Utilizarea tehnicilor nuantate (fuzzy) si de dinamica neliniara pentru sinteza adaptiva a vorbirii

Horia-Nicolai L. Teodorescu
Academia Româna, Sectia Stiinta si Tehnologia Informatiei,
Calea Victoriei 125, Bucuresti
E-mail: hteodor@etc.tuiasi.ro

1. Introducere

În timp ce masina realizeaza tipic transmisie de date, omul comunica. Diferenta consta în participarea intelectuala si afectiva a persoanei la actul comunicarii, participare reflectata atât la nivelul limbajelor neverbale (gestica, mimica etc.), cât si la nivelul vocal. Aceasta participare afectiva da varietate, coloratura si sensuri suplimentare, nu neaparat pe plan semantic, semnalului vocal. Sinteza vocii, în prezent, este limitata de lipsa afectului, varietatii si sensurilor suprapuse în planuri multiple. Vocea masinii ramâne astfel cantonata într-o regiune "moarta" a comunicarii, este monotona si obositoare pe termen lung.

În aceasta lucrare, reluând unele idei din [1-12], precum si în contextul unor dezvoltari recente [13-27], în special legate de e-Voice si VXML, prezentam si dezvoltam unele concepte si tehnici care ar putea permite masinii atingerea dezideratelor mai sus mentionate. Realizarea unor masini capabile sa mimeze calitatile vocii umane si sa *dialogheze* cu oamenii, sau macar sa comunice într-o maniera similara în care omul o face, este un deziderat în numeroase domenii, de la dialogul om-calculator, la sistemele auto si la sistemele de învatare asistata de calculator [13-15]. Rezolvarea acestei probleme are implicatii semnificative pentru acceptarea sintezei vocii într-o varietate de aplicatii, de la robotica la realitate virtuala, la industria de jocuri electronice si la protezare.

Prozodia, adica structura acustica ce se extinde pe mai multe segmente de semanl vocal, chiar peste mai multe cuvinte sau propozitii, implica ritm, accent, intonatie, timbru, afect si alte caracteristici ale vocii înca insuficient întelese sau vag definite în literatura. Informatia paralingvistica ce este continuta de prozodie nu este nicaieri regasita la nivelul "spus" prin cuvinte, dar – asa cum am subliniat în [2] – aceasta informatie poate fi chiar mai importanta pentru ascultator decât informatia lingvistica propriu-zisa. Incapacitatea sistemelor actuale de sinteza vocala de a reda prozodia naturala este evidentiata chiar de marii producatori de aplicatii [25] si este bine cunoscuta în mediul cercetatorilor în domeniul sintezei vorbirii: "*One of the most difficult problems in speech to date is prosodic modeling*" [25].

2. Solutii pentru sinteza adaptiva si varietala

Cele doua calitati ale vocii naturale, adaptivitatea – în sens larg – si variabilitatea se pot realiza, cu costuri nu neaparat mari, la nivelul sintetizoarelor actuale, cu adaptari minimale (sau deloc) la nivel hardware si cu îmbunatatiri ale programelor de control. Sinteza adaptiva se refera la adaptarea la:

- ? Conditiile sonore ambientale [1, 4];
- ? Contextul semantic-afectiv al cuvintelor si frazelor sintetizate [2, 3].
- ? Interlocutorul sistemului de sinteza automata, atunci când acesta este recunoscut [2].

Sinteza varietala se refera la modificarile inter-pronuntie, la repetarea unor fraze, chiar si în cazul în care conditiile ambientale si contextul (si interlocutorul) ramân neschimbate. Aceasta

variabilitate elimina monotonia si personalizeaza vocea (naturala sau sintetizata), în masura în care variabilitatea se face dupa reguli adesea proprii individului (cum este cazul în realitate) – si nu doar aleatoare.

Variabilitatea intrinseca a vorbirii deriva din mecanismele fizice de producere a semnalului vocal (curgere turbulenta a aerului prin organul fonator), precum si din mecanismele neurologice de control al producerii semnalului vocal (controlul neuronal este cunoscut ca având o dinamica cu o importanta componenta neliniara). Aceste caracteristici au fost documentate de mai multe grupuri de cercetare, inclusiv de noi si colaboratorii [5-9].

Adaptabilitatea si variabilitatea în sensurile de mai sus vor fi prezentate sumar în sectiunile urmatoare, sintetizând lucrarile citate si unele cercetari mai noi, nepublicate înca.

3. Adaptabilitate la mediu

Una dintre cele mai elementare adaptari ale semnalului vocal generat de om este cea de adaptare la conditiile de mediu. Adaptarea la un mediu real, cu fond de zgomot, se realizeaza pe patru cai principale: prin modificarea amplitudinii semnalului (mai mare în mediul de zgomot ridicat), prin modificarea spectrului (creste contributia frecventelor înalte), prin modificarea ritmului (scaderea ritmului, cresterea duratei vocalelor), si prin cresterea duratei dintre cuvinte, care devin separate, segmentate în timp. Adaptarile realizate – instinctiv de un vorbitor uman – se opereaza deci la un nivel relativ elementar, cu modificari de prozodie minimale.

Realizarea acestei adaptari este esentiala în multe aplicatii de sinteza a vocii, incluzând sinteza vocala pentru aplicatii în medii industriale si în mijloace de transport, sau sinteza vocala pentru proteze laringiene. Este remarcabil ca aceasta adaptare se poate realiza, la pretentii reduse, cu foarte putin hard suplimentar si/sau cu un soft minimal, aducând însa o îmbunatatire esentiala în utilizare. În privinta hardului, este necesar unul sau mai multe canale de culegere a semnalului de zgomot (semnal sonor ambiental).

Procesarea semnalului de zgomot, în vederea realizarii controlului sistemului de sinteza automata, presupune determinarea puterii zgomotului ambiental într-o fereastra temporala si determinarea componentei spectrale a semnalului ambiental. Primul parametru de caracterizare a zgomotului se obtine ca medie aritmetica a patratului semnalului *s*, într-o fereastra data, de largime de *W* esantioane si caracterizata de momentul actual de timp, *n*:

$$P(n) = \sum_{k=0}^{W} s_{n-k}^2 \qquad (1)$$

Caracterizarea spectrala se poate realiza sumar prin raportul *HL* dintre puterea la frecvente "înalte" (frecventele înalte corespunzând în mare benzii de frecventa ce include formantii nr. 2, 3, 4 si 5 din spectrul vocal) si puterea la frecventele "joase" (pâna la aproximativ al doilea formant, deci pâna la frecventa de cca. 400 –500 Hz, tinând cont si de vorbitorii feminini):

$$HL = \int_{0}^{500} S^2(f)\,df \Big/ \int_{500}^{10000} S^2(f)\,df \qquad (2)$$

Deoarece parametrii respectivi sunt relationati cu impactul pe care îl au asupra inteligibilitatii vorbirii, deci sunt dati de calitati subiective, este natural sa abordam o definire probabilista sau fuzzy a lor. Data fiind simplitatea controlului nuantat[1] (fuzzy), vom prefera a doua varianta. Un exemplu de

[1] Desi nu este larg acceptat si are o traducere mai dificila în alte limbi, vom utiliza aici termenul "nuantat", propus de Grigore C. Moisil, în locul englezescului "fuzzy".

definire[2] a functiilor de apartenenta respective este prezentat în Figura 1. Este de presupus ca aceasta definire sa constituie doar un punct de plecare, îmbunatatirea calitatii sintezei realizându-se si prin modificarea functiilor de apartenenta.

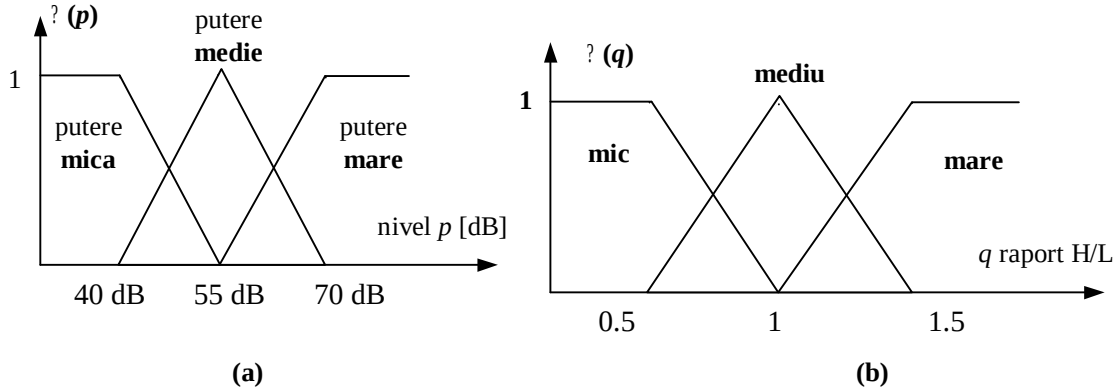

Figura 1. Functiile de apartenenta ale premiselor regulilor folosite pentru determinarea modificarilor parametrilor de control ai sintetizorului

Dupa cum s-a precizat deja, ca rezultat al aprecierii conditiilor de mediu, se controleaza patru parametri ai semnalului sintetizat:

- ? cresterea amplitudinii (parametru notat AI)
- ? cresterea continutului în frecvente înalte (HFCI)
- ? cresterea duratei vocalelor (VLI)
- ? cresterea duratei dintre cuvinte (accentuarea segmentarii pe cuvinte a frazei), notat IDBBW.

Controlul se realizeaza pe baza de reguli si poate fi rezumat în Tabelele 1-4 de mai jos.

Tabelul 1. Cresterea amplitudinii (AI – *Amplitude Increase*)

HL/P	mic	mediu	mare
mic	0,0	0,1	0,4
mediu	0,1	0,3	0,5
mare	0,4	0,5	0,6

Tabelul 2. Cresterea continutului de frecvente înalte (HFCI – *High Frequency Content Increase – F3 increase*)

HL/P	mic	mediu	mare
mic	0,0	0,1	0,4
mediu	0,1	0,3	0,5
mare	0,4	0,5	0,6

Tabelul 3. Cresterea duratei vocalelor (*Vowel Length Increase* – VLH)

HL/P	mic	mediu	Mare
mic	0,0	0,1	0,4
mediu	0,1	0,3	0,5
mare	0,4	0,5	0,6

[2] Pentru a nu încarca prezentarea, ecuatiile functiilor respective sunt date în Anexa 1.

Tabelul 4. Cresterea duratei dintre cuvinte
(*Increase of the Duration of the Break Between Words* – DBBW)

HL/P	Mic	mediu	mare
mic	0,1	0,1	0,.4
mediu	0,1	0,3	0,5
mare	0,4	0,5	0,6

Tabelele sunt interpretate în sensul uzual pentru sistemele nuantate. Preferam sistemele de tip Sugeno de ordin 0 (vezi Anexa 1), deoarece furnizeaza ca rezultat, direct, valori numerice, care vor fi interpretate ca si coeficienti de multiplicare ai valorilor nominale ale sintezei. De exemplu, prima linie si prima coloana din Tabelul 1 spun ca:

DACA *Puterea (zgomotului) este **medie** si parametrul LH este **mediu***

ATUNCI *Amplitudinea creste cu **0.3** ori.*

Toate regulile din Tabelul 1 si toate celelalte tabele se interpreteaza într-un mod similar.

Rezultatul final se obtine prin agregarea rezultatelor partiale, date de regulile respective. De exemplu, daca valoarea intensitatii sonore este de 45 dB, iar raportul HL este de 0,7, prin aplicarea fuzzificarii[3] se obtine gradul de adevar al premisei (combinate) din regula respectiva, prin

$$\min\left(\mu_{putere-mica}(P_0), \mu_{LH-mic}(LH_0)\right)$$

unde $P_0 = 45$ iar $LH_0 = 0{,}7$. Folosind expresiile functiilor (v. Anexa 1), se obtin valorile $\mu_{putere-mica}(P_0) = 0{,}67$, $\mu_{LH-mic}(LH_0) = 0{,}6$, deci valoarea minima este 0,6 si reprezinta gradul de încredere în faptul ca amplitudinea creste de 1,1 ori. Aceasta este valoarea de adevar pentru singletonul (de la iesirea sistemului) ce corespunde regulii respective, $\mu^{A}_{1,1}$. În total, sunt 9 reguli per tabel, deci exista 9 valori de singletoni. Într-adevar, în acelasi timp, valorile de intrare corespund functiilor de apartenenta „mediu" pentru „putere" si LH, deci regulii:

DACA *Puterea (zgomotului) este **mica** si parametrul LH este **mic***

ATUNCI *Amplitudinea creste cu **0.0** ori.*

cu gradul de încredere în rezultat:

$$\min\left(\mu_{putere-medie}(P_0), \mu_{LH-medie}(LH_0)\right)$$

precum si regulilor:

DACA *Puterea (zgomotului) este **mica** si parametrul LH este **mediu***

ATUNCI *Amplitudinea creste cu **0,1** ori.*

respectiv:

[3] Termenul echivalent românesc ar fi "nuantare"

DACA *Puterea (zgomotului) este **medie** si parametrul LH este **mic***

ATUNCI *Amplitudinea creste cu **0,1** ori.*

cu gradele de încredere

$$\min(\mu_{putere=mica}(P_0), \mu_{LH=medie}(LH_0))$$

si respectiv

$$\min(\mu_{putere=medie}(P_0), \mu_{LH=mic}(LH_0))$$

Celelalte cinci reguli din Tabelul 1 au gradele de încredere în rezultat nule, deoarece valorile functiilor de apartenenta „mare" ale premiselor („puterea este mare" si „LH este mare") sunt nule, pentru valorile date, $P_0 = 57$ si $LH_0 = 0{,}7$.

Prin agregare (defuzzificare), considerata aici conform formulei uzuale:

$$y = \frac{\sum_{k=1}^{9} \mu_k^A \cdot x_0^k}{\sum_{k=1}^{9} \mu_k^A(x_0)} \qquad (3)$$

se obtine valoarea de iesire (amplitudinea, cresterea continutului de frecvente înalte, cresterea lungimii vocalelor, respectiv cresterea duratei pauzei dintre cuvinte). În relatia de mai sus, μ_k^X reprezinta abscisele singletonilor de iesire din sistemele tip Sugeno respective, $\mu_k^X(h)$ reprezinta gradele de încredere în concluzia regulilor respective, iar *y* reprezinta valoarea agregata (defuzzificata) de iesire a sistemului Sugeno. Sumarea se face pentru toti singletonii de iesire (notati de la 1 la 9). Indicele "*A*" arata ca ne referim la parametrul controlat „amplitudine", controlul fiind desigur diferentiat pentru cei patru parametri discutati.

Valorile astfel obtinute sunt folosite, cum s-a precizat, ca factori de multiplicare ai parametrilor nominali[4]. De exemplu, daca amplitudinea nominala este A_0, atunci, prin aplicarea controlului, amplitudinea efectiva a semnalului va fi:

$$A = A_0 \cdot \left(1 + \frac{\sum_{k=1}^{9} \mu_k^A \cdot x_0^k}{\sum_{k=1}^{9} \mu_k^A(x_0)}\right) \qquad (4)$$

Sistemul de control este instantaneu, în sensul ca nu tine cont decât de valorile recente (din fereastra prezenta, de largime *W*) ale zgomotului, nu si de valorile anterioare. Controlul de

[4] Nominali, în sensul ca sunt valorile standard pentru sistemul de sinteza automata respectiv si pentru sunetul respectiv produs în conditiile contextuale date.

amplitudine si frecventa se poate exercita în afara sintetizorului propriu-zis, asupra unui amplificator si a unui filtru plasate la iesirea sintetizorului. Aceste doua controale se pot prevedea de altfel si în alte aplicatii, precum sisteme de sonorizare mari (eventual distribuite, ca în cazul sonorizarii unor spatii mari, gen piete sau stadioane), sau a unor sisteme de sonorizare locale (de exemplu, sisteme de interfonie). Controlul pauzelor dintre cuvinte si un control fin al spectrului vocalelor necesita comanda directa a sintetizorului.

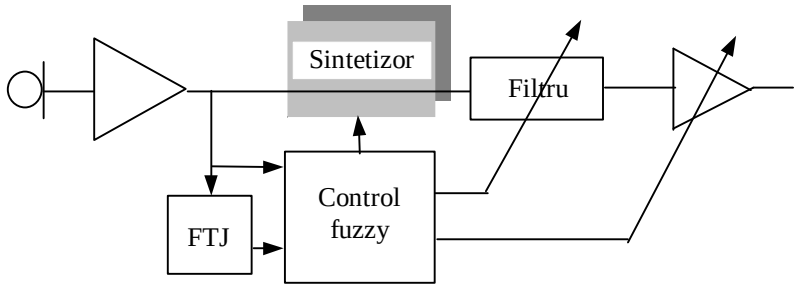

Figura 2. Schema bloc a unui sistem audio adaptiv la zgomotul ambiental

In cazul în care se utilizeaza doar primele doua tipuri de adaptare, în amplitudine si spectral, adaptarea se poate realiza si cu mijloace hardware externe sintetizorului, putând, de altfel, fi utilizata în orice aplicatie audio (de sonorizare etc.). Schema unui asemenea sistem de adaptare este cea prezentata în Figura 2, o varianta fiind initial propusa în [4].

4. Adaptare si variabilitate contextual-interpretativa

Interlocutorul uman raspunde cu afect, dupa cum considera anormala, nepotrivita, sau oricum în alt fel “departe de asteptari” întrebarea sau afirmatia facuta de partenerul la dialog. De asemenea, raspunsul este diferit atunci când vorbitorul uman este nesigur de raspuns, are un interes special în raspuns sau în topica discutiei, sau, din contra, este dezinteresat. In plus, situarea interlocutorului fata de partenerul sau partenerii de dialog, în context social sau afectiv, tonalizeaza discursul verbal si îi imprima specificitate relativa. Toate aceste caracterisitici participative, precum si altele asemenea, dau *comportamentul verbal* al omului, sunt traduse în mare masura la nivelul semnalului vocal prin prozodie, dar în prezent nu se regasesc la nivelul masinii. Privitor la elementele de baza privind prozodia, vezi [26].

Pentru a implementa un comportament verbal, masina trebuie sa dispuna de o baza de cunostinte minimala prin care sa genereze acest comportament. De exemplu, este necesar sa se interpreteze “departe de normal” într-o asertiune sau întrebare a interlocutorului uman. Deci, vom presupune ca exista o baza de cunostinte care permite o asemenea interpretare. Constructia acestei baze de cunostinte depinde de domeniul în care se poarta dialogul. În aceste conditii, accentul va fi mai puternic pe anumite parti ale frazei, sau raspunsul va depinde de asertiune sau întrebare. Modul de raspuns va fi dirijat de asemenea de o baza de cunostinte, care include regulile necesare modificarii sintezei (vezi Figura 3).

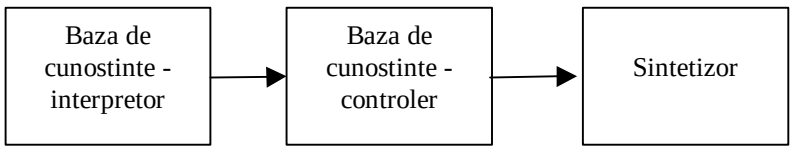

Figura 3. Schema de principiu a controlului contextual-interpretativ

Baza de cunostinte-controler poate de asemenea fi implementata cu reguli *Daca… Atunci*, de exemplu, de forma:

DACA oferta / raspunsul interlocutorului este neasteptat (negasit în baza de cunostinte – baza de asteptare/ baza de cazuri),

ATUNCI afectul sintezei este mirare / neîncredere/…/ etc.

ori

DACA oferta / raspunsul interlocutorului este neasteptat negativ (conform bazei de cunostinte),

ATUNCI afectul sintezei este mirare si/sau furie.

Folosind rezultatele regulilor de acest fel, se pot seta parametrii ierarhic inferiori, de tonalitate, ai vocii sintetizate, pe baza acestora generându-se parametrii efectivi de control ai sintezei (amplitudine, frecvente formanti etc.).

Desi acest gen de control poate parea complicat, sunt situatii destul de generale în care el se poate implementa cu un efort relativ redus. De exemplu, atunci când se determina (printr-o masuratoare relativ simpla, de frecventa medie în spectrul vocal, sau de fundamentala) ca interlocutorul este un copil sau o persoana de gen feminin, se poate selecta una sau ambele dintre alternativele:

? sistemul de sinteza automata se seteaza pe o voce de acelasi tip (copil/feminin)

? sistemul de sinteza automata se seteaza pe voce “calda” si “vorbire clara”.

Utilitatea si modalitatea de realizare a primei setari nu necesita explicatii. A doua setare (care poate fi simultana cu prima) se justifica – în cazul interlocutorului copil – prin necesitatea de a îi crea un mediu afectiv propice si linistit de dialog (voce “calda”) si prin necesitatea unei comunicari cât mai informative, usor de urmarit. Pentru a obtine o voce “calda”, se pot folosi trasee melodice cu variatii lente precum si frecvente mai joase ale formantilor si largimi mai mari (în zona spre frecvente joase) a spectrelor formantilor. “Claritatea” vocii se poate traduce prin segmentarea mai pronuntata pe cuvinte, precum si vocale mai lungi (cu sau fara accentuari ale spectrelor formantilor). Utilizarea unor asemenea adaptari – ce ramân în mare masura sa fie concepute în detaliu, implementate si testate – este neîndoielnic mare la sinteza pentru procese educative [15, 26], în aplicatii medicale (raspuns sintetic destinat pacientilor), precum si în numeroase aplicatii generale (de exemplu, sintetizoare utilizate în muzee, pentru prezentarea exponatelor).

Alte modalitati de personalizare afectiva sunt colorarea frecventiala si în amplitudine a anumitor parti din fraza sau în cadrul unui cuvânt, aceste modificari locale fiind larg documentate în literatura, de ex. [16-18] si fiind relativ usor de implementat.

5. Variabilitate prin metoda modularii de catre un sistem dinamic neliniar

Variabilitatea semnalului vocal uman este bine cunoscuta [5-9], [19-26]. Variabilitatea de tip natural a semnalului vocal sintetizat se poate obtine prin modularea diverselor controale (al amplitudinii, lungimii vocalelor, accentului, pitch-ului etc.) sau semnale lent variable, generate de sisteme care prezinta dinamica neliniara (haos). Parametrii sistemului haotic respectiv pot modela un anume subiect; consideram aici ca acesti parametri reprezinta individul vorbitor si “personalitatea” lui. Aceasta metoda, propusa de noi initial în 1992 ([28] s.a.), dar nepublicata în forma extinsa, credem ca reprezinta o metoda promitatoare de “personalizare” a vocii.

Consideram un sistem dinamic neliniar, dependent de parametri; semnalul în timp generat de acesta este de forma $x(t) = x(t, \lambda_1, \lambda_2, ..., \lambda_q)$, unde λ_h reprezinta parametrii sistemului haotic si permit modelarea specificitatii vorbitorului. Semnalul x poate fi folosit în modularea amplitudinii,

frecventei fundamentale, sau spectrului semnalului vocal sintetizat. De exemplu, spectrul poate fi modificat folosind o lege de variatie a frecventei centrale a formantilor de forma:

$$f_j(t) = (1 + x_j(t)) \cdot f_{j0}(t) \tag{5}$$

unde $f_j(t)$ este frecventa formantului numarul j la momentul t, $x_j(t)$ este semnalul haotic respectiv $(x_j(t) \in (-1,1))$, iar $f_{j0}(t)$ este frecventa "nominala" a formantului respectiv.

Un exemplu simplu de sistem haotic ce poate fi folosit în acest scop este dat de ecuatiile:

$$\begin{aligned} r_{n+1} &= \alpha_3 u_n^3 + \alpha_2 u_n^2 + \alpha_1 u_n + \alpha_0 \\ u_n &= \alpha_4 r_n + \alpha_5 \end{aligned} \tag{6}$$

unde setul de coeficienti $(\alpha_0, \alpha_1, ..., \alpha_5) \in \mathbf{R}^6$ se alege în domeniul de valori ce corespunde unui comportament haotic al sistemului (vezi Anexa 2). Setul de coeficienti $(\alpha_0, \alpha_1, ..., \alpha_5)$ se poate seta specific pentru fiecare sistem de sinteza automata, "personalizând" sistemul. Valorile de iesire ale generatorului se scaleaza corespunzator si se folosesc la modularea unuia dintre parametrii de sinteza. Pentru exemplul din sectiunea 3, amplitudinea semnalului sonor devine, prin utilizarea modulatiei haotice:

$$A_n = A_0 \left(1 + \frac{\sum_{k=1}^{9} A_k x_0}{\sum_{k=1}^{9} A_k x_0}\right) (1 + \beta r_n) \tag{7}$$

unde β este un coeficient de scalare a seriei de timp r_n. Coeficientul β se alege astfel încât contributia termenului $\beta \cdot r_n$ sa fie de ordinul procentelor ($\beta \cdot r_n = 0{,}1 \cdot n$).

Desigur, scara de timp a procesului de generare de esantioane de semnal vocal difera de scara de timp a proceselor haotice folosite în modulatie, ceasul celui de al doilea proces fiind mult mai lent (de ordinul 1/100) decât al primului proces. Pentru evitarea tranzitiilor bruste ale parametrului controlat, valorile generate pot fi interpolate si se poate realiza o variatie lenta între doua valori succesive. Considerând ca un esantion al seriei haotice r_n este generat la fiecare Q esantioane de semnal vocal, seria r_n se poate înlocui cu seria (mai "fina", dupa ceasul de generare a esantioanelor semnalului vocal):

$$r_k = r_{n+1} + \frac{r_n - r_{n+1}}{Q} k, \quad k = 0, 1, ..., Q \tag{8}$$

În scopul modularii haotice a mai multor parametri de sinteza (amplitudine, frecventa centrala a formantilor, largimea formantilor, elemente prozodice etc.), sunt necesare mai multe generatoare haotice, câte unul pentru fiecare parametru controlat. Alternativ, se poate folosi un sistem nuantat (fuzzy) haotic, aceste sisteme generând simultan un numar mare de iesiri necorelate sau slab corelate [28].

6. Concluzii si discutii

Adaptabilitatea si variabilitatea sistemelor de sinteza a vocii si ale celor audio, în general, se pot asigura prin modificari relativ simple hard si soft ale sistemelor actuale. Adaptabilitate se poate manifesta atât în raport cu mediul sonor, cât si în raport cu contextul sau cu interlocutorul. Ideea de adaptabilitate si metodele respective au fost introduse de noi în urma cu peste 20 de ani si dezvoltate continuu în lucrarile citate, atât pentru aplicatii de uz general, cât si pentru aplicatii medicale.

O aplicatie de interes medical-educational este utilizarea unor sisteme de învatare a unei limbi pentru copii de vârste mici (1 luna – 3 ani) care sufera de deficiente de auz. Utilizarea unor sintetizoare cu spectru si amplitudine controlate, astfel încât sa fie optim adaptate auzului (curbei de sensibilitate audiometrica) a fiecarui copil în parte ar ajuta asemenea copii sa învete limba la aceasta vârsta. Este, într-adevar, demonstrat ca învatarea primelor elemente ale unei limbi la aceste vârste asigura o sansa mult mai mare de învatare a limbii ulterior si de inserare sociala [24].

Lucrarea prezenta se situeaza într-un context mai larg, în cadrul cercetarilor realizate de diverse colective care cauta solutii pentru a face vocea sintetica purtatoare de informatie emotionala. Astfel, în [31] se descrie o metoda de sinteza a "vocii emotionale", capabila sa transmita trei emotii (suparare-furie, bucurie, tristete) folosind elemente de prozodie si segmente de tip vocala-consoana-vocala (specifice limbii japoneze). În [32], starea ("mood") si personalitatea sunt vazute ca elemente esentiale aparând în subsidiar în voce si necesar a fi introduse si în vocea sintetizata. Alti autori [33] vorbesc de "nivelul de placere al auditiei" (pleasantness) – dincolo de inteligibilitate – si vad naturaletea vocii sintetizate prin aceasta prisma, a utilizarii la nivel semnificativ, a prozodiei ("…we need to know more about how prosody could be utilized in human-computer interaction. We believe that we could borrow a lot from professional human speakers. Furthermore, speech applications should be built in a way that makes it possible to use prosodic features efficiently.").

Credem ca, în viitor, o metoda comoda de a genera automat prozodia, pentru o voce artificiala data si pentru o anumita stare, ar putea fi constituita de o procedura inversa celei descrise în [34].

Incheiem cu un citat din [35]: "… in spite of the long history of speech synthesis, no one speech synthesis system available today is able to produce speech that could be characterized as natural or completely pleasant. In order to improve the speech quality of current text-to-speech (TTS) systems in terms of naturalness, three areas must be addressed[5]: 1) improved linguistic analyses, 2) improved prosody modeling, and 3) improved speech synthesis models."

Multumiri. Aceasta lucrare a fost realizata cu sprijinul material al Academiei Române – Institutul de Informatica Teoretica Iasi – precum si cu sprijinul material partial al Societatii "Tehnici si Tehnologii" s.r.l. Iasi. Autorul multumeste colegilor Dragos Burileanu, Bogdan Branzila si Oana Geman pentru sugestii si corectii la o forma preliminara a lucrarii.

Referinte bibliografice

[1]. Teodorescu H.N., Chelaru M., Sofron E., Adascalitei A.: Adaptive speech synthesis. In vol. Digitale Sprach-verarbeitung - Prinzipien und Anwendungen. VDE Verlag, Berlin (W), pp. 183-188, 1988

[2]. Teodorescu H.N.: Interrelationship, Communication, Semiotics, and Artificial Consciousness. In: Kitamura, T. (Ed.): *What Should be Computed to Understand and Model Brain Functions?* FLSI Book Series, vol. 3, World Scientific, 2000

[3]. Teodorescu H.N.: Computer semiotics: understanding meanings and parallel languages (Refereed invited paper), Proc. Int. Conf. IIZUKA'98, Japan, 1998

[4]. Teodorescu H.N.: Making speech synthesisers noise-adaptabile. *Electronic Engineering* (UK), July 1987, p. 23

[5]. Rodriguez, W., Teodorescu H.N., Grigoras Fl., Kandel A., Bunke H.: A Fuzzy information space approach to speech signal nonlinear analysis. *J. of Intelligent Systems* (Wiley), Dec. 1999

[6]. Grigoras Fl., Teodorescu H.N., Apopei V.: Nonlinear Analysis and Synthesis of Speech. *Studies in Informatics and Control*, vol. 7, no. 1, March 1998, pp. 57-72

[5] Aici, autorul citat face referire la L. R. Rabiner, "Applications of Voice Processing to Telecommunications," *Proc. IEEE*, vol. 82, pp. 199–228, February 1994.

[7]. Teodorescu H.N., Grigoras Fl., Apopei V.: Nonlinear processes in speech production. *Int. J. Chaos Theory and Applications*, vol. 2, no. 2 (1997), pp. 35-52

[8]. Teodorescu H.N., Grigoras Fl.: Nonlinear Techniques in Speech Signal Analysis. Proc. International Conference on Intelligent Technologies in Human-Related Sciences, ITHURS'96. July 5-7, Leon, Spain. Vol. 2, pp. 293-298, 1996

[9]. Grigoras Fl., Teodorescu H.N., Apopei V.: Analysis of nonlinear and nonstationary processes in speech production, IEEE 1997 Workshop on Applications of Processing to Audio and Acoustics. Mohonk Mountain House New Paltz, New York, October 19-22, 1997 (IEEE Catalog # 97^{TH}8278)

[10]. Burlui V., Teodorescu H.N., Morarasu C.S.: La fonction phonatoire chez l'edente total. Analyse en frequence. *Les Cahiers de Prothese* (France), No. 88, Decembre 1994, pp. 63-68 1994

[11]. Teodorescu H.N. et al.: Fuzzy models in speech analysis and medical application, in Book of Summaries Int. Conf Modelling and Simulation, Istanbul, Turkey, July 1988, vol. 1, p. 162 (Summary)

[12]. Teodorescu H.N., L. Buchholtzer, Chelaru M., Teodorescu L.: A laryngeal prosthesis based on perilaryngean reflexes, Proc. 9th Int. EMBS Conf. IEEE, Boston. Vol. 4, IEEE, pp. 2114-2115, 1987

[13]. Anonymous Automotive Industry OEM/Supplier: Talking to computers vs. talking to humans 7/12/2000. http://www-nrd.nhtsa.dot.gov/departments/nrd-13/driver-distraction/Topics013040293.htm#A293

[14]. Anne-Marie Derouault, The Future of Speech Recognition. Evolving speech recognition technology is driving transparent computing, making it easier for people to interact with computers. http://www.advisor.com/Articles.nsf/ID/OA000107.DERO01

[15]. House D., Bell L., Gustafson K. & Johansson L. Child-directed speech synthesis: evaluation of prosodic variation for an educational computer program. Proc of Eurospeech'99, pp. 1843-1846, 1999

[16]. Heldner M., Strangert E. & Deschamps T.: Focus detection using overall intensity and high frequency emphasis. In: Andersson R, Abelin Å, Allwood J & Lindblad P, eds. Proc of Fonetik 99; pp. 73-76, 1999.

[17]. Heldner M., Strangert E. & Deschamps T.: A focus detector using overall intensity and high frequency emphasis. Proc of ICPhS-99, pp. 1491-1494, 1999.

[18]. Heldner M.: On the non-linear lengthening of focally accented Swedish words. In: W. van Dommelen & T Fretheim, eds. Nordic Prosody: Proc of the VIIIth Conference, Trondheim 2000 . Frankfurt am Main: Peter Lang. 2001

[19]. Karlsson I., Banziger T., Dankovicová J., Johnstone T., Lindberg J., Melin H., Nolan F. & Scherer K.: Within-speaker variability due to speaking manners. Mannell RH & Robert-Ribes J, eds. Proc of ICSLP98, 2379-2382. 1998

[20]. Karlsson I.: Within-speaker variability in the VeriVox database. In: Andersson R, Abelin Å, Allwood J & Lindblad P, eds. Proc. of Fonetik 99, pp. 93-96, 1999.

[21]. Karlsson I, Banziger T, Dankovicova J, Johnstone T, Lindberg J, Melin H, Nolan F, Scherer K (1998), Within speaker variation due to induced stress, Proc Fonetik-98, 150-153. www.ling.su.se/ fon/publications/fonetik98/

[22]. Gustafson-Capkova S & Megyesi B.: A Comparative Study of Pauses in Dialogues and Read Speech. Proc of Eurospeech 2001, pp. 931-935, 2001

[23]. Beskow J.: A tool for teaching and development of parametric speech synthesis. In: Branderud P & Traunmüller H (eds). Proc of Fonetik -98, pp. 162-165. 1-98, 1998

[24]. Rachel I. Mayberry, Elizabeth Lock, Hena Kazmi: Linguistic ability and early language exposure. *NATURE*, Vol. 417, 2 May 2002, p. 38, 2002

[25]. Mircrosoft Co.: Platform SDK: Agent. Characters. http://msdn.microsoft.com/library/default.asp?url=/library/en-us/msagent/deschar_8nn6.asp

[26]. Mauricio Lumbreras, Gustavo Rossi: Metaphor for the Visually Impaired: Browsing Information in a 3D Auditory Environment. CHI'95 Proc., www.acm.org/sigchi/chi95/proceedings/shortppr/ml_bdy.htm

[27]. Christophe d'Alessandro & Jean-Sylvain Liénard: 5.2 Synthetic Speech Generation. In: Survey of the State of the Art in Human Language Technology. http://cslu.cse.ogi.edu/HLTsurvey/ch5node4.html#SECTION52

[28]. Teodorescu H.N.: Chaos in fuzzy systems and signals. Vol. Proceedings of the 2nd Int. Conf. on Fuzzy Logic and Neural Networks. Vol. 1., pp. 21-50 (Jono Printing Co., 1992, Iizuka, Japan)

[29]. Teodorescu H.N., Kandel A., Jain L. C. (Eds.), Fuzzy and Neuro-Fuzzy Systems in Medicine (International Series on Computational Intelligence). CRC Press, Boca Raton, USA, 1998.

[30]. Teodorescu H.N., Mlynek D., Kandel A. (Eds.): Intelligent Systems and Interfaces (The Kluwer International Series In Intelligent Systems). Kluwer Publ., Boston, 2000.

[31]. Yasuhisa Niimi, Masanori Kasamatu, Takuya Nishimoto and Masahiro Araki: Synthesis of Emotional Speech Using Prosodically Balanced VCV Segments. http://www.ssw4.org/papers/133.pdf.

[32]. Nick Campbell: WHERE IS THE INFORMATION IN SPEECH? (and to what extent can it be modelled in synthesis?) www.slt.atr.co.jp/cocosda/jenolan/Proc/r82/r82.pdf.

[33]. Hakulinen J., Turunen, M.: Prosodic Features for Speech User Interfaces. www.cs.uta.fi/hci/spi/reports/Prosodic_Features_for_Speech_User_Interfaces.pdf.

[34]. Ansgar Rinscheid: Voice Conversion Based On Topological Feature Maps and Time-Variant Filtering. www.asel.udel.edu/icslp/cdrom/vol3/235/a235.pdf.

[35]. Syrdal A., Stylianou Y., Garrison L., Conkie A. Schroeter J.: Td-Psola Vs. Harmonic Plus Noise model in Diphone Based Speech Synthesis. www.research.att.com/projects/tts/papers/1998_ICASSP/paperSYN.ps.

Anexa 1: Sisteme nuantate de tip Sugeno, de ordin 1. Functii de apartenenta

Reamintim ca o multime (clasica) $A \subset X$, unde X noteaza universul de discurs, este definita de o functie caracteristica, de forma:

$$\mu_A : X \to \{0,1\}$$

$$\mu_A(x) = \begin{cases} 1 & daca \quad x \in A \\ 0 & daca \quad x \notin A \end{cases}$$

Prin generalizarea conceptelor de multime si de functie caracteristica, se definesc multimile nuantate (fuzzy) si functiile de apartenenta corespunzatoare astfel: o multime nuantata, notata $\tilde{A}$, peste universul de discurs X, este caracterizata unic de o functie de apartenenta:

$$\mu_A : X \to [0,1]$$

In particular, functia de apartenenta poate fi de forma:

$$\mu_A(x) = \begin{cases} 1 & pentru \quad x = a \in X \\ 0 & pentru \quad x \neq a \end{cases}$$

caz în care se numeste *singleton*.

Un sistem de tip Sugeno, de ordin 0, este descris de reguli de forma:

DACA intrarea (premisa) # 1 *SI premisa #* 2 *SI … Si premisa # n ATUNCI concluzia*

unde premisele sunt de forma: x_i *este* $\tilde{A}_{ij}$, iar $\tilde{A}_{ij}$ sunt valoari nuantate (fuzzy), de exemplu $\tilde{A}_{i1}$ = "mare", $\tilde{A}_{i2}$ = "mediu", atributelor lingvistice "mare", "mic" etc. fiindu-le atasate câte o functie de apartenenta. Specific sistemelor Sugeno este faptul ca în concluzie apar valori numerice si nu valori nuantate, concluzia fiind deci de forma "y = 0,3" (singleton).

Definitiile functiilor de apartenenta pentru intensitatea sonora din Figura 1.a sunt:

$$\mu_{Putere\,mica}(p) = \begin{cases} 1 & pentru \quad p \leq 40\ dB \\ 1 - \dfrac{p - 40}{15} & pentru \quad 40 \leq p \leq 55\ dB \\ 0 & pentru \quad p \geq 55\ dB \end{cases}$$

$$\mu_{Putere\,medie}(p)=\begin{cases}0 & \text{pentru } p \le 40\ dB\\ \dfrac{p-40}{15} & \text{pentru } 40 \le p \le 55\ dB\\ 1-\dfrac{p-55}{15} & \text{pentru } 55 \le p \le 70\ dB\\ 0 & \text{pentru } p \ge 70\ dB\end{cases}$$

$$\mu_{Putere\,mare}(p)=\begin{cases}0 & \text{pentru } p \le 55\ dB\\ \dfrac{p-55}{15} & \text{pentru } 55 \le p \le 70\ dB\\ 1 & \text{pentru } p \ge 70\ dB\end{cases}$$

Definitiile functiilor de apartenenta pentru raportul HL (Figura 1b) sunt:

$$\mu_{HL\,mica}(q)=\begin{cases}1 & \text{pentru } q \le 0.5\\ 1-\dfrac{q-0.5}{.5} & \text{pentru } 0.5 \le q \le 1.\\ 0 & \text{pentru } p \ge 1.\end{cases}$$

$$\mu_{HL\,medie}(q)=\begin{cases}0 & \text{pentru } q \le 0.5\\ \dfrac{q-0.5}{0.5} & \text{pentru } 0.5 \le q \le 1.0\\ 1-\dfrac{q-1.}{0.5} & \text{pentru } 1. \le q \le 1.5\\ 0 & \text{pentru } p \ge 1.5\end{cases}$$

$$\mu_{HL\,mare}(q)=\begin{cases}0 & \text{pentru } q \le 1.\\ \dfrac{q-1.}{0.5} & \text{pentru } 1.0 \le q \le 1.5\\ 1 & \text{pentru } q \ge 1.5\end{cases}$$

Pentru detalii asupra manipularii functiilor de apartenenta si a regulilor în sistemele nunatate, a se vedea orice manual în domeniul sistemelor fuzzy, sau volume precum [29, 30] în care se pot gasi si aplicatii specifice legate de întelegerea vorbirii, sau alte aplicatii medicale.

Anexa 2: Procesul haotic

Procesul reprezentat de ecuatiile (7) are o dinamica haotica doar pentru anumite subintervale relativ înguste din $\mathbf{R}^6$. În restul spatiului, comportamentul este asimptotic instabil (peste tot pentru valori ale coeficientilor lui r^3 mai mari ca 1, în modul, daca si coeficientul lui u este mai mare ca 1 în modul); comportamentul este stabil sau periodic pentru alte zone, relativ reduse din $\mathbf{R}^6$.

Diagrama de bifurcatie a procesului, asa cum apare în Figura A1, este obtinuta pentru: valorile coeficientilor [Q]={.1, -.17, -.18, .1}; coeff_4 = 1.1; coeff_5= -.15; conditie initiala r[0]= 0.3; numar total de puncte în diagrama de bufurcatie: 500 (punctele de la 500 la 1000); regimul tranzitoriu eliminat: primele 500 puncte; precizia tuturor coeficientilor si variabilelor: double.

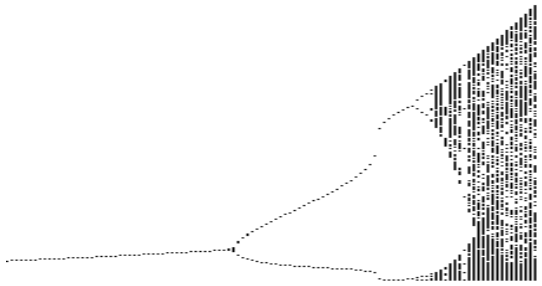

Figura A2-1. Diagrama de bifurcatie a procesului

Legile folosite (conform codului, scris în limbajul C) sunt:

u[n]= (coeff_4)*r[n]+coeff_5 -0.005*(float)k;

x=u[n]; r[n+1]=poly(x, Q, coeff);

(Q este numarul de valori în vectorul coeficientilor, Q=4)

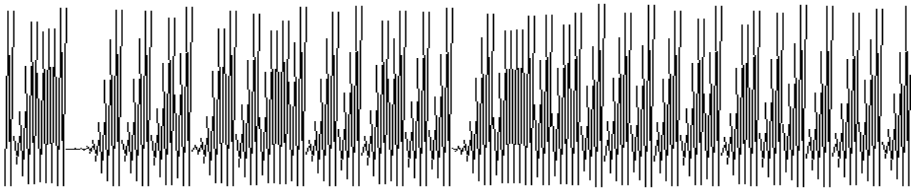

Figura A2-2

Semnalul în domeniul amplitudine-timp din Figura A2 a fost obtinut pentru ecuatiile (cod C):

u[n]= coeff_4*r[n]+coeff_5-0.05*21.;

x=u[n]; r[n+1]=poly(x, Q, coeff);

Semnalul obtinut pentru valoarea u[n]= coeff_4*r[n]+coeff_5-0.05*21. (restul programului fiind identic ca pentru cazul anterior) este ilustrat în Figura A3.

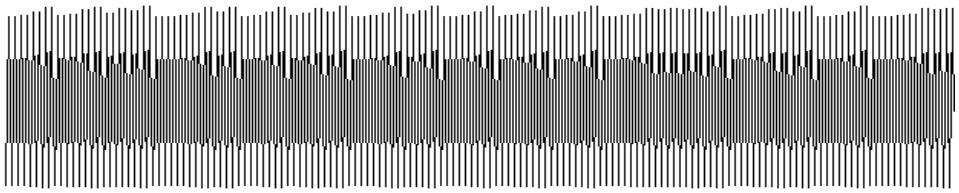

Figura A2-3

iar semnalul obtinut cu u[n]= coeff_4*r[n]+coeff_5-0.05*20.7, precum si la o scara dubla de timp, este ilustrat în Figura A4:

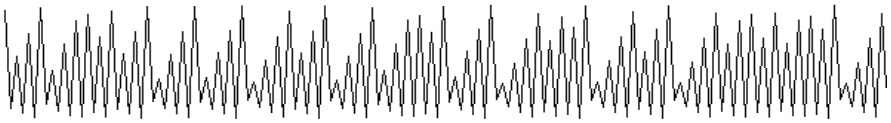

Figura A2-4

Regiunile spatiului parametrilor în care sistemul este stabil, dupa cum s-a spus deja, sunt relativ înguste. Pentru parametrii coeff_1-coeff_4 fixati si coeficientul coeff_5 variabil între –25.15 si + 4.85 (600 de pasi, cu pas 0,05), doar zona îngusta din Figura A2-5 este stabila, oscilanta sau

haotica, în rest sistemul fiind asimptotic instabil. Pentru usurinta urmaririi scarii, linia din partea de jos a figurii reprezinta intervalul mentionat, [–25.15, + 4.85]), în care s-a testat sistemul.

Figura A2-5

În figura, se poate remarca diagrama de bifurcatie a sistemului, cu zonele de stabilitate, oscilatie si haos. Pentru restul intervalului, prin program, calculele sunt abandonate, deoarece valorile de iesire ale sistemului depasesc, în valoare absoluta, 10000.