Componentele unui sistem de sinteză text-vorbire

Mihai Alexandru Ordean
iQuest Technologies
Cluj-Napoca, România
Mihai.Ordean@iquestint.com

Andrei Şaupe
iQuest Technologies
Cluj-Napoca, România
Andrei.Saupe@iquestint.com

Mihaela Ordean
iQuest Technologies
Cluj-Napoca, România
Mihaela.Ordean@iquestint.com

Dorian Gorgan
Universitatea Tehnică
Cluj-Napoca, România
Dorian.Gorgan@cs.utcluj.ro

REZUMAT
In acest articol vom descrie componentele unui sistem de sinteză text-vorbire (TTS) pentru limba română. Se prezintă arhitectura generală a sistemului TTS după care sunt detaliate componentele acestuia şi despărţirea în silabe.

Cuvinte cheie
Sinteza text-vorbire, procesare de text, despărţirea în silabe.

Clasificare ACM
H5.2. Information interfaces and presentation (HCI): Miscellaneous.

INTRODUCERE

Regulile de procesare a textului din cadrul sistemului TTS sunt conforme cu cele din DOOM II [1]. În prima etapă din sinteza text-vorbire se convertesc simbolurile ortografice lexicale într-o reprezentare fonemică , eventual utilizând informaţii diacritice, precum plasarea accentului [2]. Analiza fonetică este obligatorie într-un sistem de sinteză text-vorbire dacă obiectivele se axează pe obţinerea unei calităţi ridicate pentru sunetul final, care să acopere toate inflexiunile şi complexitatea unui limbaj.

Convertirea cuvintelor din forma scrisă in forma vorbită nu este un proces trivial şi influenţează puternic performanţele unui sistem de sinteză text-vorbire [2].

Componenta de analiză a textului într-un sistem TTS este responsabilă, în general, de determinarea structurii documentului de intrare, de conversia simbolurilor ne-ortografice şi de înţelesul textului. Această componentă indică toate cunoştinţele despre text, specifică şi codează într-un format uşor de folosit în partea de sinteză a vorbirii.

Elementul de bază în componenta de analiză a textului este parserul de text care transcrie şi dezambiguează textul de intrare luând decizii referitoare la înţeles dintr-un areal nelimitat de posibilităţi.

Pentru limba română există puţine abordări de realizare a unui sistem TTS. Astfel, Burileanu et al. [4] prezintă modalitatea de realizare a unui asemenea sistem, incluzând detalii pentru transcrierea fonetică. Pentru parsarea textului, în [4] se foloseşte o abordare clasică, Lex/Yacc. Trebuie totuşi să menţionăm că până în prezent, nu există nici un sistem de tip TTS comercial disponibil în România.

ARHITECTURA SI COMPONENTELE UNUI SISTEM DE SINTEZĂ TEXT-VORBIRE

Pentru încadrarea analizorului de text în cadrul unui sistem TTS este utilă amintirea componentelor acestuia. Astfel, un sistem de sinteză este alcătuit din (figura 1) [3]:

Procesorul de text – responsabil cu parsarea textului de intrare şi generarea unei structuri de date pe care să o poată utiliza celelalte componente.

Componenta de selecţie a unităţilor acustice – pe baza structurii text şi a informaţiilor adiţionale, se realizează selecţia unităţilor din cadrul bazei de date de unităţi necesare ulterior în procesul de procesare şi concatenare.

Baza de date de unităţi acustice – conţine toate unităţile acustice disponibile.

Procesorul de sunet – aplică procesările necesare (precum modificarea duratei, a pitch-ului, etc.) asupra listei de unităţi din baza de date de unităţi, care au fost procesate şi concatenate.

In continuare vom face referire la componentele de procesare de text.

Textul de intrare

Textul de intrare este alcătuit dintr-o secvenţă de paragrafe care pot fi descompuse în fraze iar acestea, la rândul lor, vor fi descompuse în cuvinte.

Un cuvânt nu este doar o secvenţă de litere mici ci poate conţine şi alte caractere cu excepţia spaţiului, liniei noi şi a caracterelor speciale ('@','#'). Un cuvât poate începe şi se poate termina doar cu caractere care nu sunt marcate. Aceste caractere valide pot fi separatori (spaţii), separatori de frază ('.' '!' '?'), separatori de paragraph (linie nouă), caractere de punctuaţie (',' ':' ';') sau caractere speciale ('@' '#').

Textul de intrare nu poate conţine expresii matematice, simboluri intertextuale sau secvenţe ambigue de caractere. În cazul în care textul de intrare conţine caractere ambigue (altele decât cele menţionate mai sus), acestea se înlocuiesc cu caractere speciale.

Arhitectura sistemului TTS şi procesarea textului

Procesarea se referă la un set de operaţii care includ segmentarea textului, dezambiguarea textului şi interpretarea semnelor de punctuaţie.

Segmentarea textului. Textul este parsat prima dată pentru delimitarea paragrafelor, frazelor şi cuvintelor, rezultatul fiind o structură internă de tip arbore. În timpul parsării se realizează şi dezambiguarea abreviaţiilor prin consultarea unui dicţionar de abreviaţii. Acest dicţionar este activat la apariţia simbolului '.'.

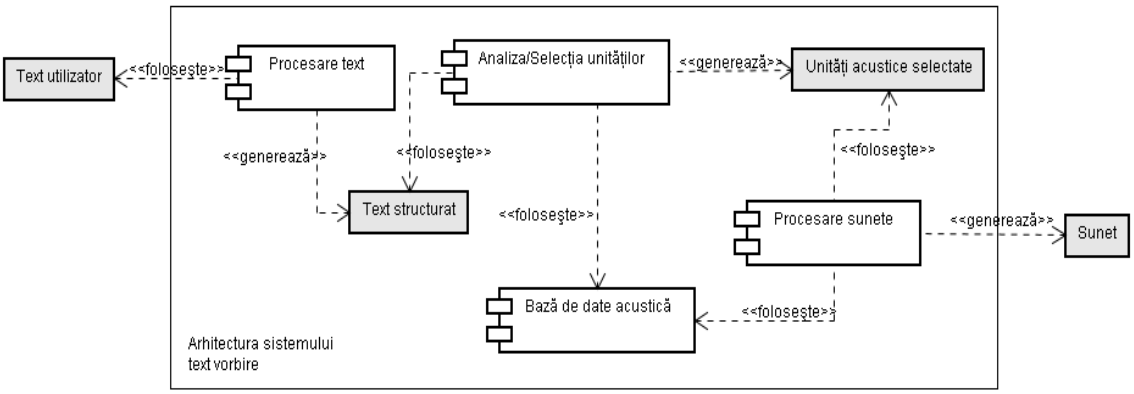

Figura 1. Arhitectura sistemului TTS

Dezambiguarea textului. Textul este parsat a doua oară pentru cuvintele de tip dată, timp şi număr. Tot în cadrul acestui pas se realizează transcrierea fonetică şi construcţia silabelor. Prima verificare în acest pas este pentru cuvintele de tip dată, timp şi număr. Dacă se întâlneşte un astfel de cuvânt acesta este expandat după care, pentru textul expandat, se realizează transcrierea fonetică şi construirea silabelor.

Interpretarea semnelor de punctuaţie. La nivelul frazelor este aplicat un algoritm de prozodie pentru interpretarea semnelor de punctuaţie şi adăugarea pauzelor la sfârşitul propoziţiilor.

Pentru procesarea textului de intrare se utilizează trei gramatici. Prima gramatică interpretează diferite predicate (cuvinte, fraze, paragrafe) activând module specifice pentru salvarea indecşilor şi construirea structurii interne de tip arbore. Pentru expandarea cuvintelor (abreviaţii, dată, timp, numere) se utilizează cea de-a doua gramatică. A treia gramatică se utilizează pentru transcrierea fonetică şi construirea silabelor.

Majoritatea modulelor componente ale gramaticilor utilizate sunt parţi statice care sunt încărcate în memorie o singură dată. O mică parte a acestor gramatici este dinamică şi se încarcă în memorie în momentul execuţiei. Gramaticile folosite la parsarea textului sunt prezentate in detaliu in [5].

Procesarea textului de intrare ar putea fi realizată într-o singură traversare prin combinarea celor trei gramatici într-una singură. Acest lucru ar determina scăderea timpului de procesare având însă drept dezavantaj creşterea costurilor de mentenanţă şi scăderea posibilităţilor de extensie. Considerând timpul câştigat prin cuplarea gramaticilor drept o îmbunătăţire minoră, s-a ales parsarea în doi paşi care permite o arhitectură modulară şi structurată.

Toate cuvintele obţinute sunt căutate în dicţionar iar, dacă nu sunt găsite, se realizează transcrierea fonetică şi despărţirea în silabe.

Gramaticile utilizate sunt construite cu toleranţă la erori astfel încât permit ca o propoziţie să înceapă cu literă mică sau să existe un text vid la intrare. Cu toate acestea, gramaticile nu rezolvă toate ambiguităţile. Similar cu [4], cu cât textul de intrare este mai clar (mai puţine ambiguităţi şi erori) cu atât se obţin rezultate mai bune la ieşire în fişierul de sunet generat.

Componentele modulului de procesare de text

Modulul de procesare a textului conţine următoarele componente (figura 2):

Dicţionarul – conţine cuvinte (atât cuvintele comune cât şi excepţiile).

Logică de Inteligenţă artificială – este componenta de inteligenţă artificială care conţine regulile de transcrire fonetică şi cele de despărţire în silabe. Se utilizează utilitarul Weka pentru determinarea regulilor pe baza unui set de antrenare. Pentru transcrierea fonetică s-au implementat regulile prezentate în DOOM II. Însă, DOOM II nu acoperă totalitatea cazurilor care pot apărea. Pentru excepţii (de exemplu “cea” sau „cia”), se generează arbori de decizie folosind Weka. Aceşti arbori de decizie sunt învăţaţi pe baza unor exemple bine selectate şi realizează cu o precizie foarte bună identificarea fonemelor în cazurile critice. Prezentarea în detaliu a modului de realizare a acestor arbori de decizie iese de sub incidenţa acestei lucrări.

Prozodia – această componentă analizează structura de date şi adaugă informaţie prozodică precum intonaţia, accentul şi pauzele de vorbire.

SELECŢIA UNITĂŢILOR DIN BAZA DE DATE

Componenta de Selectare Unităţi pe baza datelor primite de la componenta de Procesare Text va face o selectare a celor mai potrivite unităţi din Baza de Date. Acestea vor fi apoi concatenate şi modelate de componta de sinteză a sunetului [4]. Definim ca unitate o reprezentare abstractă a unui segment de vorbire.

Unele sisteme nu folosesc modelul concatenativ, ci generează sunetul. Dezavantajul la un astfel de sistem este faptul că se pierde naturaleţea semnalului sonor deoarece această naturaleţe este greu de reprezentat printr-un set de reguli de generare. În sistemele concatenative sunetul final este obţinut prin selectarea si concatenarea unor unităţi din Baza de Date care se potrivesc textului de intrare. Avantajul acestei abordări este păstrarea naturaleţii vorbitorului. Dezavantajul este că se produc concatenări de unităţi care nu se potrivesc. Segmentele din vorbire sunt foarte afectate de coarticulare astfel, dacă încercăm să concatenăm două segmente care nu au fost adiacente, putem obţine discontinuităţi prozodice sau spectrale. Discontinuităţile spectrale apar când formanţii segmentelor nu se potrivesc la punctul de concatenare. Discontinuităţile prozodice se referă la nepotrivirea pitch-ului celor două unităţi în punctul de concatenare [7].

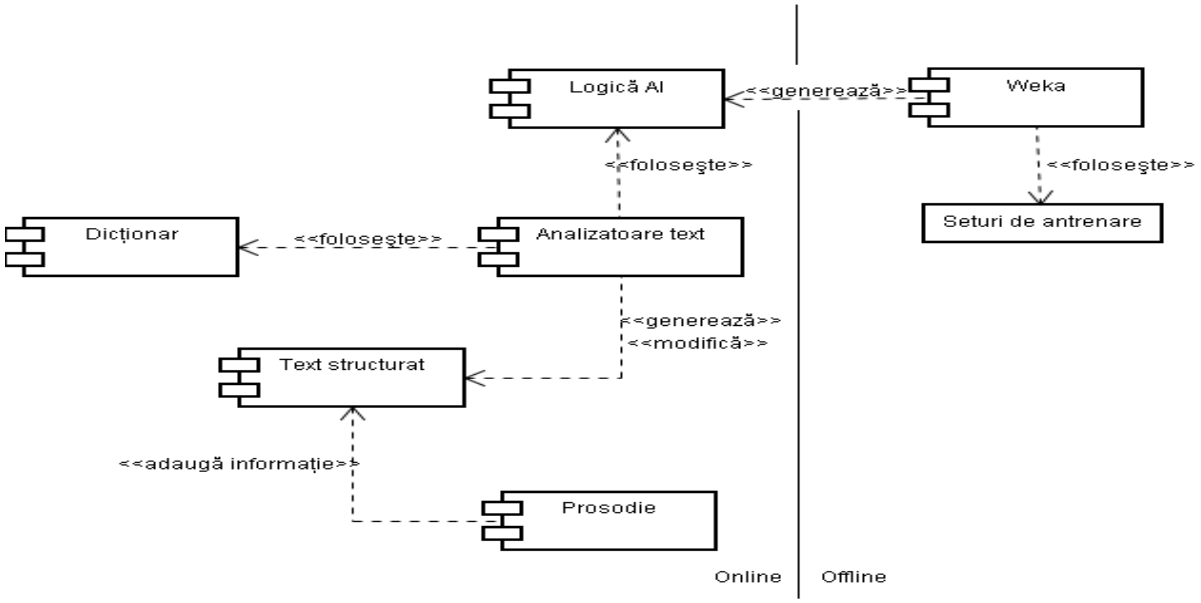

Figura 2. Componentele modulului de procesare de text

Tranzițiile între sunetele vorbirii sunt definite în domeniul articulator şi sunt exprimate prin intermediul unor relații neliniare. Când sunetele ce urmează să fie înlănțuite sunt suficient de "aproape" unul de altul, relațiile lineare dau totuşi rezultate bune [8]. Unitățile care au fost folosite în acest domeniu sunt: foneme care sunt independente de context, difoneme, foneme dependente de context, unități subfonemice, silabe, cuvinte, propoziții şi fraze. Selecția unităților se face folosind funcții de cost.

DESPĂRȚIREA ÎN SILABE

În cadrul acestei secțiuni se prezintă algoritmul de despărțire în silabe a cuvintelor din limba română. Algoritmul se aplică asupra textului de intrare după transcrierea fonetică şi transformarea literă-în-sunet. Metoda utilizată pentru construirea silabelor este împărțirea în funcție de modul de pronunție. Această metodă a fost aleasă în defavoarea metodei structurate din două motive. Primul motiv este faptul ca metoda structurată are nevoie de informații complexe referitoare la fiecare cuvânt, informații precum rădăcina cuvântului, prefix şi sufix. Alte informații necesare ar fi fost: dacă cuvântul este compus sau nu. Al doilea motiv este faptul că împărțirea textului de intrare cu metoda structurată determină crearea mai multor unități diferite decât metoda orientată pe pronunție.

Pentru despărțirea în silabe prin metoda orientată pe pronunție se construiesc în primul pas expresii comune pentru potrivirea structurilor cu fonemele consecutive. S-a identificat că în limba română (similar majorității limbilor) structurile utilizate pentru despărțirea în silabe sunt: hiați, diftongi şi triftongi. Un hiat este alcătuit din două vocale sonore în secvență. Diftongul este secvența alcătuită dintr-o semivocală şi o vocală sonoră sau o vocală sonoră urmată de o semivocală. Triftongul este secvența a trei sunete dintre care unul este vocală sonoră iar celelalte sunt semivocale. În limba română se întâlnesc două cazuri de triftongi. Primul caz este când două sunete semivocalice sunt urmate de o vocală sonoră, iar al doilea caz când un sunet vocalic este precedat şi urmat de câte o semivocală.

Prezentăm mai jos abstractizarea acestor structuri.

V = "a" | "ă" | "î" | "e" | "i" | "o" | "u"

S = "ĕ" | "ĭ" | "ŏ" | "ŭ"

C = "b" | "k" | "č" | "k'" | "d" | "f" | "g" | "g'" | "g" | "h" | "j" | "l" | "m" | "n" | "p" | "r" | "s" | "ş" | "t" | "ţ" | "v" | "z"

H = V, V

D = (S, V) | (V, S)

T = (S, S, V) | (S, V, S)

unde

"V" – sunetele vocalice

"C" – sunetele consonantice

"H" – regula de obținere a unui hiat

"D" – regulile de obținere a unui diftong

"T" – regulile de obținere a unui triftong

Pentru grupurile de consoane consecutive am identificat patru cazuri în care despărțirea în silabe este realizată diferit de regula generală. Două dintre excepții apar la întâlnirea unui grup alcătuit din patru consoane (FCE1 – prima excepție pentru grup patru consoane; FCE2 – a doua excepție pentru grup patru consoane). O excepție apare la grupul de trei consoane consecutive (THCE - excepție pentru trei consoane) şi ultimul caz pentru două consoane consecutive (TWCE - excepție pentru două consoane).

FCE1 = "stşk" | "ldsp" | "ngst" | "rnbl" | "nsfr" | "nsgr" | "nspl" | "rtkh" | "rtdr" | "rtsk" | "rtst" | "stpr" | "stsk"

FCE2 = "rst", C

THCE = "ltč" | "nsč" | "rgş" | "rtţ" | "mpţ" | "nkş" | "nkţ" | "ldm" | "lpn" | "ndb" | "ndk" | "nsb" | "nsk" | "nsd" | "nsf" | "nsh" | "nsl" | "nsm" | "nsn" | "nsp" | "nss" | "nsv" | "ntl" | "rtb" | "rtk" | "rth" | "rtj" | "rtm" | "rtp" | "rts" | "rtt" | "rtv" | "stb" | "stk" | "std" | "stf" | "stg" | "stl" | "stn" | "stp" | "sts" | "stt" | "stv" | "lpt" | "mpt" | "nkt" | "ndv" | "rkt" | "rtf" | "stm"

TWCE = ("b" | "k" | "d" | "f" | "q" | "h" | "p" | "t" | "v" | "g"), ("l" | "r")

unde

FCE1 – prima excepție patru consoane

FCE2 – a doua excepție patru consoane

THCE – excepție trei consoane

FWCE – excepție două consoane

În continuare prezentăm regulile generale de despărțire în silabe. Aceste reguli au fost generate utilizând informațiile din DOOM II [1].

Pentru grupurile de consoane există cinci cazuri, pornindu-se de la existența unei singure consoane până la existența a cinci consoane. Pentru grupurile de vocale şi semivocale consecutive există patru cazuri, pornindu-se de la o singură vocală până la patru vocale şi semivocale consecutive.

C, C, C, C, C $\rightarrow$ "CC-CCC"

FCE1 $\rightarrow$ "CCC-C"

FCE2 $\rightarrow$ "CC-CC"

C, C, C, C $\rightarrow$ "C-CCC"

THCE $\rightarrow$ "CC-C"

C, C, C $\rightarrow$ "C-CC"

TWCE $\rightarrow$ "-CC"

C, C $\rightarrow$ "C-C"

C $\rightarrow$ "-C"

S, V, S, V $\rightarrow$ "SV-SV"

V, T $\rightarrow$ "V-T"

T $\rightarrow$ nicio acțiune

V, D $\rightarrow$ "V-D"

D $\rightarrow$ no action

H $\rightarrow$ "V-V"

"[^{i}]" $\rightarrow$ "[^{i}]-" şi legătură

V $\rightarrow$ nicio acțiune

unde

" $\rightarrow$ " – acțiunea care va avea loc la depistarea secvenței respective după ce textul de intrare a fost transformat în foneme.

CONCLUZII

În cadrul acestui articol s-a descris arhitectura unui sistem de sinteză text-vorbire pentru limba română, componentele acestui sistem şi despărțirea în silabe.

Acest sistem TTS utilizează o abordare concatenativă a unităților de selecție după cum este recomandat în [2]. Textul de intrare este inițial procesat de componenta de analiză a textului care are rolul de a determina structura şi de a realiza normalizarea textului (expandarea numerelor, dezambiguarea, etc.). Modulul de selecție a unităților caută în dicționarul extins şi selectează unitățile corespunzătoare şi modul de pronunție. În cazul în care nu s-a gasit cuvântul în dicționar, are loc analiza fonetică care realizează transformare literă-în-sunet prin utilizarea utilitarului Weka [5]. Transcrierea fonetică este succedată de despărțirea în silabe prin care se determină unitățile de formare a sunetului. Aceste unități sunt selectate din baza de date acustică şi trimise modulului de procesare de sunet.

Articolul s-a axat pe componentele de procesare de text şi pe despărțirea în silabe.

REFERINȚE

1. Romanian Academy, Iorgu Iordan - Al. Rosetti Lingvistic Institute, Dicționarul Ortografic, Ortoepic şi Morfologic al Limbii Române , 2nd Edition. Univers Enciclopedic Bucureşti, 2005
2. Huang, X., Acero, A., Hsiao-Wuen, H.: Spoken Language Processing: A Guide to Theory, Algorithms, and System Development. Prentice Hall (2001)
3. M. A. Ordean, A. Saupe, L. Teodorescu, M. Boldizsar, M. Ordean, and G. C. Silaghi, "Efficient parsing of romanian language for text-to-speech purposes," in submitted to 12th Intl. Conf. on Text, Speech and Dialogue (TSD2009), Czech Republic
4. D. Burileanu, "Basic research and implementation decisions for a text-to-speech synthesis system in romanian," International Journal of Speech Technology, vol. 5, no. 3, pp. 211–225(15), 2002
5. A.Saupe, M.A. Ordean, L.R. Teodorescu, M. Ordean, and G. C. Silaghi, "Efficient Parsing of Romanian Language for Text-to-Speech Purposes", Proceedings of the 12th International Conference on Text, Speech and Dialogue, Pilsen, Czech Republic, 2009, to appear in LNAI, Springer Verlag
6. W. Daelemans and A. van den Bosch, "Language-independent data-oriented grapheme-to-phoneme conversion," in Progress in Speech Processing. Springer Verlag, 1997, pp. 77–89
7. Sproat, R.: Multilingual text analysis for text-to-speech synthesis. Natural Language Engineering 2(4) (1996)
8. Hunt, A.J., Black, A.W.: Unit selection in a concatenative speech synthesis system using a large speech database. In: ICASSP '96: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE Computer Society (1996)