Revista Română de Interacţiune Om-Calculator 8(1) 2015, 45-64

O descriere generală a arhitecturii sistemelor automate de recunoaştere vocală

Valentina Sofroni, Alexandru Stan

Universitatea Babeş-Bolyai – Departamentul de Informatică Economică

Strada Teodor Mihali nr. 58-60, Cluj-Napoca 400591 România
E-mail: valentinasofroni@gmail.com , alexandru.stan@econ.ubbcluj.ro

Rezumat. Recunoaşterea vocală automată reprezintă un aspect important în sistemele inteligente ce tind să faciliteze interacţiunea omului cu calculatorul. Pentru a fi eficiente, aceste sisteme trebuie să permită procesări precise în timp real. În atingerea acestui deziderat, arhitectura sistemului de recunoaştere a vorbirii (ASR) joacă un rol esenţial. În acest articol oferim o perspectivă actualizată asupra arhitecturii unui sistem de recunoaştere vocală automată şi a componentelor constitutive. Articolul prezintă paradigmele de modelare dominante la ora actuală în acest tip de sisteme (modelele Markov invizibile, modelele de amestecuri Gaussiene, clasificatorii Bayes, modelul de limbaj n-gram, etc.) împreună cu constrângerile arhitecturale pe care fiecare dintre acestea le impun asupra design-ului sistemului informatic. Acest studiu reprezintă o etapă intermediară într-un proces mai amplu ce urmăreşte conceperea şi realizarea unui sistem ASR cu un înalt nivel de acurateţe, independent de locutor pentru recunoaşterea limbii române într-un domeniu restrâns de aplicabilitate, precum justiţia.

Cuvinte cheie: interacţiune om-calculator, recunoaştere vocală automată.

1. Introducere

Chiar dacă pe parcursul ultimilor ani tehnologia ASR a înregistrat progrese încurajatoare, problematica generală a recunoaşterii vocale computerizate pentru orice limbă sau vorbitor este departe de a fi soluţionată. Realizarea unui sistem performant de recunoaştere vocală automată este complexă (Duma, Giurgea, Ordean, & Zălhan, 2015) şi presupune implementarea unui ansamblu de subsisteme ce folosesc diverşi algoritmi statistici, de tratare a semnalelor şi de optimizare. Scopul acestui lucrări este acela de a oferi o perspectivă actualizată asupra arhitecturii unui astfel de sistem şi a componentelor sale constitutive. De asemenea, ea trece în revistă paradigmele de modelare dominante la ora actuală în acest tip de sisteme:

modele Markov invizibile, modelele de amestecuri Gaussiene, clasificatorii Bayes, modelul de limbaj n-gram etc.

Articolul este organizat după cum urmează: secțiunea 2 oferă o descriere generală a arhitecturii unui sistem ASR, secțiunile 3 şi 4 descriu modelele acustice şi de analiză acustică, secțiunea 5 prezintă modelul de limbaj, iar secțiunea 6 procesul de decodare/căutare propriu-zisă.

2. Descriere generală a arhitecturii ASR

Sistemul de recunoaşterea automată a vorbirii (*Automatic Speech Recognition* sau ASR) converteşte secvențe de semnale acustice în format text. Oamenii convertesc cuvintele în semnal audio utilizând mecanismul lor de producere a vorbirii. Un sistem ASR încearcă să deducă acele cuvinte originale reprezentate prin intermediul unor unde de vorbire. Arhitectura generală a sistemului este reprezentată în figura 1, aceasta cuprinde cele trei interfețe fundamentale (input, soluție şi output) cu subcomponentele acestora. (Gruhn, Minker, & Nakamura, 2011)

Interfața input reprezintă sursa semnalului, adică grupul țintă de oameni cu caracteristici vocale diferite, a căror voce este captată prin intermediul unor dispozitive, astfel încât aceste sunete să poată fi stocate în format electronic. Acestea din urmă prezintă baza prelucrărilor viitoare în sistem.

Soluția sau modelul de recunoaştere a vorbirii descrie totalitatea proceselor aplicate inputului astfel încât convertirea din semnal acustic în format text să se facă cu rate de erori minimale. Soluția cuprinde mai multe modele, şi anume: modelul de analiză acustică, modelul acustic, modelul de limbaj, modelul căutare.

Modelul de analiză acustică are rolul de a parametriza inputul într-o secvență de vectori acustici. În această etapă se extrag mai multe caracteristici acustice din semnalul de vorbire. Metodologiile utilizate în cadrul acestei etape pentru determinarea parametrilor acustici sunt: *linear predictive coding* (LPC), *mel-frequency cepstral coefficients* (MFCC) şi *perceptual linear prediction* (PLP) (Deng & Li, 2013). Analiza acustică cuprinde două componente (Deng & Li, 2013):

- achiziția datelor *(data acquisition)*: această componentă cuprinde elemente hardware utilizate pentru generarea datelor digitale.
- extragerea caracteristicilor *(feature extraction)*: componentă responsabilă de procesarea semnalului.

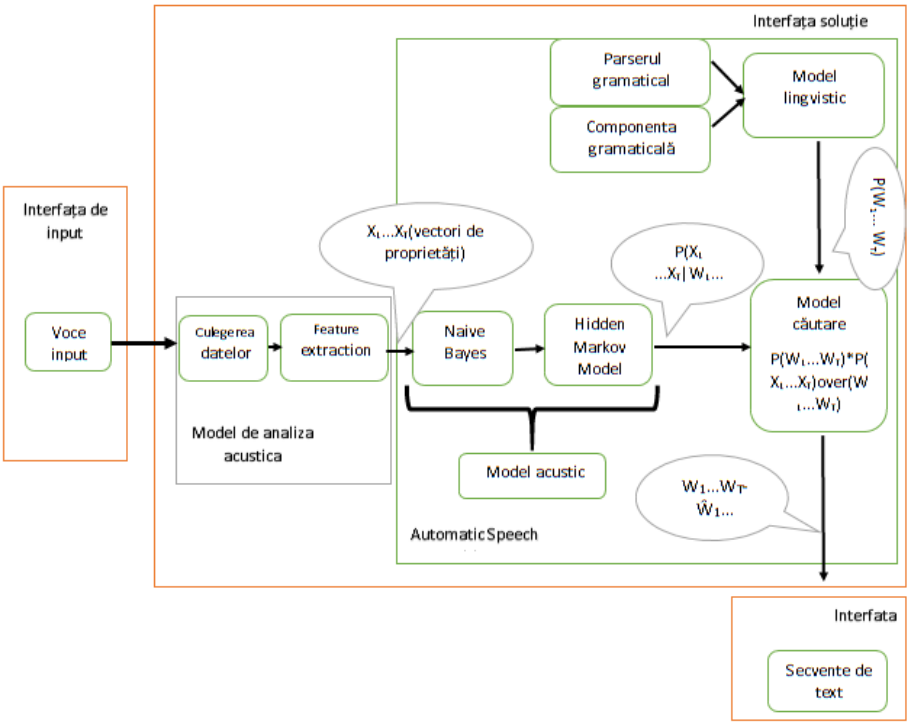

Figura 1. Arhitectura generală a sistemului ASR.

Procesarea semnalului, la rândul său, se subdivide în alte etape sau procese descrise detaliat în capitolele ce urmează.

În scopul determinării probabilităților statistice se utilizează modelul HMM (*Hidden Markov Models*), iar pentru gruparea caracteristicilor în clase de caracteristici se aplică clasificatori de tipul Naive Bayes.

Modelul de limbaj specifică proprietățile lingvistice (sintaxă, semantică şi pragmatică) şi furnizează probabilitatea a-priori a unei secvențe de cuvinte. În cadrul acestei etape, în scopul analizei lingvistice se utilizează modelul lingvistic n-gram împreună cu HMM.

Modelul de căutare se bazează pe regulile de decizie ale lui Bayes. Acest model utilizează programarea dinamică pentru a determina construcții gramaticale posibile, care corespund cel mai bine rezultatului oferit de modelul acustic.

Un algoritm general de funcționare a sistemului ar fi următorul (Deng & Li, 2013):

1. Achiziția datelor (*Data acquisition*) – se generează semnalul.

2. Extragerea caracteristicilor (*Feature extraction*) este realizată într-o perioadă foarte scurtă de timp. Semnalul obținut este divizat în cadre de lungime fixă. Din fiecare cadru sunt extrase mai multe domenii de frecvență şi parametri care formează vectorii de trăsături.

3. **Modelul acustic** urmăreşte să măsoare similaritatea dintre un semnalul input şi un model obținut deja în urma unei sesiuni de antrenament (training). Se va obține un alt model sau referință, care se potriveşte cu semnalul de input, astfel încât eroarea obținută din diferența dintre input şi output să fie minimală. Aici se utilizează modele Markov cu stări invizibile pentru determinarea trăsăturilor statistice ale unor clase de cuvinte.

4. **Modelul de limbaj**. În procesul de conversie a sunetelor în secvențe de cuvinte, modelul de limbaj oferă contextul necesar distincției între cuvintele şi frazele cu pronunții asemănătoare. În acest caz se utilizează modelul lingvistic n-gram.

5. **Căutarea propriu-zisă**. Modelele sunt utilizate pentru a calcula probabilitatea oricărei secvențe de cuvinte corespunzătoare unui semnal acustic observat. Decodarea vizează găsirea secvenței de cuvinte W care maximizează expresia matematică $P(W)\sum_H P(H \mid W)P(X \mid H)$, unde $P(W)$ este modelul de limbaj, $P(H \mid W)$ modelul de pronunție, iar $P(X \mid H)$ modelul acustic. În concordanță cu modelele specificate se va produce secvența de text ce maximizează această probabilitate. Procesul de decodare trebuie să fie precis şi rapid.

3. Procesarea semnalului

Prima etapă în procesul de recunoaştere automată a vorbirii este analiza acustică. Ea are ca scop reducerea erorilor, înlăturarea zgomotului, îmbunătățirea inteligibilității atunci când semnalul este degradat, şi extragerea caracteristicilor acustice utile din undele vocale. În ieşire, analiza front-end produce un set compact şi eficient de parametrii ce înglobează proprietățile observate ale semnalului vocal de intrare. Aceşti parametrii sunt utilizați ulterior în modelul acustic. (Huang, Acero, & Hon, 2001) (Meng, 2004) Principalele discipline care contribuie la tehnologiile de

analiză acustică sunt fonetica, fonologia şi metodele generale de prelucrare a semnalelor. În această etapă pot fi utilizați codificatori şi analizatori vocali. Codificatorii reduc debitul semnalului vocal fără reduceri semnificative ale informației transmise. Analizatorii transformă semnalul vocal într-o serie de caracteristici ce servesc mai apoi la cuantificarea nivelului de inteligibilitate şi de zgomot.

3.1. Achiziția semnalului

Cu ajutorul calculatoarelor din ziua de astăzi putem să gestionam achiziția de semnal (*speech signal acquisition tasks*) prin software. Spre exemplu, majoritatea plăcilor de sunet din componența calculatoarelor au acces direct la memorie şi discursul poate fi digitalizat în memorie fără încărcarea procesorului. (Huang, Acero, & Hon, 2001)

În scopul creşterii performanței în cadrul sistemului de recunoaştere a vorbirii, un număr de componente- cum ar fi sunetul digital, extragerea şi transformarea caracteristicilor, sincronizarea acustică, şi modelul lingvistic bazat pe căutare sunt ilustrate în figura arhitecturii sistemului JustASR. Mai multe sisteme de operare oferă mecanisme pentru organizarea secvențelor într-un mediu multitasking. Memoria tampon trebuie alocată în aşa mod încât să permită procesarea sincronă a fiecărei componente. Sistemele de calcul lente au nevoie de memorie tampon de dimensiuni mari din cauza potențialelor încetiniri în procesarea pe componente. (Huang, Acero, & Hon, 2001)

Un sistem de recunoaştere a vorbirii are un necesar de memorie tampon cuprins între 4 şi 64 kB cu o rată de eşantionare a vorbirii de 16-kHz şi cu o precizie A/D de 16-bit. O rată de eşantionare de 16-kHz este suficientă însă pentru lățimea de bandă a semnalului specific vorbirii. Lățimile de bandă înguste, cum ar fi canalele de telefon, cresc rata erorii în recunoaşterea vorbirii. Pentru rate de eşantionare de 8 kHz mărite până la 11 kHz rata de eroare este redusă la 10%, la fel şi pentru o rată de eşantionare de 16Hz (Huang, Acero, & Hon, 2001).

3.2. Extragerea caracteristicilor

Extragerea caracteristicilor constituie faza esențială a procesului de recunoaştere a vorbirii. Ea presupune extragerea informației relevante din semnalului vocal (*speech frame*), rezultatul fiind reprezentat sub forma unor

parametrii sau vectori de caracteristici. Extracția pornește de la semnalul audio inițial şi produce valori derivate esențiale şi neredundante, destinate învățării şi generalizărilor ulterioare. Astfel, extracția caracteristicilor este o metodă eficace de reducere a dimensionalității. Parametrii comuni utilizați în procesul de recunoaştere a vorbirii sunt *Linear Predictive Coding* (LPC), şi *Mel Frequency Ceptral Coefficients* (MPCC) (Meseguer, 2009).

Datorită importanței etapelor prezentate în figura 2 în procesul de extragere a caracteristicilor, acestea vor fi detaliate în subsecțiunile ce urmează.

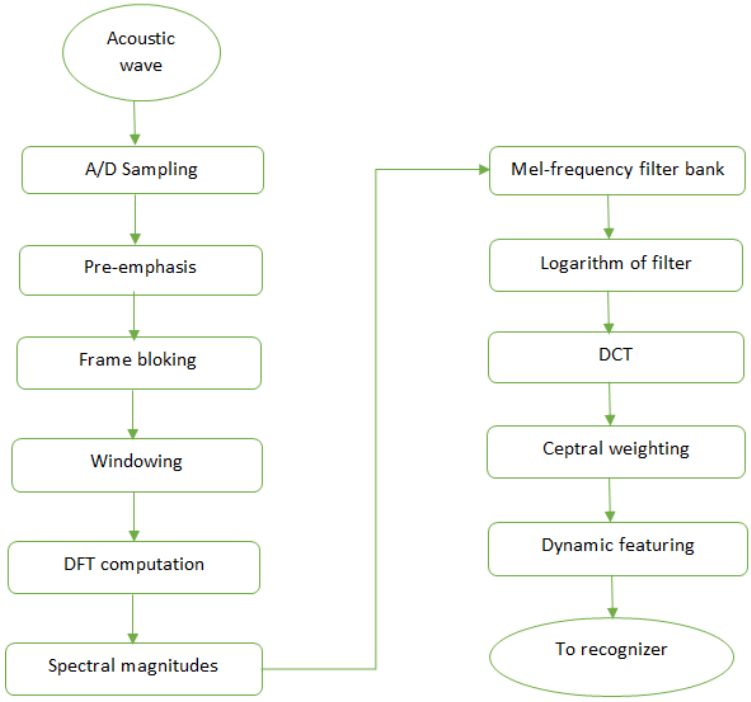

Figura 2. Etapele procesului de extragere a caracteristicilor (Meseguer, 2009).

3.2.1. Pre-procesare

Înainte de analiza spectrală sau analiza Fourier este utilizat un filtru pre-procesat, care îngustează spectrul semnalului. Acesta trebuie să compenseze frecvențele înalte ale semnalului, care au fost diminuate în timpul producerii sunetului de către mecanismul uman. Cele mai utilizate filtre sunt cele de tip

FIR (*Finite Filter Response*). Un astfel de filtru este descris în ecuația următoare: (Meseguer, 2009)

$$H(z) = 1 - \alpha * z^{-1}, 0.9 \leq \alpha \leq 1.0$$

3.2.2. Fixarea cadrelor şi windowing

În scopul efectuării unei analize eficiente a semnalului, acesta este divizat într-o secvență de cadre, unde fiecare cadru poate fi analizat independent şi reprezentat de un singur vector de proprietăți. În această etapă sunt utilizate ferestre de 20-25 ms aplicate unor intervale de 10 ms, astfel încât cadrele să devină fixe, cum este prezentat în figura 3.

Operația de *Windowing* se realizează prin compunerea semnalului cu o funcție care ia valoarea zero în afara intervalului de timp dorit. Acest tip de filtru induce discontinuități la capetele intervalului. De asemenea, pentru a reduce discontinuitatea unui semnal la extremitățile unui cadru sunt utilizate ferestrele variabile. Cel mai folosit tip de fereastră este fereastra Hamming, definită prin ecuația următoare: (Meseguer, 2009)

$$w(n) = 0.54 - 0.46\cos(\frac{2\pi(n-1)}{N-1})$$

Acest tip de filtru are avantajul de a avea doar trei coeficienți ai transformatei Fourier discrete diferiți de zero, şi induce o reprezentare puțin densă în domeniul frecvențelor.

Mai apoi, analiza spectrală se face prin aplicarea transformatei Fourier. Transformata Fourier descompune semnalul acustic în frecvențele care îl compun, sunetele vor fi astfel exprimate ca amplitudine a (sau intensitate sonoră) din sub-unitățile constitutive.

Extracție caracteristicilor are ca obiectiv principal reducerea dimensionalități semnalelor acustice prin eliminarea elementelor redundante. Acest proces va facilita strategiile de generalizare şi învățare ulterioare.

4. Modelul acustic

Modelul acustic este creat pornind de la un ansamblu de înregistrări audio şi transcrierile corespunzătoare acestora. După etapa de extragere a caracteristicilor se obține o secvență de vectori de caracteristici, X. Trebuie mai apoi estimate probabilitățile acustice a-posteriori ale acestor

caracteristici, P(X|W), dat fiind modelul fonetic sau modelul pentru subunitățile limbajului (cuvinte), W, astfel încât să poată fi recunoscute semnalele corespunzătoare cuvintelor (subunităților) corecte.(Huang, Acero, & Hon, 2001)

Un model acustic conține un set de reprezentări statistice pentru fiecare fonem în parte. El este creat prin aplicarea unor algoritmi de învățare pe o bază de date (numită speech corpus). Pentru reprezentarea statistică a fonemelor sunt utilizate modelele Markov cu stări invizibile (HMM) deoarece acestea pot fi uşor instruite, şi sunt foarte eficace din punct de vedere computațional. De asemenea, modelele sunt bine adaptate semnalelor vocale, caracterizate de stationaritate pe scurte intervale de timp. Fiecare fonem are propriul HHM. De aceea, un HMM pentru o secvență de foneme sau cuvinte se obține prin concatenarea modelelor individuale ale fonemelor şi cuvintelor separate. În scopul clasificării caracteristicilor extrase în urma procesului de analiză a semnalului se utilizează clasificatorii Bayes naivi, a căror rol în cadrul sistemului de recunoaştere vocală va fi descris în secțiunea următoare.

4.1. Clasificatorii Bayes

Sistemul de recunoaştere a vorbirii prezintă anumite probleme legate de clasificarea multi-claselor de caracteristici, iar pentru rezolvarea acestora se aplică clasificatorii liniari Bayes naivi (Sonia, David, & Poulose, 2013) (Goel, Kumar, & Byrne, 2004). Pentru utilizarea clasificatorului Bayes naiv este necesar un set relativ mic de date. Clasificatorul Bayes naiv are la bază teoria lui Bayes, care descrie o tehnică de clasificare eficientă pornind de la probabilitățile unor caracteristici. Pentru fiecare instanță reprezentată de un vector de caracteristici $X = (x_1, x_2, \dots, x_n)$ se calculează o probabilitate $p(C_i | x_1, x_2, \dots, x_n)$ de apartenență la fiecare din cele I clase posibile. Folosind teorema lui Bayes, probabilitățile condiționale pot fi rescrise ca $p(C_i \mid x_1, x_2, \dots, x_n) = p(C_i) * p(x_1, x_2, \dots, x_n \mid C_i) / p(x_1, x_2, \dots, x_n)$. Valorile caracteristicilor unei instanțe sunt condițional independente de valorile celorlalte caracteristici dată fiind clasa de apartenență C_i, principiu numit independența condițională a claselor $(p(x_k|C_i, x_j) = p(x_k|C_i))$. Independența condițională a claselor determină următoarea egalitate $p(C_i | x_1, x_2, \dots, x_n) = p(C_i) * \prod_{k=1}^{n} p(x_k|C_i)$. Clasificatorul Bayes naiv

asociază fiecărei instanțe ipoteza cea mai probabilă ce maximizează probabilitatea a-posteriori de apartenență:

$$\underset{i \in \{1, \dots, I\}}{argmax}\, p(C_i) * \prod_{k=1}^{n} p(x_k|C_i)$$

Astfel, vectorul de caracteristici X va fi asociat clasei C_i cu cea mai mare probabilitate.

4.2. Modelul Markov cu stări invizibile aplicat în recunoaşterea vorbirii

După eşantionarea semnalului vocal, acesta poate fi aproximat pe intervale scurte de timp cu un proces stohastic staționar. Astfel, semnalul vocal poate fi reprezentat sub forma unui HMM în vederea analizelor ulterioare. Un model Markov cu stări invizibile este de fapt un lanț Markov a cărui ieşiri reprezintă o variabilă aleatoare X generată de o funcție probabilistică de ieşire asociată fiecărei stări.

Un astfel de HMM produce (Adami, 2010) la fiecare câteva milisecunde secvențe succesive de vectori n-dimensionali numiți şi coeficienți Cepstral. Coeficienții sunt obținuți prin aplicarea transformatelor Fourier şi cosinus. Fiecare stare a modelului are o distribuție statistică de tipul unui amestec de distribuții normale corelate care va indica probabilitatea de apariție a fiecărui vector în parte. Astfel, fiecare fonem sau cuvânt va avea o distribuție de emisie diferită.

Pentru a aplica eficient modelele Markov cu stări invizibile în rezolvarea problemelor de recunoaştere a vorbirii, acestea trebuie mai întâi antrenate pentru a detecta ansamblul unităților de vorbire selecționate. Pentru estimarea parametrilor HMM cele mai des utilizate tehnici de învățare sunt cele, aşa numite, discriminative , ce încearcă să optimizeze criterii de clasificare a datelor de antrenament. Printre aceste criterii regăsim: criteriul informației mutuale maxime (MMI, *Maximum Mutual Information*), *Minimum Clasification Error* (MCE), *Minimum Word Error* (MWE), *Minimum Phone Error* (MPE).

4.2.1. Învățarea discriminativă

Probabilitatea maximă standard (ML, *Standard Maximum Likelihood*) maximizează probabilitatea de a defini corect modelul HMM λ în felul următor:

$$\lambda_{ML} = \underset{\lambda}{argmax}\, P(O|\lambda)$$

În cazul sistemului de recunoaştere a vorbirii, fiecare clasă acustică c dintr-un set de clase C este reprezentată de un model HMM, cu un parametru $\lambda_c, c = 1,2,\dots,C$. Criteriul ML pentru estimarea modelului λ_c utilizând secvența de observații O^c pentru clasa c poate fi definit în felul următor:

$$(\lambda_c)_{ML} = \underset{\lambda}{argmax}\, P(O^c|\lambda)$$

Criteriul ML nu garantează faptul că metodele estimate reprezintă soluțiile optimale pentru minimizarea probabilității erorilor de recunoaştere, întrucât fiecare model este estimat independent. De asemenea, criteriul ML nu ia în calcul posibilitatea de discriminare a fiecărui model (abilitatea de a distinge observațiile generate de un model corect de cele generate de alte modele). Un criteriu alternativ care maximizează această discriminare este criteriul informației mutuale maxime.

Implementarea criteriului MMI este bazată pe o variantă a algoritmului Braum-Welch (Pylkkonen & Kurimo, 2012) care maximizează valoarea expresiei de mai sus.

4.2.2. Selectarea unităților de vorbire

Un aspect esențial în contextul modelului acustic este selectarea unităților de vorbire, care prezintă informația acustică şi lingvistică pentru limbaj. Unitățile de vorbire trebuie să derive din cuvintele vocabularului şi să poată fi antrenate (adică să existe volumul de date necesar pentru estimarea modelului).

Unitățile de vorbire pot lua valori pornind de la sub-foneme şi mergând până la cuvinte, acestea din urmă fiind utilizate în cadrul recunoaşterii digitale. Un avantaj în cazul selectării cuvintelor este faptul că acestea înglobează şi coarticulația fonetică. O astfel de strategie îşi pierde însă viabilitatea practică în cazul vocabularelor de mare volum. De aceea, sunt frecvent selectate ca unități de vorbire fonemele şi sub-fonemele. Totuşi, selectarea fonemelor depinde de restul fonemelor (coarticulație fonetică).

Dependența între foneme poate fi redusă prin modelarea contextului unde acestea sunt generate. Această abordare, numită modelare fonetică dependentă de context (*context-depedent phonetic modeling*) este aplicată în cazul sistemelor de recunoaştere a vorbirii. Cel mai utilizat tip de model

dependent contextual este modelul Markov cu stări invizibile trifoneme. Contextul fonemului central este format din cele două foneme aflate la stânga şi la dreapta sa. Modelele fonetice dependente de context prezintă şi inconveniente. Ele induc o explozie combinatorică a numărului teoretic de n-foneme. De exemplu, in cazul utilizării a 24 de foneme, numărul teoretic al trifonemelor este egal cu $24^3 = 13824$ iar numărul pentafonemelor cu $24^5 = 7962624$. Astfel, apar probleme de antrenare insuficientă a HMM-urilor asociate acestora.

5. Determinarea modelului lingvistic şi căutare

Modelul de limbaj specifică proprietățile lingvistice (sintaxă, semantică şi pragmatică) şi furnizează probabilitatea a-priori a unei secvențe de cuvinte $w_1, w_2, \ldots, w_n$ cu ajutorul unei distribuții de probabilitate $P(w_1, w_2, \ldots, w_n)$. În procesul de conversie a sunetelor în secvențe de cuvinte, modelul de limbaj oferă contextul necesar distincției între cuvintele şi frazele cu pronunții asemănătoare. Acest model cuprinde două componente principale: componenta gramaticală si cea de parsare.

5.1. Componenta gramaticală

Componenta gramaticală este folosită în principal pentru a determina constrângeri de structură gramaticală părților unei propoziții sau fraze, în contextul dependențelor şi raporturilor dintre acestea.

5.1.1. Teoria Limbajelor Formale

În stabilirea regulilor sintactice a unei limbi, este important să se țină cont de caracterul general, selectiv şi comprehensibil a gramaticii respective. Generalitatea şi selectivitatea determină setul de propoziții care respectă sau nu regulile gramatice definite. În cazul sistemelor de înțelegere a limbii vorbite trebuie să avem o componentă gramaticală care să reglementeze structura celor mai uzuale propoziții. De asemenea, sistemul trebuie să facă diferența între anumite tipuri de propoziții ce sunt utilizate în situații distincte. (Huang, Acero, & Hon, 2001)

Cel mai uzual mod de reprezentare a structurii propozițiilor este utilizarea arborilor. Spre exemplu, în cazul propoziției "*Mary loves that person*", arborele gramatic arată în felul următor:

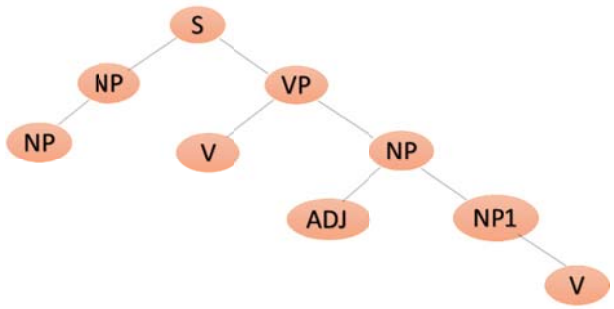

Figura 3. Reprezentarea sub formă de arbori şi componentele gramaticale corespunzătoare.

Nodul S este o abstracţiune a întregii propoziţii. S este părintele nodurilor NP şi VP, ce corespund părţilor de vorbire substantiv şi verb. La rândul său, nodul VP este părintele nodurilor V şi N şi aşa mai departe. Fiecare nod terminal este asociat cu câte un cuvânt din propoziţia supusă analizei. Pentru a genera arborele unei propoziţii trebuie cunoscută structura limbii şi determinat setul de reguli de retranscriere ce transforma nodul rădăcină S în ansamblul corespunzător de noduri terminale. Din figura de mai sus se poate observa că orice simbol poate fi extins într-un set de alte simboluri. Astfel, structura gramaticală ne ajută să determinăm raporturile dintre cuvintele unei propoziţii.(Huang, Acero, & Hon, 2001)

5.1.2. Ierarhia Chomsky

În teoria lingvistică formală descrisă de Chomsky, un sistem gramatical este definit ca un ansamblu G=(V,T,P,S), unde V şi T reprezintă seturi finite de simboluri neterminale şi, respectiv, terminale. În exemplul prezentat anterior, S, NP, NP1, VP, NAME, ADJ, N şi V sunt simboluri neterminale. Setul T de simboluri terminale cuprinde , care în cele mai multe cazuri sunt notate cu litere mici. P este un set finit de reguli de retranscriere, iar S este un simbol neterminal special, numit simbol iniţial. (Huang, Acero, & Hon, 2001). Ierarhia lui Chomsky indică corespondenta între diferite tipuri de automate şi gramaticile pe care acestea le acceptă. Astfel, o maşină Turing poate accepta orice structura gramaticală a unei propoziţii, un automat linear acceptă o gramatică dependentă de context, un automat *push-down* o gramatică independentă de context, iar un automat cu stări finite o gramatică regulară.

Limbajul analizat este, în cele mai multe cazuri, o secvență de simboluri terminale, cum ar fi "*Mary loves that person*". Acesta este produs pornind de la simbolul inițial prin aplicarea iterativă a unor reguli de retranscriere asupra secvențelor ce conțin simboluri neterminale. Regulile de retranscriere sunt de forma $\alpha \rightarrow \beta$, unde α şi β sunt secvențe de simboluri terminale sau neterminale de lungime arbitrară, iar α trebuie să fie diferit de nul. În teoria limbajelor formale, cele patru tipuri de limbaje majore şi sistemele gramaticale asociate sunt structurate ierarhic. Aşa cum am indicat mai sus, există patru tipuri de automate, care pot accepta limbajele generate de aceste sisteme gramaticale. Aceste automate nu sunt numai nişte instrumente matematice utilizate pentru recunoaşterea sistemelor gramaticale, ele fiind instrumente informatice fundamentale. În aceste categorii de automate regăsim HMM-urile şi modelul lingvistic n-gram, acesta din urmă fiind descris în paragrafele următoare.

În cadrul acestei etape, pentru realizarea analizei lingvistice se utilizează modelul n-gram deoarece este bine adaptat recunoaşterii vocale.

5.2. Parserul gramatical

Parsarea reprezintă un caz special în problemele de căutare, din cadrul sistemelor de recunoaştere vocală. Un algoritm de parsare caută setul de reguli gramaticale utilizate în scopul generării arborelui ce prezintă structura propoziției. Procedura de căutare poate fi de tip *top-down* sau *bottom-up*. Procedura *top-down* începe de la rădăcina arborelui, adică cu simbolul inițial S şi încearcă să genereze secvența dorită de simboluri terminale. Procedura *bottom-up* începe de la cuvintele propoziției input şi încearcă să identifice o secvență care corespunde unui set de simboluri neterminale intermediare. Procedura *bottom-up* poate fi repetată atât timp cât rădăcina arborelui nu este explorată. Metoda *data-directed search* este cea mai utilizată în sistemele SLU. (Huang, Acero, & Hon, 2001)

5.3. Modelul lingvistic n-gram

Un model n-gram este un tip probabilistic de model de limbaj folosit pentru estimarea elementului următor într-o secvență de lungime $n-1$ (Huang, Acero, & Hon, 2001). Modelul lingvistic poate fi formulat ca o distribuție

de probabilitate $P(w_1, w_2, ..., w_n)$ a apariției secvenței $w_1, w_2, \dots, w_n$ într-o propoziție.

$P(w_1, w_2, \dots, w_n)$ poate fi descompus în felul următor:

$$P(w_1, w_2, \dots, w_n) = \prod_{i=1}^{n} P(w_i | w_1, w_2, \dots, w_{i-1})$$

unde $P(w_i | w_1, w_2, \dots, w_{i-1})$ este probabilitatea de apariție a unității w_i, cunoscându-se secvența de cuvinte $w_1, w_2, \dots, w_{i-1}$. Aşadar, în acest caz, apariția unității w_i depinde de toată secvența precedentelor $i-1$ unități. În practică aceste probabilități sunt greu sau chiar imposibil de estimat chiar şi pentru valori mici ale lui i. O soluție practică pentru acestă problemă ar fi să se presupună că elementul curent depinde doar de un număr redus de $n-1$ elemente anterioare:

$$P(w_1, w_2, \dots, w_n) = \prod_{i=1}^{n} P(w_i | w_1, w_2, \dots, w_{i-1}) \approx \prod_{i=1}^{n} P(w_i | w_{i-(n-1)}, w_2, \dots, w_{i-1}) .$$

De aceea modelul n-gram este un model Markov de ordinul $n-1$. Dacă unitatea w_i depinde numai de ultimele două elemente, atunci avem un model trigram: $P(w_i | w_{i-1}, w_{i-2})$. La fel, putem avea modele lingvistice unigrame $P(w_i)$, sau bigrame $P(w_i | w_{i-1})$. În majoritatea cazurilor cuvintele depind doar de precedentele două cuvinte, astfel modelul cel mai utilizat este modelul trigram. Modelul trigram poate fi estimat calculând raportul dintre numărul perechilor $C(w_{i-2}, w_{i-1})$ şi numărul tripletelor $C(w_{i-2}, w_{i-1}, w_i)$.

Textul necesar pentru crearea unui model n-gram este numit corpus lingvistic. Pentru un model n-gram mărimea setului de antrenament este de ordinul milioanelor de cuvinte. Estimarea este bazată pe principiul verosimilității maxime.

5.4. Algoritmi de căutare

Continuous speech recognition (CSR) presupune atât o problemă de recunoaştere cât şi una de căutare pornind de la modelele acustice şi lingvistice pre-construite prin procese de învățare. În acest cadru, decodarea implică alegerea unei soluții optimale în contextul oferit de aceste modele.

Procesul de decodare (Huang, Acero, & Hon, 2001) vizează găsirea acelei secvențe de cuvinte al cărui model lingvistic şi acustic corespunde cel mai bine unui semnalul de input. Astfel, procesul de decodare utilizând modele lingvistice şi acustice învățate se mai numeşte şi proces de căutare

(*search process*). La baza modelului stau algoritmii de căutare pe grafuri a căror complexitate depinde de mărimea spațiului de căutare, determinată de constrângerile impuse de modelul de limbaj.

Modelul de căutare presupune explorarea spațiului de căutare, examinarea stărilor acestuia şi luarea unor decizii optimale. Problematica căutării este expusă de paradigma *state-space search*. Aceasta este definită de tripletul (S, O, G), unde S este setul de stări inițiale, O este setul de operatori (sau reguli de tranziție între stări), iar G este setul de stări care se doresc a fi obținute. O soluție constă în determinarea unui drum de la o stare inițială la una finală. Algoritmii utilizați în cadrul modelului căutare sunt: căutarea în adâncime (DFS), căutarea în lățime (BFS)**,** căutările *best-first* şi *beam-search*.

6. Măsurarea performantei unui sistem de recunoaştere a vorbirii

Performanța unui sistem de recunoaştere a vorbirii este caracterizată de trei factori esențiali: precizia de recunoaştere, complexitatea şi robustețea.

6.1. Precizia de recunoaştere

Precizia de recunoaştere (*Recognition accuracy*) este un criteriu simplu şi foarte important în măsurarea performanței unui sistem de recunoaştere a vorbirii. Inițial datele colectate sunt împărțite în două seturi : un set pentru învățare şi unul pentru testare (Meng, 2004). Setul pentru învățare, care conține cea mai mare parte a datelor esențiale, este utilizat pentru estimarea parametrilor din cadrul modelului acustic. Setul pentru testare este utilizat, cel mai des, pentru măsurarea performanței sistemului. Precizia de recunoaştere pentru ambele seturi de date este calculată aplicându-se rata de eroare a cuvântului. Totuşi, calcularea preciziei de recunoaştere pentru seturi noi de date nu este valabilă într-un sistem real, cel mai des modelele reale sunt orientate spre a minimiza rata de eroare a cuvântului şi diferența de performanță dintre setul de învățare şi cel de testare.

6.2. Complexitatea

Complexitatea (Meng, 2004) este o altă caracteristică care trebuie să fie luată în considerație când se discută performanța unui sistem de recunoaştere a vorbirii. Această trăsătură este şi mai importantă când achiziția de componente hardware prezintă o problemă pentru succesul sistemului. În termeni generali, complexitatea unui sistem ASR se referă la complexitatea procesării şi complexitatea modelului proiectat. Complexitatea de procesare cuprinde costul dat de timpul de procesare a fiecărei subcomponente a modelului. În cazul sistemelor reale de recunoaştere a vorbirii, în care execuția operațiilor trebuie realizată într-un timp eficient, complexitatea procesării prezintă o importanță crucială. Complexitatea modelului este măsurată în numărul de parametri distincți ai modelului.

Există o relație strânsă între complexitatea modelului şi precizia de recunoaştere.

6.3. Robustețea

În timp ce acuratețea este esențială pentru performanța unui sistem ASR, la fel şi robustețea joacă un rol la fel de important (Meng, 2004). În prezent, mai multe sisteme de recunoaştere a vorbirii sunt antrenate pe un set de date colectate în anumite situații prevăzute. Aceste sisteme ar funcționa corect dacă condițiile de procesare ar fi echivalente cu situațiile în care au fost colectate datele. Din păcate, de cele mai multe ori, condițiile date nu sunt respectate întrucât aceste diferențe sunt inevitabile. Aspectele esențiale referitoare la condițiile de procesare a sistemelor ASR cuprind nivelul zgomotului de fundal, denaturarea canalului de comunicație, zgomotul de pe canalul de comunicație, diferența dintre stilurile de vorbire, deviația sintactică, vorbirea spontană etc. În practică, deviația acestor condiții de la trăsăturile stabilite în faza de proiectare a sistemului poate afecta substanțial performanța acestuia. Astfel, robustețea reprezintă o preocupare esențială, devenind un indicator de performanță critic al tuturor sistemelor de recunoaştere a vorbirii.

6.4. Metode de măsurare a performanței

Evaluarea performanței este esențială în procesul de dezvoltare a unui sistem de recunoaştere a vorbirii. Una dintre cele mai utilizate metode de măsurare a performanței este calculul ratei erorii de recunoaştere a cuvântului (*word recognition error rate* sau *WER*) . În cazul analizei mai multor algoritmi de modelare acustică , este foarte important de comparat reducerea relativă a erorii (*relative error reduction* (Huang, Acero, & Hon, 2001)). Aşadar, pentru a estima parametrul WER este necesar un set de date care conține cel puțin 500 de propoziții, colectate de la 5 până la 10 vorbitori diferiți. Cu ajutorul acestor date trebuie să se obțină o reducere relativă a erorii de 10% astfel încât să considerăm un algoritm optim. Un alt caz ar fi utilizarea unui eşantion de cuvinte mai mic (extras din datele pentru învățare) pentru măsurarea performanței setului de învățare, ceea ce e mai rațional decât utilizarea datelor din setul pentru testare. Performanța setului de învățare este utilă în procesul de dezvoltare a sistemului, întrucât aceasta permite identificarea potențialelor erori. În caz contrar, ar trebui utilizate un set pentru dezvoltare (*development set*) care, cel mai des, conține date care nu sunt niciodată utilizate în procesul de învățare .

În contextul sistemelor de recunoaştere a vorbirii se evidențiază trei categorii de erori:

- **Substituția** : un cuvânt corect a fost înlocuit cu un cuvânt fonetic asemănător
- **Ştergerea** : un cuvânt corect a fost omis în secvența recunoscută
- **Inserarea** : un cuvânt în plus a fost adăugat în secvența recunoscută

Rata minimă de eroare (*minimum error rate*) nu poate fi calculată doar prin compararea a două secvențe de cuvinte, cuvânt cu cuvânt. Formula pentru calcularea ratei de eroare a cuvântului (*Word Error Rate*) fiind următoarea:

$$Word\ Error\ Rate = 100\% * \frac{Substitutii + Stergeri + Inserari}{Numarul\ de\ cuvinte\ din\ propozitia\ corecta}$$

Această metodă mai este numită potrivirea maximă a subşirurilor, care poate fi manipulată cu ajutorul programării dinamice.

7. Concluzii

Recunoaşterea vocală automată reprezintă un aspect important în sistemele

inteligente ce tind să faciliteze interacțiunea omului cu calculatorul. Pentru a fi eficiente, aceste sisteme trebuie să permită procesări precise în timp real. În atingerea acestui deziderat, arhitectura sistemului ASR joacă un rol esențial. In această lucrare am oferit o perspectivă actualizată asupra arhitecturii unui astfel de sistem şi a componentelor sale constitutive. Credem că analiza intermediară prezentată în acest articol va facilita atingerea obiectivului mai general pe care ni l-am propus: conceperea şi realizarea unui sistem ASR cu un înalt nivel de acuratețe, independent de locutor pentru recunoaşterea limbii române într-un domeniu restrâns de aplicabilitate, precum justiția.

Confirmare

This work was financed by UEFISCI, under PN-II-PTPCCA-2013-4-1644, and cofinanced from the European Social Fund through the project POSDRU/159/1.5/S/134197.

Referințe

Adami, A. G. (2010). Automatic Speech Recognition: From the beginnning to the Portuguese Language. Caxias do Sul: Universidade de Caxias do Sul.

Deng, L., & Li, X. (2013). Machine Learning Paradigms for Speech Recognition: An Overview. IEEE Transactions on Audio, Speech, and Language Processing, 1060-1089.

Duma, M., Giurgea, C., Ordean, M., & Zălhan, P. (2015). Premise pentru o tehnologie de recunoaştere automată a vorbirii în limba română aplicată domeniului juridic. Revista Română de Interacțiune Om-Calculator 8(1), 1-20.

Goel, V., Kumar, S., & Byrne, W. (2004). Segmental minimum Bayes-risk decoding for automatic speech recognition. IEEE Trans. Speech Audio Process, 234-249.

Gruhn, R., Minker, W., & Nakamura, S. (2011). Statistical Pronunciation Modeling for Non-Native Speech Processing. Hardcover: Springer.

Huang, X., Acero, A., & Hon, H.-W. (2001). Spoken Language Processing: a guide to theory, algorithm, and system development. United Sate of America: Prentice Hall PTR.

Meng, Y. (2004). Speech Recognition on DSP: algorithm optimization and performance analysis. Hong Kong.

Meseguer, N. A. (2009). Speech Analysis for Automatic Speech Recognition. Norway.

Pylkkonen, J., & Kurimo, M. (2012). Analysis of extended Baum-Welch and constrained optimization for discriminative training of HMMs. EEE Audio, Speech, Lang. Process., 2409-2419.

Sonia, S., David, P., & Poulose, J. (2013). Combined feature extraction technique and naive bayes classifier for speech recognition. Kochi, India.