Revista Română de Interacţiune Om-Calculator 6 (2) 2013, 155-168

Recunoaşterea automată a limbii cu aplicaţie în sinteza vocală diferenţiată lingvistic

Paul Fogarassy-Neszly[1], Vasile Gherheş[2]

[1]BAUM Engineering
Str. Traian Moşoiu nr. 8, 310175 Arad
E-mail: pf@baum.ro

[2]Universitatea „Politehnica" Timişoara
P-ţa Victoriei nr. 2, 300006 Timişoara,
E-mail: vasile.gherhes@cls.upt.ro

Rezumat. Lucrarea prezintă o trecere în revistă a unor considerente privind recunoaşterea automată a limbii în care este scris un text şi continuă cu descrierea particularităţilor algoritmilor folosiţi la sinteza vocală diferenţiată lingvistic. Sunt prezentate metodele de optimizare a algoritmilor, criteriile de performanţă la recunoaşterea limbii, precum şi particularităţile legate de cititoarele de ecran folosite în medii multiculturale.

Cuvinte cheie: recunoaşterea limbii, sinteză vocală, accesibilitate, tehnologii asistive.

1. Introducere

Recunoaşterea automată a limbii (RAL) în care este scris un text face parte din categoria generală a algoritmilor de clasificare a textelor şi are numeroase aplicaţii. În general, algoritmii pentru recunoaşterea limbii sunt utilizaţi pentru analiza automată a unor volume mari de documente în vederea căutării şi extragerii de informaţii şi în general pentru clasificarea documentelor. Unele aplicaţii recente ale RAL au în vedere sinteza vocală diferenţiată lingvistic şi traducerea automată (Kłosowski, P., Dustor, A. 2013; Devadoss, 2010).

În comparaţie cu alte operaţii de prelucrare a textelor, cum ar fi extragerea de informaţii sau traducerea automată, RAL implică algoritmi mai puţin sofisticaţi; problema devine însă complexă atunci când apar restricţii privind încărcarea procesorului, viteza algoritmilor implicaţi şi dimensiunea redusă a textului analizat (Dunning, 1994; Zissman, 1996).

Sinteza vocală este utilizată în mod special de către persoanele cu deficienţe de vedere sau de citire (dislexie) pentru accesibilizarea

documentelor în format electronic. Algoritmii de sinteză vocală se bazează pe particularităţile fonetice ale limbii în care este scris textul, precum şi pe regulile de scriere specifice limbii respective. Atunci când există posibilitatea ca texte scrise în limbi diferite să alterneze, este necesară identificarea limbii în care este scris fragmentul de text. În funcţie de aceasta, este folosită sinteza vocală corespunzătoare; acest lucru este numit *sinteză vocală diferenţiată lingvistic*.

În continuare acest articol este organizat astfel: în secţiunea următoare sunt trecute în revistă metodele de recunoaştere automată a limbii; în secţiunea 3 sunt prezentate particularităţile identificării limbii în contextul specific al aplicaţiilor de sinteză vocală diferenţiată lingvistic şi particularităţile algoritmilor pentru identificarea limbii care răspund cerinţelor specifice aplicaţiilor de sinteză vocală utilizate în tehnologiile asistive. În ultima secţiune sunt prezentate criteriile de performanţă a algoritmilor de RAL şi considerente privind particularităţile cititoarelor de ecran folosite de nevăzători şi dislexici, precum şi modul în care acestea influenţează sinteza vocală diferenţiată.

2. Metode pentru Recunoaşterea Automată a Limbii (RAL)

2.1 Identificarea intervalului de codificare a caracterelor

Identificarea limbii în care este scris un text poate fi făcută uneori cu uşurinţă prin simpla identificare a codificării caracterelor (character encoding detection) pentru textul analizat (Baldwin şi Lui, 2010). Pentru tratarea unitară, textul este convertit (dacă este cazul) în format Unicode. Acesta este un format definit de către Unicode Consortium pentru codarea, stocarea şi interpretarea textelor în mediul informatic; acesta este standardul de codificare de facto utilizat la interpretarea datelor binare în format text (The Unicode Consortium, 2006).

Pentru ca discriminarea limbilor prin identificarea intervalului Unicode să poată fi aplicată, trebuie ca limbile avute în vedere să folosească intervale Unicode diferite. Astfel, metoda se pretează (de exemplu) pentru discriminarea între limba română şi rusă, între ebraică şi arabă sau între greacă şi engleză. Desigur, metoda nu poate fi utilizată (de exemplu) pentru discriminarea limbilor engleză şi franceză sau a limbilor rusă şi bulgară,

deoarece aceste perechi de limbi folosesc caractere din acelaşi interval Unicode. În Tabelul 1 sunt prezentate intervalele Unicode pentru câteva limbi mai mult folosite.

Tabelul 1. Intervalele Unicode pentru câteva limbi (The Unicode Consortium, 2006)

Tip caractere	Nume Unicode	Interval Unicode
Arabe	Arabic	U+0600-U+06FF
	Arabic Supplement	U+0750-U+077F
	Arabic Extended-A	U+08A0-U+08FF
	Arabic Presentation Forms-A	U+FB50-U+FC3F
	Arabic Presentation Forms-B	U+FE70-U+FEFF
Chineze	CJK Unified Ideographs	U+4E00-U+9FFF
	CJK Extension A	U+3400-U+4DBF
	CJK Extension B	U+20000-U+2A6DF
	CJK Extension C	U+2A700-U+2B73F
	CJK Extension D	U+2B840-U+2B81F
Chirilice	Cyrillic	U+0400-U+04FF
	Cyrillic Supplement	U+0500-U+052F
	Cyrillic Extended-A	U+2DE0-U+2DFF
	Cyrillic Extended-B	U+A640-U+A69F
Ebraice	Hebrew	U+0590-U+05FF
	Hebrew Presentation Forms	U+FB00-U+FB4F
Greceşti	Greek and Coptic	U+0370-U+03FF
	Greek Extended	U+1F00-U+1FFF
Japoneze	Kanji	U+4E00–U+9FBF
	Hiragana	U+3040–U+309F
	Katakana	U+30A0–U+30FF
Latine	Basic Latin	U+0000-U+007F
	Latin-1 Supplement	U+0080-U+00FF
	Latin Extended-A	U+0100-U+017F
	Latin Extended-B	U+0180-U+024F

2.2 Identificarea limbii prin metode statistice

Probabil că cea mai intuitivă metodă de a identifica o limbă constă în căutarea cuvintelor (inclusiv a formelor flexionare) în dicționare specifice limbilor respective. Desigur, această metodă, deşi are avantajul considerabil de a fi precisă chiar şi pentru texte foarte scurte, nu este practică datorită volumului foarte mare al dicționarelor; acest lucru, împreună cu identificarea formelor flexionare implică algoritmi lenți.

O altă metodă intuitivă şi simplă, metodă folosită încă la sortarea manuală a documentelor scrise cu acelaşi tip de alfabet, se bazează pe identificarea unor caractere speciale tipice unei limbi. Acestea sunt de

obicei caracterele cu diacritice şi ligaturile. În Tabelul 2 sunt prezentate caracterele speciale folosite în câteva limbi europene care se scriu cu grafie latină.

Tabelul 2. Caractere speciale în limbi cu grafie latină

Limba	Caractere specifice
Albaneză	ç ë
Cehă	á é í ó ú ý č ď ě ň ř š ť ž ů
Daneză	ø å
Estonă	õ ä ö ü
Franceză	à â æ ç é è ê ë î ï ô œ ù û ü ÿ
Finlandeză	ä ö
Germană	ä ö ß ü
Irlandeză	á é í ó ú
Letonă	ā ē ī ū ŗ ļ ķ ņ ģ š ž č
Lituaniană	č š ž ą ę į ų ū ė
Maghiară	ö ü á é í ó ú ő ű
Poloneză	ą ć ę ł ń ó ś ź ż
Română	ă â î ş ţ
Sârbă	č ć đ š ž
Slovacă	á é í ó ú ý Í ŕ č ď ľ ň š ť ž ä ô
Slovenă	č š ž
Spaniolă	ñ

Desigur, folosirea caracterelor speciale pentru discriminarea limbilor are aplicabilitate limitată, deoarece în cazul unor texte scurte acestea pot să nu apară. De asemenea, unele dintre caracterele speciale sunt comune mai multor limbi.

O altă idee care a circulat la începutul anilor 1990 a plecat de la identificarea unor secvenţe de caractere specifice exclusiv unei singure limbi (Churcher, 1994; Churcher et al., 1994, Dunning, 1994). Tabelul 3 prezintă câteva asemenea secvenţe de caractere.

Tabelul 3. Secvenţe de caractere specifice unei singure limbi

Limba	Secvenţă
Olandeză	„vnd”
Engleză	„ery ”
Franceză	„eux ”
Galeză	„mh”
Germană	„ der ”
Italiană	„cchi ”
Sârbă	„lj ”
Spaniolă	„ ir ”

Deşi ideea pare tentantă prin simplitatea sa, asemenea secvenţe de caractere izolate nu oferă suficientă încredere în identificarea limbii pentru

care se presupune că sunt specifice, datorită unui număr mare de excepții posibile, precum şi din cauza cuvintelor preluate în alte limbi şi scrise conform regulilor din limba de origine.

2.3 Identificarea limbii pe baza analizei n-gramelor

Metoda secvenței de caractere specifice a condus la cea mai utilizată metodă şi anume metoda *n-gramelor (*Dunning, 1994). Deşi analiza frecvenței n-gramelor este tot o metodă statistică, aceasta este prezentată separat, datorită faptului că este cea mai utilizată şi are mai multe variante, fiecare cu avantajele şi dezavantajele sale.

O n-gramă este o sub-secvență de n elemente dintr-o secvență dată; în general, secvența de elemente poate fi orice, de la caractere şi până la cuvinte. În analiza lingvistică n-gramele sunt utilizate mai mult pentru cuvinte sau pentru caractere. În această lucrare, prin n-gramă se înțelege o secvență de n-caractere succesive dintr-un text. Atunci când este vorba de două caractere (n = 2) se mai foloseşte termenul de bigramă (sau digramă), iar când este vorba de succesiuni de trei caractere (n = 3) termenul consacrat este trigramă.

În Tabelul 4 este prezentat un exemplu de descompunere în n-grame a cuvântului „analiză".

Tabelul 4. Bigramele, trigramele şi 4-gramele cuvântului „analiză"

	analiză
2-grame	_a , an , na , al , li , iz , ză, ă_
3-grame	_an , ana , nal , ali , liz , iză , ză_
4-grame	_ana , anal , nali , aliz , liză, iză_

După cum se poate vedea în Tabelul 4, începutul de cuvânt şi sfârşitul de cuvânt au fost marcate printr-un caracter special, ne-literal (_). Este relevantă din punct de vedere lingvistic această marcare, deoarece localizarea unui grup de caractere la începutul sau la sfârşitul cuvântului este statistic semnificativă pentru caracterizarea unei limbi.

Discriminarea limbilor pe baza n-gramelor pleacă de la observația că pentru fiecare limbă anumite n-grame apar mai frecvent decât altele. Studii experimentale realizate de Cavnar and Trenkle (1994) au arătat că utilizarea trigramelor conduce la cele mai bune rezultate. Identificarea limbii se face

prin compararea frecvenței de apariție a trigramelor în textul analizat cu frecvența acestora în corpusurile limbilor care sunt avute în vedere.

În faza de „învățare" a unei limbi se construieşte spectrul de frecvențe al n-gramelor pentru fiecare limbă în parte. Acesta se bazează pe un corpus relevant pentru limba avută în vedere şi domeniul de aplicare (dacă este cazul). Din studii realizate de (Dunning, 1994; Ljubeši et al., 2007) rezultă că un corpus de circa 50.000 de cuvinte oferă o precizie foarte bună care nu mai creşte semnificativ prin mărirea volumului. Frecvențele relative ale n-gramelor reprezintă caracteristica fiecărei limbi, iar calculul frecvenței n-gramelor din textul analizat se realizează în timpul rulării, în acelaşi mod. În funcție de modul de construire a acestui spectru, metoda poate fi mai rapidă sau mai lentă, mai precisă sau mai puțin precisă.

Corpusul trebuie să fie omogen din punctul de vedere al limbii caracterizate de acesta şi trebuie să fie corect gramatical şi sintactic; calitatea corpusului are o influență hotărâtoare asupra preciziei de identificare a limbii.

Compararea spectrului analizat cu cele de referință se poate face în diverse feluri. Cel mai simplu criteriu îl reprezintă suma abaterilor absolute (ecuația 1) sau suma abaterilor pătratice (ecuația 2).

$$A_L = \sum_{i=1}^{m} \left| f_{ai} - f_{Li} \right| \tag{1}$$

$$A_L = \sum_{i=1}^{m} \left(f_{ai} - f_{Li} \right)^2 \tag{2}$$

unde A_L este abaterea frecvențelor pentru limba L, m este numărul de n-grame din textul analizat, f_{ai} este frecvența n-gramei i din textul analizat, iar f_{Li} este frecvența n-gramei i din spectrul de frecvențe al limbii L.

Limba identificată cu cea mai mare probabilitate este limba L pentru care A_L este minim.

Criterii mai sofisticate de calcul al abaterii A_L țin cont (printr-un coeficient de pondere) de probabilitatea mult mai mare sau mult mai mică (chiar zero) a unor n-grame pentru o anumită limbă. În acest caz, abaterea A_L se calculează conform relației (3).

$$A_L = \sum_{i=1}^{m} k_{Li} \left(f_{ai} - f_{Li} \right)^2 \tag{3}$$

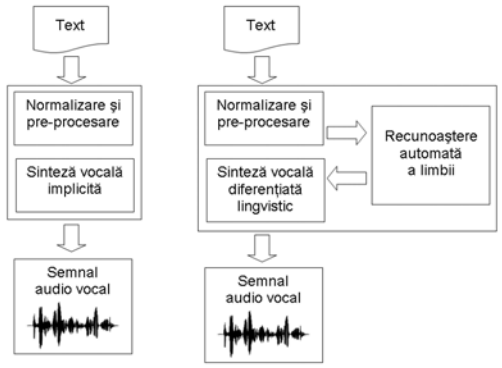
unde k_{Li} este coeficientul de pondere pentru n-grama i din limba L; spre deosebire de ecuațiile (1) şi (2), $k_{L\text{i}}$ poate fi pozitiv sau negativ, aşadar A_L poate avea valori pozitive sau negative.

3. Particularități ale RAL pentru aplicații de sinteză vocală diferențiată

3.1 Identificarea limbii în contextul specific al aplicațiilor de sinteză vocală diferențiată

În cazul cel mai simplu, sinteza vocală se realizează plecând de la un text de intrare; acesta, în urma unei pre-procesări, este trimis unui modul de sinteză vocală care produce rezultatul audio.

În cazul în care există posibilitatea ca textul care trebuie sintetizat să fie în limbi diferite, se impune în prealabil recunoaşterea limbii acestuia, pentru a se folosi sinteza vocală specifică acestei limbi. Acest lucru este prezentat schematic în Figura 1.



a) Sinteza vocală cu limbă implicită *b) Sinteza vocală diferențiată lingvistic*

Figura 1. Sinteza vocală cu limbă implicită şi cu diferenţiere lingvistică

Eficiența şi precizia algoritmilor de RAL depind direct de trei factori: numărul de limbi posibile, volumul şi calitatea corpusului limbilor folosite pentru caracterizarea limbilor şi dimensiunea textului analizat. În cazul

particular a discriminării limbii pentru sinteza vocală diferențiată, numărul de limbi este în general două, mai rar trei şi numai în cazuri speciale sunt avute în vedere mai mult de trei limbi. Restricțiile majore pentru acest tip de aplicație sunt lungimea mică a textului analizat (uneori de numai câteva cuvinte) şi eficiența algoritmului din punctul de vedere al vitezei de calcul.

Sinteza vocală a unui text cu porțiuni în limbi diferite apare în special în mediile multi-culturale, în țările în care se vorbesc mai multe limbi. De exemplu în Republica Moldova se vorbeşte în mod curent limba română (72%) şi limba rusă (27%), în Belgia se vorbeşte olandeza (58%), franceza (32%) şi germana (10%), în Elveția se vorbeşte germană (63%), franceză (20%) şi italiană (7%) etc. Este firesc să existe texte care cuprind fragmente scrise în limbi diferite în asemenea zone multi-culturale, motiv pentru care înainte de redarea unui asemenea text prin sinteză vocală, fiecare fragment de text trebuie analizat pentru identificarea corectă a limbii în care este scris şi doar apoi redat cu vocea sintetică corespunzătoare. În lipsa unui asemenea demers, rezultatul poate fi incomprehensibil.

La ora actuală există un număr mare de aplicații care permit recunoaşterea limbii; aceste aplicații sunt deosebit de performante şi au o rată de succes foarte bună, mai ales pentru texte suficient de lungi. Aceste aplicații pot discrimina chiar zeci de limbi. Din păcate, algoritmii utilizați de aceste aplicații nu pot fi folosiți în aplicațiile care implică sinteza vocală. În aplicațiile de accesibilizare pentru deficienți de vedere, sinteza vocală rulează în paralel atât cu cititorul de ecran, cât şi cu aplicația folosită de utilizator. Acest lucru impune pentru sinteza vocală o responsivitate cât mai bună, adică un timp cât mai scurt între momentul în care un text este trimis spre sinteză şi momentul în care începe redarea vocală a acestuia; în mod normal, 100 de milisecunde este limita maxim admisibilă a timpului în care aplicația trebuie să răspundă la orice intervenție a utilizatorului.

3.2 Particularitățile algoritmilor de sinteză vocală diferențiată

Ca în orice situație în care viteza unui algoritm este critică, trebuie lăsate la rulare doar calculele care nu pot fi efectuate în timpul analizei informatice.

Spre deosebire de aplicațiile de RAL cu caracter general, condițiile de eficiență a algoritmilor folosiți la sinteza vocală diferențiată sunt favorizate prin limitarea la minimum a limbilor posibile; în mod normal, mediile multiculturale trebuie să distingă între două, mai rar trei şi foarte rar patru

limbi. Prima etapă de optimizare constă în configurarea aplicației doar pentru limbile care trebuie discriminate, conform specificațiilor beneficiarului. Această alegere permite o viteză de răspuns mult mai mare decât a algoritmilor cu caracter general.

A doua etapă de optimizare constă în selecția metodei de discriminare în funcție de intervalul Unicode caracteristic limbilor specificate. Dacă cele două limbi folosesc intervale Unicode diferite (vezi tabelul 1) atunci acesta este singurul criteriu de discriminare. În cazul în care cel puțin două dintre limbile specificate folosesc acelaşi interval Unicode, atunci se va folosi un algoritm bazat pe spectrul n-gramelor specific acestor limbi.

A treia optimizare, care se realizează tot prin configurarea inițială a aplicației, constă în definirea unei limbi implicite. Acest lucru este util pentru situațiile în care rezultatul analizei este incert, mai ales datorită fragmentelor de text foarte scurte, sau datorită amestecului de cuvinte din limbi diferite în interiorul aceluiaşi fragment de text; tot rezultat incert poate să apară şi în cazul în care textul este scris într-o altă limbă decât cele pentru care algoritmul a fost configurat.

În sfârşit, a patra optimizare care se poate face constă în realizarea unor liste de n-grame specifice pentru fiecare limbă analizată. În aceste liste vor fi păstrate doar n-gramele caracteristice care apar cu o pondere semnificativă în limba respectivă, de exemplu prin eliminarea n-gramelor care apar cu o frecvență relativă mai mică de 10% din media frecvenței tuturor n-gramelor identificate în corpusul limbii respective; valoarea maximă a acestui procent poate fi determinată experimental.

În cazul în care discriminarea urmează să fie făcută între două limbi, se realizează o nouă listă care va conține diferența frecvențelor relative a n-gramelor din listele caracteristice celor două limbi. Din nou, dacă această diferență este nesemnificativă, adică mai mică în valoare absolută decât 10% (de exemplu) din media diferențelor absolute, lista poate fi restrânsă, obținându-se o listă de diferențe semnificative ale n-gramelor caracteristice celor două limbi.

Pentru a identifica limba în care este scris textul analizat, pentru fiecare din cele *m* n-grame ale acestuia se va calcula suma dată de ecuația (4)

$$A = \sum_{j=1}^{m} p_j \tag{4}$$

în care p_j este diferența frecvențelor n-gramelor caracteristice celor două limbi; diferențele p_j pentru n-gramele caracteristice cel puțin uneia dintre cele două limbi se calculează conform ecuației (5).

$$p_j = f_{L1j} - f_{L2j} \tag{5}$$

în care f_{L1j} și f_{L2j} sunt frecvențele n-gramei j în lista de n-grame caracteristice limbii L_1 respectiv L_2.

După cum se poate observa, aceasta este o optimizare foarte puternică, deoarece atât întocmirea listei de frecvențe ale n-gramelor caracteristice, cât şi calculul diferențelor acestora în cele două limbi se fac în faza de pregătire a aplicației, la rulare rămânând de calculat doar suma dată de ecuația (5), pentru un număr relativ mic de numai m valori (numărul de n-grame a textului analizat).

Din rezultatul final al sumei date de ecuația (4) se va folosi doar semnul rezultatului, Sign(A) şi eventual raportul dintre numărul total de n-grame ale textului analizat şi suma calculată, ca o măsură a certitudinii de discriminare. Dacă semnul sumei A este pozitiv, atunci cel mai probabil textul analizat este scris în limba $L1$, altfel este mai probabil ca acesta să fie scris în limba $L2$. Rezultatul ecuației (4) este incert dacă valoarea sa absolută este mai mică decât un prag limită, care poate fi calculat în funcție de media frecvențelor relative ale n-gramelor în cele două limbi şi numărul de n-grame ale textului analizat. De exemplu, pragul de incertitudine poate fi stabilit pentru valori mai mici de 10% din frecvența medie ale n-gramelor caracteristice înmulțite cu m, numărul de n-grame ale textului analizat.

4. Performanța algoritmilor de RAL

4.1 Considerente privind particularitățile cititoarelor de ecran

Cel mai comun context în care se foloseşte sinteza vocală diferențiată este utilizarea cititoarelor de ecran. Cititorul de ecran este o aplicație informatică care identifică proprietățile obiectelor care alcătuiesc interfața grafică a unei aplicații şi le transformă în informație de tip text, care poate fi în continuare transmisă unui program de sinteză vocală şi/sau unui afişaj Braille. Deşi acest gen de aplicații se adresează în principal nevăzătorilor, acestea sunt

utile şi persoanelor cu deficiențe de vedere, precum şi utilizatorilor care nu pot sau nu ştiu să citească (Fogarassy, 2011).

Cele mai performante cititoare de ecran permit pre-procesarea textului înainte de a fi trimis către sinteza vocală sau afişajul Braille. Acest lucru, deşi este util în general, poate interfera cu algoritmii de RAL; din acest motiv, utilizatorul va trebui să-şi configureze corespunzător cititorul de ecran, pentru a evita erorile de citire.

De exemplu, majoritatea cititoarelor de ecran au posibilitatea de a interpreta numerele, transformându-le în cuvinte care sunt ulterior trimise la sinteza vocală (de exemplu, „12” poate fi transformat în „doisprezece” sau în „twelve”, în funcție de modul în care este setată aplicația). În cazul în care contextul lingvistic este diferit de cel implicit, se obține un rezultat necorespunzător, în general incomprehensibil în urma sintezei vocale.

Alte exemple sunt legate de interpretarea unor abrevieri, a datelor calendaristice, a expresiilor matematice, a simbolurilor etc. Având în vedere că sintezele vocale performante pot gestiona corespunzător acest gen de situații este recomandată dezactivarea interpretării lor de către cititorul de ecran; astfel, acestea vor fi citite în limba corespunzătoare contextului în care se găsesc.

O altă problemă tipică cititoarelor de ecran constă în segmentarea unui text în fragmente de numai câteva cuvinte; uneori, numerele, datele calendaristice sau expresiile matematice sunt separate de contextul lor, ceea ce poate face dificilă pronunțarea corectă a acestora. În acest caz, algoritmul RAL trebuie să identifice incertitudinea şi fie să păstreze neschimbată limba identificată anterior, fie să concatenze textul anterior cu cel curent pentru a elimina incertitudinea.

4.2 Criterii de performanță ale RAL

Datorită faptului că în aplicațiile de sinteză vocală diferențiată textele analizate sunt scurte şi foarte scurte (uneori doar unul sau două cuvinte), este de aşteptat ca rezultatele să nu fie întotdeauna corecte. Aceste situații nu pot fi gestionate decât în cazul când rezultatul analizei este identificat ca incert. În această situație, fie se consideră că limba identificată este cea implicită, fie se foloseşte textul analizat anterior împreună cu textul curent pentru a creşte volumul textului analizat; eventual, se poate considera că în cazul unor rezultate incerte limba nu se schimbă.

Cel mai frecvent context defavorabil pentru algoritmii de RAL constă în analiza textelor foarte scurte. Pentru asemenea situații, în cazul unui rezultat incert, se recomandă reanalizarea textului curent la care se adaugă textul analizat anterior; păstrarea limbii identificate anterior este o altă soluție pentru acest caz.

O altă situație defavorabilă în RAL constă în analiza unui text care conține cuvinte şi/sau fragmente din mai multe limbi. În cazul în care textul este suficient de lung, o analiză statistică ar putea identifica o distribuție bi-modală a frecvențelor n-gramelor caracteristice; în urma unui asemenea rezultat, textul analizat poate fi segmentat pentru reluarea analizei, în speranța unui rezultat mai bun. Desigur, analiza statistică nu poate fi realizată cu un efort de calcul suficient de mic, astfel încât să nu fie introduse întârzieri datorate încărcării procesorului, inadmisibile pentru utilizator, aşa că în cazul în care textul analizat este suficient de lung, dar rezultatul obținut este incert este aproape sigur cazul unui text mixt şi se recomandă menținerea limbii identificate anterior sau presupunerea că textul este scris în limba implicită.

În cazul în care algoritmii de RAL sunt neperformanți, utilizatorul constată întreruperi ale fluxului vocal şi funcționarea greoaie a aplicațiilor; acest lucru se poate întâmpla şi în cazul sistemelor de calcul depăşite moral.

În cazul erorilor de identificare a limbii utilizatorul va auzi unele cuvinte pronunțate într-o altă limbă; acest lucru poate face fragmentul de text neinteligibil.

O altă deficiență de concepție a unei aplicații de sinteză vocală diferențiată, bazată pe RAL constă în pronunția unor cuvinte izolate în limba în care au fost scrise, deşi contextul este într-o altă limbă. Un exemplu de acest fel este citirea izolată a unor cuvinte împrumutate din alte limbi, nume de persoane, toponime etc. Schimbarea limbii în care se face sinteza vocală în interiorul unui fragment de text relativ scurt este supărătoare pentru utilizator.

5. Concluzii şi direcții de cercetare

Prezenta lucrare trece în revistă metodele de recunoaştere automată a limbii unui text. De asemenea, sunt prezentate particularitățile aplicării algoritmilor la sinteza vocală în timpul rulării împreună cu cititoarele de

ecran, pe dispozitive mobile, relativ sărace în resurse hardware (memorie şi procesor), precum şi faptul că textul care trebuie analizat este scurt.

Modificarea unor algoritmi existenți, după cum s-a arătat, trebuie însoțită de analiza performanțelor în diverse condiții, cum sunt de exemplu texte scurte dar care conțin totuşi cuvinte din limbi diferite. Un amplu program experimental semi-automatizat urmează să valideze algoritmii care vor fi propuşi de analişti. De asemenea, tot experimental, se vor identifica pragurile optime de incertitudine.

Confirmare

Această lucrare a fost elaborată ca şi analiză primară în cadrul propunerii de proiect „*Aplicație pentru Conversia din Text în Voce Sintetică cu Recunoaşterea Automată a Limbii*”, în cadrul Programului Inovare, Dezvoltare Sisteme-Produse-Tehnologii, declarat eligibil pentru finanțare, în urma evaluării de către UEFISCDI.

Referințe

Baldwin, T. and Lui M. (2010) *Language Identification: The Long and the Short of the Matter*. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 229–237, Los Angeles, California

Cavnar, W., and Trenkle, J. (1994). N-gram-based text categorization. Proc. 3rd Symp. on Document Analysis and Information Retrieval (SDAIR-94)

Churcher, G. (1994) *Distinctive character sequences*. Personal Communication by Ted Dunning

Churcher, G., Hayes, J., Johnson, S. and Souter, C (1994) *Bigraph and trigraph models for language identification and character recognition*. in Proceedings of 1994 AISB Workshop on Computational Linguistics for Speech and Handwriting Recognition, University of Leeds

Devadoss, J.M. (2010) *Advanced Natural Language Translation System*, Global Journal of Computer Science and Technology, Vol 9, No 5

Dunning, T. (1994) *Statistical Identification of Language*. Technical Report MCCS 94-273, New Mexico State University

Fogarassy-Neszly, P. (2011) *Aplicații informatice şi dispozitive cu interfață vocală*. Revista Română de Interacțiune Om-calculator 4 (Numar special RoCHI 2011), 53-58.

Kłosowski, P., Dustor, A. (2013) *Automatic Speech Segmentation for Automatic Speech Translation*, Computer Networks Communications in Computer and Information

Science Volume 370, 2013, pp 466-475

Ljubeši´c, N., Mikeli´c, N., and Boras, D. (2007) *Language identification: How to distinguish similar languages*. In Lužar-Stifter, V. and Hljuz Dobri´c, V., editors, Proceedings of the 29th International Conference on Information Technology Interfaces, pages 541–546, Zagreb SRCE University Computing Centre

The Unicode Consortium (2006) *The Unicode Standard, Version 5.0*. Fifth Edition, Addison-Wesley Professional, 27

Zissman, M. A. (1996) *Comparison of four approaches to automatic language identification of telephone speech.* IEEE Transactions on Speech and Audio Processing, vol. 4.