(19) OFICIUL DE STAT
PENTRU INVENŢII ŞI MĂRCI
Bucureşti

ROMÂNIA

(11) **RO 130883 A2**
(51) Int.Cl.
G10L 15/08 (2006.01),
G10L 15/04 (2006.01)

(12) # CERERE DE BREVET DE INVENŢIE

(21) Nr. cerere: **a 2014 00347**

(22) Data de depozit: **07/05/2014**

(41) Data publicării cererii:
29/01/2016 BOPI nr. **1/2016**

(71) Solicitant:
• **BUZO ANDI**,
BD. GENERAL VASILE MILEA NR. 8, BL. B2, SC. 3, ET. 4, AP.57, SECTOR 6, BUCUREŞTI, B, RO;
• **CUCU HORIA**, *ALEEA POLITEHNICII NR. 6, BL. 3, SC. 4, ET. 2, AP. 42, SECTOR 6, BUCUREŞTI, B, RO;*
• **PETRICĂ LUCIAN**, *INTRAREA VÂSLEI NR. 1, BL. PM63, SC. 2, ET. 5, AP. 77, SECTOR 3, BUCUREŞTI, B, RO;*
• **BURILEANU DRAGOŞ**,
STR. JOHANNES KEPLER NR. 2 BL. 2, SC. 2, AP. 61, SECTOR 2, BUCUREŞTI, B, RO

(72) Inventatori:
• **BUZO ANDI**,
BD. GENERAL VASILE MILEA NR. 8, BL. B2, SC. 3, ET. 4, AP.57, SECTOR 6, BUCUREŞTI, B, RO;
• **CUCU HORIA**, *ALEEA POLITEHNICII NR. 6, BL. 3, SC. 4, ET. 2, AP. 42, SECTOR 6, BUCUREŞTI, B, RO;*
• **PETRICĂ LUCIAN**, *INTRAREA VÂSLEI NR. 1, BL. PM63, SC. 2, ET. 5, AP. 77, SECTOR 3, BUCUREŞTI, B, RO;*
• **BURILEANU DRAGOŞ**,
STR. JOHANNES KEPLER NR. 2, BL. 2, SC. 2, AP. 61, SECTOR 2, BUCUREŞTI, B, RO

(74) Mandatar:
CABINET D.NICOLAESCU, *STR.TURDA, NR.102, BL.30A, ET.7, AP.28, BUCUREŞTI*

(54) # METODĂ ŞI SISTEM PENTRU DIARIZARE ÎN TIMP REAL A SEMNALELOR AUDIO, UTILIZATE PENTRU RECUNOAŞTEREA AUTOMATĂ A VORBIRII ŞI A VORBITORULUI

(57) **Rezumat:**

Invenţia se referă la o metodă şi la un sistem pentru diarizarea în timp real a semnalelor vocale şi este destinată a fi utilizată în domeniul sistemelor de procesare a semnalului audio pentru recunoaşterea automată a vorbirii şi identificarea automată a vorbitorului. Metoda conform invenţiei constă în citirea periodică a unui număr de vectori de caracteristici audio, segmentarea vectorilor, stocarea modelelor audio în istoria de diarizare, gestionarea periodică a istoriei şi folosirea unei funcţii de cost bazată pe vechimea modelului şi ponderea sa în istoria de diarizare. Sistemul conform invenţiei constă dintr-o memorie tampon (**303**, **304**) pentru vectori de caracteristici audio, o memorie (**201**) pentru segmente, o memorie (**202**) pentru grupuri de segmente, o memorie pentru modele (**203**), un serviciu (**102**) extern de segmentare şi automate de control (**301**), care gestionează istoria de diarizare.

Revendicări: 5
Figuri: 3

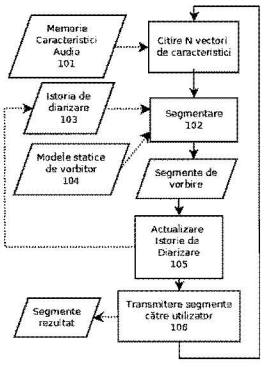

Fig. 1

RO 130883 A2

METODĂ ŞI SISTEM PENTRU DIARIZARE ÎN TIMP REAL A SEMNALELOR AUDIO, UTILIZATE PENTRU RECUNOAŞTEREA AUTOMATĂ A VORBIRII ŞI A VORBITORULUI

Invenţia aparţine domeniului sistemelor de procesare a semnalului audio pentru recunoaşterea automată a vorbirii şi identificare automată a vorbitorului.

Un sistem de recunoaştere a automată a vorbirii are ca scop transcrierea unui semnal audio, de cele mai multe ori înregistrarea unei conversaţii sau a unui monolog. Prin recunoaştere se obţine un fişier text care conţine cuvintele rostite. În cazul în care semnalul audio conţine segmente de muzică, zgomot, efecte speciale, sau în cazul în care proprietăţile semnalului audio se schimbă în timp, rezultatele procesului de recunoaştere sunt afectate în mod negativ, prin introducerea de erori. De exemplu, sistemul de recunoaştere va încerca transcrierea unui semnal muzical, rezultând text fără sens gramatical. Probleme similare sunt întalnite şi în cazul recunoaşterii vorbitorului, care nu este posibilă atunci când înregistrarea audio conţine muzică sau zgomot, sau când vorbirea înregistrată conţine semnal vocal de la mai mulţi vorbitori.

Diarizarea este procesul prin care semnalul audio este analizat şi segmentat, astfel încât fiecare segment are proprietăţi uniforme din punct de vedere audio şi conţine un singur fel de semnal, care poate fi linişte, muzică, vorbire sau alte tipuri de semnal audio. Recunoaşterea se poate face apoi doar pe segmentele de vorbire.

Segmentele de vorbire identificate pot fi suplimentar procesate în cadrul procesului de diarizare pentru separarea vorbitorilor, folosind proprietăţi audio cum ar fi frecvenţa fundamentală a semnalului vocal. În urma acestui pas, fiecare segment de vorbire aparţine unui singur vorbitor şi poate fi folosit pentru identificarea vorbitorului respectiv.

În cele ce urmează vom enunţa terminologia folosită în descrierea prezentei invenţii, cu menţiunea că se foloseşte terminologia consacrată specifică domeniului, în limba engleză:

- Frame – Fereastră audio – un număr de eşantioane ale semnalului audio digital, reprezentând în mod obişnuit un interval de timp fix, de ordinul milisecundelor (10-20ms)

- Speech Features – Vector de caracteristici audio - coeficienții cepstrali de frecvență (MFCC) şi a alte măsuri ale semnalului din o fereastră audio, folosite pentru procesul de diarizare. Semnalul audio este reprezentat în procesul de diarizare ca o înşiruire de vectori de caracteristici consecutivi.

- Segment – Segment – un număr de vectori de caracteristici audio consecutivi, care au proprietăți similare

- Cluster – Grup – un număr de segmente consecutive, care au proprietăți similare, de exemplu aparțin aceluiaşi vorbitor

- Speaker model – Model audio - un model de mixtură Gaussiană (GMM) care aproximează caracteristicile vorbitorului. Un model de vorbitor poate fi antrenat folosind vectorii de caracteristici audio ai unui grup. Un GMM poate de asemenea să aproximeze clase mai largi de semnal audio, cum ar fi muzica, linişte, vorbire, voce de bărbat, voce de femeie, şi altele.

Stadiul cunoscut al tehnicii, în ceea ce priveşte sistemele de diarizare automată, este analizat în [S. Galliano, G. Gravier, L. Chaubard, "The ester 2 evaluation campaign for the rich transcription of French radio broadcasts," In Proc. Interspeech, pp. 2583-2586, 2009]. Unul din sistemele cele mai performante este dezvoltat de laboratoarele LIUM [S. Meignier, T. Merlin, "LIUM SpkDiarization: An Open Source Toolkit For Diarization," in Proc. CMU SPUD Workshop, 2010.] Acesta procesează semnalul vocal în următoarele etape:

1. Extragerea vectorilor de caracteristici audio din semnalul audio
2. Segmentare
3. Agregarea segmentelor în grupuri
4. Antrenarea modelelor audio
5. Re-segmentarea Viterbi, folosind toate modelele identificate
6. Identificarea segmentelor care conțin vorbire
7. Identificarea bărbat/femeie, folosind modele pre-antrenate
8. Agregarea finală în grupuri a segmentelor care aparțin aceluiaşi vorbitor.

Aşa cum se prezintă sistemele de diarizare cunoscute, dezavantajele lor sunt inabilitatea de a segmenta semnalul audio pe măsură ce acesta este primit de către

sistemul de diarizare. Sistemele cunoscute de diarizare nu pot determina caracteristicile segmentului (vorbitor, zgomot de fundal, etc) dacă nu au la dispoziție toate eşantioanele segmentului, astfel încât să poată antrena modelele audio necesare. Mai mult, sistemul de diarizare nu poate determina graniţa între două segmente dacă nu are la dispoziție toate eşantioanele pentru ambele segmente. În cazul ideal, diarizarea se face pe tot semnalul vocal, eliminând problemele enunțate, dar această metodă introduce intarzieri mari, inacceptabile pentru unele aplicatii.

De exemplu, rezultatele diarizării sunt utile procesului de recunoaştere a vorbirii prin filtrarea vorbire/linişte sau informația despre vorbitor (pentru adaptarea modelului acustic), şi este de dorit ca recunoaşterea să aibă loc abia după terminarea diarizării. În multe aplicații, de exemplu cele care implică fluxuri audio foarte lungi, sau care necesită răspuns în timp real al sistemului de recunoaşterea vorbirii, diarizarea pe tot semnalul vocal nu este o soluție viabilă şi este necesară o soluție de diarizare în timp real, care să segmenteze semnalul audio pe măsură ce acesta este primit.

Invenția se referă la o metodă de diarizare în timp real a semnalelor vocale, care se realizează folosind, atât modele statice pre-antrenate de vorbitor, cât şi o istorie de segmente şi modele de vorbitor create dinamic, istoria fiind gestionată periodic prin actualizarea modelelor de vorbitor şi prin eliminarea modelelor, folosind o funcție de cost bazată pe vechimea modelului şi ponderea sa în istoria de diarizare, atunci când istoria de diarizare depăşeşte o dimensiune prestabilită. Invenția se mai referă şi la un sistem de diarizare în timp real, pentru implementarea metodei, care constă din module funcționale ce pot fi implementate ca programe software executabile pe un calculator sau circuite integrate, care mențin istoria de diarizare şi comunică cu servicii externe de extragere a vectorilor de caracteristici audio, recunoaştere automată a vorbirii, actualizare a modelelor GMM.

Se prezintă în continuare, în detaliu, principiile şi realizarea invenției, în legătură şi cu figiurile de la 1 la 3, care reprezintă:

Fig.1 prezintă metoda de diarizare propusă, ce presupune folosirea unei istorii de diarizare, conținând modele dinamice de vorbitor, şi modele statice de vorbitor, pentru segmentarea unui flux de caracteristici de vorbire.

Fig. 2 prezintă metoda de gestiune a istoriei de diarizare, prin care segmente noi sunt adăugate şi, atunci când istoria depăşeşte o dimensiune dată, sunt

eliminate segmente şi modele de vorbitor.

Fig. 3 prezintă un sistem de diarizare construit pe baza metodei propuse, constând din memorii pentru istoria de diarizare, şi componente pentru gestiunea acestei istorii.

Invenţia se referă la o metodă pentru separarea unui semnal audio în segmente omogene din punctul de vedere al proprietăţilor audio (diarizare), incluzând separarea segmentelor de vorbire de segmentele audio de linişte şi, separarea segmentelor de vorbire în funcţie de vorbitor. Metoda propusă are la bază faptul că segmentarea se face fără a aştepta primirea întregului fişier audio, iar segmentele rezultate dintr-o anumită porţiune a fluxului audio sunt calculate şi livrate utilizatorului în timp real, cu o întârziere fixă, relativ la fluxul audio.

Metoda propusă pentru diarizare în timp real este prezentată în Figura 1. Metoda se bazează pe citirea, la fiecare T secunde, a unui număr N de vectori de caracteristici audio dintr-o memorie tampon 101. Vectorii sunt segmentaţi folosind un serviciu extern de segmentare 102. Modelele audio folosite pentru segmentare sunt stocate în istoria de diarizare 103 sau în memoria statică de modele pre-calculate 104. Istoria 103 acoperă ultimele S secunde de semnal audio, şi conţine, atât segmentele, cât şi grupurile de segmente împreună cu modelele audio asociate grupurilor. Memoria statică 104 de modele pre-calculate conţine modele pre-calculate pentru vorbitori consideraţi a fi importanţi, a căror recunoaştere este esenţială (persoane publice cunoscute, , VIP)

Segmentele rezultate din diarizarea de la pasul curent sunt adăugate la istoria de diarizare 103 şi modelele vechi sunt eliminate din istoria 103 conform unei metode 105 de gestiune a istoriei de diarizare. Actualizarea modelelor de vorbitor din istoria 103, folosind informaţia audio de la pasul curent, este realizată de un serviciu extern. Vectorii audio corespunzători tuturor segmentelor, în afară de ultimul, sunt apoi transmise către utilizator. Vectorii audio din ultimul segment sunt păstraţi şi folosiţi ca parte a următorului set de N vectori de caracteristici audio ce vor fi procesaţi.

Metoda propusă pentru gestiunea istoriei de diarizare este ilustrată în Figura 2. Istoria de diarizare este compusă din trei memorii distincte:

- memoria pentru segmente (MS) 201, ce conține caracteristicile de vorbire corespunzătoare segmentelor identificate anterior prin diarizare,
- memoria pentru grupuri de segmente (MG) 202, ce conține grupurile identificate anterior prin diarizare, şi
- memoria pentru modele GMM (MM) 203, ce conține modelele de mixtură gaussiană calculate pentru fiecare grup de segmente în parte.

În urma diarizării unei ferestre audio, segmentele sunt adăugate istoriei de diarizare 103. Se încearcă asocierea fiecărui segment cu unul din modelele de vorbitor existente, la pasul 204. Dacă asocierea există, se actualizează grupurile de segmente la pasul 205 şi se actualizează modelele de vorbitor ale grupurilor respective, la pasul 206. Dacă asocierea nu există, se crează un grup nou la care segmentul este adăugat, şi se generează modelul de vorbitor pentru grupul nou creat. Atât noul grup, cât şi modelul său de vorbitor, sunt adăugate istoriei de diarizare.

Dacă istoria de diarizare depăşeşte o dimensiune D prestabilită de către utilizator, se execută o procedură de curățare a acesteia:

- Se verifică dacă există grupuri în care toate segmentele au vechime mai mare de S secunde, unde S este o valoare specificată de utilizator. Aceste grupuri sunt eliminate din istoria de diarizare la pasul 207.

Dacă pasul anterior nu a dus la scăderea dimensiunii istoriei de diarizare sub dimensiunea D, se calculează o funcție de cost la pasul 208 pentru fiecare grup/model, în felul următor:

- Se calculează o valoare de cost CV a vechimii modelului, proporțională cu numărul de secunde de la ultima actualizare a modelului
- Se calculează o valoare de cost CD a dimensiunii grupului asociat modelului, invers proporțională cu numărul de segmente conținute de grupul respectiv
- Se calculează o valoare de cost total CT prin mediere ponderată a CV şi CD, cu ponderi alese de utilizator

Modelele sunt ordonate în funcție de valoarea de cost total CT asociată fiecăruia, în lista 209. În mod repetitiv, modelul cu valoarea cea mai mare de cost este eliminat din istoria de diarizare, împreună cu segmentele şi grupul asociate modelului, până când dimensiunea istoriei de diarizare scade sub pragul D.

Sistemul propus pentru diarizare în timp real a unui flux audio este ilustrat în Figura 3. Istoria de diarizare este conținută în memorii RAM. Gestiunea istoriei de diarizare, incluzând scrierea şi citirea memoriilor 201, 202, şi 203 care formează istoria de diarizare, este realizată de un automat finit de control 301 ce poate fi implementat ca un circuit sau ca un program executat pe un microcontroler. Sistemul include o memorie RAM suplimentara pentru vectori audio, 302, care conține vectori de caracteristici audio citiți dintr-o memorie tampon 303, atât timp cât este necesar pentru diarizare, vectori care apoi sunt scrişi în memoria tampon 304. Întregul proces este controlat de un automat finit de control ce poate fi implementat ca un circuit sau ca un program executat pe un microcontroler sau un microprocesor.

Invenția prezentată are multiple avantaje față de stadiul tehnicii:

- Metoda propusă, folosind istoria de diarizare, permite execuția în timp real a procesului de diarizare, prin menținerea unui număr relativ mic de modele de vorbitor şi actualizarea acestora pe măsură ce rezultatele diarizării sunt produse. Efortul computațional pentru recunoaşterea vorbitorului prezent în fiecare segment de vorbire este proporțional cu numărul de modele de vorbitor avute în vedere, prin urmare reducerea numărului de modele reduce efortul computațional.

- Metoda propusă ocupă o cantitate mai mică de resurse de memorie, prin mecanismul de gestiune care elimină modelele de vorbitor, împreună cu segmentele asociate, folosind o funcție de cost ce ține cont de vechimea modelului şi ponderea sa în istoria de diarizare. Datorită faptului că necesarul de memorie pentru metoda de diarizare propusă este fix, indiferent de lungimea fluxului audio diarizat şi numărul de vorbitori din acest flux, iar dimensiunea efectivă poate fi setată arbitrar de mic, metoda propusă se pretează în special sistemelor încorporate şi implementărilor folosind resurse limitate.

Revendicări

1. Metodă de diarizare în timp real a semnalelor vocale prin identificarea şi marcarea de segmente din fluxul audio vocal ce aparţin aceluiaşi vorbitor sau aceleiaşi clase audio, metoda fiind **caracterizată prin aceea că** diarizarea este realizată folosind, atât modele statice pre-antrenate de vorbitor, cât şi o istorie de segmente şi modele de vorbitor create dinamic, istoria fiind gestionată periodic atunci când istoria de diarizare depăşeşte o dimensiune prestabilită, menţinându-se astfel un necesar fix de memorie indiferent de lungimea fluxului audio şi de numărul de vorbitori.

2. Metodă de diarizare conform revendicării 1, în care se realizează gestionarea istoriei de diarizare, ce conţine modele statistice caracteristice pentru vorbitorii identificaţi, modele ce pot fi create sau actualizate în procesul de diarizare, metoda fiind **caracterizată prin aceea că**, atunci când istoria de diarizare depăşeşte o dimensiune prestabilită D, sunt eliminate toate modelele statistice pentru vorbitorii care nu au mai apărut în fluxul audio de cel putin S secunde.

3. Metodă de diarizare conform revendicării 1, **caracterizată prin aceea că**, atunci când istoria de diarizare depăşeşte o dimensiune prestabilită D, se calculează o funcţie de cost CT pentru modelele statistice dinamice de vorbitor şi se elimină iterativ câte un model statistic din istoria de diarizare, alegându-se de fiecare dată modelul care are costul cel mai mare, până când dimensiunea istoriei de diarizare scade sub pragul D.

4. Metodă de diarizare conform revendicării 3, folosită pentru eliminarea selectivă a modelelor statistice de vorbitor din istoria de diarizare, metoda fiind **caracterizată prin aceea că** funcţia de cost CT reprezintă o medie ponderată între costul CV dat de vechimea modelului în istoria de diarizare, unde costul CV este direct proporţional cu vechimea modelului, şi costul CD dat de ponderea segmentelor asociate modelului în istoria de diarizare, unde costul CD este invers proporţional cu această pondere.

5. Sistem de diarizare în timp real ce implementează metoda conform revendicării 1, **caracterizat prin aceea că** sistemul constă din memorii şi automate finite de control, ce pot fi implementate ca programe software executabile pe un calculator sau circuite integrate, care mențin istoria de diarizare şi comunică cu servicii externe de extragere a vectorilor de caracteristici audio, recunoaştere automată a vorbirii, actualizare a modelelor GMM şi diarizare preliminară.

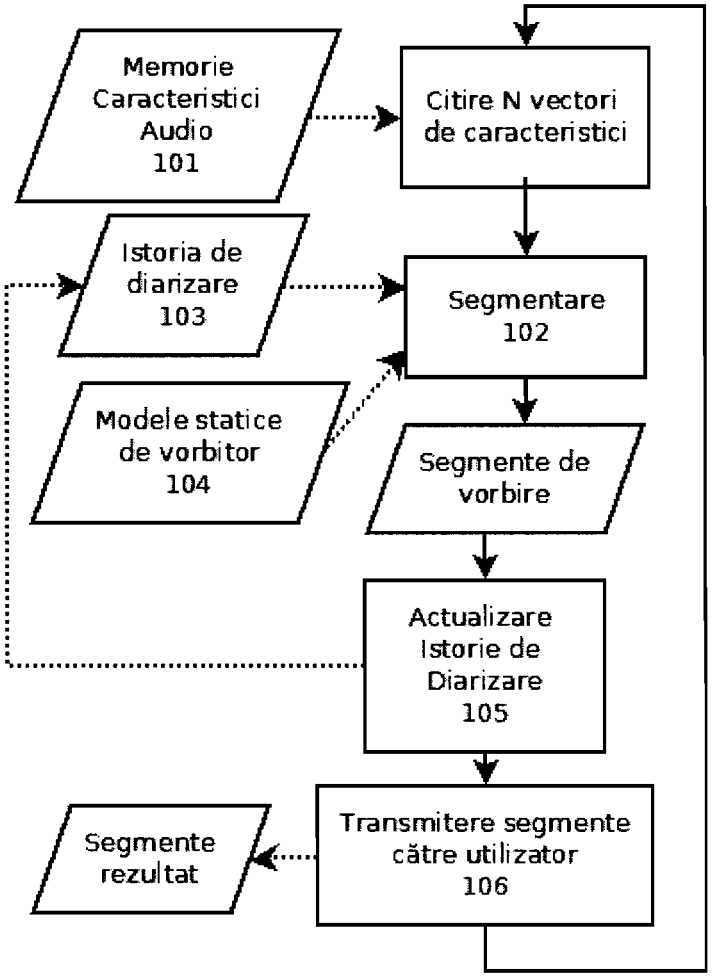

Figura 1: Metodă de diarizare

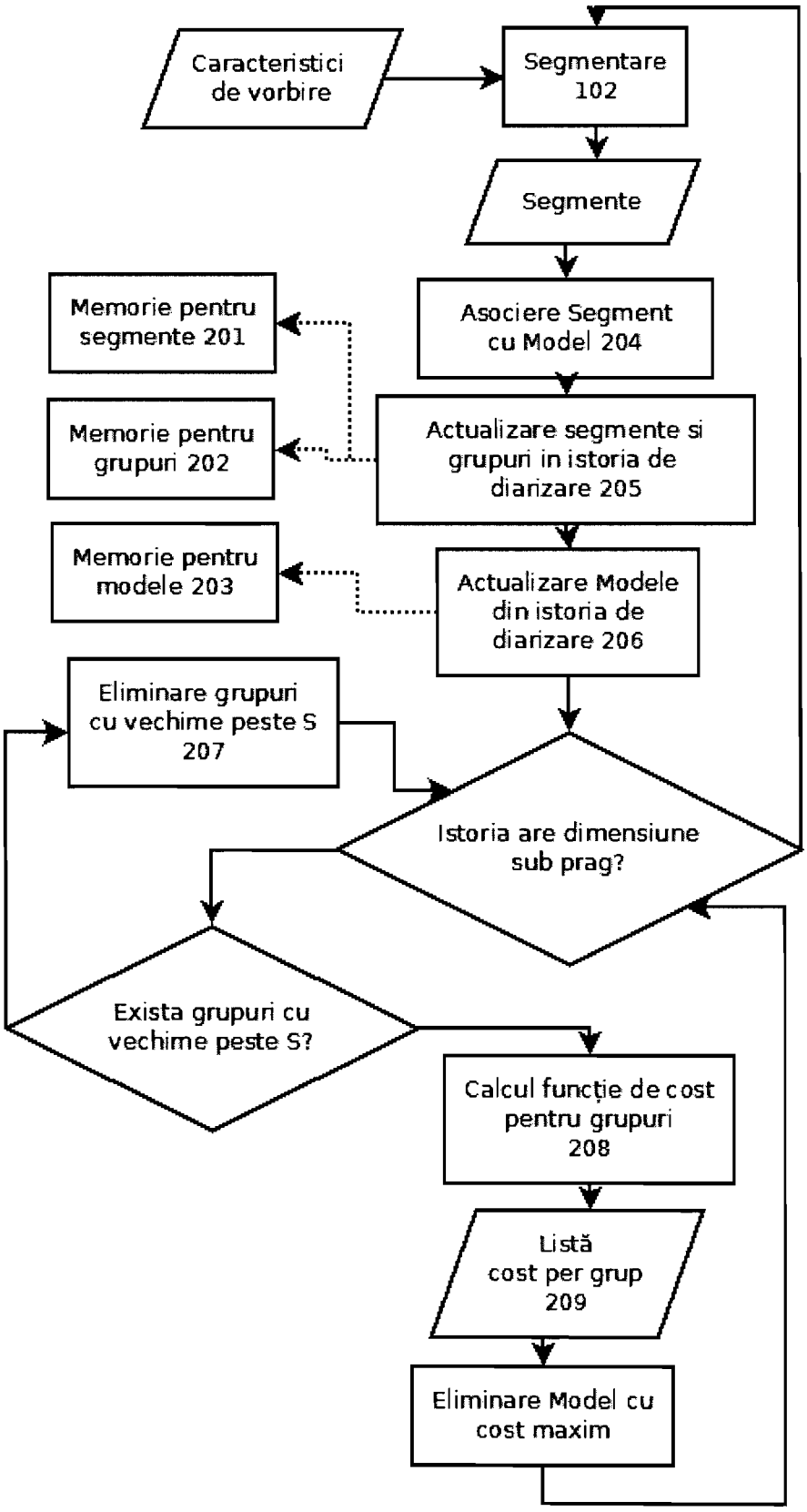

Figura 2: Gestiunea istoriei de diarizare

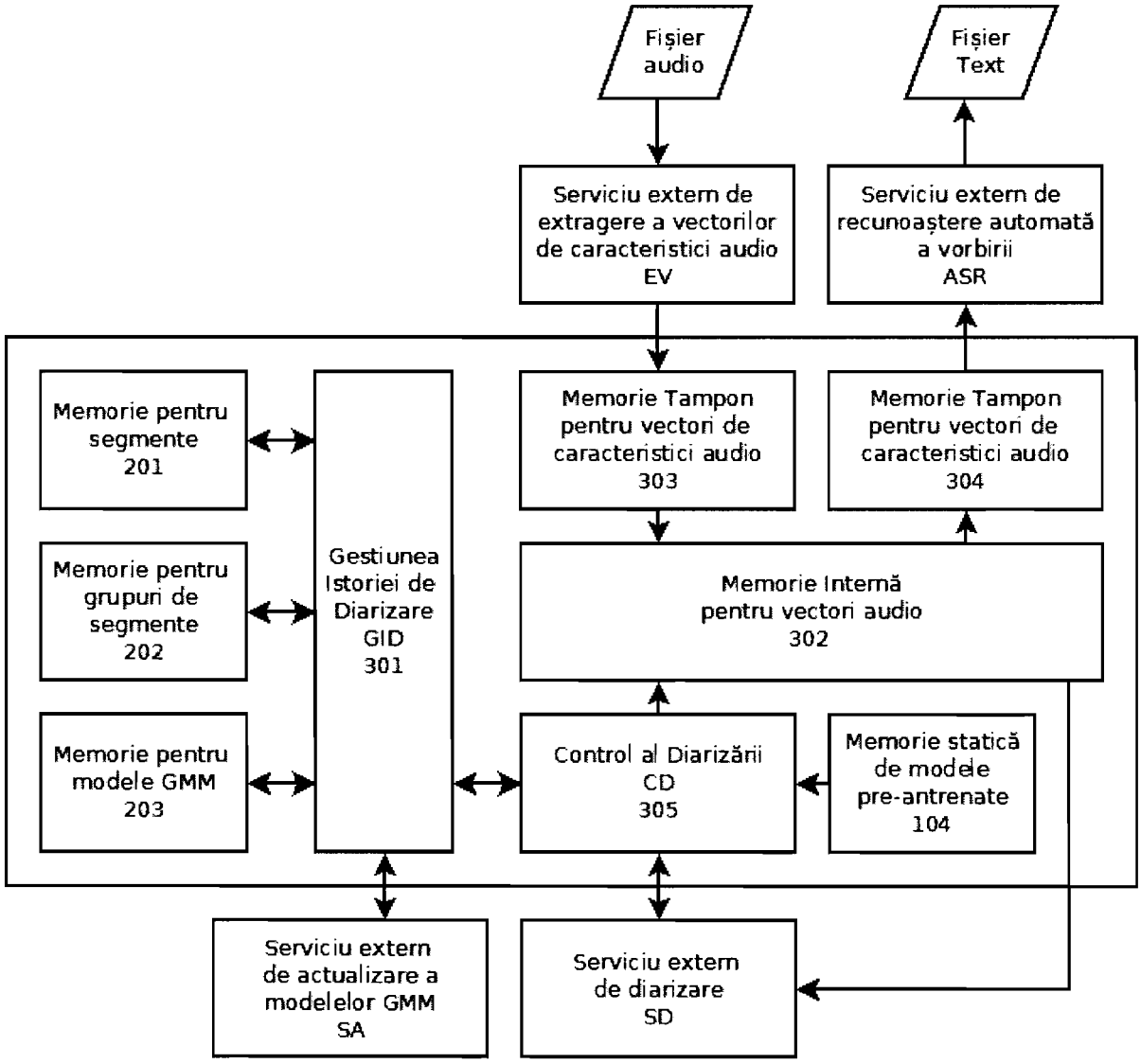

Figura 3: Sistem de diarizare în timp real