras Fl.,(2002b),* An ANN-Based Method to Improve the Phonetic Transcription Module of a TtS System for the Romanian Language*, în CD ROM Proc. ECIT2002 - European Conference on Intelligent Technologies, 2002.*
40. *Jitcă D., Apopei V., Grigoraş Fl., (2002c)* Text-to-Speech System for Romanian Language based on Formantic Synthesis*, in CD-ROM Proc. ECIT'2002 - European Conference on Intelligent Technologies.*
41. *Jitcă D., Apopei V., Grigoraş Fl., (2002d),* Elemente de prozodie a limbii române în analiza şi sinteza vocalã*, Zilele Academice Ieşene, Academia Romana, Filiala Iasi, Septembrie, 2002.*
42. *Jitcă D., Apopei V., (2002e),* Basic Romanian Language Prosody Analysis and Synthesis*, Buletinul Institutului Politehnic Iaşi*
43. *Jitcă D., Apopei V., Grigoraş Fl.,(2001),* Sistem TtS pentru limba română*, Zilele Academice Ieşene, Academia Romana, Filiala Iasi, Octombrie, 2001.*
44. Jitcă D., Teodorescu H. N., Apopei V., Grigoraş Fl., (2000a) *Naturalness in Speech Synthesis by Fuzzy Control of the Glottal Parameters*, IIZUKA'2000-Int. Conf. on Fuzzy Logic and NN, CD-ROM Proceedings, Iizuka, Japan, 2000.
45. Jitcă D., Apopei V., Teodorescu H. N., Grigoraş Fl., (2000b), *Soft computing based speech analysis and synthesis for the Romanian language*, Memoriile Secțiilor Ştiințifice ale Academiei Române, Vol. 23, pp.203-230
46. Kohler K. J. (2005), *Timing and communicative functions of pitch contours*, Phonetica 62 pp. 88-105.
47. Kohler K. J. (1997), *Modelling prosody in spontaneous speech*. In Y. Sagisaka, N. Cambell, N. Higuchi (eds.) Computing prosody. Computational models for processing spontaneous speech. N.Y.: Springer, pp. 187-210.
48. Potamianos A., Maragos P. (1997), *Speech analysis and synthesis using an AM-FM modulation model,* Proceedings of EUROSPEECH-1997, pp. 1355-1358.
49. Rabiner L.R, Schafer R.W. (1978), *Digital processing of speech signals, Pentice Hall International*, London.
50. Raport 2008a, Teodorescu H.N, Apopei V, Jitcă D, *Stabilirea unui set de parametri pentru descrierea prototipurilor formelor de contur ale unităților de accentuare*, Raport de cercetare, Institutul de Informatică Teoretică Iasi ,Academia Română, iunie 2008.
51. Raport 2007b, Teodorescu H.N, Apopei V, Jitcă D, *Modul de generare automată a frecvenței F0*

pentru implementarea intonației în sinteza vocală în limba română, Raport de cercetare, Institutul de Informatică Teoretică Iasi ,Academia Română, noiembrie 2007.
52. Raport 2006a, Teodorescu H.N, Apopei V, Jitcă D, *Analiza modului de corelare a unor evenimente intonaționale cu structura morfologică*, Raport de cercetare, Institutul de Informatică Teoretică Iasi, Academia Română, iunie 2006.
53. van Santen J.P.H. (2002), *Quantitative Modeling of Pitch Accent Alignment*, Speech Prosody Conference, Aix-en-Provence, France, 11-13 April 2002
54. van Santen J. P. H., Pols L. C. W., Abe M., Kahn D., Keller E., Vonwiller J. (1998), *Report on the Third ESCA TTS Workshop Evaluation Procedure*, Third ESCA Workshop on Speech Synthesis 98.
55. Schröder M. (2004), *Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis*, PhD thesis, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University, 2004
56. Schröder M., Trouvain J. (2003), *The German text-to-speech synthesis system MARY: A tool for research, development and teaching*, Intl.J. Speech Technol., vol. 6, pp. 365–377, http:// mary.dfki.de, 2003
57. Sun X (2002), *The Determination, Analysis, and Synthesis of Fundamental Frequency*, Phd thesis, NorthWestern University
58. Teodorescu H.N., *Aproposed theory in prosody generation and perception: th multi-dimnsional contextual integration principle of prosody,* SpeD 2005 - 3th Conference on Speech Technology and Human Dialogue, Eds. C. Burileanu, Trend in Speech Technology, Editura Academiei Române, 2005, ISBN 973-27-1178-7, pp. 109-118
59. Teodorescu H.N., Ceauşu A., Apopei V., *Îmbunătățirea aspectelor prozodice în sinteza text-to-speech pentru limba română*, Revista de Inventică, Nr.4, pp. 11-17, 2003
60. Teodorescu H.N (2002), Grant CNCSIS TIP A, *Îmbunătățirea aspectelor prozodice în sinteza Text-to-Speech pentru limba română*, Responsabil Grant: H. N. Teodorescu, m.c. Colectiv de realizare: Dan Cristea, Vasile Apopei, Alexandru Ceauşu, ş.a.
61. Teodorescu H. N., Grigoras Fl., Apopei V., *Nonlinear processes in speech production*, Int. J. Chaos Theory and Applications, vol. 2, no. 2, pp. 35-52, 1997
62. Tesnière L., *Éléments de Syntaxe structurale*, Paris, 1959, Klincksieck, 670 p.
63. Tran T.H., Ha Q.P., G. Dissanayake,(2004), *New Wavelet-Based Pitch Detection Method for Human-Robot Voice Interface*, Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems September 28 - October 2, 2004, Sendai, Japan
64. Tufiş D., Ion R.: *Parallel Corpora, Alignment Technologies and Further Prospects in Multilingual Resources and Technology Infrastructure,* Proc. of the 4th Conference on Speech Technology and Human Computer Dialogue "SpeD 2007", Iasy, Romania, May 10-12, pp. 183-194, 2007,
65. Tufiş Dan, *Using a large set of EAGLES-compliant Morpho-Syntatctic Descriptors as a Tagset for Probabilistic Tagging*, Proceedings of LREC 2000, Athens May, 2000, pp. 1105-1112
66. Turculeț A., Apopei V., Jitcă D. (2006), *Aspecte ale intonației propozițiilor interogative totale cu structura VO(adj)*, Anuar de lingvistică şi istorie literară 2004-2006, Editura Academiei Române, XLIV-XLIVI, pp. 85-105.
67. Xu Y. (2007), *Speech as articulatory encoding of communicative functions*, In Proceedings of The 16th International Congress of Phonetic Sciences, Saarbrucken, August, 2007, pp. 25-30
68. Xu Y. (2004b), *Transmitting tone and intonation simultaneously—the parallel encoding and target approximation (PENTA) Model*. In: Proceedings of International Symposium Symposium on Tonal Aspects of Languages: with Emphasis on Tone Languages, Beijing, pp. 215–220.